Transunion-data-science-analytics-hiring-challenge-2022-Approach

Machine learning model to classify the credit score based on people bank details and credit related information.

1.Exploratory Data Analysis

- Pandas, seaborn, matplotlib libraries are used in Exploratory data analysis.

2.Data Pre-Processing

- The following columns have non_numeric values. So replaced all the numeric values and convert the column to numerical data type. Also, removed the unusual values.
  - Age
  - Annual income
  - Num_of_loan
  - Num_of_delayed_payment
  - Changed_credit_limit
  - Amount_invested_monthly
  - Monthly_balance
- The following column has special characters. So replaced all the special characters.
  - Occupation
  - Credit_mix
  - Payment_behaviour
- The credit history age column has text and numbers. converted into numeric column by extracting numbers.
- The type of loan column has comma-separated text. So using the TF-IDF vectorizer convert the columns values to numeric format.
- Missing values are replaced by mean and group mean.

- Create a missing row indicator for the columns which have missing values.

.

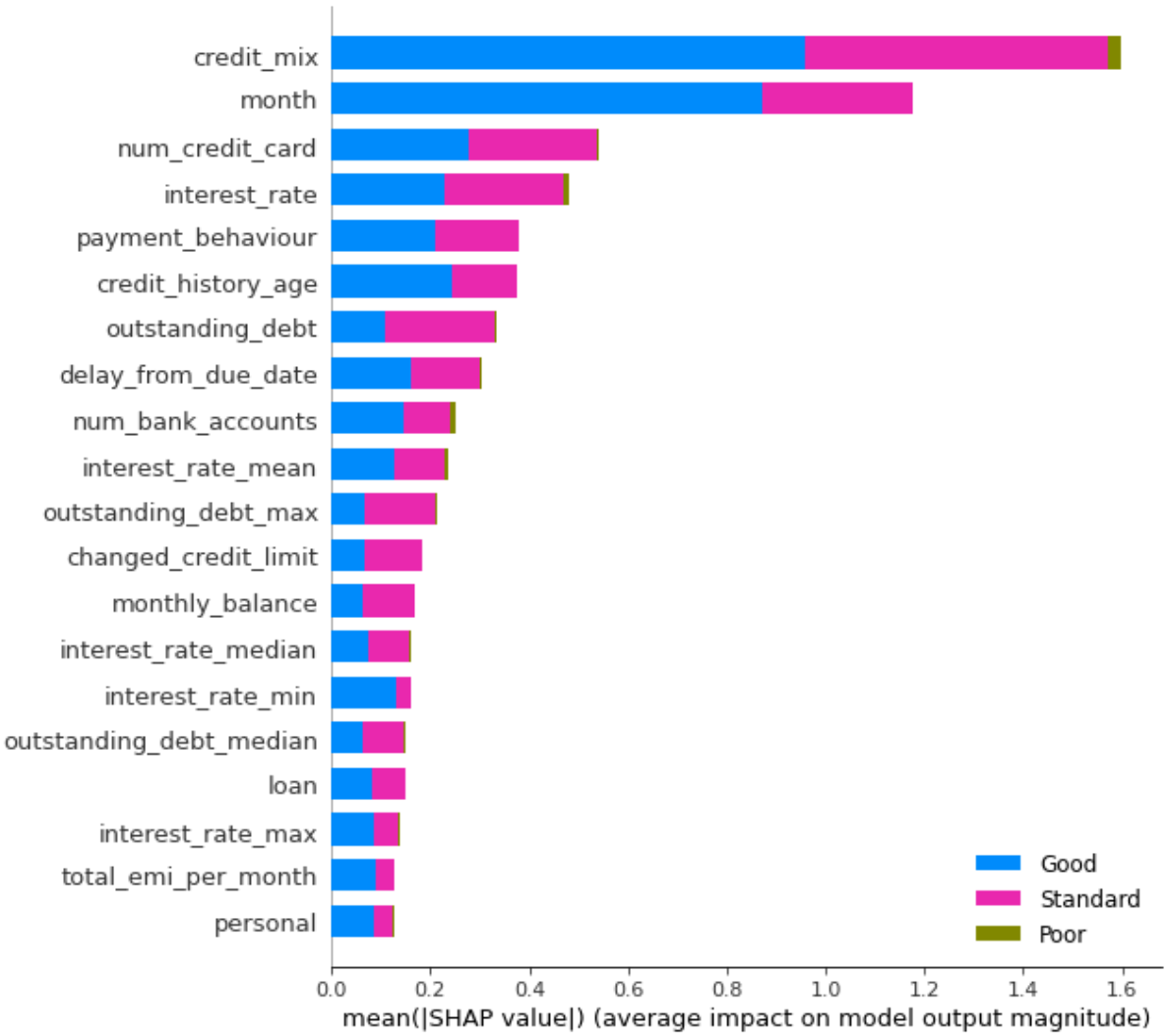
3.Model

- After pre-processing finally 113 columns are selected for the classification model.
- Selected columns are,
    - 0_month
    - 1_age
    - 2_annual_income
    - 3_monthly_inhand_salary
    - 4_num_bank_accounts
    - 5_num_credit_card
    - 6_interest_rate
    - 7_num_of_loan
    - 8_delay_from_due_date
    - 9_num_of_delayed_payment
    - 10_changed_credit_limit
    - 11_num_credit_inquiries
    - 12_outstanding_debt
    - 13_credit_utilization_ratio
    - 14_credit_history_age
    - 15_payment_of_min_amount
    - 16_total_emi_per_month
    - 17_amount_invested_monthly
    - 18_payment_behaviour
    - 19_monthly_balance
    - 20_name_isnull
    - 21_age_isnull
    - 22_ssn_isnull
    - 23_occupation_isnull
    - 24_monthly_inhand_salary_isnull
    - 25_num_bank_accounts_isnull
    - 26_num_credit_card_isnull
    - 27_interest_rate_isnull

    - 28_num_of_loan_isnull
    - 29_type_of_loan_isnull
    - 30_num_of_delayed_payment_isnull
    - 31_changed_credit_limit_isnull
    - 32_num_credit_inquiries_isnull
    - 33_credit_mix_isnull
    - 34_credit_history_age_isnull
    - 35_amount_invested_monthly_isnull
    - 36_payment_behaviour_isnull
    - 37_monthly_balance_isnull
    - 38_credit_score_isnull
    - 39_ssn
    - 40_occupation
    - 41_type_of_loan
    - 42_credit_mix
    - 43_auto
    - 44_builder
    - 45_consolidation
    - 46_credit
    - 47_debt
    - 48_equity
    - 49_home
    - 50_loan
    - 51_mortgage
    - 52_payday
    - 53_personal
    - 54_specified
    - 55_student
- Created catboost classifier with 5-fold stratified cross validatation and tuned the hyperparameters with optuna framework.
- Fitted catboost classifier with tuned parameters

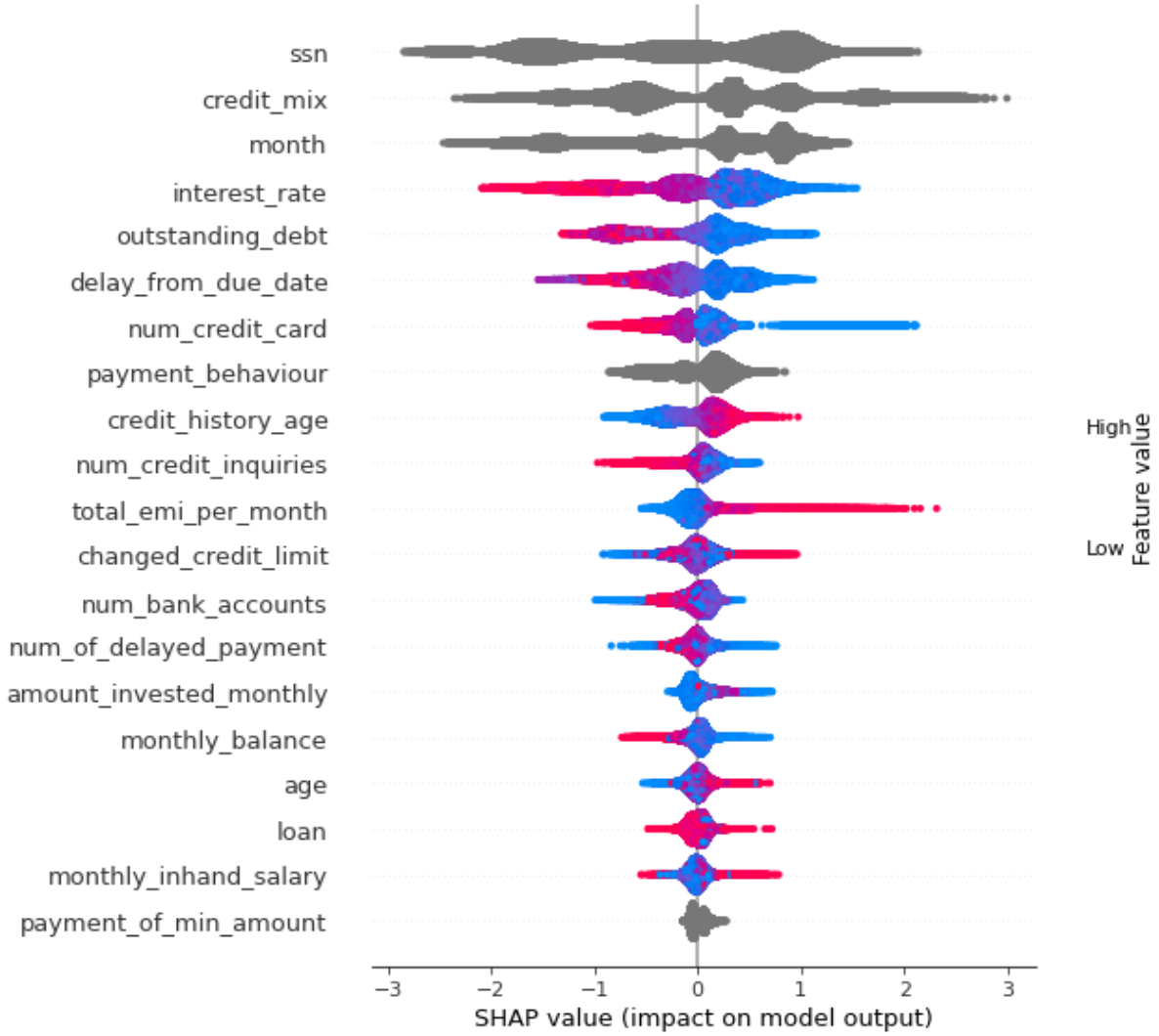
- Parameters are,
  {'reg_lambda': 0.9138676838297956,
  'learning_rate': 0.08135287540629096,
  'n_estimators': 887,
  'max_depth': 10,
  'random_state': 2020,
  'boosting_type': 'Plain',
  'bootstrap_type': 'Bernoulli',
  'subsample': 0.9414567528861059}

- Model Explaination with shap library
- Feature importance plot

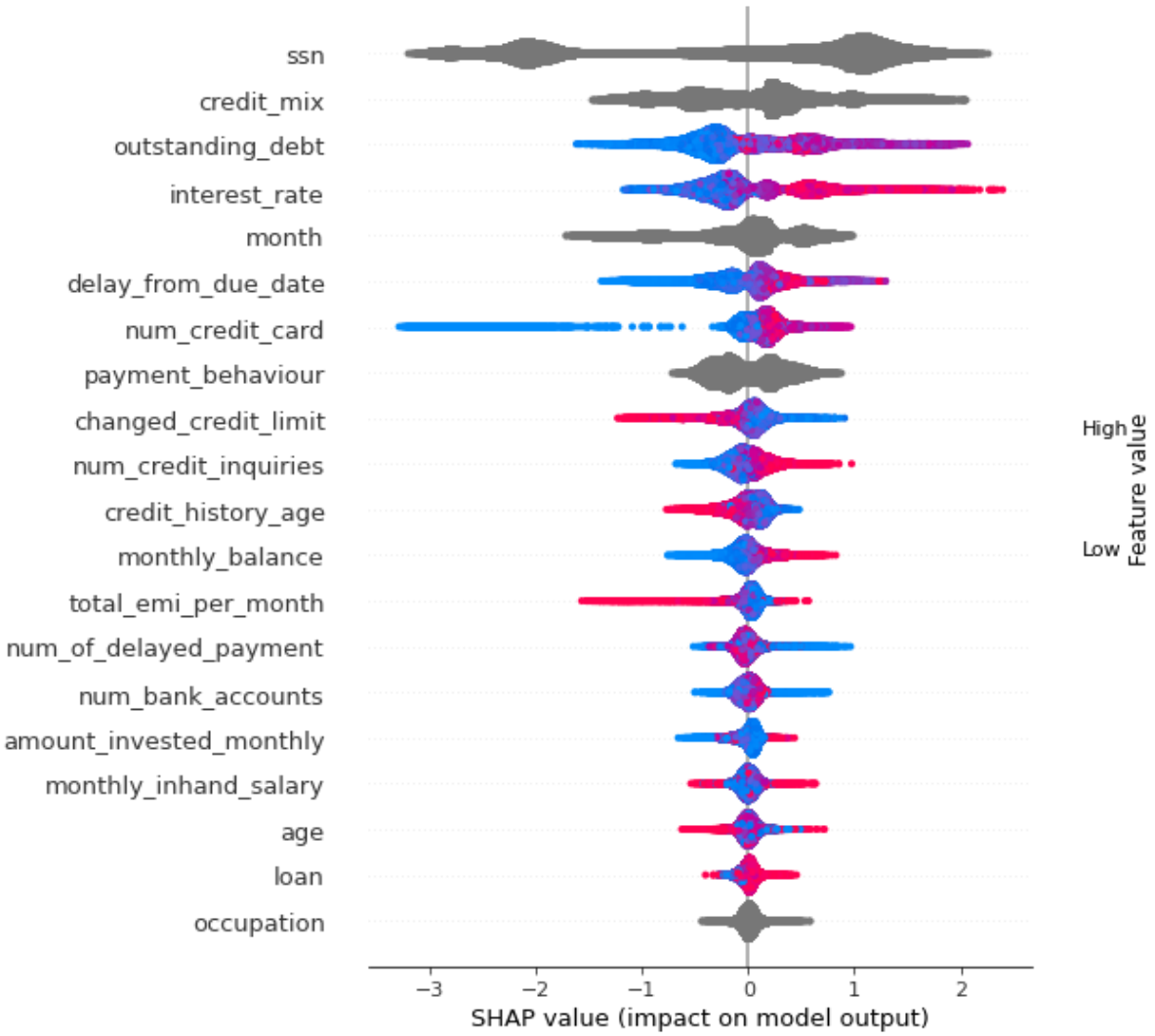


Top features impact the model

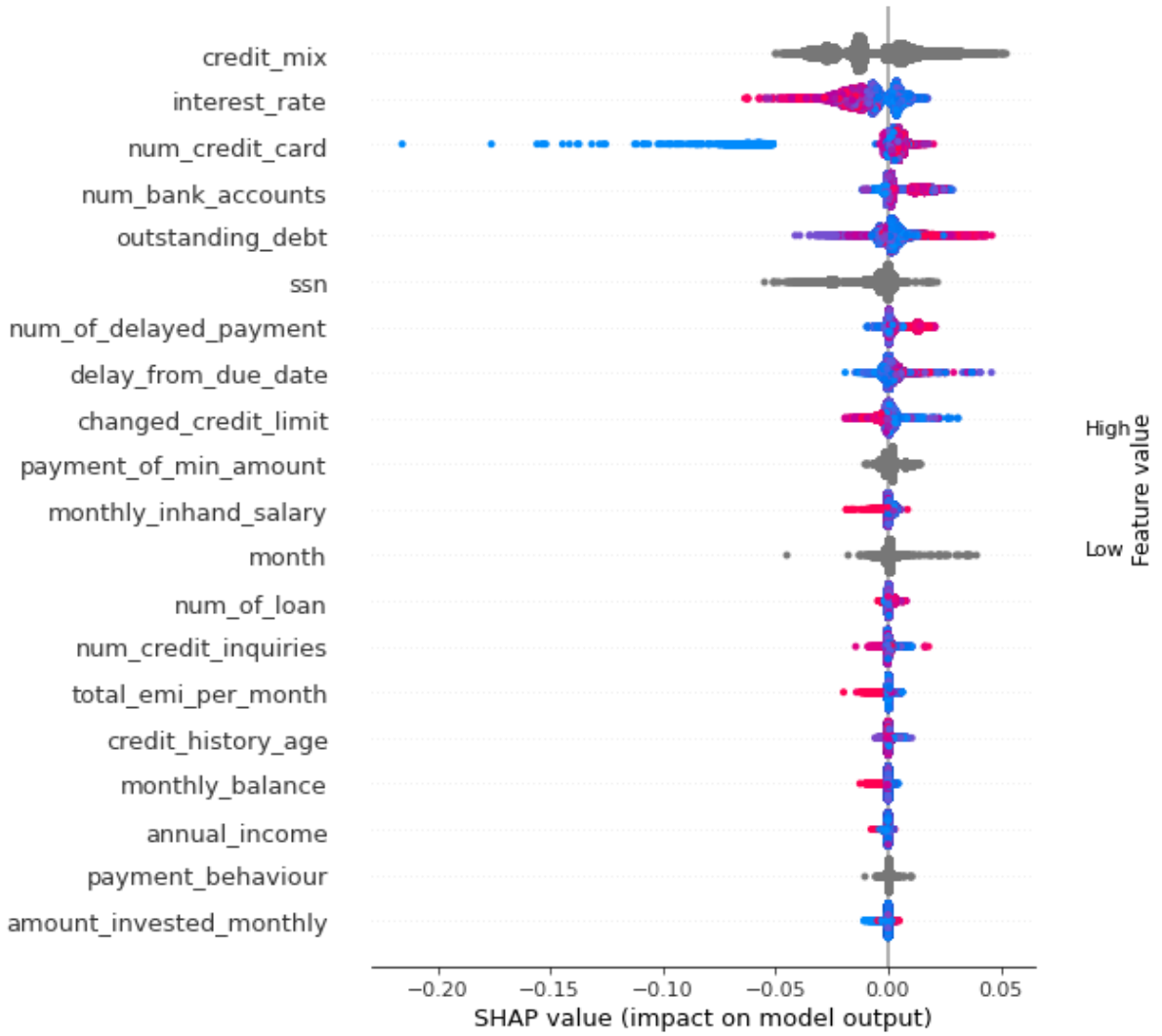
- For Good Class



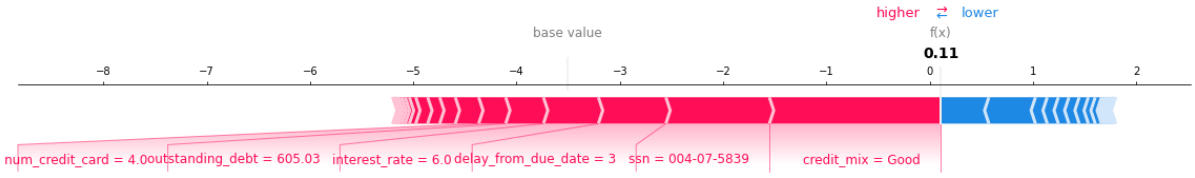
- For Standard Class



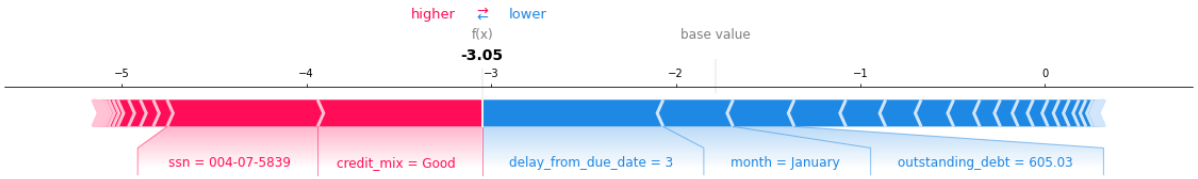
- For Poor Class



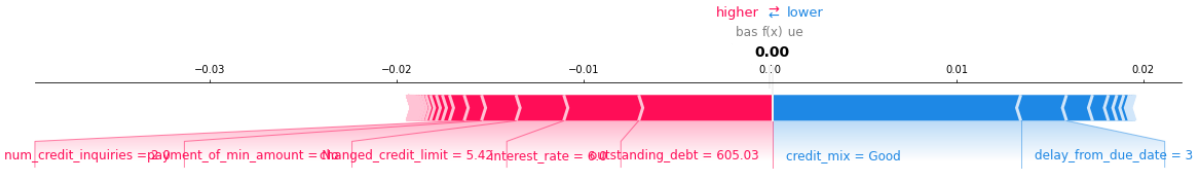
- Feature influences the model prediction for a single observation in each class
  - For Good Class



- For Standard Class



- For Poor Class



- Final score is 78.03