1. **Project Overview:**

There are several banks, which gives loan to the customers for E-Bike purchase.

One such bank is facing profit challenges due to escalating incidences of non-payment in their E-Bike financing division. The firm's objective is to ascertain the loan reimbursement capabilities of their clients and comprehend the relative significance of each factor that contributes to a borrower's propensity to honor the loan repayment.

Objective:

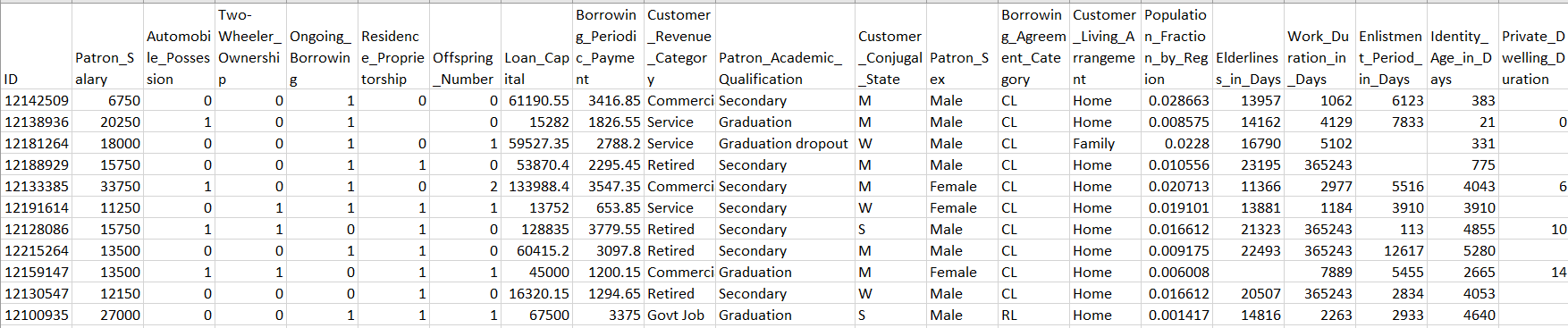
The aim of the problem is to foresee whether a client will fail to honor their loan repayment obligation or not. For each identifier in the attached dataset (refer column "ID"), the task is to predict the "Non-Payment" risk level. Suggest the optimal Credit Risk Model by using the attached dataset. Feel free to give any proxy indicator/ any analysis which can guide bank to find out potential Loan Defaulter. Define the objective function (Evaluation Metric) for Model based on your own intuition.

Data:

Column ID as an Identifier

Column "Default" as Y variable

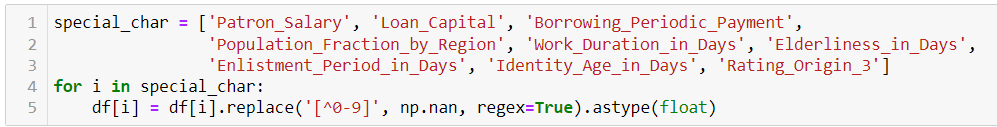
Rest of the columns as X Variable



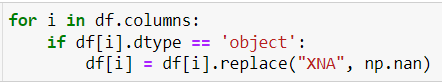
1. **Data Preparation**

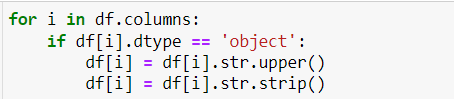
Cleaning the data

* In the data some of the columns were numeric but their data type was object cause they were having some special characters in the data.
* For those cases replacing anything other than numbers to NAN.

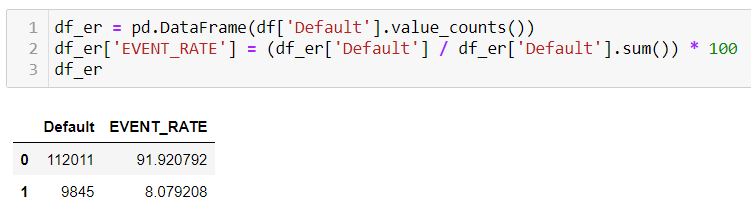


* For categorical columns there were XNA values.
* Replacing those XNA value to NAN. Also, making categorical column values as upper case and removing any leading or trailing white spaces from categories values

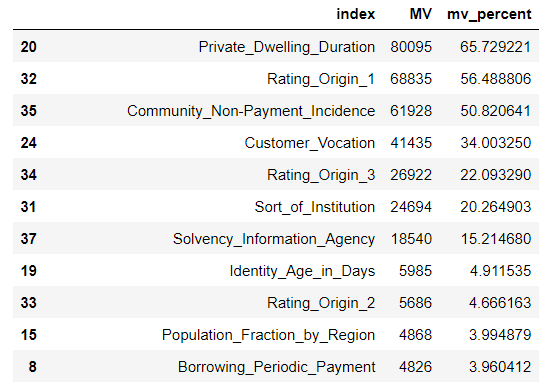




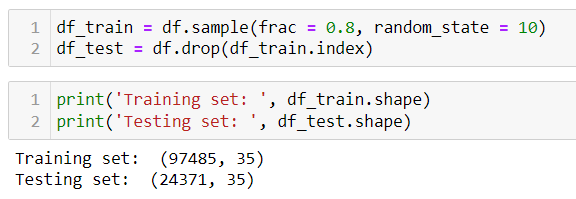
* Checking the event rate of the model



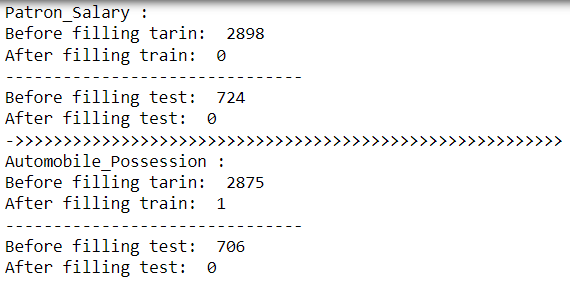
* Checking the count of missing values



* Keeping only those columns which are having less than 25% missing values.
* Then Splitting the data into train & test. Keeping the ratio as 80:20 where training set is 80% and testing set is 20%.

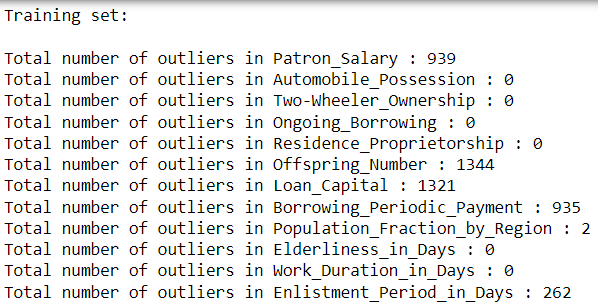


* Filling missing value based on median and mode value of Customer revenue category column.

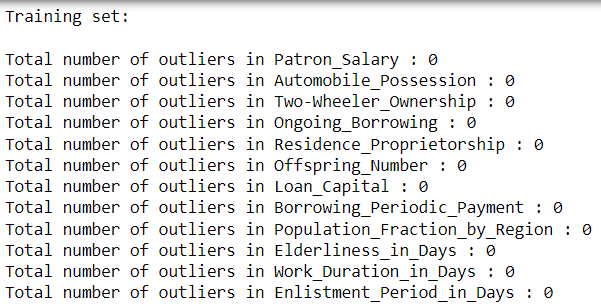


* Outlier treatment:

First identifying the outliers using z-score method and then treating it using 5th and 95th percentiles.



After treatment,



1. **Exploratory Data Analysis (EDA)**

For this step, I have created bivariate graph based on the target, where the binning of the variables has been done on the basis of weight of evidence.

You can refer to the excel where all the graphs of each variable are present in different sheets.

File name: **‘bivariate\_base\_on\_default.xlsx’**

This bivariate has been created in different jupyter notebook.

File name: **‘bivariate\_code.ipynb’**

1. **Variable Selection**

Variable selection is done basis on information gain as I have used tree-based algorithms.

For checking the multi collinearity between variables I have used VIF.

The variables are also selected based on the trends which they are following in the bivariate graph.

|  |  |  |
| --- | --- | --- |
| **VARIABLE** | **Information Gain** | **VIF** |
| Rating\_Origin\_2 | 0.1826 | 1.188 |
| Elderliness\_in\_Days | 0.0789 | 2.027 |
| Enlistment\_Period\_in\_Days | 0.0641 | 1.141 |
| Borrowing\_Periodic\_Payment | 0.05 | 2.671 |
| Work\_Duration\_in\_Days | 0.0449 | 1.855 |
| Identity\_Age\_in\_Days | 0.0325 | 1.146 |
| Loan\_Capital | 0.0221 | 2.503 |
| Telecommunication\_Switch | 0.0186 | 1.085 |
| Rating\_Origin\_3 | 0.0142 | 1.094 |
| Patron\_Salary | 0.0031 | 1.446 |
| Sort\_of\_Institution | 0.0027 | CATG |
| Population\_Fraction\_by\_Region | 0.0024 | 1.328 |
| Patron\_Academic\_Qualification | 0.0021 | CATG |
| Customer\_Revenue\_Category | 0.0017 | CATG |
| Customer\_Urban\_Area\_Ranking | 0.0016 | 1.492 |
| Patron\_Sex | 0.0012 | CATG |
| Patron\_Constant\_Correspondence\_Marker | 0.0008 | CATG |
| Customer\_Living\_Arrangement | 0.0005 | CATG |
| Customer\_Conjugal\_State | 0.0004 | CATG |
| Borrowing\_Agreement\_Category | 0.0004 | CATG |
| Customer\_Professional\_Communication\_Marker | 0.0004 | CATG |
| Automobile\_Possession | 0.0003 | 1.091 |
| Offspring\_Number | 0.0003 | 3.623 |
| Residential\_Phone\_Marker | 0.0003 | 1.199 |
| Employment\_Phone\_Operation | 0.0003 | 1.126 |
| Request\_Submission\_Hour | 0.0003 | 1.106 |
| Patron\_Kin\_Count | 0.0002 | 3.54 |
| Solvency\_Information\_Agency | 0.0002 | 1.052 |
| Request\_Submission\_Day | 0.0001 | 1.001 |
| Two-Wheeler\_Ownership | 0 | 1 |
| Ongoing\_Borrowing | 0 | 1.001 |
| Residence\_Proprietorship | 0 | 1.043 |
| Cellphone\_Marker | 0 | 119699.1 |

1. **Trends of Variables**

**Rating\_Origin\_2**

If Rating\_Origin\_2 of a customer increases, the chance of non-payment risk decreases;

Rating\_Origin\_2 and non-payment risk are negatively correlated.

**Work\_Duration\_in\_Days**

If Work\_Duration\_in\_Days of a customer increases, the chance of non-payment risk decreases; Work\_Duration\_in\_Days and non-payment risk are negatively correlated.

**Elderliness\_in\_Days**

If Elderliness\_in\_Days of a customer increases, the chance of non-payment risk decreases; Elderliness\_in\_Days and non-payment risk are negatively correlated.

**Identity\_Age\_in\_Days**

If Identity\_Age\_in\_Days of a customer increases, the chance of non-payment risk decreases; Identity\_Age\_in\_Days and non-payment risk are negatively correlated.

**Telecommunication\_Switch**

If Telecommunition\_Switch of a customer increases, the chance of non-payment risk decreases; Telecommunition\_Switch and non-payment risk are negatively correlated.

**Rating\_Origin\_3**

If Rating\_Origin\_3 of a customer increases, the chance of non-payment risk decreases; Rating\_Origin\_3 and non-payment risk are negatively correlated.

**Sort\_of\_Institution**

Customers of Group 3 has higher chances of non-payment risk.

|  |  |
| --- | --- |
| Descriptions | Sort\_of\_Institution |
| Group 1 | INSURANCE,INDUSTRY: TYPE 6,TRADE: TYPE 6,TRADE: TYPE 4,UNIVERSITY,POLICE,SECURITY MINISTRIES,MILITARY,INDUSTRY: TYPE 12,TRANSPORT: TYPE 1,BANK,ELECTRICITY,SCHOOL,INDUSTRY: TYPE 2,TRADE: TYPE 2,TRADE: TYPE 1,SERVICES,BUSINESS ENTITY TYPE 2,INDUSTRY: TYPE 5,RELIGION,INDUSTRY: TYPE 9,KINDERGARTEN,CULTURE,HOTEL,GOVERNMENT,TELECOM,TRANSPORT: TYPE 2,INDUSTRY: TYPE 7,INDUSTRY: TYPE 13,ADVERTISING,MEDICINE |
| Group 2 | OTHER,HOUSING,REALTOR,BUSINESS ENTITY TYPE 3,POSTAL,BUSINESS ENTITY TYPE 1,INDUSTRY: TYPE 11,TRADE: TYPE 7 |
| Group 3 | INDUSTRY: TYPE 3,MOBILE,TRADE: TYPE 3,SELF-EMPLOYED,CLEANING,TRANSPORT: TYPE 4,LEGAL SERVICES,RESTAURANT,SECURITY,AGRICULTURE,CONSTRUCTION,INDUSTRY: TYPE 10,INDUSTRY: TYPE 4,INDUSTRY: TYPE 1,EMERGENCY,TRANSPORT: TYPE 3,INDUSTRY: TYPE 8,TRADE: TYPE 5 |

**Patron\_Academic\_Qualification**

Customers of Group2 has higher chances of non-payment.

|  |  |
| --- | --- |
| Descriptions | Sort\_of\_Institution |
| Group 1 | POST GRAD,GRADUATION |
| Group 2 | GRADUATION DROPOUT,SECONDARY,JUNIOR SECONDARY |

**Customer\_Revenue\_Category**

Customers of Group 4 have higher chances of non-payment.

|  |  |
| --- | --- |
| Descriptions | Sort\_of\_Institution |
| Group 1 | BUSINESSMAN,MATERNITY LEAVE,STUDENT,RETIRED |
| Group 2 | GOVT JOB |
| Group 3 | COMMERCIAL |
| Group 4 | SERVICE |

1. **Model Evaluation**

Below is the AUC comparison of different techniques applied on the model base.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model KPI | Model Techniques Explored | | | |
| Logistic Regression | Decision Tree | Random Forest | XG Boost |
| AUC Train | 71.47% | 68.27% | 70.22% | 76.98% |
| AUC Test | 71.90% | 68.03% | 70.84% | 73.75% |

XG Boost is performing well than rest of the other algorithms. Hence, XG Boost is recommended for predicting the non-payment risk of the customers.

1. **Model Performance of XG Boost**

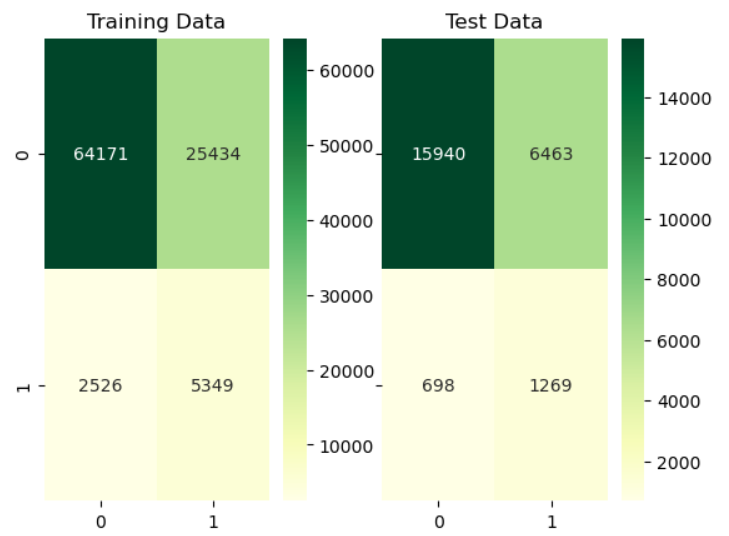
|  |  |  |  |
| --- | --- | --- | --- |
| Model Set | Top 3 Decile Captures | Model Customer Lift  (Top 3 vs Bottom 7) | AUC |
| Train Set | 66.23% | 4.58 | 76.98% |
| Test Set | 62.89% | 3.95 | 73.75% |

High Lift signifies – By targeting 30% of the population, we can higher propensity to prevent non-payment risk. For random model, lift is 1 and AUC is 50%.

Calculation of Lift:

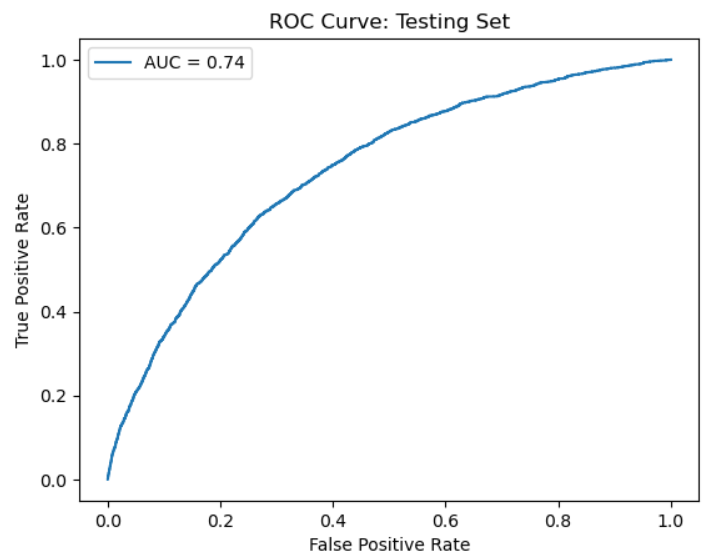
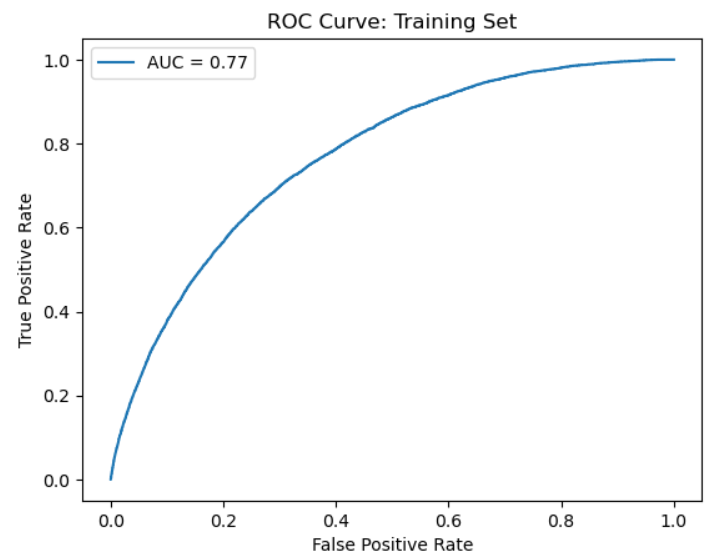
1. Split the model base into 10 deciles.
2. Based on the model output, calculate % Event capture (No. of events/Total decile population) that have been captured in Top 3 decile vs Bottom 7 deciles.
3. Calculate % Lift – (% Event Capture at Top 3 deciles) / (% Event Capture at Bottom 7 deciles)
4. **Error Analysis**

* Confusion Matrix:



Lower the false negative, higher is the recall of the model.

* ROC Curve & AUC



1. **Feature Importance:**

|  |  |
| --- | --- |
| Features | Gain |
| Sort\_of\_Institution | 17.714 |
| Rating\_Origin\_3 | 17.331 |
| Customer\_Revenue\_Category | 15.613 |
| Rating\_Origin\_2 | 15.424 |
| Patron\_Academic\_Qualification | 14.653 |
| Work\_Duration\_in\_Days | 6.175 |
| Elderliness\_in\_Days | 4.503 |
| Telecommunication\_Switch | 4.502 |
| Identity\_Age\_in\_Days | 4.085 |

1. **Final Output**

* After getting the probabilities of each customer from the model we can sort the probabilities in descending order.
* After sorting, divide the data into 10 equal deciles from (0 to 9) where 9 represents the highest probabilities and 0 represents the lowest probabilities.
* Then, basis on the deciles giving High, Medium & Low bucketing to customer where High represents 7th, 8th & 9th deciles (i.e.; these customers have high chances of non-payment), Medium represents 3rd, 4th, 5th & 6th deciles (i.e.; these customers have medium chances of non-payment) and Low represents 0th, 1st & 2nd deciles (i.e.; these customers have low chances of non-payment).

|  |  |
| --- | --- |
| **ID** | **ACTION\_BUCKET** |
| 12154986 | B.MEDIUM |
| 12178336 | C.LOW |
| 12125779 | B.MEDIUM |
| 12198699 | B.MEDIUM |
| 12215293 | C.LOW |
| 12174821 | B.MEDIUM |
| 12154811 | A.HIGH |
| 12207402 | B.MEDIUM |
| 12173377 | B.MEDIUM |
| 12207691 | C.LOW |
| 12161737 | C.LOW |
| 12205207 | A.HIGH |
| 12173521 | A.HIGH |