Credit Risk Prediction, Customer Segmentation and Roll Rate Analysis

Capstone Final Project Report

Table of Contents

Executive Summary

Credit risk profiling is very important for banks and other lending institutions. Profiling risky segments can reveal useful information for credit risk management. They decide who is creditworthy and who is not based on the individuals/companies demographic information, credit history with the bank (if available) and the information available at Credit Bureaus. Credit providers often collect a vast amount of information on credit users. Information on credit users (or borrowers) often consists of dozens or even hundreds of variables, involving both categorical and numerical data with noisy information.

This is a typical classification project where the company seeks to identify which clients are creditworthy. The project also seeks to provide the following solutions to the company:

- Segment customers such that the company can develop marketing strategies to target specific segments which will increase the potency of the marketing strategy
- Compute roll rates to predict credit losses based on delinquency

The data available to develop the model and achieve other project objectives is as follows:

- Demographic details of customers such as age, income, marital status, family size etc.
- Nature of occupation
- Details of area in which they live
- Information provided by customers in their application

The above information is provided in the main training data. Along with this the company has also provided the following details:

- Details of previous loan applications made to the company
    - Demographic details
    - Type of loan requested for
    - Whether the application was accepted or rejected and reasons for rejection
- List of loans from other financial institutions
- Month on month credit card balance
- Payment history for previous loans at the bank

A. Approach Adopted

a. Preliminary data analysis and data cleaning

Initial analysis revealed that main training data set has over 3 lakh observations with 52 categorical and 69 numeric variables. The original training data set had over 8 lakh missing values. Variables with over 40% of missing values were removed; however due consideration was given to nature and meaning of the data. For others, missing values were imputed. Features with zero variance were dropped and outliers were capped between 5^{th} and 95^{th} percentiles.

b. Feature Selection

Redundant features were identified using tools such as Pearson Correlation and this enabled removal of 18 highly correlated features. To confirm these features Recursive Feature Elimination (RFE) was adopted using Random Forest.

c. Data Visualization

Data visualization was conducted to get a deeper insight into the data. The key findings were as follows:

- Data set is highly imbalanced
- Increased debt ratio leads to higher defaults
- Applicants with previously refused applications tend to have higher default rates
- Individuals staying with parents had higher default rates

B. Model Development

Post cleaning of data, models were developed using Decision Tree (C 5.0) and Logistic Regression approach.

The decision tree algorithm offered the following rules:

- Clients with external rating > 0.08 from source 3, having not more than 4 high interest rate prior loans and having not more than 1 active credit card are very unlikely to default
- Clients with external rating > 0.08 from source 3, having not more than 4 high interest rate prior loans
- Clients with external rating between less than equal to 0.54 from source 3 and having credit card limit between 16k and 18k are highly likely to default

The Logistic Regression model: Highly correlated features were removed and logistic regression model was run. The key variables were as follows:

- Normalized score from external source 2
- Normalized score from external source 3
- Count of credit cards
- Debt ratio at the bureau
- Count of family members

The table below provides the parameters for both models:

	Decision Tree (C5.0)	Logistic Regression
Accuracy	92.22%	84.71%
Error Rate	7.78%	15.29%
Precision / Positive Predictive Value	90.86%	86.71%

Negative Predictive Value	99.65%	74.25%
Recall / Sensitivity / TPR	99.93%	94.63%
FPR	33.50%	48.34%
Specificity / TNR	66.50%	51.66%
AUC	83.22%	85.37%

We recommend that the company proceed with the logistic regression model as the accuracy of the data on entirely unseen new data (different from the train and test data used for modelling) is significantly higher 68.43% vs. 57.24% for the decision tree model. Also, as logistic regression is a probability-based model, it will offer greater flexibility to the organisation i.e. with minimal modification the company can vary the benchmark default risk probability which will be acceptable to them. This will enable the company to manage its strategy while responding to changing economic and industry landscape.

C. Customer Segmentation

Cluster analysis was conducted to identity different customer segments to develop marketing strategies. As the customer information consisted of both categorical and numeric variables, the traditional distance measures i.e. Euclidian, Manhattan etc. could not be applied. Therefore, Gower distance was used to calculate distance. Summary of the two clusters developed is as follows

- Cluster 1 has got higher % of clients with payment difficulties
- Cluster 1 has applicants who are currently employed with average mean income of approx. 31K higher than the applicants in Cluster 2
- Even though Cluster 1 has applicants who are currently employed, the probability of defaulting the loan is higher in this cluster when compared to Cluster 2.
- Cluster 2 has applicants who are mostly pensioners.
- Both the clusters are dominated by female applicants.
- Revolving loan type is applied mostly by applicants in Cluster 1 when compared to those in Cluster 2.
- External Source ratings are higher for clients in Cluster 2
- The average amount of loan credited to Cluster 1 applicants are higher than that of Cluster 2 applicants and Cluster 1 applicants also pledge higher annuity amount

D. Roll Rate Analysis

Roll rate model is a loan level state transition where the probability of transitioning to a new state is dependent on information in current state and does not depend on prior states. It is an effective way to predict future losses based on delinquency. The objective of roll rate analysis is to review overall trends and estimate future performance of loans moving from one state to the other.

Roll rate analysis was conducted for cash loans. The findings of the analysis are as follows

- The delinquency risk is minimal as no accounts go to 90 DPD status

- The credit institution may safely promote cash loan type to its customers with minimal risk of losses due to non-payment

Title & Objective

Project: Credit Risk Prediction, Customer Segmentation & Roll Rate Analysis

Scope and Objectives

The objective of the analysis is to develop a model which helps the lending agency to identify borrowers who have a higher likelihood of repaying a loan. This will help it to broaden the market and boost sales.

To achieve the objective, we shall analyse the borrowing behaviour of a group of customers. The lending agency has provided the following data with respect to the customers.

- List of loans from other financial institutions
- History of previous applications for loans made at the bank
- Credit card balances
- Payment history for previous loans at the bank

Demographic data for customers such as age, income, marital status, education level, employment status etc is also available for the analysis.

We strive to achieve the following objectives through the analysis:

1. Predict if the applicant is a credit risk (how capable is the applicant of repaying a loan):
   Target: 1 (client with 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), 0 (all other cases)
2. Segment customers into clusters basis the demographic data, prior credit requests, kind of credit requirement to devise marketing strategies and enable cross-selling
3. Compute Roll Rates to predict credit losses based on delinquency. Roll rate model is a loan level state transition where the probability of transiting to a new state is dependent on information in current state and does not depend on prior states. The portfolio is classified into distinct and mutually exclusive loan states based on Days Past Due (DPD)

Data Source & Description

The prime and sole source of the data for this project is from Kaggle. The dataset provided has the data from the Credit Bureau as well as the Customer Relationship data available with the bank, viz. past loans, credit cards, repayment history etc.

The data includes several features providing data for the following:

- Income, Education, Family Status, Age, Gender
- Assets owned such as house, car etc.
- Location/Region of living

- Occupation Details with Organization Type
- Credit Details such as Credit Active, Overdue, Amount annuity, Credit Card Balance, Past Loans etc.
- Past Loan application data

Dataset	Description	Observations	Features
application_train	Current application data with our lending company	307511	122
POS_CASH_balance	Previous Point of Sales and Cash loans with the lending c ompany	10001358	8
bureau	Credit Bureau data	1716428	17
bureau_balance	Credit Bureau monthly balance data	27299925	3
previous_application	Prior loan application data with the lending company	1670214	37
installments_payments	Repayment History	13605401	8
credit_card_balance	Credit Card History	3840312	23
application_test	New data to test the model	48744	121

Refer to the **Annexure** for more details on the datasets and the data dictionary.

Statistical Tools & Techniques used

Tools: R, Tableau

Techniques:

- EDA using summary statistics & data visualization (using ggplot2 and Tableau)
- Outlier treatment using univariate approach through boxplots & capping
- Missing value treatment using Hmisc package and manual imputations basis the feature definition (Note: the dataset is huge hence applying predictive techniques for imputation viz. mice, rpart, knn etc. wasn't computationally feasible). Additionally, features with over 40% missing data were dropped
- New Feature creations using dplyr package
- SMOTE (Synthetic Minority Over-sampling Technique) for handling class imbalance (DMwR package)
- Feature selection using Pearson Correlation, Learning Vector Quantization (LVQ), Recursive Feature Elimination (RFE), Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA)
- Clustering using Gower distance & Partitioning Around Medoids (PAM)
- Roll Rate Analysis using Markov Chain algorithm
- Predictive Modeling using various techniques before arriving at a final model (Tree-based viz. C5.0, CART, Random Forest, GBM; Logistic Regression & Naïve Bayes; LDA & SVM; Neural Network). Caret package was used to train models
- ROC (Specificity vs. Sensitivity | FPR vs. TPR), AUC, Predicted Probabilities for model performance & evaluation

Domain: Finance and Risk Analytics

Limitations

- Very large datasets which requires high-end computational powers to analyse and train
- Highly imbalanced class distribution
- Many black-box normalized features

- A very high percentage of missing data

Literature Review

The complete work can be divided into the following logical steps:

Step 1: Preliminary Data Analysis to understand the features & the dataset

Step 2: Data Preparation which involved Data Cleaning (Outlier treatment, Missing value treatment, Reducing/Correcting Categories) and Feature Engineering (To derive new features from existing features to make better sense of the data as well as to help reduce dimensions). This step also involved handling of class imbalance.

Step 3: Data Visualization

Step 4: Feature Selection using various techniques to remove redundant features as well as to rank them to derive important features to use in the final model

Step 5: Creating base models on all features, tune the models to arrive at ideal hyper-parameters and then re-run the model on selected features to achieve improvement as well as to reduce complexity and make the final model explainable

Step 6: Finally test the model on an entirely new & unseen data to evaluate its performance

We have got 7 different datasets and all of them were treated & feature engineered separately first and then merged together and then treated again to finally derive the final dataset for training the model. In order to achieve our credit risk prediction objective, we evaluated various models considering different parameters such as speed, accuracy, stability & explainable. Some of these models tried (viz. SVM, NN) were high on accuracy even on a small sample drawn out of full dataset but computationally not feasible for the complete dataset. Considering the complexity of the problem and its applicability & implementation, we zeroed down on modeling techniques which would be explainable. We evaluated a C5.0 Decision Tree model & Logistic Regression as our final models. While C5.0 helped draw out rules, the Logistic Regression model had far better accuracy on the unseen new data.

For the other objective of segmentation, we used PAM clustering technique, however, applied it on a smaller sample drawn out of the complete dataset to make the clustering algorithm computationally feasible.

For our last objective we leveraged Markov-chain algorithm to perform Roll Rate analysis for Cash Loan accounts as well as Credit Loan accounts. These analysis models are experimental in nature given the complexity of the given datasets and has a scope of improvement if provided with appropriate additional data points.

Data Preparation

The dataset covers all aspects of the lending company's customers:

- Personal Information (Age, Gender, Income, Family Size, Employment etc.)
- Personal Assets (Real Estate, Vehicle etc.)
- Credit Information (Past & Present)
- Environmental (Social Surroundings etc.)

Key notes unravelled through preliminary data analysis:

- Value 365243 denotes infinity in DAYS variables in the datasets, therefore we can consider them NA values
- All amounts in all datasets are in the same currency
- XNA/XAP in categories denote NA values

We start with the main dataset *application_train* and gradually bring in more features from other supporting datasets.

- No. of observations: 307511 new loan applications
- No. of features: 122 (TARGET being the response/classification variable)
- There are 52 categorical features and 69 numerical features, apart from 1 response/class variable.

Refer to Annexure for more details.

The *application_train* dataset was cleaned for:

- Outliers were treated using capping between 5th percentile and 95th percentile
- Missing values:
    - We cannot afford to delete observations as we have an imbalanced dataset and in doing so we might lose out on minority class representation
    - Features with >40% missing data were dropped however due consideration was given to the meaning & significance of these features before dropping, e.g.
        - Further looking into the definition of these features, we realized most of them were basically the normalized information about the building where the client lives. Also, it maybe noted that these features could very well be MNAR (missing not at random) as probably the data gathering process, about the buildings the customers reside in, could either be incomplete or infeasible. Hence, we decided to drop these features and not attempt a meaningless imputation
        - EXT_SOURCE_1 is an important normalized score, hence we did not drop it even though it has got a very high missing %. Same was the case for OCCUPATION_TYPE, EXT_SOURCE_3, ORGANIZATION_TYPE and other such features. We have applied appropriate imputations for these.
        - Rest were imputed appropriately, categorizing them as MAR, MNAR, MCAR
- Features with zero or almost zero variance were dropped
- Incorrect values viz. 365243 days, 'XNA', 'XAP' were marked as NAs and then imputed
- For categorical data, incorrect/invalid levels were corrected and some categories were binned together to reduce the number of levels, e.g. for OCCUPATION_TYPE feature, 'Core staff', 'Cleaning staff', 'Cooking staff', 'Security staff', 'Waiters/barmen staff', 'Medicine staff' were grouped together as Core staff
- Categories with minority observations were either dropped or merged with other categories

[Click here for the complete R code]

Next was Credit Bureau data which comprised of all client's previous credits provided by other financial institutions that were reported to Credit Bureau (*bureau*) along with the monthly balances of these accounts as reported to the Credit Bureau (*bureau_balance*).

- No. of observations: 1716428 bureau accounts with 27299925 monthly balances (for only 817395 bureau accounts)
- No. of features: 20
- There are 6 categorical features and 14 numerical features

These 2 datasets were first cleaned the same way as described above (note that the imputations done were manual basis the feature definition) and then the most recent credit data from *bureau_balance* was merged with the main *bureau* data. Note that every unique customer will have multiple rows in this dataset pertaining to all their credit accounts. Hence, we used feature engineering to derive new features out of the existing ones – mainly we created new features through summary aggregates & ratios to associate with each client. We also created a new feature, RISK_SCORE, which is a weighted risk score calculated basis their monthly statuses as reported in Credit Bureau.

Also, note that there are only 305811 unique client IDs in the bureau data which we can use to associate with the main application data.

[Click here for the complete R code]

Similarly, we cleaned the other datasets – *previous_application, credit_card_balance, POS_cash_balance and instalments_payments* – and derived new features, mainly aggregates & ratios, e.g. payment to instalment ratio for which we desire value >=1; higher the ratio, better is the credit health of the customer.

[Click here for the complete R code]

The final cleaned datasets have the following dimensions:

Dataset	Description	Observations	Features
application_train	Current application data with The bank	307511	74
POS_CASH_balance	Previous Point of Sales and Cash loans with the bank	337252	11
bureau_data	Credit Bureau data	305811	22
previous_application	Prior loan application data with The bank	338857	16
installments_payments	Repayment History	339587	5
credit_card_balance	Credit Card History	103558	18

Next we merge all the datasets and clean the merged dataset and then apply SMOTE to handle the class imbalance (92:08) and achieve 77:23 class balance. Following is the code snippet for SMOTE:

```
### now using SMOTE to create a more "balanced problem"
library(DMwR)
application_train <- SMOTE(TARGET ~ ., application_train, perc.over = 200, perc.under = 500)
```

[Click here for the complete R code]

Following are some of the new features we created:

```
PAYMENT_RATIO_HIST - Actual Payment to Instalment Ratio
CC_BALANCE - Total Credit Card Balance
CC_INTEREST_DUE - Total Interest Due
POS_CANCELED_CNT - Number of Canceled POS/Cash loans
POS_INSTALMENT_FUTURE_CNT - Total Future Installments
PREV_CREDIT_TO_APP_RATIO - Credit Amount to Application Amount Ratio
PREV_REFUSED_CNT - Number of Refused loans
PREV_HIGH_INTEREST_GROUP_CNT - Number of High Interest Rate Group loans
PREV_INSURED_RATIO - Insured to Non-Insured loan ratio
RISK_SCORE – Weighted Risk Score for Bureau Credits
BUREAU_BAD_DEBT_RATIO - Bad Loan Ratio for Bureau loans
BUREAU_ACTIVE_DEBTS_RATIO - % Open debts as per Bureau
BUREAU_TOTAL_AMT_OVERDUE - Total Overdue for Bureau loans
BUREAU_OVERDUE_DEBT_RATIO - Overdue Debt Ratio
```

Refer to **Annexure** for the complete list of new features

Feature Selection

Feature Selection is one of the most crucial step which can help reduce training times significantly and still be able to improve the performance of the algorithm.

We identified redundant features for numerical features using Pearson Correlation and could drop 18 highly correlated features, using 0.75 as the threshold, e.g. Number of Children is correlated with Family size, Pledged Annuity amount is correlated with the Goods price for which the client is seeking a loan.

Next we applied Learning Vector Quantization (LVQ) to rank the features by importance. LVQ is an ANN algorithm. We train a 10-fold repeated cross validation LVQ model and then rank the features by importance using varImp().

Below is the code snippet:

```
# prepare training scheme
control <- trainControl(method="repeatedcv", number=10, repeats=3)
# train the model - Learning Vector Quantization (lvq)
model <- train(TARGET~., data=application_train[,-c(1)], method="lvq", preProcess="scale", trControl=control)
# estimate variable importance
importance <- varImp(model, scale=FALSE)
```

In order to confirm these feature selections reliably we built another model to derive important features. We have used Recursive Feature Elimination (RFE) using Random Forest.

Below is the code snippet:

```
# define the control using a random forest selection function
control <- rfeControl(functions=rfFuncs, method="cv", number=10)
# run the RFE algorithm
results <- rfe(application_train[,-c(1,2)], application_train[,c(2)], sizes=c(3:74), rfeControl=control)
```

By comparing the results of both LVQ and RFE we derive 49 features as important to train our predictive models with.

[Click here for the complete R code]

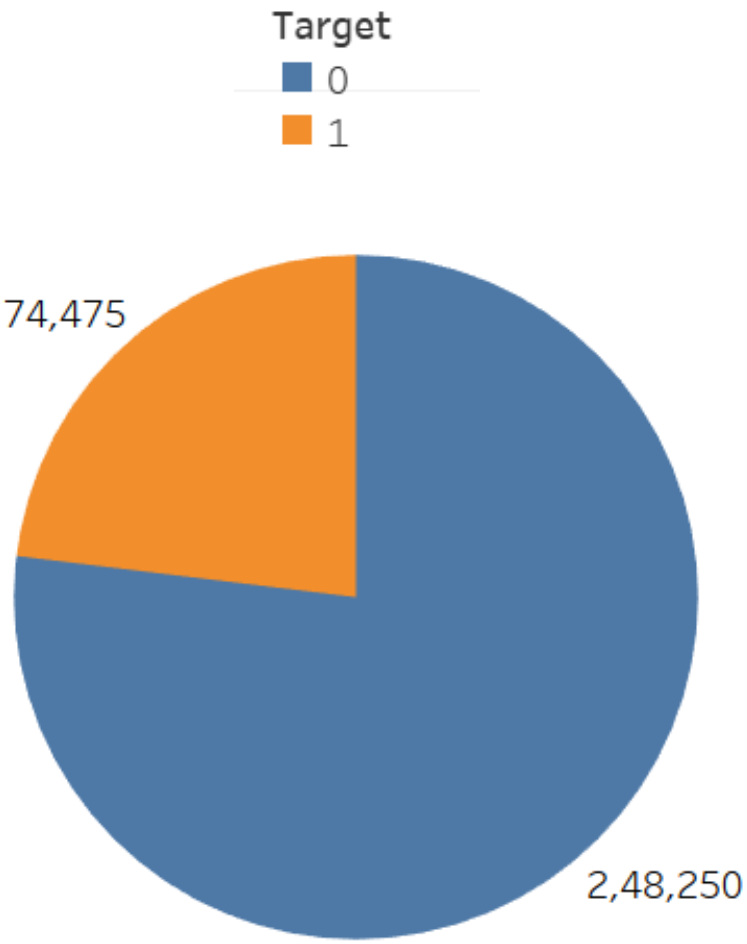
We also attempted PCA with numerical features and MCA with categorical features to evaluate dimensionality reduction. However, PCA components & MCA dimensions didn't actually lead to dimensionality reduction and hence we decided to stick with the features as is.

MCA (Multiple Correspondence Analysis) is an extension of Correspondence Analysis (CA) which is an extension of Principal Component Analysis (PCA) which works on 2 categorical variables. CA uses a contingency table, a table with frequencies. This contingency table provides factor scores (coordinates) and helps evaluate whether there is a significant dependency between the categories. MCA is suited to handle more than 2 categorical variables.

[Click here for the R code]

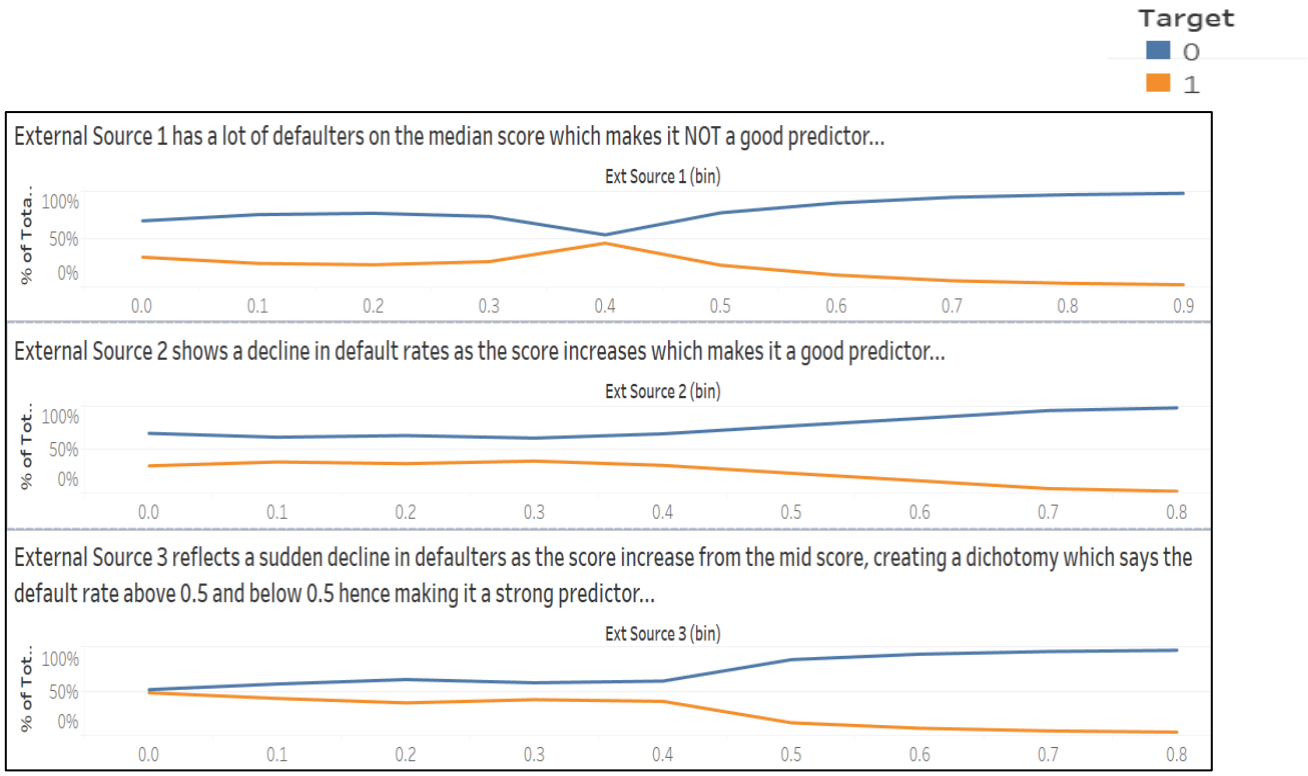
Data Visualizations

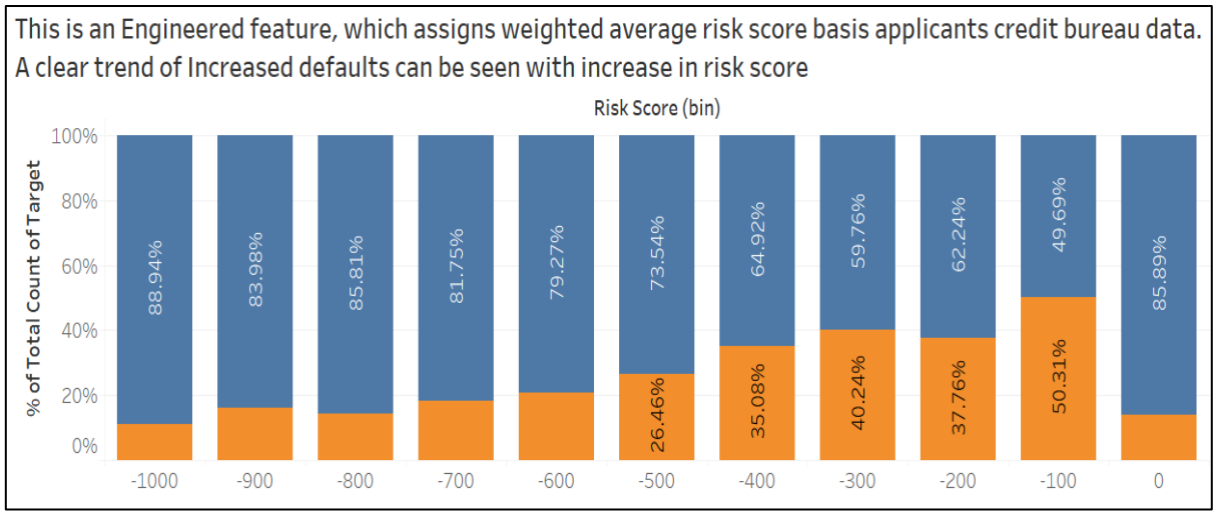
Following are some key visualizations:

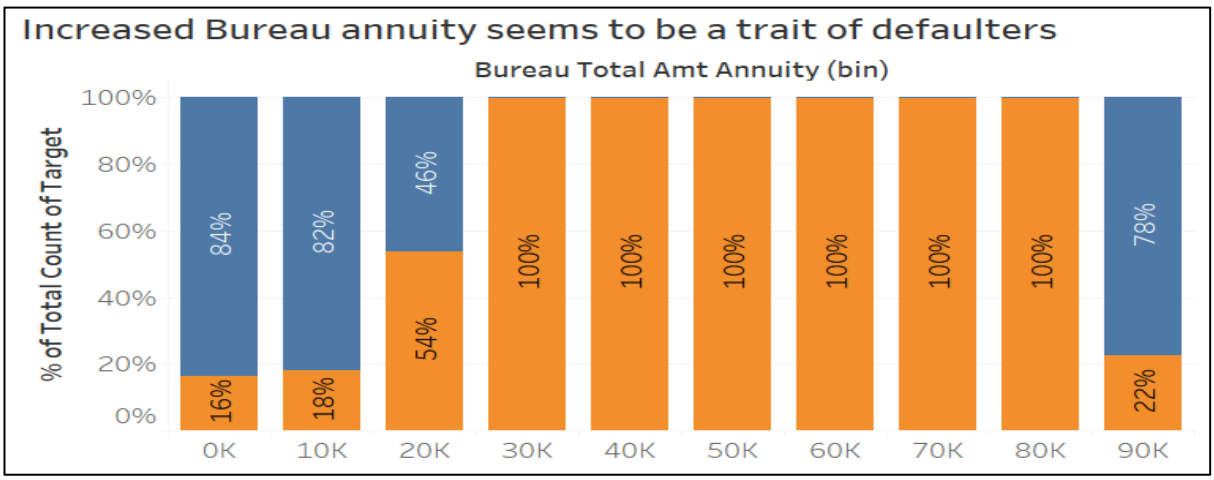


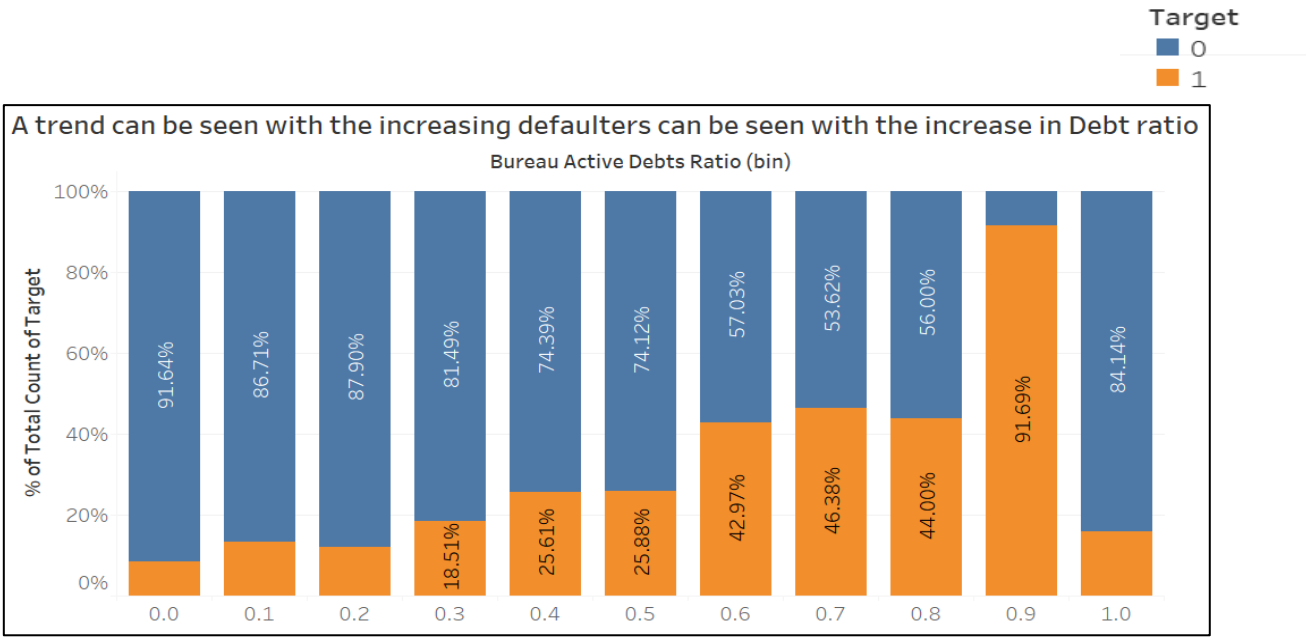
Target = 1 are the ones with payment difficulties

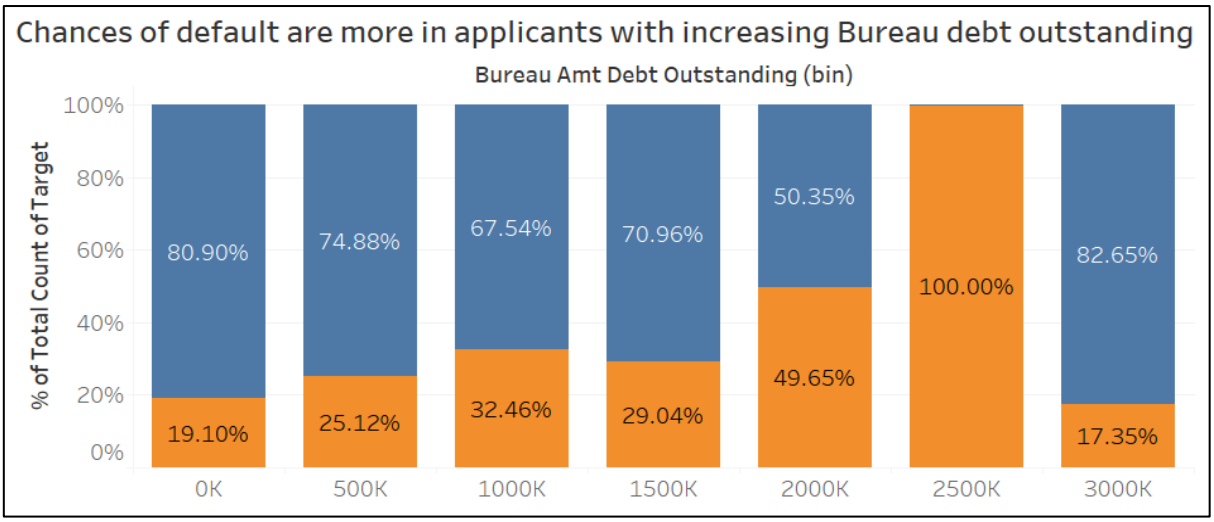
Target = 0 are the ones with non-delinquency in payments

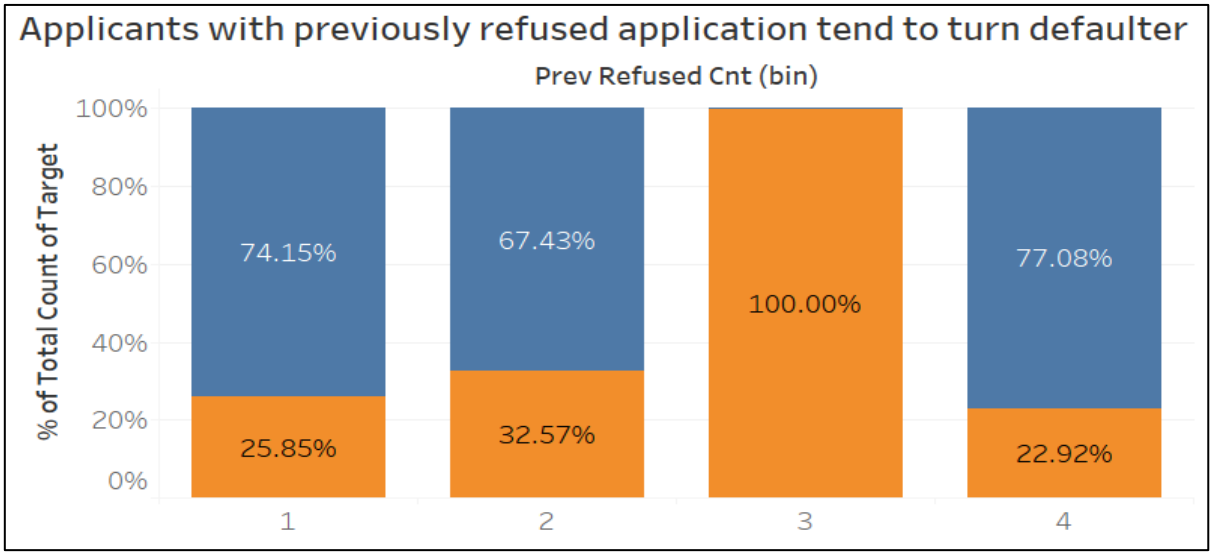

Target
0
1
External Source 1 has a lot of defaulters on the median score which makes it NOT a good predictor...
Ext Source 1 (bin)
% of Tota..
100%
50%
0%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
External Source 2 shows a decline in default rates as the score increases which makes it a good predictor...
Ext Source 2 (bin)
% of Tot..
100%
50%
0%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
External Source 3 reflects a sudden decline in defaulters as the score increase from the mid score, creating a dichotomy which says the default rate above 0.5 and below 0.5 hence making it a strong predictor...
Ext Source 3 (bin)
% of Tot..
100%
50%
0%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

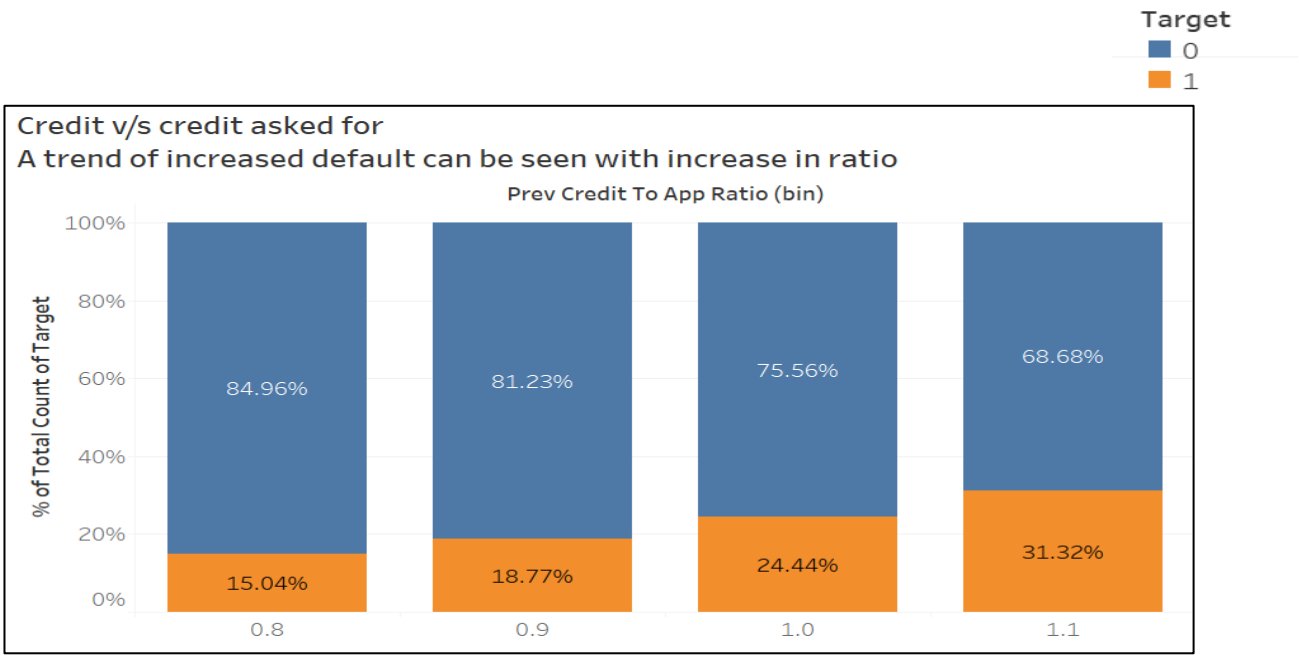

This is an Engineered feature, which assigns weighted average risk score basis applicants credit bureau data.
A clear trend of Increased defaults can be seen with increase in risk score
Risk Score (bin)
% of Total Count of Target
100%
80%
60%
40%
20%
0%
88.94%
83.98%
85.81%
81.75%
79.27%
73.54%
64.92%
59.76%
62.24%
49.69%
85.89%
26.46%
35.08%
40.24%
37.76%
50.31%
-1000
-900
-800
-700
-600
-500
-400
-300
-200
-100
0

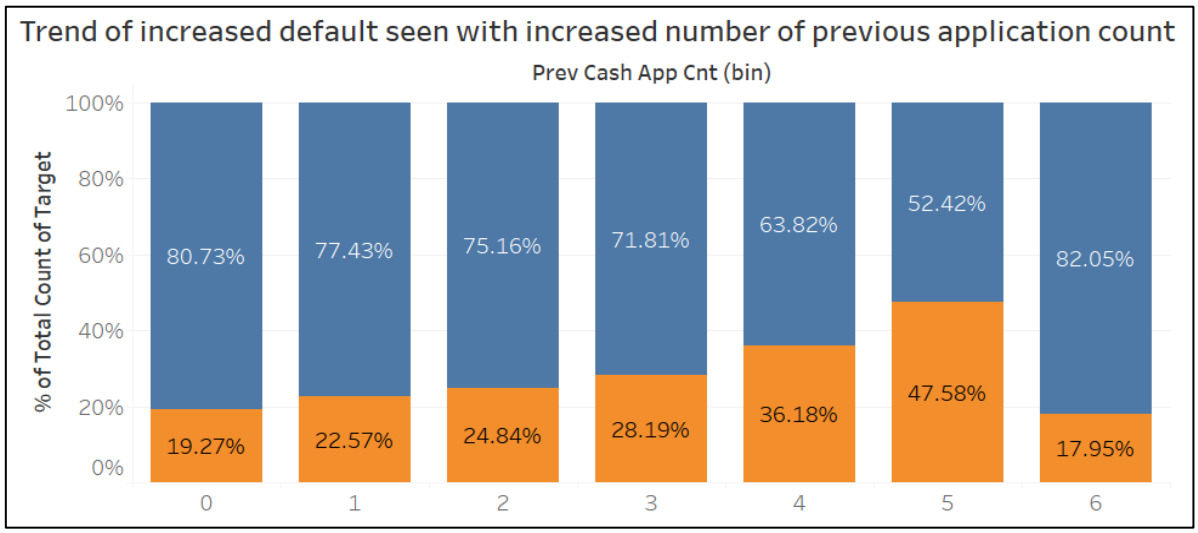

Increased Bureau annuity seems to be a trait of defaulters
Bureau Total Amt Annuity (bin)
% of Total Count of Target
100%
80%
60%
40%
20%
0%
84%
82%
46%
100%
100%
100%
100%
100%
100%
78%
16%
18%
54%
22%
0K
10K
20K
30K
40K
50K
60K
70K
80K
90K

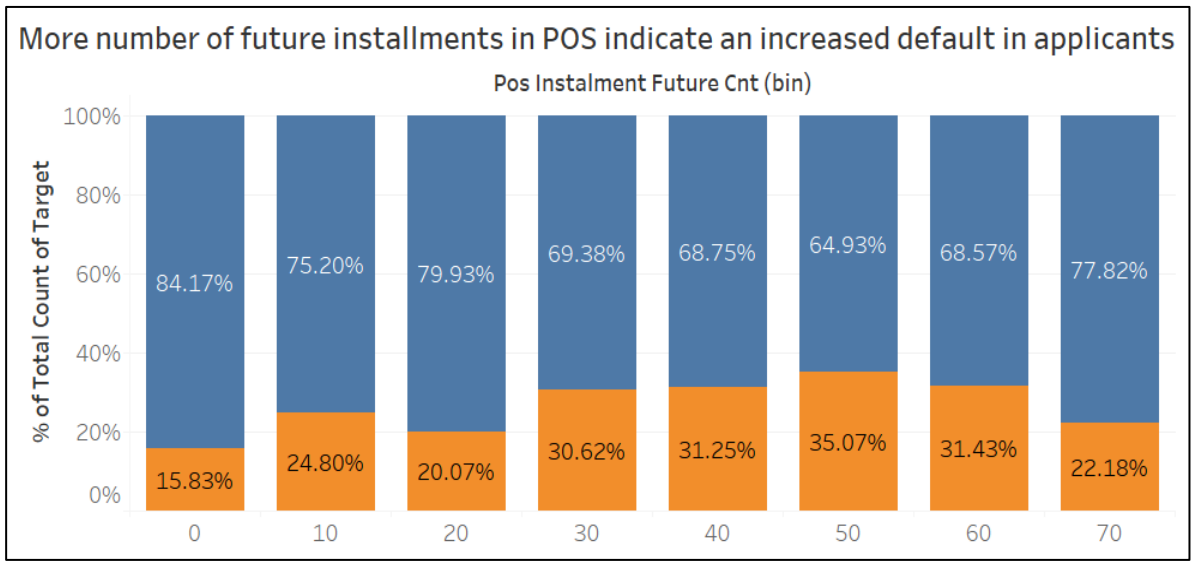

Target
0
1
A trend can be seen with the increasing defaulters can be seen with the increase in Debt ratio
Bureau Active Debts Ratio (bin)
% of Total Count of Target
100%
80%
60%
40%
20%
0%
91.64%
86.71%
87.90%
81.49%
74.39%
74.12%
57.03%
53.62%
56.00%
84.14%
18.51%
25.61%
25.88%
42.97%
46.38%
44.00%
91.69%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

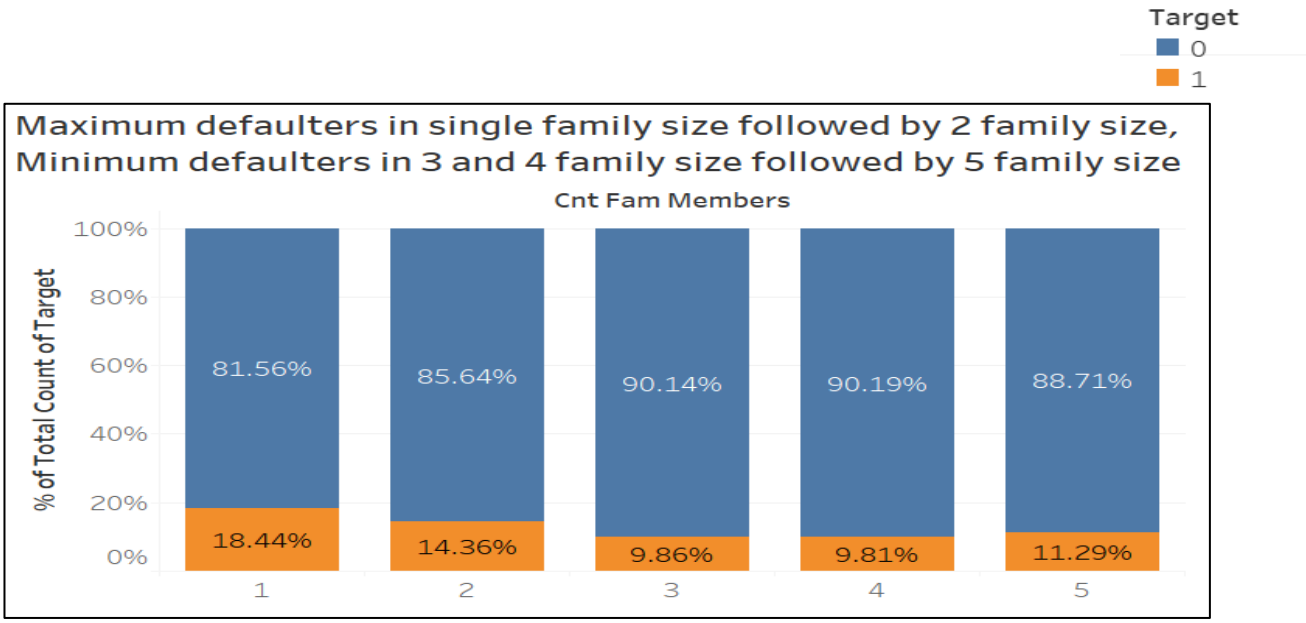

Chances of default are more in applicants with increasing Bureau debt outstanding
Bureau Amt Debt Outstanding (bin)
% of Total Count of Target
100%
80%
60%
40%
20%
0%
80.90%
74.88%
67.54%
70.96%
50.35%
100.00%
82.65%
19.10%
25.12%
32.46%
29.04%
49.65%
17.35%
0K
500K
1000K
1500K
2000K
2500K
3000K

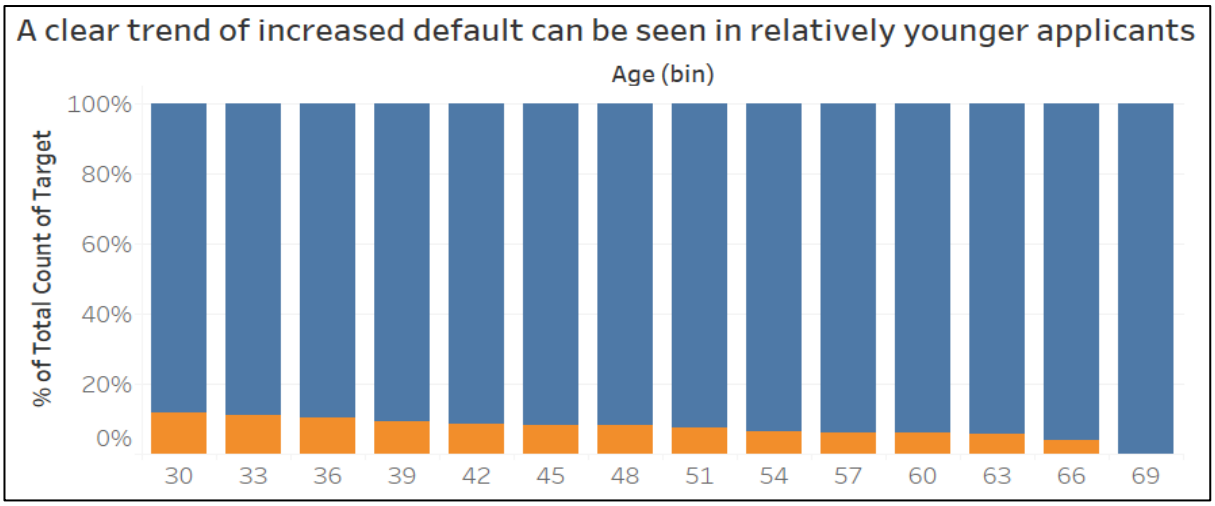

Applicants with previously refused application tend to turn defaulter
Prev Refused Cnt (bin)
% of Total Count of Target
100%
80%
60%
40%
20%
0%
74.15%
67.43%
100.00%
77.08%
25.85%
32.57%
22.92%
1
2
3
4

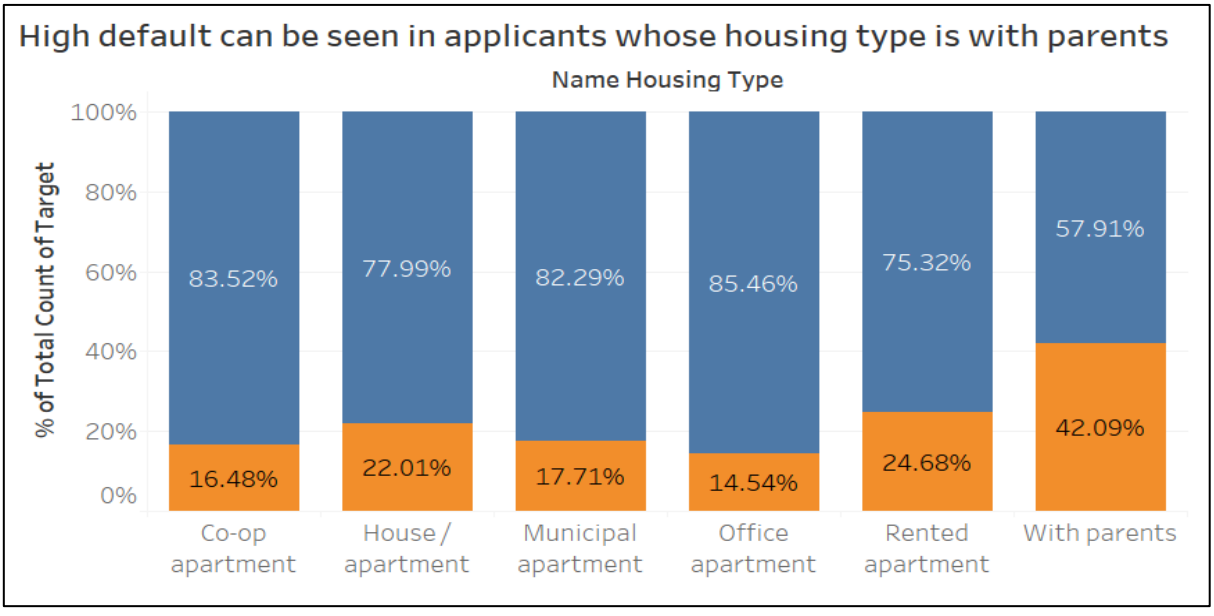

Target
0
1
Credit v/s credit asked for
A trend of increased default can be seen with increase in ratio
Prev Credit To App Ratio (bin)
% of Total Count of Target
100%
80%
60%
40%
20%
0%
84.96%
81.23%
75.56%
68.68%
15.04%
18.77%
24.44%
31.32%
0.8
0.9
1.0
1.1

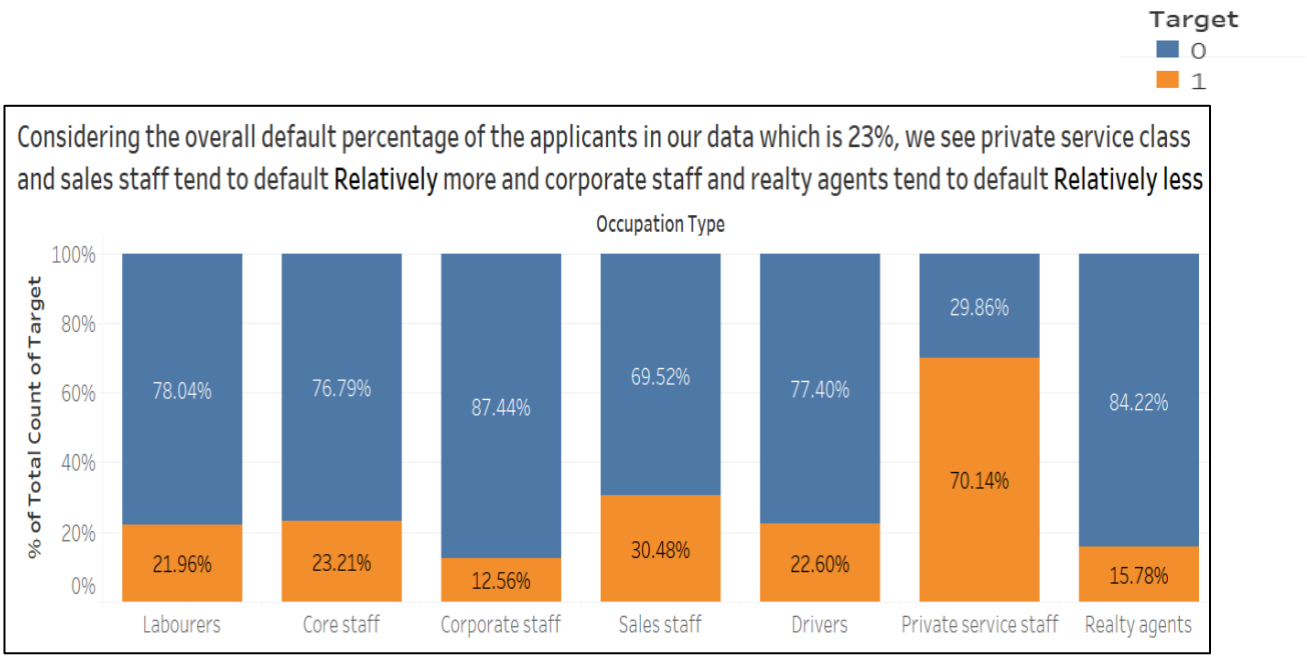

Trend of increased default seen with increased number of previous application count
Prev Cash App Cnt (bin)
% of Total Count of Target
100%
80%
60%
40%
20%
0%
80.73%
77.43%
75.16%
71.81%
63.82%
52.42%
82.05%
19.27%
22.57%
24.84%
28.19%
36.18%
47.58%
17.95%
0
1
2
3
4
5
6

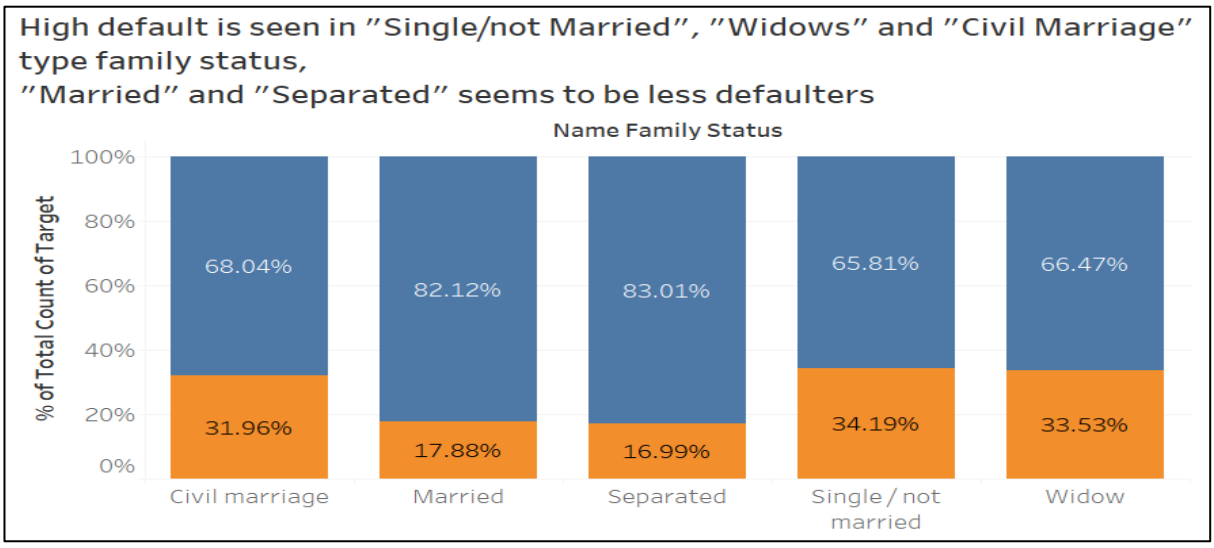

More number of future installments in POS indicate an increased default in applicants
Pos Instalment Future Cnt (bin)
% of Total Count of Target
100%
80%
60%
40%
20%
0%
84.17%
75.20%
79.93%
69.38%
68.75%
64.93%
68.57%
77.82%
15.83%
24.80%
20.07%
30.62%
31.25%
35.07%
31.43%
22.18%
0
10
20
30
40
50
60
70

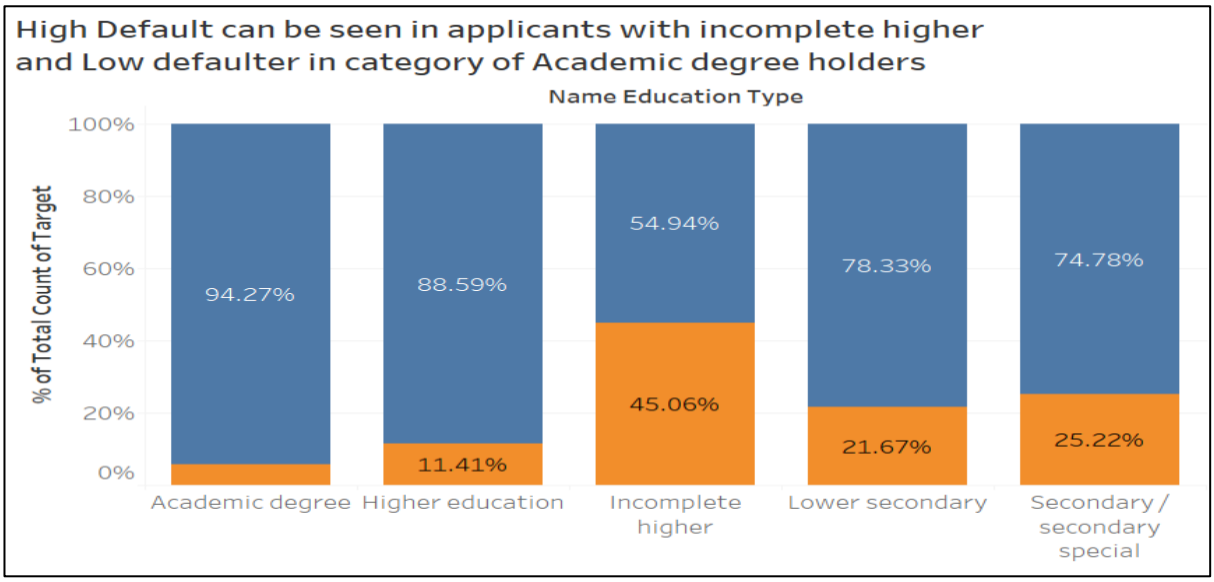

Target
0
1
Maximum defaulters in single family size followed by 2 family size,
Minimum defaulters in 3 and 4 family size followed by 5 family size
Cnt Fam Members
% of Total Count of Target
100%
80%
60%
40%
20%
0%
81.56%
85.64%
90.14%
90.19%
88.71%
18.44%
14.36%
9.86%
9.81%
11.29%
1
2
3
4
5

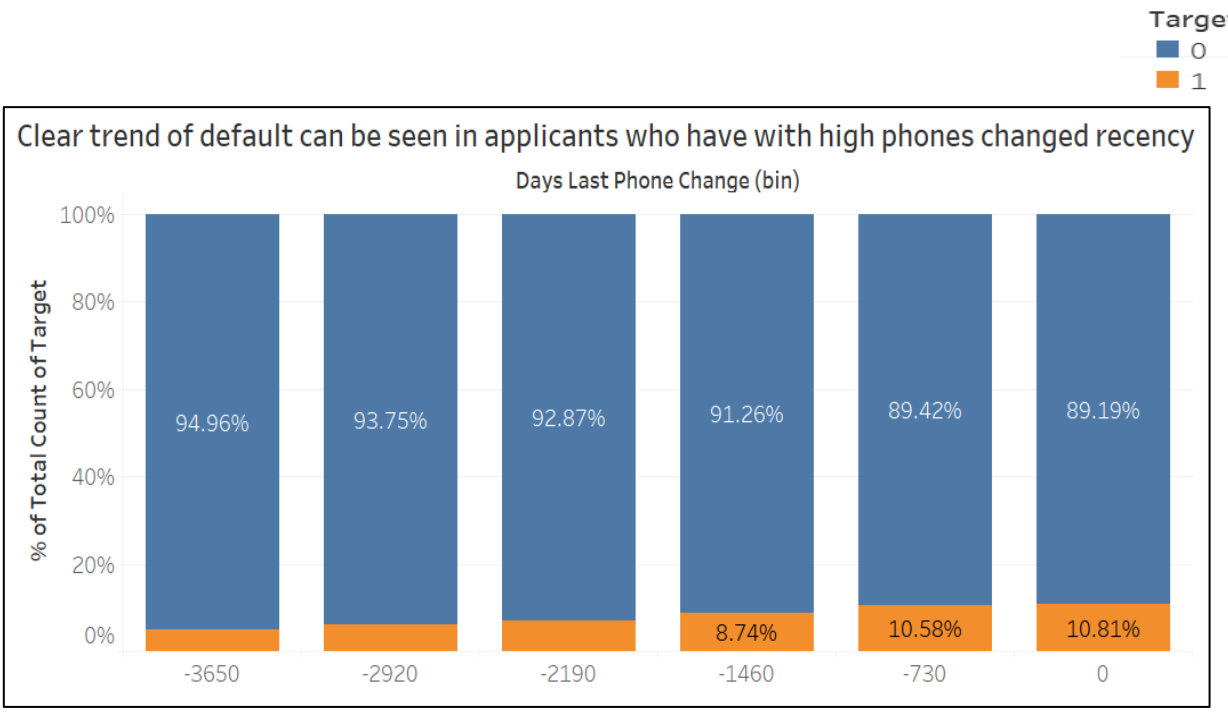

A clear trend of increased default can be seen in relatively younger applicants
Age (bin)
% of Total Count of Target
100%
80%
60%
40%
20%
0%
30
33
36
39
42
45
48
51
54
57
60
63
66
69

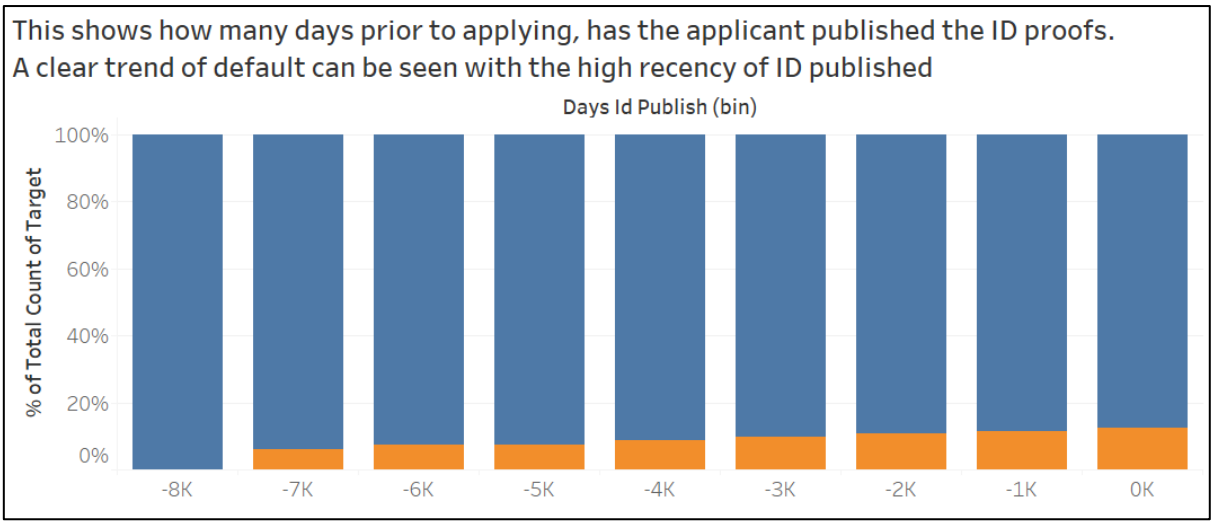

High default can be seen in applicants whose housing type is with parents
Name Housing Type
% of Total Count of Target
100%
80%
60%
40%
20%
0%
83.52%
77.99%
82.29%
85.46%
75.32%
57.91%
16.48%
22.01%
17.71%
14.54%
24.68%
42.09%
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents


Target
0
1
Considering the overall default percentage of the applicants in our data which is 23%, we see private service class and sales staff tend to default Relatively more and corporate staff and realty agents tend to default Relatively less
Occupation Type
% of Total Count of Target
100%
80%
60%
40%
20%
0%
78.04%
21.96%
76.79%
23.21%
87.44%
12.56%
69.52%
30.48%
77.40%
22.60%
29.86%
70.14%
84.22%
15.78%
Labourers
Core staff
Corporate staff
Sales staff
Drivers
Private service staff
Realty agents


High default is seen in "Single/not Married", "Widows" and "Civil Marriage" type family status,
"Married" and "Separated" seems to be less defaulters
Name Family Status
% of Total Count of Target
100%
80%
60%
40%
20%
0%
68.04%
31.96%
82.12%
17.88%
83.01%
16.99%
65.81%
34.19%
66.47%
33.53%
Civil marriage
Married
Separated
Single / not married
Widow


High Default can be seen in applicants with incomplete higher and Low defaulter in category of Academic degree holders
Name Education Type
% of Total Count of Target
100%
80%
60%
40%
20%
0%
94.27%
88.59%
11.41%
54.94%
45.06%
78.33%
21.67%
74.78%
25.22%
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special





[Click here to see the visualisations in Tableau]

[Click here for the density/distribution of the selected features]

Objective 1 - Credit Risk Prediction through Predictive Modeling

After evaluating various models & considering the complexity of the problem and its applicability & implementation, we zeroed down on modeling techniques which would be explainable. We evaluated a C5.0 Decision Tree model & Logistic Regression as our final models. While C5.0 helped draw out rules, the Logistic Regression model had far better accuracy on the unseen new data (different from train and test data used for modeling).

C5.0 Decision Tree Model

We drop the highly correlated features and keep the features selected after Learning Vector Quantization and Recursive Feature Elimination in the dataset. The dataset is split in 70:30 ratio to create training and testing samples distribution of which is shown below.

```
# train - class distribution
table(train_data$TARGET)
```

```
     0      1
173775  52133
```

```
# test - class distribution
table(test_data$TARGET)
```

```
    0     1
74475 22342
```

Model is now built with C5.0 where the number of trials used is 3 and in the control parameter predictor winnowing is enabled so that feature selection will be used during modelling.

```
library(C50)
dtC50 = C5.0(TARGET ~ ., data = train_data[,-c(1)], rules=TRUE, trials=3,
             control = C5.0Control(winnow = TRUE))
summary(dtC50)
```

```
Call:
C5.0.formula(formula = TARGET ~ ., data = train_data[, -c(1)], rules =
 TRUE, trials = 3, control = C5.0Control(winnow = TRUE))


C5.0 [Release 2.07 GPL Edition]        Tue Jan  1 12:31:20 2019
-------------------------------

Class specified by attribute `outcome'

Read 225908 cases (50 attributes) from undefined.data

21 attributes winnowed
```

Output below shows the no of rules derived and the classification error for each trial. Since C5.0 uses adaptive boosting and the line labelled boost shows that. When the weighted voting of all the classifiers are combined, the final predictions have a lower error rate of 7.7%. The final predictions on the train dataset gives accuracy of 92.30%.

```
Evaluation on training data (225908 cases):

Trial           Rules
-----     ----------------
            No      Errors

   0        47 17334( 7.7%)
   1        85 29610(13.1%)
   2        23 18476( 8.2%)
boost         17392( 7.7%)   <<


            (a)    (b)    <-classified as
           -----  -----
          173745     30    (a): class 0
           17362  34771    (b): class 1
```

Attribute usage is as shown below:

```
        Attribute usage:

         98.87% CC_PAYMENT
         96.39% EXT_SOURCE_3
         93.17% BUREAU_CREDIT_CARD_COUNT
         86.67% CC_PAID_INSTALMENT_CNT
         64.01% DAYS_LAST_PHONE_CHANGE
         55.21% DPD_HIST
         49.73% POS_ACTIVE_CNT
         46.47% CC_LIMIT
         46.07% DAYS_EMPLOYED
         42.63% NAME_EDUCATION_TYPE
         39.80% CC_SENT_PROPOSAL_CNT
         35.48% CC_DRAWINGS
         35.32% PREV_HIGH_INTEREST_GROUP_CNT
         34.89% POS_COMPLETED_CNT
         31.56% PREV_MIDDLE_INTEREST_GROUP_CNT
         30.54% BUREAU_ACTIVE_DEBTS_RATIO
         30.17% BUREAU_MAX_AMT_OVERDUE
         29.39% CC_ACTIVE_CNT
         24.34% CNT_FAM_MEMBERS
         23.19% BUREAU_TOTAL_AMT_ANNUITY
         22.92% DBD_HIST
         21.08% PREV_POS_APP_CNT
         17.72% BUREAU_CONSUMER_CREDIT_LOAN_COUNT
         13.79% PREV_LOW_NORMAL_INTEREST_GROUP_CNT
         11.45% CC_SIGNED_CNT
         11.13% DAYS_ID_PUBLISH
          8.43% AMT_REQ_CREDIT_BUREAU_YEAR
          2.65% PREV_UNUSED_CNT
```

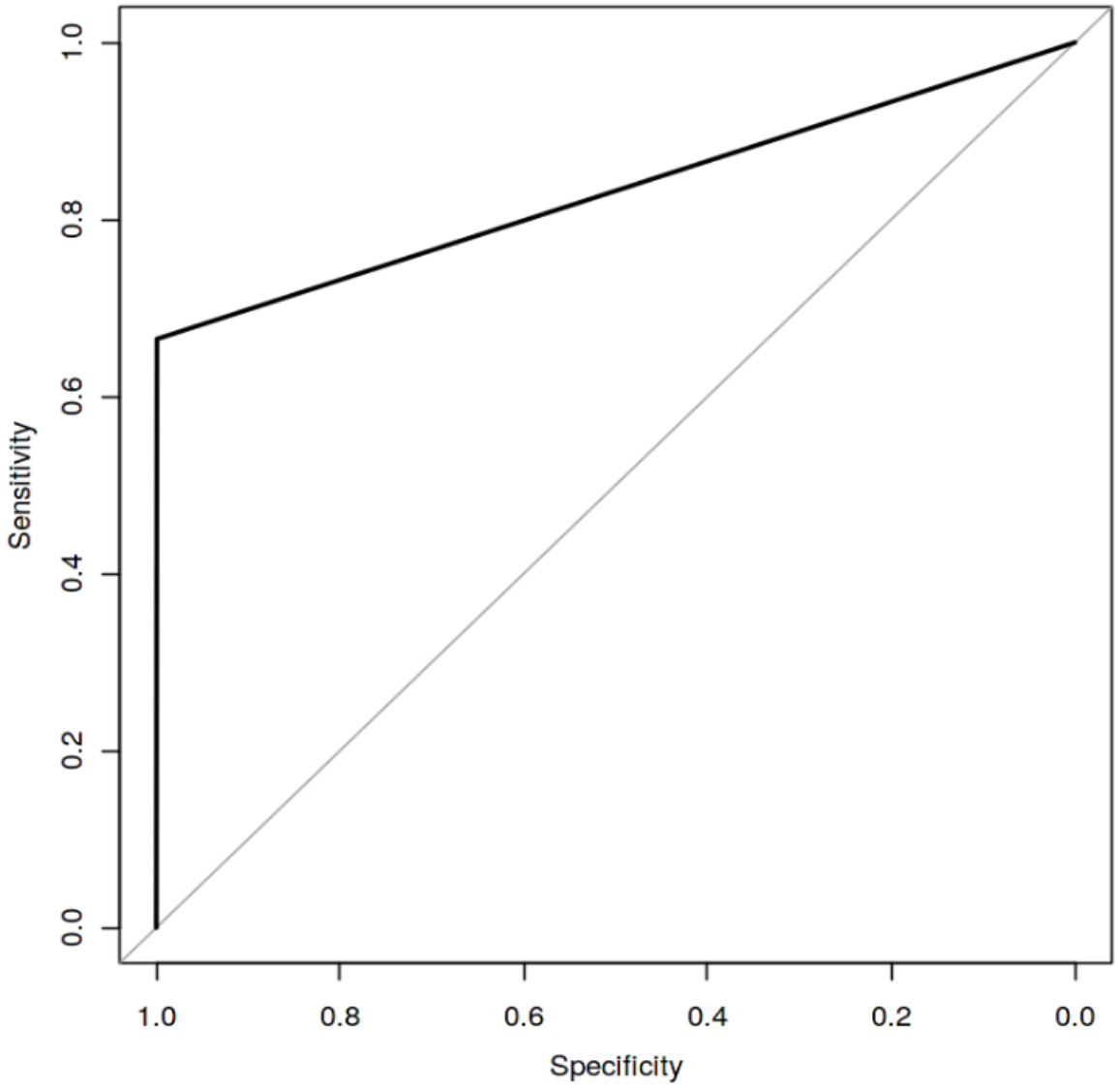
Prediction accuracy on test data is 92.22%
Sensitivity = 99.93%, Specificity = 66.5%

```
             pred_model_test
actual _values     0     1
             0 74423    52
             1  7484 14858
```

92.216243015173

Area under the Receiver Operating Characteristic curve is 0.83 which is considered as “good” in separating defaulters from non defaulters in the system.

0.832163645804101



Let’s now look at some of the rules that have been generated:

Rule 0/1: (780/44, lift 1.2)
EXT_SOURCE_3 > 0.08503369
PREV_HIGH_INTEREST_GROUP_CNT <= 4.007714e-05
CC_ACTIVE_CNT <= 0.4944755
-> class 0 [0.942]

Clients with external rating > 0.08 from source 3, having not more than 4 high interest rate prior loans and having not more than 1 active credit card are very unlikely to default.

```
Rule 0/2: (11815/772, lift 1.2)
        EXT_SOURCE_3 > 0.08503369
        CC_PAYMENT <= 6149.048
        PREV_HIGH_INTEREST_GROUP_CNT <= 4.007714e-05
        CC_DRAWINGS <= 73753
        DBD_HIST > 63.49452
        ->  class 0  [0.935]
```

Clients with external rating > 0.08 from source 3, having not more than 4 high interest rate prior loans, making total credit card payments not exceeding 6149, having not made more than 73753 credit card drawings and making instalment payments in advance by over 60 days are very unlikely to default.

```
Rule 0/34: (3254/13, lift 4.3)
        EXT_SOURCE_3 <= 0.5361825
        CC_LIMIT > 157557
        CC_LIMIT <= 179898.9
        ->  class 1  [0.996]
```

Clients with external rating between less than equal to 0.54 from source 3 and having credit card limit between 16k and 18k are highly likely to default.

```
Rule 0/39: (27, lift 4.2)
        EXT_SOURCE_3 > 0.08503369
        EXT_SOURCE_3 <= 0.3288566
        CC_PAYMENT <= 6149.048
        BUREAU_CREDIT_CARD_COUNT > 1.000342
        CNT_FAM_MEMBERS <= 2.499993
        CC_LIMIT > 22500
        BUREAU_MAX_AMT_OVERDUE > 10517.19
        DBD_HIST <= 63.49452
        NAME_EDUCATION_TYPE in {Incomplete higher,
                Secondary / secondary special}
        ->  class 1  [0.966]
```

Clients with external rating between 0.08 and 0.33 from source 3, making total credit card payments not exceeding 6149, having more than 1 credit card from other banks, with total credit card limit over 22500, having maximum overdue amount above 10.5k, not making advance instalment payments by over 60 days, with family size of at most 2 and having not completed higher education are highly likely to default.

[Click here for the complete R code]

Logistic Regression Model

We drop the highly correlated features and keep the features selected after Learning Vector Quantization and Recursive Feature Elimination in the dataset. The dataset is split in 70:30 ratio to create training and testing samples. Model summary shows most of the selected features as significant.

```
logit_model1<-glm(TARGET~., data = train_data[,-c(1)], family = binomial(link = 'logit'))
summary(logit_model1)
```

```
EXT_SOURCE_3                                          < 2e-16 ***
EXT_SOURCE_2                                          < 2e-16 ***
CC_INTEREST_DUE                                      2.07e-10 ***
CC_PAYMENT                                            < 2e-16 ***
CC_SENT_PROPOSAL_CNT                                  < 2e-16 ***
BUREAU_CREDIT_CARD_COUNT                              < 2e-16 ***
CC_DPD                                               0.048082 *
BUREAU_TOTAL_AMT_ANNUITY                              < 2e-16 ***
CC_SIGNED_CNT                                        0.005392 **
BUREAU_DEBT_CREDIT_RATIO                              < 2e-16 ***
CNT_FAM_MEMBERS                                      2.43e-12 ***
PREV_HIGH_INTEREST_GROUP_CNT                         5.85e-06 ***
CC_ACTIVE_CNT                                         < 2e-16 ***
PREV_REFUSED_CNT                                      < 2e-16 ***
DPD_HIST                                              < 2e-16 ***
CC_DRAWINGS                                           < 2e-16 ***
BUREAU_ACTIVE_DEBTS_RATIO                            0.001237 **
EXT_SOURCE_1                                          < 2e-16 ***
PREV_LOW_NORMAL_INTEREST_GROUP_CNT                    < 2e-16 ***
CC_PAID_INSTALMENT_CNT                               5.33e-15 ***
POS_ACTIVE_CNT                                        < 2e-16 ***
DAYS_BIRTH                                           4.33e-07 ***
CC_LIMIT                                              < 2e-16 ***
BUREAU_AMT_DEBT_OUTSTANDING                           < 2e-16 ***
POS_INSTALMENT_FUTURE_CNT                             < 2e-16 ***
PREV_POS_APP_CNT                                      < 2e-16 ***
POS_COMPLETED_CNT                                     < 2e-16 ***
ORGANIZATION_TYPEBusiness Entity Type 1              4.63e-05 ***
ORGANIZATION_TYPEBusiness Entity Type 2              3.16e-05 ***
ORGANIZATION_TYPEBusiness Entity Type 3              0.805298
ORGANIZATION_TYPEEducation                           4.28e-12 ***
ORGANIZATION_TYPEGovernment                          5.00e-10 ***
ORGANIZATION_TYPEHospitality                         1.41e-05 ***
ORGANIZATION_TYPEIndustrial                           < 2e-16 ***
```

Additionally, the residual deviance is lesser than the null deviance, which again is a good sign:

Null deviance: 244074 on 225907 degrees of freedom
Residual deviance: 167883 on 225830 degrees of freedom
AIC: 168039

Multicollinearity is verified via VIF. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.
We observe below that there is 1 feature 'PREV_CASH_APP_CNT' with VIF>9 and might be problematic. All others are within our threshold.

	GVIF	Df	GVIF^(1/(2*Df))
EXT_SOURCE_3	1.423889	1	1.193268
EXT_SOURCE_2	1.109789	1	1.053465
CC_INTEREST_DUE	1.633940	1	1.278257
CC_PAYMENT	1.543534	1	1.242390
CC_SENT_PROPOSAL_CNT	1.335212	1	1.155514
BUREAU_CREDIT_CARD_COUNT	1.413208	1	1.188784
CC_DPD	1.056208	1	1.027720
BUREAU_TOTAL_AMT_ANNUITY	1.456909	1	1.207025
CC_SIGNED_CNT	1.134393	1	1.065079
BUREAU_DEBT_CREDIT_RATIO	2.047159	1	1.430790
CNT_FAM_MEMBERS	1.778616	1	1.333648
PREV_HIGH_INTEREST_GROUP_CNT	3.628103	1	1.904758
CC_ACTIVE_CNT	1.154968	1	1.074694
PREV_REFUSED_CNT	2.671622	1	1.634510
DPD_HIST	1.443609	1	1.201503
CC_DRAWINGS	1.300987	1	1.140608
BUREAU_ACTIVE_DEBTS_RATIO	1.858570	1	1.363294
EXT_SOURCE_1	1.226545	1	1.107495
PREV_LOW_NORMAL_INTEREST_GROUP_CNT	2.667213	1	1.633160
CC_PAID_INSTALMENT_CNT	1.493907	1	1.222255
POS_ACTIVE_CNT	5.021044	1	2.240769
DAYS_BIRTH	2.234778	1	1.494917
CC_LIMIT	1.229674	1	1.108907
BUREAU_AMT_DEBT_OUTSTANDING	1.937179	1	1.391826
POS_INSTALMENT_FUTURE_CNT	2.678235	1	1.636531
PREV_POS_APP_CNT	4.551360	1	2.133392
POS_COMPLETED_CNT	1.489776	1	1.220564
ORGANIZATION_TYPE	2.166724	14	1.028000
BUREAU_MAX_AMT_OVERDUE	1.165286	1	1.079484
AMT_REQ_CREDIT_BUREAU_YEAR	1.702953	1	1.304972
PREV_CREDIT_TO_APP_RATIO	1.462678	1	1.209412
DBD_HIST	2.441290	1	1.562463
PREV_CASH_APP_CNT	9.132456	1	3.021995
DAYS_EMPLOYED	1.819420	1	1.348859
OCCUPATION_TYPE	2.017651	6	1.060239
PAYMENT_RATIO_HIST	1.360928	1	1.166588
BUREAU_CONSUMER_CREDIT_LOAN_COUNT	1.944266	1	1.394369
PREV_MIDDLE_INTEREST_GROUP_CNT	3.801162	1	1.949657
RISK_SCORE	1.365396	1	1.168502
NAME_HOUSING_TYPE	1.173211	5	1.016103
PREV_CANCELED_CNT	1.752052	1	1.323651
NAME_FAMILY_STATUS	2.093765	4	1.096771
PREV_INSURED_RATIO	1.597319	1	1.263851
PREV_UNUSED_CNT	1.124078	1	1.060225
AMT_INCOME_TOTAL	1.395254	1	1.181209
DAYS_LAST_PHONE_CHANGE	1.219980	1	1.104527
BUREAU_LATEST_CREDIT	1.460554	1	1.208534
DAYS_ID_PUBLISH	1.224207	1	1.106439
NAME_EDUCATION_TYPE	1.253664	4	1.028662

We now check the results of Likelihood Ratio test.

```
library(lmtest)
# Log Likelihood Test
lrtest(logit_model1)
```

#Df	LogLik	Df	Chisq	Pr(>Chisq)
78	-83941.57	NA	NA	NA
1	-122036.90	-77	76190.67	0

The lrtest is performed by estimating two models (one having the predictors and other without it). The null hypothesis is that the smaller model without predictors is the "true" model. Since p- val < 0.05 we reject null hypothesis and conclude that model with predictors fits significantly better than model without predictors.

McFadden's pseudo R-squared ranging from 0.2 to 0.4 indicates good model fit. Here value is 0.31 indicating good fit. The McFadden pseudo R2 value is ~30% indicating that 30% of the uncertainty of intercept only model is explained by the current model.

```
library(pscl)
# McFadden Pseudo RSquare Test
pR2(logit_model1)
```

llh	-83941.5669801484
llhNull	-122036.900463401
G2	76190.6669665062
McFadden	0.312162414307448
r2ML	0.286279705898542
r2CU	0.433399582475578

Odds is determined and shown below for some of the variables. It is interpreted as - For every unit change in a predictor, the log odds of defaulting increases by the odds value for the predictor.

```
# Odds Ratio
odd_model<-exp(coef(logit_model1))
odd_model
```

(Intercept)	0.236228241726096
EXT_SOURCE_3	0.0561054505654263
EXT_SOURCE_2	0.114695851177191
CC_INTEREST_DUE	0.999973697726407
CC_PAYMENT	0.999984263478604
CC_SENT_PROPOSA...	0
BUREAU_CREDIT_CA...	1.19377427997599
CC_DPD	0.792747703591515
BUREAU_TOTAL_AM...	1.00000584594667
CC_SIGNED_CNT	1.74282765836913
BUREAU_DEBT_CRE...	3.1461186686786
CNT_FAM_MEMBERS	0.934452618709895
PREV_HIGH_INTERE...	1.04694069793522
CC_ACTIVE_CNT	3.47190640043217
PREV_REFUSED_CNT	1.06759036654949
DPD_HIST	0.998404977084856
CC_DRAWINGS	0.999996282108527
BUREAU_ACTIVE_DE...	1.0997334585823
EXT_SOURCE_1	0.328651003252209
PREV_LOW_NORMAL...	0.70551376678166
CC_PAID_INSTALME...	0.993117800332415
POS_ACTIVE_CNT	0.766986435004577
DAYS_BIRTH	0.99998890499275
CC LIMIT	0.99999936666879

Model performance on training data is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 164529  24946
         1   9246  27187

               Accuracy : 0.8486
                 95% CI : (0.8472, 0.8501)
    No Information Rate : 0.7692
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5235
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9468
            Specificity : 0.5215
         Pos Pred Value : 0.8683
         Neg Pred Value : 0.7462
             Prevalence : 0.7692
         Detection Rate : 0.7283
   Detection Prevalence : 0.8387
      Balanced Accuracy : 0.7341

       'Positive' Class : 0
```

Area under the Receiver Operating Characteristic curve is 0.85 which is considered as “good” .

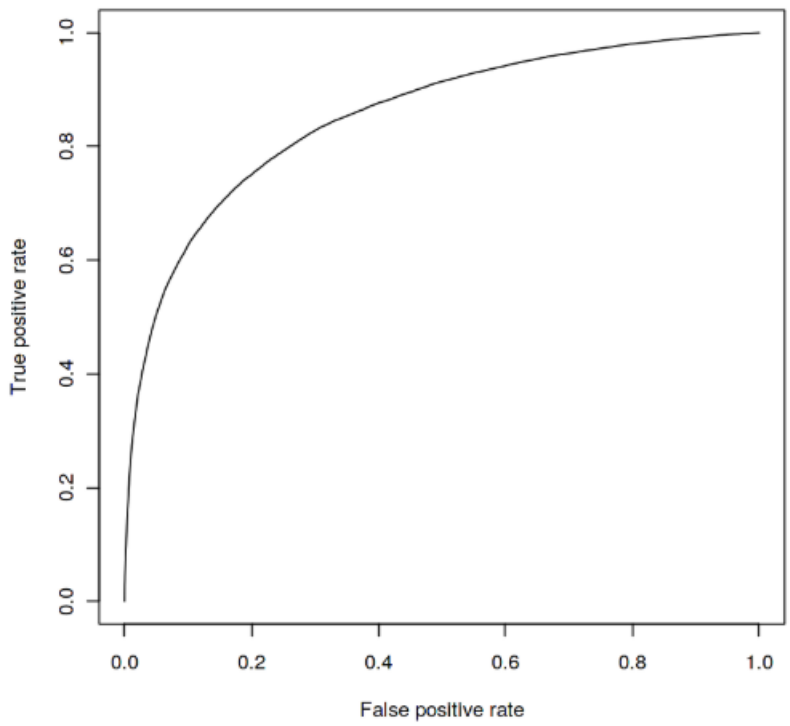
```
0.85355864927913
```

Model performance on test sample is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 70472 10800
         1  4003 11542

               Accuracy : 0.8471
                 95% CI : (0.8448, 0.8494)
    No Information Rate : 0.7692
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.518
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9463
            Specificity : 0.5166
         Pos Pred Value : 0.8671
         Neg Pred Value : 0.7425
             Prevalence : 0.7692
         Detection Rate : 0.7279
   Detection Prevalence : 0.8394
      Balanced Accuracy : 0.7314

       'Positive' Class : 0
```

Area under the Receiver Operating Characteristic curve is 0.85 which is considered as “good” in separating defaulters from non defaulters in the system.

```
0.853726423039066
```



Higher percentage of concordant pairs (85.37%), low percentage of discordant pairs (14.63%) and absence of tied pairs indicate a desirable model.

Of all combinations of actual 1s and 0s, Concordance is the percentage of pairs, whose scores of actual positives are greater than the scores of actual negatives. For a perfect model, this will be 100%. So, the higher the concordance, the better is the quality of model.

```
Concordance(test_data$TARGET, predict_prob)
```

```
$Concordance
0.853726423039034
$Discordance
0.146273576960966
$Tied
0
$Pairs
1663920450
```

[Click here for the complete R code]

Model comparison

	Decision Tree (C5.0)	Logistic Regression
Accuracy	92.22%	84.71%
Error Rate	7.78%	15.29%
Precision / Positive Predictive Value	90.86%	86.71%
Negative Predictive Value	99.65%	74.25%
Recall / Sensitivity / TPR	99.93%	94.63%
FPR	33.50%	48.34%
Specificity / TNR	66.50%	51.66%
AUC	83.22%	85.37%

As mentioned before we tried other models as well and achieved comparable performance and not much improvement, hence, we chose what's optimal for our problem and which is explainable. We also attempted Support Vector Machine (Linear as well as Non-Linear) on a smaller subset of the data as it wasn't computationally possible to run it on the entire dataset. However, the SVM model though trained on a smaller sample did very well on the unseen new data (not the train-test split). Hence, though it was possible to attain a much higher accuracy but we couldn't go for the same due to limitations in computation power.

Refer to **Annexure** for the R code of the other models we tried.

Objective 2 – Customer Segmentation

With the objective of identifying customer segmentation, a small sample was drawn out of full dataset to perform cluster analysis. The chosen sample has the same class balance as the full dataset.

As the data contains both numeric and categorical variables, Gower distance was used to calculate distance. Gower distance basically tell how different two records are. Gower distance falls between 0 and 1. We use daisy() with 'gower' as metric to compute all the pairwise dissimilarities (distances) between observations in the data set.

```
library(cluster)
gowerdist<-daisy(application_train[,-c(1)],metric = "gower") #,type = list(logratio = c(51:54))
summary(gowerdist)
```

```
Warning message in daisy(application_train[, -c(1)], metric = "gower"):
"binary variable(s) 64, 66, 70, 71, 72 treated as interval scaled"


118203000 dissimilarities, summarized :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01691 0.18991 0.22106 0.22178 0.25290 0.48800
Metric :  mixed ;  Types = N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N,
N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, I, I, I, I, I,
I, I, I, I, I, I, I, I, I, I, I, I, I, I, I, I, I, I, I, I, I
Number of objects : 15376
```

To verify the similarity & dissimilarity as calculated above, we display the most similar & most dissimilar pair.

```
## developing matrix to calculate distance between 2 most similar and 2 most dissimilar
gower_mat<-as.matrix(gowerdist)

# Output most similar pair
application_train[
  which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]),
        arr.ind = TRUE)[1, ], ]

# Output most dissimilar pair
application_train[
  which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]),
        arr.ind = TRUE)[1, ], ]
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	NAME_TYPE_
10120	334987	0	Cash loans	F	N	Y	Unaccompan
9299	316285	0	Cash loans	F	N	Y	Unaccompan

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	NAME_TYPE_S
9029	309965	0	Cash loans	M	Y	Y	Family
6407	250309	1	Cash loans	F	N	N	Unaccompanie

As can be observed from above, most of the attributes scores in similar pair are identical whereas non-identical in dissimilar pair.

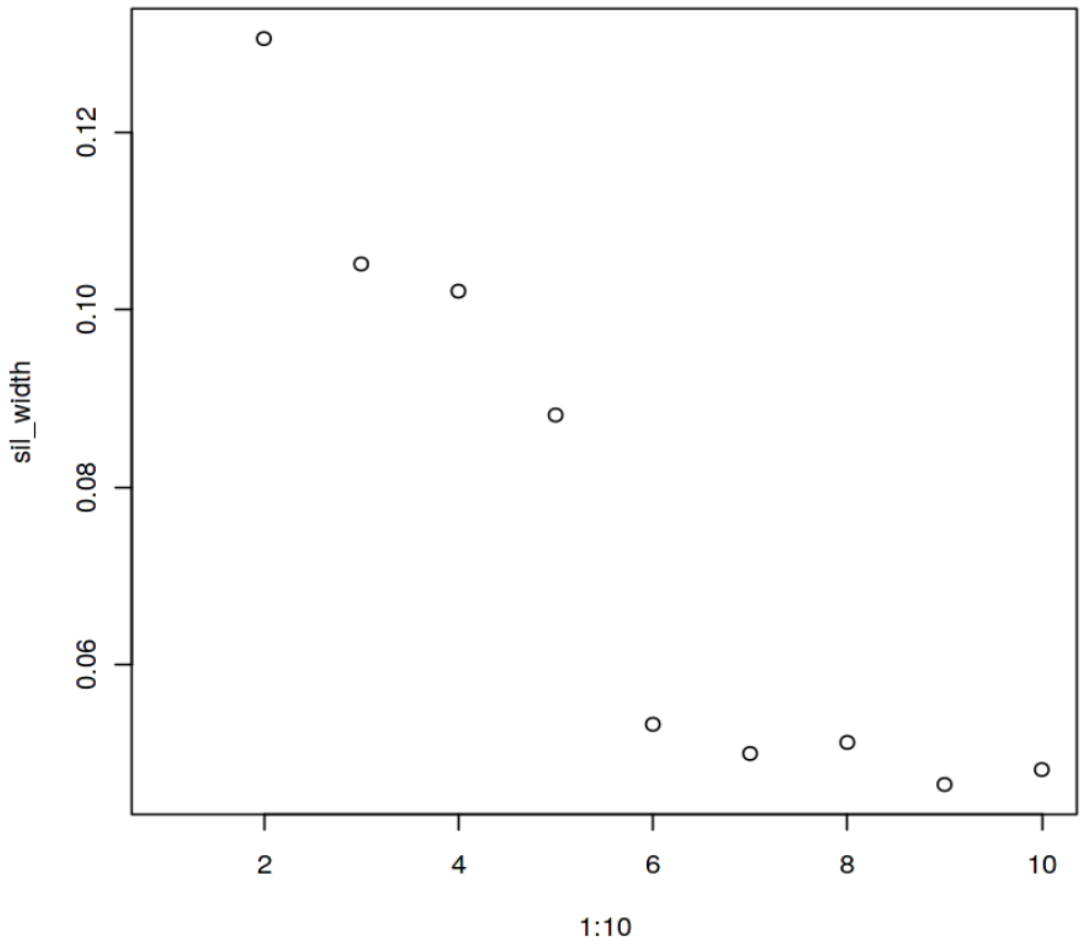
Next, we do clustering using PAM (Partitioning Around Medoids) algorithm. PAM is an iterative k-medoid clustering procedure using the custom distance matrix we provided. This process is similar to the k-means clustering. [Note however that this algorithm is highly compute intensive]

For selecting the optimal number of clusters we use silhouette width which is an aggregated measure of how similar an observation is to its own cluster compared to its closest neighbouring cluster. The value ranges between -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

Silhouette Width was calculated for clusters between 2 to 10 and plotted. We notice that 2 clusters yields the highest value.

```
## calculating silhoutte width for several k

sil_width<-c(NA)

for (i in 2:10) {
    pam_fit<-pam(gowerdist,diss = TRUE, k = i)
    sil_width[i]<-pam_fit$silinfo$avg.width
}
```

```
## Plotting Silhouette width (higher is better)
plot(1:10,sil_width)
```



Clusters were formed corresponding to the highest Silhouette width.

```
## Pick the number of cluster with the highest silhoutte width

pam_fit<-pam(gowerdist,diss = TRUE, k = 2)

mutate(application_train,cluster = pam_fit$clustering)

library(dplyr)

pam_results <- application_train %>%
    mutate(cluster = pam_fit$clustering) %>%
    group_by(cluster) %>%
  do(the_summary = summary(.))

pam_results$the_summary
```

Quick summary of the two clusters:

- Cluster 1 has got higher % of clients with payment difficulties
- Cluster 1 has applicants who are currently employed with average mean income of approx. 31K higher than the applicants in Cluster 2
- Even though Cluster 1 has applicants who are currently employed, the probability of defaulting the loan is higher in this cluster when compared to Cluster 2.
- Cluster 2 has applicants who are mostly pensioners.
- Both the clusters are dominated by female applicants. However, this is more predominant in Cluster 2 with more than 80% female applicants out of which close to 20% are widows.
- Cash loan type is mostly applied for in both the clusters. However, Cluster 2 again shows more volume when compared to Cluster 1.
- Revolving loan type is applied mostly by applicants in Cluster 1 when compared to those in Cluster 2.
- Less than 20% of the applicants base form part of Cluster 2. Hence majority of the applicants are from Cluster 1
- External Source ratings are higher for clients in Cluster 2
- The average amount of loan credited to Cluster 1 applicants are higher than that of Cluster 2 applicants and Cluster 1 applicants also pledge higher annuity amount

Medoids for the clusters:

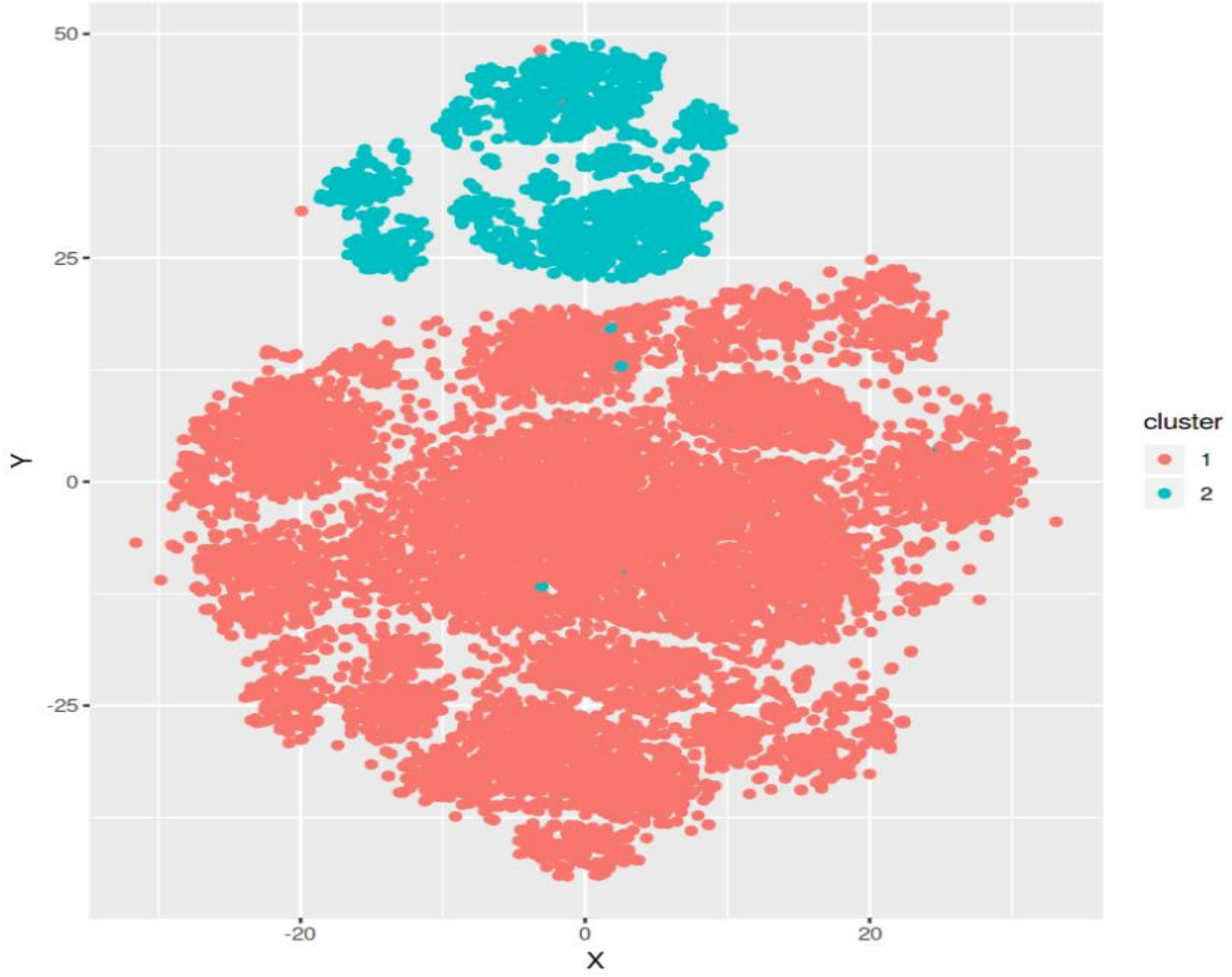
Identifying the medoids for the clusters. Medoids are representative objects of a data set or a cluster with a data set whose average dissimilarity to all the objects in the cluster is minimal. Medoids are similar in concept to means or centroids, but medoids are always restricted to be members of the data set.

```
# The medoids for the clusters
application_train[pam_fit$medoids, ]
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	NAME_TYPE_S
2294	153355	0	Cash loans	F	N	Y	Unaccompanie
6444	251235	0	Cash loans	F	N	Y	Unaccompanie

Cluster visualization: (with 15 features)

```
# Visualization
library(Rtsne)
tsne_obj <- Rtsne(gowerdist, is_distance = TRUE)

tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering),
         feature1 = application_train$TARGET,
         feature2 = application_train$NAME_CONTRACT_TYPE,
         feature3 = application_train$CODE_GENDER,
         feature4 = application_train$FLAG_OWN_CAR,
         feature5 = application_train$FLAG_OWN_REALTY,
         feature6 = application_train$NAME_TYPE_SUITE,
         feature7 = application_train$NAME_INCOME_TYPE,
         feature8 = application_train$NAME_EDUCATION_TYPE,
         feature9 = application_train$NAME_FAMILY_STATUS,
         feature10 = application_train$NAME_HOUSING_TYPE,
         feature11 = application_train$OCCUPATION_TYPE,
         feature12 = application_train$CNT_CHILDREN,
         feature13 = application_train$AMT_INCOME_TOTAL,
         feature14 = application_train$DAYS_EMPLOYED,
         feature15 = application_train$CNT_FAM_MEMBERS)

ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster))
```



Viewing first few records for both the clusters:

```
head(tsne_data[tsne_data$cluster=='1',])
```

cluster	feature1	feature2	feature3	feature4	feature5	feature6	feature7	feature8	feature9	feature10	f
1	1	Cash loans	M	N	Y	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	L
1	1	Cash loans	M	N	Y	Unaccompanied	Commercial associate	Secondary / secondary special	Married	House / apartment	L
1	0	Revolving loans	M	N	Y	Unaccompanied	Working	Secondary / secondary special	Separated	Municipal apartment	L
1	0	Cash loans	F	N	Y	Unaccompanied	Commercial associate	Higher education	Widow	House / apartment	L
1	0	Revolving loans	F	N	N	Unaccompanied	Commercial associate	Higher education	Married	House / apartment	L
1	0	Cash loans	F	N	Y	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	L

Cluster 1 – Dominated by employed category

```
head(tsne_data[tsne_data$cluster=='2',])
```

cluster	feature1	feature2	feature3	feature4	feature5	feature6	feature7	feature8	feature9	feature10	featur
2	0	Cash loans	M	Y	Y	Unaccompanied	Pensioner	Secondary / secondary special	Married	House / apartment	Labou
2	0	Cash loans	F	Y	Y	Family	Pensioner	Secondary / secondary special	Married	House / apartment	Labou
2	0	Cash loans	F	Y	N	Unaccompanied	Pensioner	Secondary / secondary special	Civil marriage	House / apartment	Labou
2	0	Cash loans	F	N	Y	Unaccompanied	Pensioner	Secondary / secondary special	Civil marriage	House / apartment	Labou
2	0	Cash loans	F	N	Y	Family	Pensioner	Secondary / secondary special	Widow	House / apartment	Labou
2	0	Cash loans	F	N	N	Family	Pensioner	Secondary / secondary special	Married	House / apartment	Labou

Cluster 2 – Dominated by pensioners category applying cash loans and are mostly females.

[Click here for the complete R code]
[Click here for the complete R code with clustering done on SMOTEd data – insights were the same except that the clusters are well defined on the original dataset]

Objective 3 – Roll Rate Analysis

Roll Rate Model

Roll rate model is a loan level state transition where the probability of transitioning to a new state is dependent on information in current state and does not depend on prior states. It is an effective way to predict future losses based on delinquency.

Objective

Banks closely monitor roll rates and credit loss provisions to gauge the risks of borrowers. Roll rates can also help credit issuers to set underwriting standards based on repayment trends for various types of products and different types of borrowers. Our objective via roll rate analysis is to review overall trends and estimate future performance of loans moving from one state of delinquency to another.

Assumptions

- Analysis is based on significant amount of data. However, it doesn't consider complete data due to computational complexities.
- The Cash Roll Rate has been calculated considering the number of applicants at the start month and it has remained consistent throughout the analysis to review exact account movement without fluctuations due to additional applicants added in successive months.
- For the purpose of model development, the portfolio is classified into 4 distinct and mutually exclusive loan states based on Days Past Due (DPD):
    - Current – Day 1 – Day 29
    - 30 DPD – Day 30 – Day 59
    - 60 DPD – Day 60 – Day 89
    - 90 DPD – Day 90 and above
- As this is for illustrative purposes only where only four roll over periods have been calculated, anything beyond 90 Days have been considered in 90 Days Past Due bucket.

Limitations

- Only Cash loan type has been considered due to limited data for credit loan type.
- Only accounts that were available during month 1, have been considered to calculate % movement in successive past due buckets. This is due to computational limitations owing to huge file size.
- While roll rate transition methodology is not perfect, the tool is well suited for top down overview approach of estimating future performance and assessing the overall health of the portfolio.
- In general, the use of historical roll rate transitions does not account for external risk factors and macro-economic conditions, so roll rate transition matrices may be better suited for short term forecasting.

Roll Rate Computation Process for Cash Loan Type

Cash loan type data was extracted and accounts at the start were considered basis month number corresponding to applicant identity number.

Total cash loan applicants at the start of month 1 was 1633.

Overall movement of accounts from one stage to another was very low. In this particular exercise only 3 accounts out of 1633 actually moved to successive past due buckets.

Below is a roll rate visualization of cash loan type expressed in percentage of total accounts:

Roll Rate Transition Matrix (Percentage of Accounts)				
Period / Bucket	Current	30 DPD	60 DPD	90 DPD
Current	99.999	0.001	0	0
30 DPD	99.999	0	0.0006	0
60 DPD	99.999	0.0006	0	0
90 DPD	100	0	0	0

Once the loan enters terminal state (in this case 90 DPD) it is locked and cannot exit. Hence the probability of transition to any other state is zero. Probability of an account to go from clean to 60 DPD, clean to 90 DPD, and 30 DPD to 90 DPD is zero, since more than one-time period is required for these jumps.

Observations and Recommendations

- Not more than 3 accounts move from one state to another during various review periods.
- The delinquency risk is minimal as no accounts actually go to 90 DPD status.
- The credit institution may safely promote cash loan type to its customers with minimal risk of losses due to non-payment.

Recommendations

- Client scores from External Source 3 and 2 are highly significant in predicting credit risk and hence the lending agency can drop External Source 1 and just rely on Source 3 and 2
- Clients with external rating > 0.08 from source 3, having not more than 4 high interest rate prior loans and having not more than 1 active credit card are very unlikely to default.
- Clients with external rating > 0.08 from source 3, having not more than 4 high interest rate prior loans, making total credit card payments not exceeding 6149, having not made more than 73753 credit card drawings and making instalment payments in advance by over 60 days are very unlikely to default.
- Clients with external rating between less than equal to 0.54 from source 3 and having credit card limit between 16k and 18k are highly likely to default.
- Clients with external rating between 0.08 and 0.33 from source 3, making total credit card payments not exceeding 6149, having more than 1 credit card from other banks, with total credit card limit over 22500, having maximum overdue amount above 10.5k, not making advance instalment payments by over 60 days, with family size of at most 2 and having not completed higher education are highly likely to default.
- Credit Bureau data is again very significant and hence before taking any action on credit applications, key features from Bureau like Annuity Amount, Current Active Debts, Current Outstanding on other existing credits from other lenders should be taken into account
- The lending agency can de-prioritize data collection for client's residence quality
- The lending agency should give priority to the past application history of its clients
- Younger/Unmarried applicants should go under more scrutiny in contrast to older applicants viz. pensioners
- Lending institution may safely support applicants who are pensioners as the risk of default is lower when compared to Working group
- Clients changing their IDs and contact details very frequently should be scrutinized more
- Few of the key attributes the lending institution may consider are: Credit card payment history, previous interest rate payment, credit card drawings, past due history, credit card limit, type of loan, family members and working status of the applicant etc.
- The customers segmentation is recommended basis two mutually exclusive groups:
    - Working Class
    - Pensioners
- Finally, for the predictive modelling we recommend Logistic Regression as it is a probability-based model, it will offer greater flexibility to the organisation i.e. with minimal modification, the company can vary the benchmark default risk probability which will be acceptable to them. We back this recommendation with the fact that when the model was tried on totally unseen & new data it performed very well

Conclusions

This project successfully met the 3 core objectives identified at the project start:

- We were able to successfully model two solutions to predict loan repayment capability of an applicant. Two model proposed are basis Decision Tree and Logistic Regression with Logistic Regression being the preferred one

- We were also able to identify the key customer segmentation basis the data availability to promote safe lending

- Lastly we were also able to prepare a transition matrix to calculate the probability of cash loan type transition to another loan level state.

References and Bibliography

References:

- Kaggle
- Investopedia
- Analytics India Mag
- Research Gate
- Analytics Vidhya
- R bloggers
- STHDA
- r-statistics.co

Bibliography:

- Machine Learning Algorithms By Giuseppe Bonaccorso (Packt)
- Machine Learning with R By Brett Lantz (Packt)
- Statistics for Machine Learning By Pratap Dangeti (Packt)
- Practical Data Analysis By Hector Cuesta (Packt)
- Learning Predictive Analytics with R By Eric Mayor (Packt)
- Applied Predictive Analytics By Dean Abbott (Wiley)
- Statistics for Data Science By James D. Miller (Packt)
- Practical Data Science with R (Manning)
- R for Data Science (O'Reilly)

Definitions/Abbreviations:

- **Roll Rates** - refers to the percentage of credit users who become increasingly delinquent on their accounts. The roll rate is the percentage of credit users who "roll" from the 60-days late to the 90-days late category, or from the 90-days late to the 120-days late category, and so on. (Source: Investopedia)
- **Revolving Loans** - Loans which allow customers withdraw, repay and again redraw any number of times are called revolving loans. The balance on the revolving loan varies between zero and the maximum allowed value. Given the high degree of flexibility, such loans have a higher interest rate as compared to other loans
- **Missing Completely at Random (MCAR):** This means that the nature of the missing data is not related to any of the variables, whether missing or observed.
- **Missing at Random (MAR):** This means that the nature of the missing data is related to the observed data but not the missing data.
- **Missing Not at Random (MNAR):** This is also known as non-ignorable because the missingness mechanism cannot be ignored. They exist when the missing values are neither MCAR or MAR.
- **SMOTE (Synthetic Minority Over-sampling Technique):** Creating Synthetic Minority classes by randomly sampling from the feature set
- **LVQ (Learning Vector Quantization):** LVQ is an ANN technique. The LVQ algorithm allows one to choose the number of training instances to undergo and then learns about what those instances look like
- **RFE (Recursive Feature Elimination):** RFE is a simple backwards selection algorithm of predictors based on predictor importance ranking. The predictors are ranked and the less important ones are sequentially eliminated
- **MCA (Multiple Correspondence Analysis):** MCA is an extension of CA (Correspondence Analysis) which works like PCA (Principal Component Analysis) on mixed predictor types (numerical + categorical)
- **Gower distance:** For each variable type, a particular distance metric that works well for that type is used and scaled to fall between 0 and 1. Manhattan distance is used for quantitative & ordinal type and for nominal features variables of k categories are first converted into k binary columns and then the Dice coefficient is used
- **Silhouette score & plot** displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters

- **PAM (Partitioning Around Medoids):** Partitions (Clusters) data into k clusters around medoids; works on a custom distance matrix and mixed predictor data types

Annexures

A brief summary of the dataset is as follows:

category	1	application_train.csv	SK_ID_CURR	ID of loan in our sample
category	2	application_train.csv	TARGET	Target variable (1 - client with 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, 0 - all other cases)
category	3	application_train.csv	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
category	4	application_train.csv	CODE_GENDER	Gender of the client
category	5	application_train.csv	FLAG_OWN_CAR	Flag if the client owns a car
category	6	application_train.csv	FLAG_OWN_REALTY	Flag if client owns a house or flat
numeric	7	application_train.csv	CNT_CHILDREN	Number of children the client has
numeric	8	application_train.csv	AMT_INCOME_TOTAL	Income of the client
numeric	9	application_train.csv	AMT_CREDIT	Credit amount of the loan
numeric	10	application_train.csv	AMT_ANNUITY	Loan annuity
numeric	11	application_train.csv	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
category	12	application_train.csv	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
category	13	application_train.csv	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,Ö)
category	14	application_train.csv	NAME_EDUCATION_TYPE	Level of highest education the client achieved
category	15	application_train.csv	NAME_FAMILY_STATUS	Family status of the client
category	16	application_train.csv	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
numeric	17	application_train.csv	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
numeric	18	application_train.csv	DAYS_BIRTH	Client's age in days at the time of application
numeric	19	application_train.csv	DAYS_EMPLOYED	How many days before the application the person started current employment
numeric	20	application_train.csv	DAYS_REGISTRATION	How many days before the application did client change his registration
numeric	21	application_train.csv	DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
numeric	22	application_train.csv	OWN_CAR_AGE	Age of client's car
category	23	application_train.csv	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)
category	24	application_train.csv	FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)
category	25	application_train.csv	FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)

category	26	application_train.csv	FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)
category	27	application_train.csv	FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)
category	28	application_train.csv	FLAG_EMAIL	Did client provide email (1=YES, 0=NO)
category	29	application_train.csv	OCCUPATION_TYPE	What kind of occupation does the client have
numeric	30	application_train.csv	CNT_FAM_MEMBERS	How many family members does client have
category	31	application_train.csv	REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
category	32	application_train.csv	REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1,2,3)
category	33	application_train.csv	WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
category	34	application_train.csv	HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
category	35	application_train.csv	REG_REGION_NOT_LIVE_REGION	Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
category	36	application_train.csv	REG_REGION_NOT_WORK_REGION	Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
category	37	application_train.csv	LIVE_REGION_NOT_WORK_REGION	Flag if client's contact address does not match work address (1=different, 0=same, at region level)
category	38	application_train.csv	REG_CITY_NOT_LIVE_CITY	Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
category	39	application_train.csv	REG_CITY_NOT_WORK_CITY	Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
category	40	application_train.csv	LIVE_CITY_NOT_WORK_CITY	Flag if client's contact address does not match work address (1=different, 0=same, at city level)
category	41	application_train.csv	ORGANIZATION_TYPE	Type of organization where client works
numeric	42	application_train.csv	EXT_SOURCE_1	Normalized score from external data source
numeric	43	application_train.csv	EXT_SOURCE_2	Normalized score from external data source
numeric	44	application_train.csv	EXT_SOURCE_3	Normalized score from external data source
numeric	45	application_train.csv	APARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

numeric	46	application_train.csv	BASEMENTAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	47	application_train.csv	YEARS_BEGINEXPLUATATION_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	48	application_train.csv	YEARS_BUILD_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	49	application_train.csv	COMMONAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	50	application_train.csv	ELEVATORS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	51	application_train.csv	ENTRANCES_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

numeric	52	application_train.csv	FLOORSMAX_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	53	application_train.csv	FLOORSMIN_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	54	application_train.csv	LANDAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	55	application_train.csv	LIVINGAPARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	56	application_train.csv	LIVINGAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	57	application_train.csv	NONLIVINGAPARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

numeric	58	application_train.csv	NONLIVINGAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	59	application_train.csv	APARTMENTS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	60	application_train.csv	BASEMENTAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	61	application_train.csv	YEARS_BEGINEXPLUATATION_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	62	application_train.csv	YEARS_BUILD_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	63	application_train.csv	COMMONAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

numeric	64	application_train.csv	ELEVATORS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	65	application_train.csv	ENTRANCES_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	66	application_train.csv	FLOORSMAX_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	67	application_train.csv	FLOORSMIN_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	68	application_train.csv	LANDAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	69	application_train.csv	LIVINGAPARTMENTS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

numeric	70	application_train.csv	LIVINGAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	71	application_train.csv	NONLIVINGAPARTMENTS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	72	application_train.csv	NONLIVINGAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	73	application_train.csv	APARTMENTS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	74	application_train.csv	BASEMENTAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	75	application_train.csv	YEARS_BEGINEXPLUATATION_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

numeric	76	application_train.csv	YEARS_BUILD_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	77	application_train.csv	COMMONAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	78	application_train.csv	ELEVATORS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	79	application_train.csv	ENTRANCES_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	80	application_train.csv	FLOORSMAX_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	81	application_train.csv	FLOORSMIN_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

numeric	82	application_train.csv	LANDAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	83	application_train.csv	LIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	84	application_train.csv	LIVINGAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	85	application_train.csv	NONLIVINGAPARTMENTS_ MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	86	application_train.csv	NONLIVINGAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
category	87	application_train.csv	FONDKAPREMONT_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

category	88	application_train.csv	HOUSETYPE_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	89	application_train.csv	TOTALAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
category	90	application_train.csv	WALLSMATERIAL_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
category	91	application_train.csv	EMERGENCYSTATE_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
numeric	92	application_train.csv	OBS_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 30 DPD (days past due) default
numeric	93	application_train.csv	DEF_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 30 DPD (days past due)
numeric	94	application_train.csv	OBS_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 60 DPD (days past due) default
numeric	95	application_train.csv	DEF_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 60 (days past due) DPD
numeric	96	application_train.csv	DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
category	97	application_train.csv	FLAG_DOCUMENT_2	Did client provide document 2
category	98	application_train.csv	FLAG_DOCUMENT_3	Did client provide document 3
category	99	application_train.csv	FLAG_DOCUMENT_4	Did client provide document 4
category	100	application_train.csv	FLAG_DOCUMENT_5	Did client provide document 5
category	101	application_train.csv	FLAG_DOCUMENT_6	Did client provide document 6
category	102	application_train.csv	FLAG_DOCUMENT_7	Did client provide document 7
category	103	application_train.csv	FLAG_DOCUMENT_8	Did client provide document 8

category	104	application_train.csv	FLAG_DOCUMENT_9	Did client provide document 9
category	105	application_train.csv	FLAG_DOCUMENT_10	Did client provide document 10
category	106	application_train.csv	FLAG_DOCUMENT_11	Did client provide document 11
category	107	application_train.csv	FLAG_DOCUMENT_12	Did client provide document 12
category	108	application_train.csv	FLAG_DOCUMENT_13	Did client provide document 13
category	109	application_train.csv	FLAG_DOCUMENT_14	Did client provide document 14
category	110	application_train.csv	FLAG_DOCUMENT_15	Did client provide document 15
category	111	application_train.csv	FLAG_DOCUMENT_16	Did client provide document 16
category	112	application_train.csv	FLAG_DOCUMENT_17	Did client provide document 17
category	113	application_train.csv	FLAG_DOCUMENT_18	Did client provide document 18
category	114	application_train.csv	FLAG_DOCUMENT_19	Did client provide document 19
category	115	application_train.csv	FLAG_DOCUMENT_20	Did client provide document 20
category	116	application_train.csv	FLAG_DOCUMENT_21	Did client provide document 21
numeric	117	application_train.csv	AMT_REQ_CREDIT_BUREAU_HOUR	Number of enquiries to Credit Bureau about the client one hour before application
numeric	118	application_train.csv	AMT_REQ_CREDIT_BUREAU_DAY	Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)
numeric	119	application_train.csv	AMT_REQ_CREDIT_BUREAU_WEEK	Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)
numeric	120	application_train.csv	AMT_REQ_CREDIT_BUREAU_MON	Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
numeric	121	application_train.csv	AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
numeric	122	application_train.csv	AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)

category	123	bureau.csv	SK_ID_CURR	ID of loan in our sample - one loan in our sample can have 0,1,2 or more related previous credits in credit bureau
category	124	bureau.csv	SK_ID_BUREAU	Recoded ID of previous Credit Bureau credit related to our loan (unique coding for each loan application)
category	125	bureau.csv	CREDIT_ACTIVE	Status of the Credit Bureau (CB) reported credits
category	126	bureau.csv	CREDIT_CURRENCY	Recoded currency of the Credit Bureau credit
numeric	127	bureau.csv	DAYS_CREDIT	How many days before current application did client apply for Credit Bureau credit
numeric	128	bureau.csv	CREDIT_DAY_OVERDUE	Number of days past due on CB credit at the time of application for related loan in our sample

numeric	129	bureau.csv	DAYS_CREDIT_ENDDATE	Remaining duration of CB credit (in days) at the time of application in Home Credit
numeric	130	bureau.csv	DAYS_ENDDATE_FACT	Days since CB credit ended at the time of application in Home Credit (only for closed credit)
numeric	131	bureau.csv	AMT_CREDIT_MAX_OVERDUE	Maximal amount overdue on the Credit Bureau credit so far (at application date of loan in our sample)
numeric	132	bureau.csv	CNT_CREDIT_PROLONG	How many times was the Credit Bureau credit prolonged
numeric	133	bureau.csv	AMT_CREDIT_SUM	Current credit amount for the Credit Bureau credit
numeric	134	bureau.csv	AMT_CREDIT_SUM_DEBT	Current debt on Credit Bureau credit
numeric	135	bureau.csv	AMT_CREDIT_SUM_LIMIT	Current credit limit of credit card reported in Credit Bureau
numeric	136	bureau.csv	AMT_CREDIT_SUM_OVERDUE	Current amount overdue on Credit Bureau credit
category	137	bureau.csv	CREDIT_TYPE	Type of Credit Bureau credit (Car, cash,...)
numeric	138	bureau.csv	DAYS_CREDIT_UPDATE	How many days before loan application did last information about the Credit Bureau credit come
numeric	139	bureau.csv	AMT_ANNUITY	Annuity of the Credit Bureau credit
category	140	bureau_balance.csv	SK_ID_BUREAU	Recoded ID of Credit Bureau credit (unique coding for each application) - use this to join to CREDIT_BUREAU table
category	141	bureau_balance.csv	MONTHS_BALANCE	Month of balance relative to application date (-1 means the information to the freshest monthly snapshot, 0 means the information at application - often it will be the same as -1 as many banks are not updating the information to Credit Bureau regularly)
numeric	142	bureau_balance.csv	STATUS	Status of Credit Bureau loan during the month (active, closed, DPD0-30,Ö [C means closed, X means status unknown, 0 means no DPD, 1 means maximal did during month between 1-30, 2 means DPD 31-60,Ö 5 means DPD 120+ or sold or written off])
category	143	POS_CASH_balance.csv	SK_ID_PREV	ID of previous credit in Home Credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loans in Home Credit)
category	144	POS_CASH_balance.csv	SK_ID_CURR	ID of loan in our sample

numeric	145	POS_CASH_balance.csv	MONTHS_BALANCE	Month of balance relative to application date (-1 means the freshest balance date)
numeric	146	POS_CASH_balance.csv	CNT_INSTALMENT	Term of previous credit (can change over time)
numeric	147	POS_CASH_balance.csv	CNT_INSTALMENT_FUTURE	Installments left to pay on the previous credit
category	148	POS_CASH_balance.csv	NAME_CONTRACT_STATUS	Contract status during the month
numeric	149	POS_CASH_balance.csv	SK_DPD	DPD (days past due) during the month of previous credit
numeric	150	POS_CASH_balance.csv	SK_DPD_DEF	DPD during the month with tolerance (debts with low loan amounts are ignored) of the previous credit

category	151	credit_card_balance.csv	SK_ID_PREV	ID of previous credit in Home credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loans in Home Credit)
category	152	credit_card_balance.csv	SK_ID_CURR	ID of loan in our sample
numeric	153	credit_card_balance.csv	MONTHS_BALANCE	Month of balance relative to application date (-1 means the freshest balance date)
numeric	154	credit_card_balance.csv	AMT_BALANCE	Balance during the month of previous credit
numeric	155	credit_card_balance.csv	AMT_CREDIT_LIMIT_ACTUAL	Credit card limit during the month of the previous credit
numeric	156	credit_card_balance.csv	AMT_DRAWINGS_ATM_CURRENT	Amount drawing at ATM during the month of the previous credit
numeric	157	credit_card_balance.csv	AMT_DRAWINGS_CURRENT	Amount drawing during the month of the previous credit
numeric	158	credit_card_balance.csv	AMT_DRAWINGS_OTHER_CURRENT	Amount of other drawings during the month of the previous credit
numeric	159	credit_card_balance.csv	AMT_DRAWINGS_POS_CURRENT	Amount drawing or buying goods during the month of the previous credit
numeric	160	credit_card_balance.csv	AMT_INST_MIN_REGULARITY	Minimal installment for this month of the previous credit

numeric	161	credit_card_balance.csv	AMT_PAYMENT_CURRENT	How much did the client pay during the month on the previous credit
numeric	162	credit_card_balance.csv	AMT_PAYMENT_TOTAL_CURRENT	How much did the client pay during the month in total on the previous credit
numeric	163	credit_card_balance.csv	AMT_RECEIVABLE_PRINCIPAL	Amount receivable for principal on the previous credit
numeric	164	credit_card_balance.csv	AMT_RECIVABLE	Amount receivable on the previous credit
numeric	165	credit_card_balance.csv	AMT_TOTAL_RECEIVABLE	Total amount receivable on the previous credit
numeric	166	credit_card_balance.csv	CNT_DRAWINGS_ATM_CURRENT	Number of drawings at ATM during this month on the previous credit
numeric	167	credit_card_balance.csv	CNT_DRAWINGS_CURRENT	Number of drawings during this month on the previous credit
numeric	168	credit_card_balance.csv	CNT_DRAWINGS_OTHER_CURRENT	Number of other drawings during this month on the previous credit
numeric	169	credit_card_balance.csv	CNT_DRAWINGS_POS_CURRENT	Number of drawings for goods during this month on the previous credit
numeric	170	credit_card_balance.csv	CNT_INSTALMENT_MATURE_CUM	Number of paid installments on the previous credit
category	171	credit_card_balance.csv	NAME_CONTRACT_STATUS	Contract status (active signed,...) on the previous credit
numeric	172	credit_card_balance.csv	SK_DPD	DPD (Days past due) during the month on the previous credit
numeric	173	credit_card_balance.csv	SK_DPD_DEF	DPD (Days past due) during the month with tolerance (debts with low loan amounts are ignored) of the previous credit

category	174	previous_application.csv	SK_ID_PREV	ID of previous credit in Home credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loan applications in Home Credit, previous application could, but not necessarily have to lead to credit)

category	175	previous_application.csv	SK_ID_CURR	ID of loan in our sample
category	176	previous_application.csv	NAME_CONTRACT_TYPE	Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application
numeric	177	previous_application.csv	AMT_ANNUITY	Annuity of previous application
numeric	178	previous_application.csv	AMT_APPLICATION	For how much credit did client ask on the previous application
numeric	179	previous_application.csv	AMT_CREDIT	Final credit amount on the previous application. This differs from AMT_APPLICATIO N in a way that the AMT_APPLICATIO N is the amount for which the client initially applied for, but during our approval process he could have received different amount - AMT_CREDIT
numeric	180	previous_application.csv	AMT_DOWN_PAYMENT	Down payment on the previous application
numeric	181	previous_application.csv	AMT_GOODS_PRICE	Goods price of good that client asked for (if applicable) on the previous application
category	182	previous_application.csv	WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for previous application
category	183	previous_application.csv	HOUR_APPR_PROCESS_START	Approximately at what day hour did the client apply for the previous application
category	184	previous_application.csv	FLAG_LAST_APPL_PER_CONTRAC T	Flag if it was last application for the previous contract. Sometimes by mistake of client or our clerk there could be more applications for one single contract
category	185	previous_application.csv	NFLAG_LAST_APPL_IN_DAY	Flag if the application was the last application per day of the client. Sometimes clients apply for more applications a day. Rarely it could also be error in our system that one application is in the database twice

category	186	previous_application.csv	NFLAG_MICRO_CASH	Flag Micro finance loan
numeric	187	previous_application.csv	RATE_DOWN_PAYMENT	Down payment rate normalized on previous credit
numeric	188	previous_application.csv	RATE_INTEREST_PRIMARY	Interest rate normalized on previous credit
numeric	189	previous_application.csv	RATE_INTEREST_PRIVILEGED	Interest rate normalized on previous credit
category	190	previous_application.csv	NAME_CASH_LOAN_PURPOSE	Purpose of the cash loan
category	191	previous_application.csv	NAME_CONTRACT_STATUS	Contract status (approved, cancelled, ...) of previous application
numeric	192	previous_application.csv	DAYS_DECISION	Relative to current application when was the decision about previous application made
category	193	previous_application.csv	NAME_PAYMENT_TYPE	Payment method that client chose to pay for the previous application
category	194	previous_application.csv	CODE_REJECT_REASON	Why was the previous application rejected
category	195	previous_application.csv	NAME_TYPE_SUITE	Who accompanied client when applying for the previous application
category	196	previous_application.csv	NAME_CLIENT_TYPE	Was the client old or new client when applying for the previous application
category	197	previous_application.csv	NAME_GOODS_CATEGORY	What kind of goods did the client apply for in the previous application
category	198	previous_application.csv	NAME_PORTFOLIO	Was the previous application for CASH, POS, CAR, Ö
category	199	previous_application.csv	NAME_PRODUCT_TYPE	Was the previous application x-sell o walk-in
category	200	previous_application.csv	CHANNEL_TYPE	Through which channel we acquired the client on the previous application
category	201	previous_application.csv	SELLERPLACE_AREA	Selling area of seller place of the previous application
category	202	previous_application.csv	NAME_SELLER_INDUSTRY	The industry of the seller

numeric	203	previous_application.csv	CNT_PAYMENT	Term of previous credit at application of the previous application
category	204	previous_application.csv	NAME_YIELD_GROUP	Grouped interest rate into small medium and high of the previous application
category	205	previous_application.csv	PRODUCT_COMBINATION	Detailed product combination of the previous application
numeric	206	previous_application.csv	DAYS_FIRST_DRAWING	Relative to application date of current application when was the first disbursement of the previous application
numeric	207	previous_application.csv	DAYS_FIRST_DUE	Relative to application date of current application when was the first due supposed to be of the previous application
numeric	208	previous_application.csv	DAYS_LAST_DUE_1ST_VERSION	Relative to application date of current application when was the first due of the previous application
numeric	209	previous_application.csv	DAYS_LAST_DUE	Relative to application date of current application when was the last due date of the previous application
numeric	210	previous_application.csv	DAYS_TERMINATION	Relative to application date of current application when was the expected termination of the previous application
category	211	previous_application.csv	NFLAG_INSURED_ON_APPROVAL	Did the client requested insurance during the previous application

category	212	installments_payments.csv	SK_ID_PREV	ID of previous credit in Home credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loans in Home Credit)
category	213	installments_payments.csv	SK_ID_CURR	ID of loan in our sample

numeric	214	installments_payments.csv	NUM_INSTALMENT_VERSION	Version of installment calendar (0 is for credit card) of previous credit. Change of installment version from month to month signifies that some parameter of payment calendar has changed
numeric	215	installments_payments.csv	NUM_INSTALMENT_NUMBER	On which installment we observe payment
numeric	216	installments_payments.csv	DAYS_INSTALMENT	When the installment of previous credit was supposed to be paid (relative to application date of current loan)
numeric	217	installments_payments.csv	DAYS_ENTRY_PAYMENT	When was the installments of previous credit paid actually (relative to application date of current loan)
numeric	218	installments_payments.csv	AMT_INSTALMENT	What was the prescribed installment amount of previous credit on this installment
numeric	219	installments_payments.csv	AMT_PAYMENT	What the client actually paid on previous credit on this installment

New Features derived through Feature Engineering:

(Note all these features are on a client level)

RISK_SCORE - weighted risk score basis the status reported to the Credit Bureau
BUREAU_LOAN_COUNT - Number of loans reported in bureau of a client
BUREAU_CONSUMER_CREDIT_LOAN_COUNT – Number of Consumer Credit loans of a client
BUREAU_CAR_LOAN_COUNT – Number of Car loans of a client
BUREAU_CREDIT_CARD_COUNT – Number of Credit Cards of a client
BUREAU_BUSINESS_LOAN_COUNT – Number of Business loans of a client
BUREAU_MICRO_LOAN_COUNT – Number of Micro loans of a client
BUREAU_MORTGAGE_LOAN_COUNT – Number of Mortgage loans of a client
BUREAU_REAL_ESTATE_LOAN_COUNT – Number of Real Estate loans of a client
BUREAU_OTHER_LOAN_COUNT – Number of any Other type of loans taken by a client
BUREAU_TOTAL_AMT_ANNUITY – Total Annuity amount pledged by a client for all credit loans
BUREAU_BAD_DEBT_RATIO – Bad Debt ratio of a client
BUREAU_ACTIVE_DEBTS_RATIO – Active Debts ratio of a client
BUREAU_LATEST_CREDIT – Latest Credit loan applied apart from applying in our lending agency (in days)
BUREAU_TOTAL_AMT_OVERDUE – Total Amount Overdue from all bureau reported loans for a client
BUREAU_MAX_AMT_OVERDUE- Maximum Overdue for a client
BUREAU_DAYS_OVERDUE – Average Overdue days for a client
BUREAU_CREDIT_PROLONGED_FREQ – Average Credit Prolonged Frequency for a client
BUREAU_AMT_DEBT_OUTSTANDING – Total Outstanding Debt for a client
BUREAU_OVERDUE_DEBT_RATIO – Overdue Debt ratio for a client
BUREAU_DEBT_CREDIT_RATIO – Debt over Credit ratio for a client
DPD_HIST - Total Days Past Due (DPD) for all loan payments done by due date for past & existing loans from our le nding agency

DBD_HIST - Total Days Before Due (DBD) for all loan payments done before due date for past & existing loans fro m our lending agency
EXCESS_PAYMENT_HIST – Excess Payments done for loans from our lending agency
PAYMENT_RATIO_HIST - Actual Payment to Instalment Ratio for loans from our lending agency
CC_BALANCE - Total Credit Card Balance for the Credit Cards from our lending agency
CC_LIMIT - Total Credit Card Limit for the Credit Cards from our lending agency
CC_DRAWINGS - Total Credit Card Drawings through various channels, viz. ATM, POS etc. for Credit Cards from our lending agency
CC_MIN_INSTALMENT – Total Minimum Due for all the Credit Cards owned by a client from our lending agency
CC_PAYMENT – Total Payments done for all the Credit Cards owned by a client from our lending agency
CC_PRINCIPAL_DUE – Total Principal amount due for all the Credit Cards owned by a client from our lending agen cy
CC_INTEREST_DUE - Total Interest amount due for all the Credit Cards owned by a client from our lending agency
CC_AVG_DRAWINGS_CNT – Average number of drawings on all the Credit Cards owned by a client from our lend ing agency
CC_PAID_INSTALMENT_CNT – Total number of paid instalments for all the Credit Cards owned by a client from our lending agency
CC_ACTIVE_CNT – Number of Active Credit Cards
CC_APPROVED_CNT – Number of Approved Credit Cards
CC_COMPLETED_CNT – Number of Completed Credit Cards
CC_DEMAND_CNT – Number of Credit Cards with ‘Demand’ status
CC_REFUSED_CNT – Number of Refused Credit Cards
CC_SENT_PROPOSAL_CNT – Number of Credit Cards with ‘Sent Proposal’ status
CC_SIGNED_CNT – Number of Credit Cards with ‘Signed’ status
CC_DPD – Total DPD on all Credit Cards owned by a client from our lending agency
POS_INSTALMENT_FUTURE_CNT – Total Future Instalments for all POS/Cash loans from our lending agency
POS_DPD – Total DPD on all POS/Cash loans from our lending agency
POS_ACTIVE_CNT – Number of Active POS/Cash loans
POS_AMORTIZED_DEBT_CNT - Number of Amortized debt POS/Cash loans
POS_APPROVED_CNT - Number of Approved POS/Cash loans
POS_CANCELED_CNT - Number of Canceled POS/Cash loans
POS_COMPLETED_CNT - Number of Completed POS/Cash loans
POS_DEMAND_CNT – Number of POS/Cash loans with Demand status
POS_RETURNED_CNT - Number of Returned to the store POS loans
POS_SIGNED_CNT – Number of POS/Cash loans with Signed status
PREV_AMT_ANNUITY - Total Annuity Amount pledged for previous credits from our lending agency
PREV_CREDIT_TO_APP_RATIO - Credit Amount to Application Amount Ratio for previous credits from our lendi ng agency
PREV_APPROVED_CNT – Number of Approved credits from previous applications
PREV_CANCELED_CNT – Number of Cancelled loans from previous applications
PREV_REFUSED_CNT – Number of Refused credits from previous applications
PREV_UNUSED_CNT - Number of Unused offer loans from previous applications
PREV_CARDS_APP_CNT – Number of past Credit Card applications
PREV_CARS_APP_CNT – Number of past Car loan applications
PREV_CASH_APP_CNT – Number of past Cash loan applications
PREV_POS_APP_CNT – Number of past POS loan applications
PREV_HIGH_INTEREST_GROUP_CNT – Number of High Interest Rate loans
PREV_LOW_ACTION_INTEREST_GROUP_CNT – Number of Low Action Interest Rate loans
PREV_LOW_NORMAL_INTEREST_GROUP_CNT – Number of Low Normal Interest Rate loans
PREV_MIDDLE_INTEREST_GROUP_CNT – Number of Medium Interest Rate loans
PREV_INSURED_RATIO – Insured to Non-Insured ratio for past loans from our lending agency

Other Predictive Models Tried

Decision Tree (CART)

Random Forest

Gradient Boosting Model (GBM)

Linear Support Vector Machine (Linear SVM)

Non-Linear Support Vector Machine (Non-Linear SVM)

Linear Discriminant Analysis (LDA)

Regularized Discriminant Analysis (RDA)

Naïve Bayes

Neural Network (ANN)