MachineHack: Analytics Olympiad 2023

Hariprasath Venkatraman

Table of Contents

Introduction

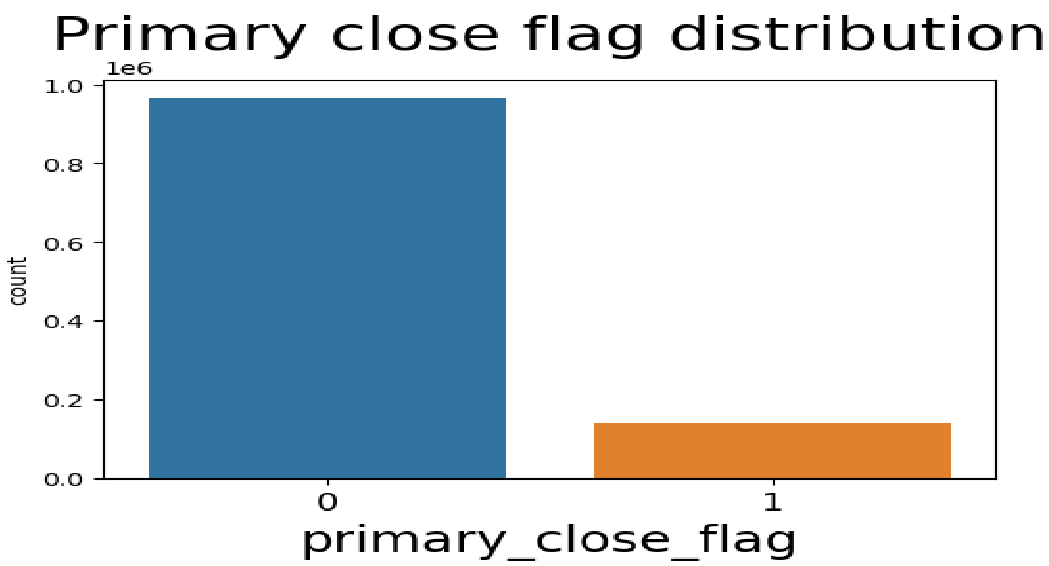
Create a machine learning models to determine the likelihood of a customer defaulting on a loan based on the credit history, payment behaviour, and account details.

Data Understanding

To predict the likelihood of a customer default, the model must predict two outcome variables.

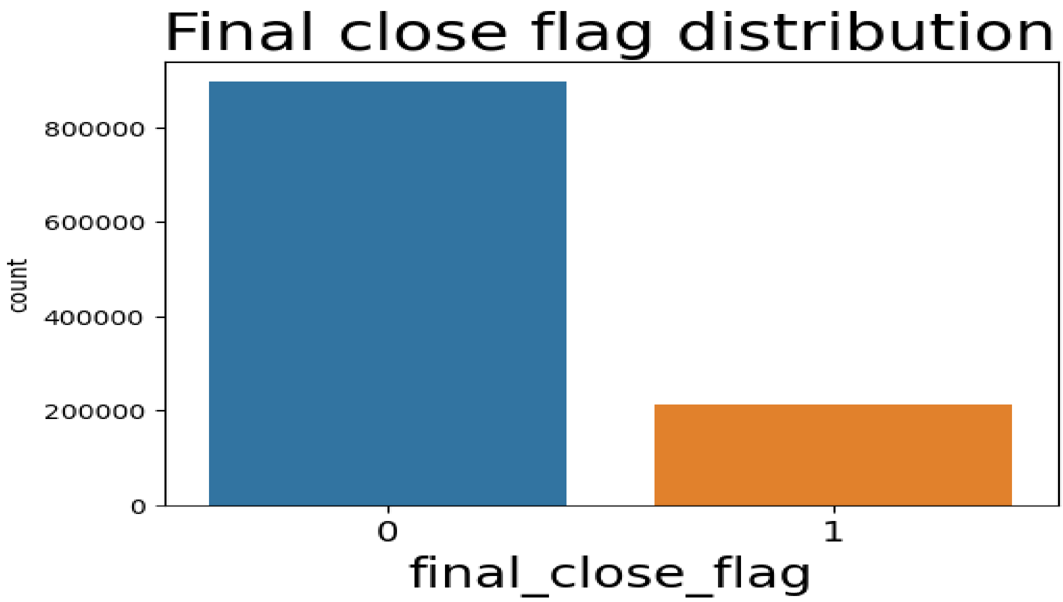
- ✓ Primary close flag
- ✓ Final close flag

- Primary close flag
    - Primary close flag target column contains binary values.
- Primary close flag data distribution



There is an imbalance between the primary close flag class distribution

- Final close flag
    - Final close flag target column contains binary values.
- Final close flag distribution



There is an imbalance between the final close flag class distribution.

- Features category
    - The features are categorized into following types,
        - ✓ Credit card information
        - ✓ Loan overdue information
        - ✓ Credit utilization and limit information
        - ✓ Encoded features
    - Credit card information features
        - ✓ record number
        - ✓ days_since_opened
        - ✓ days_since_confirmed
        - ✓ primary_term

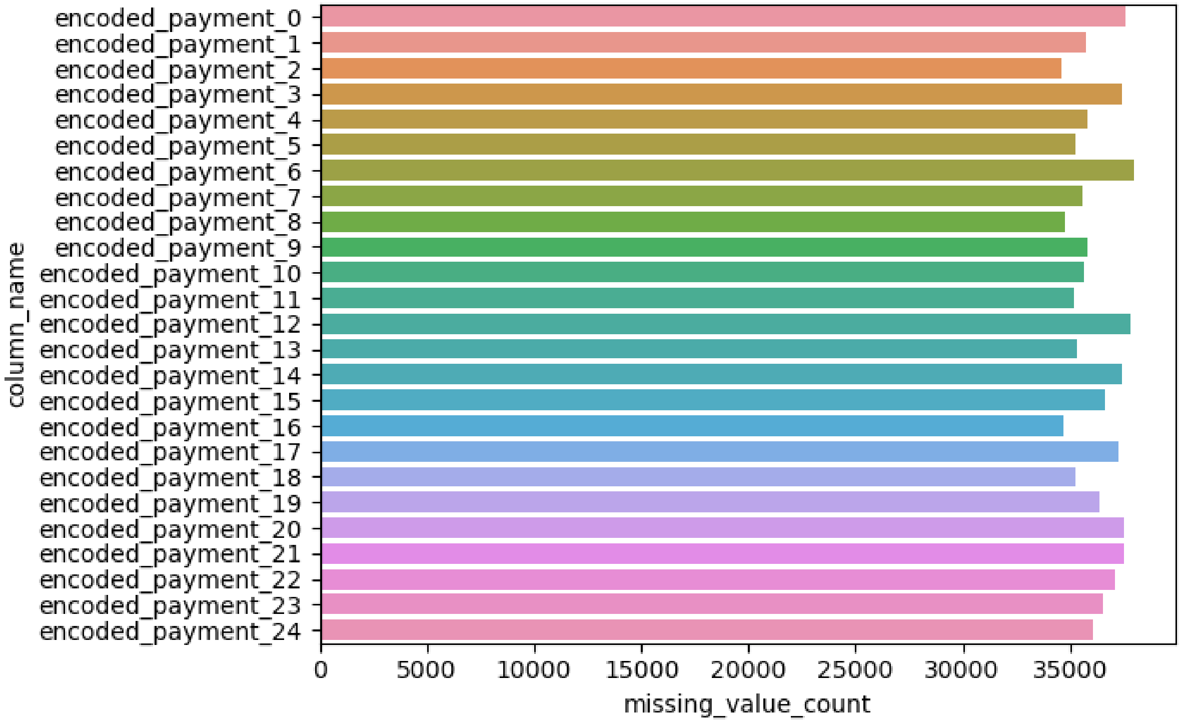
- ✓ final_term
    - ✓ days_till_primary_close
    - ✓ days_till_final_close
    - ✓ loans_credit_limit
    - ✓ loans_next_payment_summary
    - ✓ loans_outstanding_balance
    - ✓ loans_max_overdue_amount
    - ✓ loans_credit_cost_rate
- Loan overdue information related features
    - ✓ loans_within_5_days
    - ✓ loans_within_5_to_30_days
    - ✓ loans_within_30_to_60_days
    - ✓ loans_within_60_to_90_days
    - ✓ loans_over_90_days
    - ✓ is_zero_loans_within_5_days
    - ✓ is_zero_loans_within_5_to_30_days
    - ✓ is_zero_loans_within_30_to_60_days
    - ✓ is_zero_loans_within_60_to_90_days
    - ✓ is_zero_loans_over_90_days
- Credit utilization and limit information related features
    - ✓ utilization
    - ✓ over_limit_count
    - ✓ max_over_limit_count
    - ✓ is_zero_utilization
    - ✓ is_zero_over_limit_count
    - ✓ is_zero_max_over_limit_count
- Encoded features
    - ✓ Enoded_payment feature from 0 to 24
    - ✓ encoded_loans_account_holder_type

- ✓ encoded_loans_credit_status
- ✓ encoded_loans_credit_type
- ✓ encoded_loans_account_currency

- All of the features above are ordinal categories, except for the following nominal features.
    - ✓ is_zero_loans_within_5_to_30_days
    - ✓ is_zero_loans_within_30_to_60_days
    - ✓ is_zero_loans_within_60_to_90_days
    - ✓ is_zero_loans_over_90_days
    - ✓ is_zero_utilization
    - ✓ is_zero_over_limit_count
    - ✓ is_zero_max_over_limit_count

Data Preparation

There are missing values in the encoded features of both the train and test datasets.



- Feature Engineering.
    - The feature engineering process contains following task.
        - Missing value imputation
        - Category column-wise primary close flag target frequency
        - Category column-wise final close flag target frequency

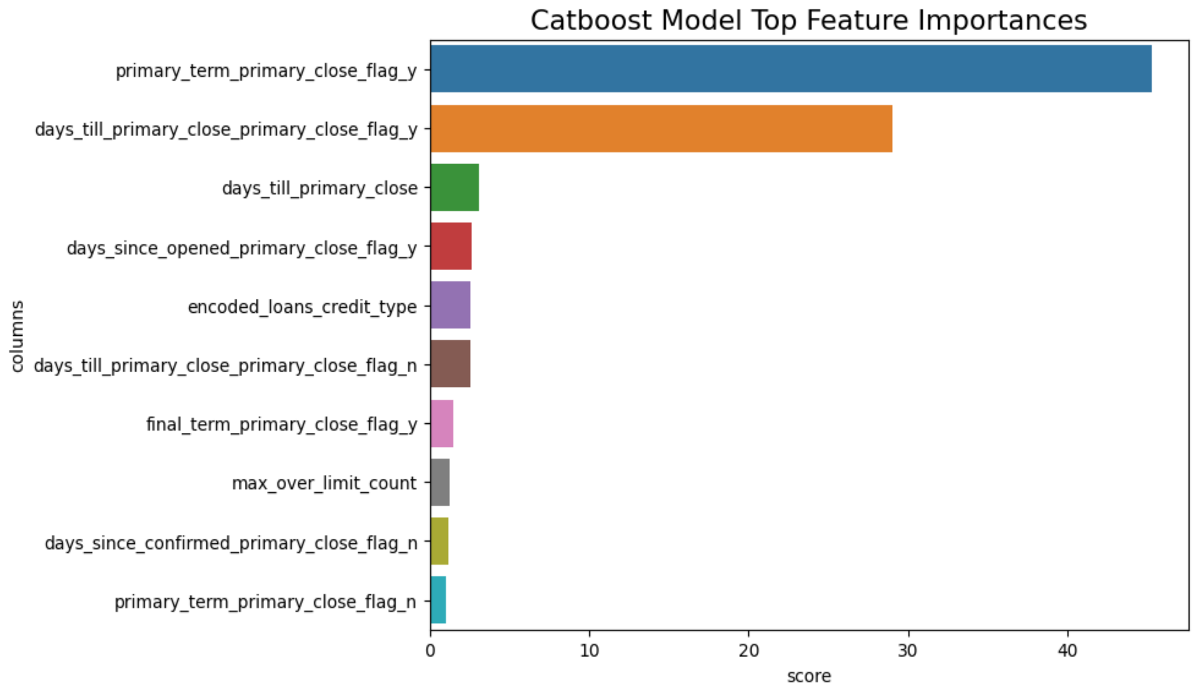
Data Partition

- The train data further split into train and validation data. The split is based on the target column (stratified split).
  - 67% of data to train the model
  - 33% of data to validate the model
- The data is split and stored separately for both target columns.

Model

- The Catboost model was trained separately for both targets, using default parameters.
- The model was evaluated at each iteration using validation data.
- The model's performance was assessed using an accuracy score.

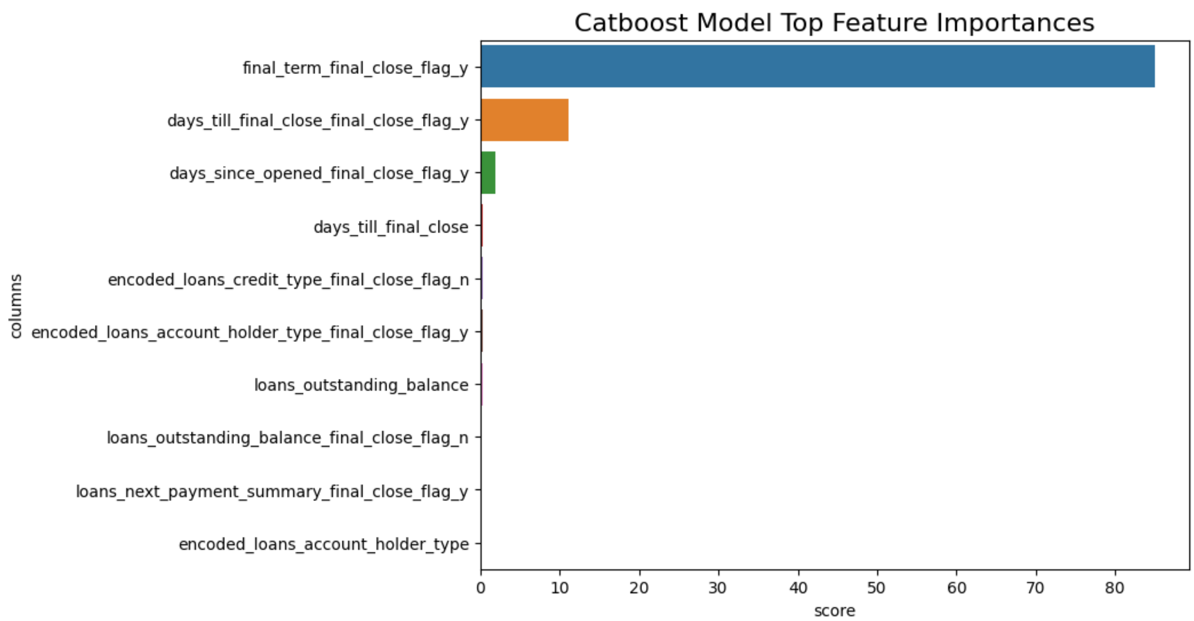
- Catboost model permutation feature importance for the target primary close flag.



The permutation feature importance plot explains that the likelihood of a customer defaulting on a loan is determined by the following features.

- Primary term feature’s category-wise primary close flag target’s positive class frequency
- Days till primary close feature’s category-wise primary close flag target’s positive class frequency

- Catboost model permutation feature importance for the target final close flag.

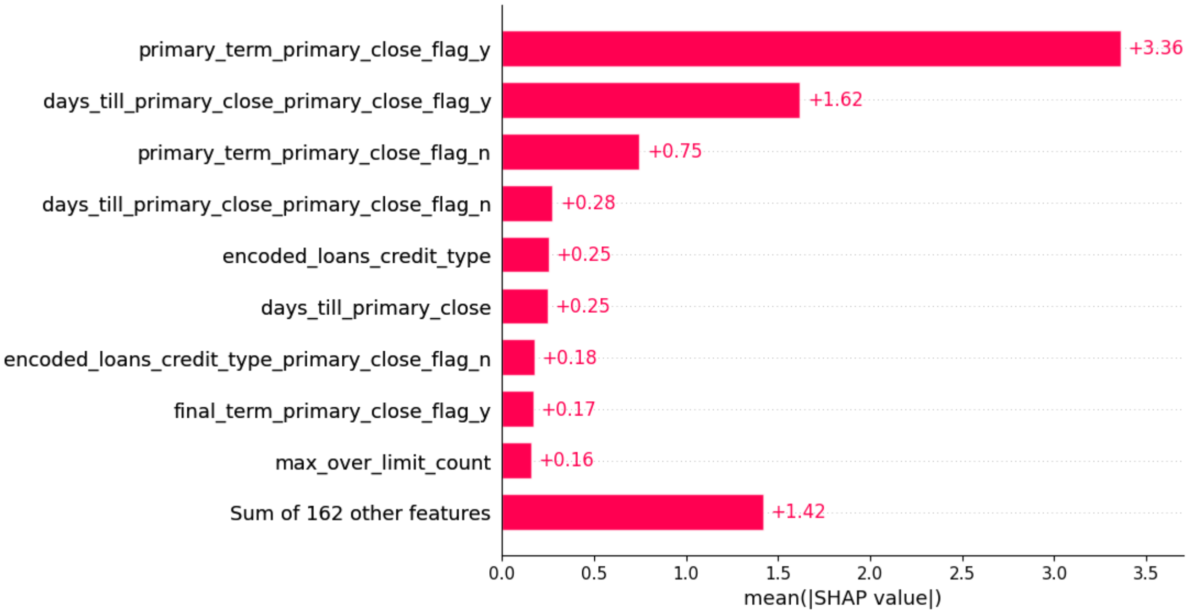


The permutation feature importance plot explains that the likelihood of a customer defaulting on a loan is determined by the following features.

  - Final term feature's category-wise final close flag target's positive class frequency
  - Days till final close feature's category-wise final close flag target's positive class frequency

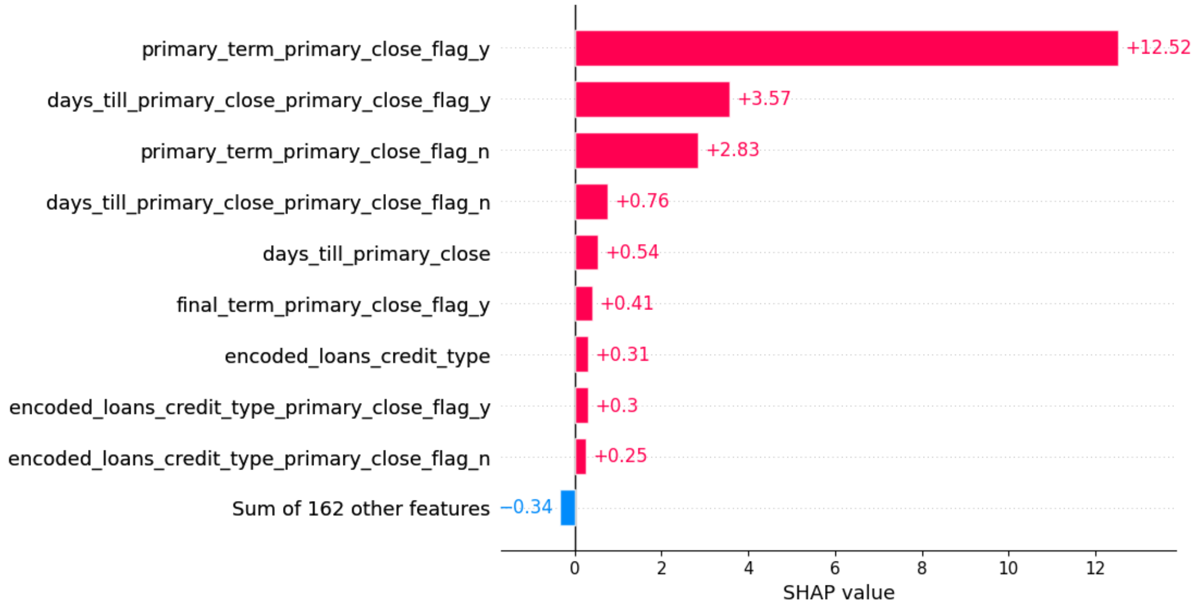
Model Interpretation

- The model further interpreted with the SHAP library.
- Primary close flag
    - SHAP - Global feature importance

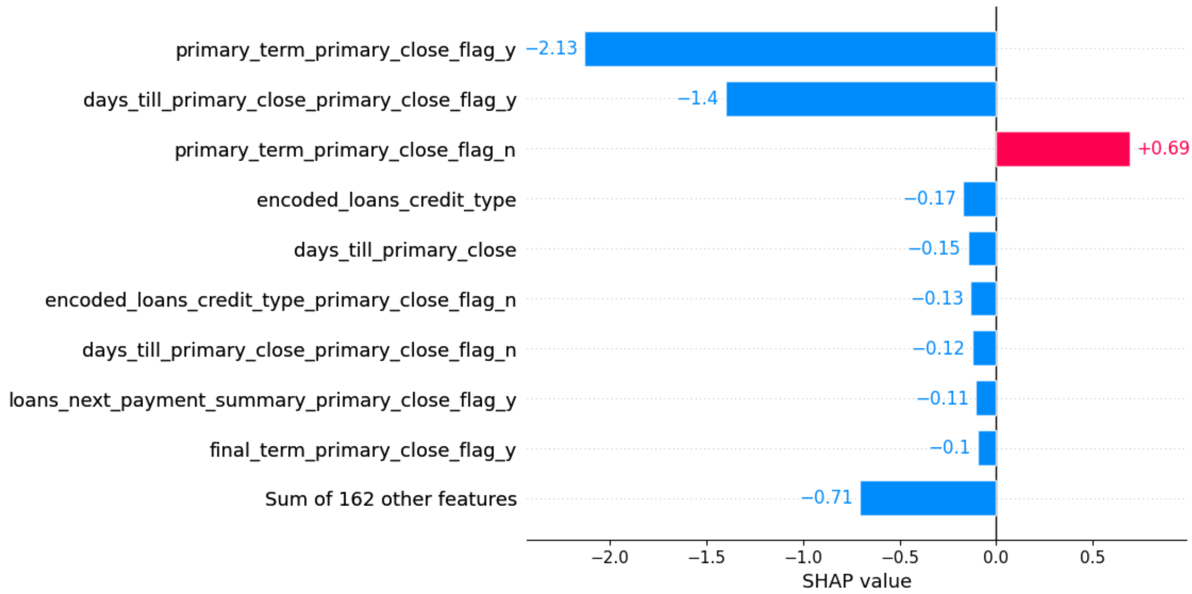


    - The SHAP global importance explain that the likelihood of a customer defaulting on a loan is determined by the following features.
        - Primary term feature's category-wise primary close flag target's positive class frequency
        - Days till primary close feature's category-wise primary close flag target's positive class frequency
        - Primary term feature's category-wise primary close flag target's negative class frequency

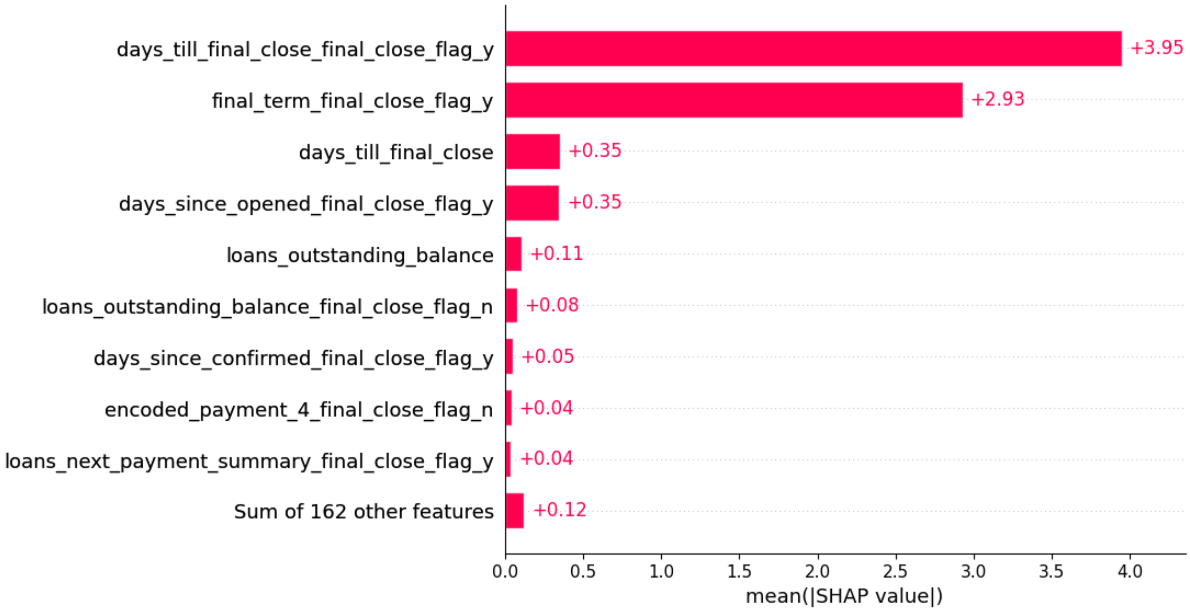
- SHAP - Local feature importance
  - For primary close flag 1



  - For primary close flag 0

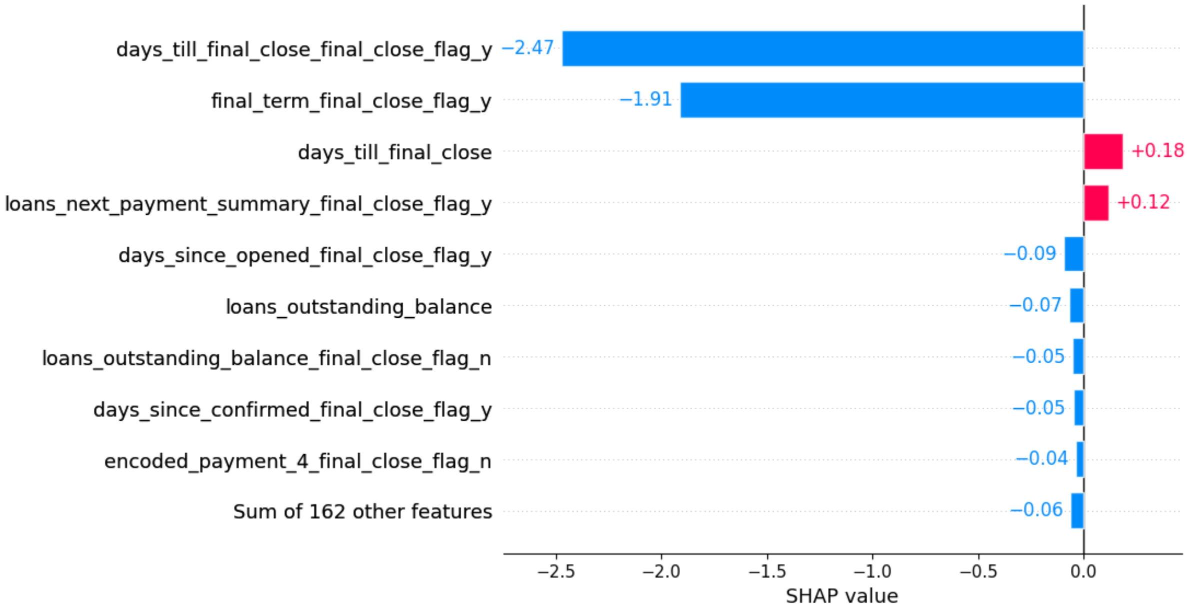


- Final close flag
    - SHAP - Global feature importance

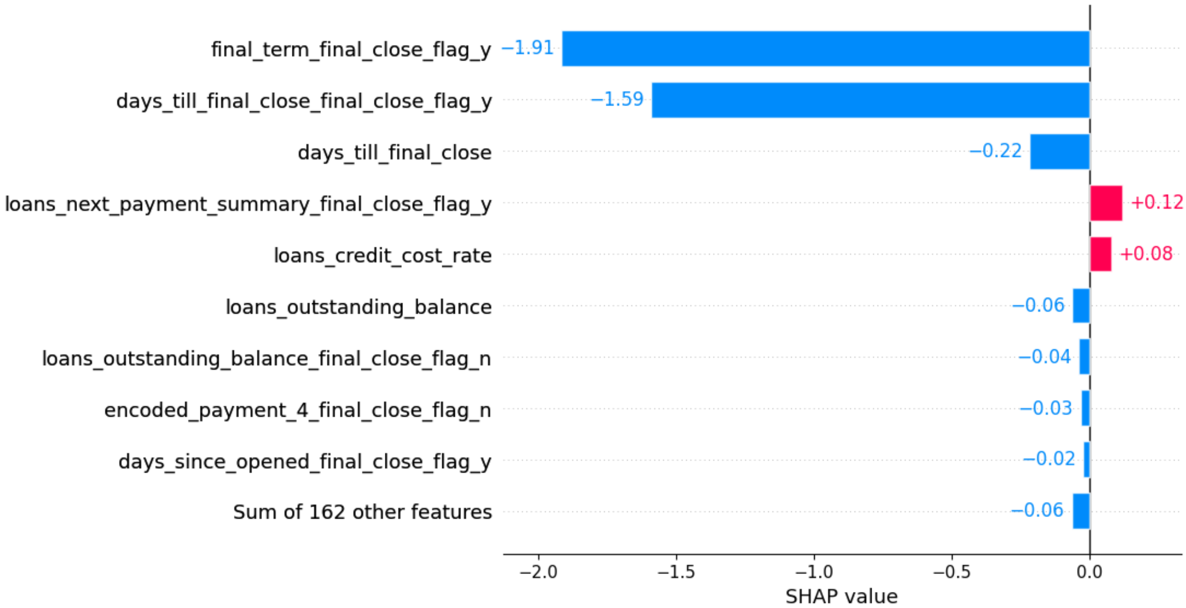


    - The SHAP global importance explain that the likelihood of a customer defaulting on a loan is determined by the following features.
        - Days till final close feature's category-wise final close flag target's positive class frequency
        - Final term feature's category-wise final close flag target's positive class frequency

- SHAP - Local feature importance
  - For final close flag class 1



  - For final close flag class 0



Result

The likelihood of a customer defaulting on a loan is determined by the following features.

- ✓ Primary term feature’s category-wise primary close flag target’s positive class frequency
- ✓ Days till primary close feature’s category-wise primary close flag target’s positive class frequency
- ✓ Primary term feature’s category-wise primary close flag target’s negative class frequency
- ✓ Days till final close feature’s category-wise final close flag target’s positive class frequency
- ✓ Final term feature’s category-wise final close flag target’s positive class frequency