Application of Clustering, Logistic Regression and Decision Tree Induction on EGM Data for Detection and Prediction of At-Risk and Problem Gamblers

by

Wenjia Ni

Submitted in partial fulfillment of the requirements
for the degree of Master of Electronic Commerce

at

Dalhousie University
Halifax, Nova Scotia
July 2014

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The use of data mining techniques for problem gambling behavior analysis has huge potential to offer players protection and to reduce the risk of gambling-related harms. In this thesis, we apply three data mining models—clustering, logistic regression and decision tree on one month EGM player data to separate players into different groups, identify which gambling behavior are highly associated with gambling addiction, and derive predictive rules for predicting potential at-risk and problem gamblers. We consequently separated all players into four groups—non-problem gambler, low-risk gambler, moderate-risk gambler, and problem gambler groups, based on their similar behavioral characteristics. Three behavioral indicators and four best predictive rules are finally obtained to predict at-risk and problem gamblers. It is hoped that this thesis will provide a useful resource for EGM manufacturers to redesign their machines to avoid risky and problem gambling behavior.

LIST OF ABBREVIATIONS USED

ANOVA	Analysis of Variance
CPGI	Canadian Problem Gambling Index
df	degrees of freedom
EGMs	Electronic Gaming Machines
MD	Mahalanobis Distance
RGC	Responsible Gambling Council of Canada
SAS E-Miner	SAS Enterprise Miner

ACKNOWLEDGEMENTS

First and foremost, I would like to express the deepest appreciation to my supervisor Dr. Vlado Keselj. Without his guidance and persistent help it would not be possible to finish the research and the thesis. Many thanks to Dr. Evangelos E. Milios and Dr. Qiguang Gao on my thesis committee.

Wenjia Ni

Halifax

July 22, 2014

CHAPTER 1 INTRODUCTION

1.1 BACKGROUND

Problem gambling is any gambling behavior that negatively impacts personal life, family members, social network, and the society (Griffiths, 2009). Unlike recreational players, problem gamblers are usually unable to control their own gambling behavior and therefore spend much more time and money than they initially intended. The uncontrollable gambling behavior results in the large financial losses and ultimately causes the significant harm to their life (Productivity Commission, 2010).

Although problem gambling prevalence rates are dissimilar in different jurisdictions, the overall rate has increased quickly in recent years globally. The Responsible Gambling Council of Canada (RGC) reported that the rate of moderate-risk and problem gamblers among Canadian adults was 3.7% in 2011-12 (Responsible Gambling Council, 2013), nearly 1% higher than ten years ago (Responsible Gambling Council, 2004). In many European countries, the prevalence rates have been above 3%, which is higher than the average world rate (typically 0.5%-2%) (Griffiths, 2009).

The growth of problem gambling is primarily being driven by the introduction of Electronic Gambling Machines (EGMs), which is the most accessible and predominant form of gambling in casinos. The RGC pointed out that EGM play was one of the strongest predictors of problem gambling (Responsible Gambling Council, 2006). The Australian Productivity Commission reported that the frequency of gambling on EGMs was highly related to the risks of problem gambling. They estimated that approximately 15% of all EGM frequent players were problem gamblers, and the other 15% were moderate-risk gamblers (Productivity Commission, 2010).

Griffiths (2009) also mentioned that most problem gamblers who seek treatment in a lot of European countries were EGM players.

Given the rapid growth of EGM-related problem gambling and its significantly potential harms, researchers have been conducting studies in an attempt to find out what causes problem gambling and how to prevent gambling addiction. Researchers typically first use questionnaires or interviews method to collect data, and then apply statistical methodologies to analyze the collected data and to understand the characteristics of gamblers.

But recently, the method of collecting gambler behavioral data through the casino loyalty program has been proposed, inasmuch as an increasing number of casinos have been implementing the loyalty program to track gamblers gambling behavior. Like loyalty programs used by retailers, the casino loyalty program encourages players to use their loyalty card when gambling by offering bonus rewards or cash back. When gamblers insert their card into the EGM, their behavioral data such as duration, money spent and game types are recorded by the customer tracking system. By using behavioral data in conjunction with demographical data, researchers are able to completely analyze a gambler, and accordingly to identify at-risk and problem gamblers (Schellinck & Schrans, 2011).

Although this approach provides more reliable data compared with questionnaires or interviews, several potential difficulties need to be noticed. If customers share their card, the data will be unreliable and cannot be used for individual behavior analysis. If customers lost their card and use a new one, the data of the new card must be connected to their existing account; otherwise the data will be useless (Schellinck & Schrans, 2011). Moreover, some casinos have never

implemented the loyalty program, resulting in the loyalty tracking data are not available (Keselj, 2011).

In these situations, data mining methods have been proposed and used to detect patterns of gambling behavior (Keselj, 2011). In this research, all data were collected from the EGMs directly, so the customer tracking data and demographical information are unavailable. Researchers therefore applied page-stay-time-based session identification method to identify gambling behavior sessions based on the individual events recorded by the EGMs. Each identified session refers to the group of gambling activities performed by a player from the moment he started playing the EGM to the moment he stopped it. Thus, each session corresponds to one player and the analysis of sessions will reveal the gambling behavioral characteristics of individual players (Keselj, 2011).

1.2 THESIS OBJECTIVES

After the sessions are identified and collected, the data is prepared for analysis. Instead of only using traditional statistical methods, we employ both statistical and data mining techniques in our research to interpret the data, and more importantly to discover the hidden information and relationships in the data. The thesis objectives are:

- Use the clustering to distinguish the different levels of EGM players and understand their gambling behavioral characteristics.
- Apply the logistic regression to identify which gambling behavior is highly associated with gambling addiction.
- Adapt the decision tree to derive rules that can be used to assign new players into different player groups and to predict at-risk as well as problem gamblers.

We first clean the raw data by using the statistical methods in SPSS to improve the data quality. Then, the k-means clustering technique in SPSS is adapted to partition all players into four player groups. Finally, we employ the logistical regression and decision tree models in SAS E-Miner to identify which gambling behavior is highly associated with gambling addiction and to obtain the predictive rules.

Actually, the ultimate goal of this thesis is to provide the information to EGMs manufactures that may redesign their EGMs and set up the warning system by applying the findings and rules obtained from this research to avoid risky and problem gambling.

1.3 OUTLINE OF THE THESIS

Chapter 2- Literature Review presents knowledge carried out by researchers in previous studies regarding data mining, and its use on customer behavior analysis and problem gambling. The data mining methodologies discovered and discussed are referred in designing and developing experiments.

Chapter 3 – Methodology illustrates the experiment design and the research methodology. The experimental framework with brief description is presented to explain the data analysis procedure.

Chapter 4 – Results and Discussion presents the results of the data analysis, including the detailed profiles of each gambler cluster, the relationship between gambling behavior changes and the development of gambling problem, and the predictive rules to predict at-risk and problem gamblers. Some discussions about the results are included in this chapter.

Chapter 5 – Conclusion is used to indicate the overall objectives of the research, to discuss the limitations, and to summarize both empirical contributions and practical implications. Some thoughts and suggestions for future research are also given in this chapter.

CHAPTER 2 LITERATURE REVIEW

Although data mining has been widely used in many fields, the application of it for problem gambling behavior analysis is still in infancy as very few empirical studies have been found. For this reason, we start the chapter with a general data mining overview in Section 2.1. Follow this, all data mining methods and models applied in our research are discussed in Section 2.2. In Section 2.3, the use of data mining techniques for customer behavior analysis is reviewed since gambling behavior analysis is the application of customer behavior analysis in the gambling industry. Finally, we review researches in relation to data mining for gambling behavior analysis in Section 2.4, though very few papers are found.

2.1 DATA MINING OVERVIEW

Data mining is the application of several specific algorithms to analyze large quantities of data, extract patterns or correlations hidden in data set, and transform data into an understandable structure for further analysis and decision making (Fayyad, Shapiro, & Smyth, 1996). Data mining is popularly treated as a synonym or an essential element of knowledge discovery in databases, which refers to "the process of identifying valid, novel, potentially usefully and ultimately understandable patterns in data" (Remondino & Correndo, 2005).

Today, data mining has been widely used in business, as increasing number of organizations have already realized benefits of data-driven decision making. By collecting and analyzing data, organizations are able to uncover hidden patterns in historical data to forecast sales, generate new marketing campaigns, and accurately analyze customer behavior (Alexander, 1997).

2.2 DATA MINING PROCESS

Data mining is an iterative process, which typically involves several important aspects including problem definition, data preparation, modeling and deployment (Linoff & Berry, 2011).

2.2.1 Problem Definition

Each data mining project starts with understanding the project objectives and identifying the problems that need to be solved. Once the problems have been specified clearly, it is necessary to translate them into data mining problems (Linoff & Berry, 2011).

For example, we aim at solving three problems in this research, which are:

- How to distinguish the different levels of EGM players.
- How to identify which gambling behavior is highly associated with gambling addiction.
- How to obtain predictive rules to identify and predict potential at-risk and problem gamblers.

When they are translated into data mining problems, they become:

- How to separate session data into different clusters in a reasonable way.
- How to find out the hidden relationship between the input variables and the target variable.
- Which predictive model is more appropriate for generating predictive rules and which rules are the best for differentiating at-risk and problem gamblers from their recreational counterpart.

2.2.2 Data Preparation

The purpose of data preparation is to examine and transform raw data in order to make them mean more and improve the quality of data. Without data preparation, the hidden information is not easily accessed by data mining models (Pyle, 1999). On the other hand, some data errors, particularly outliers or unreasonable data, can have a negative impact on results. Data preparation process is composed of different parts, but we only review three related portions that are creation of derived variables, detection of outliers, and transformation of data.

1. Derived Variables

The creation of new derived variables is about generating new variables or converting existing variables to make the presented information more visible and to express more hidden information in the data set. Transformation methods, such as turning numeric values into percentile or replacing categorical variables with numeric ones, are commonly used by researchers to deal with single variable. But more derived variables are generated by combining two existing variables, since more hidden information in both variables can be uncovered (Linoff & Berry, 2011).

2. Outliers Detection

In our research, outlier detection and removal is the most important task in the stage of data preparation as we identify several unreasonable items in the raw data set. The existence of those unreasonable data points will introduce complexity into data models, and finally reach erroneous conclusions. Due to this reason, we conduct more detailed review and compare different methods in order to find out the most suitable technique.

Outliers refer to those data points that are considerable dissimilar in a data set. Although various outlier detection methods have been proposed, most of them can be classified into four categories, which are distribution-based, density-based, distance-based and clustering-based categories. With regard to our research, the distribution-based as well as density-based approaches are not suitable. Distribution-based methods, such as Standard Deviation and Boxplot, are mainly applied to deal with univariate data set (Jayakumar & Thomas, 2013), but our data set is multivariate with several variables. Density-based approaches are usually used in a data set that is not as large as our data set (Patra, 2012). We thereby review the other two categories that are appropriate in our research.

(1) Clustering-based Approaches

Clustering-based outlier detection approaches involve a clustering step, which partitions each observation into different clusters based on the similarity of characteristics. These techniques rely on a key assumption that normal observations gather to form large clusters, while those observations that are highly different from normal instances belong to small clusters (Pachgade & Dhande, 2012).

However, some researchers argued that the clustering algorithms, particularly the k-means algorithm, are inappropriate for detecting outliers and should be avoided to use since they are sensitive to outliers (Jayakumar & Thomas, 2013).

(2) Distance-based Approaches

Among all distance-based methods, a classical and commonly used approach to detect outliers in a multivariate data set is the Mahalanobis distance (MD) technique. The MD describes the distance between each observation and the center of mass. If a data point is located on the center

of mass, its MD is equal to 0; otherwise, the MD is more than 0. If the MD of an observation is more than 0 and exceeds the threshold value, this observation can be regarded as an outlier (Matsumoto, Kamei, Monden, & Matsumoto, 2007). Then the next important step is to determine the threshold value.

The MD analysis follows the Chi-Square distribution, so the Chi-Square critical values table is used as a means to determine the threshold. The threshold is decided by the significance level (p) and degrees of freedom (df). The significance level is usually set at 0.05 (p=0.05), which is the most commonly used number and has already been accepted as a standard by researchers (Taylor, 2013). Unlike the Chi-Square test, the MD is evaluated with the degrees of freedom equal to the number of independent variables involved in the calculation (df=n) (Northern Arizona University, 2002).

Although the MD approach has been commonly used, some researchers pointed out that it is not appropriate to deal with outliers in a large data set, since the distance between observation and the center of the whole data set needs to be calculated which increase the computation time but decrease the accuracy (Pachgade & Dhande, 2012).

In order to overcome the disadvantage, Pachgade and Dhande (2012) proposed a robust method, named hybrid approach, which combines the clustering-based technique and distance-based approach together.

(3) Hybrid Approach

In this method, the clustering-based technique is applied first to segment all observations into several clusters and to calculate the centroid of each cluster. Then the MD approach is employed

to calculate the distance between each observation and the centroid of the cluster that the observation belongs to (Pachgade & Dhande, 2012).

3. Data Transformation

Data transformation technique refers to the use of a mathematical function *f* to replace each data point *x* to the transformed value $y = f(x)$, ensuring that highly skewed data can be transformed to be less skewed and more symmetric (Ambrosius, 2007). It is necessary to transform the data set in our research by using this technique as the MD method requires the symmetrically distribution of data (Franklin & Thomas, 2000). There are numbers of data transformation approaches have been developed, but the most frequently used one is log transformation or log base-10 technique.

2.2.3 Data Mining Models

Data mining models are usually divided into two categories: descriptive and predictive models. Descriptive models are applied to get an understanding of general prosperities of a data set and to find out subgroups in the bulk of data based on the similar characteristics of observations (Shodhganga, 2013). Descriptive models mainly include clustering, summarization, association rule and sequence analysis (Remondino & Correndo, 2005). We review the clustering technique in more detail later since it is applied in our research to separate sessions into subgroups.

Predictive models are carried out to perform inference and to make prediction for future outcomes based on the current data and historical records. Predictive models are mainly composed of decision tree, regression, neural network, estimation, and time series analysis (Shodhganga, 2013). The regression, decision tree and neural network are reviewed as they are selected to conduct predictive analysis in our research.

1. Clustering

Clustering is a well-known descriptive model, which partitions data into a set of clusters or subgroups, such that those observations with similar characteristics are gathered together. Therefore, a cluster is composed of those data items which are similar to each other but dissimilar to those points in other groups (Correa, González, Nieto, & Amezquita, 2012). A good cluster model is supposed to ensure that the intro-cluster similarity is high but the inter-cluster similar should be low (Han & Kamber, 2006).

Many different cluster algorithms have been developed, but the most widely used is the k-means algorithm that is also used in our research to segment all sessions into different player groups. K-means algorithm attempts to partition *n* observations in a data set into a *k* number of clusters in which each observation belongs to a cluster with the nearest centroid. (Correa, González, Nieto, & Amezquita, 2012).

Furthermore, clustering is often an important starting point to other forms of data modeling (Linoff & Berry, 2011). In our research, for example, we apply the clustering technique to separate players, but more importantly, the cluster label generated by the cluster analysis is used as the target variable in the predictive models.

2. Decision Tree Model

The decision tree model is one of the most powerful and widely used predictive models. The decision tree involves a two-step process to form a predictive model. First, the decision tree selects the most important predicators from all input variables to progressively split observations into smaller and smaller subsets that have similar values. Then, the model generates and uses the splitting rules to partition each observation into different classes (Linoff & Berry, 2011).

Therefore, the decision tree can not only be used to determine the most important predicators but also be applied to predict the future by using the rules.

3. Regression Model

Regression models are usually adapted to find out which predictors are highly related to the target variable, and how the changes of predicators affect the target variable. Regression models are most effective when they are used to predict a data set having a large amount of observations but small number of variables. Furthermore, regression models work well to predict a data set when predicators and the target variable have causal relationship and the changes between them are expected to be predictable (Armstrong, 2012).

Linear regression and logistic regression models are two commonly used regression models. In our research, we apply the logistic regression rather than linear regression as the latter is mainly applied to predict the relationship between one input variable and the target variable with one category (Linoff & Berry, 2011).

4. Neural Network Model

Neural Network models are also commonly used for classification and prediction. However, compared with decision tree and regression models, neural network models have a limitation. The decision tree generates a set of rules that can be easily applied to predict future events efficiently and are interpretable by researchers. The regression model assists us to judge which input variables are critical and how they have an impact on the target variable. However, it is impossible to obtain clear rules and to understand the relationship between predicators and the target variable by using the neural network model (Linoff & Berry, 2011).

2.3 DATA MINING FOR CUSTOMER BEHAVIOR ANALYSIS

Customer behavior analysis conducts the study of purchasing behavior, churn behavior, customer satisfaction and loyalty. By analyzing customer behavior, companies are able to deeply understand their customers, identify profitable customers and accordingly retain them with unique and target-specific marketing campaigns (SAP AG, 2014). Recently, customer behavior analysis has evolved from simply describing presented information to discover hidden patterns in behavioral data by using data mining techniques (Intoweb, 2014). Different data mining techniques are applied to achieve different purpose of analysis.

With cluster analysis, companies can easily separate customers into distinct but internally homogeneous segments based on their similar characteristics (Hsieh & Chu, 2009). For instance, customers are usually clustered into three categories that are non-users, light-users, and heavy-users based on the frequency of use for products and services. Those heavy-users who account for less than 2% of all customers contribute more than 25% of company's revenue (Perfetto & Woodside, 2009). Thus, the deep understanding of each customer segment profile aids a company in generating unique strategies and marketing campaigns to effectively and successfully target as well as retain those profitable customers.

Besides clustering technique, predictive analysis models such as decision tree, neural networks, and regression models are frequently applied to predict potential behavior and future activity (Person, 2012). For example, companies use decision tree models to analyze customer churn behavior, apply association rules to analyze purchasing behavior, and adapt regression models to analyze customer satisfaction and loyalty (Intoweb, 2014).

2.4 DATA MINING FOR GAMBLING BEHAVIOR ANALYSIS

Like many other companies, casinos commonly utilize data mining techniques to segment players, identify the best customers, predict their future worth and measure the performance of promotional campaigns. The most widely used data mining models in gambling industry are clustering, regression and decision tree. Cluster analysis is typically used to segment players based on their similar gambling behavior as well as demographic information. By separating players into different segments, casinos are easily to identify best-of-breed players and therefore create targeted marketing campaigns to effectively retain those profitable players and to increase their return on investment (Person, 2012). Decision tree model is widely used by casinos to follow up and measure the performance of marketing campaigns. By using all related factors to construct a decision tree model, casinos can quickly find those factors that are highly related to player's response. In terms of the prediction of player's future worth, the most common and effective technique is regression analysis. Regression models, particularly multiple regression models, can be built by using a large amount of historical data and a variety of predictors such as theoretical and actual win, time on device, average bet, gender and age range (Sutton, 2011).

Although data mining techniques have been used in gambling industry, there are very few studies can be seen in terms of using data mining for identifying at-risk and problem gamblers. We intended to review studies about data mining for EGM gambling behavior analysis, but it was impossible to find any published paper about this topic, though EGM play has been regarded as a main cause of gambling addiction. Therefore, we reviewed some related researches about data mining for identifying online at-risk and problem gamblers.

Braveman and Shaffer (2010) used the k-means clustering technique on online gambling data to separate all players into four gambler groups based on their similar patterns of gambling behavior. Four gambling behavior indicators that are "frequency (total number of active days), intensity (total number of bets per day), variability (standard deviation of bets) and trajectory (the tendency of bets change)" were applied in the analysis. Finally, they demonstrated that at-higher-risk gamblers were those players who gambled more intensively and frequently than others.

Based on their research, Dragicevic, Tsogas and Kudic (2011) also used the k-means clustering model and similar risk indicators to analyze different Internet gambling data set and draw similar conclusions. However, they pointed out that their results might be inaccurate since they did not conduct outlier detection and the k-means technique has difficulty dealing with the data set with a lot of outliers. They also mentioned that some alternative techniques such as regression analysis could be more appropriate for handling this type of data set including a lot of outliers. Furthermore, they suggested using behavioral or risk indicators in regression models to predict players with which behavioral characteristics are easier to self-exclude themselves from online gambling.

Relying on the data obtained by Braveman and Shaffer (2010), Philander (2013) evaluated nine data mining algorithms used in classification and regression models to determine which one is the most effective model for identifying online problem gamblers. He found that many algorithms were able to correctly identify problem gamblers when they were applied on training sample data, but they showed decreased performance when they applied on new samples. Therefore, he pointed out that it is necessary to separate the data set into different sub-samples and to have validation and test samples when classifying gamblers.

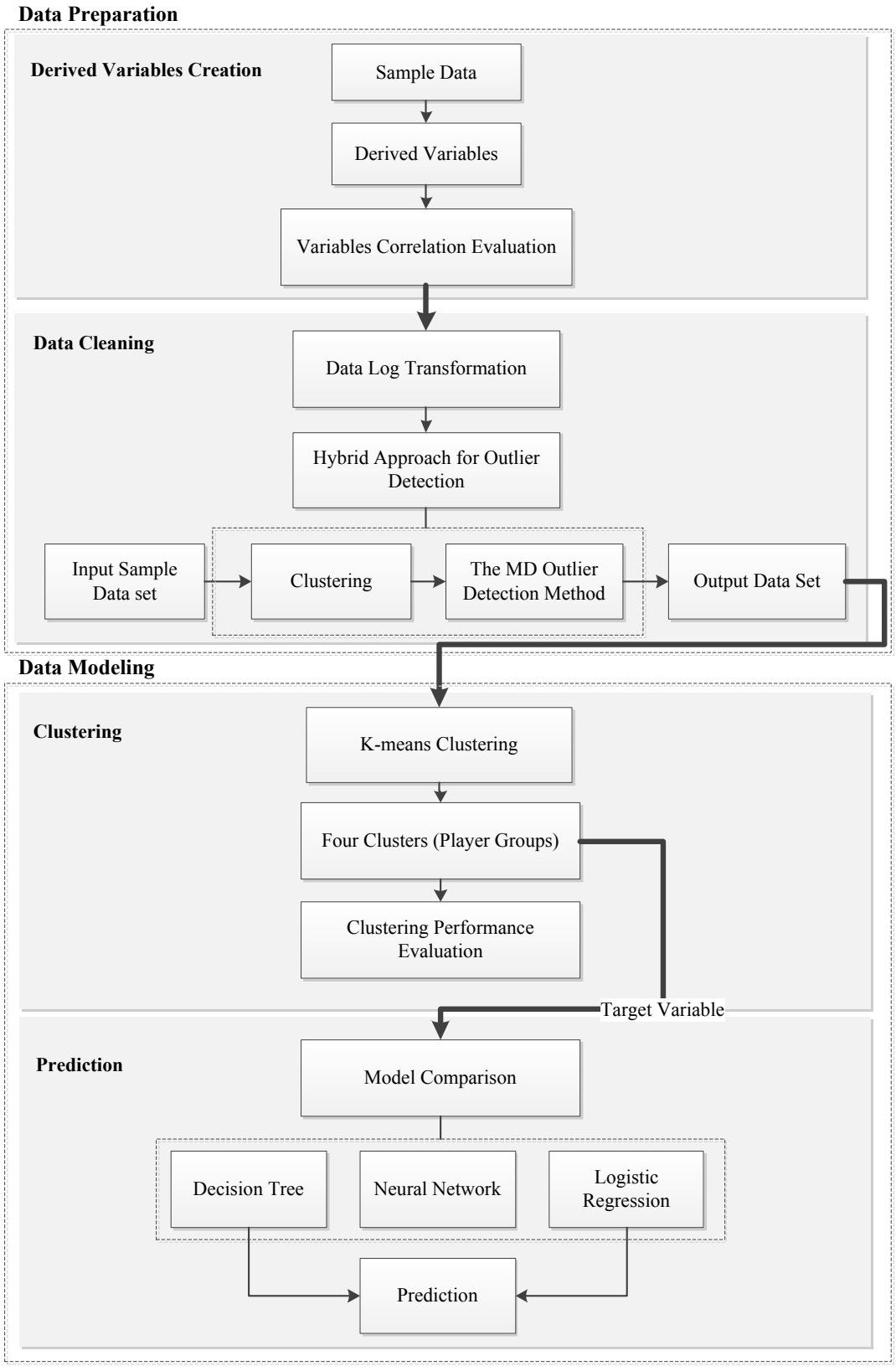
CHAPTER 3 METHODOLOGY

In this thesis, we employ data mining techniques along with statistical methods to prepare and analyze the session data in an attempt to distinguish and understand different levels of EGM players, to identify which gambling behavior is highly associated with gambling addiction, and finally to derive predictive rules for predicting at-risk and problem EGM gamblers.

Figure 1 illustrates the overall flow of data analysis, which consists of two major modules: data preparation and data modeling. In this schema, we start with the data preparation to uncover more hidden information in the data by creating derived variables, and to clean the raw data set by using the hybrid outlier detection method. When the data set is prepared for modeling, we adapt the clustering and predictive models to perform the analysis. First, the k-means clustering technique is adapted to partition sessions into four player groups, at the same time, the target variable used in the predictive models is also generated. Then, the logistic regression, decision tree and neural network models are compared in order to select the effective ones. Finally, the logistic regression model is applied to identify which behavioral indicators are highly associated with gambling addiction because of its highest overall accuracy. Then the decision tree model is constructed with the cluster label as the target variable to develop predictive rules that will be used to predict potential at-risk and problem gamblers.

Two data mining tools, SPSS and SAS Enterprise Miner (E-Miner), are selected to perform the analysis. SPSS is mainly used for preparing and clustering data. SAS E-Miner is applied for carrying out predictive analysis.

Figure 1: Research Schema of Data Analysis



3.1 SAMPLE DATA

The population is usually too large to research all values, thus researchers typically collect sample data with manageable size from the population in order to conduct the analysis in a fast and effective manner (Columbia Cnmtl, 2013). In our work, researchers used the page-stay-time-based session identification technique to collect one month of data from 288 EGMs and to form the sample data set which is composed of 46,514 anonymous sessions and four variables:

- Duration – duration of a session in seconds
- Bets per minute – the average number of bets per minute
- Redeemed – the total amount of money put per gambling session
- Vouchers – the total amount obtained from the EGM in the voucher form

3.2 DATA PREPARATION

In the raw data set, not all variables and their values are valid for data modeling. We only want to keep the most relevant information and correct data. Thus, it is necessary to prepare and clean the raw data set to increase the relevance of information and the accuracy of data.

3.2.1 Derived Variables

Although the four variables (‘Duration’, ‘Bets per minute’, ‘Redeemed’, and ‘Vouchers’) involved in the raw data set are indicative of behavioral characteristics of a player to some extent, we consider that they are not sufficient. Therefore, based on the previous researches we create three derived variables by using the transformation and combination methods.

DurationMin. In almost each gambling related research, duration was regarded as one of the most important indicators to distinguish problem gamblers from recreational players. Thus, it is necessary to maintain this variable in our work. However, we notice that in most papers, researchers measured duration or time spent by minutes rather than by seconds. So, we transform 'Duration' to 'DurationMin' by using the 'Duration' divided by 60 (DurationMin = Duration/60).

TotalBets. Braveman and Shaffer (2010) found out that betting activity can serve as a main behavior marker to predict the development of online problem gambling and to differentiate high-risk gamblers from their low-risk counterparts. They found that at-higher-risk gamblers gambled more intensively (total number of bets per day) than others. Building on this research, Dragicevic, Tsogas and Kudic (2011) applied betting activity as a key indicator and put forward 'Total number of bets' and 'Number of bets per day' as two main variables to separate online players into different groups.

In our research, the indicators of betting activity need to be involved as problem gamblers typically present common behavioral characteristics no matter what types of gaming they are engaged in. Given the 'Bets per minute' variable has already been generated; we create the 'TotalBets' variable in order to evaluate the total number of bets a gambler played in a session. The 'TotalBets' is created by using the 'Bets per min' times 'DurationMin' (TotalBets = Bets per min × DurationMin).

Loss. We combined the 'Redeemed' along with the 'Vouchers' to create a new variable named 'Loss' in order to further evaluate gamblers' total money spent at the end of each visit. Some previous studies pointed out that the behavior of chasing losses is a key sign of problem

gambling. When players lost money, at-risk and problem gamblers choose to play more rather than stop in order to win their losses back. The increased gambling is more likely to result in more losses, consequently leading to gambling addiction. On the other hand, recreational players know when to stop (Gambling: Help and Referral, 2013). Thus, comparing losses of different levels of players assists us to differentiate at-risk and problem gamblers from their recreational counterparts. The 'Loss' variable is generated by using the 'Vouchers' minus the 'Redeemed' (Loss = Vouchers - Redeemed).

By generating three derived variables, we totally obtain six variables:

- DurationMin—duration of a session in minute
- BetsPerMin—the number of bets per minute
- TotalBets—the total number of bets per session
- RedeemedPerS—the total amount of money put per session
- VouchersPerS—the amount obtained from EGM in the voucher form
- Loss—the total amount of loss (or win) per session

It is necessary to perform the relevance analysis before importing all six variables into the data mining models to avoid using highly correlated variables which impart nearly exactly the same information. If those highly correlated variables are imported into the models, the regression coefficients in regression models will be unreliable and unstable (Allison, 2012) and the clusters in clustering model will be indistinct (Mooi & Sarstedt, 2011).

Pearson's correlation coefficient analysis is usually applied to measure the relation between two variables. According to Pearson's theory, if the correlation value between two variables is 1.0 or -1.0, these two variables are totally positive or negative correlated. In this regard, if the absolute

correlation value between two variables is above 0.9, they are highly correlated and problematic (Mooi & Sarstedt, 2011). Table 1 shows that the 'Loss' and 'VouchersPerS' are highly correlated variables since the value between them is 0.962. Thus, one of them needs to be eliminated to ensure that all the clusters are unique.

Table 1: Result of correlation detection analysis

Correlations

		DurationMin	BetsPerMin	TotalBets	RedeemedPerS	VouchersPerS	Loss
DurationMin	Pearson Correlation	1	-.241**	.093**	.050**	-.013**	-.031**
	Sig. (2-tailed)		.000	.000	.000	.005	.000
	N	46514	46514	46514	46514	46514	46514
BetsPerMin	Pearson Correlation	-.241**	1	.410**	.240**	.108**	.013**
	Sig. (2-tailed)	.000		.000	.000	.000	.004
	N	46514	46514	46514	46514	46514	46514
TotalBets	Pearson Correlation	.093**	.410**	1	.567**	.093**	-.125**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	46514	46514	46514	46514	46514	46514
RedeemedPerS	Pearson Correlation	.050**	.240**	.567**	1	.107**	-.276**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	46514	46514	46514	46514	46514	46514
VouchersPerS	Pearson Correlation	-.013**	.108**	.093**	.107**	1	.926**
	Sig. (2-tailed)	.005	.000	.000	.000		.000
	N	46514	46514	46514	46514	46514	46514
Loss	Pearson Correlation	-.031**	.013**	-.125**	-.276**	.926**	1
	Sig. (2-tailed)	.000	.004	.000	.000	.000	
	N	46514	46514	46514	46514	46514	46514

**. Correlation is significant at the 0.01 level (2-tailed).

We decide to maintain the 'Loss' instead of using the original 'VouchersPerS' variable because of two reasons. First, as previously mentioned, chasing-loss is a key sign to identify at-risk and problem gamblers. The more money a gambler lost, the more likely he or she will develop a gambling problem. Second, the 'Loss' variable is generated by combining the 'Redeemed' and 'Vouchers' variables. The combination of two related variables makes a description of behavioral profiling and a prediction of problem gamblers more effective and accurate (Schellinck & Schrans, 2011).

After eliminating one variable, we consequently obtain five variables that are 'DurationMin', 'BetsPerMin', 'TotalBets', 'RedeemedPerS', and 'Loss'. These five variables can be divided into three behavioral indicator types that are duration, betting activity, and money spent.

Table 2: Summary of risk indicator types and indicators

Indicator Type	Indicator	Meaning
Duration	DurationMin	total time spent in a session
Betting Activity	BetsPerMin	the number of bets per minute (intensity of gambling)
	TotalBets	the total number of bets per session
Money Spent	RedeemedPerS	the total amount of money put in a session
	Loss	the total amount of lose (or win) at the end of a session

3.2.2 Preliminary Data Analysis

The main purposes of preliminary data analysis are to clean the raw data set by investigating and deleting obviously incorrect or unreasonable values, and to understand the general properties of the data set by describing key features of the data (Blischke, Rezaul Karim, & Prabhakar Murthy, 2011).

We first sort the data set by each variable to investigate whether some obviously unreasonable data exist. When sorting all data by the 'BetsPerMin' variable in ascending order, we find that the first five data points of this variable seem to be unreasonable. The value of these five data points are equal to 0 which indicate that EGMs were not played during the whole session. It is possible that they are generated from errors occurring at session identification stage. To avoid their negative impact on the further analysis, we remove them and obtain a new data set including 46,509 sessions.

Table 3: Noisy data in the *BetsPerMin* variable

DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
1246.05	0.00	0.00	10.00	-10
1051.02	0.00	0.00	5.00	-5
1184.85	0.00	0.00	10.00	-10
721.92	0.00	0.00	10.00	-10
1496.37	0.00	0.00	10.00	-10
1510.90	0.01	15.11	5.00	-5
5501.55	0.01	55.02	10.00	-10
753.38	0.01	7.53	5.00	-5

After removing unreasonable observations, we carry out the descriptive data summarization analysis to measure the central tendency, dispersion, and skewness of the data set to investigate whether outliers exist (Han & Kamber, 2006).

Central Tendency. The measure of central tendency focuses on measuring the mean and median (Han & Kamber, 2006). As the mean is calculated by adding each item up and divided by the number of total items, it is sensitive to extreme values. On the other hand, the median is the value of the data point that is in the middle of the sorted list in ascending order, so the median is relatively unaffected by the extreme data involved in a data set. If the mean and median are nearly the same, the data set is roughly symmetric, otherwise, it indicates that this data set consists of extreme data or outliers (Police Analyst, 2012).

As can be seen, the mean and median of each variable, particularly the 'DurationMin', 'TotalBets', and 'RedeemedPerS', is different, that is, extreme data or outliers exist in the data set especially in these three variables.

Table 4: Central tendency of the data set

	DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
Mean	45.15	16.29	373.21	70.45	-20.11
Median	15.71	18.65	169.03	30	-19.9

Dispersion. The measure of dispersion of data focuses on measuring standard deviation, maximum and minimum (Han & Kamber, 2006). Standard deviation is regarded as a significant indicator of the presence of outliers because it is largely influenced by extreme data. A data set with a small standard deviation has a narrow spread of observations around the mean and accordingly has comparatively few outliers. But if a data set has a high standard deviation, it consists of extreme values (Statistic Canada, 2013). The high standard deviation of each variable, except the 'BetsPerMin', in Table 5 indicates the existence of outliers.

When looking at the maximum and minimum values of each variable, we notice that the maximum value of the 'DurationMin' (16,784 minutes or 279 hours) seems unreasonable since the longest gambling time in one session was 115 straight hours based on the Guinness World Record (Jørgensen, 2013).

Table 5: Dispersion of the data set

	DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
Std. Deviation	152.4	9.87	605.68	150.65	397.05
Minimum	0.167	0.01	1.86	5	-6424.69
Maximum	16784.28	68.32	13281.8	6425	40783.05

Skewness. Skewness is designed to measure the degree of symmetry in the variable distribution. If the skewness is 0, the variable is symmetrical; if the skewness is too large or too small with value far away from 0, the variable is asymmetrical (TexaSoft, 2008). Table 6 shows that the skewness of each variable expect the 'BetsPerMin', is far away from 0. Thus, it is necessary to perform the transformation to make the data set less skewed and more symmetrical before carrying out the outlier detection by using the MD approach.

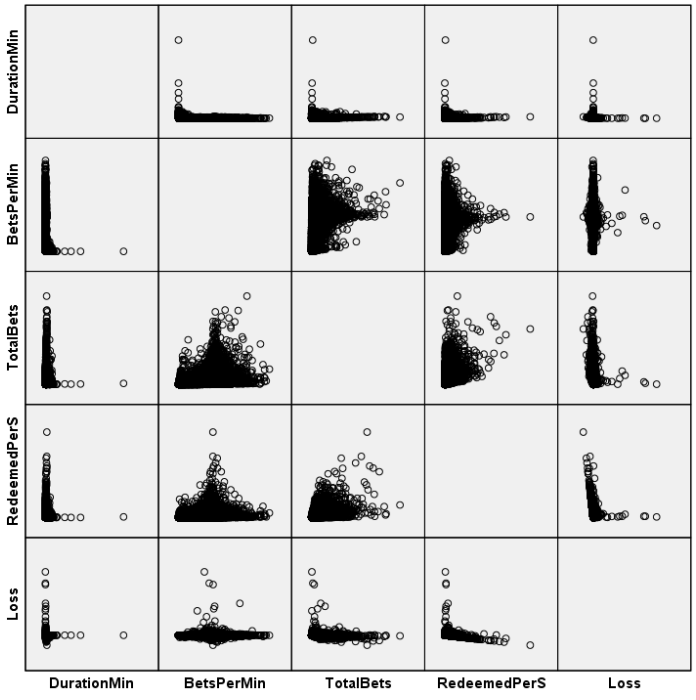
Table 6: Skewness of the data set

	DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
Skewness	36.001	-0.12	5.11	10.46	57.71

The descriptive data summarization technique confirms the existence of outliers or extreme data in each variable (except 'BetsPerMin'). However, it does not show the location of outliers. We thereby apply the graphic displays in attempt to find out where they are.

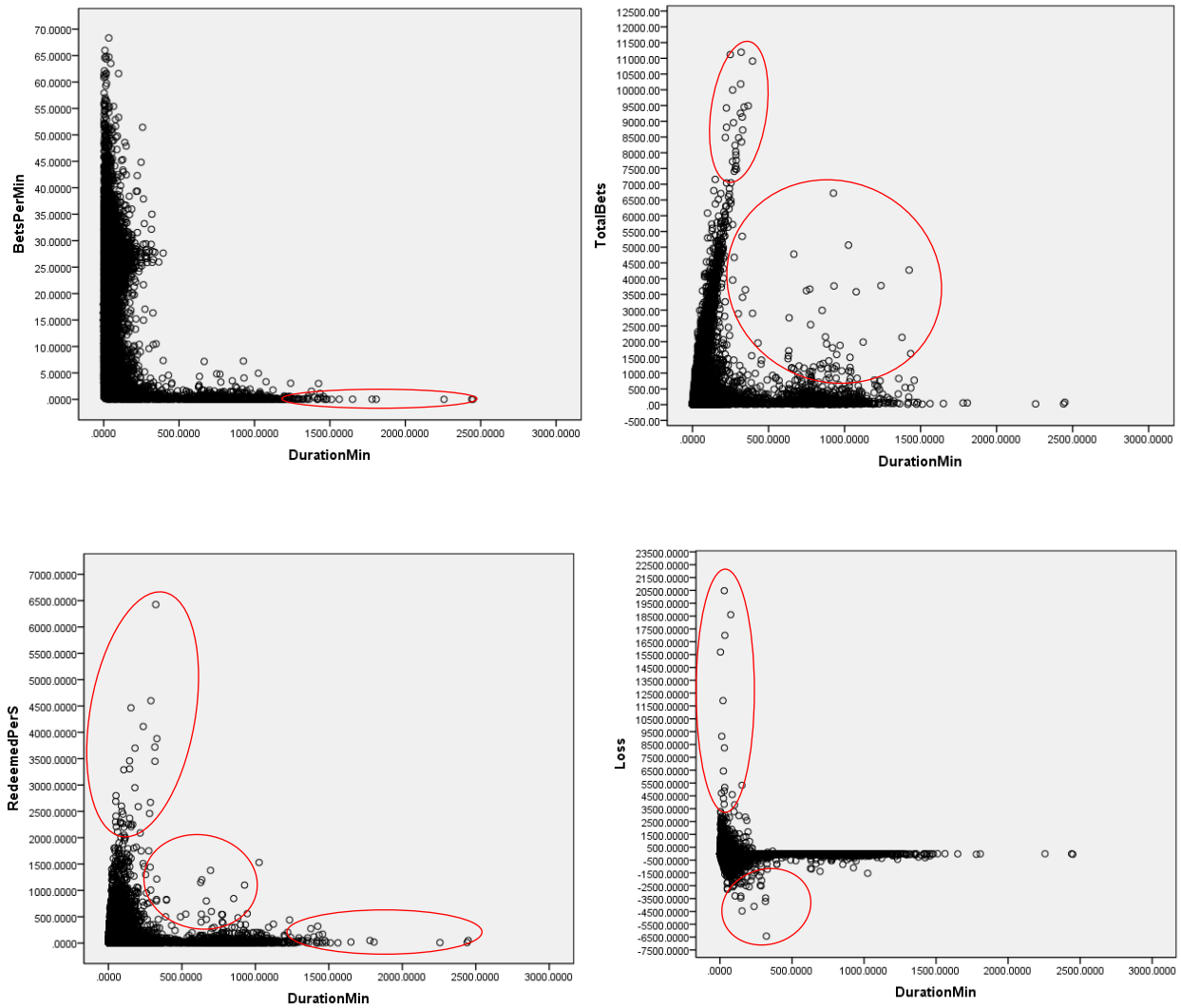
Scatter plot is an effective graphical method for identifying whether outliers are involved in a data set and where they are. In a scatter plot, the data points that are not close to the majority of data can be considered as outliers. Figure 2 is the scatter plot matrix which is a useful extension to the scatter plot when dealing with a data set including more than two variables as it shows the scatter plots of multivariable simultaneously, providing us a general visual impression of outliers (Han & Kamber, 2006).

Figure 2: Scatter plot matrix



However, if there are too many variables in a data set, the scatter plot matrix gets so small plots between each two variable as to be relatively unclear (Psychwiki, 2008). Thus, the regular scatter plots with coordinates are further applied, since the coordinates assist us to find out the exact locations of the outliers.

Figure 3: Scatter plots showing the location of outliers



3.2.3 Data Transformation

Although different techniques have been developed, as previously mentioned, the commonly used is the log 10 method in terms of data transformation. However, this technique has a

limitation as it is only appropriate for dealing with strictly positive data (Wicklin, 2011) but the 'Loss' consists of a few non-positive values. Therefore, all those non-positive data need to be transformed into positive ones; otherwise, the log 10 cannot be applied to deal with our data set. Researchers usually add a constant value in each data item to move the minimum into 1.00 (Osborne & W, 2002). The minimum value of the 'Loss' is -6424.69, therefore we add (6424.69+1) to each item in this variable to ensure that the minimum value is equal to 1.00. Finally, the log 10 formula used for the 'Loss' is $y=log_{10}(x+6424.69+1)$.

We compare the mean and median of each log-transformed variable to evaluate the performance of log transformation. In a perfectly symmetrical distribution, the mean and median are exactly the same. If a distribution is roughly symmetrical, these two values are similar (Gravetter & Wallnau, 2009). It is obvious that the data set has been successfully transformed to be more symmetrical than before since the mean and median of each variable are similar.

Table 7: Descriptive statistics of the log-transformed data set

Statistics

		lgDuration	lgBetsPerMin	lgRedeemed PerS	lgTotalBets	lgLoss
N	Valid	46509	46509	46509	46509	46509
	Missing	0	0	0	0	0
Mean		1.2165	1.0147	1.5118	2.2312	3.8061
Median		1.1964	1.2707	1.4771	2.2280	3.8066

In addition, the histogram is another useful method, which assists us to quickly understand the distribution of a variable. In a symmetrical distribution, most of data are located in the center of the distribution and the histogram is commonly shaped like a bell curve, otherwise, one tail of the histogram is longer than the other (Grasso, 2012). By comparing the histograms of original along with log-transformed variables, we are more certain that the variables have been

transformed to be less skewed and more symmetrical than before, as more data in each variable are now located in the center of the distribution.

Figure 4: Histograms of each variable (left: original variable, right: log-transformed variable)

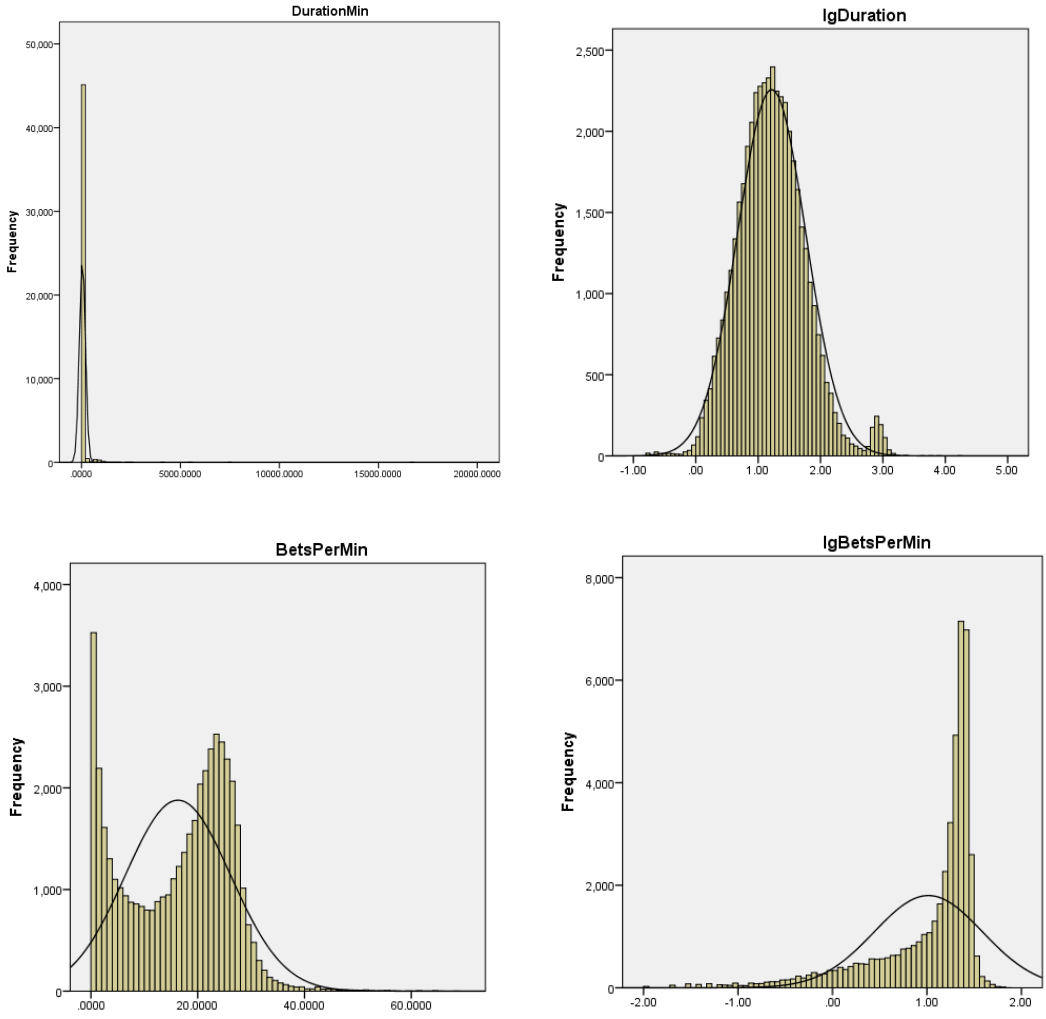
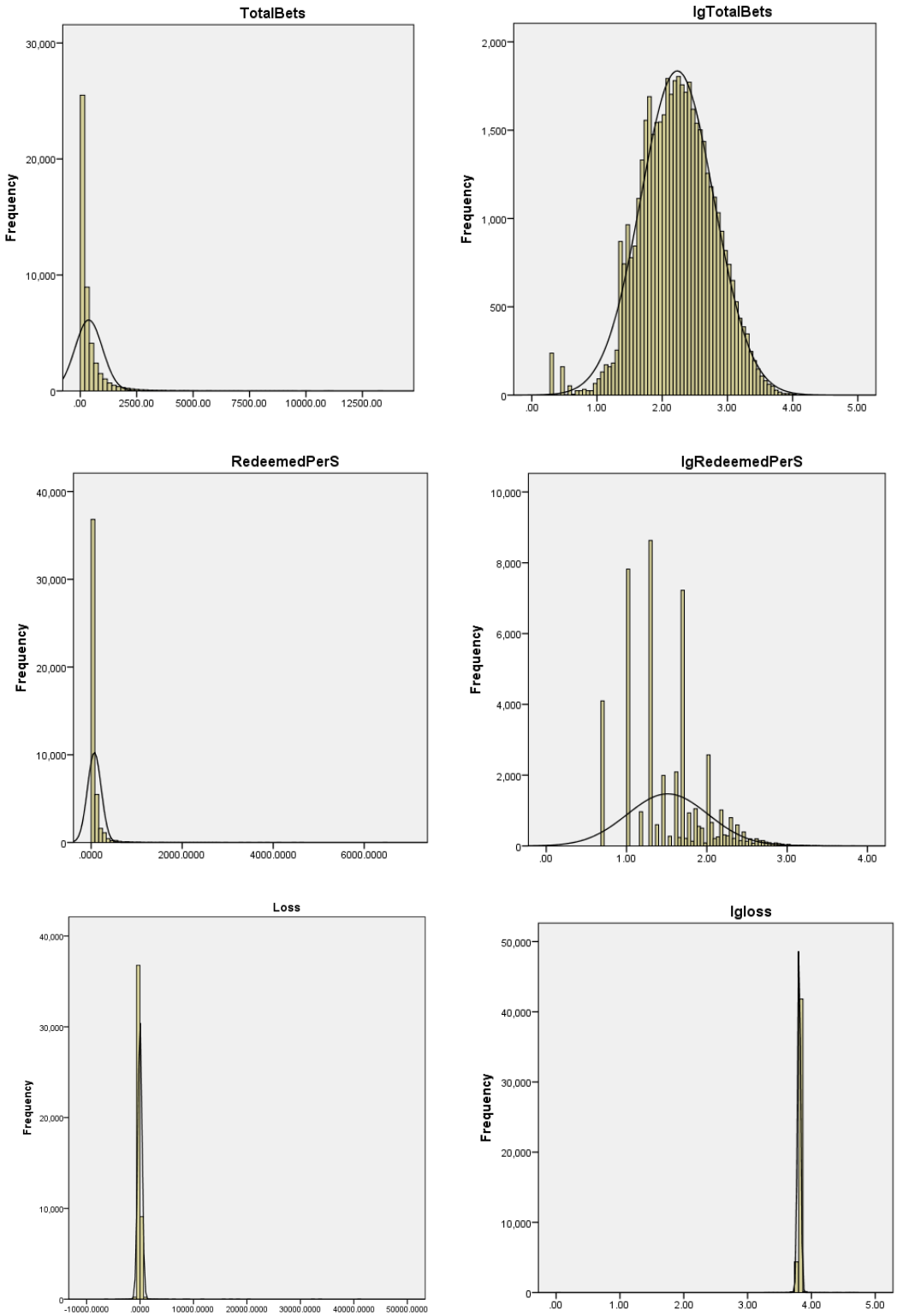


Figure 4: Histograms of each variable (continued)



3.2.4 Hybrid Approach for Outlier Detection

When all requirements of the hybrid outlier detection approach have been met, we carry out this method to detect and remove the outliers from the data set. This approach is composed of two steps that are the clustering-based technique and distance-based technique.

1. Clustering-Based Technique

In the first phase, a clustering method is utilized to partition 46,509 observations into several clusters and to calculate the centroid of each cluster. Instead of using the k-means clustering techniques, we chose the two-step clustering technique in SPSS since we do not know how many clusters should be appropriate and this technique can automatically determine the optimal number of clusters. Figure 5 shows that the two-steps approach automatically separates all sessions into three good-quality clusters and the cluster centers are also calculated.

Figure 5: Result of the two-steps clustering analysis

2. Distance-Based Technique

After adapting the MD technique to calculate the distance between each log-transformed observation and the centroid of the cluster that the observation belongs to, we need to determine the threshold of outliers.

The threshold is determined by two factors that are the significance level (*P*) and degrees of freedom (*df*), which are equal to 0.05 and 5 (five variables were involved in the MD calculation), respectively. Finally, the threshold is determined to be 11.07 by using the *P* and *df* in the Table of the Chi-square distribution. Therefore, any observation having the MD value larger than 11.07 is regarded as an outlier and removed from the data set. After removing 2,093 outliers, we consequently obtain a new data set including 44,416 sessions (95.5% of the sample data).

Table 8: Table of the Chi-square distribution (Marson, 2011)

TABLE D Chi-square distribution critical values

	p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11

The descriptive data summarization technique is applied again to investigate whether the outliers have been removed. The measure of data cleaning performance by checking the statistics of data is considered as a critical step in data quality control process (Chapman, 2005).

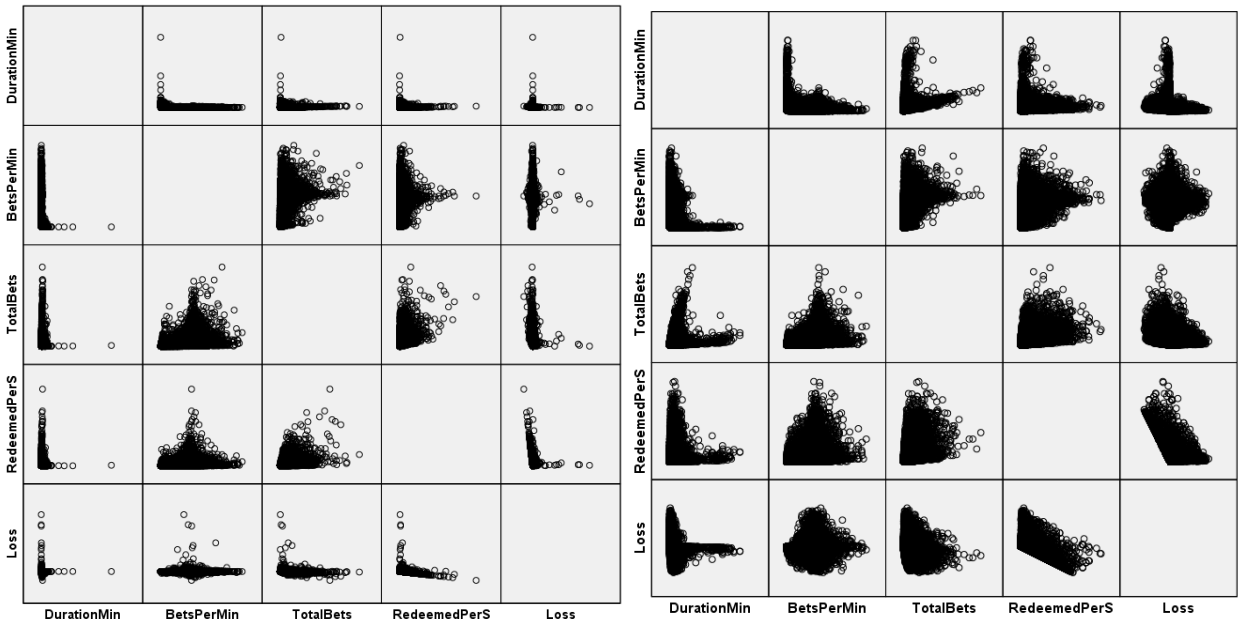
According to Table 9, we observe that the difference between mean and median, standard deviation, as well as skewness of each variable have been decreased, indicating that outliers which had an impact on the data set have been removed successfully. In addition, the most obvious indicator is the maximum of the 'DurationMin', which has been decreased to a reasonable value (1,036 minutes or 17.3 hours).

Table 9: Descriptive statistics of the data set before and after outlier removal

			DurationMin	BetsPerMin	TotalBets	Redeemed PerS	Loss
Central Tendency	Old data set (46,509 sessions)	Mean	45.15	16.29	373.21	70.45	-20.11
		Median	15.72	18.65	169.03	30	-19.9
		Difference between Mean and Median	**29.43**	**-2.36**	**204.18**	**40.45**	**-0.21**
	New data set (44,416 sessions)	Mean	29.14	16.69	360.53	62.39	-23.4
		Median	15.23	18.97	173.98	30	-19.9
		Difference between Mean and Median	**13.91**	**-2.28**	**186.55**	**32.39**	**-3.5**
Dispersion	Old data set (46,509 sessions)	Std. Deviation	**152.4**	**9.87**	**605.68**	**150.65**	**398.05**
		Minimum	0.17	0.1	1.86	5	-6424.69
		Maximum	16784.28	68.32	13281.8	6425	40783.05
	New data set (44,416 sessions)	Std. Deviation	**49.32**	**9.62**	**532.49**	**93.44**	**122.89**
		Minimum	0.47	0.14	3	5	-800
		Maximum	**1036.03**	68.32	9452.62	1230	1170
Skewness	Old data set (46,509 sessions)	Skewness	**36**	**-0.12**	**5.11**	**10.46**	**57.71**
	New data set (44,416 sessions)	Skewness	**7.94**	**-0.151**	**4.07**	**3.91**	**1.02**

The scatter matrix in Figure 6 visually demonstrates that the outliers have been removed from the data set successfully, as those data points that far away from the bulk of data in the old matrix disappear in the new one.

Figure 6: Scatter plot matrix of the data set before and after outlier removal



Consequently, the data set that is composed of five behavioral indicators ('DurationMin', 'BetsPerMin', 'TotalBets', 'RedeemedPerS', and 'Loss') with 44,416 observations is ready for data analysis modeling.

3.3 DATA ANALYSIS MODELING

Two types of data mining tasks are carried out successively to analyze the data set. The k-means cluster analysis is first applied to segment gamblers based on their similar behavioral characteristics and to form the target variable. Then, three most frequently used classification and prediction models: decision tree, logistic regression and neural network are compared in order to choose the best one. Finally, decision tree and logistic regression models are employed to perform predictive analysis.

3.3.1 Cluster Analysis

1. Building the Cluster Model

We adapt the k-means clustering technique in SPSS to separate the observations into different player groups. Although other clustering techniques, such as hierarchical and two-steps can be considered, k-means is the most appropriate technique in our research because of two main reasons.

Firstly, k-means is less computationally demanding than hierarchical in the face of such a large data set containing more than 500 sessions. Secondly, hierarchical and two-steps approaches are normally used in the case of un-predefined number of clusters in order to automatically select the number of clusters. If the number of clusters is pre-determined, k-means is more suitable and highly recommended by researcher (Mooi & Sarstedt, 2011). With regard to our research, we attempt to divide all sessions into four player groups that are non-problem gamblers, low-risk gamblers, moderate-risk gamblers and problem gamblers based on the categories developed by the Canadian Problem Gambling Index (CPGI).

When running the k-means clustering is SPSS, we use the 'iteration and classify' function which automatically run the model 74 times until the stability as well as good quality of the clusters are reached (Correa, González, Nieto, & Amezquita, 2012).

2. Cluster Performance Evaluation

Investigating whether the four clusters are distinguishable is of importance in cluster analysis. Only if they are distinguishable does the cluster analysis perform successfully (Mooi & Sarstedt, 2011).

Typically, the performance of cluster analysis is measured by one-way analysis of variance (ANOVA). However, it is important to notice that one-way ANOVA is only appropriate when all the required assumptions are met; otherwise the result is invalid. The most important assumption in relation to our research is that the data must be or approximately normally distributed (Laerd Statistics, 2013). Three commonly used ways to test the normality of data are graphical, numerical and formal normality test approaches (Razali & Wah, 2011).

Graphical Method. Histograms along with Q-Q plots provide us a quick visual impression of the data and aids in identifying whether the data is normally distributed. If the data is in normal distribution, histograms are supposed to be symmetric with two equal tails and the data points in Q-Q plot are close to the straight diagonal line (Katenka, 2010). But the following histograms and Q-Q plots clearly demonstrate that the data set is non-normally distributed.

Figure 7: Histograms and Q-Q plots of each variable

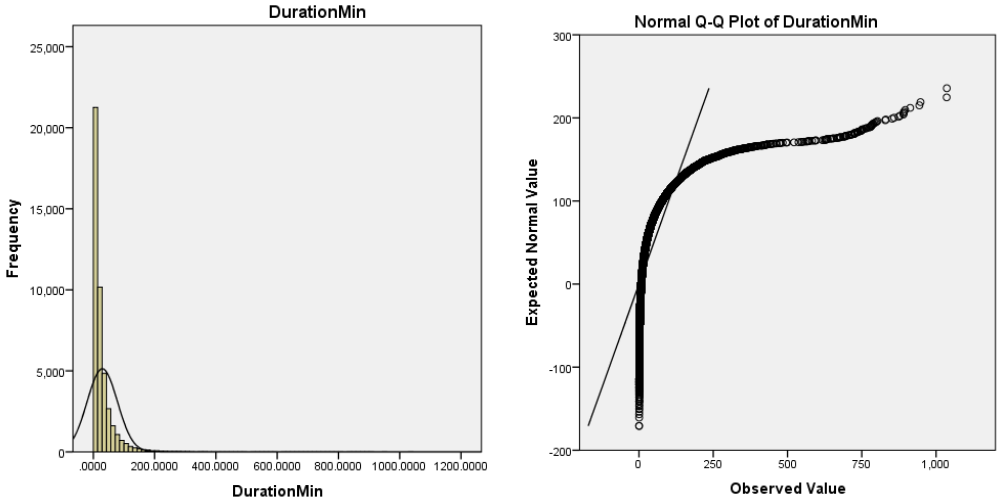


Figure 7: Histograms and Q-Q plots of each variable (continued)

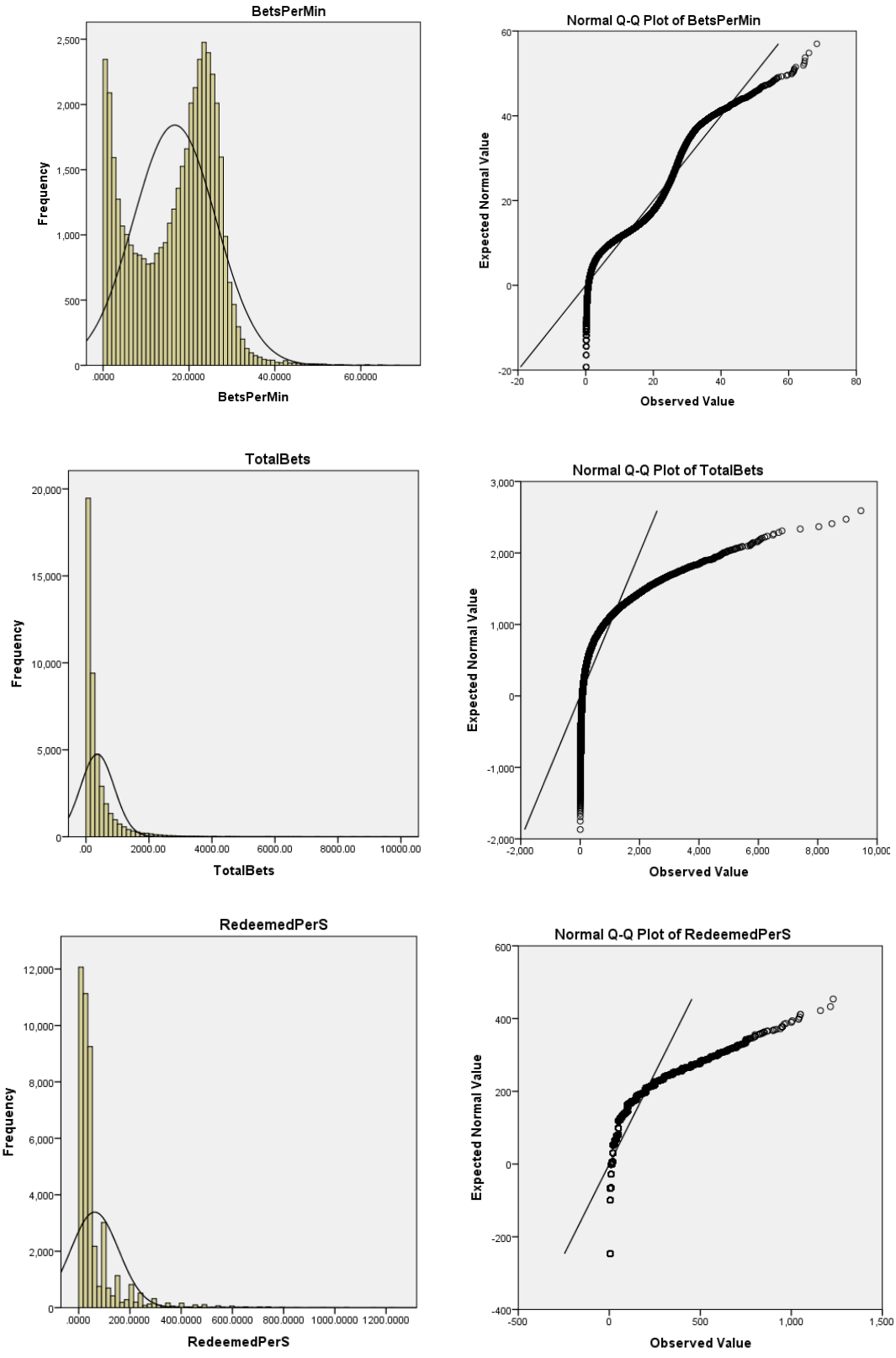
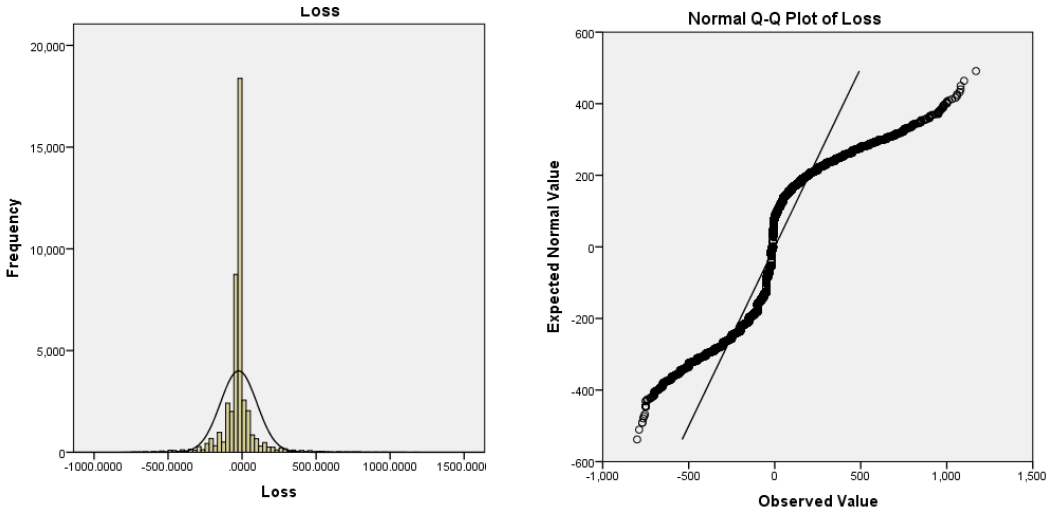


Figure 7: Histograms and Q-Q plots of each variable (continued)



Although the graphical method is the easiest way to get an idea whether the data is normal, some researchers argued that this method is insufficient to support the conclusion. Therefore, numerical approach and formal normality test are needed to provide further proofs (Razali & Wah, 2011).

Numerical Approach. The numerical approach mainly refers to the measure of skewness and kurtosis. If the data is normally distributed, the skewness is 0 and the kurtosis is 3; otherwise, they depart from 0 and 3, respectively (Brown, 2008). It is obvious that the data set is non-normally distributed, as the skewness and kurtosis of each variable are far away from 0 and 3 respectively.

Table 10: Skewness and kurtosis of each variable

Statistics

		DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
N	Valid	44416	44416	44416	44416	44416
	Missing	0	0	0	0	0
Skewness		7.944	-.151	4.071	3.908	1.016
Std. Error of Skewness		.012	.012	.012	.012	.012
Kurtosis		100.651	-.603	26.720	20.987	14.855
Std. Error of Kurtosis		.023	.023	.023	.023	.023

Formal Normality Tests. Shapiro-Wilk test and Kolmogorov-Smirnov test are two most commonly used formal normality tests. The Shapiro-Wilk test has been proved to be the most powerful test for checking the normality of all sample size, whereas the Kolmogorov-Smirnov test is only appropriate for the large data set (more than 2,000 data points) (Razali & Wah, 2011). Since our data set is so large that both two tests are powerful and appropriate in our research.

The null hypothesis of the formal normality tests is that the data is normally distributed. In order to determine whether the null hypothesis is true, we have to calculate the p-value to determine whether the result is statistically significant. If the p-value is no more than 0.05, the null hypothesis is rejected and the data is proved to be non-normally distributed (Maths-Statistics-Tutor, 2010). Table 11 obviously demonstrates that the data is non-normally distributed since the significance values of the tests are far away from 0.05.

Table 11: Results of tests of normality

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Cluster Number of Case	Statistic	df	Sig.	Statistic	df	Sig.
DurationMin	1	.327	8466	.000			
	2	.140	434	.000	.702	434	.000
	3	.304	33088	.000			
	4	.271	2428	.000	.292	2428	.000
BetsPerMin	1	.119	8466	.000			
	2	.151	434	.000	.870	434	.000
	3	.091	33088	.000			
	4	.114	2428	.000	.896	2428	.000
TotalBets	1	.099	8466	.000			
	2	.164	434	.000	.808	434	.000
	3	.113	33088	.000			
	4	.101	2428	.000	.926	2428	.000
RedeemedPerS	1	.199	8466	.000			
	2	.133	434	.000	.917	434	.000
	3	.227	33088	.000			
	4	.143	2428	.000	.862	2428	.000
Loss	1	.153	8466	.000			
	2	.063	434	.000	.988	434	.002
	3	.292	33088	.000			
	4	.091	2428	.000	.958	2428	.000

a. Lilliefors Significance Correction

Based on the results obtained from the three normality assessment approaches, we confidently conclude that the data set is non-normal distributed, thus the one-way ANOVA is not appropriate to evaluate the cluster performance. Instead, an equivalent-parametric (distribution free) test – Kruskal-Wallis H test is applied (Kruskal-Wallis H Test using SPSS, 2013).

The Kruskal-Wallis H test is based on the statistic H that is approximately Chi-Square distributed, so the significance value is set to 0.05 ($p = 0.05$). If the calculated value is less than 0.05 ($p< 0.05$), there exists enough evidence that the clusters are significantly different from each other (VCU, 2013). Table 12 shows that the Sig. value of each variable are less than 0.05, which demonstrates that the cluster technique successfully partitions the sessions into different clusters.

Table 12: Result of Kruskal-Wallis H test

Test Statistics[a,b]

	DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
Chi-Square	12997.101	9188.837	25709.404	13940.481	3821.605
df	3	3	3	3	3
Asymp. Sig.	.000	.000	.000	.000	.000

a. Kruskal Wallis Test

b. Grouping Variable: Cluster Number of Case

3.3.2 Prediction Methods

Before applying predictive models to perform the analysis, we need to determine the target variable, partition data into three subsets to ensure model accuracy, and compare different algorithms of the same model as well as different predictive models in order to select best ones.

1. Target Variable

The clustering technique does not only successfully partition all sessions into four different player groups, but more importantly, generate the target variable for predictive models. Thereby, the target variable used to construct the predictive models is the cluster label including four categories, which are non-problem gambler group, low-risk gambler group, moderate-risk gambler group, and problem gambler group.

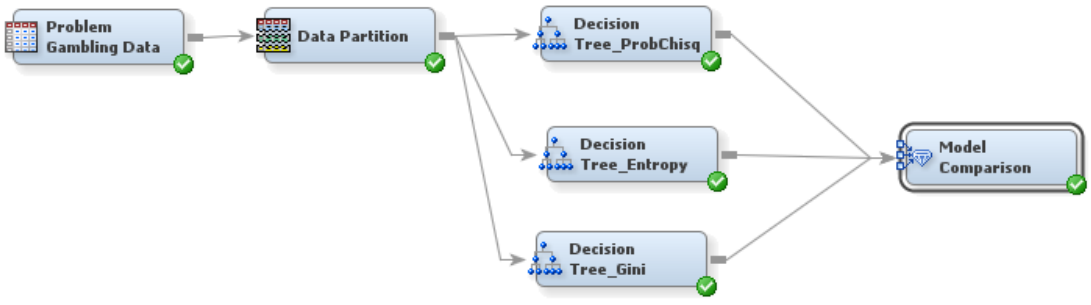
2. Data Partition

It is necessary to partition the data set into three sub-samples: training, validation and test sample. The data partition ensures the patterns generated by the predictive models can occur in the wider data sets (Linoff & Berry, 2011). Based on the general rule, we specify the percentage of 60%, 20% and 20% for the training, validation and test sub-samples, respectively.

3. Selection of Split for Decision Tree

One of the most important tasks in constructing a decision tree model is to select the best split, which determines the best way to separate observations into individual classes (Linoff & Berry, 2011). Accurately selecting the best split has a big impact on the performance of the decision tree. SAS E-Miner provides three multi-split classification rules that are ProbChisq (p-value of Pearson Chi-square statistic), Entropy (information gain) and Gini (population diversity) (Ghoson, 2010). Although SAS E-Miner suggests using Gini to split a categorical target variable, we compare the performance of each split criterion to select the best one.

Figure 8: Decision tree models with different splitting criterion



Performance of a predictive model is commonly evaluated by looking at its accuracy. A classification matrix is typically used by researchers, inasmuch as it is easier to be understood (Schellinck & Schrans, 2011). A classification matrix sorts all cases into different categories by determining whether predicted value matches the actual value (Microsoft, 2013).

From the following classification matrixes, we observe that the decision tree model with Gini rule reaches the highest overall accuracy. Considering the recommendation by SAS E-Miner, we finally determine to use Gini rule to construct the decision tree.

Table 13: Classification matrices of decision tree with different splitting criteria

Gini	Predicted					
Actual		Cluster 1	Cluster 2	Cluster 3	Cluster 4	% Correct
	Cluster 1	1666		20	7	98.41%
	Cluster 2		86			100%
	Cluster 3	9		6609		99.86%
	Cluster 4	3	3		480	98.76%
	% Correct	99.29%	96.63%	99.69%	98.56%	**99.53%**

ProbChisq	Predicted					
Actual		Cluster 1	Cluster 2	Cluster 3	Cluster 4	% Correct
	Cluster 1	1667		19	7	98.46%
	Cluster 2		86			100%
	Cluster 3	10		6608		99.84%
	Cluster 4	3	3		480	98.76%
	% Correct	99.22%	96.62%	99.71%	98.56%	**99.52%**

Entropy	Predicted					
Actual		Cluster 1	Cluster 2	Cluster 3	Cluster 4	% Correct
	Cluster 1	1672		14	7	98.76%
	Cluster 2		86			100%
	Cluster 3	16		6602		99.76%
	Cluster 4	3	3		480	98.76%
	% Correct	98.88%	96.63%	99.78%	98.56%	**99.52%**

4. Selection of Neural Network Hidden Units

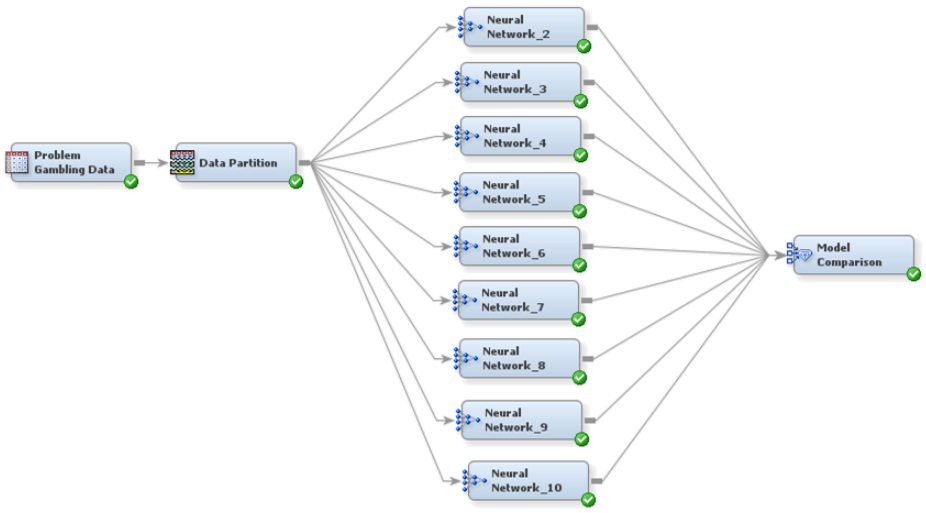
The determination of the number of hidden units plays a critical role in constructing a good performance neural network model. We run the model by using the default value (three hidden units), and the overall accuracy is 99.67%.

Table 14: Classification matrix table of neural network with three hidden units

	Predicted					
Actual		Cluster 1	Cluster 2	Cluster 3	Cluster 4	% Correct
	Cluster 1	1692			1	99.94%
	Cluster 2		86			100%
	Cluster 3	6		6612		99.99%
	Cluster 4	18	4		464	95.4%
	% Correct	98.6%	95.5%	100%	99.79%	**99.67%**

Although the overall accuracy is pretty high, we still run the model nine times by using the different number of hidden units (from 2 to 10, except 3) as it is necessary to run the model many times with different numbers of hidden units to select the number with the best result (Matignon, 2005).

Figure 9: Neural network models with different hidden units



When we calculate the overall accuracy of each model with different hidden units, we find that it is difficult to determine which model is the best since the difference of overall accuracy between each other is too small to be identified. We further investigate the misclassification rate table in the face of this situation. Table 13 clearly shows that the model reaches the highest overall accuracy when seven hidden units are involved in the model. Based on this result, we determine to use the neural network of seven hidden units.

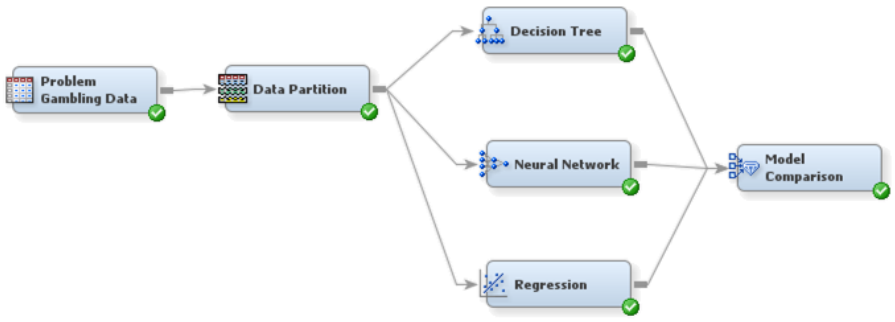
Table 15: Misclassification rate of neural network models with different numbers of hidden units

Model Name	Misclassification Rate	
	Validation Subset	Training Subset
Neural 7	0.00146	0.00191
Neural 2	0.00169	0.00225
Neural 4	0.00191	0.0018
Neural 8	0.00202	0.00158
Neural 5	0.00248	0.00296
Neural 10	0.00259	0.00266
Neural 6	0.00293	0.00289
Neural 9	0.00315	0.00285
Neural 3	0.00326	0.0285

5. Comparison of Three Predictive Models

To determine the best models, we construct and compare three models—decision tree using Gini splitting rule, neural network using seven hidden units, and logistical regression models.

Figure 10: Comparison of predictive models



Besides the overall accuracy, some researchers evaluate the effectiveness of models by measuring the sensitivity (true positive rate) and specificity (true negative rate) in classification matrix (Schellinck & Schrans, 2011). Table 16 shows that the neural network and logistic regression models reach almost the same outcomes in terms of sensitivity, specificity and overall accuracy. On the other hand, the decision tree reaches the lowest accuracy, though slightly lower than that of the other two models.

Table 16: Classification matrix of each predictive model

	Decision Tree				Neural Network		Logistic Regression	
	*TP	480	FP	7	486	6	485	0
	FN	6	TN	8390	0	8391	1	8397
* Sensitivity	98.8%				100%		99.8%	
Specificity			99.9%			99.9%		100%
False-Negative Rate	1.2%				0%		0.2%	
False-Positive Rate			0.1%			0.1%		0%
Overall Accuracy	**99.8%**				**99.9%**		**99.9%**	

* TP: true positive; FP: false positive; FN: false negative; TN: true negative

* The calculation methods: (1) sensitivity = TP/ (TP + FN); (2) specificity = TN / (FP + TN); (3) false negative rate = FN / (TP + FN); (4) false positive rate = FP / (FP + TN); (5) overall accuracy = (TP+TN)/ total.

By further looking at the misclassification matrix, we find out that the logistical regression is slightly better than the neural network model. However, it is hard to determine which model is better on the basis of the slight difference of overall accuracy between each model.

Table 17: Misclassification rate of each predictive model

Model Name	**Misclassification Rate**	
	Validation Subset	**Training Subset**
Logistical Regression	0.000338	0.0000375
Neural Network	0.00146	0.00191
Decision Tree	0.00473	0.00417

Actually, one important reason for using the logistic regression rather than the neural network is the inherent limitation of the network model as previously mentioned. The neural network is not able to tell us much about what factors will be important in arriving at the predictive model, and thus few analysis can be made from it out (Linoff & Berry, 2011). On the other hand, the logistic regression assists us to find out which gambling behavior are highly related to gambling addiction by identifying which predictor variables contribute more to the target variable.

Before making the final decision, however, we need to check whether the logistic regression is statistically significant by using the likelihood ratio test table. If the probability in the table is less than 0.05, the model is statistically significant and the predicator variables have an impact on the target variable. If the probability is more than 0.05, it is nonsense to use the model for predictive analysis (Fultz, 2012). Table 18 shows the likelihood ratio Chi-Square of 39375.9 with a probability (Pr>ChiSq) less than 0.0001, indicating that the model is statistically significant. Based on the result, the regression model is finally determined to be applied for predictive analysis.

Table 18: Likelihood ratio test result of the logistic regression model

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood		Likelihood		
Intercept Only	Intercept & Covariates	Ratio Chi-Square	DF	Pr > ChiSq
39403.675	27.762	39375.9125	15	<.0001

On the other hand, we adapt the decision tree model to perform the predictive analysis due to its unique advantages though the performance of it is slightly lower than that of the other two models. The decision tree model is easily to be evaluated and explained with interpretable

English rules (Ghoson, 2010). In our research, we construct a decision tree with the cluster label as the target variable. All rules derived from the decision tree can be used to assign new players to the correct gambler group and therefore to predict at-risk and problem gamblers.

Consequently, the logistic regression and decision tree models are adapted to carry out the predictive analysis.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 DESCRIPTION OF CLUSTERS

4.1.1 Cluster Size

By using the clustering, we separate the 44,416 sessions into four clusters that correspond to four categories: non-problem gamblers, low-risk gamblers, moderate-risk gamblers, and problem gamblers based on the categories generated by the CPGI. The problem is which cluster should correspond to which player group.

Previous studies mentioned that the majority of players are recreational players, and only a small portion of players are at-risk or problem gamblers (New Zealand Health Survey, 2012). This finding has been supported by the CPGI that surveyed 2,681 players in their research and found that 1.04% of those gamblers were problem gamblers, 2.76% were moderate-risk gamblers, 7.91% were low-risk gamblers and 82% were non-problem gamblers (Ferris & Wynne, 2001).

According to these previous researches, we can hypothesize that Cluster 1 (19.1%) is low-risk gambler group, Cluster 2 (1%) is problem gambler group, Cluster 3 (74.5%) is non-problem gambler group, and Cluster 4 (5.5%) is moderate-risk gambler group.

Table 19: Cluster size

		Number of Gamblers	Percentage
Cluster	1	8,466	19.1%
	2	434	1.0%
	3	33,088	74.5%
	4	2,428	5.5%
Total		44,416	100%

4.1.2 Comparison of Clusters

The comparison of clusters by behavioral indicators assists us to find out the unique behavioral characteristics of each cluster and therefore to differentiate them. Typically, researchers compare the cluster center (mean) of each variable within each cluster. But we consider that the comparison of the cluster center is not sufficient since the cluster center only indicates the behavior of typical players in each group and is highly affected by the extreme values within each cluster. We thereby investigate and compare other information such as mode, 90^{th} percentile, as well as maximum value of each behavioral indicator within each cluster in an attempt to completely understand behavioral characteristics of the majority of players.

1. DurationMin

Figure 20 shows that Cluster 2 players spent the longest time on playing EGMs. They typically spent 133 minutes (mean=132.9) and most of them frequently spent 108 minutes (mode=108.33) during a playing session. Half of them spent more than 122 minutes (median=121.93) but less than 184 minutes (90^{th} percentile = 183.8) on playing, and 9% of them played even longer than 184 minutes before they left. Previous researches pointed out that most problem gamblers play longer than other players (Productive Commision, 2009), and they spend over 120 or 180 minutes on gambling during a session (Dragicevic, Tsogas, & Kudic, 2011). Based on these research results, we regard Cluster 2 players as problem gamblers in terms of their time spent.

In contrast, Cluster 3 players exhibited the opposite behavior. Most Cluster 3 players spent the shortest time on playing EGMs. They averagely spent around 22 minutes (mean = 22.3) and most frequently spent only 4 minute (mode = 4) on playing, and 90% of them spent no more than

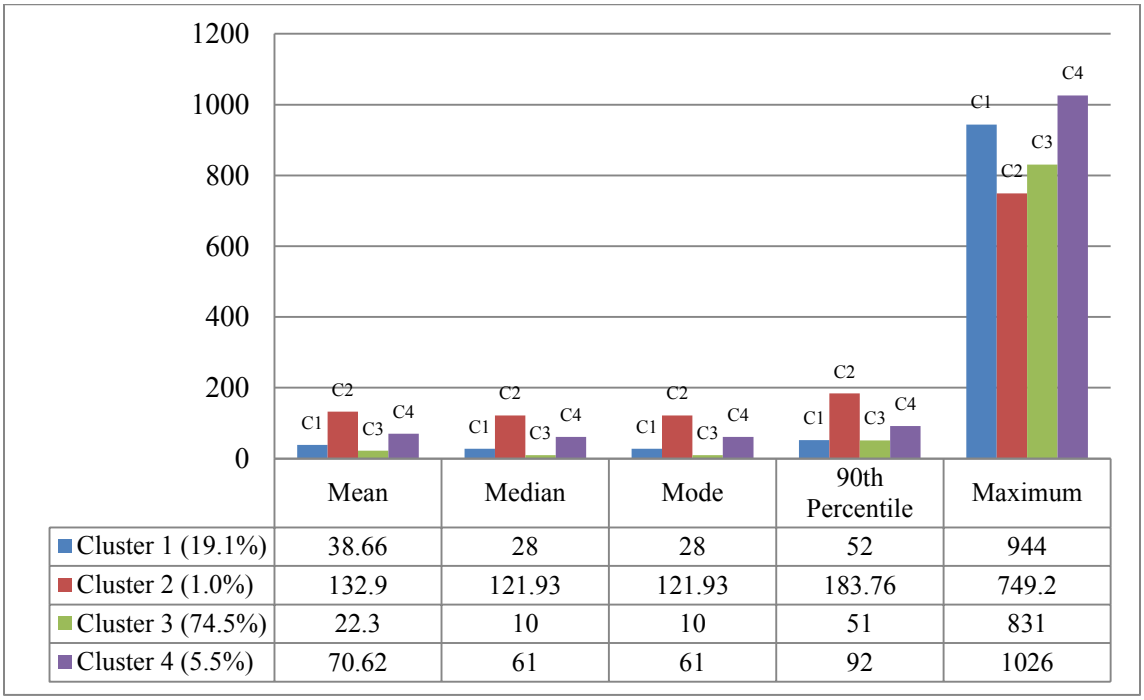
51 minutes (90th percentile = 51). With regard to time spent, those players know when to stop and treat EGMs play as a recreational activity.

When comparing time spent of Cluster 4 and 1 player, we find out that the majority of Cluster 4 players spent more time than Cluster 1 players on EGMs. Therefore, we consider that Cluster 4 players are moderate-risk gamblers and accordingly Cluster 1 is low-risk gambler group.

Although the majority of Cluster 1, 3 and 4 players did not spend as long time as Cluster 2 players on EGMs, it is noticeable that the maximum of these groups are higher than that of Cluster 2. The reason behind this may be explained by the chasing-losses behavior. Some players in these three groups spent as much time as they could rather than stop in order to win their losses back. In that sense, it is difficult to determine whether Cluster 3 is recreational player group and Cluster 1 as well as 4 are at-risk gambler groups.

Therefore, it is necessary to analyze and compare betting activity and money spent.

Figure 11: Comparison of duration

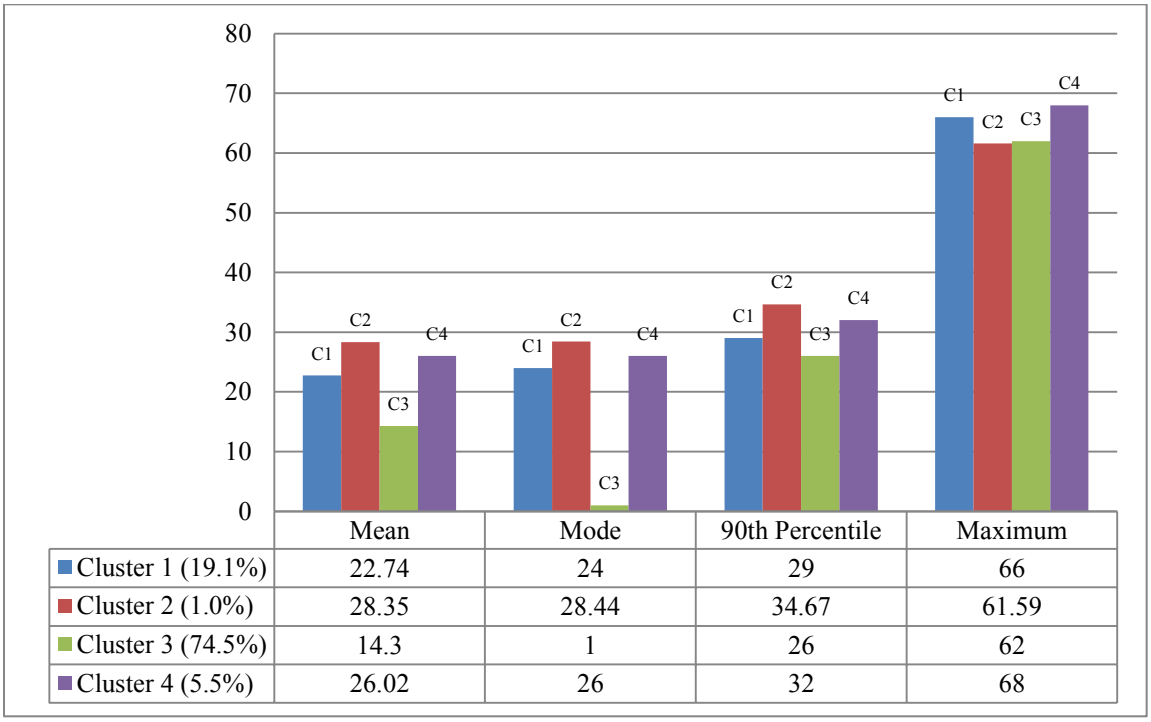


	Mean	Median	Mode	90th Percentile	Maximum
Cluster 1 (19.1%)	38.66	28	28	52	944
Cluster 2 (1.0%)	132.9	121.93	121.93	183.76	749.2
Cluster 3 (74.5%)	22.3	10	10	51	831
Cluster 4 (5.5%)	70.62	61	61	92	1026

2. Betting Behavior

When comparing the values of the 'BetsPerMin' within each cluster, we identify that Cluster 1, 2 and 4 players show the similar betting behavior. Players in these three groups typically placed around 22 to 28 bets and most frequently placed 24 to 26 bets per minute. Most of them placed approximately 30 to 35 bets per minute. Those players bet more quickly than players in Cluster 3, who typically placed around 14 bets per minute and most frequently placed only 1 bet per minute. Researchers mentioned that at-risk and problem gamblers usually bet more quickly than their recreational counterparts in order to win their money back. The more quickly a player bets, the more likely he or she is a problem gambler (Delfabbro, King, & Griffiths, 2012). Based on this finding, we thereby regard Cluster 1, 2 and 4 players as at-risk or problem gamblers and Cluster 3 as recreational players.

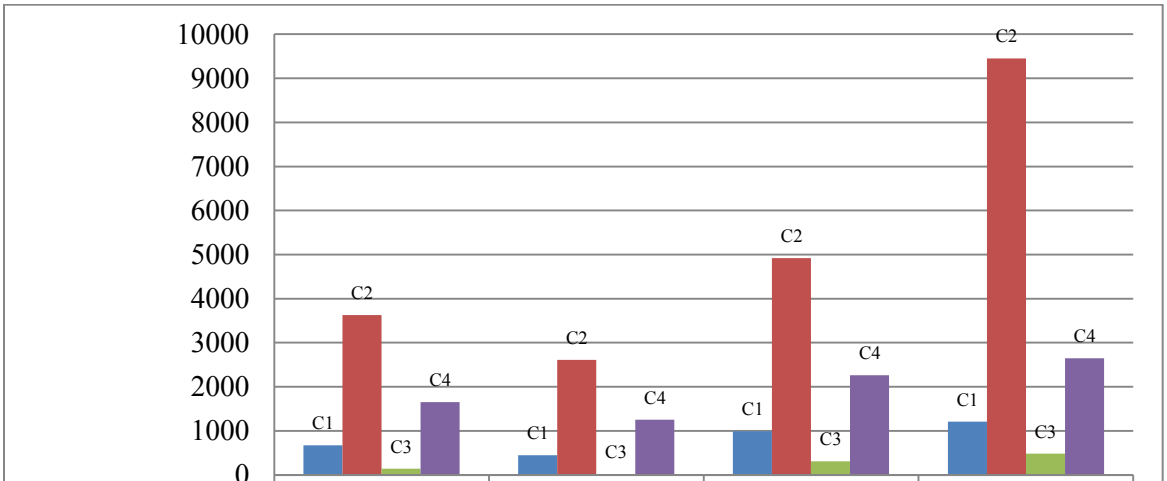
Figure 12: Comparison of betting behavior (BetsPerMin)



	Mean	Mode	90th Percentile	Maximum
Cluster 1 (19.1%)	22.74	24	29	66
Cluster 2 (1.0%)	28.35	28.44	34.67	61.59
Cluster 3 (74.5%)	14.3	1	26	62
Cluster 4 (5.5%)	26.02	26	32	68

In order to further distinguish problem gamblers from their at-risk counterparts, we compare the total number of bets those groups placed during a session. Figure 13 clearly shows that Cluster 2 players placed much more bets than the other two cluster players during their visit. Based on previous researches we previously mentioned, Cluster 2 players can be regarded as problem gamblers. Accordingly, Cluster 4 players are more likely to be moderate-risk gamblers than Cluster 3 players as players in Cluster 4 placed more bets in total.

Figure 13: Comparison of betting behavior (TotalBets)



	Mean	Mode	90th Percentile	Maximum
Cluster 1 (19.1%)	672.72	445	995	1213
Cluster 2 (1.0%)	3625.79	2609.27	4920.83	9452.62
Cluster 3 (74.5%)	143.2	22	307	485
Cluster 4 (5.5%)	1651.14	1251	2261	2649

3. Money Spent

Compared with the other players, the majority of Cluster 2 players spent as well as lost the largest amount of money during a session. Those players typically spent 316 money units* (RedeemedPerS mean=316.04) and lost 189 money units (Lost mean =188.68) at the end of playing. They most frequently spent 200 money units (RedeemedPerS mode = 200) and finally lost half of them (Lost mode = 100). According to the previous researches, the more money a

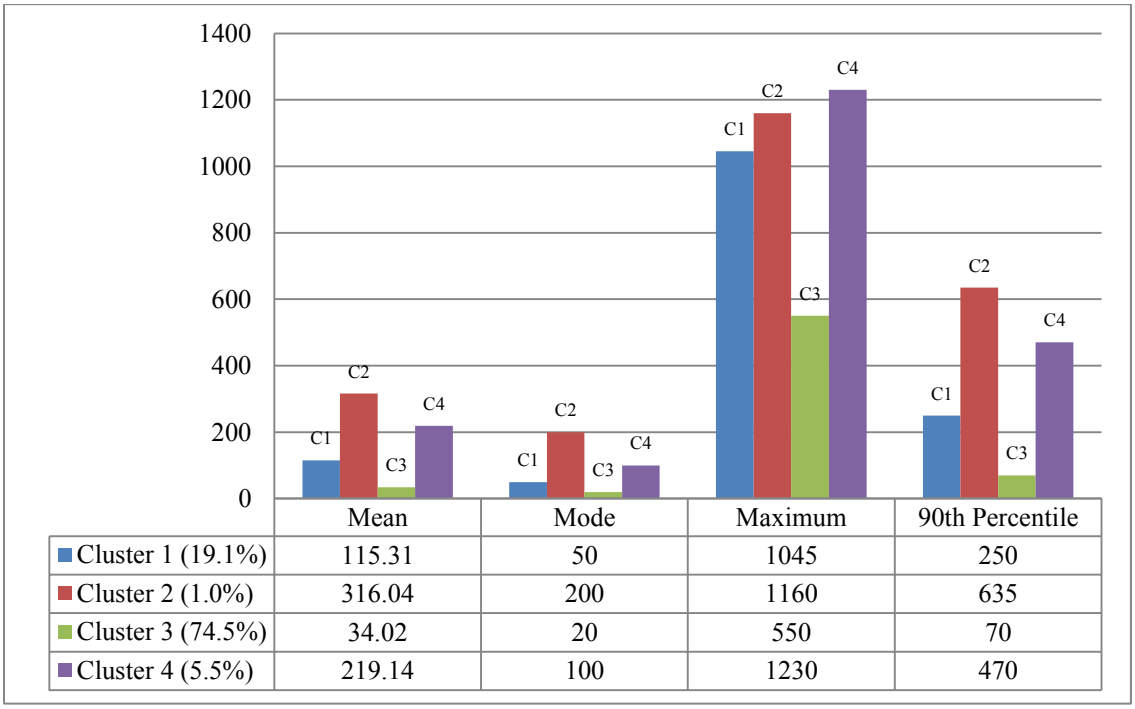
* We utilize general 'unit' to present money spent. One money unit represents the value of $1.48 CAD or €1.

gambler spends, the more likely he or she is a problem gambler (Delfabbro, King, & Griffiths, 2012) , we therefore consider Cluster 2 as the problem gambler group. On the contrary, Cluster 3 players spent and lost the least amount of money, so they can be regarded as recreational player groups.

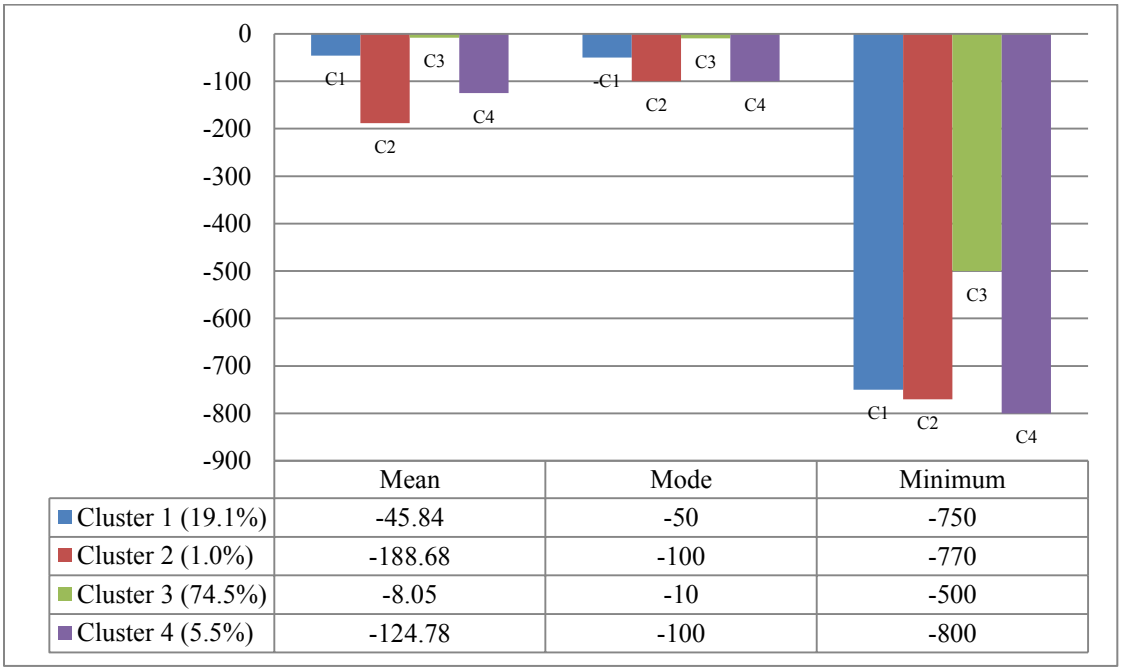
When comparing the maximum of the 'RedeemedPerS' and the minimum value of the 'Loss' variables within each cluster, we notice that some Cluster 4 players spent and lost the largest amount of money among all players. A possible explanation is that Cluster 4 players attempted to win their losses back by spending more money, but consequently may result in more losses and lead to gambling addiction (Gambling: Help and Referral, 2013). The more money a player spend, the more likely he or she is an at-risk or problem gambler.

Additionally, the majority of Cluster 4 players spent and lost much more money than Cluster 1 players, therefore Cluster 4 players are more likely to be moderate-risk gamblers than Cluster 1 players.

Figure 14: Comparison of money spent



	Mean	Mode	Maximum	90th Percentile
Cluster 1 (19.1%)	115.31	50	1045	250
Cluster 2 (1.0%)	316.04	200	1160	635
Cluster 3 (74.5%)	34.02	20	550	70
Cluster 4 (5.5%)	219.14	100	1230	470



	Mean	Mode	Minimum
Cluster 1 (19.1%)	-45.84	-50	-750
Cluster 2 (1.0%)	-188.68	-100	-770
Cluster 3 (74.5%)	-8.05	-10	-500
Cluster 4 (5.5%)	-124.78	-100	-800

In summary, Cluster 2 players played longer than the other players, placed the largest number of bets per session, spent as well as lost the largest amount of money at the end of playing. In

contrast, Cluster 3 players spent the least amount of time and money on playing EGMs, bet slowest per minute and placed the least number of bets during a session. Accordingly, they finally lost the least amount of money when they left. Based on previous research, we conclude that Cluster 2 and 3 players are problem gamblers and non-problem gamblers, respectively.

In terms of at-risk gambler groups, we need to further compare the distances between cluster centers before making the final decision, though Cluster 4 players exhibited more characteristics of moderate-risk gamblers than Cluster 3 players.

4. Distance between Cluster Centers

Table 20 provides the information of distances between cluster centers, the greater distances between clusters, the greater dissimilarities between these two clusters (IBM, 2011). It shows that the distance between Cluster 1 and Cluster 3 is the shortest, which refers that Cluster 1 players exhibited similar behavior as Cluster 3 players. Our previous analysis has proved that Cluster 3 players are non-problem gamblers, thus Cluster 1 players can be regarded as low-risk gamblers and accordingly Cluster 4 players are moderate-risk gamblers.

Table 20: Distances between cluster centers

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		2964.831	537.452	987.592
2	2964.831		3500.513	1979.044
3	537.452	3500.513		1524.626
4	987.592	1979.044	1524.626	

4.1.3 Conclusion and Profiles of Clusters

Based on previous analysis, we consequently prove the hypothesis that Cluster 3 (74.5%) is non-problem gambler group, Cluster 1(19.1%) is low-risk gambler group, Cluster 4 is moderate-risk gambler group, and Cluster 2 (1.0%) is problem gambler group.

1. Cluster 3 (74.5%): Non-Problem Gamblers

Players in Cluster 3 are non-problem gamblers, who account for the majority of EGM players. On average, non-problem gamblers spent only 22 minutes on EGMs, placed approximately 14 bets in a minute and 143 bets totally, and spent 34 money units as well as lost 8 money units at the end of playing. They most frequently spent only 4 minutes per session, place only 1 bet in minute and 22 bets in a session, spent 20 money units and finally lost half of them. The majority of non-problem gamblers (90%) spent no more than 51 minutes, placed no more than 26 bets per minute along with 307 bets during a session, spent less than 70 money units and finally won less than 40 money units (actually, 80% of them lost more than 5 money unites).

Table 21: Profile of non-problem gamblers

Cluster 3 (74.5%) Non-Problem Gamblers	Duration	Betting Behavior		Money Spent	
	DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
Mean	22.3	14.3	143.2	34	-8.1
Mode	4	1	22	20	-10
90th Percentile	51	26	307	70	40

2. Cluster 1 (19.1%): Low-risk Gamblers

Cluster 1 players are low-risk gamblers. On average, they spent about 39 minutes, placed 23 bets per minute and 673 bets per session, spent 115 money units as well as lost approximately 46 money units. Those low-risk gamblers frequently spent 19 minutes during a session, placed 24

bets per minute and 445 bets in total, spent 50 money units and finally lost them all. The majority of them spent no more than 52 minutes, placed no more than 29 and 995 bets per minute and per session, respectively, spent no more than 250 money units and won 110 money units (70% of them actually lost more than 20 money units) when they left.

Table 22: Profile of low-risk gamblers

Cluster 1 (19.1%) Low-Risk Gamblers	**Duration**	**Betting Behavior**		**Money Spent**	
	DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
Mean	38.7	22.7	672.7	115.3	-45.8
Mode	19	24	445	50	-50
90th Percentile	52	29	995	250	110

3. Cluster 4 (5.5%): Moderate-Risk Gamblers

Moderate-risk gamblers in Cluster 4 on average spent around 71 minutes, placed 26 bets in minute and 1651 bets in total, spent 219 money units along with lost 125 money units during a session. They most frequently spent 56 minutes on EGMs, placed 26 bets in minute and 1251 bets per session, spent 100 money units and finally lost them all. Most of moderate-risk gamblers spent less than or equal to 92 minutes on EMGs, placed no more than 32 bets per minute and 2261 bets per session, spent less than 470 money units and won 109 money units (70% of them lost more than 50 money units) at the end of playing.

Table 23: Profile of moderate-risk gamblers

Cluster 4 (5.5%) Moderate-Risk Gamblers	**Duration**	**Betting Behavior**		**Money Spent**	
	DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
Mean	70.6	26	1651.1	219.1	-124.8
Mode	56	26	1251	100	-100
90th Percentile	92	32	2261	470	109

4. Cluster 2 (1.0%): Problem Gamblers

Players in Cluster 2 are problem gamblers, who account for 1% of all EGM players. On average, problem gamblers spent about 133 minutes, placed 29 bets in minute as well as 3626 bets in total, spent 316 money units and lost approximately 189 money units when they left. Problem gamblers frequently spent 108 minutes, placed 28 bets per minute and 2609 bets per session, spent 200 money units and finally lost half of them. And 90% of problem gamblers spent no more than 184 minutes on EGM gambling, placed no more than 35 bets in minute and 4921 bets in total, spent less than or equal to 635 money units and finally won 180 money units (80% of them lost more than 5 money units).

Table 24: Profile of problem gamblers

Cluster 2 (1.0%) Problem Gamblers	**Duration**	**Betting Behavior**		**Money Spent**	
	DurationMin	BetsPerMin	TotalBets	RedeemedPerS	Loss
Mean	132.9	28.4	3625.8	316	-188.7
Mode	108.3	28.4	2609.3	200	-100
90th Percentile	183.8	34.7	4920.8	635	180.3

4.2 PREDICTION OF AT-RISK AND PROBLEM GAMBLERS

4.2.1 Importance of Risk Indicators

The Type 3 Analysis of Effects is used to test the statistical significant of each coefficient in the model, therefore it assists us to understand the importance of each risk indicator. In the Type 3 analysis table, if the p-value (Pr>ChiSq) of a variable is less than 0.0001, the input variable is extremely significant (SAS Institute Inc, 1999). Figure 15 displays that the p-value of the 'Loss', 'RedeemedPerS', and 'TotalBets' is less than 0.0001, indicating that these three risk indicators are extremely significant when predicting gamblers. In other words, we can use money spent,

loss and total bets a player placed per session to predict which gambler group he should belong to.

On the other hand, the 'DurationMin' is only significant but not extremely significant when predicting gamblers, as its p-value is 0.00257 (0.01≤p-value≤0.05 is regarded as significant) (GraphPad Software, Inc, 2014). The 'BetsPerMin' will not play a significant role in predicting EGM gamblers since its p-value is more than 0.05.

Table 25: Type 3 analysis of effects

Effect	Pr > ChiSq
BetsPerMin	0.7487
DurationMin	0.0257
Loss	<. 0001
RedeemedPerS	<.0001
TotalBets	<.0001

However, the result of indicators' importance created by the decision tree model is slightly different. The decision tree model indicates that the 'TotalBets', 'Loss', 'RedeemedPerS' and 'BetsPerMin' are four important indicators to predict gamblers.

Table 26: Variable importance determined by the decision tree

Name	Importance	Vimportance
TotalBets	1.0000	1.0000
Loss	0.1103	0.0802
RedeemedPerS	0.0418	0.0463
BetsPerMin	0.0265	0.0000

Therefore, by using a combination of these two results, we determine that total bets, loss, and money spent are three most important indicators than can be applied to predict new EGM

gamblers. Duration also plays a significant role in predicting gamblers, though it is less significant than the previous three indicators.

The maximum likelihood table of the logistic regression model shows the relationship for each input varaible with the target variable. If Exp (Est) in the table is less than 1, increasing values of the input variable leads to decreasing odds of the target variable; if Exp (Est) is greater than 1, then increasing values of the predictor variable results in increasing odds of the target variable (BGSU, 2006). By using this theory, we are able to understand the relationship between the changes of gambling behavior and the development of gambling addiction. Howevere, we need to look at the p-value of each variable before studying the Exp value. If the p-value is more than 0.05, it is not necessary to study the Exp value, since the variable is not significant for predicting the target variable.

Table 27: Analysis of maximum likelihood estimates

Parameter	Player Groups	Pr > ChiSq	Exp (Est)
BetsPerMin	4	0.4889	1.0920
	3	0.4288	1.0770
	2	0.6801	1.5820
DurationMin	4	0.1047	1.0320
	3	0.0103	0.9400
	2	0.6937	1.1150
Loss	4	0.0006	0.9340
	3	< 0.0001	1.1710
	2	0.0459	0.9350
RedeemedPerS	4	0.0006	1.1030
	3	< 0.0001	0.7000
	2	0.0086	1.1110
TotalBets	4	0.0004	2.359
	3	<.0001	0.102
	2	<.0001	2.965

* The Cluster 1 (low-risk gambler) is arbitrarily selected as baseline by SAS E-Miner.

BetsPerMin. As can be seen, the p-values of the 'BetsPerMin' are more than 0.05, so this variable is not significant in predicting gamblers.

DurationMin. This risk indicator is significant to predict or assign potential non-problem gamblers (p-value=0.01), increasing duration correspondes with decreasing odds of being non-problem gamblers (Exp = 0.94). In other words, the more time a player spends on EGM gambling, the more likely he or she will become a at-risk or problem gambler.

Loss. First, the 'Loss' value is extremely signigicant for predicting non-problem and moderate-risk gamblers (p-value < 0.0001 and p-value = 0.0006, respectively), and significant for predciting problem gamblers (p-value = 0.0459). Since the 'Loss' variable includes negative and non-negative values, it is necessary to consider two situations. If the value of the 'Loss' is non-negative, increasing values result in decreasing odds of being moderate-risk and problem gamblers. In other words, the more money a player wins, the more likely he or she will not develop problem gambling. On the other hand, if the value of the 'Loss' is negative, increasing values leads to increasing odds to being moderate-risk and problem gamblers, but decreasing odds of being non-problem gamblers. Thus, the more money a player lose, the more likely he or she will become at-risk or problem gambler.

RedeemedPerS. This variable plays an extremely significant role in predicting each gambler group. Increasing money spent corresponds to increasing odds of being moderate-risk and problem gamblers, and decreasing odds of being non-problem gamblers. The more money a gambler spend, the more likely he or she will become an at-risk or problem gambler.

TotalBets. The 'TotalBets' also plays an extremely important role in predicting each gambler group. Increasing the number of bets per session corresponds to increasing odds of being

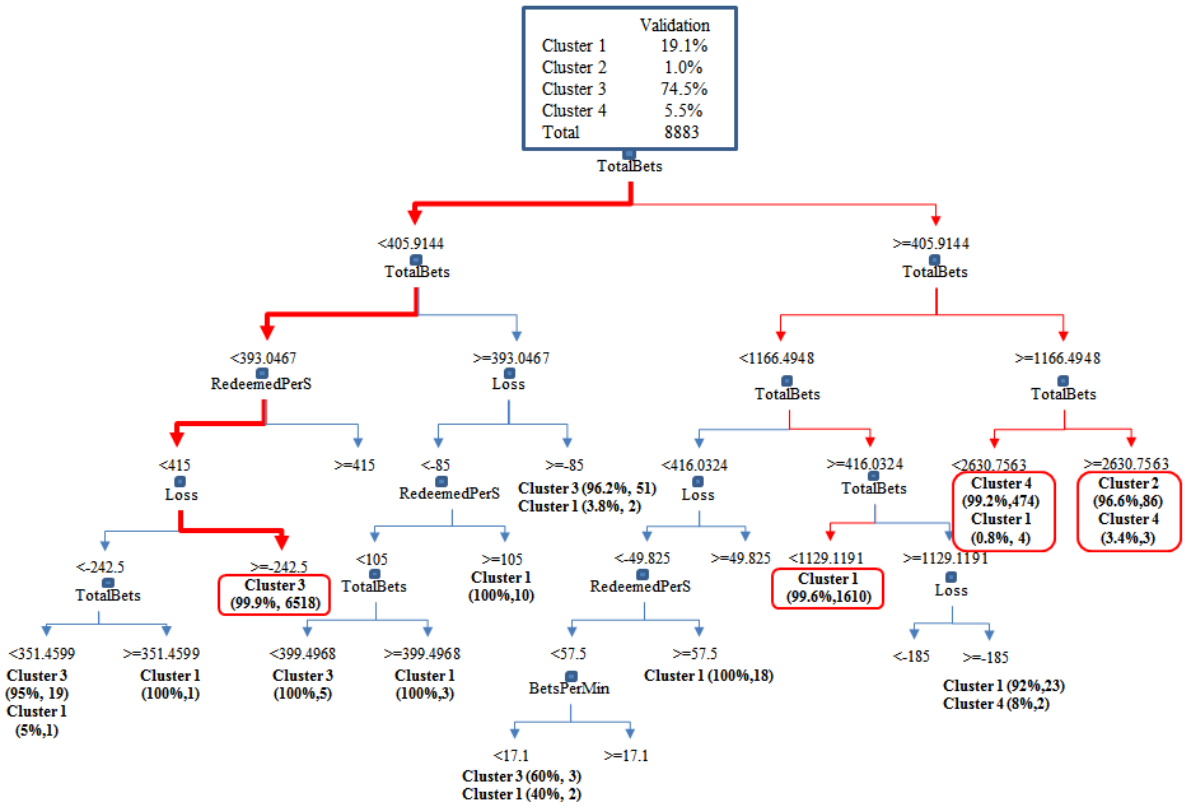
moderate-risk and problem gamblers, and accordingly decreasing odds of being non-problem gamblers. So the more bets a player places per visit, the more likely he or she will become an at-risk or problem gambler.

In summary, the more money spent, loss, and bets a player place, the more likely he or she will become an at-risk or problem gambler.

4.2.2 Prediction of At-Risk and Problem Gamblers

We use the cluster label as the target variable and the behavioral indicators as input predictors when constructing the decision tree model, thus, the splitting rules generated by the tree are available to assign new players into the correct groups and accordingly to predict at-risk and problem gamblers based on their gambling behavior.

Figure 15: Decision tree



The tree mainly uses three behavioral indicators including 'TotalBets', 'RedeemedPerS' and 'Loss' to split players into different gambler groups. Table 28 shows that thirteen rules are generated and most of them are used to classify non-problem gamblers and low-risk gamblers.

Table 28: Rules for predicting potential players

Gambler Group	**Rules**
Non-Problem Gambler	**If TotalBets<393.047 AND RedeemedPerS<415, AND Loss≥-242.5**
	If TotalBets<405.914 AND TotalBets≥393.047, AND Loss≥-85
	If TotalBets<351.456 AND RedeemedPerS<415, AND Loss<-242.5
	If TotalBets<399.497 AND TotalBets≥393.047, AND RedeemedPerS<105, AND Loss<-85
	If TotalBets<416.032 AND TotalBets≥405.914, AND RedeemedPerS<57.5, AND Loss<-49.825 AND BetsPerMin<17.1
Low-Risk Gambler	**If TotalBets<1129.12 AND TotalBets≥416.032**
	If TotalBets<1166.49 AND TotalBets≥1129.12, AND Loss≥-185
	If TotalBets<416.032 AND TotalBets≥405.914, AND RedeemedPerS≥57.5, AND Loss<-49.825
	If TotalBets<405.914 AND TotalBets≥393.047, AND RedeemedPerS≥105, AND Loss<-85
	If TotalBets<405.914 AND TotalBets≥399.497, AND RedeemedPerS<105, AND Loss<-85
	If TotalBets<393.047 AND TotalBets≥351.456, AND RedeemedPerS<415, AND Loss<-242.5
Moderate-Risk Gambler	**If TotalBets<2630.76 AND TotalBets≥1166.49**
Problem Gambler	**If TotalBets≥2630.76**

* All the rules generated from the decision tree are involved in the Appendix 1.

Although a set of rules are generated by the decision tree, it does not mean that all rules are appropriate for predicting potential players. Only are those rules separating players into leaves where a single gambling group dominates selected. Moreover, a good rule should not create a leaf containing very few players (Linoff & Berry, 2011). Based on this theory, we thereby find out four best rules predicting four gambling groups, respectively.

1. Non-Problem Gamblers (Cluster 3)

The best rule to classify the non-problem gamblers is 'If TotalBets < 393.047 AND RedeemedPerS < 415, AND Loss ≥ -242.5'. It leads to a leaf that includes 99.9% of non-problem gamblers, accounting for 98.5% of all players in this group (6,518 out of 6,618 Cluster 3 players in the validation subset).

By using this rule, we are able to differentiate non-problem gamblers from at-risk and problem gamblers, since recreational players place less than 393 bets, spend less than 415 money units as well as lose less than 242.5 money units.

Table 29: The best rule to predict non-problem gamblers

Rules	If TotalBets<393.047 AND RedeemedPerS<415, AND Loss≥-242.5
Number of Observations	6518
Predicted	Non-problem gambler: 99.9%
	Low-risk gambler: 0.1%

2. Low-risk Gamblers (Cluster 1)

The best rule to predict low-risk gamblers is 'If TotalBets<1129.12 AND TotalBets≥416.032', which creates a leaf that is composed of 99.6% of low-risk gamblers accounting for 95.3% of all gamblers in the group (1,616 out of 1,696 players in Cluster 1). Thus, those players who place between 416 and 1,129 bets can be predicted as low-risk gamblers.

Table 30: The best rule to predict low-risk gamblers

Rules	If TotalBets<1129.12 AND TotalBets≥416.032
Number of Observations	1616
Predicted	Low-risk gambler: 99.6%
	Non-problem gambler: 0.4%

3. Moderate-risk Gamblers (Cluster 4)

The rule 'If TotalBets<2630.76 AND TotalBets≥1166.49' creates a leaf that contains 99.2% of moderate-risk gamblers accounting for 98% of players in this group (478 out of 488 Cluster 4 players). Thus, this rule is the best rule for predicting moderate-risk gambles. Based on this rule, we are able to predict that players who place between 1,166 and 2,631 bets are most likely to be moderate-risk gamblers.

Table 31: The best rule to predict moderate-risk gamblers

Rules	If TotalBets<2630.76 AND TotalBets≥1166.49
Number of Observations	478
Predicted	Moderate-risk gambler: 99.2%
	Low-risk gambler: 0.8%

4. Problem Gamblers (Cluster 2)

Only one rule is generated to predict problem gamblers. The leaf generated by the rule 'If TotalBets≥2630.76' consists of 96.6% of problem gamblers and 3.4% of moderate-risk gamblers. Players who place more than 2,631 bets can be predicted as problem gamblers.

Table 32: The best rule to predict problem gamblers

Rules	If TotalBets≥2630.76
Number of Observations	86
Predicted	Problem gambler: 96.6%
	Moderate-risk gambler: 3.4%

By applying these four best rules, we are able to successfully assign most of the potential players into the correct gambler groups based on their gambling behavior. But more importantly, they

assist us to distinguish at-risk as well as problem gamblers from their recreational counterparts and to differentiate between at-risk and problem gamblers.

We classify players who place less than 393 bets, spend less than 415 money units and lose less than 242.5 money units as recreational players, that is, players who do not exhibit these behavior can be regarded as at-risk or problem gamblers.

With regard to the differentiation of at-risk and problem gamblers, the total number of bets plays a critical role. Players who place less than 1,129 bets are predicted as low-risk gamblers, that is to say, players who place more than these bets should be predicted as moderate-risk or problem gamblers.

In terms of moderate-risk and problem gamblers, 2,631 is the cut-off point to distinguish problem gamblers from moderate-risk gamblers. If players place more than 2,631 bets, they can be regarded as problem gambler; otherwise they are classified as moderate-risk gamblers.

Therefore, we can conclude that at-risk and problem gamblers can be differentiated from their recreational players by judging from their money spent and the number of total bets. But the difference between at-risk and problem gamblers is only determined by the number of total bets. By calculating the total number of bets, we are able to successfully predict problem and moderate-risk gamblers.

CHAPTER 5 CONCLUSIONS

5.1 CONCLUSION

The objectives of this thesis are using data mining techniques to distinguish different levels of EGM players based on their gambling behavior, to identify which gambling behavior are highly associated with gambling addiction, and finally to derive rules that can be utilized to predict potential at-risk and problem gamblers.

The data analysis procedure is composed of three stages, which are data preparation, clustering and prediction analysis. In the first data preparation stage, we create three derived variables by transforming and combining the existing variables in an attempt to find out more hidden behavioral characteristics information. The correlation coefficient analysis is conducted to finally determine the five behavioral indicators for creating data mining models. Then the descriptive data summarization technique is applied to investigate whether outliers exist, and the hybrid outlier detection method is adapted to detect and remove outliers.

When the data set is ready for modeling, we apply the k-means clustering technique to separate players into four gambler groups that are non-problem gamblers, low-risk gamblers, moderate-risk gamblers and problem gamblers, based on the categories created by the CPGI. Accordingly, the detailed profiles of groups are provided.

Finally, by applying the logistic regression and decision tree models, we identify that money spent, loss, and total bets play a critical role in predicting at-risk and problem gamblers. Moreover, the decision tree generates a set of rules that are able to assign new players into correct gambler groups and to predict at-risk and problem gamblers. Among all of those rules,

four of them are most important as they classify the majority of players into each group correctly. Based on these four rules, we conclude that at-risk and problem gamblers are able to differentiate from their recreational counterparts by judging their money spent and the number of total bets. At-risk and problem gamblers are easier to be distinguished by calculating the number of total bets.

5.2 CONTRIBUTIONS

It is hoped that this thesis will provide usable information on studies about utilizing data mining techniques on EGM gambling behavior analysis since there is very few published research results about this topic. We also hope that findings and rules generated from this thesis will be a guide and a potentially valuable tool for EGMs manufacturers that may design their EGMs and set up the warning system to monitor player behavior and to give advice when the machine identifies risky gambling behavior.

5.3 LIMITATIONS AND FUTURE WORK

Typically, the data used for customer behavior analysis, particularly for the prediction of customer behavior, should be composed of both behavioral and demographic data. However, in our research, the demographic data of players are not available since the data were obtained from EGMs directly rather than from the customer tracking systems. Without the basic demographic data, we thereby are not able to build more comprehensive profiles on different levels of gamblers. Furthermore, a lack of demographic data decreases the predictability of models (Nisbet, Elder, & Miner, 2009).

Thus, we suggest that more variables should be involved in future research. More variables will lead to more meaningful and natural clusters and generate models with higher predictability. If

possible, demographic data and variables should be collected and involved in the analysis, particularly in the predictive analysis, since demographic data assist researchers to know basic information and characteristics of a gambler and therefore detect what types of people are more likely to become at-risk and problem gamblers. Besides demographic variables, other variables may include the payment methods (e.g., bill acceptors, cash, and direct electronic fund transfer), payout methods (e.g., tickets, tokens), type of games, and sensory effects (e.g., lights and types of sounds) (Responsible Gambling Council, 2006).

The other limitation is that the data set is composed of one month anonymous data, thus we are not able to identify the problem gambling development progress over time. If more than one month data could be involved in the future research, we would be able to understand the problem gambling trajectories.

BIBLIOGRAPHY

New Zealand Health Survey. (2012). *Problem Gambling in New Zealand.*

SAS Institute Inc. (1999). *The GENMOD Procedure.* Retrieved May 10, 2014, from SAS OnlineDoc: Version 8: http://www.math.wpi.edu/saspdf/stat/chap29.pdf

Kruskal-Wallis H Test using SPSS. (2013). Retrieved December 21, 2013, from Laerd Statistics: https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

Alexander, D. (1997). *Data Mining*. Retrieved May 4, 2014, from The University of Texas at Austin: http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/

Allison, P. (2012). *When Can You Safely Ignore Multicollinearity?*. Retrieved July 27, 2014 from Statistical Horizons: http://www.statisticalhorizons.com/multicollinearity

Ambrosius, W. T. (2007). *Topics in Biostatistics.* Humana Press Inc.

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting*, 689–694.

Andrew, P. (2012). *Predictive Analytics In the Gaming Industry.* Retrieved August 2,2014, from Academic Edu https://www.academia.edu/2467733/Predictive_Analytics_in_the_Gaming_Industry

BGSU. (2006). *Logistic Regression*. Retrieved February 2, 2014, from Bowling Green State University: http://www2.bgsu.edu/downloads/cas/file36803.pdf

Blischke, W. R., Rezaul Karim, M., & Prabhakar Murthy, D. N. (2011). Preliminary Data Analysis. In *Warranty Data Colection and Analysis* (pp. 158-189). Springer.

Brown, S. (2008). *Measures of Shape: Skewness and Kurtosis*. Retrieved January 21, 2014, from Tompkins Cortland Community College: http://www.tc3.edu/instruct/sbrown/stat/shape.htm

Chapman, A. D. (2005). Principles and Methods of Data Cleaning – Primary Species and Species. *Global Biodiversity Information Facility.* Copenhagen.

Columbia Cnmtl. (2013). *Why Sample?* Retrieved November 20, 2013, from QMSS e-Lessons: http://ccnmtl.columbia.edu/projects/qmss/samples_and_sampling/why_sample.html

Correa, A., González, A., Nieto, C., & Amezquita, D. (2012). Constructing a Credit Risk Scorecard using Predictive Clusters. *SAS Global Forum* (p. 128). SAS Institute Inc.

Delfabbro, P., King, D. L., & Griffiths, M. (2012). Behavioural profiling of problem gamblers:a summary and review. *International Gambling Studies*, pp. 349-366.

Dragicevic, S., Tsogas, G., & Kudic, A. (2011). Analysis of casino online gambling data in relation to behavioural risk markers for high-risk gambling and player protection. *International Gambling Studies*, 377-391.

Fayyad, U., Shapiro, G. P., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 40.

Ferris, J., & Wynne, H. (2001). *The Canadian Problem Gambling Index: Final Report* . Retrieved July 3, 2013, from The Canadian Centre on Substance Abuse: http://www.ccsa.ca/2003%20and%20earlier%20CCSA%20Documents/ccsa-008805-2001.pdf

Franklin, S., & Thomas, S. (2000). *Robust Multivariate Outlier Detection Using Mahalanobis' Distance and Modified Stahel-Donoho Estimatiors.* Retrieved December 8, 2013, from American Statistical Association: http://www.amstat.org/meetings/ices/2000/proceedings/S33.pdf

Fultz, N. (2012). *SAS Annotated Output Proc Logistic*. Retrieved February 2, 2014, from idre: http://www.ats.ucla.edu/stat/sas/output/sas_logit_output.htm

Gambling: Help and Referral. (2013). *10 signs of problem gambling*. Retrieved Janunary 7, 2014, from Gambling: Help and Referral: http://www.jeu-aidereference.qc.ca/www/signs_problem_gambling_en.asp?cmpt=2#chasing_losses

Ghoson, A. A. (2010). Decision Tree Induction & Clustering Techniques In SAS Enterprise Miner, SPSS Clementine, And IBM Intelligent Miner –A Comparative Analysis. *International Journal of Management & Information Systems*, 57-70.

GraphPad Software, Inc. (2014). *Extremely significant?* Retrieved May 10, 2014, from GraphPad Software: http://www.graphpad.com/guides/prism/6/statistics/index.htm?extremely_significant_results.htm

Grasso, F. (2012). Retrieved January 5, 2014, from http://cgi.csc.liv.ac.uk/~floriana/COMP106/20ps.pdf

Gravetter, F. J., & Wallnau, L. B. (2009). *Statistics for the Behavioral Sciences.* Belmont: Wadsworth.

Griffiths, M. (2009). *Problem gambling in Europe: An overview.* Retrieved July 30, 2013, from Gambling Awareness Nova Scotia:

http://www.nsgamingfoundation.org/uploads/Problem%20Gambling%20in%20Europe.pdf

Han, J., & Kamber, M. (2006). *Data Mining:Concepts and Techniques (Second Edition).* San Francisco: Morgan Kaufmann.

Hsieh, N.-C., & Chu, K.-C. (2009). Ehancing Consumer Behavior Analysis by Data Mining Techniques. *Intermational Journal of Information and Management Sciences*, pp. 39-53.

IBM. (2011). *Distances between Final Cluster Centers*. Retrieved January 29, 2014, from IBM: http://pic.dhe.ibm.com/infocenter/spssstat/v20r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Fidh_quic_ite.htm

Intoweb. (2014). *Customer Behaviour Analysis*. Retrieved May 5, 2014, from Intoweb: http://www.intoweb.co.za/archivedarticles/et_customer_behaviour_analysis.html

Jayakumar, D. S., & Thomas, B. J. (2013). A New Procedure of Clustering Based on Multivariate Outlier Detection. *Journal of Data Science*, 69-84.

Jørgensen, H. (2013). *Gambling in the Guinness Book of World Records!* Retrieved February 7, 2014, from online-gambling.co.uk: http://www.online-gambling.co.uk/articles/Gambling-Guinness-Book-World-Records.asp#.UvUEc3Ckrr5

Katenka, N. (2010). *How to check if the distribution is indeed Normal?* Retrieved January 21, 2014, from Boston University Arts & Sciences Mathematics & Statistics: http://math.bu.edu/people/nkatenka/MA115_FALL2010/QQPlot.pdf

Keselj, V. (2011). *Use of Data Mining on Electronic Gaming Machine Session Variables for Implementation of Responsible Gaming Support.* Dalhousie University, Faculty of Computer Science. (Inner Report)

Laerd Statistics. (2013). *One-way ANOVA in SPSS.* Retrieved December 21, 2013, from Laerd Statistics: https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

Linoff, G. S., & Berry, M. A. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* Indianapolis: Wiley Publishing, Inc.

Marson, S. M. (2011). *Chi-square distribution critical values table.* Retrieved May 30, 2014, from http://www2.uncp.edu/home/marson/syllabi/3600chisquareTableD.pdf

Maths-Statistics-Tutor. (2010). *Performing Normality in PASW (SPSS)*. Retrieved January 23, 2014, from Maths-Statistics-Tutor.com : http://www.maths-statistics-tutor.com/normality_test_pasw_spss.php

Matignon, R. (2005). *Neural Network Modeling Using Sas Enterprise Miner.*

Matsumoto, S., Kamei, Y., Monden, A., & Matsumoto, K.-i. (2007). Comparison of Outlier Detection Methods in Fault-proneness Models. *IEEE Computer Society*, pp. 461-463.

Microsoft. (2013). *Classification Matrix (Analysis Services - Data Mining)*. Retrieved December 22, 2013, from Microsoft SQL Server: http://technet.microsoft.com/en-us/library/ms174811.aspx

Microsoft. (2014). *Training and Testing Data Sets.* Retrieved February 4, 2014, from Microsoft SQL Server: http://technet.microsoft.com/en-us/library/bb895173.aspx

Mooi, E., & Sarstedt, M. (2011). *A Concise Guide to Market Research.* Heidelberg: Springer.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications.* Elsevier Inc.

Northern Arizona University. (2002). *Multiple Regression*. Retrieved December 8, 2013, from Northern Arizona University: http://oak.ucc.nau.edu/rh232/courses/EPS625/Handouts/Regression/Multiple%20Regression%20-%20Handout.pdf

Osborne, & W, J. (2002). *Normalizing Data Transformations. ERIC Digest.* Retrieved January 14, 2014, from ericdigests.org: http://www.ericdigests.org/2003-3/data.htm

Pachgade, S. D., & Dhande, S. S. (2012). Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach. *International Journal of Advanced Research in Computer Science and Software Engineering, 2*(6), 12-16.

Patra, B. K. (2012). Using the triangle inequality to accelerate Density based Outlier Detection Method. *Procedia Technology*, 469–474.

Perfetto, R., & Woodside, A. G. (2009). Extremely Frequent Behaviour in Consumer Research: Theory and Empirical Evidence for Chronic Casino Gambling. *J Gambl Stud*, 297-316.

Philander, K. S. (2014). Identifying high-risk online gamblers: a comparison of data mining procedures. *International Gambling Studies*, 53-63.

Police Analyst. (2012). *Understanding the Difference Between Mean and Median*. Retrieved January 8, 2014, from Police Analyst: http://policeanalyst.com/understanding-the-difference-between-mean-and-median/

Productive Commision. (2009). *Game features and machine design*. Retrieved January 31, 2014, from Productivity Commission: http://www.pc.gov.au/__data/assets/pdf_file/0009/95697/14-chapter11.pdf

Productivity Commission. (2010). *Productivity Commission Inquiry Report Volume 1.* Retrieved February 5, 2014, from Australian Government Productivity Commission: http://www.pc.gov.au/__data/assets/pdf_file/0010/95680/gambling-report-volume1.pdf

Psychwiki. (2008). *Lab #3 – Correlation*. Retrieved November 11, 2013, from PsychWiki: www.psychwiki.com/images/1/16/Lab3Correlation.doc

Pyle, D. (1999). *Data Preparation for Data Mining.* Academic Press.

Razali, N. M., & Wah, Y. B. (2011). Power compaisions of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Stastical Modeling and Analytics*, 21-33.

Remondino, M., & Correndo, G. (2005). DATA MINING APPLIED TO AGENT BASED SIMULATION . European Conference on Modelling and Simulation.

Responsible Gambling Council. (2004). *Canadian Gambling Digest.* Retrieved February 2, 2014, from Responsible Gambling Council: http://www.cprg.ca/articles/canadian_gambling_digest_2004.pdf

Responsible Gambling Council. (2006). *Electronic Gaming Machines and Problem Gambling.* Retrieved July 30, 2013, from Saskatchewan Liquor and Gaming Authority: http://www.slga.gov.sk.ca/Prebuilt/Public/EGM%20Study%20Full%20Report.pdf

Responsible Gambling Council. (2013). *Canadian Gambling Digest 2011-2012.* Retrieved July 30, 2013, from Canadian Partnership for Responsible Gambling: http://cprg.ca/articles/Canadian%20Gambling%20Digest%202011-12.pdf

SAP AG. (2014). *Customer Behavior Analysis*. Retrieved May 5, 2014, from SAP Help Portal: http://help.sap.com/saphelp_sm40/helpdata/EN/3d/b1663b109a4a27e10000000a114084/content.htm

SAS. (2013). SAS® Enterprise Miner™ 12.3. USA. Retrieved January 31, 2014, from http://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf

Schellinck, T., & Schrans, T. (2011). Intelligent design: How to model gambler risk assessment by using loyalty tracking data. *Journal of Gambling Issues*, 51-65.

Scott, S. (2011). *Patron Analytics in the Casino and Gaming Industry: How the House Always Wins.* Retrieved August 3, 2014, from SAS Global Forum 2011 http://support.sas.com/resources/papers/proceedings11/379-2011.pdf

Statistic Canada. (2013). *Variance and standard deviation*. Retrieved November 20, 2013, from Statistic Canada: http://www.statcan.gc.ca/edu/power-pouvoir/ch12/5214891-eng.htm

Statistics Solutions. (2013). *Kruskal-Wallis Test*. Retrieved December 2013, from Statistics Solutions: http://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/kruskal-wallis-test/

Taylor, C. (2013). *What Is the Difference Between Alpha and P-Values?* Retrieved November 20, 2013, from About.com: http://statistics.about.com/od/Inferential-Statistics/a/What-Is-The-Difference-Between-Alpha-And-P-Values.htm

TexaSoft. (2008). *Descriptive Statistics Using Microsoft Excel*. Retrieved January 9, 2014, from Tutorials for Statistical Data Analysis: http://www.stattutorials.com/EXCEL/EXCEL-DESCRIPTIVE-STATISTICS.html

VCU. (2013). *Kruskal-Wallis Tests in SPSS.* Retrieved October 10, 2013, from Statistical Sciences and Operations Research: http://www.stat.vcu.edu/help/SPSS/SPSS.KruskalWallis.PC.pdf

Wicklin, R. (2011). *Log transformations: How to handle negative data values?* Retrieved January 13, 2014, from SAS Blogs: http://blogs.sas.com/content/iml/2011/04/27/log-transformations-how-to-handle-negative-data-values/

APPENDIX 1 RULES OF DECISION TREE

```
*-------------------------------------------------------------*
 Node = 9
*-------------------------------------------------------------*
if TotalBets < 393.047 or MISSING
AND RedeemedPerS >= 415
then
 Tree Node Identifier   = 9
 Number of Observations = 18
 Predicted: PG=3  = 0.50
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.50

*-------------------------------------------------------------*
 Node = 11
*-------------------------------------------------------------*
if TotalBets < 405.914 AND TotalBets >= 393.047
AND Loss >= -85 or MISSING
then
 Tree Node Identifier   = 11
 Number of Observations = 163
 Predicted: PG=3  = 0.96
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.04

*-------------------------------------------------------------*
 Node = 14
*-------------------------------------------------------------*
if TotalBets < 2630.76 AND TotalBets >= 1166.49 or MISSING
then
 Tree Node Identifier   = 14
 Number of Observations = 1452
 Predicted: PG=3  = 0.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.98
 Predicted: PG=1  = 0.02

*-------------------------------------------------------------*
 Node = 15
*-------------------------------------------------------------*
if TotalBets >= 2630.76
then
 Tree Node Identifier   = 15
 Number of Observations = 260
 Predicted: PG=3  = 0.00
 Predicted: PG=2  = 1.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.00
```

```
*-------------------------------------------------------------*
 Node = 17
*-------------------------------------------------------------*
if TotalBets < 393.047 or MISSING
AND RedeemedPerS < 415 or MISSING
AND Loss >= -242.5 or MISSING
then
 Tree Node Identifier   = 17
 Number of Observations = 19492
 Predicted: PG=3  = 1.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.00

*-------------------------------------------------------------*
 Node = 21
*-------------------------------------------------------------*
if TotalBets < 405.914 AND TotalBets >= 393.047
AND RedeemedPerS >= 105 or MISSING
AND Loss < -85
then
 Tree Node Identifier   = 21
 Number of Observations = 39
 Predicted: PG=3  = 0.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 1.00

*-------------------------------------------------------------*
 Node = 25
*-------------------------------------------------------------*
if TotalBets < 416.032 AND TotalBets >= 405.914
AND Loss >= -49.825 or MISSING
then
 Tree Node Identifier   = 25
 Number of Observations = 106
 Predicted: PG=3  = 0.81
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.19

*-------------------------------------------------------------*
 Node = 26
*-------------------------------------------------------------*
if TotalBets < 1129.12 AND TotalBets >= 416.032 or MISSING
then
 Tree Node Identifier   = 26
 Number of Observations = 4834
 Predicted: PG=3  = 0.01
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.99
```

```
*------------------------------------------------------------*
 Node = 30
*------------------------------------------------------------*
if TotalBets < 351.456 or MISSING
AND RedeemedPerS < 415 or MISSING
AND Loss < -242.5
then
 Tree Node Identifier   = 30
 Number of Observations = 75
 Predicted: PG=3  = 1.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.00

*------------------------------------------------------------*
 Node = 31
*------------------------------------------------------------*
if TotalBets < 393.047 AND TotalBets >= 351.456
AND RedeemedPerS < 415 or MISSING
AND Loss < -242.5
then
 Tree Node Identifier   = 31
 Number of Observations = 12
 Predicted: PG=3  = 0.08
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.92

*------------------------------------------------------------*
 Node = 34
*------------------------------------------------------------*
if TotalBets < 399.497 AND TotalBets >= 393.047 or MISSING
AND RedeemedPerS < 105
AND Loss < -85
then
 Tree Node Identifier   = 34
 Number of Observations = 8
 Predicted: PG=3  = 1.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.00
```

```
*------------------------------------------------------------*
 Node = 35
*------------------------------------------------------------*
if TotalBets < 405.914 AND TotalBets >= 399.497
AND RedeemedPerS < 105
AND Loss < -85
then
 Tree Node Identifier   = 35
 Number of Observations = 8
 Predicted: PG=3  = 0.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 1.00

*------------------------------------------------------------*
 Node = 39
*------------------------------------------------------------*
if TotalBets < 416.032 AND TotalBets >= 405.914
AND RedeemedPerS >= 57.5 or MISSING
AND Loss < -49.825
then
 Tree Node Identifier   = 39
 Number of Observations = 56
 Predicted: PG=3  = 0.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 1.00

*------------------------------------------------------------*
 Node = 44
*------------------------------------------------------------*
if TotalBets < 1166.49 AND TotalBets >= 1129.12
AND Loss < -185
then
 Tree Node Identifier   = 44
 Number of Observations = 22
 Predicted: PG=3  = 0.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.91
 Predicted: PG=1  = 0.09

*------------------------------------------------------------*
 Node = 45
*------------------------------------------------------------*
if TotalBets < 1166.49 AND TotalBets >= 1129.12
AND Loss >= -185 or MISSING
then
 Tree Node Identifier   = 45
 Number of Observations = 83
 Predicted: PG=3  = 0.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.08
 Predicted: PG=1  = 0.92
```

```
*-------------------------------------------------------------*
 Node = 56
*-------------------------------------------------------------*
if TotalBets < 416.032 AND TotalBets >= 405.914
AND RedeemedPerS < 57.5
AND Loss < -49.825
AND BetsPerMin < 17.1 or MISSING
then
 Tree Node Identifier   = 56
 Number of Observations = 12
 Predicted: PG=3  = 0.00
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 1.00

*-------------------------------------------------------------*
 Node = 57
*-------------------------------------------------------------*
if TotalBets < 416.032 AND TotalBets >= 405.914
AND RedeemedPerS < 57.5
AND Loss < -49.825
AND BetsPerMin >= 17.1
then
 Tree Node Identifier   = 57
 Number of Observations = 8
 Predicted: PG=3  = 0.88
 Predicted: PG=2  = 0.00
 Predicted: PG=4  = 0.00
 Predicted: PG=1  = 0.13
```