MINING THE RIDE-HAILING SERVICE: A MALAYSIA CASE

SHARON WONG HE YONG

BACHELOR OF SCIENCE (HONS)

STATISTICAL COMPUTING AND OPERATION RESEARCH

FACULTY OF SCIENCE UNIVERSITY TUNKU ABDUL RAHMAN

OCTOBER 2017

MINING THE RIDE-HAILING SERVICE: A MALAYSIA CASE

By

SHARON WONG HE YONG

A project report submitted to the

Department of Physical and Mathematical Science

Faculty of Science University Tunku Abdul Rahman

in partial fulfilment of the requirements for the degree of Bachelor of Science (Hons)

Statistical Computing and Operation Research

October 2017

ABSTRACT

In recent years, e-hailing services such as Uber and GrabCar have earned a lot of popularity in conjunction with the advance smartphone technology. However, this innovative mode of transport has threatened the glory of taxi industries. In this research, a preliminary attempt is to utilise the available data set to discover some hidden information and insights brought by ride-hailing services. A secondary data sets with a sample size of about 400 respondents was adopted where this data set was collected in Klang Valley. Data mining process will be performed through SAS Enterprise Miner and models will be built via logistic regression, decision tree and neural network. Also, clustering and segmenting customers are involved to examine customer's behaviour. Customers are segmented into three cluster, those who preferred e-hailing service, favoured taxi service and without any preference. Result shows that experience one-hailing service, present or absent of e-hailing apps, availability of technology and customer satisfaction on e-hailing service are significantly affecting customer preferences. While comparing among the three models built, neural network appears to be fitter in multiple-step predictive process.

Keywords: Data mining, Ride-hailing service, Taxi service, Customer preferences

ACKNOWLEDGMENT

It is a pleasure to express my deep sense of thanks and gratitude to everyone who assistances and supports me throughout this research project. First of all, the completion of this research could not have been possible without the persistence guidance and advice from my supervisor, Mr. Looi Sing Yan. I sincerely appreciated him for sparing his precious time for sharing his knowledge, opinions and personal experience regrading this project. I would also extend my gratitude to Mr. Mohan a/l Selvaraju who provided the secondary dataset to me.

Also, I would like to thank SAS Institute for its great invention, SAS Enterprise Miner. SAS Enterprise Miner is the main software adopted in this research, its user friendly feature has provided much ease for me to complete this research. Moreover, I thank profusely to University Tunku Abdul Rahman (UTAR), Kampar, for bearing the license cost of SAS Enterprise Miner as well as offering academic resources to me.

Lastly, to all friends, family and others who in one way or another shared their support, either financially or physically, thank you and I am very grateful.

DECLARATION

I hereby declare that the project report is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

_____

SHARON WONG HE YONG

DATE:

APPROVAL SHEET

This project report entitled "**MINING THE RIDE-HAILING SERVICE: A MALAYSIA CASE**" was prepared by SHARON WONG HE YONG and submitted as partial fulfilment of the requirements for the degree of Bachelor of Science (Hons) Statistical Computing and Operation Research at University Tunku Abdul Rahman.

Approved by:

_____

(Mr. Looi Sing Yan) Date: ……

Supervisor

Department of Physical and Mathematical Science

Faculty of Science

University Tunku Abdul Rahman

FACULTY OF SCIENCE

UNIVERSITY ABDUL RAHMAN

Date: _____

PERMISSION SHEET

It is hereby certified that **SHARON WONG HE YONG** (ID No: 14ADB03805) has completed this final year project entitled "**MINING THE RIDE-HAILING SERVICE: A MALAYSIA CASE**" under the supervision of Mr. Looi Sing Yan from Department of Physical and Mathematical Science, Faculty of Science.

I hereby give permission to the University to upload the softcopy of my final year project in pdf format into the UTAR Institutional Repository, which may be made accessible to the UTAR community and public.

Yours truly,

_____

(SHARON WONG HE YONG)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST IF ABBREVIATIONS

ANN	Artificial Neural Network
ANOVA	Analysis of Variance
CART	Classification and Regression Tree
CRM	Customer Relationship Management
DT	Decision Tree
DWT	Discrete Wavelet Transform
KDD	Knowledge Discovery in Databased
KNN	K Nearest Neighbour
LDA	Linear Discriminant Analysis
MLP	Multilayer Perceptron
NN	Neural Network
PNN	Probabilistic Neural Network
QDA	Quadratic Discriminant Analysis
SAS	SAS Enterprise Miner Workstation
SERVQUAL	Service Quality
SMD	Scaled Mean Deviation
SPAD	Land Public Transport Commission
SPSS	Statistical Package for the Social Sciences
SVM	Support Vector Machine
SWE	Shear Wave Electrography
TMC	Taxi Market with Competition
UDMT	Unified Data Mining Theory

CHAPTER 1

INTRODUCTION

1.0 Introduction

In Chapter 1, it starts off with background of the studies along with its motives. It included problem statements and objectives of this research as well.

1.1 Background of Study

The conveyance of human or goods from one place to another is known as transportation or travel (Merriam-Webster, n.d.). There are wide range of travel options and in recent years, e-hailing services such as Uber and GrabCar have earned a lot of popularity in conjunction with the advance smartphone technology- mobile application. According to Commercial Vehicles Licensing Board, a formal definition for e-hailing service is "*a motor vehicle having a seating capacity of four persons and not more than eleven persons (including the driver) used for the carriage of persons on any journey in consideration of a single or separate fares for each of them, in which the arrangement, booking or transaction, and the fare for such journey are facilitated through an electronic mobile application provided by an intermediation business.*" (Parliament of Malaysia, 2017). Note that the term "ride-hailing" and "e-hailing" imply different concept and meaning. The National Aging and Disability Transportation Center (2016) had defined ride-hailing service as "*when an individual personally summons or hails a vehicle, is immediately picked up and driven to their destination, usually alone*." In a word, ride-hailing encompasses both taxi and e-hailing service.

The e-hailing industry in Malaysia is now dominated by Uber Technologies Inc. and Grab Malaysia. Uber was launched in Malaysia in January 2014 (Ee, 2014) whereas GrabCar is in May 2014 (Chi, 2014). Both companies penetrated into Malaysia market for less than five years, however, over this short period, both had established over 10 million subscribers and users (SimilarWeb LTD, 2017). The growing of this industry is undeniably significant. In addition, the popularity of this service has allow it to gain reorganization from Land Public Transport Commission (SPAD) soon as legalized public service vehicle subject to an intermediation business licence despite the strong oppose from taxi drivers (Nik Imran Abdullah, 2017).

This innovative transportation mode has definitely catered to more drivers and mass consumer base. Unfortunately, as e-hailing expends, problems and doubts grow along. E-hailing has threatened the glory of taxi industry in term of demands and revenues. Protest against e-hailing service by a group of 50 taxi drivers outside of the parliament was held in March 2017. The group spokesman, Zailani Sausudin has cited that the unfair legal procedures applied on taxi drivers are the main trigger of this protest (Akil Yunus, 2017). Besides, a hapless pregnant women was robbed by Uber driver and suffered from miscarriage (Jo Timbuong, 2017). After the incident, SPAD responded that they will fine-tune necessary actions to be taken to ensure consumer's safety and comfort during the trip (Yusof M., 2017). Also, car ownership in Malaysia is the third highest in the world, with 93% of owning, leasing or renting a car which causes vehicle to be a symbol of Malaysia (Emma Howard, 2016). Congestion in capital of Malaysia, Kuala Lumpur is notorious. Apparently, Uber and GrabCar claimed

that they will be able to resolve the jam or queue in Kuala Lumpur (Emma Howard, 2016). But, is the claim or promise reliable? After revealing all these e-hailing related issues, it is understood that the outputs from e-hailing service is uncertain and vague, yet these outputs might adversely impact on the industry itself as well as causes unnecessary issues to the society.

Thus, to unseal information associated with e-hailing, a studies on this industry is conducted conjointly with data mining techniques. Kindly be noted that taxi industry acts as a complementary case in this research where it is used to compare with e-hailing service. There are several definitions for data mining which more or less carried similar meaning. At times, some might see data mining as knowledge discovery in databased (KDD), but in fact, data mining is part of the KDD process (Vicen¸c Torra, Josep Domingo-Ferrer and Angel Torres, 2001). The differences between data mining and KDD are not the main concern of this research, hence both terms are equivalent within this research context. According to Wee Keong Ng, et al. (2006), *data mining is the process of posing queries and extracting pattern, often previously unknown from large quantities of data using pattern matching or other reasoning techniques*. A deeper discussion on data mining will be tender in the methodology section.

1.2 Problem Statement

A common practice from many organizations, they tend to employ analytical tools to have a better understand on their customers or gain insight for their business, especially on new industry. So, what issues normally posed during this analysis? In this sub section, potential questions aiding in mining e-hailing service are posed.

1.2.1 E-hailing Users in Malaysia

In business perceptive, understanding their customers is fundamental. Paying extra attention to customers is a strategic move that enable business to establish loyalty from customers that will advocate for its brand and products (Mark L. Blazey, 2009). Without this level of insight, organization is exposed to frustrate and high chance of annoying customers with well-intention of serving them (Harley Manning, Kerry Bodine and Josh Bernoff, 2012). Further, it may cause unnecessary cost to incur.

1.2.2 Factors Affecting the E-hailing Industry

As an infant industry, E-hailing service is subjected to wide range of variabilities. However, not all variables will guarantee a significant impact towards e-hailing. Commuters or trip makers are rational being who make choices based on maximum benefits. By choices, it refers to mode of transportation.

1.3 Objective of Research

Four objectives were proposed to pursue each problem statement from the previous section and listed as below:

i. To identify significant factors associated with choice of transportation mode which refers to e-hailing or taxi service.

ii. To discover ride-hailing users' characteristic by segmentation and profiling attribute

iii. To build ride-hailing predictive modelling by using different data mining techniques which include decision tree, logistic regression and neural network.

iv. To compare the predictive performance among these models as well as between single-step and multiple-step data mining.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

Several works and papers were reviewed in this chapter in order to present the overview of ride-hailing service and the possible significant factors that affecting the preference of transportation mode. This chapter was divided into two sub-sections, studies associated to ride-hailing service as well as data mining general framework and applications.

2.1 Related Studies on Ride-hailing Service

As the e-hailing industry blooming, a lot of recent research effort has been considered. Most of the research attentions are on the performance of e-hailing service. As a part of the service industry, measuring service quality or performance are arduous as it is intangible and can be very subjective at times. Anderson (2016) has illustrated how e-hailing (drivers) performance was examined through interviewing three e-hailing drivers. Basically, three aspects were monitored, namely, control of pay and work, control of information and monitoring performance. A short description for each aspect were as follow:

i. *Control of work and pay*. Price were pre-determined based on time (normal or peak hour) and distance. Company also actively engaged in changes in pricing, pay incentives, and income guarantee as to test the impact on drivers and ridders.

ii. *Control of information*. Companies strictly controlled information (either drivers or ridders) through the mobile app. Information of ridders were privately disclosed to the matched drivers and vice versa via their personal electronic gadget.

iii. *Monitoring performance*. It depends on multi-dimensions including notably acceptance and cancellation rates, and five stars rating system.

Factors affecting ones' preference towards public transport are also reviewed in this section as a referencing factors which might influences ride-hailing's performance. Sam, Adu-Boahen and Kissah-Korsah (2014)'s study found that fare, perceived safety and accident record, comfort and quality of vehicle as well as reliability in terms of timing schedule are the main consideration of students for transportation mode selection. A study of investigating factor driving commuters' intention and ride-sharing (e-hailing) in Jakarta by Krontalis (2016) through three antecedents of behavioral intention: Attitude towards ride-sharing; Perceived Behavior Control, and Personal Norm. The result has shown that these three antecedents of behavioral had a significant and positive relationship on intention to participate in ride-sharing. The research also found that ride-sharing behavior is affected by ride-sharing habit and intention to participate in ride-sharing. Another study on taxi market equilibrium with third-party hailing service (e-hailing) is exemplified by Qian and Ukkusuri in 2017, where the characteristics of taxi and e-hailing was analysed, the taxi market as a multiple-leader-follower game at the network level was modelled and the equilibrium of taxi market with competition (TMC Equilibrium) was investigated through the use of network modelling. The numerical results indicated that pricing policy

and fleet size were highly associated with the level of competition in the market and existed significant impact on total passengers cost, average waiting time, and fleet utilization.

2.2 Data Mining General Framework and Applications

Data mining has been extensively applied in different areas like marketing, optimizing inventories, mining criminals and so on. In fact, some problems might favor a particular or specific set of algorithms. For example, Classification and Regression Tree (CART) is very useful in mixed categorical and continuous data with high-dimensional space but contexts where the true decision boundaries between classes are described by a second order polynomial is less effective when CART is used. An important point is that, there is no 'universal' data mining method or framework (Fayyad, Piatetsky-Shapiro and Smyth, 1996). As discussed in Chapter 1, there are several definitions for data mining and the same applies to the framework of data mining. Data mining is said to be very dynamic, it does not have a consistent or fix body of theory. In most of the time, the definition and focus of data mining differ primarily as a matter of experience (most of the time referring to prior knowledge), context and necessity. The lack of consistent theory and universal method for data mining process has become one of the major challenges for most researcher in applying data mining techniques as a single context can be analyzed through various methods with different outputs. Hence, numerous criteria that should be fulfil in an approach to develop a model-theoretic for data mining were listed by Mannila (2000) as follow:

- *Model typical data mining tasks (clustering, rule discovery, classification)*
- *Describe data and the inductive generalizations derived*
- *Express data in diverse form*
- *Encourage iterative and iterative process*
- *Express comprehensible relationships*
- *Incorporate users in the process*
- *Incorporate multiple criteria for defining an 'interesting' discovery*

Mannila (2000) also cited that, none of the approach developed satisfied all the listed criteria. Khan, Mohamudally and Babajee (2013) argued that data mining is a multiple-step process rather than a single-step process and proposed a theoretical framework that unifies clustering, classification and visualization- Unified Data Mining Theory (UDMT). These individual data mining techniques (clustering, classification and visualization) are able to generate outputs independently by means of single-step process. While UDMT is a composition of clustering, classification and visualization, labelled as multiple-process. To be precise, clustering algorithms are first applied, then classification algorithms, subsequently visualization and the results are based on the interpretation of visualization. The composition of functions is illustrated in Figure 2.1

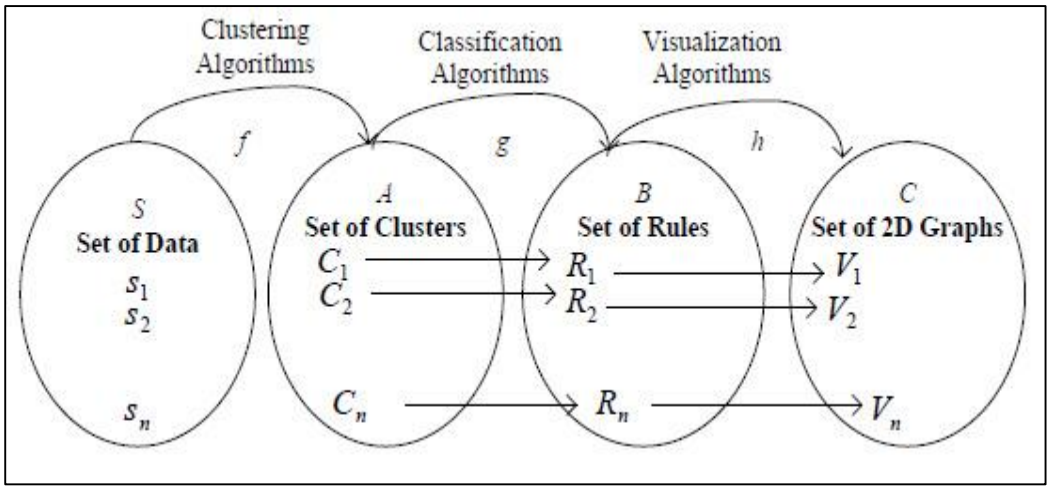


Figure 2.1 The Composition of Functions (Khan, Mohamudally and Babajee, 2013. p.3.)

Bahari, T.F. and Elayidom, M.S. (2015) from Cochin University of Science and Technology, Kochi proposed an effective Customer Relationship Management (CRM) Data Mining framework. CRM is essential since it provides a meaningful communication and improve customer acquisition, customer loyalty, customer profitability and customer retention. In their research, two classification models (or data mining techniques), Neural Network and Naïve Bayes were studied as to monitor CRM. The result of the studies indicated that Neural Network is much better comparatively. The CRM-data mining framework aids in identifying valuable customers, predicts the future customer's behavior and eventually allows organizations to develop their business strategy based on the results obtained. The first phase of the CRM-data mining framework is to understand the business goals and requirements. Next, data pre-processing such as cleaning, data transformation, attribute selection was performed followed by models construction and evaluation. The last phase is visualization that illustrate the result obtained.

In medical field, Acharya, et al. (2017) narrated a unique algorithm for an automated characterization of the benign and malignant breast lesions using Shear wave electrography (SWE) images. SWE *is an imaging technique using ultrafast ultrasound (20k fps) to measure tissue elasticity* (David A. Jaffray, 2015). The framework proposed data pre-processing as the initial phase. As data obtained in this research are images, the initial phase involved merely standardization of images size and colour conversion. Next phase is to extract quantifiable measurement from these images known as feature extraction, which are three levels of Discrete wavelet transform (DWT), Second order statistics (Run Length Statistics) and Hu's moments followed by feature selection and ranking so that supremely resourceful features with useful information is recognized. Lastly, classification such as decision tree (DT), K nearest neighbor (KNN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), support vector machine (SVM), and probabilistic neural net-work (PNN) are implemented.

In facing the challenges of big data, a distributed framework of artificial neural network (ANN) is introduced by Chang and Liou (2016) to manage big data real time analysis generate acceptable outcome in a very short delay. In his paper, the main focus was on the data mining process adopted which is a three simple stages process, per-processing, mining and post-processing. Data partition and transformation was performed in the first stage. In the mining stage, ANN algorithm was executed. Finally, two steps were involved in the last stage- post-processing, namely formulating and back-testing. A trading strategy was

formulated into mathematical model through ANN later it was reused for back-testing propose on the simulation set (validation set).

In short, formulation of data mining framework for a specific context is complicated and tortuous. However, there are a few overlapping steps in these reviewed papers which is sufficient to provide an outline for mining the e-hailing service. A further discussion on the framework for e-hailing service is held in the following chapter.

CHAPTER 3

METHODOLOGY

3.0 Introduction

Instruments or software used in this research is revealed in this section. Yet, introducing data mining and machine learning concepts, specific framework adopted as well as algorithms and techniques used in this research. The framework describes all of the process involved in this research.

3.1 Deployed Software

Two main softwares are involved in this research: Statistical Package for the Social Sciences (SPSS) and Statistical Analysis System (SAS) Enterprise Miner Workstation. The formal software is used for simply data cleaning process while the latter one is used for data analysis. Also, Microsoft Excel is deployed for visualization of data.

3.2 Data Mining Process

Data mining can be categorised into two different type, namely verification-oriented and discovery-oriented. The verification type utilised common verification techniques like *t*-test, goodness-of-fit test, analysis of variance (ANOVA) and so on to evaluate or verify the hypothesis proposed by an external user, such as expert of a particular field. While the discovery type is about

discovering new rules and patterns from a dataset. In this research, the main focus will be on discovery type. Both descriptive and predictive methods are included in the discovery type. These two methods are also known as unsupervised and supervised method respectively in the traditional machine learning context (Lior and Oded, 2014).

Descriptive method (unsupervised) is often used to extract understandable and interpretable information from data in order to reveal concealed knowledge, including dependencies and characteristics from a particular data set. Hence, descriptive adaptation normally provides pattern that is understandable, straightforward but least accuracy to fulfil compared to predictive method. In contrast, predictive method (supervised) aims to learn models that are subsequently use for classification like predicting a certain class (Martin, 2006). While explaining from the machine learning perceptive, unsupervised learning transpired when there is only input variables (*X*) to be studied. Whereas, supervised learning is where the present of both input variables (*X*) and output variable (*Y*), and through algorithms to learn the mapping from input to output (John, 2016). Refer to Figure 3.1 for a finer illustration on the data mining concept or sub-sections.

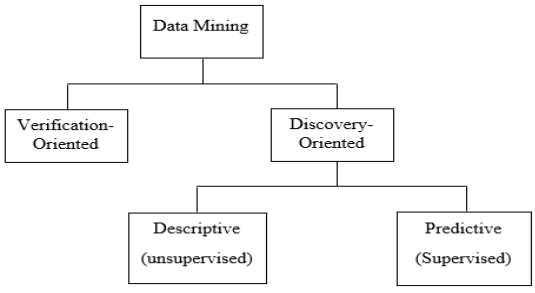


Figure 3.1: Data Mining Concept.

3.3 Machine Learning

The use of data mining is often overlaps with machine learning. All data mining, predictive techniques employed in this research are developed within a field known as machine learning. Particularly, machine learning can be defined as "*a set of methods that can automatically detect patterns in data, and then use the uncovered pattern to predict future data, or to perform other kinds of decision making under uncertainly*." (Murphy, 2012). Its algorithm requires split-sample test, where the raw dataset is split or partitioned into three sets: training set, validation set and test set through random sampling. The training set is a set of samples used to train or fit the models; validation set is to estimate prediction error for model selection and test set is to determine the generalization error of the final model chosen (Priddy and Keller, 2005). The usage of these three sets will be explained in detail on the following section.

3.4 Framework Adopted

The general procedure for this research is shown in Figure 3.2. Before performing discovery, there are some preliminary steps have to be taken. These steps are often known as data pre-processing and are explained in the following sections together with the references on Figure 3.2. Data pre-processing includes data selection, data cleaning and data transformation. Evaluation and interpretation as well as model comparison will discuss on Chapter 4.

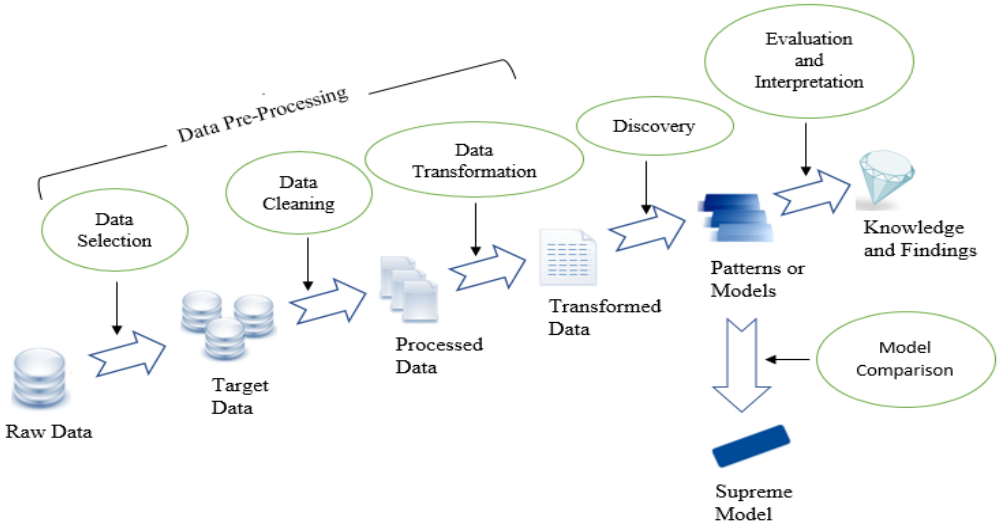


Figure 3.2: Framework adopted in this research.

3.4.1 Data Selection

Initially, there are 400 observations (also refer as samples or customers in this context) with 34 variables in the raw data. The foremost step in mining a dataset is to identify and select the target or predict variable. The target variable selected in this study is customers' transportation mode preference, either favour taxi service or e-hailing service. The target is a binary variable and it is coded as 0 if customer favoured taxi service, 1 if customer chose e-hailing service. Later, a brief exploration over the selected target variable is done. The result is shown in Figure 3.3.

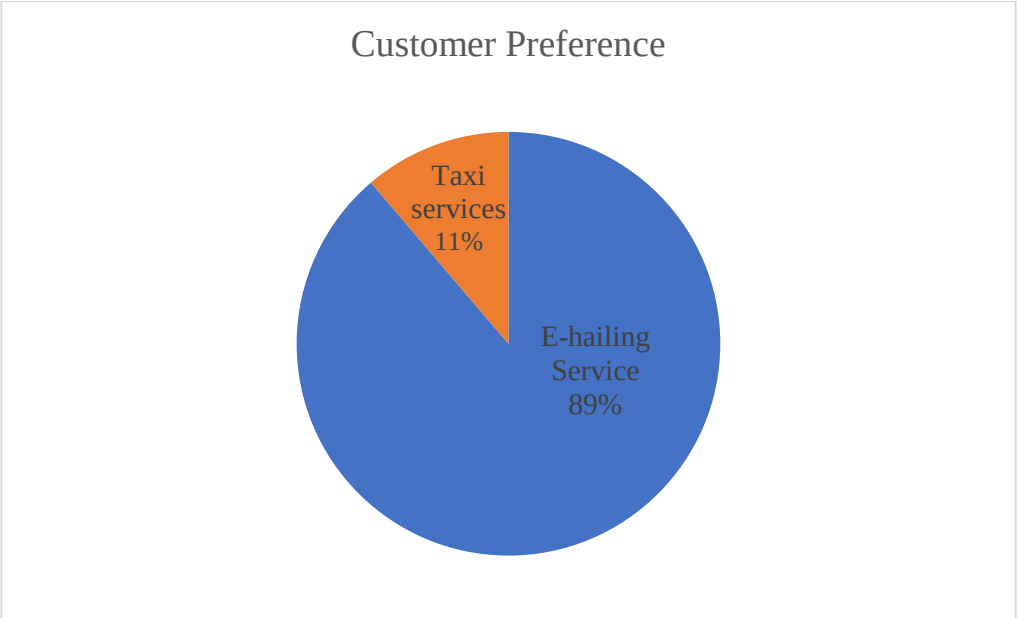


Figure 3.3: Preferences on either taxi service or e-hailing service.

Observed that the blue portion in Figure 3.3 is much larger than the orange portion. The blue portion is the majority group whereas orange portion is the minority group. In precise, the blue portion represents 335 customers who choose e-hailing service over the taxi service (red portion) with merely 45 customers. The ratio of taxi versus e-hailing is about 0.1125: 0.8875. This situation arises as imbalanced target. In most cases, performance of classifiers like decision tree, logistic regression as well as neural network is affected by distribution of the target variable (Bee Wah, et al., 2014). Since unbalanced target is likely to cause the standard machine learning algorithms to classify all observations into the majority class (Andrea, et al., n.d.). Also, the majority group has much higher possibility to be selected. Random under-sampling is one of the most commonly used method to deal with this situation. It is defined as technique used to reduce the frequency of apparent majority group by randomly deleting observations from the majority group (Rhupesh and Kiren, 2016). The random selection and deletion of observations is done through SPSS. The event

rate for minority group, taxi service improved from 11.25% to 50% after performing randomly under-sampling. At this rate, the ratio of taxi service versus e-hailing service is even or fair and the sample size is now reduced to 90. Although random under-sampling is not the only method to address this problem and it has some drawbacks on loss of information and biasness, but it is the simplest yet effective way of doing so. In addition, Japkowicz (2000) also commented that, sophisticated under-sampling and over-sampling appears to be unnecessary. Besides, other method like over-sampling (synthetic observations for minority group) also causes problem such as overfitting, increases run-time as well as complexity of algorithm used (Cranefield and Nayak, 2013).

3.4.2 Data Cleaning

The main purpose of cleaning data is to smooth out noise and inconsistent data and eventually produces quality data for discovery. Data cleaning's regular practice includes impute missing values, remove noise, recognize outliers as well as correct inconsistencies within the data (Han, Pei and Kamber, M., 2011). In the apparent dataset, there is 12.22% and 4.44% of missing values on variable customer satisfaction on e-hailing service and customer satisfaction on taxi service. Imputation of these missing value is done by SAS Enterprise Miner through the utilization of the impute node. The impute method used is mean which replaces missing values with average of all non-missing values (SAS, n.d.). Observe that imputed variables are indicated through a prefix, IMP.

Noted that with the use of Likert Scale yet without including any open-ended questions, there should not be any interval scale. However, all ordinal data is treated as interval in this research. Reason of confronting ordinal data as interval is to allow the use of more powerful statistical methods. Treatment of ordinal scale as interval is not uncommon and has long been controversial (Knapp, 1998). Moreover, Leach (2004) supported this dilemma with the following condition:

i. Scale contain sequential integer. These sequential integers imply the existence of interval values and also prompt respondents that they are allowed to answer on an interval scale.
ii. Scale lies on an evenly divided line or box. These format strong encourages respondents to provide interval answers.
iii. Scale with less than five categorises or the rating level anchors are chosen (multiple choice questions), worded (open-ended questions), or formatted in a way that does not hint any kind of even spacing to the scale, then treat the data as ordinal or nominal.

Serendipitously, the questionnaire used in this research has satisfied the above three conditions and hence the treatment of ordinal scale as interval scale is likely to be reliable and logic.

3.4.3 Data Transformation

Frequently, data violates a lot of fundamental statistic assumption, nonlinearity or inconsistent variance, for instance. Data transformation is a process of transforming or consolidating data into appropriate forms, at least meeting some underlying assumptions. Transform node is use in SAS Enterprise Miner to

undergo transformation. The transform method used in this case is multiple. With this setting, the system makes several transformations for each input and passes them to the following node. Later, a regression node is connected to the transformation node which then utilizes stepwise selection method to select the best transformation for each variable and best variables to be included in the regression model.

Despite the fact that transformation is essential in data pre-processing, there are several downsides of this process. Consider the following problems addressed by Van Bommel (2005):

i. Loss of data. The original dataset is now transformed into a new set of data though only the necessary inputs are transformed. The resulted data might not adequately describe the original dataset.
ii. Incomprehensibility. The effect of the transformation is ambiguous.
iii. Focus on instance. Data instances are transformed, without incorporation of data types.
iv. Focus on types. Data types are transformed, without incorporation of data instances.
v. Correctness. There is no exact definition of a set of correctly transformed data, no provable correctness.

Heeding the matters above, the originality of the dataset is to be maintained at its finest. Therefore, input variables transformation is applied merely on logistic regression since it is a parametric statistic model and it has a strong obligation in satisfying its statistical assumptions. Whereas, no transformation is done on

decision tree. Decision tree is categorised as non-parametric statistic model which can be used with less assumptions. Lastly, artificial neural network has comprised transformation in its algorithm.

3.5 Discovery: Techniques and Algorithms Adopted

The mean of discovery is to undergoes data mining techniques as to obtain a supportive decision model. The following sections are disseminating the conceptual algorithms used in this research (based on SAS algorithms), where decision tree, logistic regression and neural network (NN) is categorised as predictive type whilst clustering and profiling is descriptive type.

3.5.1 Decision Tree

A decision tree is metaphor through the structure of tree, represents a hierarchical segmentation of the data. The entire data set is the original segment and it is known as the root node (sometimes called as parent node) of the tree. At first, the tree is partitioned into at least two segments by applying a series of simple rules. Then, each resulting segment is further partitioned into sub-segments. Again, each resulting sub-segments is further branched into more sub-segments until partitioning is no longer feasible or when the stopping criteria is met. This process is called recursive partitioning and is illustrated in Figure 3.4. The final segments are normally named as terminal nodes or leaf nodes. Noted that all terminal nodes are disjoint subset of the root nodes, meaning to say that there is no overlap among root nodes (Kattamuri, 2013). In general, there are two type of decision tree, classification tree and regression tree. If a predict or output

variable is a continuous type, then regression tree is adopted and if a predicted variable is a discrete type, classification tree is used (Shan, 2015). In this research, classification tree is adopted as the predict variable is a binary type, either favour towards taxi or e-hailing.

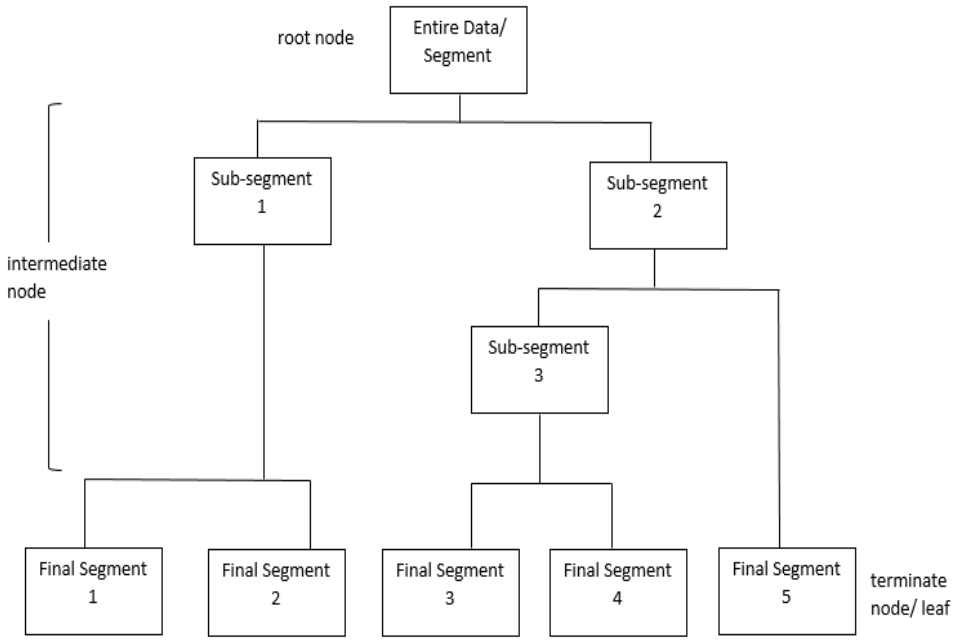


Figure 3.4.: Illustration on decision tree's recursive partitioning.

3.5.2 Logistic Regression

Since a binary target (predict variable) is used in this research, logistic regression is employed. Interpretation of regression coefficient in logistic regression has a slight different compare to the normal regression. They indicate the increase or decrease (in term of positive or negative sign respectively) in the predicted probability of meeting some defined criteria, due to a one-unit change in the input variables. The predict variable itself takes either value of 1 or 0, but the predicted regression coefficients take the form of probabilities conditional on the

values of the input variables (Fred, 2000). In SAS Enterprise Miner, when the software detected the target variable as binary, the regression node will provide a logistic regression with logit link function by default. The estimation of event (in this case, is either favour taxi service or e-hailing) is computed through logit link function which indicate the relationship between the probability of the event and a linear predictor, which is linear combination of inputs. The link function maintains the predicted regression coefficient between 0 and 1 (Kattamuri, 2013). The logit link function and be expressed as follow:

$$log\left(\frac{\Pr(y=1|x)}{1-\Pr(y=1|x)}\right) = \beta' x \tag{3.1}$$

where $\beta' x$ is known as the linear predictor since it is a linear combination of inputs, x is a vector of inputs, β is the vector of predicted regression coefficients. Then, by solving the logit link function (Equation 3.1), probability of response, $\Pr(y=1|x)$ is able to obtain as follow:

$$\Pr(y=1|x) = \frac{\exp(\beta' x)}{1+\exp(\beta' x)} \tag{3.2}$$

3.5.3 Neural Network (NN)

Neural Network emulates the analogy of biological neural network. Specifically, NN extracts important features from the entire dataset with an identified input variable and perform a pattern recognition task by learning from instances without the need of explicitly stating the rule for execution (B. Yegnanarayana, 2009). Similar to biological neural network that takes in inputs and produces outputs. SAS Enterprise Miner offers a neural network node to generate neural network models. It uses mathematical functions to maps inputs to outputs. The

output produced in this case will be probabilities of favour e-hailing and favour taxi service.

Unlike regression, combinations, transformations of inputs as well as model estimation (estimation of weights) can be done simultaneously within the neural network framework and minimizes a specific error function. A NN model is built up by a number of layers, where each layer comprises computing elements known as units or neurons (refer to Figure 3.5).

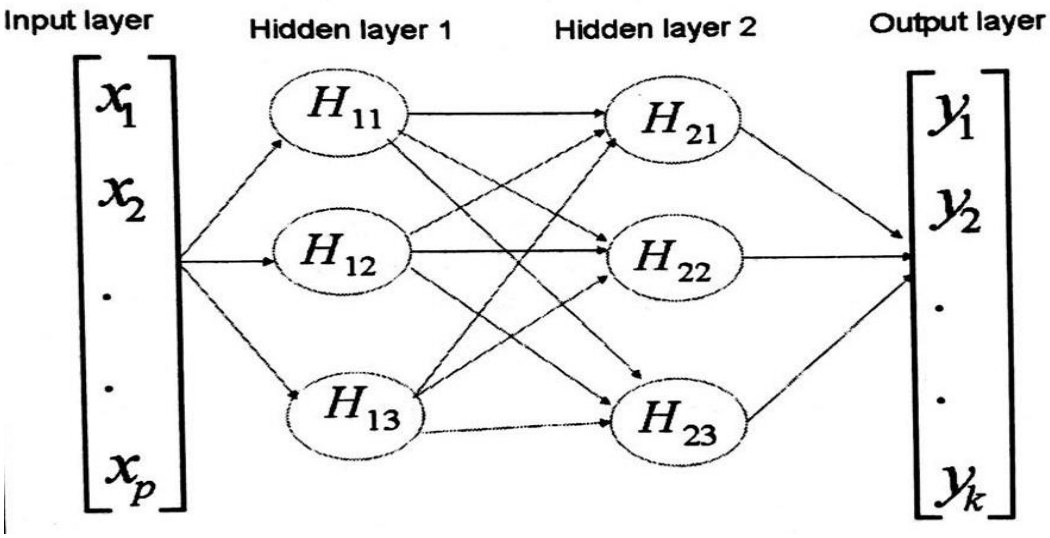


Figure 3.5: A neural network with two hidden layers (Kattamuri, 2013. P.242).

The first layer is the input layer and the last layer is the output layer. The first layer does not involve any combinations and transformations process but it does standardize the inputs. A number of hidden layers lies in between the input and output layers. The units in these hidden layers are known as hidden units which act as an intermediate for calculation. Each unit in a layer receives inputs from the preceding layer and computes outputs for the following layer. By means of computing outputs, hidden units combine inputs (from pervious layer) through

Hidden Layer Combination Functions and transform the combined values using Hidden Layer Activation Functions. Both Hidden Layer Combination Functions and Hidden Layer Activation Functions are options provided by SAS Enterprise Miner. In general, a linear combination function is used to combine inputs. The function is as follow:

$$\eta_k = w_{0k} + \sum_{j=1}^{p} w_{jk}\, x_j \tag{3.3}$$

where η_k represents weighted sum, w_{0k} is the bias coefficient, w_{jk} is the weight of j^{th} input in the k^{th} hidden unit and x_j is the inputs (independent variables). Notice that Equation 3.3 is similar to the linear regression function. Thus, the weight in ANN serves the same purpose as the regression coefficient. Then, transformation is performed through hyperbolic tangent activation function which describe as:

$$H_k = \tanh(\eta_k) = \frac{\exp(\eta_k) - \exp(-\eta_k)}{\exp(\eta_k) + \exp(-\eta_k)} \tag{3.4}$$

The aim of transforming combined values (weighted sum) are to force the fall into the range of -1 to +1. As for the output layer, similar procedures are applied to calculate the weighted sum (through Equation 3.3). Yet, different activation function is used to compute the final outputs. The selection of activation function on the output layer has to be specific and precise as this is the layer that will provide us with predicted values of the targets. As stated earlier, target variable for this research is a binary type, thus logistic activation function is selected. The probability of preferences is calculated as:

$$\pi_j = \frac{\exp(\eta_k)}{1 + \exp(\eta_k)} = \frac{1}{1 + \exp(-\eta_k)} \tag{3.5}$$

By substituting Equation 3.3 into Equation 3.5, an explicit nonlinear function of the weights and the inputs (independent variables) can be denoted by:

$$\pi_i(W, X_i) \tag{3.6}$$

where W is the vector of weights and X_i is the vector whose elements are inputs. In fact, the NN demonstrated above is named as Multilayer Perceptron (MLP), the utilization of linear combination functions together with sigmoid activation functions in the hidden layers.

The procedures for estimation of weights (W) will be examined through a two-steps iterative search on minimized error function- Bernoulli function:

$$E = -2\sum_{i=1}^{n}\left\{y_i ln\frac{\pi_i(W,X_i)}{y_i} + (1-y_i) ln\frac{1-\pi_i(W,X_i)}{1-y_i}\right\} \tag{3.7}$$

where y_i is the preferences of the i^{th} sample or person. To be specific, Equation 3.7 can be simplified as:

$$E = \begin{cases} -2\log\left(1-\pi_i(W,X_i)\right), & y_i = 0\ (favour\ in\ taxi\ service) \\ -2log\pi_i(W,X_i), & y_i = 1\ (favour\ in\ e-hailing) \end{cases} \tag{3.8}$$

Firstly, a set of weights that minimize the error function (Equation 3.7) is generated through iterative procedures. In each iteration, the system alters the weight by a small amount and revises the error function. This procedure terminates when the error cannot be further reduced. Subsequently, *n* set of weights is produced. The system then selects one of the n sets of weights based on a specific Model Selection Criteria. In this research, the Model Selection

Criteria is set as Misclassification. This selection is executed based on the validation dataset.

3.5.4 Clustering and Profiling

The objective of clustering is to examine whether the dataset composed of natural subclasses while segmentation is merely to partition data in a way that is meaningful and useful. Both clustering and segmenting are done without any predefined class (unsupervised). Often, clustering and segmentation is distinguished based on the aim of research (Randall, 2017). Considering one of the objective of this research is to identify natural subclasses that is meaningful or equivalently customer profiling, thus segmenting and clustering are assumed to be unvaried in this study. To be particular, segmentation is done by clustering analysis. SAS Enterprise Miner provides cluster node to create clusters of customer (samples) based on similarity. By mean of similarity, it is the measure of distance among each customer. Noted that cluster node only creates clusters from the input variables without references to any target variable. Meaning to say that if a variable role is set as target, this particular variable will be excluded during the cluster analysis. And hence, cluster analysis does not uncover the relationship between inputs and target variable (Kattamuri, 2013).

Aforementioned, the core of clustering analysis is the similarity among samples. Here, rise of an important question “What unit of measurement should be used when all input variables are varied in terms of their unit?”. Measurements that are widely different in units tend to be incomparable and provide inaccurate

insights. To address this issue, SAS Enterprise Miner allow users to compute scaling through altering the Internal Standardization. This research deployed Standardization method (another method is Range) where the variable values are divided by the standard deviation (SAS, n.d.).

CHAPTER 4

RESULTS AND DISCUSSION

4.0 Introduction

This chapter provide the source of data deployed as well as unveils and elaborates the outcomes generated through both descriptive and predictive methods.

4.1 Data Description

The data set used in this research is a data collection on users' preferences on ride-hailing service and taxi services with a sample size of 400. These samples were collected through conduction of self-administered survey in Klang Valley region. A brief description of the questionnaire structure is showed in Table 4.1. Sample of the questionnaire is attached in Appendix A. All sections adopted 5-point Likert scale except for section A, demographic profile.

Table 4.1: Brief description of questionnaire structure.

Section	Description/ Variables	Scale Technique
Section A: Demographic profile	Gender	Multiple choice
	Age Group	Multiple choice
	Profession	Multiple choice
	Monthly personal income	Multiple choice
Section B: Transportation characteristics	User of Taxi Service	Multiple choice
	Frequency of using taxi service	Multiple choice
	User of e-hailing service	Multiple choice
	Frequency of using e-hailing service	Multiple choice
	Barriers of using e-hailing service	5-point Likert scale
	Customer preference	Multiple choice
	Installation of application	Multiple choice
	Type of application installed	Multiple choice
	Type of preferred payment system	Multiple choice
Section C: Variables affecting taxi and e-hailing service	Comfort	5-point Likert scale
	Fare	5-point Likert scale
	Reliability	5-point Likert scale
	Safety	5-point Likert scale
Section D: Customer satisfaction	Customer satisfaction on taxi service	5-point Likert scale
	Customer satisfaction on e-hailing service	5-point Likert scale

The four main determinants or factors involved are comfort, fare, reliability and safety. All determinants passed reliability analysis of Cronbach's alpha which are greater or equal to 0.7 (Refer to Table4.2). According to Clare Bradley, Cronbach's alpha normally varies from 0 to 1, and the higher the greater

reliability and internal consistency. However, an alpha coefficient exceeding 0.9 may indicates occurrence of redundancy. Thus, an alpha coefficient of 0.7 or greater should be good enough (Clare Bradley, 2013).

Table 4.2. Result of Reliability Analysis (Liew and Yu, 2016. p.30)

Variables	Cronbach's alpha
Comfort	0.743
Fare	0.869
Reliability	0.855
Safety	0.724

4.2 Descriptive Result

4.2.1 Demographic Profile of Respondents

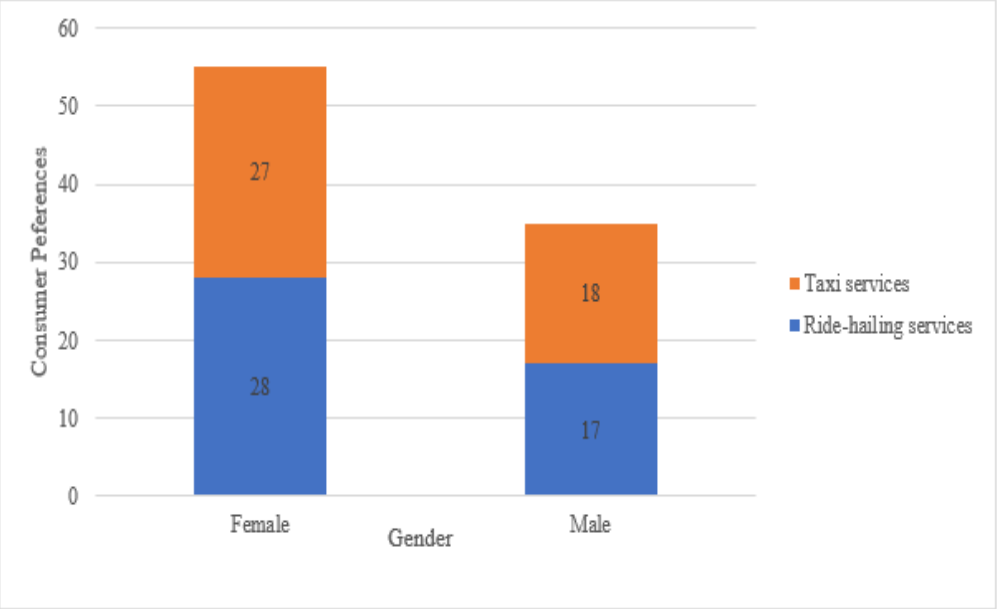


Figure 4.1: Plot of gender by target (Q6).

Figure 4.1 shows the proportion of female and male with their respective preferences on transportation mode. There are 55 females and 35 males which are 61.11% and 38.88% accordingly. Respondent's preferences in each gender is approximately uniformed.

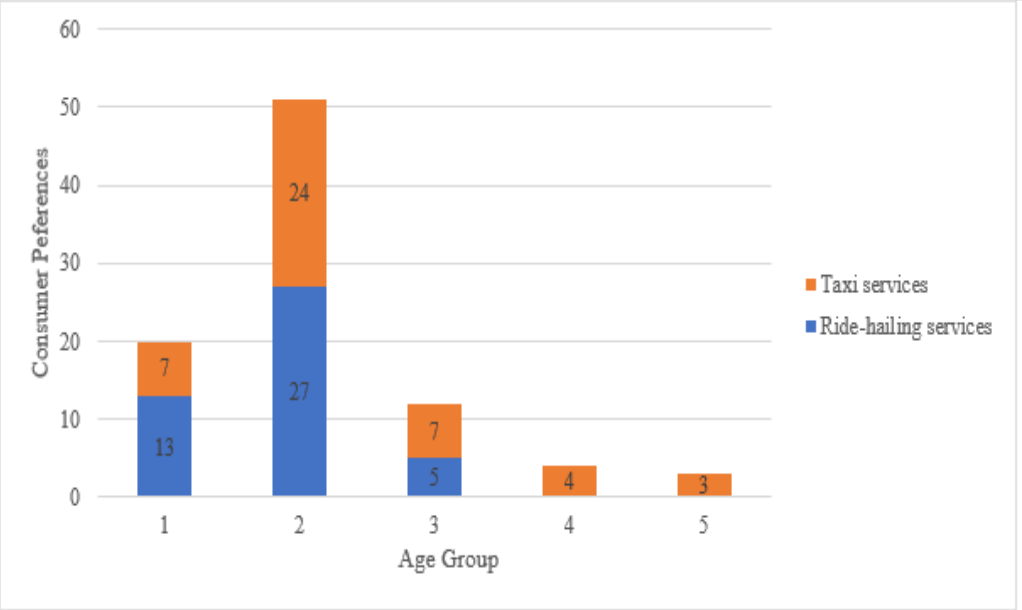


Figure 4.2: Plot of age group by target (Q6).

Age group percentage for group 19 and below, 20 to 29, 30 to 39, 40 to 49 as well as 50 and above are 22.22%, 56.67%, 13.33%, 4.44% and 3.33% respectively (refer to Figure 4.2). Most of the respondents are aged from 20 to 29. Heed that taxi service is dominant in group 4 and 5, which are older adults and senior citizens. While respondent's preference in group 1, 2 and 3 are about uniformed.

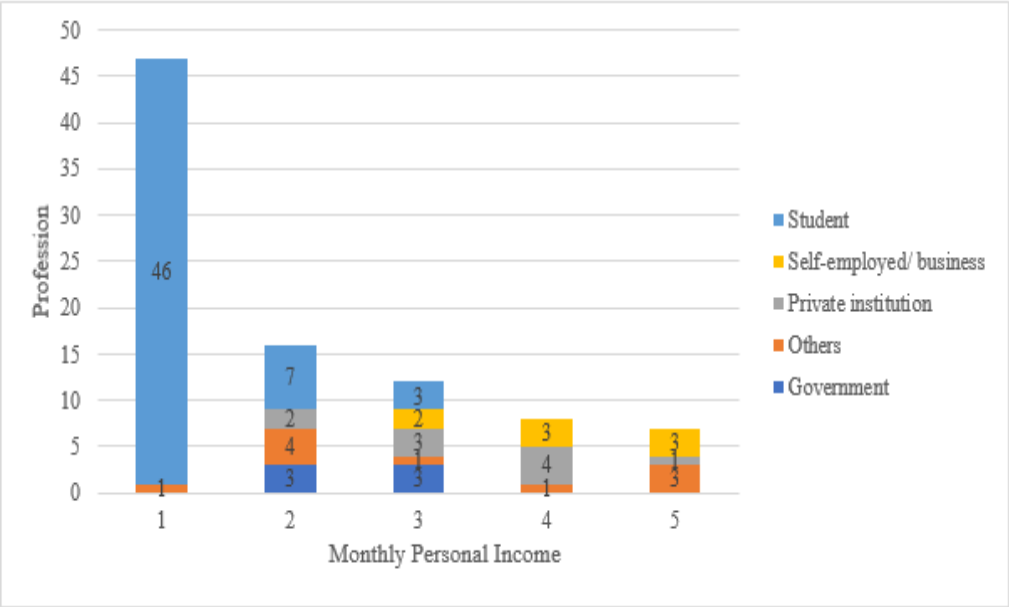


Figure 4.3: Plot of income by professions.

The proportion for student, government sector, private institution, self-employed or business and others are 62.22%, 6.67%, 11.11%, 8.89% and 11.11% respectively. Since more than half of the respondents are students which normally have no or low income, the plot in Figure 4.3 is highly positive skewed. Respondents with higher income are mostly from private institution, self-employed or business or other professions.

4.2.2 Exploration on Inputs Variables

StatExplore and MultiPlot nodes are used as an initial inspecting tools. Illustrated in Figure 4.4.

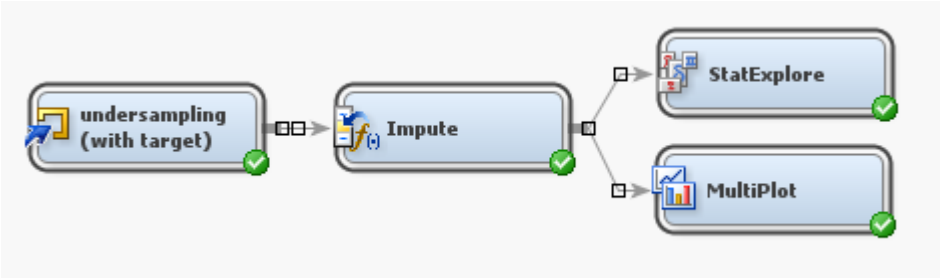


Figure 4.4: Inspection through StatExplore and MultiPlot node.

Exploration on the dependency relationship between the target variable and each input variable based on Chi-Square statistic can be done by using StatExplore node. Note that Chi-Square statistic only anticipate categorical inputs (nominal or ordinal scale), hence, in order to comprehend continuous inputs (interval or ratio scale), one additional step has to be taken before executing the StatExplore node, conversion of continuous input variables to categorical variables. This conversion also known as binning, where continuous input variables are grouped or partitioned into desired number of bin. In this case, number of bin is selected based on the "2 to the *K* rule" (Macfie. and Nufrio, 2006). The rule proposed that, the selection of bins should begin with the smallest number of bin (*K*) such that 2^k is greater than the number of observation. The bins are customized to have equal bin intervals, hence 10 was chosen. The Chi-square statistic of each input variables are calculated using Equation 4.1.

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad 4.1$$

where, O_i is observed frequency and E_i denotes expected frequency. Input variables with a p-value that is less than 0.05 will be considered as important inputs. Thus, the significant input variables are listed in Table 4.3.

Table 4.3: Chi-Square statistic of significant input variables.

Variables	Chi-Square Statistic	*P*-value
Frequent use of e-hailing service	40.5442	<0.0001
Type of application installed	36.4717	<0.0001
Present or absent of e-hailing application	33.4574	<0.0001
Experience on e-hailing service	32.6580	<0.0001
Customer satisfaction on e-hailing: Comfort	28.7782	<0.0001
Customer satisfaction on e-hailing: Fare	24.7694	0.0002
Customer satisfaction on e-hailing: Reliability	26.3846	<0.0001
Customer satisfaction on e-hailing: Safety	21.4212	0.0007
Barrier: Availability of technology	14.0077	0.0073
Payment system- ease of payment	12.7015	0.0128
Customer satisfaction on taxi service: Reliability	11.7557	0.0383
Preferable payment system	8.6631	0.0032

Besides, StatExplore node also provides the worth of each input variable. Likewise, another method to address dependency between interval input variables and target is through scaled mean deviation (SMD) plot. The scaled mean deviation corresponds to Chi-Square statistic in term of calculation and interpretations as follow:

SMD(x,0) = [mean (x, where target=0)-mean(x)]/mean(x)] 4.2

SMD(x,1) = [mean (x, where target=1)-mean(x)]/mean(x)] 4.3

Where x represents a particular input variable. A SMD plot for this study is shown in Figure 4.6.

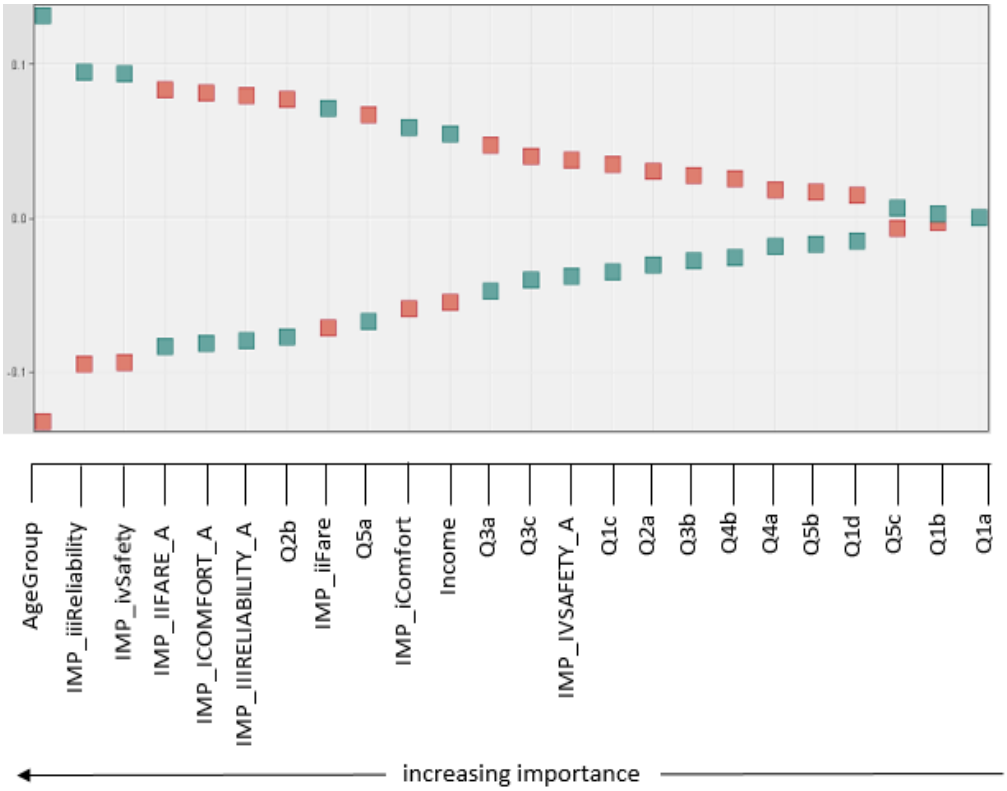


Figure 4.5: Scaled mean deviation (SMD) plot.

Based on Figure 4.5, the vertical axis represents interval input variables and each variable has two points (green and red) that denote its scaled mean deviation. The green point symbolizes SMD of a variable given that target is taxi whereas red is the SMD of the same variable when target is e-hailing. In other words, the green and red points are plotted through Equation 4.2 and Equation 4.3 (Ricardo G, n.d.). The larger the gap between the green and red points for a variable, the higher the residual and hence yield an important or dependent input variable.

Later, the worth for each variable is computed from the p-value (p) corresponding to the identified Chi-Square statistic through the formula:

$$worth\ of\ input = -2\log(p) \qquad 4.4$$

The worth indicates weight or strength between each input variables and target (Q6) and the result is shown in the variables worth plot (Figure 4.6).

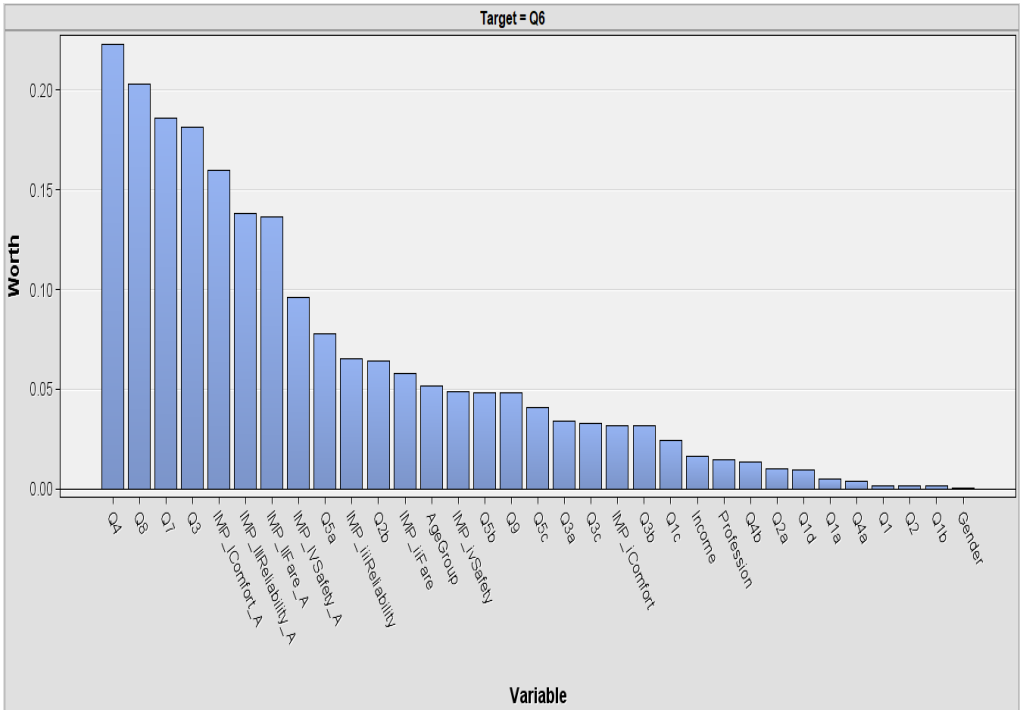


Figure 4.6: The variables worth plot.

Variable Q4 which hold the highest Chi-Square value and worth implies that it is the most important input, followed by Q8, Q7 and so on (refer to Figure 4.6). Ultimately, Chi-Square statistic, SMD and variable worth suggested similar significant inputs. And consequently, investigation over these important variables are mandatory. MultiPlot node is used to visualise these important inputs identified by StatExplore node.

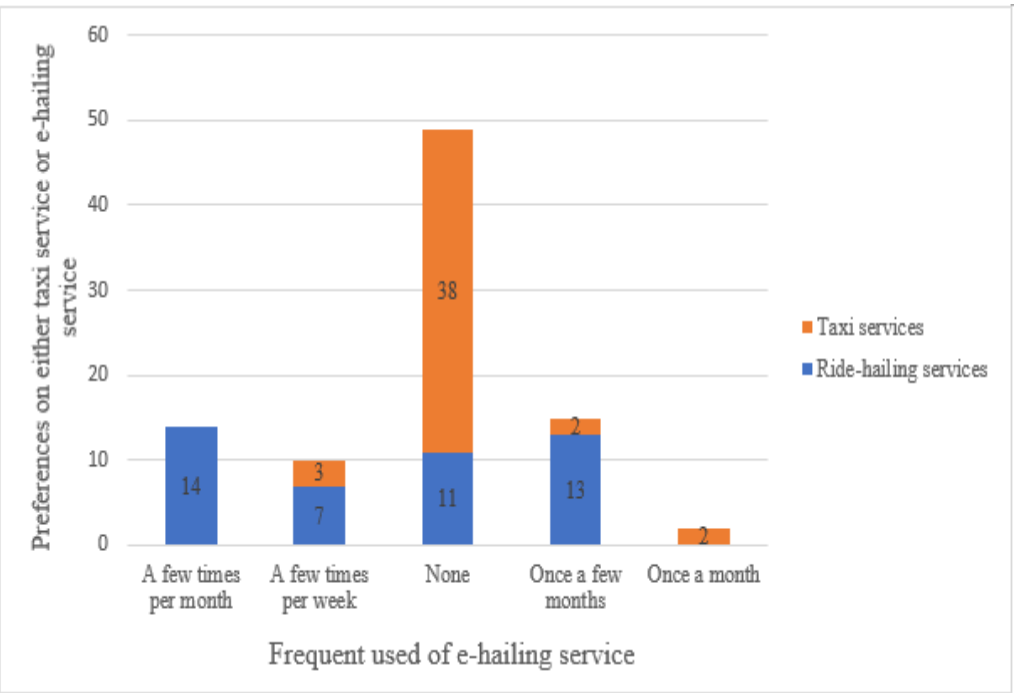


Figure 4.7: Plot of frequent used of e-hailing service (Q4) by target (Q6).

Table 4.4: Mode of the frequent used of e-hailing service (Q4) for each target class (Q6).

Target Level	Mode	Mode Percentage	Second Mode	Second Mode Percentage
E-hailing Service	A few times per month	31.11	Once a few months	28.89
Taxi Service	None	84.44	A few times per week	6.67
Overall	None	54.44	Once a few months	16.67

According to Figure 4.7, more than half of the respondents (54.44%) have never tried of using e-hailing service and out of this 54.44%, only 22.45% of them could choose e-hailing over taxi service as a form of transportation. In other words, 77.55% of them choose not to venture in to e-hailing service although they have not experience it. The second highest frequency falls on "once a few months", which is about 16.67% of overall response. From these two scenarios, it appears that respondents rarely uses e-hailing service. Also, in Table 4.4, the first and second modal for e-hailing service are "a few times per month" and "once a few months", which indicated that even experienced e-hailing users utilize e-hailing service on occasional basic.

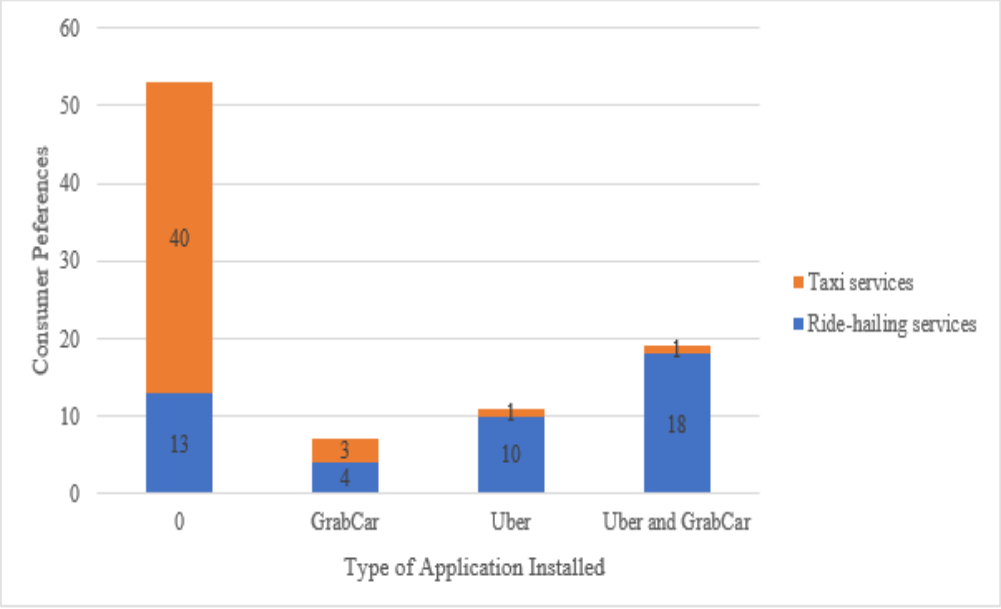


Figure 4.8: Plot of type of application installed (Q8) by target (Q6).

Referring to Figure 4.8, the mode (58.89%) happens to be 0, which coded as "did not install any e-hailing related application". Within the context of not having any e-hailing related application, there is still 24.53% of respondents who

preferred e-hailing service. The reasons for this phenomenon might due to incompatible technology, peer influences and more. However, it is not within my scope of studies and more research has to be done to validate these reasons. Whereas, in circumstances where respondents have e-hailing related application installed (either GrabCar, Uber or both), their preference is toward e-hailing service which represent in blue.

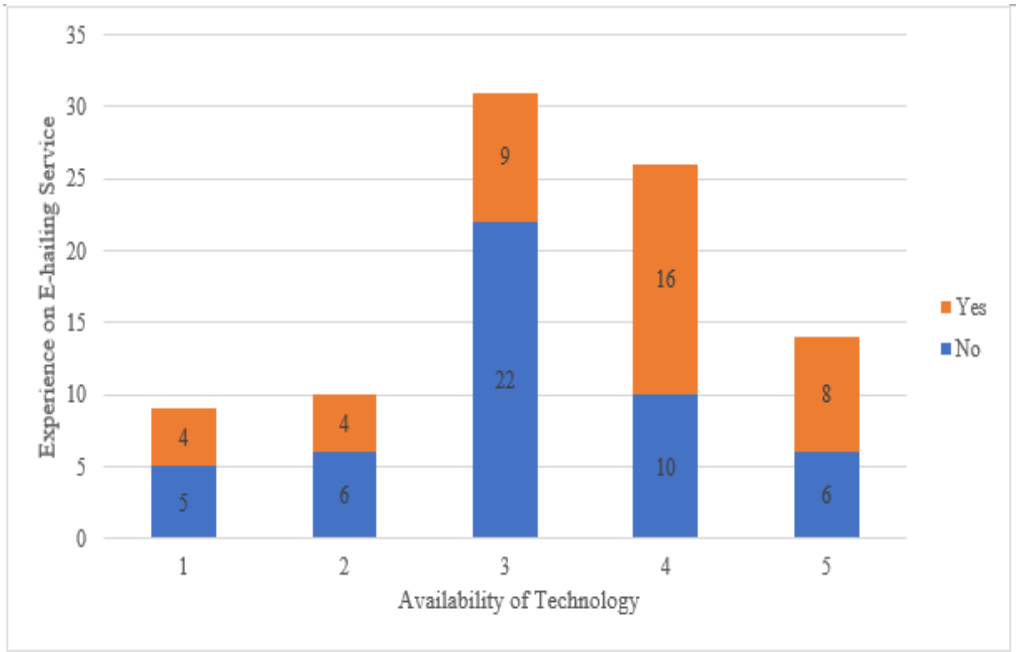


Figure 4.9: Plot of barrier-availability of technology (Q5a) by experience on e-hailing service (Q3).

Most of the respondents take either neutral stand or consider availability of technology as a relevant barrier of using e-hailing service since from Figure 4.9, the plot skewed to left. About 34.44% of respondents have taken the neither relevant nor irrelevant position. Out of this 34.44% only 29.03% respondents have an experience on e-hailing service (red portion), this figure indicates that most respondents choose a neutral stand due to the lack of experience and

understanding on e-hailing service. Respondents who never travel through e-hailing service might not know how it works and the role of technology in e-hailing service.

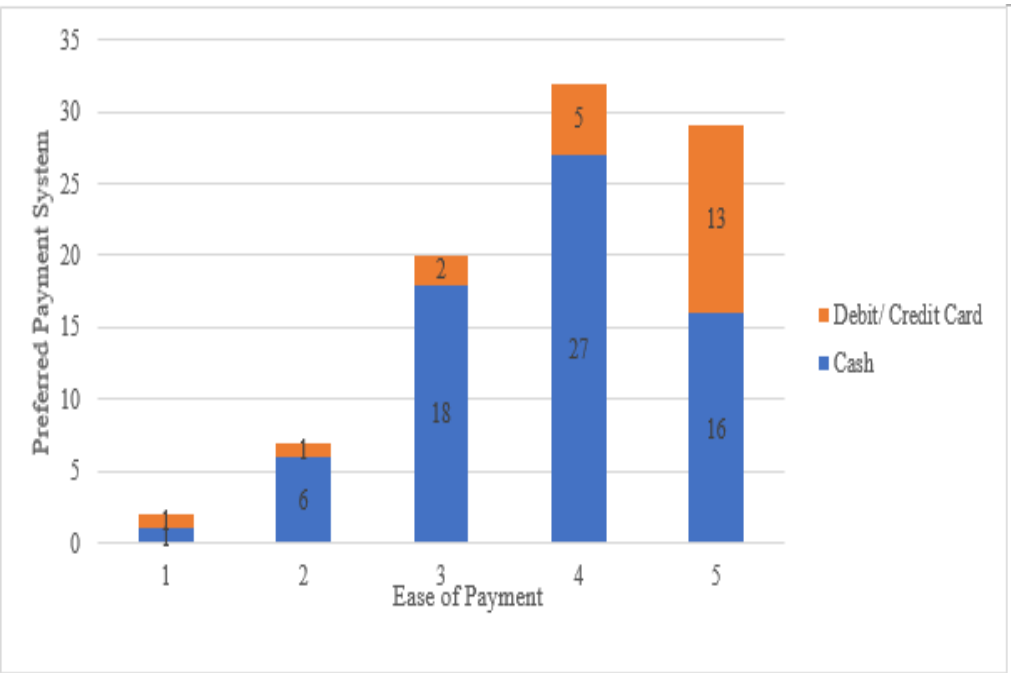


Figure 4.10: Plot of ease of purchasing (Q2b) by preferred payment system (Q9).

Observing Figure 4.10, it clearly shows a deviate toward “most important” and the average rate for importance of ease of purchasing is 3.8778 over 5. Also, notice that from point 2 to 5, blue portion is dominating. This scenario explains that cash payment system is preferable regardless of the rate on importance of ease of purchasing. Additionally, 90% out of the 22.22% of neutral stand respondents do favour in chase payment system. Overall, ease of payment is one of the significant input variable identified and to customers, cash payment is a form of easy payment system.

Figure 4.11: Average rate of customer satisfaction on taxi service and e-hailing service.

Figure 4.11 uncovers that respondents or customers' satisfaction on e-hailing service outranges taxi service on all aspects (safety, reliability, fare and comfort), despite there is 54.44% (refer to Table 4.4) of respondents who have never travel through e-hailing service. This provides a positive signal that e-hailing industry has much spaces for development since it makes good impressions.

4.2.3 Clustering and Segment Profiling

Three segments are generated through the Clustering node and examination on these three segments are done through Segment Profile node, refer to Figure 4.12 for illustration.

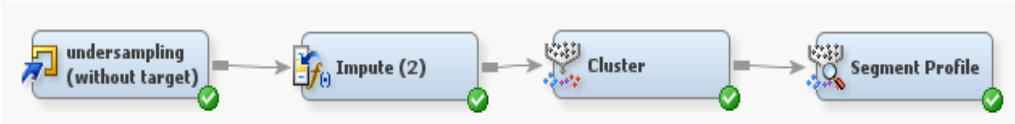


Figure 4.12: Clustering analysis through Cluster node and Segment Profile node.

The prima objective of clustering analysis in marketing context is to produce meaningful clusters that would provide insights. Therefore, labelling segments based on statistic or output generated by Clustering node is essential in determining the meaning of these segments. In this case, target variable (Q6) is used as a benchmark to differentiate segments generated.

Table 4.5: Target (Q6) proportion in each segment.

Segment ID	Label	Target= Taxi Service	Target= E-haling Service
1	Taxi	0.783784	0.216216
2	E-hailing	0.037037	0.962963
3	Without preferences	0.576923	0.423077

From Table 4.5, notice that 78.38% of respondents from segment 1 prefer taxi service over e-hailing service, therefore it is named as "Taxi". Contrary in segment 2, there is 96.30% of respondents favour e-hailing service which then labelled as "E-hailing". Whereas for segment 3, the portions for target are about the same and hence it is labelled as "Without preferences".

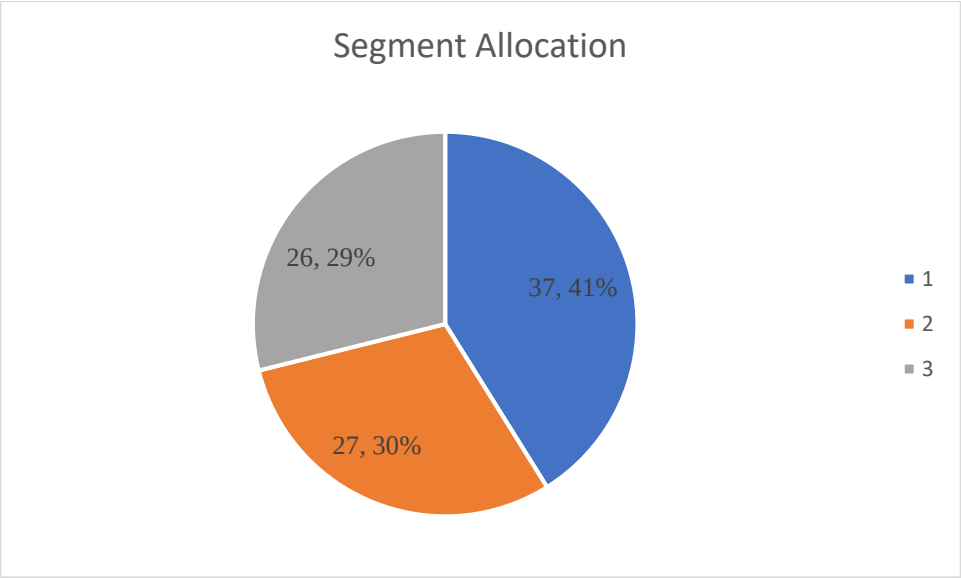


Figure 4.13: Proportion of three segments generated by Clustering node.

Also, referring to Figure 4.13, the segment's count 37, 27 and 26 for segment 1,2 and 3 accordingly are almost balanced or it can say to be closed to uniform, which is what a marketing manager would normally preferred (Collica, 2011). This supported the segments generated to be valid and meaningful.

Next, results from Segment Profile node will be discussed. It unwraps top ten most important factors or variables for each segment and the result is displayed in Table 4.6. Variable worth plot for each segment is posed in Appendix C.

Table 4.6: Top ten most significant variables for each segment.

Segment ID	Segment Preferences	Significant Variables
1	Taxi	Frequent use of e-hailing service, Experience on e-hailing service, Type of application installed , Present or absent of e-hailing application, Age Group, Customer preference, Waiting time, Customer satisfaction on e-hailing service based on comfort, fare and reliability
2	E-hailing	Frequent use of e-hailing service, Type of application installed, Experience on e-hailing service, Customer preference , Present or absent of e-hailing application, Customer satisfaction on e-hailing service based on comfort, fare and reliability, Air conditioning-temperature and ventilation
3	Without preferences	Air conditioning- temperature and ventilation, Punctuality, Waiting time, Driver's behaviour and attitude, Comfortable seats, Safety assurance, Cleanliness and hygiene of vehicle, Frequency of service, Trip information, Value for money of the fare

It was found that all of the important variables for segment 1 and 2 are parallel to the significant variables identified through Chi-Square statistic, except for waiting time and air conditioning- temperature and ventilation. These exceptional case (waiting time and air conditioning) might be the attributes that differentiate segment 1 and segment 2. Significant variables in segment 3 arise with least consistent with pervious findings (based on Chi-Square statistic and Scaled Mean Deviation) and segment 1 as well as segment 2. A profound analysis and discussion is held next, to explain these significant variables regarding its segment.

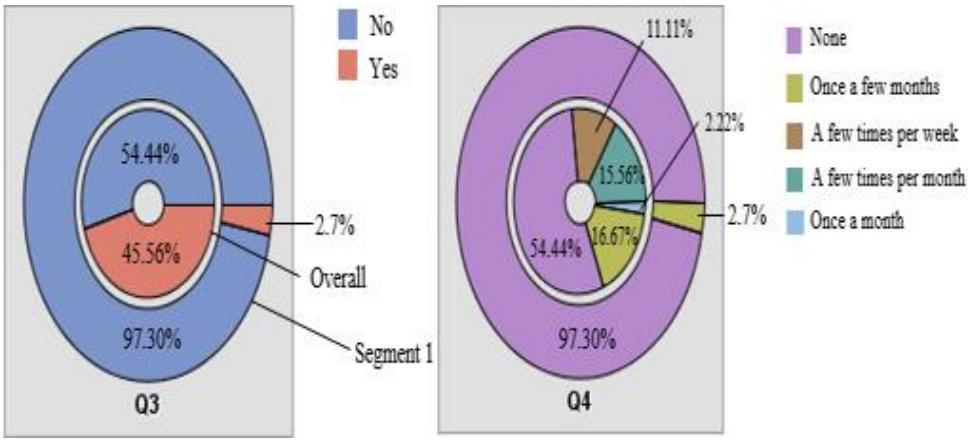


Figure 4.14:Charts of experience on e-hailing service (Q3) and frequent use of e-hailing service (Q4) for segment 1 (Taxi) versus overall.

Note that the inner pie charts represent the overall position of respective variables whereas the outer pie charts are the distribution of the segment. In Figure 4.14, variable experience on e-hailing service (Q3) shows a highly bias trend towards no experience on e-hailing service (97.30%) in segment 1, conversely, the overall position is rather fair. While merely 2.70% of respondents who experienced e-hailing service, disclosed that they utilized e-

hailing service once a few months. It implies that respondents from Taxi segment barely traval through e-hailing service.

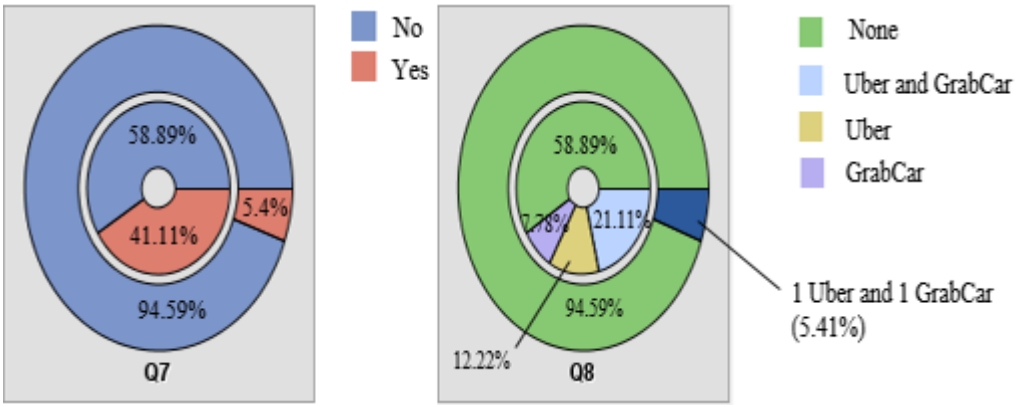


Figure 4.15: Charts of present or absent of e-hailing related application (Q7) and type of e-hailing related application installed (Q8) for segment 1 (Taxi) versus overall.

About 94.59% or 35 respondents in this segment 1 did not install any e-hailing related apps (addressing variable Q7). The remaining two respondents each installed Uber and GrabCar (refer to Figure 4.15).

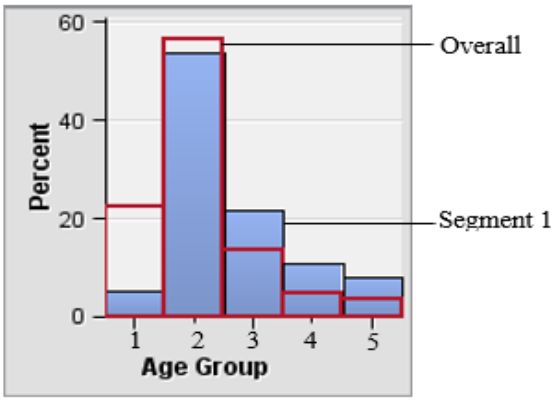


Figure 4.16: Charts of age group for segment 1 (Taxi) versus overall.

Note that red bar charts represent the overall position of respective variables whereas the blue bar charts are the distribution of the segment. The age group of respondents in Taxi segment appears to be much older since the blue bars for AgeGroup variable is below the overall position on left while exceed the overall on the right side. This also agreed with findings in section 4.2.1, where greater age groups are dominated by taxi preference.

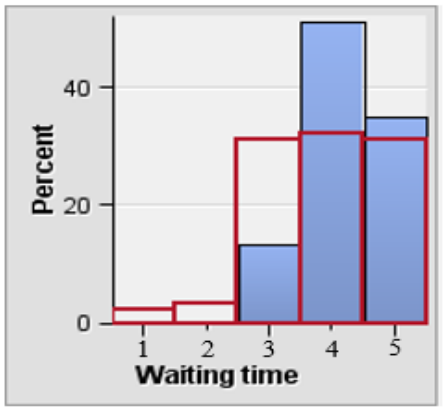


Figure 4.17: Charts of waiting time (Q3c) for segment 1 (Taxi) versus overall.

From Figure 4.17, observed that blue bars are much denser on the right end compare to overall (red bars). This signals that segment 1 emphasizes on waiting time to be served.

Table 4.7: Frequency of customer satisfaction on e-hailing service for segment 1(Taxi) and overall.

Customer satisfaction criteria	Segment 1 (Taxi)	Overall
Comfort		
1 (Poor)	0	0
2	3	5
3	15	25
3.76 (imputed)	8	11
4	9	33
5 (Excellent)	2	16
Fare		
1 (Poor)	0	1
2	3	4
3	14	25
3.84 (imputed)	8	11
4	9	26
5 (Excellent)	3	23
Reliability		
1 (Poor)	0	0
2	2	3
3	14	26
3.85 (imputed)	8	11
4	11	30
5 (Excellent)	2	20

Although variable customer satisfaction on e-hailing service (comfort, fare and realibility) is a significant input for segment 1 (Taxi), but respondents in segment 1 tends to be neutral toward these varaible since the mode for these three variables in segment 1 (yellow highlighted) are all 3, which represent a neutral stand in Table 4.7. Whereas in overall dataset, customer statisfaction on e-hailing service trends to be appraised with a higher rate (green highlighted). Also, refer to Appendix D for a better illustrastion on this situation.

To sum up the attributes of Taxi segment, it is a segment which mostly have not experience e-hailing service and might be caused by the weak technology adaptation among eldery. Also, this segment stresses on waiting time to be served.

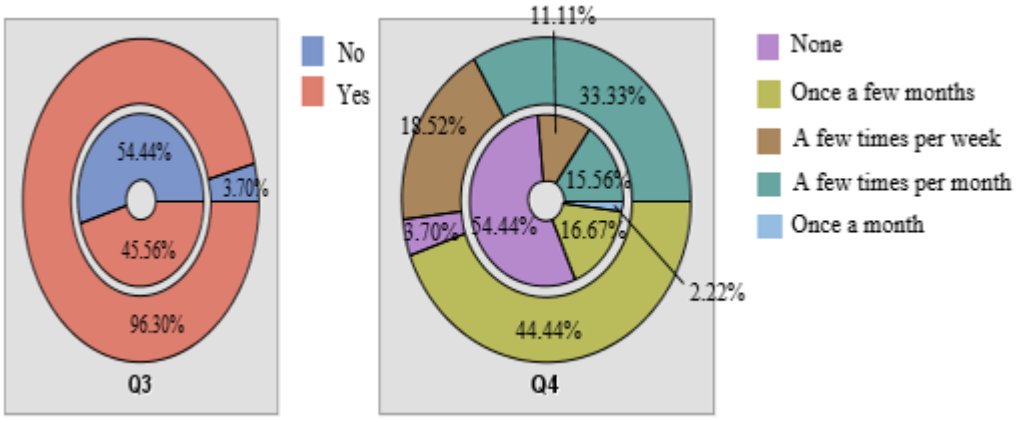


Figure 4.18: Charts of experience on e-hailing service (Q3) and frequent use of e-hailing service (Q4) for segment 2 (E-hailing) versus overall.

Referring to Figure 4.18, variable frequent use of e-hailing service (Q4), the usage of e-hailing service is definitely much frequency than overall since the "none" portion (indicated in purple) is much smaller in the segment with merely 3.73%. Additionally, this stament also supported by variable experience on e-hailing service (Q3), where it illustrate the 3.73% in blue portion. The other three shares involved are "a few times per week" (brown), "a few times per month" (dark green) and "once a few months" (light green) with portion of 18.52%, 33.33% and 44.44% accordingly.

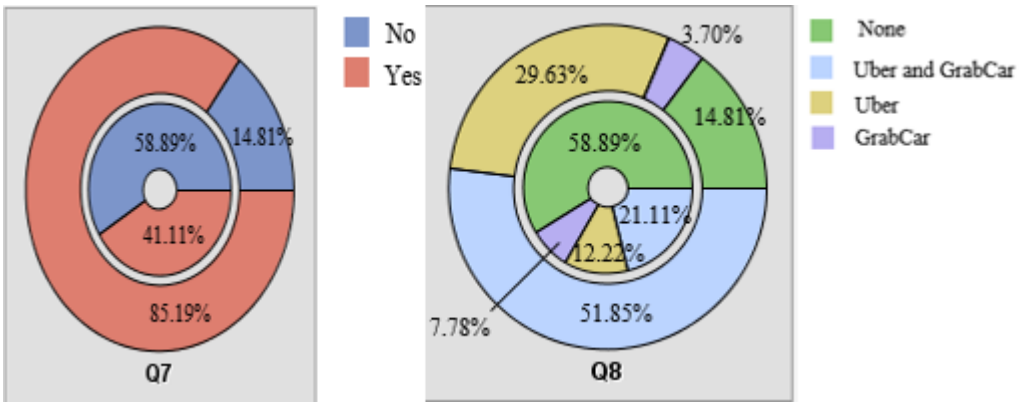


Figure 4.19: Charts of present or absent of e-hailing related application (Q7) and type of e-hailing related application installed (Q8) for segment 2 (E-hailing) versus overall.

As anticipated, for variable type of application installed (Q8), respondents in E-hailing segment are more likely to have both Uber and GrabCar applictions installed (51.85%). Also, the green portion for "no e-hailing related application installed" is dropped from 58.89% (overall) to 14.81% (E-hailing segment). The other allocation for Uber and GrabCar application installed are 29.63% and 3.70% respectively.

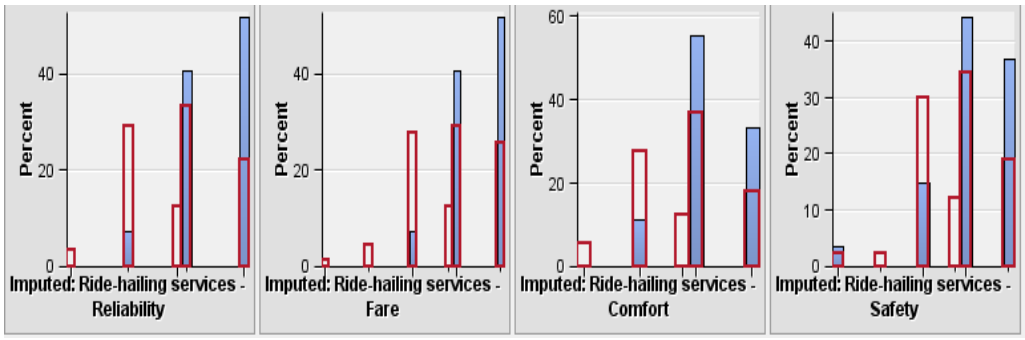


Figure 4.20: Charts of customer satisfaction on e-hailing service based on reliability, fare, comfort and safety for segment 2 (E-hailing) versus overall.

From Figure 4.20, notice that in all aspects (reliability, fare, comfort and safety), the blue bar charts tend to skewed to the left which reveals that respondents from this segment are very statisfied with the service provided by e-hailing.

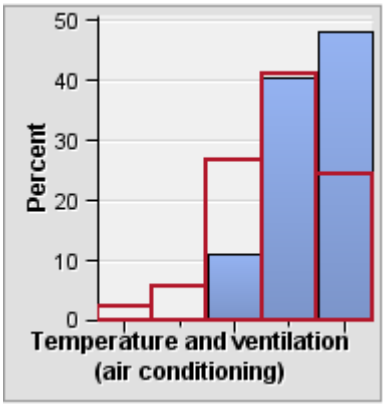


Figure 4.21: Chart of air conditioning- temperature and ventilation for segment 2 (E-hailing) versus overall.

Lastly, it was found that the chart shows in Figure 4.21 are densed on the right end which leads to the conclusion that respondents from this segment emphazised air conditioning since the right end is associated with greater importance.

To wrap up this segment, it is mostly represented by respondents who utilize e-hailing service in a frequen manner and the bridge or intermedia used to connect respondents with e-hailing service, Uber or GrabCar applications are presumably installed. Extra attention has to be pay on air conditioning in order to retain or surpass the current good performance.

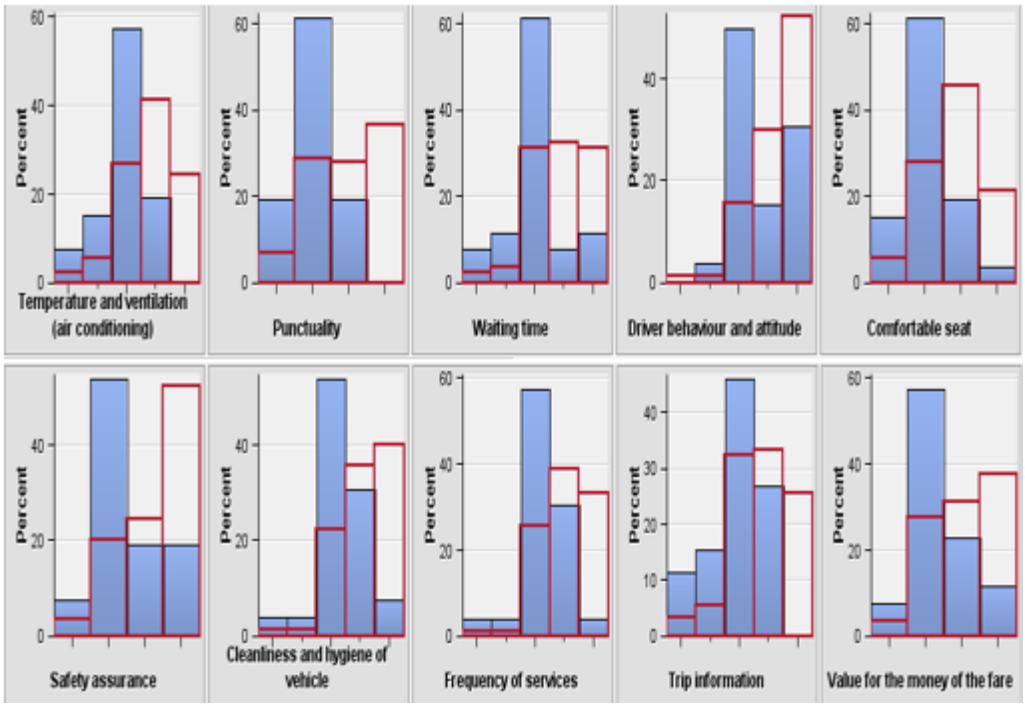


Figure 4.22: Chart of significant varibles for segment 3 (Without Preference) versus overall.

Perceive in Figure 4.22, the mode value for all the important variables are "3" which represents a neutral stand and they are far beyond the overall position. This could be a bright spot for most marketers. According to Ryals (2009), customers or segments with neutral stand should be consider a top priority for sales and marketing because there is high potential by winning business from competitors and generates a positive payoff from the marketing efforts made. Therefore, either party (Taxi or E-hailing industry) should be industrious and endeavour to draw customers to its side by scheming marketing strategic based on the significant variables proposed.

4.3 Predictive Results

Predictive analysis is performed through two different process, single-step process and multiple-step process. In the single-step process, data mining techniques like clustering, classification and visualization are used to generate output independently. Whereas, multiple-step process is a composition of clustering, classification and visualization. Comparison among the performance of the three models generated (Decision Tree, Logistic Regression and Neural Network) is first examined within each process. Later, performance evaluation between process is done.

4.3.1 Single-step Predictive Analysis

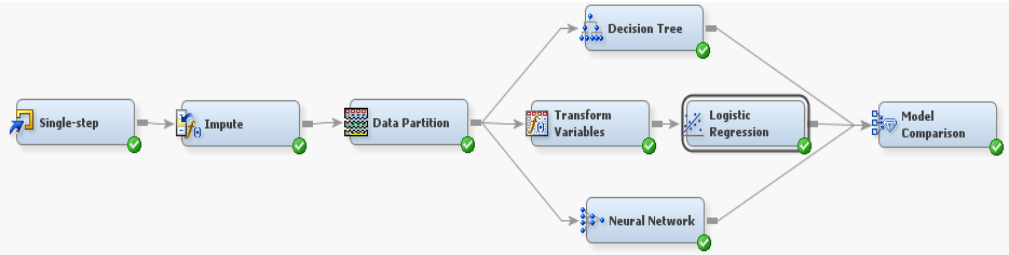


Figure 4.23: Framework of single-step predictive analysis.

Only results will be discussed in this and the following section since the framework is already revealed in Chapter 3. Decision Tree node, Regression node and Neural Network node are used to create models. The model generated will then passed to the succeeding node, Model Comparison node.

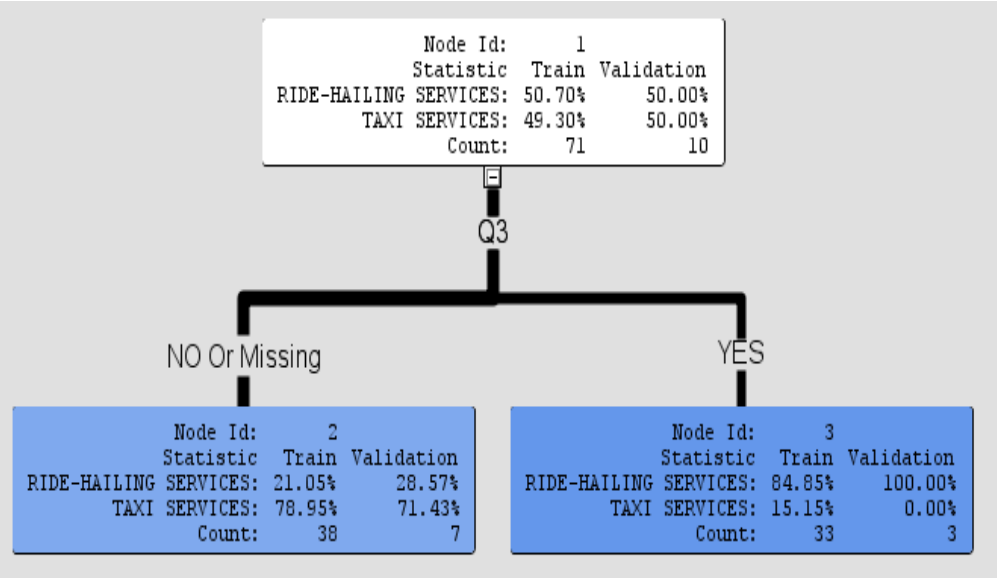


Figure 4.24: Output from Decision Tree node.

In Figure 4.24, the root node (top) represents the original data set with its respective train and validation portion. Variable experience on e-hailing service (Q3) is the only significant variable to the tree. Again, the selection of important variables are bosed on Chi-Square statistic and log worth. Based on the validation set, about 28.57% of respondents who have not experience e-hailing service would perfer e-hailing over taxi serive. On the other hand, when respondents have tried traveling via e-hailing service, they will definitely prefer e-hailing over taxi. These remarkable figures have encouraged e-hailing industry to scheme its strategic plan based on variable experience on e-hailing service. In other words, e-hailing industry should pay extra attention on new or potential users as to increase and retain subcribers.

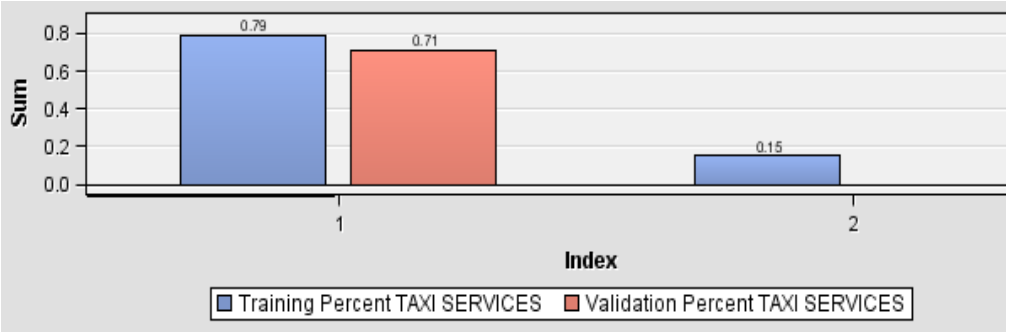


Figure 4.25: Output from Decision Tree node- Leaf statistic.

Based on Figure 4.25, the *x*-axis of the plot denotes number of leafs while the *y*-axis is the predicted value assigned to the node. In this case, only two leafs are produced and it is predicted that about 78.95% (training data set) as well as 71.43% (validation data set) of respondents would prefer taxi service given that they have never experience e-hailing service (first leaf). Expecting 15.15% (training data set) and none (validation data set) would choose taxi service over e-hailing service if then have ever tried e-hailing. The differences between the training set and validation is presumed to be small, the smaller the better. However, there is not benchmark or standard definition of how small is sufficient to be good. Thus, further exploration on the tree is carried out.

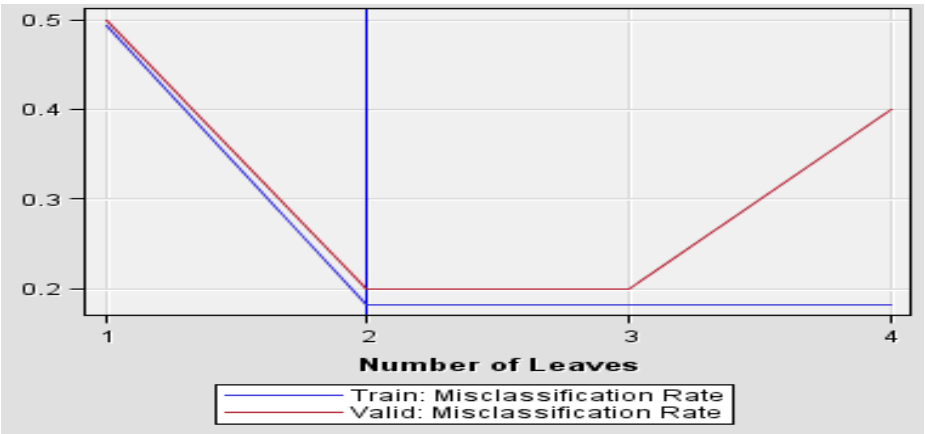


Figure 4.26: Output from Decision Tree node- Misclassification Rate.

On the horizontal axis, it shows the number of leaves while the misclassification rate on the vertical axis. Referring to Figure 4.26, observed that number of leaves is up to four while in pervious findings there is merely two leaves as shown in Figure 4.24. This phenomenon is due to the maximal decision tree or the full tree generated includes four leaves instead of two. In spite of that, misclassification rate for both training and validation data are optimal or lowest when number of leaves appeared to be two and three. Two leaves are then selected for generating the tree as simplicity concerned.

Next, model generated from Logistic Regression node is examined. Variables selected (significant variables) in the final logistic regression model and its respective estimated coefficients are listed in Table 4.8.

Table 4.8: Coefficients for each selected variable.

Label	Parameter/ Variable	Estimated Coefficients
1	Intercept	-2.1325
2	Q7	1.9384
3	Q3	1.4501
4	Q1	-1.4428
5	Q5c	0.1512
6	Q5a	-0.0516

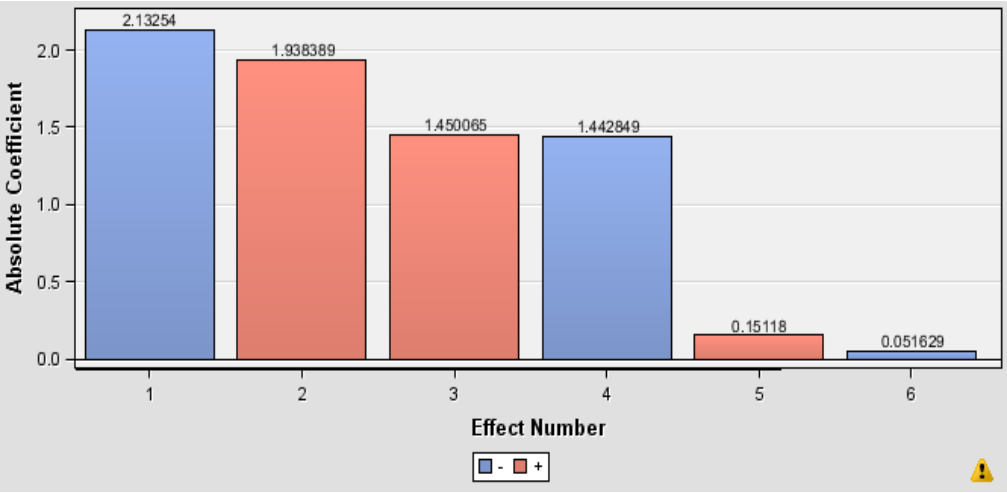


Figure 4.27: Output from Logistic Regression node- Effect Plot.

Noticed that variables barrier- availability of technology as barrier (Q5a) and experience on taxi service (Q1) have a negative relationship with target, respondent's preferences (Q6).

The following model to be discussed is neural network. Note that neural network does not provide any significant input variables but rather it takes in all input variables to its algorithm and generates numerous outputs.

Parameter Estimates	143		
	Optimization Start		
Active Constraints	0	Objective Function	0.0005706524
Max Abs Gradient Element	0.0003481969		
	Optimization Results		
Iterations	0	Function Calls	4
Gradient Calls	2	Active Constraints	0
Objective Function	0.0005706524	Max Abs Gradient Element	0.0003481969
Slope of Search Direction	-0.000075144		

Convergence criterion (ABSCONV=0.0013864928) satisfied.

Figure 4.28: Output from Neural Network node.

Figure 4.28 shows that 143 parameters are used in the neural network for estimation process and the convergence criterion is satisfied or equivalent to the neural network is feasible. However, observed that the number of iteration in this case is zero signals that the network is a redundant model.

Table 4.9: Result from neural network: Classification table for training data set.

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
E-hailing	E-hailing	50.7042	100	36	50.7042
Taxi	E-hailing	49.2958	100	35	49.2958

Table 4.10: Result from neural network: Classification table for validation set.

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
E-hailing	E-hailing	50	100	100	50
Taxi	E-hailing	50	100	100	50

Considering the misclassification rate based on Table 4.9 and Table 4.10, the rate is about 49.30% for training data set and 50.00% for validation data set. This figure indicates that the neural network is least useful. This condition may due to the convergence criteria met without any iteration.

After all, a Model Comparison node is used to identify the best model generated among the three methods, decision tree, logistic regression and neural network. The result is displayed in Table 4.11

Table 4.11: Misclassification rate for each model.

Model	Selection Criterion: Misclassification Rate
Logistic Regression	0.070423
Decision Tree	0.183099
Neural Network	0.492958

Logistic regression (about 7.04%) surpassed the other two models as it has the lowest misclassification rate. Followed by decision tree with misclassification rate of 18.10% and lastly neural network (49.30%).

4.3.2 Multiple-steps Predictive Analysis

In order to perform multiple-steps predictive analysis, a preceding step has to be taken, clustering. Thus, outputs from Cluster node is exported through Save Data node and used in this section (refer to Figure 4.29 and Figure 4.30).

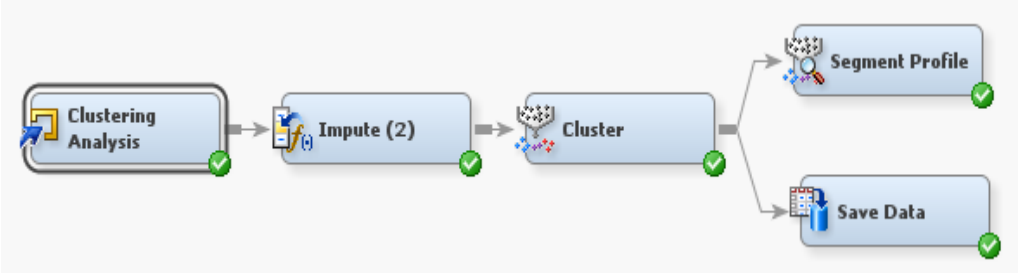


Figure 4.29: Exporting outputs from Cluster node.

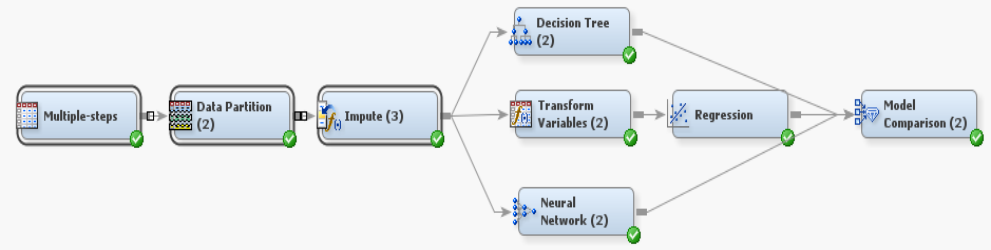


Figure 4.30: Framework of multiple-steps predictive analysis.

Figure 4.29 is similar to Figure 4.30, while the only different is that the source node (first node) in Figure 4.32 is and import data from Cluster node.

Behold that the output of decision tree and logistic regression via multiple-steps predictive analysis is exactly identical as the decision tree and logistic regression produced through single-step predictive analysis. Therefore, no discussion on decision tree and logistic regression in this section. Whereas, outcome from neural network is surprisingly outstanding.

Parameter Estimates 139

Optimization Start

Active Constraints 0 Objective Function 0.0803014902
Max Abs Gradient Element 0.0073004248

Iter	Restarts	Function Calls	Active Constraints	Objective Function	Objective Function Change	Max Abs Gradient Element	Step Size	Slope of Search Direction
1	0	12	0	0.05072	0.0296	0.0113	0.699	-0.0540
2	0	16	0	0.04916	0.00155	0.0122	0.00669	-0.418
3	0	21	0	0.04573	0.00344	0.0122	0.0400	-0.141
4	0	24	0	0.04428	0.00145	0.00923	0.0100	-0.214
5	0	29	0	0.04075	0.00353	0.0390	0.0316	-0.193
6	0	32	0	0.03177	0.00898	0.0363	0.0100	-1.190
7	0	36	0	0.02528	0.00649	0.0171	0.00364	-2.998
8	0	40	0	0.00817	0.0171	0.0550	0.246	-0.146
9	0	43	0	0.00499	0.00317	0.00752	0.0107	-0.626
10	0	48	0	0.00270	0.00230	0.00498	0.729	-0.0052
11	0	50	0	0.00109	0.00160	0.00455	0.337	-0.0171

Optimization Results

Iterations 11 Function Calls 52
Gradient Calls 20 Active Constraints 0
Objective Function 0.0010919908 Max Abs Gradient Element 0.0045519194
Slope of Search Direction -0.017061488

Convergence criterion (ABSCONV=0.0013864928) satisfied.

Figure 4.31: Output from Neural Network node.

This time, only 139 parameters are used for estimation process and the convergence criterion is also satisfied. Moreover, the estimation process required 11 iterations.

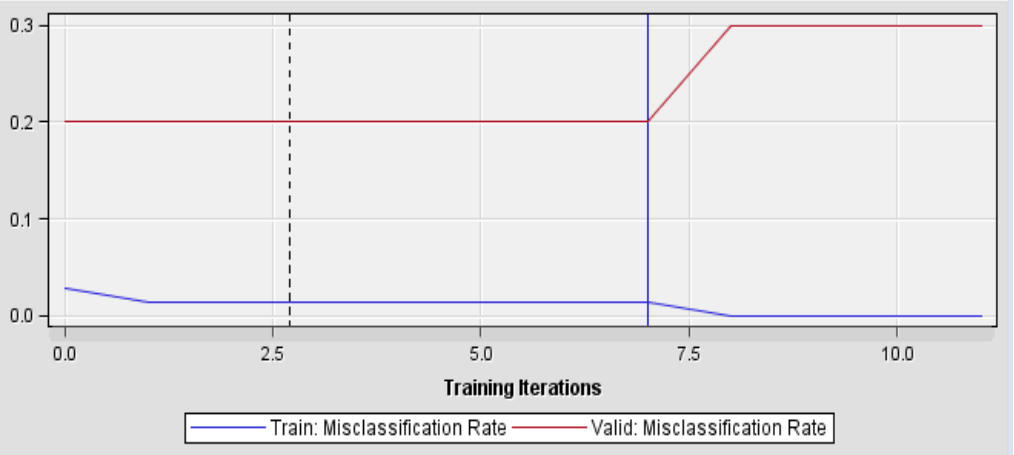


Figure 4.32: Output from Neural Network node- Iteration Plot.

Figure 4.32 shows the number of required iteration requires to optimize both training and validation set based on misclassification rate. Notice that training data set maintained a constantly low misclassification rate while misclassification rate for validation data set spike after iteration 11, hence, the optimum iteration is 11.

Then, utilise the Model Comparison node to recognize the better performing model. Result is shown in Table 4.12.

Table 4.12: Misclassification rate for each model.

Model	Selection Criterion: Misclassification Rate
Logistic Regression	0.070423
Decision Tree	0.183099
Neural Network	0.014085

The foremost model identified through multiple-steps predictive analysis is neural network with merely 1.48%, followed by logistic regression (7.04%) and decision tree (18.3%). It undoubtedly indicates that performance of artificial neural network can be improved via multiple-steps predictive analysis. To be precise, ANN performance is refined by clustering analysis which grouped and reduced input set to be learned by neural network. The computation intensity of the neural network is then reduced and eventually improves its performance (Shrivastava and Chaudhari, 2011). This statement also supported by the number of parameter estimated as the number of parameter estimated in multiple-steps predictive analysis is lesser (refer to Figure 4.28 and Figure 4.31).

CHAPTER 5

CONCLUSION

5.0 Introduction

The last section of this research poses summary of findings and discoveries. It also provides limitations and improvements that could enhance the quality of this study as well as recommendations for future related work.

5.1 Summary of Findings and Discoveries

As a prima objective of this research, identifying significant input variables that poses high possibility in effecting the e-hailing industry with its respective method is summarized in Table 5.1.

Table 5.1 Summary of significant input variables identified.

Method	Significant Variables
Chi-Square Statistic	frequent use of e-hailing service, type of application installed, present or absent of e-hailing application, experience on e-hailing service, customer satisfaction on e-hailing service hinged on comfort, fare, reliability and safety, availability of technology as barrier, payment system- ease of payment, customer satisfaction on taxi service based on reliability and preferable payment system
Scaled Mean Deviation(SMD)-covers only interval inputs	age group, customer satisfaction on taxi service hinged on comfort, fare, reliability and safety, customer satisfaction on e-hailing service hinged on comfort, fare and reliability, payment system- ease of payment, availability of technology as barrier, punctuality
Log Worth	frequent use of e-hailing service, type of e-hailing related application installed, present or absent of e-hailing application, experience on e-hailing service, customer satisfaction on e-hailing service hinged on comfort, fare and reliability
Decision Tree	experience on e-hailing service
Logistic Regression	availability of technology as barrier, experience on taxi service, experience on e-hailing service, present or absent of e-hailing application, user friendly as barrier

Although variables identified tends to varied from methods, however they did provide some common variables like experience one-hailing service, present or absent of e-hailing apps, availability of technology, customer satisfaction on e-hailing service hinged on comfort, fare, reliability, safety and so on. These common significant inputs identified through different methods pose a high

credibility in the statistic perceptive as it implies that the importance of these variables are regardless of method or algorithm used and signals an internal consistency. Hence, these variables should take priority over others during the development of strategic and smart marketing scheme.

Besides, to cater the second objective in this research, segmenting and profiling user, three segments are identified namely, Taxi segment, E-hailing segment and Without Preferences segment through clustering analysis. Taxi segment and E-hailing segment can be sort as segments with preferences and the summary of result is shown in Table 5.2 and Table5.3.

Table 5.2: Frequency (%) of Taxi segment and E-hailing segment (Categorical case).

	Taxi Segment	E-hailing Segment
Experience on e-hailing service		
Yes	2.70	96.30
No	97.30	3.70
Frequent use of e-hailing service		
None	97.30	3.70
A few times per week	0.00	18.52
A few times per month	0.00	33.33
Once a month	0.00	44.44
Once a few months	2.70	0.00
Type of e-hailing related application installed	94.59	14.81
None	0.00	29.63
Uber	2.70	3.70
GrabCar	2.70	51.85
Uber and GrabCar		
Present or absent of e-hailing related application	94.59	85.19
Yes	5.41	14.81
No		
Preferences on either taxi service or e-hailing service	78.38	3.70
Taxi	21.62	96.30
E-hailing		
Age Group		
19 and below	5.45	25.93
20 to 29	54.05	25.93
30 to 39	21.62	11.11
40 to 49	10.81	0.00
50 and above	8.10	0.00

Table 5.3: Mean rating of Taxi segment and E-hailing segment (Interval case).

	Taxi Segment	E-hailing Segment
Waiting time	4.22	4.19
Temperature and ventilation (air conditioning)	4.03	4.37
customer satisfaction on e-hailing service		
Comfort	3.43	4.22
Fare	3.50	3.82
Reliability	3.53	4.44
Safety	3.43	4.11

Taxi segments manifestly dominated by respondents who favour taxi service and mostly has less or no experience on e-hailing service. This segments appears to be older compared to overall data and about 94.59% of them have not installed e-hailing related application. The waiting time to be served is also one of the focus of this segment. Thus, marketers who intend to draw customer from this segment has to concentrate on these loopholes. Whereas for E-hailing segment, it is dense with younger respondents who frequently utilise e-hailing service although there is 14.81% of them have not installed e-hailing application. This segment emphasizes on air conditioning which is where marketers should pay extra attention in. The last segment is a thoroughbred to all marketers. The Without Preferences segment is build up from respondents who have no preferences on mode of transportation. This segment rated most of the significant criteria as neutral. The task for marketers are to improve these criteria to a higher rating by constant efforts.

Lastly, multiple-step predictive process is strongly recommended based on this research. The use of clustering analysis as a preceding step for model building has enhanced the redundant neural network to the finest model among Decision Tree, Logistic Regression and NN. The fittest model selected in this case is NN with only 1.48% of misclassification rate.

5.2 Limitations

This research imposes numerous limitations. The survey of this research is conducted in merely within Klang Valley area, which might not be able to generalize behaviour of all Malaysian. Also, the data collected is highly biased toward E-hailing service and respondent of age group of 20 to 29 as well as students. These flaws have caused results that are least meaningful and hence, under-sampling over the target was forced to be applied. The sample size was then reduced from 400 to 90. The under-sampling procedure is undeniably essential for data mining, however, it might lead to loss of information. These biasness is potentially caused by the sampling method- convenience sampling. Besides, during the data selection process, instead of under-sampling, another possible option, over-sampling should also be taken into consideration and compare its performance with under-sampling. Unfortunately, due to time limitation, only under-sampling is carried out. Also, most of the settings and arrangement in SAS Enterprise Miner requires a lot of try and error work. For example, the proportion for data partition were arbitrary set and the number of layer as well as neurons used is in default setting. To obtain the finest information, by right, these setting have to be revised and regenerate its result,

until the optimal was found. However, again, due to time constrict, this iterative work is omitted. Lastly, the invasion of e-hailing service to Malaysia is considerably short, which induces the lack of study materials, respondents attend to questionnaire with misperception or even without any ideas.

5.3 Recommendation for Future Related Work

Researchers are strongly recommended to re-conduct the survey by using different sampling method like systematic sampling. Systematic sampling is a type of probability sampling method which theoretically poses higher creditability over convince sampling. It is used when there is a sequence of unit occurring naturally in space or time (Ranjan, 1995). In this situation, the sequence of unit occurring will be the respondents and space is the area of study. The aim is to obtain a less bias data. But, if the data obtained remains bias, either under-sampling or over-sampling has to be applied and it is encouraged to use both and select the superior one. Of course, researchers should revise the setting in SAS Enterprise Miner rather than arbitrary whim. Last but not least, further research can involve more or other input variables for study. For instance, researcher can consider variables that fall under the SERVQUAL model developed by Parasuraman, Zeithaml and Berry. The term SERVQUAL stands for “service quality”. This model proposed five service quality dimensions that is able to determine customers’ perception of any service business, tangibles, reliability, responsiveness, assurance and empathy (Hill and Alexander, 2017).

REFERENCES

Acharya, U.R., Ng, W.L., Rahmat, K., Sudarshan, V.K., Koh, J.E., Tan, J.H., Hagiwara, Y., Yeong, C.H. and Ng, K.H., 2017. Data mining framework for breast lesion classification in shear wave ultrasound: A hybrid feature paradigm. *Biomedical Signal Processing and Control*, *33*, pp.400-410.

Agatz, N., Erera, A.L., Savelsbergh, M.W. and Wang, X., 2011. Dynamic ride-sharing: A simulation study in metro Atlanta. *Procedia-Social and Behavioral Sciences*, *17*, pp.532-550.

Akil Y., 2017. Cabbies protest against GrabCar and Uber outside Parliament. *The Star Online,* [online] 21 March. Available at: < http://www.thestar.com.my/news/nation/2017/03/21/cabbies-protest-against-grabcar-and-uber-outside-parliament/> [Accessed 14 May 2017]

Anderson, D.N., 2016. Wheels in the Head: Ridesharing as Monitored Performance. *Surveillance & Society*, *14*(2), p.240.

Andrea D. P., Olivier C., Reid A. Johson and Gianluca B., n.d.. *Calibrating Probability with Undersampling for Unbalanced Classification*. Brussels, Belgium

B. Yegnanarayana, 2009. *Artificial Neural Network.* PHI Learning Pvt. Ltd

.

Bahari, T.F. and Elayidom, M.S., 2015. An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia Computer Science*, *46*, pp.725-731.

Bee Wah Y., Khatijahhusna A. R., Hezlin A. A. R., Simon F., Zuraida K. and Nik Nairan A., An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. In: T. Herawan, M. M. Deris and J. Abawajy, eds., Proceedings of the First *International Conference on Advanced Dataand Information Engineering*. Malaysia, 2014. Singapore: Springer

Chang, J.Y., Liou, W.C. and Ho, S.H., 2016. An Implementation of Distributed Framework of Artificial Neural Network for Big Data Analysis. *圖書館學與資訊科學*, *42*(2), pp.45-64.

Chi, M., 2014. Uber changes the _taxi scene‘ in KL and the winner is… us!. *Malaymail Online*, [online] 11 August. Available at: <http://www.themalaymailonline.com/money/article/uber-changes-the- taxi-scene-in-kl-and-the-winner-is-us> [Accessed 14 May 2017].

Clare B., 2013. *Handbook of Psychology and Diabetes: A Guide to Psychological Measurement in Diabetes Research and Practice*, New York: Routledge

Collica, R., 2011. *Customer segmentation and clustering using SAS Enterprise Miner*. SAS Institute.

Cranefield, S. and Nayak, A.C. eds., 2013. *AI 2013: Advances in Artificial Intelligence: 26th Australian Joint Conference, Dunedin, New Zealand, December 1-6, 2013. Proceedings* (Vol. 8272). Springer.

David A. J., 2015, *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada.* Toronto, Canada: Springer.

Ee, A.N., 2016. Uber Malaysia views challenges as opportunities. *The Sun Daily*, [online] 4 January. Available at: http://www.thesundaily.my/news/1654779 [Accessed 14 May 2017].

Emma H., 2016. Will Uber and Grab help or hinder Malaysia's congestion crisis? *Guardian Sustainable Business- Transport*, [online] 22 June. Available at:< https://www.theguardian.com/sustainable-business/2016/jun/22/uber-grab-malaysia-kuala-lumpur-congestion-crisis-car-sharing> [Accessed 15 May 2017]

Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P., 1996, August. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD (Vol. 96, pp. 82-88).

Fred C. P., 2000. *Logistic Regression: A Primer.* California: Sage Publication, Inc.

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

Harley M., Kerry B. and Josh B., 2012. *Outside in: The Power of Putting Customers at the Center of Your Business.* Boston, New York: Houghton Mifflin Harcourt Hill, N. and Alexander, J., 2017. *The Handbook of Customer Satisfaction and Loyalty Measurement*. Routledge.

Japkowicz, N., 2000, June. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence.*

Jo T., 2017. Women robbed on Uber ride suffers miscarriage. *The Star Online,* [online] 27 May. Available at: < http://www.thestar.com.my/news/nation/2017/05/27/woman-robbed-on-uber-ride-suffers-miscarriage/> [Accessed 14 May 2017]

John B., 2016. *Supervised and Unsupervised Machine Learning Algorithms.* [online] Available at: < http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/> [Accessed 25 Jun 2017]

Kattamuri S. S., 2013. *Predictive Modeling with SAS Enterprise Miner: Practical Solution for Business Application.* 2nd Edition. Cary, NC, USA: SAS Institute Inc.

Khan, D.M., Mohamudally, N. and Babajee, D.K.R., 2013. A unified theoretical framework for data mining. *Procedia Computer Science*, *17*, pp.104-113.

Knapp, T.R., 1998. *Quantitative nursing research*. Sage Publications.

Krontalis, A.K., 2016. Why ridesharing? Investigating factors driving commuters' intention and ridesharing behavior in Jakarta.

Leach, R.A., 2004. *The chiropractic theories: a textbook of scientific research*. Lippincott Williams & Wilkins.

Liew, Y.Y. and Yu, X.Y., 2016. *Malaysian's choice criterion on ride-hailing services against tai services: Survey in Klang Valley, Malaysia*. Degree. University Tunku Abdul Rahman.

Ling, 2008. *Tutorial: Pearson's Chi-square Test for Independence.* [online] Available at: < http://www.ling.upenn.edu/~clight/chisquared.htm > [Accessed 31 July 2017]

Lior R. and Oded M., 2014. *Data Mining with Decision Trees: Theory and Applications*. Second Edition. Singapore: World Scientific.

Macfie, B.P. and Nufrio, P.M., 2006. *Applied statistics for public policy*. ME Sharpe.

Mannila, H., 2000. *Theoretical frameworks for data mining*. SIGKDD Exploration, 1(2), 30-32.

Mark L. Blazey, 2009. *Insights to Performance Excellence 2009-2010: An Inside Look at the 2009-2010 Baldrige Award Criteria.* United State of America: William A. Tony.

Martin, A., 2006. *Knowledge-intensive Subgroup Mining: Techniques for Automatic and Interactive Discovery*. Germany: Inflix

Merriam-Webster, n.d.. *Trasnportation.* [online] Available at: <https://www.merriam-webster.com/dictionary/transportation> [Accessed 14 May 2017]

Murphy, K.P., 2012. *Machine learning: a probabilistic perspective*. MIT press.

National Aging and Disability Transportation Center, 2016. *NADTC 2016 Transportation Trends Report.* United State: NADTC.

Nik I. A., 2017. Legit rides: Uber, GrabCar to be recognized soon. *New Straits Times*, [online] 4 April. Available at: < https://www.nst.com.my/news/2017/04/227180/legit-rides-uber-grabcar-be-recognised-soon > [Accessed 14 May 2017]

Parliament of Malaysia, 2017. *Commercial Vehicles Licensing Board (Amendment) Act 2017*. [online] Availble at: < https://labourlawbox.com/files/bills/pdf/2017/MY_FS_BIL_2017_16.pdf > [Accessed 14 May 2017]

Priddy, K.L. and Keller, P.E., 2005. *Artificial neural networks: An Introduction* (Vol. 68). SPIE press.

Qian, X. and Ukkusuri, S.V., 2017. Taxi market equilibrium with third-party hailing service. *Transportation Research Part B: Methodological*, *100*, pp.43-63.

Randall S. C., 2017. *Customer Segmentation and Clustering Using SAS Enterprise Miner*. 3rd Edition. Cary, NC, USA: Institute Inc.

Ranjan K. S., 1995. *Practical Sampling Techniques.* Second Edition. United State, America: Marcel Deeker.

Ricardo G., n.d.. *Improving the Performance of Data Mining Models with Data Preparation Using SAS® Enterprise Miner*. Brazil: SAS Institute Brazil

Robert N., John E. and Gary M., 2009. HANDBOOK OF: Statistical Analysis & Data Mining Application. Oxford: Elsevier Inc.

Rpupesh D. G. K. and Kiren R. M. J. M., 2016. Predicting Rare Events Using Specialized Sampling Techniques in SAS®. Paper presented at SAS Institute Inc.. USA, 2016

Ryals, L., 2009. *Managing customers profitably*. John Wiley & Sons.

Sam, E.F., Adu-Boahen, K. and Kissah-Korsah, K., 2014. Assessing the factors that influence public transport mode preference and patronage: Perspectives of students of University of Cape Coast (UCC), Ghana. *Journal of Development and Sustainability*, *3*(2), pp.323-335.

SAS Enterprise Miner: Profiling Segments. n.d. [video] Directed by SAS instructor Cat Truxillo. SAS Institute Inc.

SAS Institute Inc., n.d. *Impute Missing Values.* [online] Available at: <http://support.sas.com/documentation/cdl/en/emgsj/62040/HTML/default/viewer.htm#a003124620.htm> [Acessed 20 July 2017]

SAS Institute Inc., n.d.. *Cluster node standardization when the Number of Clusters Specification Method is Automatic.* [online] Available at: < http://support.sas.com/kb/53/805.html> [Accessed 20 July 2017]

Shan S., 2015. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. New York: Springer.

Shrivastava, V., Khan, M. and Chaudhari, V.K., 2011, April. Neural network learning improvement using K-means clustering algorithm to improve the performance of web traffic mining. In *Electronics Computer Technology (ICECT), 2011 3rd International Conference on* (Vol. 1, pp. 78-82). IEEE.

SimilarWeb LTD, 2017. *GrabCar Mobile App*. [online] Available at :< https://www.similarweb.com/website/grabcar.com> [Accessed 14 May 2017]

SimilarWeb LTD, 2017. *Uber Mobile App*. [online] Available at :< https://www.similarweb.com/app/google-play/com.ubercab/statistics> [Accessed 14 May 2017]

Van Bommel, P. ed., 2005. *Transformation of knowledge, information and data: theory and applications*. IGI Global.

Vicen¸c Torra, Josep D. and Angel T., 2001. *Data Mining Methods for Linking Data Coming from Several Sources.* Institut d'Investigaci´o en Intel·lig`encia Artificial - CSIC, Dept. Computer Engineering and Maths (ETSE), Universitat Rovirai Virgili. Available at <http://www.iiia.csic.es/~vtorra/publications/unrestricted/confUNECE.2003.143.150. pdf> [Accessed 18 May]

Ng W. K., Masaru K., Jianzhong L. and Chang K., 2006. *Advances in Knowledge Discovery and Data Mining: 10th Pacific-Asia Conference, PAKDD 2006, Singapore, April 9-12, 2006, Proceedings*. Germany: Springer-Verlag Berlin Heidelberg

Yusof M., 2017. After robbery case, SPAD and Uber to discuss passenger safety, comfort. *Malaymail Online,* [online] 29 May. Available at: < http://www.themalaymailonline.com/malaysia/article/after-robbery-case-spad-and-uber-to-discuss-passenger-safety-comfort> [Accessed 14 May 2017]

APPENDICES

Appendix A Research Questionnaire

UNIVERSITI TUNKU ABDUL RAHMAN
Faculty of Science

MALAYSIANS' CHOICE CRITERION ON RIDE-HAILING SERVICES AGAINST TAXI SERVICES: SURVEY IN KLANG VALLEY, MALAYSIA

Questionnaire

Dear Respondents,

- We are final year students from Bachelor of Science (Hons) Logistics and International Shipping, UTAR. We are currently doing a research to study the determinants affecting why Malaysians prefer ride-hailing services in Klang Valley, Malaysia.
- This questionnaire contains four sections.
- We would highly appreciate if you could take a few minutes to complete this questionnaire.
- The information or details provided by you will be kept confidential.

Section A: Demographic Profile [Please place (√) where appropriate]

1. Gender:

☐ Male ☐ Female

2. Age Group (in years):

☐ 19 and below ☐ 20 to 29
☐ 30 to 39 ☐ 40 to 49
☐ 50 and above

3. Professions:

☐ Student ☐ Government
☐ Private institution ☐ Self-Employed/ Business
☐ Others

4. Monthly personal income:

☐ None ☐ RM2000 and below
☐ RM2001 – RM4000 ☐ RM4001 – RM6000
☐ RM6001 and above

Section B: Transportation Characteristics [Please place (√) where appropriate]

1. Do you use taxi services before? **[If NO, skip question 2 and proceed to question 3]**

☐ Yes ☐ No

2. How often do you use the taxi services?

☐ A few times per week ☐ A few times per month
☐ Once a month ☐ Once a few months

3. Do you use ride-hailing services (Uber or GrabCar) before? **[If NO, skip question 4 and proceed to question 5]**

☐ Yes ☐ No

4. How often do you use the ride-hailing services?

☐ A few times per week ☐ A few times per month
☐ Once a month ☐ Once a few months

5. Please circle your answer to indicate your opinion on barriers of using ride-hailing services (Uber or GrabCar).

[(1) = Least relevant; (2) = Less relevant; (3) = Neutral; (4) = More relevant; and (5) = Most relevant]

Barriers	Least relevant	Scale			Most relevant
Availability of technology	1	2	3	4	5
Reliability of application	1	2	3	4	5
User friendly	1	2	3	4	5

6. Do you prefer taxi services or ride-hailing services?

☐ Taxi Services ☐ Ride-hailing Services (Uber or GrabCar)

7. Do you install any applications of ride-hailing services (Uber or GrabCar)? **[If NO, skip question 8 and proceed to question 9]**

☐ Yes ☐ No

8. Which of the following applications did you install? **[Answer can be more than one]**

☐ Uber ☐ GrabCar
☐ Others, please specify _____

9. What type of payment system do you prefer?

☐ Cash ☐ Debit/ Credit Card

Section C: Determinants Affecting Taxi and Ride-hailing Services

Please rate the factors below to indicate your opinion regarding the importance of such factor towards taxi and ride-hailing services.

[(1) = Least important; (2) = Less important; (3) = Neutral; (4) = More important; and (5) = Most important]

Comfort

	Least important	Scale			Most important
Comfortable seat	1	2	3	4	5
Cleanliness and hygiene of vehicle	1	2	3	4	5
Temperature and ventilation (air conditioning)	1	2	3	4	5
Trip information	1	2	3	4	5

Fare

	Least important ← Scale → Most important				
Value for money of the fare	1	2	3	4	5
Payment system – ease of purchasing	1	2	3	4	5

Reliability

	Least important ← Scale → Most important				
Punctuality	1	2	3	4	5
Frequency of services	1	2	3	4	5
Waiting time	1	2	3	4	5

Safety

	Least important ← Scale → Most important				
Safety assurance	1	2	3	4	5
Driver behaviour and attitude	1	2	3	4	5

Section D: Customer Satisfaction on Taxi Services and Ride-hailing Services

Please rate the taxi services and ride-hailing services based on following criteria.

[(1) = Poor; (2) = Unsatisfactory; (3) = Satisfactory; (4) = Good; and (5) = Excellent]

Taxi Services Poor ← → Excellent					Factors	Ride-hailing Services Poor ← → Excellent				
1	2	3	4	5	**Comfort**	1	2	3	4	5
1	2	3	4	5	**Fare**	1	2	3	4	5
1	2	3	4	5	**Reliability**	1	2	3	4	5
1	2	3	4	5	**Safety**	1	2	3	4	5

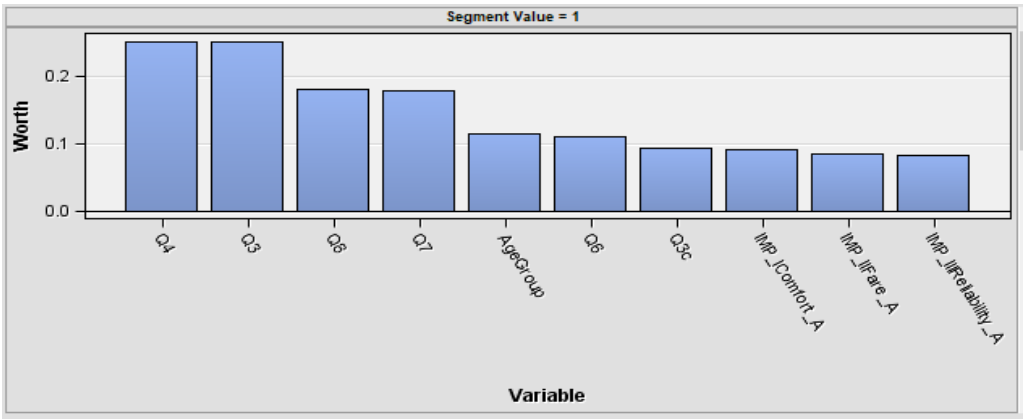
Thank you very much for your cooperation!

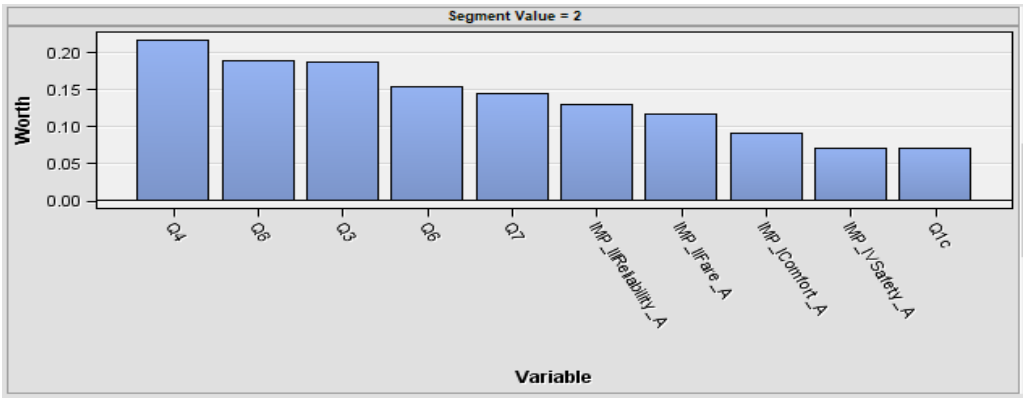
Appendix B Summary of Variables

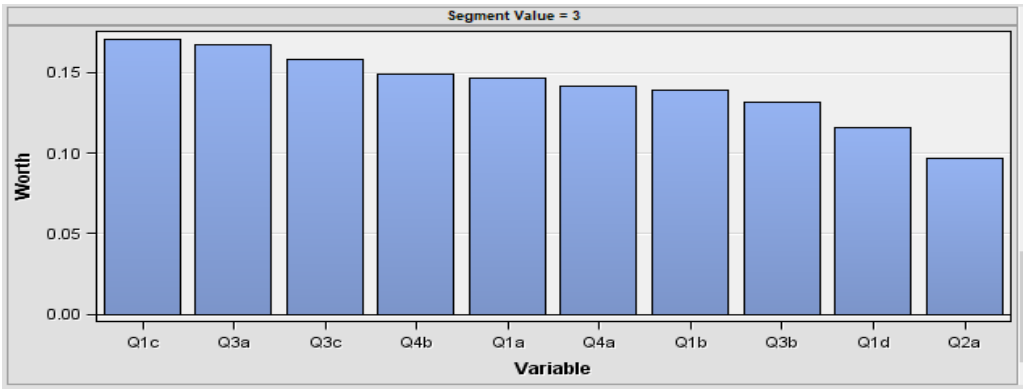
Variable	Label/ Description	Role	Level of Measurement
AgeGroup	Age group	Input	Interval
Gender	Gender	Input	Binary
iComfort	customer satisfaction on taxi service based on comfort	Input	Interval
IComfort_A	customer satisfaction on e-hailing service based on comfort	Input	Interval
iiFare	customer satisfaction on taxi service based on fare	Input	Interval
IIFare_A	customer satisfaction on e-hailing service based on fare	Input	Interval
iiiReliability	customer satisfaction on taxi service based on reliability	Input	Interval
IIIReliability_A	customer satisfaction on e-hailing service based on reliability	Input	Interval
Income	Monthly personal income	Input	Interval
ivSafety	customer satisfaction on taxi service based on safety	Input	Interval
IVSafety_A	customer satisfaction on e-hailing service based on safety	Input	Interval
Profession	Profession	Input	Nominal
Q1	Experience on taxi service	Input	Binary
Q1a	Comfort: Comfortable seat	Input	Interval
Q1b	Comfort: Cleanliness and hygiene of vehicle	Input	Interval
Q1c	Comfort: Temperature and ventilation (air conditioning)	Input	Interval
Q1d	Comfort: Trip information	Input	Interval
Q2	Frequent use of taxi service	Input	Nominal
Q2a	Fare: Value for money	Input	Interval
Q2b	Fare: Payment system- ease of purchasing	Input	Interval
Q3	Experience on e-hailing service	Input	Binary
Q3a	Reliability: Punctuality	Input	Interval
Q3b	Reliability: Frequency of service	Input	Interval
Q3c	Reliability: Waiting time	Input	Interval
Q4	Frequent use of e-hailing service	Input	Nominal
Q4a	Safety: Safety assurance	Input	Interval
Q4b	Safety: Driver behaviour and attitude	Input	Interval
Q5a	Barrier: Availability of technology	Input	Interval
Q5b	Barrier: Reliability of application	Input	Interval
Q5c	Barrier: User friendly	Input	Interval

Q6	Preferences on either taxi service or e-hailing service	Target	Binary
Q7	Present or absent of e-hailing related application	Input	Binary
Q8	Type of e-hailing related application installed	Input	Nominal
Q9	Preference on type of payment system	Input	binary

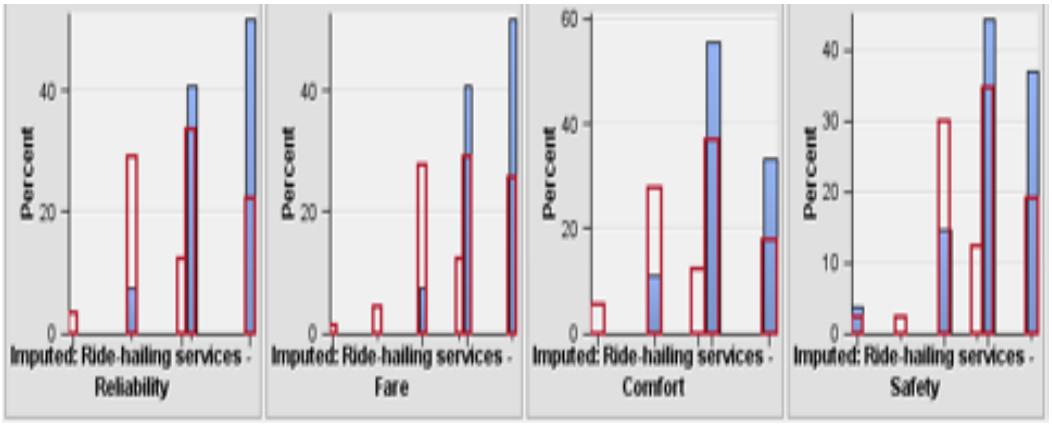
Appendix C Variable Worth Plot for Each Segment







Appendix D customer satisfaction on e-hailing service for segment 1(Taxi) and overall



Appendix E Output for Logistic Regression

The selected model is the model trained in the last step (Step 9). It consists of the following effects:

Intercept EXP_Q5a Q1 Q3 Q7 SQR_Q5c

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood		Likelihood		
Intercept Only	Intercept & Covariates	Ratio Chi-Square	DF	Pr > ChiSq
98.413	33.165	65.2479	5	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
EXP_Q5a	1	9.3262	0.0023
Q1	1	4.7584	0.0292
Q3	1	7.9588	0.0048
Q7	1	9.1656	0.0025
SQR_Q5c	1	5.2643	0.0218

Appendix F Classification Table for Neural Network

Classification Table

Data Role=TRAIN Target Variable=Q6 Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
RIDE-HAILING SERVICES	RIDE-HAILING SERVICES	50.7042	100	36	50.7042
TAXI SERVICES	RIDE-HAILING SERVICES	49.2958	100	35	49.2958

Data Role=VALIDATE Target Variable=Q6 Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
RIDE-HAILING SERVICES	RIDE-HAILING SERVICES	50	100	5	50
TAXI SERVICES	RIDE-HAILING SERVICES	50	100	5	50