South German credit dataset

Content:

Introduction to Dataset:

Loan is one of the vital asset for the banks. As planning the credit risk efficiently is good solution to the financial distresses, which is a major factor for the banks, in case of failure of repayments. As there can many reasons for the failure of repayment of loan or credits borrowed from the bank, we are going to research how certain features affect the credit risk of the individual. Based on it, two research question were formed:

Research Question 1: how amount of loan affect the credit risk of different age individuals, along with their current checking account status?

Research Question2: How the non-bank related variables affect the credit risk of the individual?

This dataset 'SouthGermanCredit'[1] has been taken from the UCI Machine learning respiratory. The dataset was updated last on 20-June-2020. The dataset contains financial and non-financial attributes to determine the credit risk of individuals from the 1000 rows and 21 variables.

Methodology:

Data Preprocessing & Exploratory Data Analysis:

Before proceeding in our research, we started cleaning and preprocessing our data. As we don't have any missing or empty data, so we didn't need to handle them. There was 3 numeric variable and 17 categorical features. But all our categorical data was given in numeric term 0-5 for all columns and where in every column, each express different category. Along the csv file, there was text file given which explained all categorical features values. We converted the numeric category data to their real values and factors. The column description can be found in **Appendix A, table 1.**

As exploring the dataset, we plot the for 'amount' vs. 'duration'(**Plot 1**) of credit with best fit line, which showed the constant increasing linear relation between them. It can been seen that they was highly correlated, which lead us to remove the variable 'duration'.

As further exploration of our dataset brings us to our target variable, which is 'credit risk'. Upon checking the instances of the binary response variable, we came to know that our dataset is imbalances to 700:300 for response variable.

Model Preface:

We split the data into 7:3 ratio for train and test set respectively.

After data processing, the final dataset was named 'cleaned_credit_risk'. For each research question we choose different model i.e. for research question 1 we implied Logistic Regression as well as SVM (which is our third model choice) and for research question 2 we implied KNN model.

Results:

Note: results from R output can found in Appendix B

Model 1: Logistic regression

As we planned to implement the model logistic regression for research question 1 to predict the dependable variable ‘credit risk’ for individuals based on their age, checking account status and credit history. We decided to see data through visualization to gain in-depth knowledge of what our data represents. We plot the age variable against the amount variable and filled it with credit risk (**Appendix A: plot2: age vs amount**). From plot, it is visible that as age grows, the credit risk get better. Whereas, with high amount of credit in young age people have high proportion of bad credit risk. It is safe to assume that the young people are more risky than old aged people in terms of credit risk for banks. One of the reasons can be the unstable job or low saving and income.

Whereas, if we check the checking account status against credit risk, it shows that proportion of bad or good credit risk differs in every subcategory of checking account status.(**plot3:Checking Account status vs Credit risk**)

Next we check the age column against the credit risk, by plotting violin plot. (**plot4: Age vs Credit risk**). It is visible that median, first quartile and third quartile of bad credit risk is lower than good credit risk. Which describe the same that young people seem to be more risky than mature people.

Next we implement the model logistic regression using package caret. As our target variable is imbalance, we use the sampling ‘up’ in train control. Along with it we set 5 fold cross validation and used Precision & Recall summary function. While we also implied the preprocessing to center and scale the numeric data in model fitting. And we optimized the model for AUC. The train model is called ‘Credit_risk_LR’.

The model resulted in 64 percent on test set. Due to ‘up’ sampling techniques, our model not result in predicting mostly credit risk as good can been seen in confusion matrix ‘ConfusionM_logit’.

As our Balanced accuracy is 0.673 % and our Precision & Recall is 0.849 and 0.59 respectively. As from summary, the coefficient of age and amount is 0.21631 and -0.45652 respectively, which show each unit grow in age make the credit risk is less, whereas when amount grow the each unit it negatively affect the credit risk which make it riskier as high amount mean higher risk.

Model2: KNN

Implement the KNN model to our research question 2: how non-bank related variables affect the credit risk of an individual? To predict credit risk based on the non-bank related variable (‘age’, ‘personal_status_sex’, ‘employment_duration’, ‘property’, ‘housing’, and ‘job’, ‘foreign_worker’)

We need to check by visualization which variable affect the credit risk. We checked first by plotting for personal status sex against credit risk (**Plot5: personal_status_sex Vs. credit_risk**) as it can been seen, that proportion of credit risk categories is very much equal in male:divorced/separated and where as in every category of personal status sex, the proportion of category credit risk differs, which represent that every category affect in different way.

As when we plot Employment duration against credit risk (**plot6: Employment_duration vs. Credit risk**), it is clearly visible how being unemployed or employment duration being low increase the risk of credit loss for the bank. As we see constant increase in number of years in employment reduce the risk. Next we checked for how being foreign worker affect the credit risk to bank. As from plot foreign worker Vs. Credit risk (**plot7: Foreign worker vs. credit risk**), it can been seen that being foreign worker makes you likely the credit risk to bank, but it can biased as the number of instances for being foreign worker is really less compare to resident.

As we implement Knn, we set 10 fold repeated cross validation with repeat of 3 and applied ‘down’ sampling to deal with class imbalance. We tested k values of range from 1 to 28 to tune the hyper parameter K and we optimized the model for AUC with summary function of Precision and Recall. The train model is called ‘knn_credit_risk_prediction’. After hyper parameter tuning, the best value for K is selected 21 by our model. The accuracy of the model is 0.55 where the balanced accuracy is 0.5325. Following Precision and Recall is 0.724 and 0.576 respectively.

Model3: SVM (Support vector Machine)

For model 3, we implement SVM(support vector machine) on research question 2 to predict the To predict credit risk based on the non-bank related variable (‘age’, ‘personal_status_sex’, ‘employment_duration’, ‘property’, ‘housing’, ‘job’, ‘foreign_worker’).

Where we tuned the hyper parameter C by testing model on different values of it via tune grid and tune length of 10. We set 5 repeated fold cross validation with repeat of 3 and implied ‘up’ sampling technique. The fitted model is called ‘svm_Linear’, where we see the best accuracy of 0.647 at the value of C (0.95) **Plot8.** The test set accuracy is 0.55 and balanced accuracy is 0.5262 which is very similar to our KNN model. Where Precision and Recall is 0.719 and 0.5857 respectively, also very similar to our KNN precision and recall.

Conclusion:

In conclusion, we would like to include that age, amount of the credit to take does affect the credit risk of the individual. Whereas, along with them there are other situations for example employment duration, personal status as well as savings and checking account status affect the credit risk negatively. Other important thing to notice is that our data is being sensitive to class imbalance as in test results, if sampling was not used, the model mostly predict 80-90%of cases to majority class ‘good’ in our target variable credit_risk. Due to limit of packages and dataset as majority were categorical variables, model could not perform better and we could not improve the sampling technique. As other packages i.e. smote from Smote family require numeric data.

References:

1. **Wickham, H. [2021]. CRAN - Package tidyr . link: https://cran.r-project.org/web/packages/tidyr/index.html**
2. **Kuhn.M[2021].CRAN – package caret, link: https://cran.r-project.org/web/packages/caret/index.html**
3. **Takahashi, K, Wickham, H. CRAN - Package ggplot2. Link: https://cran.r-project.org/web/packages/ggplot2/index.html**
4. **Henry, L., & Müller, K. [August 2020] CRAN - Package dplyr, Link: https://cran.r-project.org/web/packages/dplyr/index.html**

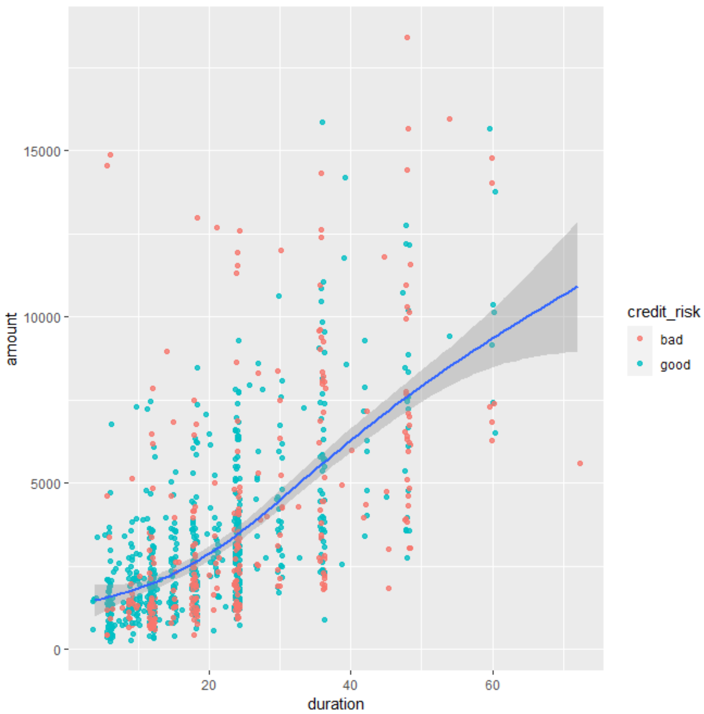
Appendix A:

Variables	Description

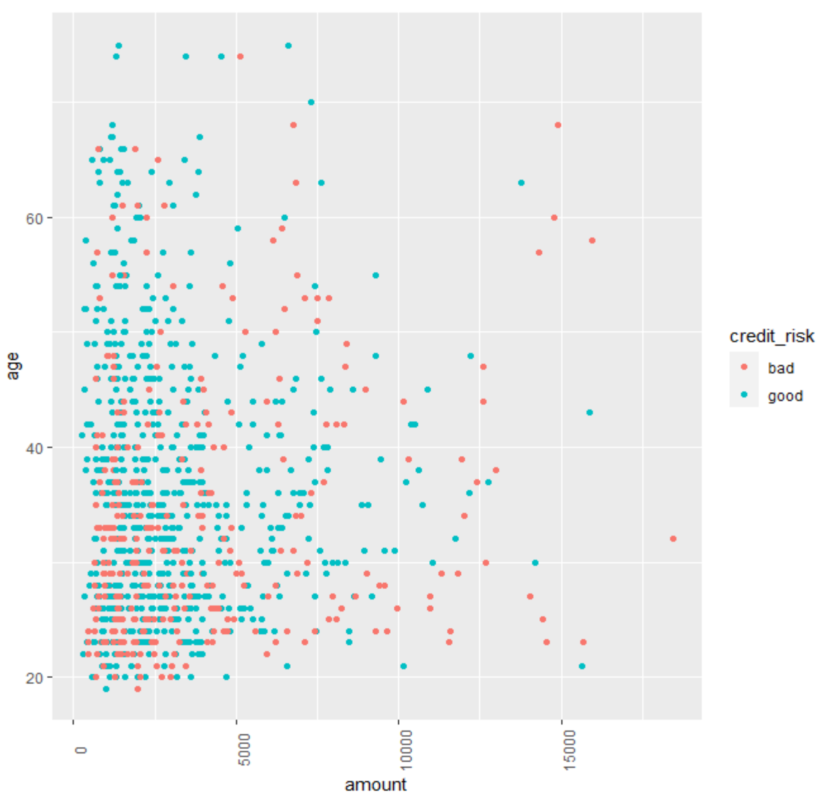
checking_acc_staus	status of the debtor's checking account with the bank
Duration	credit duration in months
Credit_history	history of compliance with previous or concurrent credit contracts
Purpose	purpose for which the credit is needed
Amount	credit amount in DM
Savings	debtor's savings
Employment Durtion	duration of debtor's employment with current employer
Installment_ Rate	credit installments as a percentage of debtor's disposable income
Personal_status_sex	Status according to male and females
Other_debtors	Is there another debtor or a guarantor for the credit?
Present_residence	length of time (in years)
Property	the debtor's most valuable property
Age	age in years
Other_installment_ Plans	installment plans from providers other than the credit-giving bank
Housing	type of housing the debtor lives in
Number_credits	number of credits including the current one the debtor has (or had) at this bank
Job	quality of debtor's job
People_liable	number of persons who financially depend on the debtor
Telephone	Is there a telephone landline registered on the debtor's name?
Foreign_worker	Is the debtor a foreign worker?
Credit_risk	Has the credit contract been complied with (good) or not (bad) ?

Table1: Data Description

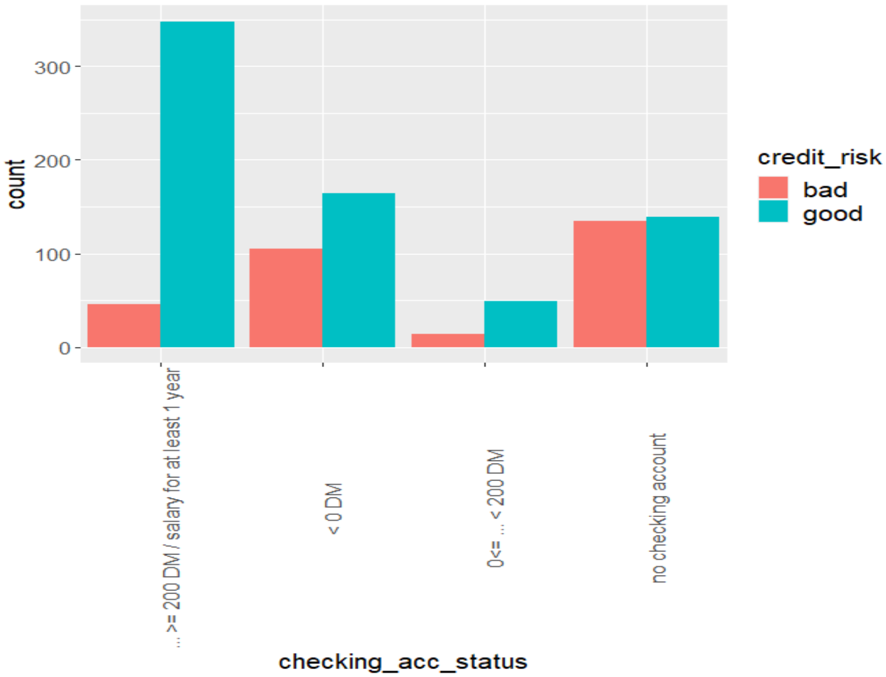
Plot1: Amount Vs Duration by Credit Risk



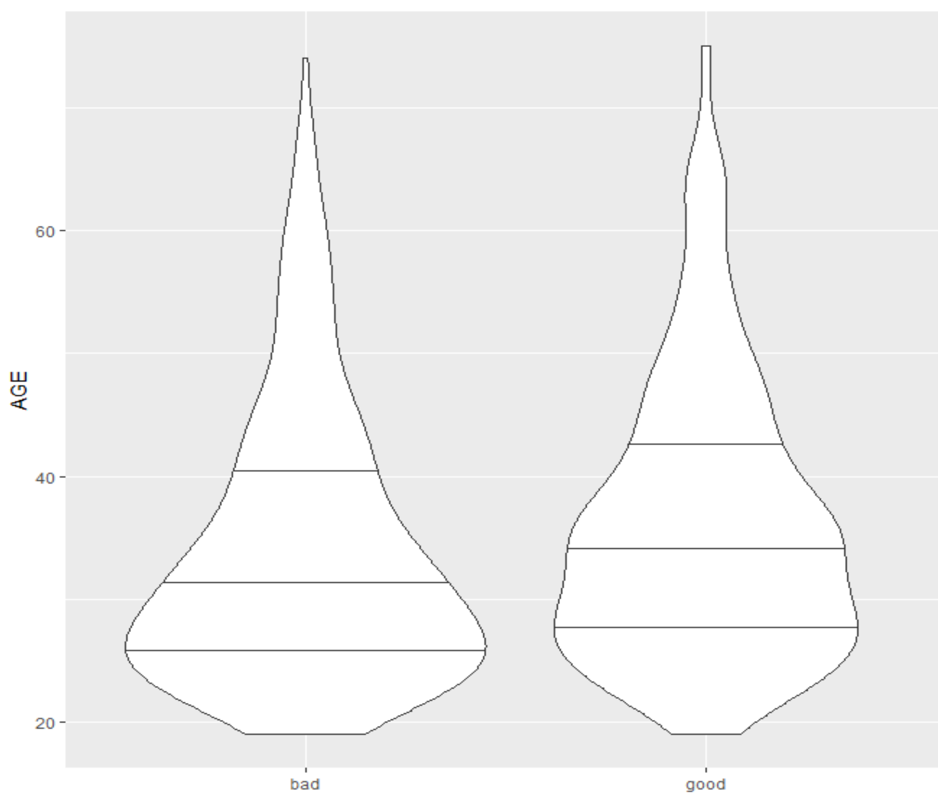
Plot2: Age Vs. Amount by Credit risk



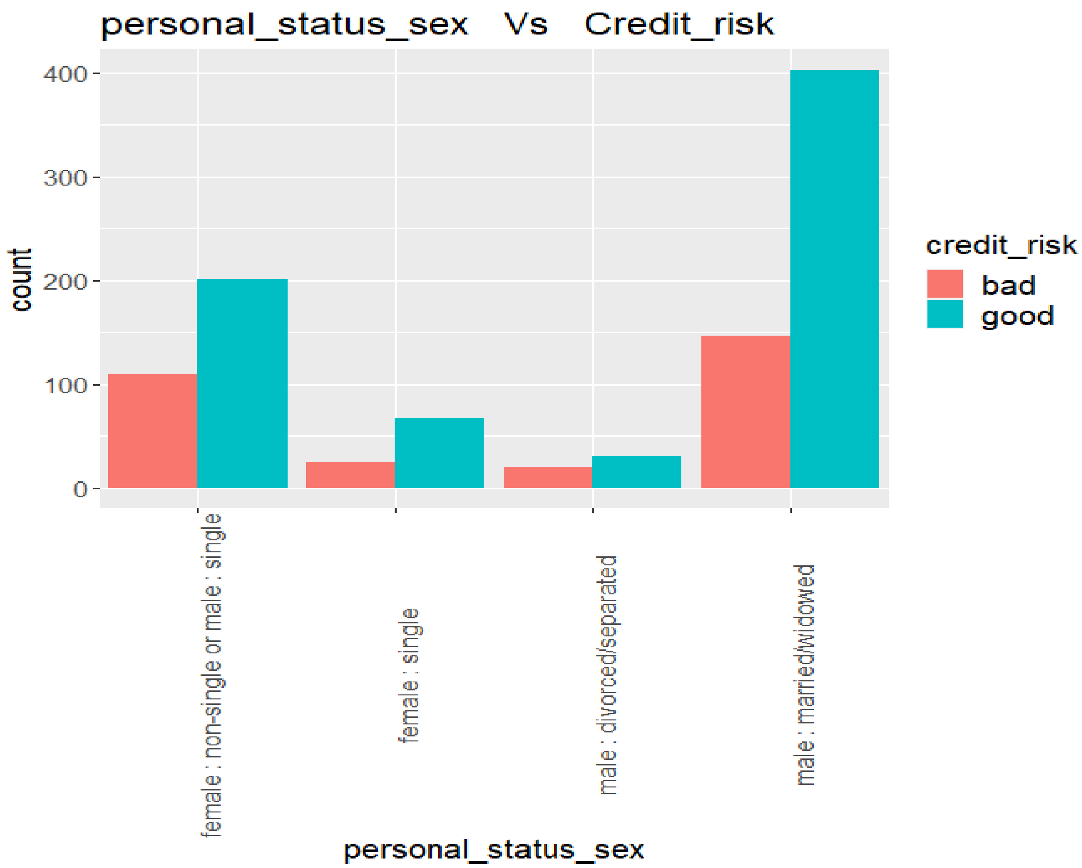
Plot3: Checking account status vs. Credit Risk



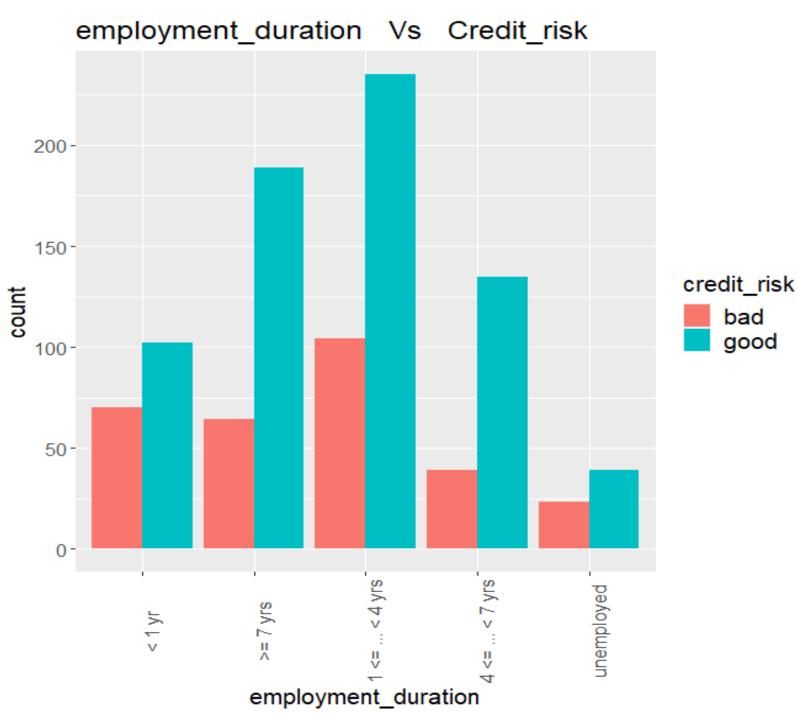
Violin Plot4: Age Vs. Credit Risk



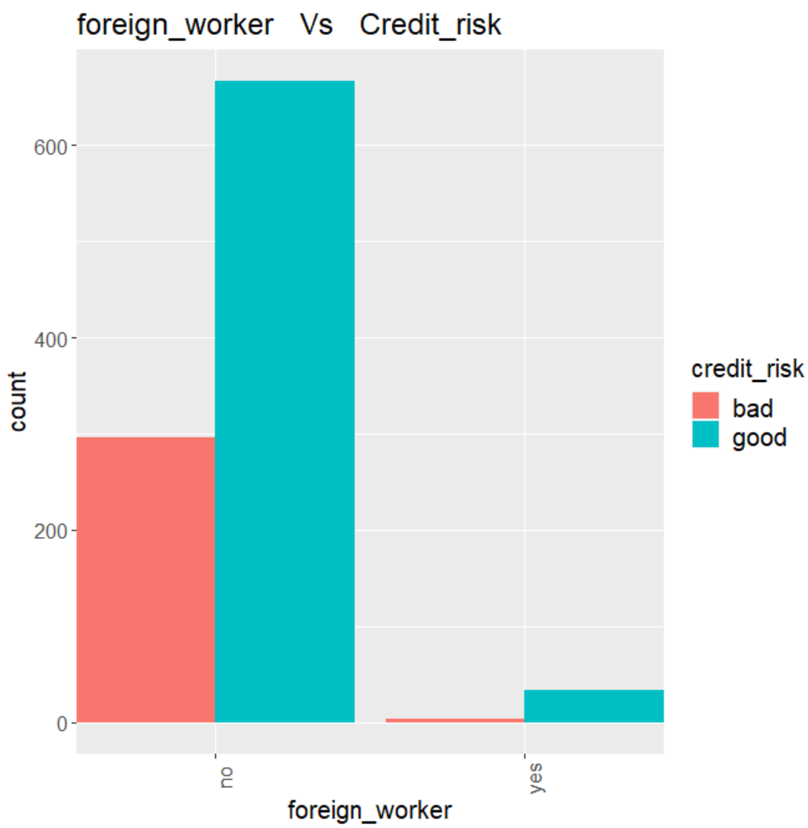
Plot5: Personal Status Sex Vs. Credit Risk



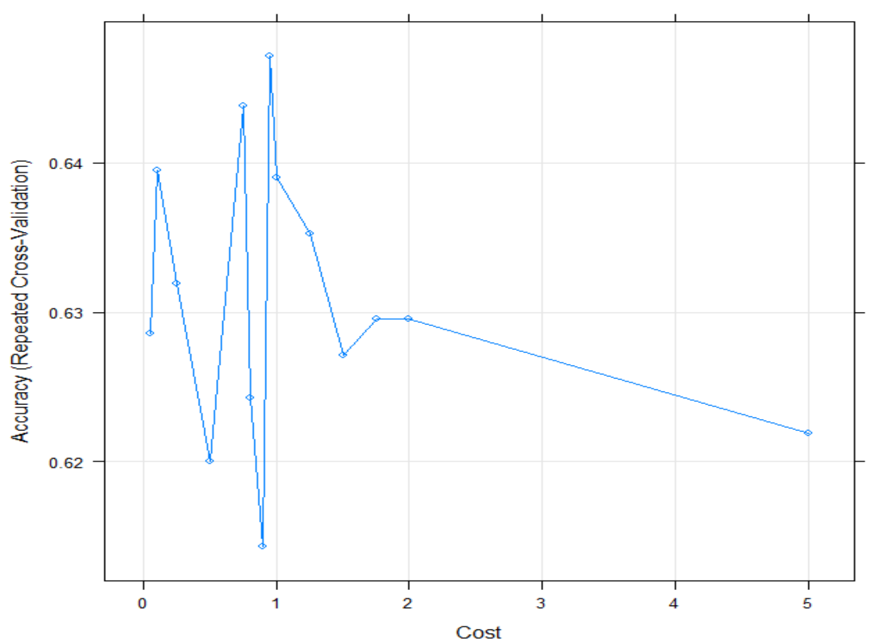
Plot6: Employment Duration Vs. Credit Risk



Plot7: Foreign Worker vs. Credit Risk



Plot8: SVM hyper parameter tuning



Appendix B: Output Results

Model 1: Logistic Regression

Table1: Summary

```
Call:
NULL

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.12708  -0.89093   0.02707   0.86043   2.28544

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                    0.03031    0.07437   0.408 0.683602
age                                            0.21631    0.07510   2.880 0.003972 **
`checking_acc_status< 0 DM`                   -0.81180    0.08867  -9.155  < 2e-16 ***
`checking_acc_status0<= ... < 200 DM`         -0.23626    0.07380  -3.201 0.001367 **
`checking_acc_statusno checking account`      -0.95470    0.08955 -10.661  < 2e-16 ***
amount                                        -0.45652    0.08484  -5.381 7.42e-08 ***
`credit_historycredits paid back`             -0.29136    0.08204  -3.552 0.000383 ***
`credit_historycritical account`              -0.25625    0.07874  -3.255 0.001136 **
credit_historydelays                          -0.19119    0.08454  -2.262 0.023727 *
`credit_historynocredits taken/all paid back fully` -0.28958    0.08854  -3.270 0.001074 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1358.6  on 979  degrees of freedom
Residual deviance: 1101.9  on 970  degrees of freedom
AIC: 1121.9

Number of Fisher Scoring iterations: 4
```

Table2: Confusion Matrix & test set results

```
Confusion Matrix and Statistics

          Reference
Prediction bad good
      bad   68   86
      good  22  124

               Accuracy : 0.64
                 95% CI : (0.5828, 0.6944)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.9892

                  Kappa : 0.2876

 Mcnemar's Test P-Value : 1.343e-09

            Sensitivity : 0.5905
            Specificity : 0.7556
         Pos Pred Value : 0.8493
         Neg Pred Value : 0.4416
             Prevalence : 0.7000
         Detection Rate : 0.4133
   Detection Prevalence : 0.4867
      Balanced Accuracy : 0.6730

       'Positive' Class : good

> confusionM_logit$byClass[c('Precision','Recall')]
Precision    Recall
0.8493151 0.5904762
```

Model 2: KNN

Table3: summary KNN

```
k-Nearest Neighbors

700 samples
  7 predictor
  2 classes: 'bad', 'good'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 630, 630, 630, 630, 630, 630, ...
Addtional sampling using down-sampling

Resampling results across tuning parameters:

  k   AUC        Precision  Recall     F
   1  0.1971815  0.3386011  0.5142857  0.4060033
   2  0.2808129  0.3622267  0.5555556  0.4367086
   3  0.3288918  0.3645896  0.5904762  0.4489641
   4  0.3527727  0.3613586  0.5714286  0.4413139
   5  0.3624111  0.3777790  0.5984127  0.4605562
   6  0.3582585  0.3868407  0.5936508  0.4664605
   7  0.3667280  0.3727274  0.5714286  0.4494239
   8  0.3615268  0.3781953  0.5587302  0.4495453
   9  0.3673400  0.3847014  0.5682540  0.4563811
  10  0.3671246  0.3855567  0.5650794  0.4558476
  11  0.3551861  0.3622450  0.5365079  0.4312873
  12  0.3569107  0.3790358  0.5507937  0.4476455
  13  0.3758037  0.3795043  0.5587302  0.4498613
  14  0.3656405  0.3671803  0.5349206  0.4326392
  15  0.3670034  0.3667006  0.5222222  0.4288674
  16  0.3812195  0.3742928  0.5507937  0.4434033
  17  0.3736653  0.3753825  0.5539683  0.4443877
  18  0.3626428  0.3780748  0.5476190  0.4450571
  19  0.3694288  0.3655310  0.5190476  0.4263176
  20  0.3700991  0.3621181  0.5333333  0.4287269
  21  0.3820340  0.3811323  0.5571429  0.4504439
  22  0.3680978  0.3709877  0.5285714  0.4315080
  23  0.3615890  0.3553016  0.5253968  0.4216255
  24  0.3625203  0.3718531  0.5380952  0.4363066
  25  0.3627623  0.3658599  0.5365079  0.4325778
  26  0.3675870  0.3639773  0.5253968  0.4264191
  27  0.3621267  0.3559288  0.5190476  0.4179556
  28  0.3686095  0.3594748  0.5301587  0.4256020

AUC was used to select the optimal model using the largest value.
The final value used for the model was k = 21.
```

Table4: Confusion Matrix & test set results

```
Confusion Matrix and Statistics

          Reference
Prediction bad good
      bad   44   89
      good  46  121

               Accuracy : 0.55
                 95% CI : (0.4918, 0.6072)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.0000000

                  Kappa : 0.0573

 Mcnemar's Test P-Value : 0.0003006

            Sensitivity : 0.5762
            Specificity : 0.4889
         Pos Pred Value : 0.7246
         Neg Pred Value : 0.3308
             Prevalence : 0.7000
         Detection Rate : 0.4033
   Detection Prevalence : 0.5567
      Balanced Accuracy : 0.5325

       'Positive' Class : good

> confusionM_knn$byClass[c('Precision','Recall')]
Precision    Recall
0.7245509 0.5761905
```

Model3: SVM

<u>Table5: summary SVM</u>

```
Support Vector Machines with Linear Kernel

700 samples
  7 predictor
  2 classes: 'bad', 'good'

Pre-processing: centered (17), scaled (17)
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 560, 560, 560, 560, 560, 560, ...
Addtional sampling using up-sampling prior to pre-processing

Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.05  0.6285714  0.2071241
  0.10  0.6395238  0.2135220
  0.25  0.6319048  0.2155792
  0.50  0.6200000  0.1812637
  0.75  0.6438095  0.2149937
  0.80  0.6242857  0.1881393
  0.90  0.6142857  0.1840474
  0.95  0.6471429  0.2247666
  1.00  0.6390476  0.2165095
  1.25  0.6352381  0.2095536
  1.50  0.6271429  0.2038208
  1.75  0.6295238  0.2071241
  2.00  0.6295238  0.2070327
  5.00  0.6219048  0.1847691

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was C = 0.95.
```

<u>Table6: Confusion Matrix & test set results</u>

```
Confusion Matrix and Statistics

          Reference
Prediction bad good
      bad   42   87
      good  48  123

               Accuracy : 0.55
                 95% CI : (0.4918, 0.6072)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.000000

                  Kappa : 0.0466

 Mcnemar's Test P-Value : 0.001074

            Sensitivity : 0.5857
            Specificity : 0.4667
         Pos Pred Value : 0.7193
         Neg Pred Value : 0.3256
             Prevalence : 0.7000
         Detection Rate : 0.4100
   Detection Prevalence : 0.5700
      Balanced Accuracy : 0.5262

       'Positive' Class : good

> confusionM_SVM$byClass[c('Recall','Precision')]
   Recall Precision
0.5857143 0.7192982
```