Credit Card Approval Prediction

Presented By

Samuel Adedokun
Uzwa Mazhar
Sakshi Joshi
Maria Valentina
Srinidhi Lakshminarayanan
Shakthi Viswanathan

Table of Contents

OBJECTIVE

Objective of this project is to enable a bank to decide whether to issue a credit card to an applicant or not. It is achieved by predicting the probability of future defaults and credit card borrowings using personal information and data submitted by credit card applicants.

BACKGROUND

Financial industry has for a long time used Credit score cards to determine the approval of a loan. Banks use personal information and data submitted by credit card applicants to assess the risk of applicants defaulting. This parameter in turn influences the approval of a future loan.

Logistic model is a common method for credit scoring. Logistic is suitable for binary classification tasks and can calculate the coefficients of each feature.
At present, with the development of machine learning algorithms. More predictive methods such as Boosting, Random Forest, and Support Vector Machines have been introduced into credit card scoring. However, these methods often do not have good transparency. It may be difficult to provide customers and regulators with a reason for rejection or acceptance.

In this study a few apart from Logistic models, a few other models have been discussed and the prediction of each model is presented.

DATA

SOURCE:

- **Source**: Kaggle link
- **Data files**:
  - application_record.csv - Personal information of the applicant
  - Credit_record.csv – Credit behavior of the applicant

Total Observations: 631765

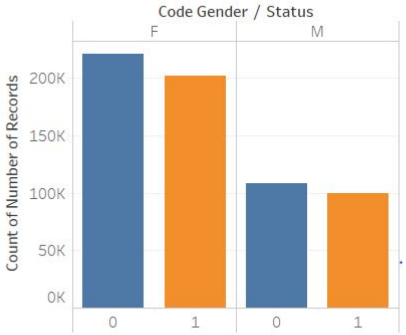
Total Features: 17

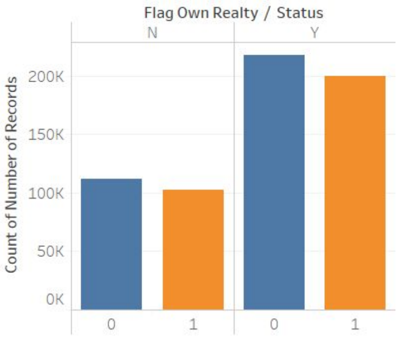
TYPES OF PREDICTOR VARIABLES:

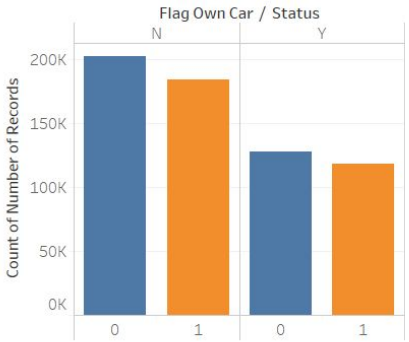
Binary Variables	Categorical Variables	Continuous Variables
• Gender • Having a car or not • Having house reality or not • Having a phone or not • Having an email or not • Having a Work Phone or not	• Income Type • Occupation Type • House Type • Education • Marriage Condition	• Number of Children • Annual Income • Age • Working Years • Family Size • Monthly Balance

EXPLORATORY DATA ANALYSIS

BINARY VARIABLES

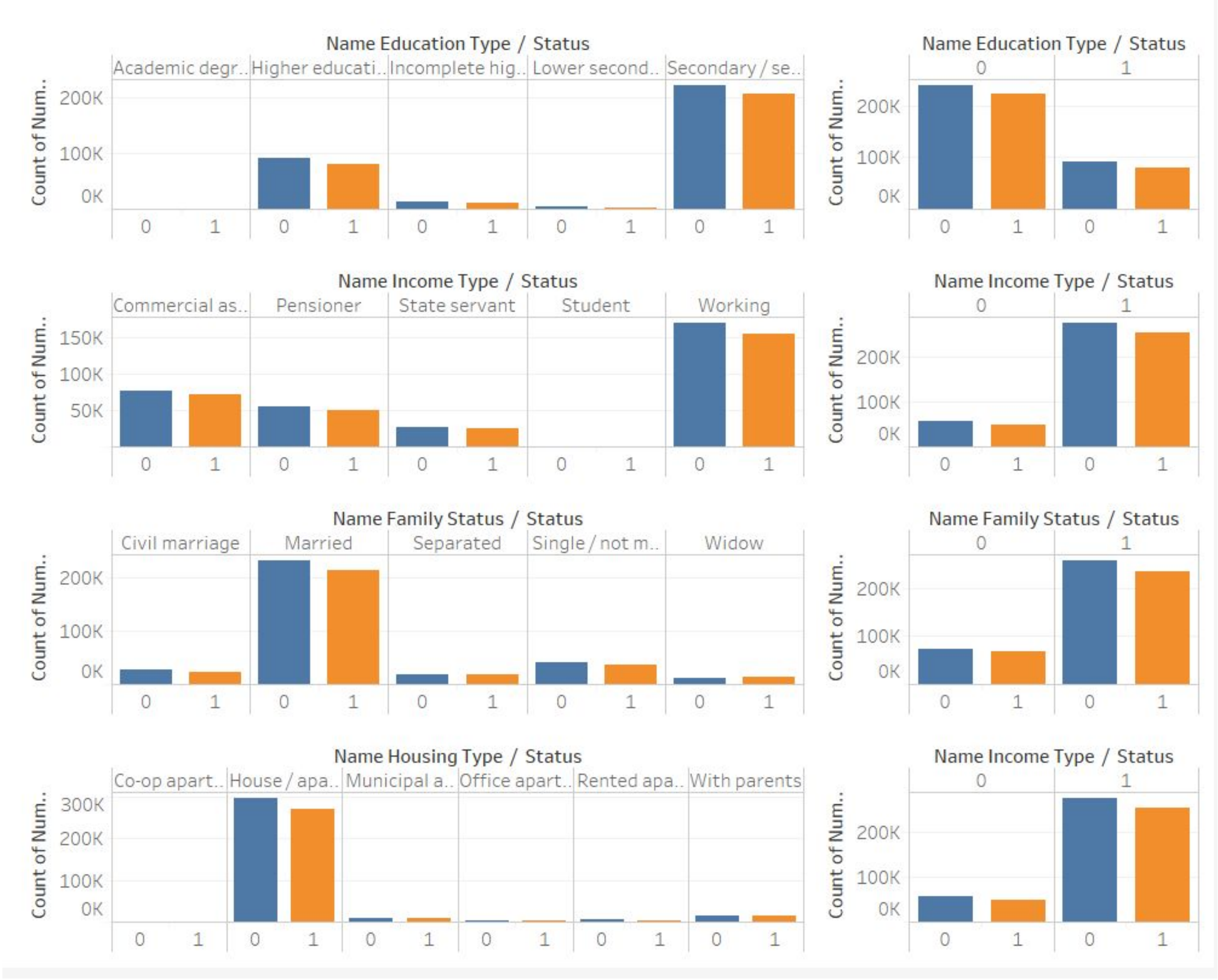






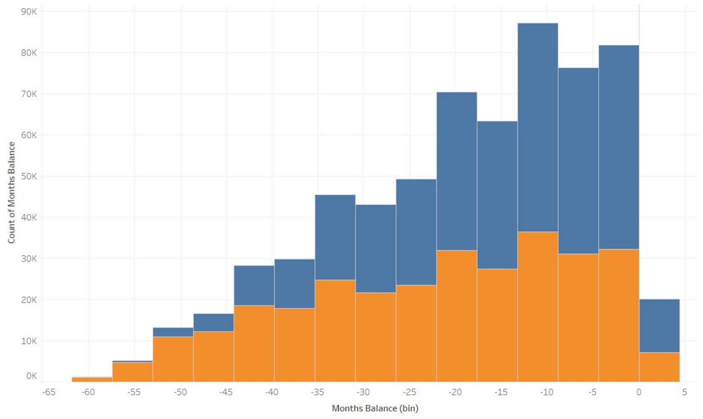
- **Gender vs Status**
  - There are more female customers compared to male customers
  - More percentage of women default their credit card payment compared to men
- **Own Realty vs Status**
  - There are more number of customers who a realty estate
  - The customers with a realty are more likely to default compared to customers who do not
- **Own car vs Status**
  - There are more customers who do not own a car
  - The customers who own a car are less likely to default compared to customers who do not

CATEGORICAL VARIABLES

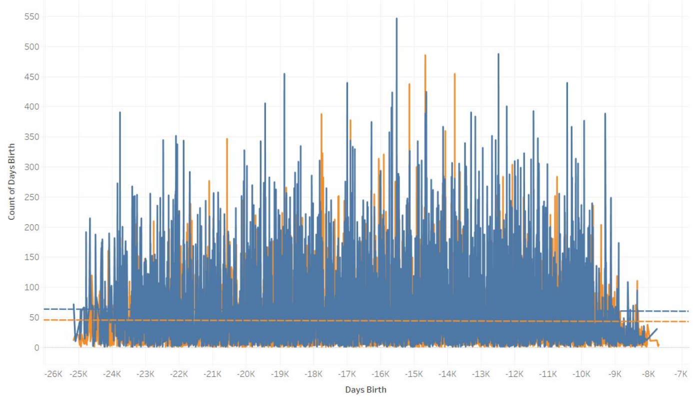


The multiclass variables shown here have more than 2 categorical values and the distribution is shown in the graph. These variables are converted into binary variables to ease the process of modelling and interpretation.

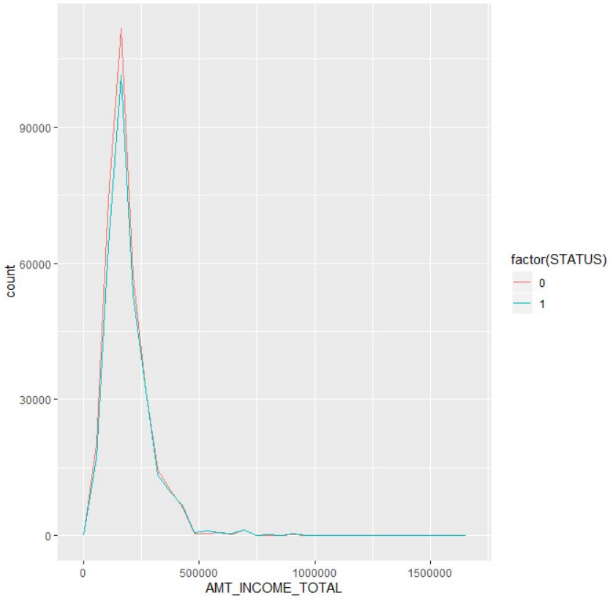
CONTINUOUS VARIABLES

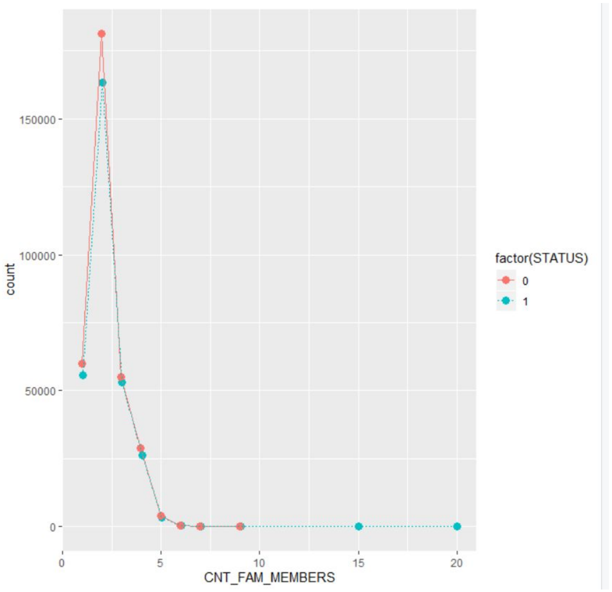


The stacked histogram for balance of the month clearly shows that the share of higher monthly balance is higher for customers who are defaulting their credit card payment.



The graph clearly shows the average age of customers who are defaulting in the credit card payment is lower than the average age of customers who pay on time.





The graphs above show that less number of defaulters have a small family compared to customers who did not default. Also, the total income of defaulters is slightly higher than the income of customers who repay on time.

HYPOTHESIS TESTING

Using the generalized linear models (glm), we looked at each predictor to see how much of an effect it has on the status of future credit card borrowings and to decide whether the individual should receive a credit card. Of the 17 predictors we initially had, 13 of them were found to have a significant effect on the status. The following predictors and their p-values were:

- The applicant's gender
    - P-Value if the applicant is a male: 0.00575
- Whether the users own a car
    - P-Value if the applicant does own a car: 3.52 * 10^(-6)
- The number of children they have
    - P-Value: 0.000132
- Their annual income:
    - P-Value: less than 2 * 10^(-16)
- Whether they have a work phone
    - P-Value: 0.0118
- Whether they have a phone
    - P-Value: less than 2 * 10^(-16)
- Whether they have an email:
    - P-Value: 1.92 * 10^(-15)
- The size of their family:
    - P-Value: 0.0238
- Their income category (comparison against commercial associate)
    - P-Value for Pensioneer: 2.03 * 10^(-14)
    - P-Value for State Servant: 0.0196
    - P-Value for Student: 2.30 * 10^(-13)
    - P-Value for Working: 1.49 * 10^(-13)

- Their education level (comparison against having an academic degree):
    - P-Value for Higher Education: 3.21 * 10^(-16)
    - P-Value for Incomplete Higher Education: 3.69 * 10^(-16)
    - P-Value for Lower Secondary: 2.55 * 10^(-6)
    - P-Value for Secondary/Secondary Special: less than 2 * 10^(-16)
- Their marital status (comparison against having a Civil Marriage)
    - P-Value for Married: 6.30 * 10^(-11)
    - P-Value for Separated: 4.80 * 10^(-5)
    - P-Value for Single/Not Married: 9.57 * 10^(-6)
    - P-Value for Widow: less than 2 * 10^(-16)
- The number of days they were employed:
    - P-Value: 3.98 * 10^(-6)
- The monthly balance:
    - P-Value: less than 2 * 10^(-16)

CORRELATION ANALYSIS:

After extensive EDA and hypothesis testing of individual predictor variables on the outcome variable, correlation analysis between each predictor variable was done.

The multi categorical variables were binned to dichotomous variables in order to perform Pearson's correlation analysis.

The binning logic was as follows,

BINNING LOGIC:

NAME_INCOME_TYPE

Level 1: Working, Commercial Associate, State Servant

Level 0: Student, Pensioner

NAME_FAMILY_STATUS

Level 1: Civil Marriage, Married

Level 0: Divorced, Separated, Single/Not Married

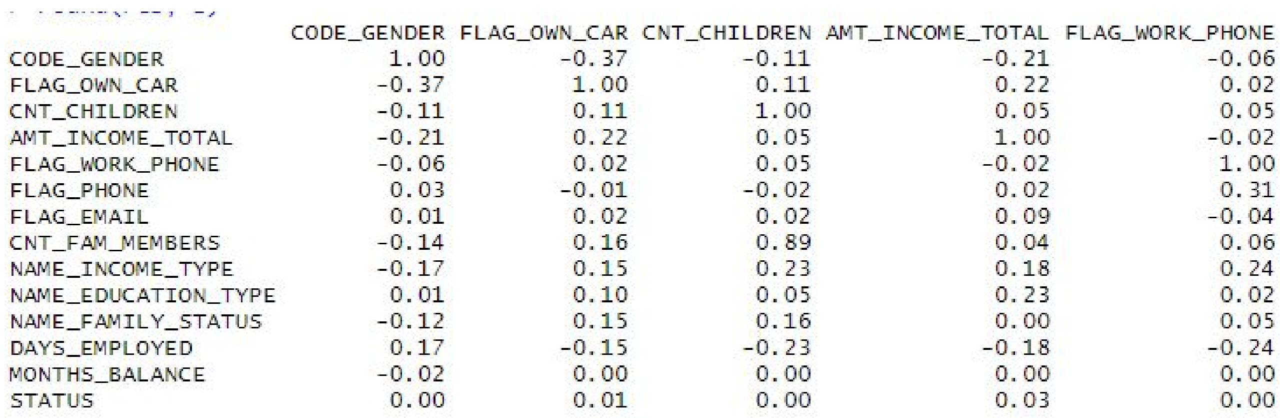
NAME_EDUCATION_TYPE

Level 1: Academic Degree, Higher

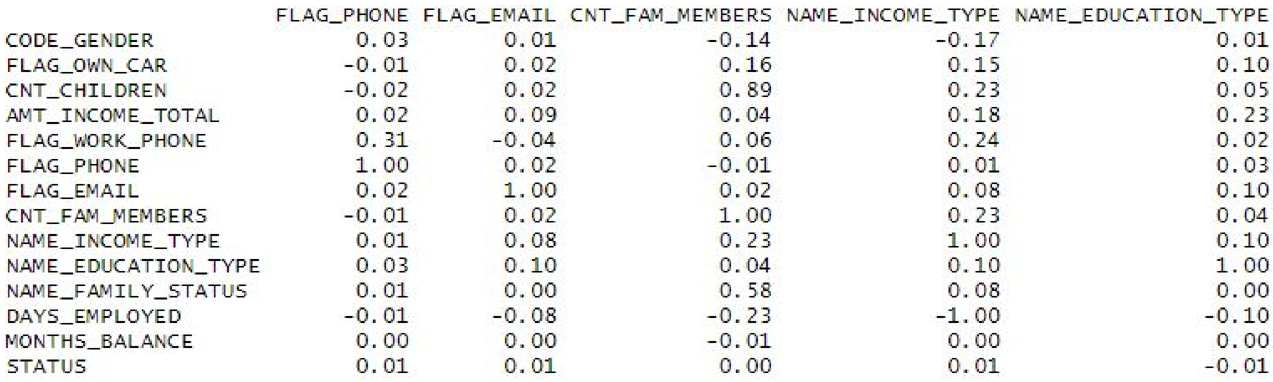
Level 0: Lower Sec, Secondary, Incomplete Higher

CORRELATION MATRIX

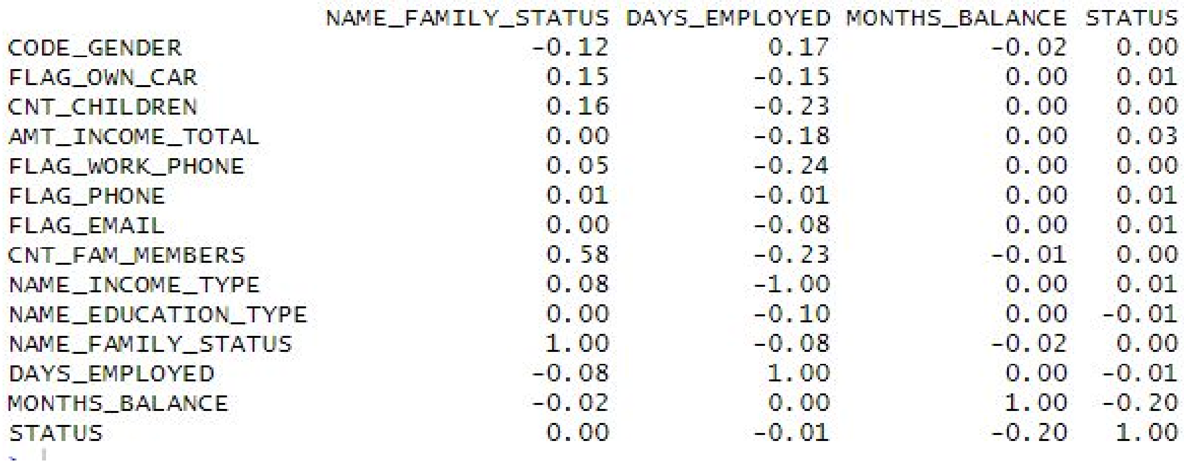
The correlation matrix indicating the correlation factor between each variable with all other variables is displayed below.



```
                     CODE_GENDER FLAG_OWN_CAR CNT_CHILDREN AMT_INCOME_TOTAL FLAG_WORK_PHONE
CODE_GENDER                 1.00        -0.37        -0.11            -0.21           -0.06
FLAG_OWN_CAR               -0.37         1.00         0.11             0.22            0.02
CNT_CHILDREN               -0.11         0.11         1.00             0.05            0.05
AMT_INCOME_TOTAL           -0.21         0.22         0.05             1.00           -0.02
FLAG_WORK_PHONE            -0.06         0.02         0.05            -0.02            1.00
FLAG_PHONE                  0.03        -0.01        -0.02             0.02            0.31
FLAG_EMAIL                  0.01         0.02         0.02             0.09           -0.04
CNT_FAM_MEMBERS            -0.14         0.16         0.89             0.04            0.06
NAME_INCOME_TYPE           -0.17         0.15         0.23             0.18            0.24
NAME_EDUCATION_TYPE         0.01         0.10         0.05             0.23            0.02
NAME_FAMILY_STATUS         -0.12         0.15         0.16             0.00            0.05
DAYS_EMPLOYED               0.17        -0.15        -0.23            -0.18           -0.24
MONTHS_BALANCE             -0.02         0.00         0.00             0.00            0.00
STATUS                      0.00         0.01         0.00             0.03            0.00
```



```
                    FLAG_PHONE FLAG_EMAIL CNT_FAM_MEMBERS NAME_INCOME_TYPE NAME_EDUCATION_TYPE
CODE_GENDER               0.03       0.01           -0.14            -0.17                0.01
FLAG_OWN_CAR             -0.01       0.02            0.16             0.15                0.10
CNT_CHILDREN             -0.02       0.02            0.89             0.23                0.05
AMT_INCOME_TOTAL          0.02       0.09            0.04             0.18                0.23
FLAG_WORK_PHONE           0.31      -0.04            0.06             0.24                0.02
FLAG_PHONE                1.00       0.02           -0.01             0.01                0.03
FLAG_EMAIL                0.02       1.00            0.02             0.08                0.10
CNT_FAM_MEMBERS          -0.01       0.02            1.00             0.23                0.04
NAME_INCOME_TYPE          0.01       0.08            0.23             1.00                0.10
NAME_EDUCATION_TYPE       0.03       0.10            0.04             0.10                1.00
NAME_FAMILY_STATUS        0.01       0.00            0.58             0.08                0.00
DAYS_EMPLOYED            -0.01      -0.08           -0.23            -1.00               -0.10
MONTHS_BALANCE            0.00       0.00           -0.01             0.00                0.00
STATUS                    0.01       0.01            0.00             0.01               -0.01
```

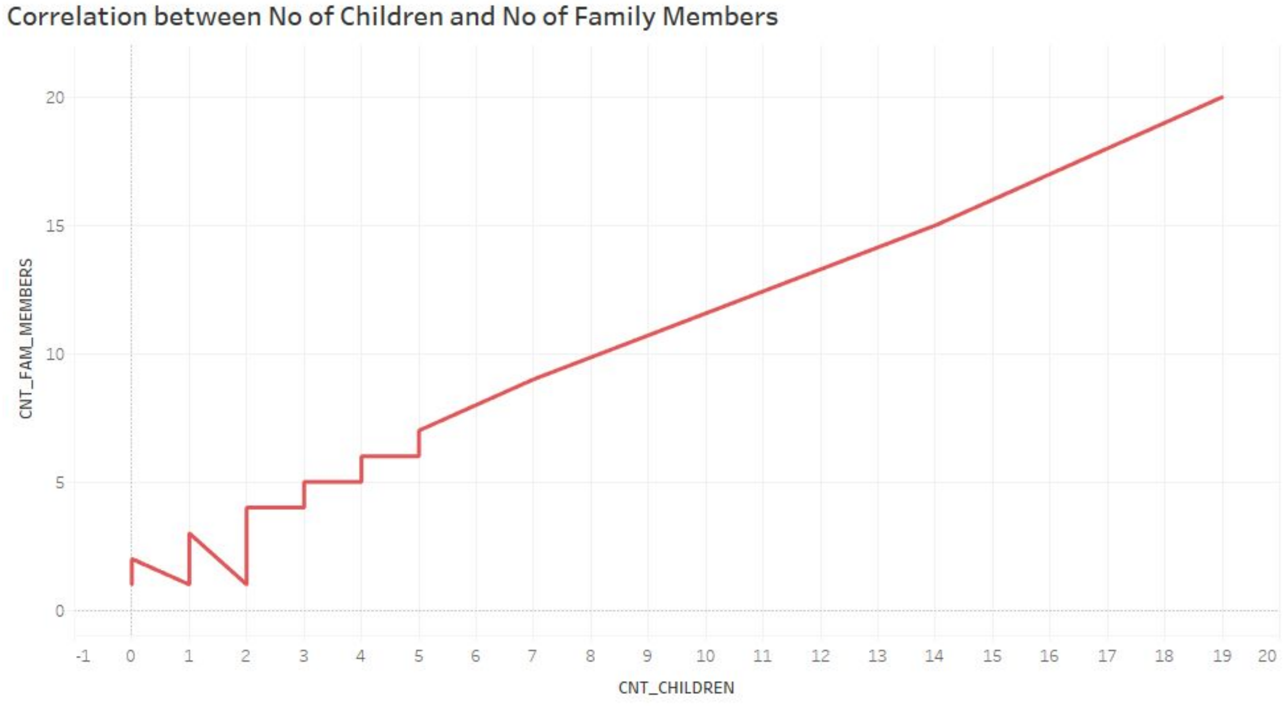


```
                      NAME_FAMILY_STATUS  DAYS_EMPLOYED  MONTHS_BALANCE  STATUS
CODE_GENDER                        -0.12           0.17           -0.02    0.00
FLAG_OWN_CAR                        0.15          -0.15            0.00    0.01
CNT_CHILDREN                        0.16          -0.23            0.00    0.00
AMT_INCOME_TOTAL                    0.00          -0.18            0.00    0.03
FLAG_WORK_PHONE                     0.05          -0.24            0.00    0.00
FLAG_PHONE                          0.01          -0.01            0.00    0.01
FLAG_EMAIL                          0.00          -0.08            0.00    0.01
CNT_FAM_MEMBERS                     0.58          -0.23           -0.01    0.00
NAME_INCOME_TYPE                    0.08          -1.00            0.00    0.01
NAME_EDUCATION_TYPE                 0.00          -0.10            0.00   -0.01
NAME_FAMILY_STATUS                  1.00          -0.08           -0.02    0.00
DAYS_EMPLOYED                      -0.08           1.00            0.00   -0.01
MONTHS_BALANCE                     -0.02           0.00            1.00   -0.20
STATUS                              0.00          -0.01           -0.20    1.00
```

SIGNIFICANT RELATIONSHIPS:

CNT_FAM_MEMBERS and CNT_CHILDREN had significant correlation factor 0.89

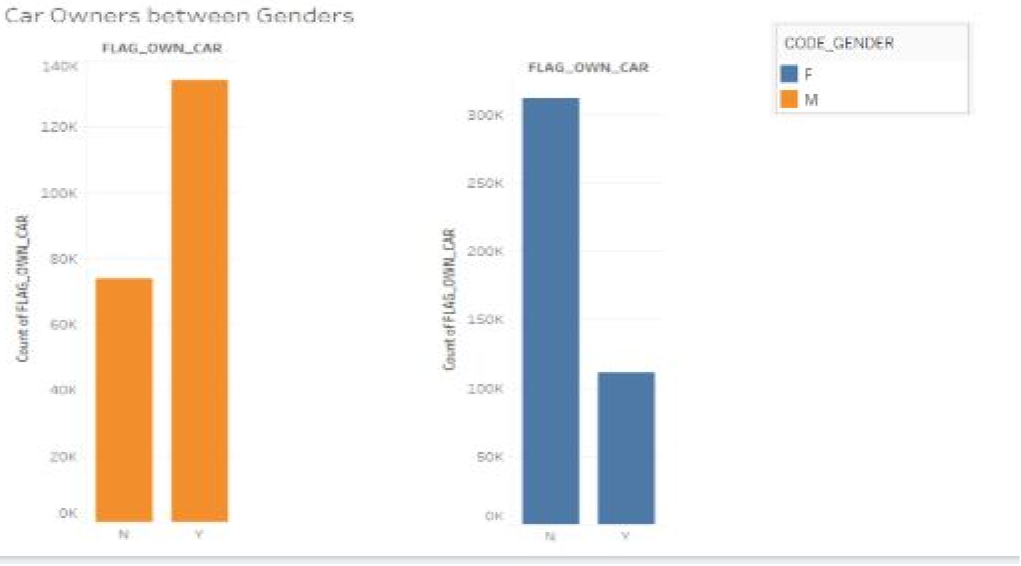
Plotting the relationship between these two variables we get the following graph, that clearly indicates a linear relationship between these two predictor variables. Since we could statistically prove this relationship in our data, and the number of children in an applicant's family is a subset of the number of family members of the applicant, the number of children in the family was no longer considered in further analysis.



Correlation between No of Children and No of Family Members

CODE_GENDER and FLAG_OWN_CAR had a correlation factor of 0.58 (so we kept them)

The bar plot below shows that clearly, more car owners are men and women. But the exact influence of this effect seems weaker than the individual effect of each predictor variable on the outcome variable. Hence both the predictor variables were considered for further analysis.



FEATURE SELECTION

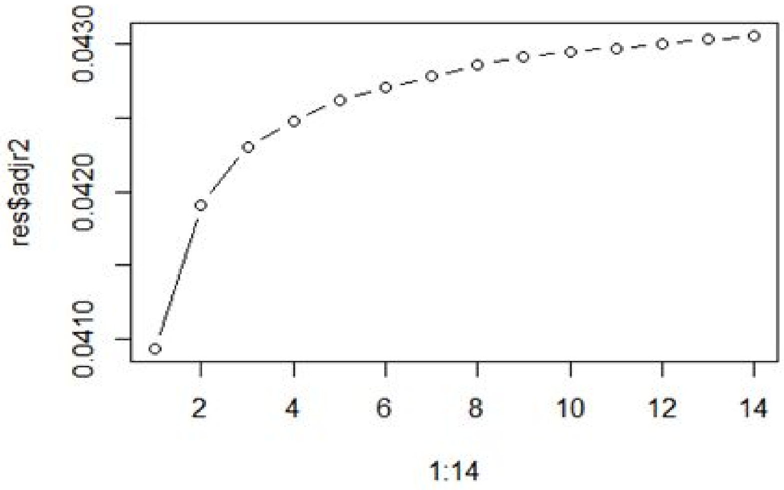
The features that were sorted as significant in the hypothesis testing were further analyzed to select the best set of features for models.

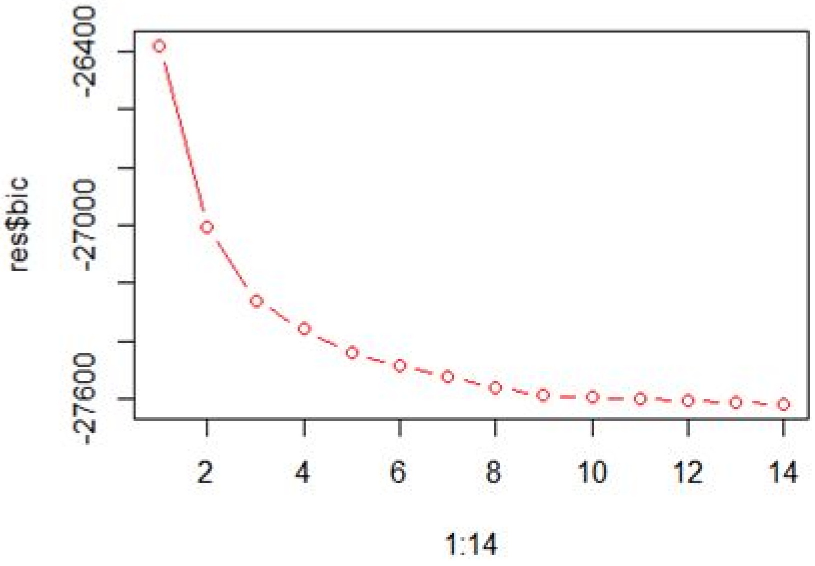
The following methods were used for feature selection

1. Exhaustive
2. Stepwise

EXHAUSTIVE SELECTION OF FEATURES

Based on the Adjusted R square and BIC, the best model was chosen





From the above graphs the 9th feature set was chosen as the best set, since after the 9th set the values of the adjusted R square and BIC begins to saturate.

The adjusted r square value for this particular model was **0.04291905** and corresponding BIC value was **-27589.26**.

Exhaustive method provided the following set of features as the best model

1. AMT_INCOME_TOTAL
2. FLAG_PHONE
3. FLAG_EMAIL
4. NAME_INCOME_TYPE Student
5. NAME_EDUCATION_TYPE Higher
6. NAME_EDUCATION_TYPE Incomplete
7. NAME_EDUCATION_TYPE Secondary
8. NAME_FAMILY_STATUS Widow
9. MONTHS_BALANCE

STEPWISE SELECTION OF FEATURES

Stepwise selection of features resulted in a different feature set compared to that of the Exhaustive method. The model suggested by stepwise in both forward and backward mode in step wise was the same. The predictor values suggested were

1. CODE_GENDER
2. AMT_INCOME_TOTAL
3. FLAG_WORK_PHONE
4. FLAG_PHONE
5. FLAG_EMAIL
6. CNT_FAM_MEMBERS
7. NAME_INCOME_TYPE
8. NAME_EDUCATION_TYPE
9. NAME_FAMILY_STATUS
10. DAYS_EMPLOYED
11. MONTHS_BALANCE

The AIC values for this model is **AIC=846955.2**

TRAINING AND TESTING DATA

The data was split into training and testing data before modelling.

Almost two third of the data was used for training the model and remaining one third of the data was used for prediction.

MODELLING:

The feature set chosen as best from the above feature selection process was modelled using different types of classifier models such as,

1. Logistic Regression
2. Classification Decision Tree
3. Random Forest
4. Support Vector Machine

The models are discussed individually and their accuracy measures are presented below.

LOGISTIC REGRESSION

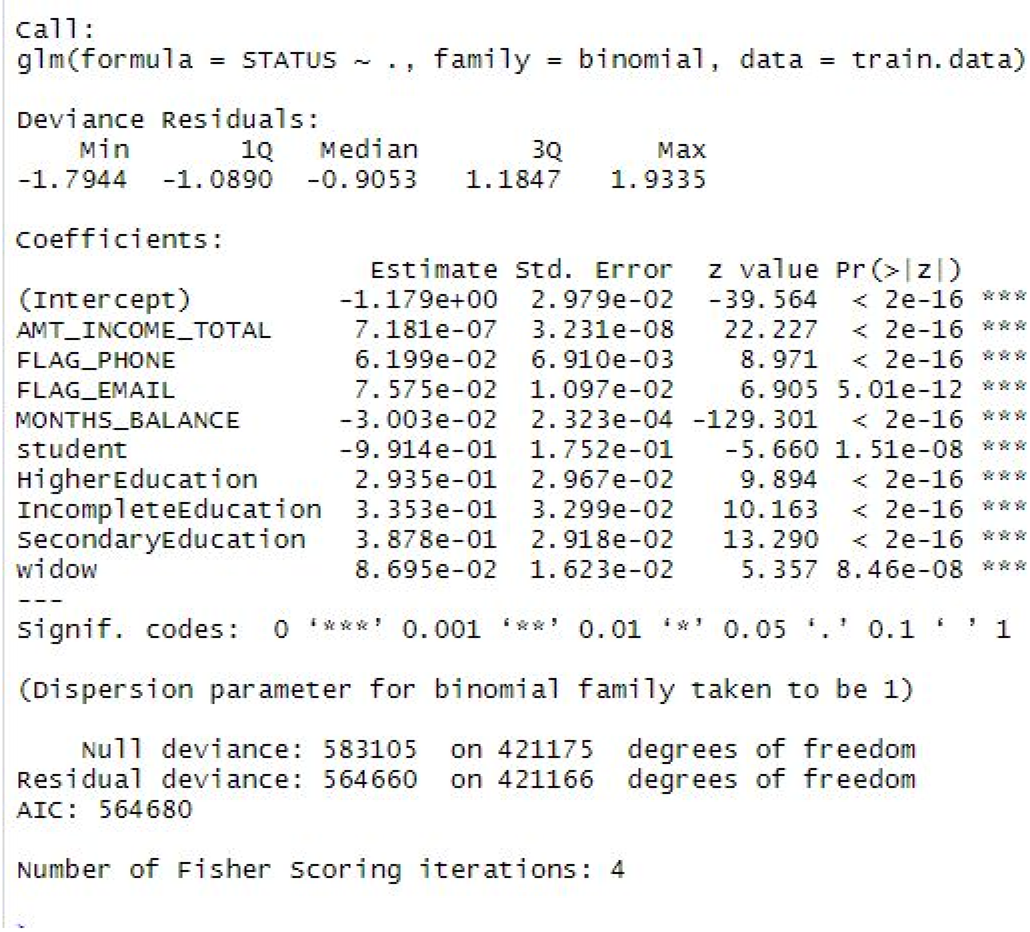
MODEL 1 (BEST)

The first model took the feature set chosen from the exhaustive feature selection method. Below is the formula of the model

STATUS ~ AMT_INCOME_TOTAL + FLAG_PHONE + FLAG_EMAIL + MONTHS_BALANCE + student + HigherEducation + Incomplete Education + Secondary Education + widow

Note: All dichotomous categorical variables were modified as 0's and 1's before modelling.

Following is the summary of the above model



```
Call:
glm(formula = STATUS ~ ., family = binomial, data = train.data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7944  -1.0890  -0.9053   1.1847   1.9335  

Coefficients:
                      Estimate Std. Error  z value Pr(>|z|)    
(Intercept)         -1.179e+00  2.979e-02  -39.564  < 2e-16 ***
AMT_INCOME_TOTAL     7.181e-07  3.231e-08   22.227  < 2e-16 ***
FLAG_PHONE           6.199e-02  6.910e-03    8.971  < 2e-16 ***
FLAG_EMAIL           7.575e-02  1.097e-02    6.905 5.01e-12 ***
MONTHS_BALANCE      -3.003e-02  2.323e-04 -129.301  < 2e-16 ***
student             -9.914e-01  1.752e-01   -5.660 1.51e-08 ***
HigherEducation      2.935e-01  2.967e-02    9.894  < 2e-16 ***
IncompleteEducation  3.353e-01  3.299e-02   10.163  < 2e-16 ***
SecondaryEducation   3.878e-01  2.918e-02   13.290  < 2e-16 ***
widow                8.695e-02  1.623e-02    5.357 8.46e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 583105  on 421175  degrees of freedom
Residual deviance: 564660  on 421166  degrees of freedom
AIC: 564680

Number of Fisher Scoring iterations: 4
```

Based on the above co-efficient obtained the following equation can be written

STATUS = -1.179 + e (0.00000071 * Income) + e(0.062 * FLAG_PHONE) +
e(0.078 * FLAG_EMAIL) + e(-0.03* MONTHS_BALANCE) + e(-0.99*Student) +
e(0.29*HigherEducation) + e(0.335*IncompleteEducation) +
e(0.387*SecondaryEducation) + e(0.87*Widow)

STATUS = -1.179 + 1 (Income) + 1.064(FLAG_PHONE) + 1.08(FLAG_EMAIL) +
0.97(MONTHS_BALANCE) + 0.37(Student) + 1.34(HigherEducation) +
1.4(IncompleteEducation) + 1.47(SecondaryEducation) + 1.09(Widow)

The model was built on the training data set. The prediction for the same was done using the testing data set. The prediction accuracy with the **cut off of 0.5** was as follows,

	0 (No Default)	**1 (Default)**
0	76382	53954
1	33752	46681

Accuracy:

76382 + 46681/ (76382 + 53954 + 33752 + 46681)

Accuracy = 58.3 %

Sensitivity:

46681 / (46681 + 53954)

Sensitivity = 46.3 %

Specificity:

76382 / (76382 + 33572)

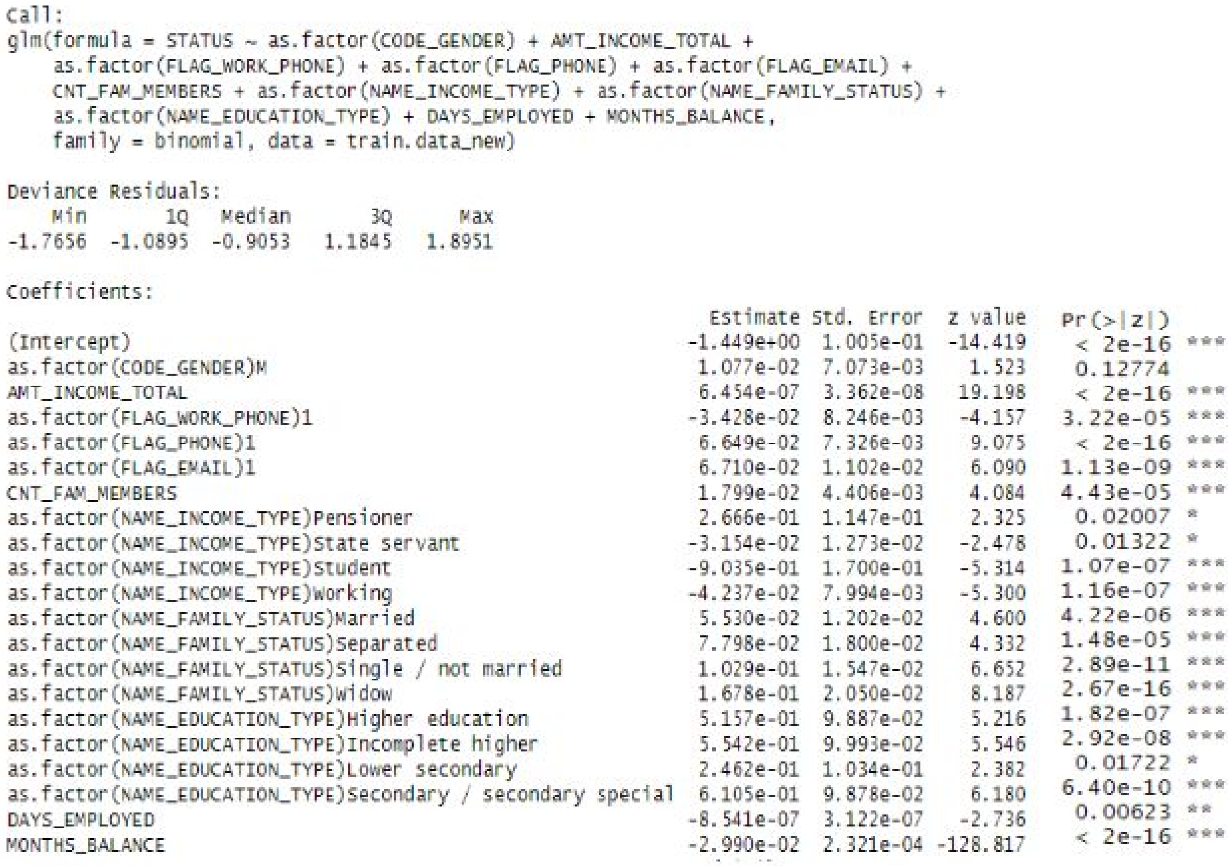
Specificity = 69.4 %

MODEL 2: (FEATURES BASED ON STEP WISE METHOD)

The stepwise feature selection method yielded the same feature set in both backward and forward method.

STATUS ~ AMT_INCOME_TOTAL + FLAG_WORK_PHONE + FLAG_PHONE + FLAG_EMAIL + CNT_FAM_MEMBERS + NAME_INCOME_TYPE + NAME_FAMILY_STATUS + NAME_EDUCATION_TYPE + DAYS_EMPLOYED + MONTHS_BALANCE

Note: as.factor function was applied to all the categorical variables



```
Call:
glm(formula = STATUS ~ as.factor(CODE_GENDER) + AMT_INCOME_TOTAL + 
    as.factor(FLAG_WORK_PHONE) + as.factor(FLAG_PHONE) + as.factor(FLAG_EMAIL) + 
    CNT_FAM_MEMBERS + as.factor(NAME_INCOME_TYPE) + as.factor(NAME_FAMILY_STATUS) + 
    as.factor(NAME_EDUCATION_TYPE) + DAYS_EMPLOYED + MONTHS_BALANCE, 
    family = binomial, data = train.data_new)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7656  -1.0895  -0.9053   1.1845   1.8951  

Coefficients:
                                                        Estimate Std. Error  z value  Pr(>|z|)    
(Intercept)                                           -1.449e+00  1.005e-01  -14.419   < 2e-16 ***
as.factor(CODE_GENDER)M                                1.077e-02  7.073e-03    1.523   0.12774    
AMT_INCOME_TOTAL                                       6.454e-07  3.362e-08   19.198   < 2e-16 ***
as.factor(FLAG_WORK_PHONE)1                           -3.428e-02  8.246e-03   -4.157  3.22e-05 ***
as.factor(FLAG_PHONE)1                                 6.649e-02  7.326e-03    9.075   < 2e-16 ***
as.factor(FLAG_EMAIL)1                                 6.710e-02  1.102e-02    6.090  1.13e-09 ***
CNT_FAM_MEMBERS                                        1.799e-02  4.406e-03    4.084  4.43e-05 ***
as.factor(NAME_INCOME_TYPE)Pensioner                   2.666e-01  1.147e-01    2.325   0.02007 *  
as.factor(NAME_INCOME_TYPE)State servant              -3.154e-02  1.273e-02   -2.478   0.01322 *  
as.factor(NAME_INCOME_TYPE)Student                    -9.035e-01  1.700e-01   -5.314  1.07e-07 ***
as.factor(NAME_INCOME_TYPE)Working                    -4.237e-02  7.994e-03   -5.300  1.16e-07 ***
as.factor(NAME_FAMILY_STATUS)Married                   5.530e-02  1.202e-02    4.600  4.22e-06 ***
as.factor(NAME_FAMILY_STATUS)Separated                 7.798e-02  1.800e-02    4.332  1.48e-05 ***
as.factor(NAME_FAMILY_STATUS)Single / not married      1.029e-01  1.547e-02    6.652  2.89e-11 ***
as.factor(NAME_FAMILY_STATUS)Widow                     1.678e-01  2.050e-02    8.187  2.67e-16 ***
as.factor(NAME_EDUCATION_TYPE)Higher education         5.157e-01  9.887e-02    5.216  1.82e-07 ***
as.factor(NAME_EDUCATION_TYPE)Incomplete higher        5.542e-01  9.993e-02    5.546  2.92e-08 ***
as.factor(NAME_EDUCATION_TYPE)Lower secondary          2.462e-01  1.034e-01    2.382   0.01722 *  
as.factor(NAME_EDUCATION_TYPE)Secondary / secondary special  6.105e-01  9.878e-02    6.180  6.40e-10 ***
DAYS_EMPLOYED                                         -8.541e-07  3.122e-07   -2.736   0.00623 ** 
MONTHS_BALANCE                                        -2.990e-02  2.321e-04 -128.817   < 2e-16 ***
```

The model was built on the training data set. The prediction for the same was done using the testing data set. The prediction accuracy with the cut off of 0.5 was as follows,

	0 (No Default)	**1 (Default)**
0	76178	53381
1	33938	47092

Accuracy:

(68127+38264)/ (68127+38264+41989+62209)

Accuracy = 58.53 %

Sensitivity:

47092/ (47092+53381)

Sensitivity = 46.8 %

Specificity:

76178/ (76178+33938)

Specificity = 69.1 %

INTERPRETATIONS

STATUS = -1.179 + 1 (Income) + 1.064(FLAG_PHONE) + 1.08(FLAG_EMAIL) + 0.97(MONTHS_BALANCE) + 0.37(Student) + 1.34(HigherEducation) + 1.4(IncompleteEducation) + 1.47(SecondaryEducation) + 1.09(Widow)

From the above equation the following interpretations can be made for individual predictor variables

INFLUENCE OF CATEGORICAL VARIABLES

INFLUENCE OF HIGHER EDUCATION

When all other factors (such as income, phone status, email status, student status, widow status) remain constant, for a person who has completed Higher education the odds of getting a loan rejected is 1.34 times than that of a person without a Higher education.

INFLUENCE OF SECONDARY EDUCATION

When all other factors (such as income, phone status, email status, student status, widow status) remain constant, for a person who has completed Secondary education the odds of getting a loan rejected is 1.4 times than that of a person without a Secondary education.

INFLUENCE OF WIDOW STATUS

When all other factors (such as income, phone status, email status, student status, Higher Education status) remain constant, for a person who is a widow, the odds of getting a loan rejected is 1.09 times than that of a person who is not a widow.

INFLUENCE OF STUDENTSTATUS

When all other factors (such as income, phone status, email status, Education status, widow status) remain constant, for a person who is a not a student has the odds of getting a loan rejected is 2.7 times than that of a person who is a student

INFLUENCE OF HAVING A PHONES

When all other factors (such as income, education status, email status, widow status, Student Status) remain constant, for a person who has a phone the odds of getting a loan rejected is 1.064 times than that of a person who does not have a phone.

INFLUENCE OF HAVING AN EMAIL

When all other factors (such as income, education status, phone status, widow status, Student Status) remain constant, for a person who has an email the odds of getting a loan rejected is 1.08 times than that of a person who does not have an email.

INFLUENCE OF INCOMPLETE EDUCATION

When all other factors (such as income, phone status, email status, widow status, Student Status) remain constant, for a person with incomplete education the odds of getting a loan rejected is 1.4 times than that of a person without incomplete education.

INFLUENCE OF CONTINUOUS VARIABLES

INFLUENCE OF NO OF MONTHS OF DELAYED CREDIT PAYMENT

When all other factors (such as education status, phone status, email status, widow status, Student Status) remain constant, for every unit increase in the number of months of delayed credit payment the odds of not getting a loan rejected increases 1.03 times.

DECISION TREE CLASSIFIER

The best model chosen from the exhaustive feature selection was used and a Decision Tree Classifier model was built on the training data.

The prediction on the model was done using the test data.

The prediction yielded the following results

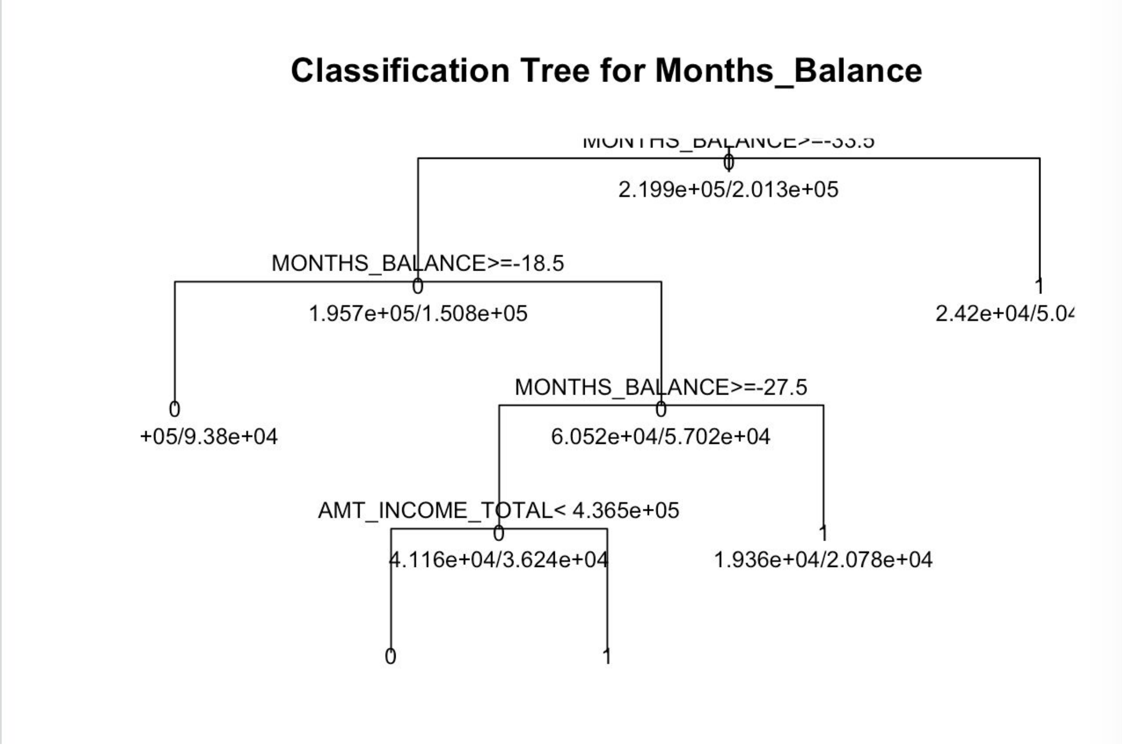
	0 (No Default)	**1 (Default)**
0	87520	22128
1	64973	35968

Specificity:61.9%

Sensitivity:57.4%

Accuracy:58.6%

The below graph shows the Classification Tree for the model developed.



RANDOM FOREST CLASSIFIER

The best model chosen from the exhaustive feature selection was used and a Random Forest model was built on the training data.

The prediction on the model was done using the test data.

The prediction yielded the following results.

	0 (No Default)	**1 (Default)**
0	147549	43407
1	13452	6181

Specificity:77.2%

Sensitivity:31.4%

Accuracy:73.3%

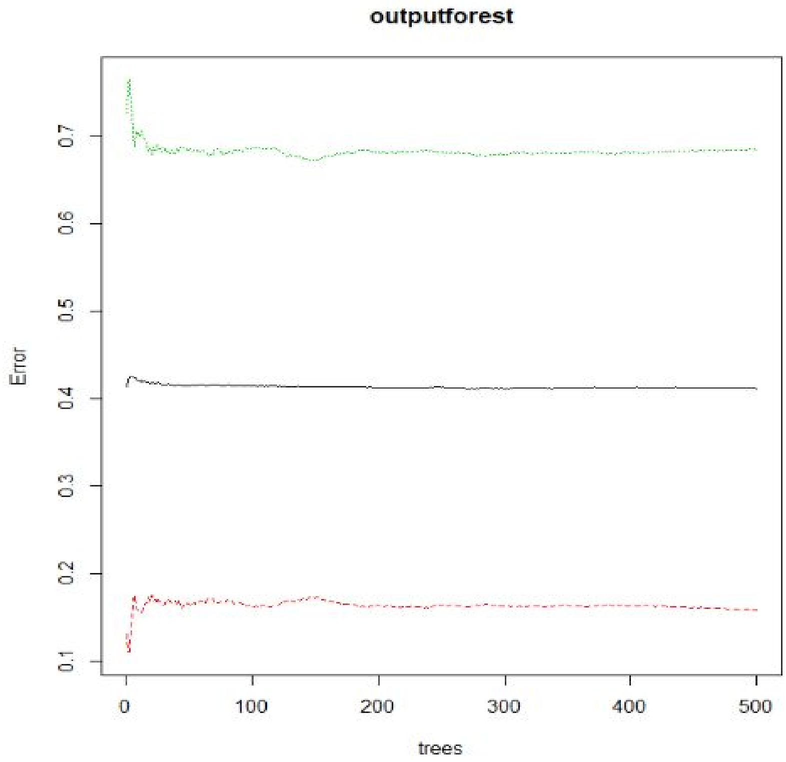
Although Random forest had a better accuracy rate than Logistic and Decision Tree, the model yielded lower sensitivity than the previous models.

Hence for overall accuracy, random forest is a better model for this data.

As far as sensitivity is concerned the other models have better prediction power for this data.

NOTE: Due to limitations in computing power, fewer observations were used for training and testing

The below graph shows the error for different trees.



The errors seem to be the same throughout all the trees.

SUPPORT VECTOR MACHINE

The best model chosen from the exhaustive feature selection was used and a Support Vector Machine model was built on the training data.

The prediction on the model was done using the test data.

The prediction yielded the following results.

	0 (No Default)	1 (Default)
0	13264	2034
1	12647	2055

NOTE: Due to limitations in computing power, fewer observations were used for training and testing

Accuracy:

(13264+2055) / (13264+2055+12647+2034)

Accuracy = 51 %

Sensitivity:

2055 / (2055+2034)

Sensitivity = 50.2 %

Specificity:

13264 / (13264+12647)

Specificity = 51.1 %

ACCURACY TABLES

Following table summarizes the results from all the models:

MODEL	FEATURE SET	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
GLM	BEST	58.3	46.3	69.4
GLM	BACKWARD	58.53	46.8	69.1
DECISION TREE	BEST	58.4	57.4	61.9
SVM	BEST	51	50.2	51.1
RANDOM FOREST	BEST	73.3	31.4	77.2

The table clearly shows that most models have consistent accuracies, sensitivity and specificity.

CONCLUSION

Based on the interpretation obtained from the Logistic models, the following can be concluded.

- Some valid insights obtained were, influence of being a widow and being a student impedes the approval of an applicant (Here valid means the real time behaviour seen in banks while approving loans)
- Contrary to some general belief, our data suggests people with higher, secondary education have a higher probability of defaulting more than others. Our EDA clearly shows that the data is well balanced, and an equal representation of each category is taken under consideration. Thus, more data and analysis are necessary to understand the odd influence of certain predictor variables
- Strictly for this data, heuristically, the credit history may not be the best deciding factor for approving a loan. The decision makers will have to determine the approval of a loan based on other factors as well.