# A model to predict churn

Hilda Cecilia Lindvall

April 18, 2014

**Abstract**

This Master Thesis has been performed at Svenska Spel with the aim to detect playing customers probability to churn, i.e. quit their gambling. The first part of the Thesis consists of some previously work done within the field, some facts about Svenska Spel and explanations of used software. The next part describes how the work has been done and the third part give the reader the theory behind the prediction model. The model used for prediction churn is Logistic Regression with related statistical test to investigate and verify the model. Finally, a prediction model and verification results are presented.

# Contents

# 1 Introduction

During the recent years, due to the advancement in technology, the relationship marketing has become even more important and available. Companies can gain competitive advantage when investing in their customer relationship management. With help from tools as data warehousing, data mining and campaign management software, deep insight of the customers' behavior can be extracted and analyzed. Large databases can hide important information of their customers and with different techniques hidden patterns and information can be exposed. All this gathered knowledge, helps the company to make proactive and smart knowledge driven decisions.

The business culture is dynamic and always changing. This also means also that the economics of customer relationship are changing and companies are now facing the need of new solutions and strategies to meet the changing demand. Before, concepts as mass production and mass marketing was in focus but nowadays the focus have, in many ways, changed from the product oriented view (design-build-sell) to a more customer oriented way (sell-build-redesign). Nowadays, companies are not only focused on increasing the number of customers, they now have the focus to increase the customer value through Customer Lifetime Value and Customer Lifetime Cycle. The new approach can be seen as a direct customer oriented strategy and structure. Also, a so called *wallet share* is of interest, which shortly can be described as a measure for how much a specific customer spend his or her field budget at the company itself. So, if buying all food at one company, say ICA, the share of the customer's wallet is high. If the customer buys some food at ICA, some at Willys and some at COOP, the wallet share with ICA is lower. Thanks to the Internet, customers nowadays have deep knowledge about products they buy and also about the companies they buy from and they get more and more informed and sophisticated. Companies need to put huge efforts to make themself a name and this can for example be done with precision targeting. The precision targeting is based on customers' demographics and behavior data.

## 1.1 Background

To get a deep insight in customers' behavior, data mining is used. The definition of data mining is *sophisticated data search capability that uses statistical algorithms to discover patterns and correlation in data.* The data mining is the procedure of discovering the unknown so that the company can use precision targeting. Many tools are used for data mining but without knowledge of the specific area and company, the data extraction is useless. So, data mining in combination with deep insight of the market can provide business value for the company. [32]

Examples of where data mining can be used to improve business:

- The retail industry can use data mining to analyze the basket of a customer so the company can improve stocking, store layout strategies and promotions. They may also be interested in which customers who prefer shopping when sale is on so the promotions and sale events can be pinpointed for these.

Figure 1: Data mining - a modern magnifying glass [30]

- The banking industry can discover and develop specific products for the different customer segments such as different types of credit cards or to be prepared when a customer's life makes a big change such as a buying a house or having children.
- Manufactures can be interested in the change in warranties when the number of customer increase.

Of course, data mining does not on its own extract all information asked for. Statistical analyzes, analytical processing and spreadsheets are only some of many different complements that are good to use to reach the provided goals about a deeper customer insight.

With help from data mining techniques, customer churn can be investigated. In the article "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models", S. Neslin et al write about how different churn predictive models have different accuracy. The article is based on results from a tournament with the aim to create the best predictive churn model. It was administrated by the Teradata Center for Customer Relationship Management at the Fuqua School of Business at Duke University. The contestants came from both the academic field and from the industry. They had all the same data material to start with and it was up to themself to create their best predictive churn model. Many details are considered in the article and the conclusion was that the logistic regression and decision trees were good methods to start with due to their simplicity and good results. The winner of the competition used a combination of quite basic trees (not more than 8 nodes) done through a gradient tree-boosting procedure.[33]

The gambling industry is a part of the entertainment industry. If summarizing the revenue of all other forms of entertainment, the gambling industry still generates more revenue than all of them together. Additional to this, in the countries where it is legal, the gambling industry is the leading source to federal- and wealth taxes. At the consumers' product market, gambling generates the highest revenue with the highest frequency of purchase of all product purchases. [1]

Due to the saturated markets it is essential keeping existing gamblers[2] as they tend to switch to another gambling site if they do not feel satisfied with the products they buy. As a switch often is without any transaction costs it is of great importance keeping customers satisfied. Research shows that various parameters influence the satisfaction, retention and customer profitability.

Service quality such as a well-managed complaint system, the odds for winning and the friendship society have significant impact of the customer satisfaction.[3]

To meet the demand and to compete with the market, Svenska Spel wants to change from a product-oriented to a customer-oriented focus. This focus assumes strong knowledge about the customers and their behavior so due to this, deeper analysis of this is required. With this thesis, Svenska Spel wishes to receive information of which parameters that influence the lifetime of the customer and possible change in gambling. So, this thesis is a small part of the upcoming customer-oriented focus at Svenska Spel.

## 1.2 Svenska Spel

Svenska Spel is under license from the Swedish Government and arranges reliable gaming and lotteries. Their vision of the perfect customer basis consists of a large number of Swedes, who each gamble regularly but for a reasonable amount of money. [1] The import of decent and healthy choices regarding gambling intensity and level pervades the whole organization.

One reason of why many companies emphasis retaining their customers is due to the Loyalty Effect. The expression is derived from a study presented in year 1996 which showed that by retaining 5% of a company's customers, the profitability of the company could increase by between 25% and 100%. [31] Svenska Spel aims to share the fun of gambling but with a high level of responsibility. Svenska Spel has the ambition to have loyal customers who enjoy the products offered and whose mentality goes in line with the one at Svenska Spel. Even though Svenska Spel has an interest in retaining customers and making a profit, Svenska Spel emphasis detecting customer with gambling problems. Those customers contribute to the revenue at Svenska Spel but are not in line with the mentality at Svenska Spel.


Figure 2: The logo of Svenska Spel

Customers with gambling issues represent approximately 2% of the existing customers. It is difficult to estimate how much the customers costs the community but research estimates the cost to approximately SEK2.3-4.5 billion[2] per year. Verdict per customer results in SEK 17 000-33 000 per person with gambling problems. This it though a rough estimation and there are reasons to believe hidden statistics. Research claims that 7-17 persons are negatively affected by a customer's gambling problems, for example family, colleagues and friends. [4]

In year 2008, Svenska Spel was appointed the *the Most Responsible Gambling Company* by the World Lottery Association.[4] Before revenue comes responsibility and this pervades all within the company. An example of this is *Playscan*[3],

[1] A reasonable amount of money can be seen as an amount that does not affect the customer's economy in a negative way

[2] In comparison with a community without any gaming players

[3] Available for online gamers

a tool to facilitate the customer's consciousness of their own gambling. Playscan can identify significant changes in how a customer plays and in that way informs the customer if his or her gambling behavior indicates a risk of developing gambling problem. Further, the customers are forced to state a weekly maximum amount of money that can be spent on gambling. If the weekly amount of money is used, the customer will have to wait until the next week to insert additional money and continue gambling. The weekly amount can be changed but the change is done for the upcoming week so it is not possible to play the weekly amount, change the weekly amount and then, during the same week, continue gambling. This creates a limitation of how much a customer can lose each week. In year 2008, two months after the weekly amount was introduced to the customers, the gambling at svenskaspel.se decreased with approximately 5-10%, which in revenue means a decrease of SEK2.5-3 millions per week. The benefit of the limitation is that no customer was lost, instead the decrease in gambling was a result of the weekly limitation of how much a customer could spend on gambling.[4] There is no guarantee that the customer ends his or her gambling but in many cases the warning indications contribute to the customer's knowledge about his or her gambling problems. Unfortunately there is a risk that the customer decreases or quits gambling at Svenska Spel but increases their gambling with a competitor instead.

Gambling problems has always been a dilemma and will unfortunately probably always exist. The reason of why Svenska Spel hopes to keep their status and their state of affairs is that they presumably work harder with gambling responsibility than anyone of their competitors. As a result of this Svenska Spel presumably identifies gambling problems quicker than others. Further, the profit of Svenska Spel accrues the Swedish State, in other words the money goes back to the community and is invested in the youth sections within different sport associations throughout the country.

Svenska Spel may of course improve their work with problem gambling and debates about the Gambling market are vivid. This Master Thesis is from a mathematical and statistical view and I am well-aware of the positive picture I have transmitted above. I want you to note that based on my knowledge from my time at Svenska Spel, I confirm the massive focus on responsible gambling. I can also confirm that the new decision that almost all gambling will need to be registered, is a tool for a customer to understand the amount of money he or she spends on gambling. So, if economic and social problems exist due to gambling, they hopefully arise quicker so the problem can be solved. I am biased and I believe my view of the Gambling market and Svenska Spel will be more unbiased in a while.

## 1.3 Aim of the thesis

The aim of this thesis is to present the research, findings and conclusions done during four month research at Svenska Spel. The research is regarding customers who significantly have decreased their gambling at Svenska Spel and to create a predictive model of which customers are about to decrease their gambling.

The thesis and research is performed at Svenska Spel and the results need to be in line of their expectations. The general and current specifications for a Master Thesis at The Faculty of Engineering, Lund University, are also taken in consideration. The disposition emanates from the recommended disposition in the course *'Metodikkurs for Examensarbete, MIO920'*[8] at Lund University. The level of language emanates from the article *'Hints on Layout and Style for writers of Dissertations and Theses at the Faculty of Science, Engineering and Medicine, Lund University'* by Helen Sheppard.[6]

The tool needed for the task is mainly SAS®Enterprise Guide®, henceforth called SAS®. As this program is new for the researcher, the procedure of learning the programs is also considered as a task to solve. Programs such as R, Oracle and Google Visualization has been used to give a better understanding os the data material provided and how SAS®works. Additional, the report is written in LaTeX.

Notable is that a predictive churn model classifies customers who have properties as previous churners and this does not say anything about the customers' desire or willingness of terminating the relationship.[3] It is also good to note that the significant parameters are significant but they do not explain the reason of churn, they only give indications of churn. To get the reason of churn, one needs to do, for example, surveys and questionnaire studies which are outside the aim of this Master Thesis.

## 1.4 Related and previous work

In the article *'A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques'* Abbasimehr et al.[9] present a table containing an overview of literature of churn prediction. As seen, different modeling techniques can be used and in this project focus lies on logistic regression. Those two are the most common models and they are mentioned in a large number of articles that consider churn. Other models are for example Random Forest, where different decision trees are combined; and Neural Network which is an analogous data processing structure.[10] For more information of other models, Kim et al. [3], Jolley et al. [1] and [5] can provide this.

In the telecom industry, the churn rate is very high (20-40%) and a lot of the research found comes from their point of view. Many of the companies have developed a predictive churn model that accurately identifies customers with a high probability to churn. An conduciveness reason for the well known churn models in the telecom industry might be first due to the saturated and competiative market but also thanks to their massive detailed data for and from each customer. [3] With the data provided, deep analysis can be provided and also with the parameter *social network* included. This parameter consists of the customers influence on other customers and by knowing a customers call graph (data which includes details of calls including caller identifiers, date, time, length of the call and so on), the company can take conclusions on how the customers effects each other. If the company's customers have free calls to other customers within the company, they surely influence each other if one or many changes

telecom company. Very often, when a telecom subscription is about to end, the telecom company contacts you to offer "latest deals with the best price" as they really want you to stay as a customer. In the article *Prediction of subscriber Churn using Social Network Analysis* by Chitra Phadke, she mentions the ACM Conference on Knowledge Discovery and Data mining and their competition on predicting mobile network churn using a large data set from Orange Labs. One of the models won the prize for the highest predictive accuracy. With this we can understand the huge value for retaining customers. [7]

The social network parameter is not considered. The weight of analyzing a customer by Network Value increase as the social network increase. The phenomenon is called Word-of-mouth communication and is when no commercial communicators have conversations about products or companies. [9]

## 1.5 Problem formulation

Customers have different expected behaviors and they gamble for varied amount of money. Based on recency of customer purchases, the purchase frequency and the monetary value of the historical transactions, frequency, recency and monetary, it is possible to calculate the probability if the customer stays at Svenska Spel as a customer or if the customer quits the gambling and then becomes a churn customer. This thesis will be based on the following:

1. A deep understanding of the different traits of the customers that are collected in the data base. This is done with work in Oracle that includes data base theory and MySQL oriented language.

2. When choosing many parameters that might influence the churn probability, the data material is divided into two groups (Churn group and Active group) so the parameters can visually be compared with each other. The parameters that seem to be significant will qualify into the model building. The parameters will then be statistically investigated to see if they are significant or not.

3. With help from previous points, try to create a predictive model to detect churn.

## 1.6 Concepts

Customer churn can be described as a customer whose propensity to buy has decreased. The definition is universal but within different areas and markets the definition is slightly modified. At Svenska Spel churn is defined as a customer who has not gambled the last 52 weeks i.e. if last gambling date is 1th of January and no gambling has been done during next 52 upcoming weeks, the customer churns on the 53rd week.

# 2 The Data and Statistical Tools

During the ongoing evolvement some new software has been introduced. The concept of Database has also been new for the author so it has been a process also to get an understanding for that.

## 2.1 Database

A *database* is data sorted in some complex way due to their relationship. All this data needs to be built up and investigated which can be done with a Database Management System, for example Microsoft Access, Microsoft SQL Server, MySQL and Oracle. In this Master Thesis, Oracle has been used to give us an understanding of the content of the database at Svenska Spel. The reason for a database is to have a structure in how data is sorted and by this, huge amounts of data can easily be understood, investigated and manipulated, even when facing complex tasks. The database can be built up in many different ways and the most common structure is the relation model which contains the data in different *fact tables* and where different *dimension tables* are joined to describe the data in the fact tables. The relation database structure is seen in figure (3). [25] p. 8

The database at Svenska Spel is an *active* one with triggers as new data is added each day. A trigger is an event that occurs due to some condition and makes an action. An example can be when new data is added each night. This means that the database is continuously changing. As we are looking at data from 2012 and 2013, our material is not directly affected of the changes done in 2014. As long as the codes are correct, no problem would occurs due to an active database. [25] p. 276

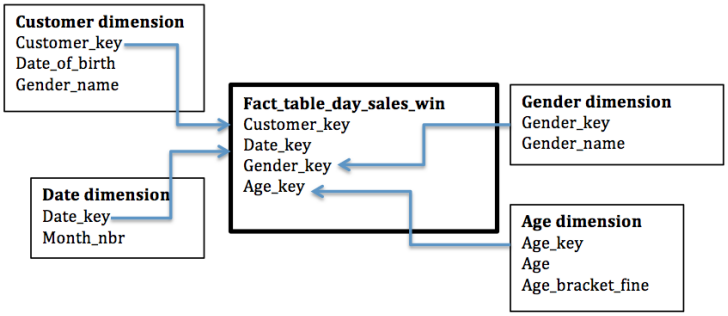


Figure 3: The figure illustrates how the relation database is built up. The attributes are some of many found in the database at Svenska Spel and they are, together with many others, used in this Master Thesis.

In the database at Svenska Spel there are approximately 50 fact tables and approximately 100 dimension tables. Many of them are not directly in use and in this Master Thesis we have used only a few of these. Much of the data in the dimension tables are information that can be extracted from the fact tables

or from other dimension tables so that is one of the reasons why we have been satisfied working only with a smaller amount of tables. In figure (3) some attributes of many used are shown. In this Master Thesis and particular in this report, no descriptions of the used variables are given. We will see the output and results with their correct values but we will not give any information on which variables are used or which is which.

### 2.1.1 Oracle SQL Developer

To get an understanding and overview of the data material collected in Svenska Spel's database, some research with Oracle SQL Developer has been done. The software is free of use and can be described as a graphical tool for database development. SQL statements and scripts can be run and editing and debugging SQL can be done. Provided reports can also be created. Although, in this Thesis, focus has been to understand how fact tables and dimension tables are connected and to understand how to obtain the requested data. Oracle has composed a good tool for facilitate the understanding of the data provided and to get a feeling of how SQL coding works. [20]

### 2.1.2 SAS

SAS®Enterprise Guide®is a statistical tool that facilitates the exploring of large amount of data and to distribute the results to target users. By knowing the data material provided and the chosen statistical methods, one can reach deep insight on patterns within the data material and to create models. Many different test are predefined in SAS so it is a nice way to do statistical tests. Some different statistical tests for goodness of fit are used in SAS and the theory for those tests are found in Appendix. [21] When the data material has been known for the author, the work in SAS®has been done. SAS®has been a good tool for selecting the parameters and to understand how these are related to each other.

### 2.1.3 R

R can be seen as an environment embedded with implemented statistical techniques. Through additional *packages*, R can easily be extended to yet more statistical tools. Another property is the simplicity to connect R with other software and tools, for example Google Visualization. Some work has been done i R but aimed mostly, just as Oracle, to get an understanding of the data in the database at Svenska Spel. Google Visualization has also only been used in the same aim. [22]

# 3 Method

Based on previous events, cases and experiences, a prediction model aims to describe what will happen in the future for the target variable. In this thesis, the target variable is the probability of churn. To create a predictive model, two steps are needed. The first step is called the training step and is the identification step. In this step, the history and properties of customers are considered and investigated so that distinctive properties can be selected. Since the target variable is the probability of whether a customer will churn or not the selected variables aim to describe and determine the relationship to the target variable. The second part is called prediction or test step in which the developed relationship is tested. If the relationship is significant, the prediction model can be more or less accurate. If the relationship is weak, the prediction model maybe useless. [2]

The work with this Master Thesis started by getting an introduction to Svenska Spel and all data provided in the data base. The Author was not well-grounded with the gambling market or Svenska Spel so research and background about this was done. Also, effort was done to get an understanding of a data base and especially the data base at Svenska Spel. The invesigation part was done mainly with Oracle SQL Developer but also some work was done in R. Also, a lot of research about previously churn models was done. The report writing has been done continously during the whole process.

The next part of the Master Thesis has been to learn about working in SAS®. Thanks to the newly knowledge with SQL, the work in SAS®was facilitated (but still difficult!). Also, many embedded functions was found during this period.

When knowing SAS®better, the work to create a big table of customers and their attributes was done. The method was to chose all playing customers during year 2012 and many of their corresponding, eventually significant for churn, attributes. When this work was done, we could divide the data into two groups - customers who had continued their playing during year 2013 (so called active customers) and customers who not had been playing anything during year 2013 (these are our churn customers). Then major work was done to get a view of these two groups different attributes and habits. During this work, Excel was used to create comparing histograms, intervals and so on. No one of these figures are included in the report as they expose too much of the data and the eventually significant parameters. Some of the tables that has been the underlying data for these figures are however included but with no explanation of what the numbers mean.

After the descriptive statistics, the work in SAS®was continued. The correlation between the customer status (active or churn) was analyzed and eight of all the parameters was chosen as candidates. The correlation between these was taken in consideration. The test data consisted of 200000 customers with the two chosen attributes and the validation data consisted out of the same number of customers but non of them were already in the model. We used the procedure stepwise elimination when analyzing which and how many of our

eight parameters that was good in the model. The stepwise selection choses the parameters most significant by going forward and backward with the parameters in the model. Out of eight parameters, all of them was significant for the churn model. However, as two of them improved the model only nominal, these two were not taken into the model. So, the decision was made to use only two parameters in our final churn model. The two parameter-model was compared to another (same input) two-parameter model but with slightly different estimations due to a, for this Master Thesis, hidden restriction. This was done due to curiosity, we wanted to see if the models was improved or not. When having these two models based on two slightly different test data, we could use them together with the validation data, to see if the prediction model was working or not. Also, for the more restricted model, the validation data was restricted in the same way. When having scores in how well the two models predicted churn, the work with summarizing the results was done.

Please note that the issue of problem players is not taken in consideration in the model building in this Master Thesis. Mainly as it has been hard to define who has problem with their gambling and who don't. In the provided data material, information on each customers' status in Playscan is known but as the status is changing over time, is has been difficult to draw a line to get a definition of problem playing indication.

# 4 Algorithms

*Generalized linear models* are a broad class of models for both continuous responses and discrete responses. In this thesis, we mainly focus on Logistic regression model that is a generalized linear model, GLM. [27] p. 65 Three components are building a generalized linear model:

The *random component* consist of the output variable Y. Y has an associated probability distribution. Dependently whether Y is discrete or continuous, the distribution is assumed to binomial respective normal. Other names for the random component are the dependent variable or component. Given a discrete Y, the output can be called some different names: *Success*, *case* or *1* are equivalent for the noteworthy possible outcome. Their opposite outcomes are called *failure, noncase* and *0*.

The *systematic component*, or the *independent component*, consists of the input explanatory variables, $x_i, i = 1...k$. A linear combination of different inputs is called a linear predictor.

The *link function*, $g = g(\mu)$, contains information on how the random component is dependent of the independent component. It is a function of the mean of the random component variable and it is assumed to have a linearly relationship with the explanatory variables. The mean $\mu$ depends only on the stochastic behavior of the random variable and as the explanatory variables are assumed to be fixed, the function $g$ is the link function between the systematic, deterministic input variable and the random variable $Y$. A link function does not need to be linear. In this Master Thesis the link function $g(\mu)$ =log$(\frac{\mu}{1-\mu})$ is used. As we are interested in calculating probabilities, this *logit* link suits the aim well and is the link connected to Logistic regression. Other link functions that are widely use are the probit function, $g(p) = probit(p) = \sqrt{2}\text{erf}^{-1}(2p-1)$, which is associated with the standard normal distribution's quantile function and the log-log function, $g(p) = log(-log(1-p))$ where $p$ is the probability for a random event $Y$ given systematic components $x$.

## 4.1 Linear regression analysis

The strength of a linear relationship between the explanatory variables $(X)$ and the systematic variable $(Y)$ can be calculated with linear regression analysis. If the relationship is bivariate and linear, it is given by

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

The parameters $\beta_0$ and $\beta_1$ are estimated by the *Ordinary least Squares (OLS)* estimation method. The method can be found in Appendix. The intercept composes of $\beta_0$ and the change in Y associated with a one-unit increase in X is the best linear estimate of Y from X and composes of the slope coefficient $\beta_1$. When having several predictor variables the regression is called multiple regression with the following relationship:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{2}$$

Here, the different $\beta$'s are called partial slope coefficients as only a partial explanation or prediction of Y can be predicted by one of the explanatory variables. As the prediction of $Y$ given $x_i, i = 1...k$, contains noise, the equation (1) contains the error called $\epsilon$.

The input x may contain error and the connection between x and Y may also contain error. Those two types of errors are summarized to a *noise term*, $\epsilon$ If a good model, we assume $\epsilon$ is $\in N(0, \sigma^2)$ and independently of each other. The expectation is a normal distributed noise but if the distribution is non-normal, the model and parameter estimation may be bad.

When $Y$ composes of individual cases, say $Y_j$ for the first case and $Y_{(j+1)}$ for the second case and so on, the calculation of $Y$ for a peculiar case $j$ is given by has the appearance

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_k x_{kj} + \epsilon_j \tag{3}$$

Just as in the bivariate linear regression case, the coefficients in multiple regression models are estimated by OLS. In the bivariate linear case, the notation of estimated Y is as follows:

$$\hat{Y} = \beta_0 + \beta_1 x \tag{4}$$

For the linear multiple case we have

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{5}$$

With estimated coefficients for the intercept and partial slopes $\hat{\beta}$ . For each case, $\epsilon_j$ are equal to $Y_j - \hat{Y}_j$ where $\hat{Y}_j$ is the estimated value of $Y_j$ for case $j$. If the regression is bivariate, the residuals easily can be shown visually but in the multiple regression case, the dimensions constrain the visualization.

### 4.1.1 Evaluation of the parameters

The regressions coefficients $\beta$ are, as said before, normally estimated by *OLS*. The OLS tries to find the parameters that minimize the *Residual Sum of Squares, (RSS)*

$$RSS = \sum_{i=1}^{N} [y_i - f(x_i)]^2 = \sum_{i=1}^{N} [y_i - \beta_1 x_1 - ... - \beta_0]^2 \tag{6}$$

also called the *Sum of the Squared Error, (SSE),*. By taking the partial derivative of SSE with respect to the different $\beta$, putting each of them to zero and by solving the equation system rised, the parameters $\beta$ are estimated by OLS. Some notational work can be done to get an expression for the different $\beta$ and you find the details in appendix. As a summary, the OLS is a systematic method to estimate parameters in the linear model for the response variable y. The SSE between the true and estimated value of y rises OLS. Even though the method is simple, the method is working well as a linear model is the first order Taylor series approximation for any function with continuous derivatives.

We have previously defined the SSE and the following error measures can also

used:

$$SST = \sum_i [y_i - \overline{y}]^2 \tag{7}$$

$$SSM = \sum_i [(f(x_i) - \overline{y}]^2 \tag{8}$$

$$\tag{9}$$

where SST is an abbreviation for Total Sum of Squares and SSM the abbreviation for the Regression Sum of Squares. In SST, the prediction error composes when the average value is used as an estimate of the output variable. SSM is instead a measure of the amount of error in the regression model. Together with the SSE, these three error measures has the relationship (letting $\beta = \beta_1$):

$$SSE = \sum_i [y_i - \overline{y} + \overline{y} - f(x_i)]^2 \tag{10}$$

$$= \sum_i [y_i - \overline{y}]^2 + \sum [f(x_i) - \overline{y}]^2 + 2 \sum_i (y_i - \overline{y})(\overline{y} - f(x_i)) \tag{11}$$

$$= \sum_i [y_i - \overline{y}]^2 + \sum [f(x_i) - \overline{y}]^2 + 2 \sum_i (y_i - \overline{y}) \beta_1 (x_i - \overline{x}) \tag{12}$$

$$= \sum_i [y_i - \overline{y}]^2 + \sum [f(x_i) - \overline{y}]^2 + 2 \sum \beta_1^2 (x_i - \overline{x}) \tag{13}$$

$$= \sum [y_i - \overline{y}]^2 - \sum [f(x_i) - \overline{y}]^2 \tag{14}$$

$$= SST - SSM \tag{15}$$

where the following relationship are used:

$$\overline{y} - f(x_i) = -\beta_1 (x_i - \overline{x}) \tag{16}$$

$$\sum_i [y_i - \overline{y}][x_i - \overline{x}] = \omega_1 \sum_i [x_i - \overline{x}]^2 \tag{17}$$

To analyze the goodness of the fit, the following measure is used:

$$R^2 = \frac{SSM}{SST} \tag{18}$$

$R^2$ is called the *coefficient of determination* and ranges between 0 and 1. If the variability observed in the output can be explained by the regression model, the $R^2$ is closer to one. If, on the other hand, the variability observed in the output cannot be explained by the regression model, the $R^2$ is closer to 0. [[12] p. 730-735]

Considering the linear regression model, many more details and explanations can be given. Although, as the method not is in use in this thesis, the information and details are given only to make a good theoretical ground for the used method and where we can use parallel reasoning when describing and understanding the other method in use.

In many cases, the relationship between the explanatory variables and the target variable is not linear. In that case, it is possible to transform one or more

of the independent variables or the target variable itself. This new relationship can hopefully be a linear relationship so that OLS can be used in coefficient estimation. Note that the *substantive* relationship remains nonlinear in terms of its variables but linear in terms of its parameters.

If one are interested in extending the linear regression so it can be built on dichotomous or dummy explanatory variables, this can be done. If the dependent variable is dichotomous the variable's mean is a function of the probability where the different cases is depending on which interval the variable gets into. By coding the variable's value as 1 or 0 one can receive the proportions by calculating the mean of the variable. If Y is a probability we want the prediction to give values in the interval $[0, 1]$. This is not a restriction for a linear regression method as its output can be given any value. Another problem is the residuals as they are not normally distributed due to the dichotomous Y and also the sampling variance are wrongly estimated. More about this and how we extend the linear regression model will follow. ([24] p. 2-18)

## 4.2 Logistic Regression

A logistic regression model can retrieve information if and how a response variable is dependent of the explanatory variables. The explanatory variables can be either continuous or categorical data. The dichotomous outcome makes the approach a bit different in comparison to the outcome in linear regression. Given certain values of X, we predict the likelihood that Y equals 1 (rather than $\neq 1$). This means an existing positive relationship if the value of Y increase if the values of X increase. As mentioned, three major reasons make the logistic regression preferable rather than the linear regression:

1. We are searching for a probability so the dependent variable Y need to be restricted to the range $[0, 1]$ which not can be a restriction in a linear regression model.

2. An assumption in linear regression is that the variance of Y is constant independently of values of X. In our case, Y is a binary variable which implies a non constant variance as the variance consist of $p(p-1)$ where $p = P(Y = 1 \mid x)$.

3. The assumption of normally distributed prediction errors, $(Y - \hat{Y})$ to test the significance of the $\beta$ weights can not be applied. This is due to the dichotomous dependent variable $Y$.

The solution in logistic regression is not as printed and straight forward as the one in linear regression with OLS estimated parameters. To make things even trickier, the unstandardized solution in the logistic regression does not have a straightforward interpretation as the solution is in the regression model. To get around some of the issues, the concept *Odds* will be defined.

### 4.2.1 The Odds and the Logit

Odds is introduce to get around the issue of the restriction of which values Y can take. In this Master Thesis, Y is a probability that means values in the range 0 to 1.

As seen before, the numerical value of the dichotomous systematic variable is not intrinsically interesting. What focus instead lies on is whether the classification cases can be predicted by the explanatory variables. It may then be useful to reconceptualize the problem and instead try to predict the probability that a specific case is classified. For the dichotomous variable, the complement of one probability is the probability for the other case so by knowing one probability we know the other one. Like the probability, the odds has a minimum value of 0 but is does not has a fixed maximum value over 1. The odds ratio of Y taking a value 1 is defined as

$$\text{Odds}_{Y=1|x} \triangleq \frac{P(Y=1 \mid x)}{P(Y=0 \mid x)} = \frac{P(Y=1 \mid x)}{1-P(Y=1 \mid x)} \tag{19}$$

with the synonym name *likelihood ratio*. The Odds is the fraction between the probability of occurrence of an event and the probability of nonoccurrence. As we are searching for the probability of Y being 1, we search $P(Y=1 \mid x)$ which can be evaluated from the equation (19). The dependent variable Y in logistic regression is the natural logarithm of the odds:

$$\log(\text{odds}_{Y=1}) = \log\left(\frac{P(Y=1 \mid x)}{1-P(Y=1 \mid x)}\right) = \text{logit(P(Y=1}| \ x)) = \text{logit(p)} \tag{20}$$

where we have used $p$ for $P(Y=1 \mid x)$. By applying the assumption that the log-odds, the *logit*, of an observation y can be expressed as a linear function of the input variables, we get

$$\text{logit(p)} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{21}$$

Which is our link function and which is, by our assumption, linear. So, related to X, the logit is lineary. As logit(p)$=ln\frac{p}{1-p}$ we have

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{22}$$

and after some steps,

$$p = P(Y=1 \mid x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}} \tag{23}$$

Which is the answer we are looking for but as it is non linear, the expression looks messy and are tricky to use.

As Y is a binary variable, the mean and variance are the following:

$$E[Y] = P[Y] = p \tag{24}$$

and

$$Var[Y] = p(1-p) \tag{25}$$

where p is as in equation (23) and where $0 \leq p \leq 1$ independently of values of $\beta$ and $x_i$. As the relationship between Y and the parameters $\beta$ is non linear, the interpretation in how Y is affected is hard to explain in a straightforward way. In linear regression, this dependency is linear and easy to interpret. [24]

### 4.2.2 Estimation and Goodness of Fit

To measure how the mathematical model of data differs from the actual data, a *loss function* can contribute with information. As seen before in the linear regression model, it is possible to measure how good the model fit the data (*goodness of the fit*). With OLS the parameters in the linear case be chosen.

Unfortunately, an easy way like this for parameter estimation in the logistic regression model does not exist. The Loss function for these models are based on the *maximum likelihood (ML)*.[14] Instead of minimizing the sum squared residuals, SSR, as in OLS, the ML finds the smallest possible deviance between the predicted values and the observed ones.

The deviance is found when no better fit is possible to find. [15] The *Convergence criteria* is the procedure to investigate if evaluated parameter estimates maximize the Loglikelihood. The convergence criteria needs to be fulfilled, if not no conclusions can be taken. The iterative process and calculations are already implemented and optimized in the computer programs by numerical algorithms and in SAS®the two mentioned above can be chosen. The special case when the Maximum Likelihood estimation does not force an iterative process, the solution and method are identical to the OLS estimates for the linear regression coefficients. For the binary response data, the parameters can be estimated by the Maximum Likelihood method where methods like the Fischer-scoring algorithm or the Newton-Raphson algoritm is used. In SAS®these two methods can be chosen as estimation and optimization techniques. Both these iterative methods generate the same estimates of the parameters but as the first-named method is based on the expected information matrix and the second-named based on the observed information matrix, they slightly differ in the estimated covariance matrices. Though, in our case with a binary logit model, their respective matrices are the same. The Fischer optimization technique is equivalent to fitting by iteratively reweighted least squares and this method is found in appendix. As we choose the Fischer-technique, the Newton-Raphson is not explained.

When goodness of fit is obtained, one can create tests to see if the goodness of fit is strong enough. The *Deviance* is a measure for the Goodness of the Fit and is obtained by twice the negative loglikelihood and the Pearson goodness of fit is obtaining from $\chi^2$-value. These two methods can be evaluated through the Likelihood ratio test and the Pearson's $\chi^2$-test which both are evaluated by reference to the $\chi^2$-distribution. The tests compare how the fit of the model has increased when input parameters are added. We expect the deviance to decrease as the degree of error in prediction decreases when adding another input variable. [29]

### 4.2.3 The Hyphothesis

A model with only the intercept is a model without any independent input variables can be called the *null model*, $H_0$, as the model would be verified by the Null hypothesis. By adding $k$ input variables to create a ful model, $H_1$ we

then can investigate if the model is better with the input parameters or not.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \tag{26}$$

$$H_1 : \beta_1 \neq \beta_2 =\neq ... \neq \beta_k \neq 0 \tag{27}$$

Which can be explained as, seen from a statistical view:

$H_0$: The input variable $x_i, i = 1...k$ does not influence the response variable y
$H_1$: The input variable $x_i, i = 1...k$ does affect the response variable y

If we not statistically can prove a significant model with more than only the intercept, we must keep $H_0$. Then we better try another model for our purpose.

Up to three different methods are used in this master thesis to investigate goodness of fit. We start by supposing the model contains $s$ inputs. The estimated probability of the observed response is $\hat{p}_j$ where index $j$ stands for the $j$th observation.

### 4.2.4 -2LL and the Likelihood Ratio Test

As we are working with the logarithmic odds, we can call the expression made by ML estimation for the *log likelihood* and it gives information of the selected parameters. The log likelihood from the ML estimation is as said before, our *deviance*. It is the corresponding measure for $R^2$ in the linear regression case. To judge the overall fit of a model with $\chi^2$ distribution, the transformed deviance (-2LL) can be used. As the -2LL is $\chi^2$ distributed, we can use $\chi^2$ distribution to investigate if the model is good and by using our deviance we can make conclusions of the goodness of the fit. [13]

Starting with the Null Hypothesis and by adding $k$ input variables to create a full model, and by taking the ratio of the resulting likelihood values, we get the *likelihood ratio test*. The difference between two values of -2LL have as said before, a $\chi^2$ distribution, it is our Deviance (D). D investigates the hypotheses

$$D = \chi^2 = -2LL_0 - (-2LL_k) = -2\ln\left(\frac{LL_0}{LL_k}\right) \tag{28}$$

where $LL_0$ is the log likelihood for the constant model (meaning no impact from the explanatory variables) and where $LL_k$ is the log likelihood for the full model with k input parameters. The null hypothesis, $H_0$, is rejected if the log likelihood is statistically significant and the conclusion that the independent variables gives a better prediction of $p$ than without them. [13] With $f$ Degrees of Freedom, our G is asymptotically converging to a $\chi^{"}$ distribution under specific regularity conditions. The number of Degrees of Freedom is evaluated as the difference in number of parameters in the saturated model and in the number of parameter in the considered model.[29]

$$D^2 \sim \chi^2_\alpha \tag{29}$$

And if the Deviance is bigger than expected value from the $\chi^2_{1-\alpha}$ distribution, where $\alpha$ is the significance level, $H_0$ is rejected: [29]

$$D^2 \geq \chi^2_\alpha \tag{30}$$

### 4.2.5 Pearson and the $\chi^2 - test$

Another Goodness of Fit measure is Pearson's $\chi^2$-test statistic which is $\chi^2$ distributed. Due to this, is is easy to verify the acquire model validation. The Pearson's $\chi^2$ is given by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_{observed} - f_{expected})^2}{f_{expected}} \sim \chi^2_\alpha \tag{31}$$

Where $f$ is frequency. One can investigate if the output, given the null hypothesis, can be probable or not.

$$\chi^2 \sim \chi^2_\alpha \tag{32}$$

And if the Pearson's test statistic is bigger than expected value from the $\chi^2_\alpha$ distribution, where $\alpha$ is the significance level, $H_0$ is rejected: [[26] p. 251-252]

$$\chi^2 \geq \chi^2_\alpha \tag{33}$$

### 4.2.6 Akaike Information Criterion

Akaike Information Criterion, AIC, combines the goodness of the fit information with the complexity of the model to get an conclusion of the quality of the model. Is is evaluated by

$$AIC = -2LogL + 2(k + s) \tag{34}$$

where $k$ is one less than the number of response levels and $s$ counts the number of explanatory effects. In the Goodness of Fit above, no consideration has been taken to the complexity of the model. So, an over fitted model have a penalty if the model is too complex and that is what the AIC is about. Note that AIC does not contain any information of the model significant so no $H_0$ or $H_1$ can be investigated. [29]

### 4.2.7 Coefficient of Determination

Proposes to get a generalized formula for the coefficient of determination has been done:

$$R^2 = 1 - \left(\frac{LL_0}{LL_k}\right)^{\frac{2}{n}} \tag{35}$$

where $LL_0$ and $LL_k$ is defined as before. $n$ is the sample size.

The $R^2$ can be adjusted with a maximum value 1:

$$\tilde{R}^2 = \frac{R^2}{R^2_{max}} \tag{36}$$

where $R^2_{max}$is the maximum less than one for discrete models:

$$R^2_{max} = 2 - (LL_0)^{\frac{2}{n}} \tag{37}$$

with the same explanation as above.[29]

### 4.2.8 Confidence intervals

In SAS®two different types of confidence intervals can be done for the parameters. One is called the *Likelihood Ratio-Based Confidence Intervals* also known as the *profile likelihood confidence interval*, and the *Wald Confidence Interval*. The first named is the most accurate and it uses an iteration scheme but as we have a huge data material and samples, we can get good results with the second named which is more easy and quickly to get. Another name for the Wald is the Normal Confidence Interval as the interval i based on the asymptotic normality of the parameters estimation. For $\beta_j$, the $100(1-\alpha)\%$ Wald Confidence Interval is given by

$$\hat{\beta}_j \pm z_{(1-\alpha/2)}\hat{\sigma}_j \tag{38}$$

where $\hat{\beta}_j$ is the ML estimate of $\beta_j$, $\hat{\sigma}_j$ is the estimate of $\hat{\beta}_j$'s standard deviation and $z_p$ is the $100p$th percentile of the standard normal distribution. [29]

# 5 Results

First, the descriptive results are presented. Figures are not included in the descriptive part as they however do not give the reader more understanding of the data material. As we do not explain what the tables and numbers means, no explanation is given for the tables. Also, not all of the tables are presented as they however do not contribute with information.

## 5.1 Descriptive results

Table 1: Some basic information about the 52 weeks of playing

| | Churn | Active |
|---|---|---|
| $x_1$ | 189351059 | 6684154655 |
| $x_2$ | 1158,43 | 6089,08 |
| $x_3$ | 8,00 | 42,42 |
| $x_4$ | 28674,8 | 31344,58 |
| $x_5$ | 16,29 | 54,54 |
| $x_6$ | 28 | 185,7977 |
| $x_7$ | 1 | 2 |
| $x_8$ | 330 | 2675,50 |
| $x_9$ | 3 | 21 |
| $x_10$ | 3530 | 19983 |
| $x_11$ | 33 | 143 |

Table 2: Some secret information

| | Churn proportion | Active proportion |
|---|---|---|
| 1 | 30% | 3% |
| 2 | 15% | 3% |
| 3 | 10% | 2% |
| 4 | 7% | 2% |
| 5 | 5% | 5% |
| 6 | 4% | 3% |
| 7 | 3% | 2% |
| 8 | 3% | 2% |
| 9 | 3% | 2% |
| 10 | 2% | 4% |
| 11 | 2% | 4% |
| 12 | 2% | 5% |
| 13 | 1% | 2% |
| 14 | 1% | 1% |
| 15 | 1% | 1% |
| 16 | 1% | 1% |
| 17 | 1% | 1% |
| 18 | 1% | 1% |
| 19 | 0% | 1% |
| 20 | 9% | 52% |

Table 3: Some secret information

| | Churn proportion | Active proportion |
|---|---|---|
| 0-400 | 55% | 10% |
| 401-800 | 16% | 7% |
| 801-1200 | 10% | 7% |
| 1201-1400 | 3% | 6% |
| 1401-1600 | 2% | 3% |
| 1601-2000 | 3% | 7% |
| 2001-2400 | 2% | 7% |
| 2401-2800 | 1% | 5% |
| 2801-3200 | 1% | 4% |
| >3200 | 6% | 44% |

Table 4: Some information about a secret proportion

| Proportion played during | Churn proportion | Active proportion |
|---|---|---|
| - | 8% | 75% |
| - - | 18% | 86% |
| - - - | 27% | 91% |

## 5.2 The analysis of the model

The model fit statistics can be seen on table (5) and table (6) which gives us that the model is improved by the input variables. This can be seen due to lower values of AIC and -2LogL. Please note that DF in the tables are Degrees of Freedome.

Table 5: The model fit statistics for the mode based on the full data material

| Criterion | Intercept only | Intercepts and covariates |
|---|---|---|
| $AIC$ | 154254.27 | 101144.64 |
| $-2LogL$ | 154252.27 | 101138.64 |

Table 6: The model fit statistics for the mode based on the restricted data material

| Criterion | Intercept only | Intercepts and covariates |
|---|---|---|
| AIC | 109297.60 | 76108.941 |
| -2LogL | 109297.60 | 76102.941 |

The significance of the two input parameters are statictically proved. This can be seen in table (7) and in table (8) This means that the null hypothesis $(\beta_1, \beta_2 = 0)$ can be discarded.

Table 7: The $\chi^2$-test for the full model based on stepwise selection

| Variable entered | DF | Score $\chi^2$ | Pr>$\chi^2$ |
|---|---|---|---|
| $x_1$ | 1 | 4469.0533 | <0.0001 |
| $x_2$ | 1 | 6141.1880 | <0.0001 |

The Maximum Likelihood estimates of the two coefficients can be seen in table (9) and in table (10)

The coefficient of determination is seen in table (11) and in table (12) These might look small but we decide not to see their values as a problem. If we would use another model, the coefficient of determination would help us to compare two models. In our case, with our models, they are very similar so conclusions can be given. We can do this due to the fact that coeficcient of determination is not a good or even not valid measure of how good a logisitc model is. Also, as we have huge amount of data, the coeficcient of determination will be small as all of the data impossibly can be explained by a model. So, due to these resons, we chose not to care about the low value.

## 5.3 The final prediction model

So, finally we have come to present the prediction model:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = f(x) \tag{39}$$

Table 8: The $\chi^2$-test for the restricted model based on stepwise selection.

| Variable entered | DF | Score $\chi^2$ | Pr>$\chi^2$ |
|---|---|---|---|
| $x_1$ | 1 | 26226.8616 | <0.0001 |
| $x_2$ | 1 | 6208.9274 | <0.0001 |

Table 9: The Maximum Likelihood estimates and their Wald confidence Limits for the model based on full data material

| Parameter | DF | Estimate | Standard error | 95% Wald confidence Limits |
|---|---|---|---|---|
| Intercept | 1 | 0.9733 | 0.0141 | [0.9458, 1.0009] |
| $x_1$ | 1 | -0.4184 | 0.0305 | [-0.4237, -0.4130] |
| $x_2$ | 1 | -2.2108 | 0.00272 | [-2.2705, -2.1512] |

Table 10: The Maximum Likelihood estimates and their Wald confidence Limits for the model based on restricted data material
and their Wald confidence interval.

| Parameter | DF | Estimate | Standard error | 95% Wald confidence Limits |
|---|---|---|---|---|
| Intercept | 1 | 0.9801 | 0.0193 | [0.9424, 1.0179] |
| $x_1$ | 1 | -0.3781 | 0.00331 | [-0.3846, -0.3716] |
| $x_2$ | 1 | -3.781 | 0.0531 | [-3.846, -3.5997] |

Table 11: The coefficient of Determination for the model based on full data material

| $R^2$ | Max-rescaled $R^2$ |
|---|---|
| 0.1304 | 0.2896 |

Table 12: The coefficient of Determination for the model based on restricted data material

| $R^2$ | Max-rescaled $R^2$ |
|---|---|
| 0.1304 | 0.2896 |

Where the $\beta$-parameters for the model based on full data material are estimated as

$$\hat{\beta}_0 = 0,9733 \tag{40}$$

$$\hat{\beta}_1 = -0.4184 \tag{41}$$

$$\hat{\beta}_2 = -2.2108 \tag{42}$$

and for the model based on restricted data material the estimates are

$$\hat{\beta}_0 = 0,9801 \tag{43}$$

$$\hat{\beta}_1 = -0.37038 \tag{44}$$

$$\hat{\beta}_2 = -3.7038 \tag{45}$$

With help from equation (23) we have the following expression

$$p = P(Y = 1 \mid x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}} \tag{46}$$

and with our estimated parameters and our input variables $x_1$ and $x_2$ for each customer, we can get an estimated value of the probability that the customer will churn. When validated the model on our validation data, we had 88.4% concordant in the model based on the full data material and 87.8% concordant in the model based on the restricted data material. Note that 1.6% and 1% of the results from the prediction model was tied.

The robustness of the model can be investigater futher by creating a model of choosing another part of the huge data material.

# 6 Future work

## 6.1 Model selection and model comparison

It would be intresting to use more than one model to predict churn. I was hoping to have time to use Decision Tree as a method to compare my logistic regression model with, but due to lack of time this comparison was never done.

The final model succeeded in prediction some churn. However, it would be intereting to transform some of the parameters and include these in the model. As an example, a log-transform would help us to solve the issue with the range within some of the parameters. Another way to solve this issue could maybe be to exclude the extreme values in each data set.

Scott A. Neslin et al writes about the results from the churn model tournament and which approaches that needs to be taken in deep consideration to create a good predictive churn model [33]. These approaches would be interesting to investigate if and how much they influence the model. In the same article, a good formula for calculating the profitability of a single churn management campaign is presented. In the formula they take the Customer Lifetime Value (see below) in consideration, the different costs of the campaign and also how good the churn probability is. [33] It would be interesting to use the same formula for different churn prediction methods. To create a dynamic model that changes with customers' behavior would also be intereting.

For time-to-event data one can use Cox Regression to build a predictive model. A survival function is produced by the model and it predicts *the probability that the event of interest has occurred at a given time t for given values of the predictor variables.* From the observed subjects, the coeficcients for the predictors and the shape of the survival function are estimated. It would be interesting to use a Cox Regression model for our churn model.

Also, more detailed question formulation in how the churn model is to be used would also be a good continuation of this Master Thesis. And a continuation in how the follow up goes and if it can create a bigger profit.

## 6.2 Customer Lifetime Value

*The part with the Customer Lifetime Value was developed before we realized the time was not enough to futher go on with it*

One well known marketing metric is the *Customer Lifetime Value, (CLV).* The metric is based on a customer's history with the relationship with the company and the metric estimates the value of a customer's whole entire future relationship with the company. Some different abbreviations are found in literature such as CLV, LCV and LTV. By analyzing customers' business and behavior one can pinpoint customers with great potential net value over time as high value customers. A company has many different costs and a profitable customer is a customer with higher revenue to the company rather than the cost to the company.

As written before, the telecom industry has a far-reaching churn and CLV investigation. One reason for this is an extremely competitive market with low switch costs. To obtain the customers the customers needs to be very satisfied. By analyzing the CLV of the customers, one can receive information of high value customers and direct the marketing to these customers. A business can be defined through its products, its prices or its customers and with the latest aspect, CLV is one asset of the organization. Research shows how it is important with quality and service to retain customers. The cost of acquisition of new customers is much higher than retaining existing customers. The rule called 20 : 80 says that 20% of the company's customers account 80% of the company's profit. High value customers are valuable for many reasons:

- Higher purchase frequency
- Buy a higher number of items
- The prices of the items are higher
- Less service costs
- Less price sensitive
- The retention rate is higher

If a company makes a profit of the customer, the yielded revenue system exceeds (by an acceptable) the amount of the company's cost due to the customer. For the customer, the benefits from the relationship with the company need to include benefits. An item or service needs to be more valuable for the customer than the level of the price. Even when a good formula for calculation CLV, some underlying uncertainty exists. One can never be sure the customer, how good it even fits the model, will act as predicted. So, CLV can retrieve important information but it is impossible to be sure the customer act just as planned.

Different formulas can build up CLV. Some are more detailed then others. Let $CF_{i,t+T}$ denote the expected gross cash flow for customer $i$ during the time period $(T-t)$, $T > t$, where $t$ is the starting point and $T$ the end point. The CLV can be expressed as: [18]

$$CLV_{i,t+T} = (T-t) \cdot CF_{i,t+T} \tag{47}$$

When taking consideration to the discount factor and if one prefer to separate different product from each other, denoting each product $j$, $j = 1...q$, the following formula for CLV can instead be used: [11]

$$CLV_{i,t+T} = \sum_{k=1}^{T} \sum_{j=1}^{q} \frac{1}{(1+r)^k} CF_{i,j,t+k} \tag{48}$$

Together with the Average Gross Margin per customer lifespan, and by taking the churn rate in consideration to create the retention rate, $r$, as the 1-the churn rate. The following formula for CLV can be used: [18]

$$CLV = m \left( \frac{r}{1+i-r} \right) \tag{49}$$

Note that this last formula for CLV, does not contain information of a specific customer but if using one predicted churn rate for one type of customers, differentiation of their CLV can be done.

CLV is a good tool to find these high value customers. If a good CLV analysis together with a good churn model, the company can retreive huge insight in how much money that can be spend on relationship and marketing to retain the customers. As it is cheaper to obtain existing customers and as loyal customers are creating a higher profit than non-loyal customers, the business value of this type of customer relation management is of deep impact. [19]

## 6.3 Customer Lifetime Value based on Social Network

In this Master Thesis, we have not used CLV to investigate how worth a customer is for the company. This has not been done due to lack of time. If the time had been enough, some basic calculations would be done for estimating the CLV. With even more detailed data material, it would be interesting to calculate the CLV based on different formulas. As humans are social creatures and we are influenced by actions from others. When a field suits our interest, we pay close attention to this so we can follow the field if interested. A parameter than should be taken in consideration in predicting churn is the social network parameter. If a customer churn, there is an impact on the people in his or her social circle; the propensity to churn is presumably increasing. The provided customer data material at Svenska Spel does not contain information in how a customer is related to another so the social circle and network is extremely hard to reach. If providing the social connectivity between the customers, the problem of knowing the strength of the relationship is still impossible. A customer churns for different reasons. As said before, the service quality is important. Competitive prices, the quality of the products, discounts and promotions are also considered as significant churn parameters.[7] At Svenska Spel, some of these parameters are impossible to use due to stricted promotion rules. An example of rule is the ban of quantity discount and excessive sales campaigns. If and how the social factors affect other customers is not shown significant but as the social network data material is increasing, the conclusion of its significance will be possible to detect. It is easy to understand that if one customer churn from a company, friends and family are affected and might also churn. Also, social pressure to adopt new technology and a curiosity for new products may also affect churn or not. In many articles the telecom industry is investigated through a churn perspective. As they have data in how their customers are connected to each other, they have the possibility to make a social network analysis. [7] At the moment at Svenska Spel, this type of information is not possible to extract from existing data so even if we wanted to do a social network analysis, we have not the possible to do that. In the future, the provided data material might include this type of information.

An example in how one can use Customer Network Value:

Consider two customers, Customer A and Customer B. Let $CLV_A$ and $CLV_B$

be A's and B's CLV. Starting with the normal approach, if

$$CLV_A > CLV_B \tag{50}$$

the company values Customer A more then Customer B as A's CLV is bigger than B's CLV. Considering the aspect of CLV, Customer A would account for a larger loss for the company than B would be.

Define a customers network as a combination of the customers CLV and how much the customer influences other customers. Now, we consider the value of customer A's and customer B's networtk value instead:

$$NetworkValue_A < NetworkValue_B \tag{51}$$

Now the conclusion is vice versa. It is considered a greater loss loosing Customer B than Customer A as the Network of B is bigger then A's. By natural reasons, it is extremely difficult to measure a customers network value. It is however becoming more and more important due to all social network channels and the customers huge impact on each other. [11]

# 7 Discussion

As the developed churn model mostly will form indications of which customers who are about to churn, no targeted campaigns will be done based on the model. If it would be the base for upcoming-targeted campaigns, problem gambling would need to be deeply considered.

It has been great to get a feeling and understanding of the Gambling market and all its parameters. I believe the procedure and the my personal output has been greater than the accuracy of the churn model. If I would re do the master thesis, I would strongly prefer to work in a couple instead of doing it on my own, I would feel more comfortable with the mathematical parts if I would have had a supplement supervisor from LTH who could help me when stuck in different mathematical formulas, methods and formulas. Additional, it would be of great impact to have a structored tutorial about the Master Thesis procedure. What I have been used in this thesis, have been information from different Departments at LTH but not from the Department who examines me. Also, some time in the Mathematical Library at LTH would have facilitate the theoretical part. I find it hard to study mathematics through a small computer screen and I would have understood that more early. Another valuable insight is to never use a small laptop when working with big data. The computer is not enough and this we realized way to late (Two weeks before deadline).

# 8 Acknowledgements

I would like to thank my Supervisor Gunnar Tángring and Olov Forling for having a lot of patience with me and all my questions. Also a big thanks to all nice collegues at Svenska Spel, I have felt very welcome at the office. Last but not least, infinity of love to you who I love the most.

# References

# References


[1] Bill Jolley, Richard Mizerski & Doina Olaru
*How habit and satisfaction affects players retention for online gambling.* Journal of Business Research 59. 770-777 (2006)

[2] Kristof Coussement & Koen W. De Bock
*Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learing.* Journal of Business Research 66. 1629-1636 (2013)

[3] Yong Seog Kim & Sangkil Moon
*Measuring the success of retention management models built on churn probability, retention probability, and expected yearly revenues.* Expert Systems with Applications 39. 11718-11727 (2012)

[4] Ann-Sofie Olsson & Rebecka Johansson
*Spelets pris - en analys av samhällsekonomiska kostnader till följd av spelproblem i Sverige* `https://svenskaspel.se/img/omsvs/Spelets_pris_med_fÃűrord.pdf`

[5] Dick Mizerski
*New research on gambling theory research and practice* Journal of Business Research 66. 1587-1590 (2013)

[6] Helen Sheppard.
*Hints on Layout and Style for writers of Dissertations and Theses at the Faculty of Science, Engineering and Medicine, Lund University* (2012)

[7] Chitra Phadke, Huseyin Uzunalioglu, Veena B. Mendiratta, Dan Kushnir & Derek Doran
*Prediction of Subscriber Churn using Social Network Analysis* Bell Labs Technical Journal 17(4),63-76 (2013)

[8] *Institutionen foor Teknisk Ekonomi och Logistik* `http://www.pm.lth.se/utbildningar/examensarbeten/metodikkurs_for_examensarbete/`

[9] Hossein Abbasimehr, Mostafa Setak & Javad Soroor.
*A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques.* International Journal of Production Research, 51:4, 1279-1294, DOI: 10.1080/00207543.2012.707342, (2013).

[10] John Hadden, Ashutosh Tiwari, Rajkumar Roy & Dymtr Ruta
*Churn Prediction using Complaints Data* World Academy of Science, Engineering and Technology 19 (2006)

[11] Nicolas Glady, Bart Baesens & Christophe Croux
*Modeling Churn Using Customer Lifetime Value* Elsevier Editorial System(tm) for European Journal of Operational Research. EJOR-D-06-01563R2

[12] Pang-Ning Tan, Michael Steinbach & Vipin Kumar
*Introduction to Data Mining* Pearson International Edition IBSN 0-321 42052-7

[13] *Logistic regression* http://www.upa.pdx.edu/IOA/newsom/da2/ho_logistic.pdf

[14] *Logistic regression* ttp://luna.cas.usf.edu/~mbrannic/files/regression/Logistic.html (2014-03-10)

[15] *Logistic regression* http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf (2014-03-10)

[16] *Ordinary Least Squares* http://www.stats.ox.ac.uk/~burke/Linear%20Models/LS%20notes.pdf (2014-03-10)

[17] *Goodness of fit* http://www.strath.ac.uk/aer/materials/5furtherquantitativeresearchdesignandanalysis/unit6/goodnessoffitmeasures/

[18] *Customer Lifetime Value* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1747726 (2014-03-24)

[19] *Customer Lifetime Value* http://blog.kissmetrics.com/how-to-calculate-lifetime-value/?wide=1 (2014-03-24)

[20] *Oracle* http://www.oracle.com/technetwork/developer-tools/sql-developer/what-is-sqldev-093866.html (2014-03-25)

[21] *SAS* http://www.sas.com/en_us/software/enterprise-guide.html (2014-03-25)

[22] *R* http://www.r-project.org/ (2014-03-25)

[23] *Generalized Least Squares and Iteratively Reweighted Least Squares* http://www.stats.uwo.ca/faculty/bellhouse/Generalized%20Least%20Squares_DaeroKim.pdf (2014-04-05)

[24] Scott Menard
*Applied Logistic Regression Analysis* Second edition. Sage Publications, Inc. Online ISBN: 9781412983433 (2014-03-10)

[25] Thomas Padron McCarthy & Tore Risch
*Databasteknik* Third edition. Studentlitteratur ISBN: 9789144044491

[26] Z:W. Birnbaum
*Introduction to probability and mathematical statistics* Harper & Brothers (1962)

[27] Alan Agresti
*An Introduction to Categorical Data Analysis* Second Edition. John Wiley & Sons, Inc. eISBN: 9780470114742 (2014-03-20)

[28] Trevor Hastie, Robert Tibshirani & Jerome Friedman
*The Elements of Statistical Learning* Second Edition. Springer.

[29] SAS®
*The Logistic Procedure* SASOnlineDoc. Version 8.

[30] Data mining
*Figure* http://www.stratebi.com/image/image_gallery?uuid=70db338e-5a62-42bd-ae2f-71c94cc851b5&groupId=10157&t=1366206498870

[31] The Loyality Effect
*The Loyality Effect* http://www.loyaltyrules.com/loyaltyrules/effect_overview.html

[32] Chris Rygielski, Jyun-Cheng Wang & David C. Yen
*Data mining techniques for customer relationship management* Technology in Society 24 (2002) 483-502

[33] Scott A. Neslin, Sunil Gupta, Wagnes Kamakura, Junxiang Lu &Charlotte H. Mason
*Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models* Journal of Marketing Research Vo. XLII (May 2006), 204-211

# A Estimation Methods

The response variable within multiple linear regression is approximately linear related to the explanatory input variables. With $k$ independent variables and $n$ observations we have tho following relationship:

$$y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + \epsilon_i \tag{52}$$

for i=1,...k, n$\leq$ $k$ with $n$ number of observations in the data set. As we have $k$ independent variables we have $k$ parameters to estimate. The $x_i$ terms are not random but they contain a negligible error. As the model is linear in the parameters, transformation can be done of the response or/and the independent variables. To make a more clear view of the equations and parameters, the following notation are used:

$$y = X\beta + \epsilon \tag{53}$$

where

$$y = (y_1, y_2, ..., y_n)^t \tag{54}$$

and

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \dots & x_{k,n} \end{pmatrix}$$

and

$$\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_k)^t \tag{55}$$

and

$$\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^t \tag{56}$$

## A.1 Ordinary Least Squares (OLS)

The *Ordinary Least Squares (OLS)* is one of the most simple methods to estimate the parameters $\boldsymbol{\beta}$. Some assumptions are to be fulfilled if the OLS method is to be used:

- The parameters in the model must have a linear relatioship.
- The residuals in the random sample are statistically independent from each other. Cov$[\epsilon_i, \epsilon_j] = 0 i \neq j$.
- The collinear between the independent variables are not too strong. This means a not too strong linear relationship between two different explanatory variables.
- Measurement error does not exist for the independent variables.
- The residuals has the expected value zero, E$[\epsilon]$=0.

- The variance of the residuals is homogeneous (constant), $\text{Var}[\epsilon] = \sigma^2\mathbf{I}$.
- The residuals are Normally distributed, $\epsilon \in N(\mu, \sigma^2\mathbf{I})$.

With OLS, we choose the coefficients $\beta^T$ which minimize the Residual Sum of Squares, RSS,

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2 = \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \tag{57}$$

As long as the $y_i$'s are conditionally independent given the inputs $x_i$ or if they are compleately independent, the above criterion is reasonable. We want to minimize equation (57) and we start by denoting the $N \times (p+1)$ matrix with $X$. In $X$, an input vector is represented in each row with 1 in the first position. In similay way, $y$ is the N-vector of outputs. The RSS can then be written as

$$RSS(\beta) = (y - X\beta)^T(y - X\beta) \tag{58}$$

In the $p+1$ parameters, the function is quadratic. As we will minimize the expression, we differentiate with respect to $\beta$ and obtain:

$$\frac{\partial RSS}{\partial \beta} = -2X^T(y - X\beta) \tag{59}$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = 2X^T X \tag{60}$$

With the assumtion that $X$ has full rank (which the case not always is), $X^T X$ is positive definite, the derivative can be put to zero:

$$X^T(y - X\beta) = 0 \tag{61}$$

and the solution are given by

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{62}$$

The resulting Least Squares estimation of the target variable $y$ is given by

$$\hat{y} = X(X^T X)^{-1} X^T y \tag{63}$$

If two input variables are perfectly correlated, then $X$ does not have full rank. The result of this is not uniquely defined $\hat{\beta}$ This might also be the result of inputs coded in a overexplicit way and fashion. The best way then, is to drop some of the inpus variables that are correlated with each other.

We still have not mentioned the properties of $\hat{\beta}$. We assume the input $x_i$ are fixed and that the observations $y_i$ are uncorrelated with constant variance $\sigma^2$. With an easy derivation by using the linearity of $X$ and the properities of $/epsilon$, we get the following variance-covariance matrix of the estimated parameters:

$$var(\hat{\beta}) = (X^T X)^{-1}\sigma^2 \tag{64}$$

Where the variance $\sigma^2$ is estimated by

$$\hat{\sigma}^2 = \frac{1}{N - p - 1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{65}$$

Where $N - p - 1$ in the denominatior makes the $\hat{\sigma}^2$ unbiased. Some more assumtions can be given. With equation (52) together with the definition of expected value and variation, one can show that

$$\hat{\beta} \in N(\beta, (X^T X)^{-1}\sigma^2) \tag{66}$$

where $\epsilon \in (N(0, \sigma^2)$ is the error. Another good propertie is

$$(N - p - 1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1} \tag{67}$$

which is a chi-squared distribution with degrees of freedom $N - p - 1$. Also, $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent. The distributional properties constitute the hypothesis tests and confidence intervals for the parameters $\beta_j$. [28]

## A.2 Generalized Least Squares (GLS)

Let $y$, $X$, $\beta$ and $\epsilon$ be as above in equations (54) -(56). In the GLS, the variance does not need to be uncorrelated as in the OLS case. So, consider a correlated variance, Var[$\epsilon$]=$\sigma^2\Sigma$ where we assue unknown $\sigma^2$ and known $\Sigma$. GLS minimize

$$(y - X\beta)^T \Sigma^{-1}(y - X\beta) \tag{68}$$

Which gives us the following $\beta$

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y \tag{69}$$

By using the Choleski Decomposition (no details given about this type of decomposition) we get the triangulat matrix S as in

$$\Sigma = SS^T \tag{70}$$

So, we now have

$$(y_X\beta)^T S^{-T} S^{-1}(y - X\beta) = (S^{-1}y - S^{-1}X\beta)^T (S^{-1}y - S^{-1}X\beta) \tag{71}$$

which can be seen as regressing $S^{-1}X$ on $S^{-1}y$

$$S^{-1}y = S^{-1}X\beta + S^{-1}\epsilon \tag{72}$$

And to get the estimate of y we have the following:

$$y' = X'\beta + \epsilon' \tag{73}$$

We can now examine the variance of the new errors, $\epsilon'$

$$\text{Var}[\epsilon'] = \text{Var}[S^{-1}\epsilon] = S^{-1}\text{Var}[\epsilon]S^{-T} = S^{-1}\sigma^2 SS^T S^{-T} = \sigma^2 I \tag{74}$$

Which gives us new variables $y'$ and $X'$ with uncorrelated errors and equal variance

$$\text{Var}[\beta] = (X^T \Sigma^{-1} X)^{-1}\sigma^2 \tag{75}$$

## A.3 Weighted Least Squares (WLS)

The errors may sometimes be uncorrelated but they might have unequal variance. In this case, WLS can be used. With a diagonal $\Sigma$ the errors are uncorrelated but the variance does not need to be equal. Let the *Sigma* be a diagonal matrix with the weights w,

$$\text{Diag}(1/w_1, ..., 1/w_n) \tag{76}$$

As before, we have regress $S^{-1}X$ on $S^{-1}y$ and so on.

## A.4 Iteratively Reweighted Least Squares - Fisher Scoring

We have finaly come to the optimization method that are used by SAS®. We do not know the variance of $\epsilon$ so we may model $\Sigma$ by the use of a small number of parameters. As an example, $\text{Var}[\epsilon_i] = \gamma_0 + \gamma_1 x_1$. We will not take the Fisher Scoring into more than pseudo details but we want to mention the method as it in use in this Master Thesis through SAS®.

1. An initial guess or starting value is choosen, for example $w_i$=1
2. Least Squares is used to estimate the parameters $\beta$
3. The residuals are used to estimate $\gamma$ by regressing $x$ on $\epsilon'^2$
4. The weights are recomputed and step 2 is done again

The procedure is continuing unitl convergence is reached. [23]

## A.5 Maximum likelihood estimate

For the regression parameters in equation (**??**), the OLS estimator of $\beta$ is identical to the Maximum Likelihood estimate. Values of the parameters are choosen to maximize the likelihood so the parameters are most consistent with the data. The residuals are assumed to follow a zero mean Normal distribution with variance $\sigma^2\mathbf{I}$. Assuming a Normal Distributed response variable, the distribution lookas as follows:

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2\right] \tag{77}$$

The product of n densities from the observations in the data composes the Likelihood function $\mathcal{L}(\beta, \sigma^2)$.

$$\mathcal{L}(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2\right] \tag{78}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2\right] \tag{79}$$

By taking the logs of the likelihood, we get the following log-likelihood function: [26]

$$ln\mathcal{L}(\beta, \sigma^2) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \left[\frac{1}{2}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2\right] \tag{80}$$

Since log is a monotonically[4] increasing function, the log likelihood maximize the likelihood and as it is more easy calculations in the log case, it is preferable using the log transformation. [12] To evaluate the score function, the derivative of the log-likelihood is taken:

$$\frac{\delta \mathcal{L}}{\delta \beta} = -\frac{1}{2\sigma^2}\left(\delta[y'y - 2\beta' X'y + \beta' X'X\beta\right) \tag{81}$$

$$= -\frac{1}{2\sigma^2}\left(-2X'y + 2X'X\beta\right) \tag{82}$$

$$= -\frac{1}{\sigma^2}\left(-X'y + X'X\beta\right) \tag{83}$$

The maximum likelihood estimator is found by putting the score funciton equal to zero:

$$\frac{1}{\sigma^2}(-X'y + X'X\beta) = 0 \tag{84}$$

$$X'X\beta = X'y \tag{85}$$

$$\hat{\beta} = (X'X)^{-1}X'y \tag{86}$$

Now, this far, we can recognize the equation for $\hat{\boldsymbol{\beta}}$ as the same as in OLS. But we are not done with ML yet. In the equation (80) We take the derivative with respect to $\sigma^2$,

$$\frac{\partial ln\mathcal{L}}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}[(y - X\beta)'(y - X\beta)] \tag{87}$$

And by putting it to zero, we get

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}[(y - X\beta)'(y - X\beta)] = 0 \tag{88}$$

$$\frac{1}{2\sigma^4}[(y - X\beta)'(y - X\beta)] = \frac{N}{2\sigma^2} \tag{89}$$

$$\frac{1}{\sigma^2}[(y - X\beta)'(y - X\beta)] = N \tag{90}$$

And as we have an expression for $\hat{\boldsymbol{\beta}}$ we can use this to get an expression for $\hat{\sigma}$.

$$\frac{1}{\sigma^2}[(y - X\beta)'(y - X\beta)] = N \tag{91}$$

$$\frac{1}{\sigma^2}[(y - \hat{y})'(y - \hat{y})] = N \tag{92}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{93}$$

Good to notice: the OLS estimate had the appearance,

$$\hat{\sigma}^2 = \frac{1}{N - p - 1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{94}$$

which shown us that the OLS and ML are different. When small samples, the ML estimator is biased but when $N \to \infty$ the estimators are symptotically equivalent i.e. they converge. [26]

[4] not turning