Loan Repayment Challenge:
Loan Default Prediction Model

Author: Woon Tian Ruen

Table of Contents

Introduction

- One of the company's main products is personal loan
- Target market is millions of American consumers who are underserved by traditional banks
- Hence, it is important that the company can assess the risk of loan applicants as accurate as possible
- With high accuracy in assessing loan applicants, it will enable the company to:
  - ➢ Better price customers
  - ➢ Reduce risk of loan losses
  - ➢ Improve decision making process
  - ➢ Increase efficiency in risk management for the loan portfolio

Objectives

- Clean and transform raw data to ensure that it is accurate, complete and consistent for machine learning model training
- Analyze the data through data exploration and visualization to identify the most important features that affect the quality of a loan
- Develop binary classification models to predict the quality of a loan application and assist in loan portfolio and risk management

Methods

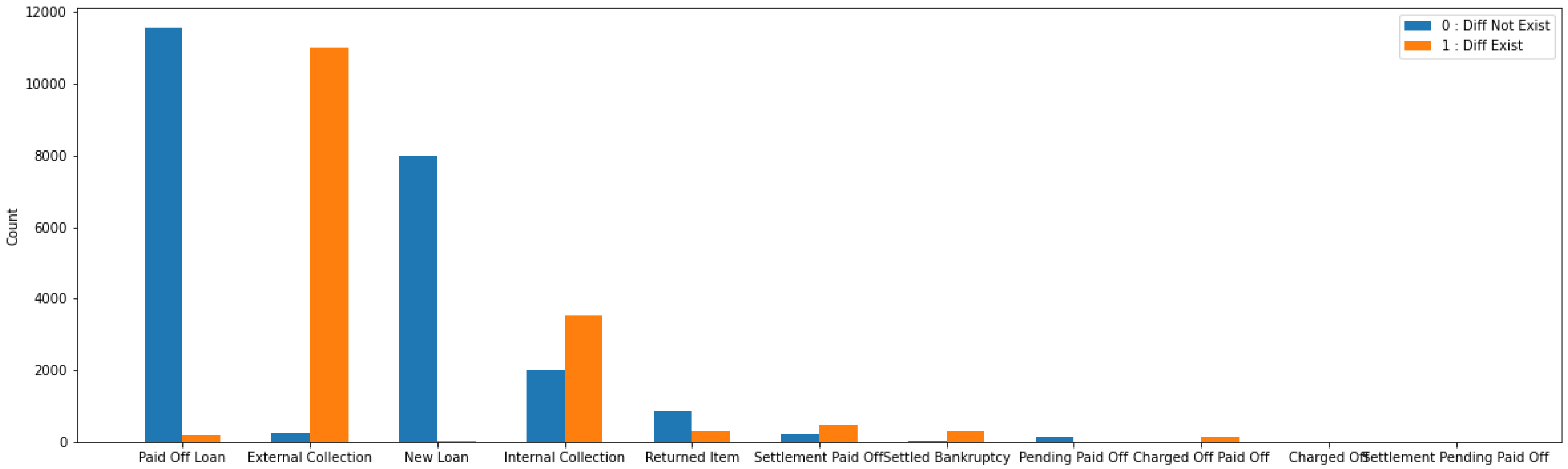
1. Transform 'Payment' table
2. Merge 'Payment' table and 'Loan' table by Loan ID and clean it
3. Classify the loan quality by definition on loan quality
4. Merge the earlier merged table with 'Clarity Underwriting Variables' table by Clarity Fraud ID and clean it
5. Conduct Exploratory Data Analysis
6. Engineer features
7. Develop Machine Learning models

Definition of Quality of Loan

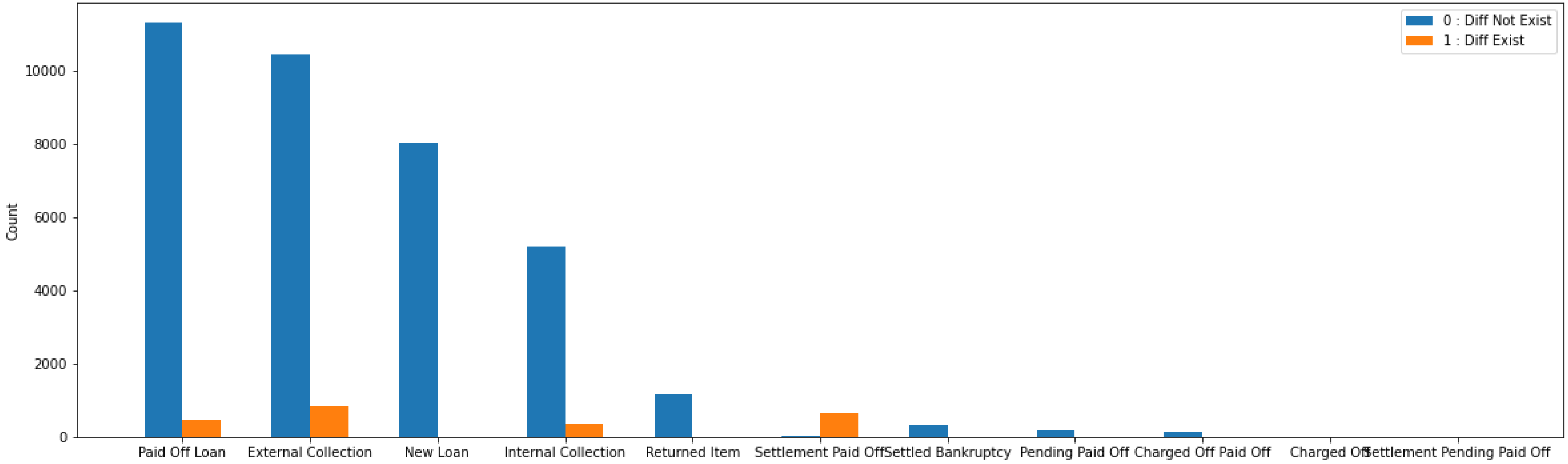
3 Criteria that determine the Quality of Loan to be Good:

1. Loan Status:
    i. 'Paid Off Loan'
    ii. 'New Loan'
    iii. 'Returned Item'
    iv. 'Pending Paid Off'
2. Difference between principal repayment amount and loan amount is 0.
    - 'diffExist' = 0
3. Loan that does not has a custom made collection plan.
    - 'isCollection' = 0

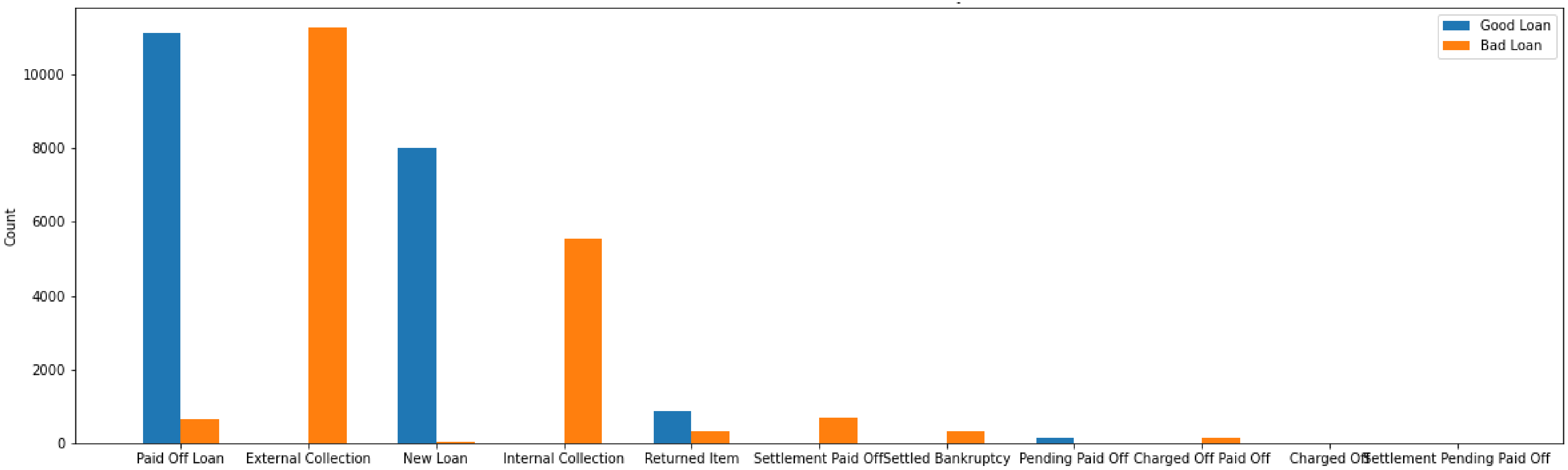
Loan Status in terms of 'diffExist'



Loan Status in terms of 'isCollection'



Loan Status in terms of 'loanQuality'



Note: 51.5% of the loans are classified as Good Loan while the remaining 48.5% of the loans are classified as Bad Loan

Data Cleaning and Transformation

Transforming the ‘Payment’ table

- Transform the ‘Payment’ table such that each ‘loanId’ only has 1 row of data instead of multiple lines

- Will be used to merge with the ‘Loan’ table later on

- New table consists of 46 columns compared to old table of 9 columns

- Methods on transforming ‘Payment’ table:
    i. loanId – unique ID from the table
    ii. installmentIndex – number of installment made
    iii. isCollection – 1 if True exist for the ID, else 0
    iv. paymentDate – first loan payment date
    v. principal – sum the amount for ‘paymentStatus’ is ‘Checked’, ‘None’, ‘Pending’
    vi. fees – sum the amount for ‘paymentStatus’ is ‘Checked’, ‘None’, ‘Pending’
    vii. paymentAmount – sum the amount for ‘paymentStatus’ is ‘Checked’, ‘None’, ‘Pending’
    viii. paymentStatus – pivot table
    ix. paymentStatusCode – pivot table

Replacing Null Values for 'Loan' table

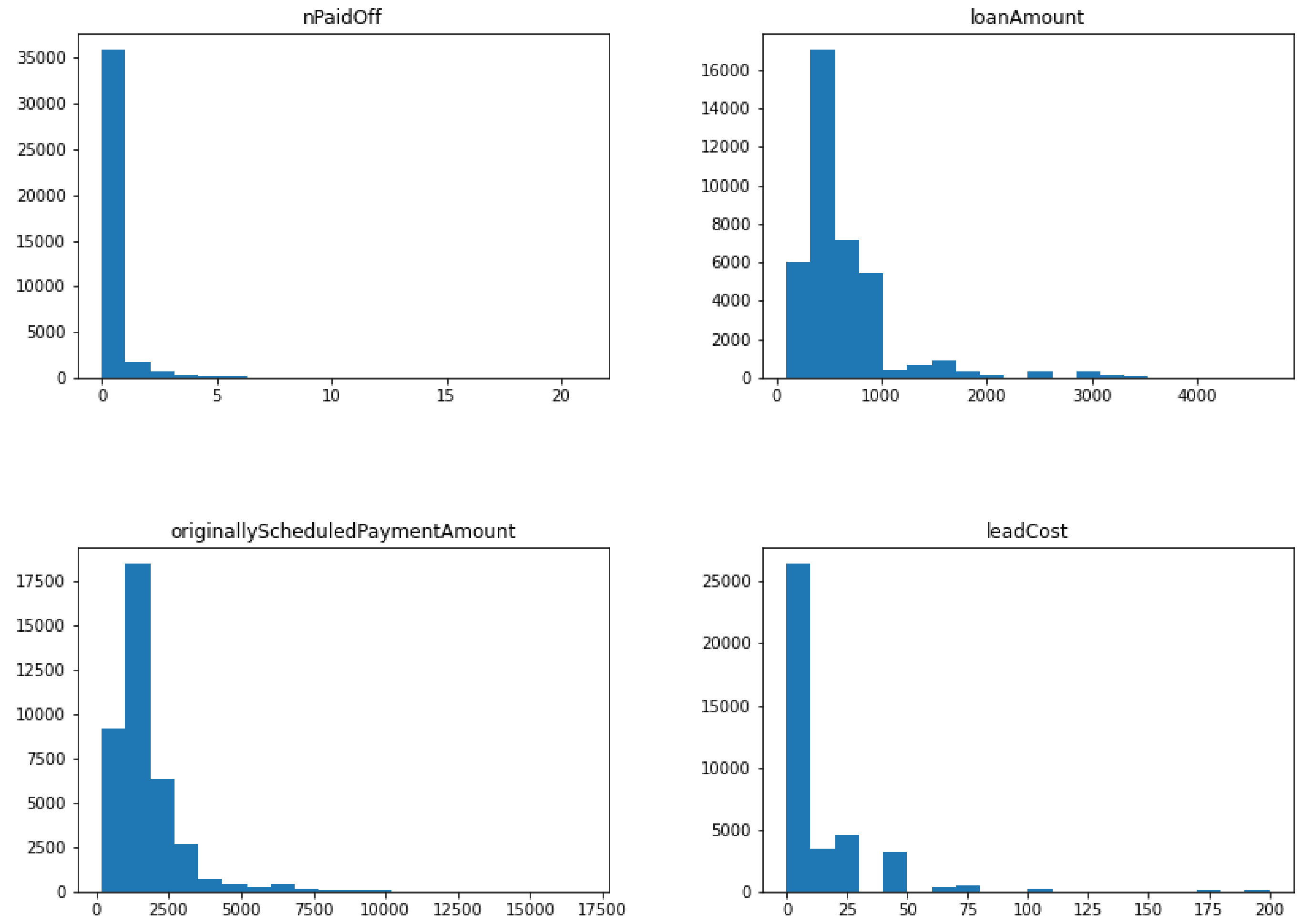
- 4 columns with null values
- Methods to replace the null values in each column:
  i. 'originatedDate' – last previous valid value
  ii. 'nPaidOff' – mode value
  iii. 'fpStatus' – mode value
  iv. 'clarityFraudId' – won't be replaced

```python
# check which columns contain null value
nullCols = lp.isnull().sum()
nullCols[nullCols > 0]
```

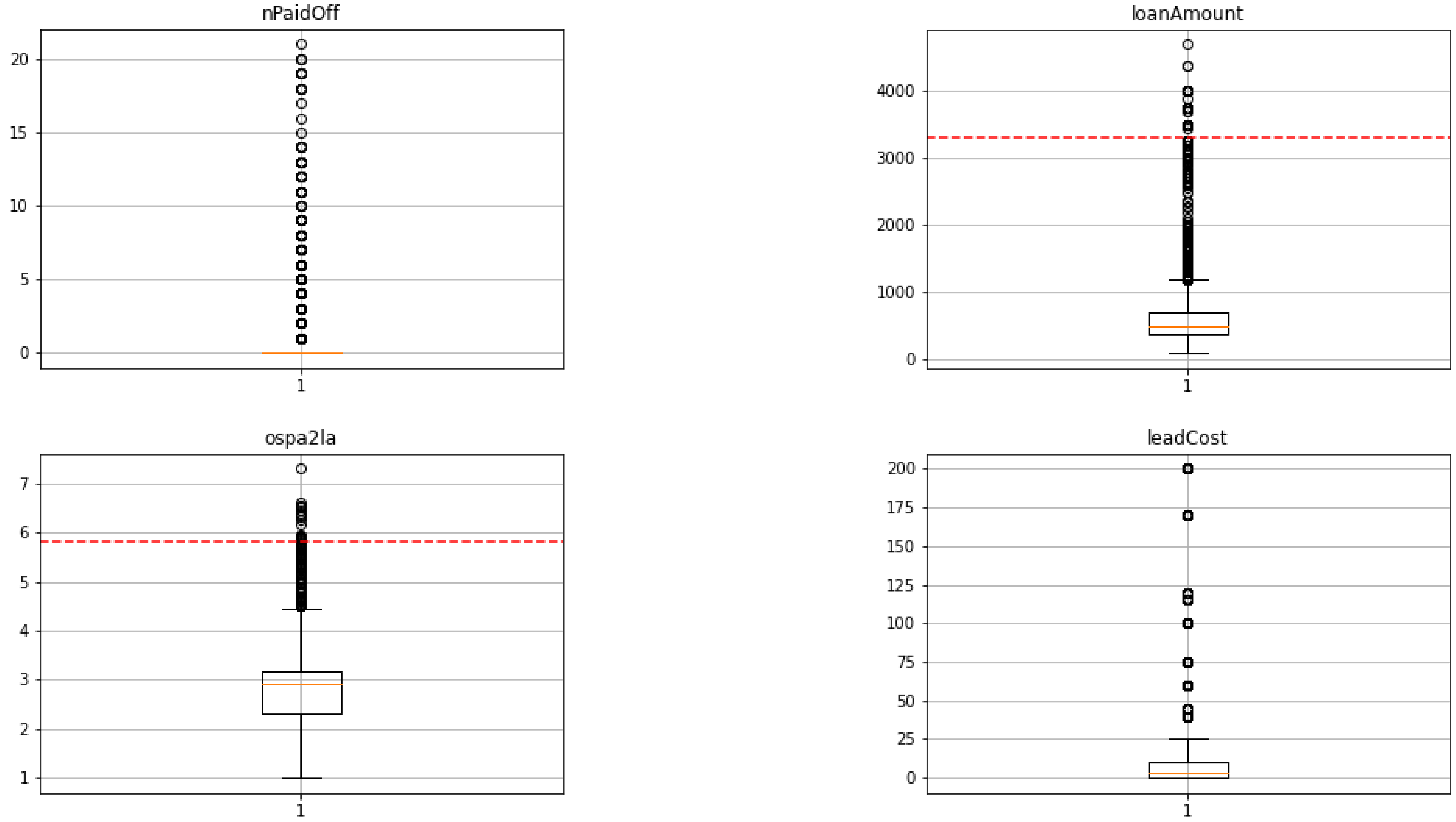
```
originatedDate      19
nPaidOff            21
fpStatus           404
clarityFraudId    6640
dtype: int64
```

Replacing Outliers for ‘Loan’ table

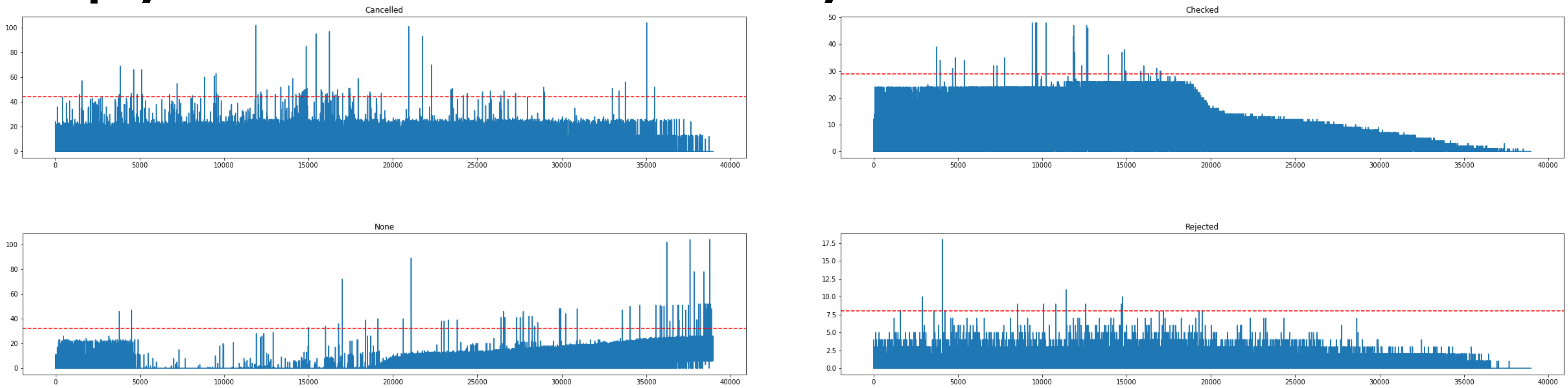
- Numerical data are positively skewed
- Methods to replace outliers for each column:
  1. ‘nPaidOff’ – reasonable large value
  2. ‘loanAmount’ – remove
  3. ‘originallyScheduledPayment Amount’ – remove
  4. ‘leadCost’ – median



Box Plot


nPaidOff
20
15
10
5
0
1
loanAmount
4000
3000
2000
1000
0
1
ospa2la
7
6
5
4
3
2
1
1
leadCost
200
175
150
125
100
75
50
25
0
1

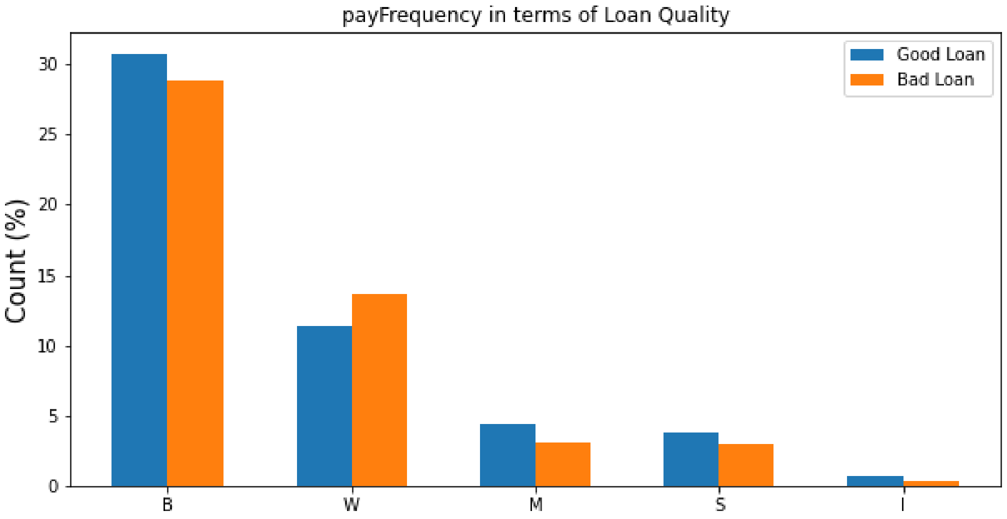
‘paymentStatus’ column from ‘Payment’ table

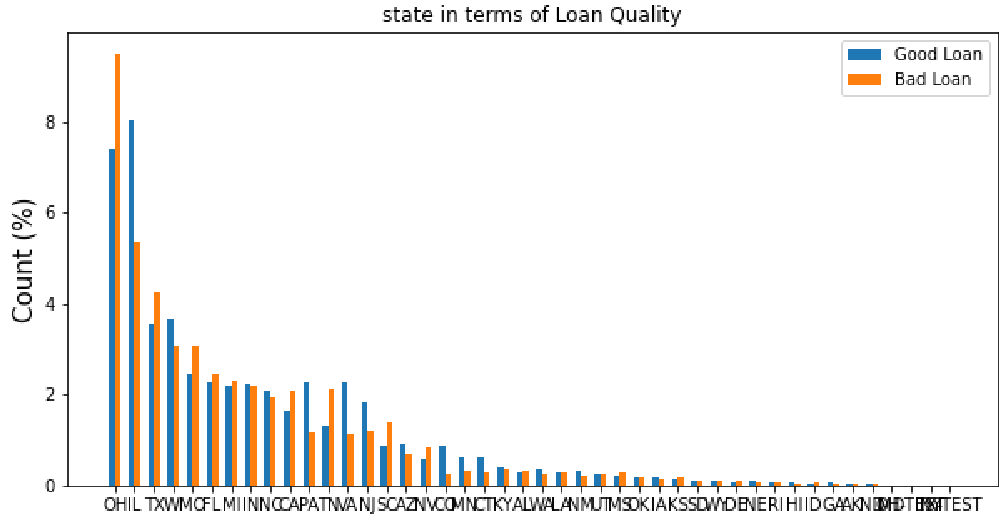


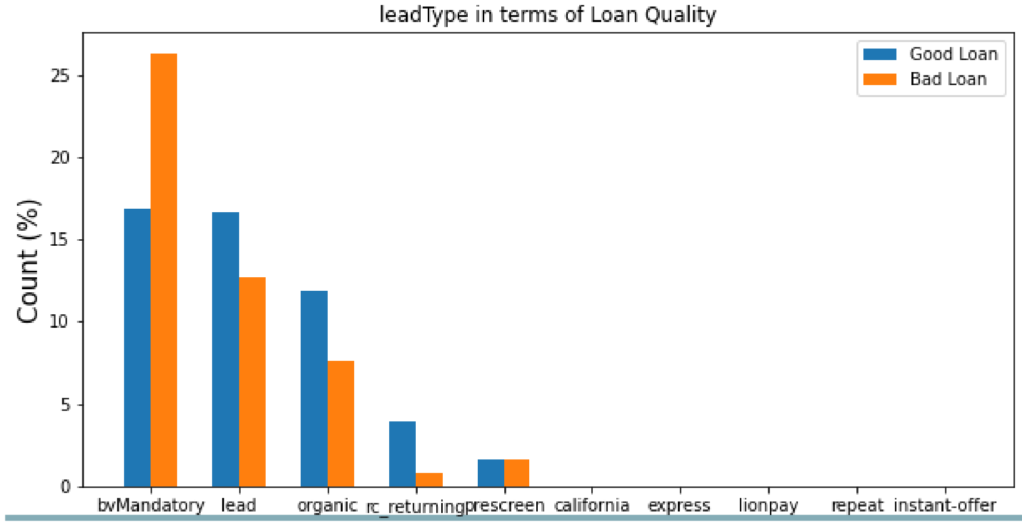
- Interpret the above graphs as time series since the loans are recorded according to payment date
- Number of ‘Checked’ gradually decrease while number of ‘None’ gradually increase
- May be due to system error
- To address this issue, we remove outliers in these 4 columns
- Definition of outliers:
    - ‘None’ – 3 x std dev + mean (to be more conservative)
    - Others – 3 x IQR + 3rd Quartile

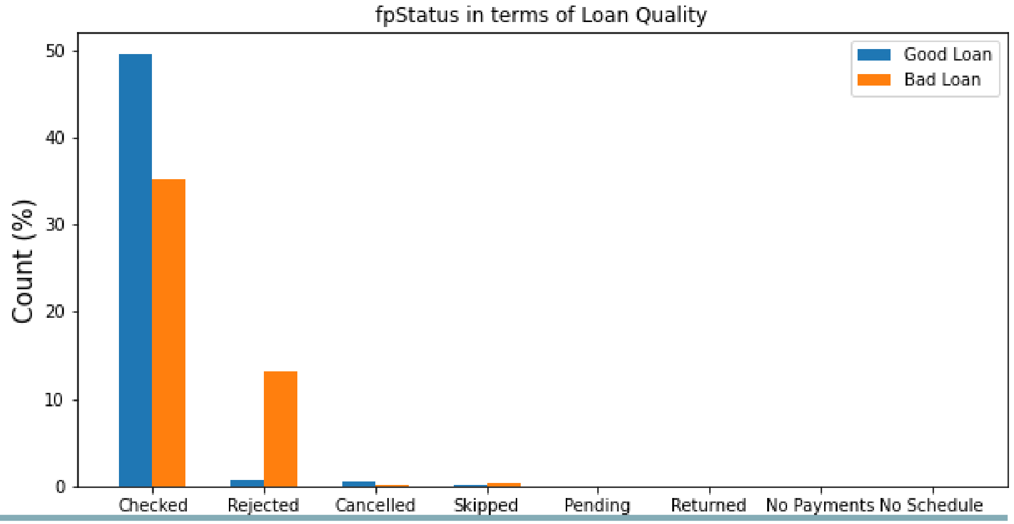
Exploratory Data Analysis (EDA)

Categorical Data from ‘Loan’ table

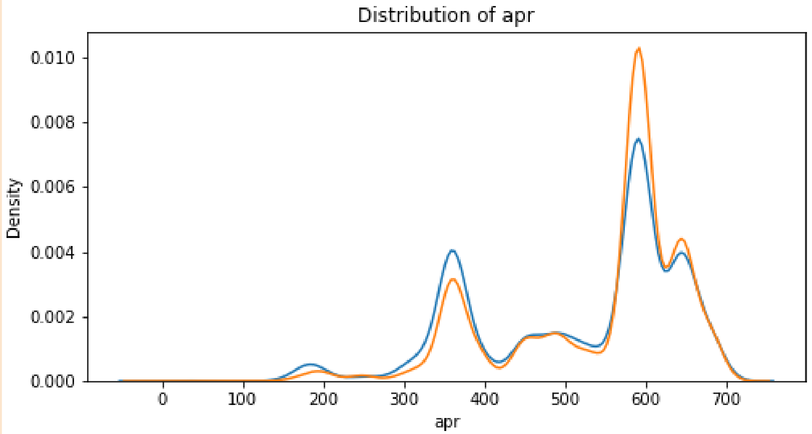


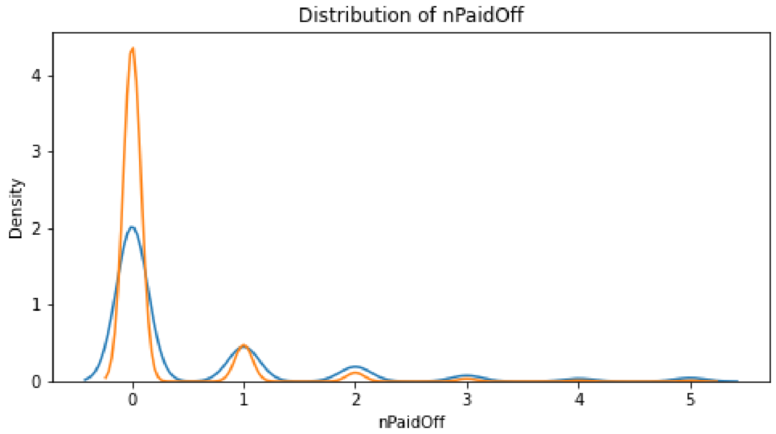


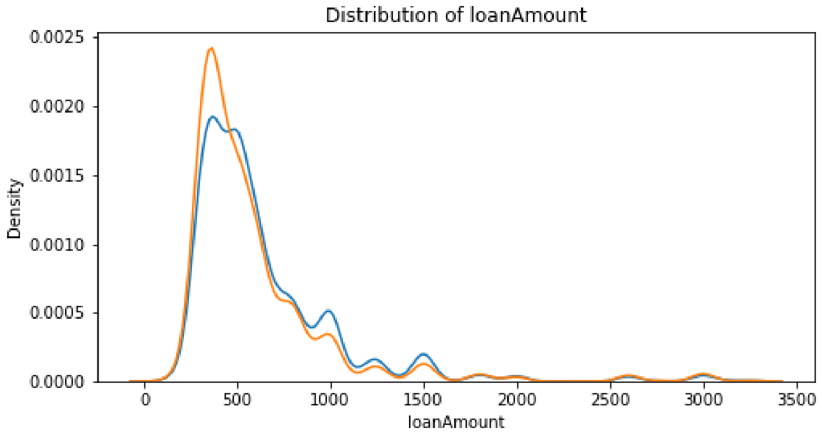


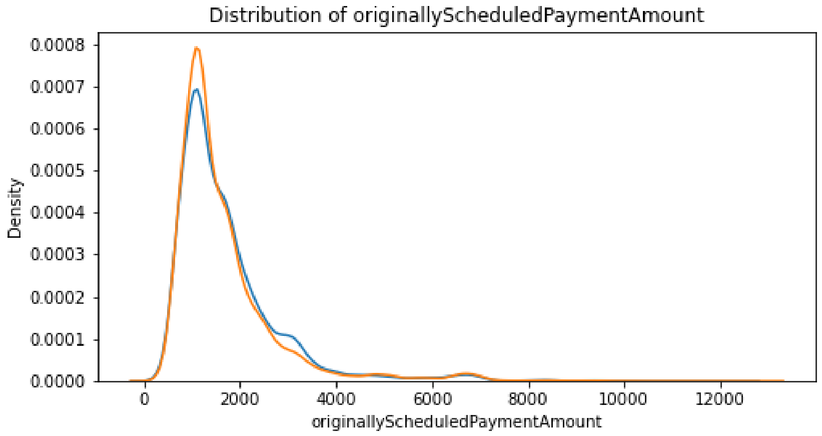


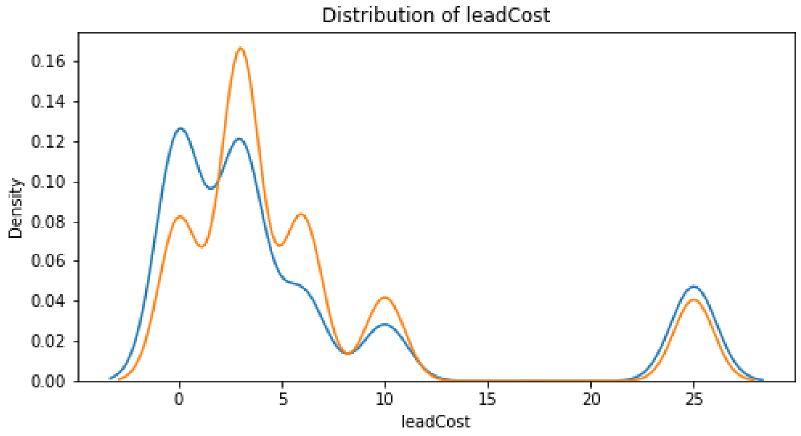
Numerical Data from ‘Loan’ table











Top 5 Most Positively & Negatively Correlated Features w.r.t Loan Quality

Positively Correlated:

1. Rejected – 0.72
2. R01 – 0.55
3. Cancelled – 0.44
4. R02 – 0.27
5. R08 – 0.25

Negatively Correlated:

1. principalPayment – -0.47
2. feesPayment – -0.32
3. Checked – -0.29
4. Pending – -0.27
5. nPaidOff – -0.19

Feature Engineering

Feature Engineering

Convert Date to Number

- Convert the year, month and day in Date to numerical data
- Date column includes:
  - First Payment Date
  - Origination Date
  - Application Date
- Add 2 new columns which record the time taken between:
  - Application and Origination
  - Origination and First Payment

Encode Categorical Data

- Convert categorical data into numerical data
- Features with a lot of different categorical values are one hot encoded while features with few categorical values are label encoded

Correlation of Features Engineered w.r.t Loan Quality

- Application, Origination & First Payment year are weakly negatively correlated with Loan Quality, with correlation coefficient of:
  - Application Year – -0.2
  - Origination Year – -0.2
  - First Payment Year – -0.2
- First Payment Status of ‘Checked’ and ‘Rejected’ are weakly correlated with Loan Quality, with correlation coefficient of:
  - Checked_fpStatus – -0.35 (negative correlation)
  - Rejected_fpStatus – 0.38 (positive correlation)

Model Development

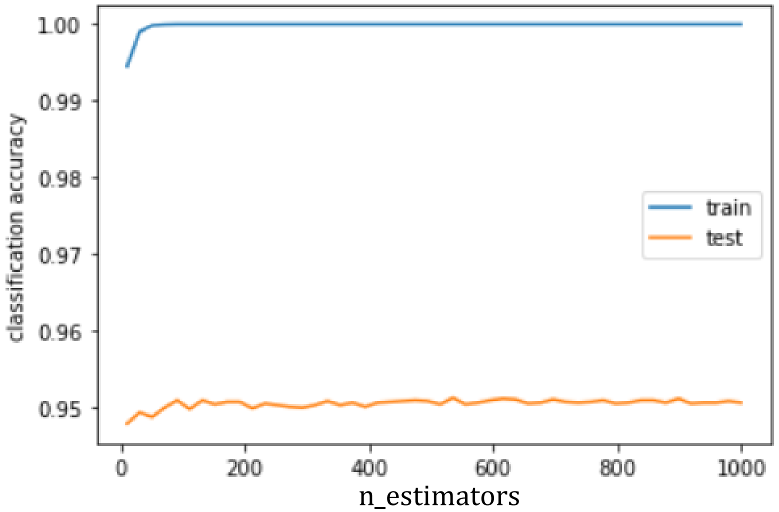
Machine Learning Model

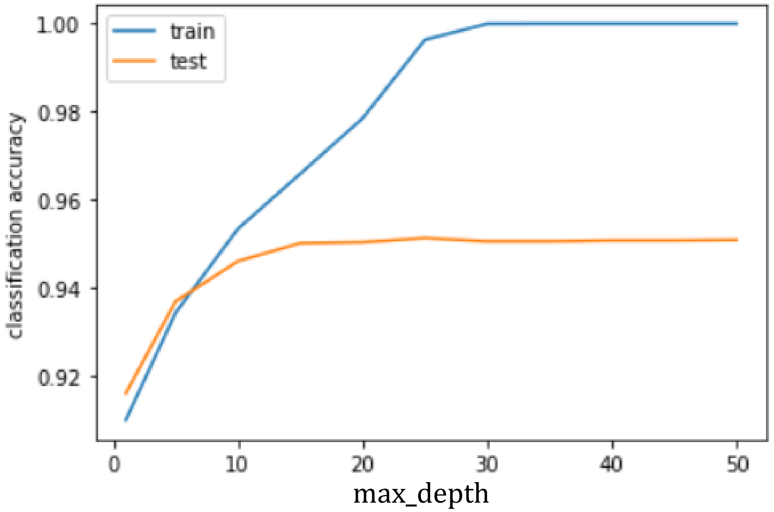
- Train 3 machine learning models:
    1. Logistic Regression
    2. Support Vector Machine (SVM)
    3. Random Forest

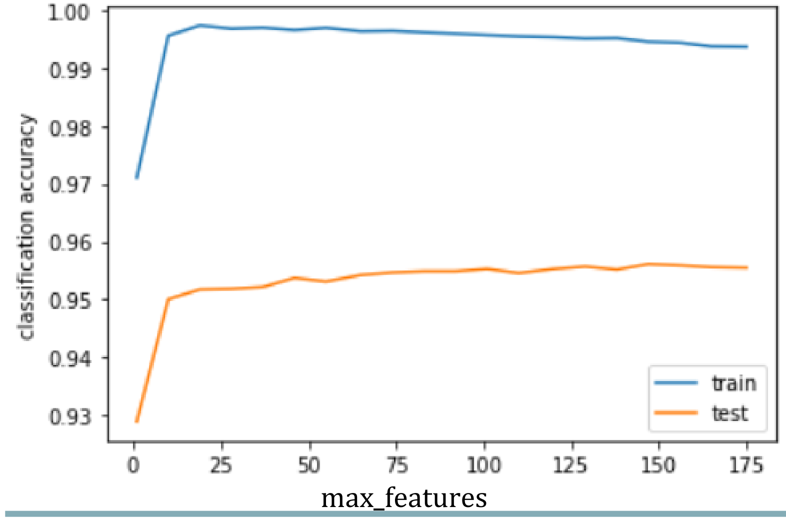
Reason & Result for using the abovementioned machine learning models

1. Logistic Regression:
    - Simple model
    - Use as a base case
    - Accuracy score: 0.93; Recall: 0.88
2. SVM:
    - Able to handle complex, nonlinear classification very well through kernel method
    - Accurate and tends to not overfit
    - Accuracy score: 0.95; Recall: 0.91
3. Random Forest:
    - Able to handle large datasets with large number of features
    - Robust to outliers and noise
    - Accuracy score: 0.96; Recall: 0.92

Hyperparameter Tuning







<u>Hyperparameter values</u>

n_estimators : 500

max_depth : 30

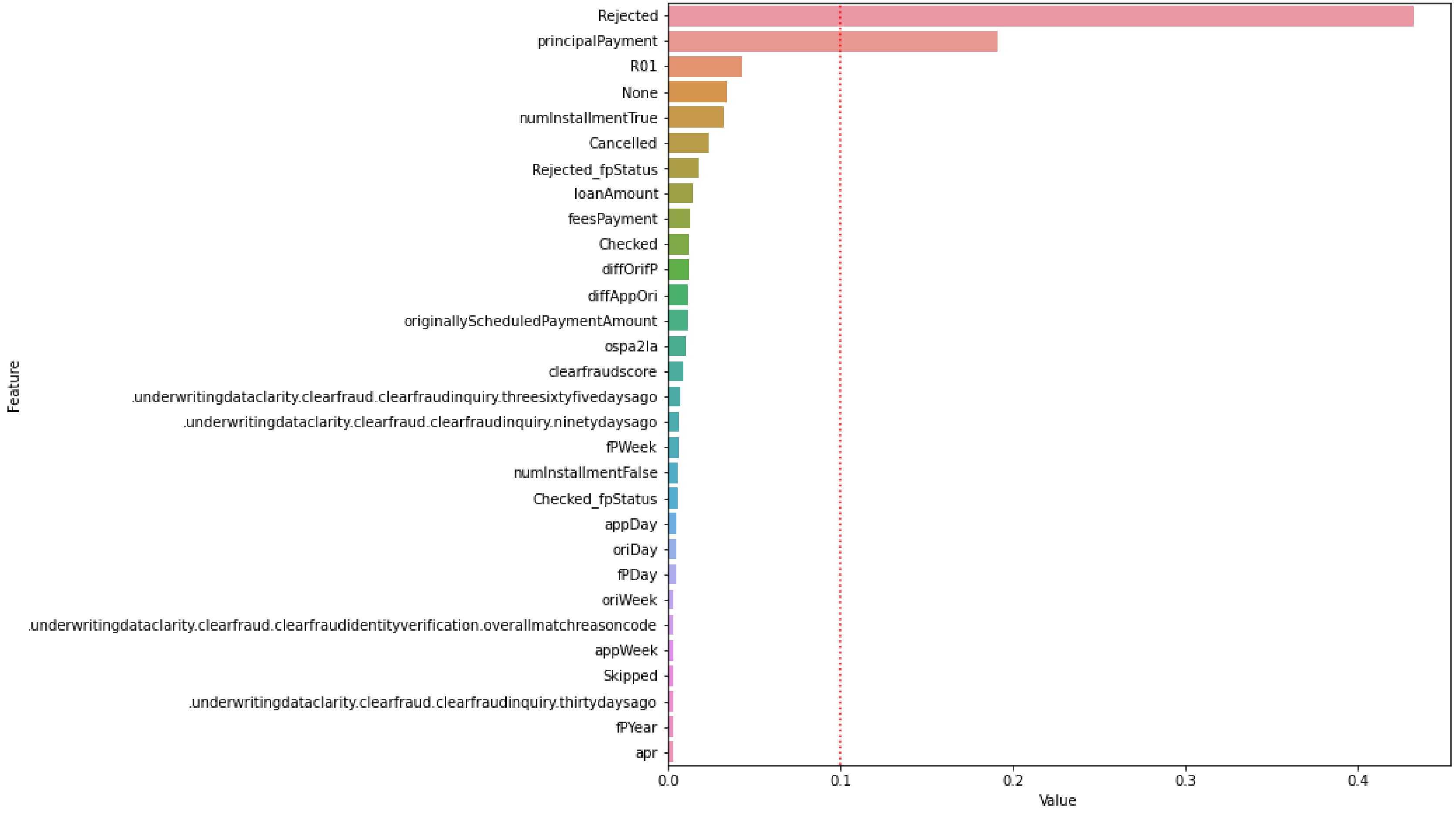
max_features : 100

criterion : 'entropy'

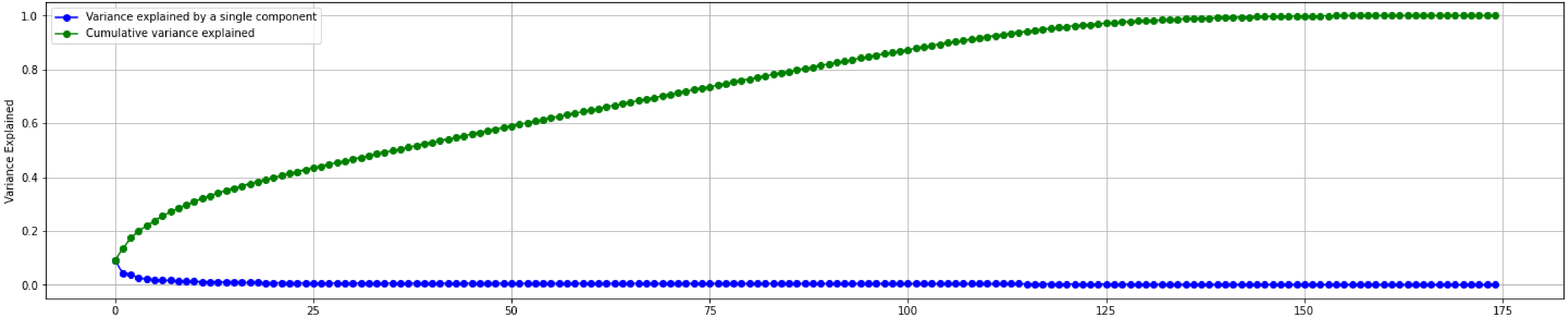
```
rfc = RandomForestClassifier(random_state = 42,
                             n_estimators = 500,
                             max_depth =25,
                             max_features = 100,
                             bootstrap = True)
rfc_gs = GridSearchCV(estimator = rfc,
                      param_grid = {'criterion' : ['gini', 'entropy', 'log_loss']})

rfc_gs.fit(X_train,y_train)
```

```
The best kernel: {'criterion': 'entropy'}
The best score from the kernel: 0.9543132315065007
```

Feature Importance



Principal Component Analysis (PCA)



Conclusion

- Random Forest produced highest accuracy. However, it is important to conduct hyperparameter tuning periodically.
- Cross validation was not carried out due to computing power limitations
- More advanced ML or AI models such as Deep Neural Network and Reinforcement Learning can be implemented
- Discuss on data quality:
  - Data from 'Payment' table:
    - Collected after loan was originated
    - Not useful during loan application process
    - Helpful in loan portfolio and risk management
  - Data from 'Clarity Underwriting Variables':
    - Provided by a third party data provider
    - Not all applicants have a Clarity Fraud ID
    - Give another level of scrutiny
  - Should request more information from applicants to improve data quality
    - E.g. Income Statement, Employment Status, Tax Code, Purpose of Loan, Debt Amount, Does the applicant has any dependants, etc.