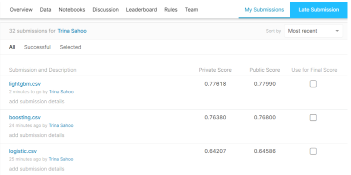
FDS Project Report

Trina Sahoo (1901254), Debodeep Banerjee (1901253)

Kaggle Score

The private leaderboard AUC score is 0.77618.



32 submissions for Trina Sahoo

Submission and Description	Private Score	Public Score	Use for Final Score
lightgbm.csv 2 minutes to go by Trina Sahoo add submission details	0.77618	0.77990	□
boosting.csv 24 minutes ago by Trina Sahoo add submission details	0.76380	0.76800	□
logistic.csv 25 minutes ago by Trina Sahoo add submission details	0.64207	0.64586	□

Salient Features

- Some features have been selected from train and test set and formed a new dataset from the selected features and merged the new dataset with the application_train dataset and application_test dataset respectively. The new application_train and application_test is obtained by applying onehot and imputer function.
- The status mentioned in bureau_balance dataset has been examined and using unstack method the status are grouped with respect to “SK_ID_BUREAU”. The unstack dataset columns are renamed and few more columns are added by calculating the size, minimum and maximum with respect to “SK_ID_BUREAU” using groupby function. The obtained new unstack dataset is merged with bureau dataset. Then onehot function is applied in the dataset bureau. Further the average is obtained with respect to “SK_ID_CURR” and “SK_ID_BUREAU” column has been deleted as “SK_ID_CURR” is enough to identify the corresponding values. This average of bureau is merged with application_train and application_test dataset.
- Onehot function is applied on previous_application dataset. The average is obtained with respect to “SK_ID_CURR” and “SK_ID_PREV” column of previous_application has been deleted as “SK_ID_CURR” is enough to identify the corresponding values. The average of previous_application is then merged with application_train and application_test dataset.
- The LabelEncoder function is used on “NAME_CONTRACT_STATUS” of the POS_CASH_balance and credit_card_balance dataset in order to convert the string into numeric form.Then two new column is introduced name “UNIQUE_STATUS_COUNT” by calculating the size with respect to “SK_ID_CURR” and “NAME_CONTRACT_STATUS”; and “UNIQUE_STATUS_MAX” by calculating the maximum with respect to “SK_ID_CURR” and “NAME_CONTRACT_STATUS”. The average is obatined using groupby function with respect to “SK_ID_CURR” newly formed POS_CASH_balance and credit_card_balance dataset which is then merged with application_train and application_test dataset.
- The average of installment_payments dataset is obtained using groupby function with respect to “SK_ID_CURR” and merged with application_train and application_test dataset. The maximum is also obtained using groupby function with respect to “SK_ID_CURR” and is merged with application_train and application_test dataset. Similarly, the minimum is obtained and merged with application_train and application_test dataset.

We have done logistic regression, gradient boosting and finally light gbm classifier is used for prediction on the obtained application train and application test dataset. We have observed improvement in the scores where logistic regression yielded the lowest score and the light gbm yielded the highest one. The obtained scores for all the predictions are mentioned in the jupyter notebook file and the final score is mentioned above.