Future Bank Wallet Size

Negin Nayeri
Elisabeth Hakim

August 25, 2016
Future Bank Wallet Size

Abstract

The main strategic aim of any bank is to maintain a large client-base that eventually could lead to business opportunities of various forms. The banking world today is very competitive where not many clients stay loyal and are more keen to search for different options. It is therefore most crucial to be able to keep and gain clients. However, too many times, mismanagement leads to the customer moving on. In this thesis the future potential of SEB’s current private customer base in the ages of 30-35 is examined. The future potential is defined as the annual revenue to SEB and its competitors in five years. To do so, the top 10% with the highest potential were selected to represent the response variable and the covariates were extracted from five years ago.

The results of this study shows that it is possible to predict the future potential with the coefficient of determination of 41% and a confidence level of 95%. Even though this model is not fit to rely on to compute the future income of the clients to the bank because in terms of risk assessment, it could be used to perform efficient segmentation.
Framtida Bankplånbok

Sammanfattning


Resultaten från denna studie visar att det är möjligt att förutsäga den framtida potentialen med determinationskoefficienten av 41 % och en konfidensnivå på 95%.

Trots att modellen inte är lämplig att förlita sig på för att beräkna kunders framtida intäkt till banken gällande riskbedömning, kan den användas för att utföra effektiv segmentering.
## Contents

1 Introduction ......................................................... 7

2 Theory ................................................................. 8
   2.1 Terminology .................................................... 8
   2.2 Assumptions .................................................... 10
   2.3 Ordinary Least Squares ........................................ 11
   2.4 Potential Problems ............................................ 11
      2.4.1 Homo- and Heteroscedasticity .......................... 11
      2.4.2 Multicollinearity ....................................... 12
   2.5 Model Validation .............................................. 13
      2.5.1 $R^2$ and adjusted $R^2$ ................................. 13
      2.5.2 F-statistics .............................................. 14
      2.5.3 Akaike Information Criterion, AIC ..................... 15
   2.6 Database/SQL ................................................. 15

3 Methodology .......................................................... 17
   3.1 Limitations ..................................................... 17
   3.2 Selection ....................................................... 18
   3.3 Data Pre-processing ........................................... 18
      3.3.1 Collection of Data ..................................... 19
      3.3.2 Software .................................................. 19
      3.3.3 Creation of new variables ............................... 19
      3.3.4 Joining tables ........................................... 21
   3.4 Analysis .......................................................... 21
      3.4.1 Variable Selection ...................................... 22
      3.4.2 Variable description ................................... 23
   3.5 Assumptions ..................................................... 23

4 Results ................................................................. 26
   4.1 Model Validation ............................................... 26
      4.1.1 $R^2$ and adjusted $R^2$ ................................ 26
      4.1.2 F-statistics .............................................. 26
   4.2 The Final Regression Model .................................... 26
   4.3 Interpretation of the Covariates ............................... 28

5 Conclusion ............................................................ 29

6 Future studies ........................................................ 30

Appendices ............................................................. 32

A Excluded Variables .................................................. 32

B Regression output .................................................... 33
List of Tables

1. Table of created variables ........................................... 20
2. Table of covariates included in the model. ....................... 24
3. Table of model fit statistics. ........................................... 26
4. Average mortgage debt amount (SEK) ............................. 28
5. Table of excluded covariates. ......................................... 32
6. Table of variables of the final model. .............................. 33
7. Table of variables of the final model. .............................. 34
List of Figures

1. Illustration of the starting point and the objective . . . . . . . . . . 7
2. An overview of the methodology . . . . . . . . . . . . . . . . . . . 17
3. Illustration of the sectioning in the selection . . . . . . . . . . . . 18
4. Venn diagrams illustrating left, inner and right join . . . . . . 22
5. The process of the analysis . . . . . . . . . . . . . . . . . . . . . . 22
6. Assumption of linearity . . . . . . . . . . . . . . . . . . . . . . . . 23
7. Histogram of the residual . . . . . . . . . . . . . . . . . . . . . . . 25
8. Scatter plot of homoscedastic error . . . . . . . . . . . . . . . . . 25
1 Introduction

Competitiveness has always been the essence of the market and the battle of clients does not exclude the banking world. It is almost a given that a business, of any form, serving a large client base will have disparate groupings of various rated clients. These clients could be rated upon their profit or such as their loyalty to the corporation over the years.

Today customers are more experimental, they are more willing to try alternative brands (Edelman 2016: p 4). This means that they are less loyal because they are more susceptible to switch to brands that better suit their needs.

In SEB’s customer base alone there is a yearly untapped income potential of 5.8 billion SEK. The bank’s customers spend such an extent on the financial products of the competitors.

A young customer at a bank can in terms of revenue be seen as an unprofitable customer. A consequence of this assumption is that the customer is placed in a less prioritized customer segment even though he or she may become a customer of greater interest to the bank. This could then end up as a loss for the bank, given that the customer could seek for other options more prerogative. This prevails the importance of effective segmentation that possibly could end up as a profit on both favors.

In this thesis a mathematical approach is applied to analyse future potential using multiple regression in order to construct a model that predicts which customers in the age of 25-30 years that will generate high revenues to the bank in five years. The variables incorporated in the model are observations and estimates from five years ago.

The aim of this thesis is to predict the total income potential in SEB’s current customer base in the age 25-30 years. By identifying future key customers the bank is able to make their customer segmentation more effective.
2 Theory

One of the most commonly applied econometric tools is linear regression. This section of the thesis is focused on showing how least-squares is used for predicting the future bank potential. Further will the assumptions, potential problems and remedies when performing regression be discussed.

2.1 Terminology

The specification of linear regression model is

\[ y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ik}\beta_k + e_i \quad i = 1, 2, \ldots, n \] (2.1.1)

In the equation 2.1.1, \( y_i \) is regarded as an observation of the dependent random variable \( y \) whose value depends on the set of deterministic variables, \( x_{i1}, \ldots, x_{ik} \).

In this study \( y \) will be referred to as the response variable and \( x \) as the covariates. The parameters \( \beta_j \) are the regression coefficients which are to be determined. \( \beta_0 \) is called the intercept which describes where the linear regression curve intersect the y-axis. The additional random variable, \( e_i \), is the residual.

The residual, also called the error term, is normally distributed (Lang 2015; pp 3-4).

The matrix notation for the regression model is:

\[ y = X\beta + e \] (2.1.2)

where

\[ y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \]

\[ X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \]

\[ \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \]

\[ e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \]
In general, $y$ is an $n \times 1$ vector of the observations, $X$ is an $n \times k$ design-matrix of the regressor variables, $\beta$ is a $k \times 1$ vector of the regression coefficients, and $e$ is an $n \times 1$ vector of random errors.
2.2 Assumptions

The large majority of statistical research and analysis trust on certain assumptions about the variables that are being used in the specific test. Therefore, certain assumptions are to be made prior to the analysis. The assumptions made may differ, depending on the parametric analysis. The outcome may violate the assumptions, thus the conclusion of the research and interpretation of the results will come to change.

OLS, Ordinary Least Squares\(^1\), is the adopted technique in this study, thus the underlying assumptions regarding the method are introduced below.

**Assumption 1** The response variable, \( y \), is a linear combination of all covariates \( X \) and the error terms \( e \) accordingly 2.1.1.

**Assumption 2** The matrix of the covariates, \( X \), must have full rank and the conditions to ensure this assumption is that the number of observations, \( y_i \), is not to be smaller than the number of covariates. Secondly, the covariates are not to be strongly collinear as it is not possible to include a variable that is a linear combination of another since it will create multicollinearity. Multicollinearity is further described in section 2.4.

**Assumption 3** All stochastic covariates, \( X \), must be exogenous, meaning that the covariates, are independent of \( e_i \).

\[
E(e_i | X) = 0 \tag{2.2.1}
\]

This independence is sufficient enough to ensure an unbiased and consistent estimator.

**Assumption 4** The error terms are to be identically distributed with the expected value of zero, i.e. unbiased, and an constant variance.

\[
E(e_i) = 0 \quad (2.2.2) \hspace{1cm} V(e_i) = \sigma^2 \quad (2.2.3)
\]

Note that if the assumption \( E(e_i | X) \) is fulfilled, then this indicates that \( E(e_i) = 0 \) is fulfilled, but this is not the case for the opposite. Violation of this assumption leads to heteroscedasticity which is further explained in section 2.4.

\(^1\)See section 2.3
2.3 Ordinary Least Squares

As mentioned, ordinary least squares is a method for estimating the regression coefficients $\beta$ in such way that the sum of squared residuals are as small as possible. The estimates of the OLS are denoted with a hat, i.e. the OLS of $\beta$ are expressed as $\hat{\beta}$, that are derived by minimizing the sum of the squares $|\hat{e}|^2$, where $\hat{e} = Y - X\hat{\beta}$ (Johnson and Wichern; 1992; pp 287-288).

When the derivative of the sum of squares, 

$$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = (Y - X\hat{\beta})^T(Y - X\hat{\beta})$$

(2.3.1)

with respect to $\hat{\beta}$ is equal to zero the following is obtained 

$$\frac{\partial (Y - X\hat{\beta})^T(Y - X\hat{\beta})}{\partial \hat{\beta}} = 0$$

(2.3.2)

$$\Rightarrow -2X^TY + 2X^TX\hat{\beta} = 0.$$  

(2.3.3)

It follows that the OLS estimate of $\hat{\beta}$ is 

$$\hat{\beta} = (X^TX)^{-1}Y$$

(2.3.4)

which satisfies the normal equation 

$$X^T\hat{e} = 0.$$  

(2.3.5)

2.4 Potential Problems

When performing a regression analysis with OLS, several covariates can be used in the regression equation. However, this can lead to covariates being correleated to eachother which is a violation of Assumption 2 and could potentially result in an unreliable model. Another issue is the problem of inconsistent residual variance, which will violate the OLS assumption that the error terms are constant which is why in this study we focus on problems concerning heteroscedasticity and multicollinearity.

2.4.1 Homo- and Heteroscedasticity

The error term $e_i$ is homoscedastic if $Var(e_i) = \sigma^2$ for $i = 1, ..., n$, i.e. $E(e_i) = 0$ and $E(e_i^2) = \sigma^2$. This means that the error terms all have the same constant variance for all $i$ and does not depend on $x$ (Stock and Watson; 2011; p. 155-157).

The estimated covariance matrix for $\hat{\beta}$ is computed by the following formula under the assumption of homoscedasticity, 

$$Cov(\hat{\beta}) = (X^TX)^{-1}s^2,$$  

(2.4.1)

where 

$$s^2 = \frac{|\hat{e}|^2}{n - k - 1}.$$  

(2.4.2)
The opposite of homoscedasticity is heteroscedasticity which means that the error terms have different variance, \( \text{Var}(e_i | X_i = x) = \sigma_i^2 \) for \( i = 1, \ldots, n \). The OLS estimator in this case is still unbiased, \( E(e_i) = 0 \), consistent and asymptotically normal (Stock and Watson; 2011; p. 158).

The problem occurs when the model has heteroscedastic residuals and is misspecified as homoscedastic. As a result the \( F - test \), which is further described in section 2.5.3, will be invalid (Lang; 2015; p. 17).

**Detecting Heteroscedasticity**

The Eyeball Test, which is implied by the name, visually evaluates whether the residual is heteroscedastic or not by observing a scatter plot of the residual. If the residual do not plot well against a line and deviates differently heteroscedasticity is present (Kennedy 2008; p. 116).

**Remedies for Heteroscedasticity**

One solution to heteroskedasticity is to incorporate White’s robust estimate of the covariance matrix

\[
\text{Cov}(\hat{\beta}) = (X^T X)^{-1} X^T D(\hat{e}_i^2) X (X^T X)^{-1}
\]

where \( D(\hat{e}_i^2) \) is the standard deviation of squared residual. If heteroskedasticity is present White’s estimator will improve the regression and therefore it is advisable to always use it (Lang; 2015; p18).

### 2.4.2 Multicollinearity

Perfect multicollinearity is a phenomenon that occurs when at least one of the covariates are a perfect linear combination of the other covariates. This model specification error renders that \( X \) will not be of full rank and therefore makes the OLS estimation impossible (Lang; 2015; p. 15).

When two or more of the covariates are highly correlated, but not a perfect linear combination, imperfect multicollinearity is the problem. This do not pose any problems to the OLS estimation but the point estimates of those coefficients will be imprecise (Stock and Watson; 2011; p.199-202).

Thus imperfect multicollinearity causes the standard errors to be very large it is just a nuisance and not a specification error (Lang; 2015; p. 15).

**Detecting Multicollinearity**

One way to detect multicollinearity is to run a regression for each covariate and compute VIF. The variance inflation factor is computed according to the
follwong formula,
\[ VIF = \frac{1}{1 - R^2} \quad (2.4.4) \]
where \( R^2 \) is obtained from a regression for every individual covariate.

A thumb rule to detect a harmful mulicollinearity in the data is that \( VIF > 10 \) (Lang; 2015; p. 54-55). In this thesis, multicollinearity will be detected by using the pearson product moment correlation coefficient which is a measure of the linear correlation between two variables, i.e determines to which quantity two variables are 'proportional' to each other. The rate of the correlation coefficient is independent of the units. (Hill; Lewicki; 2006; p.18).

Similar to the detection of heteroskedasticity, mulicollinearity can be detected through a scatter plot. By putting the measurments of each covariate in separate, ordered vectors and plotting them against each other in a scatter plot, one can see if the scatters are gathered around a line there is presence of multicollinearity.

**Remedies for Multicollinearity**

When perfect mulicollinearity appears and the OLS estimation is impossible the model should be reformulated, one suggestion is to remove the covariate that is the linear combination of one or more others.

The remedy for imperfect mulicollinearity is more complex and there is no definite solution that applies to all situations. The problem is that the standards errors of the highly correlated covariate is very large. These standard errors are decreasing with \( n \), so one solution is to increase the number of observations (Lang; 2015; p. 54-55).

### 2.5 Model Validation

This section provides the process of deciding whether the results obtained from the regression analysis are acceptable. The validation process can be performed using several methods such as analyzing goodness of fit and the residuals. The following methods used in this thesis are described below.

#### 2.5.1 \( R^2 \) and adjusted \( R^2 \)

\( R^2 \) describes the discrepancy between observed values and the values expected under the model in question. It is often referred to as the measure of goodness of fit. It is defined as the fraction of the sample variance of \( y_i \) explained by the covariates or 1 minus the fraction of the variance of \( y_i \) not explained by the regressors, according to the following equation

\[ R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \quad (2.5.1) \]
where the explained sum of squares is \( ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \), the total sum of squares \( TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 \) and the sum of squares residual \( SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}e_i^2 \) (Stock and Watson; 2011; p. 193-194). It follows from equation 2.5.1 that \( R^2 \in [0,1] \). \( R^2 \) should be as large as possible since this minimizes the error term \( \hat{e} \) and therefore implies an improved estimation of the response variable.

\( R^2 \) does typically increase when adding a new covariate to the regression equation. This do not mean that adding the covariate yields an improved fit. Thus in this sense \( R^2 \) gives an inflated estimate of the fit one can reduce or adjust the value by using the reduced \( R^2 \) denoted \( \bar{R}^2 \). \( \bar{R}^2 \) is defined as

\[
\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \frac{SSR}{TSS} \quad (2.5.2)
\]

To obtained the best fit an econometrician should search for the largest \( \bar{R}^2 \) instead of the \( R^2 \) (Kennedy; 2008; p.80).

### 2.5.2 F-statistics

It is common to report a p-value with each estimated coefficient when a regression analysis is performed. The value of the F-statistic is used to obtain the p-value. If the residual is normally distributed the F-statistic is calculated according to equation 2.5.3 (DeFusco; 2007).

\[
F = \frac{ESS}{k} \cdot \frac{k}{SSR} \cdot \frac{n-k-1}{n-k-1} \quad (2.5.3)
\]

The F-value is used to calculate the p-value. A large F-statistic yields a small p-value and a small F-statistic leads to a small.

**The F-test**

An alternative method to compute the F-statistic, with the hypothesis that the coefficients of \( r \) covariates are equal to zero, is by the following formula (Lang; 2015; p.36-37)

\[
F = \frac{R^2}{1 - R^2} \cdot \frac{n-k-1}{r} \quad (2.5.4)
\]

where \( r \) is the number of coefficients tested for zero. This is done under the normality assumption. For a multiple regression model with an intercept, we want to test the following joint null hypothesis, \( H_0 \), against the alternative hypothesis \( H_1 \):

\[
H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0 \quad (2.5.5)
\]

\[
H_1 : \beta_j \neq 0 \text{ for at least one value of } j = 1, 2, \ldots, r. \quad (2.5.6)
\]

When testing a hypothesis a significance value, \( \alpha \) is determined. The last step is to compare the obtained p-value to the significance value. If \( p \leq \alpha \) the null
hypothesis is rejected and the alternative is accepted. If not, \( H_0 \) is accepted with significance level of \( \alpha \).

The F-value can also be used to determine if the model should be reduced or not. It is usually advisable to reduce the model if the standard error of the regression decreases with the number of covariates. One way to conclude if the model should be reduced or not is to examine the F-statistic for the reduced model. The fact that the standard error goes down is equivalent to a F-statistic less than one (Lang, 2015: p. 40).

2.5.3 Akaike Information Criterion, AIC

Another method to determine if one of several covariates should enter the regression equation or not is to use the Akaike information criterion. Using this test one chooses the model that minimizes

\[
AIC = n \ln(|\hat{e}^2|) + 2k. 
\]

(2.5.7)

By minimizing AIC, the estimated information loss in minimized (Lang, 2015; p.21-22).

When choosing between a full and reduced model Akaike prefers the smaller model if and only if

\[
\eta^2 < 1 - e^{-\frac{2k}{n}}
\]

(2.5.8)

where \( \eta^2 \) is called the effect size. \( \eta^2 \) describes how ”efficient” the covariate is and is defined as

\[
\eta^2 = \frac{|\hat{e}_*|^2 - |\hat{e}|^2}{|\hat{e}_*|^2}
\]

(2.5.9)

where \( \hat{e} \) and \( \hat{e}_* \) are the residuals for the full and reduced model (Lang, 2015; p. 9).

2.6 Database/SQL

Database refers to the collection of schemas, tables, queries and views etc. or formally to a set of related data and the way that it is organized. Generally, the data is systematized to model characteristics of reality in a way that supports processes requiring information. This could be something like modeling the accessibility of seats on an airplane in a way that provisions finding available ‘airplanes’ with vacancies. The access to these data is commonly provided by a “database management system” (DBMS) that is made of an integrated set of computer software that permits users to cooperate with one or various databases and arranges for access to all of the data contained in the database. DBMS provides numerous functions that allow entry, storage and retrieval of vast quantities of information and also provides ways of managing the organization of the information. Structured Query Language (SQL) is a special-purpose programming language that is designed for handling of data that is held in a relational database management system (RDBMS). SQL
consists of data definition language, data manipulation language and a data control language.
3 Methodology

In this thesis the focus is to estimate the future potential of customers between 25-30 years, limited to five years ahead. The selection that the model is based on is the top 10\% of the customers 30-35 years regarding their total income potential, their yearly revenue to SEB and their yearly estimated revenue to the competitors. The covariates from the model are selected from five years back in time that in any way indicate the potential of the selection today. After narrowing the amount of covariates a model is trained on all customers between 30-35 years today with covariates extracted five years ago. The model is then applied on customers 25-30 years to approximate their likely potential in five years ahead.

3.1 Limitations

The majority of all research will suffer with limitations and this study is no different. The model building of this project depends on the current customers of SEB. The relation between the size of the bank wallet, i.e. the customers total income potential, and the variables that makes the advances that generates the high revenue to the bank is sought for. It is understandable that if given a large amount of data, it would be possible to perform a statistical research with that would explain different relationships, if there would be any. However access of data is not a guarantee of a successful outcome, it is most crucial but not exclusive. The amount and consistency of the data should also be of a great importance. As mentioned, the covariates in this study are five years old. However, the representative would be to go back further more, but this would lead to inconsistency in the data, thus leading to that the model incorporated would not be applicable on today’s customers.

A major disadvantage for this study is that individual facts or terms (such as certain behaviours) are not recognized in this type of analysis. Another limitation in statistical analysis is the inability of statistical methods to study

Figure 2: An overview of the methodology
the nature of phenomenon, that would include health, intelligence etc, due to the fact that it’s not possible to express these in quantitative terms.

3.2 Selection

Since there is large amount of variables available; the scale of these need to be decreased to obtain relevant results. This is done by analyzing historical data of the top potential costumers today. More specifically, the significant variables that explain the position of these customers today are identified. As the total income potential is computed for the following costumers of age 30-35 years, the top 10 % are selected. To improve the selection, a number of costumers are then excluded based on the four following criteria.

1. Since the historical data must be sufficiently extensive the costumer must have a client age of five years or more.

2. Private banking services are traditionally offered to the bank’s high net worth individuals. Private banking costumers will be outliers in the data and are therefore excluded in the selection.

3. If the costumers have a small change in potential over these five years they are not of interest. Another criteria for the selection is that the costumer must have an potential increase of at least 100 %.

4. The lack of data is the reason why all fictitious costumers are excluded.

![Figure 3: Illustration of the sectioning in the selection.](image)

3.3 Data Pre-processing

The data is processed before the analysis according to the following steps:

1. Gathering relevant data.

2. Creation of new variables.

3.3.1 Collection of Data

SEB has access to an enormous amount of classified data from its customer base. The data can cover age, salary, city of residence, number of loans, savings and more of the typically known information when one is a customer of the bank. But the research department can also perform analyses to find the chance of other perspectives that are not surely known but can be calculated to fairly good odds from the other variables known. SEB also has access to Mosaic data that they have bought from a marketing company called Insight One. Insight One’s access to a variety of new and detailed information on all individuals and households in Sweden provides a comprehensive image of consumers and helps to distribute more information to enhance the customer relationships. Insight one takes every household in Sweden and divides it into 44 Mosaic-types with different descriptions. Insight one has created these types through gathering of all statistical information from SCB, Bilregistret and Svensk adressändring but also survey data from TNS-Sifo’s great consumer survey, ORVESTO® (Insightone; 2013; pp 760).

3.3.2 Software

The analysis is done in SAS (Statistical Analysis System), which is a software that is developed by SAS institute for advanced analytics. The variables that are used in this study is interval-, binary- and nominal-variables, which all can be processed in the software. SAS Enterprise Guide is used for the data preprocessing while SAS Enterprise Miner is used for the analysis.

3.3.3 Creation of new variables

Since the selection is young costumers, 25-30 years, they might not be interpreted as attractive costumers to the bank. Therefore the focus have been on creating variables that describes the costumers cash flow patterns. A persons cash flow patterns can describe their economic intelligence, which may be of interest in this analysis since the selection may not been having large capitals or loans in the bank five years ago.

Firstly, total cash flow variables were created. All transaction for a duration of 6 months were grouped by IP ID, the personal identifier, and the direction of the transaction. The direction can either be I, in, or out, O. The grouped transaction were both summed and counted to create variables. These variables both describe the volume of the costumers total cash flow and the frequency of them.

Another variable of interest is one that describe the costumers lavish. To be able to capture this from the transaction data a similar procedure as the total cash flow variables. All transactions were grouped by IP ID but with the following conditions:
1. Since the customers lavish are of interest the first condition is that the transaction must have direction O.

2. The transaction must be equal or less than 100 SEK.

3. The transaction must be a card purchase.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>transO_n</td>
<td>Number of transactions during 6 months with direction O.</td>
</tr>
<tr>
<td>transO_sum</td>
<td>Sum of transactions during 6 months with direction O.</td>
</tr>
<tr>
<td>transI_n</td>
<td>Number of transactions during 6 months with direction I.</td>
</tr>
<tr>
<td>transI_sum</td>
<td>Sum of transactions during 6 months with direction I.</td>
</tr>
<tr>
<td>diff_IO</td>
<td>The difference between the sum of transactions with direction I and O during 6 months. transI_sum-transO_sum</td>
</tr>
<tr>
<td>scard_n</td>
<td>Number of card transactions equal or smaller than 100 SEK with direction O.</td>
</tr>
<tr>
<td>scard_sum</td>
<td>Sum of card transactions equal or smaller than 100 SEK with direction O.</td>
</tr>
<tr>
<td>partner_potential</td>
<td>The revenue to SEB and the estimated revenue to SEB’s competitors of the customers partner.</td>
</tr>
</tbody>
</table>

Table 1: Table of created variables

The transaction data do not specify which kind of purchase or income it is, but every transaction do have a name. To be able to categorize the transactions the 100 most frequent transactions in the selection top potential respectively the selection young today were used form a program. Thereafter the program was runned for the selections transactions during 6 months. The transactions are categorized by matching the text strings to the 100 most frequent ones. A variable was created for each category and defined as the sum spent on the category divided by the wage. The categories are:

- Credit
- Credit card
- Transactions
- Accommodation
• Travels
• Investments and savings
• Children
• Shopping
• Groceries
• Clothes
• Phone and Internet
• Installment of student-loan.

One large source of income to a bank is mortgages. It is common that one buy accommodation together with ones partner. Therefore data describing information about the costumers partners can possibly be of interest. By joining on the partners IP ID variables with the partners salary and total potential were created.

3.3.4 Joining tables

Since the data is not always compatible joining tables is essential. The data provided from SEB is relational. This means that the tables have links to other pieces of data that show the relationship between them. The data used in this thesis have two keys. The objective is to compute the costumers future potential. Since this is bounded to a person, the costumers IP ID is used as a identifier to join tables.

As mentioned above, the tables are not always compatible. Some tables do not have information for all the costumers in the selection. This means that the tables are of different size, even though they have been matched on the IP IDs from the same selection. This is handled by using SQL joins.

There are three different kind of joins used in this thesis, right, inner and left join. When joining two tables inner join only keeps the observation that exist in both tables. When using left join all the variables from the left table are retained and those observations that do not exist in the right table become missing values in the joined table. The right join leaves the observations that exist in the right table but not the left as missing values instead.

3.4 Analysis

The data mart created in the the data pre-processing is divided into two parts, training and validate. The training data set consist of 70 % of the total data and is used to train the model how to estimate the future potential. The fitted model is then used to predict the response variable for the observations in the validation set, which consist of 30 % of the total data.
After the data partition the data is imputed. Firstly, if any observation have more than 50% missing values it is removed. The remaining missing values in the data mart are treated in two different ways. The missing values of the numerical variables are replaced by the mean. If the variable is nominal then the missing value is replaced by the modal value.

After the imputation outliers are treated by transforming the variables. The variables are inspected and those who have a positively skewed distribution are logarithmic. The typical use of a logarithmic transformation variable is to pull the outlying data from the positively skewed distributed variable closer to the bulk of the data.

### 3.4.1 Variable Selection

As mentioned, the response variable will represent the total potential of the selection. The choosing of covariates can be performed in several ways. Forward selection is the way of beginning with no variables in the model and evaluating with the addition of each variable. Backwards elimination is the case of all variables being included at first and eliminated if not contributing ([Hill; Lewicki; 2006; p.158](#)). In this study, Stepwise regression, a method that uses a combination of forward selection and backwards elimination will be applied. The criteria used for the evaluation of each variable is Akaike information criterion (AIC) computed with the validation data set.

---

**Figure 4:** Venn diagrams illustrating left, inner and right join.

**Figure 5:** The process of the analysis.
3.4.2 Variable description

The covariates listed in table 2, minimized AIC and were used in the final model. The covariates labeled observed are obtained as from the client data and not approximated by SEB, whereas the estimated data is received from models developed by SEB analysts and InsightOne².

3.5 Assumptions

As mention, certain assumptions are made prior to a statistical analysis, depending on the method of use. Before using the method of linear regression, the assumption requirement listed in section 2.2 are to be validated. To examine if the variables in the model met the requirement of linearity, scatter plots of the covariates and the response variable were made and studied. The scatter plot below represents one of the tests, where linearity is clearly present as the data follows the straight line.

Figure 6: Assumption of linearity

²See section 3.3.1
<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Description</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>gndr_code</code></td>
<td>Nominal</td>
<td>Gender code, F M</td>
<td>Observed</td>
</tr>
<tr>
<td><code>geoclass</code></td>
<td>Nominal</td>
<td>Geographical description of the area of residence</td>
<td>Observed</td>
</tr>
<tr>
<td><code>finance_incm</code></td>
<td>Numerical</td>
<td>Annual revenue from financial products excluding mortgages</td>
<td>Observed</td>
</tr>
<tr>
<td><code>finance_incm_m</code></td>
<td>Numerical</td>
<td>Annual revenue from mortgages</td>
<td>Observed</td>
</tr>
<tr>
<td><code>insur_incm</code></td>
<td>Numerical</td>
<td>Annual revenue from insurance</td>
<td>Observed</td>
</tr>
<tr>
<td><code>invest_incm</code></td>
<td>Numerical</td>
<td>Annual revenue from invest</td>
<td>Observed</td>
</tr>
<tr>
<td><code>trade_incm</code></td>
<td>Numerical</td>
<td>Annual revenue from</td>
<td>Observed</td>
</tr>
<tr>
<td><code>fxtrn_bal_base</code></td>
<td>Numerical</td>
<td>Fixed term funds volume</td>
<td>Observed</td>
</tr>
<tr>
<td><code>dmdep_bal_base</code></td>
<td>Numerical</td>
<td>Annual revenue from deposit volume</td>
<td>Observed</td>
</tr>
<tr>
<td><code>fund_n</code></td>
<td>Numerical</td>
<td>Number of fund accounts</td>
<td>Observed</td>
</tr>
<tr>
<td><code>savings_n</code></td>
<td>Numerical</td>
<td>Number of saving accounts</td>
<td>Observed</td>
</tr>
<tr>
<td><code>client_age</code></td>
<td>Numerical</td>
<td>Number of years a client been customer</td>
<td>Observed</td>
</tr>
<tr>
<td><code>life_phase</code></td>
<td>Nominal</td>
<td>The life stage of customer</td>
<td>Estimated</td>
</tr>
<tr>
<td><code>cstd_bal_base</code></td>
<td>Numerical</td>
<td>Depot volume.</td>
<td>Observed</td>
</tr>
<tr>
<td><code>pixel_mean_dept</code></td>
<td>Numerical</td>
<td>Mean of the households dept</td>
<td>Estimated</td>
</tr>
<tr>
<td><code>pixel_mean_assets</code></td>
<td>Numerical</td>
<td>Mean of the household assets</td>
<td>Estimated</td>
</tr>
<tr>
<td><code>propensityBorn_abroad</code></td>
<td>Numerical</td>
<td>The probability of the customer being foreign-born, household level</td>
<td>Estimated</td>
</tr>
<tr>
<td><code>propensityBorn_001</code></td>
<td>Numerical</td>
<td>The probability of one of the customers parents being foreign-born, household level</td>
<td>Estimated</td>
</tr>
<tr>
<td><code>advise_n</code></td>
<td>Numerical</td>
<td>Number of times customer visited an advisor of the bank</td>
<td>Observed</td>
</tr>
<tr>
<td><code>strategic_match</code></td>
<td>Binary</td>
<td>Whether the customer is has any board involment in a company</td>
<td>Observed</td>
</tr>
<tr>
<td><code>scard_sum</code></td>
<td>Numerical</td>
<td>Amount of card transactions less then 100 SEK</td>
<td>Estimated</td>
</tr>
<tr>
<td><code>liv_n</code></td>
<td>Numerical</td>
<td>Number of insurances including IPS</td>
<td>Observed</td>
</tr>
<tr>
<td><code>tb_n</code></td>
<td>Numerical</td>
<td>Number of interactions through the telephone-bank</td>
<td>Observed</td>
</tr>
<tr>
<td><code>office_n</code></td>
<td>Numerical</td>
<td>Number of interactions through office visit</td>
<td>Observed</td>
</tr>
<tr>
<td><code>mosaic_grp</code></td>
<td>Nominal</td>
<td>Socioeconomic description</td>
<td>Estimated</td>
</tr>
</tbody>
</table>

Table 2: Table of covariates included in the model.
The case of multicollinearity is studied with the variance inflation factor (VIF). To do so, the response variable is excluded and each covariate is chosen as the target in one’s own regression equation. For example the covariate \( Pixel\_mean\_dept \); the calculated value of \( R^2 \) from the regression is .

\[
VIF = \frac{1}{1 - R^2} = \frac{1}{1 - 0.2922} = 1.412828
\] (3.5.1)

This indicates that there exists no harmful multicollinearity.

In figure 8 below, a histogram of the residual of the response variable is plotted. By visually examine figure 7 one can observe that the residual is normally distributed around zero. Thus, the assumptions 3 and 4 in section 2.2 are fulfilled.

![Figure 7: Histogram of the residual](image)

The case of homogenous variance is analyzed through scatter plots of the residual with the different covariates. The scatter plot below is an example of a covariate; the data follows a straight line even though there exists some outliers. Nonetheless, it is accepted the assumption of homogeneous variance is attained.

![Figure 8: Scatter plot of homoscedastic error](image)
4 Results

The regression is performed in SAS Enterprise Miner and the results are presented in appendix B. Variables which are transformed by optimal binning will be denoted with opt. before the covariate’s name. The statistics for the model are presented in table 4.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>81695.0000</td>
</tr>
<tr>
<td>Error degrees of freedom</td>
<td>81661.0000</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td></td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.3395</td>
</tr>
<tr>
<td>Adjusted (R^2)</td>
<td>0.3393</td>
</tr>
<tr>
<td>(p)-value</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>AIC</td>
<td>-174304</td>
</tr>
<tr>
<td>F-value</td>
<td>1149</td>
</tr>
</tbody>
</table>

Table 3: Table of model fit statistics.

4.1 Model Validation

4.1.1 \(R^2\) and adjusted \(R^2\)

By interpreting the values in table 4 one can evaluate the final model. The goodness of fit, \(R^2\), is obtained to 0.3395 which means that the regression model explains 34\% of the variance in the data.

4.1.2 F-statistics

At a significance level of 5\% one can say that each covariate in the regression equation is significant since the t-statistics values are all grater than 2 or less than -2. The \(p\)-value for each covariate is less than 0.05 which also supports the hypothesis that each covariate is significant.

The null hypothesis \(H_0: \beta_1 = \beta_2 = \cdots = \beta_{34} = 0\) is tested with a F-test. The F-statistic for the regression equation is 1149 on 34 and 81661 degrees of freedom. The corresponding \(p\)-value is less than 0.0001. We can reject the null hypothesis on a significance level of 5\%.

4.2 The Final Regression Model

The final multiple regression model will have the following form:

\[
\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{34} \hat{\beta}_j x_{ij} + \hat{\epsilon}_i \quad (4.2.1)
\]
where the predicted covariates and the coefficient estimates $x_{ij}$ are listed below in Table 5 and 6.
4.3 Interpretation of the Covariates

From the regression outcome that is listed in table 5, it is observed that the estimates of all covariates are rather small. This means that the response variable depends on a combination of variables rather than solely on a few. The covariate \textit{finance\_incm} has the highest estimate, indicating that the clients of SEB that generate annual revenues from financial products to the bank, are more likely to end up in the top potential block in five years. Moreover, the covariate \textit{invest\_incm} has the second highest estimate which reveals that revenue generated from investments is of higher value to the bank. Nonetheless, all covariates that present the higher estimates, goes hand in hand with annual revenues to the bank from various financial products and understandably so. One surprising factor was that the covariate that represents annual revenues from mortgages do not have a significant influence on ending up in the top potential, unless the volumes are relatively large. For example, the numbers listed in table 4 reveal that Stockholm has an average mortgage debt that is around 1.5 $MSEK$ higher than the remaining cities and therefore have an greater impact on the future potential.

<table>
<thead>
<tr>
<th></th>
<th>Average debt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stockholm</td>
<td>2 912 510</td>
</tr>
<tr>
<td>Göteborg</td>
<td>2 397 774</td>
</tr>
<tr>
<td>Malmö</td>
<td>1 952 063</td>
</tr>
<tr>
<td>Remaining big cities</td>
<td>1 799 959</td>
</tr>
<tr>
<td>Remaining cities</td>
<td>1 359 485</td>
</tr>
</tbody>
</table>

Table 4: Average mortgage debt amount (SEK)
5 Conclusion

The main purpose of this study was to develop a predictive model for the clients of SEB between the ages of 25-30 so that they could be effectively segmented in favor of both parties. The methodology of linear regression used to predict the relationship between the top potential clients and the chosen covariates showed an correlation inferior of 41 %. This can be interpreted in several ways. However, most will agree on the fact that it is on the weaker level.

One of the reasons of this result could be the year span the study was made on since five years is a long time to predict, even in the banking world. But the largest contributing factor is the absence of covariates that tells more about the clients background and choices.

Even though an existence of a student loan reveals an academic background, the type of education remains unknown, which is of great loss. As mentioned in limitations, it is not possible to declare the human behaviour in quantitative terms. This could mean the loss of many understandings behind a clients rise to the top potential. But the likings of the clients, such as their interest in being budget-conscious, can be explained quantitatively and would be of great interest, had it existed.

The findings of this study show however that it is most possible to achieve an efficient outcome from this model. Since the sole purpose was to build a model that can be used to perform effective segmentation; the windup result reveal that some success was accomplished.

If a student of higher education, aged 25, would become a member of the bank today, he or she would usually not be of much interest. This student would probably not be in possession of many assets, have low to no income and have a large dept. Based on this model, this student could be of large interest to the bank and should be smartly segmented. Thus, leading to a profit for the bank and decrease the possibility to loose the future potential to a competitor.
6 Future studies

The aim of this study is to find indicators that explain a customer's future potential to the bank. One client who has large assets and high income is most likely already in the right customer segment. In future studies one can extend the amount of data to contain variables like \textit{yearsofeducation,typeofeducation} etc. Access to covariates who are more descriptive of the customer's past and achievements could lead to more unexpected indicators and better accuracy.

The model can be extended for all customers, not limited to customers of age 25-30. The extended model could be a framework of effective segmentation for all customers. With forceful customer relationship management strategies the bank would capture untapped potential from all age groups by a wide margin of time.
References


# Appendices

## A Excluded Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>propensityBornAbroad</code></td>
<td>Numerical</td>
<td>Foreignborn probability estimated on household level</td>
</tr>
<tr>
<td><code>propensityForeignParents</code></td>
<td>Numerical</td>
<td>Native with nonnative parents probability estimated on household level</td>
</tr>
<tr>
<td><code>pixelMeanStud</code></td>
<td>Numerical</td>
<td>Mean of the households student loans</td>
</tr>
<tr>
<td><code>paymentIncM</code></td>
<td>Numerical</td>
<td>Annual revenue from payments</td>
</tr>
<tr>
<td><code>mncpCode</code></td>
<td>Numerical</td>
<td>County code</td>
</tr>
<tr>
<td><code>ipsBalBase</code></td>
<td>Numerical</td>
<td>Volume of individual pension</td>
</tr>
<tr>
<td><code>coLoanBalBase</code></td>
<td>Numerical</td>
<td>Co-loaner Volume</td>
</tr>
<tr>
<td><code>xipsBalBase</code></td>
<td>Numerical</td>
<td>Fund volume exclusive Ips</td>
</tr>
<tr>
<td><code>livProxy</code></td>
<td>Numerical</td>
<td>Brokered customer</td>
</tr>
<tr>
<td><code>privateN</code></td>
<td>Numerical</td>
<td>Number of private accounts</td>
</tr>
<tr>
<td><code>corporateN</code></td>
<td>Numerical</td>
<td>Number of company accounts</td>
</tr>
<tr>
<td><code>accountN</code></td>
<td>Numerical</td>
<td>Number of accounts</td>
</tr>
<tr>
<td><code>simpleLoanNumber</code></td>
<td>Numerical</td>
<td>Number of Simple loans (SEB service, &quot;Enkla lanet&quot;)</td>
</tr>
<tr>
<td><code>grantedCredit</code></td>
<td>Numerical</td>
<td>Volume of granted credit</td>
</tr>
<tr>
<td><code>cardN</code></td>
<td>Numerical</td>
<td>Number of card transactions</td>
</tr>
</tbody>
</table>

Table 5: Table of excluded covariates.
### B Regression output

<table>
<thead>
<tr>
<th>Variable $x_j$</th>
<th>$\beta_j$</th>
<th>Estimate</th>
<th>SE</th>
<th>t stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$\beta_0$</td>
<td>-3.1438</td>
<td>0.4708</td>
<td>-6.68</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>finance_incm_m</td>
<td>$\beta_1$</td>
<td>0.000052</td>
<td>6.341-10^-7</td>
<td>81.99</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>geoclass, Innercity</td>
<td>$\beta_2$</td>
<td>0.0113</td>
<td>0.00293</td>
<td>3.86</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>geoclass, Small town</td>
<td>$\beta_3$</td>
<td>-0.0299</td>
<td>0.00284</td>
<td>-10.52</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>geoclass, City</td>
<td>$\beta_4$</td>
<td>-0.0113</td>
<td>0.00303</td>
<td>-3.72</td>
<td>0.0002</td>
</tr>
<tr>
<td>savings_n</td>
<td>$\beta_5$</td>
<td>0.00807</td>
<td>0.00192</td>
<td>4.20</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>office_n</td>
<td>$\beta_6$</td>
<td>0.0110</td>
<td>0.000905</td>
<td>12.13</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>client_age</td>
<td>$\beta_7$</td>
<td>0.00108</td>
<td>0.000240</td>
<td>4.49</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>life_phase, Childless singles</td>
<td>$\beta_8$</td>
<td>-0.0302</td>
<td>0.00197</td>
<td>-15.36</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>life_phase, Childless couples</td>
<td>$\beta_9$</td>
<td>0.0204</td>
<td>0.00227</td>
<td>9.01</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>liv_n</td>
<td>$\beta_{10}$</td>
<td>0.0396</td>
<td>0.00148</td>
<td>26.81</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Pixel_mean_dept</td>
<td>$\beta_{11}$</td>
<td>2.4·10^-7</td>
<td>1.056·10^-8</td>
<td>22.73</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Pixel_mean_assets</td>
<td>$\beta_{12}$</td>
<td>1.611·10^-8</td>
<td>1.633·10^-9</td>
<td>9.87</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>propensity_born_001</td>
<td>$\beta_{13}$</td>
<td>0.0962</td>
<td>0.0315</td>
<td>3.06</td>
<td>0.0022</td>
</tr>
<tr>
<td>advise_n</td>
<td>$\beta_{14}$</td>
<td>0.0753</td>
<td>0.00394</td>
<td>19.10</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>strategic_match</td>
<td>$\beta_{15}$</td>
<td>-0.0580</td>
<td>0.0157</td>
<td>-3.70</td>
<td>0.0002</td>
</tr>
<tr>
<td>tb_n</td>
<td>$\beta_{16}$</td>
<td>0.00287</td>
<td>0.00492</td>
<td>5.84</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>fund_n</td>
<td>$\beta_{17}$</td>
<td>0.00925</td>
<td>0.000812</td>
<td>11.40</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>scard_n</td>
<td>$\beta_{18}$</td>
<td>3.55·10^-6</td>
<td>8.365·10^-7</td>
<td>4.24</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>estd_bal_base</td>
<td>$\beta_{19}$</td>
<td>0.00480</td>
<td>0.000516</td>
<td>9.32</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>dmdep_bal_base</td>
<td>$\beta_{20}$</td>
<td>0.00950</td>
<td>0.000345</td>
<td>27.52</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>finance_incm</td>
<td>$\beta_{21}$</td>
<td>0.5919</td>
<td>0.0194</td>
<td>30.46</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Table 6: Table of variables of the final model.
<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta_j$</th>
<th>Estimate</th>
<th>SE</th>
<th>t stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>fxtrm_bal_base</td>
<td>$\beta_{22}$</td>
<td>-0.00512</td>
<td>0.000879</td>
<td>-5.82</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>insur_incm</td>
<td>$\beta_{23}$</td>
<td>0.1660</td>
<td>0.0211</td>
<td>7.87</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>invest_incm</td>
<td>$\beta_{24}$</td>
<td>0.5361</td>
<td>0.0155</td>
<td>34.70</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>other_incm</td>
<td>$\beta_{25}$</td>
<td>-0.1181</td>
<td>0.0335</td>
<td>-3.52</td>
<td>0.0004</td>
</tr>
<tr>
<td>trade_incm</td>
<td>$\beta_{26}$</td>
<td>0.0563</td>
<td>0.00720</td>
<td>7.82</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>mosaic_grp, A</td>
<td>$\beta_{27}$</td>
<td>0.0933</td>
<td>0.00428</td>
<td>21.79</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>mosaic_grp, B</td>
<td>$\beta_{28}$</td>
<td>0.00620</td>
<td>0.00288</td>
<td>2.15</td>
<td>0.0315</td>
</tr>
<tr>
<td>mosaic_grp, D</td>
<td>$\beta_{29}$</td>
<td>-0.0525</td>
<td>0.00333</td>
<td>-15.76</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>mosaic_grp, F</td>
<td>$\beta_{30}$</td>
<td>-0.0415</td>
<td>0.00398</td>
<td>-10.44</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>gndr_code, Male</td>
<td>$\beta_{31}$</td>
<td>0.0698</td>
<td>0.00123</td>
<td>56.77</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Table 7: Table of variables of the final model.