Utilizing User-Generated Content to Analyze Tours and Activities in Bangkok: A TripAdvisor Case Study

Miss Naina Chugh

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Industrial Engineering
Department of Industrial Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2019

การนำเนื้อหาที่ผู้ใช้สร้างขึ้นเองมาวิเคราะห์ทัวร์และกิจกรรมในกรุงเทพมหานคร กรณีศึกษาจาก TripAdvisor

น.ส.นัยนา ชุก

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมอุตสาหการ ภาควิชาวิศวกรรมอุตสาหการ
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

Thesis Title	Utilizing User-Generated Content to Analyze Tours and Activities in Bangkok: A TripAdvisor Case Study
By	Miss Naina Chugh
Field of Study	Industrial Engineering
Thesis Advisor	Associate Professor NARAGAIN PHUMCHUSRI, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Engineering

.... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

.... Chairman
(Associate Professor WIPAWEE THARMMAPHORNPHILAS, Ph.D.)

.... Thesis Advisor
(Associate Professor NARAGAIN PHUMCHUSRI, Ph.D.)

.... Examiner
(Nantachai Kantanantha, Ph.D.)

.... External Examiner
(Assistant Professor Manop Reodecha, Ph.D.)

นัยนา ชุก : การนำเนื้อหาที่ผู้ใช้สร้างขึ้นเองมาวิเคราะห์ทัวร์และกิจกรรมในกรุงเทพมหานคร กรณีศึกษาจาก TripAdvisor. (Utilizing User-Generated Content to Analyze Tours and Activities in Bangkok: A TripAdvisor Case Study) อ.ที่ปรึกษาหลัก : รศ. ดร.นระเกณฑ์ พุ่มชูศรี

วิทยานิพนธ์ฉบับนี้ตั้งอยู่บนจุดประสงค์ที่จะทำความเข้าใจความต้องการของนักท่องเที่ยวและวัดผลความพึงพอใจของนักท่องเที่ยว การเดินทางและอุตสาหกรรมการท่องเที่ยวนั้นเป็นเสมือนกระดูกสันหลังของเศรษฐกิจโลกซึ่งนับวันยิ่งมีการแข่งขันเพิ่มมากขึ้น ข้อมูลเชิงลึกที่เกี่ยวข้องจึงยิ่งมีความสำคัญเพิ่มขึ้นอย่างมีนัยสำคัญ ชุดข้อมูลที่มีผลกระทบและควรค่าแก่นำมาการวิเคราะห์อย่างมีระบบในยุคดิจิทัลคือเนื้อหาที่ผู้ใช้สร้างขึ้นเองในโซเชียลมีเดีย ดังนั้นผู้จัดทำวิทยานิพนธ์จึงนำเนื้อหาที่ผู้ใช้สร้างขึ้นเองตรงส่วนของการวิจารณ์ (รีวิว) ออนไลน์เกี่ยวกับทัวร์และกิจกรรมทางการท่องเที่ยวในเว็บไซต์ TripAdvisor มาวิเคราะห์เพื่อให้ได้มาซึ่งข้อมูลเชิงลึกที่กล่าวไปข้างต้น กระบวนการศึกษาและวิจัยเริ่มตั้งแต่การวิเคราะห์ในหลากหลายรูปแบบ เช่น การวิเคราะห์ความรู้สึก (sentiment analysis) เพื่อรวบรวมมุมมองที่หลากหลาย การหากฎความสัมพันธ์ (association rules mining) เพื่อหารูปแบบของความต้องการ และการประมวลผลภาษาตามธรรมชาติ (natural language processing) ร่วมกับการวิเคราะห์ความถี่ในการใช้อักษร (text frequency analysis) เพื่อบอกว่านักท่องเที่ยวพูดถึงประเด็นอะไรบ่อยที่สุด ยิ่งไปกว่านั้น กระบวนการวิจัยยังครอบคลุมไปถึงการทำโมเดลทำนายผลลัพธ์ผ่านการเรียนรู้ของเครื่อง (machine learning prediction model) โดยนำการอัลกอริทึม 3 รูปแบบมาใช้ ได้แก่ การจำแนกแบบถดถอยโลจิสติกส์ (logistic regression), แบบเครื่องเวกเตอร์ค้ำยัน (support vector machine), และแบบการสุ่มป่าไม้ (random forest) เพื่อคาดคะเนพฤติกรรมการวิจารณ์ที่จะนำไปสู่การให้5ดาวหรือ1ดาวในรีวิว และระบุว่าอะไรคือปัจจัยที่ส่งผลต่อความรู้พึงพอใจในทัวร์และกิจกรรมการท่องเที่ยวทั้งในแง่บวกและแง่ลบ

สาขาวิชา	วิศวกรรมอุตสาหการ	ลายมือชื่อนิสิต ....
ปีการศึกษา	2562	ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

6170199521 : MAJOR INDUSTRIAL ENGINEERING
KEYWORD: TripAdvisor User-generated content Sentiment analysis Association rules mining Natural language processing Text frequency analysis Prediction models Classification Machine learning K-fold cross validation Logistic regression Support vector machine Random forest

Naina Chugh : Utilizing User-Generated Content to Analyze Tours and Activities in Bangkok: A TripAdvisor Case Study. Advisor: Assoc. Prof. NARAGAIN PHUMCHUSRI, Ph.D.

The overarching goal of this paper is to gain visibility on tourist preferences and whether or not the needs of tourists are being met. With the Travel and Tourism (T&T) sector being the backbone to the global economy and the sector becoming more saturated and competitive, insights on T&T are vital now, more than ever. The rise of social media and user-generated content has effectuated the opportunity for a systematic analysis of tourist preferences via user-generated content. This paper is focused on gaining insights into tourism in Bangkok, Thailand through user-generated content scraped from TripAdvisor's online reviews of tours and activities. In order to develop insights on tourist preferences and tourism trends in Bangkok, various analyses were implemented, including sentiment analysis to gather tourist point-of-view, association rules mining to find patterns of preferences, and natural language processing along with text frequency analysis to understand what features tourists are most frequently talking about. This paper also developed machine learning prediction models using Logistic Regression, Support Vector Machine, and Random Forest algorithms to forecast 5-start ratings and 1-star ratings of reviews – with the purpose of identifying factors that significantly affect positive and negative sentiments on tours/activities.

Field of Study:	Industrial Engineering	Student's Signature ....
Academic Year:	2019	Advisor's Signature ....

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my advisor and professor, Aj. Naragain, who has been by my side throughout the process of writing my thesis and pushed me to learn new concepts and methodologies. Without her continuous support, advice, and guidance, this thesis would not have been brought to fruition. I'd also like to extend my appreciation to all my M.Eng. professors who have taken the time to share their valuable knowledge and shape my perception of engineering.

I am also extremely grateful to be part of the SCE 2018 batch. We were able to form a tight-knit community, supporting each other through all the late nights, early mornings, study sessions, workload frustrations, trips abroad, and all the ups and downs that come with a post-graduate degree. So glad to have had you all there!

Last but definitely not least, I'd like to show my gratitude, appreciation, and love to my family. Mom, Dad, Kris, and Sohum -- thank you for all the constant support, guidance, and kindness you've shown me through all the phases of writing my thesis. I love you all.

Naina Chugh

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1: Introduction

1.1 The Tourism Industry

1.1.1 World Tourism Trends

Travel and Tourism, one of the world's largest economic sectors, manages to successfully generate wealth and prosperity to its various industries and businesses across the globe. According to WTTC's[1] 2019 Economic Impact report, the Travel and Tourism sector accounted for 10.4% of global GDP and 10% of total employment (or 319 million jobs) in 2018 [1].

Starting with a mere 25 million international tourist arrivals in 1950 (estimated by UNWTO[2]), 2019 saw an astounding 60X increase, at 1.5 *billion* recorded international tourist arrivals [2]. The rise in tourism that year was a 4% growth from the previous year and the 10th *consecutive* year of growth in the industry. This surge in international tourist arrivals stems from the retention and acceleration of travel from current consumers as well as the enablement of travel from new demographics. A strong global economy, a growing middle class, technological advancements, affordable travel cost, and enhanced visa facilitation are just some of the factors driving the proliferation of the industry [3].

The "Leisure Travel" domain has been prevailing over all other purpose-of-travel's, growing from 50% in 2000 to 56% in 2018. This is further reinforced by WTTC's 2019 Economic Impact report stating "the division of overall spend is firmly weighted towards the leisure market, which represents 78.5% of the total". Other purpose-of-travels in 2018 included VFR (visiting friends and relatives), Health, and Religion (27%); Business and Professional (13%); and other Non-Specified (4%) [3].

A major contributor to the growth in Travel and Tourism (T&T) is none other than the "Land of Smiles", Thailand. The country placed 10th in UNWTO's Top 10 Global Destinations list in 2017 [4]. Thailand also placed 14th in WTTC's Top 15 Contributors to GDP, in terms of T&T. Furthermore, Thailand is one of the few countries that grew at a higher rate than global T&T GDP (Thailand T&T: +6% vs. Global T&T: +3.9%) [1].

1.1.2 Thailand Tourism Trends

Thailand is one of the most developed tourism markets in Asia. The country is globally known for its exceptional hospitality, enriched historical sites, central Southeast-Asia location, world-famous cuisine, good infrastructure, and affordable accommodations [5]. The country's tourism revenue reached a high 62 Billion USD in December 2019, compared to 58 Billion USD the year before.

When compared to its neighboring countries, Thailand dominates in terms of Travel and Tourism. The industry is Thailand's major economic sectors, accounting for 16.6% of

---

[1] World Travel and Tourism Council
[2] United Nations World Travel Organization

Thailand's GDP as of 2015. This greatly exceeded other countries in the region and the global average of 9.8% [1]. Tourism in Thailand has continued to grow since, now accounting for 20% of GDP (2019) and is projected to reach up to 30% by 2030[3] [6].

1.1.2.1 Thailand's Tourism Vision

According to [7], Thailand's tourism vision is very clearly stated as follows: "By 2036, Thailand will be the world's leading quality destination, through balanced development while leveraging Thainess to contribute significantly to the country's socio-economic development and wealth distribution inclusively and sustainably." Furthermore, the TAT[4] released an Action Plan for 2019 which prioritizes marketing towards "Foodie Tourism"—showcasing the country as an outstanding food destination; "Brand Value"—establishing awareness of Thai society, religion, history, and culture; "Tackling Waste"—creating awareness to CSR and waste-disposal activities; and "Travel Routes"—encouraging travelers to move from primary to secondary cities [8].

1.1.2.2 Thailand's Tourism Outlook

The 5-year outlook for Thailand's tourism industry, according to Thailand Tourism Q2 2020 report, is "bright with steady gains" [9]. The key drivers towards this favorable outlook include expansion of low-cost flight networks, the growing disposable income in emerging and established markets, and Thailand's positive tourism reputation. Moreover, strong government backing and promotional efforts towards making Thailand a "tourist hub" greatly strengthens the country's tourism outlook.

1.2 Tours and Activities in Bangkok

Every holiday in Thailand is incomplete without a visit to "the city of angels", Krung Thep a.k.a. Bangkok. With its groundbreaking 21.98 million international visitors in 2018, the city had become the top international destination for the *fourth* year in a row [10]. Bangkok is highly attractive to tourists due to its centralized location, its convenient transportation, and its extensive offering of experiences – with everything from city-life to temples and palaces, food tours to night life scenes, and workshops to day-trips and activities.

According to TripAdvisor's Experiential Travel Trends of 2019, global tourism-related bookings haven been trending towards more experiential and immersive holidays. Some of the fastest growing types of global experiences are Family-Friendly (+204%), Classes and Workshops (+90%), Wellness Experiences (+69%), and Cultural and Themed Experiences (+65%) [11]. Keeping in line with these global trends and TAT's 2019 Action Plan, tourism in Bangkok has also been gearing up towards authentic local experiences, life enrichment, and customization[3].

---

[3] According to the Thosaporn Sirisamphand, secretary-general of the Office of the National Economic and Social Development Council (NESDC)
[4] Tourism Authority of Thailand

1.3 Problem Statement

The Travel and Tourism (T&T) sector is the backbone to Thailand's economy. From the TAT releasing public statements on expecting 3 trillion Baht in tourism revenue in 2020, to the government launching stimulus measures aimed at prompting more travel [12], initiatives related to T&T are supported by all major players in the nation.

Thailand, however, is not the only country relying on tourism for economic growth. Macau, Singapore, Greece, Japan, and Turkey are just some of the countries that have been greatly investing in their tourism sector [13]. Thailand stands to face stiff tourism competition from up-and-coming destinations, particularly in the Asia-Pacific region. Some of the Asia-Pacific cities with the fastest growing number of tourists (2009-2016) are Osaka, Chengdu, Colombo, Tokyo, Taipei, and Xi'an [13].

As more destinations establish and promote tourist activities, the market is getting more saturated and competition is massively rising. In such a highly competitive sector with so much national focus, insights on T&T is vital now more than ever. Currently, there is not much visibility on tourist preferences and whether or not their needs are being met. With the rise of social media and user-generated content, we have a very effective indicator of such preferences at our disposal. At present, however, **there is a gap in the systematic analysis of tourist preferences via user-generated content**. All tourism stakeholders—whether it be the TAT, DMOs[5], NTOs[6], or tour companies—require such insights and knowledge in order to make informed data-driven decisions, customize tour/activity offerings, transcend competitors, and anticipate future trends.

1.4 Objectives

The objective of this thesis is twofold:

1. Develop insights on tourist preferences and tourism trends in Bangkok by gathering online reviews and implementing various analyses: sentiment analysis, association rules mining, natural language processing, and text frequency analysis
2. Develop machine learning prediction models that can forecast 5-star and 1-star rating of reviews in order to identify factors that significantly affect positive and negative views on Bangkok tours/activities

1.5 Scope

1. This thesis focuses on the geographic location of Bangkok, Thailand and on tours/activities within the following categories: (1) Activities, (2) Bike Tour, (3) Cooking Class, (4) Food Tour, (5) Sight Seeing, and (6) Spa – *i.e. see **Figure 2** and **Figure 3***
2. The data (user-generated content) used for all analyses in this thesis is from the 59,758 online reviews scraped from TripAdvisor and Viator (subsidiary of TripAdvisor)

---

[5] Destination Marketing Organization
[6] National Tourist Organization

3. The online reviews used for this research covers reviews post from January 2010 – January 2020
4. In order to carry out Objective 1 and learn about tourist preferences and trends, the following analyses were conducted:
    a. Insights on **tourist preferences** were derived from the proportion (in percent) of reviews from different categories (Tour/Activity and Origin) – *i.e. see **Figure 18***
    b. Insights on **tourist trends** were similarly derived from the proportion (in percent) of reviews from different categories over time – *i.e. see **Section 4.1.2***
    c. Insights on **tourist sentiment** were derived from the sentiment analyses, using ‘sentiment score’ calculated from a lexicon of ‘positive’ and ‘negative’ words – *i.e. see **Section 3.1.4***
    d. Insights on **tourist association** were derived from the Association Rules Mining, where “association combinations” were taken from the occurrence of a single, unique reviewer leaving multiple tour/activity reviews (both within the same Activity/Tour category as well as across categories) – *i.e. see*
    e. Insights on **tourist focus** were derived from the Natural Language Processing, where word frequency was counted on all words *minus* a lexicon of “stop words” – *i.e. see **Section 3.1.5***
5. In order to carry out Objective 2 and learn about feature significance, 12 prediction models were built with the following features *(see **Section 3.2.1** for more detail)*:
    a. Purpose: models 1-6 predicted whether a review was given 5-stars or not for each of the 6 tour/activity categories; models 7-12 predict whether a review was given a 1-star rating or not for each of the 6 tour/activity categories
    b. Dependent Variables: Y = Discrete variable (1/0) of whether a review has a 5-star rating or not (models 1-6) or whether a review has a 1-star rating or not (models 7-12)
    c. Independent Variables: X1 = Sentiment Score, X2 – X11 Discrete 1/0 Origin variables, X12 – X21 Discrete 1/0 “Frequent Word” variables
    d. Prediction Models: implemented the following machine learning algorithms for predictions – Logistic Regression, Support Vector Machines, and Random Forest (*see **Section 3.2** for more detail*)
6. The model’s prediction metrics (F1-score, accuracy, recall, precision, and specificity) was used to measure the model’s effectiveness, with this thesis focusing highly on accuracy and F1-score.

1.6 Thesis Benefits

Through insights and knowledge gained from this thesis, stakeholders in the tourism industry (such as DMOs, NTOs, and tour operators) can gain the following benefits:

- ***Accurate Targeting.*** Often times, consumers are bunched together into a single group – leading to across-the-board campaigns that do not achieve any effective results. Using consumer preferences to segregate customers into market segments and then targeting campaigns tailored to each segment's interests is certain to yield a higher conversion rate.

- ***Personalized Customer Service.*** A personal and interactive connection with brands is greatly valued by consumers. Businesses that understand such are able to effectively communicate offers and information that spark a personal interest with customers. With doing so, they can greatly benefit from an increased consumer experience and thus, prolonged customer retention.
- ***National Expansion:*** Insights learned from the study could be used for large national-scale tourism projects. Government organizations, such as the TAT, could leverage newly gained knowledge on tourist behavior to revamp their action plans of making and maintaining Thailand as a tourism hub.

1.7 Research Timeline

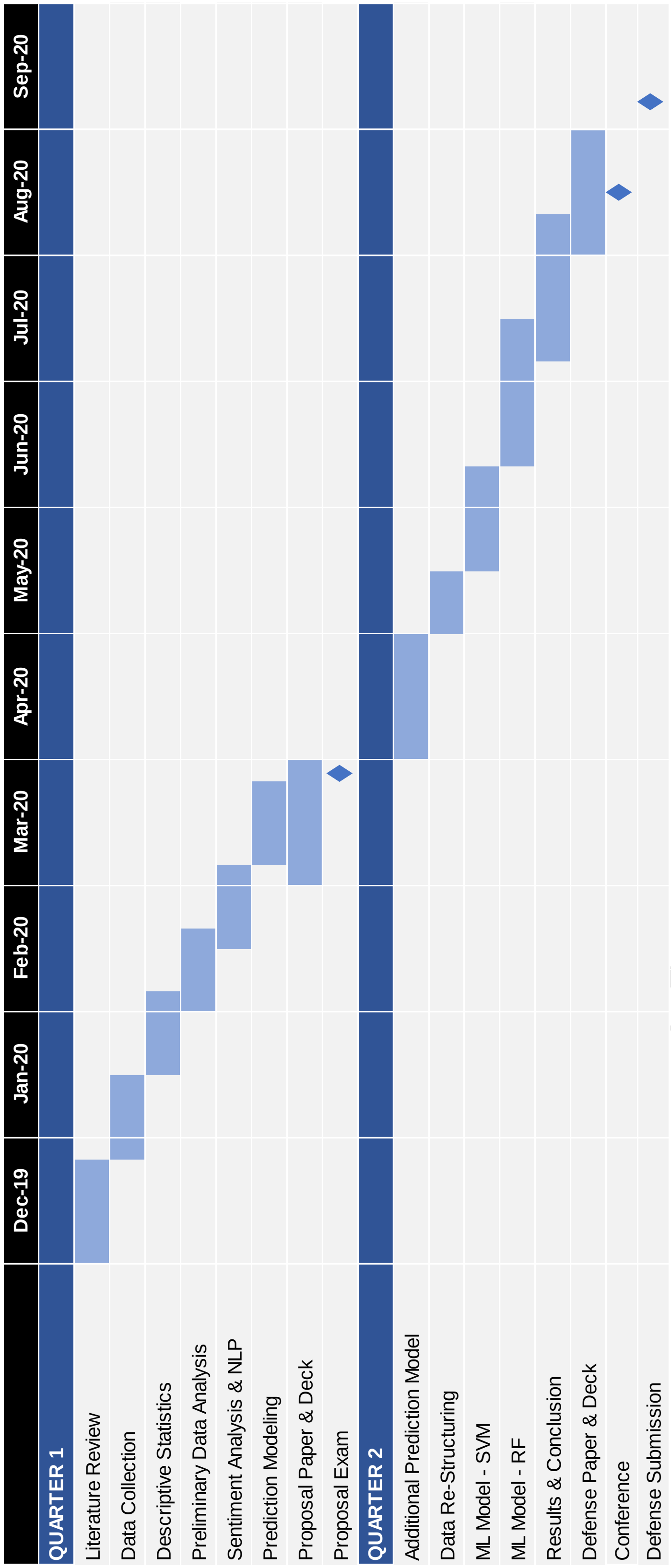


Figure 1: *Research Timeline*

Chapter 2: Literature Review

2.1 Learning about Tourist Preferences

2.1.1 Importance of Tourist Preferences

According to Lancaster's new theory of consumer demand, customer preferences about a product are fundamentally related to its features, or aspects. He further elaborated that consumer behavior is a process of choosing bundles of features of goods and services rather than the goods and services themselves [14]. Consequently, identifying such distinctive attributes and associating how customers feel about them would contribute to an improved understanding of consumer preferences.

Reasons behind learning about consumer preferences have to do with so much more than just reacting to what customers want. It is also about being forward-thinking and *anticipating* the customer's needs and *acting* before *reacting*. Knowing what your customers want and what features they find attractive allows for firms to tailor their products and services, thus, increasing their chances of conversion.

Studies on consumer preferences have been continuously conducted throughout history. A major turning-point in learning about consumer preferences was the availability of the Internet, specifically, the emergence of Web 2.0 and Big-Data (further discussed in ***Section 2.1.2.1***). Utilizing these technologies allowed for researchers to step away from relying on questionnaires and polls done on small sample sizes and move towards conducting advanced studies on massive scales.

2.1.2 Learning Consumer Preferences using Big-Data

According to Domo Inc's 6th edition report, "Over 2.5 *quintillion*[7] bytes of data are created every single day, and it's only going to grow from there. By 2020, it's estimated that 1.7MB of data will be created every second for every person on earth." [15] Furthermore, IDC[8] also stated that currently, as of 2020, there are around 40 trillion gigabytes of data available [16]. That's the size of ~3 x 10^{20} tweets! This huge bulk of data or "Big-Data" is quite important and highly insightful if handled properly.

2.1.2.1 Big-Data, Big Benefits

Big-Data (BD), just like its name suggests, refers to a large, diverse set of information that grows at an ever-increasing rate [17]. Originally coined in the late '90s in computer science literature, Big-Data was initially used as a mere scientific visualization tool [18]. The concept was properly defined in 2001 by Doug Laney, who further identified the three major characteristics of BD as the 3Vs. BD's 3Vs includes the *volume* (amount) of data, the *velocity* (speed) at which it is collected, and the *variety* of the information [17]. Over the recent years, two more V's have developed: the *value* of data, which refers to the ability to transform data

[7] Quintillion = a thousand raised to the power of six (10^{18})
[8] International Data Corporation

into business and the *veracity* of data, which refers to quality of the data (cleanliness and accuracy) [19].

[20] very perceptively stated that this growing trend of big-data is reinforced by the advent of the Internet, the proliferation of smartphones, and the Internet of Things (IoT) devices and sensors. If leveraged appropriately, big data can lead to meaningful breakthroughs, actionable insights, optimal resource allocation, and foresighted business decisions [21].

2.1.2.2 Tourism Big-Data

The field of tourism and hospitality is a key contributor for this abundance of information. [20] states that tourism destinations, firms, and consumers increasingly create and deploy large volumes of data to improve their decision-making processes and co-create value.

Tourism consumers tend to leave behind enormous amounts of online data through all phases of their travels – before travel during their planning phase, during travel through social-media sharing, and after their travel by leaving reviews and comments. It is up to tourism stakeholders to ask the right questions, gather the supporting available data, extract value from it, and transform it into applicable insights.

2.2 User-Generated Content (UGC) as Tourism Big-Data Source

2.2.1 Web 2.0: Rise of Social Media and UGC

For the first time in 40 years, TIME magazine’s 2006 Man of the Year was given not to a man, not a personality, but was given to “You” – a recognition of the millions of people who contribute to user-generated content. This is just one of the effects of the epidemic rise of User-Generated Content (UGC) and Social Media.

2.2.1.1 Web 2.0

Web 2.0 can be described as the technical infrastructure (both software and hardware) that enables and facilitates content creation, interaction, and the collection of UGC[22, 23]. Leveraging the technologies of Web 2.0, there is a clear shift of focus away from firms and companies and towards users and consumers. Web 2.0 is divided into 5 main categories: (1) Blogs such as www.huffingtonpost.com, (2) Social Networks such as www.facebook.com, (3) Content Communities such as www.youtube.com, (4) Forums/Bulletin Boards such as www.python.org, and (5) Content Aggregators such as www.google.com. Users of Web 2.0 applications are crucial, not only as consumers, but also as content creators [22].

2.2.1.2 Social Media and UGC as a Big-Data Source

Social media is the conception of applications built on the Web 2.0 technologies. It is formed through a cluster of mediums which aids the interactions between individuals. Social media, at its core, is meant to be highly accessible and scalable in nature [24]. The user-generated content that is available on social media typically consist of text, pictures, videos, and networks [22].

Social media is now becoming one and the same as big-data. The content on social media, such as tweets, comments, posts, and reviews, have contributed to the extensive creation of big data [25]. Social media is massive in size, has a high update speed, and has a vast range of content -- incorporating all the 3V characteristics that define big-data [26]. Taking Twitter as an example, the hundreds of billions of tweets give it "volume", its hundreds of millions of tweets a day give it "velocity", and its mix of text, imagery, and video offer "variety" [26].

2.2.1.3 Electronic Word-of-Mouth and Its Credibility

Electronic Word-of-Mouth (eWOM)—sometimes also referred to as word-of-mouse—is defined as "all informal communications directed at consumers through Internet-based technology related to the usage or characteristics of particular goods and services, or their sellers" [27]. eWOM is particularly important for the tourism sector because tourism and hospitality products and services are difficult to evaluate as they are intangible goods [28]. Such information plays a significant role in many aspects of tourism, especially in information search, decision-making behaviors, tourism promotion, and focusing on best practices for interacting with consumers [29].

Potential tourists highly rely on other's experiences for their decision-making, due to the experiential nature of tourism products [27]. User-generated content is many times seen as recommendations from "family and friends". It is therefore becoming a vital information source to potential tourists and is seen as more trustworthy and credible than information provided by destination or tourism service providers [30, 31]. Due to this, UGC is more inclined to direct and influence tourist choices and decisions.

[31] conducted a study to assess how much trust tourists place in different Travel 2.0 applications and how much influence they exert on tourists' perception and decisions. According to an the online survey conducted: "Respondents reported that, after having read reviews and comments posted online (UGC), they changed their hotel accommodation sometimes (64.8%), almost always (12%) or always (0.5%)". Furthermore, the study also concluded: "UGC applications quite often cause tourists to change their accommodation even once their decision has been taken and their trustworthiness is assessed by tourists as being higher when there is the same proportion of positive and negative comments and reviews".

A vast number of research and big-data analytics has been done using tourism user-generated content. This further reinforces that there is some level of trust put towards user-generated content, whether it be from traveler peers, NTOs and DMOs, or third-person researchers.

2.2.2 UGC Big-Data Applications in Tourism and Hospitality

Researchers have been able to see and understand the value of user-generated content in the tourism industry. This can be seen by the various analyses conducted over the past decade to investigate online reviews in the Travel and Tourism sector.

2.2.2.1 UGC Analyses in the Hotel Industry

Much research has been done within the Hotel industry through the analysis of user-generated content (as seen in the compiled list of past research in ***Table 1***).

[32] conducted text mining and content analysis of online hotel reviews to find determinant of customer satisfaction in hotel venues. They went through their content analysis by implementing text pre-processing (creating “bag of words” and separating “budget” hotels from “luxury” hotels), parsing (segmenting Chinese characters in order to identify words in a sentence), and frequency count. Through their research, the team was able to find factors that customers consider important (transportation convenience, F&B management, convenience to tourist destinations, and value for money).

[33] wanted to shed light on ways travelers’ rating patterns differ between independent and chain hotels. In order to do so, they categorized travelers by their profiles and hotels by their geographical location. They conducted a 5 (profiles) X 4 (regions) Two-Way ANOVA for each hotel type (chain and independent). Some of their key findings are “business travelers generally showed the most stringent rating patterns, especially for independent hotels in Asia Pacific” and “independent hotels in Europe received the highest ratings while those in Asia Pacific attracted the lowest ratings”.

[34] conducted an advanced linguistic analysis on hotel reviews in order to extract meaning from content provided by visitors. The team performed a Stepwise Regression on star-ratings (numerical data) vs. TripAdvisor’s 5-level hotel consumer rating (numerical data)— “cleanliness”, “service”, “location”, “room”, and “value”— to identify the most important dimensions to hotel consumers. They also performed a latent Dirichlet allocation (LDA) analysis on customer reviews (text data) to reveal meaningful dimensions (factors) of hotel services which otherwise would not have been known.

Table 1: *Literature Review on UGC Analysis for Hotels*

Ref	Authors	Scope	Platform	Objective	Methodology
[32]	Li, Ye & Law (2013)	Hotels (Beijing, China)	42,668 Daodao reviews	Identify determinants of customer satisfaction in hospitality venues	Content Analysis (ICTCLAS)
[35]	Barreda & Bilgihan (2013)	Hotels (Northeast USA)	17,357 TripAdvisor reviews	Identify the main themes that motivate consumers to evaluate hotel experiences in online environments	Content Analysis (NVivo 8)
[33]	Banerjee & Chua (2016)	Hotels (America, Asia Pacific, Europe, Middle East, Africa)	39,747 TripAdvisor reviews	Examine the rating patterns of hotels for different traveler profiles	ANOVA & Text Mining
[36]	Berezina, Bilgihan, Cobanoglu & Okumus (2016)	Hotels (Florida, USA)	2,510 TripAdvisor reviews	Examine underpinnings of satisfied and unsatisfied hotel customers	Text Mining: Word Categorization (PASW Modeler) & Text-Link Analysis

[37]	Geetha, Singha & Sinha (2017)	Hotels (Goa, India)	TripAdvisor reviews	Establish a relationship between review sentiment and review rating for hotels	Sentiment Analysis (Naïve Bayes) & Hierarchical Cluster Analysis
[34]	Guo, Barnes & Jia (2017)	Hotels (16 countries)	266,544 TripAdvisor reviews	Mine the sensitive and important factors influencing consumer satisfaction through UGC	Latent Dirichlet Allocation (LDA) & Perpetual Mapping
[38]	Xiang, Du, Ma & Fang (2017)	Hotels (Manhattan, NYC, USA)	438,890 TripAdvisor, 480,589 Expedia, & 30,816 Yelp reviews	Comparatively examines three major online review platforms	Latent Dirichlet Allocation (LDA), Sentiment Analysis (Naïve Bayes), Linear Regression
[39]	Ye, Luo & Vu (2018)	Hotels (Hong Kong)	115,649 TripAdvisor reviews	Understand location preferences to detect demand pattern	Time Series Analysis (TSA)
[40]	Bi, Liu, Fan & Zhang (2019)	Hotels (2 5-Star)	24,276 TripAdvisor reviews	Conduct importance-performance analysis (IPA)	Latent Dirichlet Allocation (LDA), IOVO-SVM, & Ensemble Neural Network Model (ENNM)
[41]	Cheng, Fu, Sun, Bilgihan & Okumus (2019)	Lodge Listings (New York City, USA)	1,485 and 10,000 AirBnb reviews	Investigate the effect of online review comments on potential guests' trust perception	Content Analysis & Convolutional Neural Network (CNN) Modeling
[40]	Bi, Liu, Fan & Zhang (2020)	Hotels (140 countries)	1,547,869 TripAdvisor reviews	Understanding the asymmetric effects of attribute performance (AP) on customer satisfaction (CS)	Penalty-Reward Contrast Analysis (PRCA) & Asymmetric Impact-Performance Analysis (AIPA)

2.2.2.2 UGC Analyses in Tours and Activities

Considering that the intention of this thesis is to provide insights on tours and activities in Bangkok, it's only fair to look into research and big-data analytics administered for tours and activities specifically (*see **Table 2***).

[42] conducted a highly technical analysis on tourist attractions in Phuket, Thailand. The purpose of their research, as stated in their research, is to "develop a methodology that can analyze online reviews using ML [machine learning] techniques in such a way that practitioners in the fields of tourism & destination management can understand and apply to improve their attractions". A combination of latent Dirichlet allocation (LDA)—the first ML technique—and the elbow method, and the k-means clustering algorithm, and Naive Bayes

modelling—the second ML technique—was implemented to identify and categorize dimensions of each attraction.

[43] also used the LDA algorithm to identify tourists' interests and use those insights to group (or cluster) attractions in Florida based on how well they meet these interests. The clusters were developed based on tourist origin markets: locals, out-of-state, or international. Different analyses such as network analysis, spatial analysis, and geo-visualizations were conducted in the study. Through the research, the authors were able to identify similarities and differences in attraction clusters and draw key insights on tourism trends and how the state of Florida could improve to fully utilize these trends.

Table 2: *Literature Review on UGC Analysis for Tours and Activities*

Ref	Authors	Scope	Platform	Objective	Methodology
[44]	Fang, Ye, Kucukusta, & Law (2016)	Tours/Attractions (New Orleans, USA)	41,061 TripAdvisor reviews	Investigate the effects of reviewer characteristics inferred from properties of historical rating distribution	Negative Binomial Regression & Tobit Regression Model
[43]	Kirilenko, Stepchenkova, & Hernandez (2019)	Attractions (Florida, USA)	157,285 TripAdvisor reviews	Identify attraction clusters	Latent Dirichlet Allocation (LDA)
[45]	Simeon, Buonincontri, Cinquegrani, & Martone (2017)	Tours/Activities (Naples, Italy)	12,592 TripAdvisor reviews	Analysis online reviews to explore experiences of tourists	Content Analysis & Principal Component Analysis
[42]	Taecharungroj & Mathayayomchan (2019)	Attractions (Phuket, Thailand)	65,079 TripAdvisor reviews	Analyze online reviews for DMOs to understand and apply in order to improve their attractions	Feature Extraction (LDA) and Sentiment Analysis (Naïve Bayes Modeling)
	This Thesis	Tours/Activities (Bangkok, Thailand)	59,758 TripAdvisor reviews	Develop insights on tourist preferences and trends via online reviews	Content Analysis (Association Rules Mining, Sentiment Analysis, & NLP) and Machine Learning Prediction Models (Logistic Regression, Support Vector Machine, and Random Forest)

2.2.3 Travel 2.0 Leader: TripAdvisor

Web 2.0 applications within the tourism and hospitality industry has been nicknamed Travel 2.0 by Philip C. Wolf, CEO of PhoCusWrite, a leading consultancy firm in the travel and tourism sector [46]. Just like all other, this sector is also moving away from B2C marketing towards a more peer-to-peer model – where tourism consumers are influencing one another. So much value is put in to peer comments, that information from Travel 2.0 users represent a more reliable and trustworthy source than the suppliers themselves [46].

2.2.3.1 TripAdvisor's Size

One such source of Travel 2.0 is TripAdvisor – the world's largest travel platform. The application services over 460 million unique travelers each month [47], making it the most popular online source of travel information. TripAdvisor retains an immense amount of data, with more than 859 million reviews of over 8.6 million accommodations, restaurants, experiences, airlines, and cruises [48]. The site's primary function is the collection and dissemination of user-generated content—reviews, ratings, photos, and videos—on a highly specific domain, namely travel [49].

2.2.3.2 TripAdvisor's Credibility

In the past, there has been doubt cast on the authenticity of the UGC on TripAdvisor. So much so, that one of TripAdvisor's competitors, SideStep.com, estimated that approximately 2% of the site's published reviews are "bogus" [50]. TripAdvisor has come a long way since those scandals from the early 2000's. The firm regularly posts notices prominently throughout the site warning that fake reviews will not be tolerated, and that hotels or tours attempting to manipulate the system will be penalized in their rankings and have a notice posted indicating that they post fake reviews [49]. Additionally, TripAdvisor also publicly states (on their website) that they have the technology in place and a team to screen reviews to ensure they are: family-friendly, posted to the correct business, and are in compliance with all guidelines.

Furthermore, there are policies in place to hinder organized boosting – such as the policy that reviews submitted to the site must be submitted by an individual traveler and not a third party. Lastly, in cases such as TripAdvisor where the content is so massive, the "power of the crowd" nullifies large negative ramifications of fake reviews. As the number of reviews grow, the impact of fabricated content diminish as they get overwhelmed by genuine UGC [49].

2.2.4 Machine Learning Models

From as early as 1968, [51] stated that "if computers could learn from experience their usefulness would be increased". Over the years, machine learning algorithms have evolved to break limitations, increase simplicity, and sky-rocket in accuracy. With the growth of the internet and high availability of information, the usage of machine learning algorithms has grown to encompass almost all applications, functions, and industries. As seen in ***Table 3***, today machine learning models are used in T&T, Energy & Gas, Banking, Medicine, and even Education and Food & Beverage.

***Table 3:** Research Using Machine Learning Models*

Ref	Authors	Year	Industry	Scope	ML Algorithms Used
[52]	Shafiq M., Yu X., Langhari A.A., Yao L., Karn N.K., Abdessamia F.	2016	Tele-communications / Computer Science	Analyzing and identifying different types of applications flowing in a network for internet service providers or network operations to manage overall network performance	Support Vector Machine, C4.5 Decision Tree, Naïve Bayes, Bayes Net
[53]	Singh, M.J., Girdhar, A.	2018	Computer Science	Introducing a new method of fingerprint image enhancement to increase security	Support Vector Machine
[54]	Kingsly, A.A.S., Mahil, J.	2019	Medicine	Identifying melanoma using learning base classifiers and classifying skin cancer images into cancerous and non-cancerous	Support Vector Machine
[55]	Wadhe, A.A., Suratkar, S.S.	2020	Hospitality / Travel & Tourism	Classifying sentiment analysis results to draw insights	Naïve Bayes, Support Vector Machine, Random Forest
[56]	De Nadai Fernandes, E.A., Sarriés, G.A., Bacchi, M.A., Mazola, Y.T., Gonzaga, C.L., Sarriés, S.R.V.	2020	Food & Beverage	Analyzing beef samples for their elemental content and classified according to their origin in order to increase beef traceability	Multilayer Perceptron, Random Forest, Regression Tree
[57]	Kumari, P., Toshniwal, D.	2021	Energy & Gas	Forecasting hourly global horizontal irradiance for reliable planning and efficient designing of solar energy system	Random Forest, Support Vector Machines, Extreme Gradient Boosting Forest, and Deep Neural Networks
[58]	Jemima Jebaseeli, T., Venkatesan, R., Ramalakshmi, K.	2021	Banking	Detect credit card fraud and prevent huge financial losses with more accuracy as compared to other algorithms	Random Forest
[59]	Upadhyay, A., Palival, U., Jaiswal, S.	2021	Medicine	Detecting and recognizing whether MRI scans of brain consist of tumor or not in order to avoid man-made mistakes in detection of brain tumor	Random Forest

[60]	Gajwani, J., Chakraborty, P.	2021	Education	Predicting the academic performance of a student based on certain attributes of an educational dataset – attributes are demographic, behavioral, and academic	Logistic Regression, Decision Tree, Naïve Bayes, Random Forest

In the earlier years, around 5-10 years prior, it can be seen that one of the most commonly used machine learning algorithms for classification problems was Support Vector Machine. Not only is it easy to understand and one of the most common machine learning algorithms, but the methodology also results is high accuracy and insightful findings. [52] used SVM within the Telecommunications space to analyze and identify the different types of applications flowing within a network. [53] used the algorithm to classify fingerprint images with the end goal of enhancing the image and biometric identification. [54] also used Support Vector Machine, this time within the medical space. The algorithm was used to classify melanoma images into "cancerous" and "non-cancerous" with a goal to improve skin cancer detection.

More recently, however, the Random Forest algorithm has gained popularity and is quite frequently used in prediction models – for both classification and regression models. As stated by [61], the ensemble method has "gained significant importance from researchers, owing to their stable, simple yet powerful and robust prediction algorithms". [55] used both Support Vector Machine and Random Forest within the Travel and Tourism space, in order to classify sentiment analysis. From their research, they were able to find both algorithms performing similarly, with Random Forest having slightly higher accuracy. [56] used the algorithm within the Food & Beverage space, classifying beef samples through elemental content features in order to increase beef traceability. Similarly, [57], [58], [59], and [60] also used Random Forest in their research and prediction models – further solidifying the hypothesis of Random Forest's recent increased popularity.

Chapter 3: Methodology

3.1 Preliminary Analysis

3.1.1 Data Collection

Similar to many studies done in the past, TripAdvisor's online reviews were used as a data source to learn more about Travel and Tourism consumer preferences. In order to gather the required data items in a timely manner, scraping the TripAdvisor website was necessary. After much research on tools and services that help with web scraping, ParseHub's 'Standard' package plan was chosen to do the job. ParseHub, as stated on their website, is a powerful web scraping tool that makes the task of scraping as easy as clicking on the data-items required. The company provides a GUI-based service that allows for data to be extracted from any website on to excel spreadsheets.

For the scope of this thesis, only reviews from January 2010 – Jan 2020 for Tours and Activities in Bangkok were gathered. The online reviews collected for this research were categorized into 6 groups: (1) Activities, (2) Bike Tours, (3) Cooking Classes, (4) Food Tours, (5) Sight Seeing, and (6) Spas. ***Figure 2*** shows what a review on TripAdvisor looks like for each Tour/Activity category.

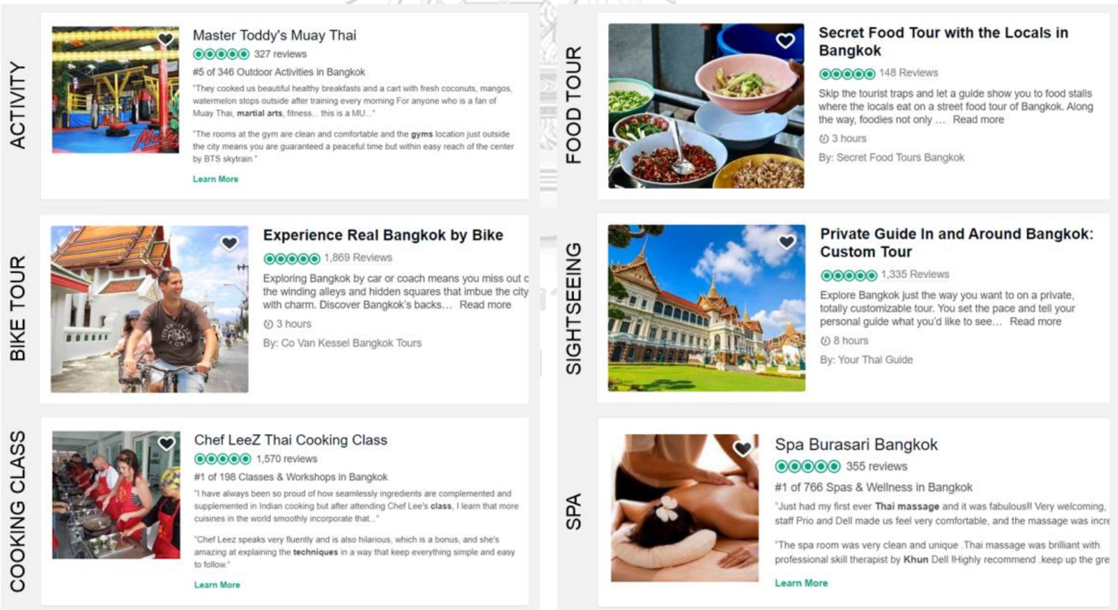


Figure 2: *TripAdvisor Reviews per Tour/Activity Category*

With the help of ParseHub, over 68,000 reviews for Bangkok Tours/Activities were gathered. However, not all reviews could be immediately used; initial data clean-up was required. This included removing reviews that didn't have a rating, reviews that were not in English, reviews that were not relevant to the scope, and reviews that were duplicates. After cleaning up, a total of almost *60,000 reviews*[9] remained to work with (see ***Figure 3***).

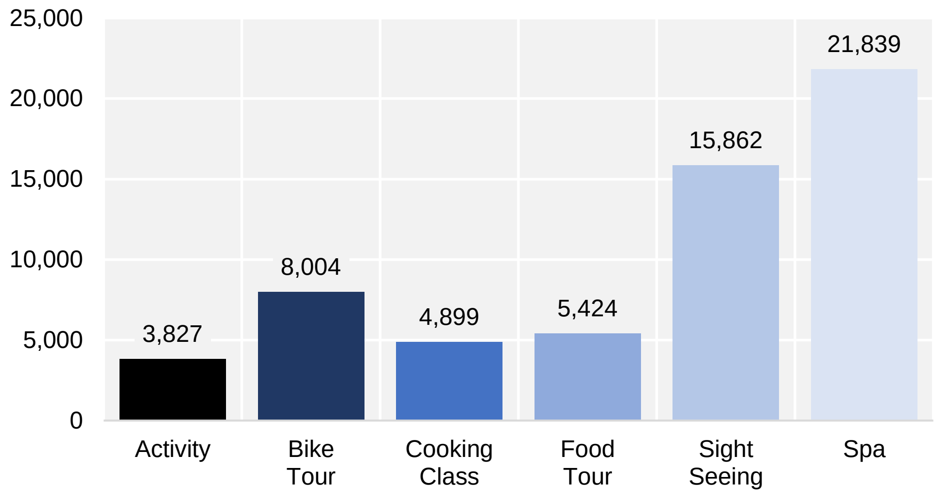


Figure 3: *Number of Collected Reviews by Tour/Activity Type*

From the dataset of around 60,000 reviews, additional cosmetic clean-up was required to further streamline the data. This included standardizing the date format across all reviews and classifying the "location" field –which, on TripAdvisor, was a free-text field where users were able to fill in anything from cities and towns, to countries and continents—to countries. Origin countries were then grouped into 10 origin 'groups'. From ***Figure 4***, you can see that most reviews on TripAdvisor come from western countries – specifically West Europe and North America, followed by Southeast Asia and Australasia.

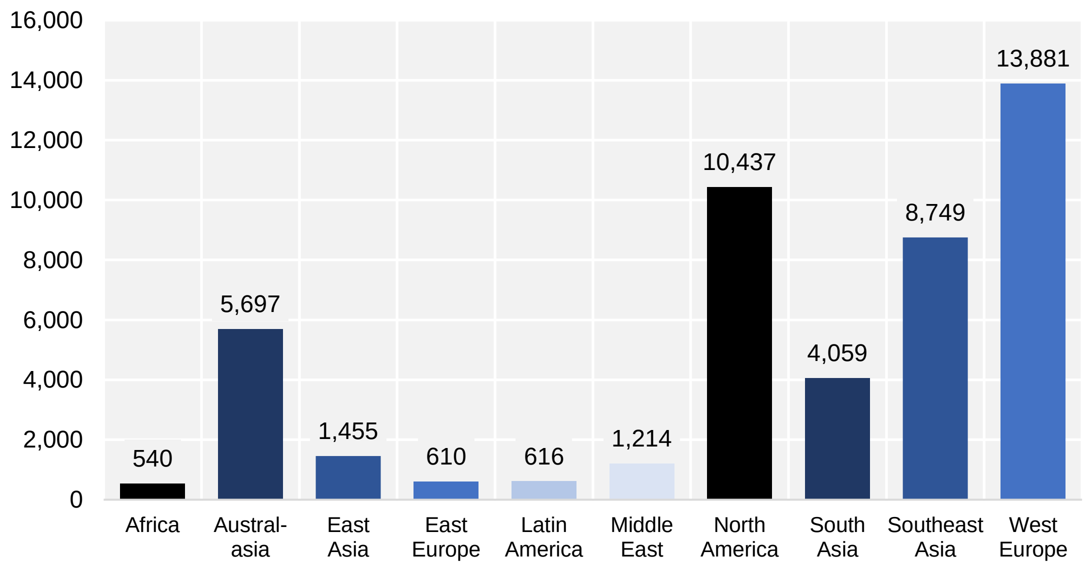


Figure 4: *Number of Collected Reviews by Reviewer Origin*

[9] 59,758 reviews remained for this research

However, not all the 60,000 reviews had a full set of origin and date information. 80%[10] of the dataset had origin information (see ***Figure 5***) and 97%[11] had date information (see ***Figure 6***).

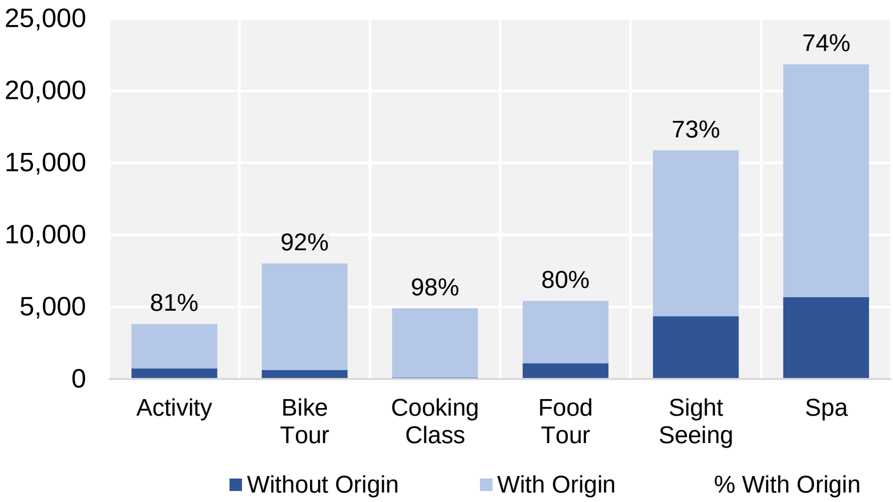


Figure 5: *Number of Collected Reviews With and Without Origin (by Tour/Activity Type)*

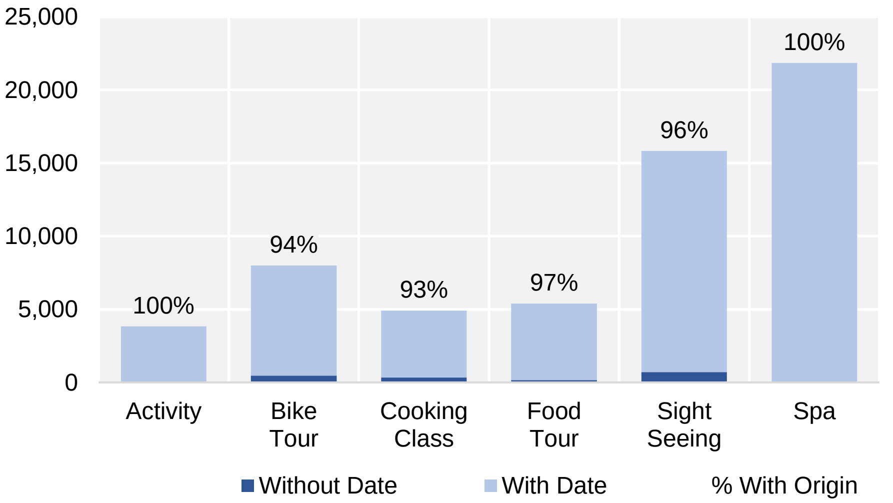


Figure 6: *Number of Collected Reviews With and Without Date (by Tour/Activity Type)*

3.1.2 Tourist Preferences and Trends via Descriptive Statistics

3.1.2.1 Chi-Square Test for Independence

In order to accredit any further insights drawn from the analyses of the collected data, it was important to prove that the features of the data items were somehow related – thus *not* independent. To do so, a Chi-Square Test of Independence was carried out. The top features of the dataset where most of the insights would be drawn from were Origin, Review Rating, and Tour/Activity Type. Thus, the chi-square test was taken for the following three

[10] 47,258 reviews of 59,758 reviews have origin information
[11] 58,091 reviews of 59,758 reviews have date information

relationships: (1) Origin & Review Rating, (2) Tour/Activity Type & Review Rating, and (3) Origin & Tour/Activity Type.

As per most chi-square tests, the null hypothesis (H0) assumed that Feature A and Feature B had no association (they were independent). The alternative hypothesis (H1) assumed that there was an association between Feature A and Feature B (they were not independent). An in depth explanation of how the test was carried out for one of the feature pairs (Origin & Review Rating) is shown in ***Appendix 1***.

From the results of the chi-square test (as seen in ***Table 4***), it is clear that all three feature pairs had some sort of relation and were not independent. Hence, further analyses on these features could be carried out.

Table 4: *Chi-Square Test for Independence Results*

H0: No Association (Independent)	H1: Association (Not Independent)			
Items Tested	*Chi-Square Statistic*	*Degrees of Freedom*	*Critical Value*	*Decision*
Origin & Review Rating	2,083	36	51	2,083 > 51; Reject H0
Tour/Activity Type & Review Rating	4,308	20	31	4,308 > 31; Reject H0
Origin & Tour/Activity Type	11,112	45	62	11,112 > 62; Reject H0

3.1.3 Tourist Association via Association Rules Mining

An interesting observation to seek out would be to find out which Tours/Activities a single tourist would repeatedly prefer. Association Rules Mining – a procedure to find patterns in data – helped with just that. Association Rules are simple if/then statements that help discover relationships, for example: If (people buy diaper), then (they buy baby powder) [62]. Similarly, an example of something this research was aiming to find out is If (people enjoy spa), then (what else do they tend to enjoy)?

3.1.3.1 Market Basket Analysis (MBA)

A very popular application of Association Rules is Market Basket Analysis (MBA), commonly used by large retailers to find associations of items that are usually bought together. Two key metrics to understand for association rules are:

1. **Support -** how much historical data supports the rule (or in terms of retail, percentage of “baskets” that contain the item set)
2. **Confidence -** how confident are we that the rule holds (or in terms of retail, percentage of times item B is purchased, given that item A was purchased)

In order to carry out the MBA application of Association Rules Mining, the data had to be prepared in a way that identified each unique reviewer (as a primary key) and associated it with the Activity/Tour type they had participated in (i.e. left a review for) – a snippet of the data is shown in ***Table 5.*** The data thrown into the model included all the reviewers and their associated Activity/Tour– whether the reviewer left a single review, multiple reviews within the same Activity/Tour category, or multiple reviews across various categories. R programming and the packages “arules” [63] and “arulesViz” [64] were used to carry out the Apriori Method of Association Rules Mining.

Table 5: *Association Rules Mining - Data Preparation*

Activity/Tour Type	Package Name	Reviewer ID
Spa	Perception Blind Massage	Alan S_Australia
Spa	Sook Sabai Health Massage	Alan S_Australia
Bike Tour	Experience Real Bangkok by Bike	Alan S_Australia
Spa	Lavana	Alvina Ho_Hong Kong
Spa	Urban Retreat Spa - Asok	Alvina Ho_Hong Kong
Sight Seeing	Private Tour of Bangkok's Temples	Alyssa C_USA
Food Tour	Bangkok Midnight Food Tour by Tuk Tuk	Alyssa C_USA

3.1.4 Tourist Sentiment via Sentiment Analysis

A very important feature to analyze when looking at customer preferences are their feelings towards the product/service offerings. Consumer feelings can be discovered through sentiment analysis – the interpretation and classification of emotions (positive, negative, and neutral) within text data using text analysis [65].

For this research, sentiment analysis was carried out with the help of the R Studio package “sentimentr” [66], the lexicon[12] of 2,006 ‘positive’ words, and a lexicon of 4,783 ‘negative’ words compiled by [67]. An example of the list of ‘positive’ and ‘negative’ words can be seen in ***Table 6*** with a more extensive list in ***Appendix 2***. For every sentence, a ‘sentiment score’ was calculated by counting the frequency of positive words (increment a positive score) and the frequency of negatives words (increment a negative score) and summing them up. Then for each review, the average of every sentence’s sentiment score was taken – leaving every review with an ‘average sentiment score’. An example of the negative sentiment score for a 1-star review is shown in ***Table 7***.

Table 6: *Positive & Negative Words (Examples)*

Positive Words		Negative Words	
accurate	affordable	abnormal	absurd
admirable	amaze	abrasive	afraid
adorable	amusing	absence	aggressive

[12] A lexicon is (a list of) all the words used in a particular language or subject

Table 7: Sentiment Score for 1-Star Rating (Example)

	Sentiment Analysis (by Sentence)				
		Element ID	Sentence ID	Word Count	Sentiment
"Firstly, a two tier pricing system. White people pay more than double what Thais pay. It's a fact. If you want to be victim of racism with firsthand experience, this place is for you. Prices, more expensive than Europe for a days fishing, that says it all. Bait - what a rip off"	1	1	1	6	0.0000
	2	1	2	9	0.0133
	3	1	3	3	0.0000
	4	1	4	16	-0.4375
	5	1	5	13	-0.1248
	6	1	6	5	-0.8944
	Sentiment Analysis (by Review)				
		Element ID	Word Count	Standard Deviation	**Average Sentiment**
	1	1	52	0.3632	**-0.2859**

3.1.5 Tourist Focus via Natural Language Processing

Although sentiment analysis is quite interesting, it is limited to interpreting just the 'positive' and 'negative' feelings of reviewers. In order to discover what reviewers are focusing on, it is necessary to look at the features they are frequently mentioning. Applying Natural Language Processing to do a Word Frequency Count was chosen to present insights on what items are most frequently written about in each Tour/Activity. Further classifying the Frequency Count by star-rating was done to additionally reveal what items reviewers liked when they were satisfied (5-star rating) and what they disliked when they are disappointed (1-star rating).

3.1.5.1 Data Pre-Processing

In order to get the best result from the Word Frequency Count, it was essential to "clean up" the text and remove any words or punctuations that could alter the results. Data pre-processing for natural language processing was done in 3 main steps (as seen in ***Figure 7***):

1. *Data Segregation.* Review content was categorized into 12 sub-categories (6 Tour/Activity Types x 2 Levels of Satisfaction – 5-star & 1-star)
2. *Corpus Creation*. For each sub-category, all the sentences from all the reviews were collapsed into a corpus[13].
3. *Bag-of-Words Creation*. Several processes took place in order to bring the corpus to be a list of essential words:
    a. *Clump Negatives* to make phrases such as "not worth" → "notworth" so that negatives won't lose their connotation when words are separated by spaces later
    b. *Remove Punctuations* (such as . , ! ;)
    c. *Remove Numbers*

[13] A corpus represents a collection of (data) texts; in machine learning area, it is referred to as a body (collection) or writings

d. *Convert to Lowercase* in order to accurately count frequency without worrying about case sensitivity
e. *Remove Stop Words* (explained in next section)
f. *Remove Remaining Short-Character "Words"* (such as "a", "i", "ve", etc.)
g. *Remove Excess Space*

An essential part of preparing the data for Word Frequency Count was to remove unimportant words that don't provide any meaningful insight, also known as "stop words" in NLP and text mining applications. Examples of stop words include "the", "is", "and", "him", etc. There is no single universal list of stop words used by all natural language processing tools. On account of this, a list of 704 stop words (***Appendix 3***) was manually put together using three reliable sources [68] [69]. A snippet of what the list of compiled stop words looked like can be seen in ***Table 8***.

Table 8: *List of Stop Words (Example)*

a	be	came	do	eight
able	became	can	does	either
about	because	cases	doesn't	else
above	become	cause	doing	elsewhere

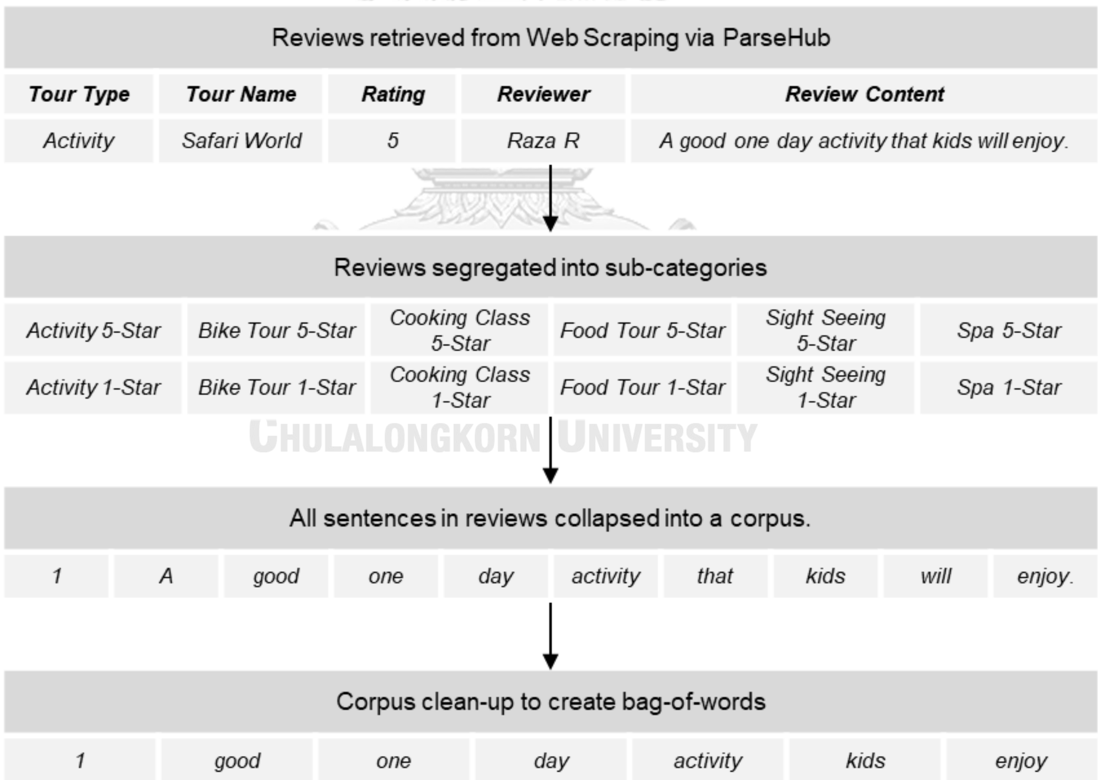


Figure 7: *Data Pre-Processing Flow (Example)*

3.1.5.2 Data Processing

The tedious part of the Word Frequency Count was the pre-processing of the text. The actual Data Processing was quite simple in comparison. For each sub-category's bag-of-words all unique words were taken into a data-frame and the number of times the word occurred in the "bag" was counted and associated with the word, resulting in the Word Frequency Count. All of these pre-processing and data-processing methods were accomplished using R Studio.

3.2 Machine Learning Models

3.2.1 Prediction Models

3.2.1.1 Predicting 5-Star Reviews

In an effort to identify which features significantly led to tourist satisfaction, models to predict 5-star ratings of reviews was built. A separate model was built for each of the six Tour/Activity categories. The features of each category were considered distinct enough to require different models.

The models in question were set up to predict whether a review was given 5 stars ("success", Y=1) or not ("failure", Y=0). The reason a binary dependent variable was used for the models was two-fold: (1) it *could* be used due to the high proportion[14] of 5-star reviews, ensuring a semi-balanced[15] dataset of successes and failures, and (2) it is known to be of the highest importance to tourism stakeholders – who consider a 5-star review to be a proxy for ultimate consumer satisfaction.

The independent variables used in the models were a combination of features of the original dataset and new features developed from prior analyses in this research. The independent variables (X) of the model were (1) average sentiment score (continuous data), (2-11) origin Boolean of reviewer origin (discrete data), and (12-31) frequent words Boolean of the top 10 highest-occurring words of 5-star reviews (discrete data). X1 originated from the Sentiment Analysis (***Section 3.2.4***), X2-X11 were features from the original dataset, and X12-X31 originated from the Frequent Word Count of NLP (***Section 3.2.5***) – see ***Appendix 6*** for more information. The process of collecting data and formatting into data frames that was used in the 5-star prediction models is shown in ***Figure 8***.

The reason that only the top 10 highest-occurring words were used in the prediction models was that with using more than 10 words – the model's effectiveness was not improved as well as the model's run-time increased, see ***Appendix 12*** for more information.

3.2.1.2 Predicting 1-Star Reviews

Similar to the previous prediction models, six additional models (models 7-12) were built to predict 1-star ratings. This was implemented in order to study which features significantly affect the dissatisfaction of tourists. Learning about which features make consumers unhappy could bring about a great opportunity for tourism stakeholders to make positive changes within the sector. For these models, the independent and dependent variables were the same as the previous models, apart from the independent variables X12 – X31. The frequent words for these models were the top 20 highest-occurring words for 1-star reviews instead of 5-star

[14] 75% of all reviews are 5-star (44,913 of 59,758)
[15] Only a "mild" degree on imbalance if minority class is 20-40% of the dataset

reviews. The process of collecting data and formatting into data frames that was used in the 5-star prediction models is shown in ***Figure 9.***

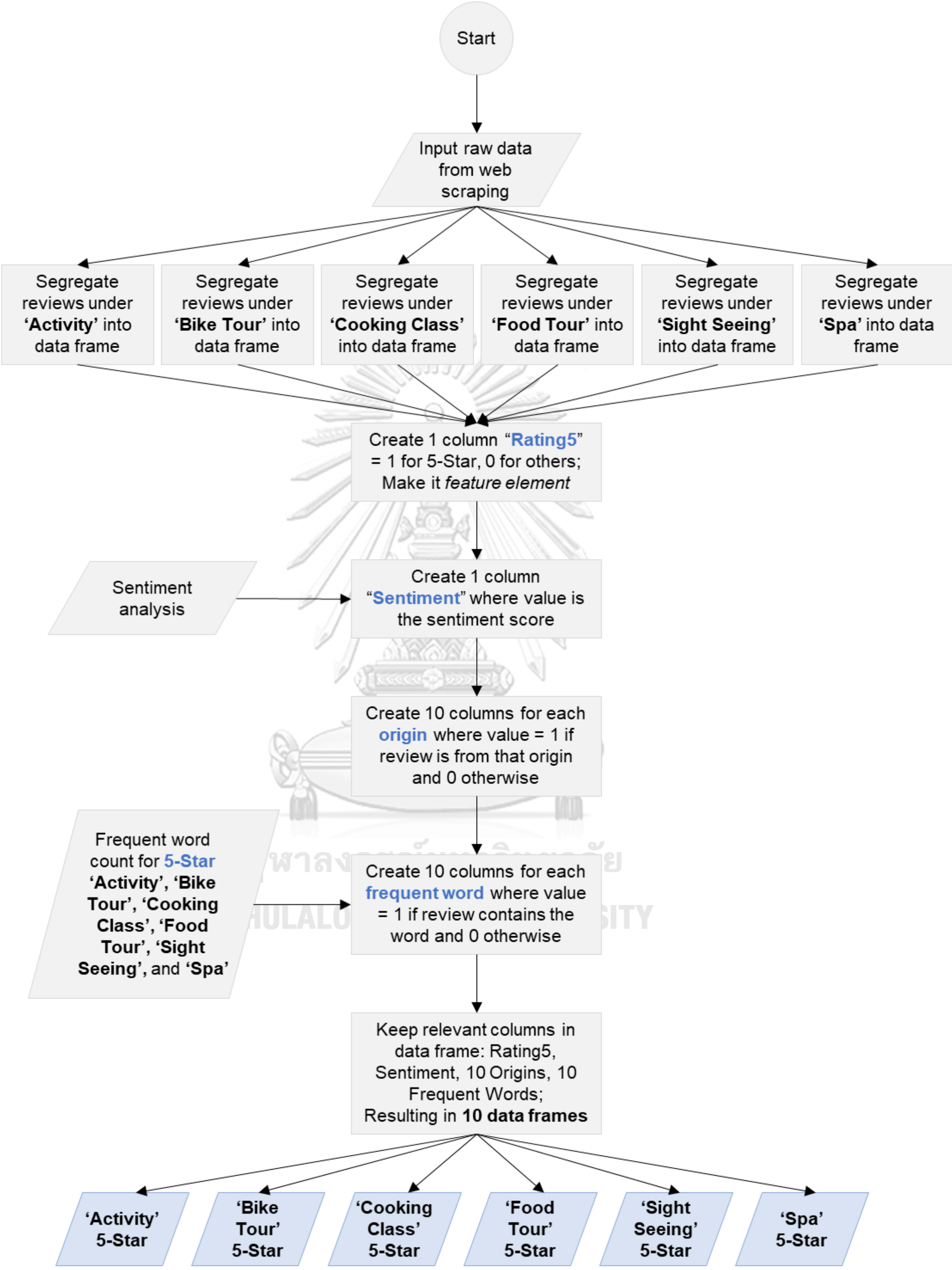


Figure 8: *Data frame creation for 5-Star Prediction Models*

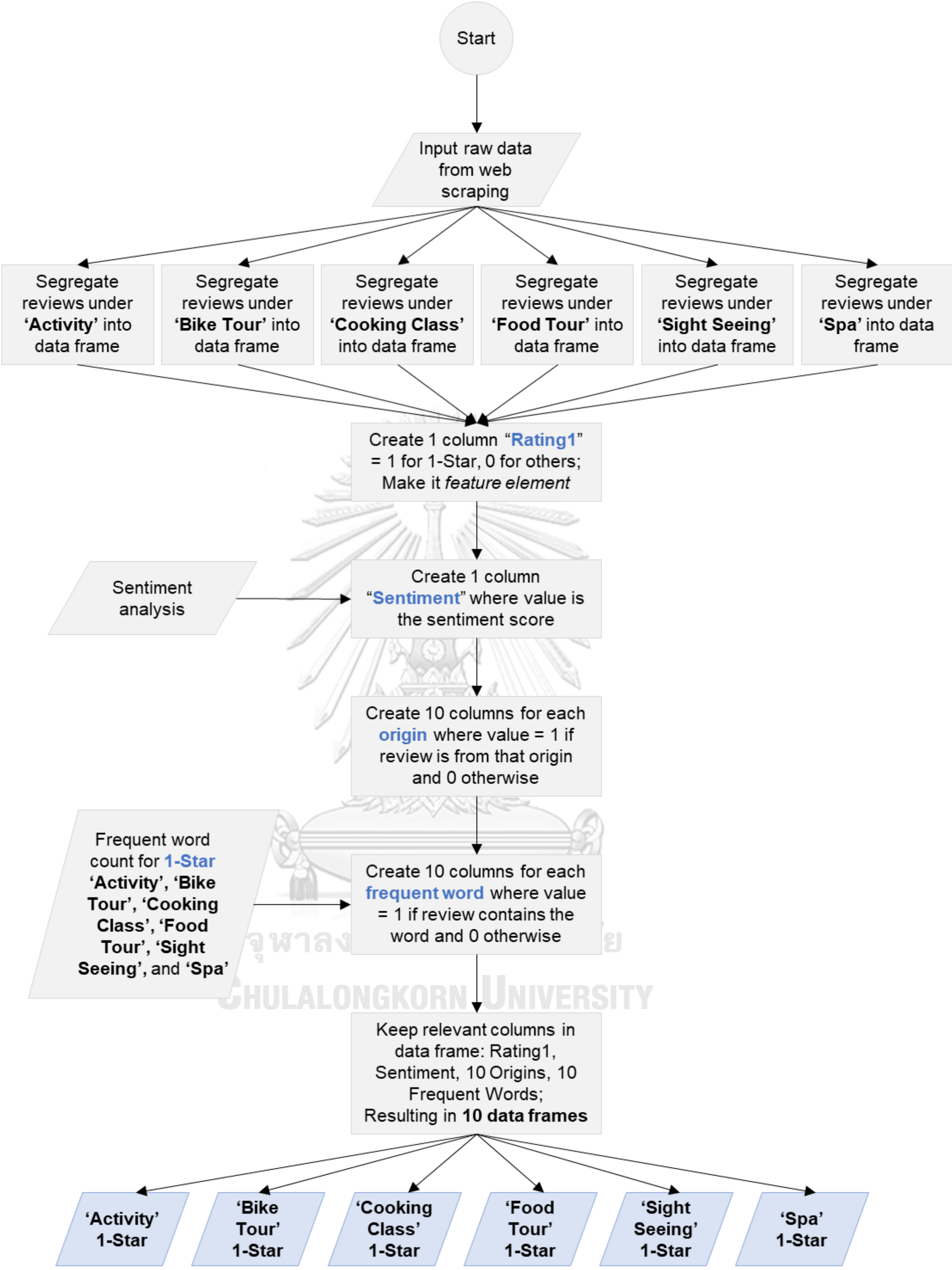


Figure 9: *Data frame creation for 1-Star Prediction Models*

3.2.2 Training and Testing Sets using K-Fold Cross Validation

An important part of evaluating the effectiveness of the prediction models was to *test* how well the models predicted the dependent variable. In order to do so, the data had to be split into training and testing sets. *Training Datasets* are known as the sample of data used to fit the model (usually a larger proportion of the data). *Testing Datasets* are known as the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset [70].

For this research, the split of the data, the creation of the model, and the testing of the model was carried out in R Studio, with the help of the "caret" [71] package. The data split was carried out using the k-fold cross validation method – a methodology where a given dataset is split into *k* number of sections (or "folds") and each fold is used as a testing set at some point [72], exactly *once*. For this these, the data was chosen to be split into 10 folds (10-Fold). This means that the entire dataset was randomly divided into 10 sections, where each fold/section was used for testing once against the 9 other folds that were used for training the model.

Additionally, the process of 10-fold re-sampling was further repeated 3 times (using the "repeatedvc" method) to ensure no biases and a robust model, finally ending up with 30 resamples[16]. The average accuracy was taken from all resamples to return the metrics for the entire model (see ***Figure 10*** for a diagram on the methodology). The sampling, splitting, modeling, and testing process was then repeated 12 times, for each distinct Tour/Activity type and 5-star/1-star combination.

An interesting point to note is that for each training set, the proportion of "success" data items was maintained as the same proportion for the entire subset used for modeling. For example, for model 1 ('Activity', predicting 5-stars) the proportion of 5-star reviews for the 'Activity' subset was 47.8% (see ***Figure 21***), thus around 47% of 5-star reviews was similarly maintained in each training set as well. Similarly, for model 2 ('Bike Tour', predicting 5-stars) the proportion of 5-star reviews for the 'Bike Tour' subset was 86.2% (***Figure 21***), thus the same proportion of 5-star reviews was also maintained in each training set for the prediction model. Going about the sampling this way ensures no distortion in predictions.

[16] It is important to note that 3 repeats of 10-Fold is *not* the same as 30-Fold.

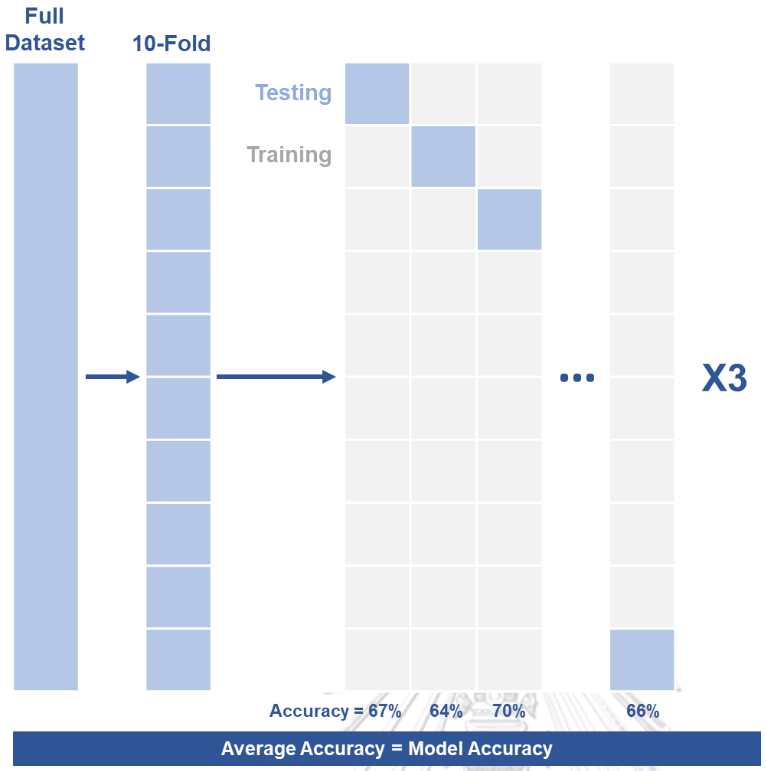


Figure 10: *10-Fold Cross Validation*

3.2.3 Model Effectiveness using Confusion Matrix and Prediction Metrics

A confusion matrix was used to test the accuracy of the Logistic Regression prediction model. A confusion matrix is a table that is used in machine learning to represent the performance of a classification model on a set of test data for which the true values are known [73]. With the confusion matrix, performance of the algorithm can be visualized by comparing the model's prediction ("Prediction") with the actual value ("Reference").

Concepts to understand regarding the confusion matrix are: True Positives (TP) are when both the prediction and the reference is positive, True Negatives (TN) are when both the prediction and the reference is negative, False Negatives (FN) are when the prediction is negative while the reference is positive, and False Positives (FP) are when the prediction is positive while the reference is negative (as seen in ***Figure 11***).

Important machine learning metrics (in %) derived by these concepts are:

- The **accuracy** of a model is given by (TP + TN) / (TP + TN + FP + FN); in other words, the percentage of correct predictions over the entire dataset. Accuracy provides the best measure for *symmetric* datasets, where the values of false positives and false negatives are almost the same.
- The **recall (sensitivity)** of a model is given by TP / (TP + FN); or the percentage of correct positives from the entire dataset of positives.

- The **precision** of a model is given by TP / (TP + FP); or the percentage of correct positives from all the predicted positives. High precision relates to low false positive rate.
- The **F1 score** of a model is given by 2*(Recall*Precision) / (Recall + Precision); or the weighted average of precision and recall. F1 is usually more insightful than accuracy, especially the dataset has an uneven distribution.
- The **specificity** of a model is given by TN / (TN + FP); in other words, the correctly labeled negatives over the entire dataset of negatives.

The dataset for predicting 5-star ratings, although not considered imbalanced, wouldn't be classified as balanced either. The proportion of "success" data items—or 5-star reviews—was 75%. Even worse, was the proportion of "success" for the models predicting 1-star ratings, which was 4% of the dataset. For these reasons, the metric of F1-score was applied to measure the effectiveness of the prediction models. However, upon the balancing of the datasets, as seen further in Section 3.2.4, the metric of accuracy also does an adequate job of measuring effectiveness, with the added ease of understanding the value. Thus, both **F1-score** and **accuracy** are used to measure the effectiveness of the models.

		References	
		No	**Yes**
Prediction	**No**	TN	FN
	Yes	FP	TP

Figure 11: *Confusion Matrix (Example)*

3.2.4 Classification Imbalance Correction

Before jumping into modeling the data, it is important to look out for data imbalances. Ideally, for optimal model results, the proportion of events and non-events in the Y variable should approximately be the same [74]. Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class [75].

As seen in ***Figure 21*** and ***Figure 22***, it is quite clear that review ratings were biased toward the upper end with most Tours/Activities rated at 4-stars and 5-stars. Naturally, this bias would translate to the data, causing there to be a data imbalance. For models predicting 5-star review ratings, the imbalance is around 30-70 – where the higher proportion of data have output Y=1 (5-star rating). For models predicting 1-star ratings, the imbalance is even more severe, at 97-3 – where the higher proportion of data have output Y=0 (*not* 1-star rating), as you can see in ***Table 9***.

Table 9: *Classification Imbalance (5-Star Ratings & 1-Star Ratings)*

Data for Predicting 5-Star Ratings			
	#Reviews	**Y=0**	**Y=1**
Activity	3,825	52.2%	47.8%
Bike Tour	7,988	13.8%	86.2%
Cooking Class	4,900	9.7%	90.3%
Food Tour	5,388	13.1%	86.9%
Sight Seeing	15,827	21.7%	78.3%
Spa	21,839	32.7%	67.3%
		23.9%	**76.1%**

Data for Predicting 1-Star Ratings			
	#Reviews	**Y=0**	**Y=1**
Activity	3,825	92.2%	7.8%
Bike Tour	7,988	99.6%	0.4%
Cooking Class	4,900	99.4%	0.6%
Food Tour	5,388	99.4%	0.6%
Sight Seeing	15,827	96.8%	3.2%
Spa	21,839	93.8%	6.2%
		96.9%	**3.1%**

In order to create a balanced dataset that would accurately predict 5-star and 1-star ratings, a random sample of reviews were removed in order to leave behind a balanced dataset. Ideally, a 50-50 balance would remain, however, due to constraints in the number of data items, data was removed in order to leave behind a 60-40 balance of data, as seen in ***Error! Reference source not found.***.

Table 10: *Classification Balanced (5-Star Ratings & 1-Star Ratings)*

Data for Predicting 5-Star Ratings			
	#Reviews	**Y=0**	**Y=1**
Activity	3,824	52.2%	47.8%
Bike Tour	2,757	40.0%	60.0%
Cooking Class	1,185	40.0%	60.0%
Food Tour	1,770	40.0%	60.0%
Sight Seeing	8,570	40.0%	60.0%
Spa	17,837	40.0%	60.0%
		42.0%	**58.0%**

Data for Predicting 1-Star Ratings			
	#Reviews	**Y=0**	**Y=1**
Activity	742	60.0%	40.0%
Bike Tour	88	60.2%	39.8%
Cooking Class	73	60.3%	39.7%
Food Tour	78	60.3%	39.7%
Sight Seeing	1,258	60.0%	40.0%
Spa	3,400	60.0%	40.0%
		60.1%	**39.9%**

Although there seems to be enough data items in the models predicting 5-star rating reviews, that wasn't the case for models predicting 1-star rating reviews, with some models having as few as 73 data items. Due to this data limitation, models predicting 1-star reviews for the categories Bike Tour, Cooking Class, and Food Tour are considering imprecise and thus not used for drawing further insights.

3.2.5 Hyperparameter Tuning

For most machine learning algorithms, certain parameters within the model can be optimized and adjusted. These values which control the model's learning process are called **hyperparameters**. Unlike parameters that are learned during training, hyperparameters have to be set *before* training. Choosing the right hyperparameter ensures an accurate machine learning model. The value helps with the tradeoff between bias and variance, making sure models aren't over- or under-fitted.

In order to find the optimal value of each hyperparameter, a tuning method using *grid search* was implemented. A grid search is a method where a subset of hyperparameters are pre-defined and used in an exhaustive search for the optimal values. For this thesis, hyperparameter tuning is implemented for the Support Vector Machine algorithm and the Random Forest algorithm. The values that the grid search varied through was found by looking at research done in the past that used Support Vector Machine and Random Forest, examples which are shown in the Literature Review section.

The grid search ran the machine learning algorithm on a random sample of 200 data points from the training set of each prediction model. The algorithm created all possible combinations of varying values of the hyperparameters (within the pre-defined range). Tuning then chose the values of the hyperparameters that resulted in the most effective model performance and returned a value called 'Best Performance'—a classification error where the *lower* the value the better.

3.2.6 Logistic Regression (LR)

Logistic Regression (LR), one of the most common classification prediction models, is carried out to understand a binary response (Y, dependent variable) on the basis of one or more predictors [77]. Simply put, logistic regression is a statistical model that uses a logistic function to model the probability of a random binary variable Y being either 0 or 1 (as seen in ***Figure 12***), given the independent variables (which can be either binary or continuous).

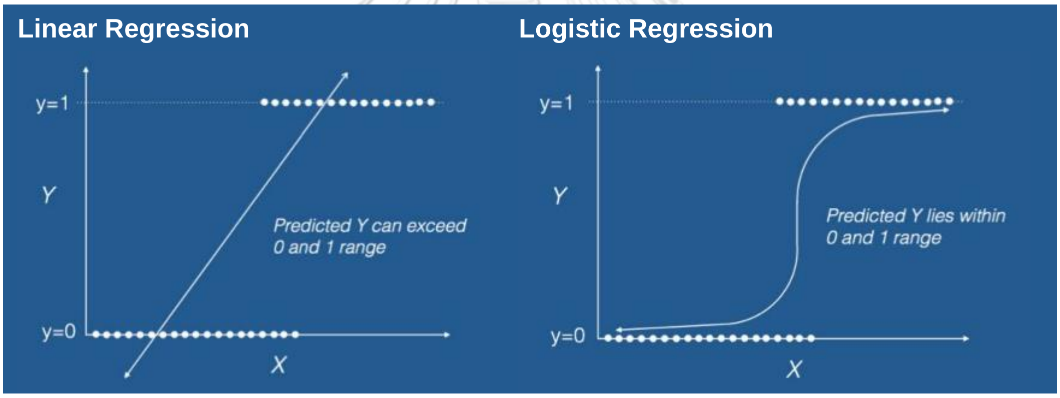


Figure 12: *Linear Regression vs Logistic Regression Graph*[17]

Running the Logistic Regression model, unlike the other two machine learning models in this study, did not have hyperparameter tuning. Modeling of logistic regression for this thesis was carried out using the **glm** method within the in R Studio [78]. GLM, or Generalized Linear Model, is a generalization of ordinary linear regression that allows for response variables to have error distribution models. The methodology flow of running the Logistic Regression algorithm can be seen in ***Figure 13.***

[17] Image courtesy Data Camp

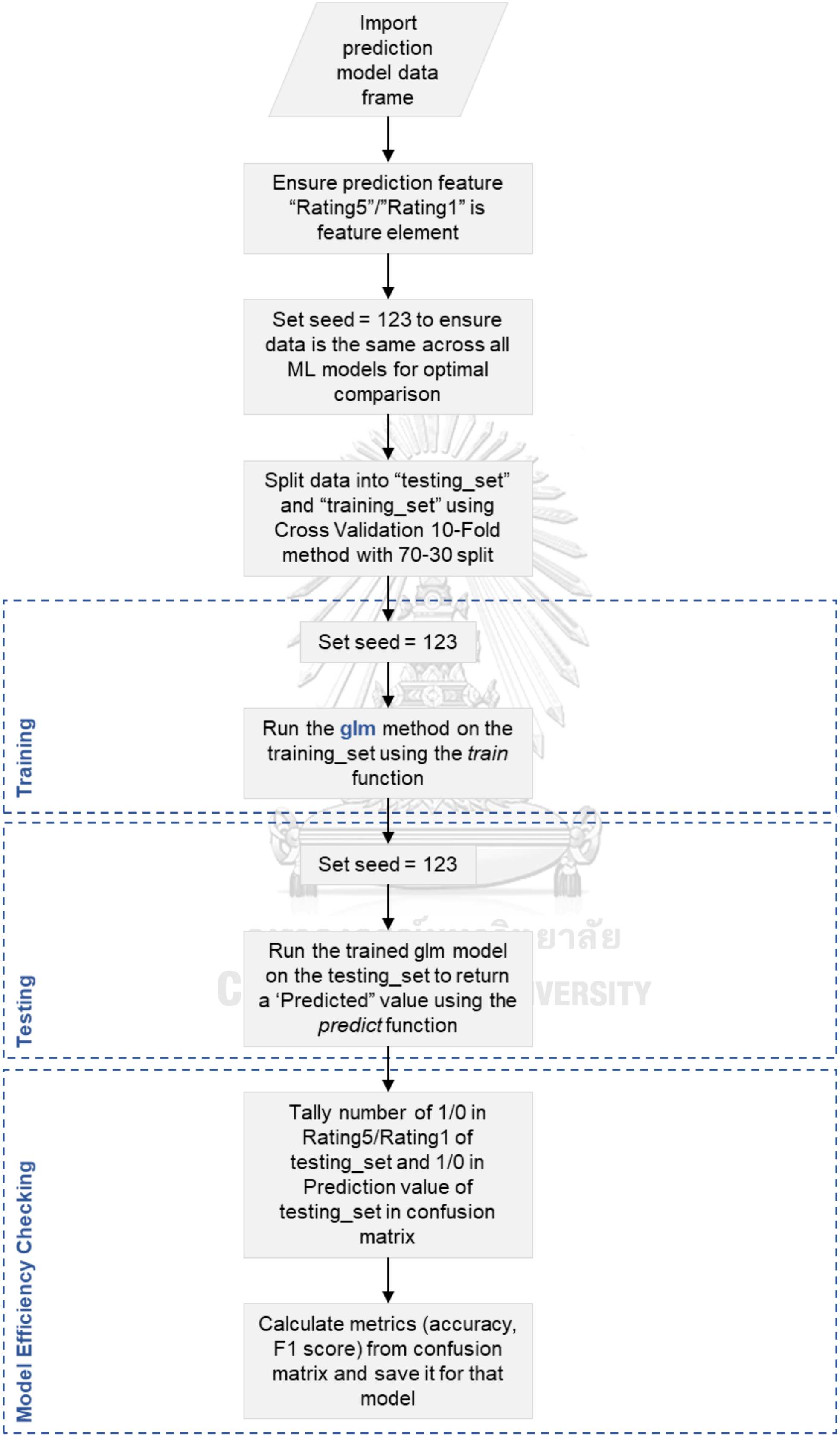


Figure 13: *Logistic Regression Methodology Flow Diagram*

3.2.7 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning algorithm greatly used for classification modeling (although it can be used for both regression and classification models). The goal of an SVM algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. The objective of SVM is to find a hyperplane that can maximize the distance between data points for the different classes – has the maximum margin. Hyperplanes – which can be a 1D line, a 2D plane, and so far – are boundaries that separate data points (as seen in ***Figure 14***). The dimension of the hyperplane depends on the number of independent variables, or input features.

Modeling of the support vector machine was carried out using the svm function within the “caret” package of R Studio [79]. The methodology flow of running the Support Vector Machine algorithm can be seen in ***Figure 15***.

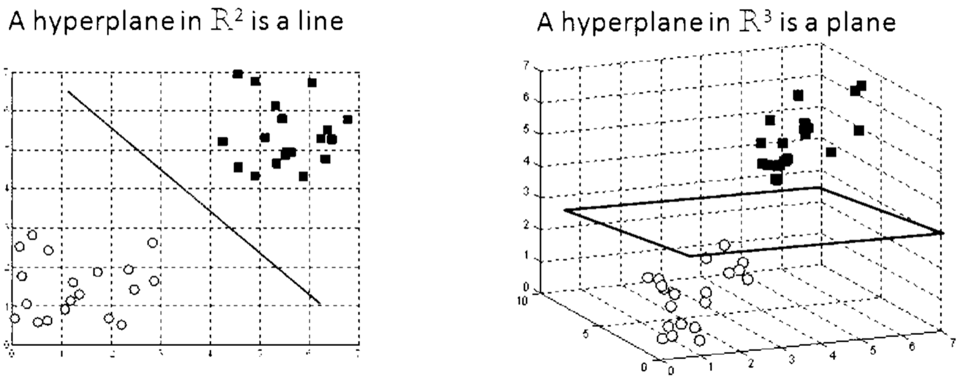


Figure 14: *SVM Hyperplanes*[18]

3.2.6.1 Support Vector Machine Hyperparameters

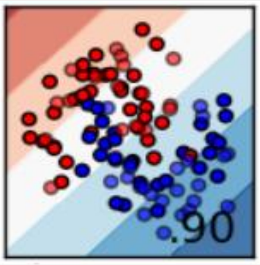
Hyperparameters considered within this thesis for the Support Vector Machine algorithm are kernel, cost, gamma, and degree.

A ***kernel*** is a function that takes data as an input and transforms it into the required higher dimension. Kernel functions only calculation relationships as if they are in the higher dimension, they don’t actually do the transformation. This method – called the “Kernel Trick” – allows SVM to go up to an infinite number of dimensions, thus allowing SVM to work effectively in high dimensional spaces (many features). Although there are many more types of kernels, for this paper, the four most commonly used kernels were implemented in the SVM models (see ***Table 11***). Linear Kernels will use a linear hyperplane (a line in the case of 2D data) while Radial Kernels (RBF), Polynomial Kernels, and Sigmoid Kernels use a nonlinear hyper-plane.

[18] Image courtesy towardsdatascience.com

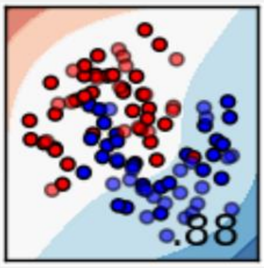
Table 11: SVM Kernel Functions

1. Linear Kernel



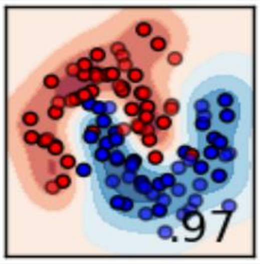
Formula: $u'v$
Hyperparameter: Cost

2. Polynomial Kernel



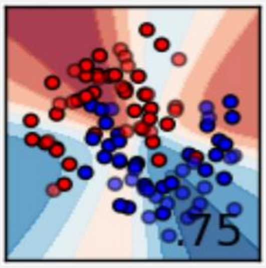
Formula: $(\gamma u'v + coef0)^{degree}$
Hyperparameters: Cost, Gamma, Degree

3. Radial Basis Function (RBF) Kernel



Formula: $e^{(-\gamma|u-v|^2)}$
Hyperparameters: Cost, Gamma

4. Sigmoid Kernel



Formula: $\tanh(\gamma u'v + coef0)$
Hyperparameters: Cost, Gamma

Cost or C is a parameter signifying penalty of the error term . C is a parameter that controls the tradeoff between correctly classifying data points and having a smooth hyperplane boundary. The cost parameter is used across all Kernel Functions. With a default value of 1, as C increases the penalty for non-separable points increases – leading to overfitting. A low value of C could then lead to underfitting and an inaccurate model. For this thesis cost was varied between 0.1 – 2 by a step of 0.25.

Gamma is a parameter used with nonlinear hyperplanes – Polynomial, Radial, and Sigmoid Kernels. The higher the gamma value, the most exact the model tries to fit the dataset. Very high gamma values would then lead to overfitting. For this thesis, gamma was varied between 2^-1 to 2^1.

Degree is a parameter used specifically with Polynomial Kernels. It represents the degree of the polynomial that is used to find the hyperplane that separates the data points. When degree = 1, the Polynomial Kernel is the same as the Linear Kernel. Increasing the degree increases the dimension of the polynomial, thus increasing the time it takes to run the model. For this thesis, degree was varied between 1 – 5 by a step of 1.

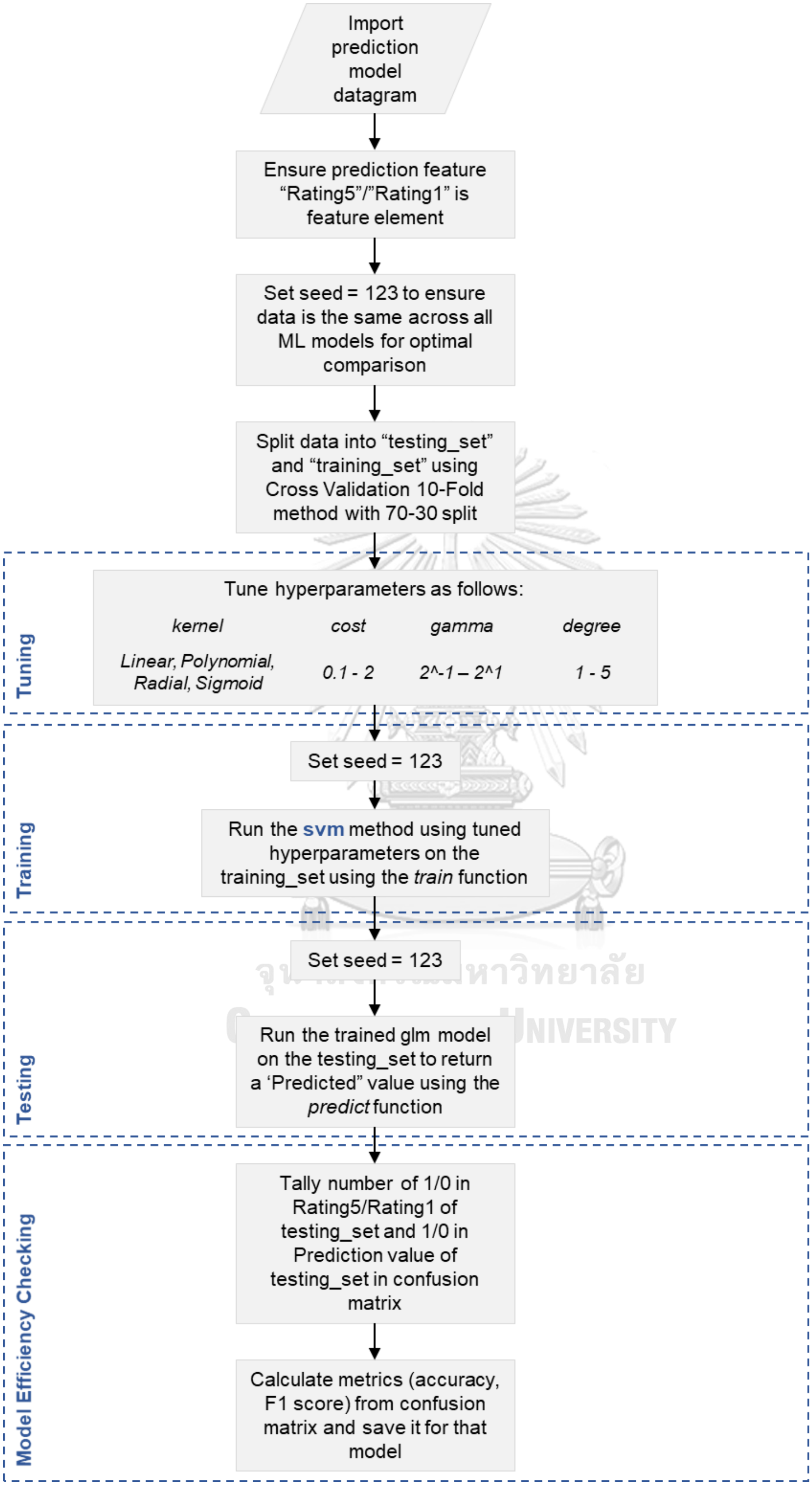


Figure 15: *Support Vector Machine Methodology Flow Diagram*

3.2.8 Random Forest

Random Forest, one of the most popular tools for classification models, is a supervised machine learning algorithm consisting of a large number of individual decision trees, just as its name suggests. Simply put, the random forest algorithm creates numerous decision trees from randomly selected variables within a sample of the dataset. It then collects the prediction from each tree to form the final prediction of the model (as seen in ***Error! Reference source not found.***). A plethora of decision trees are created and predictions are collected, where the highest prediction outcome then becomes the model's final prediction. This is where Random Forest outshines decision trees, through the "wisdom of crowds". The methodology flow of running the Random Forest algorithm can be seen in ***Figure 17***.

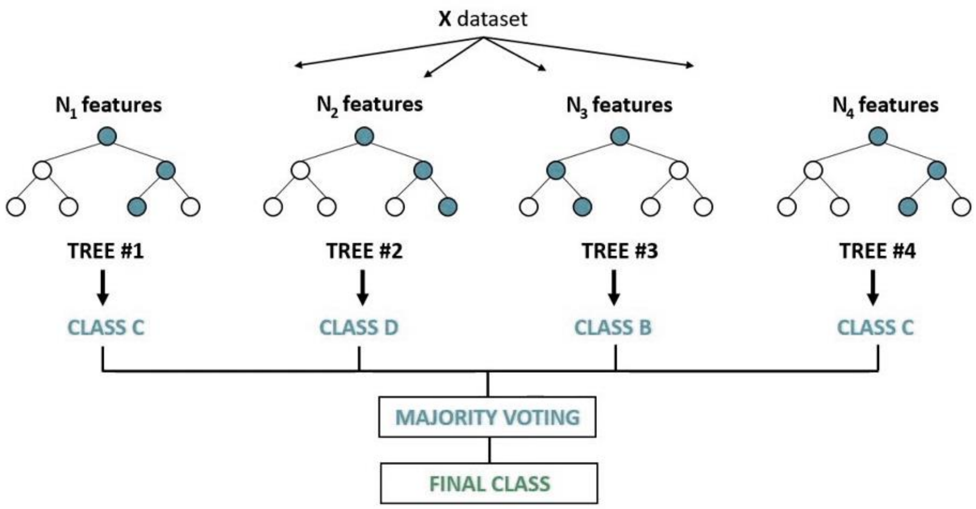


Figure 16: *Random Forest Prediction Collection*[19]

3.2.8.1 Random Forest Hyperparameters

Hyperparameters considered within this thesis for the Random Forest algorithm are ntree and mtry.

Ntree is a hyperparameter that specifies the number of trees within a Random Forest model. The number of trees within the model needs to be relatively large, in order to effectuate the "wisdom of crowds" and stabilize the error rate. The default value for this parameter is 500. The larger the number of trees, the more robust the model becomes. The tradeoff, however, is that the computational time of the model increases in a linear fashion along with the increase in ntree. For this thesis, ntree was varied between 500 – 2000, by a step of 250.

Mtry is a parameter that specifies the split-variable, the number variables sampled at each split of the tree. Mtry balances balance the tradeoff between tree correlation with predictive strength. The default value for mtry is 3 or the square root of the number of variables in the model. For this theis, mtry was varied between 1 – 10, by a step of 1. The hyperparameters ntree and mtry were tuned within these pre-defined ranges to find the optimal value that resulted in the lowest error rate.

[19] Image Courtesy TechTour

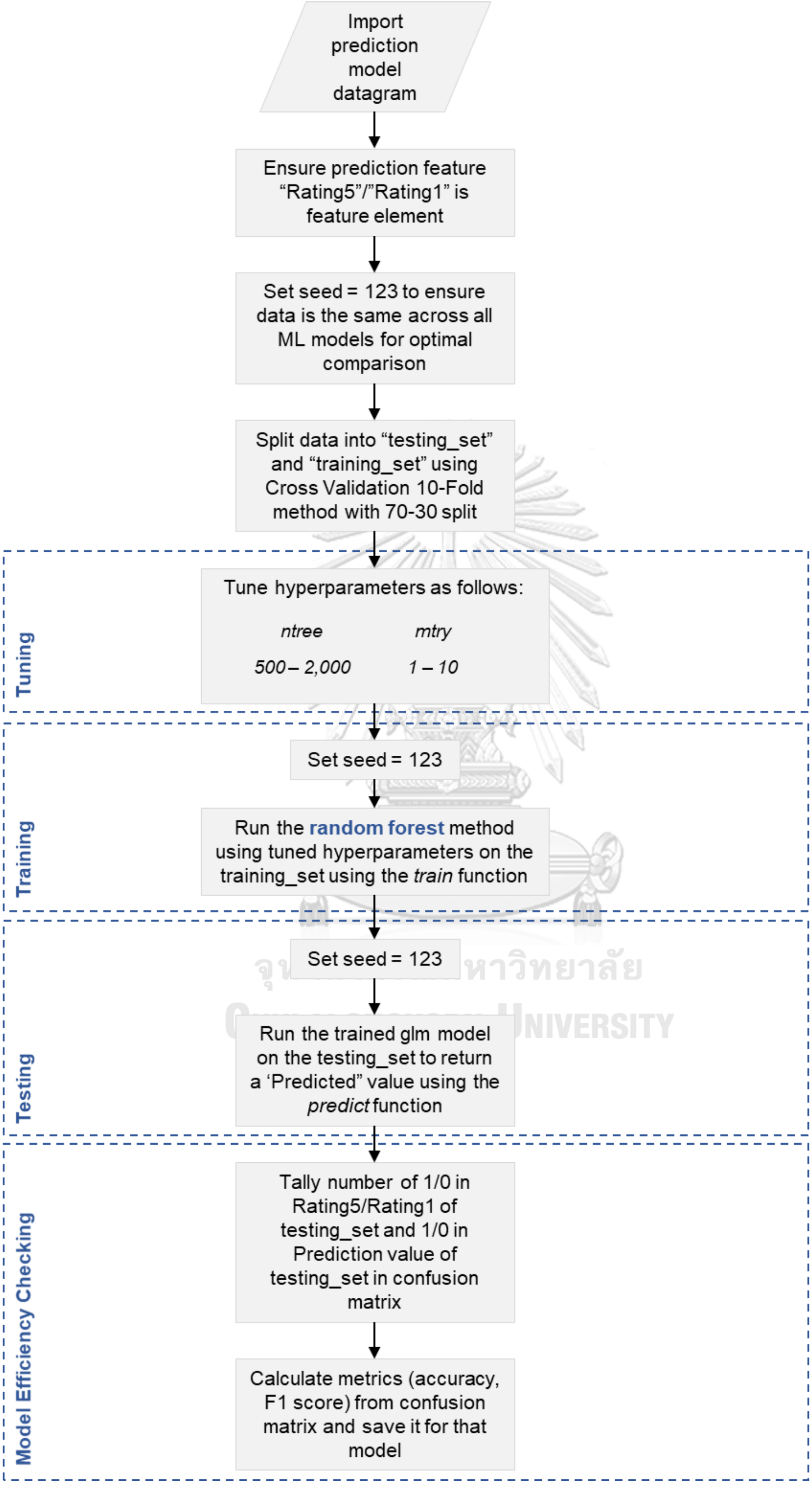


Figure 17: *Random Forest Methodology Flow Diagram*

Chapter 4: Result and Discussion

4.1 Preliminary Analysis Results

4.1.1 Insights on Tourist Preferences via Descriptive Statistics

4.1.1.1 Insights Driven by Proportions of Reviews

For the sake of discovering novel insights on consumer preferences, the proportion of reviews across different feature categories was examined. Of all the features of the dataset, the two most interesting features to examine together were Tour/Activity Type against Origin (as seen in ***Figure 18***).

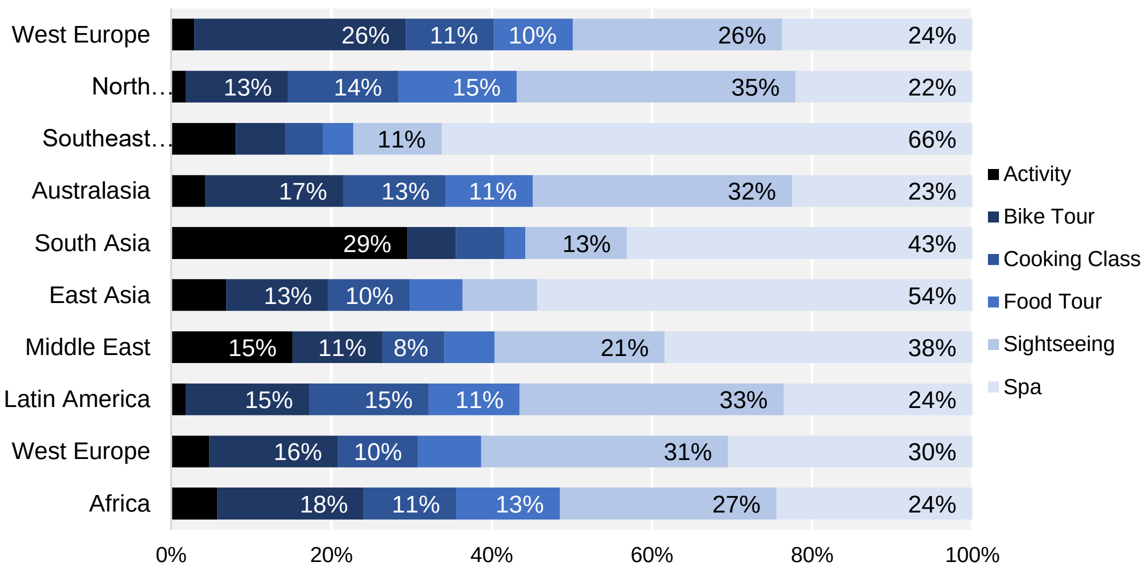


Figure 18: *Percent of Collected Reviews per Tour/Activity Type by Origin*

Through a simple visual inspection, several key insights could be drawn regarding consumer preferences:

- "Spa" is preferred by Asian countries
  Southeast Asia (60%) and East Asia (54%) compared to 20-30% from other origins
- "Sight Seeing" is greatly disfavored by Asia countries
  Southeast Asia (11%) and East Asia (9%) compared to 20-30% from other origins
- "Bike Tour" is preferred by western countries
  Specifically West Europe (20%), Africa (18%), and Australasia (17%) compared to Asian countries (<10%)
- "Activity" is preferred by middle eastern and surrounding countries
  South Asia (29%) and Middle East (15%) compared to ~5% from other origins

4.1.1.2 Insights Driven by Review Ratings

An effective method of developing insights was to look at review ratings and examine how they were across different feature categories and over time. ***Figure 19*** plotted out the average review ratings across different Tour/Activity and Origin categories. From the plot, it can be seen that there is not much disparity of review ratings across Origin categories. However, there is a clear distinction of review ratings across Tour/Activity categories, where “Bike Tour”, “Cooking Class”, and “Food Tour” consistently had the highest average ratings. “Activity”, on the other hand, had the most varied as well as the lowest ratings.

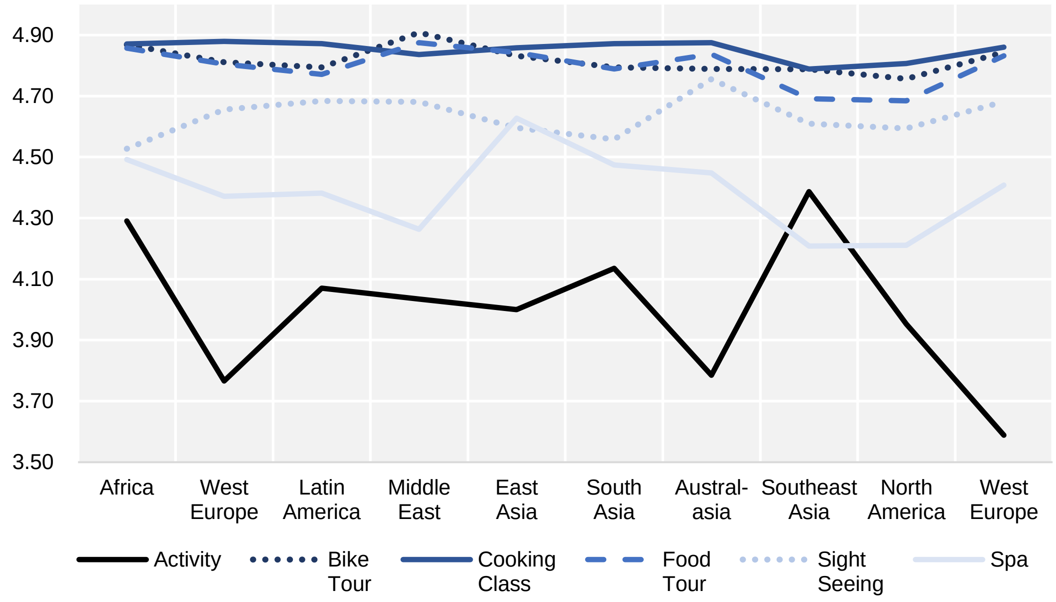


Figure 19: *Average Star-Rating per Tour/Activity Type Across Origins*

Playing off of the insights drawn from ***Figure 19***, ***Figure 20*** was a plot attempting to further inspect the disparity of ratings across Tour/Activity categories, by looking at the distribution of ratings. Given that review ratings were discrete data points, constructing a box-plot did not yield any useful results. Instead, a simple plot of mean, standard deviation, and a distribution of 2SD (±1 SD from the mean) was used as a proxy of viewing the distribution of the data. From the plot, we could be concluded that all categories had an average review rating of over 4.0. “Activity” and “Spa” had the most varied ratings, with standard deviations greater than 1.0.

Figure 20: *Distribution of Star-Rating per Tour/Activity Type*

Figure 20 also revealed an interesting piece of information that warrantied further analysis. *All categories had an average review rating of over 4.0.* The high average review ratings indicated a bias towards higher-star ratings. In order to confirm the fact, a plot of proportion of review ratings was plotted against Tour/Activity categories (***Figure 21***), and Origin categories (***Figure 22***). From the two plots, it was clear that review ratings were biased (across both features) toward the upper end with most Tours/Activities rated at 4-stars and 5-stars.

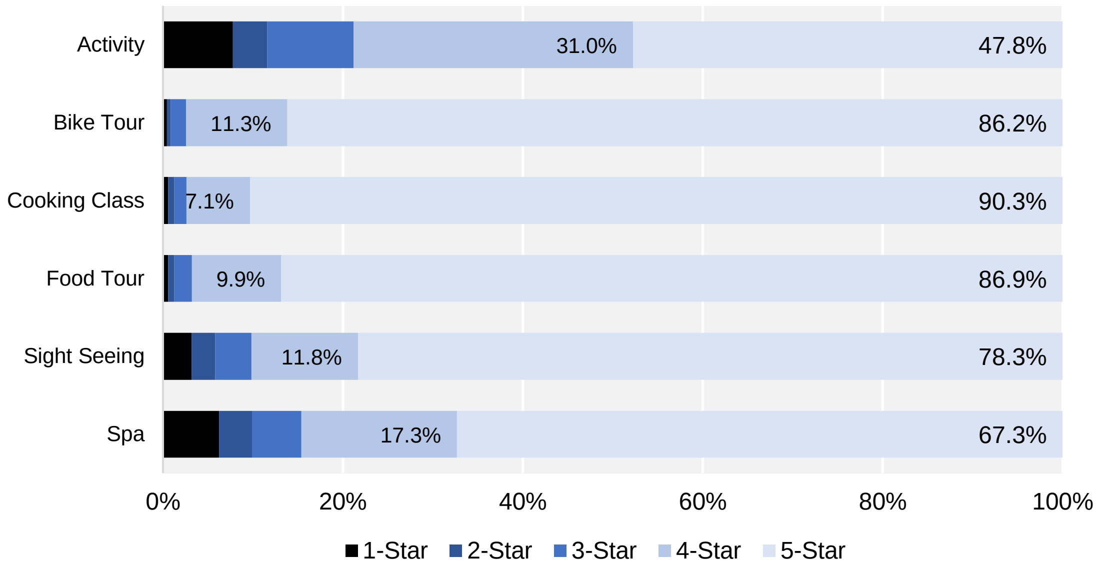


Figure 21: *Percent of Star-Rating per Tour/Activity Type*

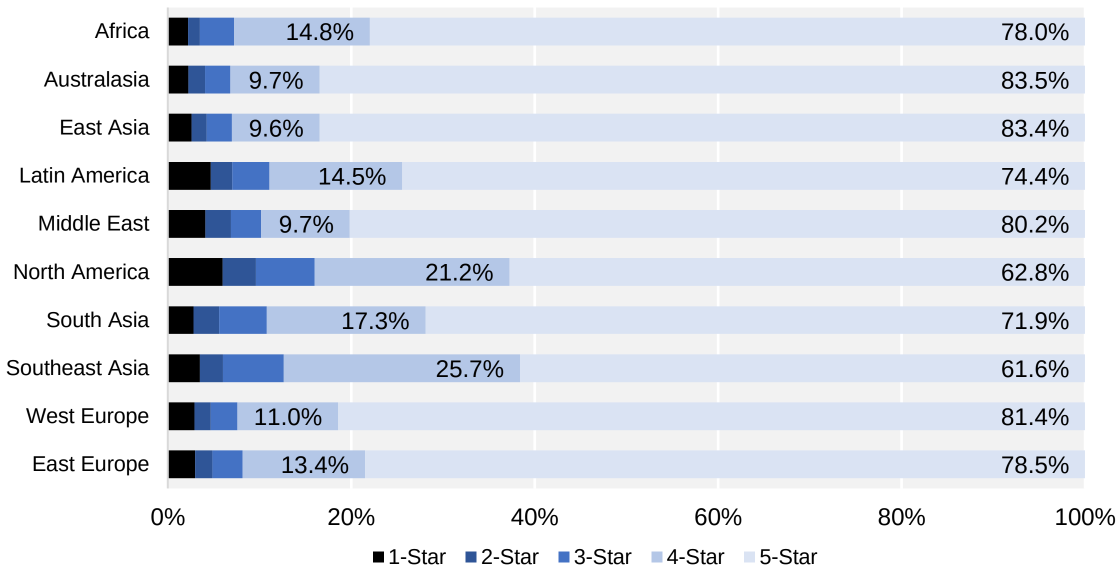


Figure 22: *Percent of Star-Rating per Origin*

To further investigate whether this bias was a new trend or has been this way all along, the proportion of average annual review ratings was plotted over time, from 2010 – 2019 (***Figure 23***). It was found that over time, there was an increasing percent of 5-Star ratings overall and a decreasing percent of mid-level ratings (3- and 4-star), signifying increasing polarization.

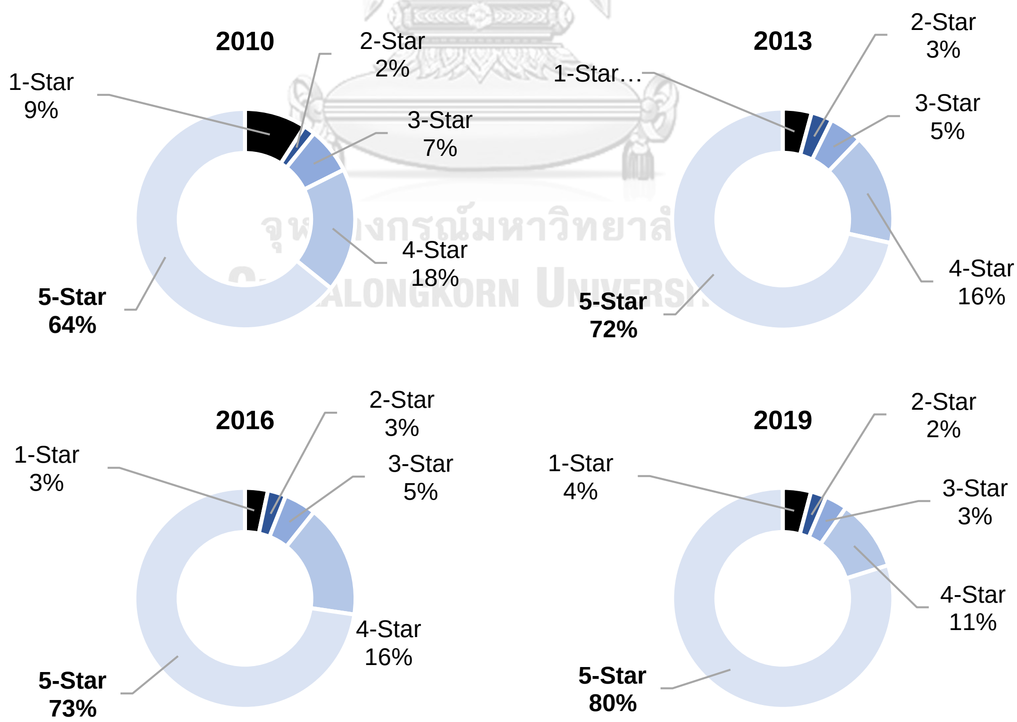


Figure 23: *Percent of Star-Rating Over the Time*

4.1.1.3 Insights Driven by Tour/Activity Prices

Tour/Activity prices were a good feature to examine when looking at consumer satisfaction. For this research, however, analysis on prices was quite limited due to the shortage of data (only 14%[20] of the dataset had price information. From the information that was available, a box-plot was set up to examine the distribution of prices of different Tour/Activity categories (***Figure 24***). From the plot, it could be concluded that lower-priced tours (low average) also had a smaller distribution of prices (such as "Activity" and "Bike Tour"). Higher-priced tours (high average) had a larger distribution of prices (such as "Cooking Class" and "Food Tour"). Sight Seeing had a very wide distribution of prices (probably due to the large variety of offerings – from Hourly Boat Tours to Full-Day Ayutthaya Tours).

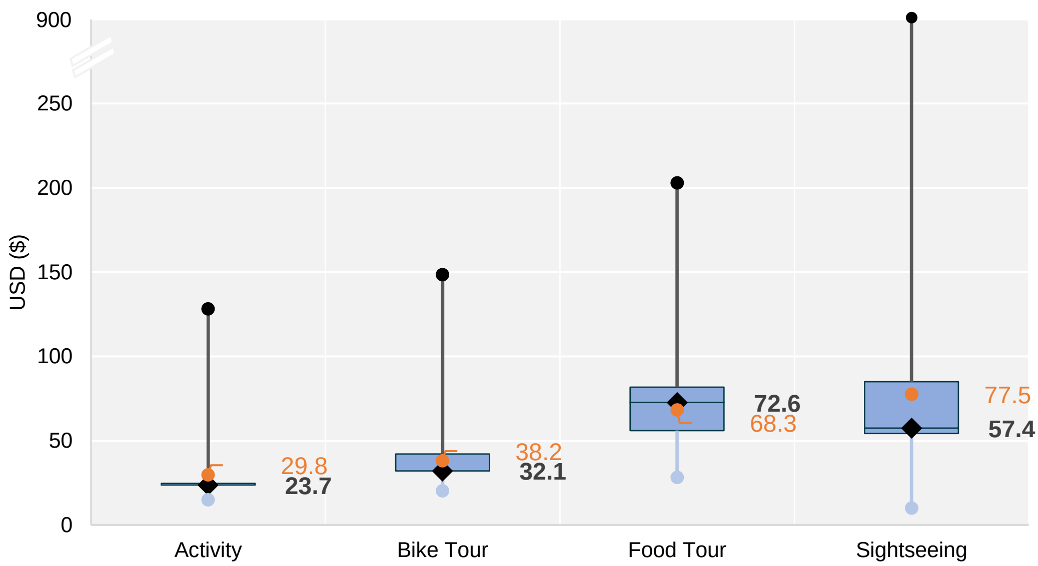


Figure 24: *Boxplot of Prices per Tour/Activity Type*

4.1.2. Insights on Tourist Trends via Descriptive Statistics

4.1.2.1 Number of Reviews Over Time

The number of reviews on Bangkok Tours/Activities had been greatly increasing over the past 9 years, almost exponentially (see ***Figure 25***). At the beginning of the decade in 2012, the number of annual incoming reviews per origin group was in the range of 23 (Africa) – 731 (West Europe). More recently in 2018, the number of annual incoming reviews per origin group grew to the range of 96 (Africa) – 2,313 (West Europe). A paralleled upwards trend was maintained across all origin groups. Additionally, the proportion of reviews per origin was maintained over time; that is, origin groups that hold the maximum proportion of reviews (West Europe, North America, Australasia, and Southeast Asia) have been doing so since the beginning of the decade.

[20] 8,616 of 59,758 are reviews from Tours/Activities that have price information

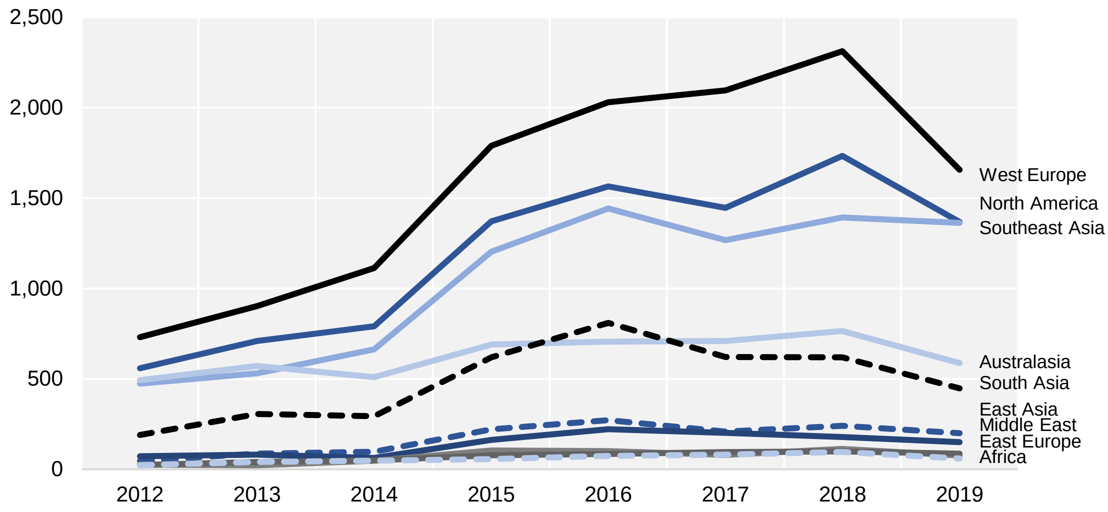


Figure 25: *Number of Reviews Over Time (By Origin)*

4.1.2.2 Review Ratings Over Time

From ***Figure 19***, it was shown that 'Activity' had the lowest as well as the most varied average review ratings. However, when looking at the *trend* of this sub-group, it can actually be seen that the average rating for 'Activity' had been steadily increasing over time (see ***Figure 26***). Starting at an average rating of 3.61 in 2012, ratings of 'Activity' had been steadily increasing to reach a high 4.25 in 2017, then slightly dropping to 4.12 in 2019. Apart from 'Activity', it can also be seen that 'Sight Seeing' ratings had been slightly decreasing over time, starting at a high 4.73 in 2012 to a low 4.52 in 2018, then to slightly increase to 4.59 in 2019.

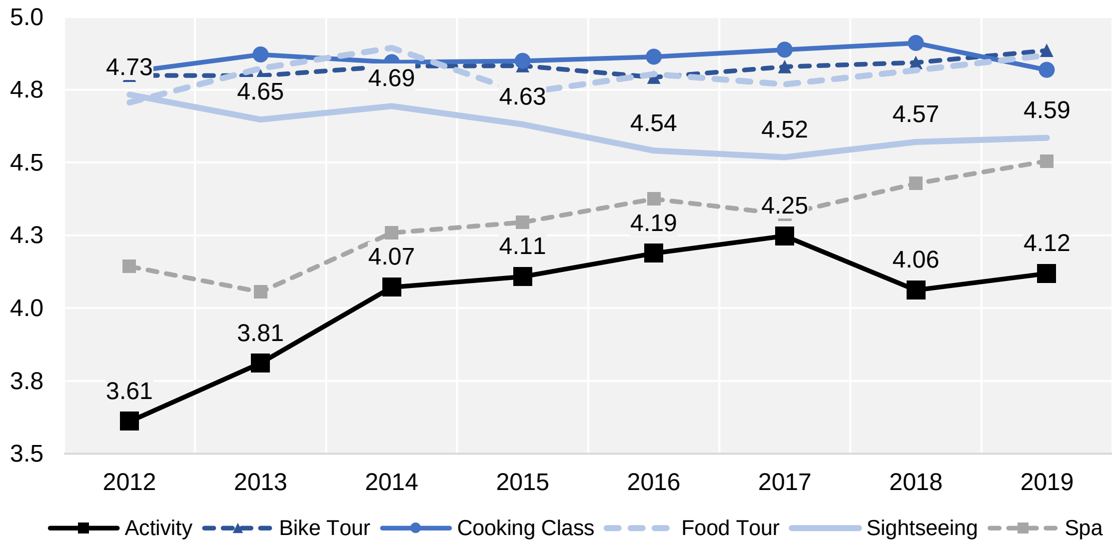


Figure 26: *Review Rating Over Time (By Tour/Activity Type)*

From ***Figure 23***, it was concluded that over time, there was an increase in 5-star reviews. ***Figure 27*** confirmed that fact by showing a similar increase of 5-star reviews. However, visual inspection of ***Figure 27*** additionally revealed that the increase in 5-star reviews was likely due to the decrease in 4-star reviews. 1-star, 2-star, and 3-star reviews remained quite steady and of low proportion comparatively.

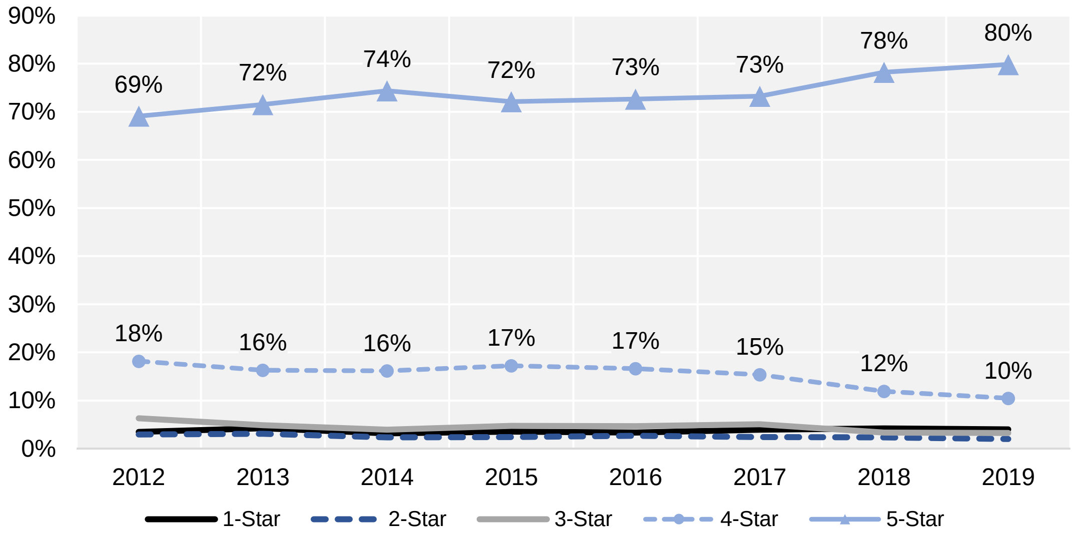


Figure 27: *Proportion of Review Ratings Over Time*

4.1.2.3 Tour/Activity Preference Over Time

In this thesis, Tour/Activity preferences were proxied by the proportion of reviews within the category (carried out in the same way as in ***Figure 18***). Within this section, the proportion of Tour/Activity categories was plotted over time (see ***Figure 28***). The plot revealed an increasing preference for 'Spa' – from 27% in 2013 to 42% in 2016, then slightly dropping to 40% in 2019. Within recent years, it could also be seen that the preference for 'Bike Tour' was decreasing – dropping from 15% in 2017 to a low 6% in 2019. 'Food Tour', however, seemed to be hiking up in terms of preference – increasing from a low 3% in 2013 to an ultimate high 13% in 2019.

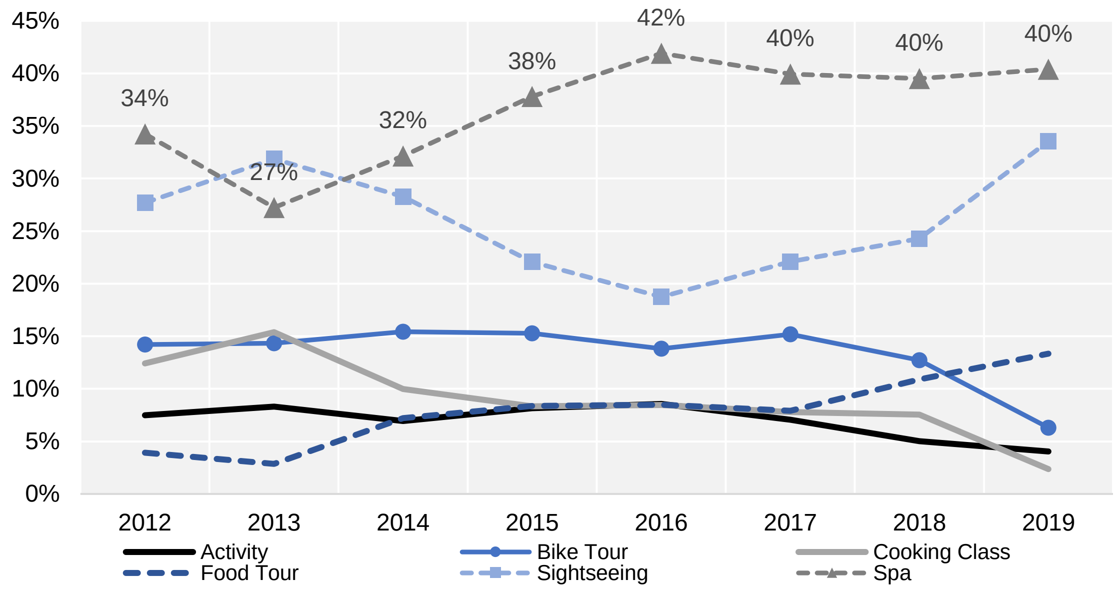


Figure 28: *Proportion of Tour/Activity Reviews Over Time*

4.1.3 Insights on Tourist Associations via Association Rules Mining

4.1.3.1 Association Insights

Preliminary association insights were drawn from the data as prepared per ***Figure 5***. The data was prepared to find the most common Tour/Activity combinations by counting the frequency of each combination. It was then plotted as per ***Figure 29***[21] to find that the most common combination of Tours/Activities was repeated “Spa”, repeated “Sight Seeing”, “Activity” with “Spa”, and “Food Tour” with “Sight Seeing” (See ***Appendix 5*** for full list of combinations).

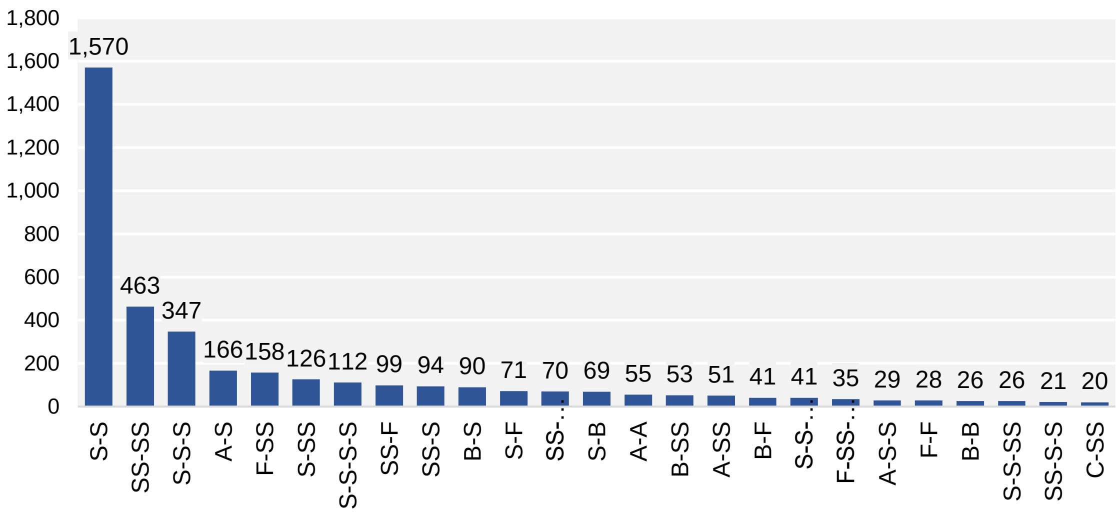


Figure 29: *Frequent Tour/Activity Combinations*

[21] Only combinations with frequency >=20 shown; 175 other combinations exist with frequency 0-19

4.1.3.2 Association Rules Mining: MBA Insights

From the Market Basket Analysis (as explained in ***Section 3.1.3.1*)**, 16 Association Rules were found (as seen in ***Table 12***), which could also be visualized as seen in ***Figure 30***. The Association Rules reflected similar insights to ***Figure 29***, where repeated "Spa", repeated "Sight Seeing", "Activity" with "Spa", and "Food Tour" with "Sight Seeing" were the associations/combinations with the highest probability of occurring.

Table 12: *Association Rules*

	lhs		rhs	support	confidence	lift	count
[78]	{}	=>	{Spa, Spa}	0.3591	0.3591	1	1570
[2]	{}	=>	{Sight Seeing, Sight Seeing}	0.1059	0.1059	1	463
[3]	{}	=>	{Spa, Spa, Spa}	0.0794	0.0794	1	347
[4]	{}	=>	{Activity, Spa}	0.038	0.038	1	166
[5]	{}	=>	{Food Tour, Sight Seeing}	0.0361	0.0361	1	158
[6]	{}	=>	{Spa, Sight Seeing}	0.0288	0.0288	1	126
[7]	{}	=>	{Spa, Spa, Spa, Spa}	0.0256	0.0256	1	112
[8]	{}	=>	{Sight Seeing, Food Tour}	0.0226	0.0226	1	99
[9]	{}	=>	{Sight Seeing, Spa}	0.0215	0.0215	1	94
[10]	{}	=>	{Bike Tour, Spa}	0.0206	0.0206	1	90
[11]	{}	=>	{Spa, Food Tour}	0.0165	0.0165	1	72
[12]	{}	=>	{Sight Seeing, Sight Seeing. Sight Seeing}	0.016	0.016	1	70
[13]	{}	=>	{Spa, Bike Tour}	0.0158	0.0158	1	69
[14]	{}	=>	{Activity, Activity}	0.0126	0.0126	1	55
[15]	{}	=>	{Bike Tour, Sight Seeing}	0.0121	0.0121	1	53
[16]	{}	=>	{Activity, Sight Seeing}	0.0117	0.0117	1	51

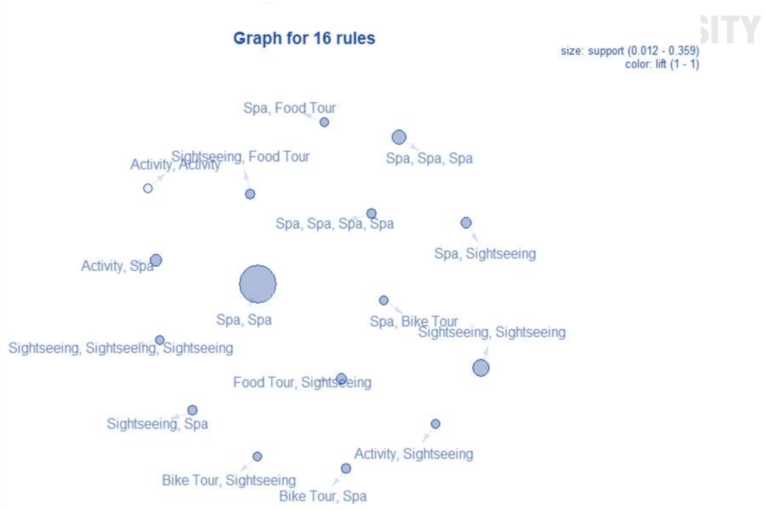


Figure 30: *Association Rules Circle Graph*

4.1.4 Insights on Tourist Sentiment via Sentiment Analysis

A boxplot was built from the findings of the Sentiment Analysis (as explained in ***Section 3.1.4***). Similar to what we would believe, sentiment score was directly proportional to star-ratings across all Tour/Activity categories. Low-star ratings had low sentiment scores and for each increment in star-rating, there is also an increment in average sentiment score (as seen in ***Figure 31***). "Activity" reviews had the greatest overall distribution of sentiment scores as well as the lowest average sentiment scores for their 1-star ratings.

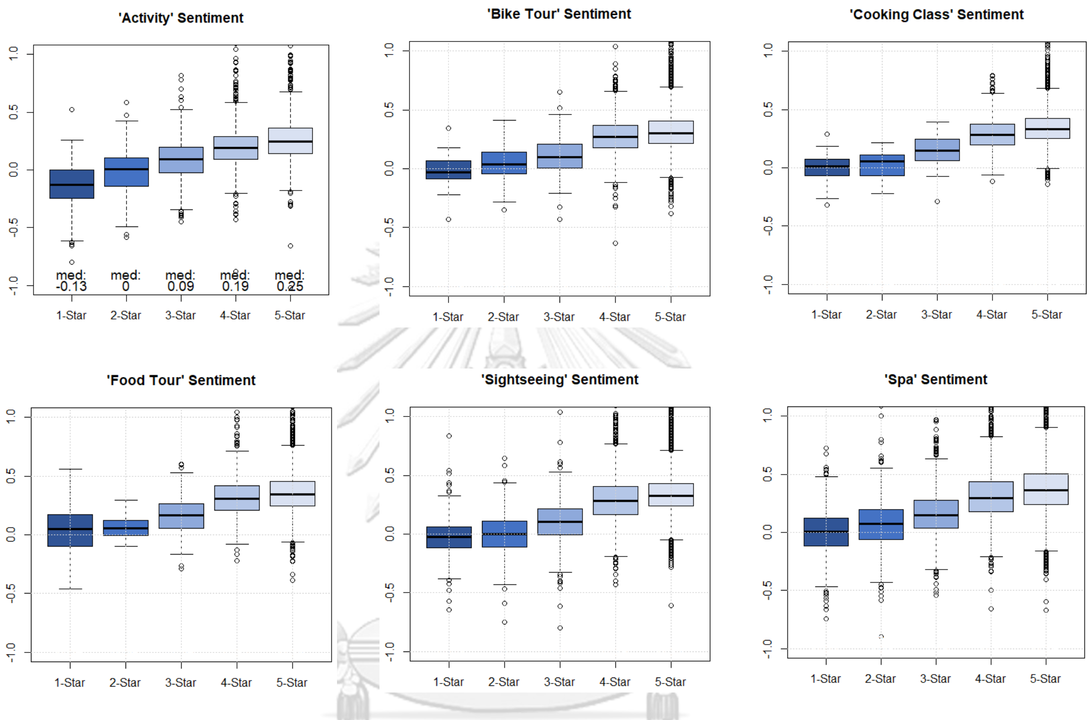


Figure 31: *Boxplot of Sentiment Score by Tour/Activity Type*

4.1.5 Insights on Tourist Focus via Natural Language Processing

Through NLP and Frequency Word Count, the top 20 most frequently occurring words for the 12 sub-categories were found (6 Tour/Activity Types x 2 Levels of Satisfaction – 5-star & 1-star) – as seen in ***Appendix 6***. Key insights could be drawn on what consumers focused on by examining these high-frequency words.

3.2.5.1 'Activity' Insights from NLP

From words such as "animals", "safari", and "zoo" that were mentioned in both 1-star and 5-star reviews, it could be assumed that activities related to zoos and safaris were very popular activities in Bangkok (***Figure 32***). The frequent mention of "kids" positively indicated that such activities were great for kids and families. The frequent mention of "cages", "conditions", and "sad" negatively indicated that tourists were dissatisfied with the upkeep of animals within these zoos and safaris.

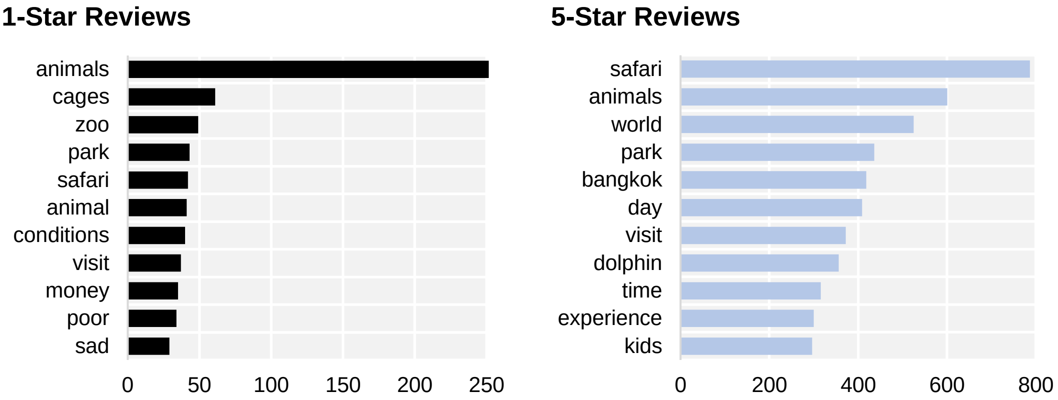


Figure 32: *'Activity' Most Frequent Words (1-Star & 5-Star Reviews)*

3.2.5.2 'Bike Tour' Insights from NLP

Upon first glance of the frequent words (***Figure 33***), it can be seen that "guide" is mentioned both positively and negatively – showing that the service provided by guides highly influenced whether tourists were satisfied or dissatisfied with the tour. Similarly, the mention of "time" both positively and negatively indicated that the time allotted for the tour also highly affected customer satisfaction. Satisfied customers often mentioned "recommend", indicating that the tour was so good they'd recommend it further.

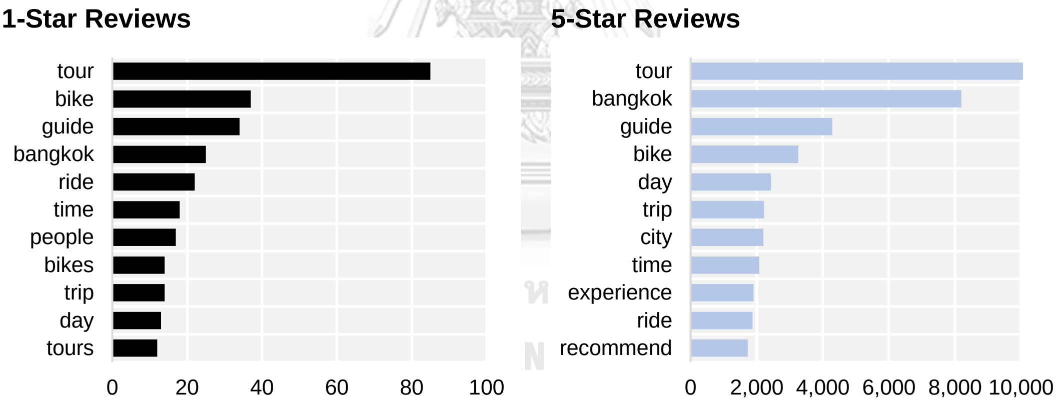


Figure 33: *'Bike Tour' Most Frequent Words (1-Star & 5-Star Reviews)*

3.2.5.3 'Cooking Class' Insights from NLP

For 'Cooking Class', highly frequent words mentioned positively included "chef", "experience" and "ingredients" – indicating that tourist were satisfied when the chef was capable, ingredients were of good quality, and they had an overall nice experience (***Figure 34***). The word "market" was mentioned quite often, both positively and negatively, indicating markets influence tourists both positively and negatively. Certain words, such as "Thai", although mentioned both ways, didn't give any further insights apart from the fact that most cooking classes in Bangkok were for Thai cuisine.

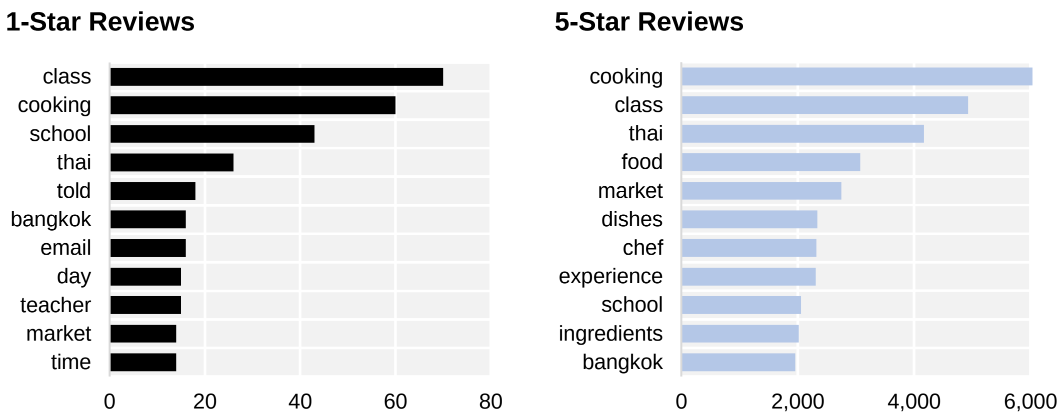


Figure 34: *'Cooking Class' Most Frequent Words (1-Star & 5-Star Reviews)*

3.2.5.4 'Food Tour' Insights from NLP

Similar to 'Bike Tour', the word "guide" was quite frequently mentioned both positively and negatively, indicating a high influence towards tourist satisfaction (***Figure 35***). "Street" and "stops" were words frequently mentioned negatively, indicating some dissatisfaction towards the streets of the food tour and the number of stops. "Tuk" mentioned positively indicated that tourists were fond of tuk-tuks (or 3-wheelers). "Night" mentioned positively indicated that night-time food tours were also positively viewed. Again, like previous categories, the word "recommend" showed up quite often for 5-star ratings.

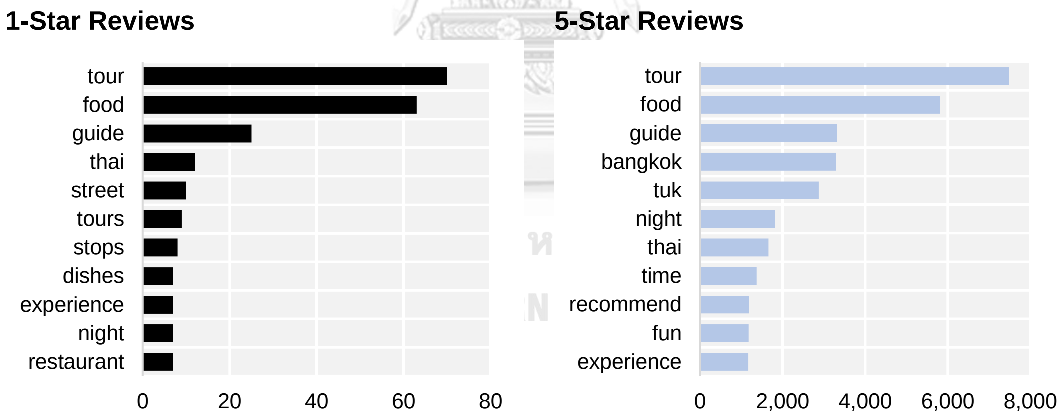


Figure 35: *'Food Tour' Most Frequent Words (1-Star & 5-Star Reviews)*

3.2.5.5 'Sight Seeing' Insights from NLP

Words such as "amazing" and "recommend" were frequently mentioned in 'Sight Seeing' 5-star ratings, solidifying the fact that tourist who were rating 5 stars are satisfied and happy with the sight seeing tour (***Figure 36***). "Market" mentioned positively indicated there is a fondness for markets. Similar to previous tours, "time" and "guide" was mentioned again both positively and negatively. This reinforced the fact that, no matter what the tour type, guides were a crucial factor towards customer satisfaction.

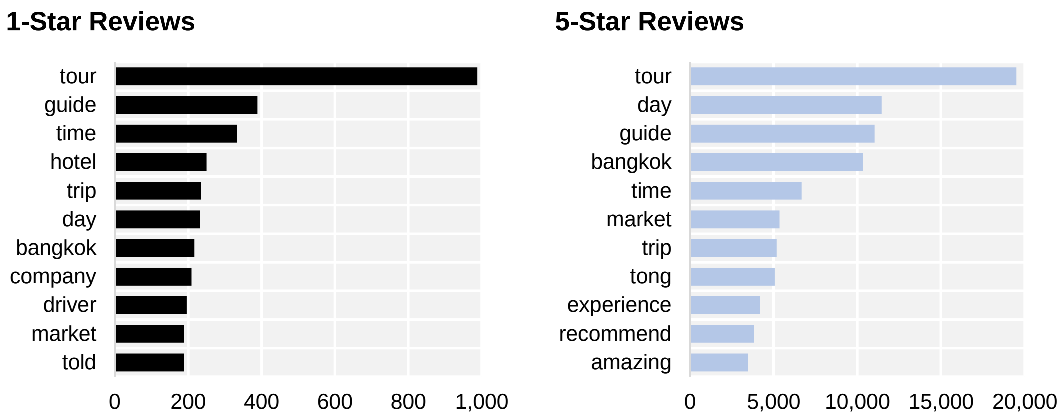


Figure 36: *'Sight Seeing' Most Frequent Words (1-Star & 5-Star Reviews)*

3.2.5.6 'Spa' Insights from NLP

"Service" was a key influencer of customer satisfaction for 'Spas' – being mentioned frequently both positively and negatively (***Figure 37***). Customers were oftentimes dissatisfied by foot massages, indicated by the negative mention of "foot". Customers were satisfied with spas when the staff was nice and professional and the area was clean – indicated by the mention of the words "staff", "clean", "professional", and "nice". "Experience" was mentioned both positively and negatively, indicating that a good or bad overall experience influenced a high or low rating.

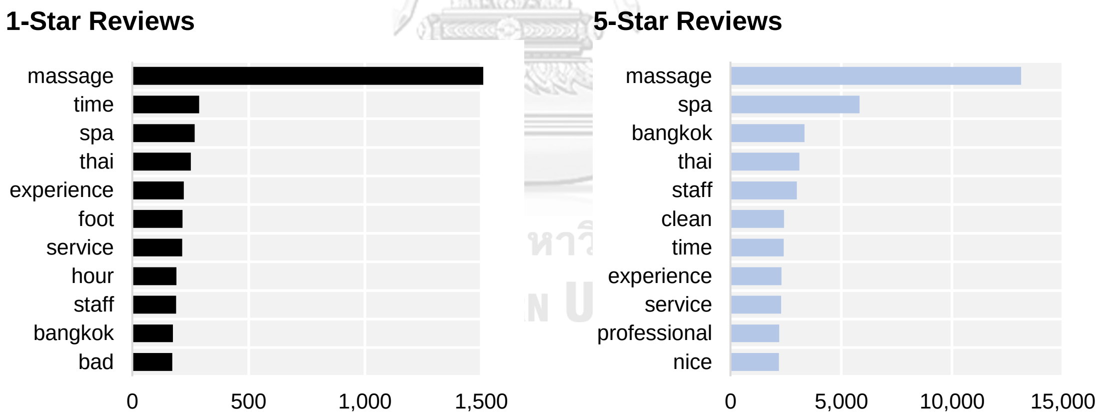


Figure 37: *'Spa' Most Frequent Words (1-Star & 5-Star Reviews)*

4.2 Machine Learning Results

4.2.1 Logistic Regression Results

4.2.1.1 Logistic Regression Effectiveness for Predicting 5-Star Reviews

For this research, the effectiveness of the prediction model was focused on *F1-score* and *accuracy*, as mentioned in ***Section 3.1.6.4***. The prediction metrics (including accuracy, precision, recall, specificity, and F1-score) for each model can be seen in following tables: ***Table 13*** (Activity), ***Table 14*** (Bike Tour), ***Table 15*** (Cooking Class), ***Table 16*** (Food Tour),

Table 17 (Sight Seeing), and ***Table 18*** (Spa). The F1-score for these logistic regression models range from 62.9% - 80.1% and have an average score of 73.8% and the accuracy range from 62.7% – 74.5% with an average score of 67.8%, indicating these models did an adequate job at predicting 5-star ratings for reviews. However, there does seem to be room for improvement in order to increase accuracy and F1 Score.

Table 13: *LR 'Activity' 5-Star Model Effectiveness*

Activity 5-Star Prediction		
	No	Yes
No	399	205
Yes	200	343

F1 Score	**62.9%**
Accuracy	**64.7%**
Precision	63.2%
Recall	62.6%
Specificity	66.6%

Table 14: *LR 'Bike Tour' 5-Star Model Effectiveness*

Bike Tour 5-Star Prediction		
	No	Yes
No	106	84
Yes	224	412

F1 Score	**72.8%**
Accuracy	**62.7%**
Precision	64.8%
Recall	83.1%
Specificity	32.1%

Table 15: *LR 'Cooking Class' 5-Star Model Effectiveness*

Cooking Class 5-Star Prediction		
	No	Yes
No	73	47
Yes	69	166

F1 Score	**74.1%**
Accuracy	**67.3%**
Precision	70.6%
Recall	77.9%
Specificity	51.4%

Table 16: *LR 'Food Tour' 5-Star Model Effectiveness*

Food Tour 5-Star Prediction		
	No	Yes
No	84	44
Yes	128	274

F1 Score	**76.1%**
Accuracy	**67.5%**
Precision	68.2%
Recall	86.2%
Specificity	39.6%

Table 17: *LR 'Sight Seeing' 5-Star Model Effectiveness*

Sight Seeing 5-Star Prediction		
	No	Yes
No	594	221
Yes	434	1321

F1 Score	**80.1%**
Accuracy	**74.5%**
Precision	75.3%
Recall	85.7%
Specificity	57.8%

Table 18: *LR 'Spa' 5-Star Model Effectiveness*

Spa 5-Star Prediction		
	No	Yes
No	1092	549
Yes	1048	2661

F1 Score	**76.9%**
Accuracy	**70.1%**
Precision	71.7%
Recall	82.9%
Specificity	51.0%

4.2.1.2 Logistic Regression Results for Predicting 5-Star Reviews

Results from the Logistic Regression model showed that "sentiment" was the most significant factor in predicting 5-star ratings for reviews –having very low p-values and high estimates across all 6 models. The p-value for "sentiment" was ~0.00 for all 6 models and estimates ranged from 2.57 ('Bike Tour') to 4.73 ('Sight Seeing'). Following that, Origin and Frequent Words were either a hit or a miss, some having p-values as low as 0 and some as high as 0.9. The result of the first model, predicting 5-Star ratings for Activities is shown in ***Table 19***. The full result of the logistic regression can be seen in ***Appendix 7***.

Table 19: *Logistic Regression Results ('Activity' 5-Star Model)*

Activity 5-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.3706	0.099611	-3.72048	0.000199
2	Sentiment	**3.712192**	0.200803	18.48672	**2.64E-76**
3	West_Europe	-0.39979	0.137886	-2.89939	0.003739
4	North_America	-0.29948	0.179207	-1.67112	0.094698
5	Southeast_Asia	-0.61643	0.115749	-5.32556	1.01E-07
6	Australasia	-0.56015	0.163548	-3.42499	0.000615
7	South_Asia	0.000545	0.102971	0.005296	0.995774
8	East_Asia	-0.36977	0.232646	-1.58942	0.111966
9	Middle_East	-0.51291	0.18125	-2.82983	0.004657
10	Latin_America	-0.44838	0.653569	-0.68605	0.492683
11	East_Europe	0.528938	0.413901	1.277934	0.201273
12	Africa	-0.19514	0.396822	-0.49176	0.622891
13	W1	-0.19289	0.094676	-2.03731	0.041619
14	W2	-0.22029	0.154975	-1.42143	0.155193
15	W3	0.198222	0.104282	1.90082	0.057326
16	W4	-0.21538	0.09364	-2.3001	0.021443
17	W5	0.146377	0.084175	1.738958	0.082042
18	W6	-0.14091	0.084118	-1.67511	0.093913

19	W7	-0.03935	0.09343	-0.42112	0.673667
20	W8	0.332076	0.103937	3.194978	0.001398
21	W9	-0.3051	0.10535	-2.89609	0.003778
22	W10	-0.17949	0.148476	-1.20891	0.226698

4.2.1.3 Logistic Regression Effectiveness for Predicting 1-Star Reviews

Similar to the previous 6 models, the metric of focus was the F1-score, which is considered useful for datasets that are not completely balanced. The effectiveness of each model can be seen in following tables: ***Table 20*** (Activity), ***Table 21*** (Sight Seeing), and ***Table 22*** (Spa). As mentioned in ***Section 3.2.4***, models predicting 1-Star ratings for Bike Tour, Cooking Class, and Food Tour have a limited number of data points, causing inaccuracies in the model. Thus, the results for those models are not considered for further analyses and insights.

The average F1 Score across all 1-Star prediction models is 80.6% and the average accuracy across all is 80.93%, indicating that logistic regression does an adequate job of predicting 1-Star reviews – an even better job than 5-Star prediction models.

Table 20: *LR 'Activity' 1-Star Model Effectiveness*

Activity 1-Star Prediction		
	No	Yes
No	121	18
Yes	12	71

F1 Score	**82.6%**
Accuracy	**86.5%**
Precision	85.5%
Recall	79.8%
Specificity	91.0%

Table 21: *LR 'Sight Seeing' 1-Star Model Effectiveness*

Sight Seeing 1-Star Prediction		
	No	Yes
No	202	28
Yes	24	122

F1 Score	**82.4%**
Accuracy	**86.2%**
Precision	83.6%
Recall	81.3%
Specificity	89.4%

Table 22: *LR 'Spa' 1-Star Model Effectiveness*

Spa 1-Star Prediction		
	No	Yes
No	1092	549
Yes	1048	2661

F1 Score	**76.9%**
Accuracy	**70.1%**
Precision	71.7%
Recall	82.9%
Specificity	51.0%

4.2.1.4 Logistic Regression Results for Predicting 1-Star Reviews

Similar to the previous section, “sentiment” was the most significant factor in predicting 1-star ratings of reviews as well. Like the previous model, the p-value for “sentiment” was ~0.00 for all 6 models and estimates were even higher than models 1-6, ranging from -15.35 (‘Cooking Class’) to -6.9 (‘Spa’). Just like the previous models, the independent variables for Origin and Frequent Words were a mix of significant and insignificant values, with p-values ranging from as low as 0.00 to as high as 0.95. The result of the first model, Predicting 1-Star ratings for Activities is shown in ***Table 23***. The full result of the logistic regression can be seen in ***Appendix 8***. This shows that further improvements would be required within these features to make them good predictors of review ratings.

Table 23: *Logistic Regression Results (‘Activity’ 1-Star Model)*

Activity 1-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.00126	0.30153	-0.00416	0.996678
2	Sentiment	-10.201	0.831952	-12.2616	1.46E-34
3	West_Europe	0.428447	0.365882	1.170996	0.2416
4	North_America	-0.24903	0.545427	-0.45658	0.647971
5	Southeast_Asia	-0.00267	0.370473	-0.00721	0.994249
6	Australasia	0.506637	0.440776	1.149421	0.250382
7	South_Asia	-1.62865	0.424098	-3.84028	0.000123
8	East_Asia	-0.78521	0.861234	-0.91173	0.361913
9	Middle_East	-1.16864	0.698264	-1.67363	0.094203
10	Latin_America	-11.9485	561.1954	-0.02129	0.983013
11	East_Europe	0.256723	1.006203	0.255141	0.798615
12	Africa	-0.7573	1.655686	-0.45739	0.64739
13	W1	0.49436	0.54988	0.899032	0.368636
14	W2	-0.34361	0.346738	-0.99098	0.321695
15	W3	0.004808	0.335369	0.014337	0.988561
16	W4	-0.94504	0.368646	-2.56355	0.010361
17	W5	0.44448	0.540966	0.821641	0.411281
18	W6	0.425774	0.300172	1.418437	0.156063
19	W7	1.024268	0.526613	1.945009	0.051774
20	W8	-0.586	0.40148	-1.45961	0.144398
21	W9	0.214674	0.475054	0.451894	0.651345
22	W10	0.103885	0.322301	0.322323	0.747208

4.2.2 Support Vector Machine Results

4.2.3.1 Support Vector Machine Hyperparameter Tuning

As mentioned in ***Section 3.2.5***, hyperparameters for machine learning algorithms have to be determined prior to training the data. The four hyperparameters that were considered for this research’s Support Vector Machine algorithm were kernel, cost, gamma, and degree. In order to find the best value for each parameter, the tuning methodology using grid search was carried out.

This tuning was run on each of the four chosen kernel functions across all models. The 'Best Performance' metric was compared across all models predicting 5-star reviews and models predicting 1-star reviews in order to find which kernel function would be most suitable for 5-star models and 1-star models.

Table 24: *SVM Classification Error Across Kernel Functions for 5-Star Prediction Models*

Classification Error Across Kernel Functions (5-Star Prediction Models)							
	Activity	Bike Tour	Food Tour	Cooking Class	Sight Seeing	Spa	Average
Linear	0.35	0.395	0.355	0.37	0.425	0.34	0.373
Polynomial	0.35	0.395	0.355	0.35	0.375	0.335	**0.360**
Radial	0.325	0.39	0.355	0.38	0.39	0.35	0.365
Sigmoid	0.375	0.405	0.37	0.385	0.415	0.36	0.385

Table 25: *SVM Classification Error Across Kernel Functions for 1-Star Prediction Models*

Classification Error Across Kernel Functions (1-Star Prediction Models)				
	Activity	Sight Seeing	Spa	Average
Linear	0.13	0.115	0.24	**0.162**
Polynomial	0.185	0.11	0.235	0.177
Radial	0.165	0.35	0.255	0.257
Sigmoid	0.375	0.415	0.36	0.383

From ***Table 24*** and ***Table 25***, it can be seen that the kernel function with the lowest value for 'Best Performance', a.k.a. the lowest classification error for 5-star prediction models is **Polynomial** and the kernel function with the lowest classification error for 1-star prediction models is **Linear**. Thus, those kernel functions were chosen for the respective prediction models. The cost, gamma, and degree parameters used for each model are tuned to the optimal value for each model and used as tuned (as seen in ***Table 26*** and ***Table 27***).

Table 26: *Hyperparameters for SVM Polynomial Kernel (5-Star Prediction Models)*

Hyperparameters Used for SVM Polynomial Kernel (5-Star Models)							
	Tuning Range	Activity	Bike Tour	Food Tour	Cooking Class	Sight Seeing	Spa
Cost	0.1 – 2	0.6	1.35	1.35	0.1	0.1	0.85
Gamma	0.5 - 2	1	1	1	0.5	0.5	0.5
Degree	1 - 5	1	1	1	1	3	1

Table 27: *Hyperparameters for SVM Linear Kernel (1-Star Prediction Models)*

Hyperparameter Used for SVM Linear Kernel (1-Star Models)				
	Tuning Range	Activity	Sight Seeing	Spa
Cost	0.1 - 2	0.35	0.1	0.6

4.2.3.2 Support Vector Machine Effectiveness for Predicting 5-Star Reviews

The Support Vector Machine algorithm using a Polynomial Kernel does a decent job in predicting 5-star reviews. As seen in ***Table 28*** (Activity), ***Table 29*** (Bike Tour), ***Table 30*** (Cooking Class), ***Table 31*** (Food Tour), ***Table 32*** (Sight Seeing), and ***Table 33*** (Spa), the F1 Score across all 6 models is at an average of 73.7%, ranging from 62.9% (Activity) to 78.2% (Sight Seeing). The accuracy of the models is at an average of 66.2%, ranging from 60.0% (Food Tour) to 73.4% (Sight Seeing). This shows that the SVM model does a decent job in predicting 5-star reviews, a better job than Logistic Regression.

Table 28: *SVM 'Activity' 5-Star Model Effectiveness*

Activity 5-Star Prediction		
	No	Yes
No	388	200
Yes	211	348

F1 Score	**62.9%**
Accuracy	**64.2%**
Precision	62.3%
Recall	63.5%
Specificity	64.8%

Table 29: *SVM 'Bike Tour' 5-Star Model Effectiveness*

Bike Tour 5-Star Prediction		
	No	Yes
No	56	31
Yes	274	465

F1 Score	**75.3%**
Accuracy	**63.1%**
Precision	62.9%
Recall	93.8%
Specificity	17.0%

Table 30: *SVM 'Cooking Class' 5-Star Model Effectiveness*

Cooking Class 5-Star Prediction		
	No	Yes
No	66	44
Yes	76	169

F1 Score	**73.8%**
Accuracy	**66.2%**
Precision	69.0%
Recall	79.3%
Specificity	46.5%

Table 31: *SVM 'Food Tour' 5-Star Model Effectiveness*

Food Tour 5-Star Prediction		
	No	Yes
No	0	0
Yes	212	318

F1 Score	**75.0%**
Accuracy	**60.0%**
Precision	60.0%
Recall	100.0%
Specificity	0.0%

Table 32: *SVM 'Sight Seeing' 5-Star Model Effectiveness*

Sight Seeing 5-Star Prediction		
	No	Yes
No	663	318
Yes	365	1224

F1 Score	**78.2%**
Accuracy	**73.4%**
Precision	77.0%
Recall	79.4%
Specificity	64.5%

Table 33: *SVM 'Spa' 5-Star Model Effectiveness*

Spa 5-Star Prediction		
	No	Yes
No	1038	496
Yes	1102	2714

F1 Score	**77.3%**
Accuracy	**70.1%**
Precision	71.1%
Recall	84.5%
Specificity	48.5%

An interesting observation to make from the effectiveness data is that models with smaller datasets (Activity, Cooking Class, Bike Tour, and Food Tour with all under 4,000 observations) tend of have lower accuracy and F1 Scores – average Accuracy of 63% and average F1 Score of 72%. Models with larger datasets (Sight Seeing and Spa with around 8,600 and 18,000 observations respectively) have higher accuracy and F1 Scores – average Accuracy of 72% and average F1 Score of 78% (see visualization in ***Figure 38***).

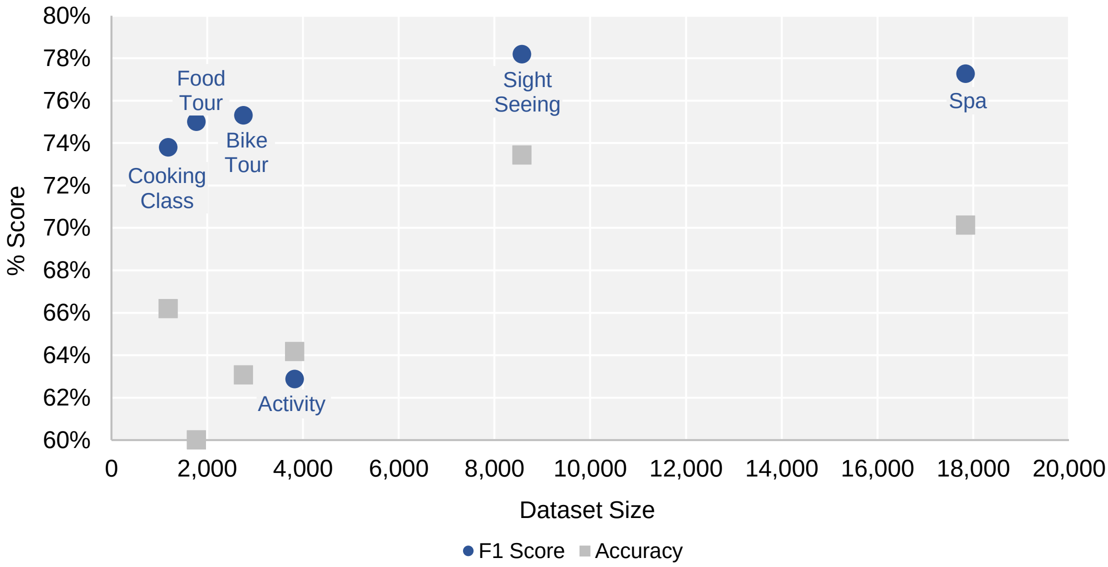


Figure 38: *Accuracy and F1 Score Compared to Dataset Size*

4.2.3.3 Support Vector Machine Effectiveness for Predicting 1-Star Reviews

The Support Vector Machine algorithm using a Linear Kernel does an even better job in predicting 1-star reviews than predicting 5-star reviews. As seen in ***Table 34*** (Activity), and ***Table 35*** (Sight Seeing), and ***Table 36*** (Spa), the F1 Score across all 3 models is at an average of 80.3% (compared to LR predicting 5-star reviews average F1 Score of 73.7%). The accuracy across all 3 models is at an average of 84.9% (compared to LR average accuracy of 80.9%). This shows that the SVM model is more effective than Logistic Regression and SVM predicting 5-star reviews.

Table 34: *SVM 'Activity' 1-Star Model Effectiveness*

Activity 5-Star Prediction		
	No	Yes
No	126	23
Yes	7	66

F1 Score	**81.5%**
Accuracy	**86.5%**
Precision	90.4%
Recall	74.2%
Specificity	94.7%

Table 35: *SVM 'Sight Seeing' 1-Star Model Effectiveness*

Sight Seeing 5-Star Prediction		
	No	Yes
No	205	31
Yes	21	119

F1 Score	**82.1%**
Accuracy	**86.2%**
Precision	85.0%
Recall	79.3%
Specificity	90.7%

Table 36: *SVM 'Spa' 1-Star Model Effectiveness*

Spa 5-Star Prediction		
	No	Yes
No	523	94
Yes	89	314

F1 Score	**77.4%**
Accuracy	**82.1%**
Precision	77.9%
Recall	77.0%
Specificity	85.5%

4.2.3 Random Forest Results

4.2.3.1 Random Forest Hyperparameter Tuning

As mentioned in ***Section 3.2.5***, the hyperparameters that were tuned for the Random Forest algorithm were ntrees and mtry, varying from 500 – 2000 and 1 – 10, respectively. The hyperparameters were tuned across all 12 prediction models, with the average taken over the 5-star prediction models and 1-star prediction models to find the best hyperparameters to predict 5-star reviews and 1-star reviews. The ntree value that resulted in the lowest classification error for predicting 5-star reviews (as seen in ***Table 37***) was 500, which is also the default value. The ntree value that resulted in the lowest classification error for predicting 1-star reviews (as seen in ***Table 38***) was 750. Thus, those values were the ones used in the final Random Forest algorithms for 5-star and 1-star prediction models.

Table 37: *RF Classification Error Across ntrees for 5-Star Prediction Models*

Classification Error Across #Trees (5-Star Prediction Models)							
	Activity	Bike Tour	Food Tour	Cooking Class	Sight Seeing	Spa	Average
500	0.380	0.451	0.372	0.368	0.426	0.375	**0.395**
750	0.449	0.535	0.411	0.426	0.371	0.386	0.430
1000	0.507	0.476	0.493	0.397	0.608	0.493	0.496
1250	0.391	0.461	0.411	0.479	0.387	0.397	0.421
1500	0.520	0.370	0.507	0.543	0.465	0.438	0.474
1750	0.464	0.475	0.427	0.438	0.512	0.368	0.447
2000	0.443	0.548	0.554	0.507	0.465	0.479	0.499

Table 38: *RF Classification Error Across ntrees for 1-Star Prediction Models*

Classification Error Across #Trees (5-Star Prediction Models)				
	Activity	Sight Seeing	Spa	Average
500	0.221	0.314	0.323	0.286
750	0.224	0.143	0.329	**0.232**
1000	0.237	0.306	0.278	0.274
1250	0.173	0.265	0.394	0.277
1500	0.284	0.190	0.250	0.241
1750	0.312	0.127	0.370	0.269
2000	0.143	0.265	0.268	0.225

The mtry value that resulted in the lowest classification error for predicting 5-star reviews (as seen in ***Table 39***) was 3, which again, is also the default value. The mtry value that resulted in the lowest classification error for predicting 1-star reviews (as seen in ***Table 40***) was 4. Thus, those values were the ones used in the final Random Forest algorithms for 5-star and 1-star prediction models.

Table 39: *RF Classification Error Across mtry for 5-Star Prediction Models*

Classification Error Across #Variables/Split (5-Star Prediction Models)							
	Activity	Bike Tour	Food Tour	Cooking Class	Sight Seeing	Spa	Average
1	0.470	0.395	0.380	0.435	0.405	0.425	0.418
2	0.350	0.415	0.370	0.390	0.380	0.385	0.382
3	0.340	0.440	0.340	0.375	0.395	0.350	**0.373**
4	0.375	0.465	0.365	0.365	0.390	0.405	0.394
5	0.360	0.475	0.380	0.370	0.395	0.365	0.391
6	0.405	0.460	0.370	0.360	0.390	0.395	0.397
7	0.360	0.455	0.385	0.375	0.400	0.420	0.399
8	0.370	0.485	0.365	0.380	0.410	0.400	0.402
9	0.395	0.500	0.375	0.380	0.415	0.445	0.418
10	0.390	0.490	0.360	0.385	0.395	0.405	0.404

Table 40: *RF Classification Error Across mtry for 1-Star Prediction Models*

Classification Error Across #Variables/Split (1-Star Prediction Models)				
	Activity	Sight Seeing	Spa	Average
1	0.325	0.380	0.345	0.350
2	0.175	0.135	0.300	0.203
3	0.155	0.110	0.220	0.162
4	0.150	0.105	0.220	**0.158**
5	0.150	0.105	0.220	0.158
6	0.160	0.115	0.210	0.162
7	0.165	0.105	0.220	0.163

8	0.175	0.110	0.225	0.170
9	0.180	0.110	0.215	0.168
10	0.190	0.110	0.200	0.167

4.2.3.2 Random Forest Effectiveness for Predicting 5-Star Reviews

The Random Forest algorithm does a good job in predicting 5-star reviews. As seen in ***Table 41*** (Activity), ***Table 42*** (Bike Tour), ***Table 43*** (Cooking Class), ***Table 44*** (Food Tour), ***Table 45*** (Sight Seeing), and ***Table 46*** (Spa), the F1 Score across all 6 models is at an average of 74.9%, ranging from 65.3% (Activity) to 81.5% (Sight Seeing). The accuracy across all 6 models is at an average of 67.6%, ranging from 63.1% (Bike Tour) to 75.4% (Sight Seeing).

Table 41: *RF 'Activity' 5-Star Model Effectiveness*

Activity 5-Star Prediction		
	No	Yes
No	379	176
Yes	220	372

F1 Score	**65.3%**
Accuracy	**65.5%**
Precision	62.8%
Recall	67.9%
Specificity	63.3%

Table 42: *RF 'Bike Tour' 5-Star Model Effectiveness*

Bike Tour 5-Star Prediction		
	No	Yes
No	77	52
Yes	253	444

F1 Score	**74.4%**
Accuracy	**63.1%**
Precision	63.7%
Recall	89.5%
Specificity	23.3%

Table 43: *RF 'Cooking Class' 5-Star Model Effectiveness*

Cooking Class 5-Star Prediction		
	No	Yes
No	56	37
Yes	86	176

F1 Score	**74.1%**
Accuracy	**65.4%**
Precision	67.2%
Recall	82.6%
Specificity	39.4%

Table 44: *RF 'Food Tour' 5-Star Model Effectiveness*

Food Tour 5-Star Prediction		
	No	Yes
No	65	35
Yes	147	283

F1 Score	**75.7%**
Accuracy	**65.7%**
Precision	65.8%
Recall	89.0%
Specificity	30.7%

Table 45: *RF 'Sight Seeing' 5-Star Model Effectiveness*

Sight Seeing 5-Star Prediction		
	No	Yes
No	544	148
Yes	484	1394

F1 Score	**81.5%**
Accuracy	**75.4%**
Precision	74.2%
Recall	90.4%
Specificity	52.9%

Table 46: *RF 'Spa' 5-Star Model Effectiveness*

Spa 5-Star Prediction		
	No	Yes
No	912	358
Yes	1228	2852

F1 Score	**78.2%**
Accuracy	**70.4%**
Precision	69.9%
Recall	88.8%
Specificity	42.6%

4.2.3.3 Random Forest Effectiveness for Predicting 1-Star Reviews

The Random Forest algorithm seemingly does the best job in predicting 1-star reviews. As seen in ***Table 47*** (Activity), ***Table 48*** (Sight Seeing), and ***Table 49*** (Spa), the F1 Score across all 3 models is at an average of 81.8% (compared to SVM predicting 1-star reviews at 80.3%). The accuracy across all 3 models is at an average of 85.6% (compared to SVM predicting 1-star reviews at 84.9%. This shows that the Random Forest model is more effective than Logistic Regression and SVM predicting both 5-star reviews and 1-star reviews.

Table 47: *RF 'Activity' 1-Star Model Effectiveness*

Activity 5-Star Prediction		
	No	Yes
No	123	18
Yes	10	71

F1 Score	**83.5%**
Accuracy	**87.4%**
Precision	87.7%
Recall	79.8%
Specificity	92.5%

Table 48: *RF 'Sight Seeing' 5-Star Model Effectiveness*

Sight Seeing 5-Star Prediction		
	No	Yes
No	204	26
Yes	22	124

F1 Score	**83.8%**
Accuracy	**87.2%**
Precision	84.9%
Recall	82.7%
Specificity	90.3%

Table 49: *RF 'Spa' 5-Star Model Effectiveness*

Spa 5-Star Prediction		
	No	Yes
No	516	85
Yes	96	323

F1 Score	**78.1%**
Accuracy	**82.3%**
Precision	77.1%
Recall	79.2%
Specificity	84.3%

4.3 Model Evaluation

4.3.1 Best Performing ML Model

4.3.1.1 F1-Score Comparison

Comparing the F1-score across all the machine learning models, as seen in ***Figure 39***, the first thing to note is that all three models have similar performance and effectiveness. All moving high and low depending on the model in question. However, upon a more in-dept look, we can see that the machine-learning algorithm with the highest F1-score overall seems to be Random Forest – indicating that may be the best performing algorithm.

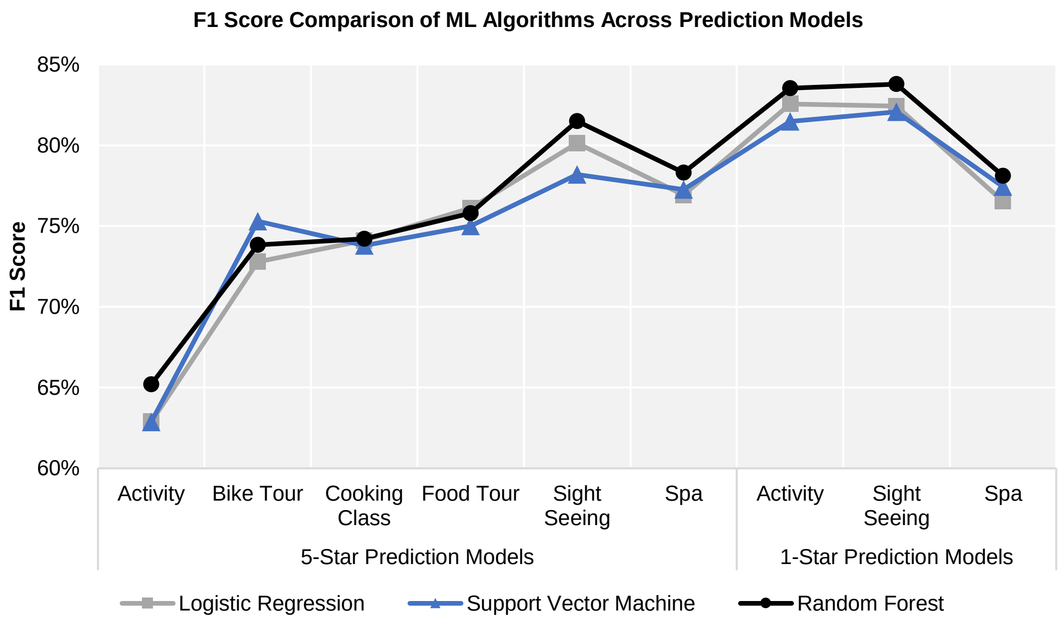


Figure 39: *F1 Score Comparison of ML Algorithms Across Prediction Models*

4.3.1.2 Accuracy Comparison

The second prediction metric of focus, accuracy, is also compared across all machine-learning models. Similar to the previous chart, the values of accuracy, as seen in ***Figure 40***, seem to also be quite close together, following a similar high and low trend depending on the model. For just one model – Food Tour – a large disparity of performance can be seen across the models. Unlike comparing F1-score, there is no clear best-performing model. However, from visual inspection, it can concluded that the top performing machine learning algorithms are either Logistic Regression or Random Forest.

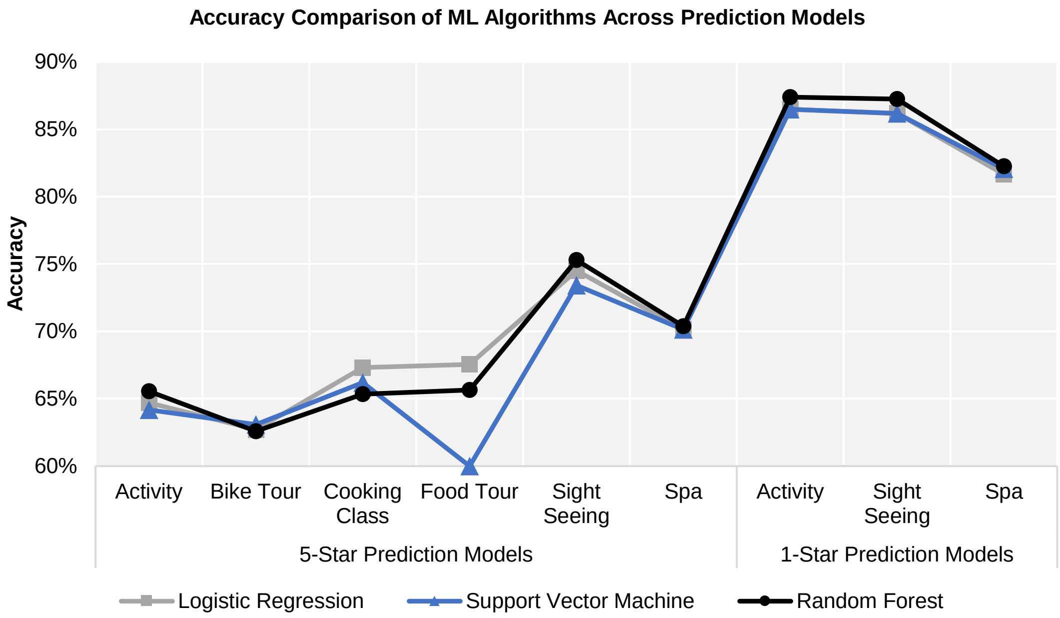


Figure 40: *Accuracy Comparison of ML Algorithms Across Prediction Models*

4.3.1.3 Run Time Comparison

The run time for the three machine learning models differed quite drastically. Run time was taken for hyper-parameter tuning and running the model itself, with the average run time being 6.15 seconds for Logistic Regression, 30.09 seconds for Support Vector Machine, and 11.83 seconds for Random Forest. The shortest run time was for Logistic Regression – mostly having to do with the fact that for this paper, this model did not include any hyper-parameter tuning.

4.3.1.4 Best Machine-Learning Model

As mentioned in the previous sections, all three models have similar performances in terms of effectiveness. For the purposes of predictions done for this thesis, all 3 models would work comparably. However, when trying to find the “best” model, it is important to foresee future work and applications. In real-life applications, having a model be highly scalable while maintaining a short run-time is quite imperative. With high effectiveness scores, reasonably low run-time comparatively, and the nature to adapt to large volumes of data, it can be concluded that the Random Forest Algorithm is the best machine-learning model.

4.3.2 Feature Importance

In order to gain insights on which feature of the prediction models are the best predictors, it is important to look at feature importance. For Random Forest algorithms, feature importance is measured using the mean decrease in Gini. **Gini Impurity** or **Gini Index** is the probability that a random sample from a particular node is misclassified. Thus, the lower the Gini index, the purer the split, the better. The *mean decrease* in Gini, then, is the average of the variable's decrease in impurity. Thus, the *higher* the mean decrease in Gini, the higher the importance of the feature.

4.3.2.1 Feature Importance for 5-Star Prediction Models

Across all 5-Star prediction models, the 'sentiment' feature – continuous variable for calculated sentiment score – has the highest importance, significantly higher than all the other features. Individual feature significance for each of the 5-star prediction models can be seen from ***Figure 41*** (Activity), ***Figure 42*** (Bike Tour), ***Figure 43*** (Cooking Class), ***Figure 44*** (Food Tour), ***Figure 45*** (Sight Seeing), and ***Figure 46*** (Spa).

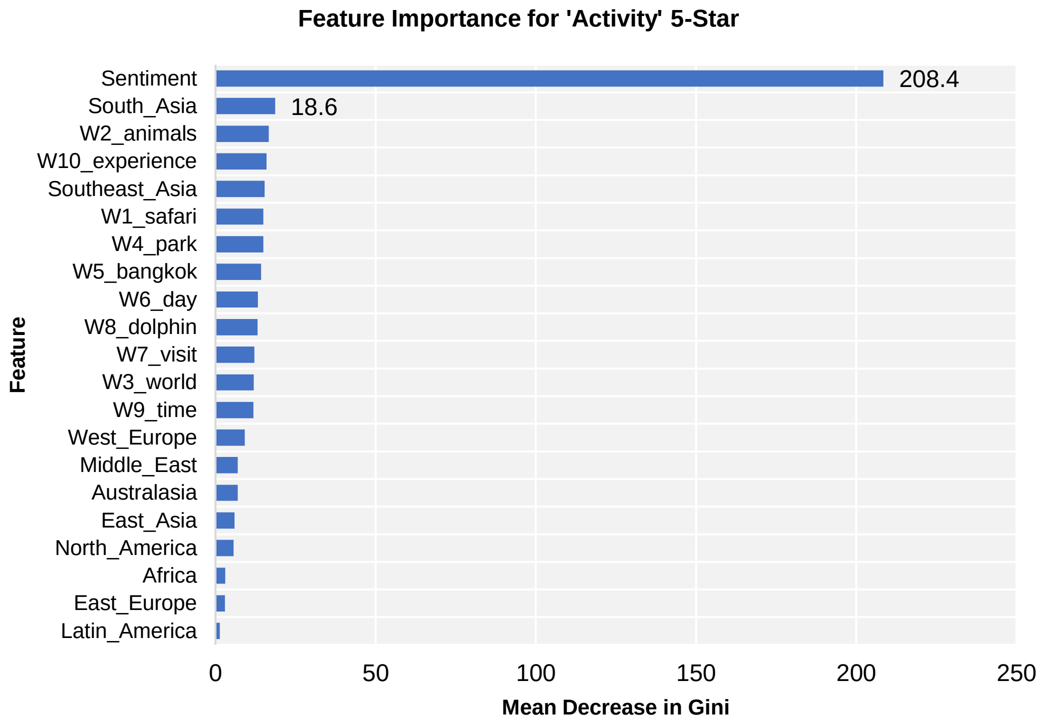


Figure 41: *Feature Importance for 'Activity' 5-Star*

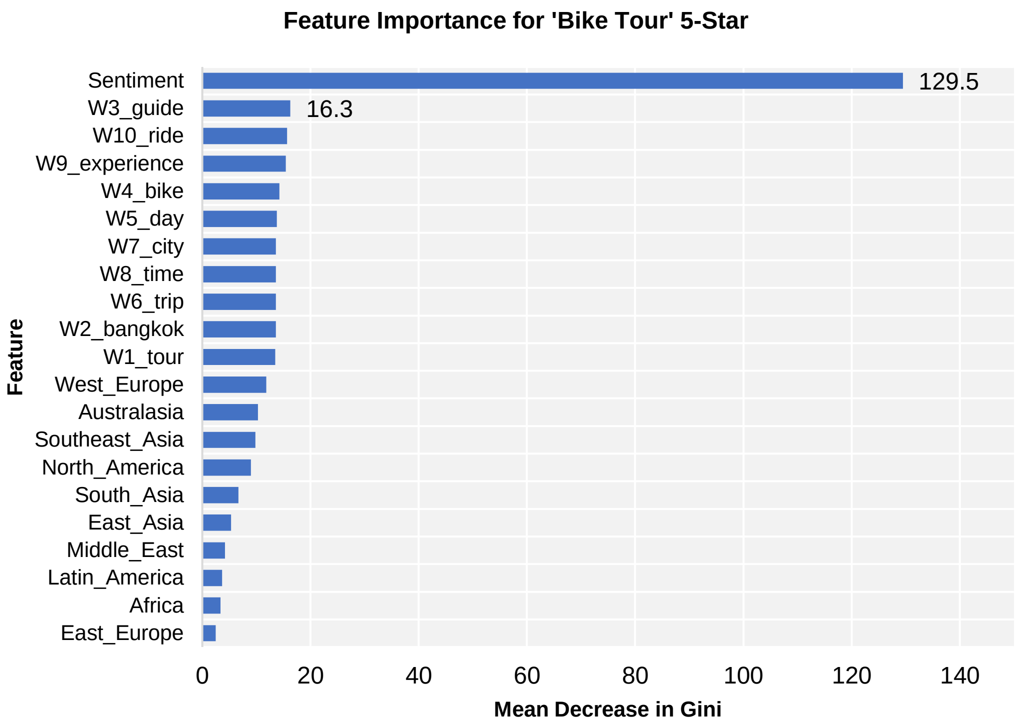


***Figure 42:** Feature Importance for 'Bike Tour' 5-Star*

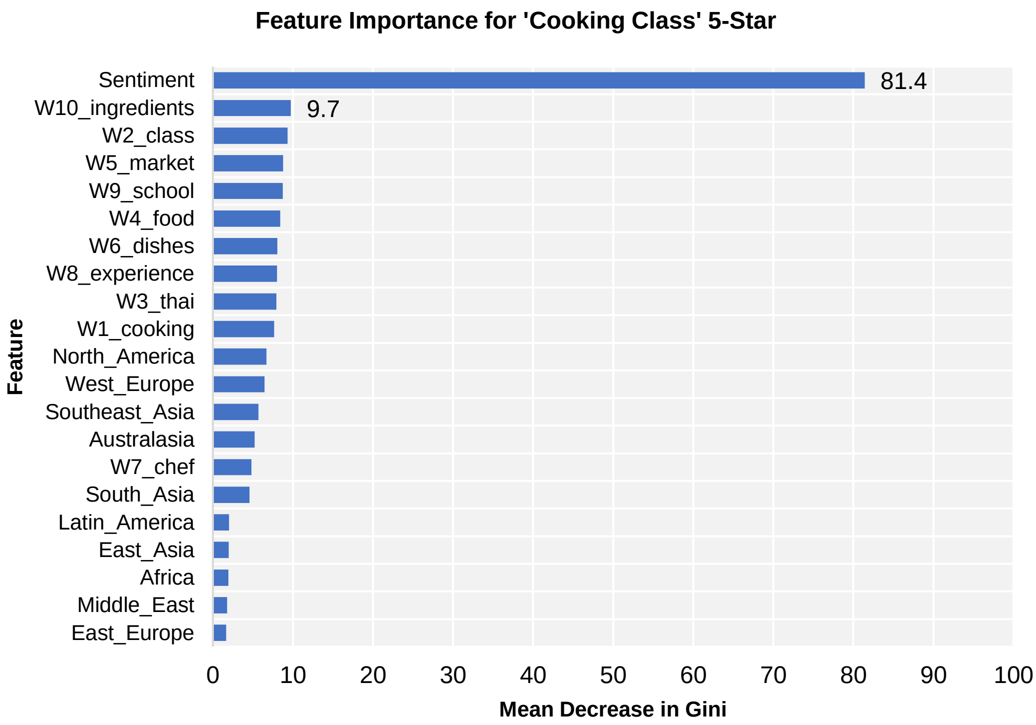


***Figure 43:** Feature Importance for 'Cooking Class' 5-Star*

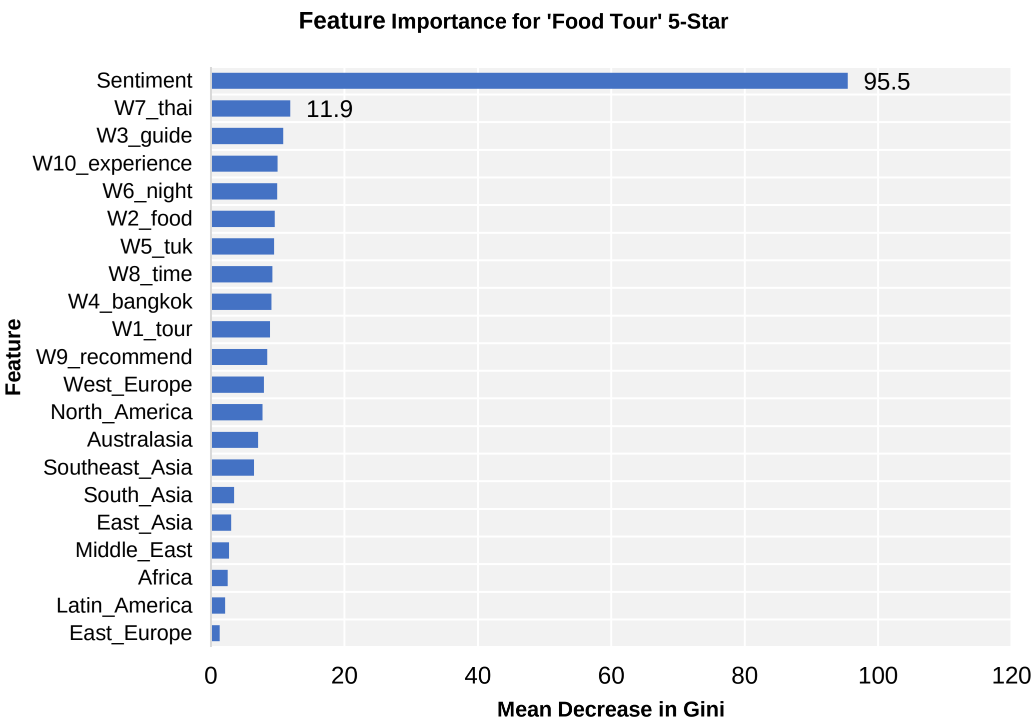


Figure 44: *Feature Importance for 'Food Tour' 5-Star*

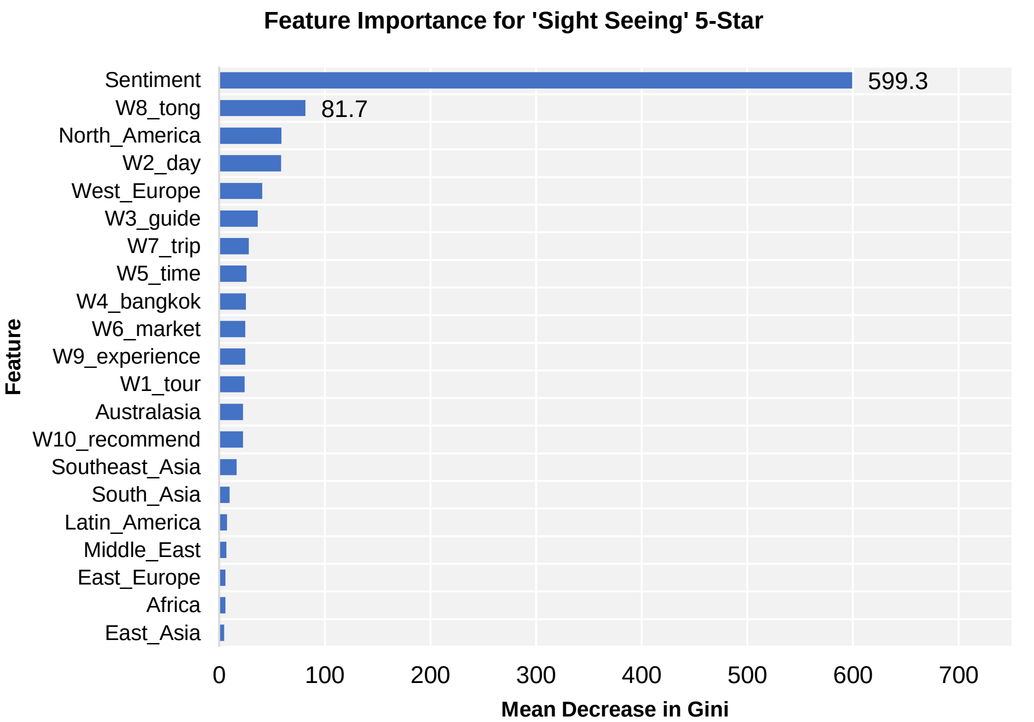


Figure 45: *Feature Importance for 'Sight Seeing' 5-Star*

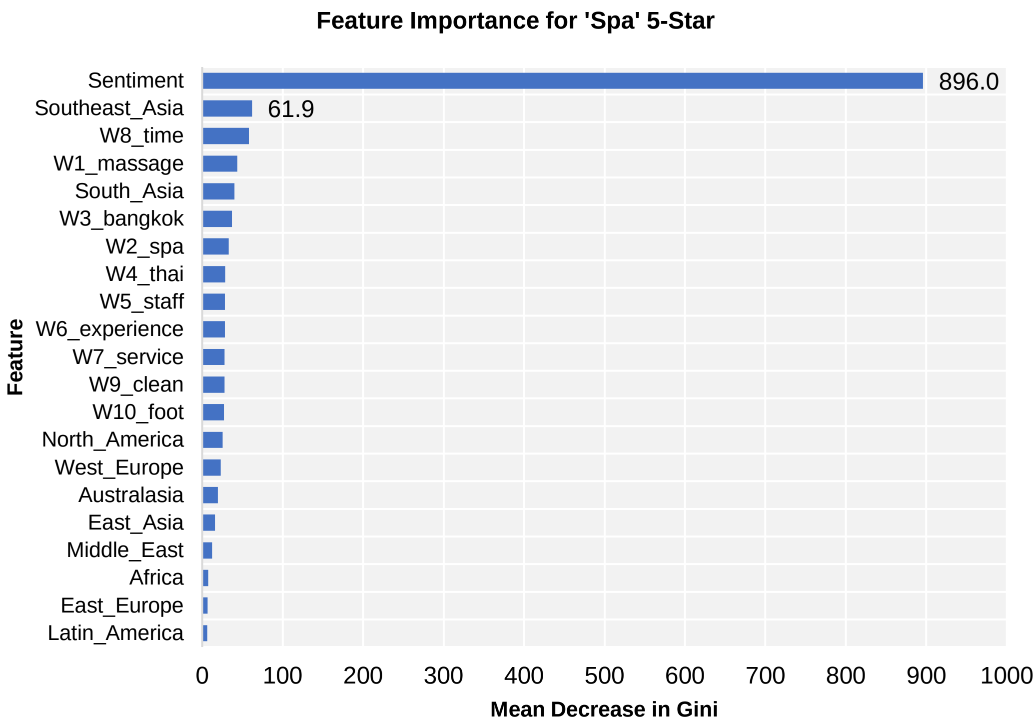


Figure 46: *Feature Importance for 'Spa' 5-Star*

Although it is extremely apparent from the individual models that the 'sentiment' feature has the most significance across all models, it is still quite unclear whether the features related to origin or the features related to frequent words are more significant. In order to find out, the average and weighted average (weighted by size of dataset) were taken and plotted (as seen in ***Figure 47***). From the plot, it can be seen that the features related to frequent words are slightly more significant than origin.

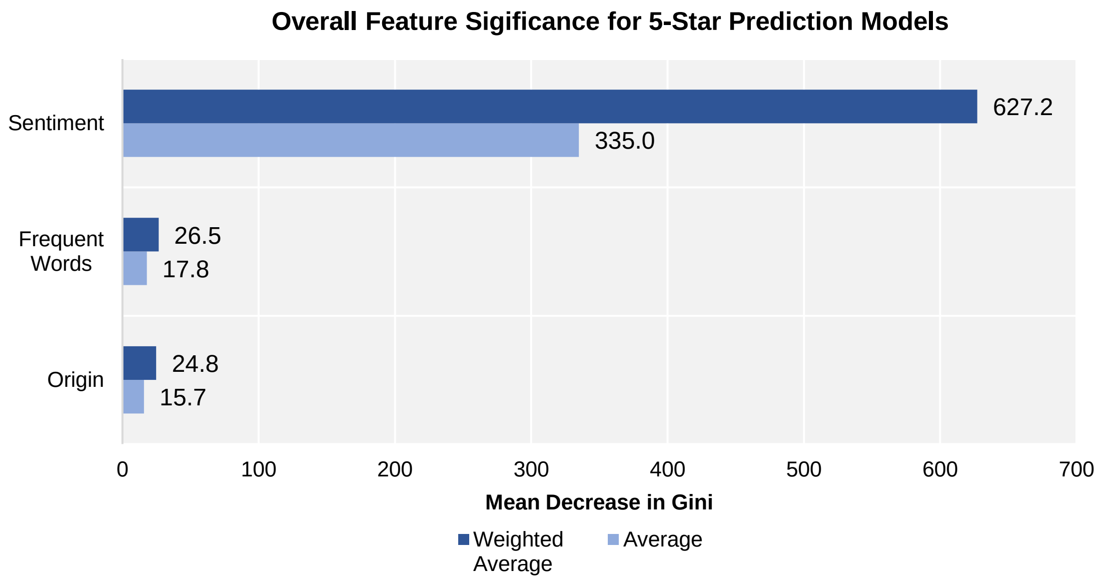


Figure 47: *Overall Feature Significance for 5-Star Prediction Models*

4.3.2.2 Feature Importance for 1-Star Prediction Models

Similar to the 5-star prediction models, across all 1-Star prediction models, the 'sentiment' feature – also has the highest importance. The difference in importance is even more substantial for the 1-star prediction models. Individual feature significance for each of the 1-star prediction models can be seen from ***Figure 48*** (Activity), ***Figure 49*** (Sight Seeing), ***Figure 50*** (Spa).

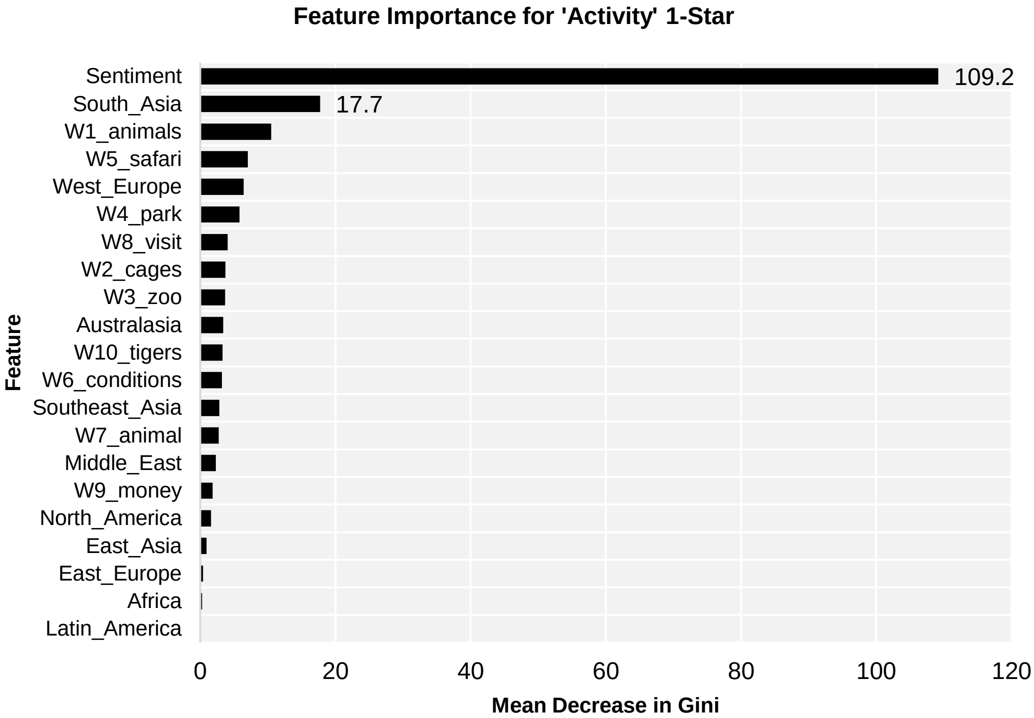


Figure 48: *Feature Importance for 'Activity' 1-Star*

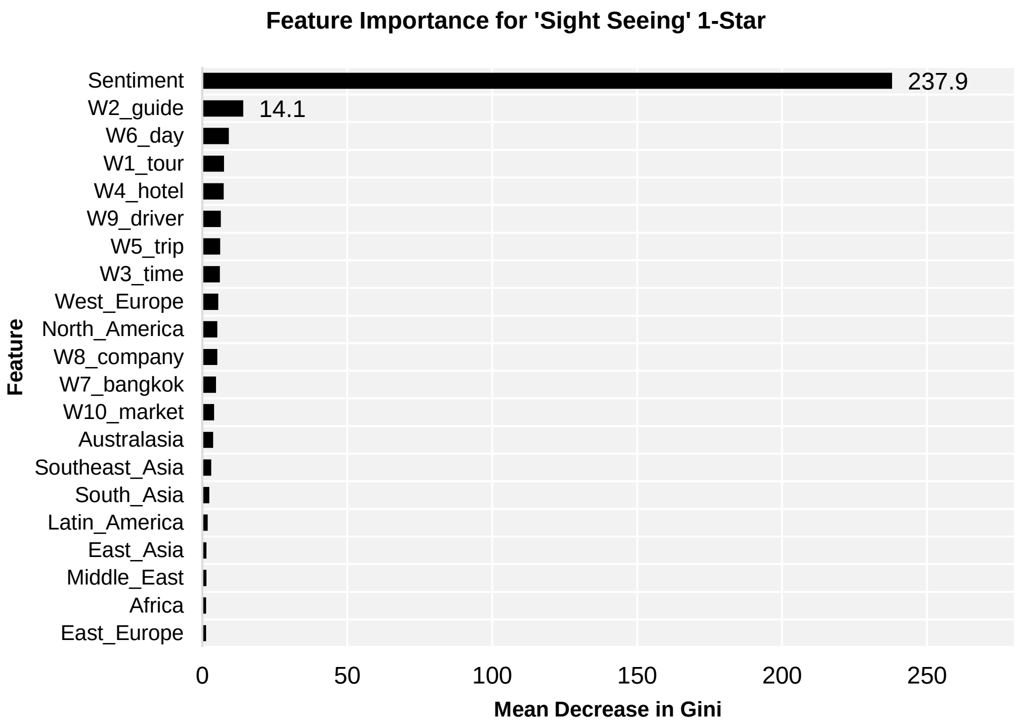


Figure 49: *Feature Importance for 'Sight Seeing' 1-Star*

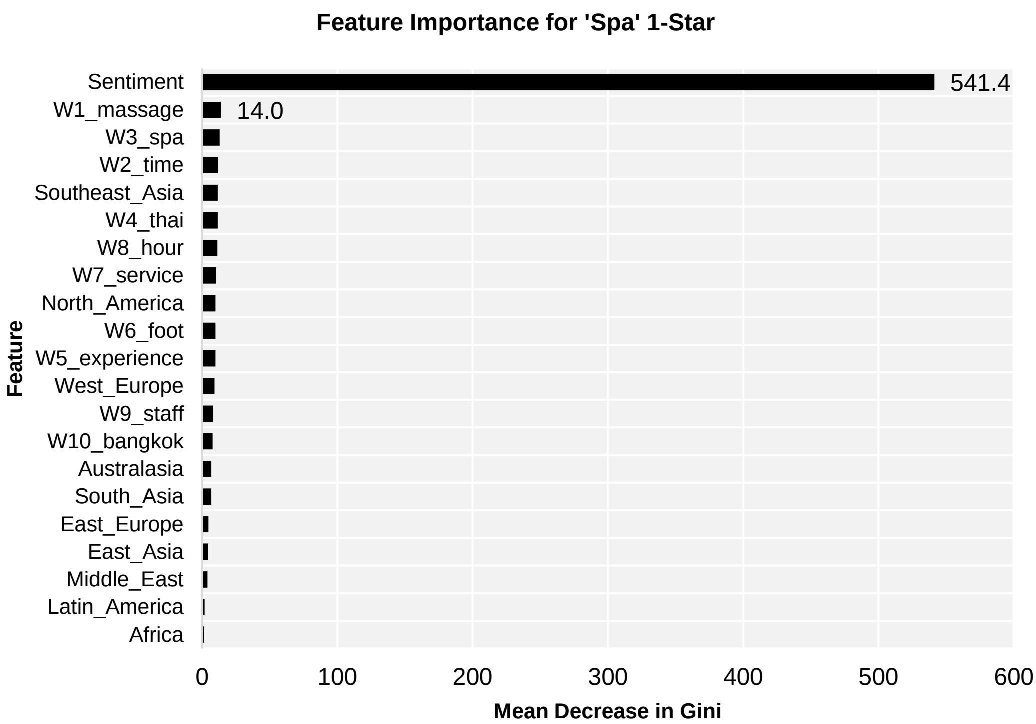


Figure 50: *Feature Importance for 'Spa' 1-Star*

For the 1-star prediction models the importance of the 'sentiment' feature is quite apparent, even more so than the 5-star models. Similar to the 5-star models, in order to understand the importance of the other features, the average and weighted average of the Mean Decrease in Gini were taken and plotted (as seen in ***Figure 51***). From the plot, it can be seen that the other features are all equally poor when compared to the 'sentiment' feature.

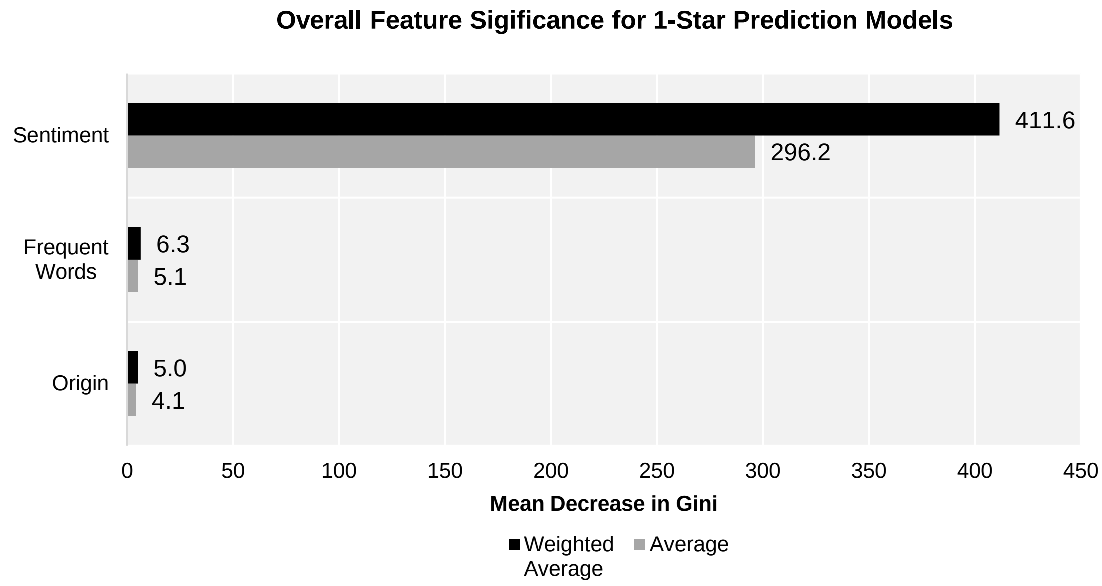


Figure 51: *Overall Feature Significance for 1-Star Prediction Models*

Chapter 5: Conclusion and Future Work

5.1 Conclusions

5.1.1 Learnings on Travel and Tourism

The importance of travel and tourism is undeniable. From the pleasure it brings to travelers to the economic benefits it provides to host nations, there is no debating the value of the sector. Due to its robust growth and vast range of positive impact, T&T continuously attracts new players yearly. With competition rising and saturation within the sector growing, a solid understanding of tourism preferences and trends are crucial to remain competitive.

Another undeniable movement has been the rise of social media. The constant and infinite supply of data through social media attest it to be the perfect source of big-data. Social media has now become the leading supply for travel and tourism information. Whether it be researching activities, accommodations, flights, and more prior to travel; status and photo updates during travel; or even reviews left post travel; the T&T sector is a great data source and provides the perfect opportunity for a systematic analysis of tourist preferences via user-generated content.

Thus, the overarching goal of this research is to gain visibility of Bangkok's tourism preferences and whether or not tourist needs are being met. This study was able to leverage the benefits of big-data and prediction models to uncover significant insights on tourist preferences, trends, and focus areas.

Throughout the study, it was uncovered that there is a preference for different tour/activity types based on tourist origin. This information could greatly influence and benefit the marketing and communication efforts within the Travel and Tourism sector. Through sentiment analysis, natural language processing, and word count frequency, it was discovered that features such as guides, cleanliness, and service-level greatly affect the experience of tourists – being almost the deciding factor of whether tourists have positive or negative perceptions towards the tour/activity.

The prediction models were able to reveal that the features (independent variables) that were predicted to affect the experience of tourists—gauged by the star-rating given—are mostly significant. For example, frequently occurring words, such as "delicious", accurately predict 5-star ratings, further solidifying the fact that features such as taste impact the positive experience of tourists.

5.2.1 Learnings on Machine Learning Algorithms

Apart from learnings and insights into Bangkok's travel and tourism sector, this research also revealed characteristics and capabilities of different machine learning algorithms.

The first machine learning model run for this research was the Logistic Regression model. The Logistic Regression algorithm is one of the most popular classification models and widely used. Due to its low complexity and the fact that it does not require any tuning, Logistic Regression was a good place to start as the first machine-learning model. As done in many other studies, for this study, Logistic Regression was initially run and used as a benchmark against other, more complicated algorithms.

The second machine learning model run for this research was the Support Vector Machine model. This algorithm was much more complex than LR, both in terms of concept understanding as well as hyperparameter tuning. Support Vector machine is known to be effective in high-dimension spaces – which is quite necessary for the purposes of this study (each model has 21 features). However, from ***Figure 39***, it is quite apparent that the SVM algorithm did not do a superior job to LR algorithm for predicting both 5-star and 1-star reviews. This is most likely due to the nature of the input data. The dataset could be pegged as noisy – with overlapping classes and no clear segregation. When thinking of future work, in terms of scalability, Support Vector Machine may not be the best choice as well. The algorithm is known to be lacking in terms of computational efficiency. The larger the dataset becomes, the exponentially longer the algorithm could take. Thus, another machine-learning algorithm would be required.

The final machine learning model run for this research was the Random Forest model. Known for its efficiency of working with large volumes of data, this model would provide something SVM could not – scalability. The Random Forest is also known to currently be the most accurate algorithm available.

From ***Figure 39***, it can be seen that all three models are seemly close in terms of performance (judged by F1 scores). However, the model that did perform the best was the last model, the Random Forest. More likely than not, this is due to the fact that Random Forest uses the "ensemble learning" technique, which is a process of building multiple machine learning models and combining the predictions into one final model prediction. The ensemble learning technique reduces variance and overfitting and thus improves accuracy and F1 score.

In conclusion, this research has been able to draw multiple insights, both for travel and tourism and for machine learning algorithms. The study was able to find that Random Forest provides the best prediction performance. For a sample of incoming tourists, stakeholders can use a similar algorithm as this study presented to map tourist features with a binary inference of whether tourists could be satisfied or dissatisfied with their travel experience. Giving stakeholders this power to predict tourist approval and view the motive behind tourist enjoyment (5-star) and complaints (1-star) could further accelerate and drive the Travel and Tourism sector forward into the future.

5.2 Research Limitations

In a perfect world, this study would be able to gain insights from a limitless supply of data, covering all geographical locations. However, to maintain feasibility, the research was only limited to retrieving data from one source – TripAdvisor, within one location – Bangkok, Thailand, and covering only six tour/activity types – Activity, Bike Tour, Cooking Class, Food Tour, Sight Seeing, and Spa.

Due to the nature of data storage on TripAdvisor, the independent variables tested against the dependent output variable were also limited to only consumer features. If other features – such as gender or traveler type – were available, this model could have had increased accuracy or more insights drawn.

This research was also limited to building only three machine learning prediction models – Logistic Regression, Support Vector Machine, and Random Forest. The modeling process across all machine learning algorithms was also limited to include only the top most important hyperparameters for tuning.

Another limitation within this research is the way data was split into only training and testing sets – with testing sets used to find the effectiveness of the models. There was no split of a validation set – a set that is held back from the training which is used to tune the parameters and provide an unbiased evaluation of the model fit.

5.3 Future Work

As the years advance and machine learning algorithms get more advanced and accurate, there could be a greatly increased performance in the prediction models covered in this research. For the future of this study, there are three key areas that would be optimal to implement.

5.3.1 Increased Features

As mentioned in ***Section 5.2***, one of the limitations for this research was the available consumer features. For the future of this research, it would be very beneficial for both tourism insights and model accuracy to have a more exhaustive list of features. These features could be gender, traveler type (single, family, couple), travel type (business, pleasure, meeting friends/family), income, and more. A larger list of features could be effective in determining what are driving factors to tourism decisions. In order to obtain such, it might be necessary in the future to explore tourist data sources outside of TripAdvisor to a more detailed information-dense social media source.

5.3.2 Further-Developed Natural Language Processing

Although the natural language processing done in this research was quite insightful, there were certain limitations to what was found out. Most of the top frequent words were just names of the Activity/Tour or other words synonymous to the category. Upon further development of the research, a process would be implemented to remove such redundant words that provide any insight (such as the word “food” and “tour” in the Food Tour category).

Apart from that, there are certain words that require further examination. For example, the word “market” begs to question whether tourists are talking about “food market” or “night market” or even “flower market”. Further development of this research would enable the frequency count of phrases that are of particular interest within the Travel and Tourism sector of Bangkok.

5.3.3 Advanced Machine Learning Models

Again, as mentioned in ***Section 5.2***, this study was limited to building three machine learning algorithms. There are, however, more algorithms or more variations of the algorithms within this study that could further increase the quality of prediction. Some methods that would be great in further iterations of this study are boosting methods such as XGBoost, LightGBM, and CatBoost, to name a few.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable [79]. LightGBM has all the benefits of XGBoost, without the load and extensive time it takes to train with large volumes of data. Lastly, CatBoost is an algorithm built using gradient boosting on decision trees. CatBoost has had chatter of being much more superior to XGBoost in terms of prediction time – some mentioning it to be up to 8X faster[80].

A combination of all Future Work recommendations could undoubtedly fine-tune the work done in this study to draw out even more insights and generate the highest predictive performance achievable.

5.3.4 Adopting to Tour Operators

This thesis was able to uncover several learnings on tourist preferences across a variety of different tour categories. This was done to learn of the tourism industry in Bangkok overall. However, this type of study could also be adopted to individual tour operators, in order to uncover tourist preferences, tourist sentiment, tourist focus, and tourism trends on a single operator. For example, if a Bike Tour operator were to apply this study, they could find out which origin prefers their bike tour, what type of highly-occurring words are used to see what features tourists are focusing on. They can conclude tourist sentiment to see whether tourists are overall satisfied or dissatisfied with the service they are providing. Upon future work, with a more exhaustive list of features, more advanced NLP, and more advanced machine learning models, this study has the potential of uncovering a limitless range of learnings within the tourism industry – whether that be for tourism overall, or for an explicit tour operator.

Appendices

Appendix 1: Chi-Square Test for Independence Methodology

Step 1: State the Null and Alternative Hypotheses

H0: No Association between Feature A and Feature B (Independent)

H1: Is Association between Feature A and Feature B (Not Independent)

Step 2: Gathered data for the two features tested

	1-Star	2-Star	3-Star	4-Star	5-Star	R Total
Africa	12*	7	20	80	421	540
Australasia	170	107	186	762	4,472	5,697
East Asia	68	34	59	211	1,083	1,455
East Europe	25	17	20	59	489	610
Latin American	16	10	17	59	514	616
Middle East	34	34	63	210	873	1,214
North America	235	188	290	1,013	8,711	10,437
South Asia	142	102	269	1,045	2,501	4,059
Southeast Asia	522	318	560	1,858	5,491	8,749
West Europe	404	247	401	1,525	11,304	13,881
C Total	1,628	1,064	1,885	6,822	35,859	47,258

Rows (r) = 10

Columns (c) = 5

Alpha (α) = 0.05

**data in cell = #Reviews per group*

Step 3: Calculated Expected Frequency Count for each Feature A against all Feature B ('Africa' shown)

	1-Star	2-Star	3-Star	4-Star	5-Star	
Africa (Exp)	18.6	12.2	21.5	78.0	409.7	
(Obs - Exp)	(6.6)	(5.2)	(1.5)	2.0	11.3	
(Obs - Exp)2	43.6	26.6	2.4	4.2	126.6	
(Obs - Exp)2 / E	2.3	2.2	0.1	0.1	0.3	5.0

$$E_{r,c} = \frac{(n_r \times n_c)}{n}$$

$$\chi^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}$$

Step 4: Calculated individual Chi-Square values for each Feature A, then summed to find final
Chi-Square Statistic

Africa	5.0	Middle East	16.9
Australasia	24.1	North America	331.7
East Asia	6.8	South Asia	540.6
East Europe	13.4	Southeast Asia	842.2
Latin America	19.4	West Europe	238.2
Chi-Square Statistic		**2,038.3**	

Step 5: Found degrees of freedom using rows (r = number of feature A categories) and columns (c = number of feature B categories)

DF = (r – 1) x (c – 1) = (10 – 1) x (5 – 1) = 36

Step 6: Calculated Critical Value

Using the excel function, found critical value = CHISQ.INV.RT(alpha,df) = CHISQ.INV.RT(0.05,36) = 62

Step 7: Decision

Compared the Chi-Square Statistic to the Critical Value to find that Chi-Square Statistic (2038.3) > Critical Value (62), thus rejected the null hypothesis and concluded that there is an association between the two features (Origin Group & Review Rating) and that they are not independent.

Step 8: Repeat

Repeated the same process for two other feature groups and got the following results.

H0: No Association (Independent) H1: Association (Not Independent)				
Items Tested	*Chi-Square Statistic*	*Degrees of Freedom*	*Critical Value*	*Decision*
Origin & Review Rating	2,083	36	51	2,083 > 51; **Reject H0**
Tour/Activity Type & Review Rating	4,308	20	31	4,308 > 31; **Reject H0**
Origin & Tour/Activity Type	11,112	45	62	11,112 > 62; **Reject H0**

Appendix 2: Lexicon of Positive and Negative Words

(incomplete list)

Positive Words					Negative Words				
a+	bravo	delicious	excellent	gorgeous	abnormal	bore	dark	fear	gall
abound	breakthrough	delight	exemplar	gorgeously	abolish	bored	darken	fearful	galling
abounds	breakthroughs	delighted	exemplary	grace	abominable	boredom	darkened	fearfully	gallingly
abundance	breathlessness	delightful	exhilarate	graceful	abominably	bores	darker	fears	galls
abundant	breathtaking	delightfully	exhilarating	gracefully	abominate	boring	darkness	fearsome	gangster
accessible	breathtakingly	delightfulness	exhilaratingly	gracious	abomination	botch	dastard	feckless	gape
accessible	breeze	dependable	exhilaration	graciously	abort	bother	dastardly	feeble	garbage
acclaim	bright	dependably	exonerate	graciousness	aborted	bothered	daunt	feeblely	garish
acclaimed	brighten	deservedly	expansive	grand	aborts	bothering	daunting	feebleminded	gasp
acclamation	brighter	deserving	expeditiously	grandeur	abrade	bothers	dauntingly	feign	gauche
accolade	brightest	desirable	expertly	grateful	abrasive	bothersome	dawdle	feint	gaudy
accolades	brilliance	desiring	exquisite	gratefully	abrupt	bowdlerize	daze	fell	gawk
accommodative	brilliances	desirous	exquisitely	gratification	abruptly	boycott	dazed	felon	gawky
accommodative	brilliant	destiny	extol	gratified	abscond	braggart	dead	felonious	geezer
accomplish	brilliantly	detachable	extoll	gratifies	absence	bragger	deadbeat	ferociously	genocide
accomplished	brisk	devout	extraordinarily	gratify	absent-minded	brainless	deadlock	ferocity	get-rich
accomplishment	brotherly	dexterous	extraordinary	gratifying	absentee	brainwash	deadly	fetid	ghastly
accomplishments	bullish	dexterously	exuberance	gratifyingly	absurd	brash	deadweight	fever	ghetto
accurate	buoyant	dexterous	exuberant	gratitude	absurdity	brashly	deaf	feverish	ghosting
accurately	cajole	dignified	exuberantly	great	absurdly	brashness	dearth	fevers	gibber
achievable	calm	dignify	exult	greatest	absurdness	brat	death	fiasco	gibberish
achievement	calming	dignity	exultant	greatness	abuse	bravado	debacle	fib	gibe
achievements	calmness	diligence	exultation	grin	abused	brazen	debase	fibber	giddy
achievable	capability	diligent	exultingly	groundbreak	abuses	brazenly	debasement	fickle	gimmick
acumen	capable	diligently	eye-catch	guarantee	abusive	brazenness	debaser	fiction	gimmicked
adaptable	capably	diplomatic	eye-catching	guidance	abysmal	breach	debatable	fictional	gimmicking
adaptive	captivate	dirt-cheap	eye catch	guiltless	abysmally	break	debauch	fictitious	gimmicks
adequate	captivating	distinction	eye-catching	gumption	abyss	break-up	debaucher	fidget	gimmicky
adjustable	carefree	distinctive	fabulous	gush	accidental	break-ups	debauchery	fidgety	glare
admirable	cashback	distinguished	fabulously	gusto	accost	breakdown	debilitate	fiend	glaringly
admirably	cashbacks	diversified	facilitate	gutsy	accursed	breaking	debilitating	fiendish	glib
admiration	catchy	divine	fair	hail	accusation	breaks	debility	fierce	glibly
admire	celebrate	divinely	fairly	halcyon	accusations	breakup	debt	figurehead	glitch
admirer	celebrated	dominate	fairness	hale	accuse	breakups	debts	filth	glitches
admiring	celebration	dominated	faith	hallmark	accuses	bribery	decadence	filthy	gloatingly
admiringly	celebratory	dominates	faithful	hallmarks	accusing	brimstone	decadent	finagle	gloom
adorable	champ	dote	faithfully	hallowed	accusingly	bristle	decay	finicky	gloomy
adore	champion	dotingly	faithfulness	handier	acerbate	brittle	decayed	fissures	glower
adored	charisma	doubtless	fame	handily	acerbic	broke	deceit	fist	glum
adorer	charismatic	dreamland	famed	hands-down	acerbically	broken	deceitful	flabbergast	glut
adoring	charitable	dumbfounded	famous	handsome	ache	broken-hearted	deceitfully	flabbergasted	gnawing
adoringly	charm	dumbfounding	famously	handsomely	ached	brood	deceitfulness	flagging	goad
adroit	charming	dummy-proof	fancier	handy	aches	browbeat	deceive	flagrant	goading
adroitly	charmingly	durable	fascinating	happier	ache	bruise	deceiver	flagrantly	god-awful
adulate	chaste	dynamic	fancy	happily	aching	bruised	deceivers	flair	goof
adulation	cheaper	eager	fanfare	happiness	acrid	bruises	deceiving	flairs	goofy
adulatory	cheapest	eagerly	fans	happy	acridly	bruising	deception	flak	goon
advanced	cheer	eagerness	fantastic	hard-working	acridness	brusque	deceptive	flake	gossip
advantage	cheerful	earnest	fantastically	hardier	acrimonious	brutal	deceptively	flakey	graceless
advantageous	cheery	earnestly	fascinate	hardy	acrimoniously	brutalizing	declaim	flakieness	gracelessly
advantageously	cherish	earnestness	fascinating	harmless	acrimony	brutalities	decline	flaking	graft
advantages	cherished	ease	fascinatingly	harmonious	adamant	brutality	declines	flaky	grainy
adventuresome	cherub	eased	fascination	harmoniously	adamantly	brutalize	declining	flare	grapple
adventurous	chic	eases	fashionable	harmonize	addict	brutalizing	decrement	flares	grate
advocate	chivalrous	easier	fashionably	harmony	addicted	brutally	decrepit	flareup	grating
advocated	chivalry	easiest	fast	headway	addicting	brute	decrepitude	flareups	gall
advocates	civility	easiness	fast-growing	heal	abnormal	brutish	decry	flat-out	galling
affability	civilize	easing	fast-paced	healthful	abolish	bs	defamation	flaunt	gallingly
a+	clarity	easy	faster	gorgeous	abominable	buckle	defamations	flaw	galls
abound	classic	delicious	fastest	gorgeously	abominably	bug	defamatory	flawed	gangster
abounds	classy	delight	fastest-growing	grace	abominate	bugging	defame	flaws	gape
affable	clean	delighted	faultless	graceful	abomination	buggy	defect	flee	garbage
affably	cleaner	delightful	fav	gracefully	abort	bugs	defective	fleed	garish
affectation	cleanest	delightfully	fave	gracious	aborted	bulkier	defects	fleeing	gasp
affection	cleanliness	delightfulness	favor	graciously	aborts	bulkiness	defensive	fleer	gauche
affectionate	cleanly	dependable	favorable	graciousness	abrade	bulky	defiance	flees	gaudy
affinity	clear	dependably	favored	grand	abrasive	bulkiness	defiant	fleeting	gawk
affirm	clear-cut	deservedly	favorite	grandeur	abrupt	bull****	defiantly	flicering	gawky
affirmation	cleared	deserving	favorited	grateful	abruptly	bull----	deficiencies	flicker	geezer
affirmative	clearer	desirable	favor	gratefully	abscond	bullies	deficiency	flickering	genocide
affluence	clearly	desiring	fearless	gratification	absence	bullshit	deficient	flickers	get-rich

Appendix 3: Lexicon of Stop Words

(incomplete list)

a	better	ever	hereafter	little	ones	seem	there's	wells
able	between	every	hereby	long	only	seemed	thereupon	went
about	beyond	everybody	herein	longer	onto	seeming	these	were
above	big	everyone	here's	longest	open	seems	they	we're
according	both	everything	hereupon	look	opened	seen	they'd	weren't
accordingly	brief	everywhere	hers	looking	opening	sees	they'll	we've
across	but	ex	herself	looks	opens	self	they're	what
actually	by	exactly	he's	ltd	or	selves	they've	whatever
after	came	example	hi	made	order	sensible	thing	what's
afterwards	can	except	high	mainly	ordered	sent	things	when
again	cannot	face	higher	make	ordering	serious	think	whence
against	cant	faces	highest	making	orders	seriously	thinks	whenever
ain't	can't	fact	him	man	other	seven	third	when's
all	case	facts	himself	many	others	several	this	where
allow	cases	far	his	may	otherwise	shall	thorough	whereafter
allows	cause	felt	hither	maybe	ought	shan't	thoroughly	whereas
almost	causes	few	hopefully	me	our	she	those	whereby
alone	certain	fifth	how	mean	ours	she'd	though	wherein
along	certainly	find	howbeit	meanwhile	ourselves	she'll	thought	where's
already	changes	finds	however	member	out	she's	thoughts	whereupon
also	clear	first	how's	members	outside	should	three	wherever
although	clearly	five	i'd	men	over	shouldn't	through	whether
always	c'mon	followed	ie	merely	overall	show	throughout	which
am	co	following	if	might	own	showed	thru	while
among	com	follows	ignored	more	part	showing	thus	whither
amongst	come	for	i'll	moreover	parted	shows	to	who
an	comes	former	i'm	most	particular	side	today	whoever
and	concerning	formerly	immediate	mostly	particularly	sides	together	whole
another	consequently	forth	important	mr	parting	since	too	whom
any	consider	four	in	mrs	parts	six	took	who's
anybody	considering	from	inasmuch	much	per	small	toward	whose
anyhow	contain	full	inc	must	perhaps	smaller	towards	why
anyone	containing	fully	indeed	mustn't	place	smallest	tried	why's
anything	contains	further	indicate	my	placed	so	tries	will
anyway	corresponding	furthered	indicated	myself	places	some	truly	willing
anyways	could	furthering	indicates	name	please	somebody	try	wish
anywhere	couldn't	furthermore	inner	namely	plus	somehow	trying	with
apart	course	furthers	insofar	nd	point	someone	t's	within
appear	c's	gave	instead	near	pointed	something	turn	without
appreciate	currently	general	interest	nearly	pointing	sometime	turned	wonder
appropriate	definitely	generally	interested	necessary	points	sometimes	turning	won't
are	described	get	interesting	need	possible	somewhat	turns	work
area	despite	gets	interests	needed	present	somewhere	twice	worked
areas	did	getting	into	needing	presented	soon	two	working
aren't	didn't	give	inward	needs	presenting	sorry	un	works
around	differ	given	is	neither	presents	specified	under	would
as	different	gives	isn't	never	presumably	specify	unfortunately	wouldn't
a's	differently	go	it	nevertheless	probably	specifying	unless	year
aside	do	goes	it'd	new	problem	state	unlikely	years
ask	does	going	it'll	newer	problems	states	until	yes
asked	doesn't	gone	its	newest	provides	still	unto	yet
asking	doing	good	it's	next	put	sub	up	you
asks	don	goods	itself	nine	puts	such	upon	you'd
associated	done	got	i've	no	que	sup	us	you'll
at	don't	gotten	just	nobody	quite	sure	use	young
available	down	great	keep	non	qv	take	used	younger
away	downed	greater	keeps	none	rather	taken	useful	youngest
awfully	downing	greatest	kept	noone	rd	tell	uses	your
back	downs	greetings	kind	nor	re	tends	using	you're
backed	downwards	group	knew	normally	really	th	usually	yours
backing	during	grouped	know	not	reasonably	than	uucp	yourself
backs	each	grouping	known	nothing	regarding	thank	value	yourselves
be	early	groups	knows	novel	regardless	thanks	various	you've
became	edu	had	large	now	regards	thanx	very	zero
because	eg	hadn't	largely	nowhere	relatively	that	via	

Appendix 4: Logistic Regression Model Data

Example for Activity:

Y (dependent variable) = Rating5 (discrete variable)

X1 (independent variable) = Sent (continuous variable)

X2 – X11 (independent variables) = West Europe, North America,…Africa (discrete variable)

X12 – X31 (independent variables) = W1, W2, …, W20

Rating5	Sent	West Europe	North America	Southeast Asia	Australasia	South Asia	East Asia	Middle East	Latin America	East Europe	Africa	W1	W2	W3
0	-0.10124	1	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.07191	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0.060966	0	0	1	0	0	0	0	0	0	0	0	0	0
0	-0.00708	1	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.28593	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0.05671	0	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.19493	0	0	0	1	0	0	0	0	0	0	0	0	0
0	-0.04047	0	0	1	0	0	0	0	0	0	0	0	0	0
0	-0.19224	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0.002768	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0.11819	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.214645	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0.225193	1	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.12188	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.191372	1	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.24458	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.024921	0	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.26538	0	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.10375	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0.126548	0	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.23344	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	-0.20441	0	0	1	0	0	0	0	0	0	0	0	0	0

Appendix 5: Frequency of Tour/Activity Combinations

S-S	1570	B-S-S	8	B-F-F	2	A-SS-F-SS	1	S-S-C	1
SS-SS	463	SS-F-SS	8	B-F-SS	2	A-SS-S-F	1	S-S-S-S-B	1
S-S-S	347	S-S-S-S-S-S-S	7	C-F-SS	2	A-SS-S-SS	1	S-S-S-S-F	1
A-S	166	C-F	6	C-F-SS-SS	2	A-SS-SS-S-B	1	S-S-S-S-S-F	1
F-SS	158	F-B	6	C-SS-SS-S	2	A-SS-SS-SS	1	S-S-S-S-S-S-S-S-S-S-S-S-S-S	1
S-SS	126	S-S-S-S-S-S-S-S	6	F-S-B	2	A-SS-SS-SS-SS	1	S-S-S-S-S-S-S-S-S-S-S-S-S-S-S	1
S-S-S-S	112	S-SS-SS	6	F-SS-F	2	B-B-F	1	S-S-S-S-S-S-S-S-S-S-S-S-S-S-S-S	1
SS-F	99	A-SS-S	5	F-SS-S-SS	2	B-B-SS	1	S-S-SS-F-SS-F	1
SS-S	94	B-SS-SS	5	S-A-SS	2	B-F-C	1	S-S-SS-SS	1
B-S	90	C-SS-SS	5	S-B-B	2	B-S-B	1	S-S-SS-SS-F	1
S-F	71	S-SS-F	5	S-B-B-F	2	B-S-S-S-S	1	S-SS-F-F	1
SS-SS-SS	70	SS-S-S-S	5	S-B-SS-F	2	B-S-S-S-S-S	1	S-SS-SS-SS	1
S-B	69	SS-SS-S-S	5	S-C	2	B-SS-F	1	SS-A	1
A-A	55	SS-SS-S-SS	5	S-S-B-SS	2	C-C	1	SS-A-SS	1
B-SS	53	A-A-SS	4	S-S-S-B	2	C-F-SS-SS-SS	1	SS-F-F	1
A-SS	51	A-A-SS-SS	4	S-S-S-S-S-S-S-S-S-S	2	C-S-S	1	SS-S-B-SS	1
B-F	41	A-F-SS	4	S-S-S-S-S-S-S-S-S-S-S	2	C-SS-S-F	1	SS-S-F	1
S-S-S-S-S	41	B-S-S	4	S-S-S-S-S-S-S-S-S-S-S-S	2	C-SS-S-SS	1	SS-S-S-B	1
F-SS-SS	35	C-S	4	S-S-S-S-S-S-S-S-S-S-S-S-S	2	C-SS-SS-S-B	1	SS-S-S-F	1
A-S-S	29	F-S-S	4	S-S-SS-F	2	C-SS-SS-S-SS	1	SS-S-S-F-SS	1
F-F	28	F-SS-S	4	S-S-SS-F-SS	2	C-SS-SS-SS	1	SS-S-S-S-S	1
B-B	26	S-A	4	SS-B-SS	2	F-B-SS	1	SS-S-S-S-S-F	1
S-S-SS	26	SS-SS-SS-SS-SS	4	SS-F-SS-F	2	F-F-F	1	SS-S-S-S-SS	1
SS-S-S	21	A-S-S-S-S	3	SS-S-B	2	F-S-B-F	1	SS-SS-B	1
C-SS	20	B-S-SS	3	SS-SS-SS-S	2	F-S-SS	1	SS-SS-F-F	1
A-B	19	C-SS-S	3	A-A-B	1	F-SS-SS-B	1	SS-SS-F-F-SS	1
S-S-S-S-S-S	19	F-SS-SS-SS-SS	3	A-A-SS-SS-SS-SS	1	F-SS-SS-S	1	SS-SS-S-B	1
SS-SS-S	19	S-B-F	3	A-B-S	1	F-SS-SS-SS-B	1	SS-SS-S-B-SS	1
SS-B	18	S-B-SS	3	A-C	1	F-SS-SS-SS-B-SS	1	SS-SS-S-S-S-SS	1
S-S-B	16	S-F-F	3	A-F-SS-SS	1	F-SS-SS-SS-S	1	SS-SS-S-S-S-SS-F	1
F-S	15	S-S-S-F	3	A-F-SS-SS-S	1	F-SS-SS-SS-S-SS	1	SS-SS-S-S-S-SS-F-SS	1
SS-SS-SS-SS	13	S-S-S-S-S-S-S-S-S	3	A-S-B	1	F-SS-SS-SS-S-SS-SS	1	SS-SS-S-SS-A	1
F-SS-SS-SS	12	S-S-S-SS	3	A-S-B-SS	1	F-SS-SS-SS-SS-SS	1	SS-SS-S-SS-SS	1
A-F	11	SS-B-F	3	A-S-F	1	F-SS-SS-SS-SS-SS-SS	1	SS-SS-SS-F	1
S-S-F	11	SS-S-S-SS	3	A-S-F-SS	1	S-A-F	1	SS-SS-SS-F-SS	1
SS-SS-F	11	SS-S-SS-SS	3	A-S-S-S-S-S	1	S-A-SS-F	1	SS-SS-SS-S-S	1
SS-S-SS	10	SS-SS-S-S-S	3	A-S-S-S-S-S-S	1	S-B-F-F	1	SS-SS-SS-SS-S	1
A-A-S	9	A-A-A	2	A-S-S-S-S-S-S-S	1	S-B-F-SS	1	SS-SS-SS-SS-S-S	1
A-S-S-S	8	A-A-SS-SS-SS	2	A-SS-B	1	S-F-SS	1	SS-SS-SS-SS-S-S-SS	1
A-SS-SS	8	A-SS-SS-S	2	A-SS-F	1	S-S-B-SS-F	1	SS-SS-SS-SS-SS-SS	1

Appendix 6: Frequent Words for Tour/Activity Types

	Activity				Bike Tour			
	1-STAR	Freq	5-STAR	Freq	1-STAR	Freq	5-STAR	Freq
1	animals	253	safari	787	tour	85	tour	10173
2	cages	61	animals	601	bike	37	bangkok	8196
3	zoo	49	world	525	guide	34	guide	4292
4	park	43	park	436	bangkok	25	bike	3262
5	safari	42	bangkok	418	ride	22	day	2428
6	animal	41	day	409	time	18	trip	2229
7	conditions	40	visit	372	people	17	city	2205
8	visit	37	dolphin	356	bikes	14	time	2076
9	money	35	time	316	trip	14	experience	1901
10	poor	34	experience	300	day	13	ride	1873
11	sad	29	kids	296	tours	12	recommend	1732
12	tigers	29	amazing	250	city	11	night	1474
13	experience	28	zoo	219	found	11	local	1368
14	farm	28	tour	215	minutes	11	bikes	1308
15	food	28	feeding	203	riding	11	people	1280
16	time	27	fun	202	person	10	streets	1269
17	crocodiles	26	thai	194	booked	9	guides	1196
18	baht	25	nice	191	call	9	amazing	1192
19	horrible	25	food	189	company	9	food	1182
20	dirty	24	animal	188	deposit	9	fun	1171

	Cooking Class				Food Tour			
	1-STAR	Freq	5-STAR	Freq	1-STAR	Freq	5-STAR	Freq
1	class	70	cooking	6155	tour	70	tour	7495
2	cooking	60	class	4930	food	63	food	5814
3	school	43	thai	4170	guide	25	guide	3317
4	thai	26	food	3075	thai	12	bangkok	3294
5	told	18	market	2749	street	10	tuk	2879
6	bangkok	16	dishes	2340	tours	9	night	1823
7	email	16	chef	2322	stops	8	thai	1657
8	day	15	experience	2312	dishes	7	time	1369
9	teacher	15	school	2059	experience	7	recommend	1186
10	market	14	ingredients	2018	night	7	fun	1175
11	time	14	bangkok	1958	restaurant	7	experience	1168
12	pm	13	cook	1803	thailand	7	amazing	939
13	arrived	11	time	1739	average	6	local	928
14	classes	11	fun	1560	english	6	highly	889
15	dishes	11	recommend	1383	lot	6	street	810
16	food	11	home	1345	love	6	market	793
17	booked	10	day	1331	pad	6	day	782
18	chef	9	delicious	981	time	6	city	769
19	didn	9	poo	970	ate	5	knowledgeable	691
20	hotel	9	highly	963	chinese	5	trip	681

	Sight Seeing				Spa			
	1-STAR	Freq	5-STAR	Freq	1-STAR	Freq	5-STAR	Freq
1	tour	988	tour	19500	massage	1548	massage	13116
2	guide	389	day	11460	time	288	spa	5831
3	time	333	guide	11035	spa	268	bangkok	3345
4	hotel	250	bangkok	10317	thai	252	thai	3109
5	trip	235	time	6667	experience	222	staff	2995
6	day	232	market	5351	foot	216	clean	2413
7	bangkok	217	trip	5188	service	215	time	2400
8	company	209	tong	5061	hour	190	experience	2299

9	driver	196	experience	4182	staff	189	service	2286
10	market	188	recommend	3848	bangkok	175	professional	2204
11	told	188	amazing	3474	bad	172	nice	2185
12	booked	174	tours	3222	worst	164	friendly	1930
13	minutes	165	thailand	3159	masseuse	134	relaxing	1822
14	didn	154	temple	3140	oil	131	foot	1780
15	hour	149	floating	3056	told	127	recommend	1520
16	people	146	hotel	2982	mins	119	hour	1476
17	boat	145	driver	2905	booked	116	massages	1422
18	hours	137	elephant	2854	didn	111	visit	1355
19	floating	134	river	2847	terrible	111	excellent	1334
20	tours	133	thai	2643	massages	109	body	1331

Appendix 7: Dataset Size for ML Algorithms

'Activity' 5-Star	
Dataset	3,824
Sample Set	200
Training Set	2,677
Testing Set	1,147

'Bike Tour' 5-Star	
Dataset	2,757
Sample Set	200
Training Set	1,931
Testing Set	826

'Cooking Class' 5-Star	
Dataset	1,185
Sample Set	200
Training Set	830
Testing Set	355

'Food Tour' 5-Star	
Dataset	1,770
Sample Set	200
Training Set	1,240
Testing Set	530

'Sight Seeing' 5-Star	
Dataset	8,570
Sample Set	200
Training Set	6,000
Testing Set	2,570

'Spa' 5-Star	
Dataset	17,837
Sample Set	200
Training Set	12,487
Testing Set	5,350

'Activity' 1-Star	
Dataset	742
Sample Set	200
Training Set	520
Testing Set	222

'Bike Tour' 1-Star	
Dataset	88
Sample Set	63
Training Set	63
Testing Set	25

'Cooking Class' 1-Star	
Dataset	73
Sample Set	52
Training Set	52
Testing Set	21

'Food Tour' 1-Star	
Dataset	78
Sample Set	55
Training Set	55
Testing Set	23

'Sight Seeing' 1-Star	
Dataset	1,258
Sample Set	200
Training Set	882
Testing Set	376

'Spa' 1-Star	
Dataset	3,400
Sample Set	200
Training Set	2,380
Testing Set	1,020

Appendix 8: Logistic Regression Results (Predicting 5-Star Reviews)

Activity 5-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.3706	0.099611	-3.72048	0.000199
2	Sentiment	3.712192	0.200803	18.48672	2.64E-76
3	West_Europe	-0.39979	0.137886	-2.89939	0.003739
4	North_America	-0.29948	0.179207	-1.67112	0.094698
5	Southeast_Asia	-0.61643	0.115749	-5.32556	1.01E-07
6	Australasia	-0.56015	0.163548	-3.42499	0.000615
7	South_Asia	0.000545	0.102971	0.005296	0.995774
8	East_Asia	-0.36977	0.232646	-1.58942	0.111966
9	Middle_East	-0.51291	0.18125	-2.82983	0.004657
10	Latin_America	-0.44838	0.653569	-0.68605	0.492683
11	East_Europe	0.528938	0.413901	1.277934	0.201273
12	Africa	-0.19514	0.396822	-0.49176	0.622891
13	W1	-0.19289	0.094676	-2.03731	0.041619
14	W2	-0.22029	0.154975	-1.42143	0.155193
15	W3	0.198222	0.104282	1.90082	0.057326
16	W4	-0.21538	0.09364	-2.3001	0.021443
17	W5	0.146377	0.084175	1.738958	0.082042
18	W6	-0.14091	0.084118	-1.67511	0.093913
19	W7	-0.03935	0.09343	-0.42112	0.673667
20	W8	0.332076	0.103937	3.194978	0.001398
21	W9	-0.3051	0.10535	-2.89609	0.003778
22	W10	-0.17949	0.148476	-1.20891	0.226698

Bike Tour 5-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	0.287321	0.1874	1.533193	0.125228
2	Sentiment	2.096543	0.249612	8.399213	4.49E-17
3	West_Europe	-0.42996	0.171214	-2.51122	0.012031
4	North_America	-0.40107	0.190754	-2.10256	0.035505
5	Southeast_Asia	-0.89651	0.21343	-4.20051	2.66E-05
6	Australasia	-0.6327	0.200355	-3.1579	0.001589
7	South_Asia	-0.80987	0.268668	-3.01438	0.002575
8	East_Asia	-0.64879	0.298714	-2.17194	0.02986
9	Middle_East	-0.9961	0.332642	-2.99452	0.002749
10	Latin_America	-0.16458	0.382113	-0.4307	0.666685
11	East_Europe	-0.10348	0.462682	-0.22365	0.823033
12	Africa	-0.43954	0.403416	-1.08955	0.275911
13	W1	-0.09972	0.099127	-1.006	0.314415
14	W2	0.024832	0.088166	0.281649	0.778212
15	W3	-0.36342	0.084243	-4.314	1.60E-05

16	W4	0.309316	0.092502	3.343896	0.000826
17	W5	0.020928	0.099998	0.209284	0.834227
18	W6	0.180216	0.097883	1.841133	0.065602
19	W7	-0.07744	0.091675	-0.84468	0.398288
20	W8	0.176327	0.096491	1.827388	0.067641
21	W9	-0.21835	0.098811	-2.20973	0.027124
22	W10	0.552052	0.095352	5.789599	7.06E-09

Cooking Class 5-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-1.93993	0.550238	-3.52563	0.000422
2	Sentiment	4.604648	0.463062	9.943917	2.68E-23
3	West_Europe	0.982862	0.51489	1.908878	0.056278
4	North_America	1.439319	0.517946	2.778901	0.005454
5	Southeast_Asia	0.387135	0.541346	0.715134	0.474526
6	Australasia	1.208842	0.532002	2.272251	0.023071
7	South_Asia	0.638208	0.554414	1.151139	0.249675
8	East_Asia	0.535502	0.657563	0.814374	0.415431
9	Middle_East	0.887636	0.714759	1.241868	0.214285
10	Latin_America	1.900571	0.701303	2.710055	0.006727
11	East_Europe	1.346889	0.782733	1.72075	0.085296
12	Africa	0.151687	0.758978	0.199857	0.841592
13	W1	-0.06471	0.206743	-0.31301	0.754275
14	W2	-0.4298	0.143397	-2.99731	0.002724
15	W3	0.134326	0.130746	1.02738	0.304241
16	W4	0.337501	0.134241	2.514145	0.011932
17	W5	0.015986	0.135309	0.118144	0.905954
18	W6	0.148521	0.134845	1.101422	0.270713
19	W7	-0.08968	0.249348	-0.35966	0.719104
20	W8	-0.13835	0.136133	-1.01626	0.309504
21	W9	-0.12463	0.139286	-0.89476	0.370914
22	W10	0.478231	0.145822	3.279547	0.00104

Food Tour 5-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.656	0.195107	-3.36226	0.000773
2	Sentiment	2.709518	0.301606	8.983646	2.62E-19
3	West_Europe	-0.23516	0.162758	-1.44482	0.148508
4	North_America	-0.07733	0.158988	-0.48639	0.626691
5	Southeast_Asia	-0.56377	0.217059	-2.59732	0.009395
6	Australasia	-0.53036	0.194932	-2.72073	0.006514
7	South_Asia	-0.62341	0.365226	-1.70691	0.087839
8	East_Asia	-0.36132	0.351583	-1.0277	0.304091
9	Middle_East	-0.37396	0.4511	-0.829	0.407105

10	Latin_America	0.13614	0.466241	0.291994	0.770291
11	East_Europe	0.273017	0.629902	0.433428	0.664704
12	Africa	-0.07695	0.47722	-0.16125	0.8719
13	W1	0.281205	0.128571	2.187151	0.028732
14	W2	0.019256	0.117942	0.16327	0.870305
15	W3	-0.08225	0.110829	-0.74211	0.458023
16	W4	0.022264	0.139273	0.159857	0.872994
17	W5	0.12171	0.120399	1.010886	0.312071
18	W6	0.113029	0.116365	0.971329	0.331384
19	W7	0.604525	0.120822	5.003444	5.63E-07
20	W8	0.112468	0.122626	0.917162	0.359058
21	W9	0.069754	0.12762	0.546575	0.584671
22	W10	-0.30269	0.12942	-2.33883	0.019344

Sight Seeing 5-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-2.14362	0.080038	-26.7824	5.18E-158
2	Sentiment	4.463202	0.149678	29.81869	2.24E-195
3	West_Europe	1.19798	0.074184	16.14873	1.16E-58
4	North_America	1.48475	0.077715	19.10499	2.29E-81
5	Southeast_Asia	0.483469	0.106413	4.543307	5.54E-06
6	Australasia	0.964704	0.090056	10.71227	8.91E-27
7	South_Asia	0.561556	0.141721	3.962393	7.42E-05
8	East_Asia	1.351271	0.325237	4.154729	3.26E-05
9	Middle_East	0.676693	0.198998	3.400498	0.00067263
10	Latin_America	1.15758	0.232029	4.988944	6.07E-07
11	East_Europe	1.118839	0.256701	4.358539	1.31E-05
12	Africa	0.299798	0.246886	1.214318	0.22462638
13	W1	0.141088	0.059451	2.373179	0.01763571
14	W2	0.524025	0.054998	9.528097	1.60E-21
15	W3	0.096873	0.056621	1.710897	0.08710013
16	W4	-0.10952	0.054406	-2.01301	0.04411366
17	W5	-0.00023	0.06348	-0.00365	0.99708503
18	W6	-0.04757	0.058366	-0.81503	0.41505583
19	W7	0.32358	0.063793	5.07235	3.93E-07
20	W8	0.625149	0.062446	10.01104	1.36E-23
21	W9	0.125602	0.069852	1.798116	0.07215864
22	W10	0.065873	0.062306	1.057249	0.29039789

Spa 5-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.14445	0.050534	-2.85854	0.004256
2	Sentiment	3.641582	0.086591	42.05499	0
3	West_Europe	-0.17282	0.056173	-3.07661	0.002094
4	North_America	0.006879	0.0622	0.110592	0.91194
5	Southeast_Asia	-0.60816	0.047014	-12.9357	2.83E-38
6	Australasia	-0.45416	0.076668	-5.92367	3.15E-09
7	South_Asia	-0.87844	0.066661	-13.1778	1.18E-39
8	East_Asia	-0.29063	0.095601	-3.04001	0.002366
9	Middle_East	-0.19615	0.120097	-1.63322	0.102422
10	Latin_America	0.025653	0.223533	0.11476	0.908635
11	East_Europe	-0.21873	0.183968	-1.18898	0.234447
12	Africa	-0.39153	0.236502	-1.6555	0.097822
13	W1	-0.41259	0.036205	-11.3961	4.37E-30
14	W2	0.065274	0.041376	1.577559	0.114667
15	W3	0.126886	0.047947	2.6464	0.008135
16	W4	0.183777	0.047486	3.870143	0.000109
17	W5	-0.09327	0.048493	-1.92338	0.054432
18	W6	0.117927	0.043231	2.727837	0.006375
19	W7	-0.40181	0.049707	-8.08349	6.29E-16
20	W8	0.486493	0.055705	8.733418	2.47E-18
21	W9	-0.02059	0.045377	-0.45387	0.649924
22	W10	0.124312	0.050665	2.453582	0.014144

Appendix 9: Logistic Regression Results (Predicting 1-Star Reviews)

Activity 1-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.00126	0.30153	-0.00416	0.996678
2	Sentiment	-10.201	0.831952	-12.2616	1.46E-34
3	West_Europe	0.428447	0.365882	1.170996	0.2416
4	North_America	-0.24903	0.545427	-0.45658	0.647971
5	Southeast_Asia	-0.00267	0.370473	-0.00721	0.994249
6	Australasia	0.506637	0.440776	1.149421	0.250382
7	South_Asia	-1.62865	0.424098	-3.84028	0.000123
8	East_Asia	-0.78521	0.861234	-0.91173	0.361913
9	Middle_East	-1.16864	0.698264	-1.67363	0.094203
10	Latin_America	-11.9485	561.1954	-0.02129	0.983013
11	East_Europe	0.256723	1.006203	0.255141	0.798615
12	Africa	-0.7573	1.655686	-0.45739	0.64739
13	W1	0.49436	0.54988	0.899032	0.368636
14	W2	-0.34361	0.346738	-0.99098	0.321695
15	W3	0.004808	0.335369	0.014337	0.988561
16	W4	-0.94504	0.368646	-2.56355	0.010361
17	W5	0.44448	0.540966	0.821641	0.411281
18	W6	0.425774	0.300172	1.418437	0.156063
19	W7	1.024268	0.526613	1.945009	0.051774
20	W8	-0.586	0.40148	-1.45961	0.144398
21	W9	0.214674	0.475054	0.451894	0.651345
22	W10	0.103885	0.322301	0.322323	0.747208

Bike Tour 1-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	3.183655	2.078502	1.531706	0.125595
2	Sentiment	-19.3439	5.058874	-3.82376	0.000131
3	West_Europe	-0.18286	1.54298	-0.11851	0.905661
4	North_America	1.072469	1.667008	0.64335	0.519997
5	Southeast_Asia	1.797945	2.274146	0.790602	0.429176
6	Australasia	1.118577	1.770267	0.631869	0.527472
7	South_Asia	16.30749	6522.639	0.0025	0.998005
8	East_Asia	-16.6947	6522.639	-0.00256	0.997958
9	Middle_East	17.99385	6522.639	0.002759	0.997799
10	Latin_America	-13.4808	4406.51	-0.00306	0.997559
11	East_Europe	-14.7009	6522.639	-0.00225	0.998202
12	Africa	-0.31512	1.540466	-0.20456	0.837914
13	W1	-1.20681	1.091618	-1.10553	0.268931
14	W2	-0.6072	1.051873	-0.57725	0.563768
15	W3	-2.17558	1.150253	-1.89139	0.058572

16	W4	-0.57518	1.201675	-0.47865	0.63219
17	W5	-0.04712	0.979732	-0.0481	0.961639
18	W6	1.903938	1.39167	1.368096	0.171282
19	W7	0.564147	1.237896	0.45573	0.648584
20	W8	0.765896	1.040399	0.736156	0.461636
21	W9	-0.44693	1.104528	-0.40463	0.685748
22	W10	3.183655	2.078502	1.531706	0.125595

Cooking Class 1-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-26.5661	145852.6	-0.00018	0.999855
2	Sentiment	-3.13E-09	225523.5	-1.39E-14	1
3	West_Europe	53.13213	194411	0.000273	0.999782
4	North_America	53.13214	175606.9	0.000303	0.999759
5	Southeast_Asia	53.13214	235596.1	0.000226	0.99982
6	Australasia	53.13213	212095.2	0.000251	0.9998
7	South_Asia	53.13214	411458.6	0.000129	0.999897
8	East_Asia	53.13213	401594	0.000132	0.999894
9	Middle_East	53.13213	409774.4	0.00013	0.999897
10	Latin_America	53.13213	398776.1	0.000133	0.999894
11	East_Europe	-4.36E-09	103121.7	-4.23E-14	1
12	Africa	1.01E-08	105711.5	9.59E-14	1
13	W1	-3.14E-10	122059.2	-2.57E-15	1
14	W2	7.68E-09	138867.9	5.53E-14	1
15	W3	-2.45E-08	128015.8	-1.92E-13	1
16	W4	-1.28E-09	106746.8	-1.19E-14	1
17	W5	-9.49E-09	100572.5	-9.44E-14	1
18	W6	1.24E-09	102149.5	1.21E-14	1
19	W7	8.83E-09	103182.6	8.56E-14	1
20	W8	6.50E-09	131292.4	4.95E-14	1
21	W9	-26.5661	145852.6	-0.00018	0.999855
22	W10	-3.13E-09	225523.5	-1.39E-14	1

Food Tour 1-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	-21.3037	148891.4	-0.00014	0.999886
2	Sentiment	-327.73	200171.9	-0.00164	0.998694
3	West_Europe	88.92779	262447.7	0.000339	0.99973
4	North_America	194.1244	137515.6	0.001412	0.998874
5	Southeast_Asia	122.7373	136613.8	0.000898	0.999283
6	Australasia	13.82323	240515.7	5.75E-05	0.999954
7	South_Asia	88.74539	275307.3	0.000322	0.999743
8	East_Asia	140.506	452658.5	0.00031	0.999752
9	Middle_East	10.51967	151912.6	6.92E-05	0.999945
10	Latin_America	37.15047	173574	0.000214	0.999829
11	East_Europe	74.63727	77798.12	0.000959	0.999235
12	Africa	-88.684	186259.1	-0.00048	0.99962
13	W1	-60.0503	108459.8	-0.00055	0.999558
14	W2	-50.8514	65593.48	-0.00078	0.999381
15	W3	35.3376	107083.6	0.00033	0.999737
16	W4	12.37671	232908.1	5.31E-05	0.999958
17	W5	-70.6955	107850.1	-0.00066	0.999477
18	W6	-5.86259	173828.5	-3.37E-05	0.999973
19	W7	-21.3037	148891.4	-0.00014	0.999886
20	W8	-327.73	200171.9	-0.00164	0.998694
21	W9	88.92779	262447.7	0.000339	0.99973
22	W10	194.1244	137515.6	0.001412	0.998874

Sight Seeing 1-Star Prediction					
	term	estimate	std.error	statistic	p.value
1	(Intercept)	0.985727	0.243453	4.048934	5.15E-05
2	Sentiment	-11.5322	0.664027	-17.3671	1.47E-67
3	West_Europe	0.231819	0.251595	0.921394	0.356845
4	North_America	-0.12519	0.257466	-0.48624	0.626798
5	Southeast_Asia	0.235338	0.419237	0.56135	0.574559
6	Australasia	-0.31849	0.316399	-1.0066	0.314126
7	South_Asia	-0.38304	0.545767	-0.70184	0.482777
8	East_Asia	0.67191	0.711556	0.944283	0.345025
9	Middle_East	-0.1307	0.65515	-0.19949	0.841879
10	Latin_America	0.713368	0.720357	0.990298	0.322029
11	East_Europe	-0.87839	0.83955	-1.04626	0.295441
12	Africa	0.784342	0.746498	1.050696	0.293398
13	W1	0.634096	0.207143	3.061152	0.002205
14	W2	-0.02285	0.190144	-0.12017	0.904352
15	W3	0.026491	0.18306	0.144711	0.884939
16	W4	0.495077	0.215169	2.300875	0.021399
17	W5	0.118922	0.195953	0.606889	0.543925
18	W6	-0.77983	0.191774	-4.06641	4.77E-05

19	W7	0.045976	0.23042	0.199532	0.841847
20	W8	-0.30245	0.328266	-0.92136	0.356863
21	W9	-0.11346	0.228027	-0.49756	0.618795
22	W10	0.151106	0.368041	0.410569	0.681389

Spa 1-Star Prediction

	term	estimate	std.error	statistic	p.value
1	(Intercept)	0.8402	0.132629	6.334977	2.37E-10
2	Sentiment	-9.01934	0.317213	-28.4331	########
3	West_Europe	-0.20166	0.160023	-1.2602	0.207596
4	North_America	-0.84521	0.190664	-4.43297	9.29E-06
5	Southeast_Asia	-0.04695	0.12908	-0.36369	0.716088
6	Australasia	-0.40383	0.230046	-1.75545	0.079183
7	South_Asia	-0.00108	0.194011	-0.00557	0.99556
8	East_Asia	-0.20396	0.267393	-0.76278	0.445594
9	Middle_East	-0.51247	0.395987	-1.29417	0.195608
10	Latin_America	-0.63967	0.78532	-0.81453	0.415342
11	East_Europe	-0.04058	0.45949	-0.08831	0.929628
12	Africa	-0.29217	0.607235	-0.48114	0.630416
13	W1	0.669569	0.112165	5.969495	2.38E-09
14	W2	0.108461	0.116265	0.932875	0.350885
15	W3	-0.00066	0.132606	-0.00501	0.996004
16	W4	0.108597	0.131681	0.824694	0.409545
17	W5	-0.23412	0.145941	-1.60421	0.108667
18	W6	0.266302	0.144499	1.842929	0.065339
19	W7	-0.02457	0.128696	-0.19091	0.848599
20	W8	-0.27997	0.150431	-1.86113	0.062726
21	W9	0.241854	0.177345	1.363744	0.172648
22	W10	-0.33024	0.178304	-1.85211	0.06401

Appendix 10: Logistic Regression Effectiveness (Predicting 5-Star Reviews)

Activity 5-Star Prediction		
	No	Yes
No	399	205
Yes	200	343
F1 Score	**62.9%**	
Accuracy	64.7%	
Precision	63.2%	
Recall	62.6%	
Specificity	66.6%	

Bike Tour 5-Star Prediction		
	No	Yes
No	106	84
Yes	224	412
F1 Score	72.8%	
Accuracy	62.7%	
Precision	64.8%	
Recall	83.1%	
Specificity	32.1%	

Cooking Class 5-Star Prediction		
	No	Yes
No	73	47
Yes	69	166
F1 Score	74.1%	
Accuracy	67.3%	
Precision	70.6%	
Recall	77.9%	
Specificity	51.4%	

Food Tour 5-Star Prediction		
	No	Yes
No	84	44
Yes	128	274
F1 Score	76.1%	
Accuracy	67.5%	
Precision	68.2%	
Recall	86.2%	
Specificity	39.6%	

Sight Seeing 5-Star Prediction		
	No	Yes
No	594	221
Yes	434	1321
F1 Score	80.1%	
Accuracy	74.5%	
Precision	75.3%	
Recall	85.7%	
Specificity	57.8%	

Spa 5-Star Prediction		
	No	Yes
No	1092	549
Yes	1048	2661
F1 Score	76.9%	
Accuracy	70.1%	
Precision	71.7%	
Recall	82.9%	
Specificity	51.0%	

Appendix 11: Logistic Regression Effectiveness (Predicting 1-Star Reviews)

Activity 1-Star Prediction		
	No	Yes
No	121	18
Yes	12	71

F1 Score	82.6%	
Accuracy	86.5%	
Precision	85.5%	
Recall	79.8%	
Specificity	91.0%	

Bike Tour 1-Star Prediction		
	No	Yes
No	14	5
Yes	1	5

F1 Score	62.5%	
Accuracy	76.0%	
Precision	83.3%	
Recall	50.0%	
Specificity	93.3%	

Cooking Class 1-Star Prediction		
	No	Yes
No	13	1
Yes	0	7

F1 Score	93.3%	
Accuracy	95.2%	
Precision	100.0%	
Recall	87.5%	
Specificity	100.0%	

Food Tour 1-Star Prediction		
	No	Yes
No	11	1
Yes	3	8

F1 Score	80.0%	
Accuracy	82.6%	
Precision	72.7%	
Recall	88.9%	
Specificity	78.6%	

Sight Seeing 1-Star Prediction		
	No	Yes
No	202	28
Yes	24	122

F1 Score	82.4%	
Accuracy	86.2%	
Precision	83.6%	
Recall	81.3%	
Specificity	89.4%	

Spa 1-Star Prediction		
	No	Yes
No	1092	549
Yes	1048	2661

F1 Score	76.9%	
Accuracy	70.1%	
Precision	71.7%	
Recall	82.9%	
Specificity	51.0%	

Appendix 12: Logistic Regression Results Top 20 vs. Top 10 Highest-Occurring Words

Activity 5-Star Prediction (Top 20)		
	No	Yes
No	1,034	45
Yes	24	44
F1 Score	**56.1%**	
Accuracy	94.0%	
Precision	64.7%	
Recall	49.4%	
Specificity	97.7%	

Activity 5-Star Prediction (Top 10)		
	No	Yes
No	399	205
Yes	200	343
F1 Score	**62.9%**	
Accuracy	64.7%	
Precision	63.2%	
Recall	62.6%	
Specificity	66.6%	

Bike Tour 5-Star Prediction (Top 20)		
	No	Yes
No	2,380	9
Yes	5	1
F1 Score	**12.5%**	
Accuracy	99.4%	
Precision	16.7%	
Recall	10.0%	
Specificity	99.8%	

Bike Tour 5-Star Prediction (Top 10)		
	No	Yes
No	106	84
Yes	224	412
F1 Score	**72.8%**	
Accuracy	62.7%	
Precision	64.8%	
Recall	83.1%	
Specificity	32.1%	

Cooking Class 5-Star Prediction (Top 20)		
	No	Yes
No	1,942	9
Yes	6	2
F1 Score	**21.1%**	
Accuracy	99.2%	
Precision	25.0%	
Recall	18.2%	
Specificity	99.7%	

Cooking Class 5-Star Prediction (Top 10)		
	No	Yes
No	73	47
Yes	69	166
F1 Score	**74.1%**	
Accuracy	67.3%	
Precision	70.6%	
Recall	77.9%	
Specificity	51.4%	

Food Tour 5-Star Prediction (Top 20)		
	No	Yes
No	2,138	11
Yes	4	1

F1 Score	**11.8%**	
Accuracy	99.3%	
Precision	20.0%	
Recall	8.3%	
Specificity	99.8%	

Food Tour 5-Star Prediction (Top 10)		
	No	Yes
No	84	44
Yes	128	274

F1 Score	**76.1%**	
Accuracy	67.5%	
Precision	68.2%	
Recall	86.2%	
Specificity	39.6%	

Sight Seeing 5-Star Prediction (Top 20)		
	No	Yes
No	6,099	166
Yes	29	35

F1 Score	**26.4%**	
Accuracy	96.9%	
Precision	54.7%	
Recall	17.4%	
Specificity	99.5%	

Sight Seeing 5-Star Prediction (Top 10)		
	No	Yes
No	594	221
Yes	434	1321

F1 Score	80.1%	
Accuracy	74.5%	
Precision	75.3%	
Recall	85.7%	
Specificity	57.8%	

Spa 5-Star Prediction (Top 20)		
	No	Yes
No	11682	378
Yes	114	170

F1 Score	**40.9%**	
Accuracy	96.0%	
Precision	59.9%	
Recall	31.0%	
Specificity	99.0%	

Spa 5-Star Prediction (Top 10)		
	No	Yes
No	1092	549
Yes	1048	2661

F1 Score	76.9%	
Accuracy	70.1%	
Precision	71.7%	
Recall	82.9%	
Specificity	51.0%	

REFERENCES

[1] "Travel & Tourism Economic Impact 2019," 2019.
[2] "International Tourism Growth Continues to Outpace the Global Economy," ed, 2020.
[3] "International Tourism Hightlights, 2019 Edition," UNWTO, Madrid, 2019.
[4] "Thailand ranks 10th most popular for global visitors," ed, 2018.
[5] N. E. i. Bangkok, "Tourism industry in Thailand," Netherlands Embassy in Bangkok, Bangkok, 2019.
[6] C. Theparat, "Prayut: Zones vital for growth," ed, 2019.
[7] (2017, March) The Second National Tourism Development Plan (2017-2021).
[8] "TAT outlines its Action Plan 2019," ed, 2018.
[9] MarketResearch, "Thailand Tourism Q2 2020," BMI Research, 2020.
[10] "Bangkok is the world's most visited city again, fourth year in a row," ed, 2019.
[11] "New skills, wellness, and family activities are travellers' focus for 2019," ed, 2019.
[12] "TAT targets 3.18 trillion Baht in tourism revenue for Thailand in 2020," ed, 2019.
[13] S. C. B. E. I. Center, "Insight Three megatrends to change the face of the Thai tourism industry," Siam Commerical Bank, Bangkok, 2017.
[14] K. J. Lancaster, "A new approach to consumer theory," *Journal of Political Economy, vol. 74,* pp. 132-157, 1966.
[15] "Data never sleeps," ed, 2010.
[16] "40 Trillion Gigabytes of Data by 2020," ed, 2013.
[17] T. Segal, "Big Data," ed, 2019.
[18] M. Cox and D. Ellsworth, "Managing big data for scientific visualization," *ACM Siggraph, MRJ/NASA Ames Research Center,* pp. 1-17, 1997.
[19] "The five V's of big data," ed, 2017.
[20] M. Mariani, "Big Data and analytics in tourism and hospitality: a perspective article," *Tourism Review,* pp. 299-303, 2020.
[21] E. Marine-Roig and S. A. Clavé, "Tourism analytics with massive user-generated content: A case study of Barcelona," *Journal of Destination Marketing and Management,* 2015.
[22] P. R. Berthon, L. F. Pitt, K. Plangger, and D. Shapiro, "Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy," *Business Horizons 55,* pp. 261-271, 2012.
[23] E. Constantinides and S. J. Fountain, "Web 2.0: Conceptual foundations and marketing issues," *J Direct Data Digit Mark Pract 9,* pp. 231–244, 2008.
[24] D. Zarella, *The social media marketing book.* North Sebastopol, CA: O'Railly Media, 2010.
[25] K. Lyu and H. Kim, "Sentiment analysis using word polarity of social media," *Wireless Personal Communications,* pp. 941-958, 2016.
[26] K. Leetaru, "How Big Is Social Media And Does It Really Count As 'Big Data'?," ed, 2010.
[27] S. Litvin, R. E. Goldsmith, and B. Pan, "Electronic word-of-mouth in hospitality and tourism management," *Tourism Management 29,* pp. 458-468, 2008.
[28] R. C. Lewis and R. E. Chambers, "Marketing leadership in hospitality,

foundations and practices," *New York, Wiley,* 2000.
[29] B. Zeng and R. Gerritsen, "What do we know about social media in tourism? A review," *Tourism Management Perspectives,* pp. 27-36, 2014.
[30] J. Fotis, "Discussion of the impacts of social media in leisure tourism: “The impact of social media on consumer behaviour: Focus on leisure travel," ed, 2012.
[31] G. D. Chiappa, "Trustworthiness of Travel 2.0 applications and their influence on tourist behaviour: an empirical investigation in Italy," *Information and Communication Technologies in Tourism*, 2011.
[32] H. Li, Q. Ye, and R. Law, "Determinants of Customer Satisfaction in the Hotel Industry: An Application of Online Review Analysis," *Asia Pacific Journal of Tourism Research, 18:7,* pp. 784-802, 2013.
[33] S. Banerjee and A. Y. K. Chua, "In search of patterns among travellers' hotel ratings in TripAdvisor," *Tourism Management 53,* pp. 125-131, 2016.
[34] Y. Guo, S. J. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation," *Tourism Management 59,* pp. 467-483, 2017.
[35] A. Barreda and A. Bilgihan, "An analysis of user-generated content for hotel experiences," *Journal of Hospitality and Tourism Technology, 4:3,* pp. 263-280, 2013.
[36] K. Berezina, A. Bilgihan, C. Cobanoglu, and F. Okumus, "Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews," *Journal of Hospitality Marketing & Management 25,* pp. 1-24, 2016.
[37] M. Geetha, P. Singha, and S. Sinha, "Relationship between customer sentiment and online customerratings for hotels - An empirical analysis," *Tourism Management 61,* pp. 43-54, 2017.
[38] Z. Xiang, Q. Du, Y. Ma, and W. Fang, "A comparative analysis of major online review platforms: Implicationsfor social media analytics in hospitality and tourism," *Tourism Management 58,* pp. 51-65, 2017.
[39] B. H. Ye, J. M. Luo, and H. Q. Vu, "Spatial and temporal analysis of accommodation preference based on onlinereviews," *Journal of Destination Marketing & Management 9,* pp. 288-299, 2018.
[40] J.-W. Bi, Y. Liu, Z.-P. Fan, and J. Zhang, "Exploring asymmetric effects of attribute performance on customer satisfaction in the hotel industry," *Tourism Management 77,* p. 104006, 2020.
[41] X. Cheng, S. Fu, J. Sun, A. Bilgihan, and F. Okumus, "An investigation on online reviews in sharing economy driven hospitality platforms: A viewpoint of trust," *Tourism Management 71,* pp. 366-377, 2019.
[42] V. Taecharungroj and B. Mathayomchan, "Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand," *Tourism Management 75,* pp. 550-568, 2019.
[43] A. P. Kirilenko, S. O. Stepchenkova, and J. M. Hernandez, "Comparative clustering of destination attractions for different origin markets with network and spatial analyses of online reviews," *Tourism Management 72,* pp. 400-410, 2019.
[44] B. Fang, D. Kucukusta, and R. Law, "Analysis of the perceived value of online tourism reviews: Influence ofreadability and reviewer characteristics," *Tourism Management 52,* pp. 498-506, 2016.

[45] M. I. Simeon, P. Buonincontri, F. Cinquegrani, and A. Martone, "Exploring tourists' cultural experiences in Naples through online reviews," *Journal of Hospitality and Tourism Technology, 8:2,* pp. 220-238, 2017.

[46] J. Miguéns, R. Baggio, and C. Costa, "Social media and Tourism Destinations: TripAdvisor Case Study," *Advances in Tourism Research,* pp. 26-28, 2008.

[47] "About TripAdvisor," ed, 2019.

[48] "TripAdvisor Internal Logs: Average Monthly Unique Visitors," TripAdvisor, 2019.

[49] P. O'Connor, "User-Generated Content and Travel: A Case Study on Tripadvisor.Com," *Information and Communication Technologies in Tourism,* pp. 47-58, 2008.

[50] C. Reiter, "Travel Web sites clamp down on bogus reviews," *International Herald Tribune,* p. 12, 2007.

[51] D. Michie, ""Memo" Functions and Machine Learning," *Nature,* vol. 218, pp. 19-22, 1968.

[52] M. Shafiq, X. Yu, A. A. Langhari, N. K. Karn, and F. Abdessamia, "Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms," in *2nd IEEE International Conference on Computer and Communications* 2016.

[53] M. J. Singh and A. Girdhar, "Fingerprint Enhancement Using Wavelet Transformation and Differential Support Vector Machine," in *International Conference on Inventive Research in Computing Applications,* Coimbatore, India, 2018.

[54] A. A. S. Kingsly and J. Mahil, "Effective approach of learning based classifiers for skin cancer diagnosis from dermoscopy images," *International Journal of Advanced Science and Technology,* vol. 28, no. 20, pp. 1016-1026, 2019.

[55] A. A. Wadhe and S. S. Suratkar, "Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques," *International Conference on Industry 4.0 Technology (I4Tech),* pp. 1-6, 2020.

[56] E. E. Fernandes, G. A. Sarriés, M. A. Bacchi, Y. T. Mazola, C. L. Gonzaga, and S. R. V. Sarriés, "Trace elements and machine learning for Brazilian beef traceability," *Food Chemistry,* vol. 333, 2020.

[57] P. Kumari and D. Toshniwal, "Extreme gradient boosting and deep neural network based ensemblelearning approach to forecast hourly solar irradiance," *Journal of Cleaner Production,* vol. 279, 2021.

[58] J. T. Jebaseeli, R. Venkatesan, and K. Ramalakshmi, "Advances in Intelligent Systems and Computing," *Advances in Intelligent Systems and Computing,* vol. 1167, pp. 189-197, 2021.

[59] A. Upadhyay, U. Palival, and S. Jaiswal, "Early brain tumor detection using random forest classification," *Advances in Intelligent Systems and Computing,* vol. 1180, pp. 258-264, 2021.

[60] J. Gajwani and P. Chakraborty, "Students' Performance Prediction Using Feature Selection and Supervised Machine Learning Algorithms," *Advances in Intelligent Systems and Computing,* vol. 1165, pp. 347-354, 2021.

[61] Y. Feng, Cui, N., Zhang, Q., Zhao, L., Gong, D., "Comparison of artificial intelligence," *International Journal of Hydrogen Energy,* vol. 42, no. 21, pp. 14418-14428, 2017.

[62] "A free web scraper that is easy to use," ed.
[63] S. Remanan, "Association Rule Mining," ed: Towards Data Science, 2018.
[64] M. Hahslet, C. Buchta, B. Gruen, K. Hornik, I. Johnson, and C. Borgelt, "arules: Mining Association Rules and Frequent Itemsets." [Online]. Available: https://cran.r-project.org/web/packages/arules/index.html
[65] M. Hahsler, G. Tyler, and S. Chelluboina, "arulesViz: Visualizing Association Rules and Frequent Itemsets."
[66] "Sentiment Analysis," ed: MonkeyLearn, 2020.
[67] T. Rinker, "Package 'sentimentr'."
[68] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in *Discovery and Data Mining (KDD-2004)*, Seattle, Washington, USA, 2004.
[69] K. Benoit and K. Watanabe, "stopwords: Multilingual Stopword Lists." [Online]. Available: https://cran.r-project.org/web/packages/stopwords/index.html
[70] S. Bleier, "NLTK's list of english stopwords." [Online]. Available: https://gist.github.com/sebleier/554280
[71] "Onix Text Retrieval Toolkit API Reference." [Online]. Available: http://www.lextek.com/manuals/onix/stopwords1.html
[72]
[73] T. Shah, "About Train, Validation and Test Sets in Machine Learning," ed: Towards Data Science, 2017.
[74] M. Kuhn, "The caret Package." [Online]. Available: http://topepo.github.io/caret/index.html
[75] K. , "K-Fold Cross Validation," ed: Medium, 2018.
[76] S. Prabhakaran, "Logistic Regression," ed, 2016.
[77] J. Brownlee, "A Gentle Introduction to Imbalanced Classification," ed, 2019.
[78] J. M. Hilbe, *Logistic Regression Models (Chapman & Hall/CRC Texts in Statistical Science) 1st Edition*. Boca Raton: Taylor & Francis Group, LLC, 2009.
[79] "glm," ed.
[80] D. Meyer, "svm," ed: RDocumentation.

VITA

NAME	Naina Chugh
DATE OF BIRTH	28 March 1994
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	International School of Engineering of Chulalongkorn University, Sasin School of Management, Wharton School of the University of Pennsylvania
HOME ADDRESS	23 Phanitchayakan Thon Buri 9 Alley, Wat Tha Phra, Bangkok Yai, Bangkok 10600