

Predicting Default Risk(Creditworthiness)

July 30, 2021

0.0.1 Business and Data Understanding

1. What decisions need to be made?

Based on the prediction made by data analyst, the manager need to be made that the new loan applicants are creditworthy or non-creditworthy and approve them if the applicants are creditworthy.

2. What data is needed to inform those decisions?

I have selected 12 (6 categorical and 6 numerical) variables as predictor variables and **Credit-Application-Result** as a target variable to inform those decisions. I have listed the variables that are useful for our prediction below.

1. Account-Balance
2. Duration-of-Credit-Month
3. Payment-Status-of-Previous-Credit
4. Credit-Amount5-Purpose
5. Purpose
6. Value-Savings-Stocks
7. Length-of-current-employment
8. Instalment-per-cent
9. Most-valuable-available-asset
10. Age-years
11. Type-of-apartment
12. No-of-Credits-at-this-Bank

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

The model is a binary model in which there are two classes that going to be classified as creditworthy(approved) or non-creditworthy(not approved).

1 Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't need to convert any data fields to the appropriate data types.

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.

2 Answer

I have done a lot of processes to cleanup and format the dataset.

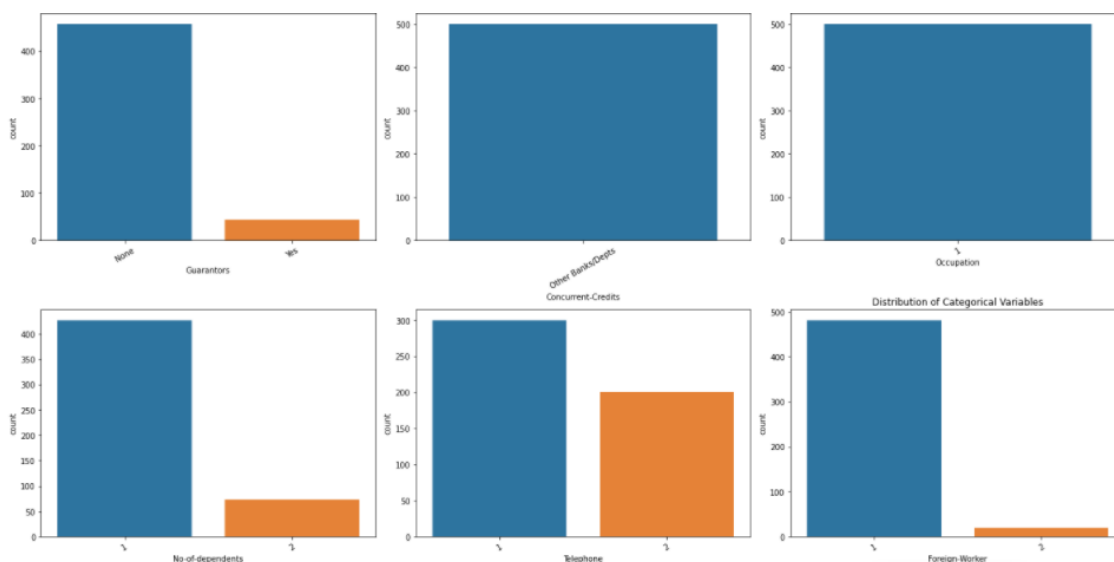
Step 1. Dealing with missing data:

As we can see in the following table there are two variables that has missing values. The **Duration-in-Current-address** has missing 344(68.8%) and the **Age-years** has missing 12(2.4%) out of 500 records. The **Duration-in-Current-address** variable has missing almost 69 percent of our data and probably our model will not have a significant changes if we drop this variable. Therefore, I have removed the **Duration-in-Current-address** variable because it is no longer needed for our prediction. But we impute the **Age-years** variable by comparing mean and median values of this variable.

	Total	%
Duration-in-Current-address	344	68.8
Age-years	12	2.4

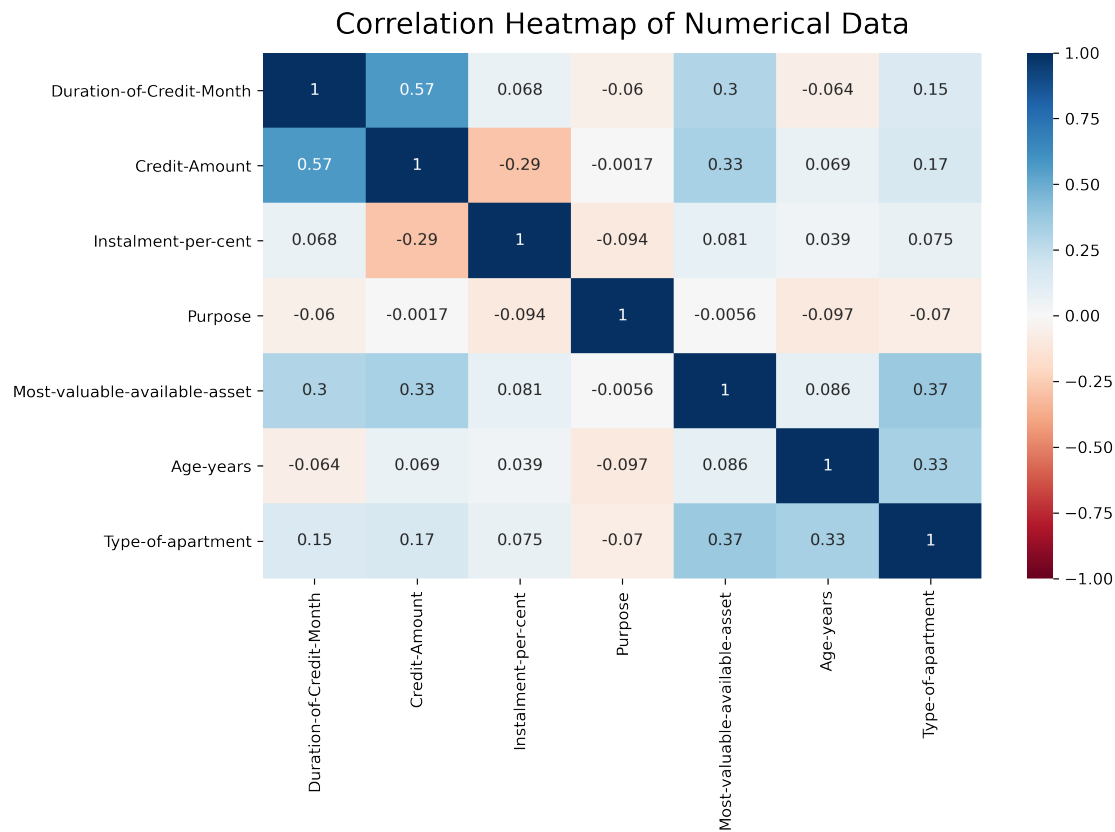
The mean and meadian values of the **Age-years** variable are **35.64** and **33** respectively. Therefore, I have imputed the **Age-years** variable with the meadian value because the mean value will have an outlier.

Step 2: Identify low variability: The variables distributions are shown below:



From the above distribution plot, we can see that the **Guarantors**, **No-of-dependents**, and **Foreign-Worker** variables are skewed the majority of the data points to one category and **Concurrent-Credits** and **Occupation** variables have one data point. These variables are low variability and no longer needed for our prediction. **Telephone** also not important because it has two unique values. Therefore, I have decided to remove these variables from the dataset.

Step 3: Plotting the correlation of numerical variables



Therefore, the above correlation graph shows that there are no variables having strongly correlated (high correlation).

After cleaning the dataset, I have found 13 columns (12 predictive variables and a target variable).

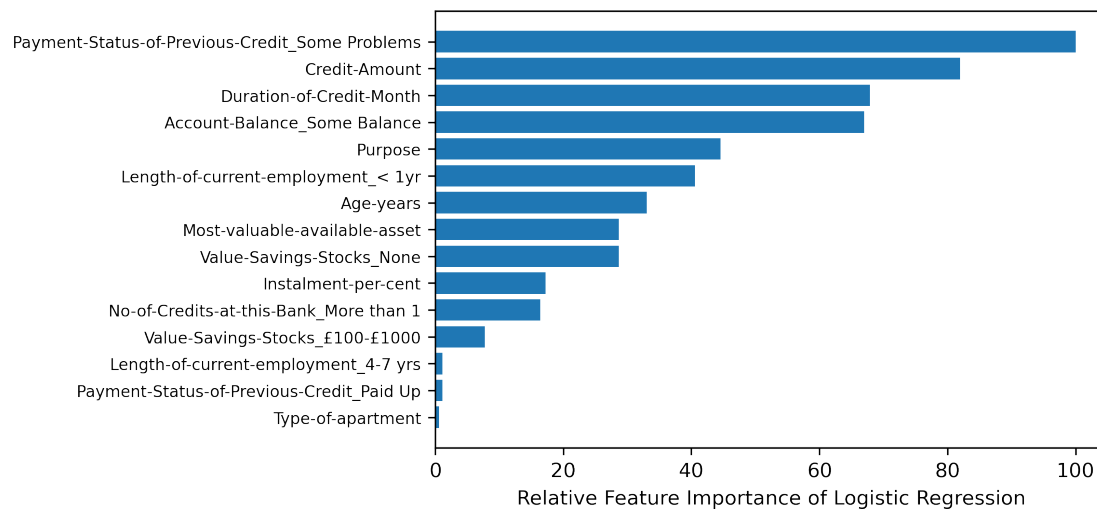
3 Step 3: Train your classification models

1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
2. 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?

I have splitted the dataset into training (70%) and testing (30%) set with random state of 1. Then, I have trained and validated the test dataset using the following classification models:

1. Logistic Regression
2. RandomForestClassifier
3. DecisionTreeClassifier
4. GradientBoosting

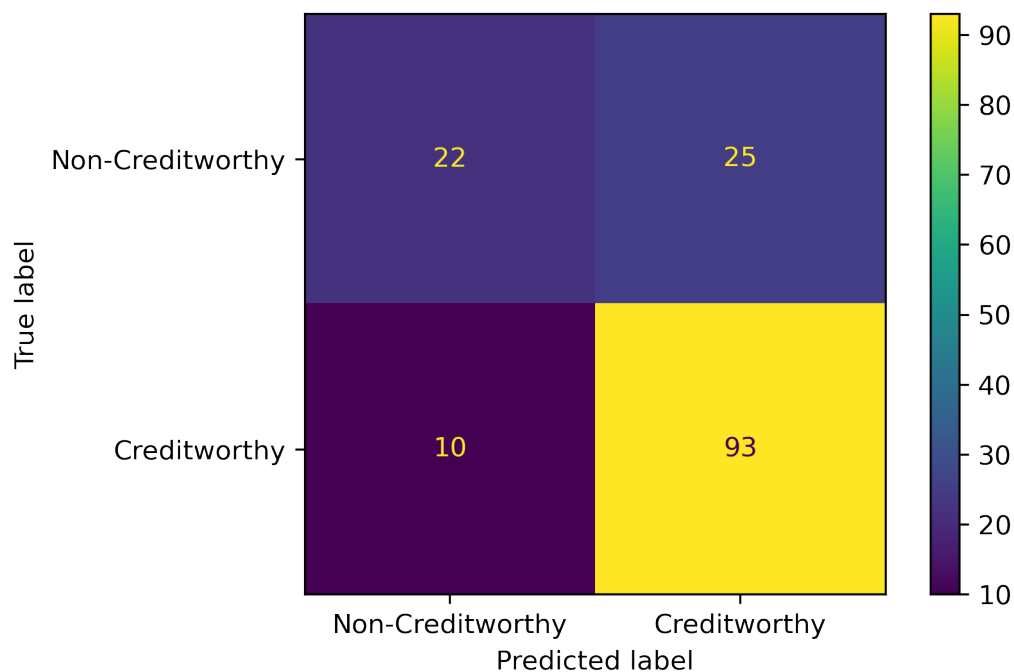
4 1. Using Logistic Regression



As per the above feature importance chart, the following predictor variables are most significant where they have higher values (which mean that the $p\text{-value} < 0.05$) - Payment-Status-of-Previous Credit_Some Problems - Credit-Amount - Duration-of-Credit-Month - Account-Balance_Some Balance - Purpose - Length-of-current-employment_<1yr

4.0.1 Confusion matrix of Logistic Regression

A confusion matrix is formed from the four outcomes produced as a result of binary classification.
 - True Positive(TP) - True Negative(TN) - False Positive(FP) - False Negative(FN)



From the confusion matrix, I have calculated the accuracy of creditworthy and Non-Creditworthy.

$$Creditworthy = TP / (TP + FP) = 93 / (93 + 10) * 100 = 90$$

$$Non - Creditworthy = TN / (TN + FN) = 22 / (22 + 23) * 100 = 47$$

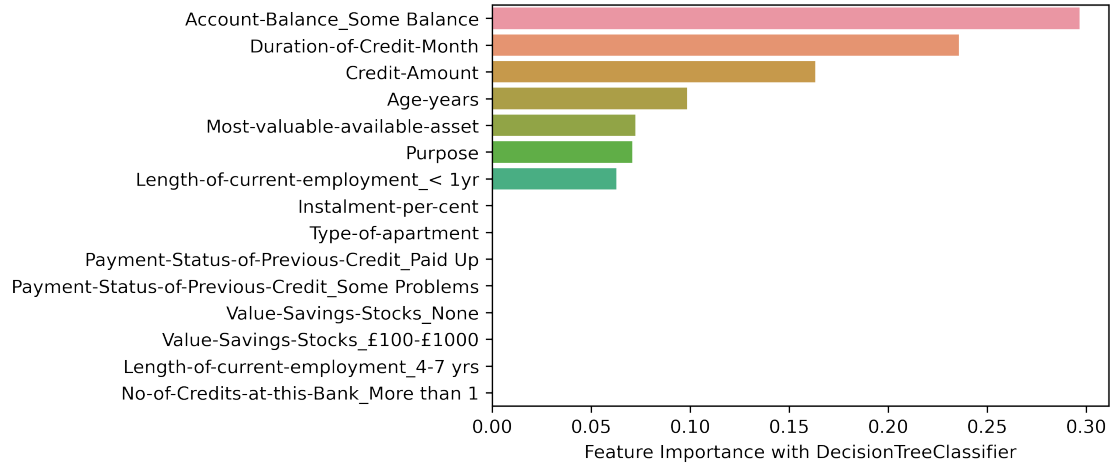
$$Overall - accuracy = (TN + TP) / (TN + TP + FP + FN) * 100 = 77$$

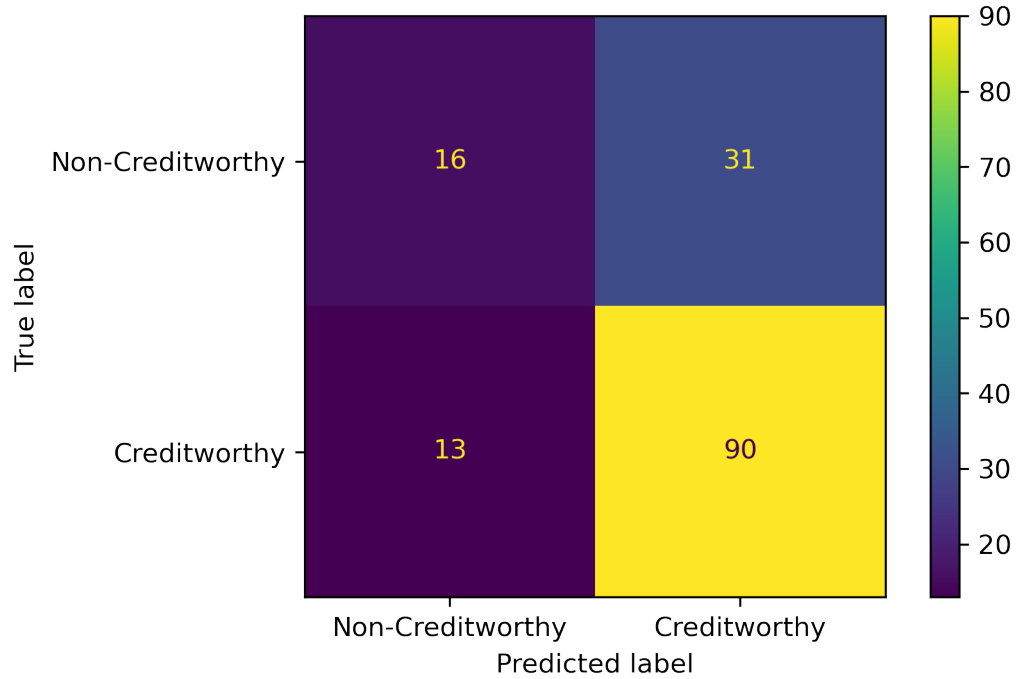
The overall accuracy of the logistic regression model against the validation set is 77%.

According to the confusion matrix, the model is slightly biased towards predicting Non-Creditworthy which mean that $100 - 47 = 53\%$ are incorrectly predicted.

4.1 2. Using Decision Tree Classifier

Account-Balance_Some Balance, Duration-of-Credit-Month, and Credit-Amount are the most significant variables that are used to classify either Creditworthy or Non-Creditworthy using decision tree classifier model.





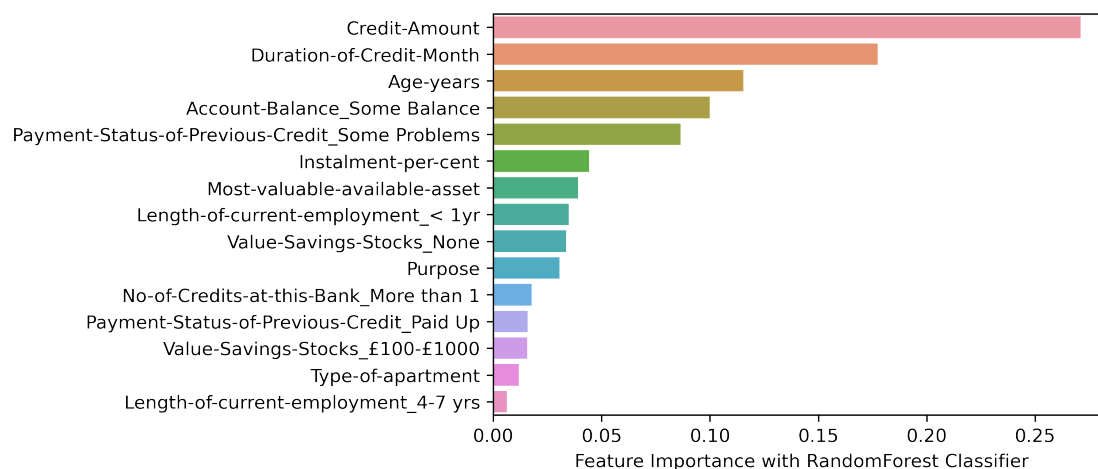
From the above confusion matrix of Decision Tree model, the accuracy of each classes are:

- Creditworthy = $TP / (TP + FP) = 90 / (90 + 13) * 100 = 84\%$
- Non-Creditworthy = $TN / (TN + FN) = 16 / (16 + 31) * 100 = 34\%$
- Overall-accuracy = $(TP + TN) / (TP + TN + FN + FP) = (90 + 16) / (90 + 16 + 31 + 10) = 71\%$

Therefore, the model is biased towards predicting Creditworthy and the model is incorrectly predicting 66% of Non-Creditworthy. In the validation data set, the model achieved accuracy of 71%.

4.1.1 3. Using RandomForest Classifier

In random forest model, there are 500 trees with 8 variables and each variables are tested at each split. The most significant features were identified and plotted in the following bar plot.

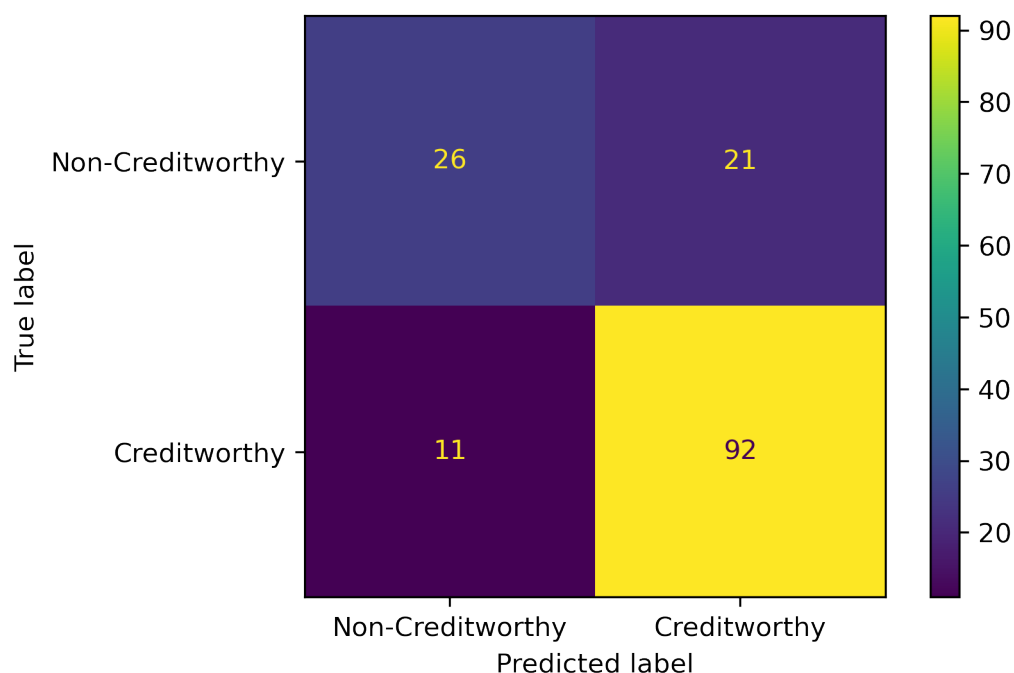


4.1.2 Confusion Matrix of Random Forest Classifier

I have used some tuning parameters to improve the performance of the model and the model is not biased towards predicting Non-creditworthy. - $\text{Creditworthy} = \frac{TP}{(TP + FP)} = \frac{92}{(92 + 11)} * 100 = 84\%$

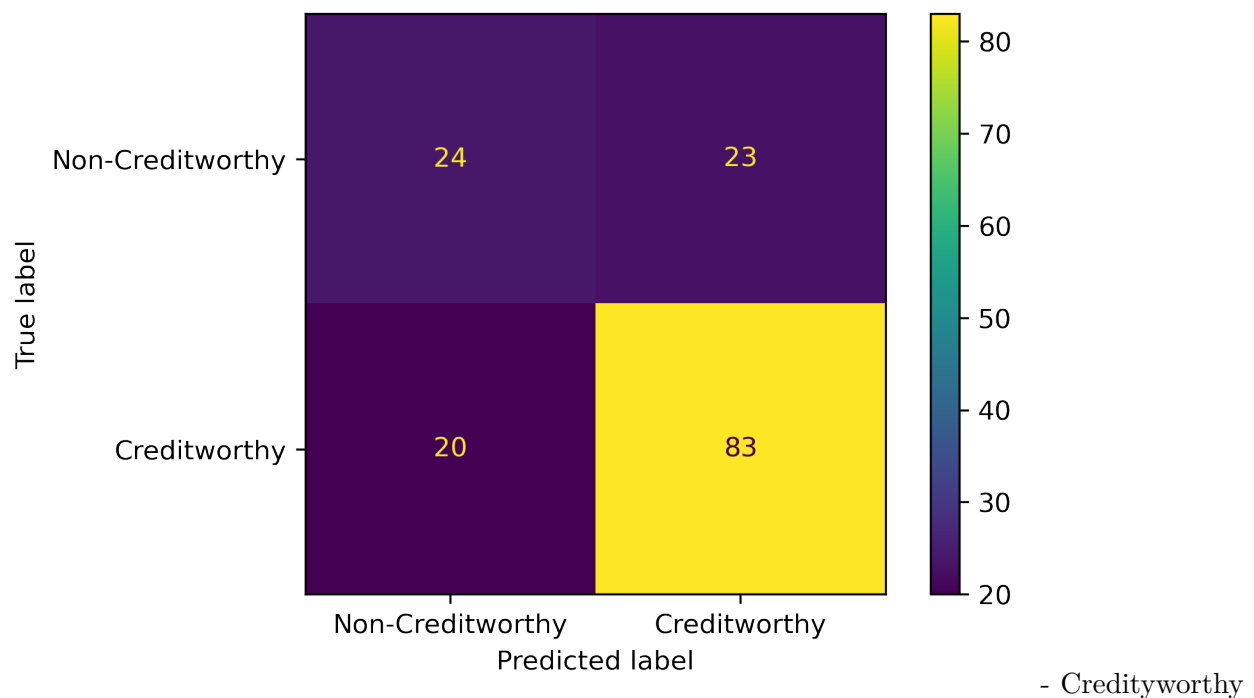
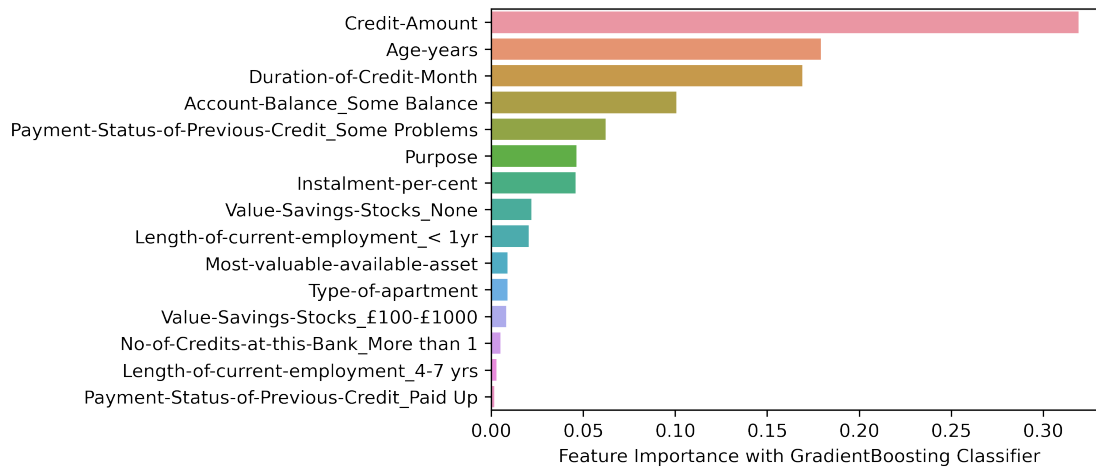
- $\text{Non-Creditworthy} = \frac{TN}{(TN + FN)} = \frac{26}{(26 + 21)} * 100 = 55\%$
- $\text{Overall-accuracy} = \frac{(TP + TN)}{(TP + TN + FN + FP)} = \frac{(92 + 26)}{(92 + 26 + 21 + 11)} = 79\%$

Therefore, from the confusion matrix the model predicting 84% as **Creditworthy** while 55% is predicting as **Non-Creditworthy**. The overall accuracy of the model is 79%.



4.2 4. Using Boosted Model(GradientBoosting Classifier)

From this model, I have got the following most significant variables as best predictors.



$$= \text{TP} / (\text{TP} + \text{FP}) = 83 / (83 + 20) * 100 = 81\%$$

- Non-Creditworthy = $\text{TN} / (\text{TN} + \text{FN}) = 24 / (24 + 23) * 100 = 51\%$
- Overall-accuracy = $(\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP}) = (83 + 24) / (83 + 24 + 23 + 20) = 71\%$

Accordingly the confusion matrix and the classification report, the model is slightly biased towards predicting Creditworthy and 49% is incorrectly predicted. The accuracy of Creditworthy is 81% while Non-Creditworthy is 51% and the percentage accuracy is 71.

4.3 Step 4. Writeup

Compare all of the models' performance against each other. Decide on the best model and score your new customers.

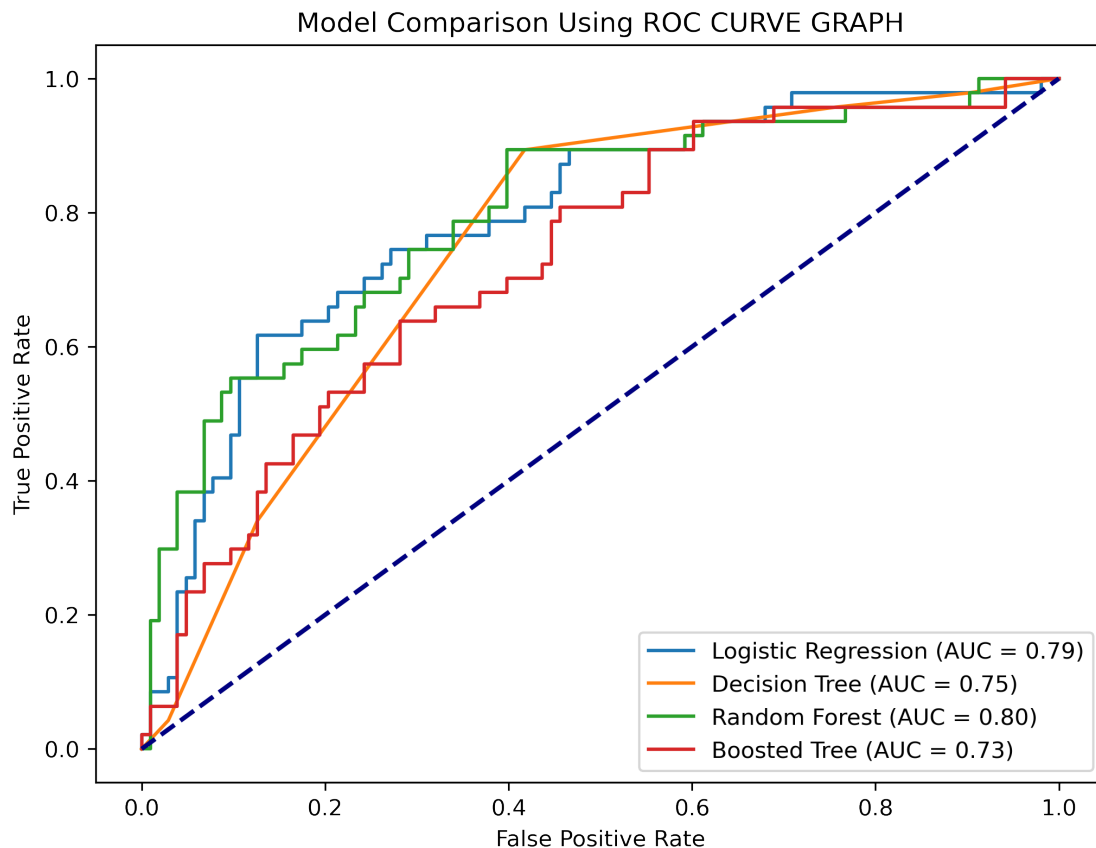
Important: Your manager only cares about how accurate you can identify people who qualify and do not qualify for loans for this problem.

Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan.

To come up with the best classification model, I have gone through the following steps:

1. cleansing, formating, and blending the dataset togther
2. Dealing with missing data
3. Identifying variables with the low variability and drop them
4. Identifying the best predictors
5. Splitting the data into training and validation set based on the criteria given(70% training, 30% test/validation dataset)
6. train the classification models with tunning parameters to improve the model performance
7. Evaluate the model performance using Precision, recall, and f1-score

4.3.1 Model Comparison



Model Comparison Table

	Overall_Accuracy	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic Regression	0.77	0.9	0.47
Decision Tree	0.71	0.87	0.34
Random Forest	0.79	0.89	0.55
GradientBoosting	0.71	0.81	0.51

The higher the **AUC**, the better the performance of the model at distinguishing between the positive and negative classes. So, the higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes. The higher the overall accuracy, the better the performance of the model.

Therefore, from the above **ROC CURVE GRAPH and Model Comparison Table**, the best performed model is **Random Forest Model**. The random forest model achieved 80% of AUC and 79% of Overall Accuracy as compared to other three models.

Then, based on the best performed model(using Random Forest Classifier) that 418 of individual loans classified as **Creditworthy** and 82 of individual loans as **Non-Creditworthy** from the new customer score data.

Note that I have used different parameters to improve the model performance and evaluated the model using Precision, Recall, and F1-score. I would like to suggest that we can have a chance to improve the model performance and get the best accurate model using different tuning parameters and cross validation mechanism.