REPORT

(KAGGLE FIRST ROUND COMPETITION)

TEAM WANDERERS (DS1079)

Highest Kaggle Submission Score: 82.583

Best F1 Score achieved in code: 58.34

```
Train accuracy
0.8278125
Test accuracy
0.8220833333333334
F1-Score
0.47413793103448276
Classification Report
              precision    recall  f1-score   support

           0       0.85      0.94      0.89      3773
           1       0.64      0.37      0.47      1027

    accuracy                           0.82      4800
   macro avg       0.75      0.66      0.68      4800
weighted avg       0.80      0.82      0.80      4800

Confusion Matrix
[[3561  212]
 [ 642  385]]
```

Classification report of best model with All Features
(Best Kaggle Score)

```
Train accuracy
0.7867708333333333
Test accuracy
0.8095833333333333
F1-Score
0.5834092980856882
Classification Report
              precision    recall  f1-score   support

           0       0.88      0.87      0.88      3730
           1       0.57      0.60      0.58      1070

    accuracy                           0.81      4800
   macro avg       0.73      0.73      0.73      4800
weighted avg       0.81      0.81      0.81      4800

Confusion Matrix
[[3246  484]
 [ 430  640]]
```

Classification report of model with SMOTE
(Best F1 Score achieved in code)

```
Train accuracy
0.821875
Test accuracy
0.82375
F1-Score
0.4784217016029593
Classification Report
              precision    recall  f1-score   support

           0       0.85      0.95      0.89      3773
           1       0.65      0.38      0.48      1027

    accuracy                           0.82      4800
   macro avg       0.75      0.66      0.69      4800
weighted avg       0.81      0.82      0.81      4800

Confusion Matrix
[[3566  207]
 [ 639  388]]
```

Classification report of best model with 10 Features

Team Members: Jathurshan Pradeepkumar (Leader)
Mithunjha Anandakumar
Vinith Kugathasan

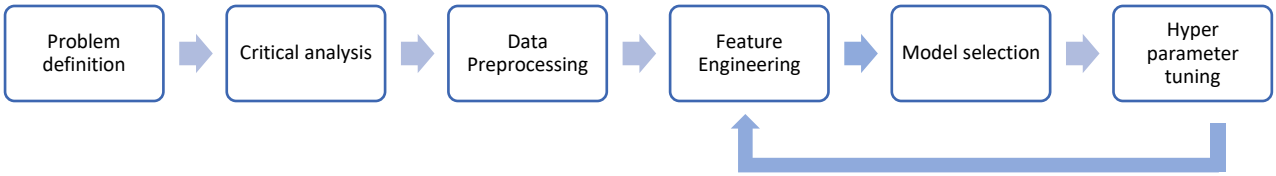
Github link:

https://github.com/Jathurshan0330/Datastorm

INTRODUCTION

The task is to use a challenging dataset consisting of Credit card default dataset to predict the credit card default for next month. The rest of the variables would be independent variables. The model should output '1' if the customer would default next month. So, this is a **Binary Classification** problem. This prediction is very important for banks to predict and prevent credit card default to prevent their bottom line.

APPROACH



First, we analysed the problem, went through literature and did brainstorming to find what features make effect on customer to default. (A default can occur when a borrower is unable to make timely payments, misses payments, or avoids/stops making payments). Then we analysed the data by processing it and plotting. From there on, we approached the problem using a feedback mechanism, consisting of Data pre-processing, Feature Engineering, Model Selection and Hyper parameter tuning.

For this business problem, it is best to have recall as evaluation metric for the model, instead of accuracy and precision. We have to do a trade-off between recall and test accuracy to suit the given problem in a practical approach. i.e. Accurate prediction of default customers is way more important than the wrong prediction of non-default customers – True positive should be high and false negative should be low as much as possible. However, for the sake of Kaggle competition, we used accuracy and F1 score as our evaluation metrics.

Following tools were used in our approach:

Pandas (data processing), Sklearn (Training the models), Imblearn (To handle unbalanced data, (SMOTE and Near-Miss)), Sklearn.ensemble (Extra Tree Classifier), Seaborn & Matplotlib (for plotting and analysing)

DATA PREPROCESSING

- First, we checked the dataset for N/A values using (isna()), but found none.
- Then the Nominal data fields in the string format such as Gender, Age, Educational status and Marital status were converted into binary variable using the one hot encoding.
- String in balance limit data field was converted into integers. i.e. 100K was converted to 100 000.

FEATURE ENGINEERING

Through our literary review and analysis, we found out that income to due ratio and credit score plays a major role in prediction of credit card default. The income of the customer is not available in the dataset. Credit score is affected by 5 main factors. Those are Payment History, Amount owed, Length of credit history, New credit and Types of credits in use.

We considered above information to identify suitable features to predict credit card default using the given dataset and features. (All the above factors are not available in the dataset)

- First, we plotted the heat map (correlation map) for all given features to find correlations and analyse the data. From the plot we found out that features which represents payment history (Pay_july, pay_aug etc) have a good positive correlation with the default

for next month. There is a good correlation between payment history-due amount for a month, and due amount-paid amount.

- Then we used all these features and trained several models to select the best model (XGBoost classifier model) with good F1 score and accuracy. Then to analyse which features affects more we plotted **feature importance** for that model. (**Submission Best Score for All features: 82.583**)

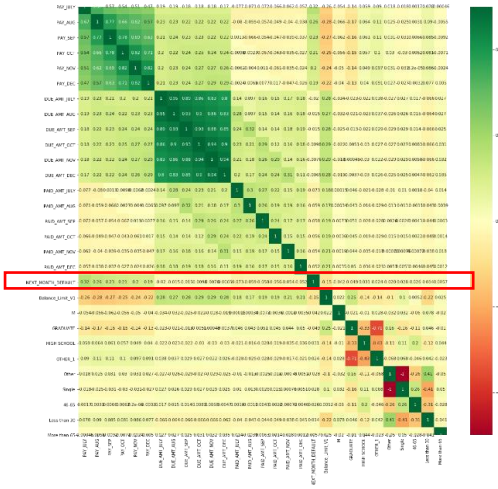


Fig.1.Heat map for all features

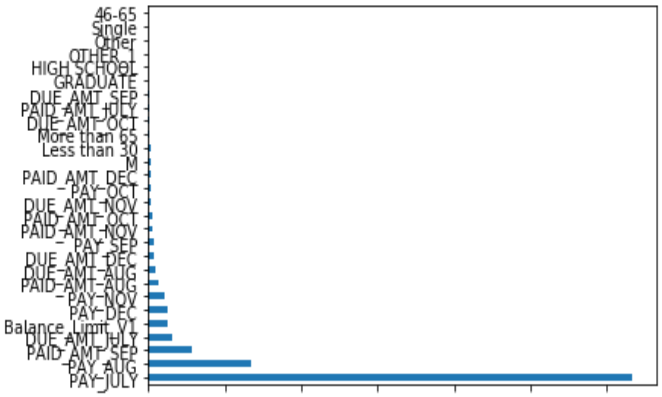


Fig.2.Feature importance for all features

- We found out that most of the features aren't making a big effect on predicting credit card default for next month. Then we removed features with low correlation and feature importance and created new features combing some of the available features.
- New features we created are:
  - **Good Pay**: Number of early and on time payments made by customer in past six months. This was used to find whether the customer is a good in terms on making payment on time or not.
  - **Mean of Paid amount/ Due amount ratio**: Ratio between paid amount and due amount for six months were found separately and then found the mean of it.This feature was created to find the ability of the customer to pay the total due amount for a month.
  - **Pay mean:** Mean of early and late payments made by the customer for past six months. This show his ability of making payment on time.
- Then we trained XGboost classifier using only 10 features and plotted **feature importance. We were able to get approximately same F1 score, accuracy and recall as previous using only 10 features.**

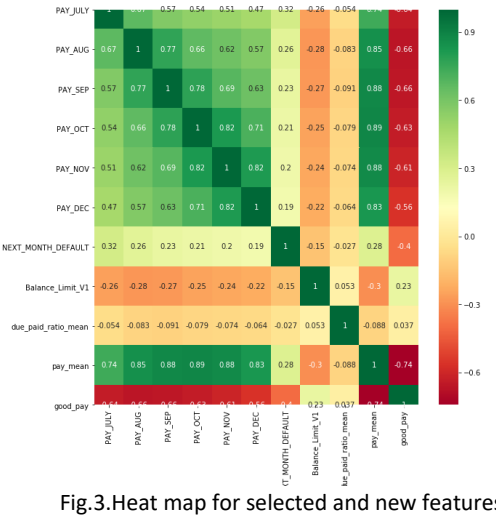


Fig.3.Heat map for selected and new features

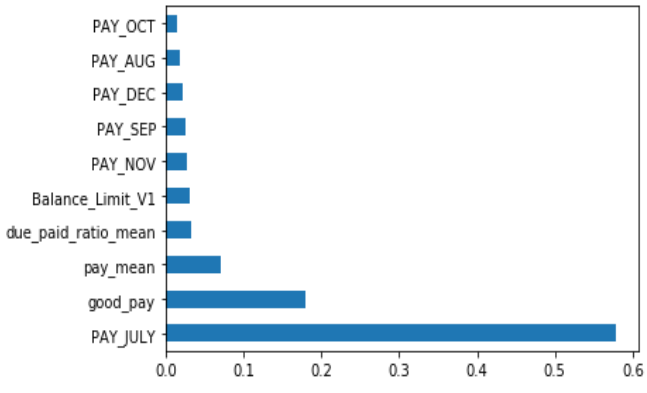


Fig.4.Feature importance for selected and new features

- Features we selected are PAY_JULY, PAY_AUG, PAY_SEP, PAY_OCT, PAY_NOV, PAY_DEC, Balance_Limit_V1, due_paid_ratio_mean, good_pay and pay_mean. **All three new features have good correlation with default and better feature importance than most. (Submission Best Score for 10 features: 82.366)**
- The data is more biased towards zero (not default), so we used SMOTE and Nearmiss techniques for oversampling to improve the recall of the model, in which SMOTE gave more prominent results. We got F1 score of **58.34 by using SMOTE**, But the accuracy of the model was reduced. This was used for one of our submission, but the accuracy was low. (Submission Score for SMOTE: 78.633).

MODEL SELECTION AND HYPER PARAMETER TUNING

After feature engineering we trained different types of classifiers, those are XGBoost classifier, Random forest classifier, Decision Tree, Logistic Regression, Support Vector Classification, Extra Tree Classifier (**ensemble approach**) and Multi-Level perceptron classifier. We chose the XGBoost Classifier and SVC as best models based on F-score, accuracy and recall. Out of these two **XGBoost classifier** gave better prediction for given dataset. Then we tuned the parameters of the model by finding the best parameters through our algorithm for this data. After, hyper parameter tuning to improve F-score we went back to data pre-processing and continued the process.

BUSINESS INSIGHTS

- According to the dataset we obtained, the lead indicators of credit card default are **PAY_JULY, PAY_AUG, PAY_SEP, PAY_OCT, PAY_NOV, PAY_DEC, Balance_Limit_V1**, which are the given features and **due_paid_ratio_mean, good_pay and pay_mean,** which are the features synthesized using the given features.
- The bank's main focus is to improve their turnover, which will get affected by the increase in credit card default. Thus, the model they should employ should focus on accurately detecting the default customers (high recall) than the accurate prediction of non-default customers (low precision). This can be achieved with the aid of the above-mentioned lead indicators hyper tuned for higher recall. Such model will help to shortlist the potentially defaultable customers (which also includes some non-default customers) and deeper insight on those customers will enable the bank to pick out the real default customers.
- The deeper insight on shortlisted potentially defaultable customers can be achieved by gathering their information related to the total amount owed, new credit obtained, other obtained loans, monthly income, frequency of the income (annual / monthly / contractual) and other banking services obtained (savings account, fixed deposit balance).
- As the given data set only contains the past payment and due history as key identifiers, the model built from that is not reliable for predicting new credit card customers as thy have no past payment and due history. Therefore, additional features such as other loans they have obtained, assets they possess, source and amount of income, bank balance have to be collected in order to evaluate the new credit card customers.
- Credit card default can be reduced among the existing customers by identifying them using the past payment history and providing promotions if they pay on time. The strategy the bank can impose on new customers is keep the maximum amount that can be used at a fixed minimum and depending on the payment history increase it so that loss incurred by default credit cards will reduce.