Project: Creditworthiness

By: Fairoza Amira Binti Hamzah

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?

  Predict the list of creditworthiness of new loan applicants based on historical data of previous loan applicants' history, to approve the new applicants' loan.

- What data is needed to inform those decisions?

  1. Account-Balance
  2. Duration-of-Credit-Month
  3. Payment-Status-of-Previous-Credit
  4. Purpose
  5. Credit-Amount
  6. Value-Savings-Stocks
  7. Length-of-current-employment
  8. Instalment-per-cent
  9. Guarantors
  10. Duration-in-Current-address
  11. Most-valuable-available-asset
  12. Age-years
  13. Concurrent-Credits
  14. Type-of-apartment
  15. No-of-Credits-at-this-Bank
  16. Occupation
  17. No-of-dependents
  18. Telephone
  19. Foreign-Worker

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

  Binary – Creditworthy (approved) or non-creditworthy (rejected)

Step 2: Building the Training Set

Answer this question:

- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

  Checking the null values in the dataset

Credit-Application-Result	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
Duration-in-Current-address	344
Most-valuable-available-asset	0
Age-years	12
Concurrent-Credits	0
Type-of-apartment	0
No-of-Credits-at-this-Bank	0
Occupation	0
No-of-dependents	0
Telephone	0
Foreign-Worker	0

Impute the Age-years by using its median to 33 and *remove* the Duration-in-Current-address as it has 344 null values.

Checking the number of unique values for each columns

Credit-Application-Result	2
Account-Balance	2
Duration-of-Credit-Month	30
Payment-Status-of-Previous-Credit	3
Purpose	4
Credit-Amount	464
Value-Savings-Stocks	3
Length-of-current-employment	3
Instalment-per-cent	4
Guarantors	2
Duration-in-Current-address	4
Most-valuable-available-asset	4
Age-years	53
~~Concurrent-Credits~~	~~1~~
Type-of-apartment	3

No-of-Credits-at-this-Bank	2
~~Occupation~~	~~1~~
No-of-dependents	2
Telephone	2
Foreign-Worker	2

Removed the Concurrent-Credits and Occupation as it only has 1 value only.

Finding the correlation between all features to Credit-Application-Result.

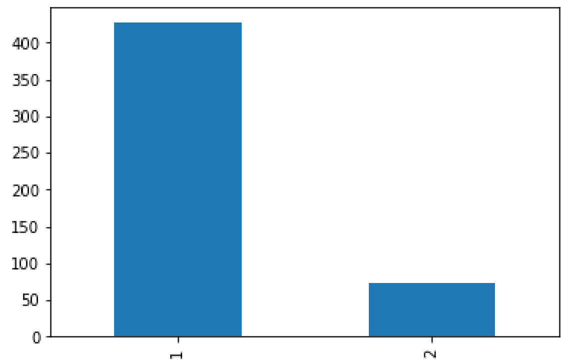
Credit-Application-Result	1.000000
Account-Balance	0.316080
Duration-of-Credit-Month	0.202504
Credit-Amount	0.2019461
Most-valuable-available-asset	0.141332
Value-Savings-Stocks	0.133424
Payment-Status-of-Previous-Credit	0.096541
Purpose	0.090912
Length-of-current-employment	0.089383
Duration-in-Current-address	0.082826
Instalment-per-cent	0.062107
No-of-Credits-at-this-Bank	0.056549
Age-years	0.052914
Guarantors	0.044105
No-of-dependents	0.041048
~~Telephone~~	~~0.028971~~
Type-of-apartment	0.026516
~~Foreign-Worker~~	~~0.009186~~

Investigate the skewness of all columns.

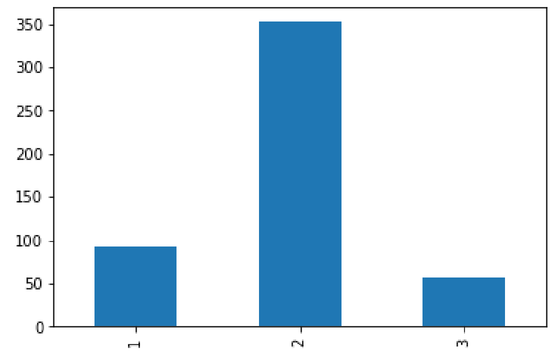
Account-Balance	0.096400
Duration-of-Credit-Month	0.991000
Payment-Status-of-Previous-Credit	-0.687677
Purpose	1.257190
Credit-Amount	2.108522
Value-Savings-Stocks	0.983026
Length-of-current-employment	0.637223
Instalment-per-cent	-0.596533
~~Guarantors~~	~~2.962197~~
Duration-in-Current-address	1.566395
Most-valuable-available-asset	0.013780
Age-years	1.102038
Concurrent-Credits	0.000000
Type-of-apartment	-0.056348
No-of-Credits-at-this-Bank	0.585090
Occupation	0.000000
No-of-dependents	2.011101
Telephone	0.409478
~~Foreign-Worker~~	~~4.847285~~

Investigate the amount of data in each feature to further understand which features need to be removed.

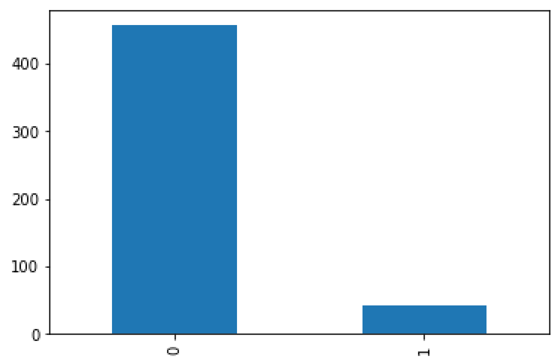
Number of dependents



Type of apartment



Guarantors



Thus, below features are removed:

- Duration-in-Current-address
- Concurrent-Credits
- Occupation
- Telephone
- Foreign-Worker
- Guarantors
- No-of-dependents

Step 3: Train your Classification Models

*Answer these questions for **each model** you created:*

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

```
OLS Regression Results
```

```
==============================================================================================
Dep. Variable:     Credit-Application-Result   R-squared (uncentered):                   0.759
Model:                                   OLS   Adj. R-squared (uncentered):              0.753
Method:                        Least Squares   F-statistic:                              128.1
Date:                       Mon, 21 Jun 2021   Prob (F-statistic):                   2.85e-142
Time:                               15:14:01   Log-Likelihood:                         -270.10
No. Observations:                        500   AIC:                                      564.2
Df Residuals:                            488   BIC:                                      614.8
Df Model:                                 12
Covariance Type:                   nonrobust
=====================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
Account-Balance                       0.2339      0.039      5.987      0.000       0.157       0.311
Duration-of-Credit-Month             -0.0005      0.002     -0.274      0.784      -0.004       0.003
Payment-Status-of-Previous-Credit     0.1608      0.030      5.404      0.000       0.102       0.219
Purpose                               0.1178      0.027      4.385      0.000       0.065       0.171
Credit-Amount                         -2e-05   9.31e-06     -2.149      0.032   -3.83e-05   -1.71e-06
Value-Savings-Stocks                  0.0998      0.029      3.488      0.001       0.044       0.156
Length-of-current-employment          0.0308      0.024      1.281      0.201      -0.016       0.078
Instalment-per-cent                  -0.0117      0.018     -0.666      0.506      -0.046       0.023
Most-valuable-available-asset        -0.0350      0.021     -1.695      0.091      -0.076       0.006
Age-years                             0.0041      0.002      2.347      0.019       0.001       0.008
Type-of-apartment                     0.0662      0.039      1.709      0.088      -0.010       0.142
No-of-Credits-at-this-Bank            0.1393      0.044      3.185      0.002       0.053       0.225
==============================================================================
Omnibus:                       39.040   Durbin-Watson:                   1.891
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               42.034
Skew:                          -0.673   Prob(JB):                     7.46e-10
Kurtosis:                       2.549   Cond. No.                     1.09e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 1.09e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The predictor variables that are significant with P value <0.05 are:

- Account-Balance
- Payment-Status-of-Previous-Credit
- Purpose
- No-of-Credits-at-this-Bank
- Value-Savings-Stocks
- Credit-Amount
- Length-of-current-employment
- Most-valuable-available-asset
- Age-years
- Type-of-apartment

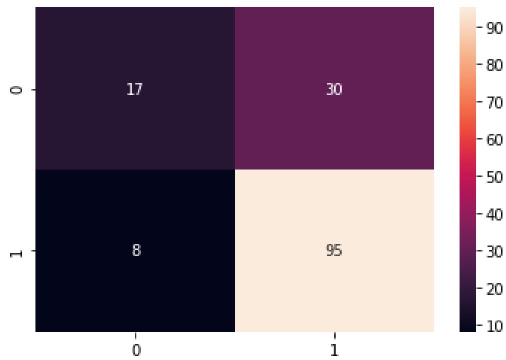
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

  By using the variables with P value < 0.05, the results are as below.
  The data is divided to 7:3 ratio for training and validation data.

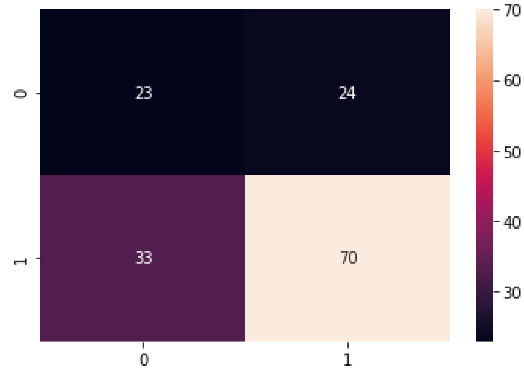
Method	Validation Accuracy
Logistic Regression	0.75
Decision Tree Classifier	0.62
Random Forest Classifier	0.76
AdaBoost Classifier	0.75
XGBoost	0.72

Below are the confusion matrix for each method:
(0: Non-creditworthy, 1: creditworthy)

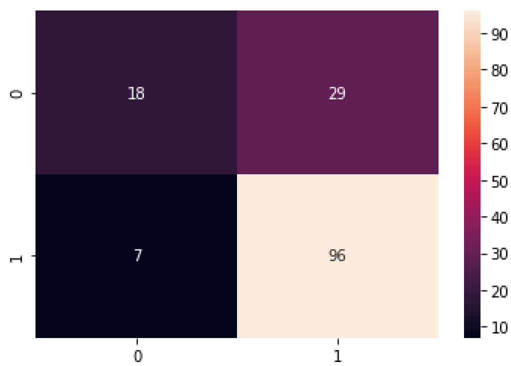
Logistic Regression



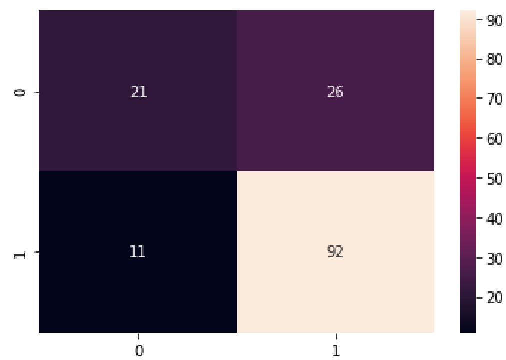
Decision Tree Classifier



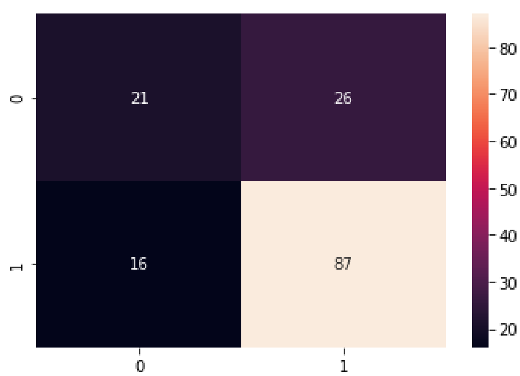
Random Forest Classifier



AdaBoost Classifier



XGBoost



Yes, there are bias in the model as some of the data is not predicted correctly.

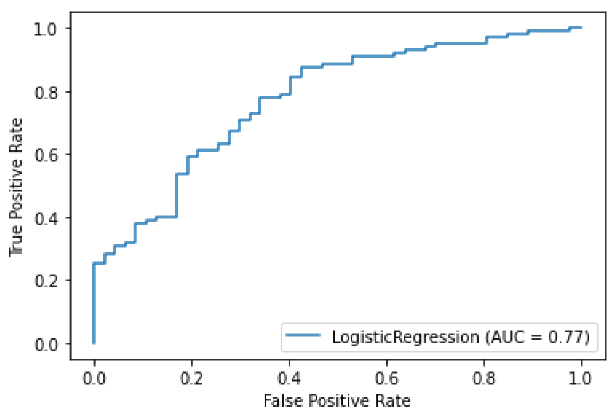
Step 4: Writeup

Answer these questions:

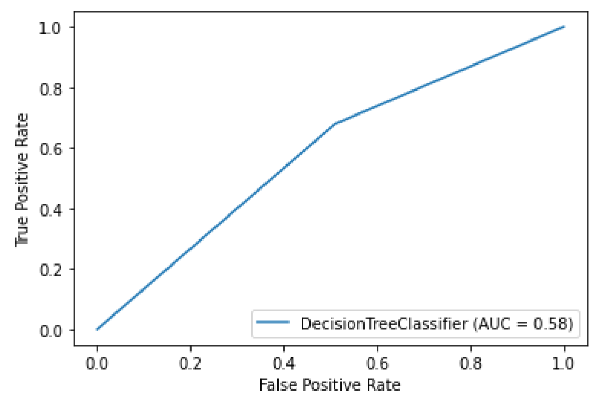
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    - ROC graph
    - Bias in the Confusion Matrices

The ROC graph for each model are as below.

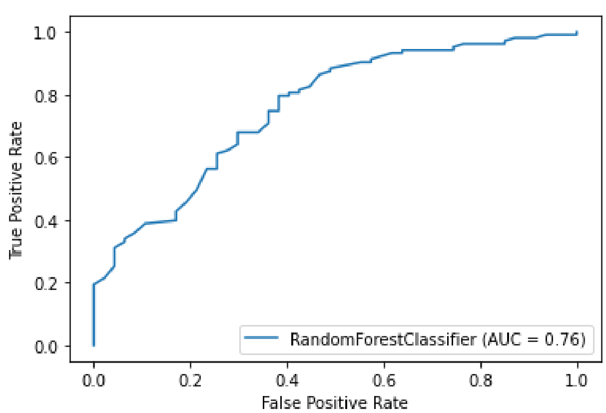
Logistic Regression



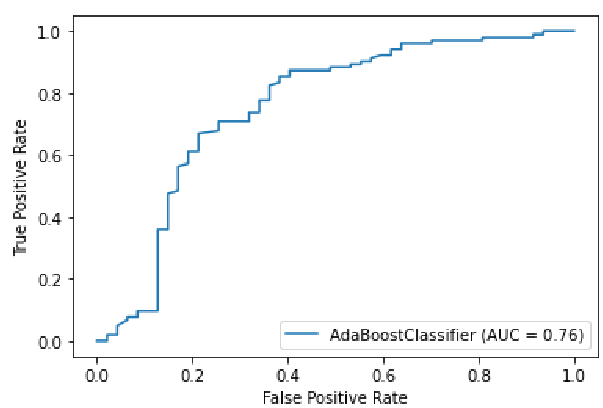
Decision Tree Classifier



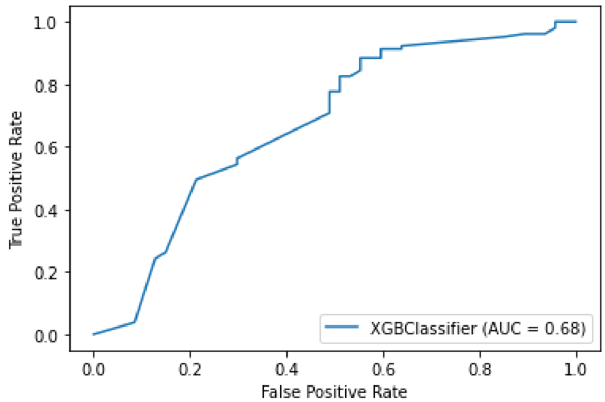
Random Forest Classifier



AdaBoost Classifier



XGBoost

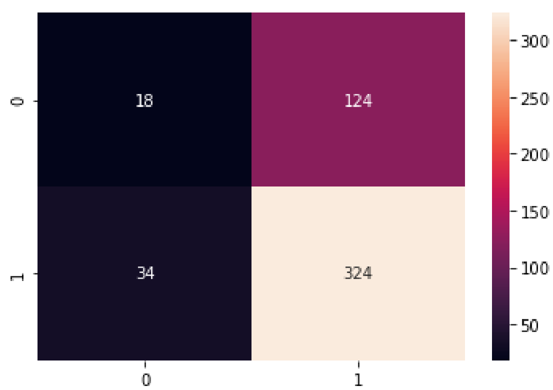


The validation and test accuracy table for each model is as below.

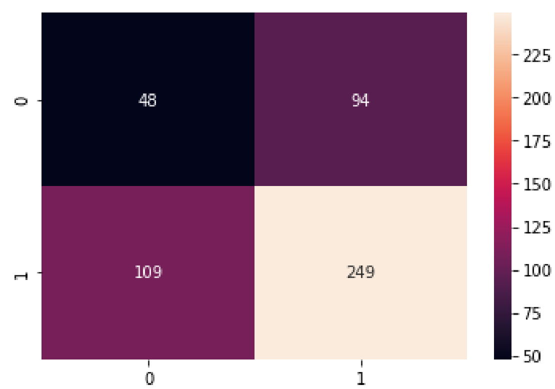
Method	Validation Accuracy	Testing Accuracy
Logistic Regression	0.75	0.68
Decision Tree Classifier	0.70	0.59
Random Forest Classifier	0.74	0.66
AdaBoost Classifier	0.64	0.62
XGBoost	0.75	0.63

Confusion matrices for each model is as below.

Logistic Regression

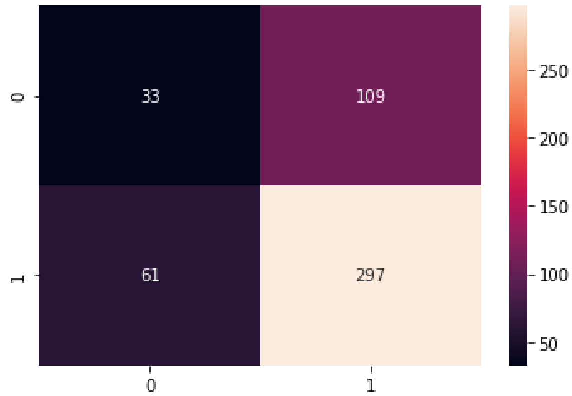


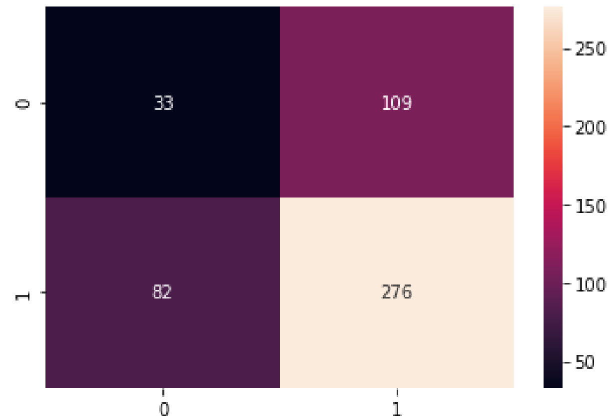
Decision Tree Classifier



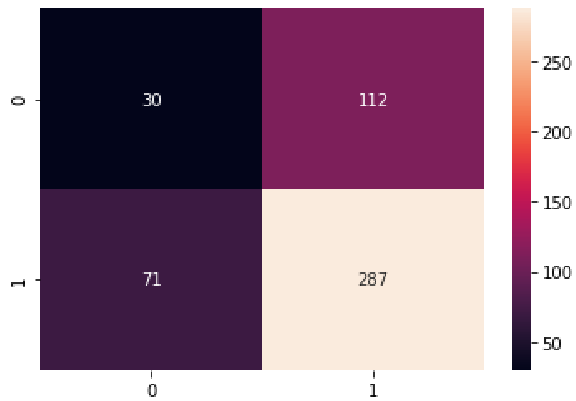
Random Forest Classifier

AdaBoost Classifier





XGBoost



From the results above, using Logistic Regression model can achieve higher accuracy and ROC.

- How many individuals are creditworthy?
  448 people predicted to be creditworthy