Expected Questions

Why were there 5 test datasets instead of only 1?

I used 5 test datasets mainly for validating my methods for using the model probability and return potential thresholds. I wanted to see if the actual returns would vary across different datasets.

Why was the model only trained on 200,000 samples if you had 2.2 million?

Training time for the models would have increased too much for the time allotment for this project. Future iterations could include more data and the model performance would likely increase.

What went in to choosing the 4 models for comparison?

In initial runs with PyCaret, the 4 models used in the comparison were the top 4 models. All models in the PyCaret model library were tested.

What do you mean by representative sample?

The representative samples were randomly drawn samples from the entire dataset and resulted with a distribution of charged off loans vs paid in full loans that were equal to that of the whole dataset. The training dataset took an equal number of charged off and paid in full loans to ensure the less frequent, charged off loans were learned.

Why was AUC used instead of Accuracy?

The model could be 81% accurate by predicting the loan would be paid in full each time. However, that provides no value since it wouldn't catch any of the loans that were charged off. AUC measures how well the model does at separating the two classes.

What would be needed to make this solution usable?

More validation would be needed to ensure high positive returns on investment. There was quite a bit of variation in returns across the test data sets. Prior to using this in the real world, we would want to be able to be confident a high return would be seen on new data. Training the model on more data would be the next step.

Why was LendingClub data used when they are no longer doing peer-to-peer loans?

The LendingClub data was used because of its public availability as a proof of concept. It would be beneficial to repeat the experiment with other platforms in which the model would be used.

Why was under-sampling used when PyCaret has a method for fixing class imbalance?

The method PyCaret uses for fixing class imbalance is SMOTE, which stands for Synthetic Minority Oversampling Technique. Since there was plenty of data available, there was no need to create synthetic training data.

Are there any other models that were not evaluated?

Yes, there are many other machine learning models outside of PyCaret's model library that were not evaluated. Additionally, neural networks were also not evaluated, which could potentially give better results.

What is a grid search?

By grid search, I mean testing all possible values for the model probability and potential return percentage to see their affects on the actual return percentage.