UDACITY

‹ Return to Classroom

Creditworthiness

REVIEW

HISTORY

Requires Changes

2 specifications require changes

Dear Friend,
Thank you for your submission, really great job so far, while working on the concepts of the Classification Models, which is not an easy task. Mostly, you have covered the project very well, just a few issues which need further improvement and rework. Please look at the comments which will help you to resolve the issue, in case you need further clarification you can contact your mentor by using the student platform https://hub.udacity.com/,https://knowledge.udacity.com/.

Good luck for the next submission,
Thank you

Business and Data Understanding

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

Good Job!!
The section is written clearly and is concise.

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.

Required: To answer Question 2, Please mention the two datasets as well the examples like credit amount, Purpose of the loan, age_years of the applicant etc. which are being used here to make the decision.

We have 2 datasets: On the one hand, we have the data of all past applications. We have used this dataset to create and train the model.
On the other hand, we have the 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.

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.

Excellent work Done!! 👌👌

You have removed 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 the reason for removal and imputation and provide the respective the visualization.

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

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.

Great work done with the section!!

- All the significant predictor variables are provided with the variable importance charts for all the models.

Suggestion: Please avoid writing the codes in the report to avoid the confusion.

Please find below the examples, on how to find the biases in the model.

Suggestion: The overall percent accuracy of the Logistic model is 76% which is strong.
PPV= true positives \ (true positives + false positives) = 92 / (92+23) =.80
NPV= true negatives\ (true negatives + false negatives) =22/ (22+13)= .63
So after checking the confusion matrix there is bias seen in the model's prediction to Creditworthy.

The accuracy of the Forest model is 79% which is strong
PPV= true positives \ (true positives + false positives) = 102 / (102+28) =.78
NPV= true negatives \ (true negatives + false negatives) = 17/ (17+3) = .85
So after checking the confusion matrix there is no bias seen in the model's prediction.

In order to know whether a model is biased or not, we should look at the accuracies on both segments. You can consider bias as a tendency of a model to predict one of its outcomes much more accurately than the others. Let's say that a model's accuracy in correctly predicting creditworthy individuals is 79% and the accuracy in correctly predicting non-creditworthy individuals is 60%. This means that this particular model has bias towards correctly predicting creditworthy individuals because its accuracy in this segment is way higher than in the other. Now another model has about 79% accuracy predicting Creditworthiness and 81% of accuracy in predicting non-creditworthiness. In this case, we say that this model is almost not biased at all, because the difference between those accuracies is very small.

For further understanding on the bias in the confusion matrix:
Confusion matrix

Writeup

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

Good Job!!
The section is written clearly and is concise.

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?**

Required: Please note that the Logistic model is not the best one to predict the creditworthiness of the loan applicant?
We need to see which model provides the highest overall accuracy among all the Four models.
Further, we have to see which model has the higher overall accuracies within the individual creditworthy and non-creditworthy segments.
Remember here your manager only cares about how accurate you can identify people who qualify and do not qualify for loans for this problem

Required: Please provide the ROC graph and justify the role of ROC graph in choosing the model to predict the creditworthiness of individuals.

Suggestion: As stated in the project report, we also need to argue if the model we choose is the best one (or among the best) according to the ROC graph. One possible way of seeing it is by noting which of the Four models reaches the top the quickest and which model is the overall highest curve of all. This means that for a given amount of false positive predictions (wrongly predicted creditworthy people),this model will give the best number of true positive predictions (correctly predicted creditworthy people). An alternative way of seeing it is by comparing the Area Under the Curve (AUC). The highest curve will also have the highest amount of area covered by it. Thus the higher the AUC, the better the model according to it.
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

technique: Precision-Recall Curve

RESUBMIT

DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

RETURN TO PATH