

Project Report for

**DATA MINING AND KNOWLEDGE DISCOVERY**

**IME 672**

**TITLE:**

**To predict how capable each applicant is of repaying a loan**

**PREPARED BY:**

|  |  |
| --- | --- |
| **Group No. : 2** | |
| **Name** | **Roll No** |
| Anurag Mishra | 17125007 |
| Archeesh Kalra | 17125009 |
| Bahul Dandona | 17125010 |
| Sandipan Bhattacharjee | 17125050 |
| Yatin Kundra | 17125069 |
| Danish Nawaz | 18114008 |

Under the Supervision of:

**Dr. Faiz Hamid**

Contents

[INTRODUCTION 3](#_Toc529647436)

[DATA DESCRIPTION 3](#_Toc529647437)

[DATA PRE-PROCESSING 4](#_Toc529647438)

[Modelling 5](#_Toc529647439)

[K Nearest Neighbours Model 5](#_Toc529647440)

[Support Vector Machine Model 5](#_Toc529647441)

[Logistic Regression Model 6](#_Toc529647442)

[Decision Tree Model 7](#_Toc529647443)

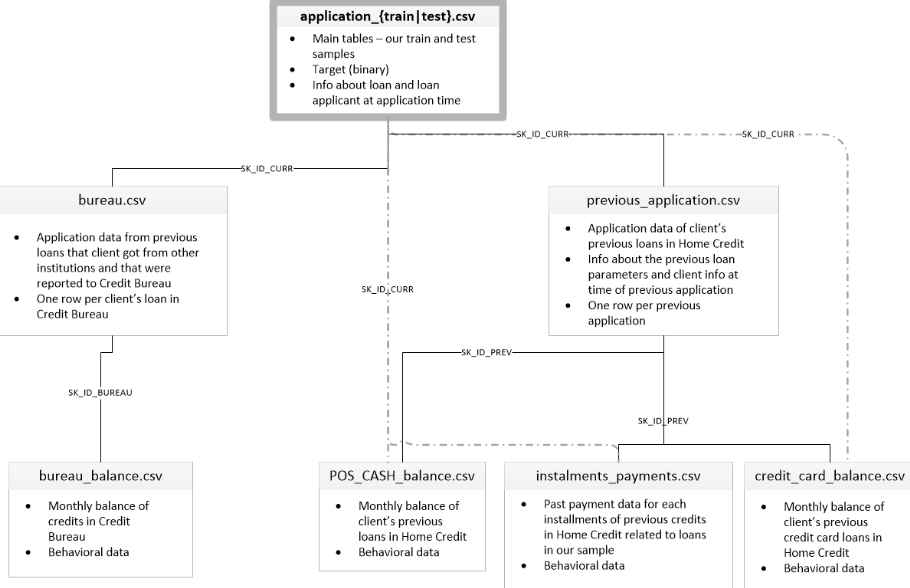
[Comparing Models and Conclusion 8](#_Toc529647444)

# INTRODUCTION

# One of the major problems people face while applying for a loan at a financial institution is their inability of getting the loan if they do not have any credit history. This gives chance to other lenders to provide a loan to these people at a higher rate of interest. Home Credit strives to improve the experience of these people by using the transactional data of the past customers which can help them to predict the repayment abilities of the loan applicants. So, the objective is to predict the value of target variable i.e. to find out whether the applicant faced the problem in repaying the loan installments based upon the various factors including the financial stability of the client, loan amount requested, any past loan history and various other factors.

# DATA DESCRIPTION

The figure below represents different data files and brief description about information presented in the data files



The application data represents demographic information about the loan applicants including their gender, income, education and family status and other attributes. The data also represents other attributes regarding the financial stability of the loan applicant like whether applicant owns car and house or not, description of the house and car in terms of the area in which house is located and age of the car of applicant. The remaining attributes give insights about the information provided by the applicant while applying for the loan.

Bureau data represents the information regarding the previous credits of all the customers that were reported to the credit bureau whereas bureau balance data represents monthly balance of previous credits in the credit bureau. Similarly, the other files represent the repayment history and the monthly balance details of previous credit cards.

# DATA PRE PROCESSING

## Data Transformation

As the application data contained more than 100 numeric predictor variables, initially graphs were plotted to understand the distribution of each variable. Out of all the predictor variables, 13 variables were skewed. So the data also needs to be transformed because of the **skewness** (right skewed or left skewed) of variables. Replacing the data with log, sqrt or inverse helped to remove the skewness. There were also some outliers in data which were removed by the transformation techniques. Missing values were also handled. The values could be missing because they are **structurally missing.** We used **K-nearest neighbour model to fill the missing values**. The advantage is that the imputed values are confined to be within the range of the training set values.

## Data Reduction

The next step of the data pre-processing involved the data reduction in which Principal Component Analysis (PCA) was done to find the linear combination of the predictor variables that captures the maximum variability of all the combinations. Before the PCA analysis the centering and scaling is done to normalize the data. To center the predictor variable, the average predictor value is subtracted from all the values. Similarly to scale the data, each value of the predictor variable is divided by its standard deviation. Then we have done Removing of variables with **zero variance.** Tree-based models are impervious to zero variance predictors but others like linear regression are not. So we have done it for implementing other models.

## Removing Multicollinearity

In the given data, there were more than 100 numerical predictor variables out of which many variables were highly correlated with one other. If all the correlated predictor variables are kept in data, they will make the model more complex. The model will be highly unstable and will be prone to numerical errors and hence degraded performance. So correlation matrix was formed. All the variables which were having correlation closer to 1 were removed from the data.

Reasons to avoid highly correlated data:

* Redundant predictors frequently add more complexity to the model than information they provide to the model
* Results in unstable model, numerical errors and degraded predictive performance

## Data Splitting and Re-sampling

The next step involved the data splitting and re-sampling of the data. The Training data is used to create the model and test data is used to qualify performance of the data. The resampling of data   
is done using the k fold cross validation technique. In the given technique the data is classified into k parts and k-1 parts of the data are used to create the model and remaining one part is the test set on which the performance of the model is evaluated . If the resampling of data is not done then it might work very well on the training data but not on the test data. So the issue of over fitting is resolved by re-sampling the data.

# Modelling

## The objective of the project was to classify whether the loan applicant will face difficulty in repaying the loan or not. So, in order to solve the given classification problem, multiple techniques were used which were:

## K Nearest Neighbours

K nearest neighbours algorithm is a supervised learning technique which is used mostly for classification problems. In this algorithm, the given data point is assigned a particular label out of the two labels based on the majority votes of its nearest k neighbours. In K nearest neighbours, Euclidean distance is used to find the distance between the two data points .So the K nearest neighbours technique was applied to the pre-processed data in order to classify the loan applicant. In order to apply the K nearest neighbours algorithm, we have used the train function from a built in package caret of R after adding appropriate parameters in the train function. The performance parameters of K nearest neighbours model were calculated for different values of K up to which the performance of the model was improving. . The Table below represents the performance parameters of model based on different value of K:

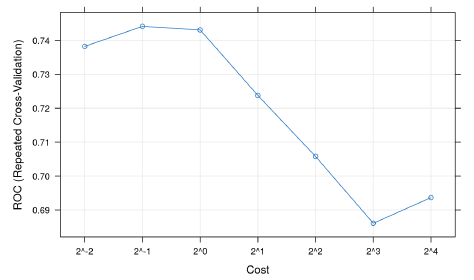
|  |  |  |  |
| --- | --- | --- | --- |
| Value of K | Sensitivity | Specificity | Area under ROC Curve |
| 3 | 0.59 | 0.63 | 0.65 |
| 4 | 0.62 | 0.63 | 0.66 |
| 5 | 0.65 | 0.64 | 0.68 |
| 6 | 0.65 | 0.64 | 0.68 |

## Support Vector Machine

Similar to the K nearest neighbours, support vector machine is a supervised learning algorithm used for the classification problems. In applying this algorithm, initially each data point is plotted in the n dimensional space where n represents the number of attributes and then find the appropriate hyper plane that differentiates the data points into one of the two labels i.e. classify the data points into different classes.

If multiple hyper planes are classifying the data points, the best hyper plane is the one which maximises the distance between the nearest data points i.e. the hyper plane having the highest margin. In order to apply the SVM algorithm, we have used the train function from a built in package caret of R after adding appropriate parameters in the train function. The ROC curve of the SVM model is shown in the figure below:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted Value | |
|  |  | No | Yes |
| Actual | No | 22871 | 9 |
| Value | Yes | 1844 | 17 |



## Logistic Regression

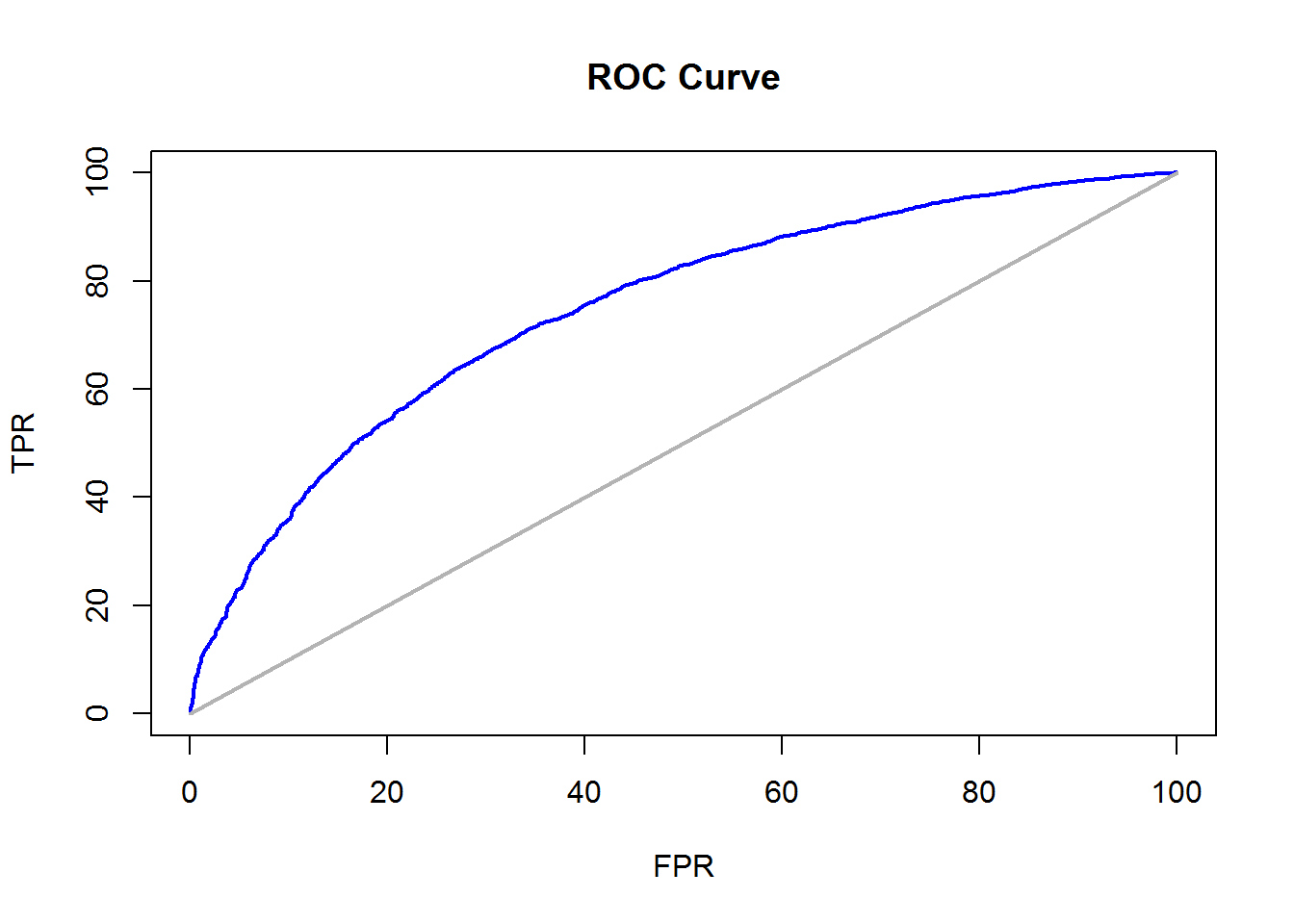
Logistic regression mathematically means finding, y = f(x), when y is a categorical variable. The use of this method is to find the label of categorical variable when know our predictors x. In our project the dependent variable y is a categorical variable and the predictors are continuous variable .In this project we are using binomial logistic regression to predict whether the customer loan will be approved or not. The input given to the model is the weighted average of attributes value and some bias. The input is processed by the logistic function called sigmoid function.

where is the weighted input given to the sigmoid function. The function processed the input data and result the output. The output is then compared with a threshold value. If the output is at least equal to the threshold value then the output is 1 otherwise 0.In this project, we compared the output with a threshold value of 0.5 if the output 0.5 we labeled the prediction as 1 and 0 otherwise. In order to apply logistic regression in R, R has inbuilt packages that makes it very easy to fit a logistic regression model. The package called glm has logistic regressor function for model fitting and prediction function for predicting the class label of target variable. In order to evaluate the performance of the logistic regression model, the confusion matrix was formed.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted Value | |
|  |  | No | Yes |
| Actual | No | 22871 | 9 |
| Value | Yes | 1844 | 17 |

The diagonal elements represent the accuracy of the given model.

Accuracy can be calculated as = (TP= true positive, TN= true negative)



The above ROC curve is for the logistic regression model. The ROC curve represents plot in which the 1- specificity of the data is the x coordinate and the sensitivity of data is the y coordinate i.e. how the true positive rates are plotted against the false positive rates.

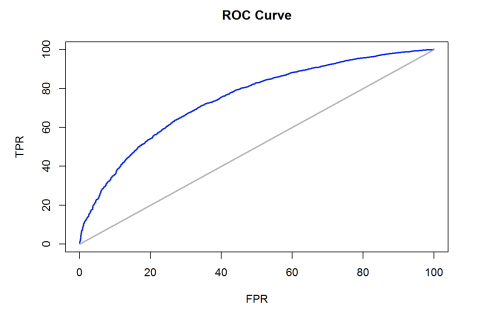
## Decision Tree

We have implemented the decision tree algorithm to predict loan default based on demography and other data of a customer. So, our main motivation for decision tree from logistic regression is that if there is a presence of non-linearity in the data set then decision tree works better than a logistic regression. The different tree-based algorithms involve stratifying or segmenting the predictor space into a number of simple regions. In order to make a prediction for a given observation, we typically use the mean or the mode of the training observations in the region to which it belongs. Since the set of splitting rules used to segment the predictor space can be summarized in a tree, these types of approaches are known as decision tree methods.

**Implementation in R:** R has inbuilt packages that make it very easy to fit a tree based model. The package called “*rpart*” has decision tree for model fitting and prediction function for predicting the class label of target variable.

After running the R code we get the following output:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted Value | |
|  |  | No | Yes |
| Actual | No | 22880 | 14 |
| Value | Yes | 1861 | 11 |

**ROC Curve**

**Limitation:** The decision tree algorithm is a greedy algorithm because at each step of the tree-building process, the best split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step. So, there is huge chance of over-fit of the model.

# Comparing Models and Conclusion

One of the most common methods used for characterizing the model’s predictive capabilities is to use the RMSE (Root Mean Square Error). RMSE is then computed by taking the square root of the MSE which is sum of square of the residuals. Another common metric used in R2. An R2 value of 0.8 means the model can explain 80% of the variation in the outcome.

The simplest metric for evaluating the predict classes is the overall accuracy (error rate). However, overall accuracy count doesn’t make any clear distinction on the type of the error being made (whether the error is a type I error or a type II error). Hence, sensitivity, which considers the true positive rate and specificity, which considers the false positive rate, can be used to evaluate the model.

Now, there is a trade-off between the sensitivity and specificity which can be resolved the ROC (Receiver Operating Characteristics) curve. Area under the ROC curve is a determinant to evaluate various models.

In our case, it is important that the model correctly find out those people who will have difficulty in repaying the loan rather than people who will not have difficulty in repaying the loan. Therefore, specificity is comparatively more important in our case than the sensitivity and accuracy.

Based on the ROC curve, SVM has the largest area under the curve (0.7565) as compared the logistic regression (0.7011). Hence, we can say that **SVM is the best suited model** according to our analysis.

