

Creditworthiness

REVIEW

HISTORY

Meets Specifications

Dear Excellent Student,
Thank you for your submission. I enjoyed reviewing your work because it was great. This submission meets all of our expectations. The Classification Model is not an easy course to grasp, however, the task was really understood. A lot of work has been done and you should be proud of yourself. Continue practicing on these projects and other projects of yours and you will become the best in your domain.

Keep up the good work and good luck in future projects!!

Business and Data Understanding

✓

The section is written clearly and is concise. The section is written in less than 250 words.

Great Job Done!!

The section is written clearly and is concise, the word limit is very well respected.

✓

All following questions have been answered:

1. What decisions need to be made?

2. What data is needed to inform those decisions?

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

Excellent work done!! answering Question 1, where the main decision was to find out how many loan applicants are creditworthy to approve the loan out of 500 applicants and will make use of the Binary model (i.e. the predicted outcome is whether the customer is creditworthy or not creditworthy) to answer Question 3.

Great Job! You have very well mentioned the data of all past applicants which we have used the dataset to create and train the model. On the other hand, we have a list of customers who have applied to get a loan. This dataset has been scored with the model to get the list and number of final customers that are creditworthy to get a loan with the list of variables like the Purpose of the loan, Age_years of the loan applicant, credit balance etc. to answer Question 2.

1. What decisions needs to be made?

The Decision that needs to be taken here is to classify the customer either as credit worthy or non-credit worthy and based on that we need classify the customer into two categories.

2. What data is needed to inform those decisions?

We need to know information that are:

• 'credit-data-training.xlsx': This file contains the data of the customers to whom bank has provided the loan to and based on that data we can make predictive model to analyse the 'credit-data-training.xlsx' data set and can categorize the customer into credit worthy or non-credit worthy.

• The variables which will be useful in deciding the credit worthiness of the customer will be: Account Balance, Credit Amount, Payment Status of Previous Credit, Purpose New Car, Value Saving Stocks, Age-Year, Duration of Credit Month.

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

The model that we need to build for this problem is Binary model as we need to decide whether the customer is credit worthy or non-credit worthy.

Building the Training Set

✓

The section is written clearly and is concise. The section is written in less than 100 words.

Awesome!!

The section is written clearly and is concise, the word limit is very well respected.

✓

The following question has been answered:

1.In your cleanup process, which field(s) did you impute or remove?

Please justify why you imputed or removed these fields. Visualizations are encouraged.

The correct fields are removed or imputed.

Awesome!!

All the variables which need to be removed and imputed are correctly identified with a reasonable justification and the correct visualization.

I like the way you have justified each of the variable which needs to be removed and imputed.

Tips: If you would like to better understand - in a very intuitive way - why the median is a good value to impute the missing values in Age field, check this site out: [measures-central-tendency](#)

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

There are several fields which are removed and one field which is imputed and they are:

- The Histogram of the variable **Guarantors, Foreign-Worker** and **No of dependents** have shown that majority of data is heavily skewed towards one type of data. These variables are removed due to heavily skewed towards one data.
- **Concurrent Credits** and **Occupation** both of them have entirely uniform data and there are no other variations in the data that's why both of them are removed due to low variability.
- **Duration in Current address** has 69% of the missing data. That's why this field is removed.
- **Telephone**, this field does not have any predictive ability to credit application results, so this field is also removed.
- **Age-Year** this field has 2% of missing value. The missing data of this variable has been imputed **using** the median, 33 of the entire data field.

Train your Classification Models

✓

The section is written clearly and is concise. The section is written in less than 500 words.

Great Job done!!

The section is written clearly and is concise, the word limit is very well maintained.

✓

All questions have been answered for each of the four models built: Logistic, Decision Tree, Forest Model, Boosted Model

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.

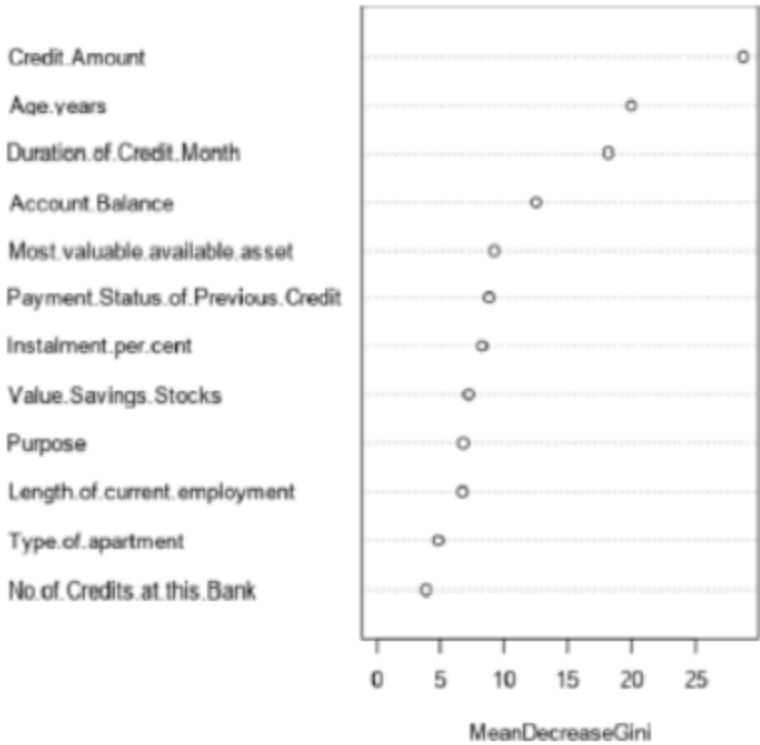
2. Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

There should be 4 sets of questions answered.

Awesome work done!!

- All the significant predictor variables are provided with the variable importance charts for all the models.
- The values in the confusion matrix and the overall percent accuracies are within range.
- The bias shown in each model is calculated and is very well justified.

Variable Importance Plot



| Variable | MeanDecreaseGini (approx.) |
|-----------------------------------|----------------------------|
| Credit Amount | 26 |
| Age.years | 20 |
| Duration of Credit Month | 18 |
| Account Balance | 10 |
| Most valuable available asset | 8 |
| Payment Status of Previous Credit | 7 |
| Instalment.per.cent | 6 |
| Value.Savings.Stocks | 5 |
| Purpose | 4 |
| Length of current employment | 4 |
| Type.of.apartment | 3 |
| No of Credits at this Bank | 2 |

Below is the model comparison report for Forest Model which shows that this model has an accuracy of 79.33%. Using Confusion Matrix,

Accuracy for creditworthy = (actual creditworthy) / (predicted creditworthy)
= 102/ (102+28) = 0.7846 = 78.46%

Accuracy for non-creditworthy = (actual non-creditworthy) / (predicted non-creditworthy)
= 17/ (3+17) = 0.85 = 85%

Since accuracies for creditworthy and non-creditworthy are comparable 78.46% and 85% respectively, this model isn't biased.

Writeup

✓

The section is written clearly and is concise. The section is written in less than 250 words.

Awesome!!

The section is written clearly and to the point.



All questions have been answered:

1. 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

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

1. How many individuals are creditworthy?

Excellent work done!!

- The business and technical justification provided to choose the Forest model for predicting the creditworthiness of the loan applicants is explained very well.
- The ROC graph is provided and the write up is in accordance with the visualization.
- The value of creditworthy individuals is within the range.

Suggestion: If you want, you can check [ROC & AUC curves](#) for a very intuitive explanation about ROC and AUC. You can also check [ROC Curve](#). There is also its "cousin" Precision-recall curves. Here is a simple explanation regarding this technique:[Precision-Recall Curve](#)

 [DOWNLOAD PROJECT](#)

[RETURN TO PATH](#)