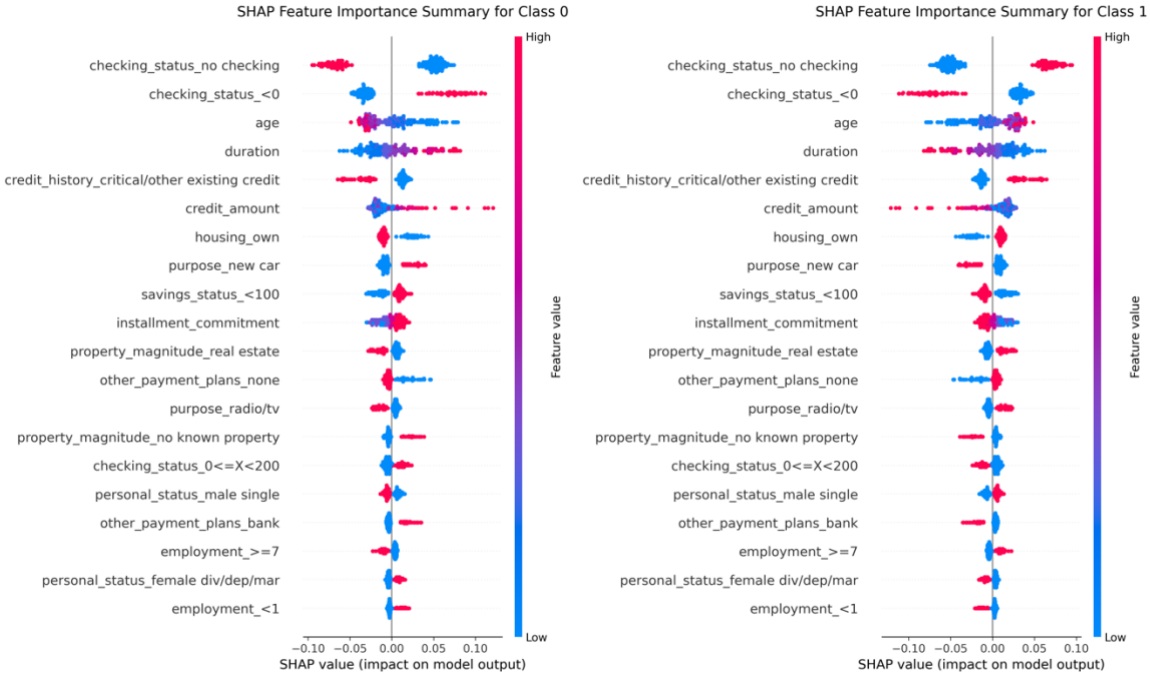
Ungraded Assignment
Week 3 - Explainability and Interpretability:

Understanding and Implementing AI Explainability

In the following analysis, below are several important notes:

- Class 0 represents “bad” loans, which will be rejected (negative class)
- Class 1 represents “good” loans, which will be approved (positive class)
- The data are encoded using one hot encoder and scaled using standard scaler

Global feature importance analysis



The two figures above show features importance on a descending order. The figures are the same, with only switching which class’s prediction we are explaining. The first figure on the left explains negative class or bad loan prediction, while the second figure on the right

explains positive class or good loan prediction. We can gather a general intuition that a loan is more likely to be predicted as a good one if:

- Having no checking account
- Having no overdrawn accounts (checking <0)
- Generally, is older in age
- Generally, having shorter loan duration
- And so on

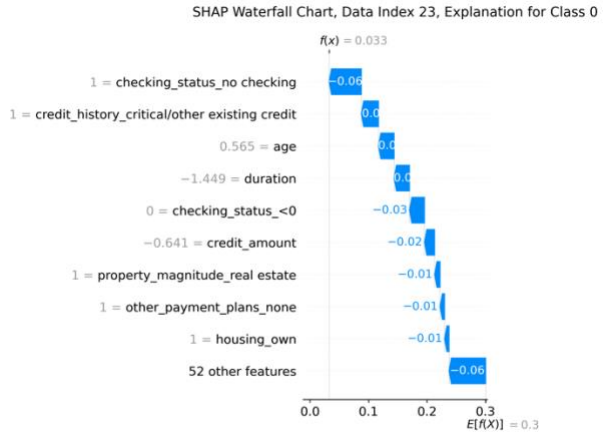
One interesting point is that we see `checking_status_no checking` is the most important feature. Because the feature is a result of one hot encoding, it only has a binary value of either 1 or 0. Therefore red (higher feature value) indicates 1 and blue indicates 0, and it shows that when the value is 1 it has a positive SHAP value, meaning it increases the probability of predicting a good loan (class 1). This may be counterintuitive as we would assume banks are less likely to provide loans for people with no accounts. But this could be due to the case of selection bias (people with no checking account rarely gets approved, and in cases where they are approved, the bank was very selective to ensure only those with low risk of delinquency is approved). But we would need further information to confirm the hypothesis.

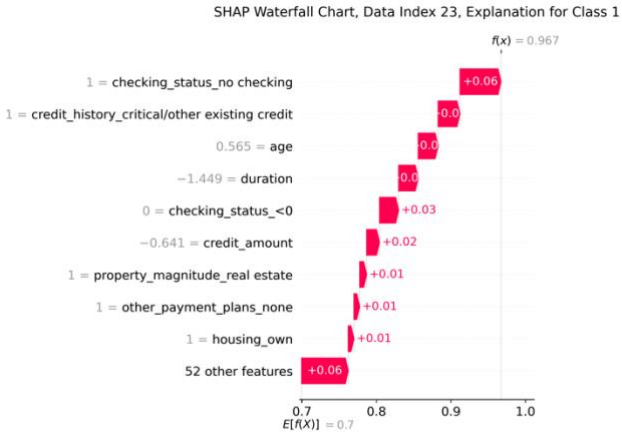
We should also consider that there are several features that are not clearly separated, such as in the case of duration and age. This is seen from the graph where the feature values are continuous (showing a gradient from blue, purple, and red) and that there is no clear separation of the values.

Detailed analysis of one approved loan

For this analysis, I took one sample data where the model predicts it with high confidence. Seen from the SHAP waterfall chart below, the figures indicates that most if not all the features go into one direction (minimizing probability for class 0 and maximizing probability for class 1).

The waterfall chart is ordered by magnitude, so we can see that the most contributing features are at the top. For this sample, not having a checking account is a factor that has the most contribution to push prediction into positive class with 0.06, followed by credit history, age, duration, and so on.

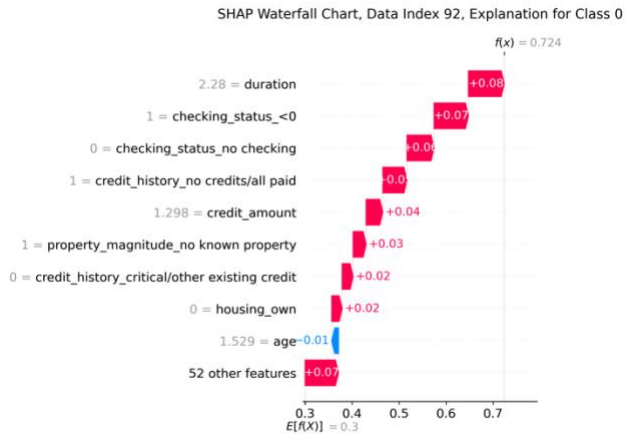


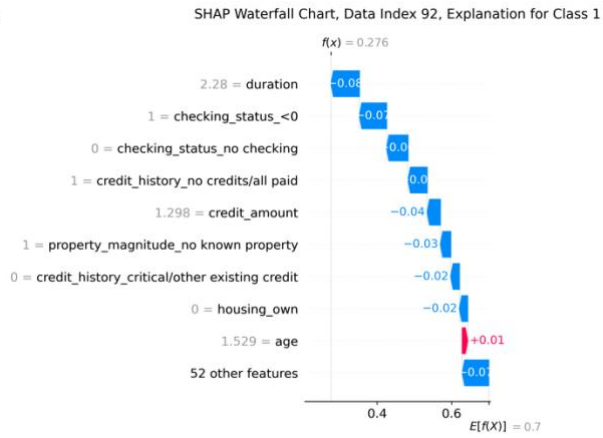


Detailed analysis of one rejected loan

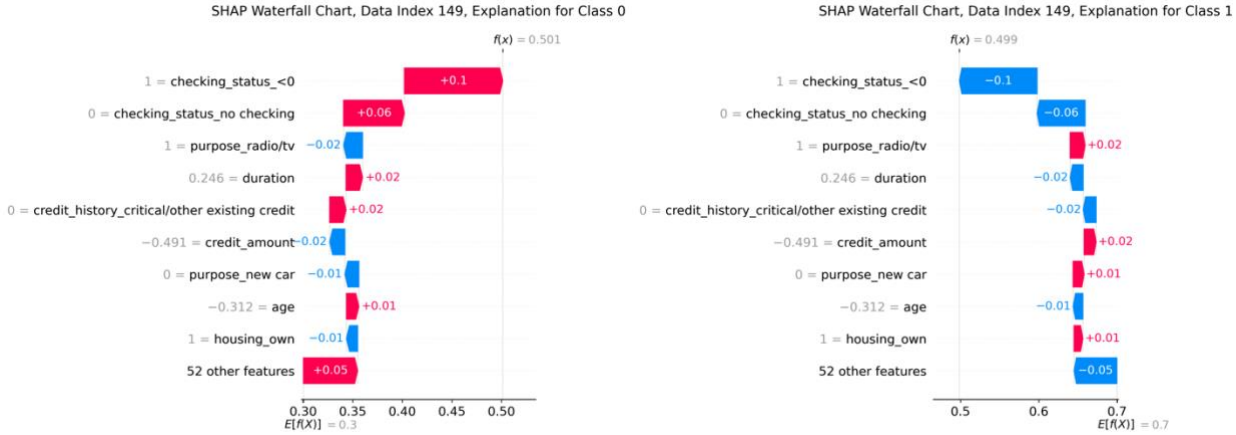
For this analysis, I examined a sample where the model predicts a rejection with high confidence. As seen from the SHAP waterfall figures below, similar with the approved case, the features predominantly push in one direction (maximizing probability for class 0 and minimizing probability for class 1), except for age.

For this sample, the longer duration is the most significant factor that push prediction to negative class, followed with having an overdrawn account, having no checking account, having critical credits, and so on.





Detailed analysis of one borderline loan



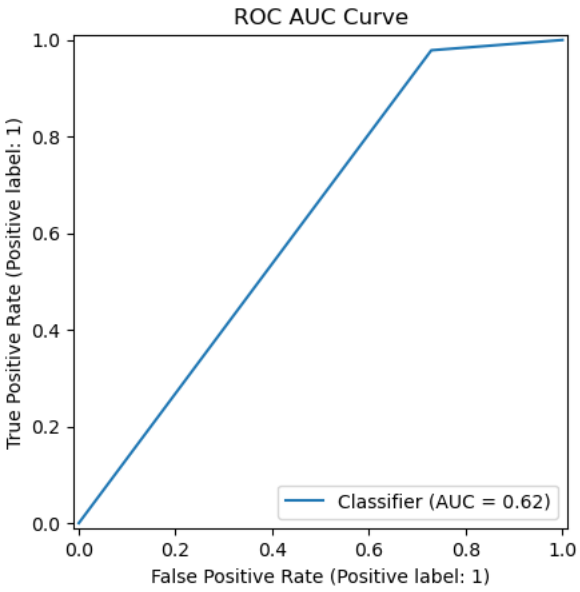
In this analysis, I selected a sample where the model shows less confidence in its prediction by choosing a sample with least difference in the class prediction probabilities. The chart in this sample shows an interesting difference with features going to both directions.

On the positive side, having the loan for getting a radio, tv, or a new car, having a relatively smaller loan amount, and owning a house work in favor of predicting positive class. However, these indicators are counterbalanced by negative factors such as having an overdrawn account, having a checking account, longer loan duration, having critical credit history, and being relatively younger. Unlike the previous two cases of high confidence approval and rejection, the features here don't align uniformly in one direction, leading to a more uncertain prediction.

Performance Metrics Analysis

The model has the following performance scores (rounded):

- Accuracy: 0.77
- Precision: 0.76
- Recall: 0.98
- F1: 0.86
- AUC: 0.62



Based on these, it can be inferred that the current model trained has the following characteristics:

1. The model is performing well against positive classes. To some extent, one of the reasons could be because the dataset is a bit imbalanced (700 positive samples and 300 negative samples).
2. The model may tend to overclassify results as positive, therefore we see the high recall score. But if we look at the ROC AUC, the score is quite low and is near 0.5. A ROC AUC score of 0.5 indicates the model isn't performing any better than a random guessing model.
3. It is further supported with the base probability presented in the SHAP analysis, where base probability for positive class is 0.7

Explanation Quality Assessment

Generally speaking, the modelling and SHAP explanation in this case seem to align well with business or industry practices. Banks would tend to provide loan to lower risk individuals which is represented by proxy metrices such as their financial wellbeing (having no overdrawn accounts, have savings, have a house or property, etc) and the loan characteristics (smaller and shorter loan indicates less risk).

SHAP explanation here is seen to be consistent with the model's confidence. When the model is clearly confident about its prediction, SHAP shows the features are uniformly

supporting that prediction. Meanwhile, when the model is not confident, the features are not aligning on their support to the prediction.

In the Github repository, I have also provided LIME results for the same 3 samples. Both LIME and SHAP are quite aligned on which features support the model's prediction too. These shows the results should be quite dependable.