Toast Data Science Case Study

Loan Default Prediction

Task

Your task is to produce a model to predict how likely a loan will default.

As you work through this case study, please know that we are far more curious in how you approach the problem as opposed to the model performance. This case will serve as a basis for our conversation in a follow-up interview.

Background

Toast Capital is one of the most important financing products that Toast offers. Eligible Toast customers can take out Toast Capital loans for any restaurant need, such as restaurant expansion and operating expenses. Loan repayments are automated as a fixed percentage of daily card sales processed through Toast. Since the product launch, many Toast customers have enjoyed the convenience and flexibility of this product to help their business grow.

In order for Toast to make the right decision about loan eligibility and pricing, it is important for us to understand the risk of lending to each customer, and Data Science plays a critical role in this effort.

For this use case we have sampled and anonymized historic customer data similar to what you could use to solve this problem internally. It includes both static features and daily revenue amounts for each customer in the year leading up to the underwriting date. This scenario is intended to simulate the modeling work we would conduct ahead of underwriting a loan on February 2nd, 2022. We're interested in your thought process in solving the problem, and we look forward to having more discussion on this topic in the interview.

Dataset

The dataset contains information about personal loans from a fictitious bank. Where are 2 csv files:

1. Lending_default_train_tx.csv: dataset with one row per date & restaurat_id.
    - Restaurant_ID: ID of customer.
    - Tx_date: The day of a customer transaction.
    - Tx_hours: The total number of hours a customer was open for transactions.
    - processing_volume: The sum of all processed transactions for a given Tx_date
2. Lending_default_train_account.csv: dataset with one row per restaurant
    - Restaurant_ID: ID of customer

    - Ownership_type: Type of ownership. Sole Proprietorship, etc.
    - Restaurant_catagory: Service style of restaurant. QSR (Quick service), etc.
    - Market_segment: Customer market segment, information about the scale of a restaurant. SMB, enterprise, etc.
3. Lending_default_train_label.csv: dataset with one row per restaurant
    - Restaurant_ID: ID of customer.
    - loan_default: target variable indicating whether the customer loan defaulted.
4. Lending_default_holdout_tx.csv: holdout dataset for scoring
5. Lending_default_holdout_account.csv: holdout dataset for scoring

Instructions

Please work in a single Jupyter notebook. You must use Python in the notebook. We encourage you to explore no more than two model approaches. Along with your notebook, please submit a csv file as follows with the predicted default probability for each loan in holdout_for_prediction.csv

restaurant_id	*pred*
….	0.1234
….	0.5678

You are free to spend as much time as you'd like on this problem, but we believe that 4-6 hours should be a reasonable amount of time investment. We are not expecting a perfect model; instead, we are interested in seeing your ability to work through different aspects of this data science problem, and understand the pros and cons of the approach. Please take time to document your thought process and code. Also feel free to jot down things you'd like to try if there wasn't a time constraint.

What to Expect Next

After you submit the notebook and prediction, our team will conduct an initial screening. If passed, we will schedule an interview for you to walk us through your notebook and have a further discussion on this topic.