Project: Creditworthiness

# Step 1: Business and Data Understanding

## Key Decisions:

* What decisions needs to be made?

The decision is whether to grant loans to new loan applicants, in doing this, we need to provide a list of creditworthy customers.

* What data is needed to inform those decisions?

Credit approvals on past applicants that the bank has completed.

New set of customers that you need to score on the classification model you will create.

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

We need a Binary classification model.

# Step 2: Building the Training Set

* For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered “high”.

With Pearson correlation analysis, there wasn’t any field that had correlation above 0.7.

* Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed.

Yes. Duration-in-Current-address should be removed as there are too many null values.

* Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability.

Cocurrent-Credits: removed due to uniform distribution.

* Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)
* In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

**I removed the following columns:**

Cocurrent-Credits: removed due to uniform distribution.

Duration-in-Current-Address: there are too many null values.

Occupation: removed as its uniform distribution

Icon

Description automatically generated

Guarantors - has 2 unique values but data is almost uniformly distributed to 1 of the values i.e it has low variability.

The following were removed because they didn’t have significant correlation with the target variable.

* No.of.dependents
* Telephone
* Foreign worker

Although Age-years doesn’t have a significant linear relationship, it has an important relationship with target variable in other aspects. Null values in age-years are replaced with median.

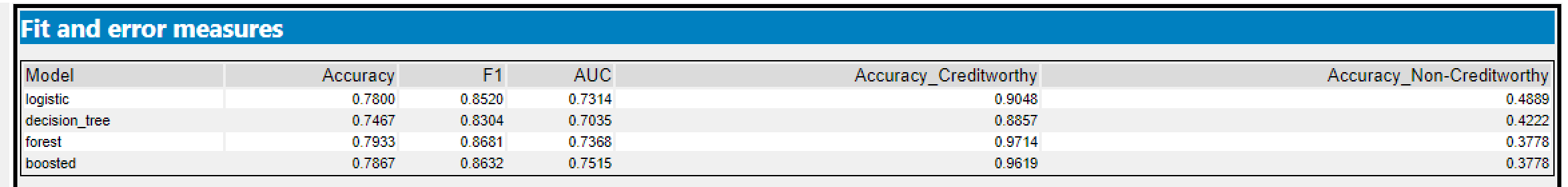
# Step 3: Train your Classification Models

*First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model*

*Effect of Bias in prediction*

Bias as a tendency of a model to predict one of its outcomes much more accurately than the others

*Accuracy for all models*

**

*Confusion matrix for all models*

*Graphical user interface, application

Description automatically generated*

*Logistic Regression*

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

Graphical user interface, table

Description automatically generated

The most significant variables are most valuable available asset, instalment per cent, length of current employment < 1 yr, payment status of previous credit some problems

* 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?

As shown above, the overall accuracy is 0.78. For the confusion matrix, there is slight bias towards predicting creditworthy. The calculation for bias is not required as per above question.

*Decision Tree*

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

Graphical user interface, application

Description automatically generated

The most important variables are account balance, value saving stocks, duration of credit month

* 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?

As shown above, the overall accuracy is 0.75. For the confusion matrix, there is slight bias towards predicting creditworthy. The calculation for bias is not required as per above question.

*Forest Model*

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

Chart

Description automatically generated

The most important variables are credit amount, age.years, duration.of.creditt.month

* 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?

As shown above, the overall accuracy is 0.79. For the confusion matrix, there is almost no bias. The calculation for bias is not required as per above question.

*Boosted Model*

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

Chart

Description automatically generated

The most important variables are account balance, and credit amount.

* 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?

As shown above, the overall accuracy is 0.7867. For the confusion matrix, there is almost no bias. The calculation for bias is not required as per above question.

# Step 4: Writeup

*Decide on the best model and score your new customers. For reviewing consistency, if Score\_Creditworthy is greater than Score\_NonCreditworthy, the person should be labeled as “Creditworthy”. 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.*

**Note:** Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

Which model did you choose to use? Please justify your decision using **all** of the following techniques.

* + Overall Accuracy against your Validation set
  + Accuracies within “Creditworthy” and “Non-Creditworthy” segments
  + ROC graph
  + Bias in the Confusion Matrices

I chose the random forest model because it had the highest overall accuracy 0.7933. It also has decent accuracies for both creditworthy and non creditworthy, 0.9714, 0.3778, with highest accuracy for creditworthy out of all models.

Chart, line chart, histogram

Description automatically generated

The random forest model also has a relatively decent ROC curve shown above, along with the boosted model.

Graphical user interface, text, application

Description automatically generated

Looking at the confusion matrix, precision 102/(102+28) =0.78 and recall of 102/(102+3)= 0.97

are calculated. The recall is quite high which means that those applicants who are creditworthy are most likely to be predicted creditworthy and get a loan.

For the confusion matrix, there is no bias for the random forest model.

410 out of 500 potential applicants are credit worthy.

Calculated by filtering predicted credit worthy probability >= 0.5.