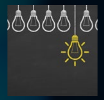
CAPSTONE DATA SCIENCE PROJECT: HOW TO USE BANKING TRANSACTIONS DATA TO DO LIQUIDITY FORECASTING, CLIENT SEGMENTATION AND LOAN DEFAULT PREDICTION

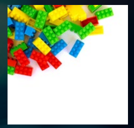
ENG SOON, WONG

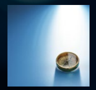
11 September 2021

CONTENTS


Problem Statement


Data Cleaning and Exploratory Data Analysis


Modelling


Conclusions & Recommendations

PROBLEM STATEMENT

- We are the **Data Scientists in AI Lab in ABC Bank** that specialises in exploratory data analysis and modelling for the bank.

- We will be using **banking transactions data** to give insights on **liquidity forecasting**, **customer segmentation** and **loan default prediction**.

- On **liquidity forecasting**, we will be using banking transaction data to **forecast the amount of liquidity which the bank needs to hold to satisfy the withdrawals** required by its borrowers.
- We will be **measure accuracy** of the **SARIMA** time-based modelling via the **mean squared error**.

- On **Customer Segmentation**, we aim to **generate leads and propose recommendations** to **increase sales and revenue** for the bank.
- We will be using **K-means clustering** to segmentise the customers and using the **silhouette score to obtain the optimal number of clusters**.

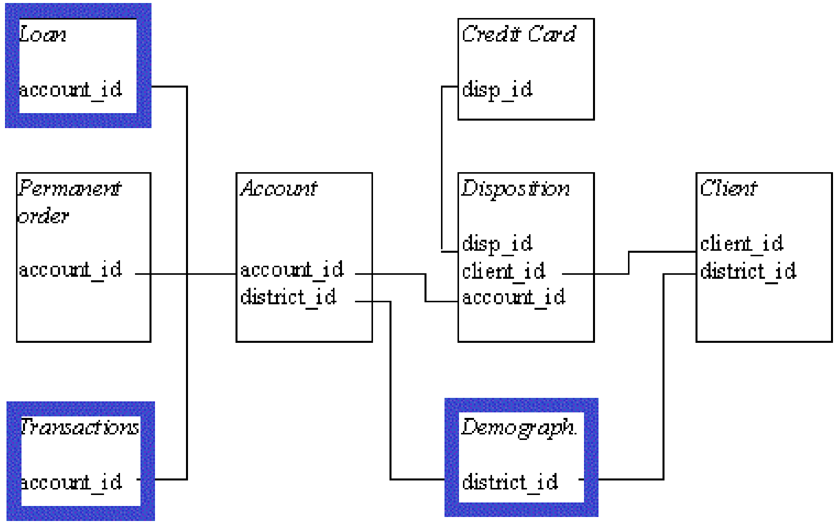
- On **Loan Default Prediction**, we will be **using banking transactions data**, together with **client demographics** data, **to enrich the loans data**.
- We will be using **various classification models** to do the **prediction** modelling and using **Accuracy metrics and ROC AUC to score** the models.

- The analysis and findings will provide valuable insights for Senior Bank Management to aid them in their decision making processes.

DATA CLEANING & EXPLORATORY DATA ANALYSIS

1. Data Collection:

- 1999 Czech Financial Dataset - Real Anonymized Transactions by Liz Petrocelli
- The dataset is a collection of financial information from a **Czech bank** that deals with over **5,300 bank clients** with approximately **1,000,000 transactions** from **1993 to 1998**.
- Additionally, the bank represented in the dataset has extended close to 700 loans and issued nearly 900 credit cards, all of which are represented in the data.



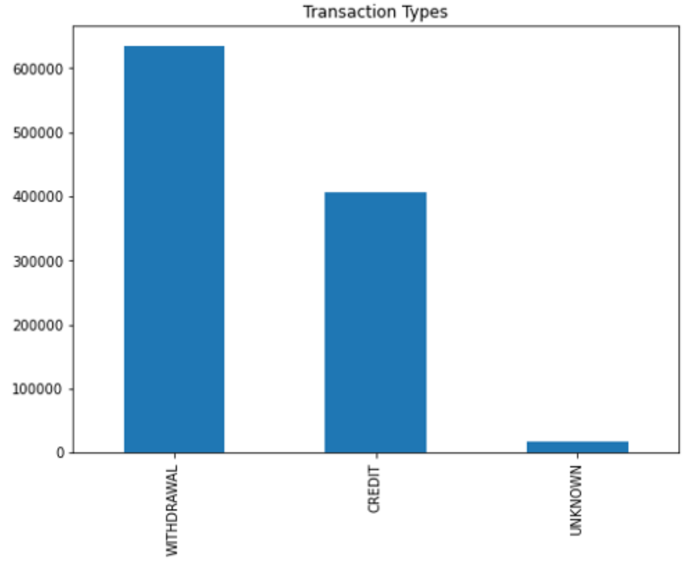
2. Data Cleaning

- Conversion of data column fields from Czechoslovak language to English
- Dates correction
- Separation of Birth Number into Birthday and Gender

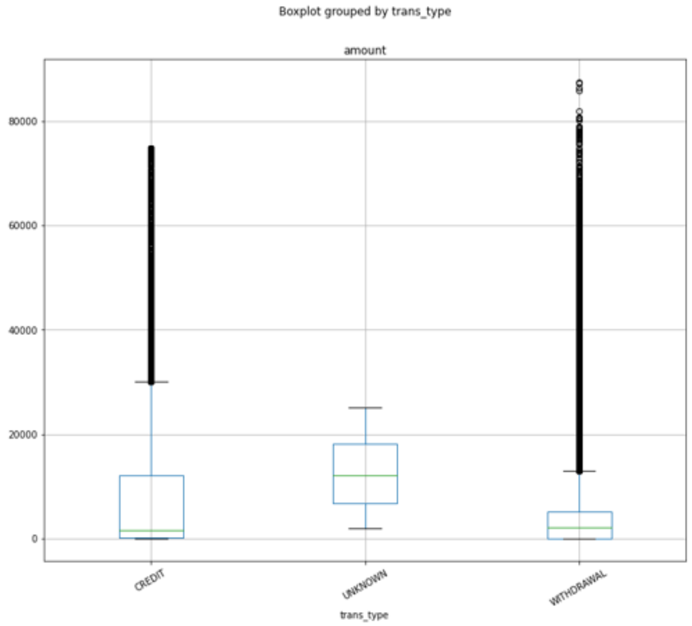
3. Exploratory Data Analysis

- On Transactions data, by Transaction Types (Withdrawal/Credit):

Number of counts of Transactions Types for Withdrawal is 200k more than Credit



In terms of Transactions Amount, we see more Withdrawals with large transactional amounts than Credit



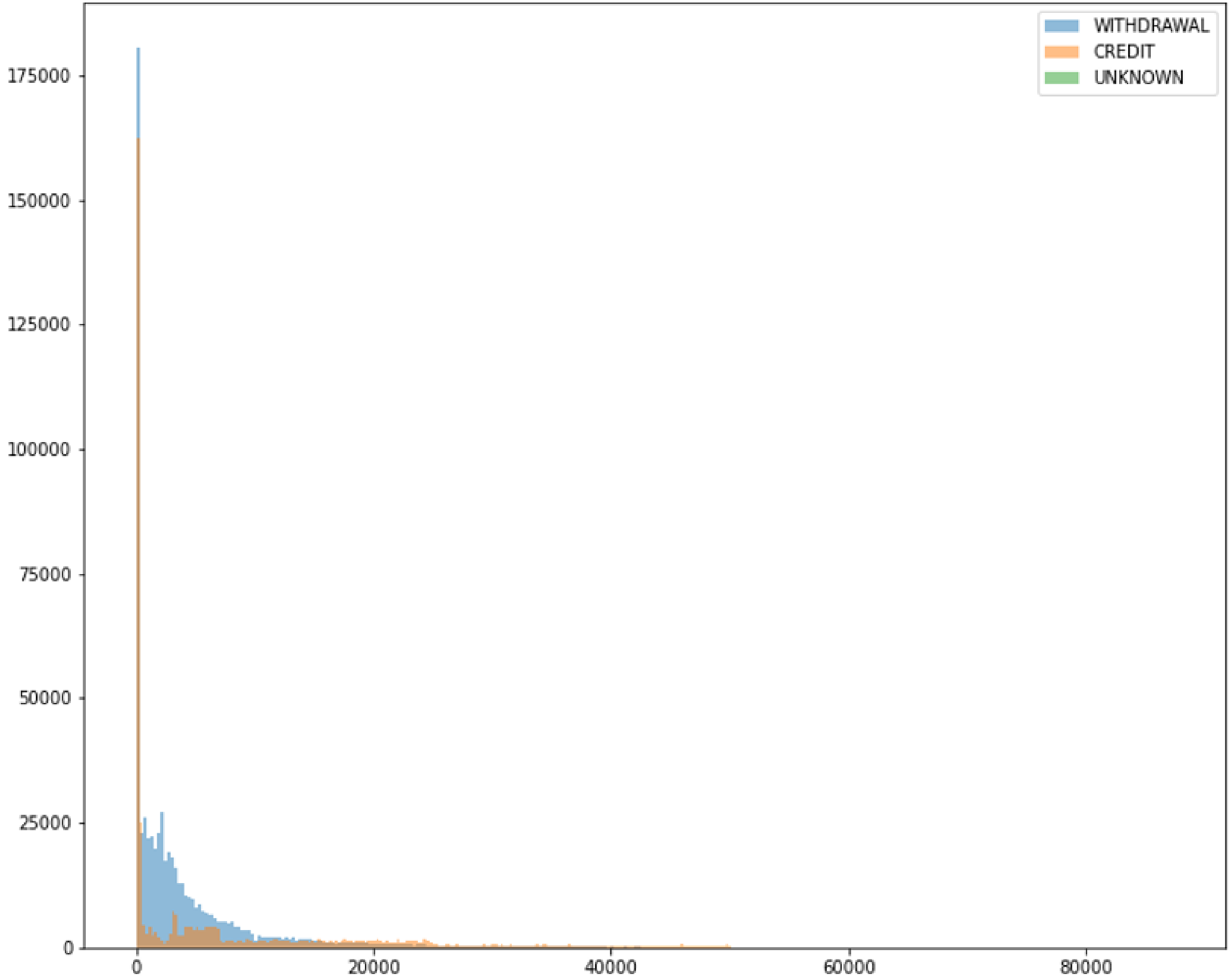
DATA CLEANING & EXPLORATORY DATA ANALYSIS

DATA CLEANING & EXPLORATORY DATA ANALYSIS

3. Exploratory Data Analysis

- On Transactions data, by Transaction Types (Withdrawal/Credit):

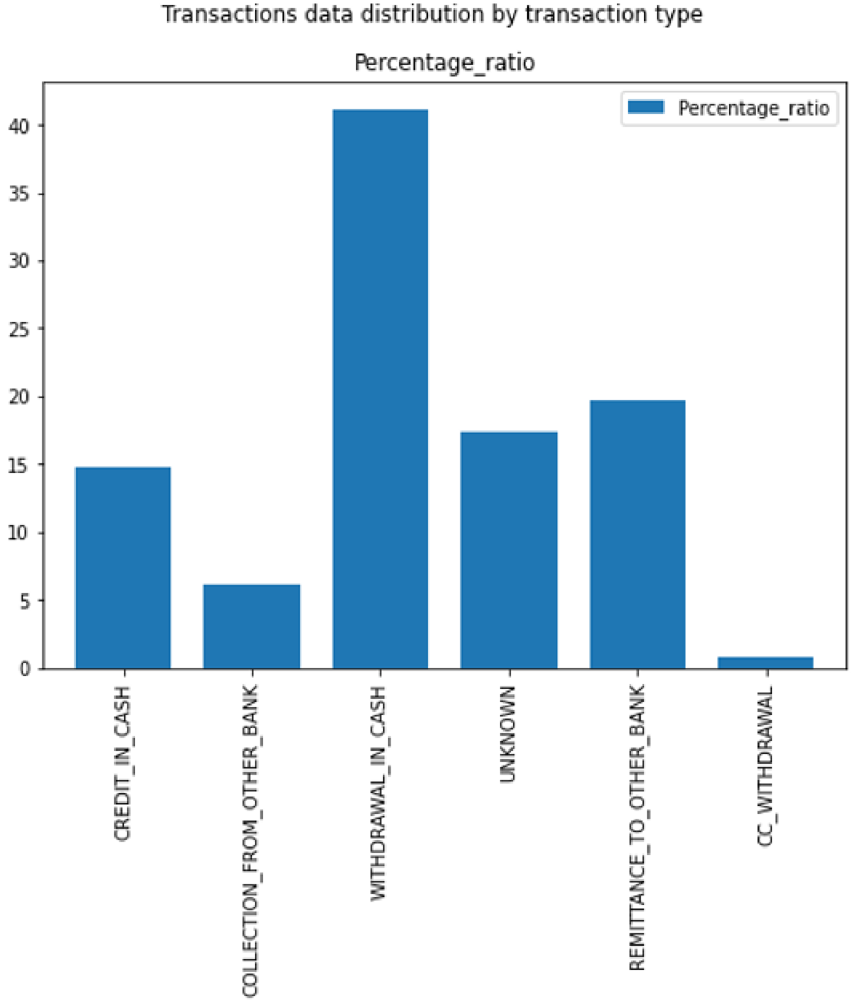
Histogram of Transaction Amounts shows Withdrawals have MORE counts of Withdrawals at almost every Transaction Amounts than Credit



3. Exploratory Data Analysis

- On Transactions data, by Transaction Operations:

Transactions Operations shows Cash Withdrawals having largest %



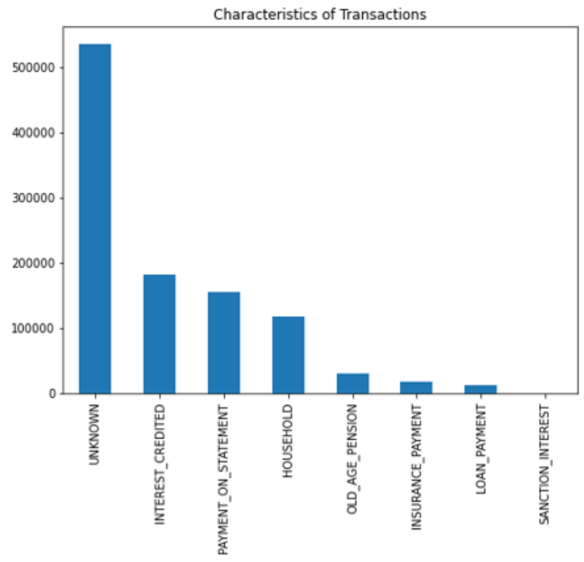
DATA CLEANING & EXPLORATORY DATA ANALYSIS

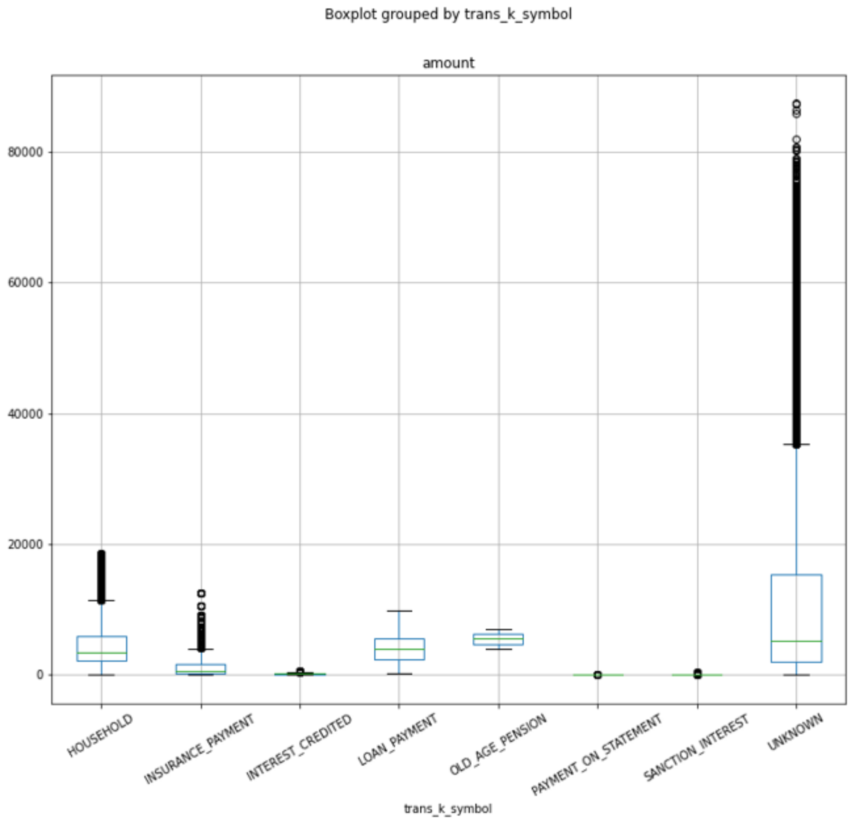
DATA CLEANING & EXPLORATORY DATA ANALYSIS

3. Exploratory Data Analysis

- On Transactions data, by Transaction Characteristics:

Transactions by Characteristics shows Unknown category are most highly skewed in their Transactions Amounts





3. Exploratory Data Analysis

- On Loans data, Loan Default is affected by big monthly loan payment, long loan duration and big loan amount

```
#correlation between loan status and monthly payments
df_loan3 = df_loan2.groupby (['loan_status_desc']).mean () ['monthly_loan_payment'].sort_values ()
```

```
df_loan3
```

```
loan_status_desc
Runing contract, OK so far                3938.535980
Contract finished, no problems            4264.137931
Runing contract, client in debt           5286.644444
Contract finised, loan was not paid       5396.258065
Name: monthly_loan_payment, dtype: float64
```

```
#correlation between loan status and duration of a loan
df_loan4 = df_loan2.groupby (['loan_status_desc']).mean () ['loan_duration'].sort_values ()
```

```
df_loan4
```

```
loan_status_desc
Contract finished, no problems         22.226601
Contract finised, loan was not paid    25.548387
Runing contract, OK so far             43.444169
Runing contract, client in debt        46.133333
Name: loan_duration, dtype: float64
```

```
#correlation between loan status and loan's amount
df_loan5 = df_loan2.groupby (['loan_status_desc']).mean () ['loan_amount'].sort_values ()
```

```
df_loan5
```

```
loan_status_desc
Contract finished, no problems          91641.458128
Contract finised, loan was not paid    140720.903226
Runing contract, OK so far             171410.352357
Runing contract, client in debt        249284.533333
Name: loan_amount, dtype: float64
```

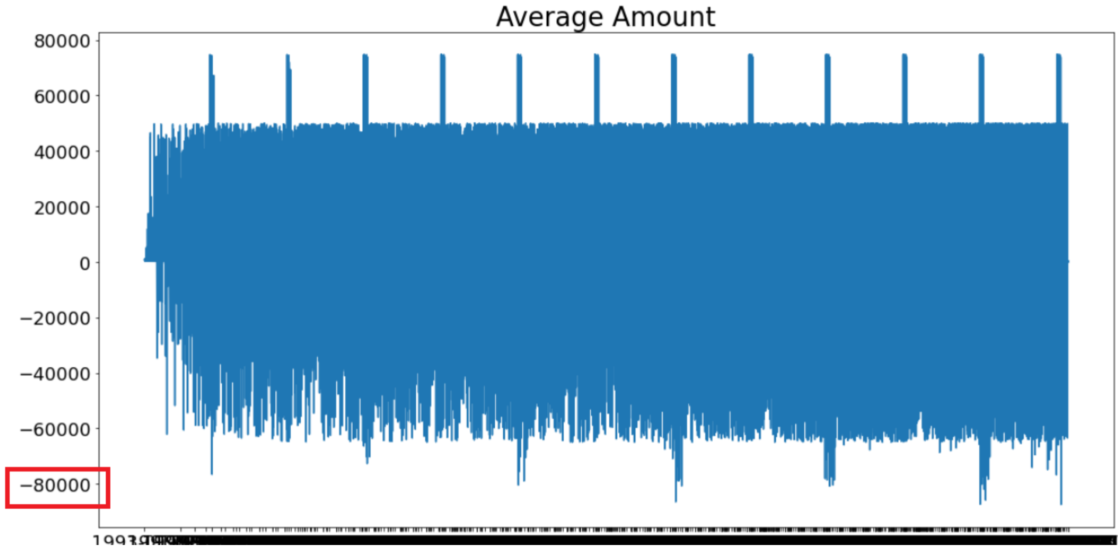
DATA CLEANING & EXPLORATORY DATA ANALYSIS

DATA CLEANING & EXPLORATORY DATA ANALYSIS

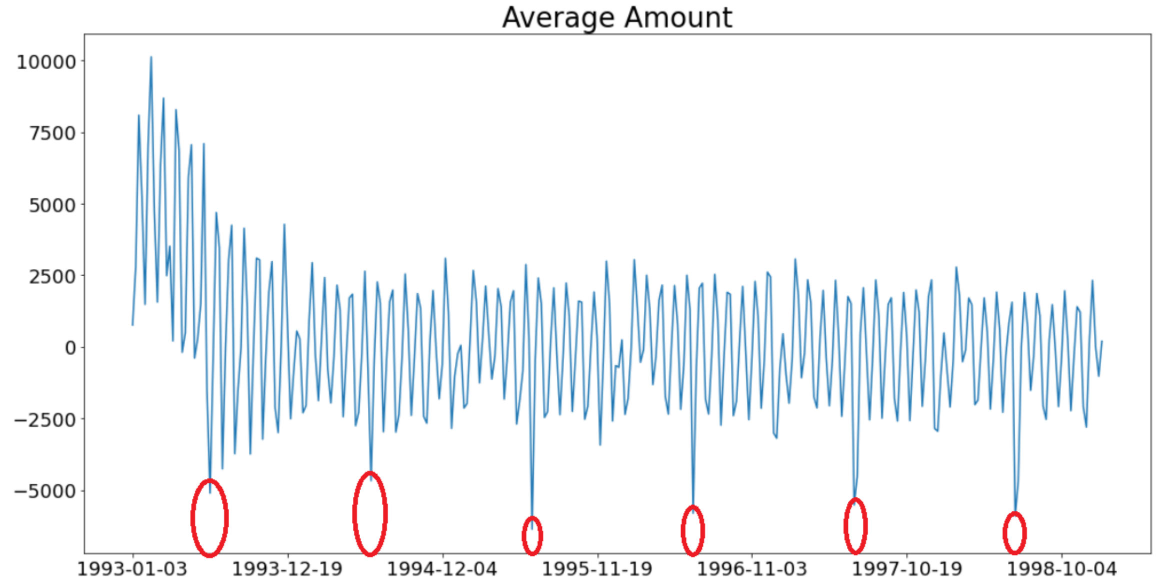
3. Exploratory Data Analysis

- On Transaction Amount, daily liquidity shortfall reaches (80,000) at its peak and this occurs during the mid-year period from 1993 to 1998.

```
# Generate a time plot of our data.
plot_series(df_trans2, ['trans_amount'], title = "Average Amount", steps=1000)
```

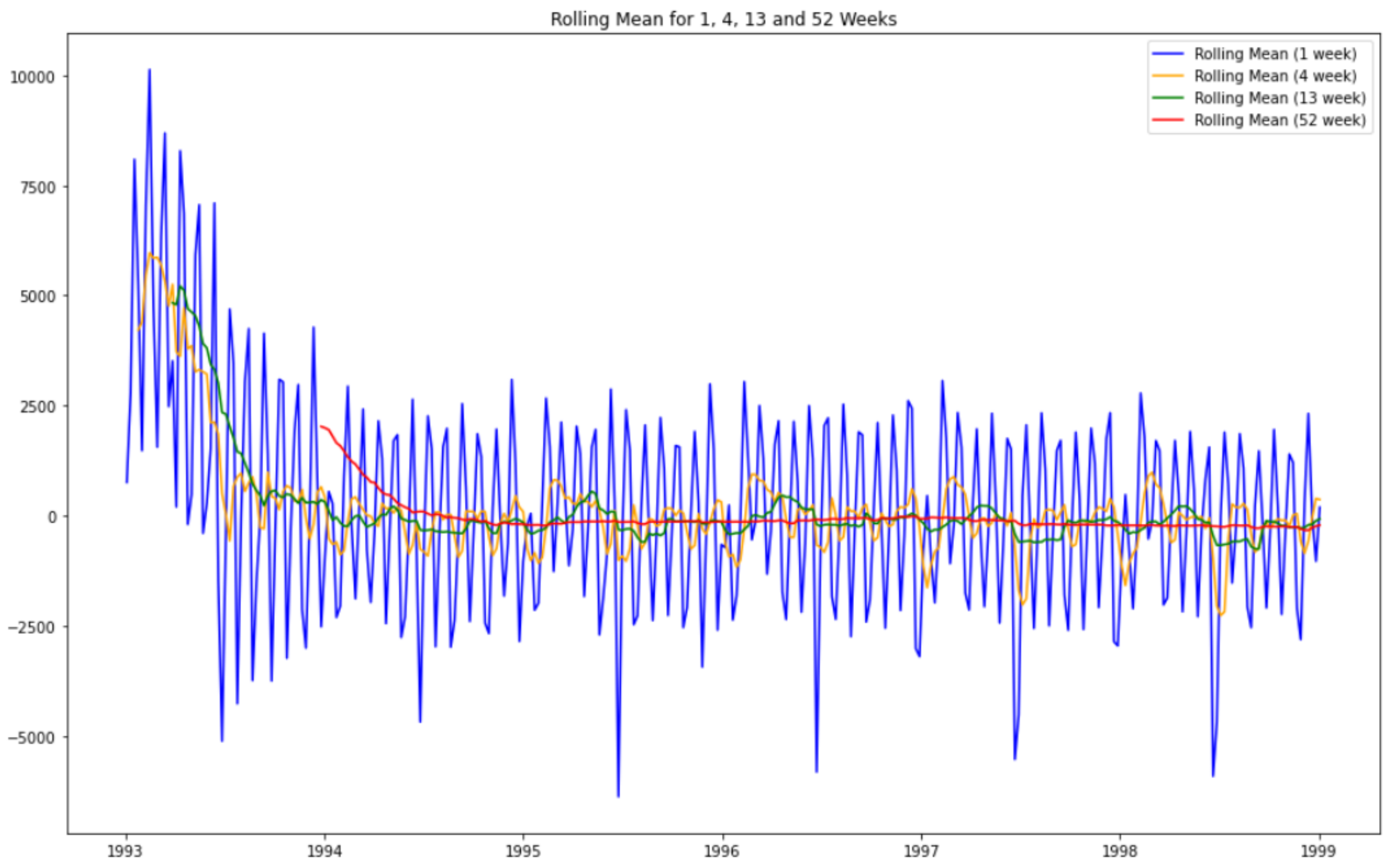



```
# Generate a time plot of our data.
plot_series(df_trans2.resample('W').mean(), cols=['trans_amount'], title='Average Amount', steps=50)
```



3. Exploratory Data Analysis

- On Transaction Amount, as time period increases from Weekly to Yearly, the Average Transaction Amount averages to zero

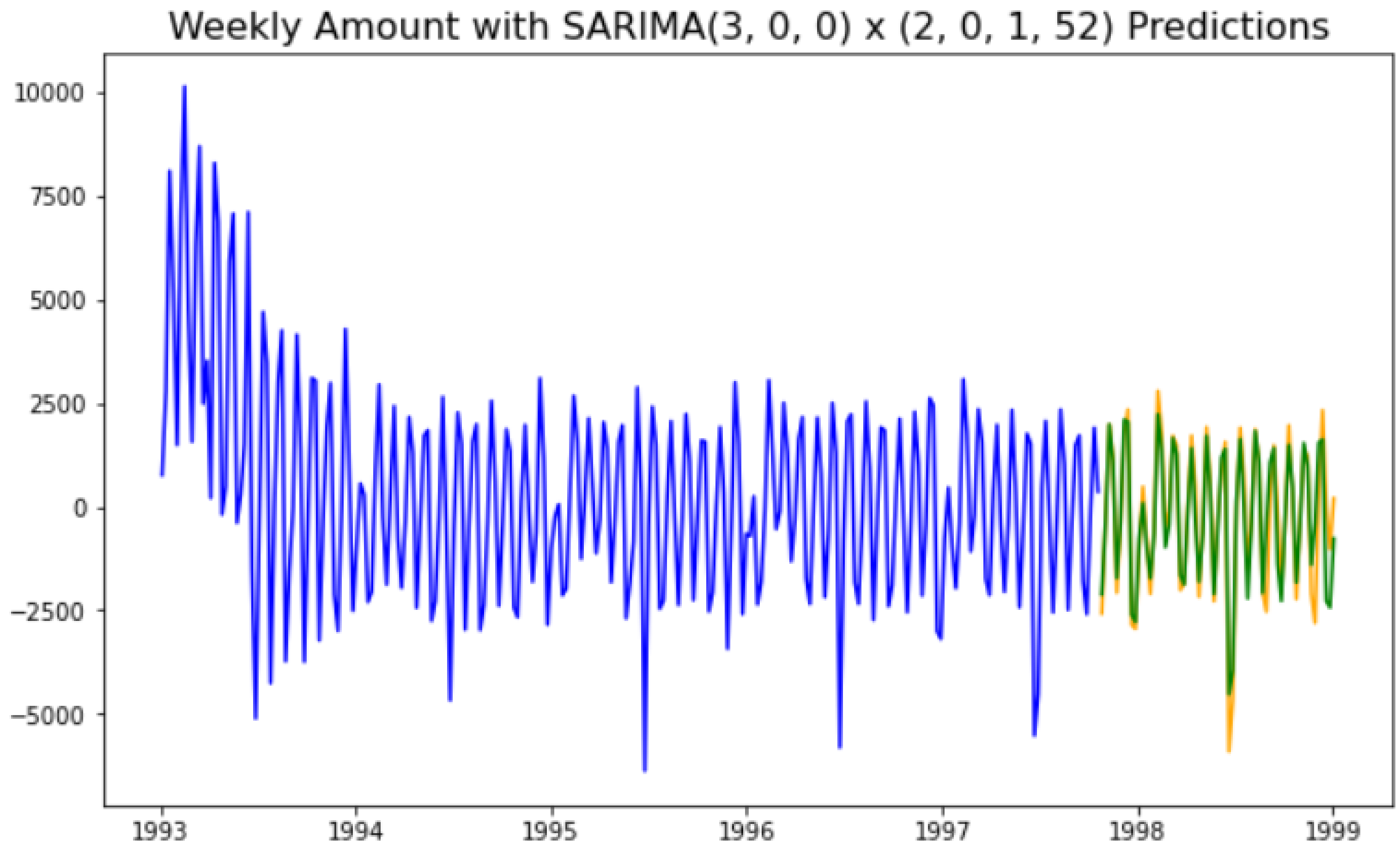


DATA CLEANING & EXPLORATORY DATA ANALYSIS

MODELLING

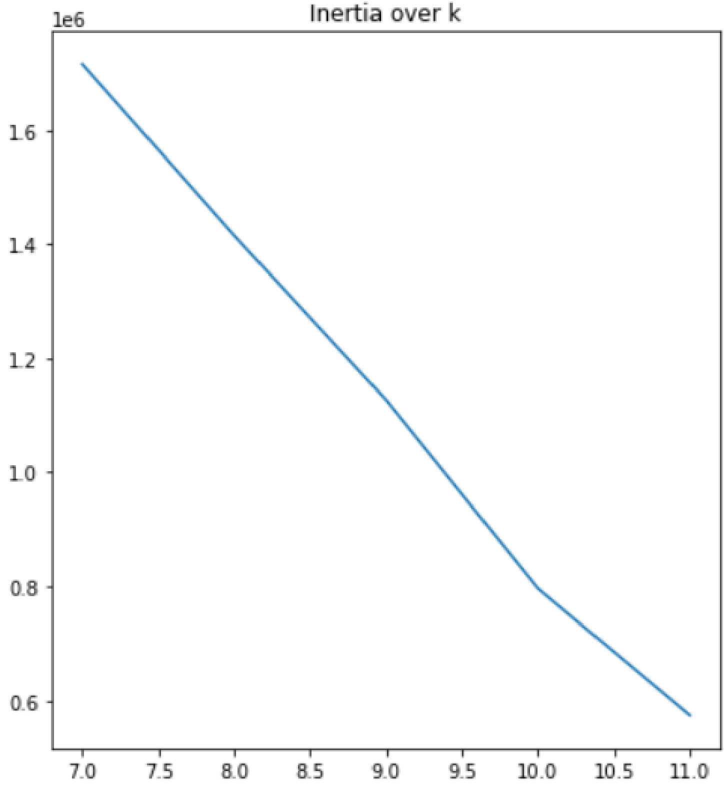
1. Time Series Forecasting using Transactions Amounts Data:

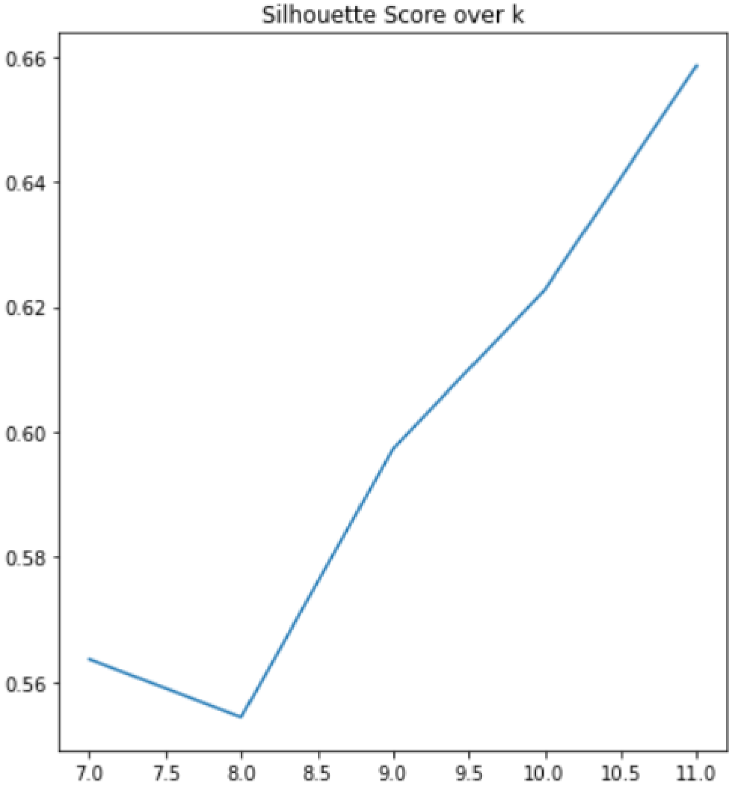
- On Liquidity forecasting, SARIMA model with below parameters is able to predict transaction amounts (green) closely with test data (yellow) with minimum mean squared error of 490k



2. Client Segmentation using K-means clustering

- Using k = 10 clusters with silhouette score of 0.6227 to cluster the clients from the Transactions Data gives an optimal number of clusters AND also sufficient number of clients per cluster with adequate Transaction Amounts and Balances.



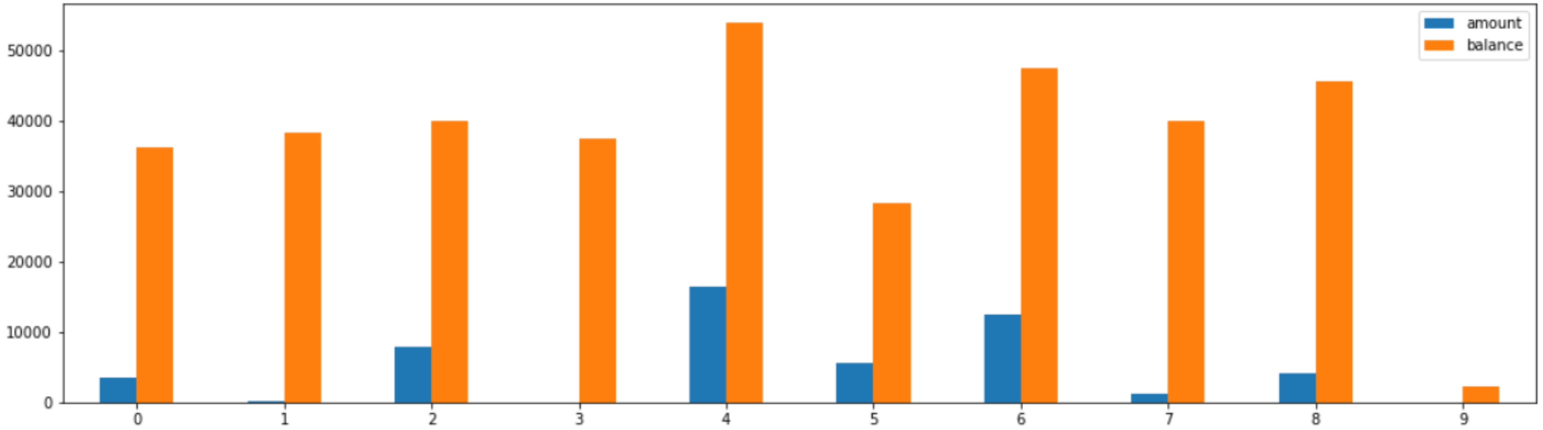


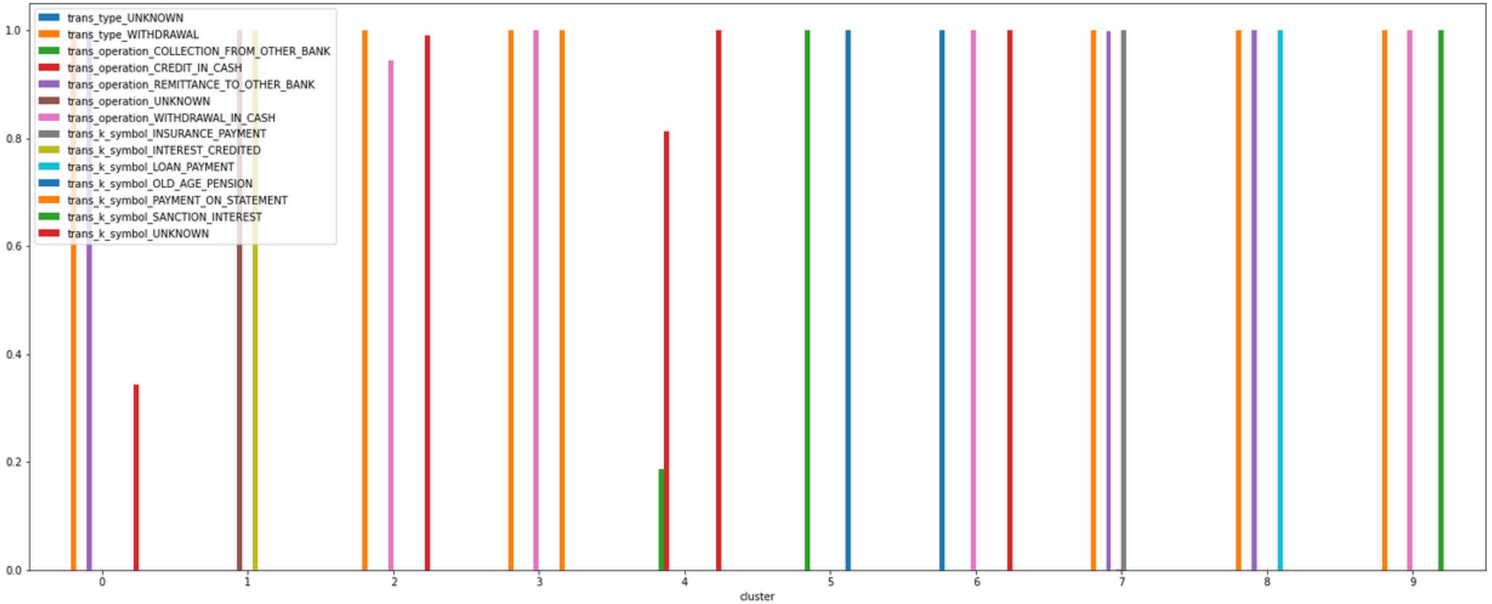
MODELLING

MODELLING

2. Client Segmentation using K-means clustering

- Cluster 4 having high transaction balance and amount has transaction operation from Collection From Other Bank and Credit In Cash.
- Product Recommendations for Cluster 4 may include Investments, Loan and Insurance Product Solutioning.

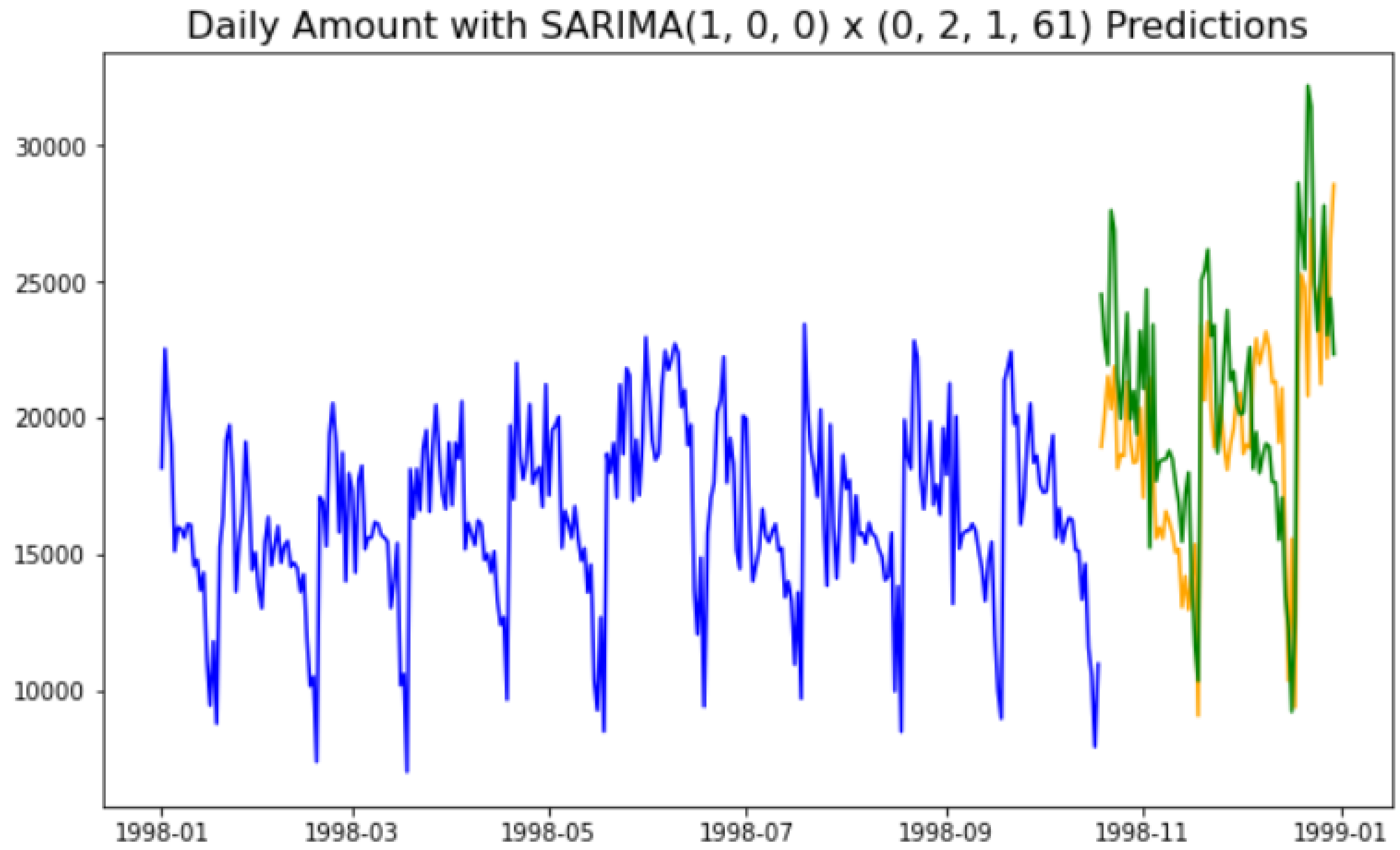




2. Client Segmentation using K-means clustering – Cluster 4

(1998 Transactions Data)

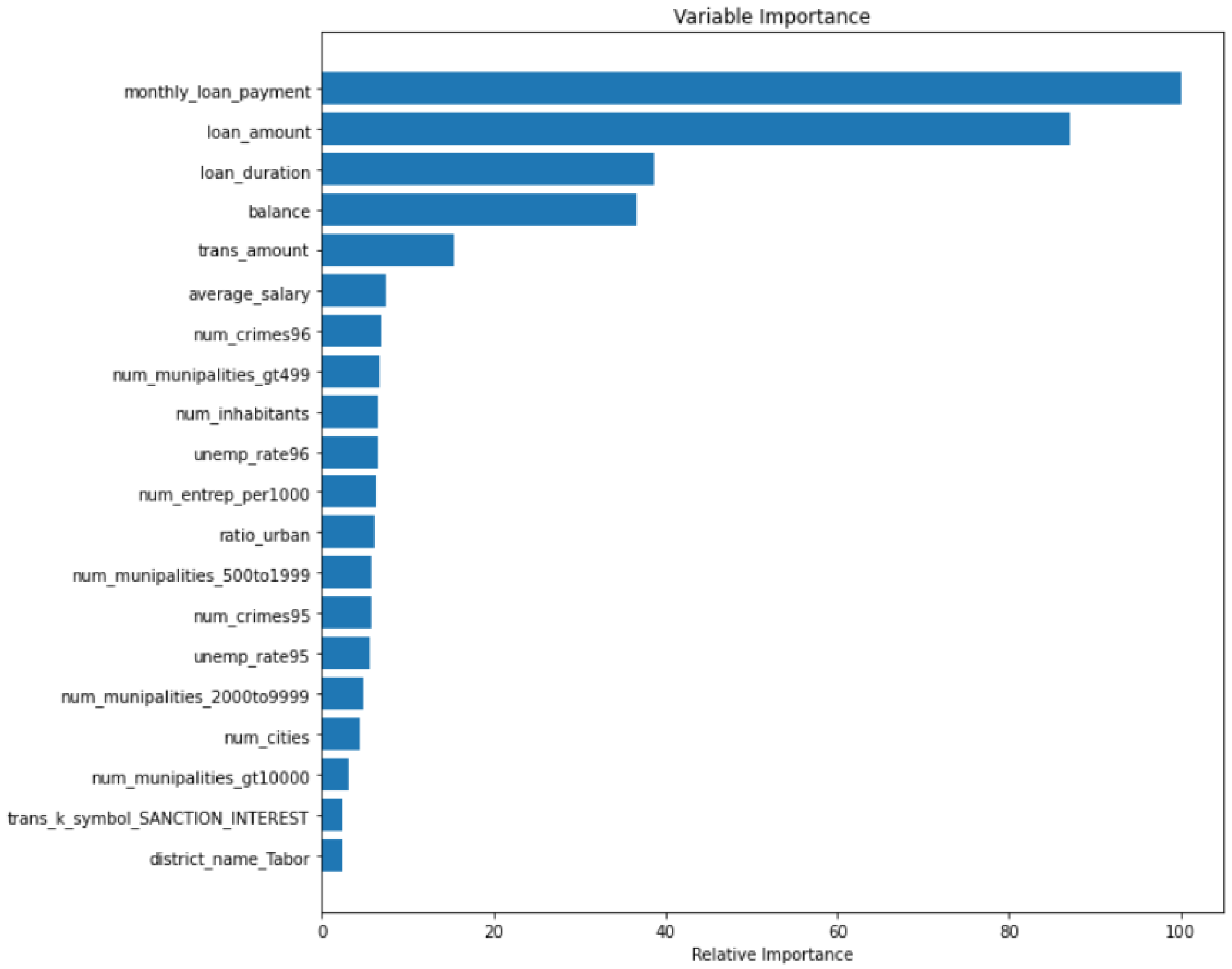
- Using time based modelling on Cluster 4 (1998 Transactions Data), we are able to forecast transaction amounts (green) closely with test data (yellow) with mean squared error of 13m
- With Leads Generation and Recommendations taken to cross-sell Banking Products, we expect Transactions Amount to be higher than below forecast.



MODELLING

3. Loan Default Prediction using Classification

Model	Model Name	Cross Validation on train	Cross Validation on test	Model Accuracy on train	Model Accuracy on test	ROC on train	Roc on test
1	Logistic Regression	0.883	0.884	0.884	0.884	0.709	0.709
2	Gradient Boosting Classifier	0.946	0.945	1	1	1	1
3	Random Forest Classifier	1	0.999	1	1	1	1
4	Decision Tree Classifier	0.999	0.996	1	0.999	1	0.998



MODELLING

CONCLUSIONS & RECOMMENDATIONS

- **Time Series Liquidity Forecasting**:
  - Transactions Amount data has shown daily liquidity shortfall reaches (80,000) at its peak in mid-year from 1993 to 1998. So adequate liquidity should be maintained to ensure sufficient cash for customer withdrawals during this period.
  - SARIMA model with parameters (3,0,0) x (2,0,1,52) on Weekly Transactions Amount data is able to forecast Transactions Amount with minimum mean squared error of 490k.
  - **Future scope for enhancement** may include **including exogenous variables** like **economic growth**, **national stock market index** or **strength of national currency** as macroeconomic factors may play a part in the transactions amount fund flows.

- **Client Segmentation using K-Means Clustering**
  - Using 10 clusters of silhouette score of 0.6227 on the Transactions Data gives an optimal number of clusters that will ensure sufficient number of clients per cluster with adequate Transaction Amounts and Balances.
  - Cluster 4 (1998 Transactions Amount data) having high transaction balance and amount with transaction operations coming from Collection From Other Bank and Credit In Cash is best potential clients to cross-sell.
  - Product Recommendations for Cluster 4 may include Investments, Loan and Insurance Product Solutioning.
  - **Future scope for enhancement** may include **including client risk rating data** so that recommendations for products may be better tailored for clients' appetite for risk.

CONCLUSIONS & RECOMMENDATIONS

- **Loans Default Prediction Classicfication**:
  - Gradient Booasting Classifier, Random Forest Classifier and Decision Tree Classifier all achieve perfect/almost perfect score of 1 in the Model Accuracy Score AND ROC Score for both train and test dataset.
  - It is worth noting that Monthly loan payment, loan amount, loan duration and balance has consistently appeared in all 3 models for Gradient Booasting, Random Forest and Decision Tree Classifier as the top 4 most important features.
  - **Future Scope for enhancement** may include **adding in macroeconomic factors** like **economic growth**, **national stock market index**, **Central Bank benchmark interest rates** and **strength of national currency** as these may also influence an individual's ability to service the loan.

THANK YOU

engsoon@hotmail.com