Business Analytics II - Homework 6

Group 1 - Ema Vargova, Miruna Bortoi, Luka Corsovic

Outline of the problem

The task that was undertaken here was detecting fraudulent transactions, their type and frequency, by using different models and methods applied to a data set that has been given. By looking at all the past transactions and finding any anomalies or patterns, we were able to obtain valuable information that can be used in risk assessment and fraud prevention.

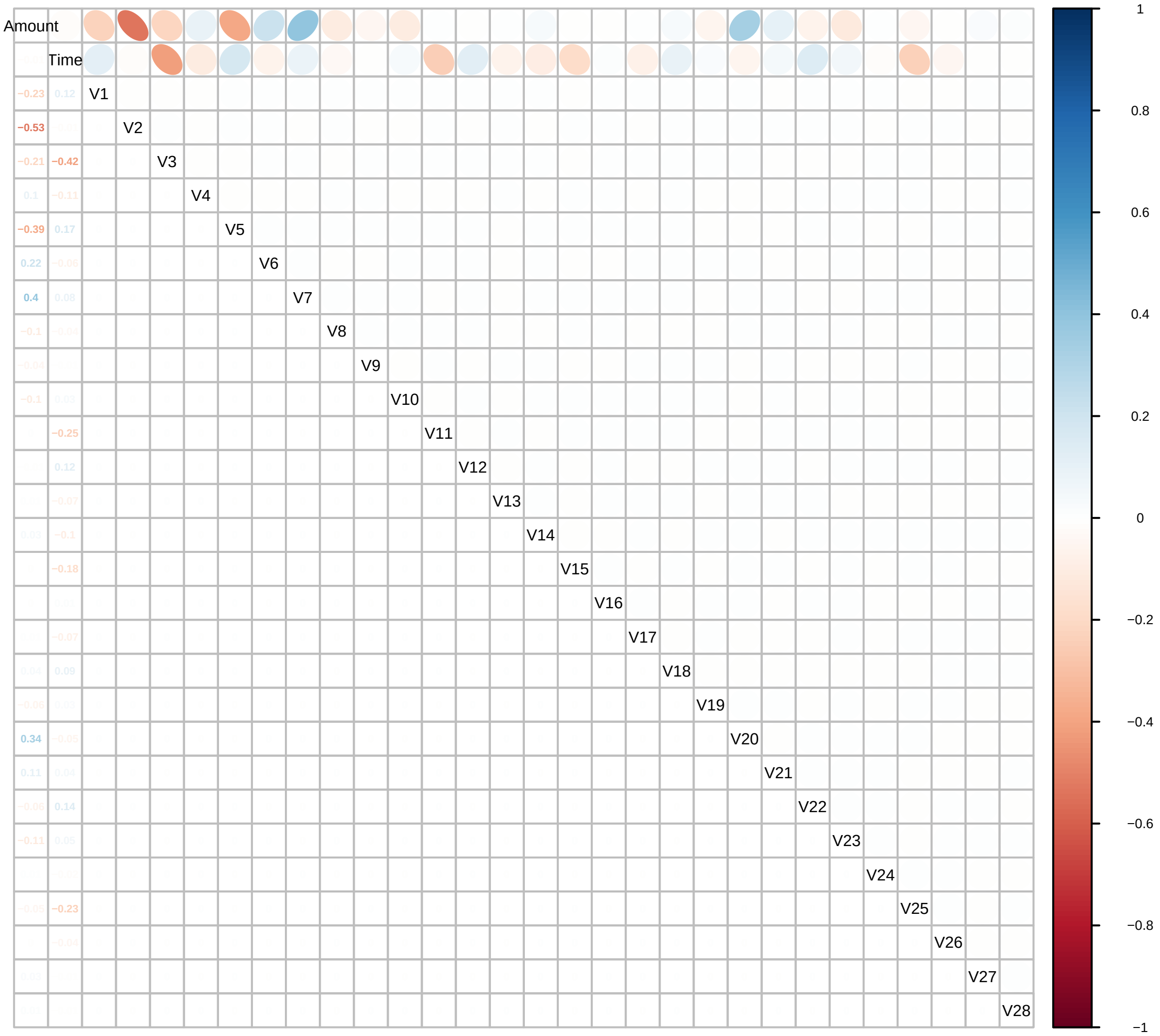
Why is fraud detection a relevant?

Billions of dollars are lost annually due to credit card fraud, and this is an issue that is constantly increasing since it has become more accessible because of the digitalization of the payment processes. Fraud is one of the biggest causes for financial losses and the first step in trying to understand and try to stop it from happening is risk assessment. The rise of these transactions is what makes it a matter that needs to be taken more seriously into consideration and an efficient fraud detection model is of the highest importance. Fraud detection is a central part in the prevention of future fraud from taking place, by analyzing the past data and finding certain patterns and information that can aid into the further understanding of the fraudulent processes. By detecting and looking at the previous fraudulent activity, we can implement certain strategies that will minimize future losses. By incorporating their findings, companies can learn from fraud that has taken place and create a system which is more able to detect fraud as soon as it happens, create a warning, or even prevent it from happening.

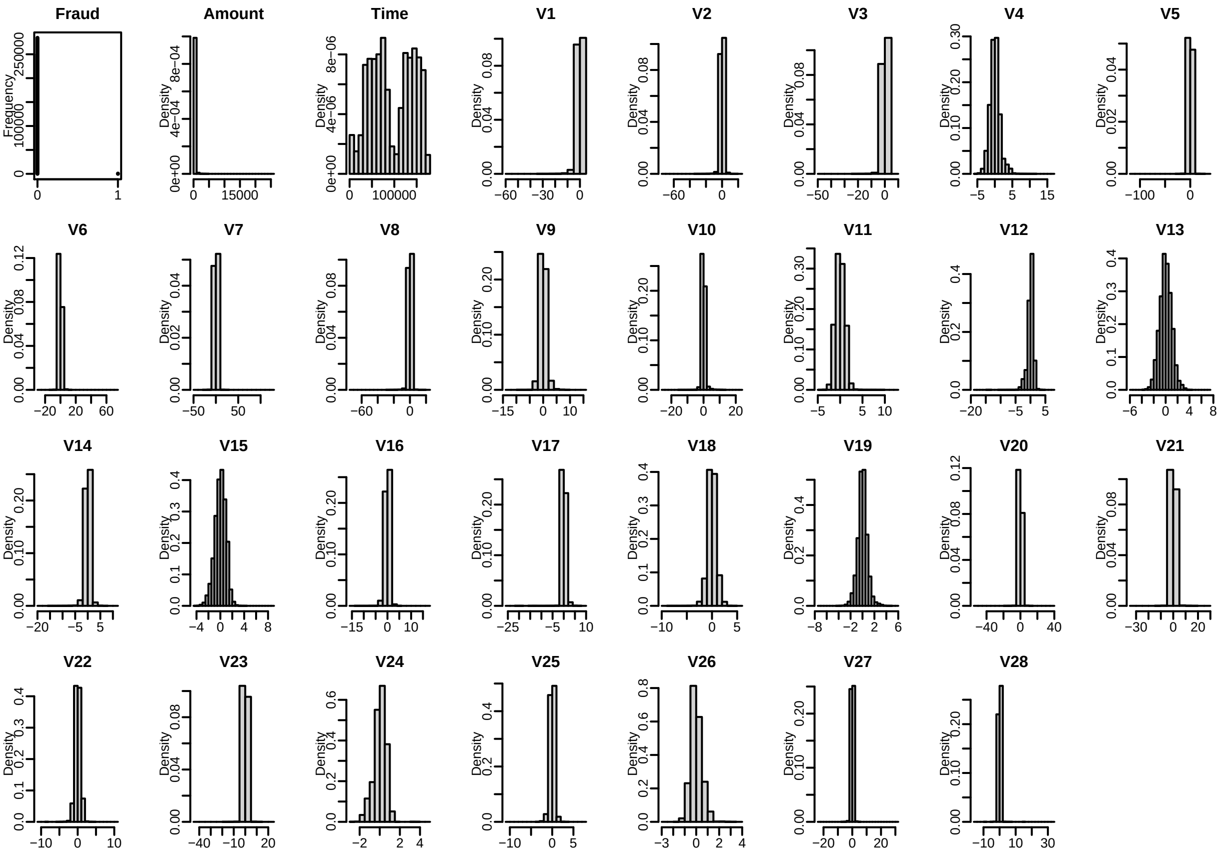
Descriptive analysis of the data set

The dataset provided is data on a number of transactions, which include data on whether or not it was a fraud, the amount of the transaction, the time at which the transaction took place (this is given as the minutes after the first transaction of the data set) as well as 28 variables which due to privacy reasons could not be disclosed. The dataset provided has some limitations which made modeling it more difficult. One of the limitations of the dataset is that we have variables which we do not know what they stand for (V1 - V28), and this makes the modeling harder as it is hard to interpret the potential relationship which the given variables have between themselves. However, we can still achieve that by looking at the correlation between the explanatory variables.

The correlation plot below provides us with insights on how the different variables correlate to each other, in a graphically very descriptive way. The numbers given on the left show us the strength of the correlation between the variables, a negative coefficient meaning that the bigger one of the variables is the smaller the other one is, and a positive coefficient meaning the opposite. On the top a nice visualization of the relationship can be observed, with the way of the relationship is portrayed by the shape of the ball and the colour corresponds to the strength of it. The correlation of the variables is also important for conducting the logistic regression, as one of the assumptions of this model is that there little or no multicollinearity among the variables, and from this plot we can observe that there is not a strong correlation between the variables, hence we can preform logistic regression.



From the marginal distribution below, several things can be observed. One of them is that the frequency of not fraudulent transactions is a lot higher than the one of fraudulent transactions. It can also be seen that the density of the amount in those transactions is usually in the same range. When it comes to timing, two peaks can be seen, with a minimum situated in between them. We can see that most variables seem to have the highest density around the value 0 and that for most of them the density does not vary that much, only when it is close to the point with the highest density. As the frequency of fraudulent transactions is very low when compared to the frequency of non fraudulent transactions, it is difficult to model predictions for it as the data is highly unbalanced. To deal with the problem of unbalanced data, we shall resample it. We will do this by implementing a function in R, that allows us to resample data in a way where we can redefine the ratio of fraudulent to non fraudulent transactions as well as the number of frauds which will be included in the training dataset. The function is also useful because it allows us to split the data into a test and a training data set, which will come useful when we create models and out of sample predictions.



Models used for testing

Since the data is highly unbalanced, we use undersampling to resample the data where we obtain training and test data sets. We use the training data sets to estimates 3 different models and predict the test data sets based on the models. For each model, we resample the data for n amount of times in order get more reliable models and predictions. However, the predictions of response variable we get from the models are continuous variables $\widehat{Y}_i^{cont}$ which we need to classify as either 0 or 1 to turn the continuous response variable into binary variable. To achieve that, we need use a threshold α based on which we classify the prediction i as 0 if $\widehat{Y}_i^{cont} < \alpha$ and we classify the prediction i as 1 if $\widehat{Y}_i^{cont} \geq \alpha$ where natural choice for the threshold is $\alpha = 0.5$.

Linear probability model

The classical model for binary outcomes is the Bernoulli distribution where we assume that $Y_i \sim Ber(p_i)$ for $i = 1, ..., n$. Based on the Bernoulli random variables we arrive to the linear probability model, given by the following formula:

$$Y_i = \beta_0 + \sum_{k=1}^{d} \beta_k X_{i,k} + \epsilon_i,$$

where β_0 is the intercept of the the linear function and β_k is the change in the probability that $Y_i = 1$ while holding the other aggressors constant. We estimate $\beta_0, \beta_1, ..., \beta_d$ via linear regression function using R. As in common multiple regression, we can use OLS estimation (ordinary least squares which is a type of linear least squares method for estimating the unknown parameters in a linear regression model) for inference about the parameters $\beta_0, ..., \beta_d$.

The benefit of this model is that it is very simple to use and understand. It is also very easy to interpret the significance of the different coefficients. One of the shortcoming of this model is that we obtain continuous random variables possibly with values in all of real numbers set $\mathbb{R}$ which is not desirable as the response variable is binary. Another crucial downside of linear regression is that outliers have a huge effect on the model, and can make it worthless.

Logistic regression model

Logistic regression model uses log-odds (also called logit) which describe the probability of response variable Y_i increasing or decreasing based on a regressor X_i where the logit is linear in X. Logit also describes the logistic regression given by:

$$\mathbb{P}(Y_i = 1 \mid X) = \frac{e^{\beta_0+\sum_{k=1}^{d}\beta_k X_{i,k}}}{1+e^{\beta_0+\sum_{k=1}^{d}\beta_k X_{i,k}}}$$

where increasing X_k by one unit changes the log-odds by β_k (it multiplies the odds by e^{β_k}). Estimation of coefficients $\beta_0, \beta_1, ..., \beta_d$ is done via Maximum Likelihood estimation (MLE).

The benefit of using the logistic regression is that the results which we get, unlike in the linear regression model, will always be between 0 and 1, so they will never fall of outside that range. It also gives us a good idea of how important a certain regressor is, and to what extent it influences the probability of the explanatory variable being 1, or in this case of a fraud occurring. It is also less affected by outliers than the linear probability model. The downside is that it assumes linearity between dependent and independent variables, and also that it is very difficult to predict multivariate relationships with a big number of regressors using the logistic regression model, such as we potentially have here because the model becomes less accurate and predicts many independent variables as insignificant.

Regularized logistic regression in Lagrangian form

The regularized logistic regression model works similarly to the logistic regression model, just that now we add a constraint to the beta coefficients of the different explanatory variables. We do this, so that we can end up with a constrained optimization problem, which will then help us to “kick out” certain explanatory variables which do not explain the response variable sufficiently enough. The formula is given by:

$$\max_{\beta_0,\beta}\left\{\sum_{i=1}^{n} Y_i(\beta_0+\beta' X_i) - \log(1+e^{\beta_0+\beta' X_i}) - \lambda\sum_{k=1}^{d}|\beta_k|\right\}$$

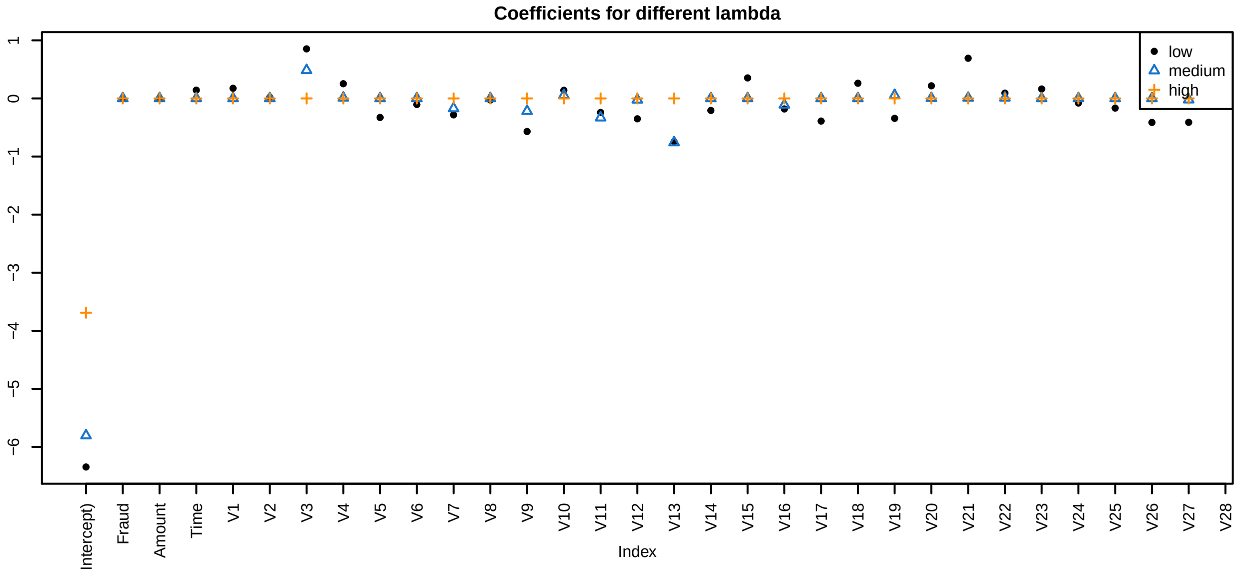
The term λ is chosen by us. As said previously, the penalty term will result in some variables being equal to zero, and this result in a variable selection.

As the penalized regression model is similar to the logistic regression model the same pros and cons apply, with the addition that the penalized regression model allows for a bigger bias in the data by kicking out certain regressors if the model deems to not be “relevant enough”. However, this can also be viewed as a benefit, as it provides us with less explanatory variables to look at, and a simpler model to predict.

Based on the lined out reasons, we believe the regularized logistic regression to be the best model to use in order to detect credit card fraud. It does take a lot of computational power to run it, but as credit card fraud is a serious issue, we believe that this model regardless of its complexity, provides the best results which will help us to identify which transactions where indeed fraudulent, by providing us only the most important variables, also defined as regressors previously, which help distinguish fraudulent transactions from non-fraudulent ones.

Estimation results

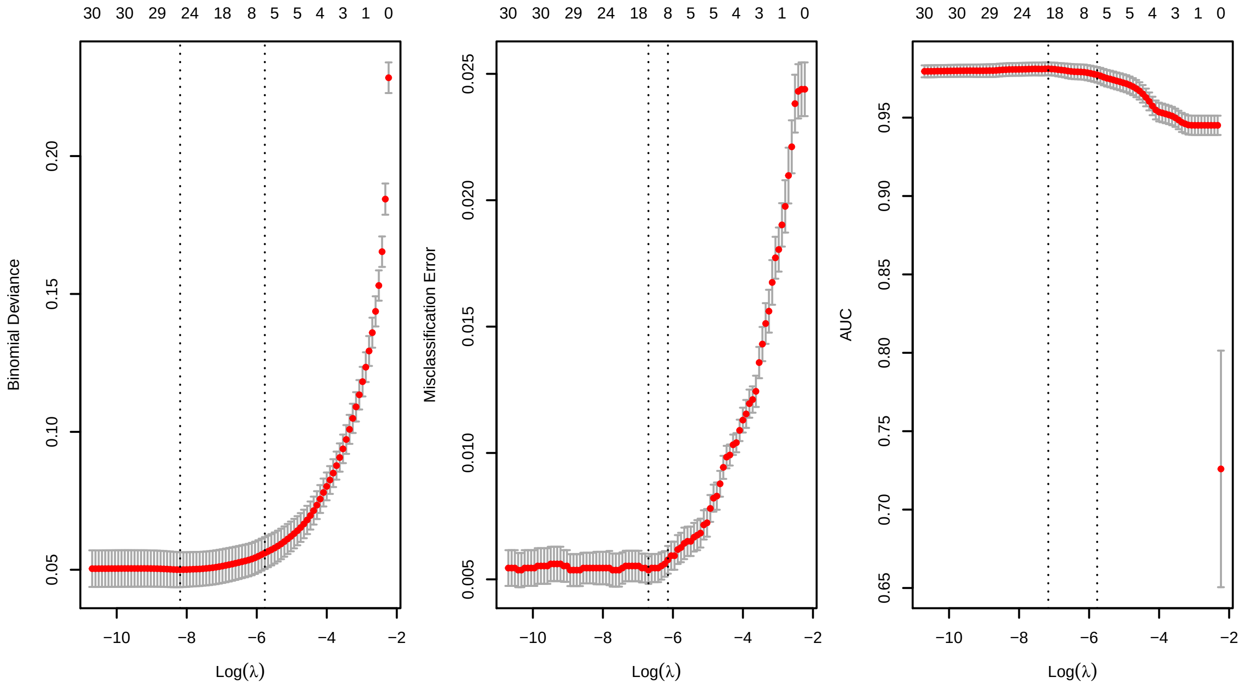
Now we are going to run the regularized logistic regression in order to show the results, explain the significance of certain variables which the model will provide us with as well as the out of sample prediction performance of the model.



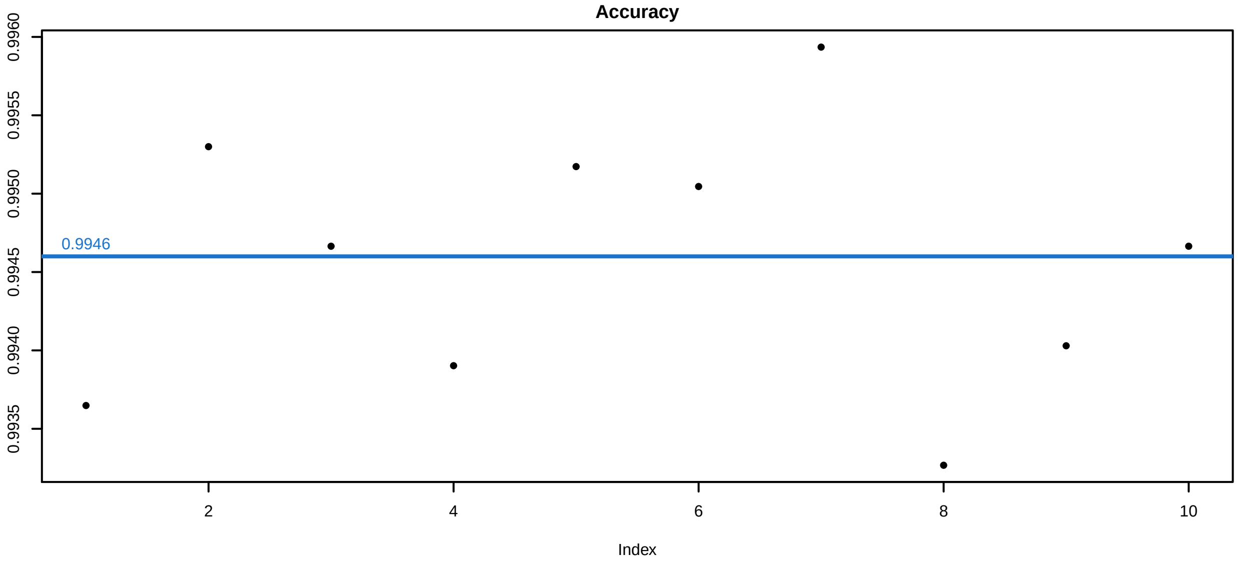
This plot highlights how for different values of λ, the number of non-zero coefficient changes. As we see, the bigger the lambda value is the more and more beta coefficients will be zero, meaning that the variables will not be important for our model. Hence, it is important to choose the right λ value so that the model is not too biased. Luckily, the *cv.glmnet* function in R provides us with the optimal λ value.

To outline the model, we shall firstly present three different methods of cross validation (using *cv.glmnet* function) with which R provides us with, and see which one preforms the best, and that one we shall implement for further analysis.

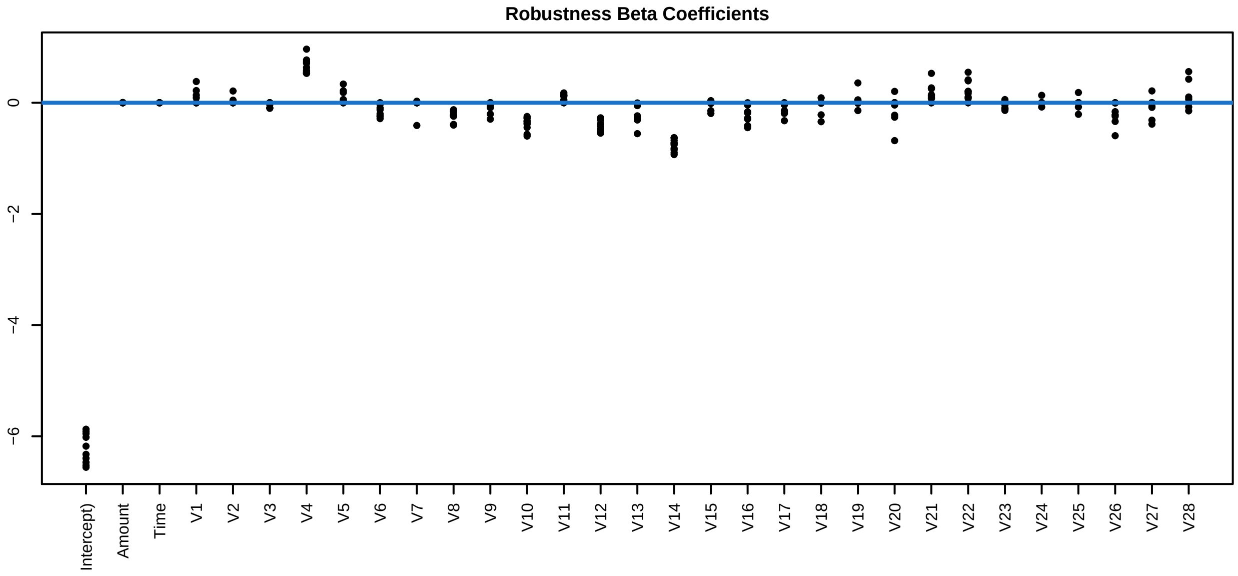
The three plots below represent three different error type measures for cross validation: binomial deviance, misclasification error and area under the curve. Binomial deviance is defined as a quality of fit statistic for model fitting and it generalizes the idea of using sum of squared residuals. The misclassification error type measure provides us with the amount of misclassification errors and area under the curve tells us how well the model can distinguish between the two different groups (0 and 1). The first plot is binomial deviance type measure, the second one is misclassification error type measure and the last one is area under the curve type measure. From the plots we can see that the all of them preform well; for misclassification error and binomial deviance the desired output is that the value is as close to zero as it can be, and we can see that for both of them the value is around 0.05. For the area under the curve type measure, the desired outcome is for the value to be as close to 1 as possible, and from the third graph we can see that our model preforms extremely well, with most of the areas being above 0.95. Another thing the graph provides us with is how many non-zero beta coefficients are present across different values of alpha; this is represented on the top of the graph by the different number values. The binomial deviance type measure has the highest variable loss out of the three, as we can see when we go from the recommended λ value (the first dotted line to the left) to the λ 1se the number of non-zero beta coefficients decreases more than with the other two models. In this respect, the misclassification error performs the best. Overall, we choose to further analyze the area under the curve model as this model preforms the best when it comes to consistent and satisfactory values, and we believe it will be the best one to choose for further model predictions.



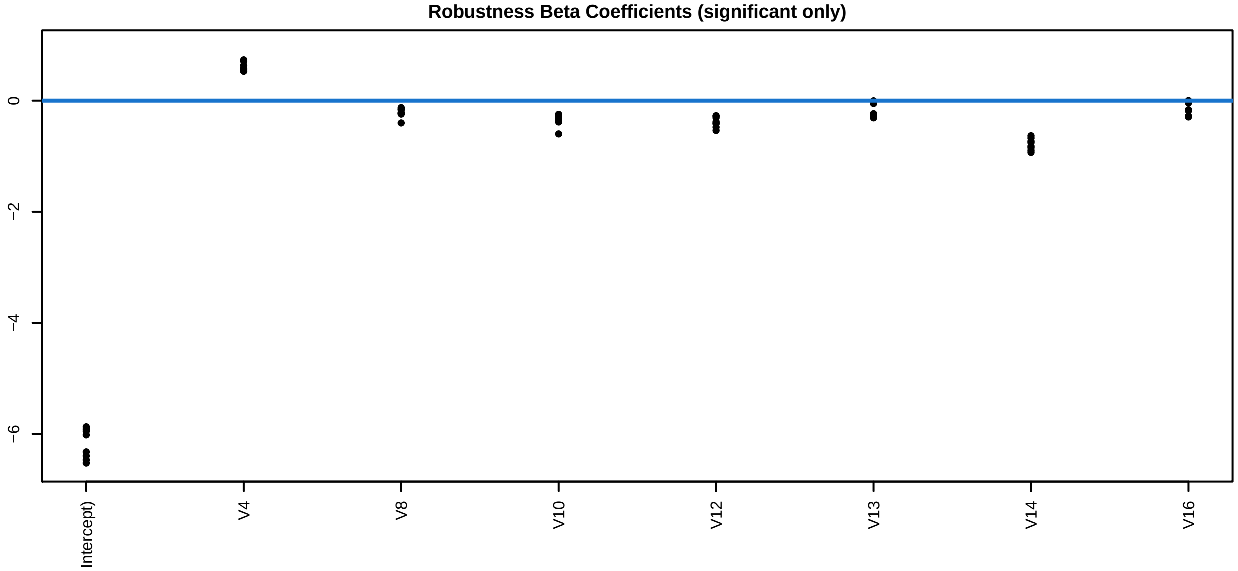
For the next step, we used the output of the resampling which we mentioned earlier. From this we received two different data sets; a train and a test data set. Using the train data set, we created a regression model, which we tested using the test data set. What we did, is we put in the values of the different explanatory variables from the test dataset into the established model, in order to receive "predictions" on what the Y value should be in this case. Then we compared the modeled Y value to the actual Y value of the test data set. We ran this 10 times, resampling the data every time, in order to ensure maximum accuracy.



From the accuracy plot, we can see that the mean accuracy is 99.46%. As mentioned above, this is the ratio to how many times we predicted the Y value correctly. The value of 100% would be ideal, however, this value is very close to 100% so we can judge that the model predicts the dataset very well.



From the robustness of the different beta coefficients, we can see how the beta coefficients for the different variables preform across the iterations. The beta coefficients tell us in what way the variable influences the probability of Y being 1, and thereby the transaction being fraudulent. The significant variables in this case will be the ones for which the beta coefficients are not equal to 0 more than 8 times.



The second plot provides us with only significant beta coefficients and we can see that the variables our model predicted to be significant are: V4, V8, V10, V12, V13, V14 and V16. Only the variable V4 has a positive beta coefficient, meaning that an increase in this variable increases the probability of our transaction being fraudulent. The others have a average beta coefficient which is negative, meaning that they decrease the probability of our transaction being fraudulent.