Final Project Report

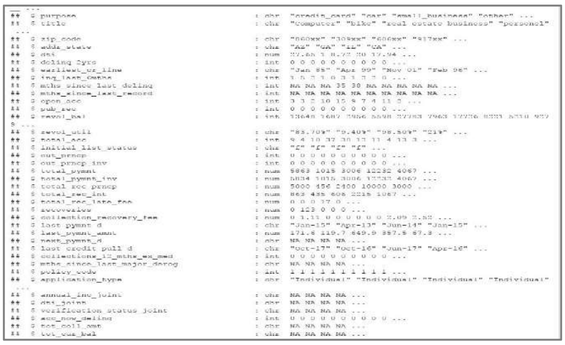
INTRODUCTION

Lending Club, a San Francisco-based fintech company, works to facilitate peer-to-peer loans through their online lending platform. Started in 2007, it is an online marketplace connecting borrowers and investors. Their website allows individuals to publicly post loan applications, which other users can then browse and choose to fund. This platform is transforming the banking system to make credit more affordable and investing more rewarding. Their website makes its historical records publicly available, leading to the interesting question:

Can we determine a way of using loan characteristics to predict which loans will be paid back in full and which will default?

DATASET

The Loan Data from 2007 to 2011 was downloaded, a file containing over 40,000 records. Each record is a distinct loan made on the Lending Club platform. For each loan, over 100 characteristics are recorded in the table. These characteristics can largely be divided into two groups - features of the loan, and features of the borrower. Let us look at the structure of the data to understand better.



```
## ...
##  $ purpose                     : chr  "credit_card" "car" "small_business" "other" ...
##  $ title                       : chr  "Computer" "bike" "real estate business" "personel"
 ...
##  $ zip_code                    : chr  "860xx" "309xx" "606xx" "917xx" ...
##  $ addr_state                  : chr  "AZ" "GA" "IL" "CA" ...
##  $ dti                         : num  27.65 1 8.72 20 17.94 ...
##  $ delinq_2yrs                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ earliest_cr_line            : chr  "Jan 85" "Apr 99" "Nov 01" "Feb 96" ...
##  $ inq_last_6mths              : int  1 5 2 1 0 3 1 2 2 0 ...
##  $ mths_since_last_delinq      : int  NA NA NA 35 38 NA NA NA NA NA ...
##  $ mths_since_last_record      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_acc                    : int  3 3 2 10 15 9 7 4 11 2 ...
##  $ pub_rec                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ revol_bal                   : int  13648 1687 2956 5598 27783 7963 17726 8221 5210 992
9 ...
##  $ revol_util                  : chr  "83.70%" "9.40%" "98.50%" "21%" ...
##  $ total_acc                   : int  9 4 10 37 38 12 11 4 13 3 ...
##  $ initial_list_status         : chr  "f" "f" "f" "f" ...
##  $ out_prncp                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ out_prncp_inv               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_pymnt                 : num  5863 1015 3006 12232 4067 ...
##  $ total_pymnt_inv             : num  5834 1015 3006 12232 4067 ...
##  $ total_rec_prncp             : num  5000 456 2400 10000 3000 ...
##  $ total_rec_int               : num  863 435 606 2215 1067 ...
##  $ total_rec_late_fee          : num  0 0 0 17 0 ...
##  $ recoveries                  : num  0 123 0 0 0 ...
##  $ collection_recovery_fee     : num  0 1.11 0 0 0 0 0 0 2.09 2.52 ...
##  $ last_pymnt_d                : chr  "Jan-15" "Apr-13" "Jun-14" "Jan-15" ...
##  $ last_pymnt_amnt             : num  171.6 119.7 649.9 357.5 67.3 ...
##  $ next_pymnt_d                : chr  NA NA NA NA ...
##  $ last_credit_pull_d          : chr  "Oct-17" "Oct-16" "Jun-17" "Apr-16" ...
##  $ collections_12_mths_ex_med  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mths_since_last_major_derog : chr  NA NA NA NA ...
##  $ policy_code                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ application_type            : chr  "Individual" "Individual" "Individual" "Individual"
 ...
##  $ annual_inc_joint            : chr  NA NA NA NA ...
##  $ dti_joint                   : chr  NA NA NA NA ...
##  $ verification_status_joint   : chr  NA NA NA NA ...
##  $ acc_now_delinq              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ tot_coll_amt                : chr  NA NA NA NA ...
##  $ tot_cur_bal                 : chr  NA NA NA NA ...
```

Some of the important variables include:

1. annual_inc: The annual income provided by the borrower during registration.
2. emp_length: Employment length in years. Possible values are between 0 and 10, where 0 means less than one year and 10 means ten or more years.
3. int_rate: Interest Rate on the loan
4. loan_amnt: The listed amount of the loan applied for by the borrower.
5. purpose: A category provided by the borrower for the loan request.

Response variable: loan_status ("Fully Paid" or "Charged Off")

DATA CLEANING

The following steps were carried out to clean the data and extract usable columns:

1. Removed columns with no data. Removed columns which consist of only one value throughout and certain other columns which had repetitive data.
2. Found out the percentage of Missing values in each column and removed columns with missing values greater than 50%.

3. Found the number of rows with missing values which approximately 3% of the entire dataset and omitted these observations. Removed loans which do not meet credit policy to avoid ambiguity

After carrying out these steps, was left with 38,971 observations.

HYPOTHESES

After gaining a basic understanding of the data I formulated hypotheses which will help me obtain variables that would determine whether a loan will be fully paid or will default.

1. The relationship between interest rate and risk of default in P2P lending is positive.
2. There is a negative relationship between loan size or loan amount and risk of default.
3. Purpose of taking loan is related to the probability of default in P2P lending. For example, higher probability of default when the loan is taken to finance a business.
4. Longer employment duration of the borrower, lower the probability of default.
5. Higher the debt-to-income of the borrower, higher the probability of default in P2P lending.
6. Lower probability of default if the borrower has a good credit history (Higher FICO rating is used to indicate this).

The last hypotheses could not be checked due to lack of data. The column representing credit history had very few observations and hence the test could not be carried out.

EXPLORATORY DATA ANALYSIS

Findings obtained from the basic summary statitics of the different attributes.

1. Most of the borrowers almost 92% are do not own homes, they are under rent or mortgage.
2. Higher number of loans are issued for a 30 month term than a 60 month term.
3. The number of 'Charged Off' loans are way lower when compared to 'Fully Paid' loans, thus we will need to oversample the data in order to balance the dataset.

Bi-variate analysis and correlation tests were carried out to understand the relationship between the different variables and loan status that helped prove or disprove the above listed hypothesis.

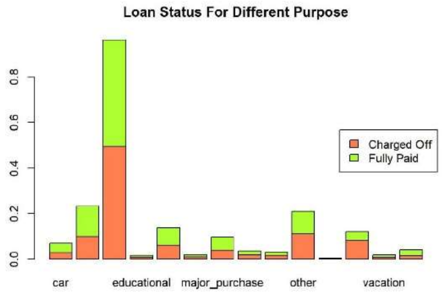


Fig1

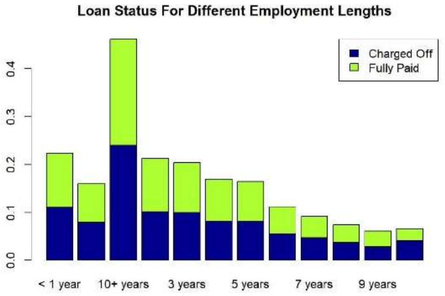


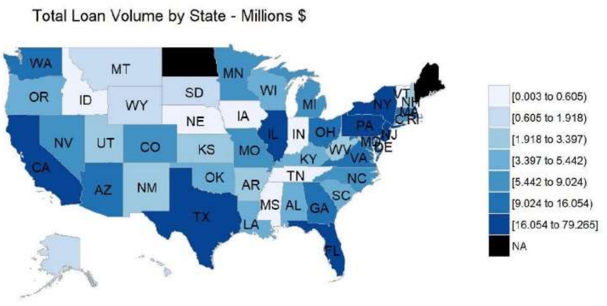
Fig2

Relationship between Loan Status and Purpose of Loan (fig 1)

From the chart above majority of the loans are taken to tackle 'debt_consolidation' of which almost half of the borrowers have defaulted.

Relationship between Loan Status and Employment Duration of the borrower (fig 2)

Maximum number of borrowers have an employment length of more than 10+ years. It can be noticed that there are quite many borrowers having <1-year employment duration



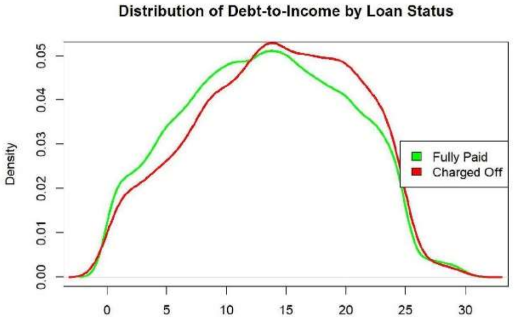


Fig 3

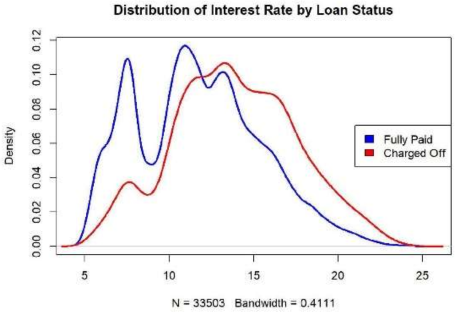
Fig 4

Total Loan Amount Volumes by State - to understand in which states the borrowers live in (fig 3)

More number of borrowers live in the states of California, Texas, Florida, Pennsylvania, New York and Illinois. This can be attributed to the fact that the population in these states are high compared to the rest of the country.

Debt to Income Ratio Distribution by Loan Status (fig 4)

Debt to Income ratio is almost normally distributed. The distribution of debt_to_income ratio is very similar for 'Fully Paid' and 'Charged Off' loans.



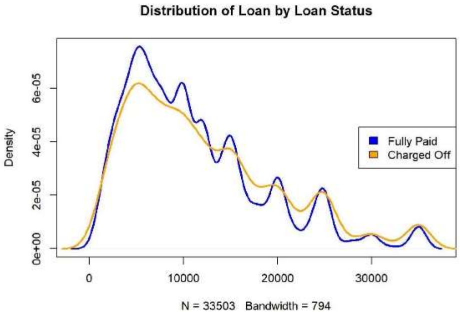


Fig 5

Fig 6

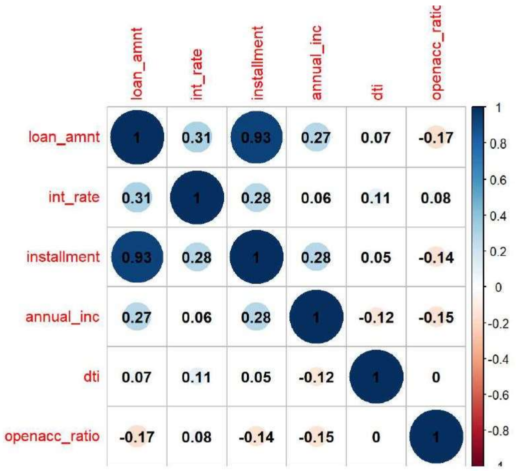
Interest Rate Distribution by Loan Status (fig 5)

Interest Rate is right skewed in the case of 'Fully Paid' loans whereas it is more spread for the Charged Off loans. This means that more number of 'Fully Paid' have lower interest rate compared to 'Charged Off' loans.

Size of Loan Distribution by Loan Status (fig 6)

It can be observed that higher number of loans have been issued for smaller loan amounts, thus the distribution is skewed. Also, the distribution dont vary a lot between 'Fully Paid' and 'Charged Off' loans.

Correlation Analysis



	loan_amnt	int_rate	installment	annual_inc	dti	openacc_ratio
loan_amnt	1	0.31	0.93	0.27	0.07	-0.17
int_rate	0.31	1	0.28	0.06	0.11	0.08
installment	0.93	0.28	1	0.28	0.05	-0.14
annual_inc	0.27	0.06	0.28	1	-0.12	-0.15
dti	0.07	0.11	0.05	-0.12	1	0
openacc_ratio	-0.17	0.08	-0.14	-0.15	0	1

From the correlation matrix we can identify that loan_amnt and installment has very high correlation, this is natural.

FEATURE ENGINEERING

1. Created an additional metric from the data that can affect the loan status. **openacc_ratio = total number of open credit lines/total number of credit lines**, this metric can assess the financial strength of the borrower, as in how much debt borrower has.

2. Merged categories of the factor variables - keeping the top five buckets (in terms of frequency) as such and combining the rest into one category 'Others'. Re-bucketed variables based on frequency as follows:

   i. emp_length variable as 10+ years, < 1 year , 1 year, 2-5 years, 6-9 years based on frequency shown below
   ii. purpose variable as debt_consolidation, credit_card, home_improvement, major_purchase, small_business, other
3. Variables created using One-Hot Encoding: The 'purpose' variable has 5 categories: debt_consolidation, credit_card, home_improvement, major_purchase, small_business, other. One hot coding will remove this variable and generate 5 new variables. Each will have binary numbers - 0 (if the category is not present) and 1(if category is present). The other variables which were used for One Hot Encoding are: emp_length having categories 10+ years, < 1 year, 1 year 2-5 years, 6-9 years and home_ownership having categories RENT, OWN, MORTGAGE and OTHER.

MODELLING

Dataset Creation

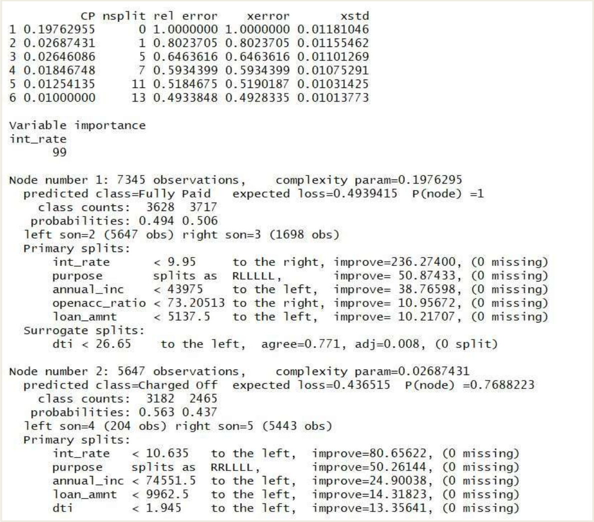
1. 70% of the data was taken for training purpose and 30% was considered for testing purpose.
2. Down sampled the 'Fully Paid' cases so that we can obtain a more balanced dataset with enough good and bad loan observations.

Decision Tree

Why Decision Tree?

Implemented decision trees first because it implicitly performs variable screening or feature selection for the given data, this wil help me identify important variables in the initial stages itself. Also, it handles categorical variables as such and it is easily interpretable.

Results from the model:



```
         CP nsplit rel error    xerror       xstd
1 0.19762955      0 1.0000000 1.0000000 0.01181046
2 0.02687431      1 0.8023705 0.8023705 0.01155462
3 0.02646086      5 0.6463616 0.6463616 0.01101269
4 0.01846748      7 0.5934399 0.5934399 0.01075291
5 0.01254135     11 0.5184675 0.5190187 0.01031425
6 0.01000000     13 0.4933848 0.4928335 0.01013773

Variable importance
int_rate
      99

Node number 1: 7345 observations,    complexity param=0.1976295
  predicted class=Fully Paid   expected loss=0.4939415  P(node) =1
    class counts:  3628  3717
   probabilities: 0.494 0.506
  left son=2 (5647 obs) right son=3 (1698 obs)
  Primary splits:
      int_rate      < 9.95     to the right, improve=236.27400, (0 missing)
      purpose       splits as  RLLLLL,       improve= 50.87433, (0 missing)
      annual_inc    < 43975    to the left,  improve= 38.76598, (0 missing)
      openacc_ratio < 73.20513 to the right, improve= 10.95672, (0 missing)
      loan_amnt     < 5137.5   to the left,  improve= 10.21707, (0 missing)
  Surrogate splits:
      dti < 26.65    to the left,  agree=0.771, adj=0.008, (0 split)

Node number 2: 5647 observations,    complexity param=0.02687431
  predicted class=Charged Off  expected loss=0.436515  P(node) =0.7688223
    class counts:  3182  2465
   probabilities: 0.563 0.437
  left son=4 (204 obs) right son=5 (5443 obs)
  Primary splits:
      int_rate   < 10.635   to the left,  improve=80.65622, (0 missing)
      purpose    splits as  RRLLLL,       improve=50.26144, (0 missing)
      annual_inc < 74551.5  to the left,  improve=24.90038, (0 missing)
      loan_amnt  < 9962.5   to the left,  improve=14.31823, (0 missing)
      dti        < 1.945    to the left,  improve=13.35641, (0 missing)
```

Inference

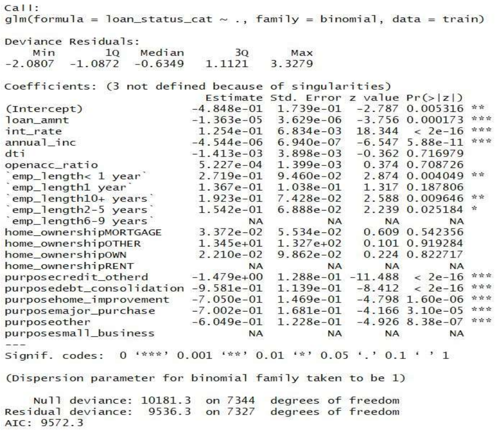
Interest_rate seems to be the most important variable which is used throught various levels of the tree, this can be inferred from tree diagram. Other important variables as indicated as in the summary include **purpose, annual_inc, openacc_ratio and loan_amnt**.

Logistic Regression

Why Logistic Regression?

On using logistic regression, I will be further able to understand the impact the various predictors on the dependent variables. Apart from knowing just the important variables, I can now understand whether there is a positive or negative relationship between the predictors and the dependent variable.

Results from the model



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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0807  -1.0872  -0.6349   1.1121   3.3279

Coefficients: (3 not defined because of singularities)
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -4.848e-01  1.739e-01  -2.787 0.005316 **
loan_amnt                 -1.363e-05  3.629e-06  -3.756 0.000173 ***
int_rate                   1.254e-01  6.834e-03  18.344  < 2e-16 ***
annual_inc                -4.544e-06  6.940e-07  -6.547 5.88e-11 ***
dti                       -1.413e-03  3.898e-03  -0.362 0.716979
openacc_ratio              5.227e-04  1.399e-03   0.374 0.708726
`emp_length< 1 year`       2.719e-01  9.460e-02   2.874 0.004049 **
`emp_length1 year`         1.367e-01  1.038e-01   1.317 0.187806
`emp_length10+ years`      1.923e-01  7.428e-02   2.588 0.009646 **
`emp_length2-5 years`      1.542e-01  6.888e-02   2.239 0.025184 *
`emp_length6-9 years`             NA         NA      NA       NA
home_ownershipMORTGAGE     3.372e-02  5.534e-02   0.609 0.542356
home_ownershipOTHER        1.345e+01  1.327e+02   0.101 0.919284
home_ownershipOWN          2.210e-02  9.862e-02   0.224 0.822717
home_ownershipRENT                NA         NA      NA       NA
purposecredit_otherd      -1.479e+00  1.288e-01 -11.488  < 2e-16 ***
purposedebt_consolidation -9.581e-01  1.139e-01  -8.412  < 2e-16 ***
purposehome_improvement   -7.050e-01  1.469e-01  -4.798 1.60e-06 ***
purposemajor_purchase     -7.002e-01  1.681e-01  -4.166 3.10e-05 ***
purposeother              -6.049e-01  1.228e-01  -4.926 8.38e-07 ***
purposesmall_business             NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10181.3  on 7344  degrees of freedom
Residual deviance:  9536.3  on 7327  degrees of freedom
AIC: 9572.3
```

Inference

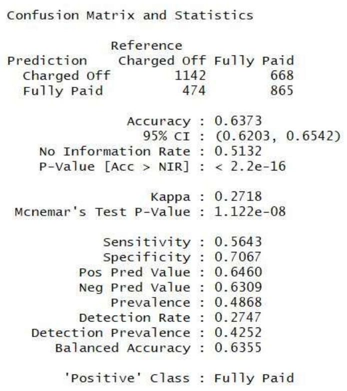
1. From the summary it can be observed that the most significant variables are loan amount, interest rate, annual_income and purpose of loan.

2. Interest rate, open_acc_ratio, employment_length and home_ownership have positive coefficients, which means the odds of loan being defaulted increases by the exponential of the corresponding beta estimate. Whereas, for annual_income, loan_amount, dti and purpose the cofficient is negative indicating that the odds of loan being defaulted is lowered by the exponential of the corresponding beta estimate. The model gives a slightly lower accuracy of 62.6% when compared to decision tree method.

Support Vector Machines

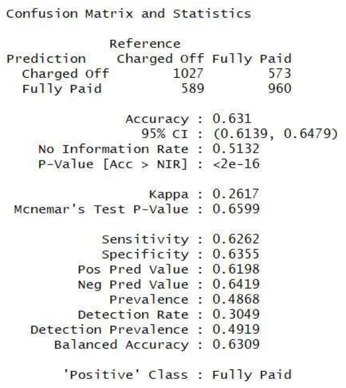
Why SVM?

SVM is the third model I chose to implement. SVM is intrinsically suited for two-class problems. Implemented classification svm with radial, linear and polynomial kernel.



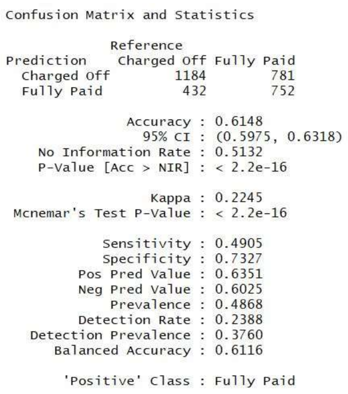
```
Confusion Matrix and Statistics

             Reference
Prediction    Charged Off Fully Paid
  Charged Off        1142        668
  Fully Paid          474        865

               Accuracy : 0.6373
                 95% CI : (0.6203, 0.6542)
    No Information Rate : 0.5132
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2718
 Mcnemar's Test P-Value : 1.122e-08

            Sensitivity : 0.5643
            Specificity : 0.7067
         Pos Pred Value : 0.6460
         Neg Pred Value : 0.6309
             Prevalence : 0.4868
         Detection Rate : 0.2747
   Detection Prevalence : 0.4252
      Balanced Accuracy : 0.6355

       'Positive' Class : Fully Paid
```

Radial



```
Confusion Matrix and Statistics

             Reference
Prediction    Charged Off Fully Paid
  Charged Off        1027        573
  Fully Paid          589        960

               Accuracy : 0.631
                 95% CI : (0.6139, 0.6479)
    No Information Rate : 0.5132
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.2617
 Mcnemar's Test P-Value : 0.6599

            Sensitivity : 0.6262
            Specificity : 0.6355
         Pos Pred Value : 0.6198
         Neg Pred Value : 0.6419
             Prevalence : 0.4868
         Detection Rate : 0.3049
   Detection Prevalence : 0.4919
      Balanced Accuracy : 0.6309

       'Positive' Class : Fully Paid
```

Linear



```
Confusion Matrix and Statistics

             Reference
Prediction    Charged Off Fully Paid
  Charged Off        1184        781
  Fully Paid          432        752

               Accuracy : 0.6148
                 95% CI : (0.5975, 0.6318)
    No Information Rate : 0.5132
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2245
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4905
            Specificity : 0.7327
         Pos Pred Value : 0.6351
         Neg Pred Value : 0.6025
             Prevalence : 0.4868
         Detection Rate : 0.2388
   Detection Prevalence : 0.3760
      Balanced Accuracy : 0.6116

       'Positive' Class : Fully Paid
```

Polynomial

Inference

Since the given dataset does not have a lot of features compared to the size of the training sample, linear SVM is not a very apt method. Radial SVM has resulted in highest accuracy which is 63.7% when compared to linear and polynomial kernels.

INFERENCE & INSIGHTS

1. From the decision tree method as well as the logistic regression method we can come to the conclusion that **Interest Rate** is the most important factor which helps us to determine whether a loan will get charged off.

2. High interest rate, open account ratio increases the odds of a loan being charged off. Also, borrowers who have housing mortgage or employment length < 1 year have higher odds of defaulting on their loan payment.

3. As per the analysis conducted using different techniques the factors that we need to consider for to detect default loan cases are interest_rate, loan_amount, employment_length, home_ownership, purpose and annual income

4. The classification model built using decision tree gives highest accuracy of 75% when compared to logistic regression (62.6%) and SVM with radial kernel (63.7%).