Erasmus University Rotterdam
Make it happen.

Master Thesis Data Science and Marketing Analytics

Does speed matter for ultrafast grocery delivery companies in customers' view? An empirical text analysis of customers' online reviews in the UK market

Student Name: Jenny Phan

Student ID number: 578079

Supervisor: Andreas Alfons

Second Assessor: Michiel Van Crombrugge

Date: August 29th 2022

ABSTRACT

Customer opinions on ultrafast grocery delivery is a fairly new research area, despite the industry's massive growth in the last few years. The goal of this study is to understand the criteria and expectations of customers in this industry by extracting the reviews, classifying the review text into different themes using the topic modeling method Latent Dirichlet Allocation (LDA), and analyzing the potential relationships between those themes and the customers' ratings with three predictive models being Logistic Regression, Random Forest, and Support Vector Machine. The outcome of LDA classifies 15 latent topics that cover 5 main themes: (1) Delivery Speed, (2) Customer Service, (3) App Convenience, (4) Price, and (5) Food Quality. Among these features, delivery speed and customer service are considered the most influential variables that can impact customers' ratings since they are consistently identified in all three models.

Keywords: ultrafast grocery delivery, e-grocery, customer online review, topic modeling, Latent Dirichlet Allocation, Logistic Regression, Random Forest, Support Vector Machine, feature importance

ACKNOWLEDGEMENT

It all started in April 2020, 4 months after the COVID-19 pandemic and two months after Europe went into a complete lockdown when my friend asked me whether I wanted to pursue a master's program next to my full-time job. With the possibility of online education back then, I decided that it could be an opportunity for me to have both, and that's how my master's journey started at Erasmus University. People have asked me how I managed to maintain both, and my first answer is: "I don't know" but deep down I know I have the answer to that question. It's the support and motivation from my beloved friends, colleagues, and family. Without them, I wouldn't be sitting here writing this piece of acknowledgment, marking the final stage of my two years journey.

First of all, my sincere appreciation goes to my supervisor, Andreas Alfons. His prompt response to my emails, his constructive answer to my questions, and his encouragement during my low time have greatly motivated me during my thesis time. Thank you! Next, I would like to thank my second reader, Michiel Van Crombrugge, for his time and effort in reviewing my paper. Also, I want to take the chance to show my genuine appreciation to my company managers, Aydin Bonabi and Peter Garcia. Thank you for having my back at work when I got busy at school. Thank you for granting me all the flexibility I need to balance my schedule between work and study. And lastly, to my family, thank you for always being my infinite source of support!

-In every obstacle, there is an opportunity-

Table of Content

List of Figures

List of Tables

1. INTRODUCTION

1.1. Introduction

The term "online grocery delivery services" was initially coined back in 2010' with the growth of digitalization and the internet of things movement. However, it was until 2020, when the COVID-19 pandemic gave a push to the adaptation of the grocery shopping habits, that grocery delivery became popular among consumers of the most fundamental industry – the grocery industry. The term ultrafast grocery delivery or rapid grocery delivery was also mentioned more often and since then has become a hot topic for different groups of people, especially economists and investors. In giant metropolitan cities in Europe and the US, it is not hard to notice those drivers wearing uniforms biking or driving with a big bag behind, rushing to the customers' delivery address within 10-15 mins. A few start-up names in the European markets include Gorillas, Getir, Zapp, Beelivery, GoPuff, Weezy, Flink, etc.

Much skepticism has arisen about the abnormal growth of such an industry. People are skeptical that the COVID-19 situation forced such a service and that it will not change people's grocery habits completely. Despite all the questions and doubts, it is undeniable that the speedy grocery delivery industry is growing unprecedently fast, with a prediction that the global online grocery market size is expected to reach $1.1 trillion by 2027, and the annual growth rate will be approximately 24.8% during that period according to a market report conducted by Grand View Research in 2020[1]. A news article from the Wall Street Journal indicated that more than $14 billion was pumped into this industry just in 2021, accounting for more than half of the total $39.3 billion invested in 2021 across the food tech sector[2]. Getir, one of the start-ups in this market founded in 2015, is now rising as one of the new unicorns. By the end of 2021, Getir was valued at $7.7 billion, and just three months later, they successfully raised another funding round of roughly $800 million in March 2022, bringing the company value to

[1] The report can be accessed via: https://www.grandviewresearch.com/industry-analysis/online-grocery-market

[2] The article can be accessed via: https://www.wsj.com/articles/food-delivery-startups-look-for-new-ways-to-sustain-growth-11644873173

about $11.8 billion according to a report from Bloomberg[3]. So, what makes the investors and market researchers have such confidence in the industry's future? One of the factors that can be considered is speed since it is one of the business values that makes part of the name of the industry: the speed online grocery delivery industry. Speed is believed to be the factor that makes a distinction between companies like Getir, Gorillas and traditional grocery companies like Amazon grocery and Picnic. The question is, do customers have the same view? Is speed one of the critical factors that add actual value for customers? This study will answer that question by examining and identifying the key service features that are deemed important according to customers by analyzing their online reviews.

1.2. Research Questions

The goal of this study is to analyze wo what extent online reviews about grocery delivery services can help us to understand customers' preferences for different services attributes and how customers make trade-offs between those factors, based on which I can come up with a conclusion of the essential factors that contribute to customers' value and satisfaction. The level of satisfaction, in this case, is measured by the review rating, and therefore, the main research question is formulated as follows:

"What key factors of an ultrafast grocery delivery company contribute to the customer's higher rating review?

To answer the main question, we break down the topic into two sub-questions:

1. Which product features are evaluated in the product reviews?
2. What are the relationships between the identified features and the customer rating?

By answering the first sub-question, the scope of features that deem important to the customers can be identified, based on which an analysis of the relationship between each feature and the customer's review rating can be performed to find out which factors among these have the most influence on the rating itself. Overall, this study aims

[3] The article can be accessed via: https://www.bloomberg.com/news/articles/2022-03-11/turkey-s-getir-nears-mubadala-led-funding-at-11-8-billion-value?srnd=technology-vp

to bring empirical insight into customer views on a highly new industry where many market questions are still not answered. The result of this study expects to be a valuable source of information for companies to understand the actual needs of customers from the rapid delivery service, based on which companies' senior management can drive their strategic planning, including operational and marketing strategies, in the direction that could bring the most value for customers and ultimately win over their satisfaction and engagement with the companies and the industry.

The remainder of the paper is structured as follows. Firstly, section 2 will elaborate on the marketing and technical theoretical background of this study. Also, the conceptual framework and the review of relevant literature is conducted in this section. In section 3, I will cover the methodology of different methods that are applied in the study. Then section 4 describes the dataset used in this study and the data preparation steps involved. Next, the result of the LDA and the predictive models is presented in section 5. Lastly, section 6 will be used for a conclusion and further discussion.

2. THEORETICAL BACKGROUND

2.1. Marketing Perspective

2.1.1. Online customer review and customer satisfaction

"What people really desire are not products but satisfying experiences" (Lawrence, 1955). This notion stated by Lawrence highlights the importance of not just cognitive but also emotional, sensory, and social aspects of customer decision making and experience. The customers focus on more than just the product to decide the purchase. According to De Keyser (2015), customer experience is described as "comprised of the cognitive, emotional, physical, sensorial, spiritual, and social elements that mark the customer's direct or indirect interaction with the supplier" (De Keyser, Lemon, & Keiningham, 2015). The customers' evaluation of their experience during the journey is believed to be a key influence on the customer satisfaction and ultimately, customer profitability (Bolton R., 1998) (Bolton, Lemon, & Verhoef, 2004). Thus, understanding which factors influence the customer experience will eventually help companies know what aspects of the business should be focused on. One way to get information on the customer experience is to ask for customer reactions to the company's product or service offerings. A customer satisfaction survey is an example of how firms reach out and ask for customer reactions, which includes measurements of customer emotions and has become a common practice in marketing (Bolton R., 1998) (Westbrook & Oliver, 1991). But over time, surveys start showing its drawback. The customer survey can potentially ask the wrong questions to customers since the questions in the survey are assumed to be important topics that are deemed to be relevant and essential to the customers. That is quite an assumption from the company, simply because not all customers share the same interest in all topics. While price and product quality might be driving factors for one customer, they might not be the key factors that drive satisfaction for another. Instead of having customers go through all the aspects, which can be very lengthy, marketing researchers have been exploring alternatives for obtaining similar data in a more efficient and thoughtful way. Research conducted by Rese (2014) suggested that surveys can potentially be replaced by analyzing publicly

available online reviews (Rese, Schreiber, & Baier, 2014). Similarly, several studies have demonstrated that user-generated content is a rich and reliable source of information for extracting and analyzing customer opinion and satisfaction (Xie, 2011) (Ye, Law, & Gu, 2009). With the growth of the internet and digitalization, online reviews have quickly become one of the most popular tools for companies to explore customer behavior and understand their customer satisfaction driving factors, especially in an industry like rapid e-grocery where everything is done via a shopping application digitally.

2.1.2. Online customer review and customer journey

If online customer review is a reflection of customer satisfaction, the satisfaction is then considered a reflection or evaluation of the customer's experience during the whole purchasing journey. Generally, a customer experience journey can be conceptualized in three phases: pre-purchase, purchase, and post-purchase (Howard & Sheth, 1969) (Pucinelli, Goodstein, & Grewal, 2009). The first phase, pre-purchase, describes the customer's experience before purchase, which could entail customers' need recognition, company research, and purchase consideration (Pieters, Baumgartner, & Allen, 1995). This is also the stage when the customer starts building up their expectation on the aspects to receive once they decide to purchase the product or service. Then the customer moves on to the second phase – purchase – which encompasses the customer interactions with the firm throughout the purchasing process and the associated environment during that purchase event (Bitner, 1990). With the expectation in mind, the customer enters the purchase stage to experience the reality. The closer the 'reality' experience compared to the anticipation, the higher the predicted satisfaction and vice versa; any significant gap that comes out during this purchase phase will be the trigger for lower satisfaction. That similarity or gap will be the input for the customer in their online customer review in the last stage – post-purchase (Doorn, 2010). The ultimate goal of the customer online review is to reflect the actual purchasing experience compared to their initial expectations. From the visualization standpoint, the customer journey, including the online customer review, can be conceptualized in Figure 1.

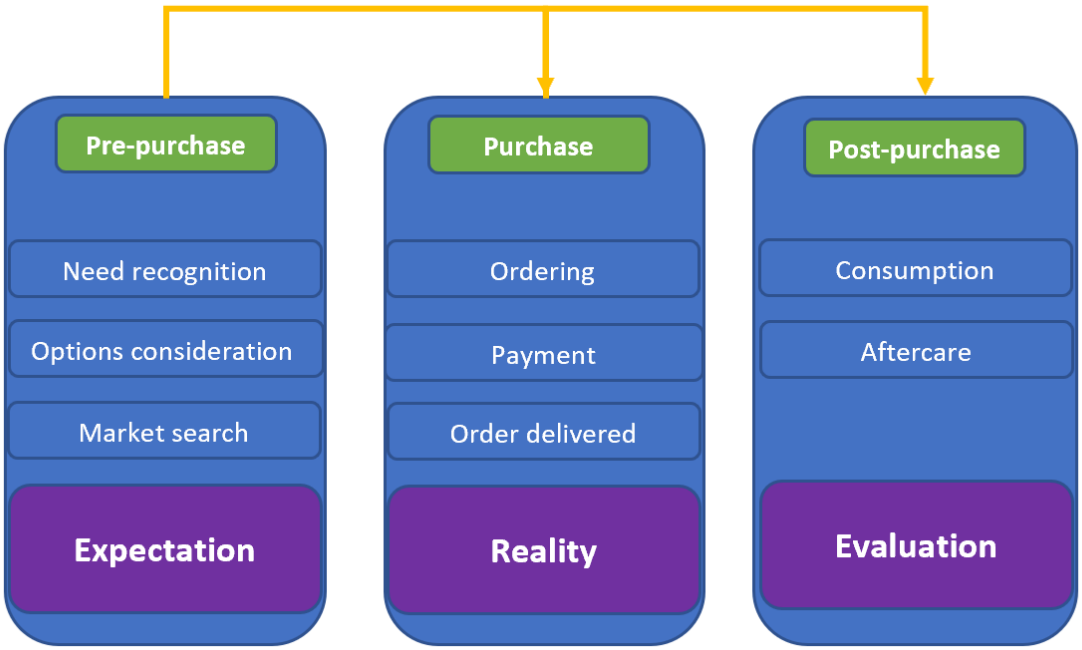


Figure 1: Conceptual framework on the customer journey for e-grocery shopping

2.1.3. Online customer review and ultrafast grocery delivery business

With the rationale behind online customer review and its connection to the concept of customer experience and customer satisfaction, this section will answer the question: what are the expected features that drive the customers' satisfaction for an ultrafast grocery delivery firm, or more generally, an e-grocery delivery business?

Several features associated with the ultrafast e-grocery delivery business were identified as important factors for customers, thus influencing their expectation and, ultimately satisfaction. The first feature mentioned is, as derived from the name of the business itself, ***delivery speed***. The hypothesis of speed being one of the critical factors that add to customers' value is supported by a paper conducted by Ramus and Asger Nielsen (2005) and further confirmed by Stanton (2018) with the statement that consumers are becoming more demanding of convenience through time and effort saved and that supermarkets or the physical grocery industry must be responsive to the consumer changes (Ramus & Asger Nielsen, 2005) (Stanton, 2018). Considering the nature of products in the grocery shopping industry, it is understandable that customers want to get the products they order as quickly as possible. You do not want to wait for the chicken to be delivered tomorrow if you want to make roasted chicken for dinner

tonight. Customers want their grocery shopping delivered quickly with minimal effort (Wolfinbarger & Gilly, 2001).

The second factor mentioned in several researches is the convenience factor, which can be associated with the app convenience in this study. The COVID-19 pandemic also changed the grocery shopping behavior; consumers are now experienced and advanced with mobile shopping. Hence, they are used to the ***convenience and time-saving*** benefits provided by online grocery services. A study by Kimberly Jensen et al. (2021) suggested that almost 60% of people in the study planned to continue online grocery shopping regardless of the pandemic conditions because of the convenient experience (Jensen, Yenerall, Chen, & Yu, 2021). Similarly, Morganosky and Cude (2000) conducted online survey studies with people who have experienced online market shopping in the UK, and the result shows that the most crucial advantage of online shopping is saving time and energy (Morganosky & Cude, 2000). Brand et al. (2020) also provided the same result in their study on UK online grocery shoppers, suggesting that UK shoppers might be attracted to online shopping for convenience or responding to time pressures and fit into daily schedules (Brand, Schwanen, & Anable, 2020). A paper from Rose et al. (2012) suggested that convenience experience during the ordering process, including the ease of navigation and product search in the grocery app, can have a significant positive impact on the customer response to the purchase interaction (Rose, Clark, Samouel, & Hair, 2012). To maximize the convenient experience, it is essential that the e-grocery app design should be user-friendly, easy to navigate, and perform smoothly.

Besides the convenience factor, the same paper from Chu (2010) suggested a third factor - the ***product price*** – which is also a sensitive feature to online grocery customers, although the sensitivity varies between different customer segments. Light online shoppers, which refer to customers with a relatively low frequency of ordering online groceries, tend to be less sensitive to the product price, while heavy online shoppers are the most susceptible to price (Chu, 2010). Other studies also identify that product price, among other features, is essential to the customer's goal-directed experience (Ramus & Asger Nielsen, 2005) (Wolfinbarger & Gilly, 2001). The overall sensitivity of product price is an indication of a negative relationship between the price and the customer opinion about the service, thus it might be reflected in a low review rating.

In addition to the price, the ***delivery fee*** charged to each online grocery order is also suggested to be a significant factor for the customers. The computational experiment conducted by Fikar et al. (2021), which investigated the importance of different product and service factors in an e-grocery setting in Austria, indicates that the delivery fee is the most crucial customer factor. In fact, e-grocery orders are increased by more than 50% in the scenario where no delivery fee is charged compared to the base setting (Fikar, Mild, & Waitz, 2021).

Another factor that is believed to influence the customer experience and satisfaction is ***customer service***. Customer service, including delivery service, is considered a key component in the customer's purchase experience. An analysis conducted by Singh and Soderlund (2020) has provided the qualitative insight and confirmed the significance of customer service to customer experience. In fact, customer service accounted for 68% of the overall experience and 42% variance in customer satisfaction (Singh & Söderlund, 2020). In an online context, customers require responsive and helpful customer service, offering the same experience they could have obtained in an offline setting (Ramus & Asger Nielsen, 2005).

Lastly, another factor mentioned is the diversity of ***product range*** available online, where Maltese suggested that customers with different lifestyle characteristics react differently to alternate grocery shopping channels. It is recommended that the best strategy to increase e-grocery should not focus on monetary, i.e., product price or service fee, but rather the attention should be paid to expanding the product range (Maltese, Le Pira, & Marcucci, 2021).

Based on the preliminary review of a subset of relevant literature, the following hypothesizes were formulated:

Hypothesis 1: Delivery Speed, App Convenience, Price, Delivery Fee, Customer Service, and Product Range will be the main service features that are evaluated in the reviews of the ultrafast grocery delivery customers.

Hypothesis 2: Each of the above features has a significant impact on the customer rating, where the direction of the relationship will differ per feature.

Table 1 below provides a general description of each feature mentioned in the hypotheses.

Features	Description
Delivery Speed	The door-to-door delivery time since the customers make an order on the app. For speedy grocery service, the average expected time is approximately between 10 and 20 minutes.
App Convenience	How user-friendly is the shopping app, and how easy the order process is for the customers
Price	The product price compared to the price of the same item in physical supermarkets
Delivery Fee	Delivery fee and any additional cost incorporated in the final order bill but are not the product price
Customer Service	The overall human service provided to customers during the whole purchase process, covering delivery service and customer support service
Product Range	The variety of products range that is offered on the app compared to physical supermarkets

Table 1: Features description on customer satisfaction

2.2. Technical Perspective

In this section, the high-level narrative of the choice of models for this study will be elaborated. The choice of research methodology for this study reflects the selection of relevant approaches to answer the sub-questions and the corresponding hypotheses mentioned in previous sections. To answer the first sub-question by proving Hypothesis 1, a topic modeling method called Latent Dirichlet Allocation (LDA) has been chosen, considering this is a flexible and robust topic modeling algorithm to generate aspects (Blei, Ng, & Jordan, 2003).

Text analytics, often referred to as part of Natural Language Processing (NLP), is an area in machine learning that focuses on creating algorithms that help to extract, process, analyze, and understand written text (Chowdhury, 2003). Along with the growth and advancement of internet technology, digitalization, and online business, NLP's application in analyzing text has become more popular. The outcome of NLP can transform pieces of text into informative insight about customer behaviors and opinions, serving as input for business management in the strategic decision-making process. One sub-topics of NLP is the analysis of words used in customer review written text to discover underlying topics, often referred to as topic modeling (Büschken & Allenby, 2016) (Blei & Lafferty, 2007). The application of topic modeling is becoming more popular for lots of businesses across various sectors, from digital entertainment

streaming (Bennett & Lanning, 2007) (Ganu, Elhadad, & Marian, 2009), hospitality (McAuley, Leskovec, & Jurafsky, 2012) (Guo, Barnes, & Jia, 2017) (Calheiros, Moro, & Rita, 2017), airlines (Korfiatis & Stamolampros, 2019), to restaurants (Dickinger, Lalicic, & Mazanec, 2017). Within topic modeling, Latent Dirichlet Allocation (LDA) is a widely used technique due to its utilization of the Dirichlet prior distributions, which helps to discover from a piece of text the hidden topics (Blei, Ng, & Jordan, 2003). On a high level, LDA assumes that every text document or sentence is a bag of words, meaning it covers only a few topics, and each topic is presented by several keywords. The output of the LDA model provides probabilities of topics appearing in a document and probabilities of words appearing in a topic. The production of LDA has proven to yield higher accuracy in further analysis due to its reliable statistical algorithm applied behind (Lu, Ott, Cardie, & Tsou, 2011) (Schouten & Frasincar, 2015). The other advantage of LDA is that the hidden topics and words can be presented in the form of a probability distribution, which can be easily interpreted and analyzed furthermore, thus LDA is a preferred topic modeling method for this study to answer the sub-question one on what are the key topics in ultrafast e-grocery delivery according to the customer's online reviews.

With the LDA latent topics as features input, the second stage is to identify the potential relationship between those features and the customers' review ratings as ways of answering sub-question two. When working on machine learning problems, including text classification, different algorithms and techniques can be used to train the text classifier. Each algorithm has its own advantages and limitations, with no confirmation upfront on which algorithm would be the best performer. Instead, the common practice is to try different techniques for training and evaluate the model's prediction power and pick out the one with the most accurate predictions possible. The same approach will be applied to answer sub-question 2. Three different models, being Logistic Regression, Support Vector Machine (SVM), and Random Forest (RF), will be used for training to understand the relationship between the features and the customers' review ratings, which will then be used for prediction. These three methods are considered well-known supervised classifiers, which have proven effective for the text classification task (Devika, Sunitha, & Ganesh, 2016). In terms of model implementation, the Logistic Regression model is performed using *stats* package built in R (v4.1.1, R Core Team, 2021).

Both Random Forest and SVM models are carried out by employing *caret* package (v6.0-89, Kuhn, 2021). The prediction of all three models will be analyzed and used for interpretation. It is noted that not all machine learning models are straightforward to interpret. Complex models, for instance, Random Forest and SVM, are considered "black box" models in the sense that it is not transparent when we look at what exactly happened within the model itself. We could evaluate how well these models predict the online review ratings based on the identified set of online grocery features; however, these models cannot answer to what extent each feature has the influence on the determination of the review rating; therefore, an extra layer is required to have a deeper look into model interpretability. For this study, a global interpretation method called the permutation feature importance will be applied to Random Forest and SVM model.

3. RESEARCH METHODOLOGY

3.1. Latent Dirichlet Allocation (LDA)

LDA is a probabilistic topic modeling method introduced by Blei (2003) that helps to discover hidden topics in the documents as well as the keywords of those topics through posterior inference (Blei, Ng, & Jordan, 2003). The fundamental idea of LDA is that every topic or document is a conditional distribution of the hidden structure, also referred to as a posterior, over words or topics representation. In this way, for each input document, a new document is created with the goal of replicating as much as possible the original document by maximizing the probability of creating the same document. The graphical model representation of the LDA model is illustrated in Figure 2 (Blei, Ng, & Jordan, 2003).

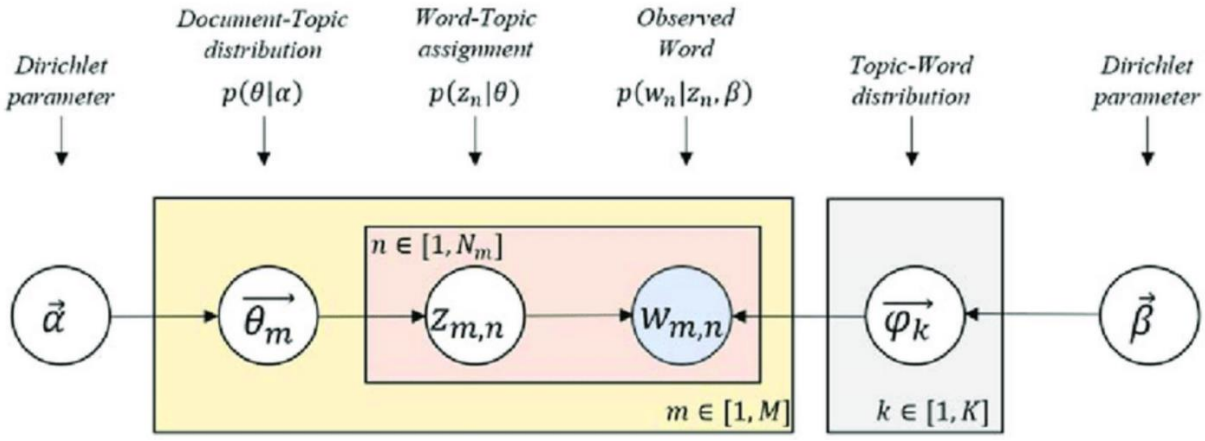


Figure 2: Latent Dirichlet Allocation process graph

In Figure 2, hyperparameters α and β define the Dirichlet distribution of topics over documents and words over topic, respectively. θ and Φ represent the topics per document and the words per topic, respectively, following the multinomial distribution. *z* is the vector with topics of all words in all documents, and *w* is the vector with all words in all documents. *M* represents the number of documents. *K* is the number of topics and *N* is the number of words.

When a document or a review is written, the first thing determined is which topics are addressed in the review. In this case it could be about the product quality, the delivery time, per suspect from the literature view, or it could potentially be about any topic that is not included earlier. These topics are drawn from all the topics using the Dirichlet

parameter α. Then, for every word in the document, the model picks a specific topic for that word from the topic distribution of θ. Finally, from the distribution of words over topics Φ, which is drawn from all words using the Dirichlet parameter β, the model picks one of the words from a specific topic. That process is repeated to pick each and every word until the document is complete. From the mathematical standpoint, the probability formula of the LDA model is presented in equation 1:

$$P(W, X, \theta, \Phi, \alpha, \beta) = \prod_{i=1}^{M} P(\theta_i, \alpha) \prod_{j=1}^{K} P(\Phi_j, \alpha) \prod_{t=1}^{N} P\left(Z_{i,t} \middle| \theta_i\right) P\left(W_{j,t} \middle| \Phi_j\right). \quad (1)$$

In equation 1, both θ and Φ are latent factors, meaning they cannot be observed directly. Instead, θ and Φ are drawn from the Dirichlet distribution of the hyperparameter α and β, respectively, where α indicates the document topic density and β represents the topic word density. Lower value of α equals to fewer topics per distribution, meaning each topic has to include more words and have more overlap since each word needs to be assigned.

From a certain perspective, LDA trades off two goals. The first goal is to allocate its words to as few topics in each document as possible. The second goal is to assign high probability to a few words for each topic. Now these goals are at odds, meaning they are conflicting with each other. Putting a document in a single topic makes it hard for that topic to assign high probability to a few words because it has to explain all the words in the document by that topic. Similarly, if the model puts very few words in each topic which satisfies the second goal, it will make the first goal difficult to accomplish because having very few words in each topic means we will need a lot of topics to somehow cover the total words of the documents. Ultimately, the task of the LDA algorithm is to find the parameters that maximize the likelihood of the corpus. The whole process of training or maximizing probability can be done using Gibbs sampling where the general idea is to make each document and each word as monochromatic as possible, meaning we want each document to have as few as possible articles and each word belongs to as few as possible topics.

In terms of implementation, there are quite a number of open sources of software packages that implement some kind of posterior inference for LDA. For this study, the

textmineR package in R is chosen to perform the LDA, in which a couple of parameters need to be defined (v 3.0.5, Jones, 2021). The first variable to tune is α, a Dirichlet distribution hyperparameter that is responsible for the topics over document density, also called the sparseness of the topic distribution. A high value of α increases the probability that a document can entail multiple topics and vice versa, a low α indicates a high probability that the document is consisted of only one or a few topics.

The second variable to consider with LDA is the perplexity, a statistical measure of how well the LDA model can predict a sample. It estimates the modeling power of the LDA with the given parameters by using the inverse probability of unobserved documents (Blei, Ng, & Jordan, 2003). The perplexity calculation equation is:

$$\text{Perplexity of set of documents} = \exp\left(\frac{-\log\left(\Pr[all\ words\ in\ docs]\right)}{Total\ number\ of\ words}\right). \qquad (2)$$

The formula suggests that the lower the perplexity, the more certainty about the unobserved documents, hence the better the model. However, low perplexity also means high number of topics and having too many topics will make it difficult to interpret the outcome (Kwartler, 2017).

The other factor that should be taken into account when building LDA model is the topic coherence. Topic coherence assesses the quality of the learned topics by measuring the semantic similarity between high scoring words in the topic and the goal is to pick the number of topics with the highest coherence value. In another word, if perplexity presents how well the model can predict an unseen document based on given input, coherence determines how well the latent topics is interpreted and whether all the representative words in the topic refer to the same topic. Recent studies have shown that perplexity and coherence are often slightly anti-correlated, meaning optimizing for perplexity may not yield interpretable topics. (Chang et al, 2009) (Mimno et al., 2011). Therefore, when identifying the optimal number of topics based on perplexity and coherence, the common practice is to find the best balance between these two values. There are several "topic coherence" metrics available in the literature (Roder, Both, & Hinneburg, 2015) (Aletras & Stevenson, 2013). In this study, the coherence measure calculation is built as part of the LDA package *textmineR* in R. Using the N highest

probability tokens for each topic, a coherence score is measured based on probability theory. The essential idea of probabilistic coherence is that if a pair of words $\{a, b\}$ in the top M words in a topic are highly coherence with each other, $\{b\}$ would be more probable to appear in a document containing $\{a\}$ than in a random document as a whole. Following that logic, the coherence is calculated as:

$$\text{Probabilistic coherence } (a, b) = \mathrm{P}(b|a) - \mathrm{P}(b). \quad (3)$$

Where $\{a\}$ is more probable than $\{b\}$ in that topic. If $\{b\}$ is not more probable in documents containing $\{a\}$, the coherence value would be close to zero.

3.2. Logistic Regression

Logistic regression is a method that describe the association between a binary or multinomial dependent variable with one or more independent variables (McCullagh & Nelder, 1989), i.e., the relationship between the customers' review rating and the review words and the LDA topics. Logistic regression is considered a simple and fast classification method yet having high interpretability, since it can measure and provide the magnitude of the relationship between the predictors and response variable through the coefficient estimate in the model output (Kwartler, 2017) (Kuhn & Johnson, Applied predictive modeling, 2013). The general logistic regression formula is presented in equation 4:

$$P(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}} . \quad (4)$$

where $P(Y = 1|X)$ defines the probability of an observation belonging to class 1 given characteristics set X_p with β_p denoting the coefficients associated to the corresponding predictor X_p. Applying to this study, $P(Y = 1|X)$ would estimate the probability of an online review belongs to a certain rating class, given the features that are mentioned in that review. The logistic regression formula is derived from equation 5 which presents the relation between the log-odds of $P(Y = 1|X)$ and the linear transformation of X:

$$\log\left(\frac{(P(Y = 1|X)}{(1 - P(Y = 1|X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p . \quad (5)$$

3.3. Support Vector Machine (SVM)

The second model used for this study is Support Vector Machine (SVM). Support vector machine (SVM) is a supervised learning method that analyzes data for classification and regression tasks and helps to find potentially very non-linear decision boundaries. Initially coined by Vapnik and colleagues in the 1990s (Vapnik & Cortes, 1995), SVM has grown in popularity since and become one of the best "out of the box" classifiers that prove to perform well in a variety of settings, including text classification task (Joachims, 1998) (Drucker, Wu, & Vapnik, 1999) (Dumais et al., 1998) (Basu, Watters, & Shepherd, 2003).

SVM's ultimate goal is to classify observations to one of the classes by a separating line called hyperplane in a multi-dimensional space. SVM is built based upon the framework of the maximal margin classifier and support vector classifier.

The margin, in a concept of maximal margin classifier, is defined as a minimal perpendicular distance between the closest data points, often referred to as support vectors, to a given hyperplane from both sides. The goal of the maximal margin classifier is to maximize the margin value and eventually pick out the optimal hyperplane such that the margin distance is at maximum.

Support vector classifier, sometimes called as soft margin classifier, is an extension of the maximal margin classifier where the classifier allows some data points to be misclassified as a trade-off for a larger margin value, thus a better hyperplane classification for most of the observations. The margin is considered soft since it can be misclassified by some of the observations. Mathematically, support vector classifier solves the following optimization equation:

$$minimize_{a_0,a_1,\dots,a_p}: \sum_{i=1}^{n} max\ \{0,1 - \ y_i\, f(x_i)\} + C\ \sum_{j=1}^{p} {a_j}^2. \qquad (6)$$

In this equation, the first term represents the maximal margin classifier while the second term is the regularization term that is added to avoid overfitting by penalizing large coefficients of the vectors. C, also called Cost or regularization parameter, controls the level of trade-off between increasing the margin size and lowering the data points that are misclassified, which is the variance-bias trade-off. When the C parameter is large, the hyperplane will separate data points with smaller error rate at cost of a narrow

margin, thus the classifier has low variance but potentially high bias. In contrast, if C is small, there will be fewer support vectors which mean the resulting classifier will have low bias but high variance.

Support vector classifier is a natural approach for classification in the two-class setting only if the boundary between the two classes is linear which is not realistic to expect in practice. This is also a limitation of other classification methods, for instance, logistic regression where the performance of linear regression can suffer when there is a nonlinear relationship between the predictors and the outcome. For that reason, SVM was proposed as an extension of the support vector classifier that can accommodate non-linear class boundaries by applying a function called kernel.

Kernel is a function that transforms low dimensional input space to a high dimensional space by quantifying the similarity of two observations. The main aim of SVM kernels to try to convert the low dimensional space into a high dimensional space. With kernel formulation, the original set of features that is given will be mapped into much higher dimensional set of features. Kernels relieves a lot of burden of manually picking features because the kernels can build an infinitely large set of features. There are many kernel functions to choose from, each of which applies a different transformation function to the data and is suitable for different situations. A support vector classifier would make use of the linear kernel. When the support vector classifier plugs in a non-linear kernel function, the classifier can be used to learn and create non-linear hyperplane, and is referred as Support Vector Machine (SVM). The common non-linear kernel functions used in SVM would include the polynomial kernel, the radial basis function (RBF) kernel, and the sigmoid kernel, among others. Some previous experiments have shown that the choice of kernel for text classification has a minimal effect on the classifier performance (Leopold & Kindermann, 2002) (Joachims, 2002), therefore, it is decided in this study to only conduct SVM with one non-linear RBF kernel.

SVM has several advantages, especially in comparison with logistic regression as a benchmark, which explains the choice to use the SVM model for this analysis. The fact that SVM can perform well in high dimensional spaces with a clear margin of separation while allowing some violations to this separation makes SVM outstanding from classical classification approaches like logistic regression. Next to that, the kernel technique which can expand the feature space to accommodate non-linear class boundaries is a

unique and valuable characteristic of SVM that not many other methods could have. A limitation of the SVM is that it is considered a 'black-box' model, meaning the interpretability is lower as compared to other classifier method like logistic regression, since there is no probabilistic explanation for the classification. In addition, SVM is not suitable for large datasets due to its computational intensiveness.

3.4. Random Forest (RF)

The third algorithm chosen in this study is Random Forest (RF). Random Forest is a decision tree method that was developed based on the foundation of a traditional decision tree, so first let's explain the concept of a decision tree before zooming in Random Forest.

3.4.1. Decision tree

Decision tree is a method that can be used for both regression and classification tasks. For a classification tree, the method involves stratifying or segmenting the predictor space into several simple regions and using the most commonly occurring class of training observations in the node to which it belongs to predict a given observation. An example of a simple decision tree is illustrated in Figure 3. This tree consists of a series of so-called splitting rules, starting at the top, which is also called the root node, and goes down to the bottom layers. The top split assigns observations with $X_1 \leq t_1$ to the left branch and the others to the right branch. Then each sub-group is then assigned to smaller sub-groups after going through the next splitting rule. Ultimately, the decision tree splits the original data set into several sub-groups R_1, R_2 , R_3, R_4, and R_5. These are called terminal nodes or leaves of the tree. When an observation is to be classified, the decision tree starts at the root, makes a decision based on these splitting rules and eventually classify the observation into one of the terminal nodes.

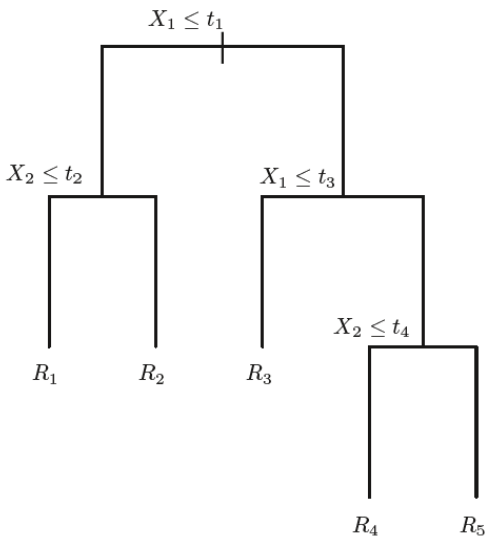


Figure 3: An example of regression tree (James, Witten, Hastie, & Tibshirani, 2021)

While a traditional decision tree has several advantages, it tends to suffers from high variance as compared to some other classification approaches due to its greedy hierarchical structure, meaning if we split the training dataset into multiple random subsets and run the decision on each of the subsets, the result we receive could be quite different, for which reason aggregated decision tree methods like the random forest were introduced (Breiman, Random Forest, 2001). By aggregating many decision trees, methods like bootstrapped aggregation or Random Forest in combination with decorrelating the trees based on a foundation concept of the decision tree can help to reduce the variance, thus improve the predictive performance of trees.

3.4.2. Bagging

Bagging, also called bootstrapped aggregation, is a very effective way to reduce the variance of a classifier in a particular context of a decision tree. The essential idea of bagging is that resampling from the original sample behaves similarly than sampling from the population, therefore, we can draw samples of the same size with replacement in order to approximate the sampling distribution using bootstrap method. The idea of bootstrap is to draw multiple datasets of the single dataset in such a way that each bootstrapped dataset has the same size as the original one and they are drawn uniformly at random with replacement. Then we build a separate prediction model for each bootstrapped training set, and average out the prediction of the observations to obtain a single low-variance statistical learning model (Breiman , 1996).

One advantage of bagging is that the method can provide a very good calibrated probabilities, meaning it is possible for an average prediction to zoom in each of these different classifiers and see how many of them have predicted that certain label. In another word, begging can tell how certain the prediction is.

3.4.3. Random Forest

Random Forest, firstly introduced by Breiman (2001) is a modification of bagging that attempts to build de-correlated trees by considering a restrictive feature for splitting (Breiman, 2001). It means at each split in the tree the algorithm is not even allowed to consider a majority of the available predictors. Instead, each node in the tree split is based on a sub-set sample of features m that is drawn randomly from the total features p. Random Forest can be described as a two phases process. In the first step, a number of bootstrapped samples are drawn from the original training data. Then a random-forest tree is built for each bootstrapped dataset by recursively repeating the following sub-steps to each terminal node of the tree, until the minimum node size is reached.

1 - Select m variables at random from the total p variables where $m < p$

2 - Pick the best variable/split-point among the m

3 - Split the node into two nodes.

Phase one is then repeated for all bootstrapped datasets, from $b = 1$ to B, and output the resemble of tree.

Since Random Forest consists of a large number of trees that are each trained separately on a random subset of the data and a random selection of features per node. This method is considered a black-box model, meaning its interpretability is lower than other classifier like Logistic Regression. Similar to SVM, one way of gathering insights into Random Forests is to compute feature importance.

When building a Random Forest, there are three tuning parameters to take into account.

(1) m which controls the number of predictors that are evaluated for each split

(2) The depth of the tree which is controlled by setting the minimum number of observations in the leaf nodes

(3) The number of trees to grow

3.5. The Confusion Matrix

To evaluate the performance of the three models, the Accuracy, Sensitivity, and Specificity measures based on the confusion matrix will be used. Confusion Matrix is a fundamental term in machine learning for a classification task that refers to a specific table layout that visualizes the performance of a prediction model (Kohavi, 1998). For the classification task, the confusion matrix consists of tables where each column of the table represents the actual classification of the observation while each row presents the observations in a predicted class which are the customer review rating.

		ACTUAL	
		FALSE	TRUE
PREDICTED	FALSE	**TN (True Negative)**	**FP (False Negative)**
	TRUE	**FN (False Positive)**	**TP (True Positive)**

Table 2: Confusion matrix

Accuracy: is the percentage of observations that are predicted correctly out of all observations. It is computed based on the confusion matrix following the formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (7)$$

Sensitivity: is also referred to as the true positive rate or the recall. It measures the number of instances from the positive class that is predicted correctly divided by the total correct prediction out of all classes, following the calculation formula:

$$Sensitivity = \frac{TP}{TP+FN}. \quad (8)$$

Specificity: is also referred to as the true negative rate. It measures the number of instances from the negative class that is predicted correctly divided by the total correct prediction out of all classes, following the calculation formula:

$$Specificity = \frac{TN}{TN+FP}. \quad (9)$$

3.6. Permutation Feature Importance

The model with the best prediction power will be chosen as the final model for further interpretation, however, not all machine learning models are straightforward to interpret. Complex models, including Random Forest and SVM, are considered "black box" models in the sense that it is not transparent when we try to look at what exactly happened within the model itself. We could evaluate how well these models predict the online review ratings based on the identified set of online grocery features, however, these models cannot answer to what extend each feature have the influence on the determination of the review rating, therefore, an extra layer is required to have a deeper look into model interpretability. For this study, the two "black-box" models, Random Forest and SVM, that were used for prediction task, will be examined further using the permutation feature importance plot.

Permutation Feature Importance was first introduced by Breiman (2001) as method that measures the increase in the prediction error of Random Forest model after permuting the feature's value (Breiman, 2001). Inspired by Breiman's idea, Fisher, Rudin, and Dominici further developed the permutation and proposed a model-agnostic version, meaning the feature importance can be calculated for any given machine learning model (Fisher, Rudin, & Dominici, 2019). The basic concept of Permutation Feature Importance is rather straightforward. The importance of one particular feature in a model is measured by calculating the increase in the model's prediction error before and after shuffling the feature value. A feature is considered "important" if permuting the value increase the model error, simply because in this case it means the model heavily relies on this feature for the prediction. Similarly, a feature is considered "unimportant" if the model prediction error remained unchanged after shuffling the values, meaning the model can perform the prediction task with the same performance power regardless of the value this feature contributes in the model. In another word, the model simply ignores this "unimportant" feature in the prediction.

For any "Black-box" model, given the trained model $\hat{f}$, feature matrix X, target vector y, and the error measure $L\ (y, \hat{f})$, the mathematical algorithm of the Permutation Feature Importance, according to Fisher, Rudin, and Dominici, can be described as follows:

1. First estimate the original model error on the original given dataset $e_{orig} = L\left(y, \hat{f}(X)\right)$where e can be any type of error scoring matrix for classification or regression models.
2. For each feature $j \in \{1, \ldots, p\}$:
   - Generate feature matrix X_{perm} by permuting feature j in data X. The purpose of doing this is to remove the existing association between the feature j and the dependent variable y.
   - Compute the permuted model error $e_{perm} = L\left(y, \hat{f}\left(X_{perm}\right)\right)$ based on the prediction of the permuted data.
   - Estimate the permutation feature importance, either by the ratio of the errors with and without permutation $FI_j = e_{perm}/e_{orig}$, or by the difference of the errors $FI_j = e_{perm} - e_{orig}$
3. Sort features by descending FI.

The result of the feature importance calculation will give us insight into how much a model's error, sometimes referred to as a loss function, will increase after the permutation, which explains the importance of each feature, thus gives insight on which variables are the most important in driving the outcome. If the model's error increases significantly after the permutation of a feature, it means that feature is important to the model behavior and vice versa, a feature that hardly influences the model's error after permutation indicates that it is not so important feature for the model.

In this study, the Permutation Feature Importance is carried out using *iml* package in R (v0.10.1, Molnar, Bischl, & Cas, 2018). Since the Random Forest and SVM perform a classification task, the classification error is chosen to measure the importance and the feature importance is measured by calculating the error ratio $FI_j = e_{perm}/e_{orig}$. The expectation is that features with FI greater than 1 are considered important where features with FI equal or less than 1 are deemed unimportant in the model prediction.

4. DATA

4.1. Data Collection

The data in this study is the online review data provided by customers living in the UK about four biggest ultrafast grocery delivery companies in the UK market which are Getir, Gorillas, Zapp, and Beelivery. The data is obtained by manually scrapping the review text from Trustpilot[4] and Appstore[5] using *rvest* package in R (v1.0.1, Wickham, 2021). *Rvest* package allows scrapping text from web page and read it into R. The original raw dataset after scrapping contains ~16000 reviews with six data attributes being (1) review text, (2) review date, (3) customer name, (4) review title, (5) language, and (6) review rating being scaled from 1 to 5. Out of these six attributes, only two are relevant and being used for this study. The first variable is the *review text* which contains the raw text written by reviewers. The second variable is, obviously, the customer rating, which is the dependent variable of interest. The dataset, after being scraped and pre-processed includes 12980 observations, which is used as input for the topic modeling LDA. One observation is that the dependent variable in this dataset is imbalanced, meaning proportion of positive rating in this dataset is substantially larger than the negative rating. As presented in Figure 4, out of the total 12980 observations, 10311 gives a 5-star rating, covering 79% of the total population.

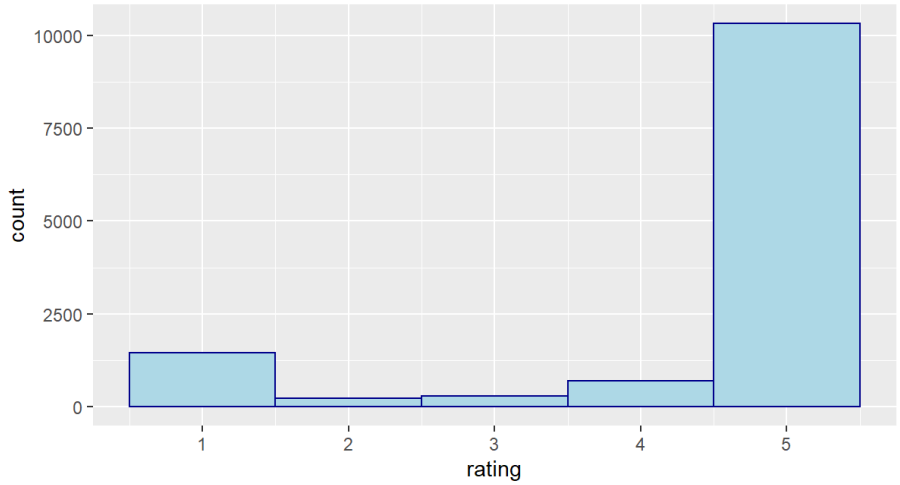


Figure 4: Reviews rating histogram

[4] Link to Trustpilot website: https://uk.trustpilot.com/

[5] Link to Appstore website: https://www.apple.com/uk/app-store/

4.2. Data Pre-processing

A number of steps were carried out in R to clean and transform the raw review text so that the data can be used further. First of all, all the reviews not written in English were removed since the study focused only on UK reviewer with English as the main language. Also, reviews that have missing value in the review field were removed as they do not provide any text needed as input for the analysis. Secondly, the raw text in the review field were broken up into individual words and converted to lowercase to obtain a standardized format using *unnest_token* function in *tidytext* package in R (v0.3.2, Silge & Robinson, 2016). A count on the frequency of each individual words was performed with *count* function from *dplyr* package (v1.0.7, Wickham et al., 2021). One observation from the word frequency table is that the top frequency words contain many non-informative words, or also called stop words. These are words like "the", "an", "a", "I", "it", etc. which do not provide much meaning to the corpus, therefore, they were removed based on a built-in stop words dataset provided by the *tidytext* package.

After that, stemming was applied, using *wordStem* function in *SnowballC* package in R to combine all the words in different forms but reflect the same meaning into one unique stemmed word (v.7.0, Bouchet-Valat, 2020). For instance, adjective word "happy", adverb word "happily", and noun word "happiness" are treated as three separate words even though they address the same meaning which is the state of emotion. By using stemming, these three words can be combined into one unique stemmed word "happi".

Using the output from stemming step, a frequency count on the stemmed words was done based on which the most frequent and most infrequent words were removed from the text. The reason for that is infrequent words, which occur less than 0.01% of the total text population, are highly likely to be case specific and does not represent the general customers' voice. Similarly, frequent words were removed since they are often just general words, like "grocery", "buy", or the company names, and carry no meaning with them. Include these frequent words will potentially increase the noise without adding meaningful interpretation with them.

Lastly, the dependent variable, customer rating, was transformed into binary variable where all the rating that are higher or equal to 4 are considered positive and the one

with rating 1, 2, or 3 are converted to negative rating. Also, irrelevant data attributes were removed since they do not add value to the further analysis. At the end of the pre-processing, the output is a dataset of 12980 observations and only two attributes, being *customer rating* and *review text*.

To train the predictive model, the dataset is split into two subsets being the training set and the test set with the proportion of 80/20 respectively. The training dataset is used to train the models and the test dataset is used to verify the predictive performance of these methods.

4.2. Data over-sampling method

Many applications of text classification task are exposed with data class imbalance challenge. Data class imbalance refers to a dataset where the number of observations in one class are outnumbered the number of the other, meaning the observations in each class is unequally distributed. The performance of many classification algorithms, such as Random Forest and SVM, is believed to suffer from the data imbalance problem (Japkowicz & Stephen, 2002).

In this study, the dataset also indicates a class imbalance where the number of reviews that give 5-star rating covers 80% where the rest of reviews with rating 1-4 counts for the rest of 20%. To confront the imbalanced data problem, the random oversampling method is applied to balance the dataset. Random oversampling is a traditional resampling technique that basically increase the number of records in the minority class to be the same size as the majority class by randomly duplicates examples with replacement from that same minority class. The random oversampling method was employed using *caret* package in R (v 6.0-89, Kuhn, 2021).

5. RESULT

In this chapter, the result of the analysis is presented, following the sections order in the methodology chapter. First, section 5.1 will demonstrate the output of the LDA, including the parameters tuning and the LDA list of features based on that tuned set of parameters. Based on the LDA topics, the outcome of Logistic Regression, SVM, and Random Forest and their corresponding performance evaluation are presented in section 5.2, 5.3, and 5.4 respectively. And lastly, the predictive performance of these three models is given in section 5.5.

5.1. LDA

Figure 5 below shows the 15 latent topics with the seven most likely words per topic. Zooming into the key words that represent each topic, it is observed that some topics are similar with each other, meaning they share similar set of key words associated to the same feature, based on which these topics were named using the logical and intuitive interpretation.

For instance, topic 3, 5, and 11 share similar key words like service, recommend, excellent, friendly. It is logical to associate these words to one main theme about Customer Service, which is a common theme that catch the customers' interest in grocery. Topic 10 and 12 are also similar where both address the app convenience topic with app, website, easy, simple, track. Similarly, topic 10 and 12 are believed to address the same feature related to app convenience. The top key words mentioned include app, easy, quick, simple, track, referring to the point of how simple and user friendly it is for customers when they use the app for ordering grocery and also how easy and convenient to track an order that is in progress of delivery.

After assigning each latent topic to a feature name based on its representative keywords, the 15 topics are classified into five main features, which are *"Delivery Speed", "Customer Service", "App Convenience", "Price", and "Food Quality"*. Four out of these 5 LDA topics are aligned with the hypothesis 1, except for "Food Quality". This is a new finding that was not addressed previously in the hypothesis. On the other hand, two topics that are

expected in the hypothesis which are “Delivery fee” and “Product range” do not seem to be important and interesting enough for the customers to mention in the reviews.

Among these features, “Delivery Speed” is the most mentioned feature with 6 out of 15 LDA topics covers this feature, being topic 2, 8, 9, 13, 14, and 15. These four topics share a set of overlapping keywords like “delivery”, “fast”, “quick”, “minute”, time”, “impressive”, etc. This finding is consistent with the hypothesis that customers are becoming more demanding of the time saving factor on traditional industry like grocery shopping, therefore, they want to get the grocery delivered at least as fast as they could achieve with the traditional grocery shopping (Ramus & Asger Nielsen, 2005) (Wolfinbarger & Gilly, 2001).

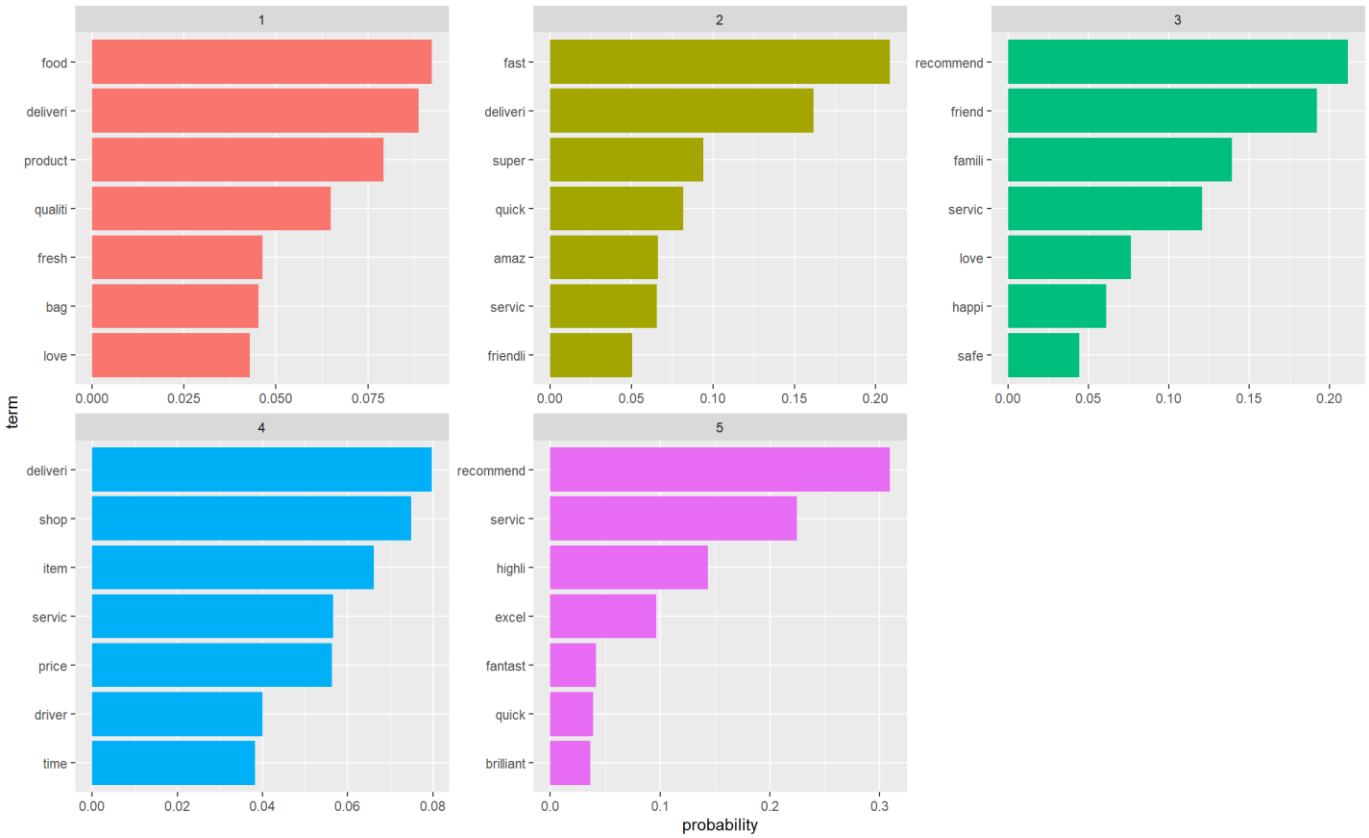
The second most frequently mentioned feature among those 15 topics belong to the “Customer Service” with 5 out of 15 LDA topics refers to this feature: topic 3, 4, 5, 7, and 11. It is noted that the definition of “Customer Service” feature in this study covers all the human interaction service that the customers could have during the whole purchasing process, which includes both the delivery service and the customer support service. The common keywords mentioned across these topics contains words like “delivery”, “service”, “driver”, “friendly”, “polite”, “excellent” etc. This outcome is aligned with other studies aforementioned that customer service is a key component in the customer’s purchase experience, thus it accounts for a significant proportion on the customer overall experience and satisfaction (Singh & Söderlund, 2020) (Ramus & Asger Nielsen, 2005). Additionally, the customer expectation in customer service to the ultrafast grocery delivery companies like Gorrilas or Getir goes beyond than just the customer aftercare service, instead, it also comes from the delivery interaction with the driver, proven by the fact what words like “drive”, “polite”, “friendly” are mentioned together in the same topic.

The third feature mentioned in the latent topics is the “App convenience” which is covered by topic 10 and 12. The key words of this feature includes “app”, “easy”, simple”, “track”, “download”, etc. This appearance of “App convenience” feature is consistent with the hypothesis where the habit of online grocery shopping in the UK has been promoted during the COVID-19 and continue after pandemic. The convenience experience in this context comes from the whole ordering process, including the ease of navigating and making order in the grocery app (Rose et al., 2012).

Lastly, “Food Quality” and “Price” are the features mentioned in topic 1 and 6 respectively. For an industry of necessity food like grocery shopping, it is understandable that price matters to customer choice and experience. This result is aligned with the hypothesis 1 where price is considered a sensitive feature to online grocery customers, hence contributes to the customers’ goal-directed experience (Ramus & Asger Nielsen, 2005). Meanwhile, “Food Quality” is a new feature that is not expected in the hypothesis. This feature also has been very limitedly addressed in research studies as an important factor to the customers. This might be an interesting finding if “Food Quality” could have a significant impact to the customer reviews rating. In term of topic frequency, Table 3 lists down all the topic names together with the average probability of each topic occurring in a document or an online review in this case. With 15 topics employed in the LDA outcome, the average probability of each topic mentioned in a document is 1/15 which equal to approximately 6.7%. Looking at the topic frequency table, topic 8 is the one with the highest topic probability of 0.1303, meaning the chance of this topic being in a document is approximately 13%, which is double the average probability. Since topic 8 is associated to feature “Delivery Speed”, it can be a sign that Delivery Speed is one of the important features that could have significant impact on the customer review, proven by the fact that it is mentioned the most in their review. However, this can only be proven/confirmed when we zoom into the relationship between these topics and the rating using predictive models. Topic 4, which cover “Customer Service” theme, comes up next at second place with the probability of this topic being mentioned in a document is 0.1061 or 10.61%. This can be considered to be remarkably higher than the means. The next topic is topic 7 (Customer Service) with the probability of approximately 7.6%. The rest of the LDA topics share a similar amount of topics probability, ranging between 5% and 7%.

Topic	Topic description	Topic probability
Topic 1	**Food quality** with key words product, food, quality, fresh	0.0659
Topic 2	**Delivery speed** with key words delivery, fast, super, quick	0.0642
Topic 3	**Customer service** with keywords service, friendly, recommend, love	0.0516
Topic 4	**Customer service** with key words delivery, shop, service, driver	0.1061
Topic 5	**Customer service** with key words service, recommend, excellent, fantastic	0.0567
Topic 6	**Price** with key words offer, promotion, discount, code, spend	0.0570
Topic 7	**Customer service** with key words delivery, driver, service, polite, friendly	0.0769
Topic 8	**Delivery speed** with key words delivery, time, hour	0.1303
Topic 9	**Delivery speed** with key words delivery, service, door, hour, amazing	0.0627
Topic 10	**App convenience** with key words app, easy, quick, simple, track	0.0555
Topic 11	**Customer service** with key words service, friendly, delivery, excellent	0.0533
Topic 12	**App convenience** with key words app, download, love, amazing	0.0536
Topic 13	**Delivery speed** with key words delivery, service, fast, quick, excellent	0.0544
Topic 14	**Delivery speed** with key words driver, time, delivery, accept	0.0574
Topic 15	**Delivery speed** with key words time, arrive, minute, delivery, quickly	0.0545

Table 3: List of LDA latent topics and the average topic probability



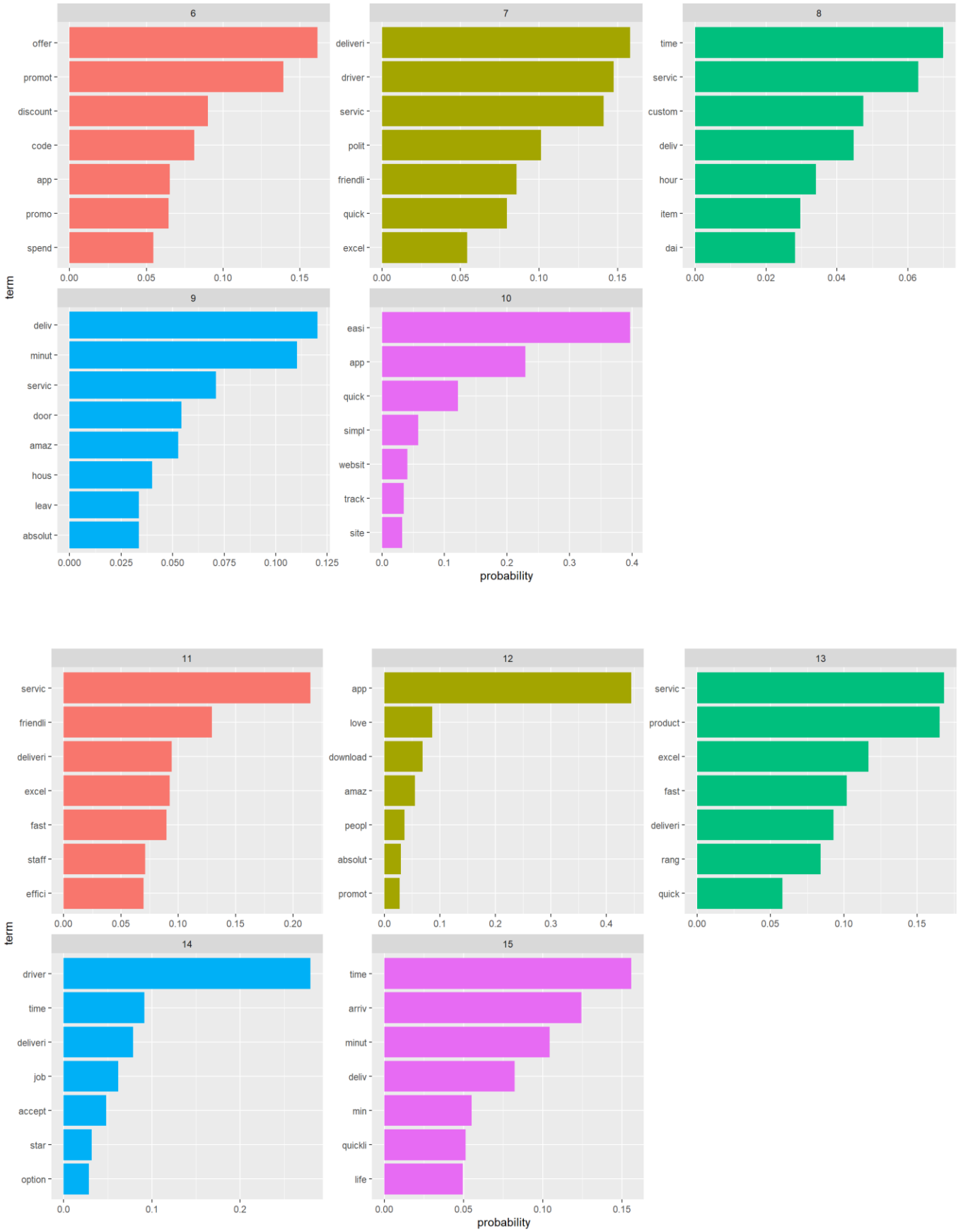


Figure 5: LDA topics and the most 7 representative words per topic

To get to this final model for LDA, a tuning for the hyperparameter α and number of topics *K* was performed to find out the combination that yields the lowest perplexity and highest coherence score. Five different values of α were chosen: 0.1, 0.5, 1, 1.5, and 2. For each value of α, *K* is iterated between 10 to 40 topics, with a 5 jumping steps between *K*. The perplexity and coherence values for all combinations of α and *K* are presented in Table 4 and Table 5 below.

	10	15	20	25	30	35	40
0.1	-20.654	-20.095	-20.016	-19.911	-19.861	-19.932	-19.976
0.5	-21.047	**-21.073**	-20.973	-21.060	-21.084	-21.193	-21.233
1.0	-21.415	-21.522	-21.504	-21.669	-21.696	-21.915	-22.002
1.5	-21.861	-21.895	-21.886	-21.819	-22.178	-22.306	-22.405
2.0	-22.018	-22.139	-22.287	-22.289	-22.424	-22.506	-22.589

Table 4: Perplexity values of the validation dataset

	10	15	20	25	30	35	40
0.1	0.041	0.055	0.053	0.049	0.053	0.052	0.053
0.5	0.046	**0.056**	0.042	0.044	0.046	0.046	0.050
1.0	0.047	0.043	0.047	0.048	0.047	0.046	0.042
1.5	0.047	0.045	0.044	0.041	0.045	0.043	0.041
2.0	0.046	0.037	0.040	0.039	0.040	0.042	0.043

Table 5: Coherence values of the validation dataset

One observation from the perplexity and coherence table is that there is not particular value of *α* that succeeds in obtaining lowest perplexity and highest coherence at the same time. This is an expected result since previous studies have proven that perplexity and coherence are anti-correlated, meaning optimizing for perplexity may not yield interpretable topics (Chang et al, 2009) (Mimno et al., 2011). For instance, the lowest values of perplexity are observed where α value is high, i.e., 1.0, 1.5, and 2, and number of topics *K* is also high, i.e., *K* = 40. However, these same combinations of α and *K* yields the lowest value of coherence, meaning the words in these topics are not coherent with each other. As a result, achieving both in this case is not possible, instead, it is a trade-off decision to find out α value that can balance the two measures. In the end, it was concluded that α value of 0.5 and the number of topics being 15 show the best value since this combination has one of the highest coherence values while able to obtain a relatively low perplexity.

5.2. Logistic Regression

Given the list of LDA latent topics, the next stage is to analyze the potential relationship between these variables and the customer ratings to identify which features among those could have a significant impact on the customer review rating. The first model outcome presented is from Logistic Regression. In this study the rating variable, after being transformed, has two values: happy (positive rating) or not happy (negative rating). The outcome of Logistic Regression is summarized in Table 8, sorted by coefficient size with the benchmark category is topic 15 – Delivery Speed. While a positive coefficient increases the likelihood of a review being a positive one, a negative coefficient indicates the opposite as compared to the benchmark topic. The values of the coefficient indicate the magnitude of the potential impact. There are a number of observations to be taken from the Logistic Regression summary.

First of all, the majority of the 15 LDA latent topics have statistically significant coefficient, except for topic 3, meaning most of the identified features could have a significant impact to the review rating. In term of direction, the number of positive and negative coefficient are quite evenly distributed with 6 positive variables and 7 negative variables. It is noted that the direction and magnitude of the coefficient are as compared to the benchmark topic 15. The top five most significantly positive variables are topic 7, 2, 5, 8, and 9 which associate with the two features of "Customer Service" and "Delivery Speed", meaning that including these topics will increase the probability of having 5-star rating in a review. This can be an indication that Customer Service and Delivery Speed are among the most influential features that could have an impact towards customer rating. The top most significantly negative variables contain topic 12, 6, and 10 which refer to two features "App Convenience" and "Price". This could indicate that as compared to the benchmark topic being "Delivery Speed", adding these two features will not increase the probability of having 5-star rating. Even though the two features still have significant impact, they are less important to the rating compared to the "Delivery Speed" feature. Based on the Logistic Regression result, it seems that most of the LDA features could have significant impact to the customers' rating, among which "Delivery Speed" and "Customer Service" are two features that seems to be the most essential to the customers.

In term of prediction performance on the test dataset, logistic regression model performs relatively well with the prediction accuracy level of 80.95%, meaning 80.95% of the observations in the validation dataset were correctly predicted by the model. The sensitivity is calculated at 82.45% and the specificity is 72.25%, which means logistic regression model performs better at predicting customers with positive rating than the ones with negative rating, which is acceptable in this case since the study focuses on analyzing and predicting the positive rating instead of the negative one. The confusion matrix and the prediction evaluation of the Logistic Regression model are presented in Table 6 and Table 7 below.

		ACTUAL	
		Nothappy	Happy
PREDICTED	Nothappy	**414**	**583**
	Happy	**159**	**2738**

Table 6: Logistic Regression confusion matrix

Model	**Logit Regression**
Accuracy	80.95%
Sensitivity	82.45%
Specificity	72.25%

Table 7: Logistic Regression Prediction

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Topics.t_7	9.39521	0.91593	10.258	< 2e-16	***
Topics.t_2	9.26396	1.02083	9.075	< 2e-16	***
Topics.t_5	8.8689	1.23045	7.208	5.69E-13	***
Topics.t_8	8.79379	1.25198	7.024	2.16E-12	***
Topics.t_9	6.04828	0.95559	6.329	2.46E-10	***
Topics.t_11	4.28136	1.31217	3.263	0.0011	**
(Intercept)	0.06568	0.73881	0.089	0.92916	
Topics.t_3	-0.15476	1.25254	-0.124	0.90167	
Topics.t_1	-1.65268	0.83035	-1.99	0.04655	*
Topics.t_4	-1.77577	0.76874	-2.31	0.02089	*
Topics.t_13	-2.75209	1.00137	-2.748	0.00599	**
Topics.t_14	-5.94884	0.9365	-6.352	2.12E-10	***
Topics.t_10	-6.28785	0.76255	-8.246	< 2e-16	***
Topics.t_6	-6.73126	0.94962	-7.088	1.36E-12	***
Topics.t_12	-9.38139	1.09548	-8.564	< 2e-16	***
Topics.t_15	NA	NA	NA	NA	

Table 8: Logistic Regression summary

5.3. Support Vector Machine (SVM)

The next method performed in this study is the SVM method. In contrast to Logistic Regression, SVM is less interpretable, thus an additional computation is required to zoom in the model and determines what are the most important features among the 15 variables. The permuted important features measure for to SVM model is presented in Table 9. A couple of observations can be highlighted here. First of all, not all of the topics seem have significant impact on the customers' rating based on the importance value. The feature is only considered to be important when the importance value, being measure by the ratio of the errors with and without permutation, is higher than one, meaning the model's error increases after the permutation of a feature. In this case only 4 out of the 15 topics are considered important with the value of higher than 1.05, which are topic 7, 8, 2, and 9. These topics cover only two features: "Delivery Speed" and "Customer Service". This is an interesting observation since these topics are the same topics highlighted in the Logistic Regression. This can reconfirm the observation that "Delivery Speed" and "Customer Service" are the important feature to customers' rating. The other topics like "App convenience", "Price", and "Food quality" do not seem to make significant impact where the importance value is close to 1. This is again aligned with the output of the Logistic Regression model.

Feature	Importance
Topics.t_7	1.14935
Topics.t_8	1.09848
Topics.t_2	1.08874
Topics.t_9	1.07035
Topics.t_5	1.04113
Topics.t_10	1.04113
Topics.t_15	1.02706
Topics.t_14	1.02165
Topics.t_3	1.01515
Topics.t_11	1.01082
Topics.t_6	1.00974
Topics.t_12	1.00866
Topics.t_13	1.00216
Topics.t_1	1.00000
Topics.t_4	0.99459

Table 9: SVM Feature Importance

To build the final SVM model, a grid search was executed to tune the following hyperparameters: (1) the cost factor with a range between 0.1 and 2; and (2) γ with a minimum value of 0.1 and maximum value of 0.5. The final model has the following parameter values: (1) cost equals 2, and (2) γ value of 0.1.

The final SVM model prediction performed on the test dataset to evaluate the prediction power is presented in Table 10 and Table 11 below. Overall, the SVM model can predict slightly better than Logistic Regression in terms of accuracy and sensitivity with 83.03% and 86.15% respectively. The specificity is, however, lower than the one in Logistic Regression with only 65.17%, which indicates that SVM tends to predict positive rating better than negative rating.

		ACTUAL	
		Nothappy	Happy
PREDICTED	Nothappy	**378**	**459**
	Happy	**202**	**2855**

Table 10: SVM Confusion Matrix

Model	**SVM**
Accuracy	83.03%
Sensitivity	86.15%
Specificity	65.17%

Table 11: SVM Prediction

5.4. Random Forest (RF)

The third and also the last model executed in this study is the Random Forest. Similar to SVM, RF model is less interpretable of a model, thus an extra step is need to understand the model and identify the most important features. The outcome of the permutation important features in RF model is presented in Table 12.

Similar to SVM, not all LDA topics have high importance value. Out of 15 topics, only two topics are considered important with the value higher than 1.05, are topic 8 – "Delivery Speed" and topic 7 – "Customer Service". This result is consistent among all three models. The two topics "Delivery Speed" and "Customer Service" are consistently mentioned in all three models as crucial features to customers' rating while the other features do not seem to make significant impact even though they might be mentioned frequently according to the LDA output.

Feature	Importance
Topics.t_8	1.2608247
Topics.t_7	1.1597938
Topics.t_5	1.04330
Topics.t_9	1.01753
Topics.t_2	1.01546
Topics.t_14	1.00928
Topics.t_4	1.00309
Topics.t_15	1.00103
Topics.t_3	1.00000
Topics.t_6	1.00000
Topics.t_12	1.00000
Topics.t_11	0.99897
Topics.t_1	0.99794
Topics.t_13	0.99794
Topics.t_10	0.99588

Table 12: Random Forest Feature Importance

To get the final Random Forest model, three hyperparameters were tuned, which are: (1) the number of trees between 100 and 1000 with a jumping step of 100 between each option; (2) the minimum number of observations in the leaf nodes with values from 15 to 25; and (3) the number of candidate variables at each split with the minimum of 5 and maximum of 10. The final model has the following parameter values: (1) number of trees = 400, (2) number of observations = 24, and (3) number of variables selected = 14. Based on the optimal hyperparameters, the RF model was trained on the train dataset and the performance of the final model is verified by running on the test dataset for prediction. As the final model tree is very complex with many layers with 400 trees, the full tree is not presentable. Instead, we look at the confusion matrix and the model prediction evaluation in Table 13 and Table 14. We can see from the result that the accuracy and sensitivity level of RF is slightly lower than SVM with 81.82% and 83.07% respectively. In addition, the RF model can predict negative ratings better than SVM with specificity level of 74.66%.

		ACTUAL	
		Nothappy	Happy
PREDICTED	Nothappy	**433**	**561**
	Happy	**147**	**2753**

Table 13: RF Confusion Matrix

Model	**Random Forest**
Accuracy	81.82%
Sensitivity	83.07%
Specificity	74.66%

Table 14: RF Prediction

5.5. Models' prediction performance comparison

Out of the three predictive models chosen in this study, SVM yields the best overall prediction performance with accuracy of 83.03%, however, the performance power is very similar between the three models. When it comes to predicting positive ratings, SVM tends to perform the best with the sensitivity of 86.15%.

Model	**Logit Regression**	**Random Forest**	**SVM**
Accuracy	80.95%	81.82%	**83.03%**
Sensitivity	82.45%	83.07%	**86.15%**
Specificity	72.25%	74.66%	**65.17%**

Table 15: Models performance comparison

An intriguing observation is that topic 7 – "Customer Service" and topic 8 – "Delivery Speed" are consistently presented in the outcome of all Logistic Regression, SVM, and RF to be the ones with significant influence to the dependent variables being the review rating. On the other hand, there is a deviation in the model result for the other topics like "App convenience", "Price", and "Food Quality". Logistic Regression model shows that all of these three features also are significant to the customers' rating while SVM and Random Forest tend to disagree with that. The feature importance of both models indicates that "Delivery Speed" and "Customer Service" are the only features that can influence the customers' review rating.

6. CONCLUSION

To answer the main research question of "What key factors of an ultrafast grocery delivery company contribute to the customer's higher rating review?", the answer for the two hypotheses is constructed.

Hypothesis 1: Delivery time, App convenience, Product price, Delivery fee, Customer Service and Product range will be the main service features that are evaluated in the reviews of the ultrafast grocery delivery customers.

In the first hypothesis, six topics are expected to the main themes evaluated by customer in their online reviews. 4 out of these 6 features are confirmed by the LDA models which are "Delivery Speed", "Customer Service", "App Convenience", and "Price". As opposed to the initial speculation, "Delivery Fee" and "Product Range" do not seem to be crucial criteria in customers' view. Instead, a new feature about Food Quality was identified by the model.

Hypothesis 2: Each of the above features has a relationship with the customer rating where the direction of the relationship will differ per feature.

Based on the outcome of the three predictive models being Logistic Regression, SVM, and Random Forest, "Delivery Speed" and "Customer Service" features are confirmed to have significant influence on the customers' review rating. These are the only two features that are consistently determined as the important features across all three models. In opposed of the hypothesis, the other topics, being "App Convenience" and "Price", do not have significant influence the ratings even though they are topics mentioned in documents according to the LDA result. Only Logistic Regression model supports the significance of these features.

To sum up, it can be concluded that "Delivery Speed" and "Customer Service" are the key factors of the ultrafast grocery delivery company that contributes to the customers' higher review rating. While delivery speed makes intuitive sense to be an important factor, customer service is equally important according to the study result. In the end, service is what makes these business companies differ from the traditional grocery shopping industry. This can be used as a preliminary suggestion that companies should focus on service, including delivery service, next to delivery speed.

REFERENCES

Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. *In Proceedings of the 10th Int. Conference on Computational Semantics*, (pp. 13-22).

Basu, A., Watters, C., & Shepherd, M. (2003). Support Vector Machines for Text Categorization. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences* (p. 7). Hawaii: IEEE.

Bennett, J., & Lanning, S. (2007, August). The netflix prize. . *In Proceedings of KDD cup and workshop*, p. 35.

Bitner, M. (1990). Evaluating Service Encounters: The Effects of Physical Surroundings and Employee Responses. *Journal of Marketing*, 69–82.

Blei, D., & Lafferty, J. (2007). A correlated topic model of science. *The annals of applied statistics*, 17-35.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993-1022.

Bolton, R. (1998). A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction. *Marketing Science*, 45–65.

Bolton, R., Lemon, K., & Verhoef, P. (2004). The Theoretical Underpinnings of Customer Asset Management: A Framework and Propositions for Future Research. *Journal of the Academy of Marketing Science*, 271–292.

Bouchet-Valat, M. (2020). SnowballC: Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library. Retrieved from https://CRAN.R-project.org/package=SnowballC

Brand, C., Schwanen, T., & Anable, J. (2020). 'Online Omnivores' or 'Willing but struggling'? Identifying online grocery shopping behavior segments using attitude theory. *Journal of Retailing and Consumer Services*, 102195.

Breiman , L. (1996). Bagging predictors. *Machine Learning*, 123-140.

Breiman, L. (2001). Random Forest. *Machine Learning*, 5-32.

Büschken, J., & Allenby, G. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, 953-975.

Calheiros, A., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, 675–693.

Chang, J., Boyd-Graber, J., Gerrish, S., Chong, W., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems* (pp. 288-296). Vancouver: Advances in Neural Information Processing Systems.

Chowdhury, G. (2003). Natural language processing. *Annual review of information science and technology*, 51-89.

Chu, J. (2010). An Empirical Analysis of Shopping Behavior Across Online and Offline Channels for Grocery Products: The Moderating Effects of Household and Product Characteristics. *Journal of Interactive Marketing*, 251-268.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society, Series B (Methodological)*(20(2)), 215–242. Retrieved from http://www.jstor.org/stable/2983890

De Keyser, A., Lemon, K., & Keiningham, T. (2015). *A Framework for Understanding and Managing the Customer Experience.* Cambridge: MA: Marketing Science Institute.

Devika, M., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, 44-49.

Dickinger, A., Lalicic, L., & Mazanec, J. (2017). Exploring the generalizability of discriminant word items and latent topics in online tourist reviews. *International Journal of Contemporary Hospitality Management*, 803–816.

Doorn, V. (2010). Customer Engagement Behavior: Theoretical Foundations and Research Directions. *Journal of Service Research*, 253–666.

Drucker, H., Wu, D., & Vapnik, V. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks,*, 1048-1054.

Dumais, S., Platt, J., Sahami, M., & Heckman, D. (1998). Inductive Learning Algorithms and Representations for Text Categorization. *7th International Conference on Information and Knowledge Management*, 148-155.

Fikar, C., Mild, A., & Waitz, M. (2021). Facilitating consumer preferences and product shelf life data in the design of e-grocery deliveries. *European Journal of Operational Research*, 976-986.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 1-81.

Ganu, G., Elhadad, N., & Marian, A. (2009, June). Beyond the stars: improving rating predictions using review text content. *WebDB*, pp. 1-6\.

Guo, Y., Barnes, S., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, 467–483.

Howard, J., & Sheth, J. (1969). *The Theory of Buyer Behavior.* New York: John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). New York: Springer. doi:10.1007/978-1-4614-7138-7

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 429-449.

Jensen, K., Yenerall, J., Chen, X., & Yu, T. (2021). US Consumers' Online Shopping Behaviors and Intentions During and After the COVID-19 Pandemic. *Journal of Agricultural and Applied Economics*, 416-434.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European conference on machine learning* (pp. 137-142). Berlin: Springer.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Norwell: Kluwer Academic Publishers.

Jones, T. (2021). textmineR: Functions for Text Mining and Topic Modeling. Retrieved from https://CRAN.R-project.org/package=textmineR

Kohavi, R. a. (1998). On applied research in machine learning. In Editorial for the special issue on applications of machine learning and the knowledge discovery process. *Machine learning*, 127-132.

Korfiatis, N., & Stamolampros, P. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*, 472–486.

Kuhn, M. (2021). caret: Classification and Regression Training. Retrieved from https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. *Springer.*, New York.

Kwartler, T. (2017). *Text mining in practice with R.* New Jersey: John Wiley & Sons.

Lawrence, A. (1955). *Quality and Competition.* New York: Columbia University Press.

Leopold, E., & Kindermann, J. (2002). Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? . *Machine Learnin*, 423–444.

Liu, B. (2012). *Sentiment analysis and opinion mining.* Califonia: Morgan & Claypool Publishers.

Lu, B., Ott, M., Cardie, C., & Tsou, B. (2011). Multi-aspect sentiment analysis with topic models. *2011 IEEE 11th international conference on data mining workshops* (pp. 81-88). IEEE.

Maltese, I., Le Pira, M., & Marcucci, E. (2021). Grocery or @grocery: A stated preference investigation in Rome and Milan. *Research in Transportation Economics*, 101096.

McAuley, J., Leskovec, J., & Jurafsky, D. (2012). Learning attitudes and attributes from multi-aspect reviews. *2012 IEEE 12th International Conference on Data Mining*, 1020-1025.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models.* London: Chapman and Hall.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262-272). Edinburg: Association for Computational Linguistics.

Molnar, C., Bischl, B., & Cas, G. (2018). iml: An R package for Interpretable Machine Learning. *Journal of Open Source Software, 3*, 786. doi:10.21105/joss.00786

Morganosky, M., & Cude, B. (2000). Consumer response to online grocery shopping. *International Journal of Retail & Distribution Management*, 17-26.

Pieters, R., Baumgartner, H., & Allen, D. (1995). A Means-End Chain Approach to Consumer Goal Structures. *International Journal of Research in Marketing*, 227-244.

Pucinelli, N., Goodstein, R., & Grewal, D. (2009). Customer Experience Management in Retailing: Understanding the Buying Process,. *Journal of Retailing*, 15-30.

R Core Team. (2021). R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org/

Ramus, K., & Asger Nielsen, N. (2005). Online grocery retailing: what do consumers think? *Internet Research*, 335-352.

Rese, A., Schreiber, S., & Baier, D. (2014). Technology acceptance modeling of augmented reality at the point of sale: Can surveys be replaced by an analysis of online reviews? *Journal of Retailing and Consumer Services*, 869-876.

Ribeiro, Tulio, M., Singh, S., & Gue, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *Knowledge Discovery and Data Mining*, 1-15.

Roder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). Shanghai: Association for Computing Machinery.

Rose, S., Clark, M., Samouel, P., & Hair, N. (2012). Online customer experience in e-retailing: an empirical model of antecedents and outcomes. *Journal of Retailing*, 308-322.

Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* (pp. 813-830). IEEE.

Schwarz, C. (2018). ldagibbs: A command for topic modeling. *The Stata Journal*, 101-117.

Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Open Journal*, 3. doi:10.21105/joss.00037

Singh, R., & Söderlund, M. (2020). Extending the experience construct: an examination of online grocery shopping. *European Journal of Marketing*, 2419-2446.

Stanton, J. (2018). A brief history of food retail. *British Food Journal*, 172-180.

Vapnik, V., & Cortes, C. (1995). Support-vector networks. *Machine Learning*, 273–297.

Westbrook, R., & Oliver, R. (1991). The Dimensionality of Consumption Emotion Patterns and Consumer Satisfaction. *Journal of Consumer Research*, 84–91.

Wickham, H. (2021). rvest: Easily Harvest (Scrape) Web Pages. Retrieved from https://CRAN.R-project.org/package=rvest

Wickham, H., Francois, R., & Henr, L. (2021). dplyr: A Grammar of Data Manipulation. Retrieved from https://CRAN.R-project.org/package=dplyr

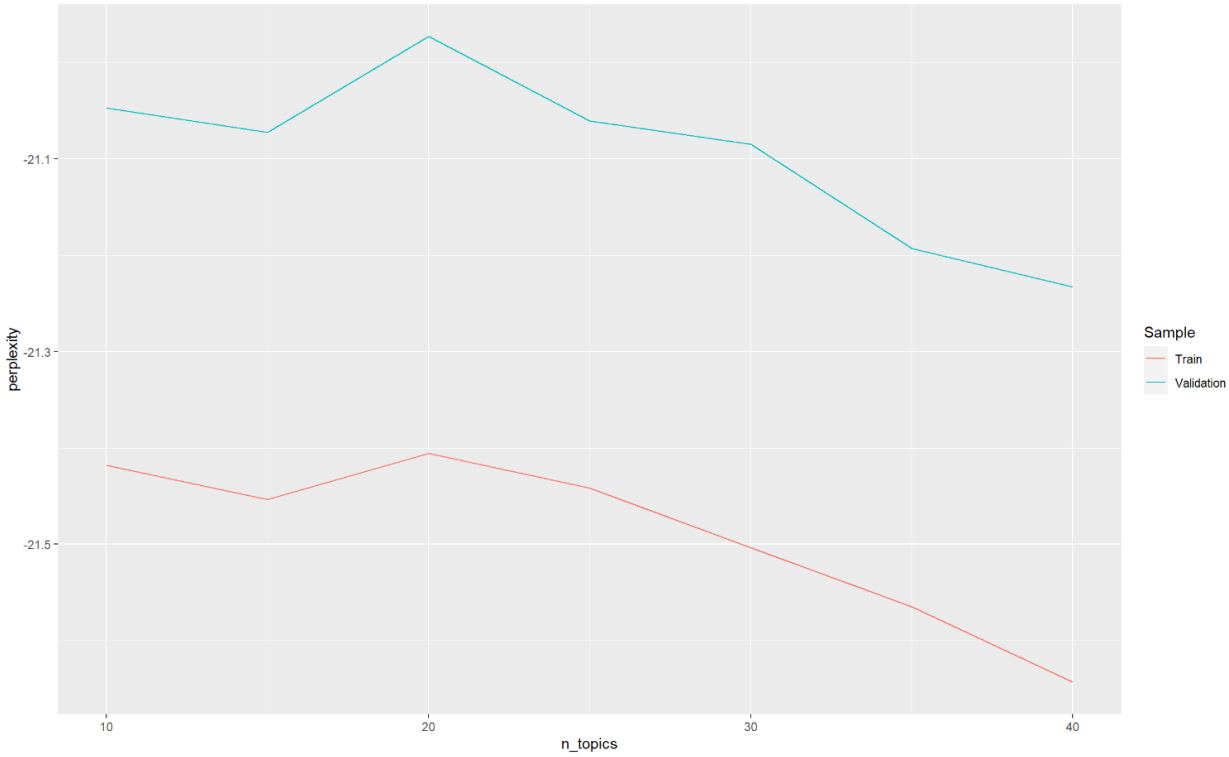
Wolfinbarger, M., & Gilly, M. (2001). Shopping online for freedom, control, and fun. *California Management Review*, 34-55.

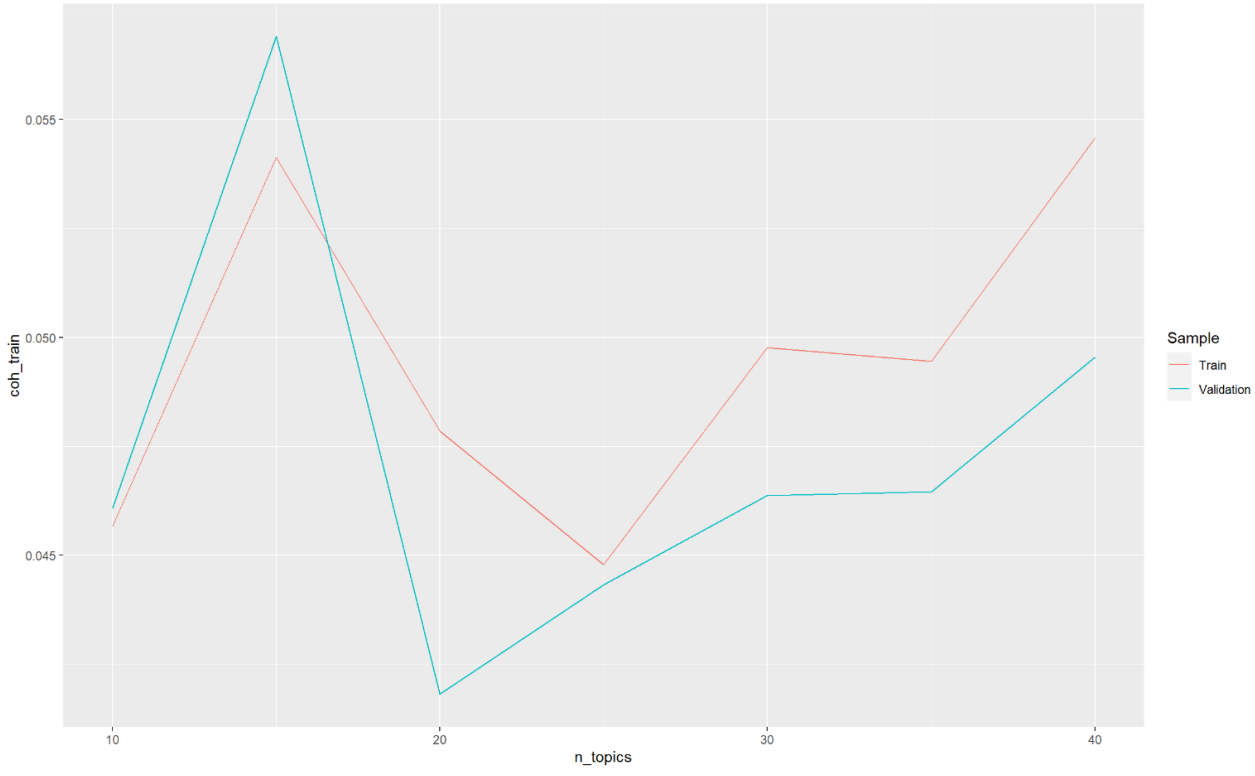
Xie, H. M. (2011). Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition. *International Journal of Hospitality Management*, 178–183.

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 180–182.

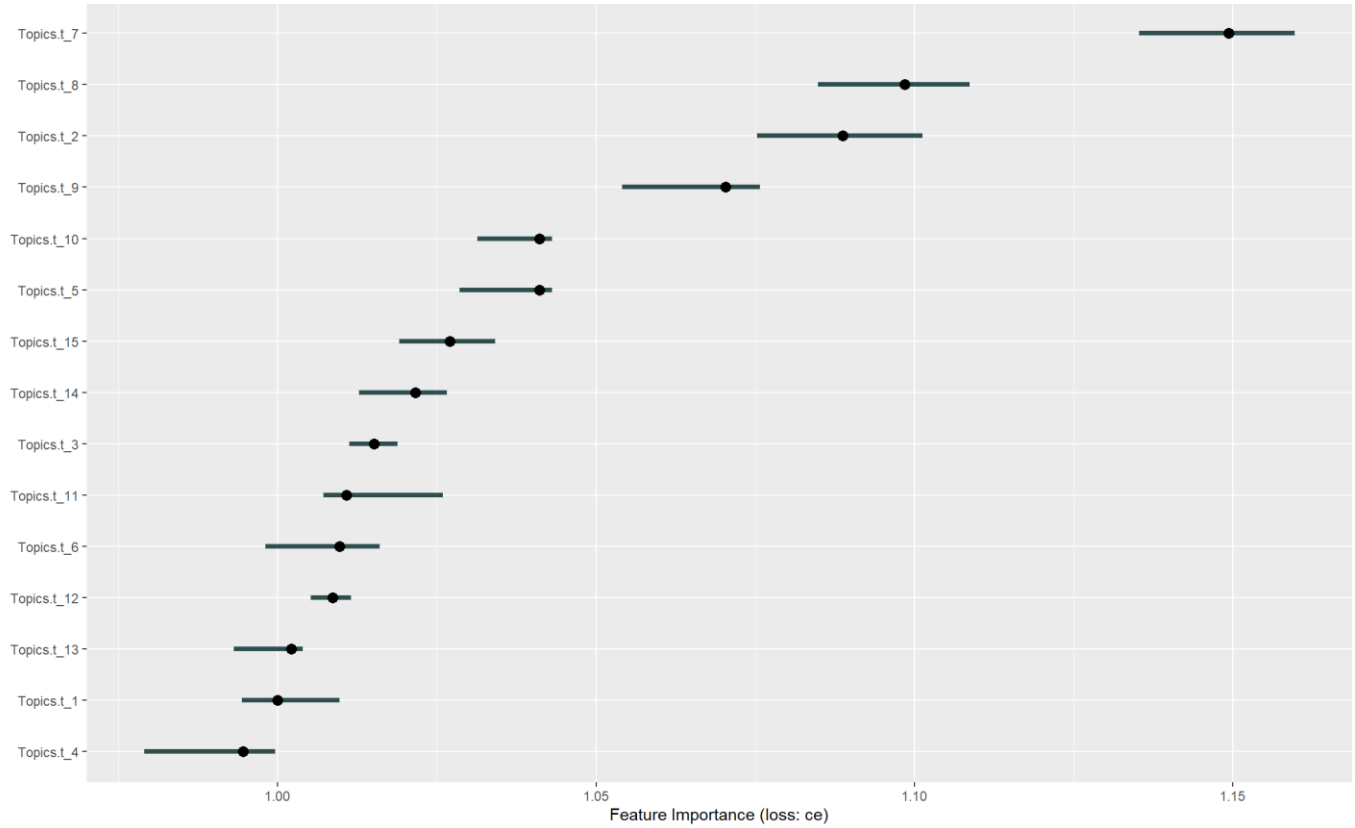
APPENDICES

Appendix 1: Final LDA perplexity and coherence





Appendix 2: SVM Feature Importance



Appendix 3: Random Forest Feature Importance