INTERNSHIP – DATA ANALYTICS

PROJECT 4
LOAN APPLICATION DATA ANALYSIS

SUBMITTED BY

HRISHIKESH KALITA (EI0123)

COLLEGE

NATIONAL INSTITUTE OF TECHNOLOGY ROURKELA

GOAL

- ❖ Checking the Customer's eligibility to get approval for the loan applied.

DATA STATS

- o Shape of the Train Data: (614, 20)
- o Columns of the Data:
  'Loanapp_ID', 'Sex', 'Marital_Status', 'first_name', 'last_name', 'email', 'address', 'Dependents', 'Qual_var', 'SE', 'App_Income_1', 'App_Income_2', 'CPL_Amount', 'CPL_Term', 'Credit_His', 'Prop_Area', 'INT_ID', 'Prev_ID', 'AGT_ID', 'CPL_Status'
- o Data info:

	App_Income_1	App_Income_2	CPL_Amount	CPL_Term	Credit_His	INT_ID
count	614.000000	614.000000	612.000000	600.00000	564.000000	6.140000e+02
mean	6484.151140	1945.494958	175.805882	342.00000	0.842199	5.055666e+09
std	7330.850008	3511.498043	102.606123	65.12041	0.364878	2.890445e+09
min	180.000000	0.000000	10.800000	12.00000	0.000000	1.788664e+07
25%	3453.000000	0.000000	120.000000	360.00000	1.000000	2.561243e+09
50%	4575.000000	1426.200000	153.600000	360.00000	1.000000	5.244783e+09
75%	6954.000000	2756.700000	200.700000	360.00000	1.000000	7.495052e+09
max	97200.000000	50000.400000	840.000000	480.00000	1.000000	9.989158e+09

COMMENTS

- o The discrepancies in the 'count' row suggest the presence of null/Nan values in the respective columns.
- o The 'Dependents' column should have been a numerical column but it is not present in the above table because of several '3+' entries in it which eventually makes it an 'object' type column. This was fixed by considering all '3+' as '3' and converting the column into a 'float' type.

DEALING WITH NULL/NAN VALUES

	Columns	Filling_Method	Difference
0	Loanapp_ID	equal	0
1	Sex	ffill	1
2	Marital_Status	ffill	1
3	first_name	equal	0
4	last_name	equal	0
5	email	equal	0
6	address	equal	0
7	Dependents	equal	0
8	Qual_var	equal	0
9	SE	equal	0
10	App_Income_1	equal	0
11	App_Income_2	bfill	1
12	CPL_Amount	equal	0
13	CPL_Term	equal	0
14	Credit_His	ffill	1
15	Prop_Area	equal	0
16	INT_ID	equal	0
17	Prev_ID	equal	0
18	AGT_ID	equal	0
19	CPL_Status	equal	0

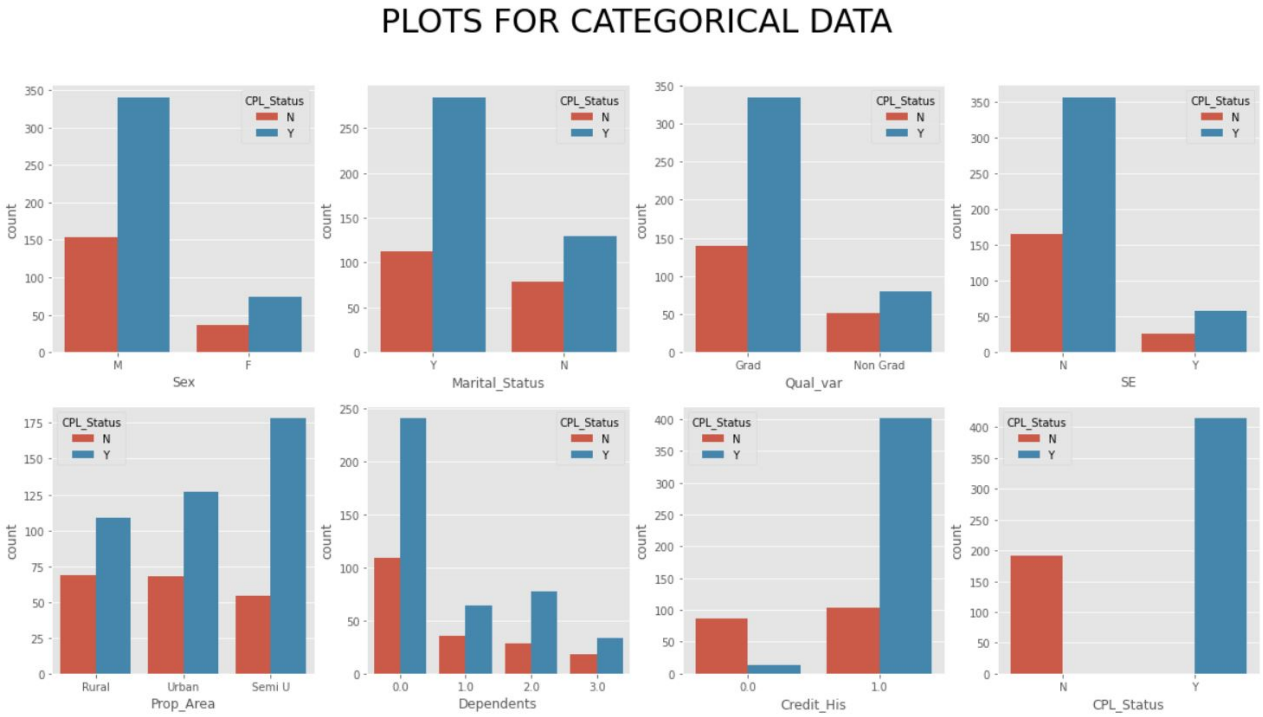
- A user-based function was used to find the trend in the data regarding the 'fill' method. It was coded with the intention of getting an overview of the fact that whether **'ffill'** or '**bfill'** should be used in Pandas 'fillna' function.

COMMENTS

However, the table obtained shows almost similar patterns for all the columns. Only 3 out of all columns had 'ffill' majority with a difference of 1, which could be ignored. However, eventually '**ffill'** was used to some extent on the training data and the remaining rows with **NaN** values were **removed**. (Train Data Shape: (605, 20) after dealing with missing values.)

EXPLORATORY DATA ANALYSIS (EDA)

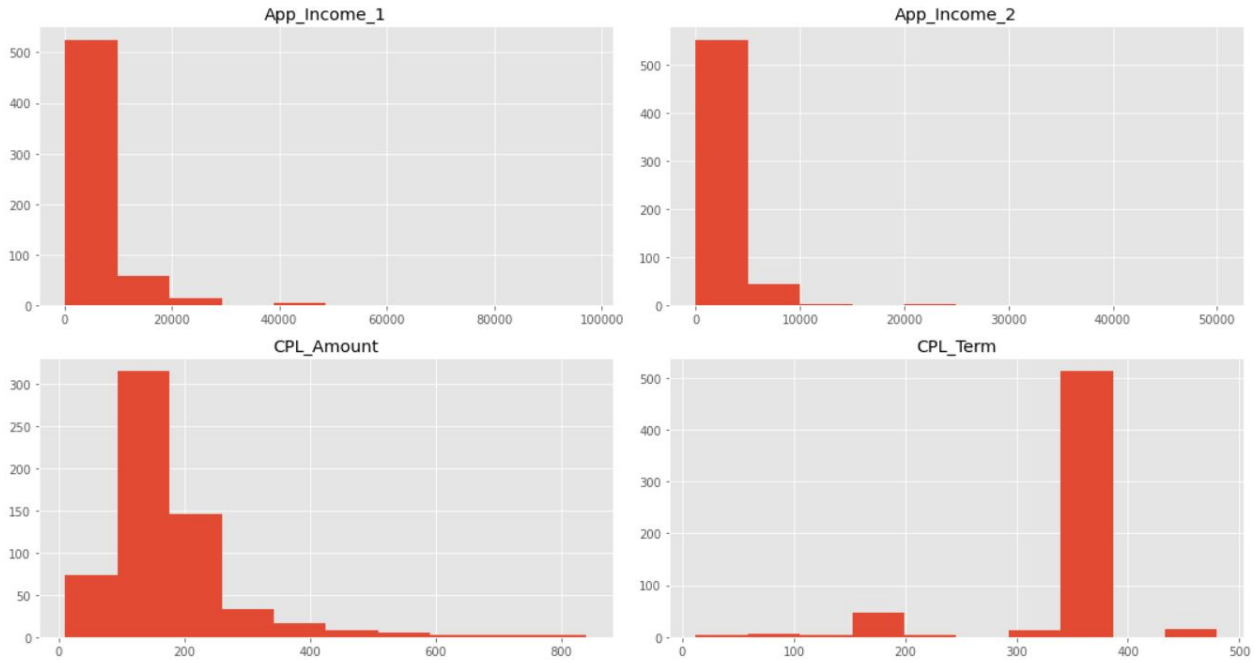
- **CATEGORICAL**



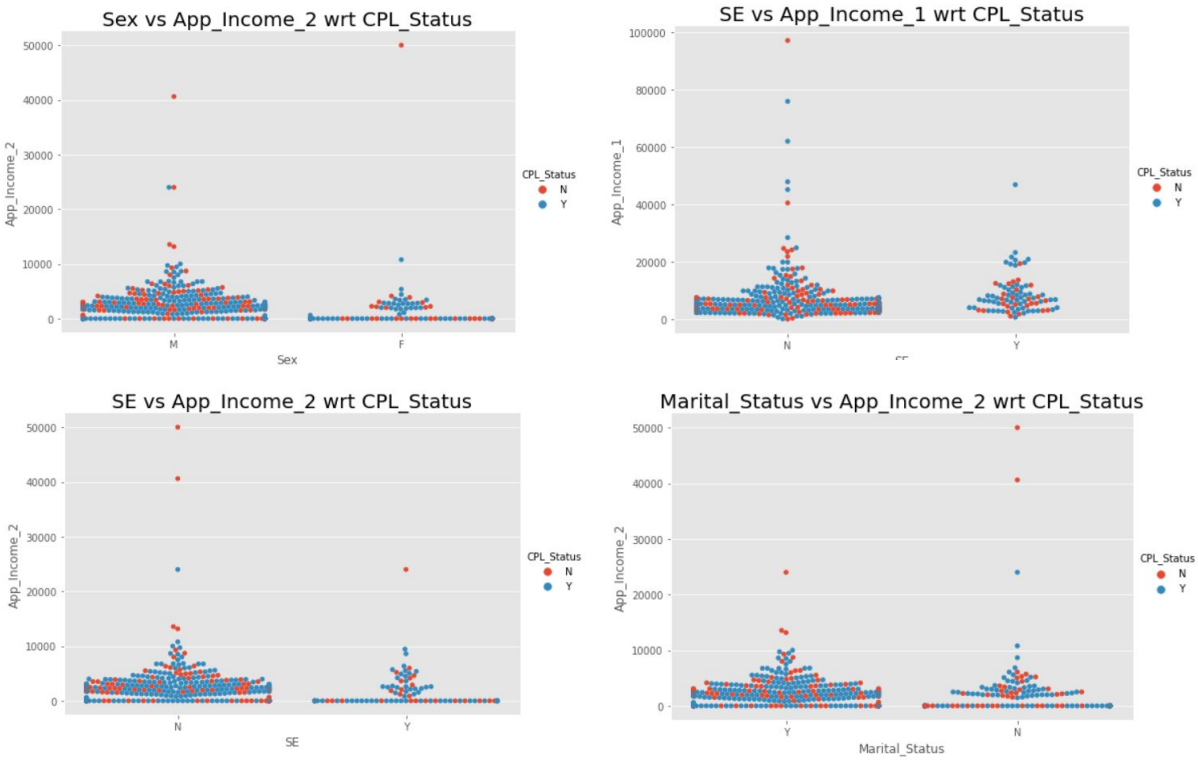
COMMENTS

- A general trend of **lesser 'Female'** applicants was observed.
- A **significant** number of **approvals** can be observed in **Semi-Urban** areas.
- In most of the cases the number of **rejected** loan applications is lesser than that of the **approved**. However, in case of applicants **without Credit History**, it is the **opposite**. There is a **greater** number of **rejections**.
- From the data observed above, it may be assumed that a **married, graduated** and **non-self-employed male** with a valid **credit history**, residing in a **semi-urban** area **without any dependents** stand a **higher chance** of being **approved** for a loan.

➢ CONTINUOUS



➢ OVERALL



COMMENTS

- From the figures above, relationship among various entities can be studied in detail.
- Presence of probable **outliers** can be seen in almost all the diagrams.
- From the visualizations, one can suggest that the dataset **doesn't contain many** people with **very high income** or **high loan amount**.
- Most 'swarms' were crowded in the **lower income and lower loan amount** regions.

FEATURE IMPORTANCE

Columns	Scores
Credit_His	28.162521
Qual_var	3.740024
Marital_Status	1.980991
Prop_Area	0.301681
Sex	0.035887
SE	0.014844
Dependents	0.000290

Obtained by using **SelectKBest** feature from **Scikit-Learn**. **Chi2** method was used to draw out the feature importance in deciding the output in the model.

Columns	Importance_Score
App_Income_1	0.408858
CPL_Amount	0.318965
App_Income_2	0.202613
CPL_Term	0.069564

Obtained using **feature importance** method in **Random Forest Classifier**.

MODELLING

MODELS USED:

- RANDOM FOREST CLASSIFIER
- DECISION TREE CLASSIFIER
- SUPPORT VECTOR MACHINES
- LOGISTIC REGRESSION
- ADABOOST CLASSIFIER

CROSS-VALIDATION was used with **cv = 10** and the **Standard Deviations** among the scores were very low implying the **stability** of the models.

PREDICTION OF TEST DATA

- The missing values were first imputed using Pandas **fillna** function (‘**ffill’**).
- For getting a better judgement, a **user-defined** function was used to **get the predictions from the models** and decide the **final prediction according to the majority rule**.

  For example, if 3 of the models used gives the prediction as “**Y**” (Approved), and other two as “**N**” (Rejected), then “**Y**” (Approved) has a 60% majority and hence the respective application is predicted to be ***approved***.
- The final predictions were attached to the test data csv file as a new column and saved as **Test_Data_Predictions.csv**