CASE STUDY

Predictive & Prescriptive Analysis

Submitted To:
Dr. Alekh Gour

Submitted By:
Rohith Kannan
Luv Saxena
Nikhil Tarala
Sivaganesh Mariappan
Prakash Yasrap

TVS Credit

Empowering India. One Indian at a Time

As part of the $8.5 billion TVS Group, we empower Indians from various socio-economic backgrounds with financial products that serve their needs.

Product Portfolio

Two-Wheeler Loans
No.1 TVS Motor financier with a market share of 49.7%

Used Car Loans
Among the top 3 players in the market

Tractor Loans
Key financier for New Tractors, Used Tractors & Agri-implements

Consumer Durable Loans
9000 dealer points touching 1200 towns.

Used Commercial Vehicle Loans
Asset size of Rs.1000 crore with 1000+ channel partners.

Business Loans
Launched in 2018 with focus on Tier2 and Tier3 customers.

Retailer Loans
Launched in 2020 with focus on small retailers.

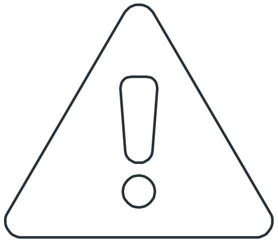
InstaCard Programme
Launched in 2020, it offers a continuous credit line to over 1 lakh+ customers

Challenge

Post Asset Verification (PAV) on Two Wheeler Customer Base

Problem

- One of the activities undertaken by TVS Credit to control fraud at the customer level is Post Asset Verification (PAV) check.



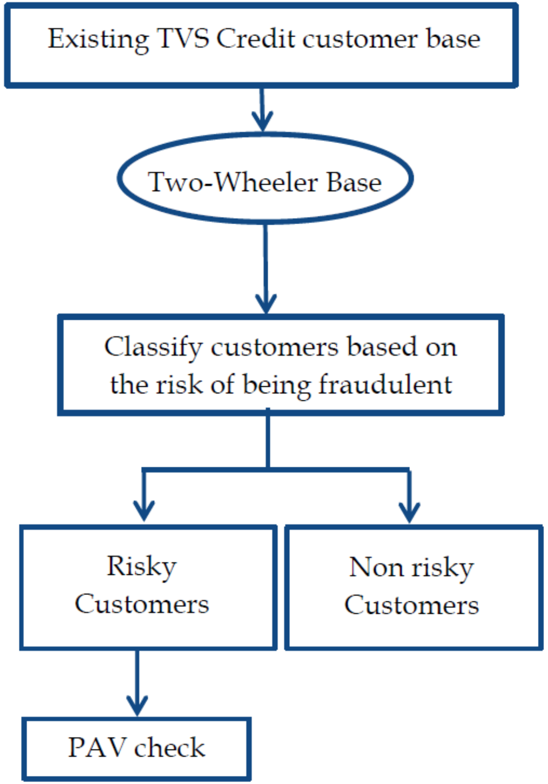
- Since this check requires an employee to visit the customer's address to verify the details, the cost of this activity is high. It is not feasible for the company to do the PAV check for each and every single customer

Mission Statement

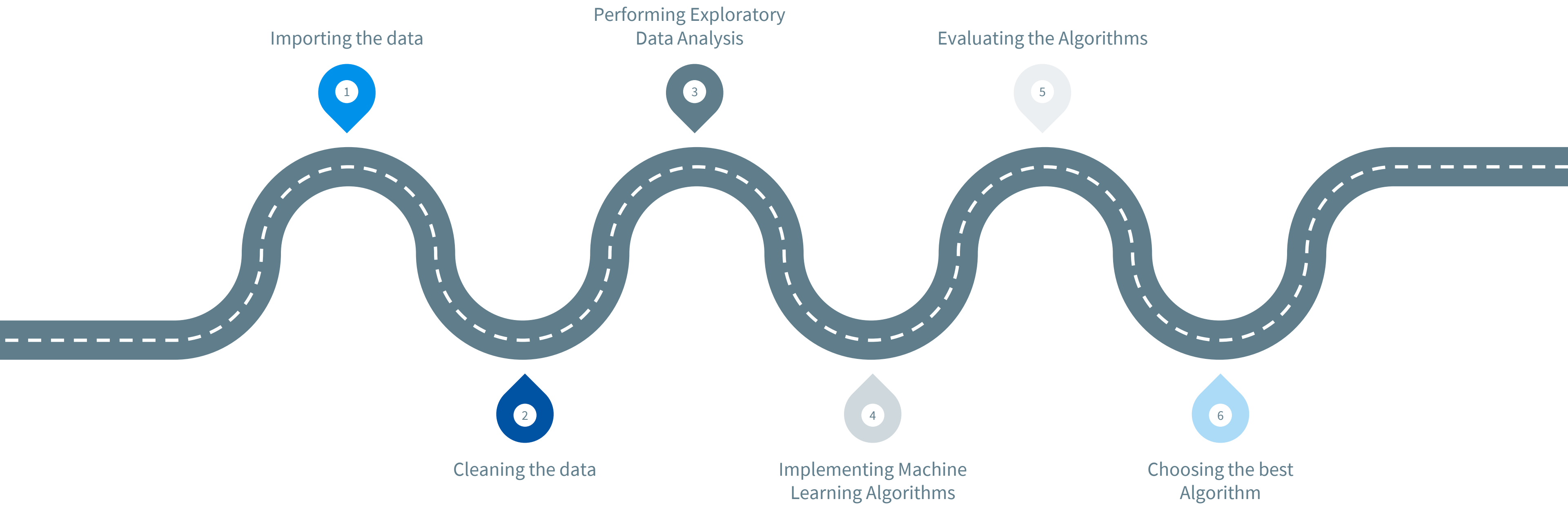
"

The objective of this case study is to use analytics to identify the segment of customers, who have a higher probability of being fraudulent. This would help the company in reducing the cases of fraud.

Approach


Existing TVS Credit customer base
Two-Wheeler Base
Classify customers based on the risk of being fraudulent
Risky Customers
Non risky Customers
PAV check

Roadmap


Importing the data
1
Cleaning the data
2
Performing Exploratory Data Analysis
3
Implementing Machine Learning Algorithms
4
Evaluating the Algorithms
5
Choosing the best Algorithm
6

Dataset Information

The dataset has 22 features with around 11000 rows of information such as :

- Customer Information
- Location of the customer
- Rate of Interest
- Loan Amount
- Qualification
- Asset Cost
- Net Salary
- Date of Disbursal of Loan
- Employment Type
- Processing Fee

Cleaning the data

- The dataset had 5 columns with missing values, but with very low percentage(0.03%).
- Those rows missing values were removed
- The 'CIBIL' score column had 60% of missing data

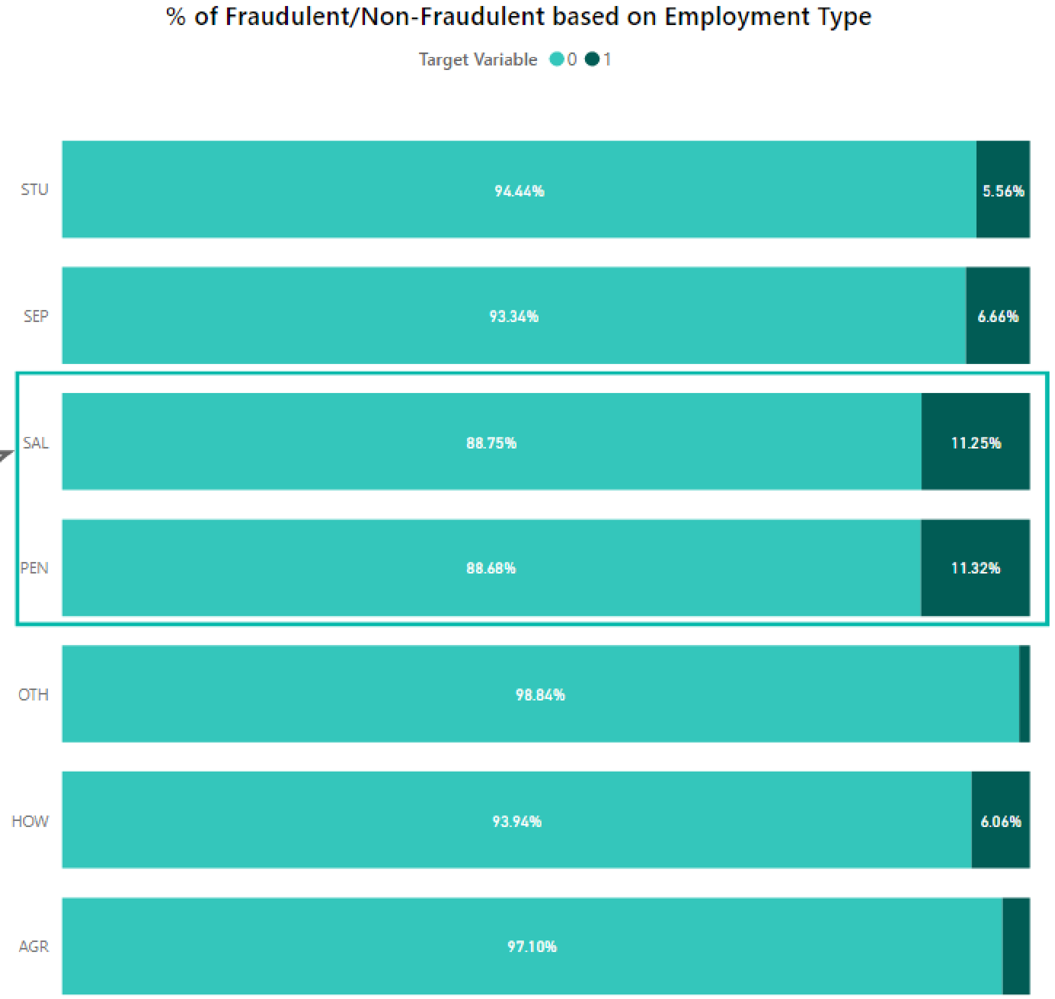
Gender	0.034182
Qualification	0.034182
Employement Type	0.034182
Residence Type	0.034182
Age	0.034182

CIBIL Score	60.681935

EDA

EMPLOYEE TYPE

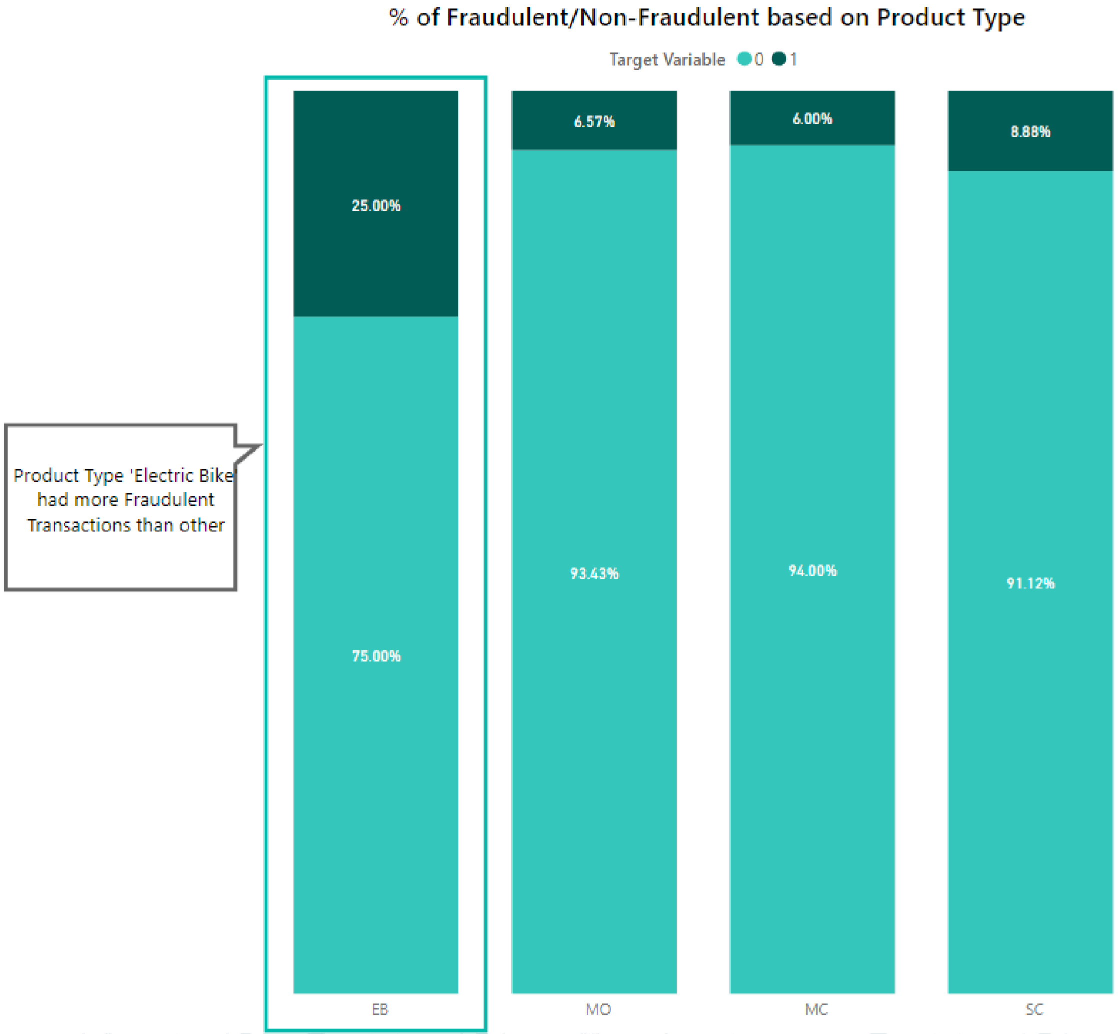
Employee Type 'Pension' and 'Salaried' had more Fraudulent Transactions than other



EDA

Product

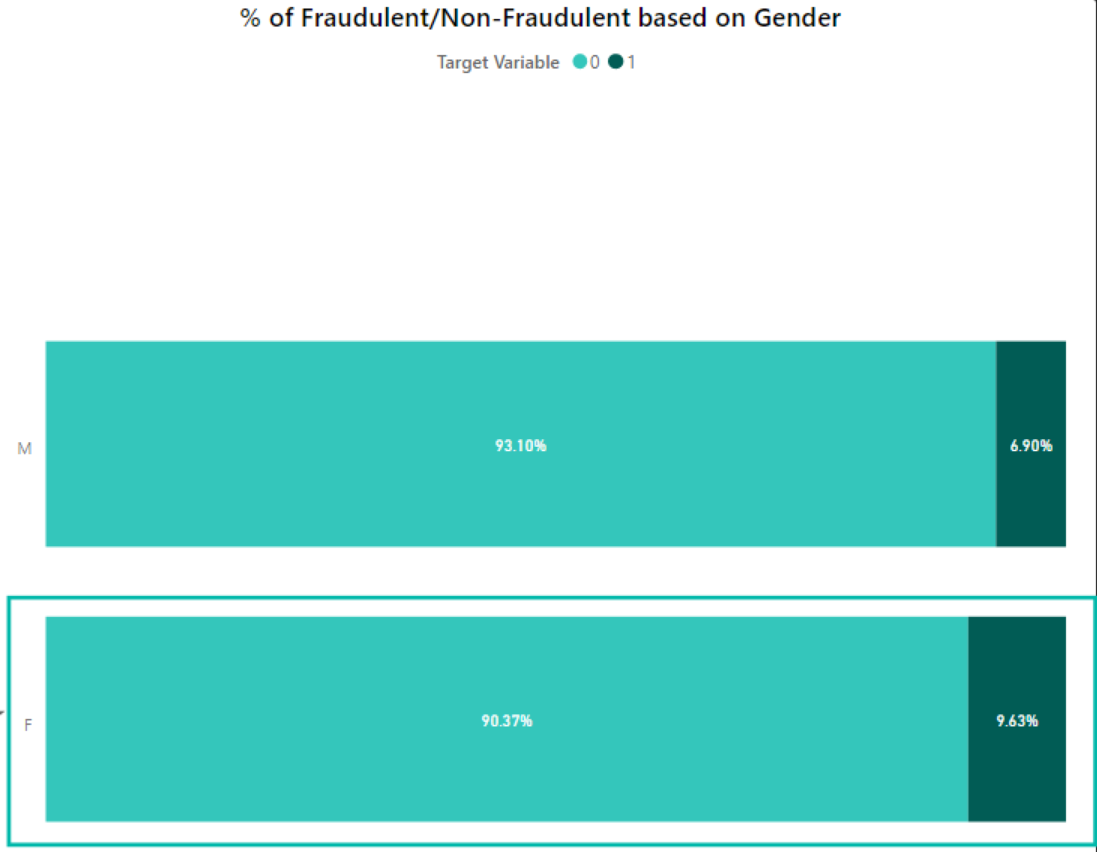
TYPE



EDA

Based on

Gender



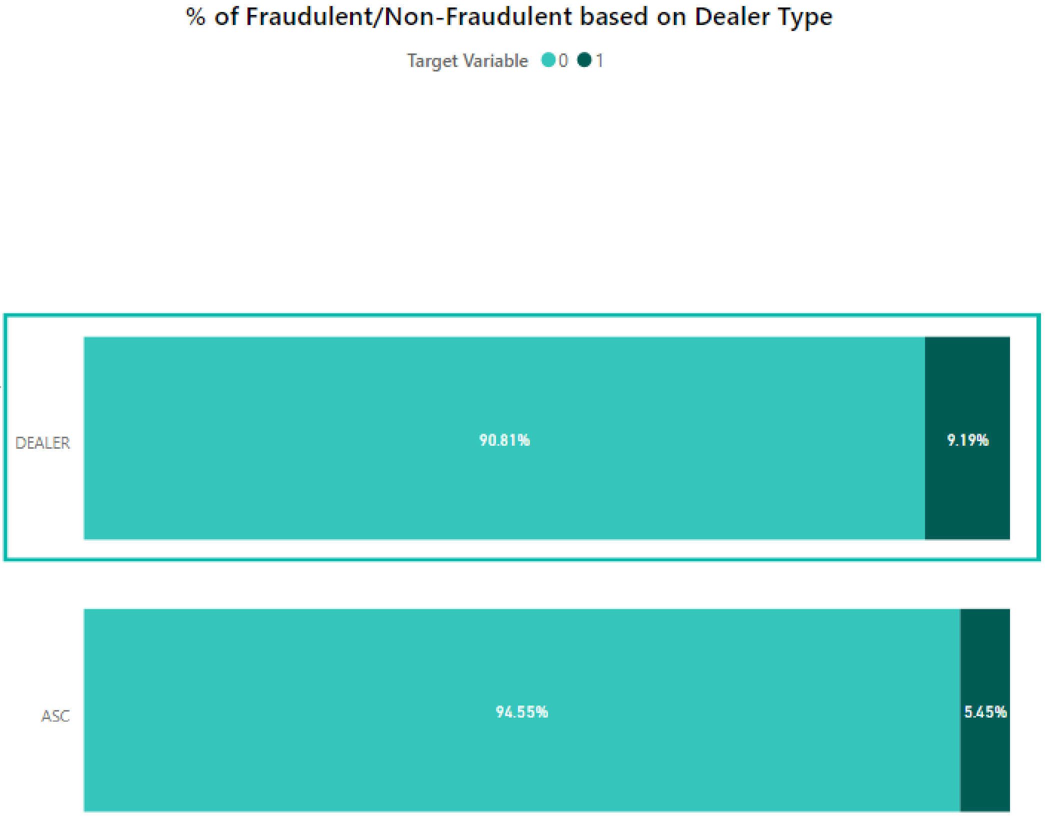
Female had more % of Fraudulent data compared to Male

EDA

Based on

Dealer Type

Dealer type 'Dealer' had more % of Fraudulent data

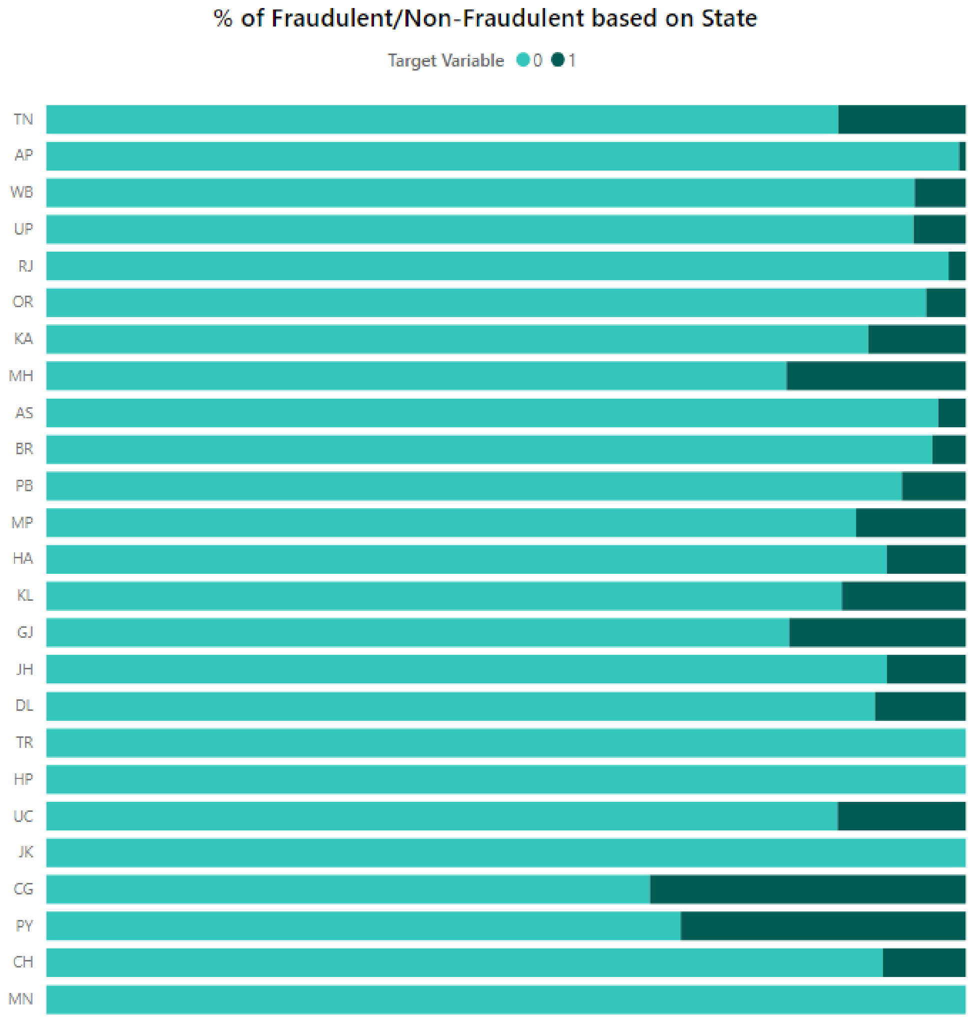


EDA

Based on

State

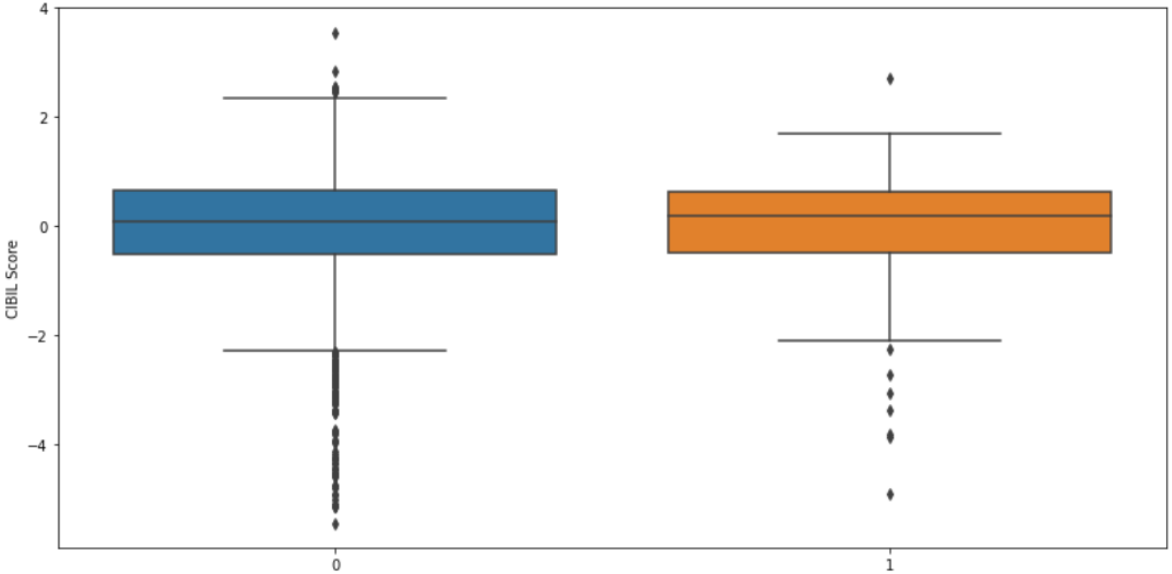
The most Riskiest States were Chatisgarh, Pondicherry, Maharashtra and Gujarat



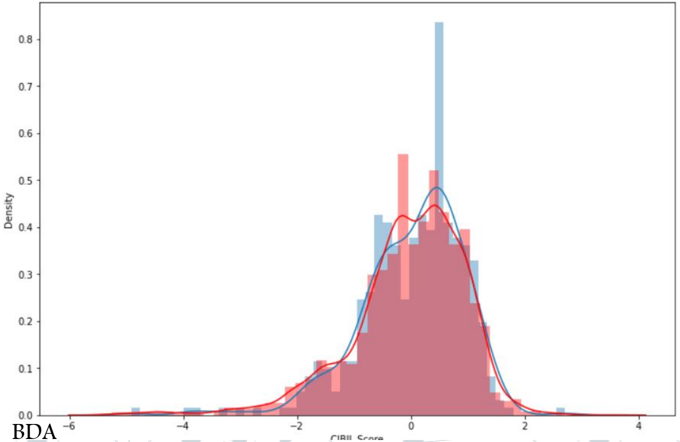
EDA

Based on

CIBIL Score



Multiple Evidences showed that CIBIL score was not a good predictor for this dataset and hence we chose to remove it



Dealing with Outliers and Skewness

Before

	Skewness
Loan Amount	**1.168**
Asset Cost	**1.76346**
EMI	**1.369546**
Net Salary	**14.300964**

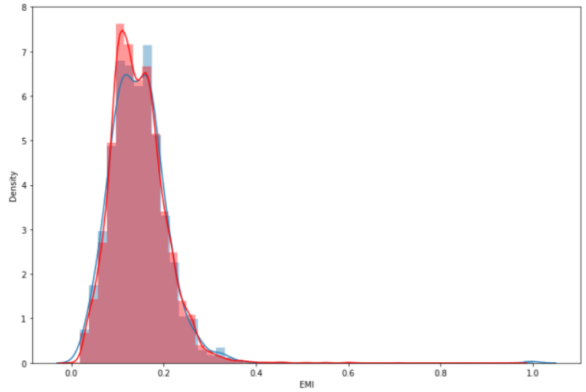
After

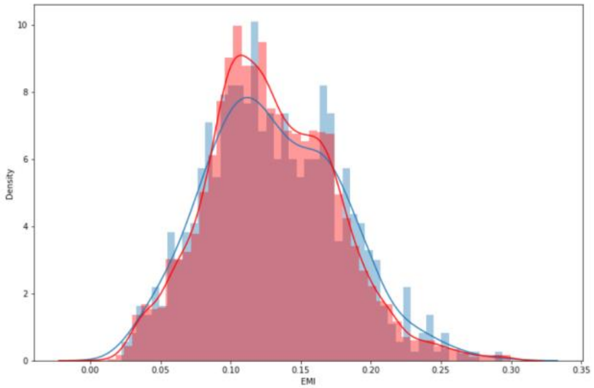
	Skewness
Loan Amount	**0.407**
Asset Cost	**-0.015599**
EMI	**0.350432**
Net Salary	**0.435927**

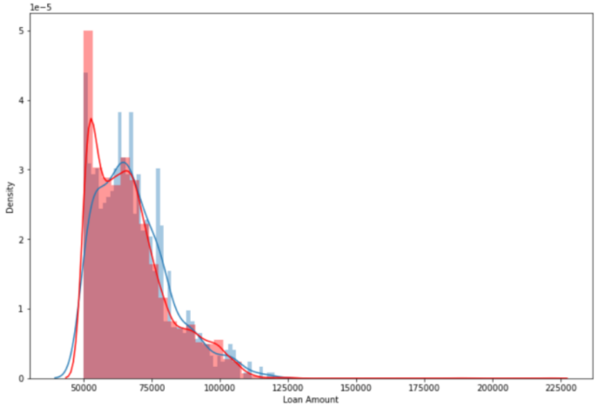
Dealing with Outliers and Skewness

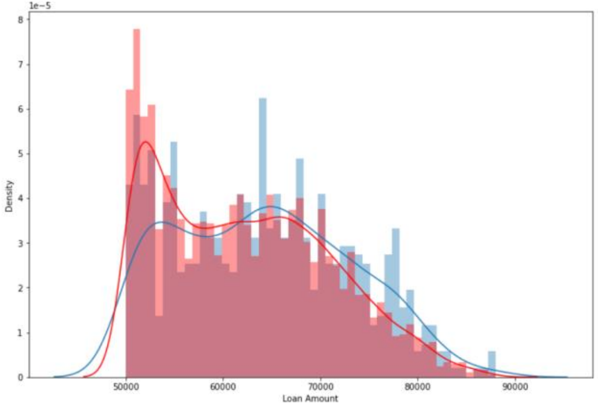
Before

After









Challenges Faced

Dealing with lot of categorical levels

5394

Pin Code

The Column ‘Pin Code’ had 5394 different Pin codes(levels). Hence it was difficult to analyze them.

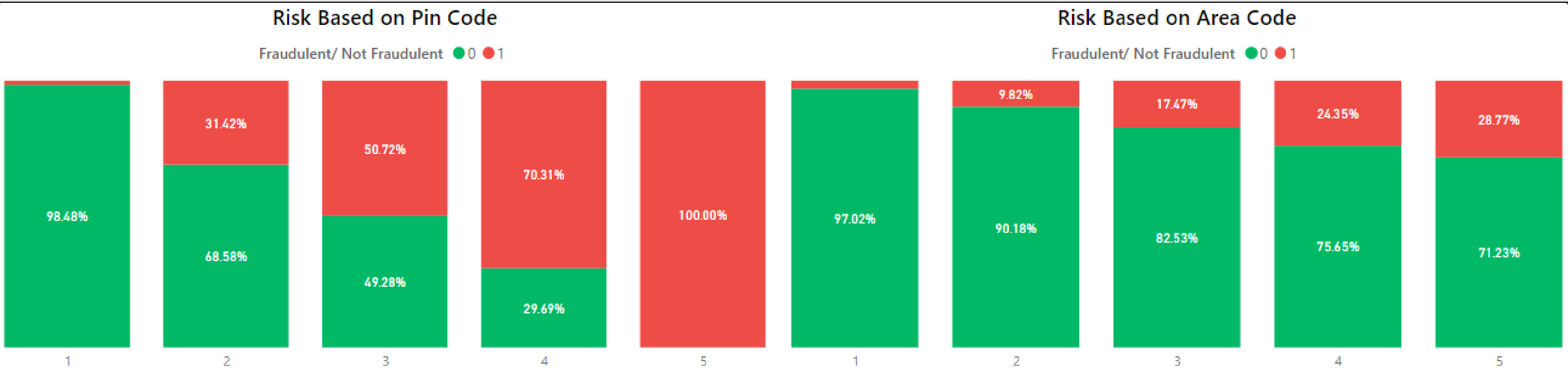
68

Area Code

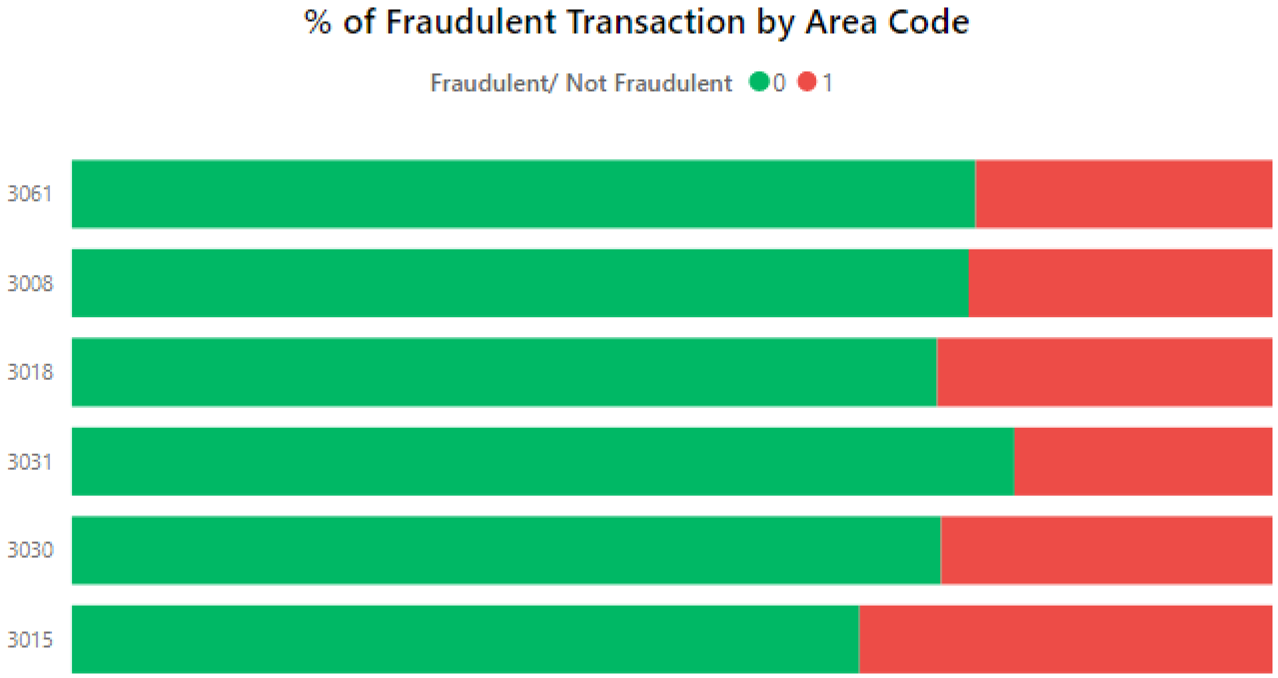
The Column ‘Area Code’ had 68 different Area codes(levels). Hence it was difficult to analyse them.

Approach – Binning and Categorization

Risk Factor : 1 – Lowest 5 - Highest



Findings

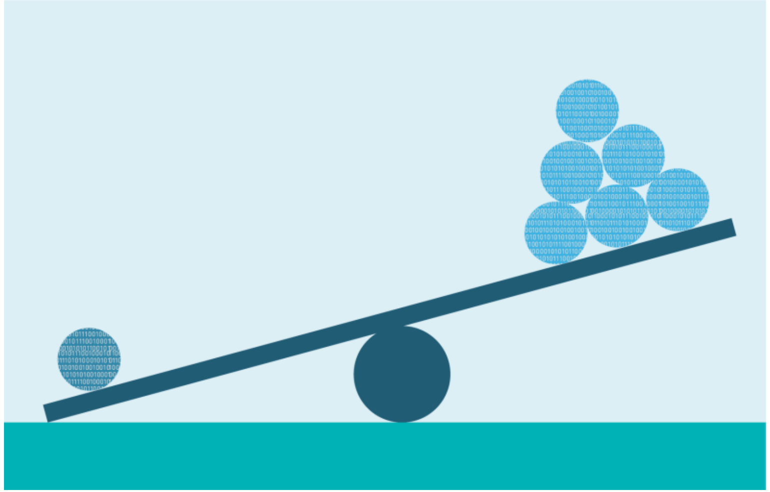


Area Codes 3061, 3008, 3018, 3031, 3030, 3015 were the riskiest of all.

Model Implementation

Im-balanced Dataset Problem

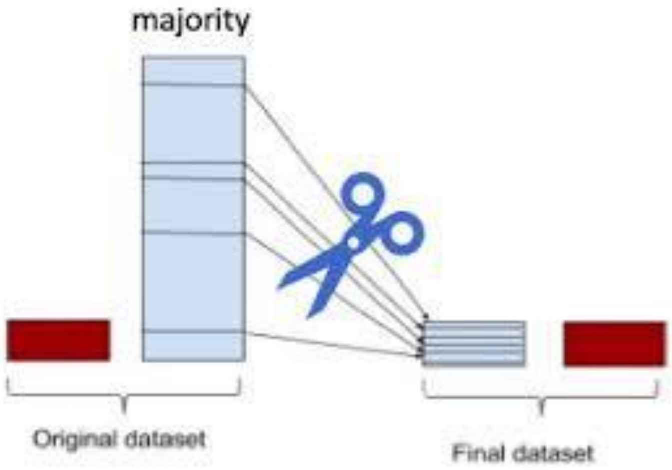
Target Variable	Count
0	7592
1	596



Model Implementation

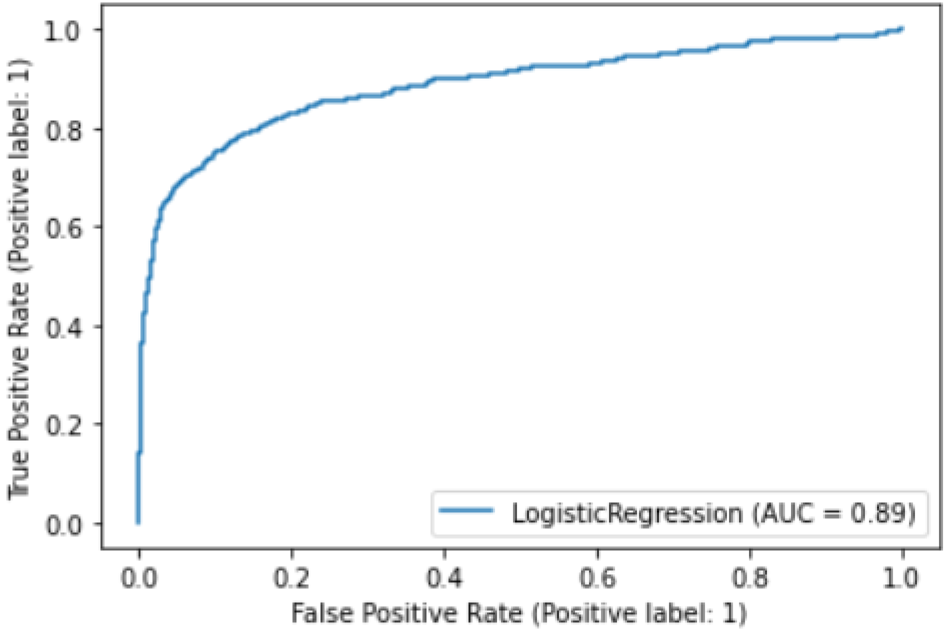
After Under-sampling and Over-Sampling

Target Variable	Count
0	4181
1	4181



Model Implementation

Logistic Regression - ROC Curve



AUC

0.89

Model Implementation

Features which contribute to the prediction

- Area Code
- State
- Dealer Type
- Rate of Interest
- Net Salary
- Employment
- And few others

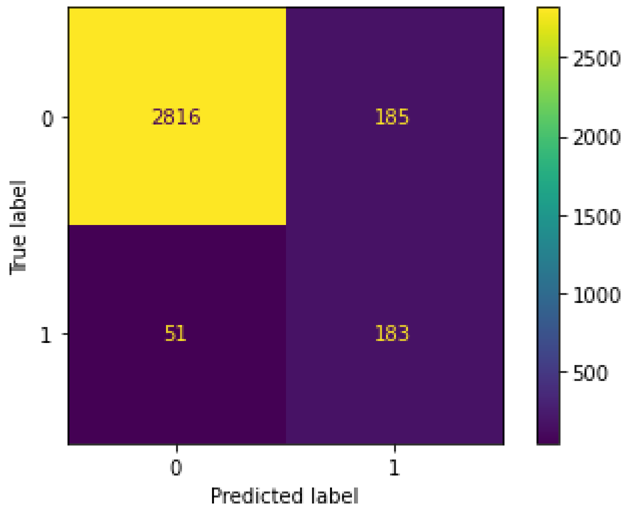
Model Implementation

Logistic Regression - Report

```
Using 0.5 as threshold:
Accuracy = 0.92705
Precision = 0.49728
Recall = 0.78205
F1 score = 0.60797
```

Classification Report

	precision	recall	f1-score	support
0	0.98	0.94	0.96	3001
1	0.50	0.78	0.61	234
accuracy			0.93	3235
macro avg	0.74	0.86	0.78	3235
weighted avg	0.95	0.93	0.93	3235



Model Implementation

Logistic Regression - Report

```
Using 0.5 as threshold:
Accuracy = 0.92705
Precision = 0.49728
Recall = 0.78205
F1 score = 0.60797

Classification Report
              precision    recall  f1-score   support

           0       0.98      0.94      0.96      3001
           1       0.50      0.78      0.61       234

    accuracy                           0.93      3235
   macro avg       0.74      0.86      0.78      3235
weighted avg       0.95      0.93      0.93      3235
```

In this particular case, our objective is to minimize the **False Negatives**, so the **Performance Metric** we are interested in is **Recall/TPR**

Model Implementation

Logistic Regression - Optimization

	Threshold	Accuracy	Precision	TPR	FPR	F1	Specificity	CM
6	0.30	0.783308	0.230681	0.854701	0.222259	0.363306	0.777741	[[2334, 667], [34, 200]]
7	0.35	0.843586	0.295796	0.841880	0.156281	0.437778	0.843719	[[2532, 469], [37, 197]]
8	0.40	0.893045	0.385714	0.807692	0.100300	0.522099	0.899700	[[2700, 301], [45, 189]]

By setting the Threshold to 0.35, we are getting an **Accuracy of 84%** and

Recall of 84%, Hence for our case **0.35** will be the **Optimal Threshold**

Model Implementation

Decision Tree - Report

```
Accuracy = 0.92519
Precision = 0.48507
Recall = 0.55556
F1 score = 0.51793
[[2863  138]
 [ 104  130]]
FPR 0.04598467177607464
Specificity 0.9540153282239253
```

		Predicted	
		0	1
True	0	2863	138
	1	104	130

Model Selection

Logistic Regression

	Value
Accuracy	**84%**
TPR	**84%**

Decision Trees

	Value
Accuracy	**92%**
TPR	**55%**

Final Decision

84%

Accuracy

84%

Recall

We are choosing **Logistic Regression** Model for this specific Case to solve the problem and hence **reduce the cost** for TVS Credit

THANK YOU

Presented to Dr. Alekh Gour || GIM PGDM ~ BDA