Creditworthiness

REVIEW

HISTORY

Meets Specifications

Congratulations! You passed in this project!

In the review, I made some comments and also suggested a few links for further reading if you want to go deeper in a few aspects seen in this project. I hope they are useful to you :)

All the best and keep on learning!

Business and Data Understanding

✓

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

✓

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?

Awesome

All your answers in Step 1 are correct! Great job!

Comment

To improve the second answer, you could have also mentioned **why** some of them are important. For instance, Age seems important because the older someone is, the more money they probably earn compared to another young person with the same social-economic status, because they had time to earn more money than the younger ones. Hence, they are more likely to pay their debts.

Building the Training Set

✓

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

✓

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 necessary variables were correctly removed and the best value to impute the Age field was used. Well done!

Comment

Indeed, the median value is the best one to use here. Alteryx only uses simple imputation methods, such as mean, median and mode values. However, if you want to go deeper in programming languages, there are even more methods! R has these modules to deal with missing data considering other fields. Since Alteryx is built on R, you may create macros in the future that use those libraries. Here is how to build R macros in Alteryx, in case you are interested.

Train your Classification Models

✓

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

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

All p-values and importance plots are correct! Moreover,all accuracies and confusion matrices are correct as well. You have also correctly identified the bias in each model!

Comment

The fact that there are more correctly predicted creditworthy individuals over correctly predicted non-creditworthy individuals reflects the high number of actual creditworthy individuals in the dataset. That way, the models are exposed to much more examples of creditworthy individuals and will be able to better identify them. This blog post shows how people deal - in general - with those class imbalances (one class that happens much more frequently than the others) in a high-level approach:
https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/

Writeup

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

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?

Awesome

Excellent work by using all those different techniques to justify your choice of Forest model as the best one. The number of creditworthy individuals is right!

Comment

Forest's F1 is also the highest among all four models, and this points out that this model is the best one as well. This quantity is not seen in this course, but if you want to learn more about it, take a look at this link: https://www.quora.com/What-is-an-intuitive-explanation-of-F-score.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

START