Prediction for Creditability

1. Summary

2. Data Preparation
2.1 Statistic Summary of Variables
2.2 Check Attribute Type
2.3 Check Categorical Variables
2.4 Check Outliers on Continuous Variables
2.5 Check Distribution of Variables with Numeric Values
2.6 Data Transformation on Variables with Skewed Distribution

3. Exploratory Analysis
3.1 Analyzing Correlation between Categorical (target) and Continuous variables
3.2 Analyzing Correlation between Categorical (target) and Categorical variables
3.3 Selection of Attributes

4. Predictive Modeling
4.1 Classification using Decision Tree
4.2 Classification using Naive Bayes
4.3 Additional Data Analysis
    4.3.1 Classification using Random Forest
    4.3.2 Classification using Gradient Boosting
4.4 Performance Metrics Comparison
4.5 Performance Comparison between “all attributes” and “selected attributes”

5. Conclusion
5.1 Major Findings

5.2 Recommendation

1. Summary

Data mining is a critical step in knowledge discovery involving theories, methodologies and tools for revealing patterns in data. It is important to understand the rationale behind the methods so that tools and methods have appropriate fit with the data and the objective of pattern recognition. There may be several options for tools available for a data set.

When a bank receives a loan application, based on the applicant's profile the bank has to make a decision regarding whether to go ahead with the loan approval or not. Two types of risks are associated with the bank's decision –

- If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank
- If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank

1.1 Objective of Analysis:

Minimization of risk and maximization of profit on behalf of the bank.

To minimize loss from the bank's perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to. An applicant's demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

The German Credit Data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants.

A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.

We implemented the steps of data preparation, exploratory analysis and predictive modeling in Python 3.6.7.

In this project, we used five different machine learning algorithms named Decision Tree Classifier Naive Bayes, Random Forest Classifier, Support Vector Classifier and Gradient Boosting Classification and compare their performance to find the best model in terms of evaluation metrics of Accuracy, Precision and Recall.

1.2 Description of Credit Card Dataset:

In order to provide loans to customers, a bank needs to make right decision in determining who should get the approval and who should not. This dataset is the German Credit Data that contains 20 attributes and the class attribute showing a good or a bad credit risk. Your team of data scientists will need to develop a data analytics based strategy for the bank managers that can help them in making a decision about loan approval for the prospective applicants/customers.

1. Creditability: The class attribute (qualitative) showing whether the credit rating is good or bad. Good credit is represented by 1 and bad credit rating is represented by 0.

2. Account Balance: Checking account status (1: < 0 DM, 2: 0<=...<200 DM, 3: > 200 DM, 4: No checking account), where DM= Deutsche Mark (qualitative attribute).

3. Duration of Credit (month): Duration of credit in months (numerical)

4. Payment Status of Previous Credit: Credit history (qualitative) 0: no credits taken, 1: all credits at this bank paid back duly, 2: existing credits paid back duly till now, 3: delay in paying off in the past, 4: critical account.

5. Purpose: Qualitative attribute showing the purpose of the loan (0: New car, 1: Used car , 2: Furniture/Equipment, 3: Radio/Television, 4: Domestic Appliances , 5: Repairs ,6: Education ,7: Vacation, 8: Retraining ,9: Business, 10: Others)

6. Credit Amount: Numerical value showing the credit amount

7. Value Savings/Stocks: Qualitative attribute showing average balance in savings and stocks (1 : < 100 DM, 2: 100<= ... < 500 DM, 3 : 500<= ... < 1000 DM, 4 : =>1000 DM, 5: unknown/ no savings account)

8. Length of current employment: Qualitative attribute showing length of employment (1 : unemployed, 2: < 1 year, 3: 1<=...<4 years, 4: 4<=...<7 years, 5:>=7years).

9. Instalment percent: Installment rate in percentage of disposable income (numerical)

10. Sex & Marital Status: Qualitative attribute showing gender and marital status (1: male : divorced/separated, 2: female : divorced/separated/married, 3 : male: single, 4: male : married/widowed, 5 : female : single)

11. Guarantors: (Qualitative) Guarantors and co-applicants: (1 : none, 2 : co-applicant, 3 : guarantor)

12. Duration in Current address: Qualitative value showing the duration in current address (1: <= 1 year, 2: 1<...<=2 years, 3: 2<...<=3 years, 4: >3 years)

13. Most valuable available asset: Qualitative attribute showing valuable assets (1 : real estate

2 : savings agreement/ life insurance, 3 : car or other, 4 : unknown / no property)

14. Age (years): Numerical value showing age in years.

15. Concurrent Credits: Installment plans (1 : bank, 2 : stores, 3 : none)

16. Type of apartment: Type of housing (1 : rent, 2 : own, 3 : for free)

17. No of Credits at this Bank: Numerical value showing number of existing credits at the bank

18. Occupation: Job (Qualitative) (1 : unemployed/ unskilled - non-resident, 2 : unskilled - resident, 3 : skilled employee / official, 4 : management/ self-employed/highly qualified employee/ officer)

19. No of dependents: Numerical value showing number of dependents

20. Telephone: Qualitative attribute for telephone number (1: yes, 2: No)

21. Foreign Worker: Qualitative attribute showing whether the person is the foreign worker or not (1: yes , 2: no)

2. Data Preparation:

2.1 Look at Attribute types:

There are a total of 21 attributes in the dataset.

- creditability: nominal (binary)
- Account Balance: nominal
- Duration of Credit: quantitative (continuous)
- Payment Status of Previous Credit: ordinal
- Purpose: nominal
- Credit Amount: quantitative (continuous)
- Value savings/Stocks: nominal
- Length of current employment: ordinal
- Instalment percent: quantitative(discrete)
- Sex & Martial Status:nominal
- Guarantors: nominal
- Duration in Current address: nominal
- Most valuable available asset: quantitative (discrete)
- Age (years): continuous
- Concurrent Credits: quantitative (discrete)
- Type of apartment: nominal
- No of Credits at this Bank: quantitative (discrete)
- Occupation: ordinal
- No of dependents: quantitative (discrete)
- Telephone: nominal

- Foreign Worker: Binary (Binary variables are nominal variables which have only two categories or levels).

2.2 Detect Missing Values

We checked and verified using python that fortunately there are no missing values in this dataset.

```
Foreign Worker                       0
Sex & Marital Status                 0
Account Balance                      0
Duration of Credit (month)           0
Payment Status of Previous Credit    0
Purpose                              0
Credit Amount                        0
Value Savings/Stocks                 0
Length of current employment         0
Instalment per cent                  0
Guarantors                           0
Telephone                            0
Duration in Current address          0
Most valuable available asset        0
Age (years)                          0
Concurrent Credits                   0
Type of apartment                    0
No of Credits at this Bank           0
Occupation                           0
No of dependents                     0
Creditability                        0
```

2.3 Statistic Summary of Variables

Below is the summary of some of the variables in the dataset:

Creditability:

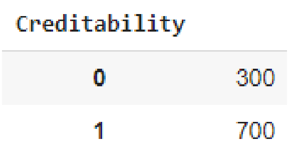
```
count    1000.000000
```

```
mean        0.700000
std         0.458487
min         0.000000
25%         0.000000
50%         1.000000
75%         1.000000
max         1.000000
Name: Creditability, dtype: float64

Duration of Credit (month)

count    1000.000000
mean       20.903000
std        12.058814
min         4.000000
25%        12.000000
50%        18.000000
75%        24.000000
max        72.000000
Name: Duration of Credit (month), dtype: float64

Credit Amount:

count     1000.00000
mean      3271.24800
std       2822.75176
min        250.00000
25%       1365.50000
50%       2319.50000
75%       3972.25000
max      18424.00000
Name: Credit Amount, dtype: float64

Age (years):

count    1000.00000
```

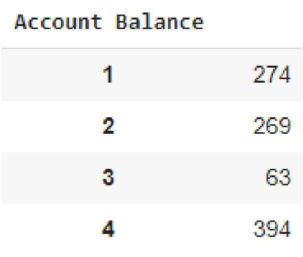
```
mean        35.54200
std         11.35267
min         19.00000
25%         27.00000
50%         33.00000
75%         42.00000
max         75.00000
Name: Age (years), dtype: float64

count     1000.000000
mean         2.577000
std          1.257638
min          1.000000
25%          1.000000
50%          2.000000
75%          4.000000
max          4.000000
Name: Account Balance, dtype: float64
```

2.4 Check Categorical Variables:

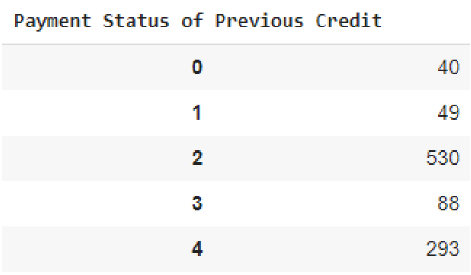
Check Categorical Variables by using Frequency Distribution Table.



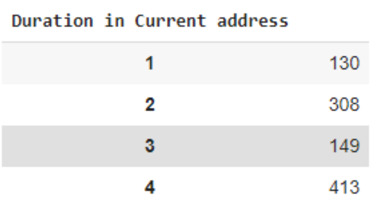
Creditability	
0	300
1	700



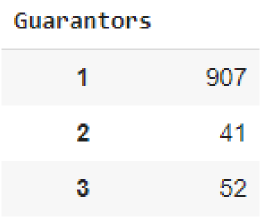
Account Balance	
1	274
2	269
3	63
4	394



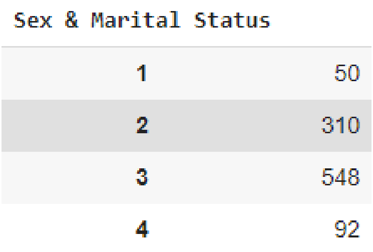
Payment Status of Previous Credit	
0	40
1	49
2	530
3	88
4	293



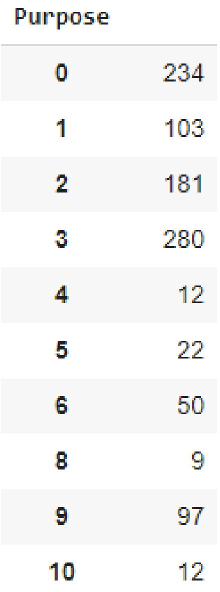
Duration in Current address	
1	130
2	308
3	149
4	413



Guarantors	
1	907
2	41
3	52

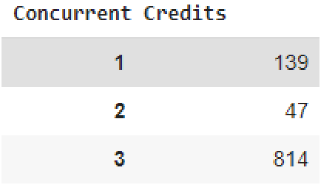


Sex & Marital Status	
1	50
2	310
3	548
4	92



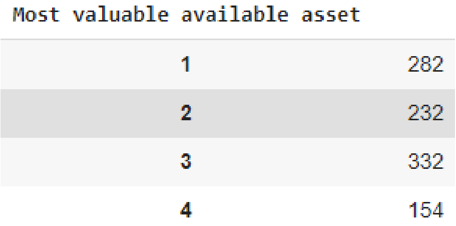
Purpose

0	234
1	103
2	181
3	280
4	12
5	22
6	50
8	9
9	97
10	12



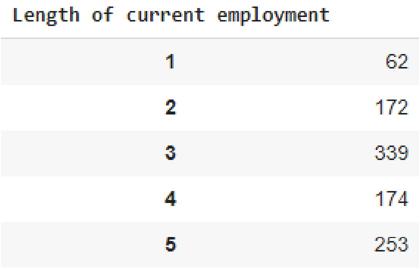
Concurrent Credits

1	139
2	47
3	814



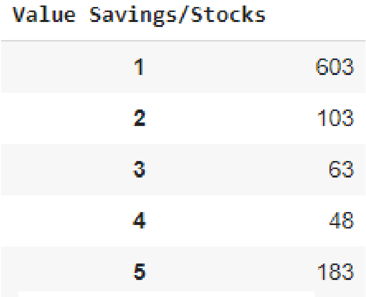
Most valuable available asset

1	282
2	232
3	332
4	154



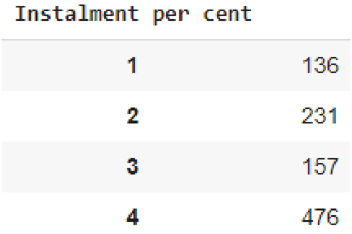
Length of current employment

1	62
2	172
3	339
4	174
5	253



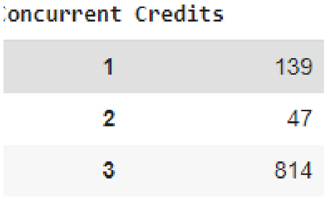
Value Savings/Stocks

1	603
2	103
3	63
4	48
5	183



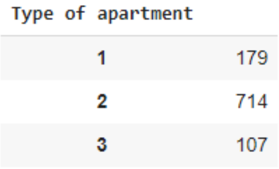
Instalment per cent

1	136
2	231
3	157
4	476



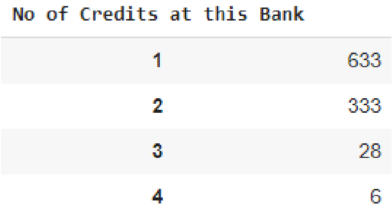
oncurrent Credits

1	139
2	47
3	814



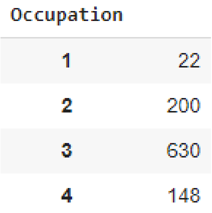
Type of apartment

1	179
2	714
3	107



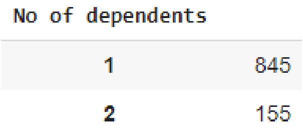
No of Credits at this Bank

1	633
2	333
3	28
4	6



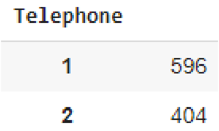
Occupation

1	22
2	200
3	630
4	148



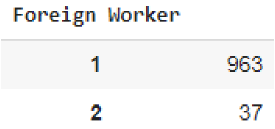
No of dependents

1	845
2	155



Telephone

1	596
2	404



Foreign Worker

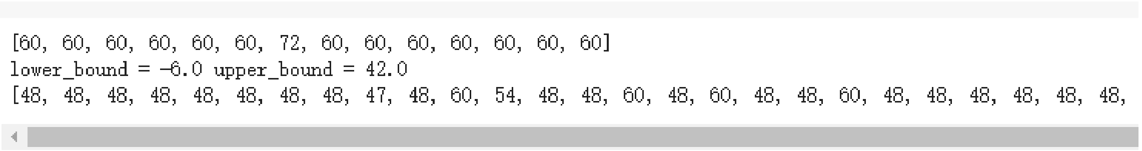
1	963
2	37

The target variable - “Creditability” is slightly imbalanced (3:7). There is no need to balance the dataset.

2.5 Detect Outliers on Continuous Variables

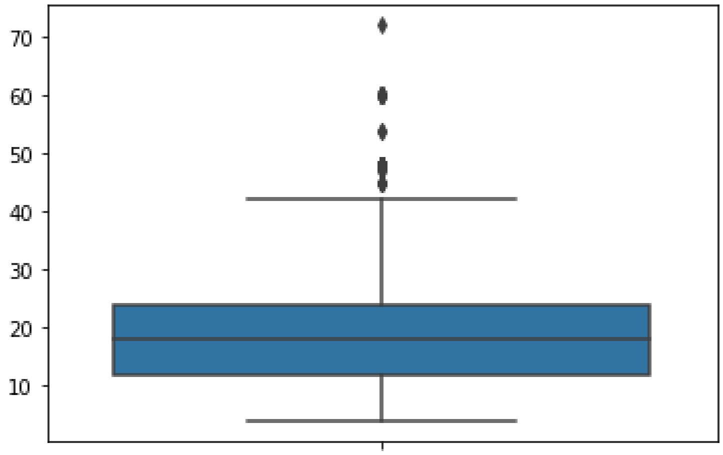
Use z-score and IQR two methods to detect outliers on continuous variables.

- **Duration of Credit (month)**

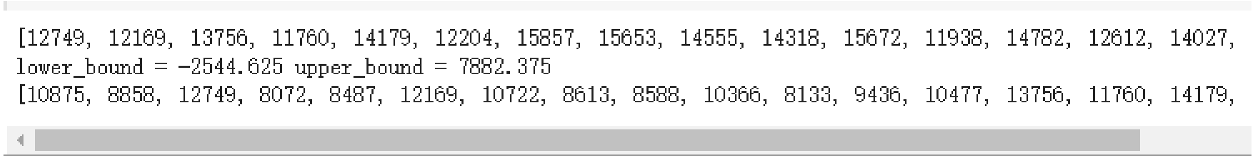


```
[60, 60, 60, 60, 60, 60, 72, 60, 60, 60, 60, 60, 60, 60]
lower_bound = -6.0 upper_bound = 42.0
[48, 48, 48, 48, 48, 48, 48, 48, 47, 48, 60, 54, 48, 48, 60, 48, 60, 48, 48, 60, 48, 48, 48, 48, 48, 48,
```

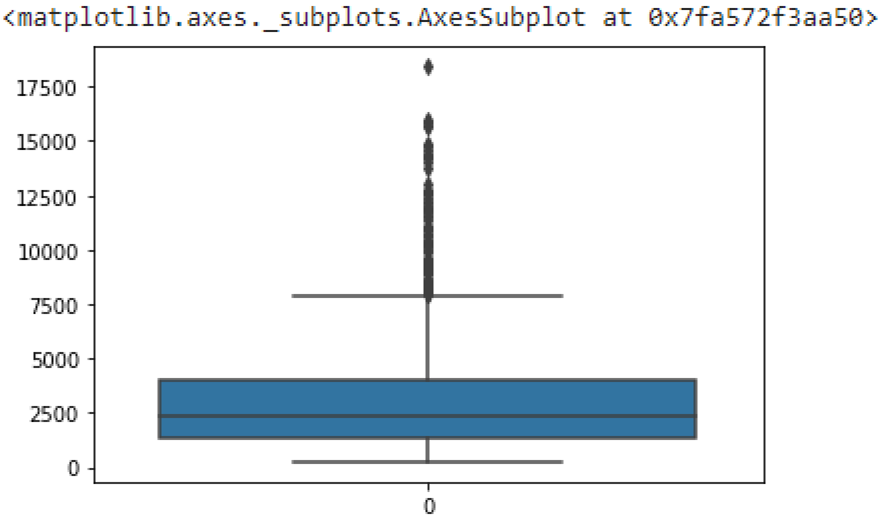
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa572fdbf10>
```



- **Credit Amount**

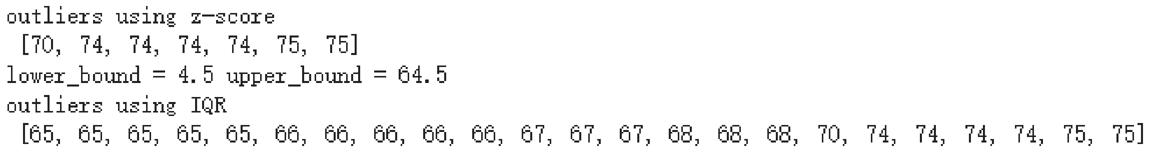


```
[12749, 12169, 13756, 11760, 14179, 12204, 15857, 15653, 14555, 14318, 15672, 11938, 14782, 12612, 14027,
lower_bound = -2544.625 upper_bound = 7882.375
[10875, 8858, 12749, 8072, 8487, 12169, 10722, 8613, 8588, 10366, 8133, 9436, 10477, 13756, 11760, 14179,
```

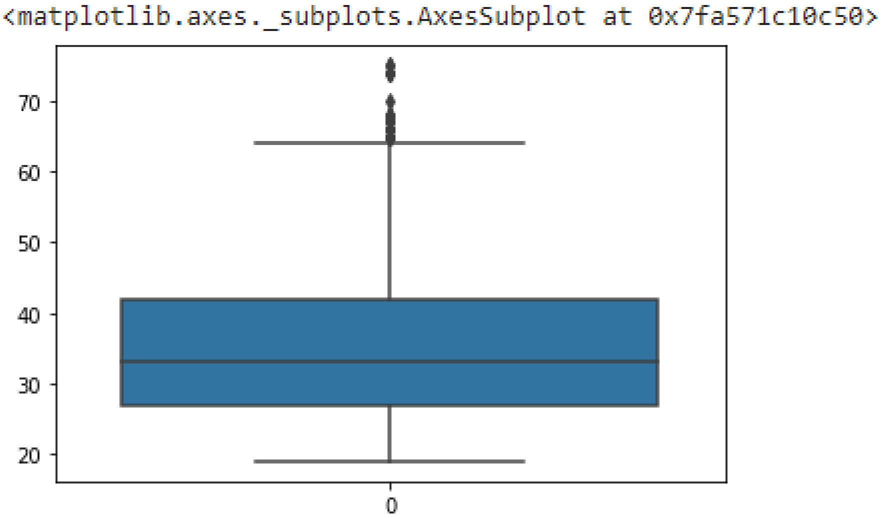


```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa572f3aa50>
```

- **Age (years)**



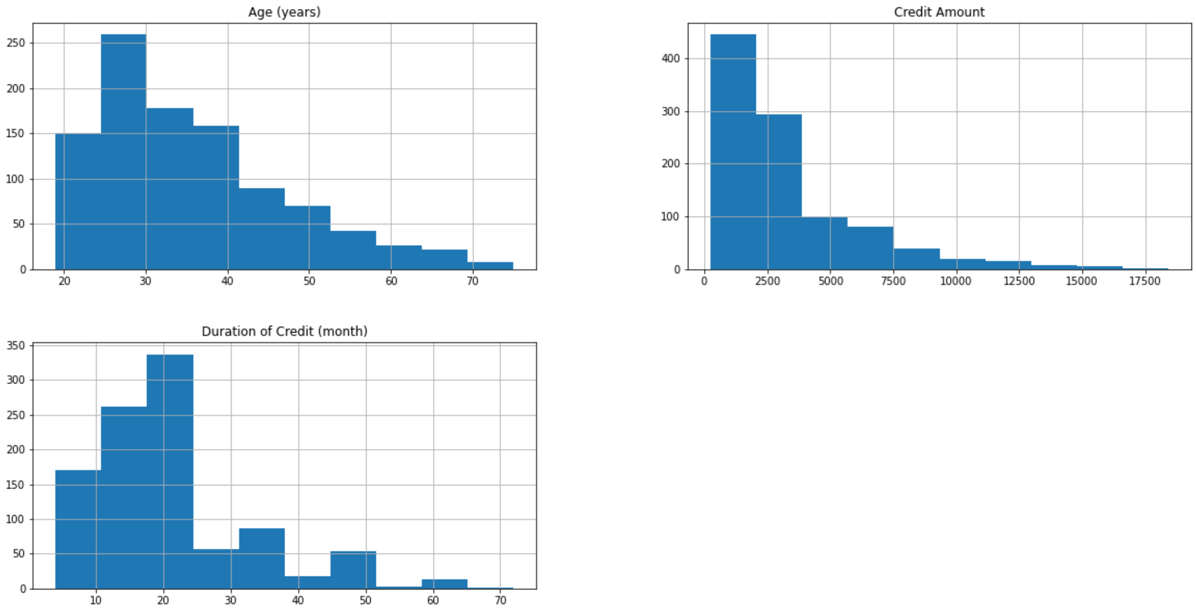
```
outliers using z-score
 [70, 74, 74, 74, 74, 75, 75]
lower_bound = 4.5 upper_bound = 64.5
outliers using IQR
 [65, 65, 65, 65, 65, 66, 66, 66, 66, 66, 67, 67, 67, 68, 68, 68, 70, 74, 74, 74, 74, 75, 75]
```



```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa571c10c50>
```

2.6 Check Distribution of Variables with Numeric Values

Plot histogram of Duration of Credit (month), Credit Amount, Age (years).

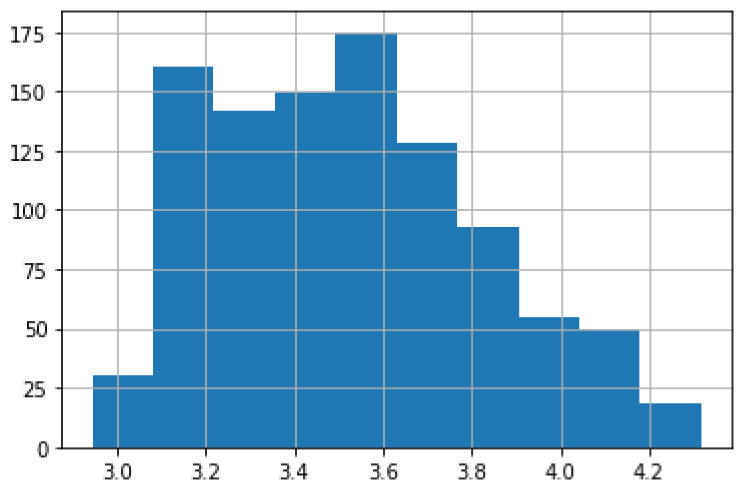


These three variables all have right skewed distribution.

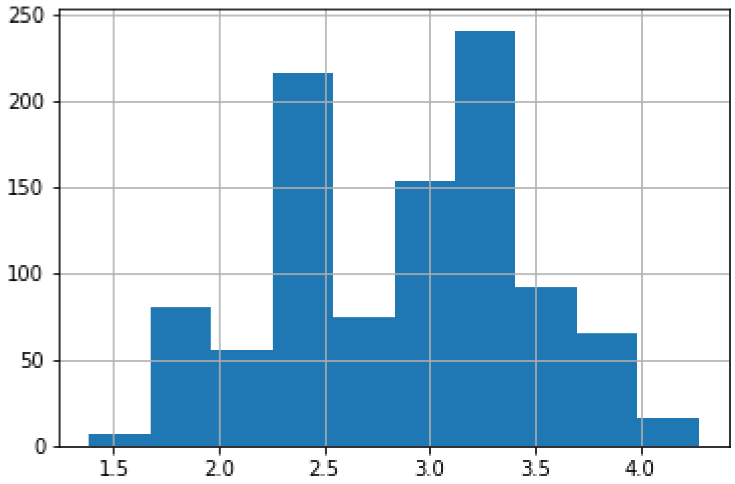
2.7 Data transformation on 'Credit Amount' , 'Age(years)' and 'Duration of Credit(month)'

Log transformation A log transformation can help to fit a very skewed distribution into a Gaussian one. After log transformation we can easily see patterns in our data.

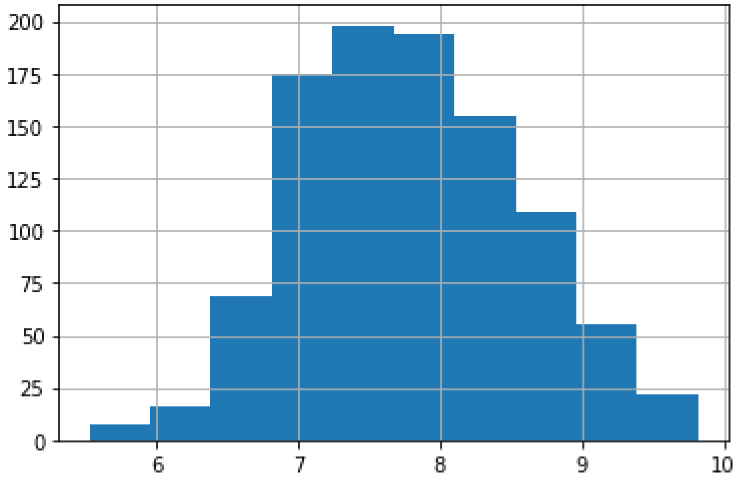
“Age(years)“ after transformation:



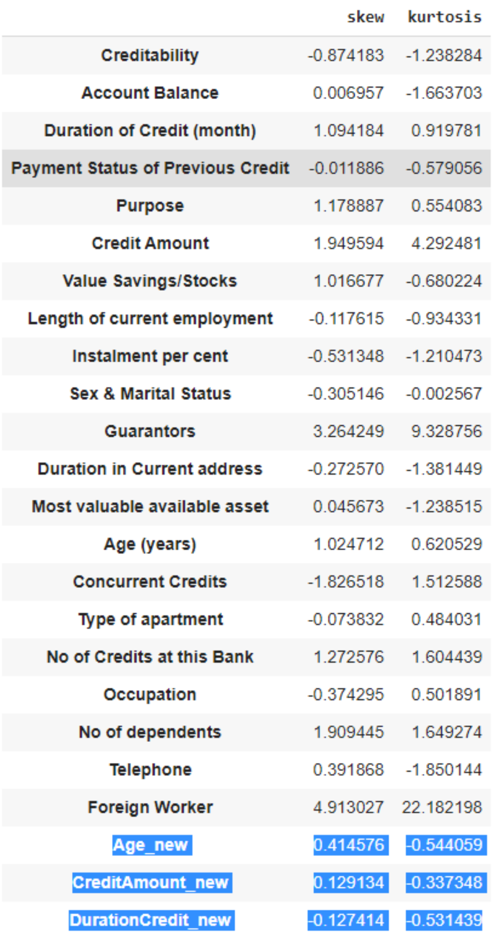
“Duration of Credit (month)” after transformation:



“Credit Amount” after transformation:



After transformation, we check the skewness which is between -0.5 to 0.5.



	skew	kurtosis
Creditability	-0.874183	-1.238284
Account Balance	0.006957	-1.663703
Duration of Credit (month)	1.094184	0.919781
Payment Status of Previous Credit	-0.011886	-0.579056
Purpose	1.178887	0.554083
Credit Amount	1.949594	4.292481
Value Savings/Stocks	1.016677	-0.680224
Length of current employment	-0.117615	-0.934331
Instalment per cent	-0.531348	-1.210473
Sex & Marital Status	-0.305146	-0.002567
Guarantors	3.264249	9.328756
Duration in Current address	-0.272570	-1.381449
Most valuable available asset	0.045673	-1.238515
Age (years)	1.024712	0.620529
Concurrent Credits	-1.826518	1.512588
Type of apartment	-0.073832	0.484031
No of Credits at this Bank	1.272576	1.604439
Occupation	-0.374295	0.501891
No of dependents	1.909445	1.649274
Telephone	0.391868	-1.850144
Foreign Worker	4.913027	22.182198
Age_new	0.414576	-0.544059
CreditAmount_new	0.129134	-0.337348
DurationCredit_new	-0.127414	-0.531439

3. Exploratory Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.It is a good practice to understand the data first and try to gather as many insights from it. Exploratory Data Analysis is all about making sense of data in hand, before getting them dirty with it.

Feature Selection Now it's time to finally choose the best columns(Features) which are correlated to the Target variable. This can be done directly by measuring the correlation values or ANOVA/Chi-Square tests. However, it is always helpful to visualize the relation between the Target variable and each of the predictors to get a better sense of data.

I have listed below the techniques used for visualizing the relationship between two variables as well as measuring the strength statistically.

Visual exploration of relationship between variables

- Continuous Vs Continuous ---- Scatter Plot
- Categorical Vs Continuous---- Box Plot
- Categorical Vs Categorical---- Grouped Bar Plots

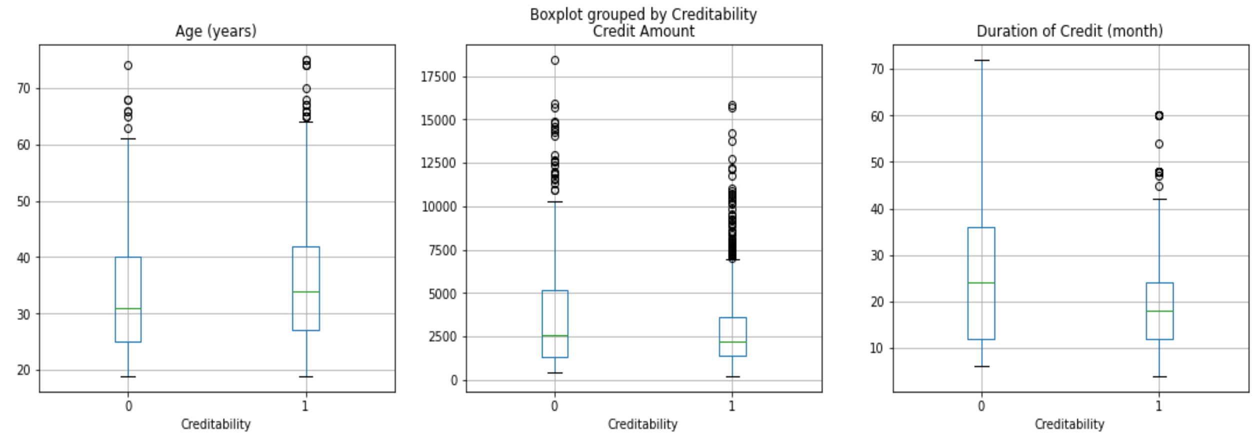
Statistical measurement of relationship strength between variables

- Continuous Vs Continuous ---- Correlation matrix
- Categorical Vs Continuous---- ANOVA test
- Categorical Vs Categorical--- Chi-Square test

In this case study the Target variable is categorical, hence below two scenarios will be present

3.1 Analyzing Correlation between Categorical (target) and Continuous variables

When the target variable is Categorical and the predictor variable is Continuous we analyze the relation using bar plots/Boxplots and measure the strength of relation using Anova test



Box-Plots interpretation What should you look for in these box plots?

These plots give an idea about the data distribution of continuous predictors in the Y-axis for each of the categories in the X-Axis.

If the distribution looks similar for each category(Boxes are in the same line), that means that the continuous variable has NO effect on the target variable. Hence, the variables are not correlated to each other.

For example, look at the first chart "Age (years)" Vs "Creditability". The boxes are in the similar line! It means that people whose loan was rejected and whose loan was approved have the same kind of age. Hence, I cannot distinguish between approval and rejection based on the age of an applicant. So this column is NOT correlated with the Creditability.

The other two charts also exhibit opposite characteristics, hence "Credit Amount" and "Duration of Credit (month)" are correlated with the target variable.

We confirm this by looking at the results of ANOVA test below

Statistical Feature Selection (Categorical Vs Continuous) using ANOVA test Analysis of variance(ANOVA) is performed to check if there is any relationship between the given continuous and categorical variable

- Assumption(H0): There is NO relation between the given variables (i.e. The average(mean) values of the numeric Predictor variable is same for all the groups in the categorical Target variable)
- ANOVA Test result: Probability of H0 being true

```
##### ANOVA Results #####


Age (years) is correlated with Creditability | P-Value:
0.003868455281308189

Credit Amount is correlated with Creditability | P-Value:
8.795399017206301e-07

Duration of Credit (month) is correlated with Creditability | P-Value:
6.488049877187189e-12

['Age (years)', 'Credit Amount', 'Duration of Credit (month)']
```

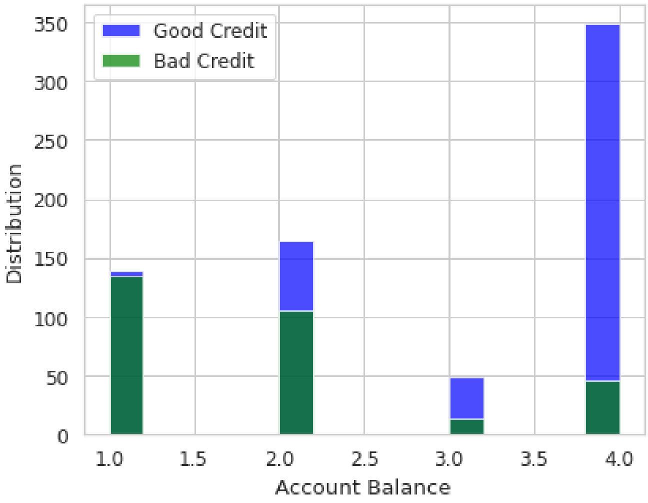
The results of ANOVA confirm our visual analysis using box plots above!

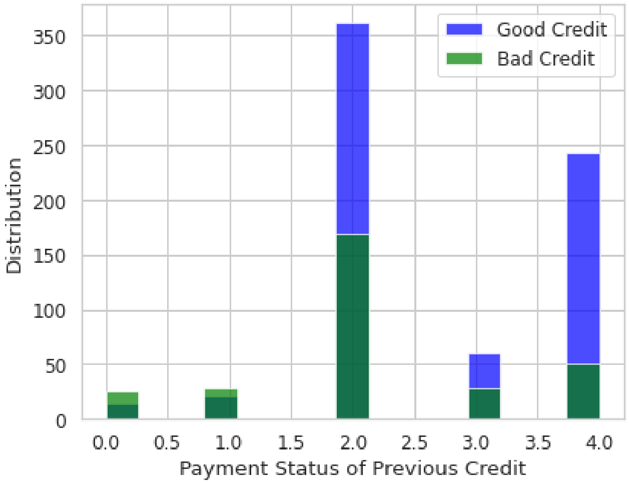
Notice the P-Value of "Age (years)", it is just at the boundary of the threshold. This is something we already doubted in the box plots section already.

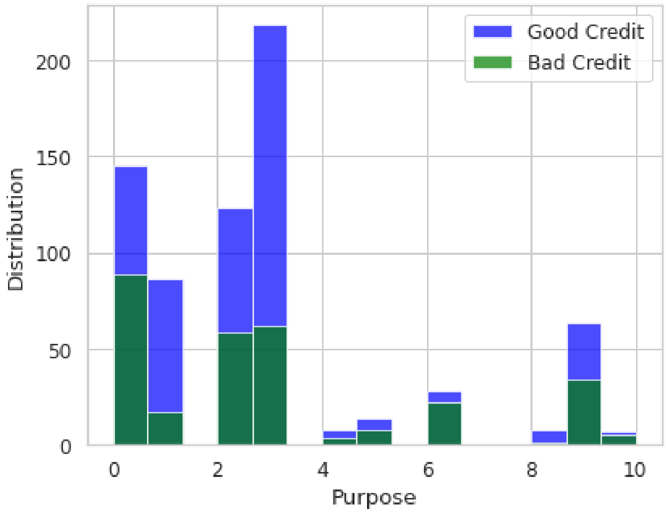
While the other two P-Values are clearly zero, hence they are correlated without doubt.

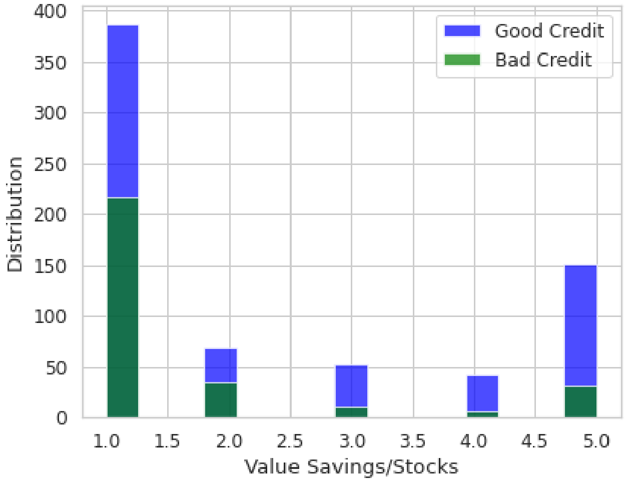
All three columns are correlated with Creditability.

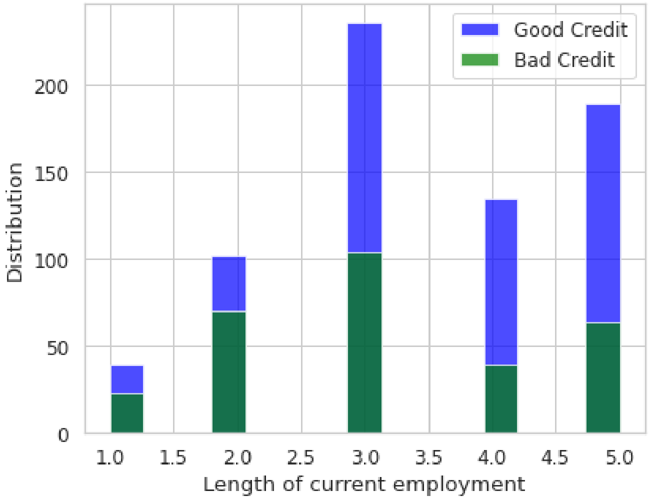
3.2 Analyzing Correlation between Categorical (target) and Categorical variables

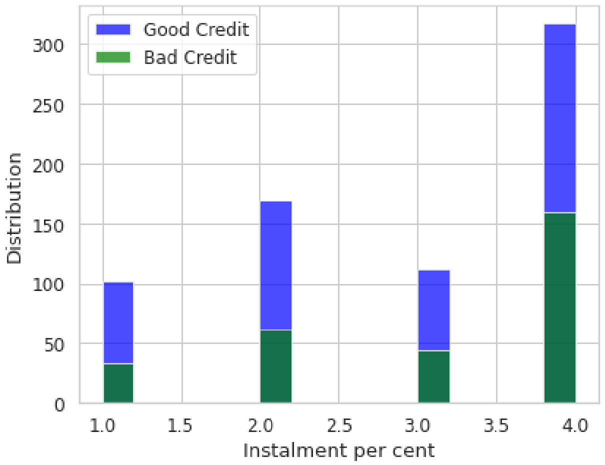


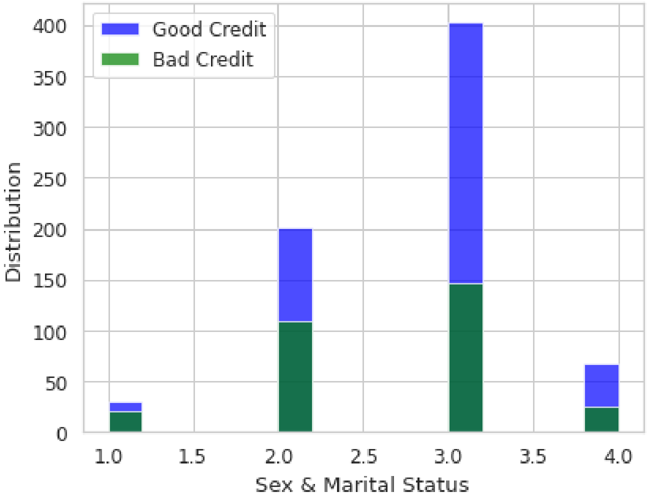


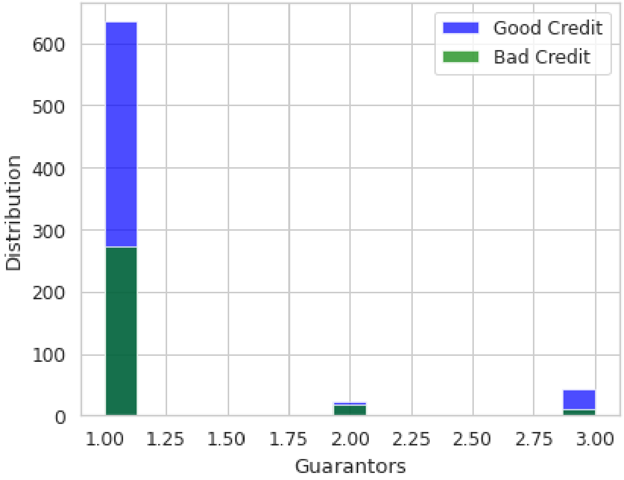

Good Credit
Bad Credit
Distribution
0
50
100
150
200
0
2
4
6
8
10
Purpose

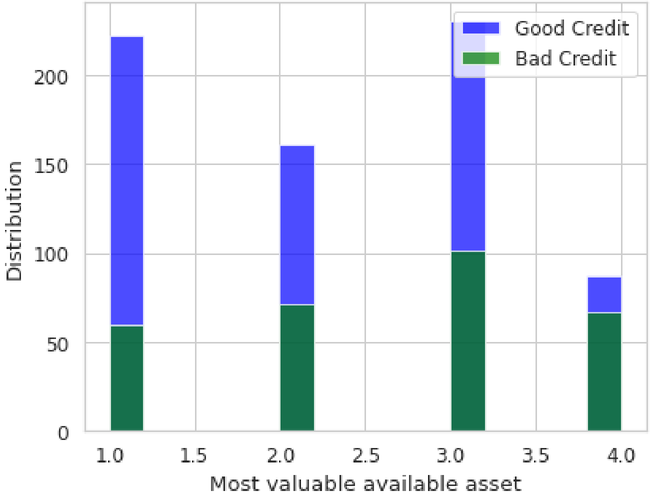

Good Credit
Bad Credit
Distribution
0
50
100
150
200
250
300
350
400
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Value Savings/Stocks

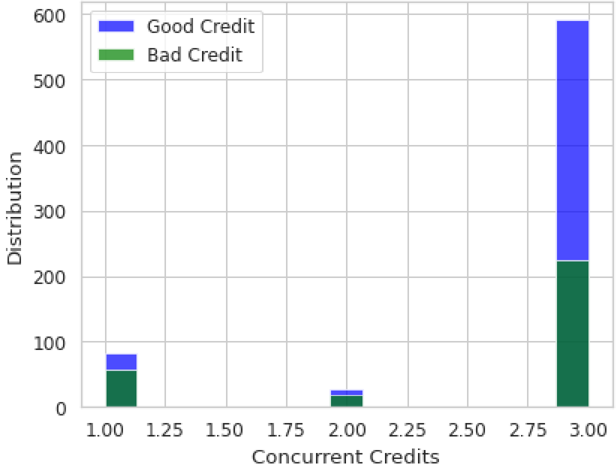

Good Credit
Bad Credit
Distribution
0
50
100
150
200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Length of current employment

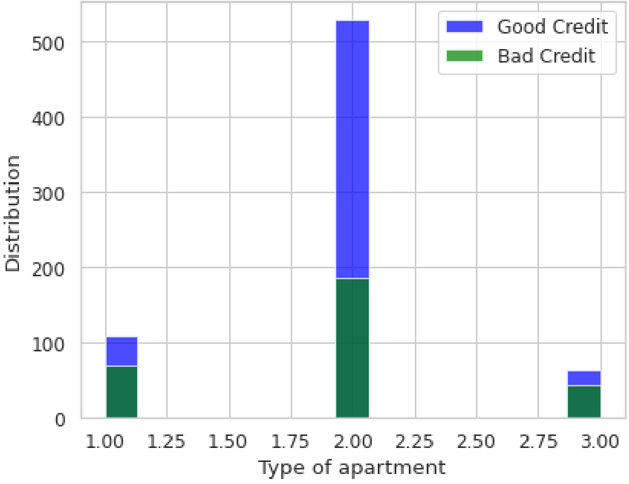

Good Credit
Bad Credit
Distribution
0
50
100
150
200
250
300
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Instalment per cent

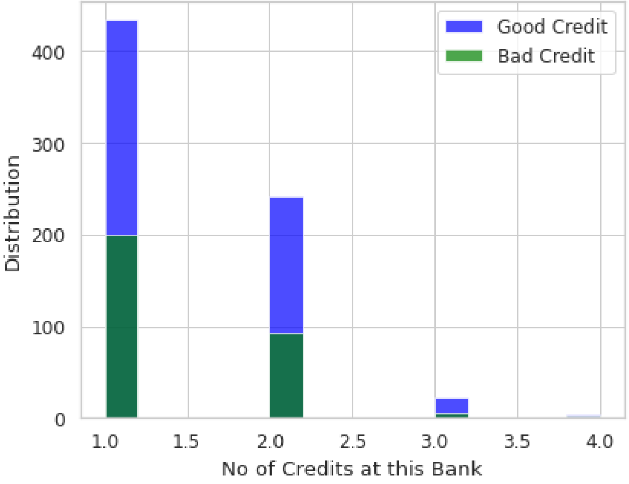

Good Credit
Bad Credit
Distribution
0
50
100
150
200
250
300
350
400
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Sex & Marital Status

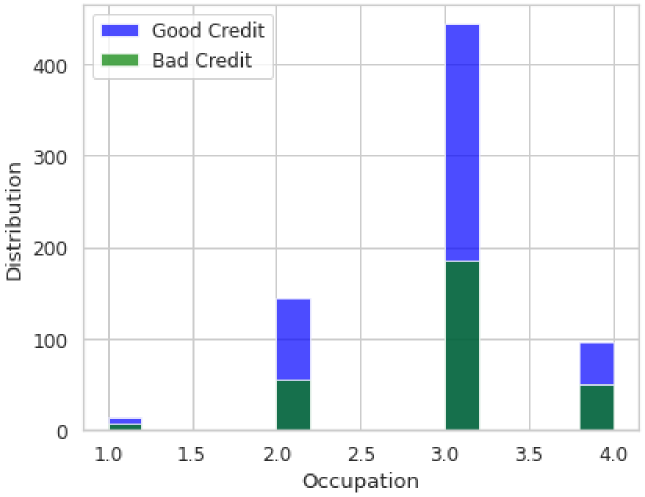

Good Credit
Bad Credit
Distribution
0
100
200
300
400
500
600
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Guarantors

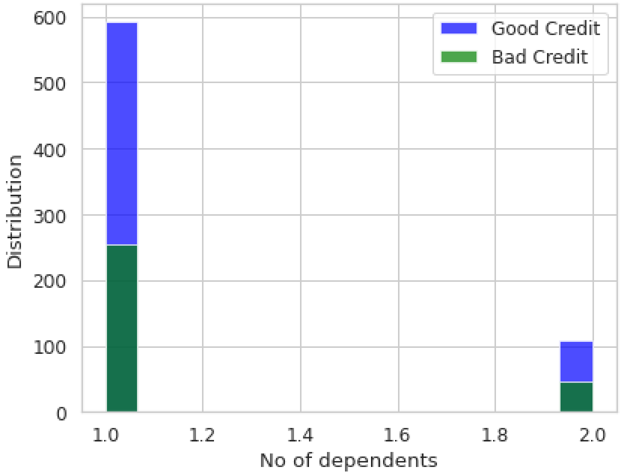

Good Credit
Bad Credit
Distribution
0
50
100
150
200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Most valuable available asset

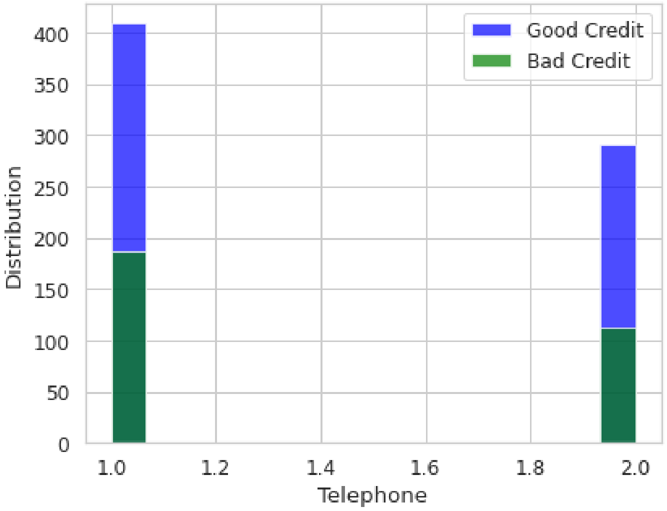

Good Credit
Bad Credit
Distribution
0
100
200
300
400
500
600
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Concurrent Credits

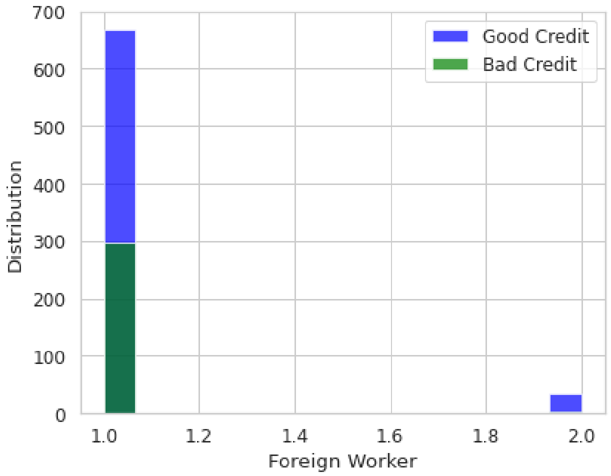

Good Credit
Bad Credit
Distribution
0
100
200
300
400
500
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Type of apartment


Good Credit
Bad Credit
Distribution
0
100
200
300
400
1.0
1.5
2.0
2.5
3.0
3.5
4.0
No of Credits at this Bank









3.3 Selection of Attributes

According to the histogram of correlations between the independent (categorical) and dependent (categorical) variables, several variables have an effect on the target class. For example, as they change between groups, the ratio of "bad credit" to "good credit" changes accordingly. They are considered to have a strong correlation with the target.

- Account Balance
- Payment Status of Previous Credit

- Purpose
- Value Savings/Stocks
- Length of current employment
- Sex & Marital Status
- Guarantors
- Concurrent Credits
- Type of apartment
- Occupation
- Foreign Worker

Together with 'Age(year)', 'Credit Amount', and 'Duration of Credit (month)', these variables are the selected features for following steps.

- The class distribution is slightly imbalanced (3:7). We don't need to deal with it.
- There are no missing values in this dataset.
- 'Credit Amount' and 'Age(years)' are highly skewed, which need to be transformed.
- 'Duration of Credit(month)' has several groups with less number, it needs to be discretized.

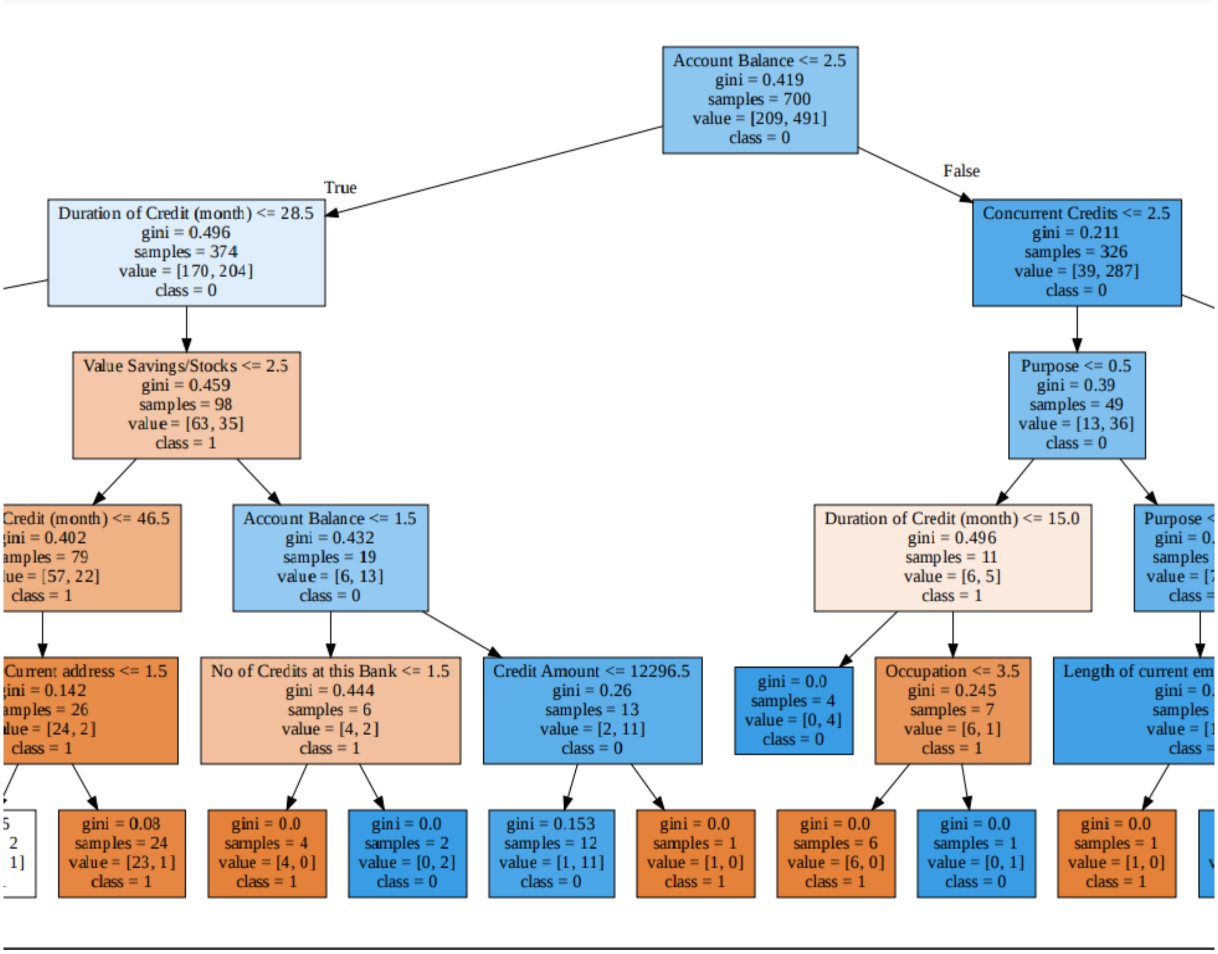
4. Predictive Modeling

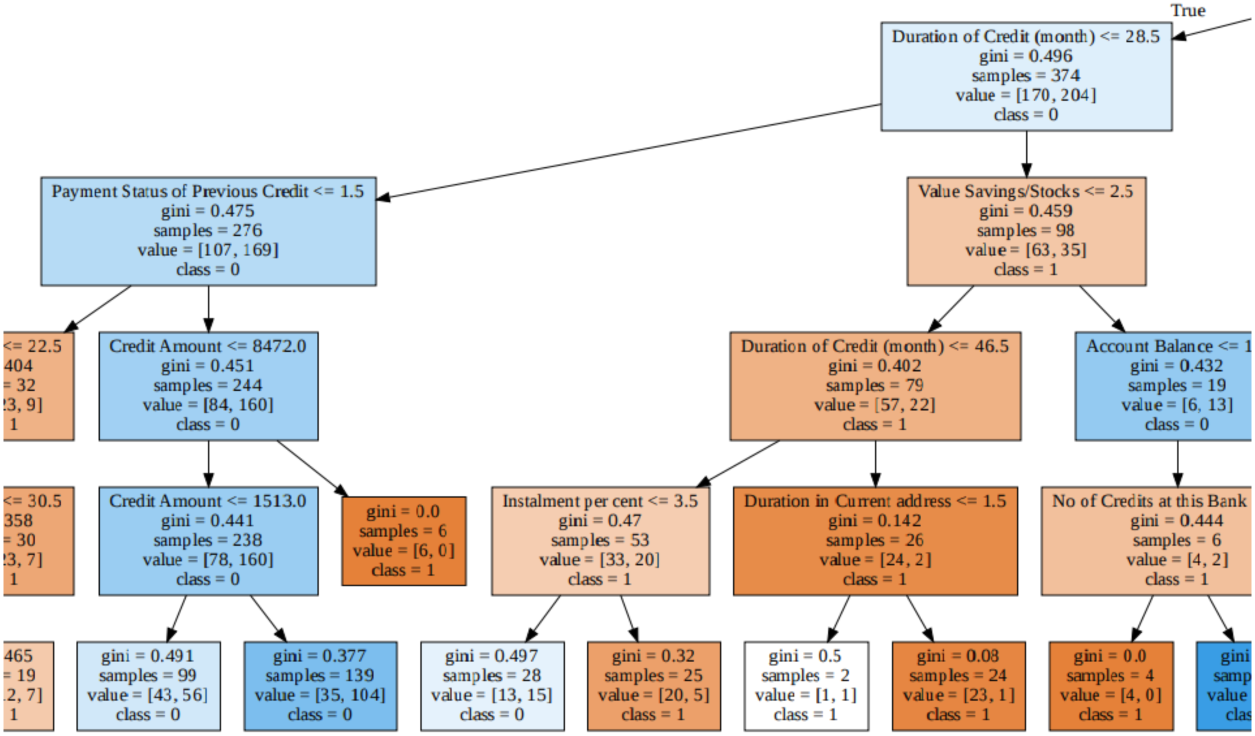
As the target variable is categorical type, which is a classification task. We choose Decision Tree and Naive Bayes as predictive models. Since these two methods are not sensitive with Normal Distribution. We can use the original data before transformation.

Before training the models, we split the dataset into training (70%) and test (30%) datasets.

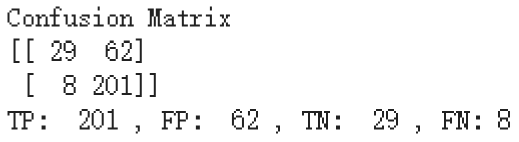
4.1 Classification using Decision Tree

By leveraging the Decision Tree model from sklearn package and setting the max depth of the tree to 5, then train the model and use grahviz package to visualize the model result.

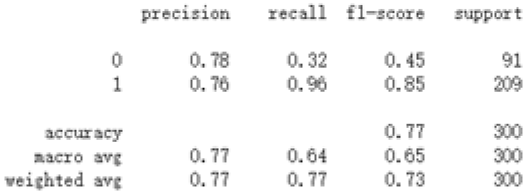




Finally, let's evaluate the prediction accuracy by checking the confusion matrix.



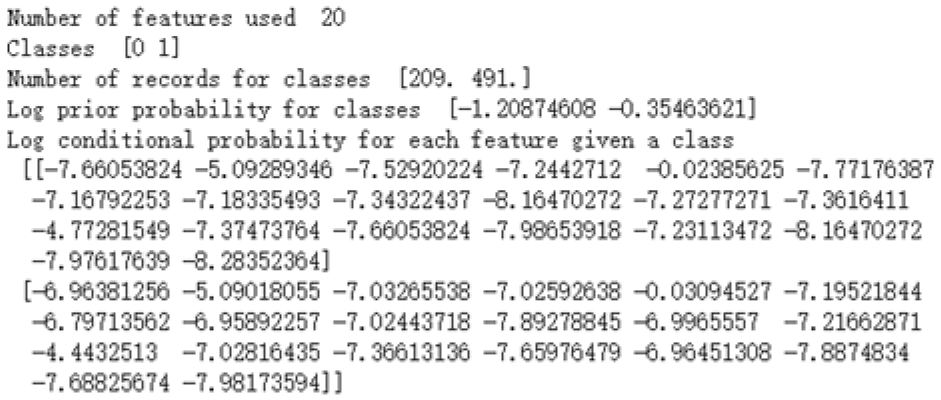
```
Confusion Matrix
[[ 29  62]
 [  8 201]]
TP:  201 , FP:  62 , TN:  29 , FN: 8
```



```
              precision    recall  f1-score   support

           0       0.78      0.32      0.45        91
           1       0.76      0.96      0.85       209

    accuracy                           0.77       300
   macro avg       0.77      0.64      0.65       300
weighted avg       0.77      0.77      0.73       300
```

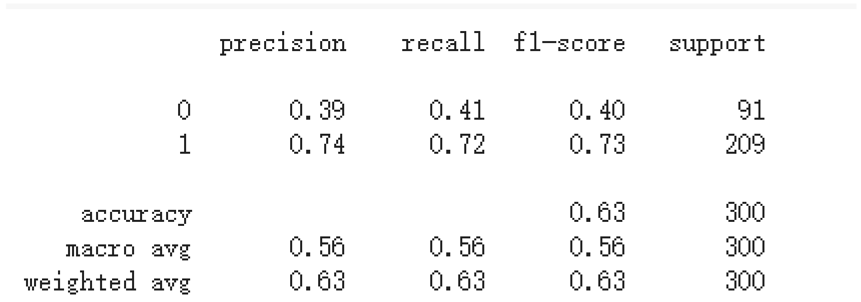
4.2 Classification using Naive Bayes

This time, we use the Naïve Bayes model on the training data.



```
Number of features used  20
Classes  [0 1]
Number of records for classes  [209. 491.]
Log prior probability for classes  [-1.20874608 -0.35463621]
Log conditional probability for each feature given a class
 [[-7.66053824 -5.09289346 -7.52920224 -7.2442712  -0.02385625 -7.77176387
  -7.16792253 -7.18335493 -7.34322437 -8.16470272 -7.27277271 -7.3616411
  -4.77281549 -7.37473764 -7.66053824 -7.98653918 -7.23113472 -8.16470272
  -7.97617639 -8.28352364]
 [-6.96381256 -5.09018055 -7.03265538 -7.02592638 -0.03094527 -7.19521844
  -6.79713562 -6.95892257 -7.02443718 -7.89278845 -6.9965557  -7.21662871
  -4.4432513  -7.02816435 -7.36613136 -7.65976479 -6.96451308 -7.8874834
  -7.68825674 -7.98173594]]
```

Evaluate the model.

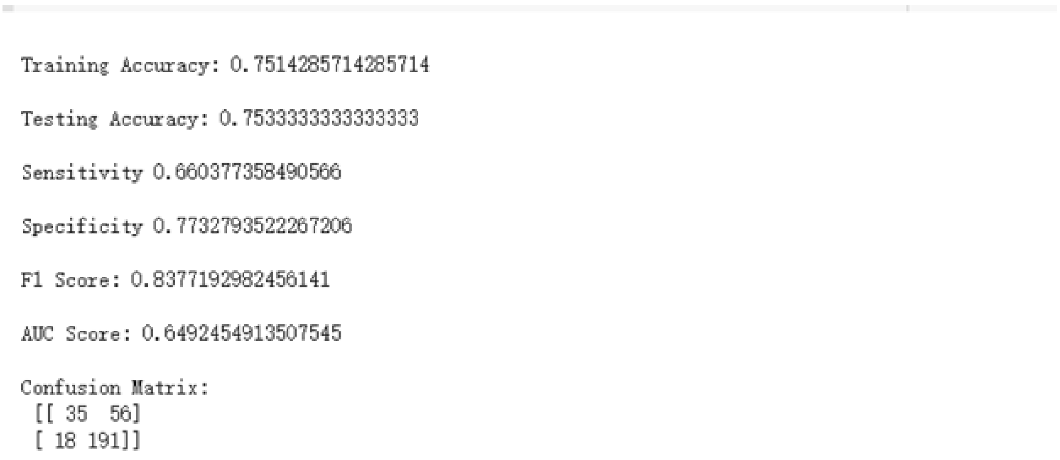


```
              precision    recall  f1-score   support

           0       0.39      0.41      0.40        91
           1       0.74      0.72      0.73       209

    accuracy                           0.63       300
   macro avg       0.56      0.56      0.56       300
weighted avg       0.63      0.63      0.63       300
```

4.3 Additional Data Analysis

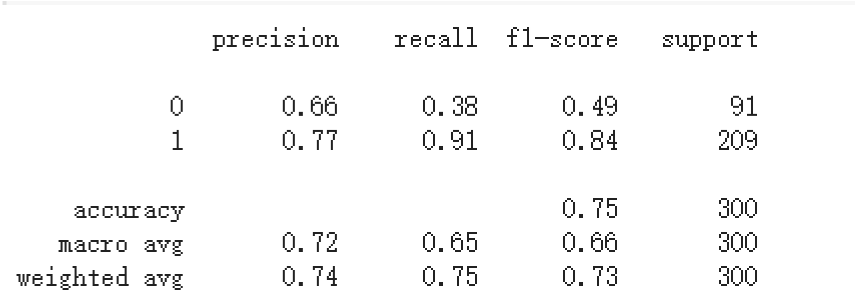
4.3.1 Classification using Random Forest

Random Forest is another classification model we use on the training data.



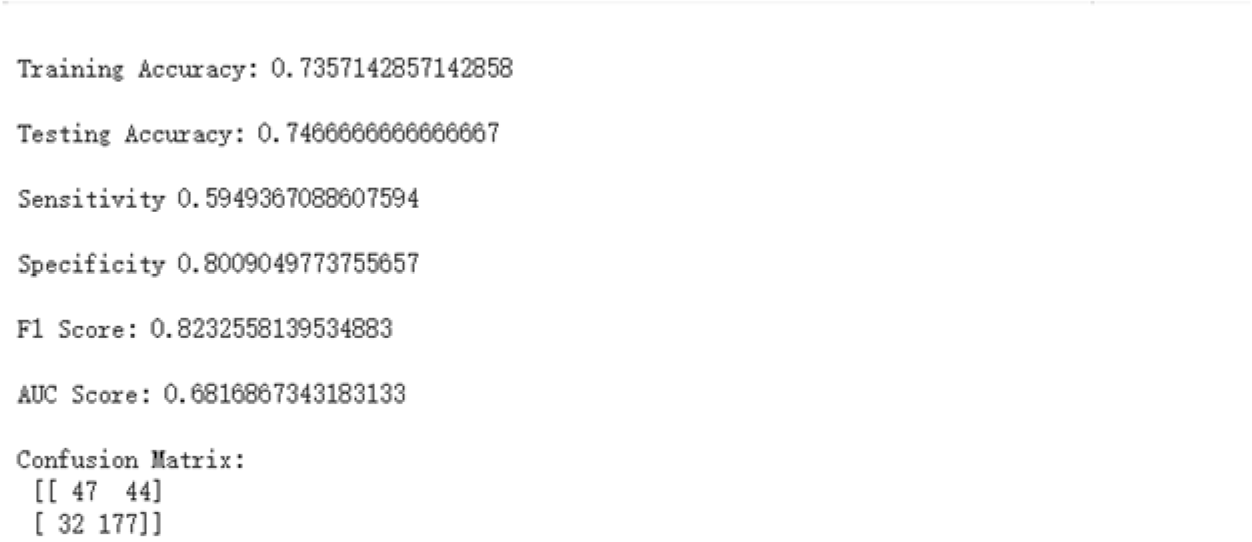
```
Training Accuracy: 0.7514285714285714

Testing Accuracy: 0.7533333333333333

Sensitivity 0.660377358490566

Specificity 0.7732793522267206

F1 Score: 0.8377192982456141

AUC Score: 0.6492454913507545

Confusion Matrix:
 [[ 35  56]
 [ 18 191]]
```

Evaluate the result.



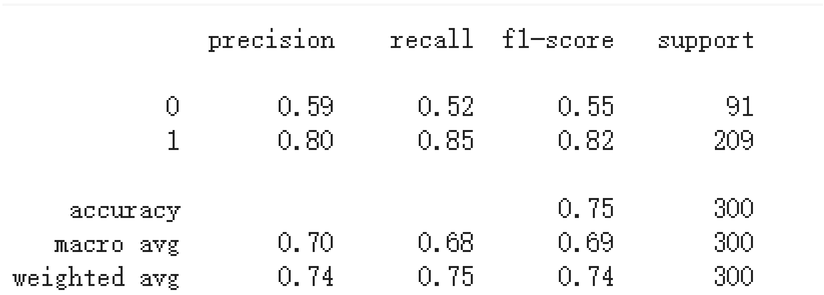
```
              precision    recall  f1-score   support

           0       0.66      0.38      0.49        91
           1       0.77      0.91      0.84       209

    accuracy                           0.75       300
   macro avg       0.72      0.65      0.66       300
weighted avg       0.74      0.75      0.73       300
```

4.3.2 Classification using Gradient Boosting



```
Training Accuracy: 0.7357142857142858

Testing Accuracy: 0.7466666666666667

Sensitivity 0.5949367088607594

Specificity 0.8009049773755657

F1 Score: 0.8232558139534883

AUC Score: 0.6816867343183133

Confusion Matrix:
 [[ 47  44]
 [ 32 177]]
```

Evaluate the result.



```
              precision    recall  f1-score   support

           0       0.59      0.52      0.55        91
           1       0.80      0.85      0.82       209

    accuracy                           0.75       300
   macro avg       0.70      0.68      0.69       300
weighted avg       0.74      0.75      0.74       300
```

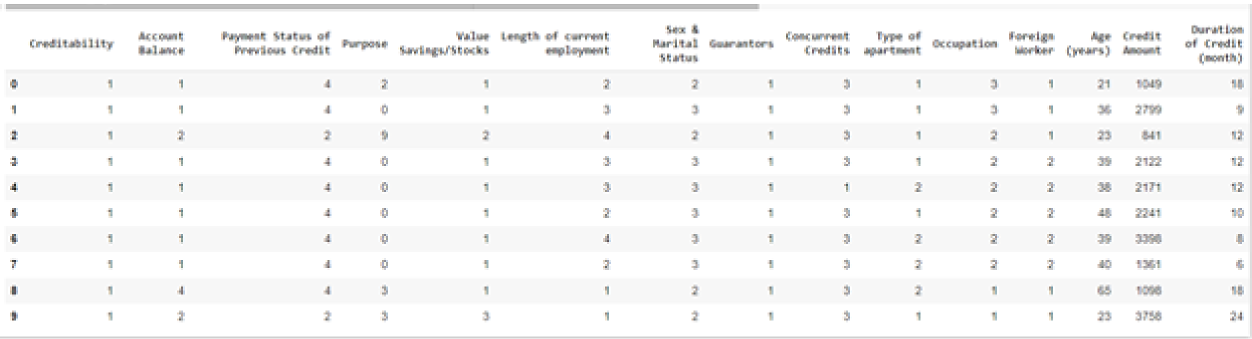
4.4 Performance Metrics Comparison

	Accuracy	Precision	Recall
Decision tree	0.77	0.76	0.96
Naive Bayes	0.63	0.74	0.72
Random forest	075	0.77	0.91
Super Vector Machine	0.75	0.80	0.85

We can see from the comparison chart that Decision Tree performs best due to high accuracy, precision and recall rates.

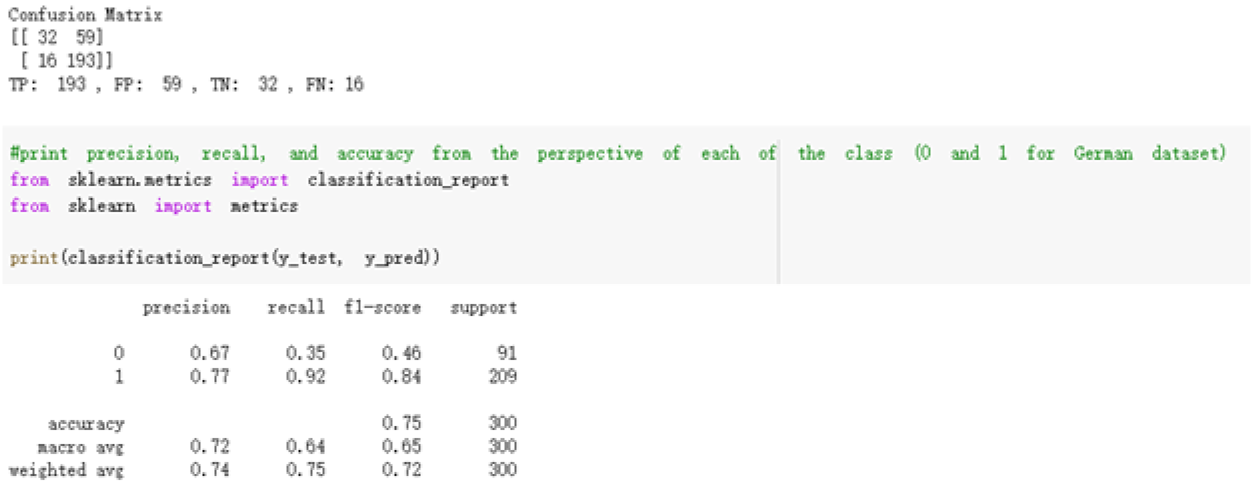
4.6 Performance Comparison between “all attributes” and “selected attributes”

Subset datasets with selected (14) features and split it into training and test datasets.



	Creditability	Account Balance	Payment Status of Previous Credit	Purpose	Value Savings/Stocks	Length of current employment	Sex & Marital Status	Guarantors	Concurrent Credits	Type of apartment	Occupation	Foreign Worker	Age (years)	Credit Amount	Duration of Credit (month)
0	1	1	4	2	1	2	2	1	3	1	3	1	21	1049	18
1	1	1	4	0	1	3	3	1	3	1	3	1	36	2799	9
2	1	2	2	9	2	4	2	1	3	1	2	1	23	841	12
3	1	1	4	0	1	3	3	1	3	1	2	2	39	2122	12
4	1	1	4	0	1	3	3	1	1	2	2	2	38	2171	12
5	1	1	4	0	1	2	3	1	3	1	2	2	48	2241	10
6	1	1	4	0	1	4	3	1	3	2	2	2	39	3398	8
7	1	1	4	0	1	2	3	1	3	2	2	2	40	1361	6
8	1	4	4	3	1	1	2	1	3	2	1	1	65	1098	18
9	1	2	2	3	3	1	2	1	3	1	1	1	23	3758	24

Training Decision Tree model on selected attributes.

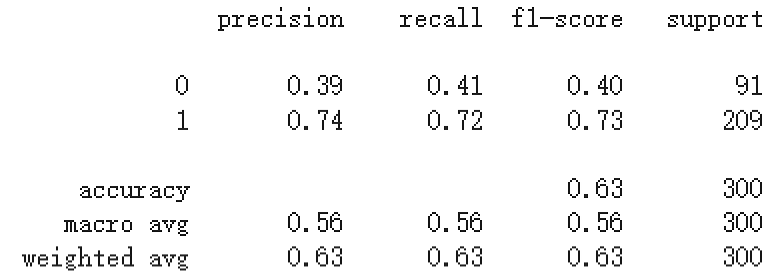


```
Confusion Matrix
[[ 32  59]
 [ 16 193]]
TP:  193 , FP:  59 , TN:  32 , FN: 16
```

```
#print precision, recall, and accuracy from the perspective of each of the class (0 and 1 for German dataset)
from sklearn.metrics import classification_report
from sklearn import metrics

print(classification_report(y_test, y_pred))
```

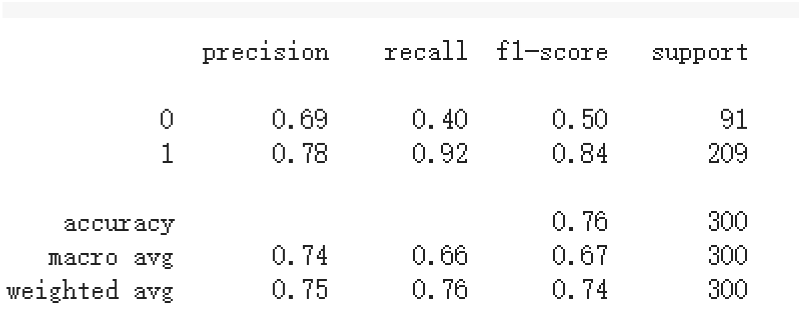
```
              precision    recall  f1-score   support

           0       0.67      0.35      0.46        91
           1       0.77      0.92      0.84       209

    accuracy                           0.75       300
   macro avg       0.72      0.64      0.65       300
weighted avg       0.74      0.75      0.72       300
```

Training Naïve Bayes model on selected dataset. There is no difference between the results of all features and selected features.



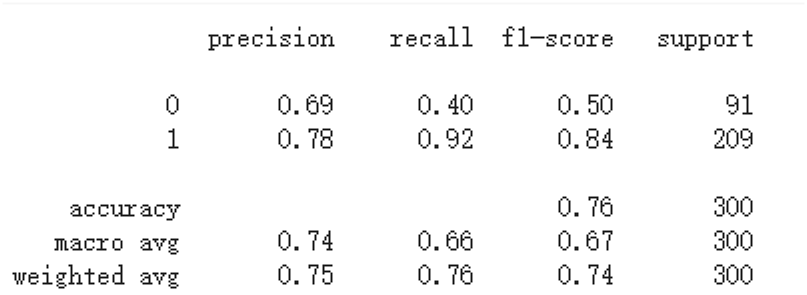
```
              precision    recall  f1-score   support

           0       0.39      0.41      0.40        91
           1       0.74      0.72      0.73       209

    accuracy                           0.63       300
   macro avg       0.56      0.56      0.56       300
weighted avg       0.63      0.63      0.63       300
```

Training Random Forest model on selected dataset.



```
              precision    recall  f1-score   support

           0       0.69      0.40      0.50        91
           1       0.78      0.92      0.84       209

    accuracy                           0.76       300
   macro avg       0.74      0.66      0.67       300
weighted avg       0.75      0.76      0.74       300
```

Training Gradient Boosting model on selected dataset.



```
              precision    recall  f1-score   support

           0       0.69      0.40      0.50        91
           1       0.78      0.92      0.84       209

    accuracy                           0.76       300
   macro avg       0.74      0.66      0.67       300
weighted avg       0.75      0.76      0.74       300
```

Comparison between the model results of all features and selected features.

	Accuracy	Precision	Recall
Decision tree (all)	0.77	0.76	0.96
Decision tree (select)	0.75	0.77	0.92
Naive Bayes (all)	0.63	0.74	0.72
Naive Bayes (select)	0.63	0.74	0.72
Random Forest (all)	075	0.77	0.91

Random Forest (select)	0.76	0.78	0.92
Super Vector Machine (all)	0.75	0.80	0.85
Super Vector Machine (select)	0.76	0.78	0.92

5. Conclusion

5.1 Major Findings

By comparing the confusion matrix of these models, we can conclude that Decision Tree outperforms the other models. After feature selection, only Random Forest and Super Vector Machine slightly improve the results, while Decision Tree has poorer predictions and Naïve Bayes has no change.

As for Exploration Analysis, ANOVA test can be used to visualize the relationship between categorical variables and continuous variables, while Grouped Bar Plots is good at demonstrating the relationship between categorical variables. Then, we can compare the impact of each feature on the target and make a selection based on the analysis of the plots.

In General, most of the classification models are not sensitive to the distribution of continuous variables. So we can use the data before transformation in this scenario.

5.2 Recommendation

There are many classification models in predicting “creditability”, such as Decision Tree, Naive bayes, Random Forest , Super Vector Machine and so on .Our recommended priority classification method is the Decision Tree, which is simple, straightforward and insensitive to feature selection, and also has the best performance.

The project could be extended to assign weights to the various variables before building and training a predictive model which will simulate if an applicant is creditable or not .

Modern credit analyses employ many additional variables like the criminal records of applicants, their health information, net balance between monthly income and expenses. A dataset with these variables could be acquired or complementary variables added to the dataset. This will make the credit simulations much realistic, similar to what is done by the banks before a credit is approved.

In order to create a practical and useful application from this study, we could develop a credit risk management tool for peer to peer lending companies. This tool could provide for instance the ideal interest rate for a loan in order to minimize its risk. A peer to peer lending company connects borrowers and lenders, the latter being investors looking for certain returns and risk ratios based on their risk profile. Predictions of credit risk of individuals could also be used to create portfolios of loans in order to diversify their risk and to help investors reach their specific return over risk target.