Data Science in Finance Capstone Project Report

Methodology, approach, and model selection rationale

The data provided by Lending Club has been cleaned and some exploratory data analysis was performed while aiming to preserve as much data as possible, opting to keep rows of data unless there is a strong reason not to.

A binary classification model which predicts the probability of default or non-default for the test data is tested first with 20% of the cleaned dataset; 80% of the dataset is used for training the model. The model's recall, which is the performance of predicting defaulters correctly, is poor even though it generalises well.

A challenger model to the baseline model includes a penalty for predicting a loan as non-default when it defaults and this penalty is set higher than the penalty for predicting a loan as a default when it is not. With this 'custom loss' added to the model, the correct identification of defaulters improves significantly. However, the model has fewer correct predictions overall. In business terms, by assigning penalties in a way that makes incorrect non-default predictions more costly, the business may be able to save more by being extra cautious even if that means it does not always predict defaults and non-defaults correctly. It instead reduces instances of approving potential default loans.

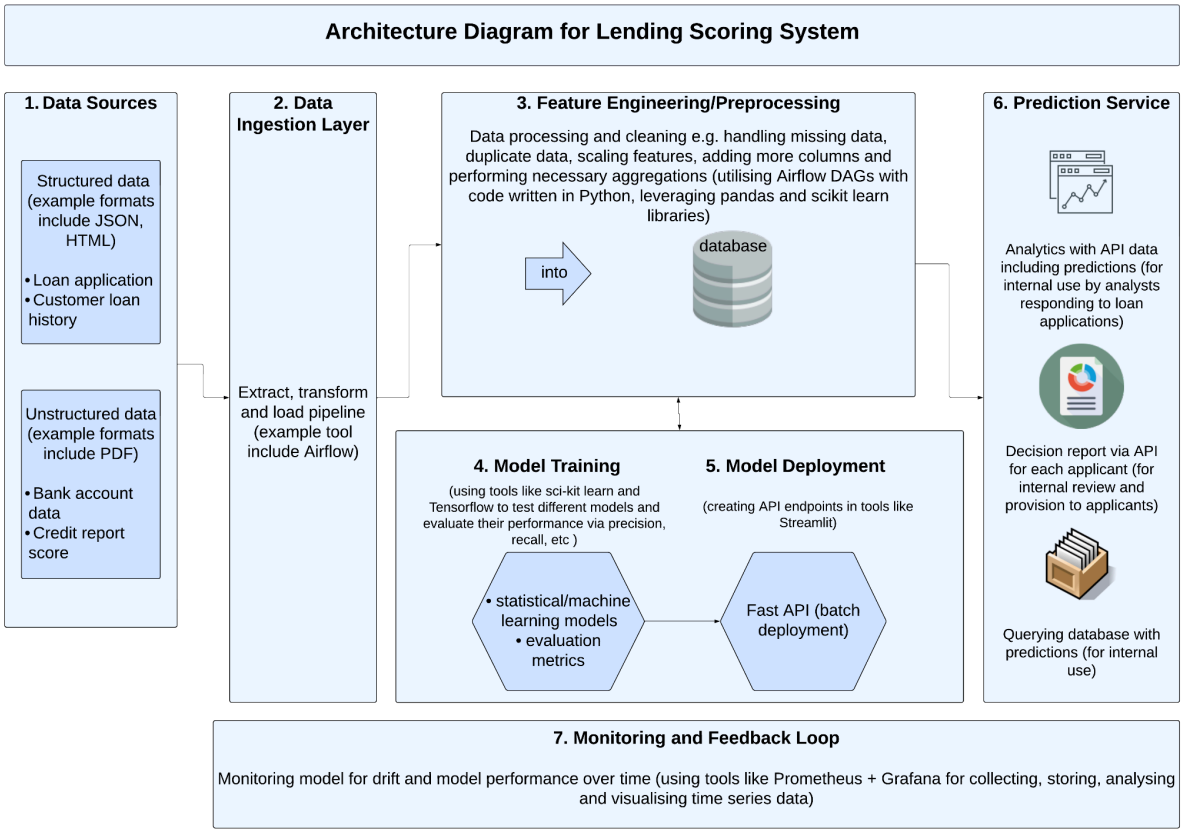
Advantages and limitations of the chosen model

While it is an improvement on the baseline model in the short term, the challenger model could still be improved for making correct predictions overall. Since it is a logistic regression, it may also perform worse when the data available on the borrowers becomes more complex and we have more features to add to the model.

Architecture of the final solution

An architect of the business solution is shown below in Figure 1.

Figure 1: Architecture diagram for lending scoring system



Considerations on deployment and scalability of the solution

As highlighted in the architecture diagram in the previous section, the model will be interacted with by analysts in charge of the loan approvals process. Among other responsibilities, these analysts will be privy to visualisations that bring together the prediction made by the model for each applicant. Analysts will also be able to see a base report that they can customise and keep for internal purposes or use as a basis for responding to applicants in the case of refusals to grant loans.

The data from the model (including the predictions made) can also be fed into a database (with unique IDs for each person) that can be queried at a later date. For example, if an applicant reapplies, the previous application stored will be useful for understanding improvements in applicants' information that have caused a change in the prediction made. Alternatively, it can serve as a baseline of information for analysts who may want to do a full review and incorporate past information in the final decision-making without solely relying on the model's new prediction.

Estimated impact/ROI of the project

The sum of the loan amount for non-defaulters in the Lending Club dataset is $1.34 trillion. By comparison, the loan amount sum for non-defaulters is $0.195 trillion. If the business makes a 60% loss from defaulted loans, that is about $117m. That is quite a substantial loss for the Lending Club.

In estimating the project impact, one of the key things to evaluate in both the medium and short term is how the loss from default loans is trending. A significant reduction in the losses paired with efficiency improvement of the application processing system would be considered project success.