DATA MINING ON RETAIL BANKING SIMULATION MODEL TO DEVISE PRODUCTIVITY IMPROVEMENT STRATEGIES

By:

LAI PIN SIEW

(Matrix No: 129357)

Supervisor:

Assoc. Prof. Ir. Dr. Chin Jeng Feng

May 2019

This dissertation is submitted to
Universiti Sains Malaysia
As partial fulfilment of the requirement to graduate with honors degrees in

BACHELOR OF ENGINEERING
(MANUFACTURING ENGINEERING WITH MANAGEMENT)

(MANUFACTURING ENGINEERING WITH MANAGEMENT)

School of Mechanical Engineering
Engineering Campus
Universiti Sains Malaysia

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently in candidature for any degree.

Signed ……...…...…..... (LAI PIN SIEW)

Date……..…...…...…....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by giving explicit references. Bibliography/references are appended.

Signed ……...…...…..... (LAI PIN SIEW)

Date……..…...…...…....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for interlibrary loan, and for the title and summary to be made available outside organizations.

Signed ……...…...…..... (LAI PIN SIEW)

Date……..…...…...…....

Witness by

Supervisor: Assoc. Prof. Ir. Dr. Chin Jeng Feng

Signed……...…...….....

Date……..…...…...…....

ACKNOWLEDGEMENT

My work would not have completed without the contribution and support from various parties. First and foremost, I would like to extend my greatest gratitude to my supervisor, Assoc. Prof. Ir. Dr. Chin Jeng Feng for his continuous assistance and divert me to the right research direction throughout the whole project execution. I truly appreciate him for guiding me step by step and explain patiently whenever I am in doubts.

Secondly, I wish to express my sincere thanks to Dr. Loh Wei Ping and the Master student, Mr. Chong Siu Hou for their precious assists and ideas which are extremely valuable for my study both theoretically and practically.

Besides, my parents are also an important inspiration for me. Thanks to my parents for giving encouragement, enthusiasm and invaluable assistance to me. Without all this, I might not be able to complete the work.

Lastly, I would like to thank my course mates and friends who have given comments, suggestions, and encouragement to allow me to generate more ideas and be more positive throughout this journey of completion of work.

Table of Contents

ACKNOWLEDGEMENT iii
ABSTRAK 1
ABSTRACT 3
CHAPTER 1 4
1.1 Chapter Overview 4
1.2 Background 4
1.3 Objective 5
1.4 Scope of Research 5
1.5 Organization of Thesis 5
CHAPTER 2 6
2.1 Chapter Overview 6
2.2 Data Mining 6
2.3 Retail Banking 8
2.4 Data Mining in Retail Banking 9
CHAPTER 3 15
3.1 Chapter Overview 15
3.2 Business Understanding 15
3.3 Data Collection 16
3.4 Data Selection 17
3.5 Data Cleaning 17
3.6 Data Transformation 17
3.7 Data Mining 18
3.8 Pattern Evaluation 19
CHAPTER 4 20
4.1 Chapter Overview 20
4.2 Business Understanding 20
4.3 Data Collection 22
4.4 Data Selection 28
4.5 Data Cleaning 29
4.6 Data Transformation 33
4.7 Data Mining 36

4.7.1 Data mining using visual analysis approach 36
4.7.2 Data mining using regression algorithm 41
4.8 Pattern Evaluation 43
CHAPTER 5 53
REFERENCES 55
APPENDICES 60

LIST OF TABLES

Table 2.1: Summary table of data mining applications in banking 13
Table 4.1: Data description 27
Table 4.2: Attributes description.....34
Table 4.3: Summarization of results 44

LIST OF FIGURES

Figure 3.1 Research methodology flow chart 15
Figure 4.1: Data mapping at macro level.....21
Figure 4.2: Data mapping at micro level 22
Figure 4.3: Four counters being simulated in WITNESS 23
Figure 4.4: Arrival profile of customer.....24
Figure 4.5: Interface to enter counter details 25
Figure 4.6: Programming the input rule for counter based on branches.....26
Figure 4.7: The transfer of data as record in database 27
Figure 4.8: Individual tables of branch info, operator info and transactions 28
Figure 4.9: Design view.....29
Figure 4.10: Select query is formed.....29
Figure 4.11: Connection of widgets in Orange.....30
Figure 4.12: Box plot of waiting duration before data cleaning 31
Figure 4.13: Box plot of servicing duration before data cleaning 31
Figure 4.14: Detection of outliers and linear projection 32
Figure 4.15: Blox plot of waiting duration after removal of outliers.....32
Figure 4.16: Blox plot of servicing duration after removal of outliers 33
Figure 4.17: Additional information obtained from given data 34
Figure 4.18: Discretization of data.....34
Figure 4.19: Bar chart in Orange 37
Figure 4.20: Widget connections for bar chart 38
Figure 4.21: Different type of data in data set (left) and display of data in Orange (Right) 38

Figure 4.22: Bar chart of number of transactions in different branches 39
Figure 4.23: Scatter plot in Orange.....40
Figure 4.24: Widget connections for scatter plot.....41
Figure 4.25: Scatter plot in Orange.....41
Figure 4.26: Widget connection for multiple linear regression 42
Figure 4.27: Details of each step in linear regression analysis 43

LIST OF ABBREVATIONS

ANN	Artificial Neural Network
ATM	Automated Operator Machine
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CART	Classification and Regression Trees
CHAID	Chi-square Automatic Interaction Detector
CRT	Chinese Remainder Theorem
DT	Decision Tree
FAHP	Fuzzy Analytic Hierarchy Process
ID3	Iterative Dichotomiser 3
LR	Logistic Regression
MLRNN	Multilayer Perceptron Neural Network
NaN	Not a number
NB	Naïve Bayes
NN	Neural Network
ODBC	Open Database Connectivity
POI	Point of Interest
QMS	Queue Management System
SQL	Structured Query Language
SVM	Support Vector Machine
TAN	Tree Augmented Naive Bayes
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
Weka	Waikato Environment for Knowledge Analysis

ABSTRAK

Perbankan runcit ialah perniagaan berorientasikan perkhidmatan. Pelanggan mengunjungi cawangan yang berbeza untuk mendapatkan perkhidmatan seperti pertanyaan transaksi, proses akaun bank dan lain-lain. Senario ini adalah pengendali kepada pelbagai jenis jawatan di setiap cawangan yang perlu membuat urusan di kaunter dan memenuhi tahap perkhidmatan pelanggan tertentu. Pelanggan datang ke kaunter dengan pelbagai corak yang berbeza. Maklumat seperti masa menunggu pelanggan, masa perkhidmatan pengendali, bilangan tiket yang diterima pada masa tertentu boleh dikumpulkan. Maklumat sedemikian membolehkan perlombongan data dilakukan untuk menemui pola yang boleh memisahkan jenis perkhidmatan. Kajian ini bertujuan untuk menunjukkan kemungkinan seperti itu melalui simulasi komputer. Khususnya, penyiasatan dibuat untuk perlombongan data bagi menentukan klustering dan klasifikasi cawangan dan produktiviti pengendali masing-masing. Metodologi penyelidikan melibatkan tujuh langkah. Langkah yang pertama, pemahaman mengenai perniagaan dilakukan untuk mendapatkan pemahaman mengenai motif memulakan latihan perlombongan data. Tahap makro dan mikro telah ditakrifkan untuk membezakan perbandingan antara cawangan. Langkah yang kedua, simulasi komputer akan disimulasikan dengan menggunakan WITNESS Horizon V.21, sebahagian besar adalah berdasarkan penerangan cawangan sebenar. Senario yang berbeza disimulasikan untuk mencerminkan tingkah laku operasi cawangan yang berlainan. Simulasi dapat menyimpan data (kira-kira enam ribu rekod) dalam pangkalan data. Langkah-langkah berikut melibatkan strategi pemprosesan pemilihan data yang berbeza (pemilihan, pembersihan, transformasi). Langkah keenam adalah perlombongan data, terutamanya menggunakan Python Orange V.3.20. Langkah ketujuh adalah penilaian corak untuk membangunkan strategi peningkatan produktiviti yang sesuai. Dalam kajian ini, pelbagai alat perlombongan data telah digunakan dan pelbagai pandangan telah dihasilkan. Terutamanya, cawangan yang menggunakan pengurusan lean telah menunjukkan peningkatan dalam produktiviti umum. Prestasi pengendali dapat dibeza berdasarkan tahun pengalaman. Penyelidikan ini memberi peluang kepada penyelidik untuk mengkaji produktiviti cawangan dengan teknik perlombongan data yang lain. Ini juga dapat membantu bank untuk memberi tumpuan kepada bidang yang betul untuk diperbaiki dan

meningkatkan keupayaan bank dalam membuat keputusan. Terdapat dua batas utama dalam penyelidikan. Pertama, model simulasi tidak dapat menunjukkan kerumitan keseluruhan operasi perbankan runcit. Kedua, batas kepada perisian yang menghalang tugas perlombongan data yang lebih kompleks untuk dilaksanakan

ABSTRACT

Retailing banking is service-oriented business. Customers visit different branches to procure services such as transaction inquiry, process bank account, etc. The scenario is operators of various positions of a branch have to render counter services and meet certain customer service level. Customers arrive at the counter at different patterns. Information can be collected, such as customer waiting time, operator on service time, number of tickets received at particular time. Such information allows data mining to be performed to discover patterns that could segregate the type of services. The study aims to demonstrate such possibility through computer simulation. Specifically, investigation is made for the data mining to determine the clustering and classification of the branches and operator productivity respectively. The research methodology involves seven steps. First, business understanding is performed to gain insight into the motive of initiating data mining exercise. Two levels, macro and micro levels were defined to differentiate inter and intra-branches comparisons. Second, a computer simulation would be constructed in WITNESS Horizon V.21, largely based on the description of a real branch. Different scenarios were built reflecting operating behaviors of different branches. The simulation stores data (about six thousands of records) in a database. The following steps involve different data selection processing strategies (selection, cleaning, transformation). Next is the data mining, primarily using Python Orange V.3.20. Last step will be pattern evaluation to develop suitable productivity improvement strategies. In this research, a variety of data mining tools have been deployed and multiple insights were generated. Notably, branches adopted lean management have shown improved in general productivity. Operator performances were able to differentiate based on years of experience. This research provides opportunities for researchers to examine the productivity of branches with other data mining techniques. It also helps bankers to focus on the right areas to be improved and increase the ability in decision making. There are two main limitations of the research. First, the simulation model does not capture the whole intricacy of retail banking front-end operations. Second, software limitations hindered more complex data mining tasks to be performed.

CHAPTER 1
INTRODUCTION

1.1 Chapter Overview

This chapter thoroughly reviews the background and rationale for the present study. It clarifies the existing challenges faced by traditional retail branches and outlines the use of data mining in banking industries. After introducing the fundamental references, the chapter subsequently describes the research objective, scope of study and provides the structure of the thesis.

1.2 Background

Banking is a business activity involved financial transactions of accepting deposits from public, investment in securities or lending as loans to the public. Over the years, the influence of technological innovation has introduced digital transformation into banks. Traditional banks with manual processes, transactions, activities are slowly replaced by digital services. When banks started to collaborate with other sectors (i.e.: mobile payment platform and e-commerce), customers are provided with more alternate channels which challenge the traditional branch method of delivering banking services (Olajide and Anthony, 2017). The service availability on the internet without the need of customers to visit branches and queue up has increased the convenience to people. Besides, online channel has taken advantage of lower average transaction costs compared to costs for a branch transaction (Luštšik, 2004). However, digital channels will ever negate the need for branches. Some customers still rely heavily on branch services. Physical interaction with bankers for investment products, loans and enquiries enable customers to choose traditional channel over online channel (Graupner *et al.*, 2015). In spite of that, it is challenging for a retail branch when it has to operate under a competitive financial atmosphere. For retail banks to have long competed on distribution, ways are needed to improve the productivity of branches.

Tremendous amount of data are generated in banks every day. These data include transaction details, personal and security information of customers. Banks have recognized

the importance of knowledge they have from their customers as their biggest asset (Hassani *et al.*, 2018). These mass data allows banks to have a clear view of the entire customer transactions, behavior and preference. One of the technologies that can be used in analyzing a huge amount of data will be data mining. Data mining has gained popularity in many industries, includes financial institutions, retails, telecommunications and insurance. It can be used to convert data collected into knowledge to reveal interesting relationships, associations and patterns within the data by processing the data. Recent applications of data mining in banking include improving customer satisfaction, marketing and optimizing strategic management (Hassani *et al.*, 2018). Banks with the understanding of customers' behavior and quality of personal service would be able to make appropriate decisions and formulate strategies for improvements. Eventually, it contributes to the improvement of performance of retail branch.

1.3 Objective

This research aims to investigate the application of descriptive and visualization techniques in data mining to improve the productivity of retail branches.

1.4 Scope of Research

In this research, it focuses on front office operations. It will consider both internal and external variables in a retail branch. External variables will be the factors which cannot be controlled by banks while internal variable will be factors which can be managed by banks such as operators and micro planning.

1.5 Organization of Thesis

Chapter 1 provides an introduction to the research. Chapter 2 reviews the retail banking and data mining literature. Chapter 3 presents the methodology planning before the execution of project. Chapter 4 demonstrates the executive methodology and data mining techniques used in the project. Results and discussion of data will be presented in this chapter. Chapter 5 will be the conclusion and future work.

CHAPTER 2
LITERATURE REVIEW

2.1 Chapter Overview

The purpose of this chapter is to build understanding of the subjects understudied. This chapter also provides reference points for new data and helping to deliver the potential impact to this research. In this chapter, the published information about data mining and retail banking will be discussed. For the first two sections, the subjects will be discussed in the structure of formal definition, background and recent developments or improvements. The third subject will be focused on the applications of data mining in banking industries and techniques used.

2.2 Data Mining

Data mining is an automated process of knowledge discovery from a large heterogeneous dataset which involves three disciplines: database, statistics, Artificial Intelligence and database (Goele and Chanana, 2012; Mejía *et al.*, 2017). First, data are extracted from data warehouses which are centralized repositories that gather data from several sources (Homayouni *et al.*, 2018). In data mining, these data will turn into an understandable format for predictive analysis and help in effective decision making (Guleria and Sood, 2014). Specifically, data mining derives patterns and extracts previously unknown, potentially useful information which cannot be obtained using conventional statistical methods (Bayer *et al.*, 2017). Cross-relations between elements in the database can be addressed to discover the elements of their indirect relationship. Data mining applications successfully implemented in various domains like health care, finance, retail, fraud detection, telecommunication etc.

Mansingh *et al.* (2013) treated data mining as a single phase in knowledge discovery. The process of knowledge discovery starts with business understanding focusing on business objectives and requirements. It is followed by data understanding which involves the collection of initial data, description, exploration and verification of data quality. Next, data preparation is performed by converting raw data to the format required by the selected

modeling tool. This paves the way to data mining which involves selecting modeling techniques, building multiple models to prepare dataset and assessing the generated dataset. Finally, outputs of the models generated are evaluated based on the previously defined business objectives.

Pre-processing tasks in data mining include data extraction, data cleaning, data integration, data reduction and data transformation. Data cleaning filters the data by filling in missing values, identifying outliers, and resolving inconsistencies. Data integration collects data from multiple sources to form a coherent data store while data reduction reduces data set into smaller volume yet produces the same analytical results. Data transformation is the consolidation of data into forms appropriate for mining. Strategies used include normalization, soothing, attribute construction etc.(Han *et al.*, 2012).

Data mining models can either be predictive or descriptive. According to Mints (2017), predictive model aims to predict the behavior of the analyzed system in a situation previously not encountered. Models under this category include classification, regression, time series analysis and prediction etc. Descriptive model includes a sense of human patterns and trend in data to discover properties of data studied and improve work efficiency (Silwattananusarn and Tuamsuk, 2012). Descriptive models include clustering, summarization, association rules and sequential pattern mining, which involve large number of marketing tasks related to the analysis of various target groups (Mints, 2017). Classification is a data mining technique that classifies new objects into predefined groups. According to Agyapong *et al.* (2016), classification deals with discrete or categorical target attributes and regression deals with numerical or continuous target attributes. Regression is used to determine the relationship between independent and dependent variables. Time series analysis is to forecast future values of certain set of data based on previously observed values. The input data will have a uniform distribution over time (Mints, 2017). Prediction is to approximate the determination of the values of some indicators in the future on the basis of the given values in the past and present (Mints, 2017). Clustering is the grouping of objects to their respective classes based on the information found in data describing of the object (A. Singh *et al.*, 2015). Unlike classification, clustering is done without defining the classes in advance. Association rules able to define the relationship of objects by discovering common patterns within a data set such as analyzing data from a

survey. Summarization is the process of finding a compact description for a subset of data (Silwattananusarn and Tuamsuk, 2012). Sequential pattern mining is used as pattern discovery for analogous trends in transaction data (A. Singh *et al.*, 2015). It implies the identification of cause-effect rules which takes into account the time factor (Mints, 2017).

Advancements in data mining have overcome various challenges include data from disparate locations, standardization of data mining languages (Goele and Chanana, 2012), construction of interactive and integrated data mining environment (Devi, 2013). The future models would expect to solve complex, high-speed data streams, noise in time series using soft computing techniques (Goele and Chanana, 2012).

2.3 Retail Banking

Retail bank is one part of classical bank, in which consumers physically come into the bank premises with their requests and needs for services (Sajic *et al.*, 2018). Services offered in retail bank include checking transactions and savings accounts, debit and credit cards, payments, investment plans, personal wealth management and loans. In this world with significantly changing consumers' behavior and technology-driven, retail banks need to derive strategies to stay competitive. Human resource management is crucial as it helps to increase the effectiveness of performance in any organization including banking. Retail banking requires multi-layer manpower for its different requirements in strategic and operational management, information technologies (IT) center, marketing, call center, back office and other support services (Sajic *et al.*, 2018). Three dimensions of high-performance work system in retail banking include quality of communications, skills and recognition and reward (Bartel, 2004). Performances of a branch are measured by sales of deposit, load products (Heskett *et al.*, 1997) and service profit chain (Bartel, 2004).

Throughout these years, improvements have been implemented to provide better customer experience in retail banks. According to Bapat (2015), branches equipped with Queue Management System (QMS) to reduce customer waiting time. Besides, extended customer hours and staggered meal breaks ensure uninterrupted customer services. For operator empowerment, roles of customer service representative and branch assist have been introduced to improve sales performance and assist customers in using high tech services

(Bapat, 2015). Moreover, process variation on branch performance has been improved by providing proper skills to operators to avoid customer dissatisfaction (Frei *et al.*, 1999).

2.4 Data Mining in Retail Banking

There is a wealth of contemporary literature related to data mining in retail banking. Application of data mining in customer acquisition and retention enable retail bank to gain new and retain existing customers. Several of them are reviewed here and have been summarized in Table 1.

Chye and Gerry (2002) applied data mining techniques to predict the churn of customers in the next six months. Data were derived and visualized using statistics tools, and Webgraph followed by clustering data into segments. After the segmentation, associate analysis was performed to identify cross-relation between demographic and transactional information with the customer churns. Logistic regression, neural network and decision tree were used as predictive modeling. As the results generated from prediction models were not identical, the training and test samples need to be compared to choose the best suitable model for a particular application.

Zadeh *et al.* (2011) proposed a model combining several techniques to profile banking customers. Important features needed for classification and segmentation of customers were listed and classified according to the profitability using NN model. Segmentation was carried out to form three segments for individual channels consist of ATM, terminal and web. It is done using K-Means algorithms: CRT, CHAID, and C5.0. The rule sets were extracted from each model and demographic characteristics of customers for each segment were listed to use as guidelines for bank executives to plan their future strategies.

A comparison study of multilayer perceptron neural network (MLPNN), tree augmented Naïve Bayes (TAN), logistic regression (LR), and Quinlan's new decision tree Model (C5.0) (Quinlan, 2002)was conducted on a real-world data of bank deposit by Elsalamony (2013). Bank direct marketing data set collected (Moro *et al.*, 2011) and turned into attributes with ranges (numerical attributes) and classes (category and binary

categorical attributes) followed by predictive modelling. Comparative results showed that MLPNN was the best for accuracy, LR took the best percentage for sensitivity, and C5.0 be the best in specificity analysis of testing samples. For analysis of training samples, C5.0 was the best in these three statistical measures.

Alizadeh and Minaei-Bidgoli (2016) proposed to use K-Means and bagging ANN to evaluate bank customers' loyalty. K-Means was used as a method of clustering for customer classification based on variables generated: age, education, gender, marriage, number of transactions, etc. The clusters were then labeled as Good and Bad using classification method- ANN. Authors suggested more influential variables such as income, employment are needed for modeling such dataset in future.

Wanke *et al.* (2016) adopted an integrated three-stage approach involving FAHP, TOPSIS, and ANN to analyse the performance of banks. In this research, performance evaluation was based on product of capital adequacy, assets quality, management, earning quality, liquidity, and sensitivity to market risk. The weights of main criteria were determined using FAHP. Then, TOPSIS was used to do the efficiency ranking. The results showed that ASEAN banking system has lower efficiency compared to US and European banks. ANN was performed on TOPSIS efficiency scores to determine the causes of inefficiency of the banking system and predict the trend on efficiency levels.

In a study by Met *et al.* (2017), efficiency of branch and location forecasting were analysed so that resources can be used effectively and economically. Financial data (profit, number of customers, etc) and non-financial data (branch age, number of employees, etc) were listed as variables in modelling phase. K- Means and LR, CHAID, C&R Tree were used as value-based and potential value-based methods. Profiles related to value segments of branches have been created using district location, number of customers and population per branch and POI variables through CHAID and C&R Tree algorithms. This provides guidelines to predict branch efficiency and facilitate evaluation. Lastly, efficiency scores for each branch and employees were established using financial variables.

The purpose of customer profiling and segmentation is to give a better understanding of customers' preferences so that better services could be provided (Moslehi *et al.*, 2018). Hassan and Tabasum (2018) illustrated the use of Naïve Bayes classification

and BIRCH clustering algorithm in customer profiling and segmentation. Customer classification and identification of segments were carried out by clustering of customer data. Then, Naïve Bayes algorithm was used to label these segments: high, medium, low and very low profitability from the analysis of behavioural, transactional, psychographic and demographic data of customers.

Risk management is one of the applications of data mining to evaluate the capability for customers to pay. Bhapkar (2018) conducted a comparative analysis of classification based data mining algorithms. WEKA tool preprocessing algorithms were used in data preprocessing operations to evacuate misrepresentation of data. The credit risk analysis model development was done by classified customers into classes using Naïve Bayes, J48 and Bagging algorithms. Based on the comparative results using cross-validation fold testing, it showed that Bagging has higher accuracy, therefore it is the best suitable for credit risk analysis.

Banking security can be reinforced through fraud detection. LR, ID3 algorithm and random forest decision tree were used to detect frauds in real time basis by analyzing incoming transactions (Patil *et al.*, 2018). Data pre-processing was done by interfacing SAS with Hadoop framework, a self-adaptive analytical framework model was then built for fraud detection. Among the three models, random forest decision tree performed the best regarding accuracy, precision and recall. Ensemble methods like random forest, bagging, boosting are very compatible with unbalanced areas (Dal Pozzolo *et al.*, 2014). However, huge amount of data will cause over-fitting of tree in memory of random forest.

Moslehi *et al.* (2018) have analysed the role of data mining techniques on identifying factors that affect electronic payments transactions. K-Means was used to categorize target variables, “Average Amount of ATM Transactions” and “Average Amount of Pin Pad Transactions”. These variables were divided into three clusters, low, medium, and high and labelled according to average transactions. Then, CART algorithm was applied to detect the hidden patterns of payment transactions. Consideration of demographic characteristics like education, salary, age will provide a better analyse of customers’ desire to use the tools and payment methods.

One concern involved in data mining will be the problem of unbalanced datasets especially it only has two classes of datasets (Fadaei Noghani and Moattar, 2017). It will greatly affect the decision costs (López *et al.*, 2012).

Table 2.1: Summary table of data mining applications in banking

Sector	**Title**	**Authors**	**Knowledge Type**	**DM Techniques**
Customer retention and acquisition	Introducing A Hybrid Data Mining Model to Evaluate Customer Loyalty	Alizadeh and Minaei-Bidgoli (2016)	Customer loyalty evaluation	Clustering (K-Medoids, X-Means; K-Means) Classification (DT, ANN, NB, KNN, SVM)
	Data Mining and Customer Relationship Marketing in Banking Industry	Chye and Gerry (2002)	Churn Modelling	Clustering Classification (LR, NN, DT)
	Bank Direct Marketing Analysis of Data Mining Techniques	Elsalamony (2013)	Compare performance of DM techniques in bank direct marketing data set using confusion matrix	Classification (MLPNN, TAN, LR, C5.0)
Customer profiling and knowledge Customer segmentation	Customer Profiling and Segmentation in Retail Banks using Data Mining Techniques	Hassan and Tabasum (2018)	Customer profiling; Customer segmentation	Classification (TAN) Clustering (BIRCH)
	Profiling Bank Customers Behaviour using Cluster Analysis for Profitability	Zadeh *et al.* (2011)	Customer profiling	Classification (NN, CRT, CHAID, C5.0) Clustering (K- Mean)

Risk management and Investment banking	Comparative Analysis of Classification Based Data Mining Algorithms for Credit Risk Analysis	Bhapkar (2018)	Compare performance of DM techniques in credit risk analysis using cross validation fold testing	Preprocessing (WEKA) Classification (TAN, J48, Bagging)
Security and fraud detection	Predictive Modelling For Credit Card Fraud Detection Using Data Analytics	Patil *et al.* (2018)	Fraud Detection	Classification (LR, ID3, Random Forest Decision Tree)
Other advanced options	Analyzing and Investigating the Use of Electronic Payment Tools in Iran using Data Mining Techniques	Moslehi *et al.* (2018)	Analyze payment tools using DM techniques	Clustering (K-Means) Classification (CART)
	Branch Efficiency and Location Forecasting: Application of Ziraat Bank	Met *et al.* (2017)	Monitor branch efficiency and location forecasting	Clustering (K-Means, Two Step)
	Predicting performance in ASEAN banks: an integrated fuzzy MCDM–neural network approach	Wanke *et al.* (2016)	Performance evaluation	MCDM (FAHP, TOPSIS) Classification (ANN)

CHAPTER 3
METHODOLOGY

3.1 Chapter Overview

Research methodology is a systematic way to resort a research problem. Its purpose is to derive work plan to the research (Rajasekar *et al.*, 2006) It focuses on how to choose appropriate methods and analysis and develop appropriate solutions for the problem (Kilani and Kobziev, 2016). In Chapter 3, all needed methods and tools to be used in this research will be outlined. The research methodology is divided into several stages, listed in Figure 3.1.

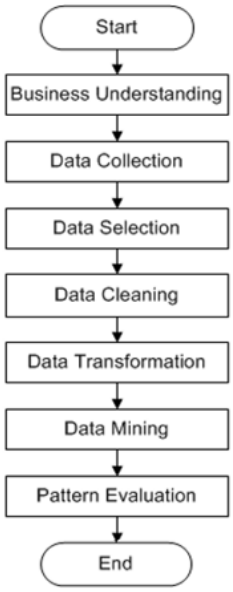

Figure 3.1 Research methodology flow chart

3.2 Business Understanding

The first step requires to understand the bank objectives from a business perspective. Such information would be obtained from suitable bank personnel. From the objectives and current situation, appropriate data mining problem types (e.g. visualization, classification or prediction) and goals could be determined to achieve the bank objectives (Mansingh *et al.*, 2013). Specifically, the objectives would be prioritized, broken down, refined and

mapped into different levels of sub-questions. This is reminiscent of a divide-and-conquer strategy, where a collection of sub-questions would be defined in relation to the business objectives and needs. They represent different facets of the business objectives and could be more effectively addressed using specific data mining tools and filtered datasets. Next, deliberation and consensus have to be made to focus on the key sub-questions, potentially leading to a smaller number of sub-questions. In the following steps, data mining would be performed separately for each sub-questions. The findings from addressing these sub-questions are then compiled and analyzed collectively to derive a conclusion to the bank objectives.

3.3 Data Collection

The source of data for mining will be generated using computer simulation based on sample data obtained by real retail banking. This is to avoid breaching of confidential information. Also, simulation model allows concepts and ideas to be more easily verified. WITNESS simulation is a flexible, proven process simulation technology widely used in industries to develop features models and simulation. It is capable for discrete event and continuous modeling to address wide range of business problems. WITNESS has open connectivity to common data sources including Microsoft Excel, databases and cloud services. Simple coding to define and structure the logic makes WITNESS a friendly simulation tool to new users. Data can also be easily exported for external analysis. The construction of the simulation would go through the proper verification process.

Next, the data generated through the simulation would be stored in the database developed using Microsoft Access. Access is a desktop database application that can store, organize, manipulate data and link related information. It consists of tables, queries, forms and reports, macros and modules. Access can work with most popular databases that support the Open Database Connectivity (ODBC) standard, including SQL Server, Oracle, and DB2. Data in Access can also import from and export to word processing files, spreadsheets, or database files directly.

3.4 Data Selection

This phase emphasizes on the integration of data into single data source and selecting only the dataset which is required to fulfil the mining goals. The selection of data is to minimize time spent on mining unimportant data and reduce the capacity required to store data. Duplicated and redundant data will also be removed in this stage.

3.5 Data Cleaning

Data cleaning is the process of detecting and correcting noisy, erroneous or incomplete data. This step is essential in data mining as quality decisions must be made based on quality data. Poor quality data will produce unreliable data mining results. In this study, data cleaning will be carried out using Orange. Orange has two outlier detection methods, one with SVM (multiple kernels) and the other with elliptical envelope. One class SVM with non-linear kernel classifies data as similar or different from the core class whileCovariance estimator fits ellipsis to central points with Mahalanobis distance metric. Both are distance-based outlier detection methods. The software would be explained further in section 3.6.

3.6 Data Transformation

Data transformation is the stage to convert data into form of specific format that is ready-to-use. Data transformation methods include smoothing noise, aggregation, normalization and feature construction (Sun *et al.*, 2018). This stage will focus on data normalization and feature construction as data has been cleaned and integrated in the previous stages. Data normalization is to scale the data into smaller and specific range so that it is readable. Feature construction is to construct new attributes from the given data. From there, additional information or features can be obtained.

In this research, data transformation will be carried out using Excel. Microsoft Excel is a spreadsheet software used to record and analyze data. It has basic features of calculation, graphing tools, pivot tables, and macro programming. It is widely used in many businesses, which use it to record plan budgets, chart data, expenditures and incomes. Other

than data records, Excel can also communicate with external sources. It can be programmed to pull in real-time data from external sources and run the data through formulas to update such information in real time.

3.7 Data Mining

This is the stage of analyzing data to discover hidden patterns. Multiple data mining techniques can be used to analyze formatted data to fulfill particular data mining goals set in previous phase. In this study, visualization techniques and regression algorithm will be chosen as the data mining method to be used. Visual data exploration is an alternative to mechanical data mining algorithms which is proven to be an effective tool in knowledge discovery (Wang *et al.*, 2000). TreeMap, Parallel Coordinates, Scatter-Plot Matrices and Spatial Visualization are some of the visualization techniques that are widely acceptable and represented in many ways (K. Yeh, 2008). The type of visualization has to be chosen based on the type of data to be presented. For sub-questions consisting of numeric data for both independent and dependent variables, line diagrams and scatter plots will be used. If only the dependent variable is numeric, bar graphs would be used (Slutsky, 2014). Regression is used as a predictive technique that could form relationship between dependent variable and set of predictors. It begins with set of data in which target values are known, then it estimates the value of target as a function of the predictors (N. Singh *et al.*, 2019; Wang *et al.*, 2000).

In this study, Orange will be used to perform data mining functions. Orange is an open source data mining and machine learning suite for data analysis through Python scripting and visual programming. Components in Orange is called as widgets, offers functionalities such as reading the data, visualizing and filtering, modeling, scoring, and model evaluation (Rangra and Bansal, 2014). Visual programming platform with graphical user interface components makes Orange an interactive data visualization and mining tool.

3.8 Pattern Evaluation

The information obtained from data mining will be analyzed and evaluated in the context of business objectives in the first phase. Useful information would help decision makers to gain understanding of their business and make better decisions.

CHAPTER 4
RESULT AND DISCUSSION

4.1 Chapter Overview

This chapter will be presenting the key findings of the research. The chapter illustrates the steps to develop the research start from business understanding to knowledge discovery. The results will then be analyzed under the discussion in order to abstract the significance of the data. In discussion, graphs will be interpreted and presented in a meaningful way. Recommendations for bankers to improve the productivity of branches will be explained in the last section of the chapter.

4.2 Business Understanding

The goal of this study is to look into the relative productivity of retail banks. Therefore, it is important to identify the variables that will affect the performance of branch.

Based on the business objectives, the study was started by breaking the business into multiple levels. Two levels which were macro and micro were defined in this case. Macro level provided a bigger picture hence a general understanding and comparison of the major entities of the business. On the other hand, micro level revealed small details of the business, such as operation performances and provided rationale to support findings in the macro level. A brainstorm session resulted in topics being defined respectively for micro level and macro level as stated below.

Macro level:

Topic 1: District location of branches in affecting the performance of the branches

Topic 2: Population serving in affecting the performance of the branches

Topic 3: Number of counter influence the performance between branches

Topic 4: Impact of micro planning on performance of branches

Micro level:

Topic 1: Influence of different operators to performance of bank.

Topic 2: Impact of age and working years of operators to the performance of a branch

Topic 3: Impact of period in a day to the number of transaction in bank

Topic 4: Impact of days in a month to the number of transaction in bank

Topic 5: Prediction for servicing time in a branch

Before proceeding to data mining, the information was organized by illustrating the concept map. Figure 4.1 and Figure 4.2 show the mapping of the retail bank. The performance of the retails banks was the focus of the concept map and served as a reference point. The specific and exclusive concepts were arranged hierarchically below. Relationship between elements was described by linking words located on the connecting lines. From the mapping, visualization was done on how items were related and the causes of the come about of certain elements.

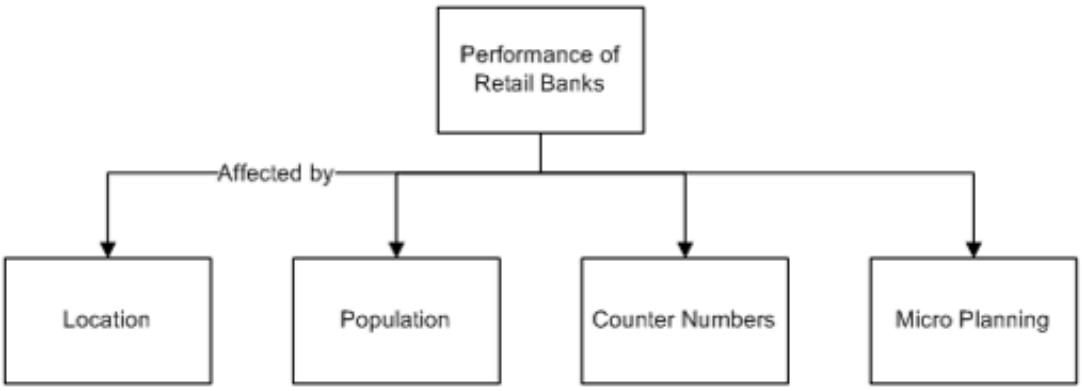

Figure 4.1: Data mapping at macro level

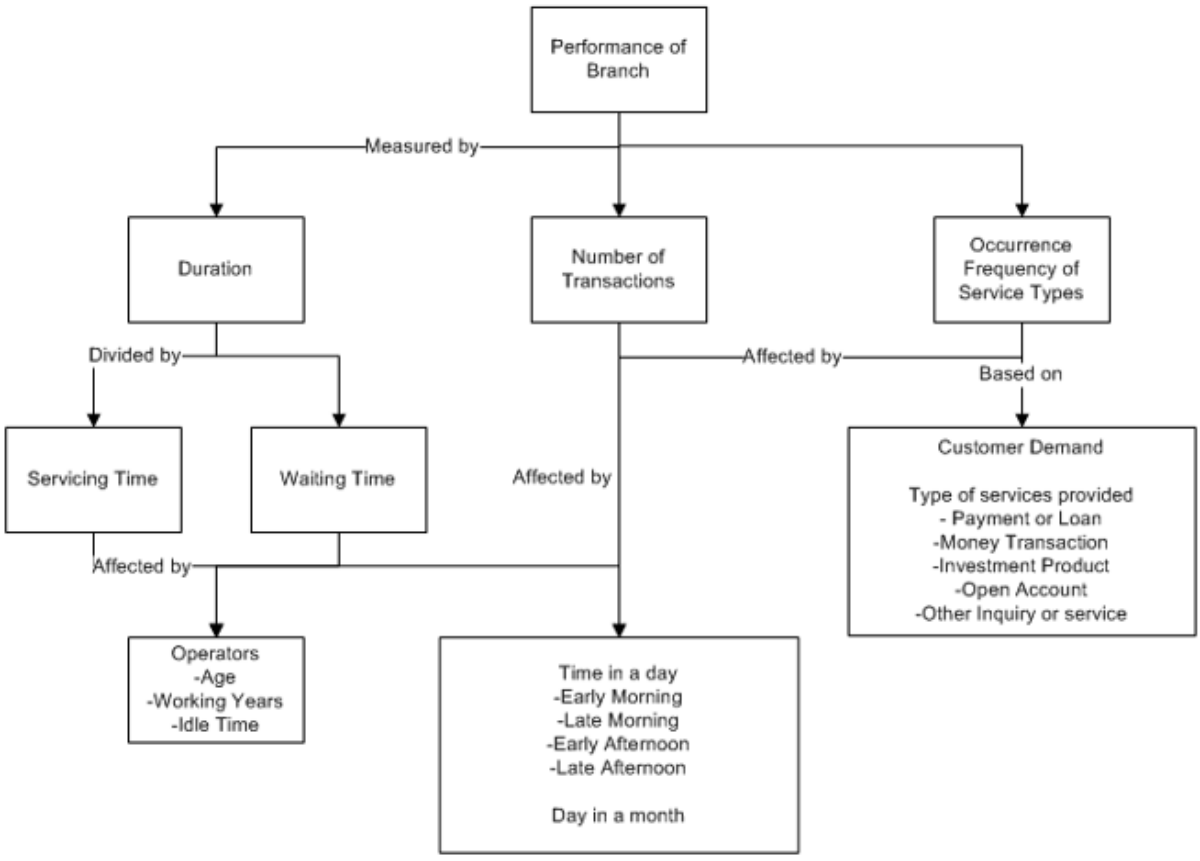

Figure 4.2: Data mapping at micro level

4.3 Data Collection

Bank branches remained as critical components in the retail banking delivery model. Retail branches were different from various perspectives. For instance, there were external factors such as geographical location and traffic measures, and the internal environment including the size of the premise, its operation layout that is highly relative to the flow of process. All these factors have to be strategic as they might contribute to the productivity of the operation. Selection of branch location was a critical decision to achieve objectives set by the banking entity. General criteria that were needed to be considered in selecting the location for a branch included demographic, socio-economic, sector of employment, banking and trade potential in that area (Cinar, 2009).

Five retail banking branches were modeled in the WITNESS Horizon V2.1 discrete-event simulation software. Only front end operations were modeled. In general, the

branches offered a number of services over the counter. The number of counters was branch-dependent. Each counter will be served by a till operator. Each service was called by a customer and can only be performed by an operator. Branches were operated from 9.00 am to 4.30 pm. Customers arrived at different hours. They would wait at the waiting area to be served. Based on the arrangement, some counters were working on general services and the others with designated services. Customers who have waited over a specific period of time would leave the branch, emulating their reaction over frustrations in waiting. After the service, the customer would exit the system. The service required by the customer base would be determined by a predetermined ratio, which was varied based on branches.

In WITNESS, a counter and a working operator were considered a set which was modeled with a single machine element. Therefore, a branch having four counters would have an array of four machine elements being programmed, as shown in Figure 4.3. On the other hand, customers were represented by products. Customers arrived at the branch by following a customized arrival profile, where peak hours somewhat fell between 12 noon to 2 pm to coincide with lunch hours. An example of arrival profile is demonstrated in Figure 4.4. Customer would hold a number of attributes, namely ID, arriving time and service required. Waiting area was modeled as a buffer (CusQue) to hold a large (30) but finite number of customers.

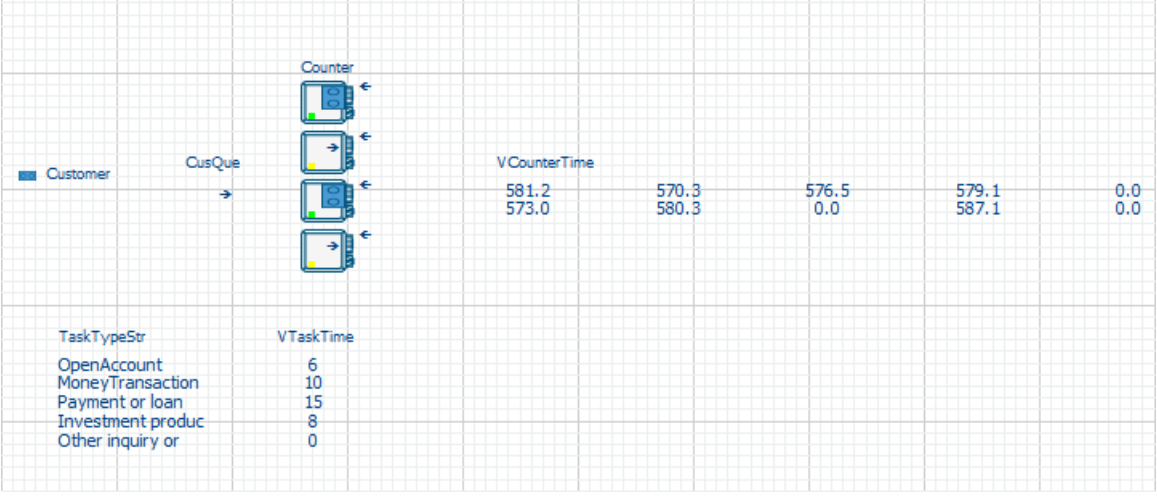

Figure 4.3: Four counters being simulated in WITNESS

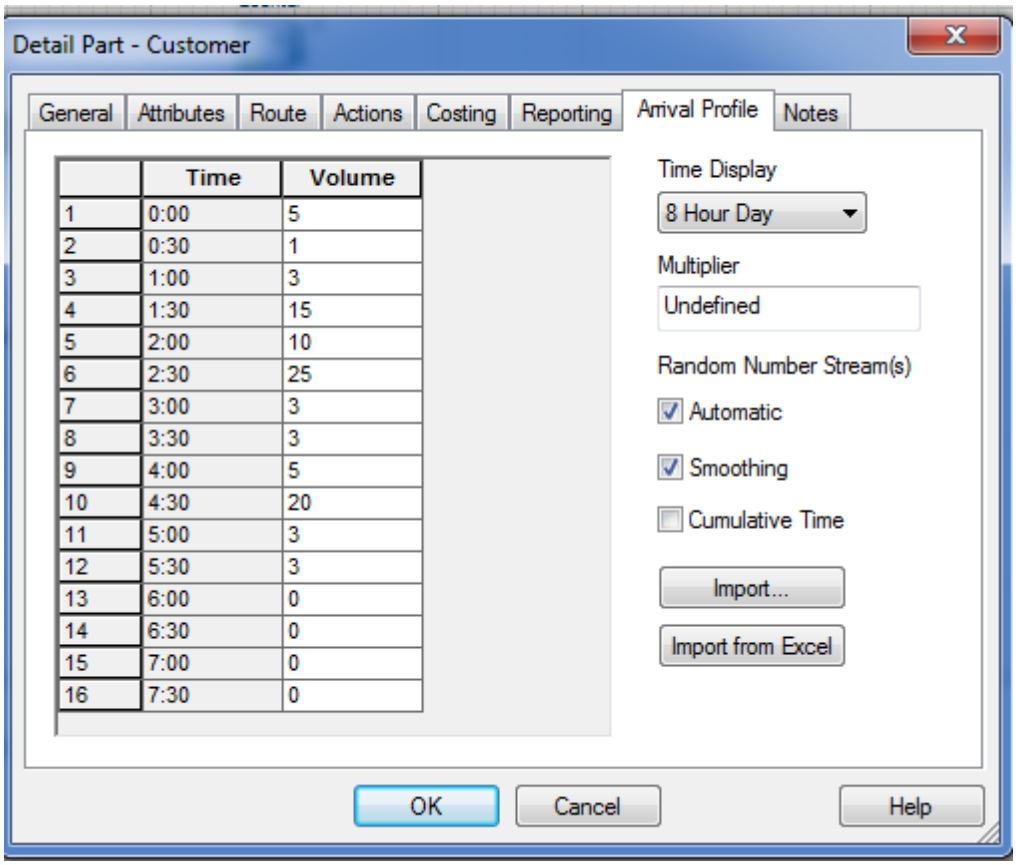

Figure 4.4: Arrival profile of customer

Five services were provided, namely open account, money transaction, payment or loan, investment product, and other inquiry or services. The interface in WITNESS to enter the counter detail is shown in Figure 4.5. The time to provide a service depended on multiple aspects: the experience of operator and the management policy. Operators having more experience would be able to close a service faster. Management policy indicated that operators who undergo training potentially help the operators to carry out services more efficiently. Truncated normal distribution was used to generate the service time, where mean would factor in the above aspects. The service time was adjusted as actions on input, meaning that it would be triggered when a new product (customer) was being loaded to the counter (machine). The assignment of customer to counter was identified as machine input rule in WITNESS being based on the first-come-first-served rule, subjected that the counter has the capability to perform such service. The specialization of the counter as such would be branch-dependent. For example, a branch would have counters to offer all types of