

Leveraging Purchase History and Customer Feedback for CRM: a Case Study on
eBay's "Buy It Now"

by

Jie Chen

B.Sc., Central China Normal University, 2013
A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in the Department of Computer Science

© Jie Chen, 2016
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by
photocopying or other means, without the permission of the author.

Leveraging Purchase History and Customer Feedback for CRM: a Case Study on
eBay's "Buy It Now"

by

Jie Chen

B.Sc., Central China Normal University, 2013

Supervisory Committee

Dr. Kui Wu, Supervisor
(Department of Computer Science)

Dr. Alex Thomo, Departmental Member
(Department of Computer Science)

Supervisory Committee

Dr. Kui Wu, Supervisor
(Department of Computer Science)

Dr. Alex Thomo, Departmental Member
(Department of Computer Science)

ABSTRACT

The rapid growth of e-commerce contributes to not only an increase in the number of online shoppers but also new changes in customer behaviour. Surveys have revealed that online shopper's brand loyalty and store loyalty are declining. Also the transparency of feedback affects customers' purchase intention. In the context of these changes, online sellers are faced with challenges in regard to their customer relationship managements (CRM). They are interested in identifying high-value customers from a mass of online shoppers, and knowing the factors that might have impacts on those high-value customers. This thesis aims to address these questions.

Our research is conducted based on an eBay dataset that includes transaction and associated feedback information during the second quarter of 2013. Focusing on the sellers and buyers in that dataset, we propose an approach for measuring the value for each seller-buyer pair so as to help sellers capture high-value customers. For a seller, the value of each of its customers has been obtained, and we create a customer value distribution for the seller so that the seller knows the majority of its customers' consumption abilities. Next, we categorize sellers based on their customer value distributions into four different groups, representing the majority of customers as being of high, medium, low, and balanced values, respectively. After this classification, we compare the performance of each group in terms of the sales, percentage of successful transactions, and the seller level labelled by the eBay system. Furthermore, we perform logistic regression and clustering to the sellers' feedback data in order to investigate whether a seller's reputation has an impact on the seller's customer

value distribution. From the experiment results, we conclude that the effect of negative ratings is more significant than that of positive ratings on a seller's customer value distribution. Also higher ratings about "Item as Described" and "Shipping and Handling Charges" are more likely to help the seller attract more high-value buyers.

Contents

Supervisory Committee	ii
Abstract	iii
Table of Contents	v
List of Tables	vii
List of Figures	viii
Acknowledgements	ix
1 Introduction	1
1.1 Motivation	1
1.2 Introduction to eBay	3
1.3 Contributions	5
1.4 Agenda	6
2 Background and Related Work	7
2.1 Customer Identification	7
2.2 Customer Retention	8
2.3 eBay Data Analytics	10
2.3.1 Academic research	10
2.3.2 Industrial Application	11
3 EBay Data Collection	13
3.1 eBay Transaction Data from Terapeak	13
3.2 Data Collection with eBay API	14
3.3 Data Storage	16

4	A Measure of Customer Value	18
4.1	Measuring Customer Value	18
4.2	The Impact of Different Weight Values	20
5	Seller Segmentation Based on Distribution of Customer Values	24
5.1	Distribution of Customer Values	24
5.2	Four Baseline Distributions	24
5.3	Categorization of eBay Merchants	29
5.3.1	Methodology	29
5.3.2	Parameter Tuning	31
5.3.3	Results of Categorization	32
5.4	Comparison of Sellers in Different Categories	32
5.4.1	Number of Transaction	32
5.4.2	Percentage of Successful Transactions	34
5.4.3	PowerSeller Level	35
5.4.4	Conclusion	36
6	Effects of Seller Reputation on the Customer Value Distribution	38
6.1	EBay Feedback System	39
6.2	Feedback Data Collection and Preprocessing	40
6.3	Analysis of Factors Using Clustering	41
6.3.1	Methodology	41
6.3.2	Result and Conclusion	45
6.4	Analysis of Factors Using Logistic Regression	45
6.4.1	Variables and Modelling	45
6.4.2	Testing Global Null Hypothesis	48
6.4.3	Analysis of Maximum Likelihood Estimates	48
6.4.4	Conclusion	49
7	Conclusion and Future Work	52
7.1	Conclusion	52
7.2	Future Work	53
	Bibliography	55

List of Tables

Table 2.1 Distributions of Articles by CRM Elements and Data Mining Techniques	9
Table 2.2 Applications for eBay Customer Relationship Management . . .	11
Table 3.1 Statistics on eBay Sellers	14
Table 3.2 Selected Output Fields	16
Table 4.1 An Example of Calculation of Customer Value	20
Table 5.1 Value Changes in Baseline Distributions ($\epsilon = 0.05$)	30
Table 5.2 Number of Transactions in Different Classes	33
Table 6.1 eBay Detailed Seller Rating System	39
Table 6.2 Variables and Shorthand	41
Table 6.3 eBay Stars and Ratings	42
Table 6.4 Distances Between Each Cluster in C_0 and Each One in C_1 . . .	44
Table 6.5 Distances Between C_0 and C_i from Smallest to Largest	47
Table 6.6 Testing Global Null Hypothesis	48
Table 6.7 Analysis of Maximum Likelihood Estimates	50

List of Figures

Figure 1.1	Number of digital buyers in Canada from 2011 to 2018 (in millions) [5]	2
Figure 1.2	Number of eBay’s total active buyers from 2010 to 2015 [5]	4
Figure 1.3	eBay’s gross merchandise volume from 2nd quarter 2014 to 3rd quarter 2015 (in billion U.S. dollars) [5]	5
Figure 3.1	Framework for eBay data collection	15
Figure 3.2	Request for feedback data on a specific seller ID	15
Figure 3.3	Database structure	17
Figure 4.1	Scatterplot of F and M values	21
Figure 4.2	Comparison on average number of customers	22
Figure 4.3	Comparison on average and standard deviation of number of customers	23
Figure 5.1	Distribution of the values of a seller’s customers. (Seller ID: 208428)	25
Figure 5.2	Five loyalty distributions proposed in [35, 34]	26
Figure 5.3	Centroids of four clusters yielded by K-means	27
Figure 5.4	New baseline distributions	28
Figure 5.5	Baseline distributions after de-zero processing	30
Figure 5.6	KL distances with change of weight (α)	32
Figure 5.7	Result of classification with $\alpha = 0.6$	33
Figure 5.8	Distributions of successful transaction percentage in different groups	34
Figure 5.9	Sellers’ Powerseller levels in different groups	35
Figure 6.1	Screen shot of the feedback profile of an eBay seller	40
Figure 6.2	Clusters based on customer value distributions and feedback factors	43

ACKNOWLEDGEMENTS

I would like to thank:

My Supervisor, Dr. Kui Wu, for his support and mentoring throughout the past two years. I am deeply grateful to him for giving me the freedom to follow my passion, and at the same time providing patient guidance on my research.

Terapeak Co-founder, Anthony Sukow, for offering a dataset and providing me with the opportunity to intern at Terapeak. This thesis would not have been possible without his help.

My Labmates, Fang Dong, Cheng Chen, Guoming Tang, et al., for their constant encouragement and insightful comments.

My Friends, for sincere friendship. I have been fortunate to make so many friends with common interests.

My Family, for always being there. I would like to express my immense gratitude to my parents who give me strength, courage and love.

*“Nothing behind me,
everything ahead of me,
as is ever so on the road.”*

Jack Kerouac

Chapter 1

Introduction

1.1 Motivation

Along with growing Internet penetration and development of e-commerce, there are new changes in customer demographics and customer behaviour. As shown in Figure 1.1 provided by Statista [5], the number of online shoppers in Canada has increased steadily, and it is forecasted to remain a growth trend in the future. Also according to a report from CBC [1], the new data from Canada Post presents that “around 76 percent of Canadian households shopped online” in 2014.

In addition to demographic changes, customer behaviour are changing as well. As the rapid growth in digital buyers contributes to plenty of business opportunities, a large number of sellers have created their online presence. As reported in [3], eBay claimed that there had been 25 million stores with more than 800 million items on sale by the end of 2014. Due to the exposure to the large quantity of choices, and the fact that everything is now transparent on the Internet, it has become easier for online customers to switch between outlets, leading to a decline in store loyalty [17, 56]. Also, brand loyalty is vanishing pointed by [41, 46]. What’s more, the visibility of customer ratings and reviews might affect customer behaviour. As the feedback mechanism is integrated with e-commerce marketplaces, buyers are able to leave feedback on their shopping experience and spread information about different sellers. The survey in [4] reveals that “61 percent of customers read online reviews before making a purchase decision”. Of that percentage, “68 percent of consumers trust reviews”. Also the case studies in [22, 36] demonstrated that the online feedback ratings have an impact on online sales.

The above changes in customers behaviour introduce following challenges to online merchants:

1. Given a large number of people shopping online, how do sellers effectively target potential online shoppers?
2. Considering the fact that a decline in store loyalty, is it still significant for sellers to pay attention to customer retention?
3. Given a large amount of feedback from customers, how do sellers learn from this information so as to attract customers?

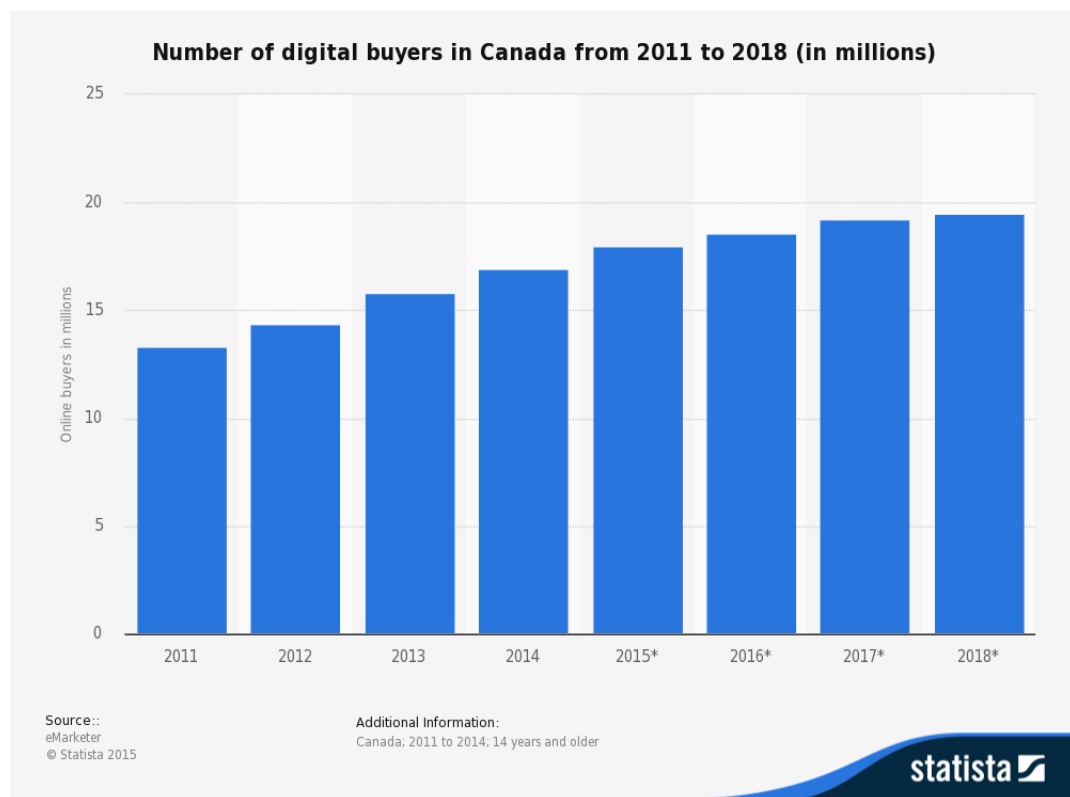


Figure 1.1: Number of digital buyers in Canada from 2011 to 2018 (in millions) [5]

These questions are normally discussed in the domain of Customer Relationship Management (CRM). There is no unified definition of CRM, and researchers define it from different perspectives. For example, from a business strategy, CRM is defined as “the strategic process of selecting customers that a firm can most profitably serve

and shaping interactions between a company and these customers” [40]. In addition to business strategy, the technology perspective is often considered from 1990s. Accordingly, CRM is defined as “the core business strategy that integrates internal processes and functions, and external networks, to create and deliver value to targeted customers at a profit. It is grounded on high quality customer-related data and enabled by information technology” [11]. Different definitions, however, are all grounded on four dimensions: customer identification, customer attraction, customer retention, and customer development [59, 53, 39]. As such, we can depend on the four dimensions for our study.

According to the survey in [49], data mining techniques work well when discovering the hidden knowledge in a mass of data, and they have been widely applied to CRM area. For example, classification model is used in [38] for customer segmentation and strategy development. In [61], researchers build a customer churn prediction model via Support Vector Machine. Based on previous literature, we apply data mining techniques to our research on e-commerce data.

In summary, this thesis aims to find a solution to solve above CRM-related questions with use of data mining techniques.

1.2 Introduction to eBay

Our research focuses on a case study with eBay online transaction and feedback data. Founded in 1995, eBay Inc. is a global e-commerce company that provides an online shopping platform. With the popularity of e-commerce, eBay was developing rapidly. According to the statistics from [5], from first quarter 2010 to fourth quarter 2014 the number of eBay active users exhibits a sustainable growth as shown in Figure 1.2. By the end of 2014, this value has reached up to 155.2 million. That critical mass of users contribute to a great deal of transaction volume. As shown in Figure 1.3 in the third quarter of 2015, the gross merchandise volume on eBay reached to 19.6 billion U.S. dollars. Due to the popularity of eBay, it would be convincing if empirical conclusions are yielded by analysis of eBay data.

eBay is primarily known for its auction-style sales in which sellers place bids and buyers compete in auctions. In the summer of 2002, eBay enables sellers to list their items at fixed prices without process of auctions, namely “Buy It Now”(BIN) service, which is like the business-to-customer mode on Amazon. According to a study on eBay data from 2003 to 2009 [24], a change in preferences towards convenient online

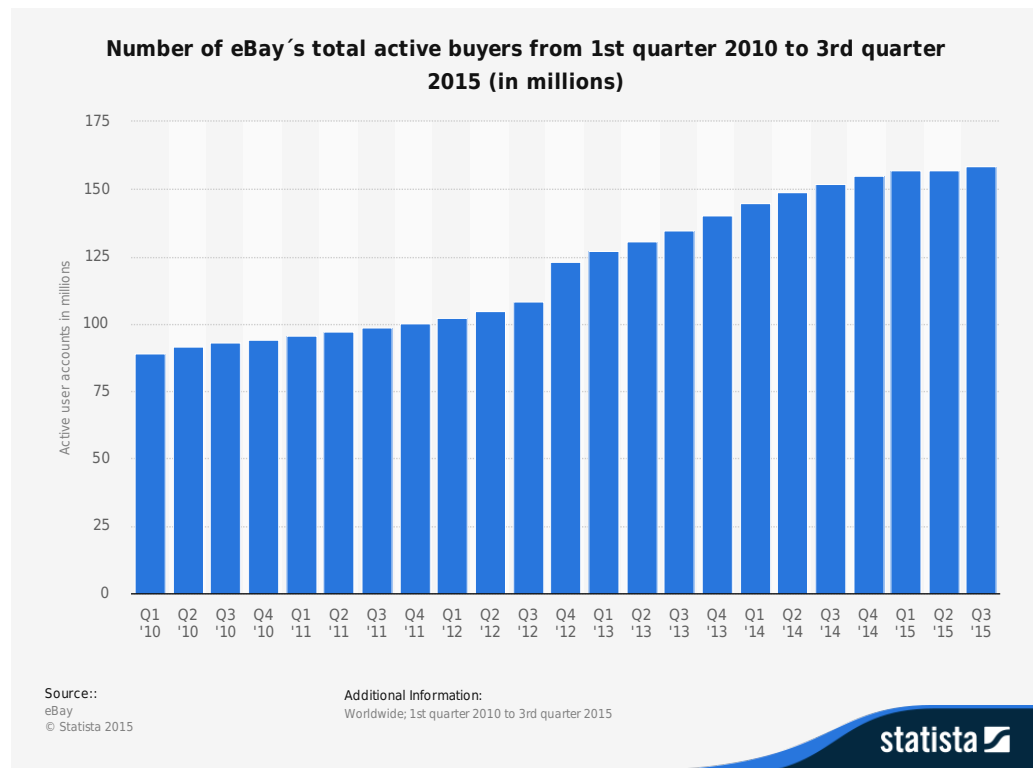


Figure 1.2: Number of eBay's total active buyers from 2010 to 2015 [5]

shopping drives a “dramatic shift of Internet retail from auctions to fixed prices”. Also eBay reported that in 2013 fixed price trading accounted for 70 percent of eBay's gross merchandise volume. Considering that shift, our study will focus on the eBay BIN data.

Feedback mechanism also plays an important role in the eBay community. eBay makes feedback profile for each eBay seller. For each transaction, the buyer can not only give overall feedback ratings but also rate the seller in four different areas: *Item as described*, *Communication*, *Shipping time*, and *Shipping and handling charges*. The multiple feedback ratings recorded by eBay feedback system make our study on feedback possible.

Overall, eBay provides robust trading and feedback systems that have fostered a large number of active sellers and buyers. More importantly, eBay offers public Application Programming Interface (API) that enables us to obtain its raw data. For these reasons, eBay would be an ideal platform for our empirical study.

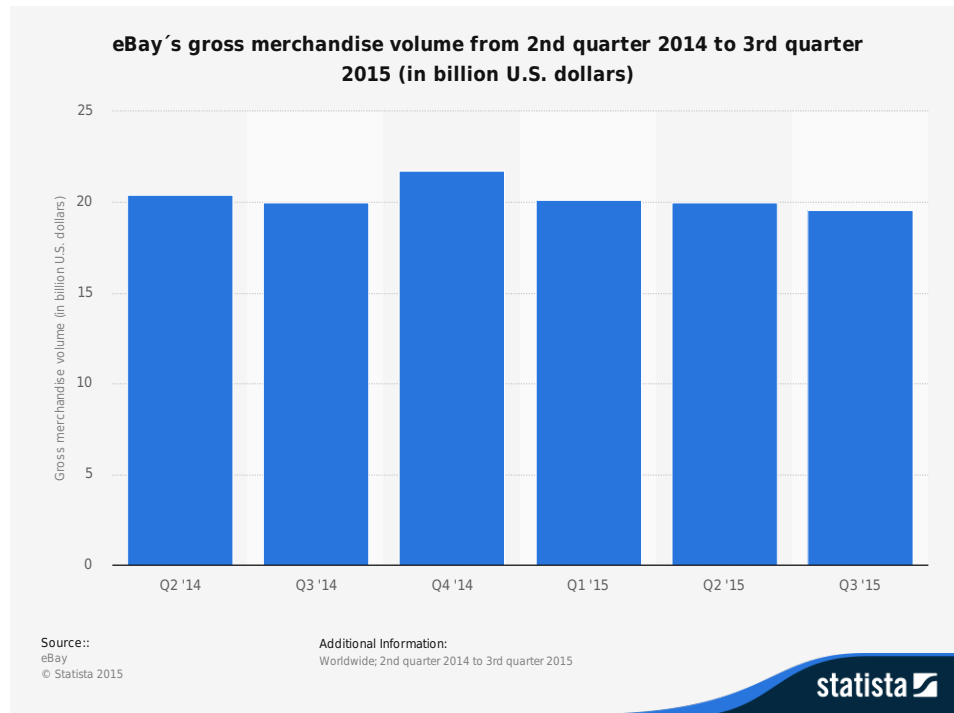


Figure 1.3: eBay’s gross merchandise volume from 2nd quarter 2014 to 3rd quarter 2015 (in billion U.S. dollars) [5]

1.3 Contributions

There have been many articles related to CRM. Nevertheless, with the rapid development of e-commerce, the behaviour and preference of customers are changing. For example, the majority of eBay users have changed their preference towards “Buy It Now” instead of auctions. Also online shopping systems are updating constantly. In the context of changed e-commerce environment, we perform a case study focusing on eBay CRM and make the following contributions.

1. Unlike the most articles that conduct case studies on eBay auctions, our case study is based on eBay “Buy It Now” that is now the most popular sales mode on eBay site.
2. We define the customer value as the combination of relative values of frequency and expenditure rather than the separate values that are often used in many previous literature. The customer value yielded in our study reflects the value of a buyer for a particular seller, which can help the seller find out potential

buyers.

3. We build a customer value distribution for each individual eBay seller. By observing the characteristics of the whole distributions, we propose four baseline distributions to capture different seller classes.
4. We classify sellers based on their customer value distributions into four baseline models, and then evaluate the performance of each group.
5. We examine the overall feedback ratings as well as the detailed seller ratings, in order to understand how a seller's reputation affects its customer value distribution.

1.4 Agenda

In the context of CRM, applying data mining techniques to analyze of eBay data, the main objective of this thesis is to provide online sellers with an approach to improving their relationship with customers from perspectives of customer identification and customer retention. A map of the remainder of this thesis is as follow:

Chapter 2 presents related work on customer value measure and examines application of data mining techniques in CRM.

Chapter 3 introduces eBay API for data collection and explains the data structure of eBay transaction and feedback data.

Chapter 4 proposes a model for measuring customer value and discusses the impact of weighted values in the model.

Chapter 5 describes the process of segmenting eBay sellers, which includes measuring customer value via RFM model, building customer value distribution for each seller, and classifying those distributions into four baseline groups, and comparing the performances of the groups.

Chapter 6 investigates whether feedbacks have impacts on the distribution of customer value.

Chapter 7 concludes the thesis.

Chapter 2

Background and Related Work

Among the four dimensions of CRM, this thesis focuses on two of them: customer identification and customer retention. In this chapter, we do literature reviews on the two dimensions and try to find out what data mining techniques are applicable to our study.

2.1 Customer Identification

To improve CRM, the first priority is to understand customers and identify potential customers. The essence of customer identification is measuring the values of customers. According to our investigation, methods of measuring customer value can be classified into a subjective way and an objective way. The subjective approach is commonly used to measure customer value from the attitudinal perspective. According to [12, 62], customer value can be defined as the emotional bond between customers and sellers and can be measured as “a form of subjective trade-off by asking the subjects”. Most research based on this definition and measurement was conducted by making questionnaire surveys or interviews. Customers are asked to give a level of overall satisfaction on a specific scale and their answers will be used to estimate their customer values. This subjective method is applied in many articles on CRM. For instance, Oh [52] obtains the customer value in the hotel industry and then explores the relationship between customer value and service quality and intentions to repurchase.

With the arrival of information age, researchers tend to utilize the objective method to quantify customer value by applying statistical analysis to the real e-

commerce dataset. In the quantitative method, Recency-Frequency-Monetary (RFM) analytic model is often used for customer value measure. It is a behaviour-based model used to analyse the purchase history of customers in the given database [31]. R, F, M denotes recency, frequency and monetary, respectively. They measure when the customer shops, how often the customer shops, and how much the customer spends in a store. As an extension, researchers proposed weighted RFM model in which relative importance weights are given to RFM variables. Liu et al. [44, 43] first employ analytic hierarchy process to give the weights to RFM, which was later used in [37, 29].

In addition to RFM, there are other ways to measure customer values in recent studies. Gale and Wood [26] discuss that the customer value can be calculated by a sum of weighted relative overall quality score and weighted relative price competitiveness score. Hwang et al. [33] calculate customer value from three perspectives: current value, potential value, and customer loyalty. They define current value as “a profit contributed by a customer during a certain period (for six months)” and potential value as a probability of cross-selling or up-selling. The customer loyalty is calculated by subtracting churn rate from one, where the churn rate represents the percentage of customers who quit from the membership list. Similarly, Yi et al. [66] define customer value as a combination of observed customer value and predicted customer value. Observed customer value represents the value of all contribution margins generated during the observation period. Predicted customer value is the predicted probability that the customer keeps active in a specified further period.

Based on the customer values estimated, customer segmentation can be performed in order to identify the most profitable segment of customers. The data mining technique that is most widely applied for this purpose is K-means clustering method. For instance, in [16, 18, 50, 29] authors applied K-means to customer segmentation according to the value of RFM for each customer. Other clustering methods, such as fuzzy clustering [63] and self-organizing map [32], are employed for this purpose as well.

2.2 Customer Retention

Customer retention is the core of CRM. According to [49] that summarizes the articles on customer retention published before 2006, its findings indicate that customer retention is the dimension that receives more attention than other three dimensions, and

Table 2.1: Distributions of Articles by CRM Elements and Data Mining Techniques

CRM elements	Data mining techniques	Objectives	References
Customer feedback	Association rule	Identifies positive comments based on the appearing frequency of positive keywords	Aguwa et al. [6]
	PageRank	Proposes a feature-based model for ranking online products by analysing customer reviews on Amazon	Zhang et al. [67]
	Regression	Discovers if negative publicity helps product awareness and then drivers purchase likelihood	Berger et al. [9]
	Visualization	Visualizes the features extracted from customer reviews on Amazon	Oelke et al. [51]
	Regression	Examines the impact of word-of-mouth volume on a movie's box office revenue	Duan et al. [23]
Customer loyalty	Regression, Random forests	Discovers if the sequence of customer's first purchase is a significant variable that affects the proposed model for partial churn detection	Migueis et al. [48]
	Regression	Discovers multivariate predictors of long-term behavioural intentions	Baumann et al. [8]
	Regression	Estimates the effect of product quality, service quality and brand image on the customer's willingness to recommend the retailer	Clottey et al. [19]
	Random forests	Applies proposed improved balanced random forests for customer churn prediction	Xie et al. [64]
	Regression	Uses the variables on customer behaviour to predict the strength of the relationship between a retailer and its customers	Buckinx et al. [10]

the articles on customer retention mainly discuss with respects to: customer loyalty and custom feedback. Resorting to the classification framework proposed by [49], we investigate the paper published since 2007 and classify them by their applications of data mining techniques, which are listed in the Table 2.1. We find that the customer retention is still a popular research topic in recent years, and data mining techniques have been widely applied to customer retention. For instance, regression models are frequently used in this area for estimating the effect of variables that are related to customer behaviour.

2.3 eBay Data Analytics

2.3.1 Academic research

As one of most well-known e-commerce websites, eBay is often treated as the research object in the area of CRM. Here we summarize the articles studying based on eBay transaction data and feedback data.

Focusing on eBay transaction data, Jank et al. [35, 34] measure the loyalty of each bidder-seller pair by the number of repeat purchases from the seller, and then construct loyalty distribution for the seller. By performing regression analysis, they find that the loyalty distribution has a strong impact on the outcome of an auction.

With the addition of eBay feedback data, researchers examine the impact of eBay overall feedback ratings on sales performance [15, 65, 13] and auction price [30, 47, 58]. For example, authors in [13] look into the impact of negative feedback ratings on the sales rate. They find that once a seller receives negative feedback, its weekly sales rate declines. In addition, the regression analysis result from [47] reveals that eBay overall positive feedback ratings are statistically positive associated to auction prices but the impact is small.

While eBay data analytics has been explored extensively, limitations of previous research exist. First, all the articles on eBay analysis we have reviewed focus on eBay auction data. However, the majority of eBay users have shifted from auction to “Buy It Now” but few study concerns this change. What’s more, the paper studying on customer feedback ratings only deals with overall ratings (i.e. positive and negative ratings) but ignores eBay detailed ratings that are important indicators for customers’ satisfactions on item description, communication, shipping time and shipping charges.

2.3.2 Industrial Application

With the increasing number of online merchants and consumers, a variety of applications have been developed for online sellers with aim to make their business more efficient. Based on our findings, current eBay tools and services on the market could be roughly categorized as follows:

1. Inventory and Order Management
2. Finance and Accounting Management
3. Shipping Management
4. Multi-channel Management
5. Customer Relationship Management
6. Reporting and Analytics

Table 2.2: Applications for eBay Customer Relationship Management

Tools	Features
Brightpearl	<ul style="list-style-type: none"> - Create a customer record - Save a customer’s demographic information - Keep track of a customer’s order, scheduled appointment, calendar activities
eBay SuperStore CRM	<ul style="list-style-type: none"> - Record customer purchase information - Categorize customers based on their purchase patterns
Simple CRM for eBay	<ul style="list-style-type: none"> - Keep track of customers - Email marketing that helps sellers tailor their campaigns to the specific interests of each customer
Ki Feedback	<ul style="list-style-type: none"> - Automate feedback actions - Feedback comment analysis - Buyer blacklist management
Feedback Pro	<ul style="list-style-type: none"> - Remind customers to leave feedback - Remind sellers when they receive negative feedback - Automated “Thank You” message - Buyer blacklist management

Among these types of analytics on eBay dataset, what we are most interested is CRM. Today there are a wide variety of professional e-commerce tools built for CRM

on eBay. Table 2.2 displays five customer support applications that are often utilized by eBay sellers. These tools mainly aim to help sellers manage their customer profiles, purchase history and to deal with the feedback from the buyers. They offer sellers data management and simple statistics rather than comprehensive data analysis, so it is not helpful for sellers to discover hidden patterns and constructive information. The lack of knowledge discovery in applications also motivates us to have an insight analysis on eBay data.

Chapter 3

EBay Data Collection

In this chapter, we introduce the dataset that we use for our thesis study. The data was collected from eBay in two different ways: collected and provided by Terapeak, an e-commerce company that supports eBay and Amazon sellers with powerful data analytical tools to grow their businesses, and collected by ourselves with eBay API calls.

3.1 eBay Transaction Data from Terapeak

Terapeak Inc. is a provider of eBay market research for sellers. With the help of Terapeak, we were able to access to a dataset consisting of all the transactions on the eBay Canada site from April 2013 to June 2013. The raw data is in XML format, and each transaction can be uniquely identified with detailed information on seller, buyer, product, and shipping.

To extract the values useful for our research, we trim down the dataset by filtering out unnecessary fields but keeping the necessary values, such as the transaction id, buyer id, seller id, quantity, price, category id, and date. Since in our research we need to build customer value distributions based on their RFM values for each seller, we ignore sellers who had less than three transactions in the dataset. This omission allows us to only focus on active sellers and draw meaningful conclusion from our study. After the above data pre-processing, our dataset includes 985,605 successfully completed transactions that involves 4,736 stores, 808,852 buyers, and 1,349,660 products over 35 categories. Table 3.1 displays the statistics of the sellers in the resulted transaction dataset.

Table 3.1: Statistics on eBay Sellers

	Min.	Max.	Mean	Std. Dev.
Num. of transactions	3	53,437	26.74	382.93
Num. of buyers	3	39,970	30.79	317.53
Revenue	0.03	1,686,141.05	1,358.52	10,162.90

From a seller’s point of view, the seller mainly cares about the number of transactions, the number of its buyers, and the total value from the transactions. From the table, we can see that on average an eBay merchant has about 31 customers within the period we studied. The standard deviation of the number of transactions and the standard deviation of transaction values are both high, indicating that there is a wide variation between stores on eBay.

3.2 Data Collection with eBay API

From the Terapeak dataset, we are able to build a network between sellers and buyers. Nevertheless, the dataset does not include all information needed for our study. In particular, we need information regarding the merchants and customers’ feedback. For this purpose, we leverage the eBay Trading API and make *getUser* and *getFeedback* calls to obtain the relevant information.

The eBay API is provided by the eBay Developers Program [2]. With the API, users can access the eBay database and perform particular tasks by sending API calls. Currently, the eBay API has a wide variety of functionalities, such as *Finding*, *Shopping*, and *Trading*. In our research, we only need to make *Trading* API calls to obtain the information about the property and feedback of a specific eBay store.

The process of making an eBay API call is shown in Figure 3.1. First of all, we need to apply for and get an authentication token on the website of eBay Developer Program. The eBay API server will check the unique token when our program tries to make a connection. Upon successful authentication, our program is allowed to submit requests and receive responses. The requests are written in XML format and are sent via the HTTP POST calls. When the eBay server receives a request, it sends back the response in XML.

Trading API consists of various calls that return details about transactions and users on eBay. In our research, we mainly use *getFeedback* and *getUser* calls to obtain

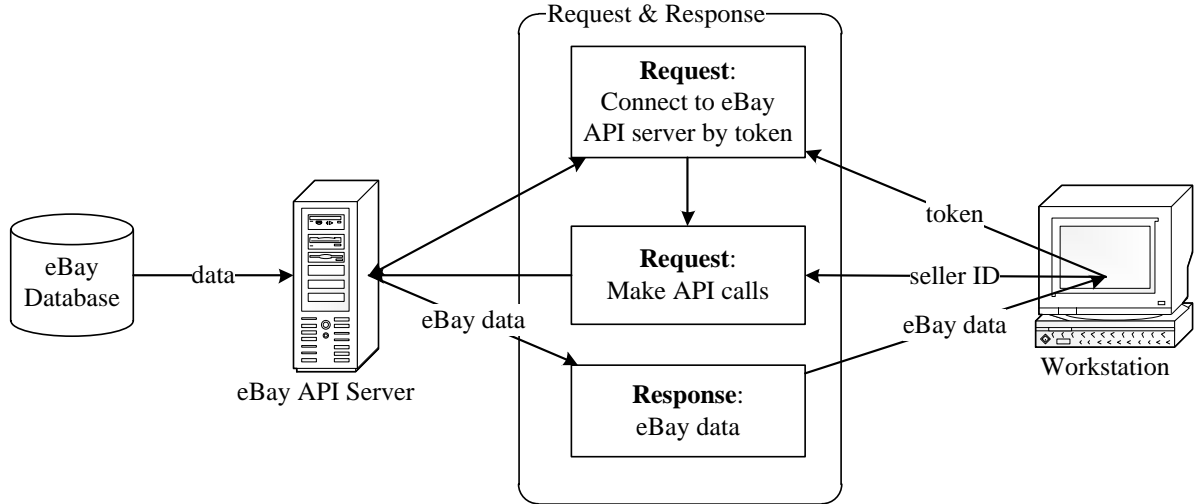


Figure 3.1: Framework for eBay data collection

information for our further analysis.

The *getFeedback* call retrieves the accumulated feedback regarding an eBay store. With the input of a store name, the call returns all the feedback data for the store. The feedback contains the information about how consumers in that specified store rate their shopping experiences. The feedback is a good indicator of the reputation of a store. Figure 3.2 displays an example of *getFeedback* request in XML format. We use the names of sellers in our transaction dataset as input and then extract the information of their features and feedback. The *getUser* call fetches the property and performance of eBay sellers. Note that the response from *getFeedback* and *getUser* calls contains hundreds of fields. We, however, only extract the ones shown in Figure 3.2 that are required by our research.

```

<?xml version="1.0" encoding="utf-8"?>
<GetFeedbackRequest xmlns="urn:ebay:apis:eBLBaseComponents">
  <RequesterCredentials>
    <eBayAuthToken>AgAAAA**AQAAAA**aAAAA**V...</eBayAuthToken>
  </RequesterCredentials>
  <UserID>SellerID</UserID>
  <DetailLevel>ReturnAll</DetailLevel>
</GetFeedbackRequest>

```

Figure 3.2: Request for feedback data on a specific seller ID

Table 3.2: Selected Output Fields

Fields	Applicable values	Description	
Fields from GetFeedback			
UniquePositiveFeedbackCount	>=0	Total # of Positive Ratings	
UniqueNegativeFeedbackCount		Total # of Negative Ratings	
UniqueNeutralFeedbackCount		Total # of Neutral Ratings	
Item as described.Rating	1-to 5-star scale	eBay Detailed Rating System (More details are presented in Chapter 6)	
Communication.Rating			
Shipping time.Rating			
Shipping and handling charges.Rating			
Item as described.RatingCount	>=0		
Communication.RatingCount			
Shipping time.RatingCount			
Shipping and handling charges.RatingCount			
Fields from GetUser			
SellerLevel	None, Bronze, CustomCode,Silver, Gold, Platinum, Titanium		Five powerseller levels based on annual sales or number of transactions
TransactionPercent	[0.0, 100.0%]	Successful transactions over the entire transactions	

3.3 Data Storage

The data provided by Terapeak and the data obtained from eBay API calls are in XML format. To ease data processing, we use SQLite, a lightweight and fast SQL database engine, and import the information in the XML files into an SQLite database. The tables we created in our database are as follows:

- **TransSeller**: This table only stores each transaction id in the dataset provided by Terapeak and its associated seller id.
- **Transaction**: This table stores the details of each transaction that appears in the “TransSeller” table. Each row in the table represents a unique transaction.
- **Seller**: This table stores the properties and feedback information of the sellers who appear in the “TransSeller” table. Each row represents a unique seller.

Figure 3.3 illustrates the data structure of the three tables created.

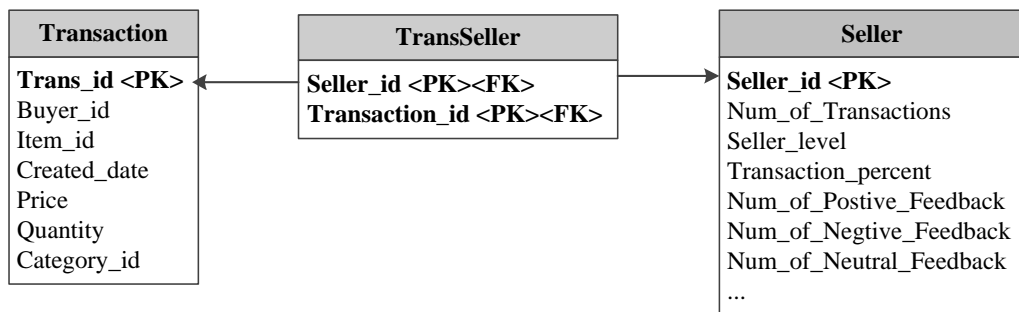


Figure 3.3: Database structure

Chapter 4

A Measure of Customer Value

4.1 Measuring Customer Value

Recency, Frequency, and Monetary (RFM) [31] are metrics commonly used to evaluate the value of a customer in a certain period of time, which are defined as follows:

- **Recency (R)** - the last date that the customer purchased in a specific store during that period.
- **Frequency (F)** - the number of times that the customer purchased in a specific store during that period.
- **Monetary (M)** - the total money amount that the customer spent in a specific store during that period.

In our study, we modify the above RFM model based on the characteristics of our dataset and our research objectives. According to the literatures [20, 25, 16, 14, 37], the time periods of data to build a RFM model are at least one year. All transactions in our dataset, however, occurred only within a period of three months, from April 2013 to June 2013. In this case, the Recency value does not disclose much information regarding the value of a customer. As such, in our study we ignore the R value and mainly focus on the values of Frequency and Monetary. Furthermore, as introduced in Chapter 2, the customer value can be evaluated from three aspects [33], the current value, the potential value, and the customer loyalty. Since there is no consensus on the definition and evaluation of customer loyalty in the context of e-business, we evaluate the value of a customer from the points of current value and potential value, with

the current value captured by the M (Monetary) measure and the potential value (implicitly) by the F (Frequency) measure.

To have a consistent evaluation, we normalize the value of a customer as one unit. We assume that the unit value of a customer is distributed into the online stores where the customer ever shopped¹. For instance, given a specific buyer i and a set of n stores in a certain category of interest, the value of buyer i for seller j , is denoted as V_{ij} . If we denote the total unit value of buyer i as V_i , we have

$$V_i = V_{i1} + V_{i2} + \dots + V_{ij} + \dots + V_{in} = 1. \quad (4.1)$$

In this thesis, V_{ij} is evaluated based on the weighted values of F and M, i.e.,

$$V_{ij} = \alpha F_{ij} + (1 - \alpha) M_{ij}, 0 \leq \alpha \leq 1, \quad (4.2)$$

where α is the weight parameter, F_{ij} and M_{ij} represent the proportion of times and the proportion of expenditures, respectively, that buyer i purchased from seller j within this buyers' whole purchases in the given category. To be more specific,

$$F_{ij} = \frac{f_{ij}}{\sum_{j=1}^n f_{ij}} \quad (4.3)$$

$$M_{ij} = \frac{m_{ij}}{\sum_{j=1}^n m_{ij}}, \quad (4.4)$$

where f_{ij} represents the number of times that buyer i purchased from seller j , m_{ij} represents the expenditure of buyer i on seller j during the period in consideration. Note that the n stores all belong to the same category. The purchase occurring in stores of other categories will not count, because it makes more sense to compare the performance of stores in the same category. The parameter α in Equation (4.2) determines the weights allocated to F or M.

We use an example to show how our method works. Assume that there are three stores in a certain category. Assume that during the period of investigation, a buyer shops in the three stores 2 times, 3 times, and 5 times, respectively. The expenditures are \$10, \$50, and \$40, respectively. Table 4.1 shows the results of customer values when the weight α is 0.5. Examining the results, we see that the customer values

¹This assumption does not necessarily mean that the customer does not shop in other stores not included in the dataset. Nevertheless, there is no enough information to infer the value and behavior of the customer regarding those stores.

Table 4.1: An Example of Calculation of Customer Value

	Category id	Category_0			
	Seller id	S_11	S_12	S_13	Total
Buyer_1	F	2	3	5	10
	M	10	50	40	100
	Relative F	$2/10 = 0.2$	0.3	0.5	1
	Relative M	$10/100 = 0.1$	0.5	0.4	1
	CV($\alpha = 0.5$)	$(2/10)0.5 + (10/100)(1 - 0.5) = 0.15$	0.4	0.45	1

reflect how a specific buyer distributes its value to different sellers.

Using the above method, we calculate the customer values in every store in each category. There are 35 categories in total.

4.2 The Impact of Different Weight Values

In the previous section, we build a model to estimate customer value, whose calculation depends on a weight parameter α as shown in Equation (4.2). In this section, we vary the value of α to investigate the impact of different weights on the F and M measures.

Using the eBay data, we draw the histogram of number of buyers vs. customer values, by changing the value of α from 0 to 1 with step size 0.1. Given an α value, we use Equation (4.2) to calculate the customer value of buyer i for seller j . Given a customer value x in Figure 4.2, the average number of customers is calculated as the total number of customers whose value is x (for some sellers) over the total number of sellers. To gain more information, we also draw the average and the standard deviation of number of customers in Figure 4.3.

From the two figures, we can observe that the histograms are similar, implying that the impact of α value on the distribution of customer values is not significant. By plotting the scatter plot of the F value and the M value, we observe strong lower tail dependence and stronger upper tail dependence between F and M values, as shown in Figure 4.1. This phenomenon explains the similar histograms shown in Figures 4.2 and 4.3.

As a final remark, although the above study suggests that our model is robust on various α values, in our later investigation we select the best α value that minimizes

the total discrepancy between benchmark and actual distributions of customer values. More details will be disclosed in Chapter 5.

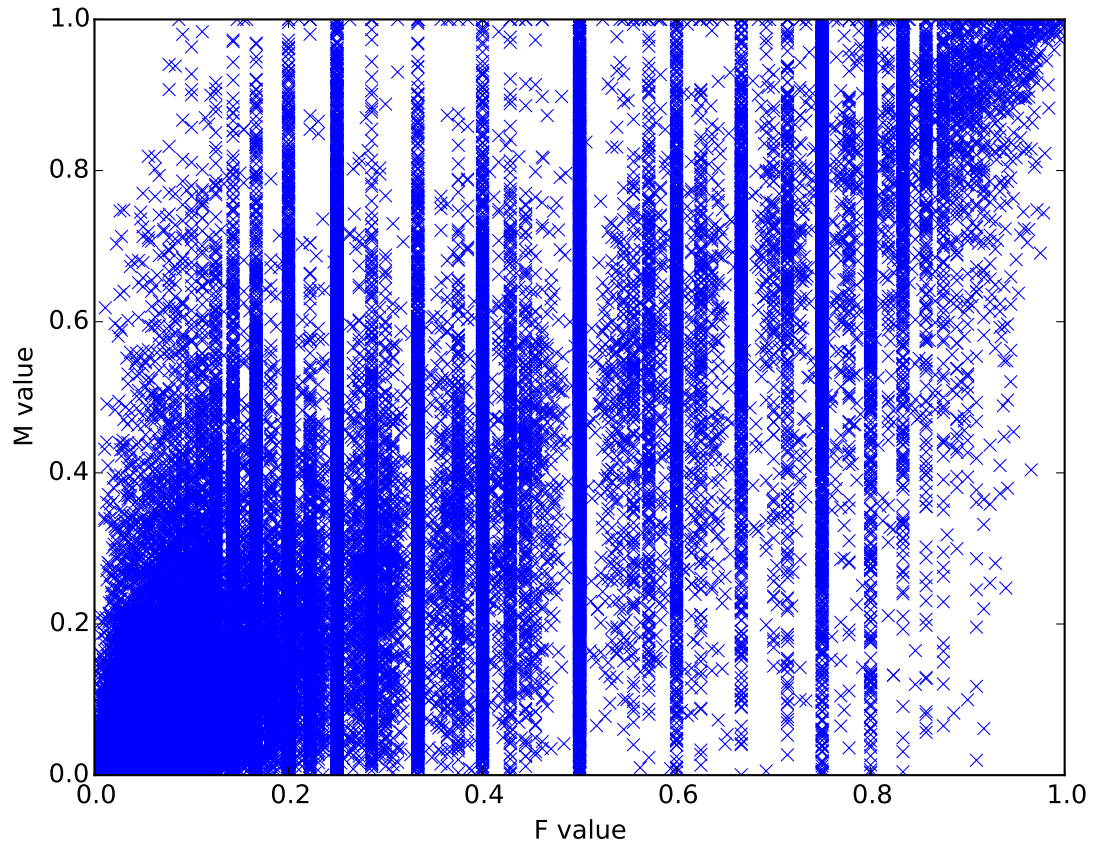


Figure 4.1: Scatterplot of F and M values

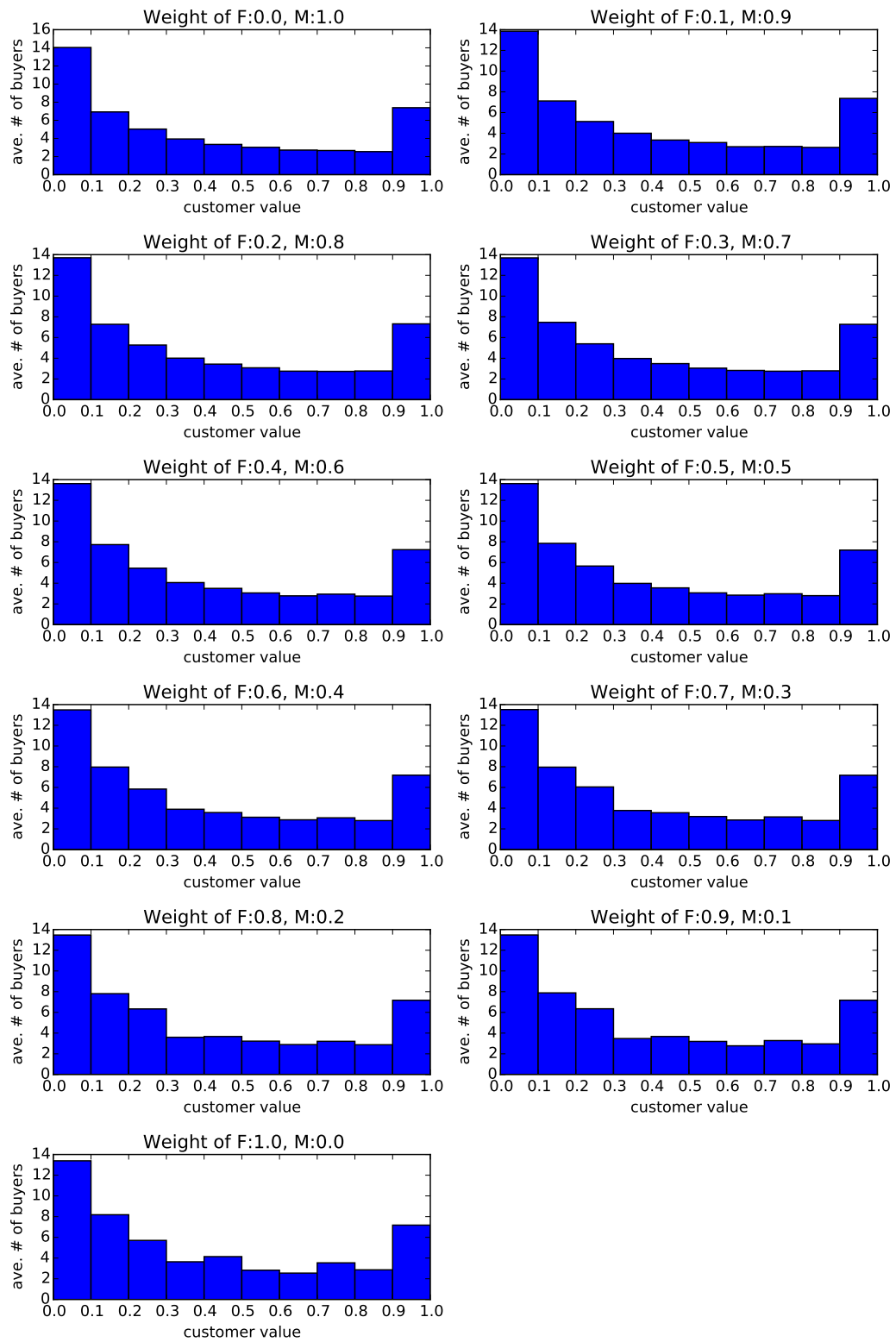


Figure 4.2: Comparison on average number of customers

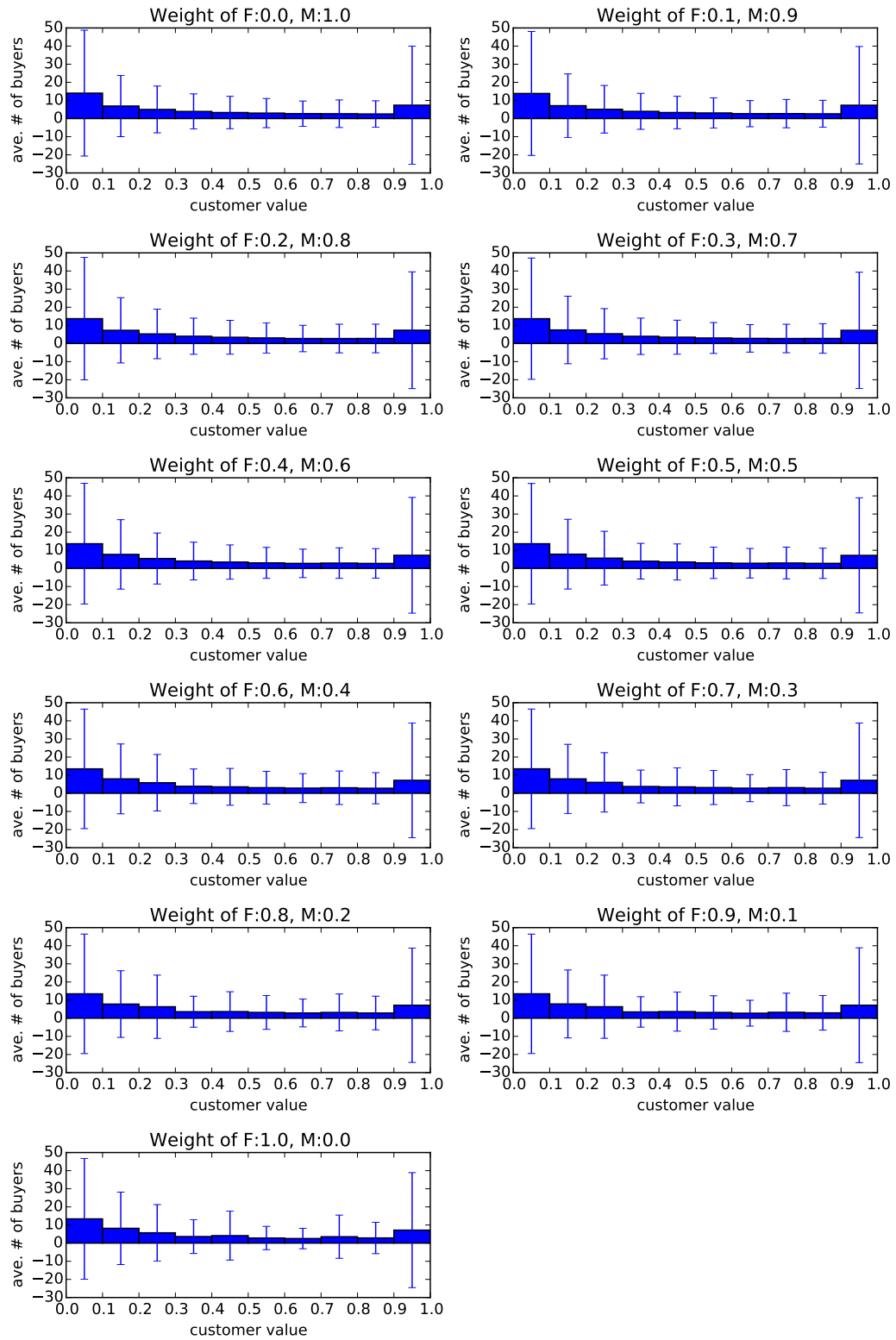


Figure 4.3: Comparison on average and standard deviation of number of customers

Chapter 5

Seller Segmentation Based on Distribution of Customer Values

5.1 Distribution of Customer Values

Given a particular online seller, the method introduced in Chapter 4 can be applied to calculate the values of its customers. The distribution of customer values provides the seller with critical information to learn the behavior of its customers.

As a concrete example, we select an eBay seller who has a total of 227 buyers in the dataset. Figure 5.1 shows the distribution of these buyers' values, calculated with the weight parameter α set as 0.5. From this figure, the seller can conclude that over 70% of its customers have a high value. Corresponding marketing strategies could be made based on this important information.

We have calculated the distribution of customer values for every seller included in our dataset. Next, we cluster the sellers based on the distributions of their customers' values, i.e., sellers having similar distributions of customer values are clustered together.

5.2 Four Baseline Distributions

There are many clustering algorithms such as k-Nearest Neighbor(k-NN) and K-means. Nevertheless, the clustering results should have intuitive and practical meaning, and due to this consideration, we first need to determine a baseline model that possesses practical meaning and is justifiable with real-world data.

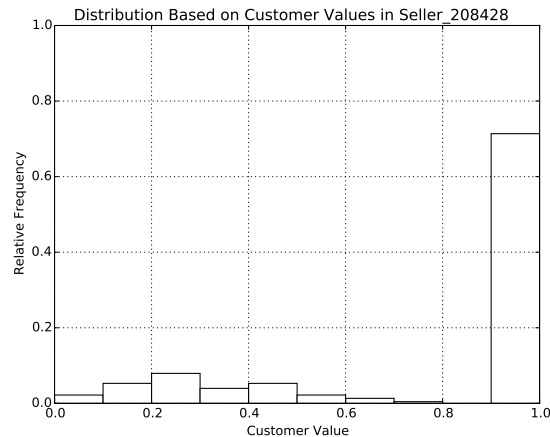


Figure 5.1: Distribution of the values of a seller’s customers. (Seller ID: 208428)

In recent research, there have been models proposed to reflect customer distributions. For instance, existing work [35, 34] presents five distributions, as shown in Figure 5.2, which are built with eBay auction data and are based on customers’ relative purchase frequency, i.e., the F value in our study. We can see that the five distributions have distinct characteristics. In particular,

- The distribution “Pure Loyalty” means that all customers have a high loyalty scores (0.8 ~ 1.0).
- The distribution of “Pure Disloyalty” means that all customers have a low loyalty scores (0.0 ~ 0.2).
- The distribution “Strong Loyalty” indicates that up to 70% of customers show high loyalty.
- The distribution “Somewhat Loyalty” means that up to 70% of customers have medium loyalty scores (0.4 ~ 0.6).
- The distribution “Two Extremes” means that half of the customers have high loyalty scores while the other half have low loyalty scores.

Since the above distribution model is based on eBay auction data, we first need to validate if our eBay transaction data (“Buy It Now” transactions) shows a similar pattern before we can use the model. To this end, we first perform clustering analysis. Figure 5.3 presents the centroids of four clusters yielded by K-means clustering.

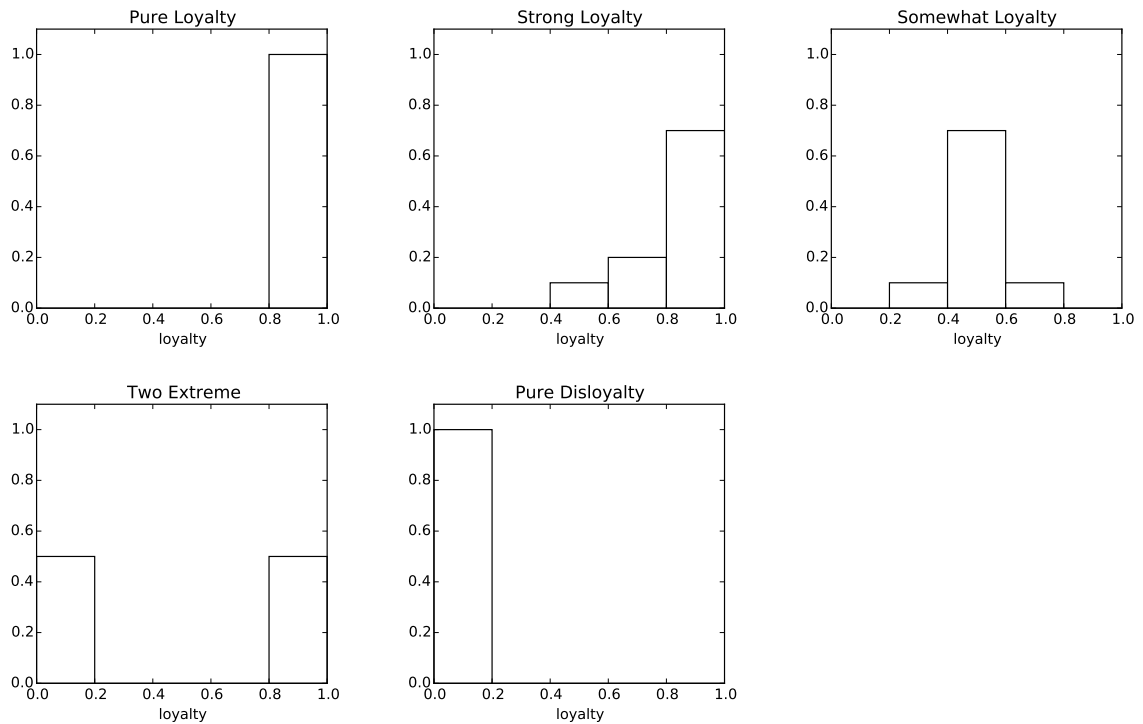


Figure 5.2: Five loyalty distributions proposed in [35, 34]

Among the 4,736 sellers, there are 1,488 and 1,843 sellers whose distributions are closest to the centroid 1 and centroid 2, respectively. Both centroids are approximate to the uniform distribution in which customer values are evenly spread across the range of the distribution. The other two centroids are skewed. Centroid 3 is a left-skewed distribution in which half of customer values are high (0.8~1.0), while centroid 4 is right skewed. The clustering results suggest that our data does not show similar patterns in the eBay auction data. To be specific, very few sellers have the distribution of customer values close to “Pure Loyalty”, “Pure Disloyalty” and “Two Extreme” distributions. In contrast, a large number of sellers whose distributions have no clear peaks and are close to the uniform distribution.

In conclusion, the traditional model developed with online auction data is not suitable to capture the pattern of customer value distributions in eBay “Buy It Now” transaction data. Due to this reason, we modify the existing models by ignoring the “Two Extremes”, “Pure Loyalty”, and “Pure Disloyalty” distributions, since we do not observe a seller have the above features in our data. We lower the peaks of “Strong loyalty” and “Somewhat loyalty” to better reflect the characteristic of our data. We

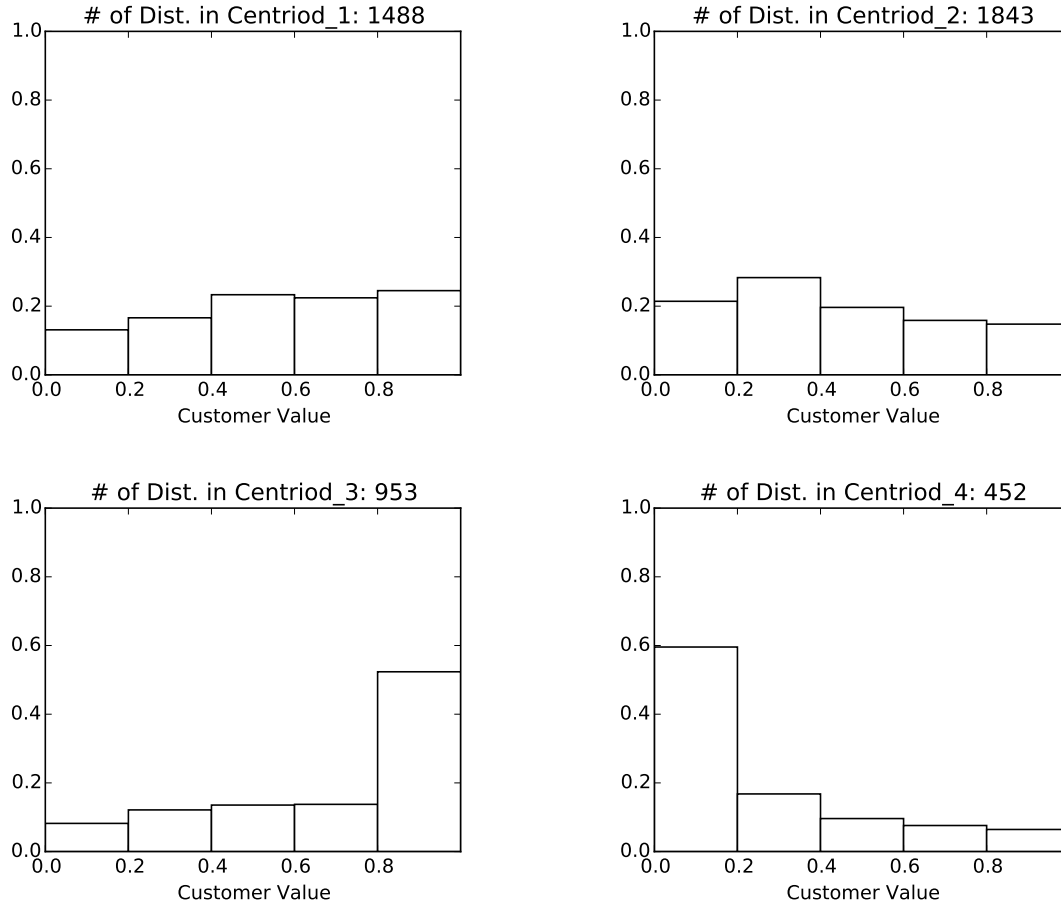


Figure 5.3: Centroids of four clusters yielded by K-means

then add the uniform distribution to model the sellers who have approximately uniform distributions of customer values. Finally, our proposed baseline distributions are shown in Figure 5.4.

These baseline distributions are denoted as High (H), Medium (M), Low (L) and Balanced (B) distributions, respectively. The baseline distributions capture the statistical features of eBay sellers and in the meantime have clear practical meanings as follows:

- **H**: it indicates that the majority of the store’s customers have high customer values (i.e, $0.6 \sim 1.0$). The high-valued customers are regular shoppers in the store. Compared to other buyers, high-valued customers shopped more often and the amount of money they spent is larger.
- **M**: it indicates that the majority of the store’s customers have medium customer

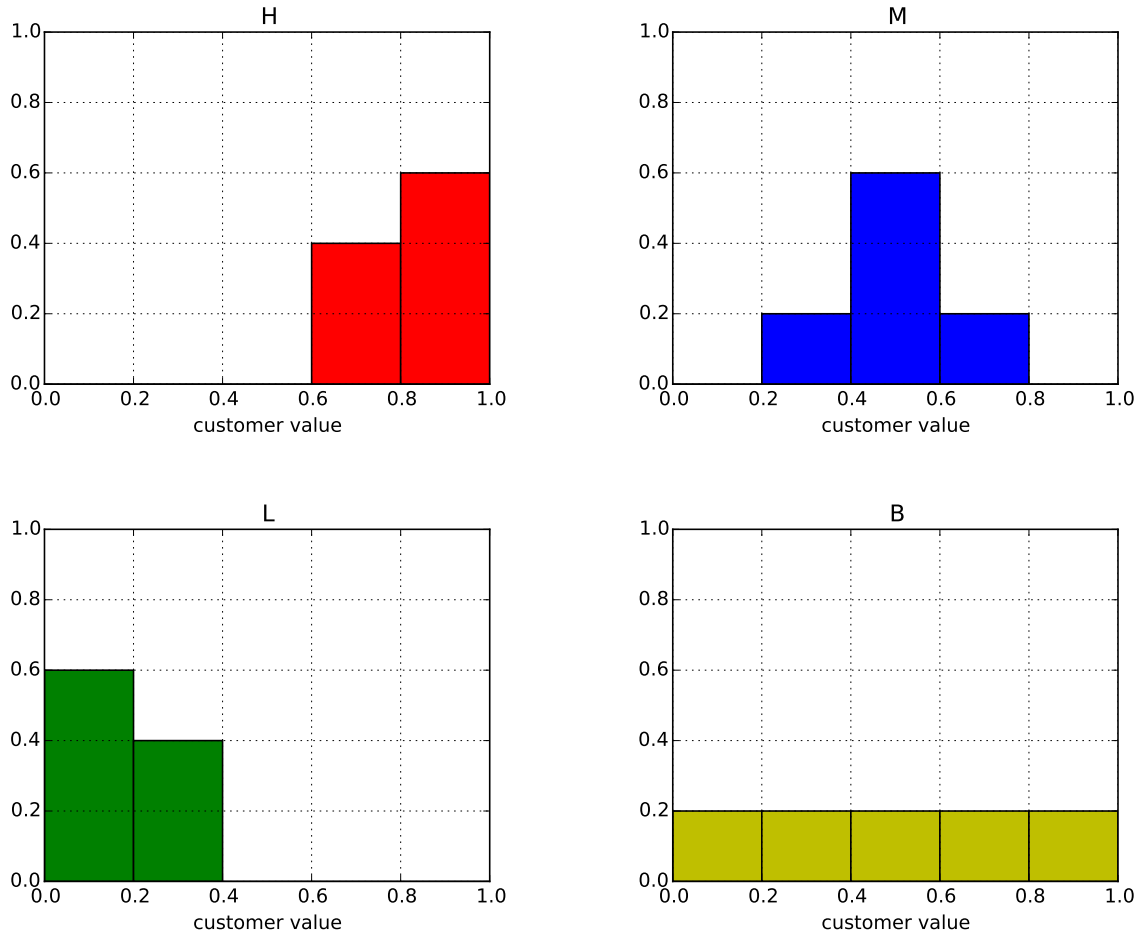


Figure 5.4: New baseline distributions

values (i.e, $0.4 \sim 0.6$). In addition, the store may have fewer shoppers having either much lower or much higher values.

- **L:** it indicates that the majority of the store’s customers have low customer values (i.e., $0.0 \sim 0.4$). In other words, the store does not have “regular” shoppers who repeatedly purchase merchandise from the store.
- **B:** it indicates that the store’s customers are diverse, with customer values evenly split in the low, medium, and high ranges.

Accordingly, we classify eBay merchants based on the above baseline distributions into four classes, labelled as C_H, C_M, C_L, C_B , respectively.

5.3 Categorization of eBay Merchants

5.3.1 Methodology

With the four baseline distributions introduced above, we next classify the eBay sellers. To this end, we first need to measure the difference between two distributions. We adopt a widely-used Kullback-Leibler Divergence (KLD) to measure the difference between two probability distributions. Given two probability distributions P and Q , the KLD of Q from P is defined as:

$$D_{KL}(P \parallel Q) = \sum_V P(V) \log \frac{P(V)}{Q(V)}, \quad (5.1)$$

where V denotes the range of customer values. In our context, we use KLD as the distance measure, i.e., $Dist(P, Q) = D_{KL}(P \parallel Q)$, where $P(V)$ and $Q(V)$ represent the probability that the customer value is within a given range. For example, $P(0.0 \leq V \leq 0.2)$ refers to the probability that the buyers' values are between 0 and 0.2.

Note that KLD is a non-symmetric measure. For consistence, we calculate $D_{KL}(P \parallel Q)$ where P represents the empirical distribution of customer values and Q represents the baseline distribution. Also note that Equation (5.1) makes sense only if $Q(V) > 0$ for all V . However, there are cases where some probability values in the baseline distributions are zero. To avoid the problem, we adjust those probability distributions by changing $Q(V)$ to:

$$Q(V) = \begin{cases} Q(V) + \epsilon & \text{if } Q(V) = 0 \\ Q(V) - n\epsilon Q(V) & \text{if } Q(V) > 0, \end{cases} \quad (5.2)$$

where ϵ denotes an arbitrarily small quantity and n represents the number of occurrence of zeros¹ in the distribution. This smoothing operation causes minor changes in our baseline distributions, as shown in Table 5.1. The above step is called “de-zero” processing. Figure 5.5 shows the histograms of adjusted baseline distributions.

After above “de-zero” processing, our probability distributions can fit into Equation (5.1). With the KLD measure, we obtain four distance values for each eBay merchant, $Dist(P, H)$, $Dist(P, M)$, $Dist(P, L)$, $Dist(P, B)$, representing the distances between the store's empirical customer value distributions and the baseline distribu-

¹Note that while in theory a customer's value could be any real number, we in this thesis only consider the discrete values of with step size 0.1. This granularity is fine enough for our case study.

Table 5.1: Value Changes in Baseline Distributions ($\epsilon = 0.05$)

Baseline		0.0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
H	Q(V)	0.0	0.0	0.0	0.4	0.6
	de-zeros	0.05	0.05	0.05	0.34	0.51
M	Q(V)	0.0	0.2	0.6	0.2	0.0
	de-zeros	0.05	0.18	0.54	0.18	0.05
L	Q(V)	0.6	0.4	0.0	0.0	0.0
	de-zeros	0.51	0.34	0.05	0.05	0.05
B	Q(V)	0.2	0.2	0.2	0.2	0.2
	de-zeros	0.2	0.2	0.2	0.2	0.2

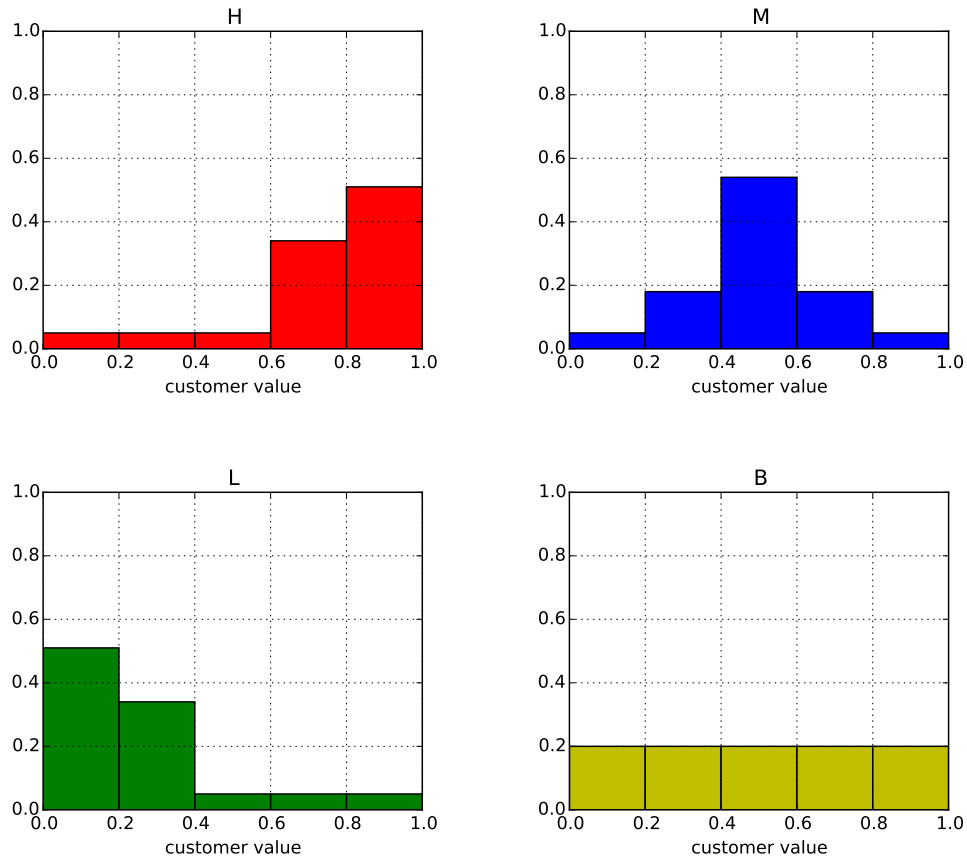


Figure 5.5: Baseline distributions after de-zero processing

tions H, M, L, B , respectively. Assume that

$$Y = \arg \min_{Q \in \{H, M, L, B\}} Dist(P, Q), \quad (5.3)$$

where P represents the merchant's empirical distribution of customer values and Q refers to one of the four baseline distributions H, M, L, B . The store is then classified into class C_Y , where $Y \in \{H, M, L, B\}$.

5.3.2 Parameter Tuning

In the previous chapter, we have demonstrated that the change of weight parameter α does not have a significant impact on the calculation of customer values. Nevertheless, to find the best model, i.e., a model that has the minimum total difference between empirical distribution and the baseline distribution, we search for an α value that can minimize the total distance between the baseline distributions and empirical distributions of customer values. Our search algorithm works in the following steps.

- **Step 1:** We vary the value of α from 0.0 to 1.0 with step size 0.1. With each α value, we build an empirical model that includes the customer value distributions of all sellers. In total, we thus have 11 empirical models, each corresponding to an α value.
- **Step 2:** For each model, we measure the total distance between the empirical distributions and the baseline distributions. To be specific, we calculate:

$$\sum_{s \in \mathbf{S}} \min_{Q \in \{H, M, L, B\}} \{Dist(P_s, Q)\} \quad (5.4)$$

where P_s denotes the empirical distribution of customer values of seller s and \mathbf{S} denote the set of all sellers.

- **Step 3:** Repeat Step 2 over all the 11 models. Each time, we obtain a total distance value with Formula (5.4). The results are shown in Figure 5.6.
- **Step 4:** Select the α value that results in the smallest overall distance between the empirical distributions and the baseline distribution. As shown in Figure 5.6, the value of α should be set as 0.6. This value will be used for our following studies.

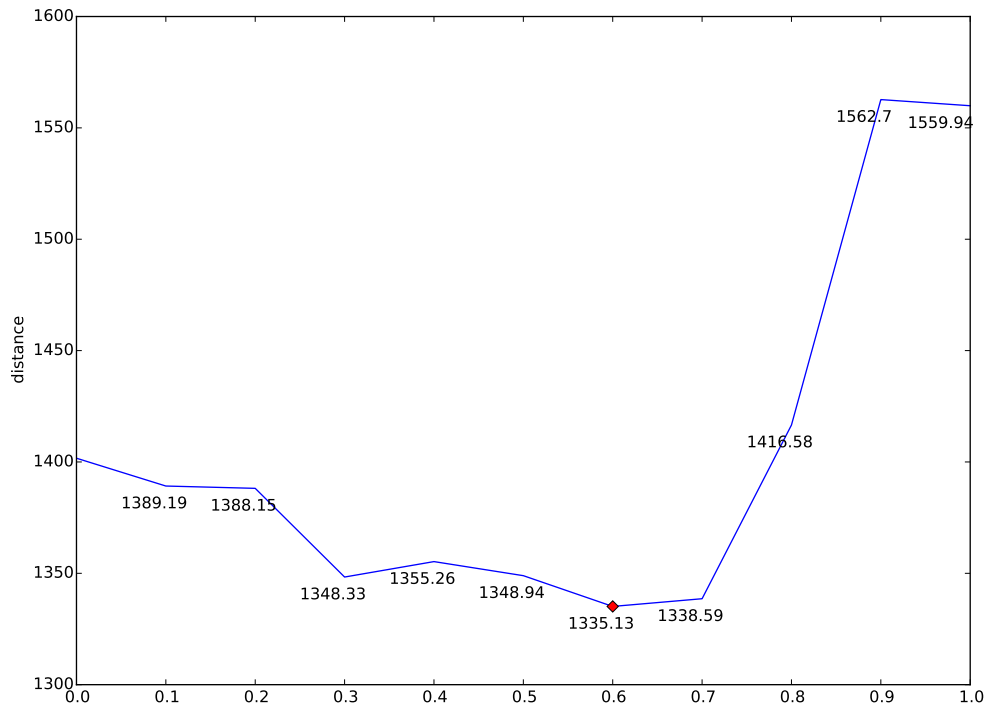


Figure 5.6: KL distances with change of weight (α)

5.3.3 Results of Categorization

With the selected α value, we categorize the distributions of customer values. The result of classification is shown in Figure 5.7. It can be seen that the cluster that has the largest proportion of sellers is C_B , with 41.47% of the total sellers. The remaining sellers are approximately evenly classified into C_H, C_M, C_L and each of them accounts for around 19%. This classification result is consistent with the result of K-means presented in the previous section, suggesting that the proposed baseline distributions capture the main features of the sellers well.

5.4 Comparison of Sellers in Different Categories

5.4.1 Number of Transaction

After categorization, our next goal is to compare the performance of sellers in different classes, C_H, C_M, C_L, C_B .

We first evaluate the number of transactions. The comparison results are shown

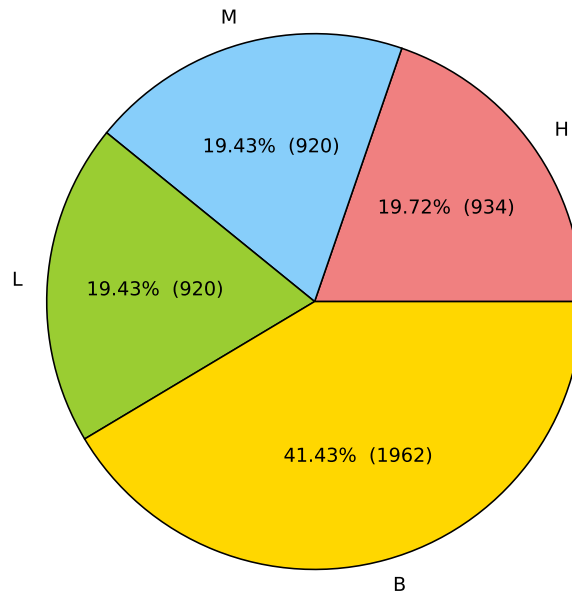


Figure 5.7: Result of classification with $\alpha = 0.6$

in Table 5.2, where the mean value is calculated by the total number of transactions over the total number of sellers in a cluster. From the results, we can see that the customers in C_H have the highest values with respect to mean and median number of transactions. In this group, sellers have 473 transactions on average and half the sellers achieve up to 124 transactions. The sellers in C_B also display good performance with around 430 transactions on average and 109 median transactions. In contrast, the average number and median of transactions associated with the sellers in C_M and C_L are much smaller.

Table 5.2: Number of Transactions in Different Classes

	mean	median
H	472.758	124
M	261.720	70
L	199.912	41
B	429.987	109

5.4.2 Percentage of Successful Transactions

In addition to the number of transactions, the percentage of successful transactions is another important measure for a seller's performance. With eBay API, as introduced in Chapter 3, we obtain this information for each seller in our dataset. Given a specific eBay seller, eBay API can return the field of `<TransactionPercent>` that contains the ratio of number of successful transactions over the total number of transactions associated with the seller. With the percentage of successful transactions of each seller, we can draw the distribution of the percentage values in each cluster, as shown in Figure 5.8.

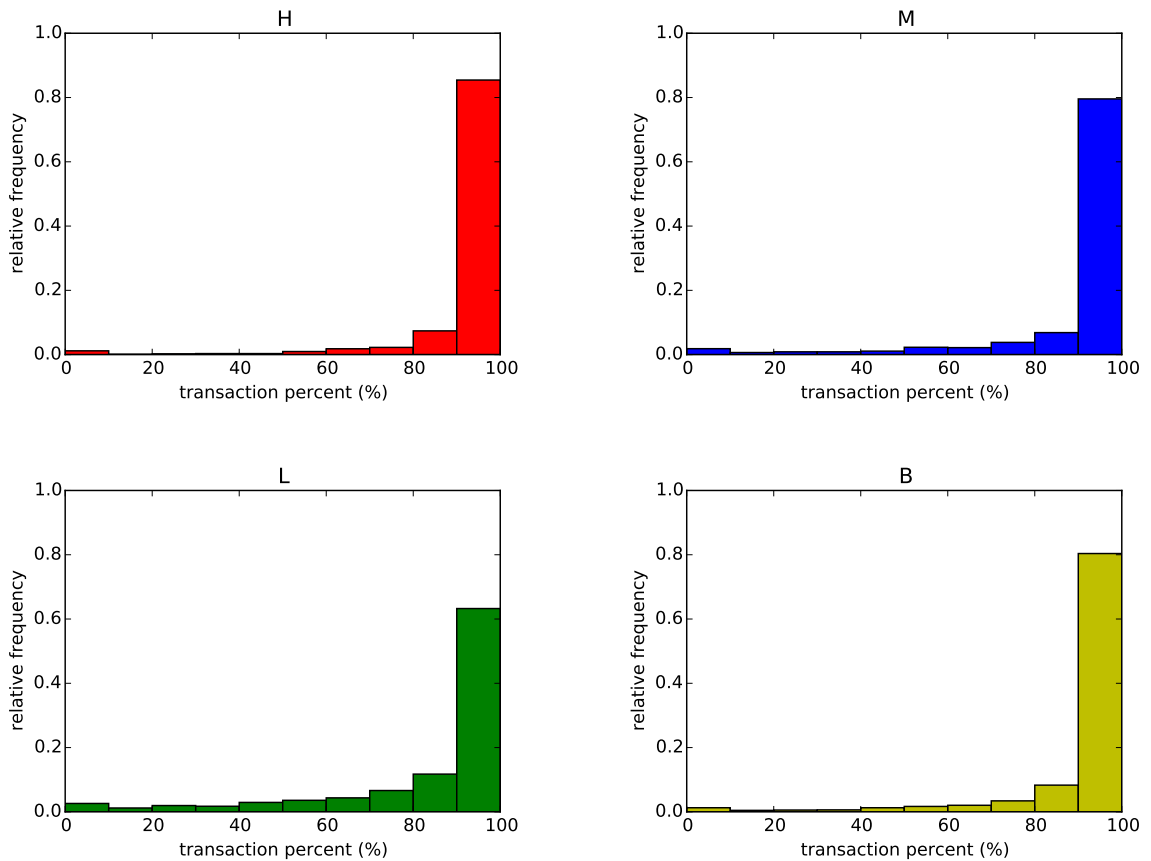


Figure 5.8: Distributions of successful transaction percentage in different groups

It can be seen that the majority of merchants in different groups have high percentages of successful transactions. Over 85% sellers in C_H have the high percentage of successful transactions (i.e., over 90%). The clusters C_M and C_B display similar performance regarding this performance measure, where there are 80% merchants who

achieve the high percentage of successful transactions. The sellers in C_L , however, show poor performance in this measure, where the percentage of sellers who achieve the high percentage of successful transactions is less than 70%.

5.4.3 PowerSeller Level

Powerseller is a program used in eBay e-commerce system for the evaluation of the overall performance of merchants on eBay. To be qualified for using this analysis, a merchant must meet multiple requirements, including at least 90 days active status, more than 98% positive feedback, a minimum of 100 transactions, a good standing account, and obeying eBay policies. PowerSeller classes sellers into different levels, from low to high including Bronze, Silver, Gold, Platinum, and Titanium levels. The eligible merchant is automatically labelled with a level depending on its overall performance that involves volume of items, customer service, feedback, and commitment.

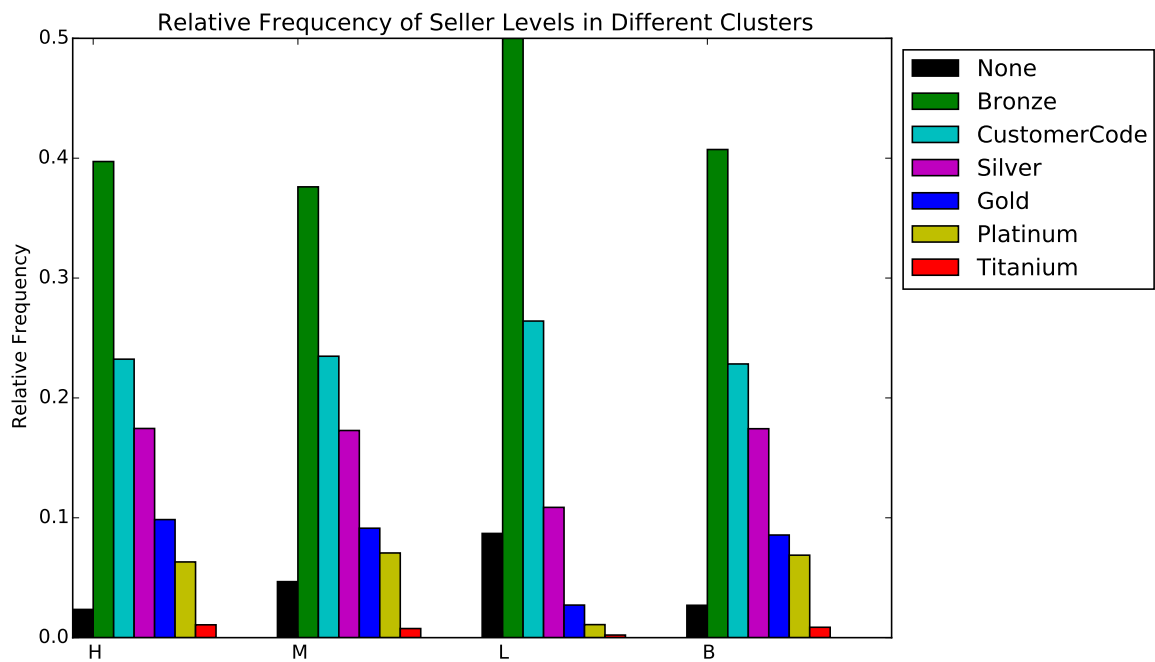


Figure 5.9: Sellers' Powerseller levels in different groups

eBay includes an API call that returns the Powerseller level of a specific seller. The returned value could be Bronze, Silver, Gold, Platinum, Titanium, None, or CustomerCode. A merchant with the latter two values is not a member of the Powerseller

program. Figure 5.9 shows how merchants in four groups are allocated to different Powerseller levels.

We consider a group having better performance regarding Powerseller level if the percentage of sellers in the group who have a high Powerseller level (i.e., Gold, Platinum or Titanium) is high. Based on this criteria, C_H, C_M, C_B show similar performance. Only C_L shows lower performance in the Powerseller level program. Nearly 50% sellers in group C_L are labelled with level of Bronze, while less than 40% sellers in other groups are labelled with Bronze. Similarly, less than 10% sellers in group C_L are labelled with level of Gold or higher, while nearly 20% sellers in other groups have a level of Gold or higher.

The performance measure regarding Powerseller level further justifies the suitability of using our four baseline distributions to classify sellers.

5.4.4 Conclusion

In this chapter, we propose four baseline distributions of customer values to classify sellers. The baseline distributions adopt the similar idea in [35, 34] but are adjusted to capture the statistical features of eBay “BIN” datasets.

Using the baseline distributions, we accordingly cluster the eBay sellers into four groups C_H, C_M, C_B, C_L . We then further evaluate the performance of sellers in different groups, measured with the number of transactions, the percentage of successful transactions, and the Powerseller level. The evaluation results show that the sellers in C_H always perform better than those in other groups. In particular,

1. The sellers in C_H have a higher average and a higher median of number of transactions;
2. Over 85% sellers in C_H have the high percentage of successful transactions (i.e., over 90%).
3. About 20% sellers in C_H have PowerSeller levels (i.e., Gold or higher).

In addition, the merchants in C_M and C_B also show relatively better performance than those in C_L . For instance, the average number and median of transactions in C_B are only slightly lower than the values in group C_H and are much higher than the values in group C_L .

The performance evaluation results are consistent with the classification of sellers, demonstrating that our proposed baseline distributions can effectively capture

the statistical features of eBay sellers and each of them can be used as an effective indicator for the seller's performance.

This chapter presents a baseline model of customer value distributions that can be used to evaluate the performance of sellers. A natural question would be why a seller has a specific customer value distribution. The answer to this question, however, is extremely difficult to find, because the customer value distribution depends on many factors for which data is hard to collect, such as the seller's investment on ads, the follow-on services after a transaction completes, the quality of products, and so on.

Due to the above reason, instead of answering the above challenging question directly, we study the customer feedback information in the next chapter. By exploring the dependency between customer feedback and the customer value distribution, we provide sellers with another insight why customer feedback is an important indicator for their business performance, which no online merchants can afford to ignore. Such an insight can help eBay merchants further improve their business strategies.

Chapter 6

Effects of Seller Reputation on the Customer Value Distribution

In the previous chapter, we have built customer value distributions for eBay sellers, from which sellers may obtain the characteristics of their customer relationships, specially, of all the customers who have ever shopped in an eBay store what the proportion of the high-value customers is. In other words, the customer value distributions can be considered as an indicator of customer retention or customer loyalty for a seller. The empirical results from our previous chapter have shown that the customer value distributions varied from store to store. Some stores have attracted and retained a large number of high-value customers while others have not. This observation raises a question: what are underlying factors that affect customer value distributions in e-commerce environment?

In contrast with the previous literature, which has mainly focused on examining the effect of product- or service-related quality [28, 60], customer satisfaction [28, 27], short-term promotions [42], and brand equity [57] on CRM, our focus is on effectiveness of online sellers' reputation. In the context of e-commerce, seller reputation reflected from consumer's feedback is available to every online shopper and it has been recognized to be one of determinants of a buyer's purchase decision and behaviour [55, 36]. In this chapter, we will leverage data mining techniques to examine how a seller's reputation is significant to its customer value distribution.

6.1 EBay Feedback System

According to previous research, an online seller's reputation measure is commonly based on the feedback ratings provided by the online feedback mechanism, also known as reputation system [54]. eBay supports a comprehensive feedback mechanism, where eBay buyers can leave a positive, neutral or negative ratings that reflect their general satisfaction with the seller. Then for that seller, eBay feedback system will record respectively the number of positive, neutral and negative ratings received. For instance, if a buyer is not satisfied with the experience with the seller, the buyer could leave a negative rating and then the number of negative ratings about that seller could be raised by one.

Table 6.1: eBay Detailed Seller Rating System

Aspects	Description	Default Value
Item as Decribed	Do the item title, description, image and performance match the item the buyer received?	N/A
Communication	Does the seller provide the contact information and timely and clearly answer any questions the customer had?	The seller would automatically receive a 5-star rating if he/she: - provided contact information - addressed questions within 1 business day - There were no refund requests
Shipping Time	Does the seller ship the item as soon as possible?	The seller would automatically receive a 5-star rating if he/she: - shipped the item within 1 day from the date of purchase - uploaded tracking information within 1 business day - updated tracking data within 4 days of payment clear
Shipping and Handling Charges	Are the shipping and handling charges reasonable?	The seller would automatically receive a 5-star rating if he/she provided free shipping.

In addition to overall feedback ratings above, eBay feedback system enables buyers

to leave Detailed Seller Ratings (DSR) up to 60 days after the transaction date. Buyers can rate four aspects of their transactions on a scale of 1- to 5-star, with the 5-star being the highest and the 1-star being the lowest. The description and automatic rating rules about those four aspects are given in Table 6.1.

Every eBay merchant owns a personal feedback profile, as shown in Figure 6.1, which contains the above feedback ratings for transactions that ended in the last 12 months. The table of “Recent Feedback ratings” displays the number of positive, number of neutral and number of negative overall feedback the seller has received. Also the table of “Detailed seller ratings” shows the average ratings and the number of ratings the customers have left on each of four additional dimensions as described in Table 6.1. Depending on the previous research on online reputation [58, 30], where the reputation of a seller was reflected by the seller’s feedback profile, our study considers the feedback ratings displayed in an eBay seller’s feedback profile as the seller’s reputation and will use them for our following analysis.

Feedback profile



Figure 6.1: Screen shot of the feedback profile of an eBay seller

6.2 Feedback Data Collection and Preprocessing

By using the eBay API as described in Chapter 3, we obtain the feedback ratings for transactions that ended in the last 12 months, including the values for number

of positive ratings, number of negative ratings, average rating and the total count of buyer ratings given for each DSR type. For notational convenience, we denote the variables as shown in Table 6.2.

Table 6.2: Variables and Shorthand

Variables	Shorthand
# of Positive Ratings	FB_pos
# of Negative Ratings	FB_neg
Aggregated Item as Describe Ratings = Average Rating * Total Count	IAD
Aggregated Communication Ratings = Average Rating * Total Count	C
Aggregated Shipping Time Ratings = Average Rating * Total Count	ST
Aggregated Shipping and Handling Charges Ratings = Average Rating * Total Count	SC













In the raw dataset, the values for the rating variables spread out in a wide range. For instance, the values of FB_pos range from the minimum 0 to the maximum 13,333,841. In order to follow the star levels aligning with the eBay practice as shown in Table 6.3, we scale the data using this table by assigning 0 to yellow star, 1 to blue star, and so on. Finally, the values of variables are limited in the range of $[0, 11]$.

6.3 Analysis of Factors Using Clustering

6.3.1 Methodology

In Chapter 5, we have categorized our collection of sellers into four clusters (i.e. C_H, C_M, C_L, C_B) based on the distances between their customer distributions and the baseline distributions. Similarly, we can re-partition these sellers into four clusters based on each individual feedback factors, as shown in Figure 6.2. Then we estimate

Table 6.3: eBay Stars and Ratings

Star	Color	Number of ratings	Scale
	Yellow	0 to 49	0
	Blue	50 to 99	1
	Turquoise	100 to 499	2
	Purple	500 to 999	3
	Red	1,000 to 4,999	4
	Green	5,000 to 9,999	5
	Yellow shooting star	10,000 to 24,999	6
	Turquoise shooting star	25,000 to 49,999	7
	Purple shooting star	50,000 to 99,999	8
	Red shooting star	100,000 to 499,999	9
	Green shooting star	500,000 to 999,999	10
	Silver shooting star	1,000,000 or more	11

the significance of a factor for the categorization result of customer value distributions by calculating the distance between the set of clusters related to customer value distributions and the clusters related to each individual feedback factors. The method is presented in detail in the following.

First of all, for notational convenience, C_H, C_M, C_L, C_B are respectively denoted as $C_{01}, C_{02}, C_{03}, C_{04}$. Also we define C_0 as a set of clusters:

$$C_0 = \{C_{01}, C_{02}, C_{03}, C_{04}\}.$$

In addition, let the set of clusters associated with each of six feedback factors, coded as C_1 for the clusters created based on FB_pos , C_2 for FB_neg , C_3 for IAD , C_4 for C , C_5 for ST , C_6 for SC . Each set includes four clusters as follows:

$$C_i = \{C_{i1}, C_{i2}, C_{i3}, C_{i4}\} \quad (1 < i < 6)$$

The clusters in C_1, C_2, \dots, C_6 are yielded by K-means, a clustering method designed to partition samples into specified k clusters. With K-means, our objective function is:

$$\arg \min_C \sum_{j=1}^k \sum_{x \in C_{ij}} \|x - \mu_{ij}\|^2 \quad (0 < i < 6), \quad (6.1)$$

where x represents a data point in cluster C_{ij} , k denotes the number of clusters, which

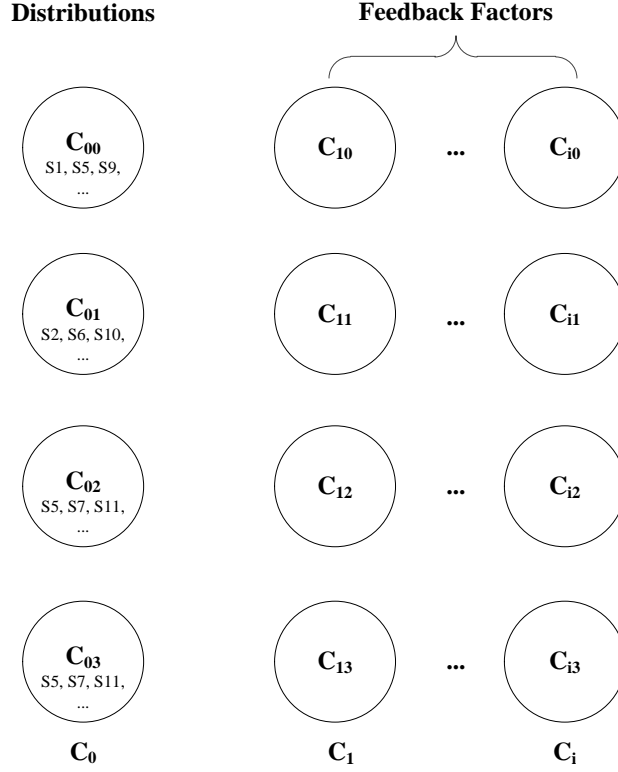


Figure 6.2: Clusters based on customer value distributions and feedback factors

is set to be 4 in our study, and μ_{ij} represents the centroid in cluster C_{ij} . Here we initialize μ_{ij} by K-means++ initialization algorithm, because in terms of accuracy and speed K-means++ works better than the traditional initialization method that is choosing k initial centers arbitrarily [7]. The steps of finding a set of initial centers (IC_i) given dataset C_i are listed as follows:

1. Choose an initial center c_{i0} uniformly at random from X_i that are the points in C_i ($0 < i < 6$) and add it to the set of IC_i ;
2. Choose the next center c_{i1} by sampling a point $x \in X_i$ with probability $\frac{D(c_{i1})^2}{\sum_{x \in X} D(x)^2}$, where $D(x)$ represents the minimum distance from x to its nearest center point from IC_i , namely $D(x) = \min(\|x - c_{ij}\|^2)$ for all $c_{ij} \in IC_i$;
3. Repeat Step 2 until IC_i has contained k elements.

Based on the above process, given C_i we get a set of initial centres $IC_i = \{c_{i0}, c_{i1}, \dots, c_{ik}\}$. Then we perform the standard K-means algorithm that works as follow:

1. Assign all points in C_i to the closest initial centers and finally formed k clusters: $C_{ij}(0 < j < k)$;
2. Recompute the centroid of each cluster by choosing the new c_{ij} to be the mean of all points in C_{ij} ;
3. Repeat Steps 1 and 2 until convergence.

After clustering, we then define the distance between each cluster in the set of C_0 and each individual cluster in the set of C_i ($0 < i < 6$) by Jaccard distance:

$$Dist(C_{0d}, C_{ij}) = 1 - \frac{|C_{0d} \cap C_{ij}|}{|C_{0d} \cup C_{ij}|} \quad (1 < i < 6, 0 < d < 4, 0 < j < 4), \quad (6.2)$$

where $|C_{0d} \cap C_{ij}|$ is calculated by the number of common sellers in both clusters while $|C_{0d} \cup C_{ij}|$ is obtained by the total number of sellers in cluster C_{0d} plus the sellers in C_{ij} .

By applying this measurement, for each C_0 vs. C_i ($0 < i < 6$) pair we can obtain 16 distances. For example, by calculating the distance between each cluster in set C_0 and each cluster in C_1 (i.e. clusters related to *FB_pos*), finally 16 distances will be obtained as listed in Table 6.4. These distances can be treated as the elements in a

Table 6.4: Distances Between Each Cluster in C_0 and Each One in C_1

	C_{10}	C_{11}	C_{12}	C_{13}
C_{01}	0.1378	0.1247	0.0726	0.0549
C_{02}	0.1481	0.0949	0.0767	0.0617
C_{03}	0.1578	0.0850	0.1080	0.0139
C_{04}	0.2325	0.1580	0.0759	0.0881

two-dimension array and will be used as input in our algorithm used for discovering the minimum distance between the set C_0 and the set C_i ($1 < i < 6$). As described in Algorithm 1, we first select four elements from the array, but the selection must require no two elements share the same row and column. After obtaining all the possible combinations of four elements, we sum up the elements in each combination and define the smallest one as the minimum distance between the two sets of clusters. For example, we input values in Table 6.4 to Algorithm 1 and then it selects a combination of four values, which are shown in bold in this table. We can see that no two of the chosen elements share the same row and column. Also, the algorithm guarantees the

sum of the four values (i.e. 0.3225) is the smallest among all the possible combinations. Thus, 0.3225 is the final distance between C_0 and C_1 .

The final distance will be treated as the indicator of how its associated factor affects the sellers' customer value distributions. The smaller the distance is, the better the clusters related to the factor would match the four clusters related to customer value distributions. In other words, the smaller distance suggests that the factor is more significant.

6.3.2 Result and Conclusion

By applying Algorithm 1, we obtain the distance between C_0 and C_i as listed in Table 6.5. We can see that the distance between C_0 and C_2 is the smallest, which indicates that the clusters based on the similarity of FB_neg best match the four clusters (i.e. C_H, C_M, C_L, C_B). Therefore we can conclude that among six feedback factors FB_neg is the most significant factor that affects the customer value distributions.

In the following sections, we perform more statistical analysis to gain deeper understanding about the relationship between a seller's reputation and the customer value distributions.

6.4 Analysis of Factors Using Logistic Regression

Regression has been widely applied for the estimation of relationship between variables. In this section, we will use logistic regression to examine whether a seller's feedback ratings have impacts on the outcome of the seller's customer value distribution.

6.4.1 Variables and Modelling

In our regression model, sellers' customer value distributions will be treated as the response/dependent variable while the feedback ratings (i.e. $FB_pos, FB_neg, IAD, C, ST$ and SC) as the predictor/independent variables. As in Chapter 5 the sellers' customer value distributions have been categorized into four clusters (i.e. C_H, C_M, C_L, C_B), the response variable can be considered as a categorical variable with four levels under the assumption that the levels have no natural ordering. In this situation, we take

Algorithm 1: Calculate distance between two sets of clusters

```

input :
  Dist[k][k] (A  $4 \times 4$  array that contains the Jaccard distances);
  k  $\leftarrow$  4;
output: minDist ;
begin
  | minDist  $\leftarrow$   $+\infty$ ;
  | elements[k];
  | /* to store the element chosen from dist[k][k] */
  | columns[k];
  | /* to store the column index of the element in array elements */
  | /*
  | chooseElement(0,k);
  | /* call the recursive function as defined in the following */

  | Function chooseElement(row, k)
  | | for col  $\leftarrow$  0 to k do
  | | | if canChoose(row, col) is true then
  | | | | elements[row]  $\leftarrow$  Dist[row][col];
  | | | | columns[row]  $\leftarrow$  col;
  | | | | if row = k - 1 then
  | | | | | sumDist  $\leftarrow$  Sum of elements[];
  | | | | | if minDist > sumDist then
  | | | | | | minDist  $\leftarrow$  sumDist;
  | | | | | end
  | | | | end
  | | | | chooseElement(row + 1, k);
  | | | end
  | | end
  | end

  | Function canChoose(row,col)
  | | /* check if the dist[row][k] can be choosen */
  | | if col  $\in$  elements then
  | | | return false;
  | | end
  | | return true;
  | end
end

```

Table 6.5: Distances Between C_0 and C_i from Smallest to Largest

Feedback factor	Distance to C_0
<i>FB_neg</i> (C_2)	0.2317
<i>SC</i> (C_6)	0.2972
<i>ST</i> (C_5)	0.3049
<i>FB_pos</i> (C_1)	0.3225
<i>C</i> (C_4)	0.3324
<i>IAD</i> (C_3)	0.3446

advantage of multinomial logistic regression that is designed for multi-class regression analysis.

With multinomial logistic regression, one of the outcomes of the response variable is designated as the reference group, and each of the remaining levels compares to the reference. According to [21], the choice of reference group can be arbitrary. From Chapter 5, we have known that the performances of C_H , C_M and C_B are close and are much better than C_L in terms of sales, so we are interested in what factors might contribute to a seller categorized as C_L than as others. If the factors do exist, it would help the sellers in C_L identify areas for improvement. We therefore choose C_L as our reference group and will compare the subjects labelled with C_H , C_M or C_B to the subjects with C_L :

1. C_H vs. C_L
2. C_M vs. C_L
3. C_B vs. C_L

Based on the above comparisons, three models will be built. Denote our independent variables $\mathbf{X} = [FB_pos, FB_neg, IAD, C, ST, SC]$. Let the three-category dependent variable coded as 0 for C_L , 1 for C_H , 2 for C_M , and 3 for C_B . Also the conditional probability that the seller is labelled with level g given the variables X is denoted by $P(D = g|\mathbf{X})$ where D is the dependent variable. Then our model is stated as follows:

$$\ln\left[\frac{P(D = g|\mathbf{X})}{P(D = 0|\mathbf{X})}\right] = \alpha_{g0} + \beta_{g1}FB_pos + \beta_{g2}FB_neg + \beta_{g3}IAD + \beta_{g4}C + \beta_{g5}ST + \beta_{g6}SC \quad (6.3)$$

where $g \in \{1, 2, 3\}$, α_{g0} and β_{gi} ($i = 1, 2, \dots, 6$) represent the parameters to be estimated.

6.4.2 Testing Global Null Hypothesis

Before running logistic regression, we need to test whether all of the independent variables are significant versus at least one of them is significant. To this end, the null hypothesis for this testing is created as follows:

$$H_0 : \beta_{g0} = \beta_{g1} = \beta_{g2} = \beta_{g3} = \beta_{g4} = \beta_{g5} = \beta_{g6} = 0,$$

where $g \in \{1, 2, 3\}$. We apply likelihood ratio, Score and Wald tests to validate H_0 and the results are displayed in Table 6.6. We can see that the all the p-values are less than 0.001. At the significance level of 0.05, we reject the above global null hypothesis, H_0 . Therefore, there is at least one variable that has a significant impact on the dependent variable.

Table 6.6: Testing Global Null Hypothesis

Test	Chi-Square	p-value
likelihood Ratio	293.5133	< .0001
Score	306.6944	< .0001
Wald	261.9860	< .0001

6.4.3 Analysis of Maximum Likelihood Estimates

Next we apply maximum likelihood estimation for estimating the parameters (i.e. intercepts and beta coefficients) in our mathematical model that has been shown in Equation (6.3). In addition, we use Chi-Square test to check whether the null hypothesis, that the regression coefficient of a specified independent variable is zero in the model, can be rejected. The estimation and the Chi-Square test results are listed in Table 6.7. The “Fun.” in this table indicates to the three models:

$$\text{Model 1} : \ln\left[\frac{P(D=C_1|\mathbf{X})}{P(D=C_0|\mathbf{X})}\right]$$

$$\text{Model 2} : \ln\left[\frac{P(D=C_2|\mathbf{X})}{P(D=C_0|\mathbf{X})}\right]$$

$$\text{Model 3} : \ln\left[\frac{P(D=C_3|\mathbf{X})}{P(D=C_0|\mathbf{X})}\right],$$

which have been described in the previous section. Each of the three logistic models requires a set of parameter estimates that are listed in the column of “Estimate”. By

exponentiating the parameter estimates, we obtain the values of “O.R.” (Odds Ratios) that indicate into which group the outcome would be more likely to be categorized with the change of the dependent variable in our model. If the odds ratio is less than one, the outcome is more likely to fall in the reference group (i.e. C_L), otherwise the outcome is more likely to be categorized into one of the remaining groups (i.e. C_H , C_M or C_B). Furthermore, the “p-value” is used to interpret the Chi-Square test results. If the p-value is less than the significance level of 0.05, the null hypothesis that a particular independent variable is non-significant cannot be accepted. We therefore focus on the parameter estimates associated with a p-value less than 0.05, namely the ones with the superscript “*”.

The estimation results for three models are different. In the Model 1, the variables of FB_pos , C , and ST are non-significant, but FB_neg , IAD , and SC are statistically significant. According to their odds ratios, given the other variables are held constant a subject with higher either IAD or SC would be more likely to be categorized as C_H than as C_L ; in contrast a subject with higher value of FB_neg would be less likely to be categorized as C_H than as C_L .

In the Model 2 that is for C_M vs. C_L , observing the parameter estimates and odds ratios we know that only FB_neg and SC are statistically significant. Specifically, a subject with higher value of FB_neg would be less likely to be categorized as C_M than as C_L , whereas a subject with higher value of SC is expected to be more likely to be categorized as C_M than as C_L .

In the Model 3 that is for C_B vs. C_L , we can conclude that FB_pos , FB_neg and SC are statistically significant. Like Model 1 and Model 2, subjects with higher FB_pos are more likely to have their customer value distributions categorized as C_B than as C_L ; in contrast, subjects with higher FB_neg are less likely to be categorized as C_B . What’s more, subjects with higher value of SC are more likely to be categorized as C_B than as C_L .

6.4.4 Conclusion

Based on the analysis of parameter estimates and Chi-Square test results, we now know how each of the variables about feedback ratings impacts the categorization result of customer value distributions. In summary, we conclude:

1. FB_pos is statistically significant for C_B vs. C_L , but not for C_H vs. C_L or C_M vs. C_L . If two sellers have identical ratings for FB_neg , IAD , C , ST and

Table 6.7: Analysis of Maximum Likelihood Estimates

Parameter	Fun.	Estimate	Std. Err.	Chi-Square	p-value	O.R.
Intercept	1	-1.0867	0.1600	46.15	< .0001***	-
	2	-0.6609	0.1446	20.89	< .0001***	-
	3	-0.4079	0.1282	10.13	0.0015**	-
FB_pos	1	0.00426	0.0531	0.01	0.9361	0.990
	2	-0.00350	0.0509	0.00	0.9452	0.992
	3	0.1347	0.0439	9.40	0.0022**	1.137
FB_neg	1	-0.4184	0.0652	41.13	< .0001***	0.667
	2	-0.1784	0.0489	13.33	0.0003***	0.840
	3	-0.2586	0.0439	34.65	< .0001***	0.778
IAD	1	0.1913	0.0974	3.86	0.0496*	1.064
	2	-0.0933	0.0989	0.89	0.3457	0.904
	3	-0.0192	0.0835	0.05	0.8183	0.954
C	1	-0.2314	0.1404	2.71	0.0994	0.793
	2	-0.1478	0.1390	1.13	0.2876	0.863
	3	-0.1996	0.1199	2.77	0.0959	0.819
ST	1	-0.1324	0.1340	0.98	0.3232	0.876
	2	0.1039	0.1286	0.65	0.4190	1.110
	3	0.1058	0.1110	0.91	0.3406	1.112
SC	1	0.5621	0.1227	21.00	< .0001**	1.754
	2	0.3859	0.1206	10.25	0.0014**	1.471
	3	0.3464	0.1056	10.76	0.0010**	1.414

*** $p < 0.001$ ** $p < 0.01$, * $p < 0.05$

SC, the seller with higher positive feedback rating would be more likely to be categorized as C_H rather than C_L .

2. *FB_neg* is statistically significant for all three models. The seller with higher negative feedback rating would be more likely to be categorized as C_L rather than as C_H , C_M or C_B .

3. *IAD* is only statistically significant for C_H vs. C_L . The seller with a higher rating on *IAD* would be more likely to be classified as C_H rather than C_L . Nevertheless, *IAD* is not significant for C_M vs. C_L or C_B vs. C_L .
4. *C* and *ST* are not significant for all three models. The ratings for *C* and *ST* do not exhibit significant impacts on the seller's customer value distribution.
5. *SC* is statistically significant for all three models. The seller with a higher rating on *SC* would be more likely to be classified as C_H , C_M or C_B than as C_L .

These conclusions summarize how each of feedback ratings is significant for a seller's customer value distribution. More importantly, the eBay sellers, especially those labelled with C_L , could obtain the following suggestions drawn from this study:

1. Make an effort to increase the ratings about *IAD* (Item as Described) and *SC* (Shipping and Handling Charges). Sellers with higher ratings of either *IAD* or *SC* would be more likely to attract high-value customers.
2. Pay attention to the overall ratings. A store with more positive ratings would be more likely to attract and retain high-value customers, whereas the higher negative ratings would have negative impacts on the customer retention.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis presents a way to tackle the challenges that online merchants are faced with in the context of rapid growth of e-commerce, as we pointed out in the beginning of this thesis.

The first challenge is about customer identification: how does a seller find out high-value customers? To this end, we propose an approach for measuring customer value based on the buyer's purchase frequency and expenditures in a particular store during our survey period so that the seller can know the value of each of its buyers. The higher the value is, the more contributions the buyer has made to the seller.

Since it is easy for online shoppers to compare and switch stores, another question that online merchants concern is whether customer retention is significant as always? Our research finds that customer retention still plays an important role in the CRM. When comparing the performance of the sellers with different customer value distributions, we discover that in general the sellers with a larger number of regular customers perform much better than the sellers with less regular customers.

Finally, should online merchants be concerned of feedback ratings and what type of feedback information could impact the customer values more significantly? By applying logistic regression and clustering to our eBay feedback data, the results reveal that the general positive and negative ratings as well as detailed ratings on "Item as Described" and "Shipping and Handling Charges" impact significantly the seller's customer value distributions, while the ratings on "Communication" and "Shipping Time" are not significant.

7.2 Future Work

While this thesis investigated several challenges related to e-CRM, the scope of this study can be further extended.

Integrate Recency to Customer Value

One limitation of our study is that it is mainly focused on eBay transactions that were made during the second quarter of 2013, provided by Terapeak. Although the eBay API enables users to communicate directly with the eBay database, it limits the daily number of calls. Due to the limitation, it is hard to collect a complete set of transaction records. If a longer-term dataset was available, we could compute the recency value (i.e. how many days have been passed since a customer's last purchase), and then consider this value in addition to the values of frequency and monetary in the process of measuring customer value.

Collect Reviews from Social Networks

In our study, we use the eBay feedback ratings to estimate a seller's reputation. In fact, with the rise of online social networks, such as Facebook, Twitter, Google Plus, buyers not only leave their ratings in the eBay feedback system, but might post their reviews through these channels. The seller's reputation thus might be captured from those web sites as well. It would be interesting to study how reviews from online social networks impact a seller's customer value distribution.

Categorize Transaction Data

The eBay transaction data we used for analysis in this thesis involves products from 35 categories. We apply logistic regression to the whole dataset without taking item category into account. Customers may have different shopping behaviour for products in different categories. In our future work, we would preprocess the data by grouping every transaction based on its associated item's category and then look into whether there are differences between different categories.

Extend Data Analysis to Prediction

Another direction is to predict the seller's future sales based on its current customer value distribution. In our research, we have defined a relative customer value for each seller-buyer pair, by comparing the buyer's purchase frequency and expenditure in that store to the buyer's purchase frequency and expenditure in other stores. This value partially reflects the buyer's loyalty to the store, therefore it could be used as an approximated probability of repeat purchase in the future. In other words, depending on a buyer's current customer value, we can estimate this buyer's future purchase intention and then predict the seller's future sales. In addition to predict the likelihood of repeat purchase, we could predict the successful rate of the transaction. With eBay API, we could know the information about shopping carts. This information could reveal if the items in a particular shopping cart have been purchased, replaced or removed. If we could keep track of the changes in shopping carts, we would explore what items are more likely to be removed from shopping cart, what items are substitutes for the removed ones, and what set of items are frequently purchased together. This study might help sellers better identify or predict popular items.

Fine-grained Analysis on eBay Customers

Based on the distributions of customer values, we partition eBay sellers into four distinct categories. With the same segmentation methodology, we could classify the customers based on their purchase behaviour. For instance, the authors in [45] examine the transaction dataset on Taobao.com and then classify the customers into six types: "economical purchasers, active-star purchasers, direct purchasers, high-loyalty purchasers, risk-averse purchasers and credibility-first purchasers". After this segmentation, they explore what promotion strategy would be appropriate to each individual group. To extend our work, we could treat those six types as the baseline and then classify the customers in our dataset. Once this categorization method works well on our dataset, the result would help sellers make targeted strategies.

Bibliography

- [1] 76 percent of canadians shopped online last year, canada post says. <http://www.cbc.ca/news/business/76-of-canadians-shopped-online-last-year-canada-post-says-1.3070651>. [Online; posted May 2015].
- [2] ebay developer program. <http://go.developer.ebay.com/>. Accessed: 2014-12-10.
- [3] ebay is 20 years old today. <http://venturebeat.com/2015/09/03/ebay-is-20-years-old-today/>. [Online; posted Sep. 2015].
- [4] Ecommerce consumer reviews: why you need them and how to use them. <https://econsultancy.com/blog/9366-ecommerce-consumer-reviews-why-you-need-them-and-how-to-use-them/>. [Online; posted July 2015].
- [5] Statista. <http://www.statista.com/>. Accessed: 2015-11-14.
- [6] Celestine C Aguwa, Leslie Monplaisir, and Ozgu Turgut. Voice of the customer: customer satisfaction ratio based analysis. *Expert Systems with Applications*, 39(11):10112–10119, 2012.
- [7] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. pages 1027–1035, 2007.
- [8] Chris Baumann, Greg Elliott, and Suzan Burton. Modeling customer satisfaction and loyalty: survey data versus data mining. *Journal of services marketing*, 26(3):148–157, 2012.
- [9] Jonah Berger, Alan T Sorensen, and Scott J Rasmussen. Positive effects of negative publicity: when negative reviews increase sales. *Marketing Science*, 29(5):815–827, 2010.

- [10] Wouter Buckinx, Geert Verstraeten, and Dirk Van den Poel. Predicting customer loyalty using the internal transactional database. *Expert systems with applications*, 32(1):125–134, 2007.
- [11] Francis Buttle. *Customer relationship management: concepts and technologies*. Routledge, 2009.
- [12] Howard E Butz and Leonard D Goodstein. Measuring customer value: gaining the strategic advantage. *Organizational dynamics*, 24(3):63–77, 1997.
- [13] Luis Cabral and Ali Hortacsu. The dynamics of seller reputation: evidence from ebay. *The Journal of Industrial Economics*, 58(1):54–78, 2010.
- [14] Hui-Chu Chang and Hsiao-Ping Tsai. Group rfm analysis as a novel framework to discover better customer consumption behavior. *Expert Systems with Applications*, 38(12):14499–14513, 2011.
- [15] Pei-Yu Chen, Shin-yi Wu, and Jungsun Yoon. The impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings*, page 58, 2004.
- [16] Ching-Hsue Cheng and You-Shyang Chen. Classifying the segmentation of customer value via frm model and rs theory. *Expert systems with applications*, 36(3):4176–4184, 2009.
- [17] Chia Chi Lin. A critical appraisal of customer satisfaction and e-commerce. *Managerial Auditing Journal*, 18(3):202–212, 2003.
- [18] Chui-Yu Chiu, Yi-Feng Chen, I-Ting Kuo, and He Chun Ku. An intelligent market segmentation system using k-means and particle swarm optimization. *Expert Systems with Applications*, 36(3):4558–4565, 2009.
- [19] Toyin A Clottey, David A Collier, and Michael Stodnick. Drivers of customer loyalty in a retail store environment. *Journal of Service Science (JSS)*, 1(1):35–48, 2011.
- [20] Richard Colombo and Weina Jiang. A stochastic rfm model. *Journal of Interactive Marketing*, 13(3):2–12, 1999.

- [21] Kleinbaum David and Klein Mitchel. Logistic regression: a self learning text, 1994.
- [22] Chrysanthos Dellarocas. The digitization of word of mouth: promise and challenges of online feedback mechanisms. *Management science*, 49(10):1407–1424, 2003.
- [23] Wenjing Duan, Bin Gu, and Andrew B Whinston. The dynamics of online word-of-mouth and product salesan empirical investigation of the movie industry. *Journal of retailing*, 84(2):233–242, 2008.
- [24] Liran Einav, Chiara Farronato, Jonathan D Levin, and Neel Sundaresan. Sales mechanisms in online markets: what happened to internet auctions? Technical report, National Bureau of Economic Research, 2013.
- [25] Peter S Fader, Bruce GS Hardie, and Ka Lok Lee. Rfm and clv: using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4):415–430, 2005.
- [26] Bradley Gale and Robert Chapman Wood. *Managing customer value: Creating quality and service that customers can see*. Simon and Schuster, 1994.
- [27] Anders Gustafsson, Michael D Johnson, and Inger Roos. The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of marketing*, 69(4):210–218, 2005.
- [28] Thorsten Hennig-Thurau and Alexander Klee. The impact of customer satisfaction and relationship quality on customer retention: a critical reassessment and model development. *Psychology & Marketing*, 14(8):737–764, 1997.
- [29] Seyed Mohammad Seyed Hosseini, Anahita Maleki, and Mohammad Reza Gholamian. Cluster analysis using data mining approach to develop crm methodology to assess the customer loyalty. *Expert Systems with Applications*, 37(7):5259–5264, 2010.
- [30] Daniel Houser and John Wooders. Reputation in auctions: theory, and evidence from ebay. *Journal of Economics & Management Strategy*, 15(2):353–369, 2006.
- [31] Arthur Middleton Hughes. *Strategic database marketing*. McGraw-Hill Pub. Co., 2005.

- [32] Chihli Hung and Chih-Fong Tsai. Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. *Expert systems with applications*, 34(1):780–787, 2008.
- [33] Hyunseok Hwang, Taesoo Jung, and Euiho Suh. An ltv model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert systems with applications*, 26(2):181–188, 2004.
- [34] Wolfgang Jank and Galit Shmueli. *Modeling online auctions*, volume 91. John Wiley & Sons, 2010.
- [35] Wolfgang Jank and Inbal Yahav. E-loyalty networks in online auctions. *Robert H. Smith School Research Paper No. RHS*, pages 06–101, 2009.
- [36] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2):618–644, 2007.
- [37] Mahboubeh Khajvand, Kiyana Zolfaghar, Sarah Ashoori, and Somayeh Alizadeh. Estimating customer lifetime value based on rfm analysis of customer purchase behavior: case study. *Procedia Computer Science*, 3:57–63, 2011.
- [38] Su-Yeon Kim, Tae-Soo Jung, Eui-Ho Suh, and Hyun-Seok Hwang. Customer segmentation and strategy development based on customer lifetime value: a case study. *Expert systems with applications*, 31(1):101–107, 2006.
- [39] Alexander H Kracklauer, D Quinn Mills, and Dirk Seifert. Customer management as the origin of collaborative customer relationship management. In *Collaborative Customer Relationship Management*, pages 3–6. Springer, 2004.
- [40] Vineet Kumar and Werner Reinartz. *Customer relationship management: concept, strategy, and tools*. Springer Science & Business Media, 2012.
- [41] Robert Kuttner. The net: a market too perfect for profits. *BusinessWeek*, May 2010.
- [42] Michael Lewis. The influence of loyalty programs and short-term promotions on customer retention. *Journal of Marketing Research*, 41(3):281–292, 2004.

- [43] Duen-Ren Liu and Ya-Yueh Shih. Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. *Journal of Systems and Software*, 77(2):181–191, 2005.
- [44] Duen-Ren Liu and Ya-Yueh Shih. Integrating ahp and data mining for product recommendation based on customer lifetime value. *Information & Management*, 42(3):387–400, 2005.
- [45] Ying Liu, Hong Li, Geng Peng, Benfu Lv, and Chong Zhang. Online purchaser segmentation and promotion strategy selection: evidence from chinese e-commerce market. *Annals of Operations Research*, 233(1):263–279, 2015.
- [46] Glenn Llopis. Consumers are no longer brand loyal. *Fobes*, Dec 2014.
- [47] Mikhail I Melnik and James Alm. Does a seller’s ecommerce reputation matter? evidence from ebay auctions. *Journal of Industrial Economics*, pages 337–349, 2002.
- [48] Vera L Miguéis, Dirk Van den Poel, Ana S Camanho, and João Falcão e Cunha. Predicting partial customer churn using markov for discrimination for modeling first purchase sequences. *Advances in Data Analysis and Classification*, 6(4):337–353, 2012.
- [49] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. Application of data mining techniques in customer relationship management: a literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.
- [50] Taher Niknam and Babak Amiri. An efficient hybrid approach based on pso, aco and k-means for cluster analysis. *Applied Soft Computing*, 10(1):183–197, 2010.
- [51] Daniela Oelke, Ming Hao, Christian Rohrdantz, Daniel Keim, Umeshwar Dayal, Lars-Erik Haug, Halldór Janetzko, et al. Visual opinion analysis of customer feedback data. pages 187–194, 2009.
- [52] Haemoon Oh. Service quality, customer satisfaction, and customer value: a holistic perspective. *International Journal of Hospitality Management*, 18(1):67–82, 1999.

- [53] Atul Parvatiyar and Jagdish N Sheth. Customer relationship management: emerging practice, process, and discipline. *Journal of Economic and Social research*, 3(2):1–34, 2001.
- [54] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. The value of reputation on ebay: a controlled experiment. *Experimental Economics*, 9(2):79–101, 2006.
- [55] Petra Schubert and Mark Ginsburg. Virtual communities of transaction: the role of personalization in electronic commerce. *Electronic Markets*, 10(1):45–55, 2000.
- [56] Venkatesh Shankar, Amy K Smith, and Arvind Rangaswamy. Customer satisfaction and loyalty in online and offline environments. *International journal of research in marketing*, 20(2):153–175, 2003.
- [57] Florian Stahl, Mark Heitmann, Donald R Lehmann, and Scott A Neslin. The impact of brand equity on customer acquisition, retention, and profit margin. *Journal of Marketing*, 76(4):44–63, 2012.
- [58] Stephen S Standifird. Reputation and e-commerce: ebay auctions and the asymmetrical impact of positive and negative ratings. *Journal of Management*, 27(3):279–295, 2001.
- [59] Ronald S Swift. *Accelerating customer relationships: using crm and relationship technologies*. Prentice Hall Professional, 2001.
- [60] Karin A Venetis and Pervez N Ghauri. Service quality and customer retention: building long-term relationships. *European Journal of marketing*, 38(11/12):1577–1598, 2004.
- [61] Wouter Verbeke, David Martens, Christophe Mues, and Bart Baesens. Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3):2354–2364, 2011.
- [62] Robert B Woodruff. Customer value: the next source for competitive advantage. *Journal of the academy of marketing science*, 25(2):139–153, 1997.

- [63] Rounq-Shiunn Wu and Po-Hsuan Chou. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, 10(3):331–341, 2011.
- [64] Yaya Xie, Xiu Li, EWT Ngai, and Weiyun Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3):5445–5449, 2009.
- [65] Qiang Ye, Yijun Li, Melody Kiang, and Weifang Wu. The impact of seller reputation on the performance of online sales: evidence from taobao buy-it-now (bin) data. *ACM SIGMIS Database*, 40(1):12–19, 2009.
- [66] Youjae Yi and Taeshik Gong. Customer value co-creation behavior: scale development and validation. *Journal of Business Research*, 66(9):1279–1284, 2013.
- [67] Kunpeng Zhang, Ramanathan Narayanan, and Alok Choudhary. Voice of the customers: mining online customer reviews for product feature-based ranking. pages 11–11, 2010.