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 Duration-in-Current-address and Age-years by using their median, 2 and 3 3, respectively.

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:                          y   R-squared (uncentered):                   0.716
Model:                            OLS   Adj. R-squared (uncentered):              0.713
Method:                 Least Squares   F-statistic:                              207.9
Date:                Fri, 18 Jun 2021   Prob (F-statistic):                    1.10e-131
Time:                        12:20:41   Log-Likelihood:                         -310.95
```

```
No. Observations:                 500   AIC:                                    633.9
Df Residuals:                     494   BIC:                                    659.2
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.3365      0.040      8.370      0.000       0.258       0.416
x2             0.0022      0.002      1.125      0.261      -0.002       0.006
x3             0.2325      0.024      9.638      0.000       0.185       0.280
x4         -1.322e-05   8.97e-06     -1.473      0.141   -3.09e-05    4.41e-06
x5             0.1043      0.031      3.403      0.001       0.044       0.164
x6             0.0502      0.018      2.792      0.005       0.015       0.085
==============================================================================
Omnibus:                       36.072   Durbin-Watson:                   1.752
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.303
Skew:                          -0.612   Prob(JB):                     1.31e-08
Kurtosis:                       2.503   Cond. No.                     8.72e+03
==============================================================================

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, 8.72e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The features are selected by finding the most correlated features to the target feature, by using correlation function and Extra Trees Classifier.
The selected features are as below.

```
'Account-Balance',
'Duration-of-Credit-Month',
'Credit-Amount',
'Most-valuable-available-asset',
'Value-Savings-Stocks',
'Payment-Status-of-Previous-Credit',
'Credit-Application-Result
```

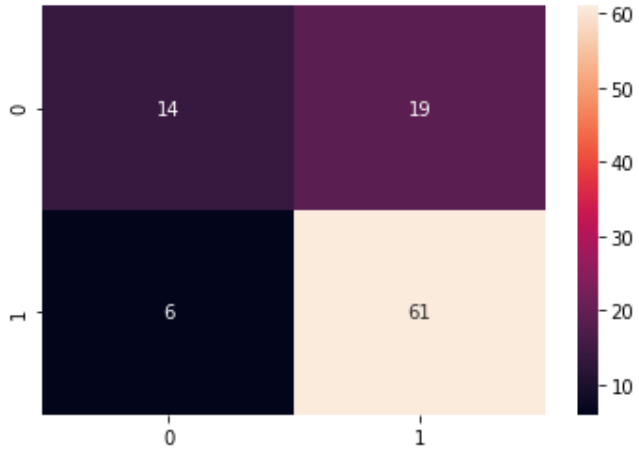
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

  The data is divided to 4:1 ratio for training and validation data.

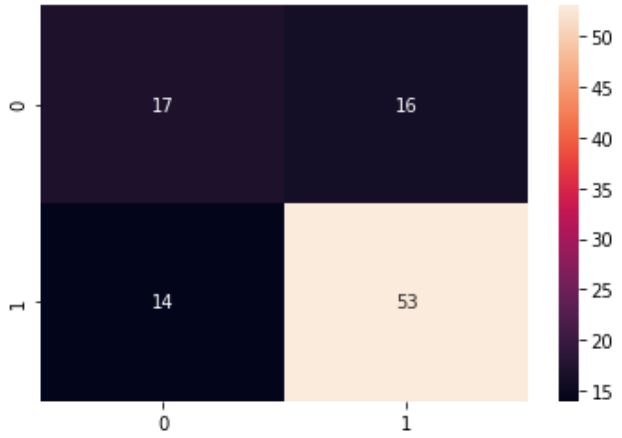
Method	Validation Accuracy
Logistic Regression	0.75
Decision Tree Classifier	0.70
Random Forest Classifier	0.74
AdaBoost Classifier	0.64
XGBoost	0.75

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

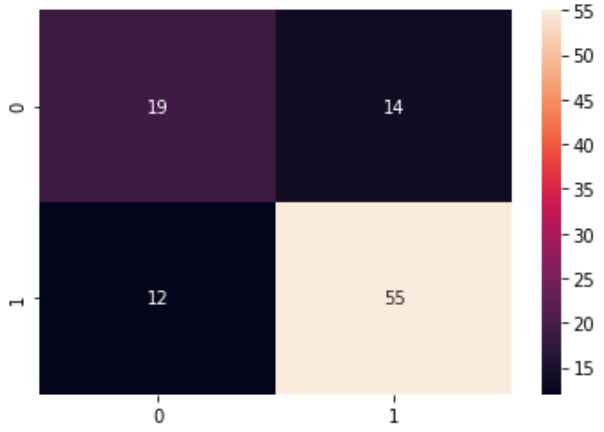
1. Logistic Regression



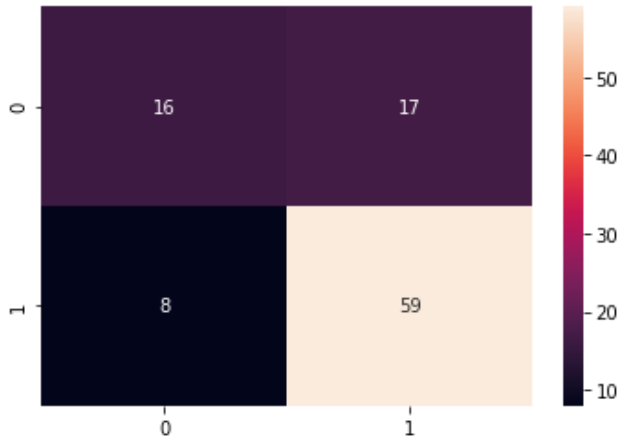
2. Decision Tree Classifier



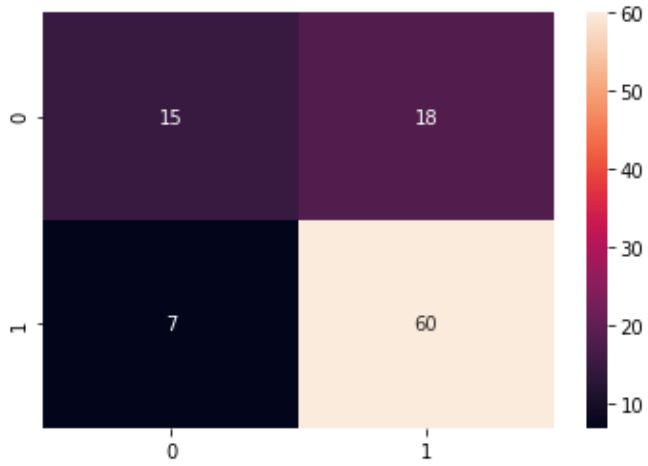
3. Random Forest Classifier



4. AdaBoost Classifier



5. XGBoost



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

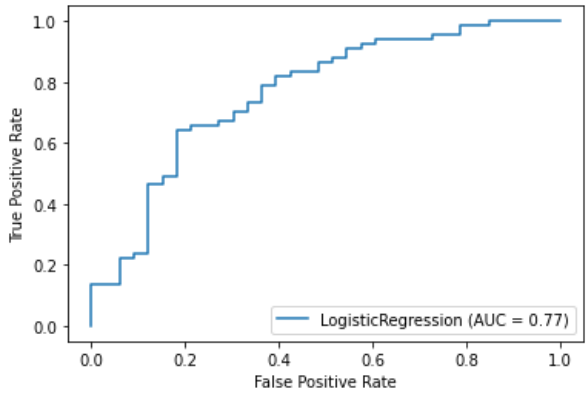
Step 4: Writeup

Answer these questions:

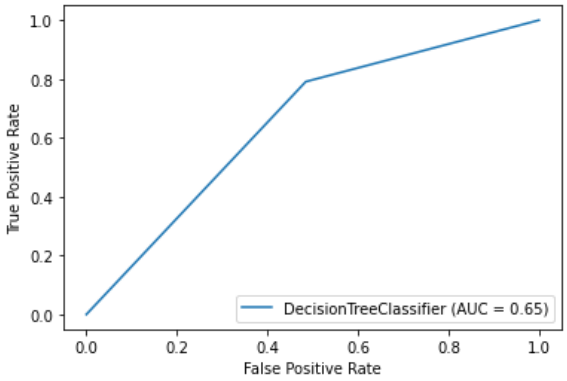
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    - ROC graph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

The ROC graph for each model are as below.

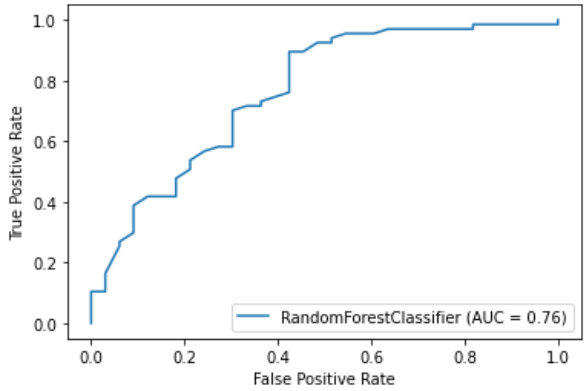
Logistic Regression



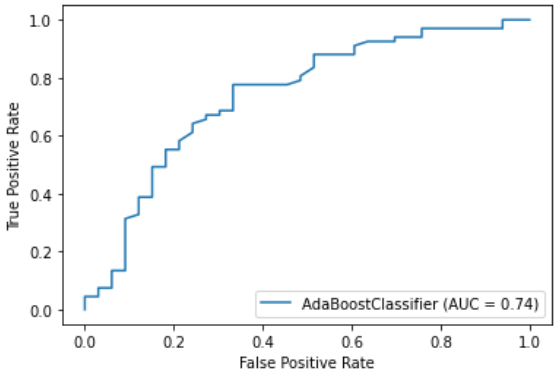
Decision Tree Classifier



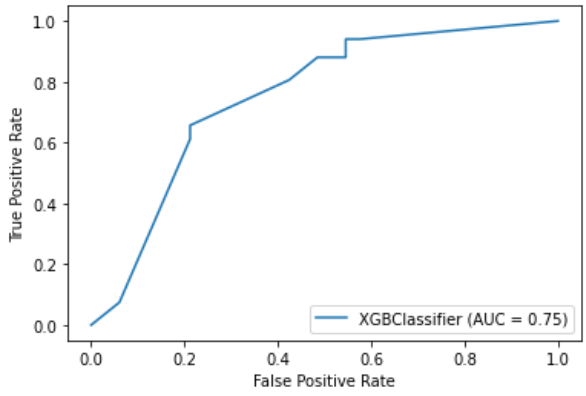
Random Forest Classifier



AdaBoost Classifier



XGBoost



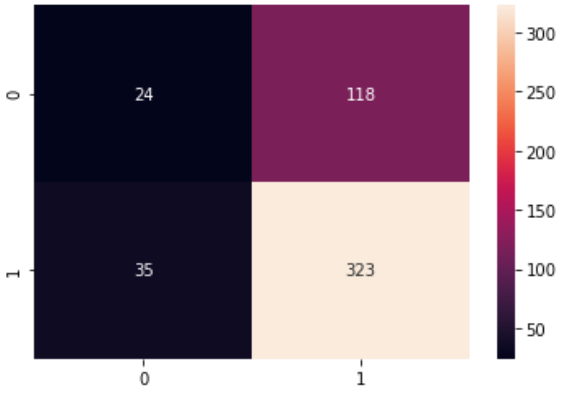
The validation and test accuracy table for each model is as below.

Method	Validation Accuracy	Testing Accuracy
Logistic Regression	0.75	0.69
Decision Tree Classifier	0.70	0.62

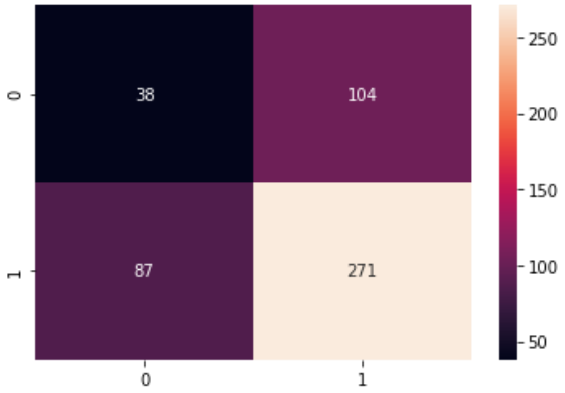
Random Forest Classifier	0.74	0.66
AdaBoost Classifier	0.64	0.64
XGBoost	0.75	0.68

Confusion matrices for each model is as below.

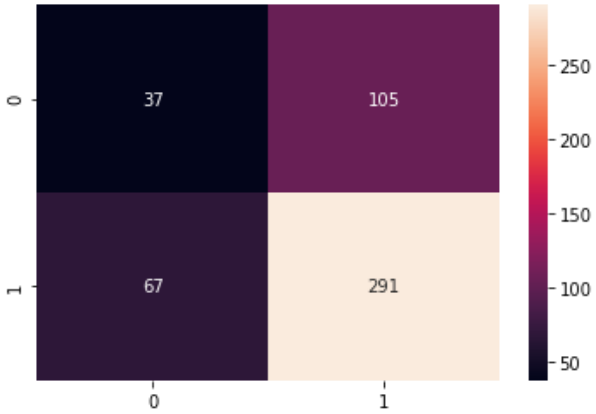
Logistic Regression



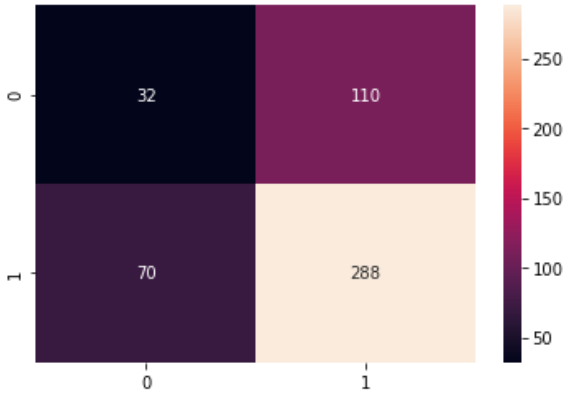
Decision Tree Classifier



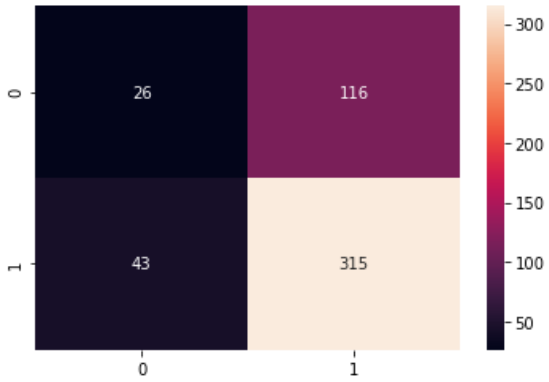
Random Forest Classifier



AdaBoost Classifier



XGBoost



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

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