STAT 387: German Credit Data Analysis

GROUP 5

Kamerin Vesajd & Louis Tu

03-01-2023

Table of Contents

1) Introduction

- Overview

When a bank receives a loan application, they need to assess whether to approve it or not based on the applicant's profile. This decision carries two risks:

- Declining a loan to a good credit risk resulting in lost business for the bank
- Approving a loan to a bad credit risk resulting in financial loss for the bank.

As a preface, we will analyze the German Credit Data set itself to validate that there's no outliers. Then we will resume analyzing the socio-economic characteristics in relation to categorical response, Default *(1 = "Defaulted", 0 = "Not Defaulted")* of the German Credit Data set. To accurately analyze loan applicants' socio-economic factors we will observe the different methods to find the optimal classification models which can help improve accuracy of loan approval decisions and minimize risk of approving loans to individuals who are more likely to default. Ultimately this will help reduce losses and improve profitability for the bank.

- Analysis Objective

To minimize the bank's potential losses by evaluating the demographic and socio-economic characteristics of loan applicants in order to determine whether to approve or reject their loan application.

- Finding the best model to help accurately assess and predict socio-economic characteristics of bank loan applicants

- Problem Statements

(a) Perform an exploratory analysis of data.

(b) Build a reasonably “good” logistic regression model for these data. There is no need to explore interactions. Carefully justify all the choices you make in building the model.

(c) Write the final model in equation form. Provide a summary of estimates of the regression coefficients, the standard errors of the estimates, and 95% confidence intervals of the coefficients. Interpret the estimated coefficients of at least two predictors. Provide training error rate for the model.

(d) Fit a KNN with K chosen optimally using test error rate. Report error rate, sensitivity, specificity, and AUC for the optimal KNN based on the training data. Also, report its estimated test error rate.

(e) Repeat (d) using LDA.

(f) Repeat (d) using QDA.

(g) Repeat (d) using Naïve Bayes Classifier.

(h) Compare the results in (b), (d)-(f). Which classifier would you recommend? Justify your answer.

- Methodologies

Exploratory Analysis

- Import data
- Check for missing values & outliers
- Examine data distribution via. GGPlot
- Check for Multi-collinearity

Building Logistic Regression Model

- Generalized Linear Model
- Step-AIC
  - Start’s with the initial model then adds or remove predictors one at a time based on its AIC value. The goal is selecting the best subset of predictor variables in the model when the number of potential predictors is large.

Data Modeling

- (KNN) K-Nearest Neighbors
  - $\hat{y}$ is the predicted class
  - y_i is the class label of the ith nearest neighbor
  - $I(\cdot)$ is the indicator function that evaluates to 1 if the argument is true, and 0 otherwise

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} I\left(y_i = y\right)$$

  - A machine learning algorithm that predicts the output of a test point based on the output of its nearest neighbors in the training set. The algorithm can be used for classification and regression tasks and the choice of k affects its accuracy.

- (LDA) Linear Discriminant Analysis
  - w represents the linear discriminant function that separates the classes
  - S_W represents the within-class scatter matrix
  - m_1 and m_2 represent the means of the two classes
  - ||·|| denotes the Euclidean norm

$$w = \frac{{S_W}^{-1}\left(m_2 - m_1\right)}{\|{S_W}^{-1}\left(m_2 - m_1\right)\|}$$

  - High Dimensionality reduction technique that works by projecting high-dimensional data onto a lower-dimensional space while maximizing the separation between the classes. Goal of LDA is to find the linear combination of features that best separates the classes.
- QDA
  - $p(y = k|x)$ represents the probability that the input x belongs to class
  - k, μ_k represents the mean vector of class k
  - Σ_k represents the covariance matrix of class k
  - |·| represents the determinant of the matrix

$$p(y = k|x) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_k\right|^{1/2}} exp\left(-\frac{1}{2}\left(x - \mu_k\right)^T \Sigma_k^{-1}\left(x - \mu_k\right)\right)$$

  - A classification algorithm that works by modeling the probability distribution of each class using a quadratic function. Unlike LDA, QDA assumes that the covariance matrix of each class is different. The goal of QDA is to find the parameters that maximize the likelihood of the training data given the class labels.
- Naive Bayes
  - $p(y = k|x)$ represents the posterior probability that the input x belongs to class k

  - $p(x|y = k)$ represents the likelihood of the input given that it belongs to class k
  - $p(y = k)$ represents the prior probability of class k
  - $p(x)$ represents the marginal probability of the input

$$p(y\ =\ k|x) = \frac{p(x|y=k)p(y=k)}{p(x)}$$

- A probabilistic classification algorithm that is based on Bayes' theorem. It is called "naive" because it assumes that the features are conditionally independent given the class label, which may not be true in reality. Despite this simplifying assumption, Naive Bayes can perform well on a wide range of classification tasks. Goal of Naive Bayes is to find the class label that maximizes the posterior probability given the input.

2) Exploratory Analysis of the Data

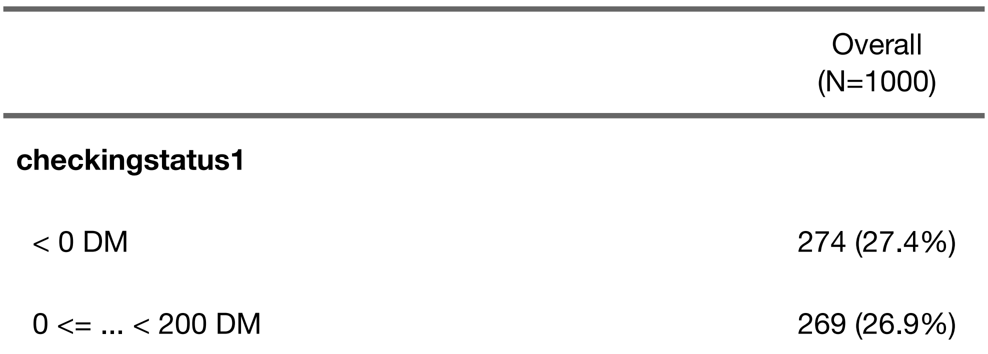
- Data description

We imported 1000 observations (1000 loan applicants) from the German Credit Data. In this data set there are 6 Numerical Variables and 15 Categorical variables where the Response 'Default' is a Categorical with the levels of 1 and 0. The response 'Default' is indicate of whether the applicant has paid their loan back yet. 1 (Defaulted) represents that they haven't paid yet and 0 (Not Defaulted) representing that they have.

- Numerical Predictors:

- Duration (Months): Length of time the individual is expected to take to repay the credit.
- Amount: Credit amount requested (DM)
- Residence: Length of time lived in current residence
- Age: Age of applicant by years
- Cards: Numbers of existing credit cards at the bank
- Liable: Number of people liable to provide maintenance for

- Categorical Predictors:



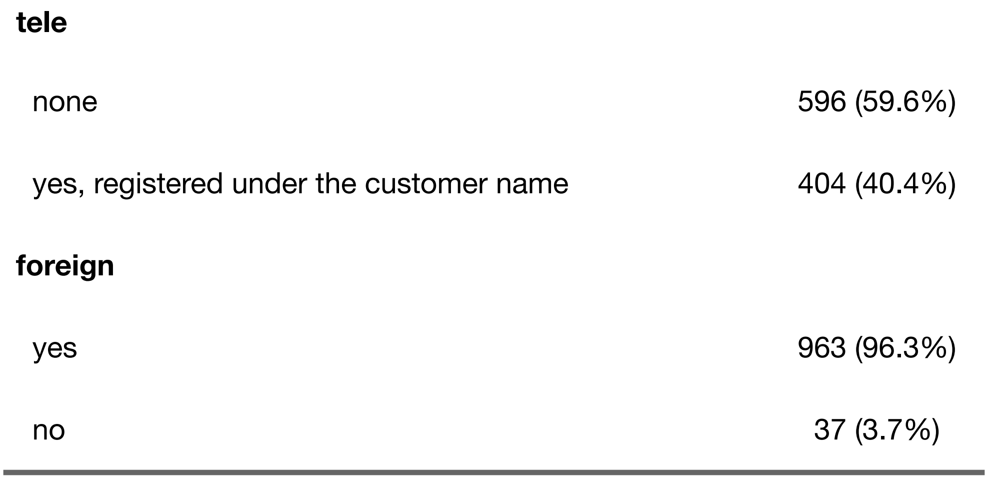
	Overall (N=1000)
checkingstatus1	
< 0 DM	274 (27.4%)
0 <= ... < 200 DM	269 (26.9%)

	Overall (N=1000)
... >= 200 DM or salary assignment	63 (6.3%)
no checking account	394 (39.4%)
history	
no credits taken/all credits paid back duly	40 (4.0%)
all credits at this bank paid back duly	49 (4.9%)
existing credits paid back duly till now	530 (53.0%)
delay in paying off in the past	88 (8.8%)
critical account/other credits existing (not at this bank)	293 (29.3%)
savings	
< 100 DM	603 (60.3%)
>= 100 ... < 500 DM	103 (10.3%)
>= 500 ... < 1000 DM	63 (6.3%)
>= 1000 DM	48 (4.8%)
unknown/no savings account	183 (18.3%)
purpose	
car (new)	234 (23.4%)
car (used)	103 (10.3%)
furniture/equipment	181 (18.1%)
radio/television	280 (28.0%)
domestic appliances	12 (1.2%)

	Overall (N=1000)
repairs	22 (2.2%)
education	50 (5.0%)
retraining	9 (0.9%)
business	97 (9.7%)
others	12 (1.2%)
employ	
unemployed	62 (6.2%)
< 1 year	172 (17.2%)
>= 1 year ... < 4 years	339 (33.9%)
>= 4 years ... < 7 years	174 (17.4%)
>= 7 years	253 (25.3%)
status	
male : divorced/separated	50 (5.0%)
female : divorced/separated/married	310 (31.0%)
male : single	548 (54.8%)
male : married/widowed	92 (9.2%)
female : single	0 (0%)
others	
none	907 (90.7%)
co-applicant	41 (4.1%)

guarantor	52 (5.2%)
property	
real estate	282 (28.2%)
building society savings agreement/life insurance	232 (23.2%)
car or other	332 (33.2%)
unknown / no property	0 (0%)
Missing	154 (15.4%)

otherplans	
bank	139 (13.9%)
stores	47 (4.7%)
none	814 (81.4%)
housing	
rent	179 (17.9%)
own	713 (71.3%)
for free	108 (10.8%)
job	
unemployed, unskilled, non-resident	22 (2.2%)
unskilled resident	200 (20.0%)
skilled, employee	630 (63.0%)
manager, self-employed	148 (14.8%)



tele	
none	596 (59.6%)
yes, registered under the customer name	404 (40.4%)
foreign	
yes	963 (96.3%)
no	37 (3.7%)

The list above is describing the levels for each categorical variable and its percentage of individuals that fall in that category (1000 applicants) provided from *germancreditDescription.docx*

Noticeable categorical labels missing:

- "A47" from "Purpose" Categorical
- "Installment" was never provided with a given label for each level.
    - *"A47" representing if the applicant is on Vacation was removed from the description table since the data set didn't contain such*
    - *Installment Rate label wasn't provided within the description, so this variable was missing in the table above due to lack of interpretation*

Categorical Predictors Definitions:

- Checkingstatus1: status of existing checking account
- History: Credit history of the applicant
- Purpose: Purpose of requesting loan
- Savings: Savings account/bonds held by the individual
- Employ: Employment status
- Status: Status of personal and family finances
- Installment: The percentage amount of money that the loan applicant has to pay back each month to repay the loan
- Others: Other debtors/guarantors on the loan
- Property: The type of property owned by the applicant
- Otherplans: Other installment plans held by the applicant

- Housing: Type of housing the applicant resides in
- Job: Type of job held by the applicant
- Tele: Indicates if applicant has a telephone registered
- Foreign: Indicates if the applicant is a foreign worker or not

Source (hyperlinked): germancreditDescription.docx

- Exploratory Analysis of the Data

By performing an exploratory analysis of data, we can identify potential issues and address them before building the logistic regression model. This can improve our model's accuracy and reliability.

To begin our exploratory analysis of the German Credit Data Set, we'll firstly import it and check how many categorical and continuous variables are present.

```
##   Default checkingstatus1 duration history purpose amount savings employ
## 1       0             A11        6     A34     A43   1169     A65    A75
## 2       1             A12       48     A32     A43   5951     A61    A73
## 3       0             A14       12     A34     A46   2096     A61    A74
## 4       0             A11       42     A32     A42   7882     A61    A74
## 5       1             A11       24     A33     A40   4870     A61    A73
## 6       0             A14       36     A32     A46   9055     A65    A73
##   installment status others residence property age otherplans housing
cards
## 1           4    A93   A101         4     A121  67       A143    A152
2
## 2           2    A92   A101         2     A121  22       A143    A152
1
## 3           2    A93   A101         3     A121  49       A143    A152
1
## 4           2    A93   A103         4     A122  45       A143    A153
1
## 5           3    A93   A101         4     A124  53       A143    A153
2
## 6           2    A93   A101         4     A124  35       A143    A153
1
##    job liable tele foreign
## 1 A173      1 A192    A201
## 2 A173      1 A191    A201
## 3 A172      2 A191    A201
## 4 A173      2 A191    A201
## 5 A173      2 A191    A201
## 6 A172      2 A192    A201
```

Total_Missing_Values
0

Total_NAs
0

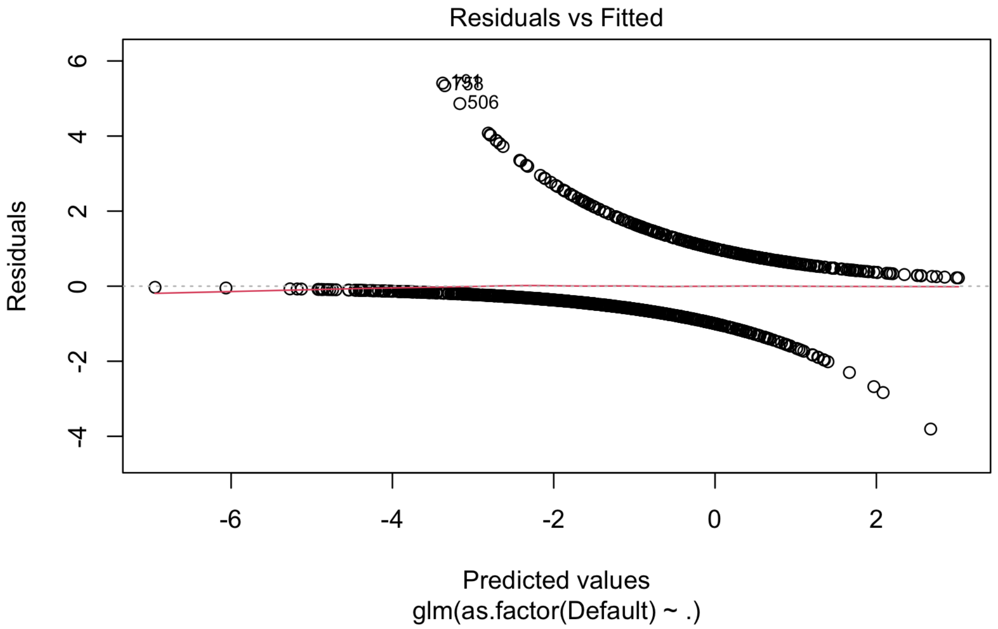
We can see that the credit data contains information about various attributes of individuals, including their Default (whether they paid required interest or installment on a debt) and other factors such as their checking status (account balance), duration of credit, payment status of previous credit, and so on.

The German credit data contains no NAs and missing values, allowing the data to be safe to resume towards the next data exploratory stages.

- Examining Unusual Observations

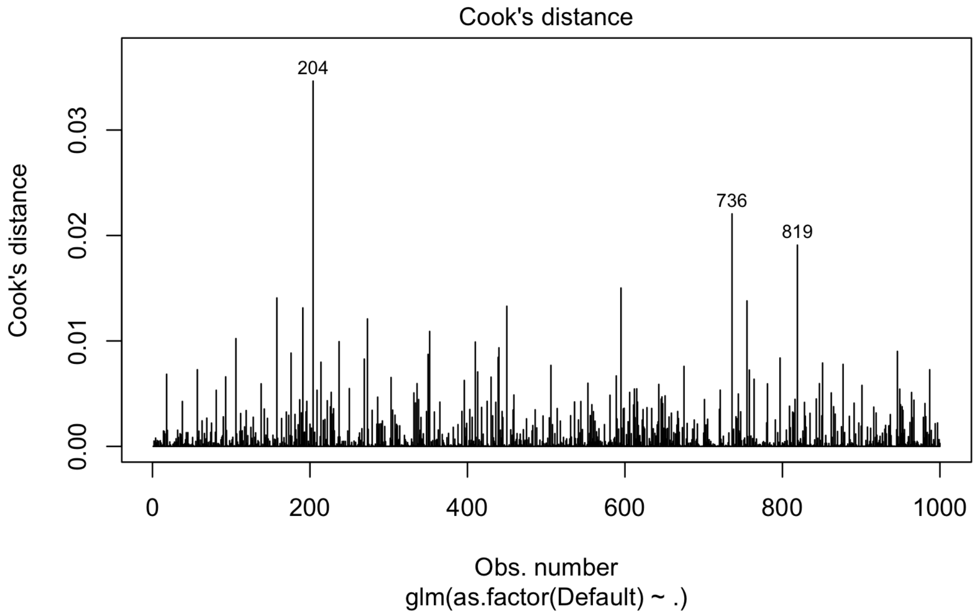
When setting up our Logistic Regression Model, we want to make sure if there are any outliers that can significantly influence the estimated coefficients and lead to a biased result. This step allows for improvement of accuracy and reliability of the model.

- Outliers



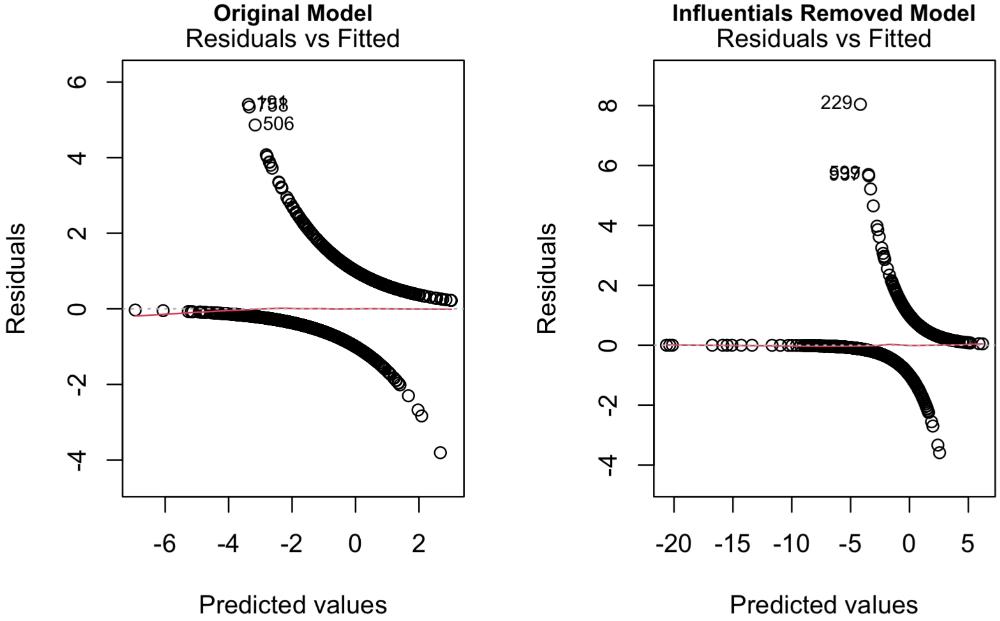
We can see that outlier points are at the observation 191, 758 and 506 with positive residuals when predicted values are negative. This may suggest that the influential outliers or observations may have an effect on the model.

- Checking Cook's Distance



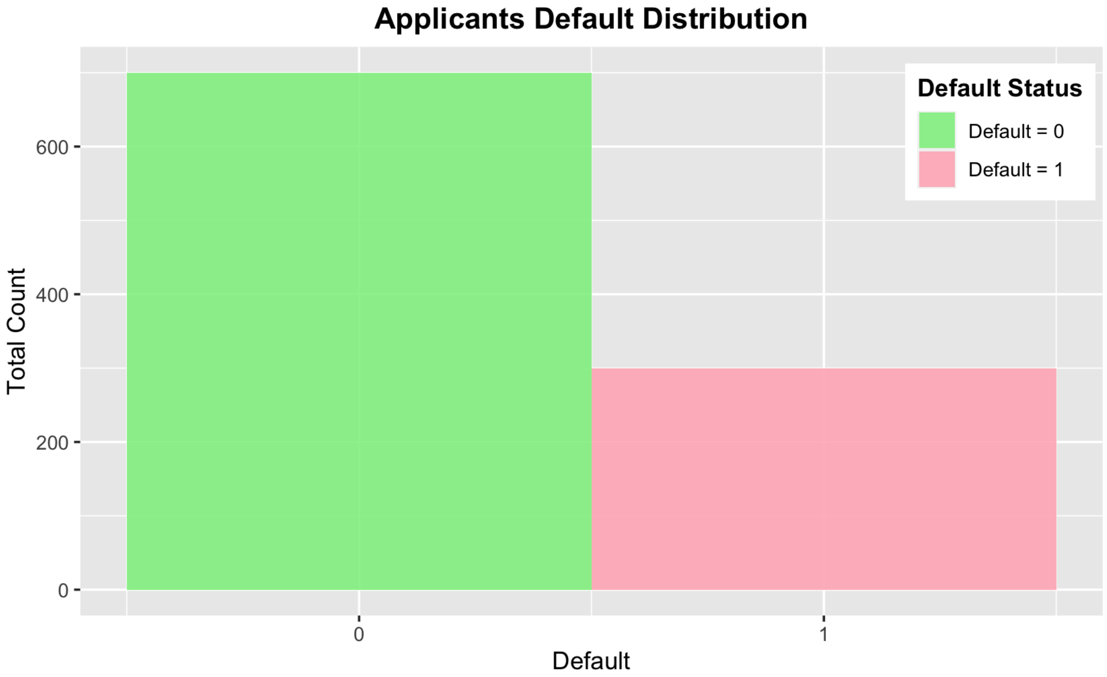
However looking further into Cook's Distance which are observations that influence the regression coefficients the most, we can see observation at 204, 736 and 819 are the points of high leverage.

- Comparing the Initial Model to the Outlier-Removed Model



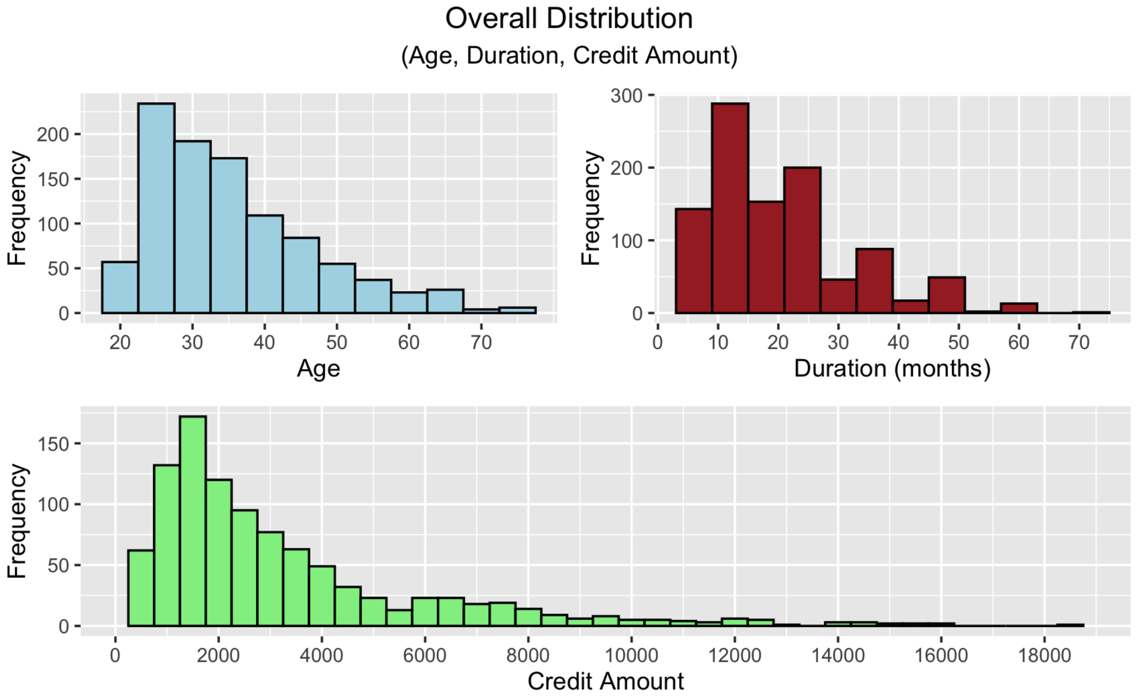
After attempting to fit the model with removed influential points, we can see the outliers are still prominent in the data with even higher residuals, thus removing the outliers were proven to show no improvement. This could tell us that the overall data may have insufficient amount of data to which causes this unusual observation. When observing a couple of the outlier rows from the credit data, there was no unexpected structure or format of the observations. Thus we will resume with the original data set.

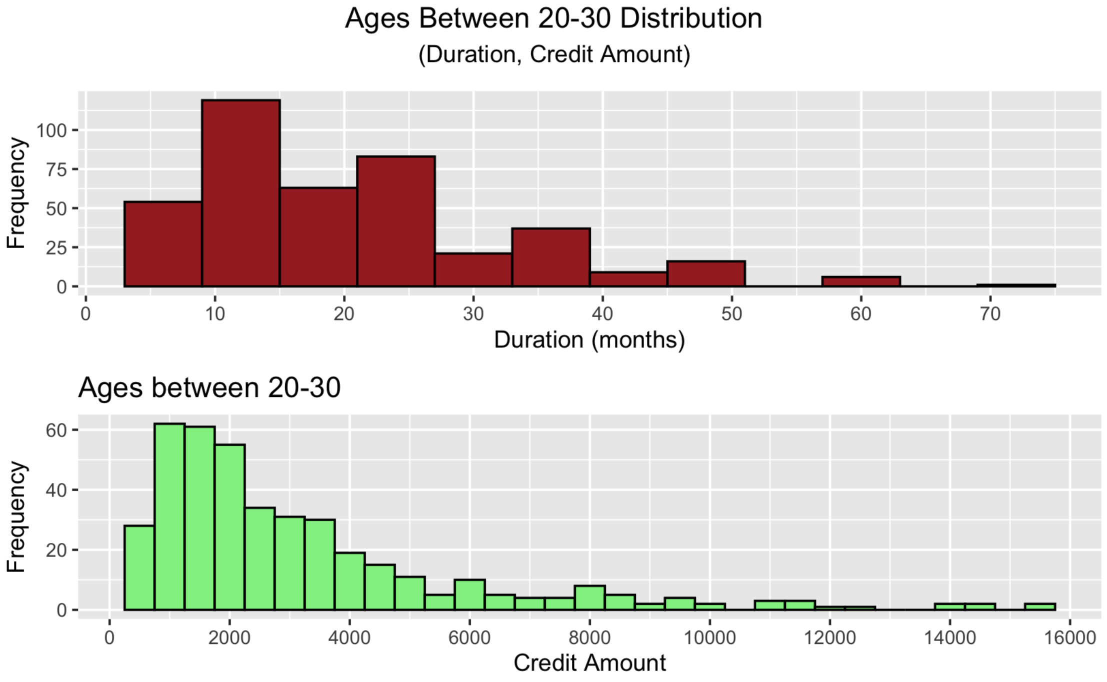
3) Examining distribution of categorical and numerical variables



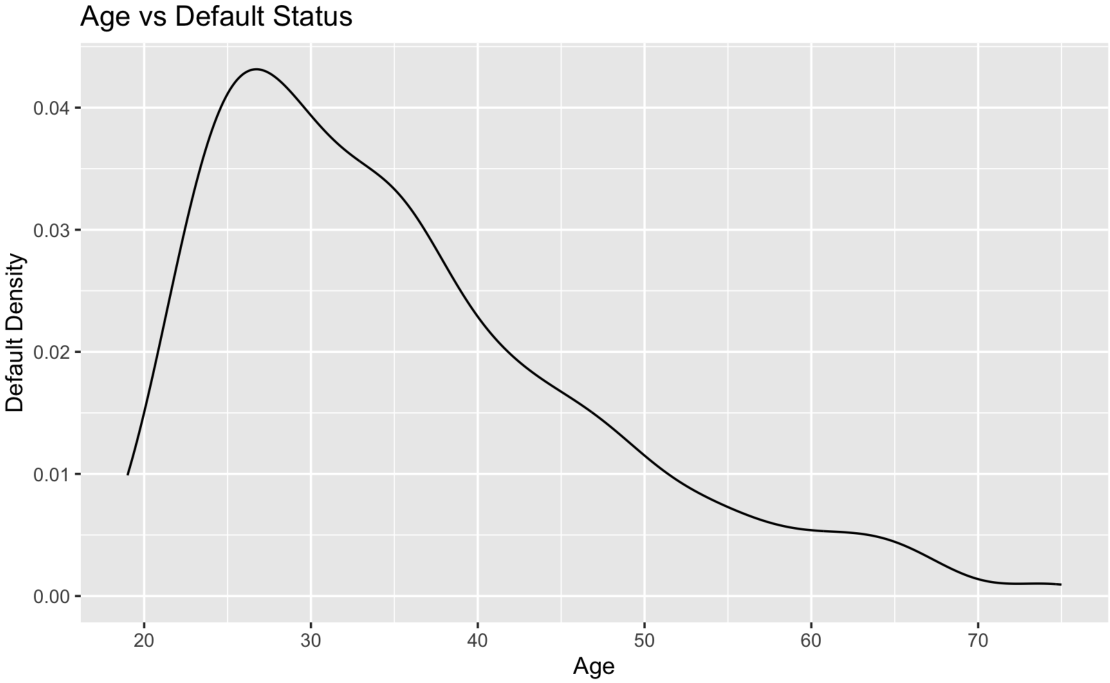
We can see that the data distribution among the response Default have a 40% more people who haven't Defaulted (0 = No) compared to those who have Defaulted (1 = Yes). This implies that 40% more individuals have been able to pay off their debt/loan on time than those who haven't.

- Observing Applicants Age in Relation to Default Status





When observing the numerical predictors (Age, Amount and Duration) in the Overall Distribution plot, we see that within around the 10-month mark of owning credit, the majority have a credit amount of 1500-2000DM at the age of 22-25. However, analyzing this from the outside is still obscure. To further analyze this relationship, the data [Figure: Ages Between 20-30 Distribution] is filtered to include only applicants within the age range of 20-30. Upon closer examination, it is found that the credit amount of 1500-2000 DM accounts for 40% of the overall credit amount frequency within this age range. This finding may indicate that a credit amount within this range is more common among younger applicants, and may be a contributing factor to the default rate observed in the data.



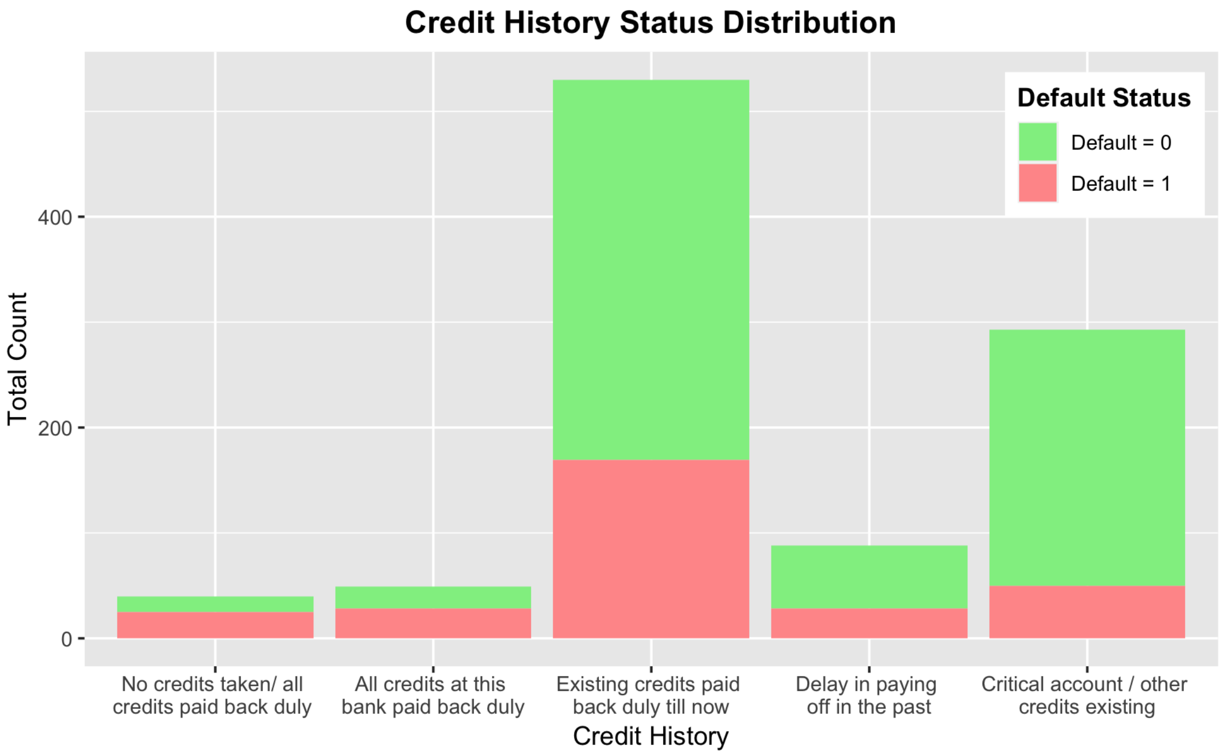
Exploring the relationship between Age and Default, we can see that at ages between 20-30, there is an increase in density of individuals that Default. This may be due to individuals in this age group may have less financial stability and may be more prone to financial struggles such as unemployment or underemployment, making it difficult to pay back their loans. Additionally, individuals in this age group may have less experience managing their finances, leading to a higher likelihood of defaulting on loans.

- Further look into Age Vs Default Case:

According to richmondfed.org, "young borrowers are considered to be less likely to have a serious delinquency than middle-aged credit card users." *(richmondfed.org, Peter Debbaut, December 2013).* Our German Credit data however was pulled from 1994 so there is a difference in economic time periods and/or regions of banking systems compared to a later time period article, as well this article is based in the United States. We also have to consider that we have 1000 observations so it may not be enough to generalize this observation, however it is an interesting insight to pull.

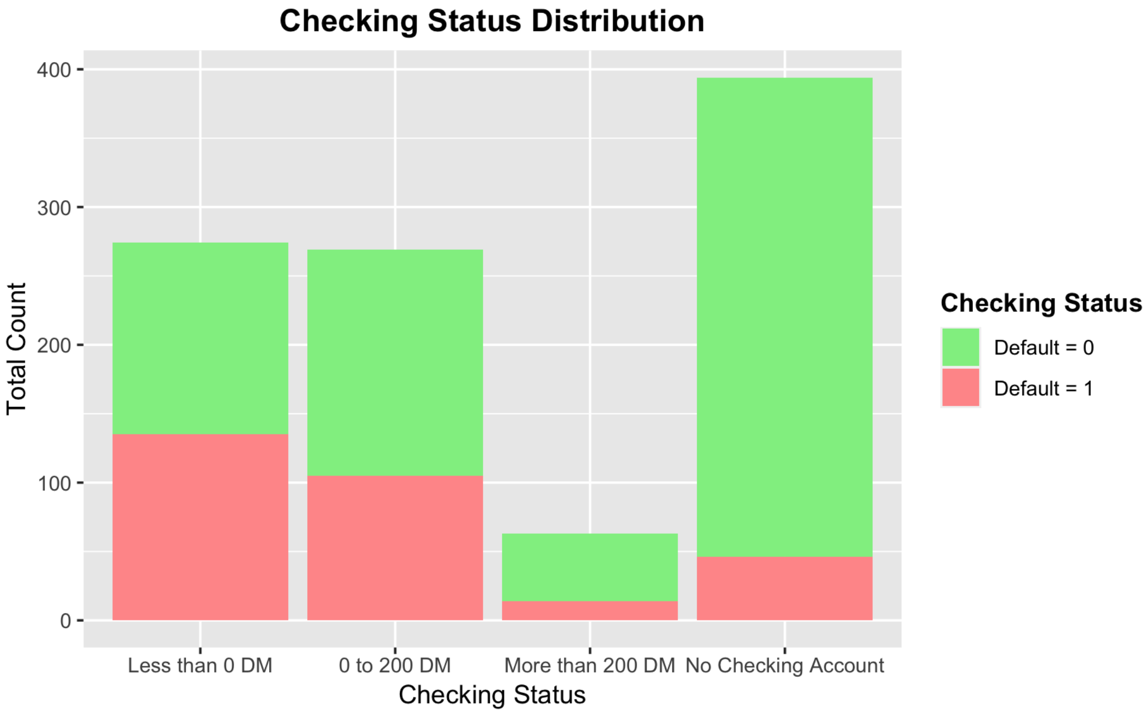
- Exploring the Socio-Economic Characteristics of Loan Applicants

Credit History



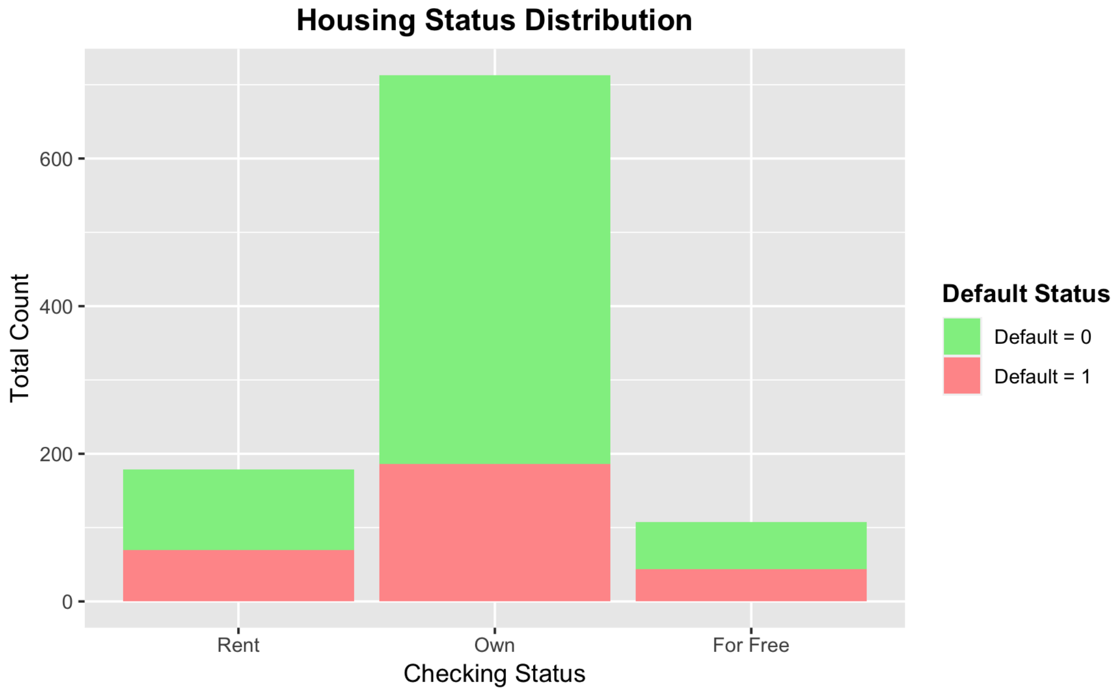
We can observe that applicants with a history of paying on time are likely Default to 0 which is clear. It is important to note that even if applicants have a history of paying their debts on time, they can still default on their loans if they are unable to make timely payments due to unforeseen financial difficulties. This is reflected in the fact that the "History" variable has a 56% share of Defaults=1, compared to other variables that Default to 1.

Checking Status



We can see that it's clear that those with no checking accounts are the least likely to Default as they are limited in access to credit and financial obligations. For those with checking account balances of less than 0DM and 0-200 DM, the ratio of Defaulting versus not Defaulting seems to be similar. However, those with balances less than 0DM have a higher likelihood of Defaulting compared to other factors. Loan applicants with a checking account balance greater than 200 DM are a smaller pool but have better odds of not Defaulting, as this score is indicative of a history of responsible financial management.

Housing Status



From the overall distribution we can see Applicants that own a home are least likely to Default, but also seem to be the most likely to Default compared to the other variables (Renting & Free).

Data Distribution Summary:

From observing the few characteristics of applicants in the German Credit Data, we can see factors where individuals who own a home, have no Checking account, and pay back their credit on time - are least likely to default as these are signs of individuals who are able to handle their money responsibly. We also found from one of the numeric components that applicants of the Ages between 20-30 are more likely to Default than older individuals

4) Logistic Regression Model Overview

Let's set up a logistic regression model to help identify the relationships between the predictor variables to the response 'Default'. This will help investigate the strength and direction of the relationships for further model diagnosis.

	Significant_P_Values
checkingstatus1A13	0.008905
checkingstatus1A14	1.664e-13
duration	0.002724
historyA34	0.001099
purposeA41	8.508e-06
purposeA42	0.002421

	Significant_P_Values
purposeA43	0.0003078
purposeA49	0.02667
amount	0.003894
savingsA64	0.01073
savingsA65	0.00031
installment	0.0001846
statusA93	0.03172
othersA103	0.02107
otherplansA143	0.006871
foreignA202	0.02609

We can see that there are 16 predictors proven to be statistically significant. The significant predictors reveal that its relationship between the response Default are unlikely to have occurred by chance. These can be factors for the bank to see which have the most effect on the response (Default).

- Checking multi-collinearity of the general data set

	Non_Collinear_Variables
checkingstatus1A12	9.338
checkingstatus1A13	8.046
purposeA410	7.146
purposeA44	6.889
purposeA45	6.509
purposeA46	7.467
purposeA49	9.767
savingsA62	7.562
savingsA63	9.498
installment	9.743
othersA102	6.614
othersA103	8.873
residence	9.086
otherplansA142	7.6
otherplansA143	8.657
liable	8.135
teleA192	9.753

Among the 48 predictors in the overall logistic regression model, we can see there are 17 predictors that aren't aren't multi-collinear. This is desirable because it can ensure that each predictor is contributing unique information to the model - not redundant with another variable.

We would like to reduce down the number of predictors by the most important predictors and to improve the accuracy of the model, thus we'll proceed with a shrinkage method (AIC) to validate the model

Correlation Plot of Numerical Variables

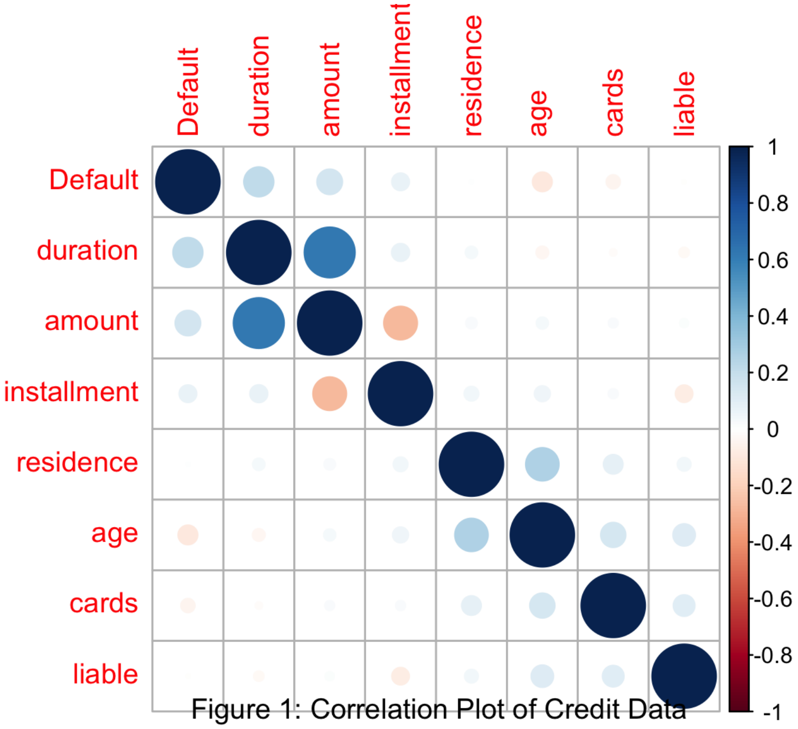


Figure 1: Correlation Plot of Credit Data

With bigger the size of the circle and darker the color, it indicates the strength of correlation. Dark blue shows high positive correlation while dark red shows high negative correlation. Upon investigating the numeric predictors in relation to the Default response, we can see that the credit amount predictor has the highest correlation.

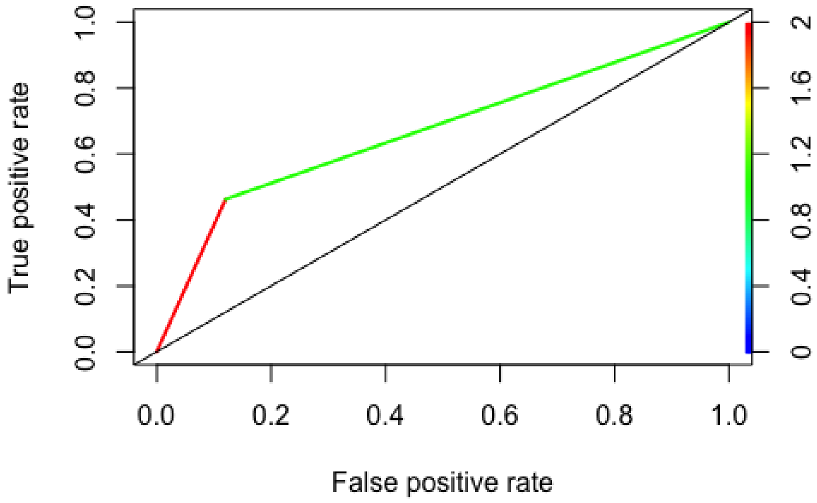
This due to credit amount being the measure of money at risk for the lender in the event of a loan default. Overall we notice moderately strong positive and negative correlations between the variables, but none are too high to be a cause for concern.

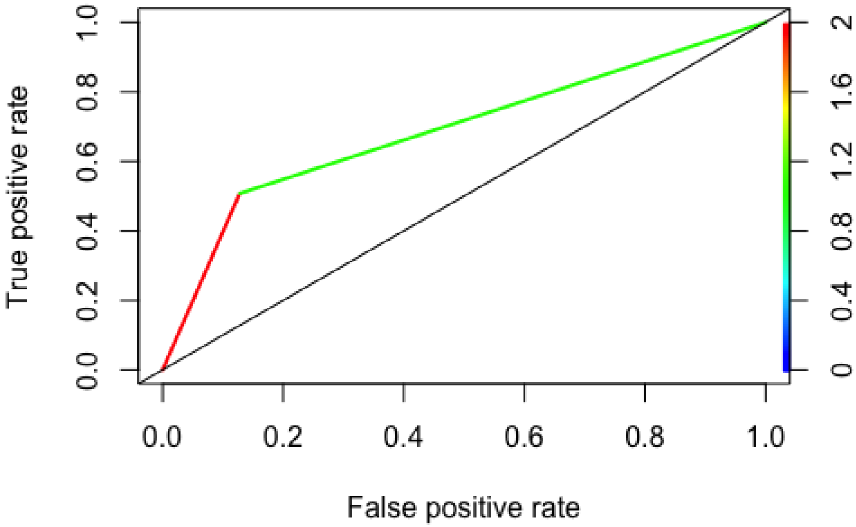
5) Building a Logistic Regression Model

Applying the stepwise function allows us to iteratively remove non-significant predictors until only significant predictors remain for the new model which will allow for higher predictive accuracy. Then apply ROC and AUC to measure overall performance of the final model in comparison to the initial model.

We created a split of data where the training set will be 80% of the data and the Testing Set will be 20% of the data which'll help us evaluate the performance of the model for new and unseen data. The method used in R is StepAIC, which will help select the most important predictors by removing the ones that don’t contribute much to the model’s ability to predict the response.

- **Initial Model (Top Chart) VS Step-AIC Model (Bottom Chart)**





Initial_Model_AUC	Step_Model_AUC
0.6712	0.6898

```
## [1] "Initial Model Confusion Matrix"

##           Reference
## Prediction   0   1
##          0 117  36
##          1  16  31

## [1] "StepAIC Model Confusion Matrix"

##           Reference
## Prediction   0   1
##          0 116  33
##          1  17  34
```

	Initial_Model_Rates	**Step_Model_Rates**
Sensitivity	0.8797	0.8722
Specificity	0.4627	0.5075

	Initial_Model_Acc	Step_Model_Acc
Accuracy	0.74	0.75

- Initial Model VS Step-AIC Model Summary

We noticed when comparing the Sensitivity, Specificity, Accuracy and AUC; StepAIC Model improves a little bit over the Initial model other than Sensitivity, showing not much of a dramatic improvement. The AUC value resulted as .679 given from the ROC curve after doing Step-AIC has shown to be considered poor performance of both models.

- ROC Plot Interpretation

To interpret our ROC Plot curve let's first understand that ROC plots measure the trade-off between the *true positive rate (sensitivity)* and the *false positive rate (1-specificity)* for different thresholds used to classify the observations.

Essentially the closer the line approaches to the upper left corner, the higher the True Positive Rate is, which is ideal to have a higher rate of True positive, since this helps us measure the performance of the model correctly identifying whether the Bank Applicant has Defaulted in their account. From a rough glance, we see that the StepAIC model is proven to be better as the ROC curve line peaks closer to the upper corner.

- Step-AIC Final Model Equation

```
##                          Estimate  Std. Error     Pr(>|z|)         2.5 %
## (Intercept)          1.9495510859 9.212269e-01 3.432311e-02  1.439796e-01
## checkingstatus1A12 -0.6056021562 2.454608e-01 1.361724e-02 -1.086697e+00
## checkingstatus1A13 -1.1627216390 3.977442e-01 3.463539e-03 -1.942286e+00
## checkingstatus1A14 -1.7577945973 2.570987e-01 8.084496e-12 -2.261699e+00
## duration             0.0244949810 1.025492e-02 1.691236e-02  4.395705e-03
## historyA34          -1.2757822802 4.840322e-01 8.395445e-03 -2.224468e+00
## purposeA41          -1.6195864492 4.234872e-01 1.310884e-04 -2.449606e+00
## purposeA410         -2.3866274255 8.897617e-01 7.311301e-03 -4.130528e+00
## purposeA43          -0.9391136006 2.760690e-01 6.695961e-04 -1.480199e+00
## amount               0.0001349117 4.679199e-05 3.936298e-03  4.320109e-05
## savingsA64          -1.2603406888 6.296195e-01 4.531164e-02 -2.494372e+00
## savingsA65          -0.9099236884 2.891744e-01 1.651661e-03 -1.476695e+00
## installment          0.3830334282 9.771461e-02 8.857842e-05  1.915163e-01
## statusA93           -0.8400832257 4.099581e-01 4.044325e-02 -1.643586e+00
## age                 -0.0209117154 1.008579e-02 3.813651e-02 -4.067950e-02
## otherplansA143      -0.8065399746 2.642372e-01 2.270697e-03 -1.324435e+00
## housingA152         -0.5126561574 2.504796e-01 4.068773e-02 -1.003587e+00
## foreignA202         -1.6132420521 6.999289e-01 2.117410e-02 -2.985077e+00
##                            97.5 %
## (Intercept)          3.7551225546
## checkingstatus1A12 -0.1245077787
## checkingstatus1A13 -0.3831572919
## checkingstatus1A14 -1.2538904193
## duration             0.0445942568
## historyA34          -0.3270965454
## purposeA41          -0.7895668160
## purposeA410         -0.6427266190
## purposeA43          -0.3980282426
## amount               0.0002266223
## savingsA64          -0.0263091980
## savingsA65          -0.3431522775
## installment          0.5745505386
## statusA93           -0.0365801516
## age                 -0.0011439356
## otherplansA143      -0.2886446214
## housingA152         -0.0217251535
## foreignA202         -0.2414066188
```

Total_predictors
36

Step_Model_Error_Rate
0.3812

We can notice that after creating our final model from the stepwise selection method - 8 predictors were removed from the original model and 36 total predictors remained as they were deemed to be the most important predictors. The summary output above however indicates only the significant predictors that were derived from the final model. From further investigation the step model contained noticed insignificant categorical predictors were left included. This can be due to the fact that these predictors still have predictive power in combination with other predictors in the model.

The 95% Confidence interval in the summary above is indicative that the Step-AIC Final Model has a narrow interval, indicating a stronger certainty in the coefficient estimates and precision. The result of the final model's training error is 38% in training error which is still high. We'll be observing different classifier methods for building a predictive model that can help better accurately classify loan applications.

For the StepAIC final model equation form we'll include only the significant predictors (18):

$$\hat{Y} = 1.95 - .61\, checkingstatus1A12 - 1.16\, checkingstatus1A13 - 1.76\, checkingstatus1A14$$

$$+ .0245\, duration - 1.27\, historyA34 - 1.62\, purposeA41 - 2.38\, purposeA410 - .94\, purposeA43$$

$$+ .0001\, amount - 1.26\, savingsA64 - .91 savingsA65 + .38\, installment$$

$$- .84\, statusA93 - .021\, age - .81\, otherplansA143 - .51\, housingA152 - 1.61\, foreignA202$$

- (Odd Ratio) Predicting Numeric Variables in Relation to Default

$$Odd\ Ratio\ Formula: \frac{odds\ in\ group\ 1}{odds\ in\ group\ 2} = \frac{p_1/\left(1-p_1\right)}{p_2/\left(1-p_2\right)}$$

Now to interpret some of the numerical predictors in the Step-AIC Final model in relation to the Response we will use a formula called Odd ratio. Odd ratio's output will help describe the odds of increase or decrease of the Loan Applicant Defaulting based on the predictor.

- When regression is **Positive**, as Predictor **increases**, **Probability of Response Increases**
- When regression is **Negative**, as Predictor **increases**, **Probability of Response Decreases**

To finding percentages of Odd Ratio we will apply in R: $(1 - exp(Coefficient\ estimate)$

For example:

Age Vs Default

- $Code\ EX:\ 1 - exp(-.021)$
- For the predictor Age scaled by Years, applying the log odds ratio we found that there's a 2% decrease in odd of Defaulting to 1 for every 1 year increase in age. Observing the age

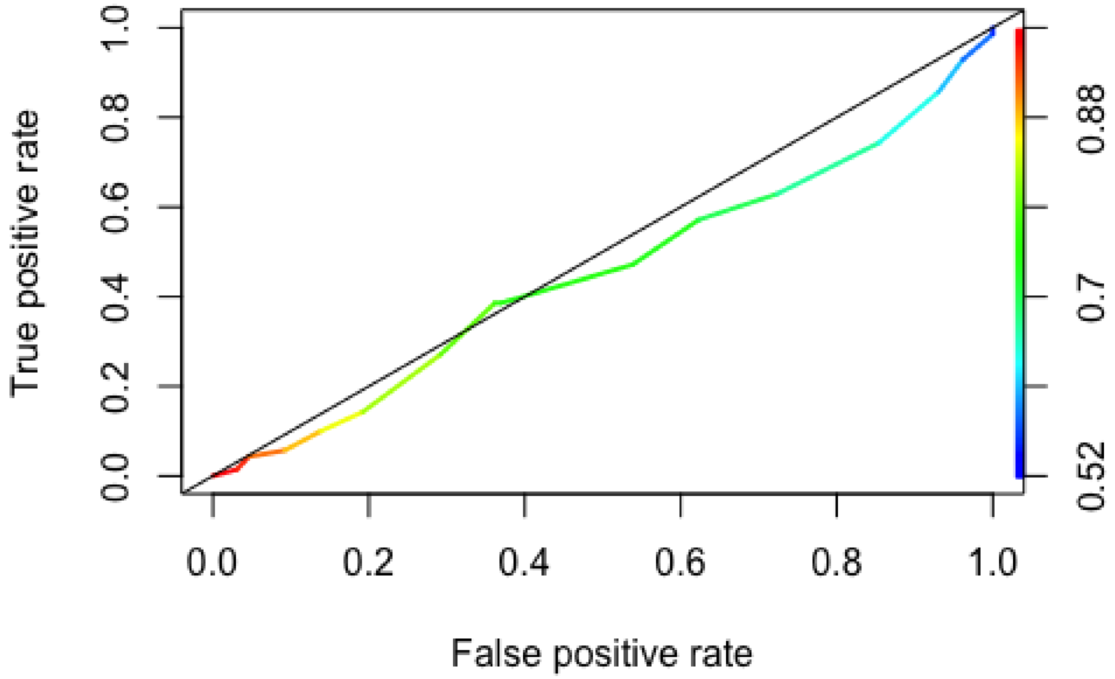
Duration Vs Default

- $Code\ EX:\ 1 - exp(.0245)$

- For the predictor Duration scale by months, we found that there's a 2% increase in the odds of the bank applicant Defaulting to 1 for every 1 month increase in length of the time the individual is expected to take to repay the credit. The duration predictor can play an important role in evaluating credit risk for loan applicants. A longer duration implies that the borrower will make payments over an extended period, which could increase their vulnerability to fluctuations in their financial status.

6) Model Selections

- K-Nearest Neighbors Result

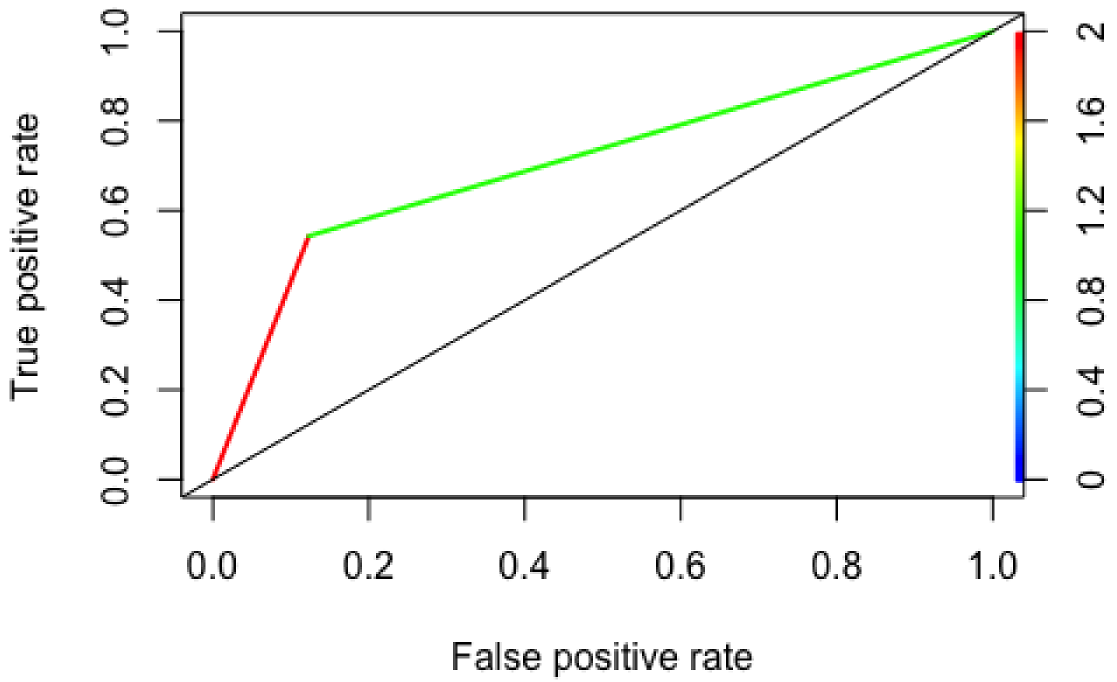


```
##           true_outcomes
## knn.model   0   1
##         0 129  68
##         1   1   2

## [1] "Error rate: 0.345"

## [1] "Sensitivity: 0.992307692307692"

## [1] "Specificity: 0.0285714285714286"
## [1] "AUC: 0.451703296703297"
```

KNN Summary

The KNN model was fit using the training set with 10-fold cross-validation to select the optimal k value. The resulting model was used to predict the test set, and the confusion matrix, error rate, sensitivity, specificity, and area under the ROC curve (AUC) were calculated. The optimal k value was found to be 33, and the AUC was 0.4517

- LDA Result

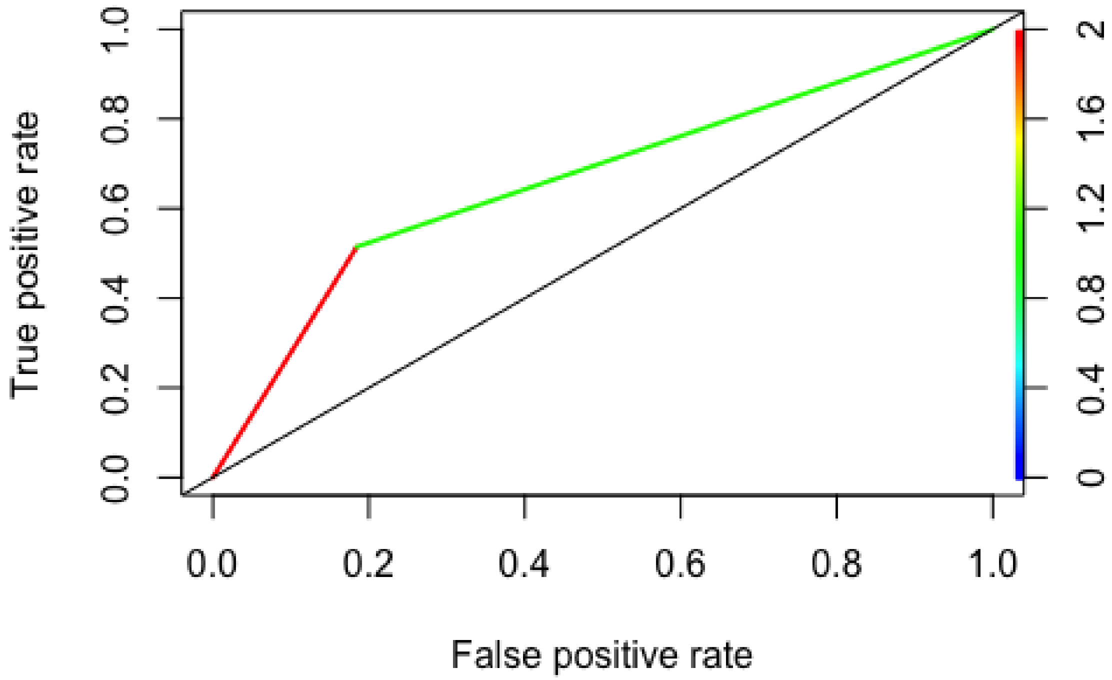


```
##            Reference
## Prediction   0   1
##          0 114  32
##          1  16  38

## [1] "Error rate: 0.24"

## [1] "Sensitivity: 0.876923076923077"

## [1] "Specificity: 0.542857142857143"

## [1] "AUC: 0.70989010989011"
```

LDA Summary

The LDA model was fit using the training set and used to predict the test set. The confusion matrix, error rate, sensitivity, specificity, and AUC were calculated. The error rate was found to be 0.24, the sensitivity was 0.87, the specificity was 0.54, and the AUC was 0.7098.

- QDA Results

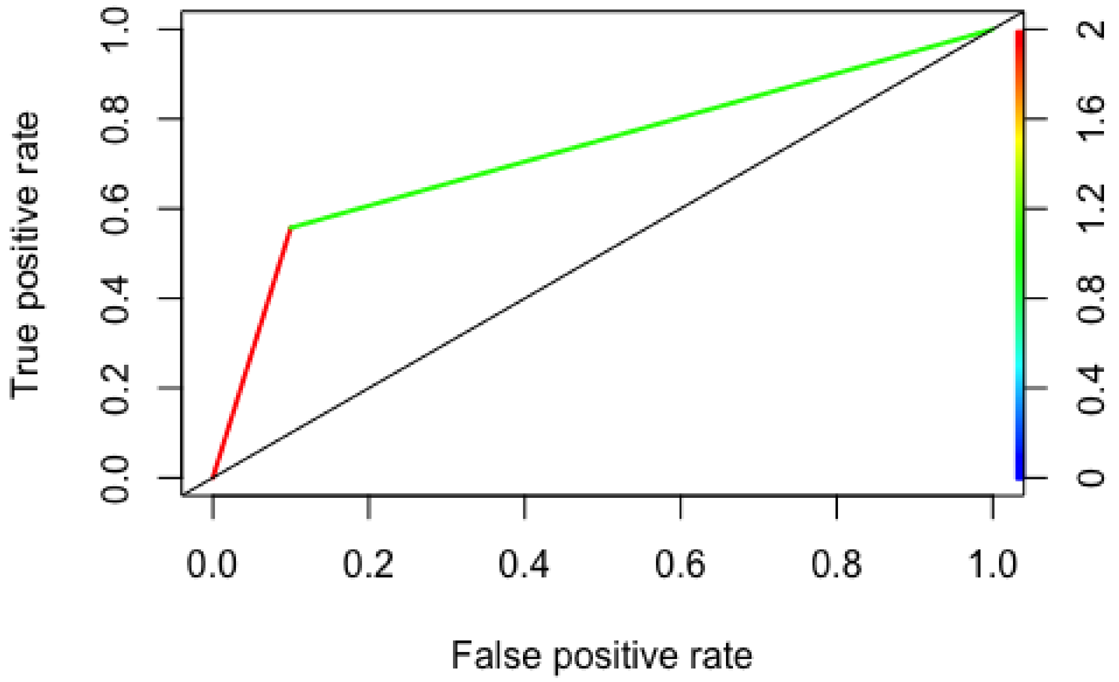


```
##            Reference
## Prediction   0   1
##          0 106  34
##          1  24  36

## [1] "Error rate: 0.29"

## [1] "Sensitivity: 0.815384615384615"

## [1] "Specificity: 0.514285714285714"

## [1] "AUC: 0.664835164835165"
```

QDA Summary

The QDA model was fit using the training set and used to predict the test set. The confusion matrix, error rate, sensitivity, specificity, and AUC were calculated. The error rate was found to be 0.29, the sensitivity was 0.81, the specificity was 0.51, and the AUC was 0.6648.

- Naïve Bayes Results



```
##             Reference
## Prediction    0   1
##          0 117  31
##          1  13  39

## [1] "Error rate: 0.22"

## [1] "Sensitivity: 0.9"

## [1] "Specificity: 0.557142857142857"

## [1] "AUC: 0.728571428571429"
```

Naive Bayes Summary

The Naive Bayes model was fit using the training set and used to predict the test set. The confusion matrix, error rate, sensitivity, specificity, and AUC were calculated. The error rate was found to be 0.22, the sensitivity was 0.90, the specificity was 0.55, and the AUC was 0.728.

7) Conclusion

```
##              ER      SENS       SPEC       AUC
## KNN      0.3450 0.9923077 0.02857143 0.4517033
## LDA      0.2400 0.8769231 0.54285714 0.7098901
## QDA      0.2900 0.8153846 0.51428571 0.6648352
## LOG(IMR) 0.2433 0.8798000 0.47830000 0.6712000
## LOG(SMR) 0.2667 0.8462000 0.47830000 0.6898000
## NAIVE-B  0.2200 0.9000000 0.55714286 0.7285714
```

- Model Selection Overview

We began by preprocessing the data by converting certain variables to factors, splits the data into a training and testing set, and fits several classification models including K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Naive Bayes.

For KNN, the optimal k value is selected using 10-fold cross-validation and the resulting model is used to make predictions on the test set. The confusion matrix, error rate, sensitivity, specificity, and area under the ROC curve (AUC) are calculated and printed.

For LDA, QDA, and Naive Bayes, the models are fit using the training set and used to predict the test set. The confusion matrix, error rate, sensitivity, specificity, and AUC are calculated and printed.

Finally, the results from all models are organized in a matrix and printed.

- Model Conclusion

In conclusion, this code demonstrates the application of several classification models to the “germancredit.csv” dataset, with the goal of predicting credit default. The KNN, LDA, QDA, and Naive Bayes models were evaluated, and their performance was assessed using various metrics such as error rate, sensitivity, specificity, and AUC.

The **Naive Bayes model** performed the best overall, with the lowest error rate and highest sensitivity. However, the other models also provided valuable insights into creditworthiness and may be useful in developing customized loan products or targeting marketing efforts.

- Further analysis

Once a bank approves a loan, there is always the risk that the borrower may default on their payments. This can lead to significant financial losses for the bank. However, by using predictive modeling techniques like the ones demonstrated in this analysis, banks can assess the creditworthiness of their borrowers and make more informed lending decisions.

For example:

- The **Naive Bayes model** performed particularly well in predicting credit Default. By utilizing this model, banks can more accurately identify high-risk borrowers and take appropriate measures to minimize the risk of default. This may include setting higher interest rates, requiring collateral, or denying the loan altogether.
- **LDA and QDA models** may also provide valuable insights into creditworthiness by identifying key variables that are strongly associated with default. Banks can use this information to develop customized loan products for different risk profiles, or to target their marketing efforts more effectively.

Using predictive modeling techniques, such as those shown in this analysis, can give banks an edge by enabling them to make better lending decisions and reduce financial losses from credit defaults. By utilizing data and analytics, banks can improve profitability and offer customized loan products that meet the unique needs and risk profiles of their customers.

- Overall Challenges

The early stages of this report was challenging as analyzing and interpreting the huge array of levels available from the Bank's categorical variables caused a fair amount of time for online research. The German Data Description had some missing labels such as for "Installment" predictor which made it difficult to interpret. Overall there weren't any clear definitions provided for the predictors in the germancreditDescription.docx, so a lot had to come from rough interpretation from other bank data articles - which was a necessary step for data analysis. As well our KNN method had trouble working properly even after extensive troubleshooting. However, our conceptual analysis was able to come together after piecing how the predictors interacted with the response and observing the statistical analysis of the important predictors and model selections.

- References Cited

German Credit Data: germancredit.csv

German Credit Description: germancreditDescription.docx

Model Selection Formulas: Bookdown.org

Bank Analysis Article: thebalancemoney.com