BML MUNJAL UNIVERSITY™

A Hero GROUP INITIATIVE

PREDICTIVE ANALYTICS
(German-Credit Dataset)

SUBMITTED TO :
Dr. Prashant Verma
SUBMITTED BY :
Devansh Sharma
Enrol. No. - 210A3010014

Contents

Executive summary

Problem statement

In banks, loan default happens when the customer is unable to repay the loan within desired time with the interest. Loan Default prediction is the process of using data to identify which customers are likely to default the loan in the future. This is important for banks because banks suffer from massive capital losses when customers are unable to repay the loan amount. Loan Default Prediction can help banks focus on identifying customers who are likely to default the loan. The German Credit project is conducted to respond to the statement "**Which existing or new customers are likely to default the loans ?** ".

Exploratory Data Analysis and Insights

The dataset contains 1000 customers and 21 features, in which **Default** is the dependent variable. The **Default** value is binary with **1** is **Default**, representing the customer default the loan and **0** is **Not Default**, representing the customer didn't default the loan. The **Default** accounts for **30%**, while the **Non Default** makes up **70%** in the dataset.

There are very few outliers in age [23 outliers], amount [72 outliers] and duration [70 outliers].

The customer profile of outliers in these variables was studied and almost the equal number of customers, default for each of these 3 variables, Hence these outliers were treated using the **Quantile based flooring method.**

In this approach, the outliers greater than the upper limit(Q3 +1.5*IQR) are masked with 100th percentile value. The outliers lesser lower limit (Q1 -1.5*IQR) are masked with 0th percentile value.

INSIGHTS

- The best model for German Credit Default prediction is Logistics Regression with an accuracy score of 77.33 %.
- Around 50 % of customers, having a checking status of less than 0 DM default the loan.

- Customers with Credit History of A30 (no credits taken/ all credits paid back duly) and A31 (all credits at this bank paid back duly) have higher percentage of default than non default. So any customer with these credit histories should be closely monitored.
- Majority of the customers take out loans for buying used cars, furniture/equipment or radio/television. Customers who have taken out a loan for buying a new car, education or business have more than 30 % probability to default.
- Out of all the customers who have employment history of either less than 1 year or are unemployed. There is around a 35 % probability that they would default. 33 % of the customers who take loans have work ex of 1 to 4 years.
- 50 % of the people who take out loans are Single males. Males or females who are divorced/ separated have the highest likelihood to default loan.
- People who don't know property or their property status is unknown have 43 % probability that they would default on the loan.
- Around 70 % of the people who take out loans own a house and 40 % of the people who live for free default the loan.

Modeling

- Decision Tree, Random Forest, Logistics Regression, and Linear Regression are preferred tools for this problem. Model performance will be conducted using Accuracy, Sensitivity, Specificity, Precision and F1-score.
- For the purpose of this problem, the model is expected to identify as many customers who are likely to Churn as possible. Hence, Sensitivity which measures the ratio between how much are correctly identified as churn to how much are actually churn, is set to be the main measure.

Model Performance

- Decision Tree provides the best insight into the features used by the algorithm to split the tree. It has the highest sensitivity among all the algorithms 96.74 %, i.e. out of 100 actual default customers, 96 were predicted correctly as default. Also the accuracy

for the train and test data are immersive with values of 65 % and 98.33 % respectively. This is the best model for predicting credit default.

- Random Forest performance is poorer than Decision tree. Having the highest specificity value for the train and test data. Also it has the highest precision among all models. But the sensitivity is extremely poor: 2.4 % for train data and 35.87 % for test data, this is the reason it wasn't considered the best model.

- The Logistics Regression model yields the relatively good result across Accuracy, Sensitivity, and Specificity in both train and test data. Logistic regression has a good accuracy score of 67 % for train data and 65 % for test data. The statistical results produced by this model are relatively consistent with the descriptive analysis.

Decision Tree will be the chosen model for implementation as it yields the highest Sensitivity and great insights into feature significance.

Recommendations

Based on the outcomes produced by the Decision Tree model, the management team is recommended to consider a number of approaches to manage customers with high likelihood of default loan, which include:

➢ The credit history of the individual and commercial credit history should be properly analyzed from past data before lending a loan.

➢ The banks should collect collateral to secure a loan like a house, so that the bank may seize that property if the customer fails to make proper payments on the loan.

➢ Myriad pieces of loan documentation that includes business and personal financial statements, income tax returns, a business plan and that essentially sums up and provides evidence for the credit history, Cash flow history and projections for the business, Collateral available to secure the loan and character.

- When reviewing the loan application, banks should consider how much experience the customer has. If he owned his business for years and has managed his company's finances responsibly, then the loan could be granted. However, if he has recently opened his business or has struggled financially, this could be detrimental.
- Customers with an age group of 20 to 40, loan amount greater than 7000 Cr or should be very carefully monitored before giving loans, as according to the analysis these people default the most.

1. Introduction

This case study is based on a very famous dataset in Machine Learning. The German Credit Risk dataset. The goal is to predict if this loan credit would be a risk to the bank or not?

In simple terms, if the loan amount is given to the applicant, will they pay back or become a defaulter?

Since there are many applications which needs to be processed everyday, it will be helpful if there was a predictive model in place which can assist the executives to do their job by giving them a heads up about approval or rejection of a new loan application.

The increasing availability of big data and the use of predictive analytics have shaped up how companies utilize data to predict whether an existing customer would Default loan or not. In another scenario whether or not a new customer is a good fit according to its past history to give a loan. The ability to predict whether or not a customer is likely to default loan becomes strategically important to banks. Failure to predict the loan default could result in huge financial losses to banks, therefore, it is imperative that bank be able to recognize well in advance customers who have high propensity to churn.

The German Credit project is conducted to respond to the statement "**Which existing or new customers are likely to default the loans ?**". Data of existing customers with credit history and profiles is used to study the current credit default pattern and predict who are likely to default. The outcomes will be used to inform the bank of high risk customers and enable the management and related teams to come up with appropriate approaches to hedge on the credit default rate.

2. Summary on data cleansing and processing

The dataset contains 1000 customers and 21 features, in which **Default** is the dependent variable. The **Default** value is binary with **1** is **Default**, representing the customer default the loan and **0** is **Not Default**, representing the customer didn't default the loan. The **Default** accounts for **30%**, while the **Non Default** makes up **70%** in the dataset.

There are very few outliers in age [23 outliers], amount [72 outliers] and duration [70 outliers].

The customer profile of outliers in these variables was studied and almost the equal number of customers, default for each of these 3 variables, Hence these outliers were treated using the **Quantile based flooring method.**

In this approach, the outliers greater than the upper limit(Q3 +1.5*IQR) are masked with 100th percentile value. The outliers lesser lower limit (Q1 -1.5*IQR) are masked with 0th percentile value.

Multicollinearity was checked among the predictors variables using the **VIF** (Varience Inflation Factor) approach. The value of VIF for each predictor variable were; **duration - 1.71**, **amount - 1.71** and **age - 1.01**. As the VIF values were below 10 which implies that Multicollinearity doesn't exist.

The data was **Normalized** and values of numeric variables duration, amount and age were standardized between -1 and +1.

Dummy Encoding was conducted on data, which converted the categorical variables into **0 and 1** encoding. A categorical variable with **n categories** was broken down into **n-1 variables**.

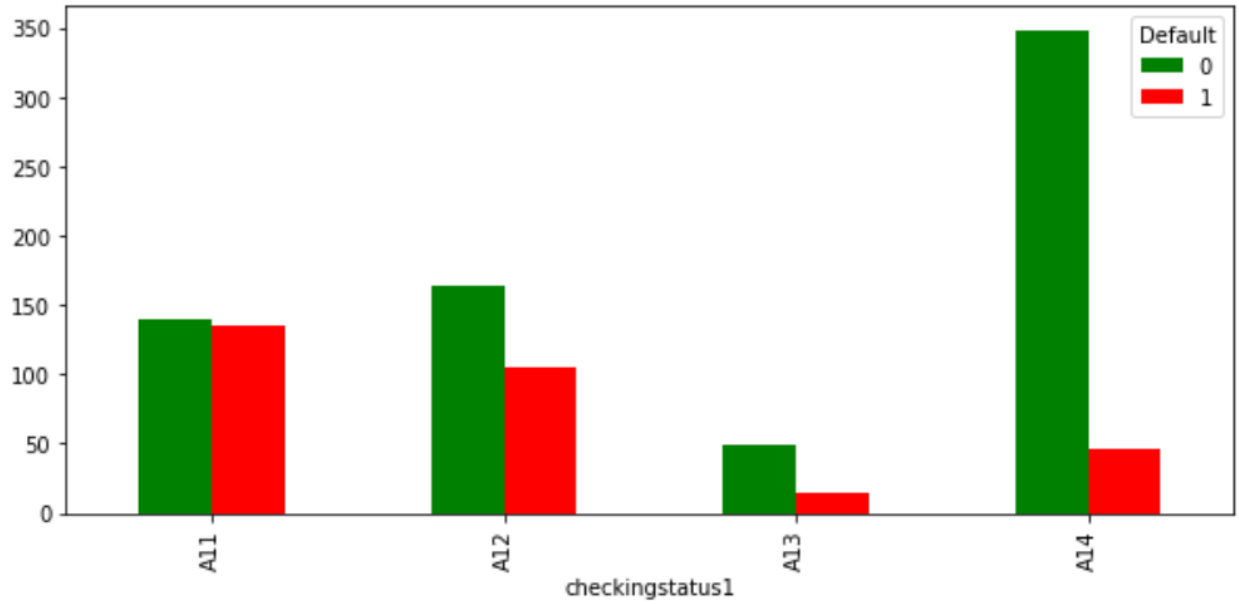
3. Descriptive analysis insights

A. Check Status

a. Status of existing checking account
   i. A11 : ... < 0 DM
   ii. A12 : 0 <= ... < 200 DM

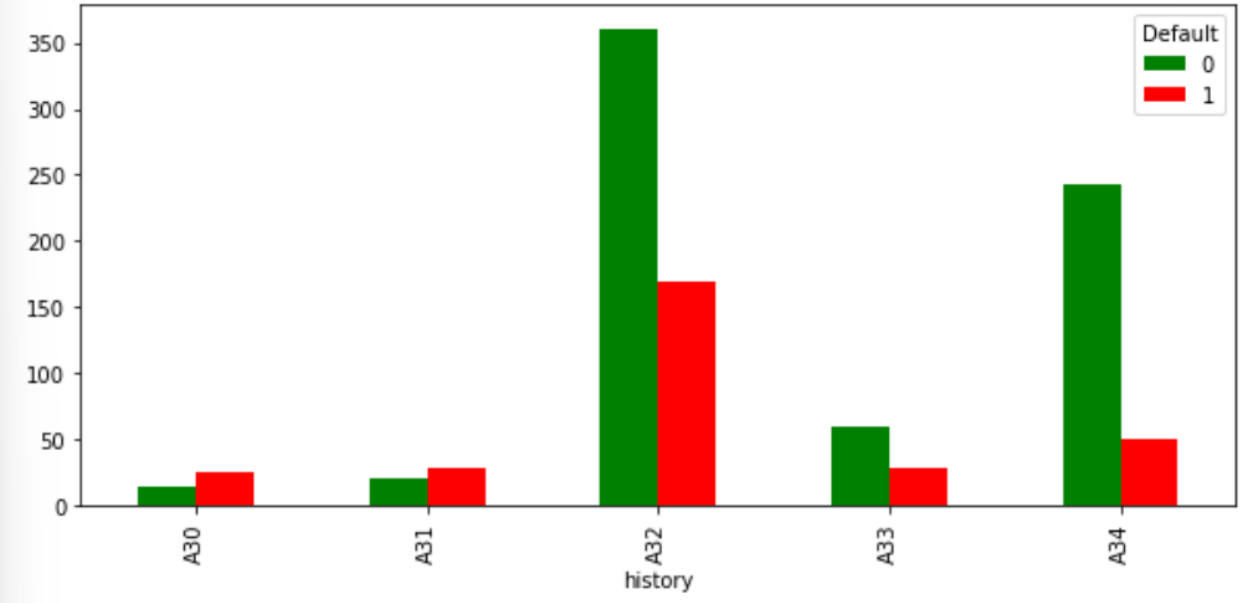
iii. A13 : ... >= 200 DM / salary assignments for at least 1 year
iv. A14 : no checking account

b. Around 40.30 % of customers do not have checking accounts, and only 12 % default credit while 88 don't.
c. Customers having checking status less than 0 DM, have the highest chance to default the credit (49 %).



B. History

a. Credit history
   i. A30 : no credits taken/ all credits paid back duly
   ii. A31 : all credits at this bank paid back duly
   iii. A32 : existing credits paid back duly till now
   iv. A33 : delay in paying off in the past
   v. A34 : critical account/ other credits existing (not at this bank)
b. 53 % of the customers have paid back existing credits back duly till now, out of which 32 % default credit loan.
c. Surprisingly, customers who all credits at this bank paid back duly have a very high default rate of 57.0 %.
   i.

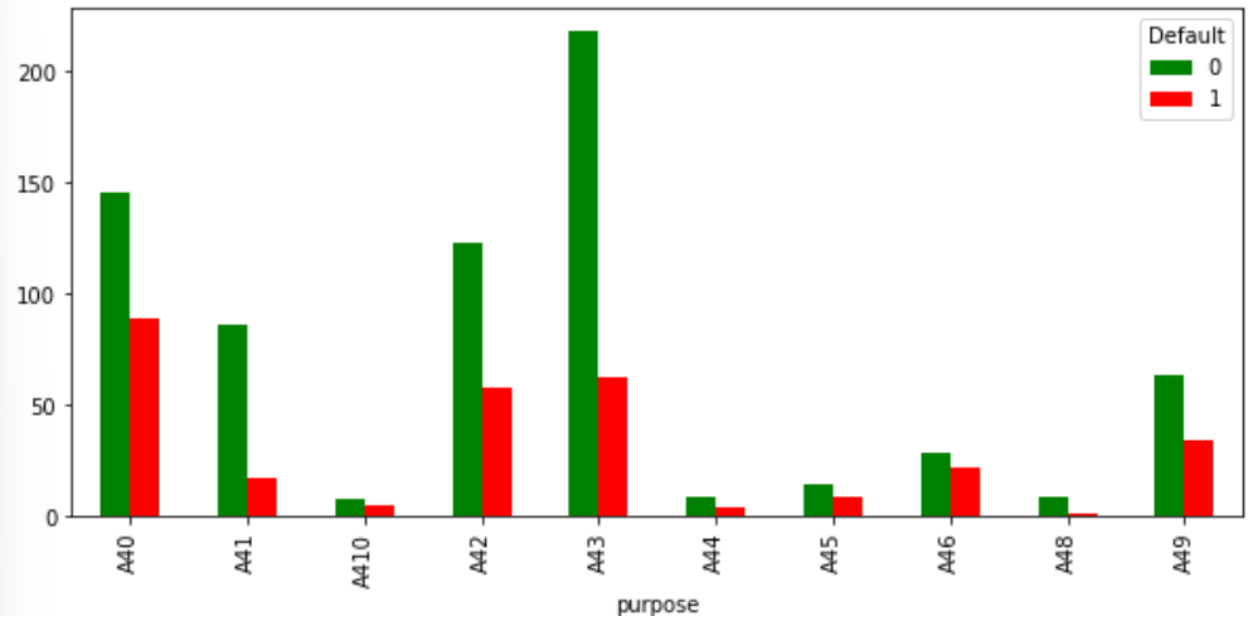


C. Purpose

a. Purpose'
   i. A40 : car (new)
   ii. A41 : car (used)
   iii. A42 : furniture/equipment
   iv. A43 : radio/television
   v. A44 : domestic appliances
   vi. A45 : repairs
   vii. A46 : education
   viii. A47 : (vacation - does not exist?)
   ix. A48 : retraining
   x. A49 : business
   xi. A410 : others

b. Around 28 % of Customers have taken credit for buying radio/ television and 23.4 % for buying a new car.

c. Customer taking loans for buying new car, education or business have the highest default rate among all categories.
   i.

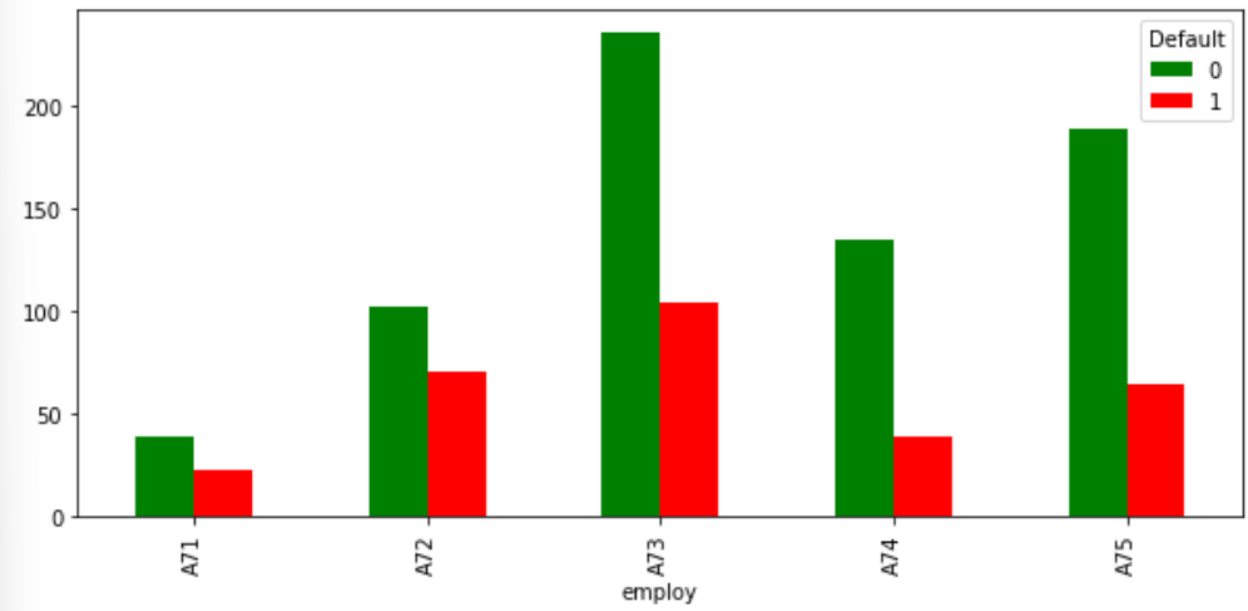


D. Employ

a. Present employment since
   i. A71 : unemployed
   ii. A72 : ... < 1 year
   iii. A73 : 1 <= ... < 4 years
   iv. A74 : 4 <= ... < 7 years
   v. A75 : .. >= 7 years

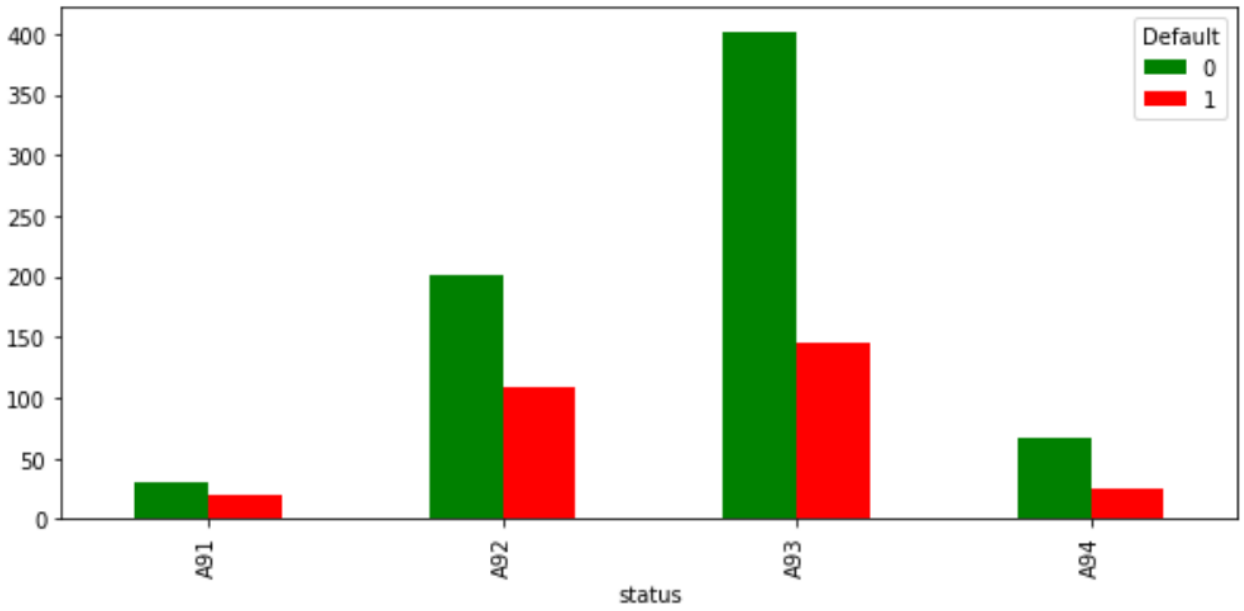
b. If a customer have number of employed years less than 1, she has 41 % to default while unemployed customers have 37 %.

c. People having work experience between 4 and 7 years have the least default percentage of 22 %.
   i.



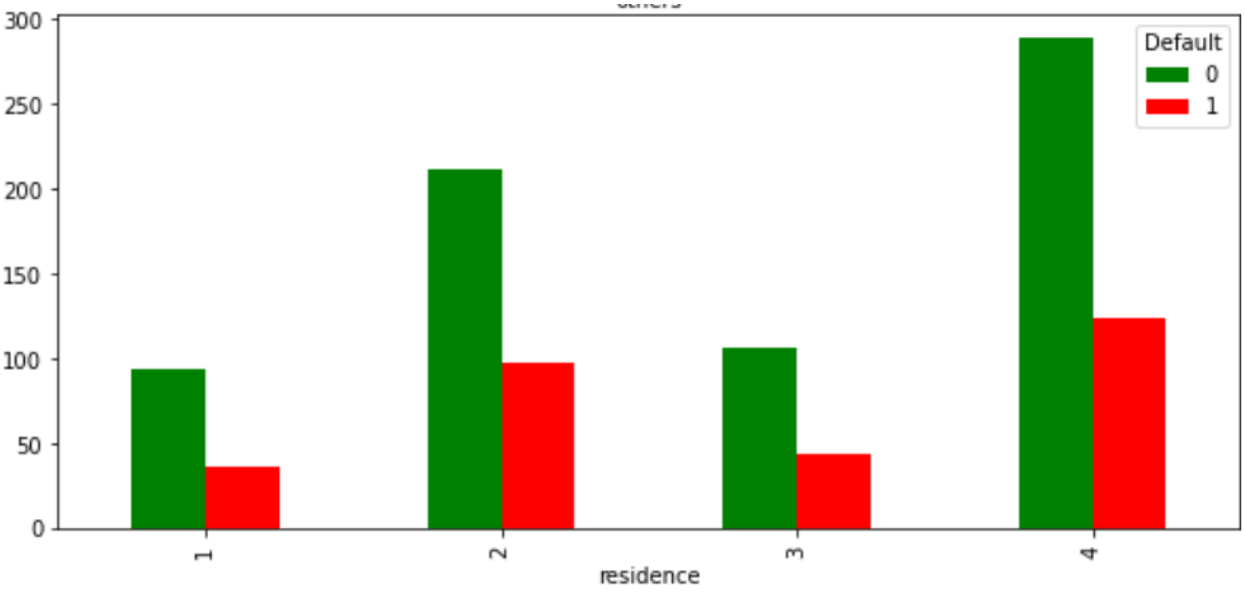
E. Status

a. Personal status and sex
   i. A91 : male : divorced/separated
   ii. A92 : female : divorced/separated/married
   iii. A93 : male : single
   iv. A94 : male : married/widowed
   v. A95 : female : single
b. Males who are divorced or separated have 40 % chance to default on the loan.
c. While, females who are divorced/ separated or married have 31 % probability to default on the loan.



F. Residence

a. Present residence since



G. Property

a. Property

i. A121 : real estate

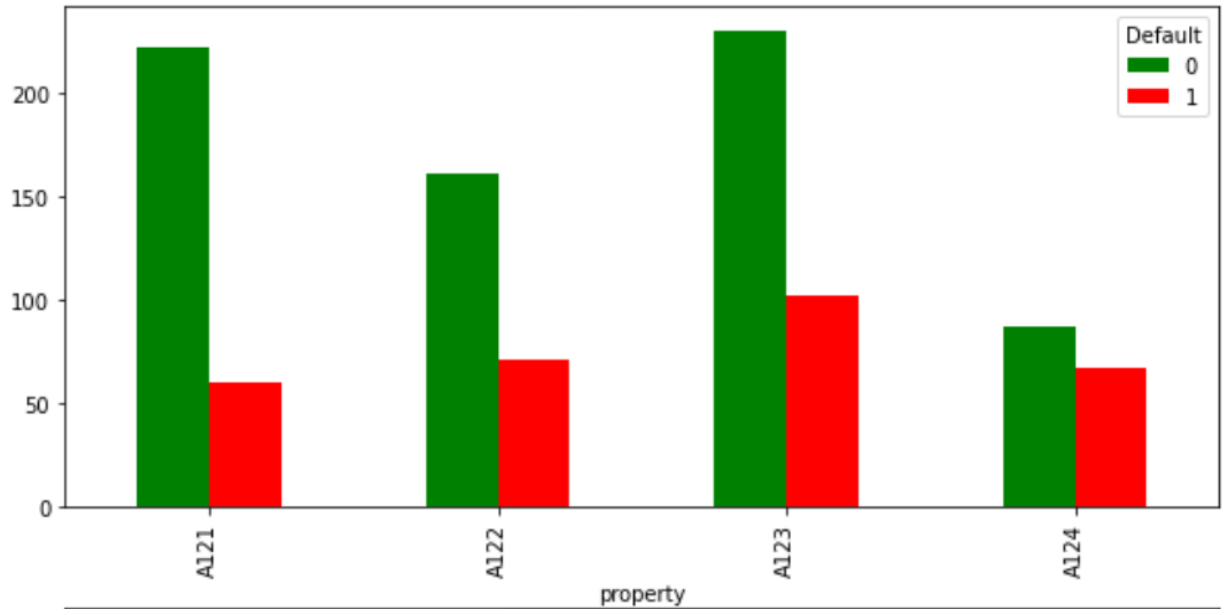
ii. A122 : if not A121 : building society savings agreement/ life insurance

iii. A123 : if not A121/A122 : car or other, not in attribute 6

iv. A124 : unknown / no property

b. Customers having unknown or no property have 46 % likelihood to default on the credit loan.

i.



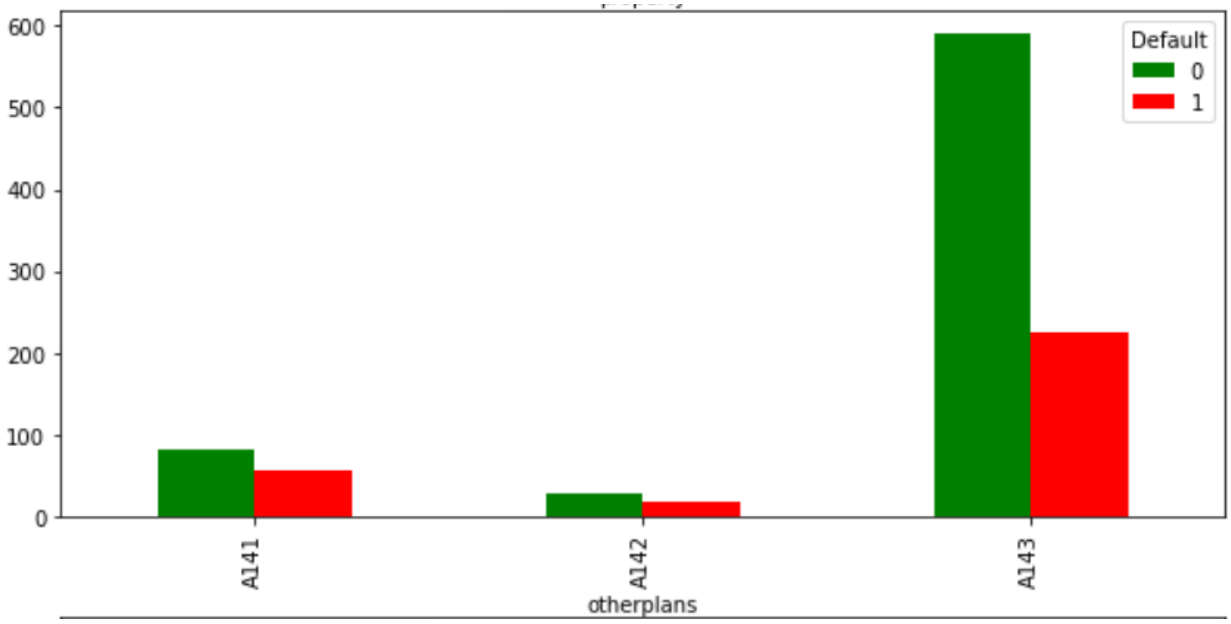
H. Other Plans

a. Other installment plans

i. A141 : bank

ii. A142 : stores

iii. A143 : none

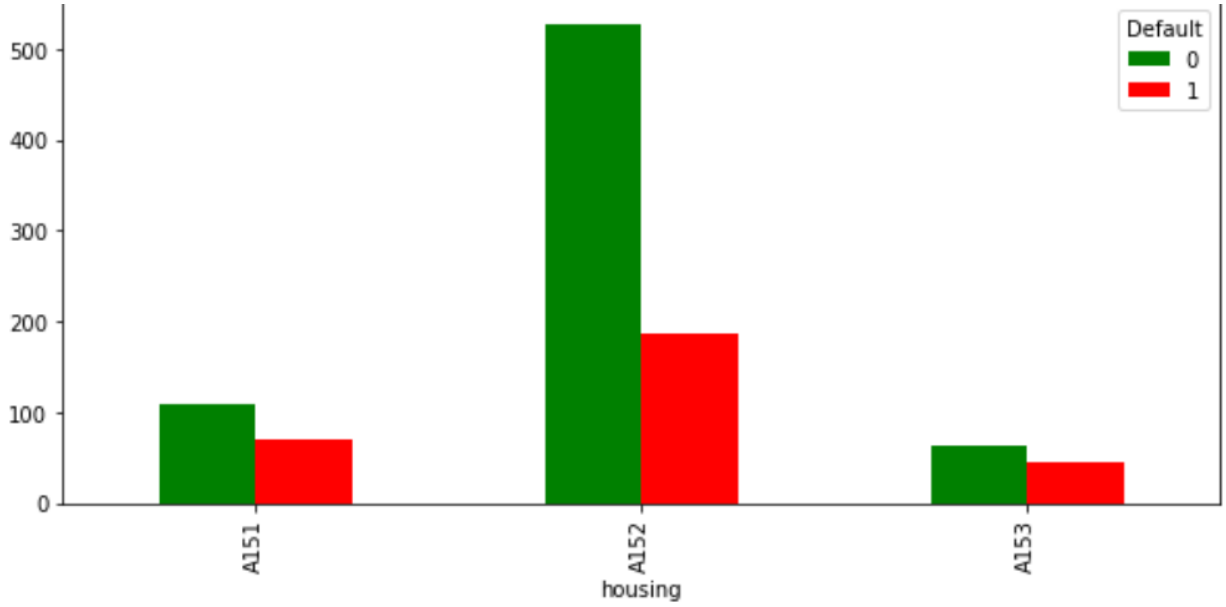


I. Housing

a. Housing

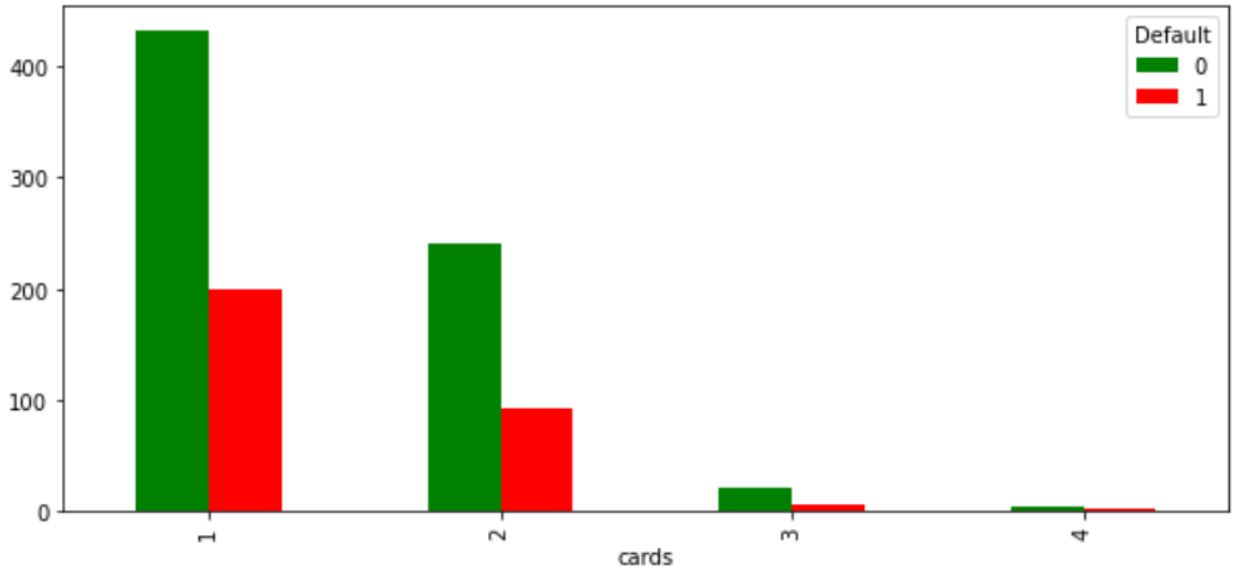
i. A151 : rent

ii. **A152 : own**
iii. **A153 : for free**



J. Cards

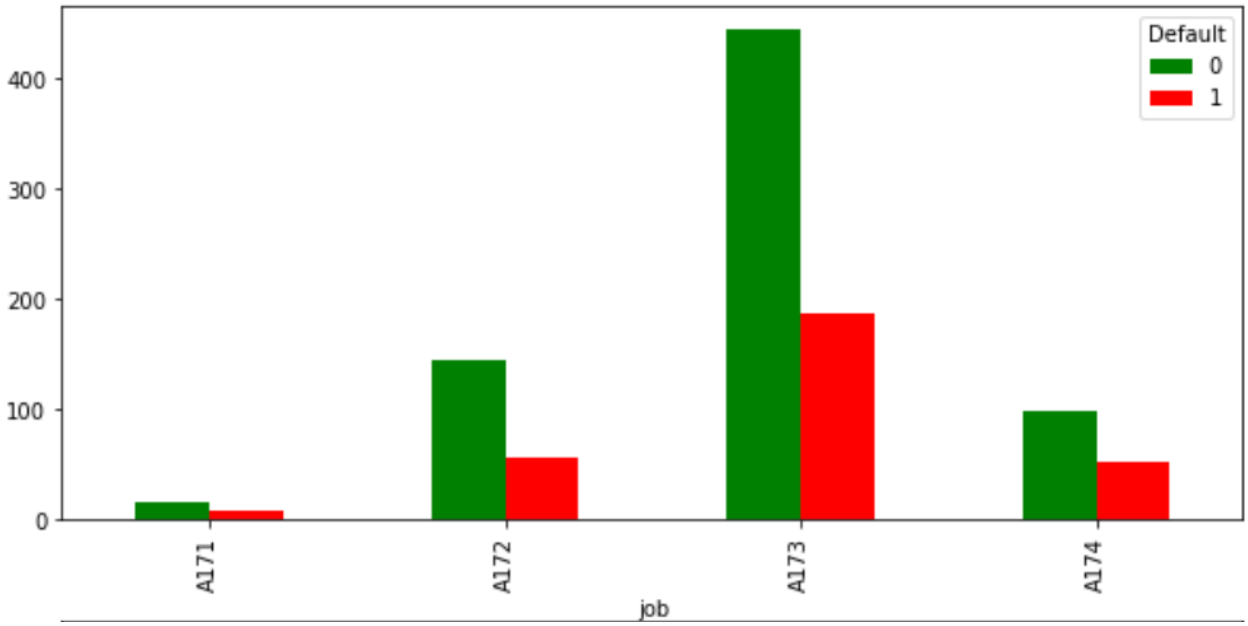
a. Number of existing credits at this bank



K. Job

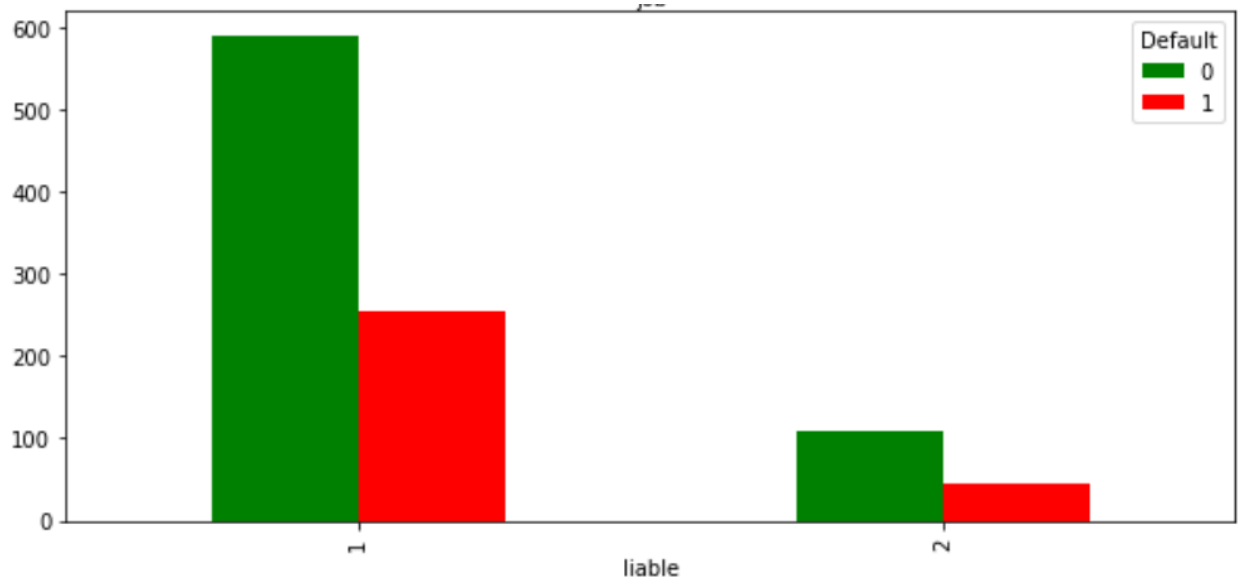
a. Job
   i. A171 : unemployed/ unskilled - non-resident
   ii. A172 : unskilled - resident
   iii. A173 : skilled employee / official
   iv. A174 : management/ self-employed/
   v. highly qualified employee/ officer

b. Highest number of customers are skilled employees or officials with a weightage of 63 %. Out of which 30 % default while 70 % do not default credit loan.



L. Liable

a. Number of people being liable to provide maintenance for

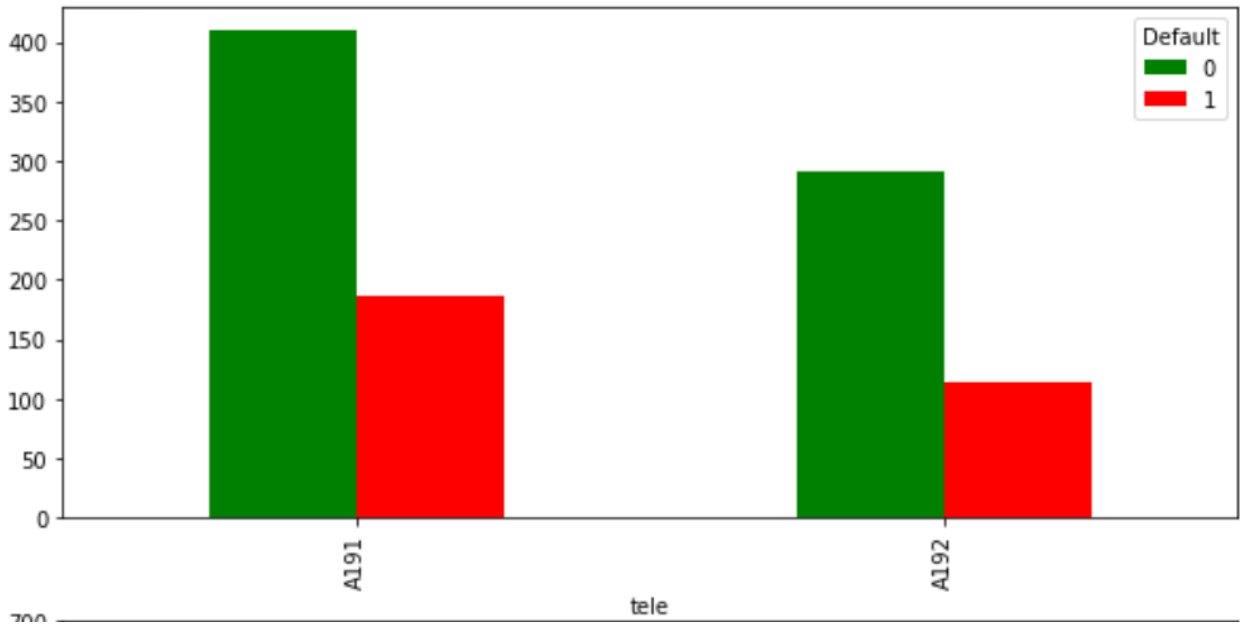


M. Tele

a. Telephone

i. A191 : none

ii. A192 : yes, registered under the customers name

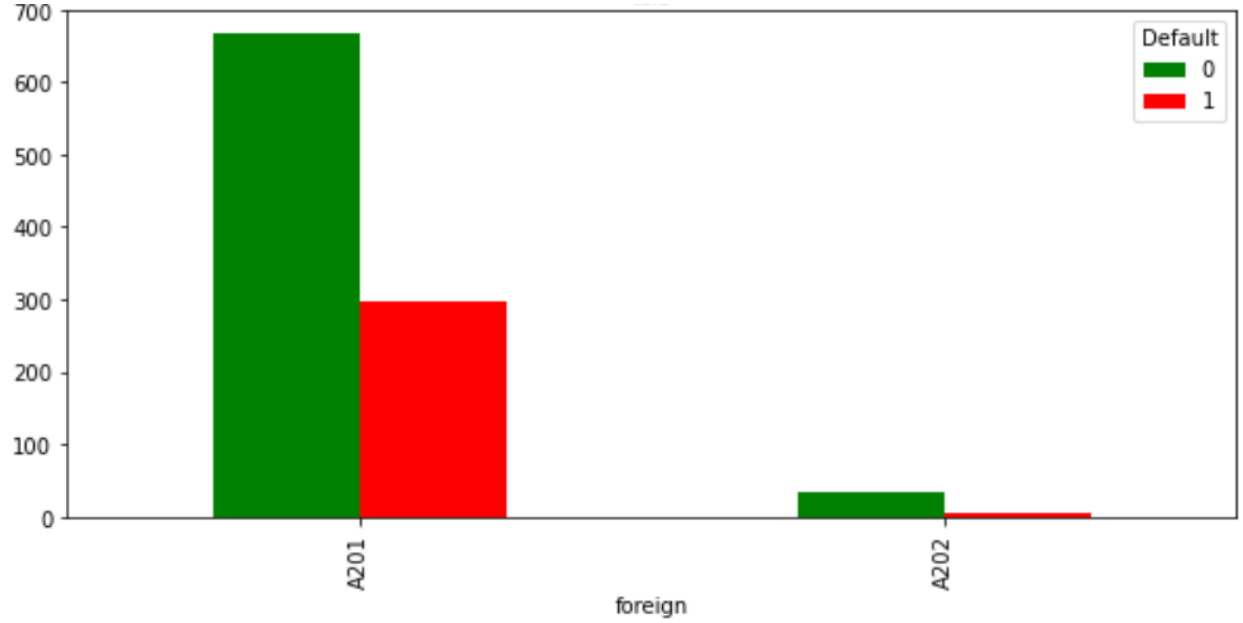


N. Foreign

a. foreign worker

i. A201 : yes

ii. A202 : no



4.) Modeling

❖ **Analytical approach and rationale**

For classification predictive modeling problems, decision tree (CART) model, Random forest and Logistic regression will be considered.

➢ The CART model is a fast and easy to understand algorithm that is great for visual representation. However, this model is unstable, meaning that a small change in the data can lead to a large change in the output.

➢ Random forest can resolve the issue of CART model and offers better accuracy. It is a “black-box" method, which foregoes the traceability of the decision tree.

➢ Logistic regression provides better insights into the weight of independent variables.

❖ **Performance Measurement**

Model performance will be conducted using Accuracy, Sensitivity and Specificity.

➢ **Accuracy**: the ratio of correct prediction to total predictions

➢ **Sensitivity**: the ratio between how much is correctly identified as positive to how much is actually positive. This ratio is important when identifying a positive class. In this problem, the positive class is 1, which represents the churn customers.

➢ **Specificity**: the ratio between how much are correctly classified as negative to how much is actually negative. In this case, the negative class is denoted as 0, representing non-churn customers.

For the purpose of this problem, the model is expected to identify as many customers who are likely to default on the loan as possible. Hence, Sensitivity is set to be the main measure for model performance.

❖ **Background into preferred approach for reporting purpose**

Several trials were run to come up with the most preferred model, these include:

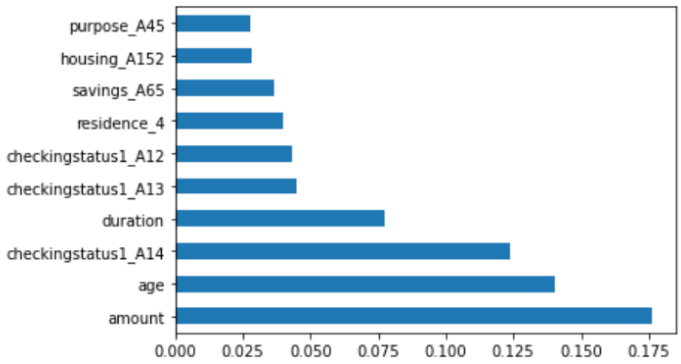
- **Originating data** vs **Outliers Capped data**. The **Capping** approach is to eliminate the influence of outliers on the accuracy of the results.
- Different split ratios of train and test dataset including 80:20 and 70:30.

In comparison between originating data and Outlier Capped data, the model performance of the latter is far better than the former. Table 2 illustrates the performance of models when run with Outliers treated data.

	CART		Random Forest		Logistic Regression	
	Train	Test	Train	Test	Train	Test
Accuracy	65.14 %	98.33 %	70.71 %	80.33 %	67.57 %	65.75%
Sensitivity	48.56 %	96.74 %	2.4 %	35.87 %	88.46 %	88.04%
Specificity	72.15 %	99.04 %	99.59 %	99.74 %	58.74 %	54.81%

5.) Predictive Models - Findings, Performance, Key takeaways

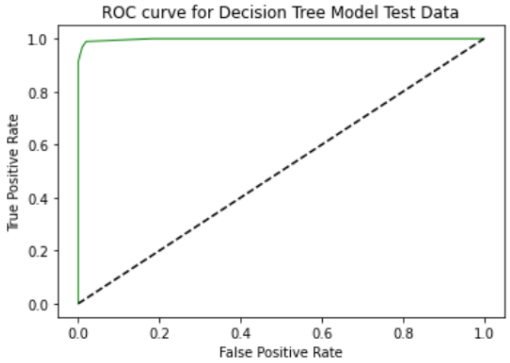
- **Decision Tree**
  - **Findings**
    - Decision tree model suggests **amount, age, checking status A14 and duration** as main predictors in the test data.



    - Please see Appendix 1 for the plots and split rules of decision tree and pruned decision trees.
  - **Model Performance**
    - **Train Data**

```
Decision Tree
Train Data
Model accuracy: 0.6514285714285715
True Positives: 101
True Negatives: 355
False Positives: 137
False Negatives: 107
-----------------------------------
Accuracy: 0.65
Mis-Classification: 0.35
Sensitivity: 0.4856
Specificity: 0.7215
Precision: 0.72
f_1 Score: 0.58
```

    - **Test Data**



-
```
Decision Tree
Test Data
Model accuracy: 0.9833333333333333
True Positives: 89
True Negatives: 206
False Positives: 2
False Negatives: 3
-----------------------------------
Accuracy: 0.9833
Mis-Classification: 0.02
Sensitivity: 0.9674
Specificity: 0.9904
Precision: 0.99
f_1 Score: 0.98
```

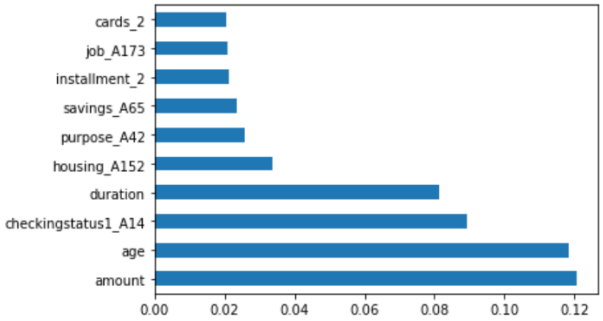
➢ **Key takeaways**

- The Accuracy, Sensitivity and Specificity rate are relatively good and consistent across the train and test dataset - **98.33 % accuracy**, **96.74 % sensitivity** and **99.83 % precision**.
- One of the pros of a decision tree model is the possibility of tracing the rules and identifying which features are main contributors to the prediction. On the other hand, this model is extremely sensitive to changes.

❖ **Random Forest**

➢ **Findings**

- Random forest model also indicates **Amount, Checking Status A14(no checking account), age** and **Duration** as important variables.

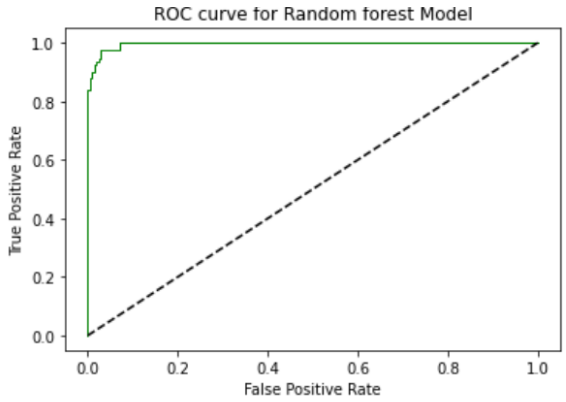


- **Model Performance**
  - **Train Data**

```
Random Forest
Train Data
Model accuracy: 0.7071428571428572
True Positives: 5
True Negatives: 490
False Positives: 2
False Negatives: 203
-----------------------------------
Accuracy: 0.71
Mis-Classification: 0.29
Sensitivity: 0.024
Specificity: 0.9959
Precision: 1.0
f_1 Score: 0.05
```

  - **Test Data**

```
Random Forest
Test Data
Model accuracy: 0.8033333333333333
True Positives: 33
True Negatives: 208
False Positives: 0
False Negatives: 59
-----------------------------------
Accuracy: 0.8033
Mis-Classification: 0.2
Sensitivity: 0.3587
Specificity: 1.0
Precision: 1.0
f_1 Score: 0.53
```



- **Key takeaways**

- Performance measurement of Random Forest is the worst among all algorithms. Although the accuracy is good, sensitivity for train and test data is very poor 2.4 % and 35.87 %.

❖ **Logistics Regression**

➢ **Findings**

-

➢ **Model Performance**

- **Train Data**

```
Model accuracy: 0.6757142857142857
True Positives: 184
True Negatives: 289
False Positives: 203
False Negatives: 24
-----------------------------------
Accuracy: 0.6757
Mis-Classification: 0.3243
Sensitivity: 0.8846
Specificity: 0.5874
Precision: 0.5874
f_1 Score: 0.706
```

- **Test Data**

```
Logistics Regression
Test Data
Model accuracy: 0.65
True Positives: 81
True Negatives: 114
False Positives: 94
False Negatives: 11
--------------------------
Accuracy: 0.65
Mis-Classification: 0.35
Sensitivity: 0.8804
Specificity: 0.5481
Precision: 0.55
f_1 Score: 0.68
```

➢ **Key takeaways**

- Logistic Regression model yields the relatively good result across Accuracy, Sensitivity, and Specificity in both Train and Test data.

6.) <u>Model comparison</u>

	CART		**Random Forest**		**Logistic Regression**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**

Accuracy	65.14 %	**98.33 %**	70.71 %	80.33 %	67.57 %	65.75%
Sensitivity	48.56 %	**96.74 %**	2.4 %	35.87 %	88.46 %	88.04%
Specificity	72.15 %	99.04 %	99.59 %	**99.74 %**	58.74 %	54.81%
Precision	72. 29 %	**99.92 %**	99.85 %	99.38 %	58.74 %	55.04 %

- In term of Accuracy, tuned Decision Tree comes out as the best;in both train and test data. Low Sensitivity means the model misses a high proportion of customers whose likelihood of default on the loan is high. For this reason, random forest will not fit the purpose of this project.
- **Decision Tree** gives a slightly better Sensitivity rate compared to Logistics Regression; hence appears to be the best model among the three. Decision Tree also produces the highest Sensitivity. Most importantly, Decision Tree provides insights into the significance of predictors and abilities to unbundle each variable to investigate the changes between sub-classes. Therefore, this model will be selected to implement.

7.) Insights

- The best model for German Credit Default prediction is Logistics Regression with an accuracy score of 77.33 %.
- Around 50 % of customers, having a checking status of less than 0 DM default the loan.
- Customers with Credit History of A30 (no credits taken/ all credits paid back duly) and A31 (all credits at this bank paid back duly) have higher percentage of default than non default. So any customer with these credit histories should be closely monitored.
- Majority of the customers take out loans for buying used cars, furniture/equipment or radio/television. Customers who have taken out a loan for buying a new car, education or business have more than 30 % probability to default.

- Out of all the customers who have employment history of either less than 1 year or are unemployed. There is around a 35 % probability that they would default. 33 % of the customers who take loans have work ex of 1 to 4 years.
- 50 % of the people who take out loans are Single males. Males or females who are divorced/ separated have the highest likelihood to default loan.
- People who don't know property or their property status is unknown have 43 % probability that they would default on the loan.
- Around 70 % of the people who take out loans own a house and 40 % of the people who live for free default the loan.

8.) Recommendation

- The credit history of the individual and commercial credit history should be properly analyzed from past data before lending a loan.
- The banks should collect collateral to secure a loan like a house, so that the bank may seize that property if the customer fails to make proper payments on the loan.
- Myriad pieces of loan documentation that includes business and personal financial statements, income tax returns, a business plan and that essentially sums up and provides evidence for the credit history, Cash flow history and projections for the business, Collateral available to secure the loan and character.
- When reviewing the loan application, banks should consider how much experience the customer has. If he owned his business for years and has managed his company's finances responsibly, then the loan could be granted. However, if he has recently opened his business or has struggled financially, this could be detrimental.
- Customers with age group of 20 to 40, loan amount greater than 7000 Cr or should be very carefully monitored before giving loans, as according to the analysis these people default the most.

9.) Appendix A

1.) Decision Tree