Summer 2018

Matthew Dobbin

Northwestern University

PREDICT 454 Advanced Modelling Techniques

Kaggle: Home Credit Default Risk

https://www.kaggle.com/c/home-credit-default-risk

Can you predict how capable each applicant is of repaying a loan?

Contents

1.0 Introduction

Home Credit Group market themselves as one of the world's largest global FinTech companies that have disrupted the traditional finance services. They have done this by using advanced algorithms that mitigate risks whilst at the same time providing fast lending decisions.

One focus of the group is responsible lending to people with little or no credit history. This demographic can struggle to get loans and can be taken advantage of by less reputable loan providers. To do this the Home Credit algorithms considers a variety of alternative data to predict the client's repayment abilities. Home Credit has used the Kaggle competition to explore the alternative datasets with the aim of discovering new insights which could lead to an increased accuracy of predicting if clients will default on loan repayments (Kaggle, 2018).

This report details the exploration of the competition data set and the implementation of several machine learning techniques that predict the probability of clients defaulting on loan repayments.

2.0 Literature Review

A review of the literature regarding the application of machine learning techniques in the field of credit risk analysis was conducted. One article, which is a review of patents in credit risk analysis and forecasting, mentions that the financial crises in 2008 and 2011 identified the need for novel solutions to support credit decisions. From the patents review, the authors found that a significantly larger number of patents were filed in 2011-2013 compared to the previous years (Danenas, P., & Garsva, G., 2014). The upward trend in patents is shown in Figure 1.

A wide variety of machine learning techniques have been applied to the field of credit risk decision analysis. Zuranda, Kunene and Guan conducted a review to compare the classification accuracy of multiple methods on multiple credit risk data sets from around the world. The methods that they considered were logistic regression (LR), support vector machine (SVM), k-nearest neighbour (kNN), decision tree (DT), neural network (NN) and radial basis function neural network (RBFNN).

A summary of their findings can be seen in Table 1. They found that model performance is contingent on the nature of the dataset and the business context of the model. For example, decision trees had poor overall performance compared to others when using the area under ROC curves metric. However, they were good at detecting bad loans at higher operating points which may be suitable for a lending institution that has high collateral requirements (Zurada, J., Kunene, N., & Guan, J. 2014).

The literature was also reviewed to determine which predictor variables have the most influence on determining if a customer is likely to default on a loan. One study surmised that "in general the financial attributes of the customers are more important than personal, social and employment ones for the prediction task" (Zurada, J., Kunene, N., & Guan, J. 2014). In the article, "Credit Risk Assessment Using Statistical and Machine Learning", the authors were able to determine the relative influence of the variables by using a decision tree and they found the most influential predictor was level of debt. (Galindo, J., & Tamayo, P. 2010). The relative influence plot can be seen in Figure 2.

While the Home Credit Group advertises that their machine learning models provide a positive and safe borrowing experience and broaden financial inclusion, an accurate model can also lead to significant financial benefits for the lender. In an article titled “Consumer credit-risk models via machine learning algorithms”, the authors estimate that the net benefits of these forecasts to be between 6-25% of total losses. This estimate was based on conservative assumptions and “summing the cost savings from credit reductions to high risk borrowers and the lost revenues from false positives” (Khandani, Kim, & Lo., 2000).

3.0 Exploratory Data Analysis

The data for the competition in contained within seven tables. The common link between the tables is the customer loan ID number (SK_ID_CURR). The full schema for the tables can be viewed in Figure 3. Information contained within includes current application data, previous applications with Home Credit including monthly balance snapshots, credit card history and payment installments. There is also external data from the Credit Bureau.

The main table (application) contains the training and test samples. The target variable is coded a 1 if the client had payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample. The training set contains 307511 rows and 122 variables. Each row represents a unique loan. The test set that requires predictions to be made for the target variable contains 92253 rows.

3.1 Data Preparation and Visualisation

Boxplots, histograms and mosaic plots were used to visually explore the distributions of the variables and to detect outliers. The plots reveal many of the variables are not normally distributed and that potentially non-parametric modelling techniques may be more accurate.

Missing values were replaced with the mean value when there were less than 20% missing values. The variables that had greater than 20% missing values were changed to dummy variables (has a value 1, missing value 0). Categorical variables of n levels were coded into n-1 dummy variables. The plots and a more detailed review of the exploratory data analysis can be seen in Appendix B.

3.2 Feature Engineering

A number of features were created by summarizing the data from the other data tables and then linking them to the application data table using the loan ID. The code from a Kaggle Kernel was referenced to create these features (BlastChar, 2018). New features that were created included the number of credit cards associated with each loan ID, if the applicant had previously been approved or refused for a loan and previous credit amounts.

3.3 Cross Validation

Prior to modelling, the numerical variables were standardized to have zero mean and a standard deviation of one. A uniform random number was used to split the sample into training and test data sets using an approximately 70/30 split respectively. The training set was used for in-sample model development and the test set used for out-of-sample model assessment.

3.4 Correlation

As the dataset had greater than 100 predictor variables, a standard correlation plot was not feasible. A table was used to show the correlation between the predictor variables and the target variable. The twenty predictor variables most correlated with the target variable are shown in Table 2.

The two most highly correlated predictor variables with the target are EXT_SOURCE_2 and EXT_SOURCE_3. These are normalized credit scores for the applicant from external data sources. A pairs scatter plot of the top five correlated variables is shown in Figure 4. It was also found that many of the predictor variables were highly correlated with each other. Variables with a correlation of greater than 0.95 were removed to prevent multicollinearity issues.

4.0 Classification Modelling

The following section describes the different classification modelling techniques that were used to predict the probability of if a customer would have loan payment difficulties. The techniques include logistic regression, linear and quadratic discriminate analysis, gradient boosted models, lasso and ridge regression and support vector machine.

4.1 Logistic Regression

The first model was created using logistic regression and it included all of the predictor variables. The variance inflation factors (VIF) for the predictor variables were checked. VIF measures how correlated each independent variable is with the other predictor variables in the model and is used to detect multicollinearity. A VIF value larger than twenty implies a large inflation of standard errors due to the variable being included in the model. Variables with VIF values greater than twenty were removed one at a time.

The regression coefficients for the full model can be viewed in Table 3. From the table it can be seen that the p-values for a large number of the predictors were not significant (greater than 0.05) and as such could have little impact on predictive accuracy. Two automated variable selection techniques (forward and backward selection) were implemented to determine which variables should be included in the model.

The automated variable selection methods use Akaike's Information Criterion (AIC) in their iterative algorithms to decide if a variable should be included in the model. AIC considers the number of parameters used in the model as well as the goodness-of-fit. Out of the 93 potential predictor variables the backward method selected 69 variables while the forward selected 68. The regression coefficients can be viewed in Table 5 and Table 5.

4.2 Gradient Boosted Model

The next model that was tested was a gradient boosted model which is an approach that improves the predictive accuracy of decision trees. The boosting method has three parameters that can be tuned. They are the number of trees, the shrinkage parameter lambda, which controls the rate at which boosting learns, and the interaction depth which determines the number of splits in each tree. (James, G., Witten, D., Hastie, T., & Tibshirani, R. 2017). The first model used 5000 as the number of trees, a lambda of 0.001 and an interaction depth of three.

A relative influence plot of the predictor variables was generated and is shown in Figure 5. Of the 92 predictor variables, 46 were found to influence the model. The two most influential predictor variables were EXT_SOURCE_2 and EXT_SOURCE_3. Other influential predictor variables included the number of days the applicant had been employed, if they had previously been refused, the age and sex of the applicant and their level of education.

A second gradient boosted model was fitted which focused on optimising the tuning parameters. To reduce computational time, only the predictor variables from the first boosted model that were found to be influential were included. The optimised tuning parameters were found to be number of trees = 150, interaction depth = 3 and shrinkage = 0.1. This led to a small improvement of predictive accuracy compared to the first boosted model.

4.3 Lasso and Ridge Regression

Ridge regression and lasso can be used to fit models containing all of the predictor variables but regularizes the coefficients such that the coefficient estimates shrink towards zero. The tuning parameter lambda in ridge regression controls the effect of the penalty term. When lambda is equal to zero, the regression will produce the least squares estimate. As lambda grows the penalty term has greater effect and the coefficient estimates will approach zero.

The best lambda value for the ridge regression and lasso were selected using cross validation. Plots of the lambda values and the mean cross validated errors are shown in Figure 6 and Figure 7. The lambda value where the minimum cross validation error occurs was selected as the best lambda. The values selected were 0.000095 and 0.004849 for lasso and ridge regression respectively.

4.4 Other Classifier Models

Three other models were fitted. They were a linear discriminate analysis (LDA) model, a quadratic discriminate analysis model (QDA) and a support vector machine (SVM). To reduce computation time, only the predictor variables that had relative influence in the gradient boosted model were included in these models. The support vector machine model which used a radial kernel and the tuning parameters of cost = 1, gamma=0.5 did not arrive at a solution after running for 36 hours.

5.0 Results Comparison

This section of the report compares the nine different classification models with the aim of selecting the best model for predicting the probability of if a client will default on a loan repayment. The out-of-sample performance (predictive accuracy) of the models was investigated by computing the area under the receiver operating characteristic curve (AUC) value. This was due to the Kaggle competition using this metric to evaluate submissions.

AUC is a value that can be used to compare the relative performance among different classifiers. It measures the trade-off between selecting as many true positives as possible while avoiding false positives. The method of using AUC to score the classification model was a common technique for scoring overall performance of the credit risk models in the articles that were reviewed.

The larger the AUC, the better the classifier. An AUC score between 0.7-0.8 represents the model having fair classifying potential. An AUC greater than 0.8 can be taken to indicate the model has good discrimination potential. The AUC scores for the models and the approximate computational processing time can be seen in Figure 8 and Table 8. The worse performing model was the quadradic discriminant analysis. The other models had very similar overall classifying performance with AUC values between 0.73-0.75. However, the computational time required differed significantly with the automated variable selection and tuned gradient boosted models taking approximately three hours to process.

6.0 Conclusion

Unfortunately, the models that were built were heavily influence by the external credit scores. This is not ideal in terms of the Home Credit Groups goal of using alternative data to help predict the capability of first time borrowers. However, the models did show that there were some non-credit history variables that are useful. These include the applicants age, sex and the number of days they have been employed.

In order to improve the classifying accuracy, more time would be required conducting the exploratory data analysis. One example would be to review EXT_SOURCE_1 which was coded as a dummy variable (1 has a value, 0 missing value) due to it missing 54% of values. Potential options could be to compute a mean value or creating a regression to try and predict the missing values to see what effect it has on classifying accuracy.

A screenshot of the Kaggle submission can be seen in Figure 9. The next project I would like to focus more time on the exploratory data analysis and implement just two or three techniques. One of which I would like to be neural networks as I have not implemented this before.

References

BlastChar. (2018). Home Credit Default Risk - Step 2 | Kaggle. Retrieved July 25, 2018, from https://www.kaggle.com/blastchar/home-credit-default-risk-step-2

Danenas, P., & Garsva, G. (2014). Intelligent techniques and systems in credit risk analysis and forecasting: A review of patents. *Journal of Food Science and Technology, 7*(1), 12-23.

Galindo, J., & Tamayo, P. (2000). Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics, 15*(1), 107-143.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*.

Kaggle. (2018). Home Credit Default Risk. Retrieved July 15, 2018, from https://www.kaggle.com/c/home-credit-default-risk

Khandani, Kim, & Lo. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance, 34*(11), 2767-2787.

Zurada, J., Kunene, N., & Guan, J. (2014). The classification performance of multiple methods and datasets: Cases from the loan credit scoring domain. *Journal of International Technology and Information Management*, *23*(1), 57-III.

Appendix A

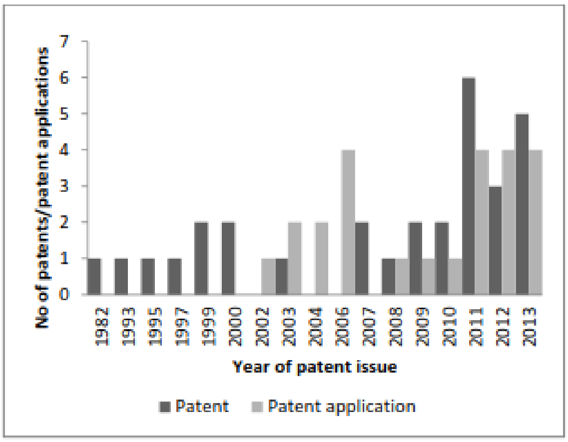

Figure 1 - Analysed patents/patent applications at each year. Retrieved from "Intelligent techniques and systems in credit risk analysis and forecasting: A review of patents" by Danenas, P., & Garsva, G. Journal of Food Science and Technology, 7(1), 12-23. 2014

Table 1 - Summary of the major findings from the Loan Credit Scoring study. Retrieved from The classification performance of multiple methods and datasets: Cases from the loan credit scoring domain by Zurada, J., Kunene, N., & Guan, J. Journal of International Technology and Information Management, 23(1), 57-III. 2014

Data Set	0.5 Cutoff Better Models	Lower cutoffs Better models	Higher cutoffs Better models	Bad Loan avg. classification (Better models)
Australian (medium sized, balanced)	SVM	Model differences indistinguishable	Model differences indistinguishable	RBFNN, DT
SAS-1 (largest, unbalanced, missing values)	NN, DT, *k*NN	*k*NN	NN, DT	NN, DT
SAS-2 (larger, unbalanced, no missing values)	*k*NN and RBFNN	*k*NN	DT, SVM, *k*NN	DT
German (large, more balanced, more attributes)	SVM	SVM	SVM	NN, SVM
Farmer (smallest, unbalanced, real values only)	NN, SVM, *k*NN comparable to LR	*k*NN	Model differences indistinguishable	NN, SVM

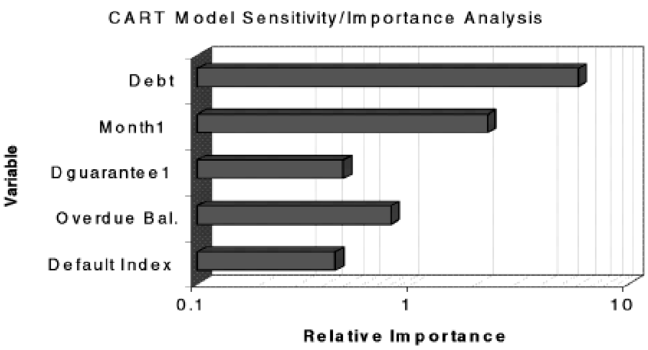

Figure 10. Relative sensitivity/importance for CART.

Figure 2 - Predictor Variables Relative Importance for CART model. Retrieved from "Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications" by Galindo, J., & Tamayo, P. Computational Economics, 15(1), 107-143. 2000

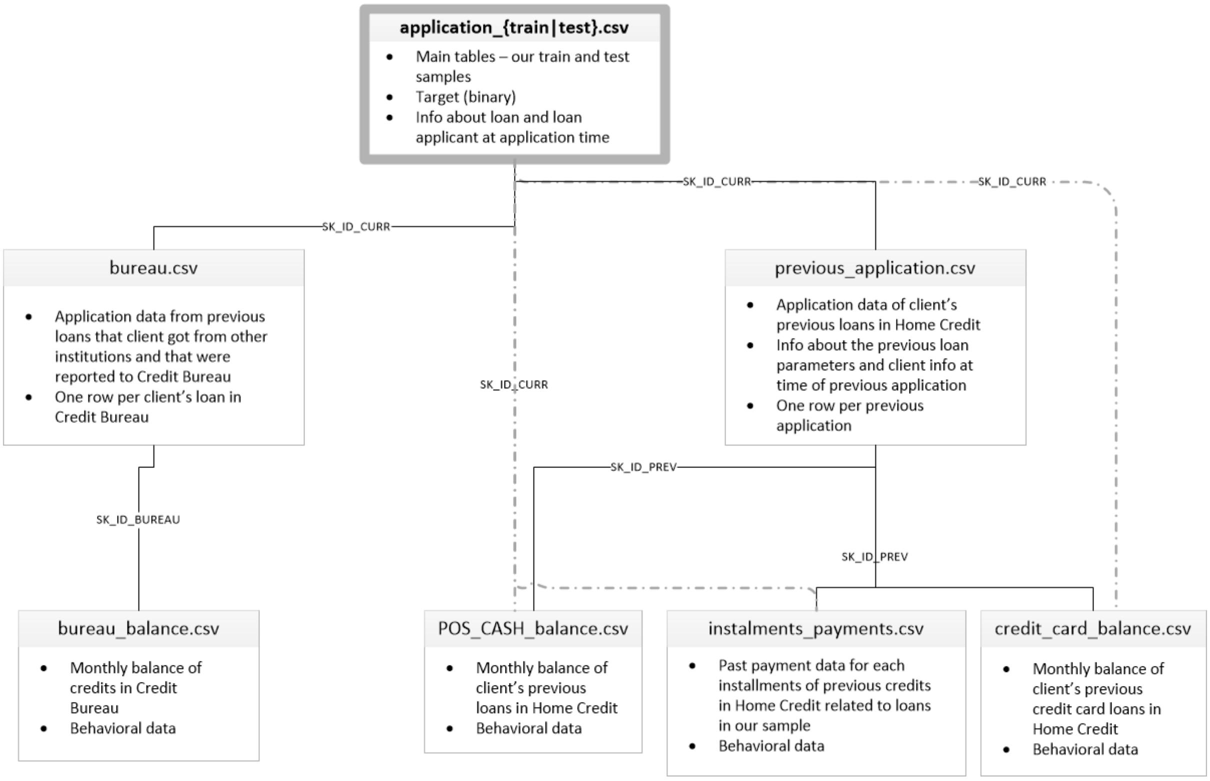

Figure 3 - Data set schema. Retrieved from Home Credit Default Risk by Kaggle, 2018 https://www.kaggle.com/c/home-credit-default-risk

Table 2 - Top 20 most correlated predictor variables with the target.

Variable 1	Variable 2	Correlation
TARGET	EXT_SOURCE_2	-0.16
TARGET	EXT_SOURCE_3	-0.16
TARGET	DAYS_BIRTH	0.08
TARGET	DAYS_EMPLOYED	0.07
TARGET	MAX_DAYS_CREDIT	-0.07
TARGET	Refused	0.06
TARGET	REGION_RATING_CLIENT_W_CITY	0.06
TARGET	REGION_RATING_CLIENT	0.06
TARGET	NAME_INCOME_TYPE_W	0.06
TARGET	NAME_EDUCATION_TYPE_HE	-0.06
TARGET	CODE_GENDER_M	0.05
TARGET	DAYS_LAST_PHONE_CHANGE	0.05
TARGET	REG_CITY_NOT_WORK_CITY	0.05
TARGET	DAYS_ID_PUBLISH	0.05
TARGET	NAME_EDUCATION_TYPE_SS	0.05
TARGET	REG_CITY_NOT_LIVE_CITY	0.05
TARGET	New	0.05
TARGET	OCCUPATION_TYPE_LA	0.04
TARGET	FLAG_DOCUMENT_3	0.04
TARGET	DAYS_REGISTRATION	0.04

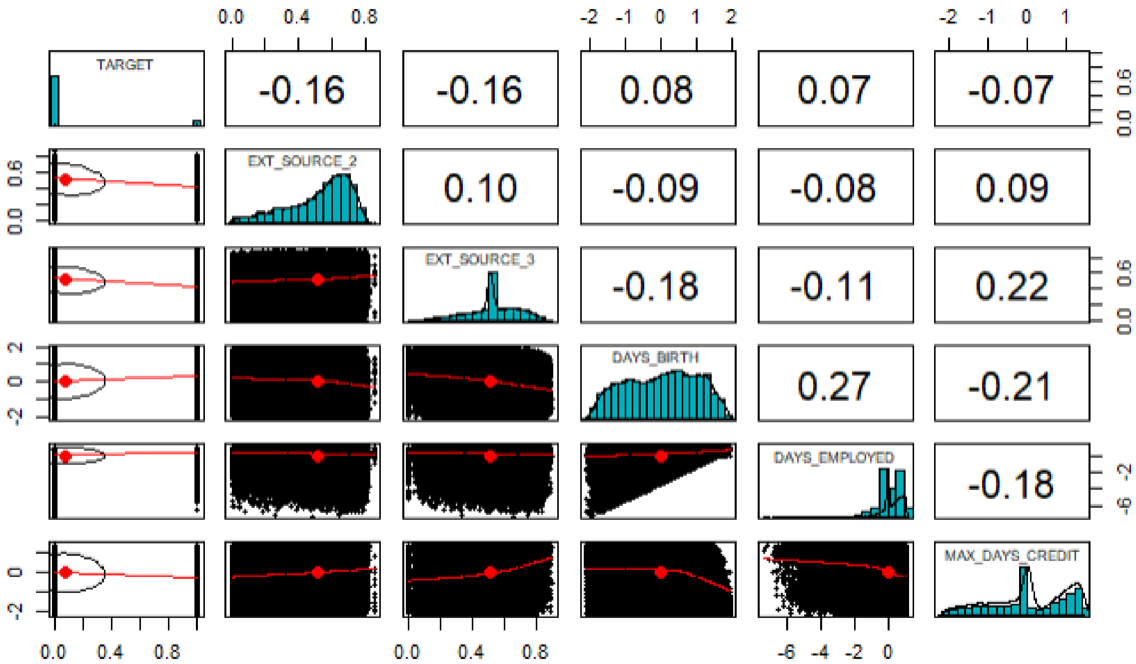

Figure 4 - Pairs scatterplot of the five predictor variables most correlated with the target.

Table 3 - Full logistic regression model coefficients.

Coefficients:

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-0.94973	0.070993	-13.378	< 0.00000000000002
CNT_CHILDREN	0.134397	0.022827	5.888	3.92E-09
AMT_INCOME_TOTAL	-0.03025	0.011141	-2.715	0.006633
AMT_CREDIT	-0.03609	0.014092	-2.561	0.010429
AMT_ANNUITY	0.129307	0.014587	8.864	< 0.00000000000002
REGION_POPULATION_RELATIVE	0.014753	0.010882	1.356	0.175166
DAYS_BIRTH	0.061862	0.012202	5.07	3.99E-07
DAYS_EMPLOYED	0.131661	0.010977	11.995	< 0.00000000000002
DAYS_REGISTRATION	0.029337	0.009365	3.133	0.001733
DAYS_ID_PUBLISH	0.058105	0.008976	6.474	9.56E-11
FLAG_WORK_PHONE	0.129024	0.021421	6.023	1.71E-09
FLAG_PHONE	-0.09167	0.020478	-4.477	7.58E-06
CNT_FAM_MEMBERS	-0.11966	0.026135	-4.579	4.68E-06
REGION_RATING_CLIENT	-0.00401	0.028089	-0.143	0.886528
REGION_RATING_CLIENT_W_CITY	0.070254	0.027898	2.518	0.011795
HOUR_APPR_PROCESS_START	-0.01863	0.008713	-2.138	0.032531
REG_CITY_NOT_LIVE_CITY	0.19632	0.040333	4.867	1.13E-06
REG_CITY_NOT_WORK_CITY	-0.05925	0.045578	-1.3	0.193584
LIVE_CITY_NOT_WORK_CITY	0.037598	0.043983	0.855	0.392653
EXT_SOURCE_1	-0.15679	0.017697	-8.86	< 0.00000000000002
EXT_SOURCE_2	-2.08559	0.043065	-48.428	< 0.00000000000002
EXT_SOURCE_3	-2.51694	0.050189	-50.149	< 0.00000000000002
LIVINGAPARTMENTS_AVG	-0.00424	0.069207	-0.061	0.951199
LIVINGAREA_AVG	0.016719	0.054376	0.307	0.758482
APARTMENTS_MODE	-0.00853	0.065575	-0.13	0.896454
BASEMENTAREA_MODE	-0.04172	0.043666	-0.955	0.339395
YEARS_BUILD_MODE	-0.05756	0.073146	-0.787	0.431319
COMMONAREA_MODE	-0.02413	0.05633	-0.428	0.668424
ELEVATORS_MODE	0.001405	0.053259	0.026	0.978958
FLOORSMIN_MODE	0.099374	0.07391	1.345	0.178776
LANDAREA_MODE	0.066791	0.040984	1.63	0.103165
NONLIVINGAREA_MODE	-0.07019	0.050395	-1.393	0.163689
OBS_30_CNT_SOCIAL_CIRCLE	0.002353	0.008685	0.271	0.786456
DEF_30_CNT_SOCIAL_CIRCLE	0.058563	0.015171	3.86	0.000113
DEF_60_CNT_SOCIAL_CIRCLE	0.02491	0.014862	1.676	0.093733
DAYS_LAST_PHONE_CHANGE	0.021436	0.009809	2.185	0.028871
FLAG_DOCUMENT_3	0.229418	0.028231	8.127	4.42E-16
FLAG_DOCUMENT_6	0.165176	0.047404	3.484	0.000493
FLAG_DOCUMENT_13	-0.74221	0.230858	-3.215	0.001304
FLAG_DOCUMENT_16	-0.47729	0.105285	-4.533	5.81E-06
AMT_REQ_CREDIT_BUREAU_MON	-0.02811	0.009617	-2.922	0.003475
AMT_REQ_CREDIT_BUREAU_YEAR	0.002166	0.009909	0.219	0.826945

MAX_DAYS_CREDIT	-0.03186	0.009296	-3.427	0.000611
CREDIT_ACTIVE	0.076949	0.007745	9.935	< 0.00000000000002
NCreditCard	-0.05298	0.009515	-5.567	2.59E-08
CNT_INSTALMENT	0.164	0.010314	15.9	< 0.00000000000002
Active	-0.16618	0.01223	-13.589	< 0.00000000000002
Completed	0.011383	0.009984	1.14	0.254271
AMT_ANNUITY_P	-0.14654	0.018204	-8.05	8.3E-16
AMT_APPLICATION_P	-0.0697	0.026524	-2.628	0.008594
AMT_DOWN_PAYMENT_P	-0.05285	0.013276	-3.981	6.86E-05
AMT_GOODS_PRICE_P	0.146757	0.029595	4.959	7.09E-07
Canceled.y	-0.02179	0.011037	-1.974	0.048347
Refused	0.113258	0.00763	14.844	< 0.00000000000002
New	0.013198	0.008657	1.525	0.127381
Refreshed	-0.00827	0.008584	-0.963	0.335405
NAME_CONTRACT_TYPE_DMY	0.159177	0.041376	3.847	0.00012
CODE_GENDER_M	0.334977	0.021505	15.577	< 0.0000000000002
FLAG_OWN_CAR_Y	-0.23471	0.019719	-11.903	< 0.0000000000002
NAME_INCOME_TYPE_W	0.113344	0.018918	5.991	2.08E-09
NAME_EDUCATION_TYPE_HE	-0.0867	0.048556	-1.786	0.07417
NAME_EDUCATION_TYPE_LS	0.335181	0.07966	4.208	2.58E-05
NAME_EDUCATION_TYPE_SS	0.219302	0.045751	4.793	1.64E-06
NAME_FAMILY_STATUS_CM	0.163171	0.027157	6.008	1.87E-09
NAME_FAMILY_STATUS_SNM	0.015352	0.034143	0.45	0.652961
NAME_HOUSING_TYPE_H	-0.04449	0.026625	-1.671	0.09476
NAME_HOUSING_TYPE_R	0.027771	0.06069	0.458	0.647244
OCCUPATION_TYPE_A	-0.19067	0.061833	-3.084	0.002045
OCCUPATION_TYPE_CK	0.114825	0.05721	2.007	0.04474
OCCUPATION_TYPE_CS	-0.0988	0.036676	-2.694	0.007062
OCCUPATION_TYPE_D	0.19221	0.037932	5.067	4.04E-07
OCCUPATION_TYPE_HS	-0.09828	0.051961	-1.891	0.058564
OCCUPATION_TYPE_LA	0.126161	0.026427	4.774	1.81E-06
OCCUPATION_TYPE_LS	0.271421	0.077042	3.523	0.000427
OCCUPATION_TYPE_M	0.015538	0.04048	0.384	0.701098
OCCUPATION_TYPE_MD	-0.08149	0.070924	-1.149	0.250566
OCCUPATION_TYPE_SS	0.051515	0.031141	1.654	0.098083
OCCUPATION_TYPE_SEC	0.183138	0.052974	3.457	0.000546
ORGANIZATION_TYPE_BT3	0.129628	0.021879	5.925	3.13E-09
ORGANIZATION_TYPE_CO	0.288133	0.050637	5.69	1.27E-08
ORGANIZATION_TYPE_I3	0.177622	0.071839	2.473	0.013417
ORGANIZATION_TYPE_MD	0.004907	0.061269	0.08	0.93617
ORGANIZATION_TYPE_RS	0.248519	0.09339	2.661	0.007789
ORGANIZATION_TYPE_SC	-0.13898	0.059398	-2.34	0.01929
ORGANIZATION_TYPE_SM	-0.32537	0.132721	-2.452	0.014225
ORGANIZATION_TYPE_SL	0.227168	0.026238	8.658	< 0.00000000000002
ORGANIZATION_TYPE_TR3	0.484112	0.10668	4.538	5.68E-06
FONDKAPREMONT_MODE_OSA	-0.18862	0.083114	-2.269	0.023241

FONDKAPREMONT_MODE_ROA	0.01068	0.04564	0.234	0.814981
FONDKAPREMONT_MODE_ROS	-0.07923	0.061875	-1.28	0.200389
HOUSETYPE_MODE_F	-0.03162	0.056378	-0.561	0.574935
WALLSMATERIAL_MODE_P	-0.08118	0.03715	-2.185	0.028876
WALLSMATERIAL_MODE_SB	0.039778	0.035857	1.109	0.267273

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 120709 on 215256 degrees of freedom
Residual deviance: 106933 on 215164 degrees of freedom
AIC: 107119

Table 4 - Forward selection logistic regression coefficients.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.956605	0.06925	-13.814	< 0.000000000002
EXT_SOURCE_3	-2.517168	0.050151	-50.192	< 0.000000000002
EXT_SOURCE_2	-2.084194	0.042845	-48.645	< 0.000000000002
DAYS_EMPLOYED	0.130846	0.010921	11.981	< 0.000000000002
NAME_EDUCATION_TYPE_HE	-0.086912	0.048523	-1.791	0.07327
CODE_GENDER_M	0.334616	0.021442	15.606	< 0.000000000002
FLAG_DOCUMENT_3	0.227997	0.028185	8.089	6E-16
Refused	0.113101	0.007539	15.002	< 0.000000000002
Active	-0.158964	0.010844	-14.659	< 0.000000000002
CNT_INSTALMENT	0.162718	0.010093	16.122	< 0.000000000002
NAME_INCOME_TYPE_W	0.11249	0.018832	5.973	2.33E-09
FLAG_OWN_CAR_Y	-0.233501	0.01967	-11.871	< 0.000000000002
DEF_30_CNT_SOCIAL_CIRCLE	0.059412	0.014792	4.016	5.91E-05
AMT_ANNUITY_P	-0.145698	0.018143	-8.031	9.71E-16
REG_CITY_NOT_LIVE_CITY	0.16308	0.027077	6.023	1.71E-09
ORGANIZATION_TYPE_SL	0.228887	0.026048	8.787	< 0.000000000002
AMT_ANNUITY	0.129135	0.014573	8.861	< 0.000000000002
REGION_RATING_CLIENT_W_CITY	0.06208	0.009182	6.761	1.37E-11
EXT_SOURCE_1	-0.155642	0.017677	-8.805	< 0.000000000002
DAYS_BIRTH	0.061062	0.012147	5.027	4.99E-07
WALLSMATERIAL_MODE_P	-0.119617	0.022406	-5.339	9.36E-08
CREDIT_ACTIVE	0.076999	0.007736	9.954	< 0.000000000002
NCreditCard	-0.051823	0.009365	-5.534	3.13E-08
DAYS_ID_PUBLISH	0.058049	0.008962	6.477	9.35E-11
OCCUPATION_TYPE_CS	-0.102961	0.035699	-2.884	0.003925
NAME_CONTRACT_TYPE_DMY	0.160264	0.041366	3.874	0.000107
FLAG_DOCUMENT_16	-0.47865	0.105181	-4.551	5.35E-06
ORGANIZATION_TYPE_BT3	0.129612	0.021735	5.963	2.47E-09
ORGANIZATION_TYPE_CO	0.286935	0.050558	5.675	1.38E-08
ORGANIZATION_TYPE_TR3	0.485095	0.106598	4.551	5.35E-06
NAME_FAMILY_STATUS_CM	0.162598	0.027133	5.993	2.07E-09
MAX_DAYS_CREDIT	-0.032781	0.009232	-3.551	0.000384
AMT_DOWN_PAYMENT_P	-0.053295	0.013289	-4.01	6.06E-05
FLAG_WORK_PHONE	0.128184	0.021313	6.014	1.81E-09
FLAG_PHONE	-0.091289	0.020434	-4.468	7.91E-06
OCCUPATION_TYPE_A	-0.194451	0.061137	-3.181	0.00147
NAME_FAMILY_STATUS_SNM	0.014847	0.034135	0.435	0.663594
CNT_CHILDREN	0.133887	0.022813	5.869	4.39E-09
CNT_FAM_MEMBERS	-0.118806	0.026106	-4.551	5.34E-06
AMT_GOODS_PRICE_P	0.147855	0.029526	5.008	5.51E-07
FLAG_DOCUMENT_13	-0.741844	0.230799	-3.214	0.001308
OCCUPATION_TYPE_HS	-0.102925	0.051214	-2.01	0.044461
NAME_EDUCATION_TYPE_SS	0.219848	0.045703	4.81	1.51E-06

NAME_EDUCATION_TYPE_LS	0.336141	0.079577	4.224	2.4E-05
AMT_INCOME_TOTAL	-0.03047	0.01106	-2.755	0.005871
DAYS_REGISTRATION	0.029814	0.009346	3.19	0.001422
AMT_REQ_CREDIT_BUREAU_MON	-0.028235	0.009604	-2.94	0.003283
FLAG_DOCUMENT_6	0.164968	0.047116	3.501	0.000463
OCCUPATION_TYPE_D	0.186762	0.036742	5.083	3.71E-07
OCCUPATION_TYPE_LA	0.12107	0.025096	4.824	1.41E-06
OCCUPATION_TYPE_LS	0.267579	0.076669	3.49	0.000483
OCCUPATION_TYPE_SEC	0.177877	0.052433	3.392	0.000693
ORGANIZATION_TYPE_RS	0.245924	0.093318	2.635	0.008406
FONDKAPREMONT_MODE_OSA	-0.199203	0.073047	-2.727	0.00639
AMT_CREDIT	-0.035559	0.014064	-2.528	0.011459
ORGANIZATION_TYPE_I3	0.177163	0.071763	2.469	0.013559
ORGANIZATION_TYPE_SM	-0.328452	0.132658	-2.476	0.013289
DAYS_LAST_PHONE_CHANGE	0.021493	0.009741	2.206	0.027352
ORGANIZATION_TYPE_SC	-0.137552	0.059263	-2.321	0.020285
HOUR_APPR_PROCESS_START	-0.018159	0.00864	-2.102	0.035574
FONDKAPREMONT_MODE_ROS	-0.091609	0.046843	-1.956	0.050503
NAME_HOUSING_TYPE_H	-0.047134	0.024812	-1.9	0.057483
OCCUPATION_TYPE_CK	0.110782	0.056763	1.952	0.050977
OCCUPATION_TYPE_SS	0.047786	0.030113	1.587	0.112532
AMT_APPLICATION_P	-0.070498	0.026456	-2.665	0.007706
Canceled.y	-0.02185	0.010528	-2.075	0.037951
DEF_60_CNT_SOCIAL_CIRCLE	0.024896	0.014831	1.679	0.093218
New	0.014185	0.00845	1.679	0.093211
OCCUPATION_TYPE_MD	-0.083777	0.057894	-1.447	0.147874

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 120709 on 215256 degrees of freedom
Residual deviance: 106947 on 215188 degrees of freedom
AIC: 107085

Table 5 - Backward selection logistic regression model coefficients.

Coefficients:

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-0.947576	0.0694	-13.655	< 0.00000000002
CNT_CHILDREN	0.140178	0.01818	7.709	1.2638E-14
AMT_INCOME_TOTAL	-0.030041	0.01107	-2.715	0.006627
AMT_CREDIT	-0.035882	0.01407	-2.551	0.010743
AMT_ANNUITY	0.129571	0.01458	8.89	< 0.00000000002
DAYS_BIRTH	0.062419	0.01186	5.261	1.43054E-07
DAYS_EMPLOYED	0.130835	0.01092	11.978	< 0.00000000002
DAYS_REGISTRATION	0.029613	0.00935	3.168	0.001534
DAYS_ID_PUBLISH	0.05816	0.00896	6.49	8.59606E-11
FLAG_WORK_PHONE	0.127502	0.02133	5.978	2.25957E-09
FLAG_PHONE	-0.090612	0.02044	-4.433	9.30997E-06
CNT_FAM_MEMBERS	-0.127442	0.01824	-6.989	2.77829E-12
REGION_RATING_CLIENT_W_CITY	0.061612	0.0092	6.695	2.15312E-11
HOUR_APPR_PROCESS_START	-0.018043	0.00864	-2.088	0.036815
REG_CITY_NOT_LIVE_CITY	0.159898	0.02732	5.853	4.82886E-09
EXT_SOURCE_1	-0.155835	0.01768	-8.814	< 0.00000000002
EXT_SOURCE_2	-2.081757	0.04288	-48.552	< 0.00000000002
EXT_SOURCE_3	-2.516887	0.05015	-50.187	< 0.00000000002
LANDAREA_MODE	0.057108	0.03489	1.637	0.101707
NONLIVINGAREA_MODE	-0.069348	0.03486	-1.989	0.046658
DEF_30_CNT_SOCIAL_CIRCLE	0.059581	0.01479	4.028	5.63093E-05
DEF_60_CNT_SOCIAL_CIRCLE	0.02483	0.01483	1.674	0.094109
DAYS_LAST_PHONE_CHANGE	0.021619	0.00974	2.219	0.026472
FLAG_DOCUMENT_3	0.227917	0.02819	8.086	6.16E-16
FLAG_DOCUMENT_6	0.164608	0.04712	3.494	0.000476
FLAG_DOCUMENT_13	-0.742344	0.23081	-3.216	0.001299
FLAG_DOCUMENT_16	-0.477574	0.1052	-4.54	5.63048E-06
AMT_REQ_CREDIT_BUREAU_MON	-0.027866	0.00961	-2.9	0.003731
MAX_DAYS_CREDIT	-0.032783	0.00923	-3.55	0.000385
CREDIT_ACTIVE	0.077047	0.00774	9.959	< 0.00000000002
NCreditCard	-0.051642	0.00937	-5.514	3.50656E-08
CNT_INSTALMENT	0.162595	0.01009	16.109	< 0.00000000002
Active	-0.158969	0.01084	-14.659	< 0.00000000002
AMT_ANNUITY_P	-0.14562	0.01814	-8.026	1.003E-15
AMT_APPLICATION_P	-0.070475	0.02646	-2.664	0.007722
AMT_DOWN_PAYMENT_P	-0.053227	0.01328	-4.008	6.12663E-05
AMT_GOODS_PRICE_P	0.147821	0.02953	5.007	5.53796E-07
Canceled.y	-0.021697	0.01053	-2.061	0.039312
Refused	0.113139	0.00754	15.005	< 0.0000000002
New	0.014092	0.00845	1.667	0.095423
NAME_CONTRACT_TYPE_DMY	0.159598	0.04137	3.858	0.000114
CODE_GENDER_M	0.335588	0.02135	15.72	< 0.0000000002
FLAG_OWN_CAR_Y	-0.234163	0.01968	-11.896	< 0.0000000002

NAME_INCOME_TYPE_W	0.111758	0.01884	5.933	2.96768E-09
NAME_EDUCATION_TYPE_HE	-0.087006	0.04852	-1.793	0.072939
NAME_EDUCATION_TYPE_LS	0.333194	0.07961	4.185	2.84859E-05
NAME_EDUCATION_TYPE_SS	0.218223	0.04571	4.774	1.80459E-06
NAME_FAMILY_STATUS_CM	0.162731	0.02713	5.997	2.00664E-09
NAME_HOUSING_TYPE_H	-0.047292	0.02481	-1.906	0.056594
OCCUPATION_TYPE_A	-0.194362	0.06114	-3.179	0.001477
OCCUPATION_TYPE_CK	0.111141	0.05676	1.958	0.050217
OCCUPATION_TYPE_CS	-0.102822	0.0357	-2.88	0.003973
OCCUPATION_TYPE_D	0.186641	0.03674	5.08	3.7688E-07
OCCUPATION_TYPE_HS	-0.102348	0.05122	-1.998	0.045688
OCCUPATION_TYPE_LA	0.121005	0.0251	4.821	1.42566E-06
OCCUPATION_TYPE_LS	0.266922	0.07667	3.481	0.000499
OCCUPATION_TYPE_MD	-0.083962	0.0579	-1.45	0.147004
OCCUPATION_TYPE_SS	0.048057	0.03012	1.596	0.110532
OCCUPATION_TYPE_SEC	0.177737	0.05243	3.39	0.0007
ORGANIZATION_TYPE_BT3	0.129893	0.02174	5.976	2.28838E-09
ORGANIZATION_TYPE_CO	0.287091	0.05056	5.679	1.35838E-08
ORGANIZATION_TYPE_I3	0.176439	0.07177	2.458	0.01396
ORGANIZATION_TYPE_RS	0.246812	0.09332	2.645	0.008176
ORGANIZATION_TYPE_SC	-0.138813	0.05928	-2.342	0.019204
ORGANIZATION_TYPE_SM	-0.328143	0.13265	-2.474	0.013367
ORGANIZATION_TYPE_SL	0.228195	0.02606	8.758	< 0.000000000002
ORGANIZATION_TYPE_TR3	0.484347	0.10661	4.543	5.54354E-06
FONDKAPREMONT_MODE_OSA	-0.195476	0.07351	-2.659	0.007833
FONDKAPREMONT_MODE_ROS	-0.0862	0.0477	-1.807	0.070732
WALLSMATERIAL_MODE_P	-0.111387	0.02523	-4.415	1.01145E-05

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 120709 on 215256 degrees of freedom
Residual deviance: 106944 on 215187 degrees of freedom
AIC: 107084

Table 6 - Gradient boosted model predictors relative influence.

```
gbm(formula = TARGET ~ ., distribution = "bernoulli", data = trainset,
    n.trees = 5000, interaction.depth = 3)
A gradient boosted model with bernoulli loss function.
5000 iterations were performed.
There were 92 predictors of which 46 had non-zero influence.

                                                      var      rel.inf
EXT_SOURCE_2                                 EXT_SOURCE_2 38.306539961
EXT_SOURCE_3                                 EXT_SOURCE_3 37.975544602
DAYS_EMPLOYED                               DAYS_EMPLOYED  4.748299613
Refused                                           Refused  2.680765787
CODE_GENDER_M                               CODE_GENDER_M  2.620365613
NAME_EDUCATION_TYPE_HE               NAME_EDUCATION_TYPE_HE  2.094787410
DAYS_BIRTH                                     DAYS_BIRTH  1.646264984
Active                                             Active  1.512420666
CNT_INSTALMENT                             CNT_INSTALMENT  0.906589706
FLAG_DOCUMENT_3                           FLAG_DOCUMENT_3  0.887121138
NAME_EDUCATION_TYPE_SS               NAME_EDUCATION_TYPE_SS  0.808604396
AMT_DOWN_PAYMENT_P                     AMT_DOWN_PAYMENT_P  0.737007881
MAX_DAYS_CREDIT                           MAX_DAYS_CREDIT  0.674964180
NAME_INCOME_TYPE_W                     NAME_INCOME_TYPE_W  0.615851674
AMT_ANNUITY                                   AMT_ANNUITY  0.551306994
FLAG_OWN_CAR_Y                             FLAG_OWN_CAR_Y  0.520218619
AMT_CREDIT                                     AMT_CREDIT  0.470121209
REG_CITY_NOT_LIVE_CITY               REG_CITY_NOT_LIVE_CITY  0.436624613
Completed                                       Completed  0.392854567
DEF_30_CNT_SOCIAL_CIRCLE           DEF_30_CNT_SOCIAL_CIRCLE  0.269253588
AMT_ANNUITY_P                               AMT_ANNUITY_P  0.209484385
New                                                   New  0.157593736
EXT_SOURCE_1                                 EXT_SOURCE_1  0.129613368
REGION_RATING_CLIENT_W_CITY REGION_RATING_CLIENT_W_CITY  0.088085872
AMT_GOODS_PRICE_P                       AMT_GOODS_PRICE_P  0.074037989
OCCUPATION_TYPE_LA                     OCCUPATION_TYPE_LA  0.067851051
ORGANIZATION_TYPE_SL                 ORGANIZATION_TYPE_SL  0.063363132
DEF_60_CNT_SOCIAL_CIRCLE           DEF_60_CNT_SOCIAL_CIRCLE  0.057951735
DAYS_ID_PUBLISH                           DAYS_ID_PUBLISH  0.043072326
AMT_APPLICATION_P                       AMT_APPLICATION_P  0.039947714
NCreditCard                                   NCreditCard  0.038990875
CREDIT_ACTIVE                               CREDIT_ACTIVE  0.030270159
WALLSMATERIAL_MODE_P                 WALLSMATERIAL_MODE_P  0.020813521
HOUSETYPE_MODE_F                         HOUSETYPE_MODE_F  0.020786563
NAME_CONTRACT_TYPE_DMY               NAME_CONTRACT_TYPE_DMY  0.019789093
ELEVATORS_MODE                             ELEVATORS_MODE  0.015620179
DAYS_LAST_PHONE_CHANGE               DAYS_LAST_PHONE_CHANGE  0.015540279
APARTMENTS_MODE                           APARTMENTS_MODE  0.012111357
NONLIVINGAREA_MODE                     NONLIVINGAREA_MODE  0.011290525
REGION_RATING_CLIENT                 REGION_RATING_CLIENT  0.008292518
OCCUPATION_TYPE_D                       OCCUPATION_TYPE_D  0.006672059
LIVINGAREA_AVG                             LIVINGAREA_AVG  0.003488381
BASEMENTAREA_MODE                       BASEMENTAREA_MODE  0.003433810
DAYS_REGISTRATION                       DAYS_REGISTRATION  0.003100556
REG_CITY_NOT_WORK_CITY               REG_CITY_NOT_WORK_CITY  0.001791467
OCCUPATION_TYPE_CS                     OCCUPATION_TYPE_CS  0.001500150
```

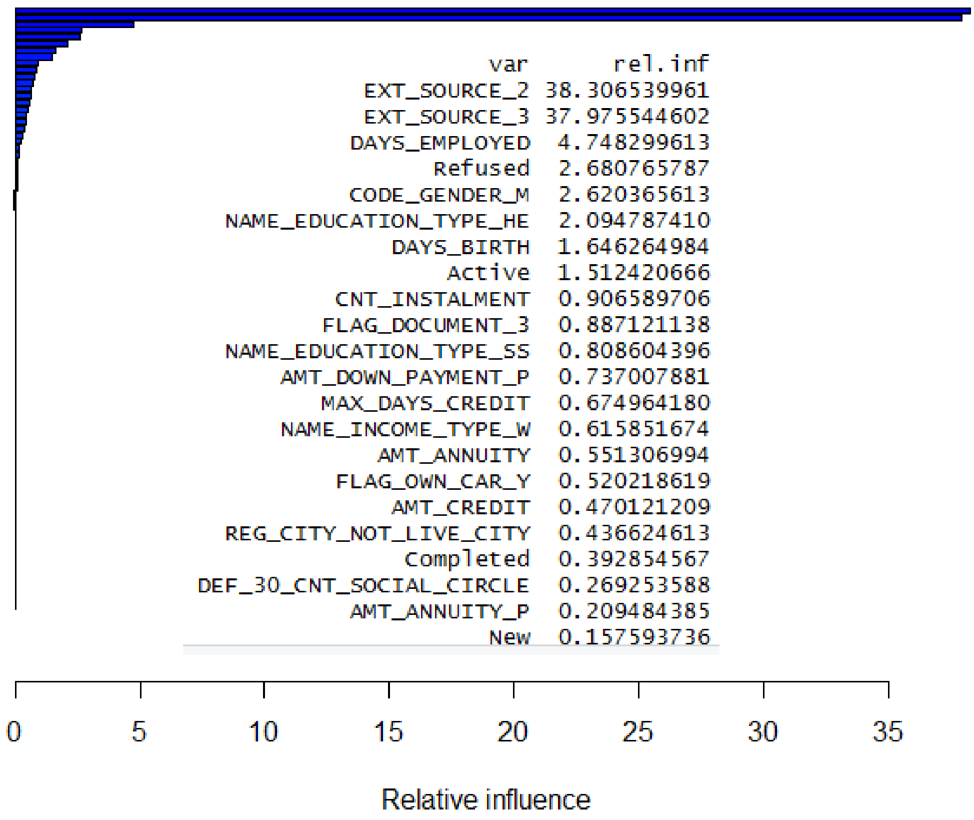

Figure 5 - Relative influence of variables in the gradient boosted model.

Table 7 - Tuned gradient boosted model predictor relative influence.

Stochastic Gradient Boosting 46 predictors

	var	rel.inf
EXT_SOURCE_3	EXT_SOURCE_3	32.18630717
EXT_SOURCE_2	EXT_SOURCE_2	30.84975007
DAYS_EMPLOYED	DAYS_EMPLOYED	4.45410031
Refused	Refused	3.21299809
CODE_GENDER_M	CODE_GENDER_M	2.81190538
DAYS_BIRTH	DAYS_BIRTH	2.63709801
Active	Active	2.42071165
CNT_INSTALMENT	CNT_INSTALMENT	2.24800440
NAME_EDUCATION_TYPE_HE	NAME_EDUCATION_TYPE_HE	2.22122941
AMT_CREDIT	AMT_CREDIT	1.53730508
AMT_ANNUITY	AMT_ANNUITY	1.26181214
MAX_DAYS_CREDIT	MAX_DAYS_CREDIT	1.11375374
CREDIT_ACTIVE	CREDIT_ACTIVE	1.08837529
AMT_DOWN_PAYMENT_P	AMT_DOWN_PAYMENT_P	1.06661283
FLAG_OWN_CAR_Y	FLAG_OWN_CAR_Y	0.99542676
AMT_ANNUITY_P	AMT_ANNUITY_P	0.94147335
FLAG_DOCUMENT_3	FLAG_DOCUMENT_3	0.81753733
DAYS_ID_PUBLISH	DAYS_ID_PUBLISH	0.76378275
NAME_INCOME_TYPE_W	NAME_INCOME_TYPE_W	0.71048466
DEF_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.71040041
REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_LIVE_CITY	0.62435973
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT_W_CITY	0.56499544
NAME_EDUCATION_TYPE_SS	NAME_EDUCATION_TYPE_SS	0.53413896
Completed	Completed	0.52818701
EXT_SOURCE_1	EXT_SOURCE_1	0.51446687

```
ORGANIZATION_TYPE_SL              ORGANIZATION_TYPE_SL  0.37548911
AMT_GOODS_PRICE_P                    AMT_GOODS_PRICE_P  0.35799633
DEF_60_CNT_SOCIAL_CIRCLE      DEF_60_CNT_SOCIAL_CIRCLE  0.35177771
NCreditCard                                NCreditCard  0.32905939
WALLSMATERIAL_MODE_P              WALLSMATERIAL_MODE_P  0.28220151
New                                                New  0.27551563
DAYS_REGISTRATION                    DAYS_REGISTRATION  0.20644043
OCCUPATION_TYPE_LA                  OCCUPATION_TYPE_LA  0.19913001
OCCUPATION_TYPE_D                    OCCUPATION_TYPE_D  0.17974125
NAME_CONTRACT_TYPE_DMY          NAME_CONTRACT_TYPE_DMY  0.16709914
APARTMENTS_MODE                        APARTMENTS_MODE  0.16374744
OCCUPATION_TYPE_CS                  OCCUPATION_TYPE_CS  0.14565051
DAYS_LAST_PHONE_CHANGE          DAYS_LAST_PHONE_CHANGE  0.11933915
AMT_APPLICATION_P                    AMT_APPLICATION_P  0.03159555


Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant at a value of 10
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth
= 3, shrinkage = 0.1 and n.minobsinnode = 10.
```

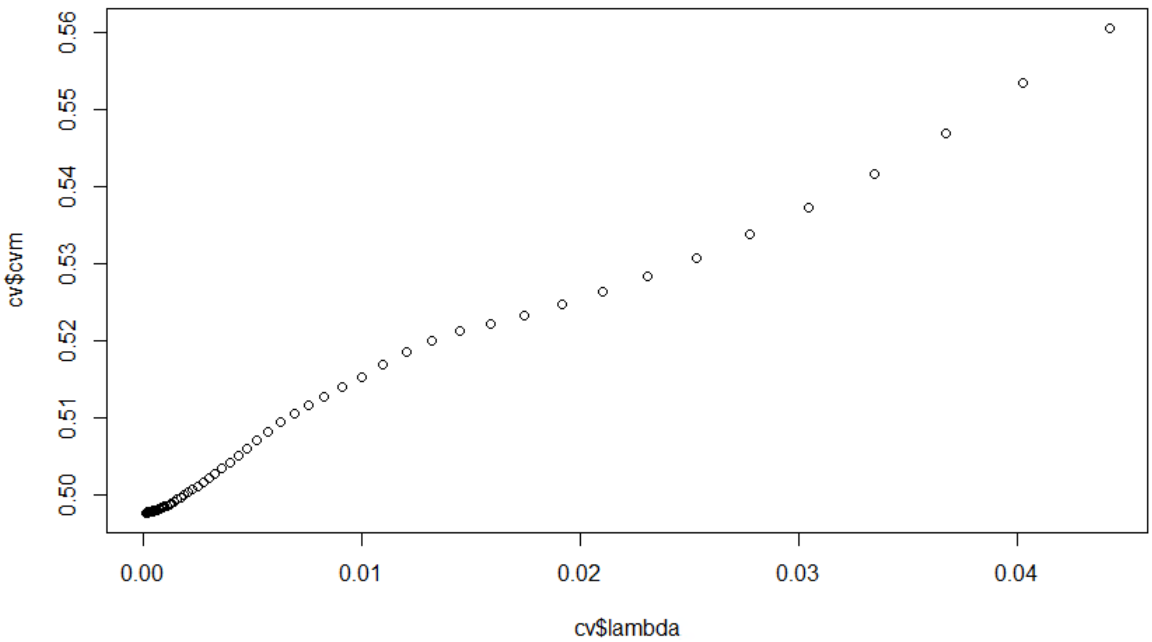

Figure 6 -Selection of lambda for Lasso model.

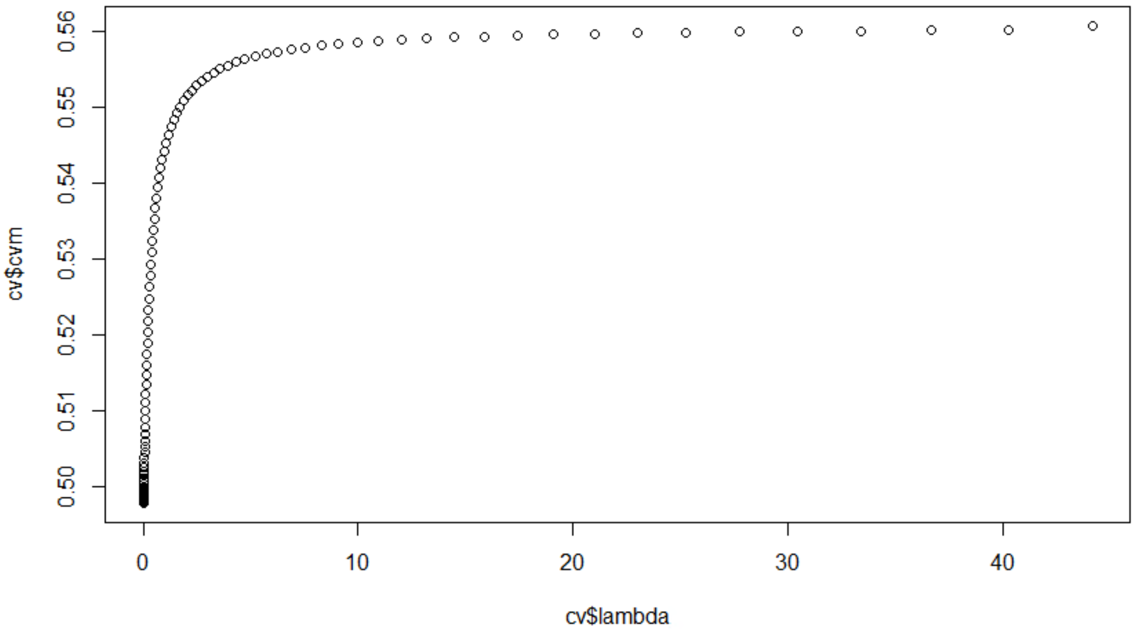

Figure 7 - Selection of lambda for ridge regression.

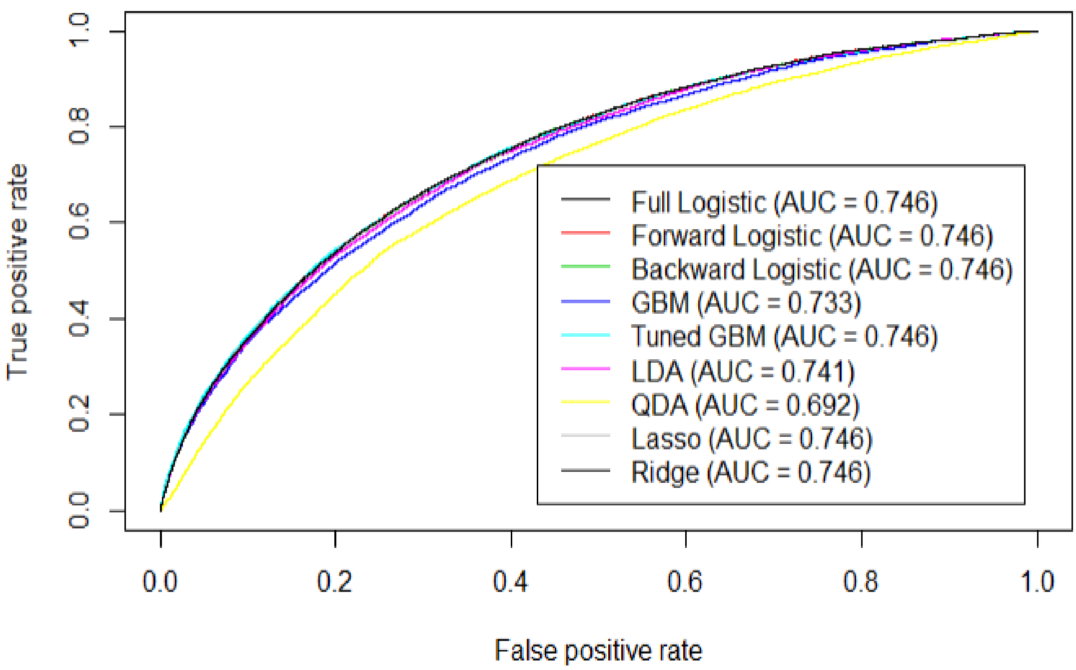

Figure 8 - ROC curves for classification models.

Table 8 - Results comparison of the different modelling techniques.

Model	**Test AUC**	**Comp. Time**
Full Model Logistic Regression	0.746	40sec
Backward Selection Logistic Regression	0.746	3hrs
Forward Selection Logistic Regression	0.746	3hrs
Gradient Boosted Model	0.733	1.2hrs
Tuned Gradient Boosted Model	0.746	3.4hrs
Linear Discriminant Analysis	0.741	24sec
Quadratic Discriminant Analysis	0.692	27sec
Lasso	0.746	5min
Ridge	0.746	9min
Support Vector Machine	NA	48+hrs
Tuned Support Vector Machine	NA	NA

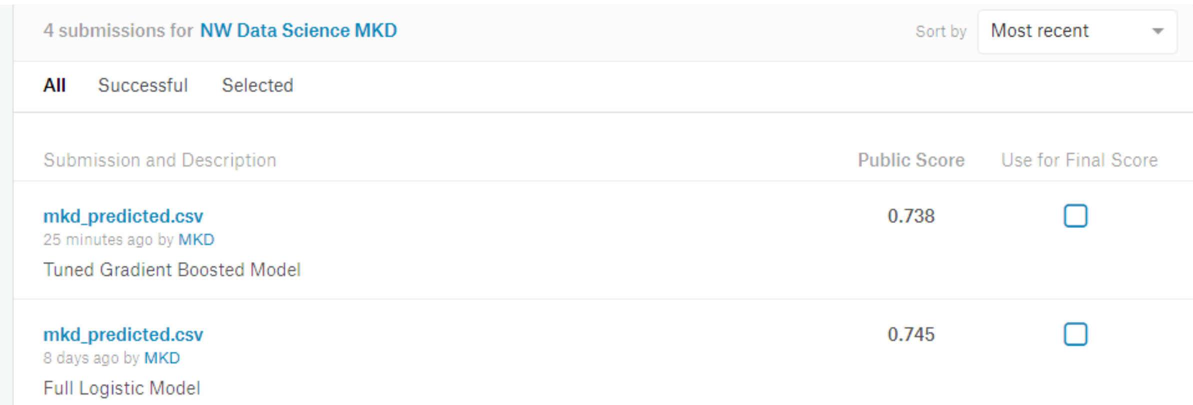

Figure 9 - Kaggle Submissions