Mid Term Project:

Credit Card Prediction Analysis:

Cause for analysis:

At present, with the development of machine learning algorithms. Predictive methods such as Random Forest, and Support Vector Machines have been introduced into credit card scoring. However, these methods often do not have good transparency. It may be difficult to provide customers and regulators with a reason for rejection or acceptance.

Data Used:

Used two datasets,

- *Credit record and*
- *Application record of the users*

We used the personal information and data submitted by credit card applicants to predict the probability of future defaults and credit card borrowings.

Source **:**

https://www.kaggle.com/datasets

Features used:

(16 columns)

- education_encoded
- gender
- car_owned
- property_owned
- income_type
- marital_status
- housing_type
- children_count
- annual_income
- days_from_birth
- days_from_employment
- owned_mobile
- owned_workphone
- owned_phone
- owned_email
- family_size

Objective of the Analysis:

Build a machine learning model to predict if an applicant is eligible to get a credit card or not.

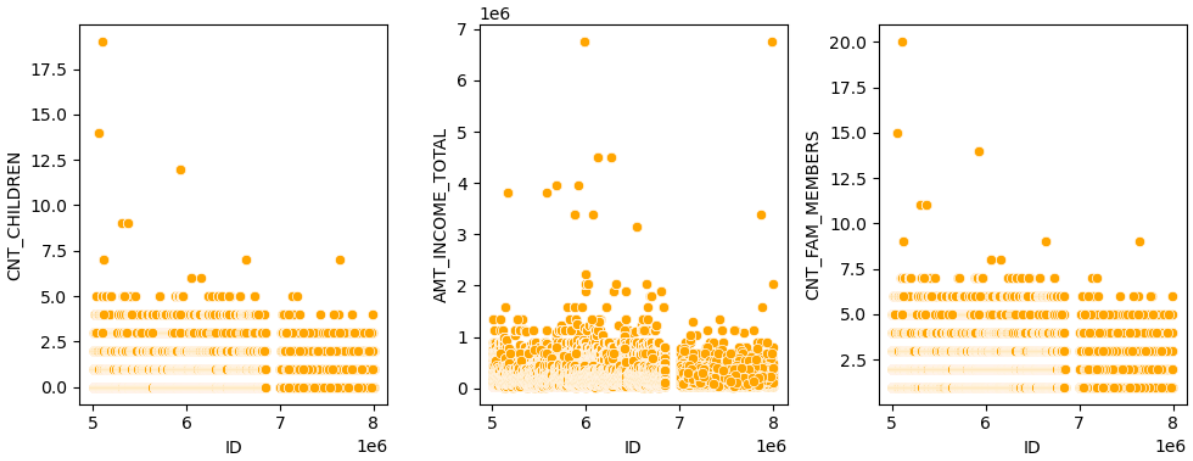
Model:

Cleaning:

1. Checked for the number of rows and columns in datasets:
   Application record= 438557 rows × 18 columns
   Credit record = 1048575 rows × 3 columns

2. Checked for unique id's and in application record and matched it with credit record
3. Changed column names and used standard lower case for naming.
4. Grouped the categorical column elements into more interpretable names.
5. Traced out the column with highest null values and removed it
   Occupation_type
6. Grouped credit record dataset on unique ID's

Tracing Outliers and removing:



The columns were important for prediction analysis hence we removed the outliers using quantile thresholds.
It's called outlier trimming or removal.

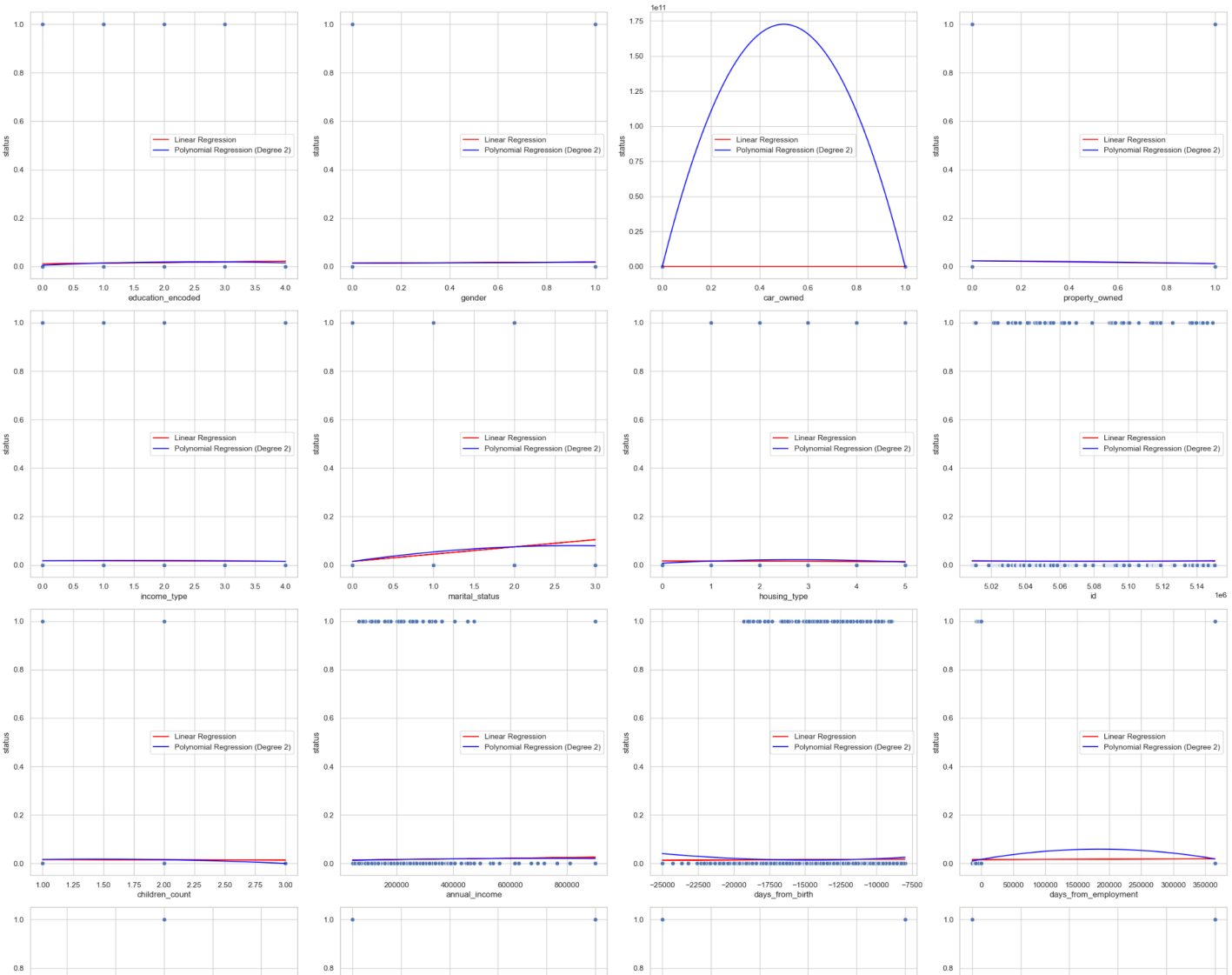
Feature Engineering:

- ☐ Changed the datatype of the columns
- ☐ Dropped duplicate rows
- ☐ Grouped columns into specific class- credit_record 'status' column into two classes ***0 and 1*** to determine the classification.
- ☐ Divided numerical and categorical columns
- ☐ Encoded categorical- ordinal and nominal to numerical columns using label/ordinal encoding
- ☐ Then merged the datasets into one final dataframe ***'credit_approval_df'***

Modelling:

- Checked for correction between the input variable and output variable
- Turned out that the correlation was non linear relationship hence we decided to go with XGBoost instead of Logistic regression as *XGBoost classifier has better classification report with non-linear correlation*

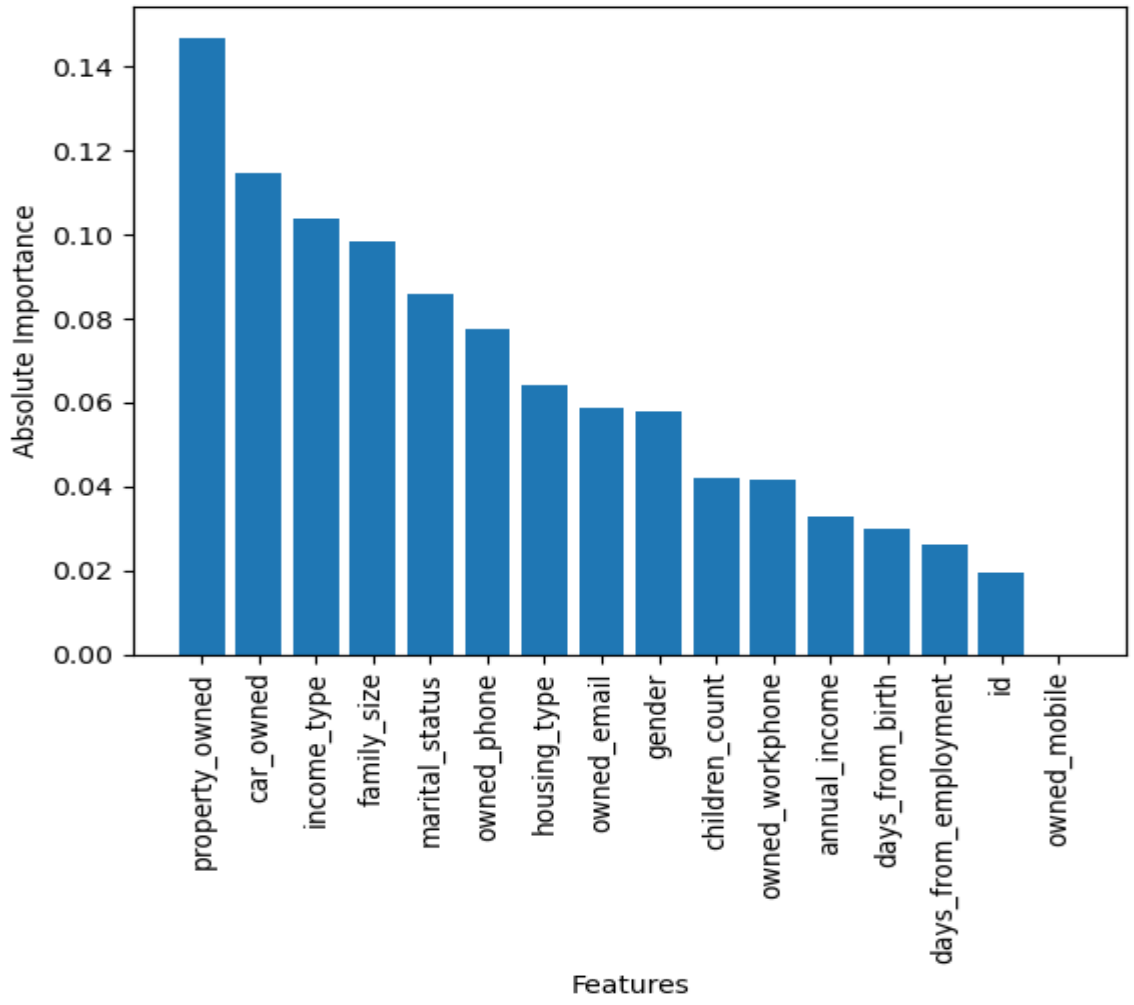
- Split **X and y train-test** and scaled the data using ***MinMaxScaler***
- Used smote() to deal with oversampling
- Tested the model with both Logistic Regression and XGBoost
- XGBoost had better results



Classification Report:

	precision	recall	f1-score	support
0	0.54	0.64	0.58	2810
1	0.56	0.46	0.50	2810
accuracy			0.55	5620
macro avg	0.55	0.55	0.54	5620
weighted avg	0.55	0.55	0.54	5620

Feature Importance:



Conclusion:

Feature Importance: Through feature importance analysis, it was identified that certain features, such as

1. income type
2. property owned
3. car_owned and
4. family size

have a significant impact on credit card approval. These features contribute the most to the predictive power of the model.

Model Performance: The predictive model used for credit card prediction demonstrated good performance classification report of

1. Accuracy : 93%
2. Precision : 99%
3. Recall : 88%

The model was able to correctly classify the majority of credit card applications as approved or rejected.

Important Factors: It was observed that factors such as income level, credit history, and employment status play crucial roles in determining credit card approval. Applicants with higher incomes, a positive credit history, and stable employment are more likely to be approved for a credit card.

Challenge:

- Using smote() to deal with the imbalance in the dataset as X and y have a 99:1 ratio, there's a possibility of having our model overfitted with the dataset.
- The synthetic samples created by SMOTE may not capture the same level of meaningful information as the original data.