

Personal Loan Prediction Model- Project



Content:

- **Business Problem and overview**
- **Data Overview**
- **Exploratory Data Analysis**
- **Model creating and Performance Analysis**
- **Comparison of Model Performance and Summary**
- **Business Insights and Recommendation**



Context:

- AllLife Bank is a US Bank with a growing customer base. The majority of these customers are liability customers (depositors) with varying sizes of deposits.
- The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans.
- In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).
- A campaign that the bank ran last year for liability customers showed a healthy conversion success rate of over 9%. This has encouraged the retail marketing department to devise campaigns with better target-marketing to increase the success ratio and asset customers



Business Solution:

- We will develop a Logistic Regression model analyzing the various features given for the customer to predict if the customer will borrow a personal loan or not.
- We will develop different Regression models and compare its overall efficiency and select the best model for the business.

Objective:

- To predict whether a liability customer will buy a personal loan or not.
- Which variables are most significant.
- Which segment of customers should be targeted more.

Key Questions:

- What are the Key variables that have a strong relationships with the dependent variable?
- Which metric is right for model performance evaluation and why?
- How accurate are the Model predictions and can it be improved?

Data Information:

Variable	Description
ID	Customer ID
Age	Customer's age in completed years
Experience	Years of work/professional Experience
ZIPCode	Home Address ZIP Code
Family	The Family size of the customer
CCAvg	Average spending on Credit Cards per month (in thousand dollars)
Education	Education level: 1. Undergraduate 2. Graduate 3. Advance/Professional
Mortgage	Value of house mortgage if any (in thousand dollars)
Personal Loan	Did the customer buy a personal loan from the previous campaign?
Securities_Account	Does the customer have a securities account?
CD_Account	Does the customer have a certificate of deposit account with the bank?
Online	Do customers use the banks online facilities?
CreditCard	Does the customer use a credit card issued by any other bank? (excluding ALLife Bank)

Data Overview

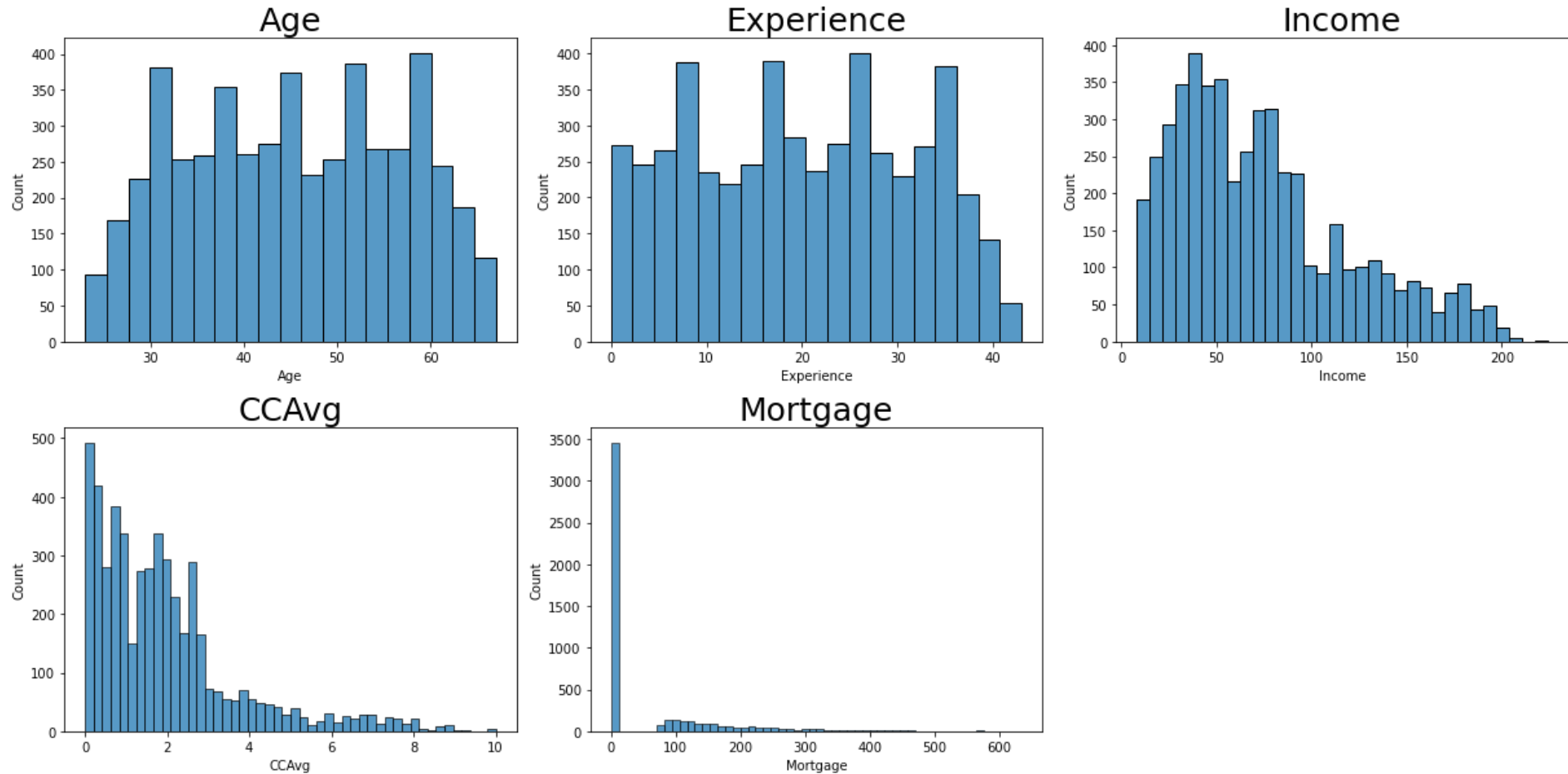
- There are total of 5000 observations of data across 14 variables.
- There were no missing values in the variables.

Data Manipulation:

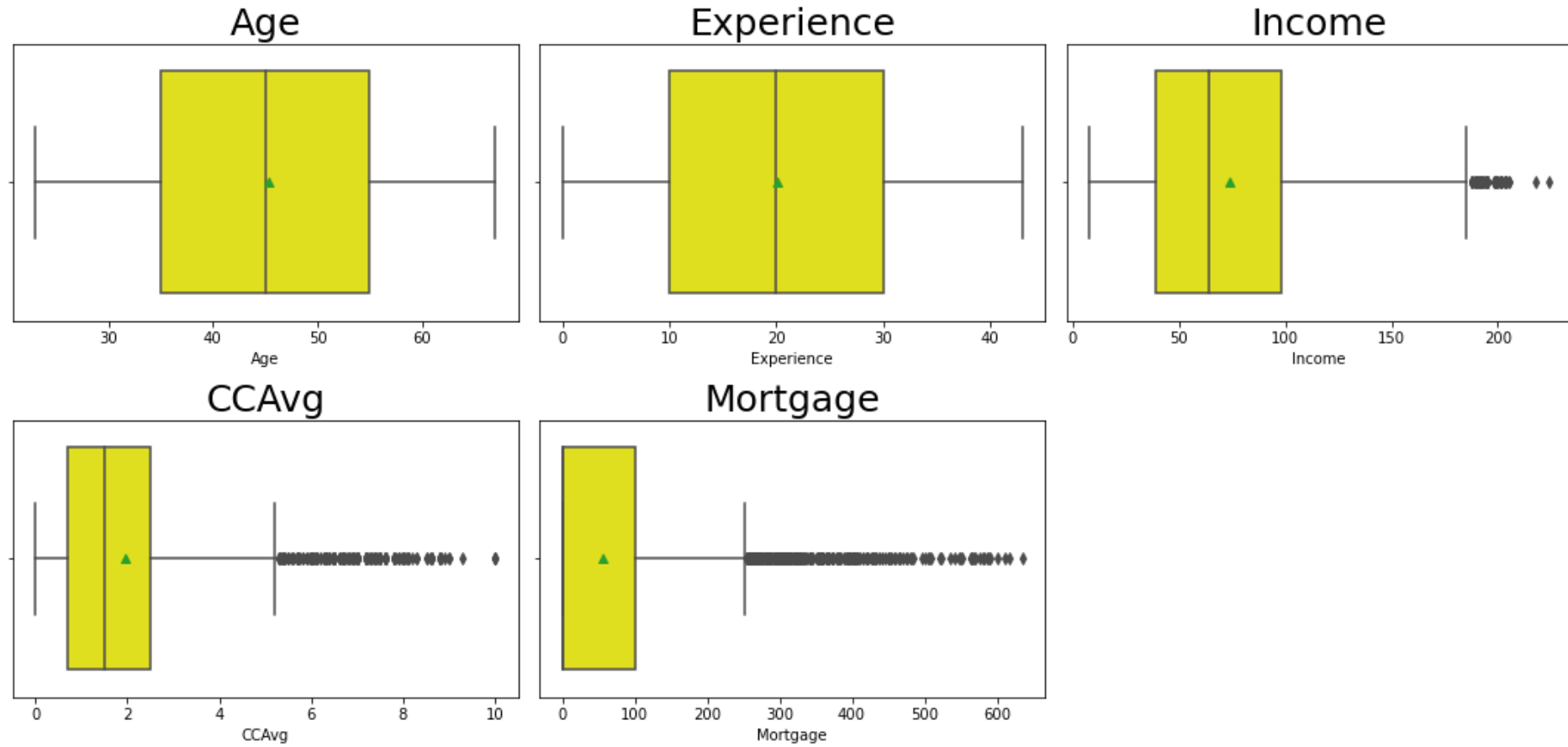
- The Dependent Variable Personal_Loan was converted to category as it only had two categories(values) 0 & 1
- Similarly, Securities_account, CD_account, Online and CreditCard were also converted from Integer to Category datatype for better model analysis
- Education, Family were also converted to category as they were label-encoded inputs of the its subsequent categories.
- Feature Engineering was applied to ZIPCode variable, and we were able to extract the county name where the customer was located. The 38 unique counties extracted were later grouped into 10 Regions, based on the 2020 California state census(<https://census.ca.gov/regions/>)
- ZIPCode, ID and county were dropped from further model building
- Experience column had negative values which were converted to positive.
- Income, Mortgage and CCAvg had several outlier values in the higher end which were treated using the Flooring and Capping method for model building

Exploratory Data Analysis: To analyze the individual variable for different cars.
In this data, there are seven numerical variables.

Univariate Analysis of Numerical Variables – Histogram Plot



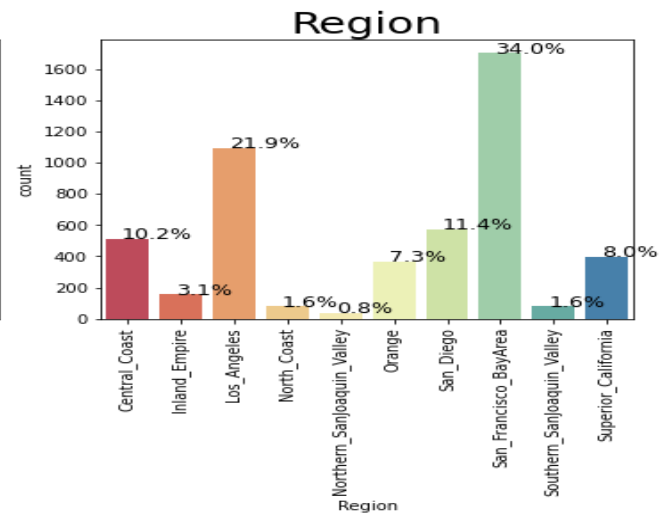
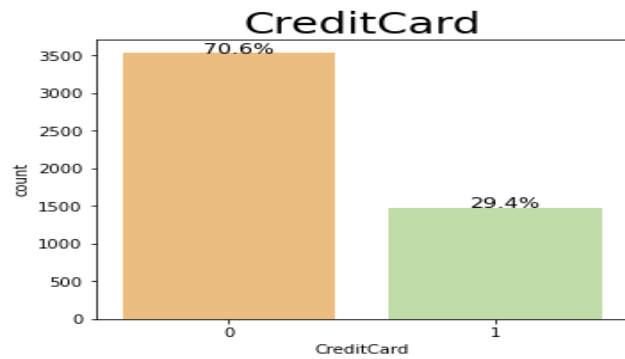
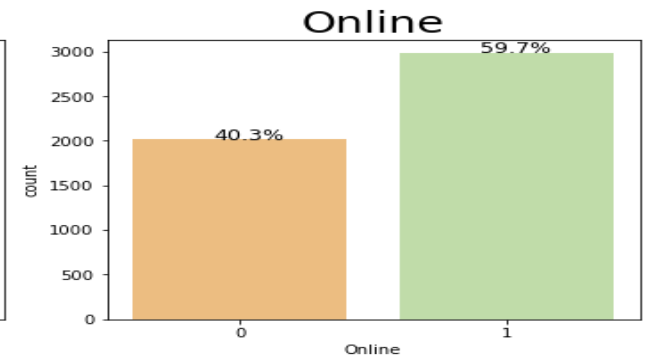
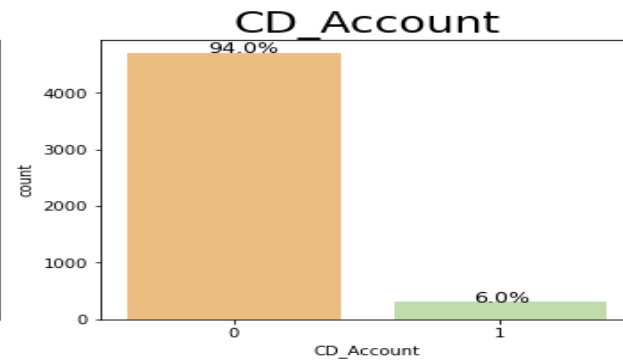
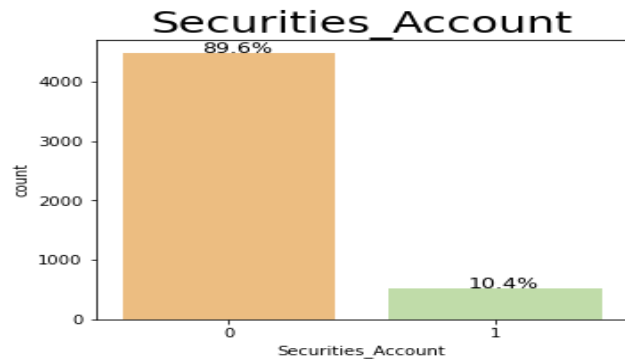
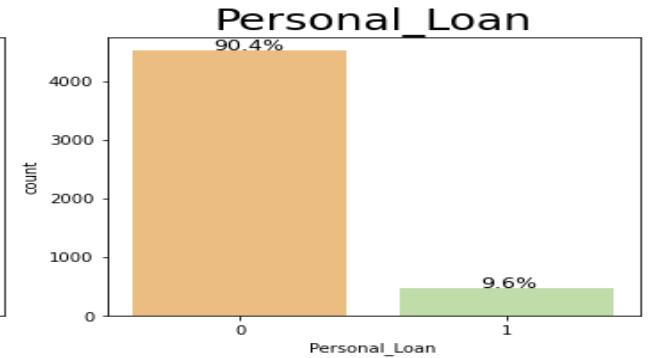
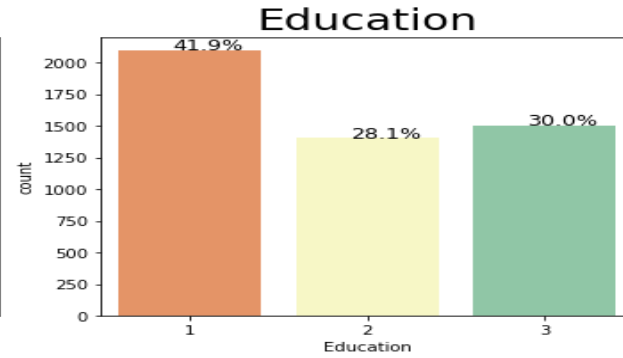
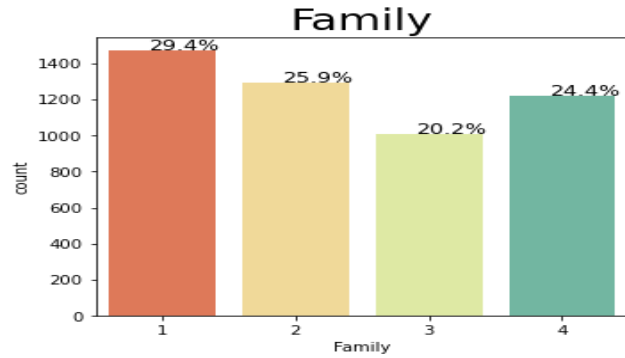
Univariate Analysis of Numerical Variables – BoxPlot



Univariate Analysis of Numerical Variables – Observations

- **Age** and **Experience** are almost normally distributed and look quite similar. This suggests a correlation between the two.
- There is skew-ness in the other three variables:
- **Income:**
 - Income shows the annual salary earned by the customer and is right-skewed in distribution. Majority of customers have income less than 100K, but there are several observations in the higher end.
- **Credit Card Average:**
 - CCAvg has several outliers in the higher end and is heavily right-skewed. Almost 75% of customers have an average of less than 2.5 (in thousand dollars). This suggests that some customers have very high charges compared to the rest.
- **Mortgage:**
 - The distribution in Mortgage variable is also heavily skewed. Almost 50 % of customers don't have a mortgage, indicating they don't own a home. We will have to analyze the mortgage for customers who only own a home to understand the distributions.

Univariate Analysis of Categorical Variables – CountPlot



Univariate Analysis of Categorical Variables – Observations

- 29.4 % of customers are of single-family household, with Family variable having four unique values.
- Education has three unique values with 41.9% of at Undergrad level(1).
- Personal_Loans is the Dependent variable and we see that there is heavy imbalance. Only 9.6% of customers in the data have accepted a loan from the previous campaign
- 89.6% of customers dont have a Securities account whereas 94% of customers dont have a CD account.
- We that 59.7 % of customers use the bank's online facilities and about 70.6% dont have credit cards issue by another bank.
- 34% of Customers are from the San_FrancisoBayArea and Los_Angeles has the next highest concentration at 21.9%

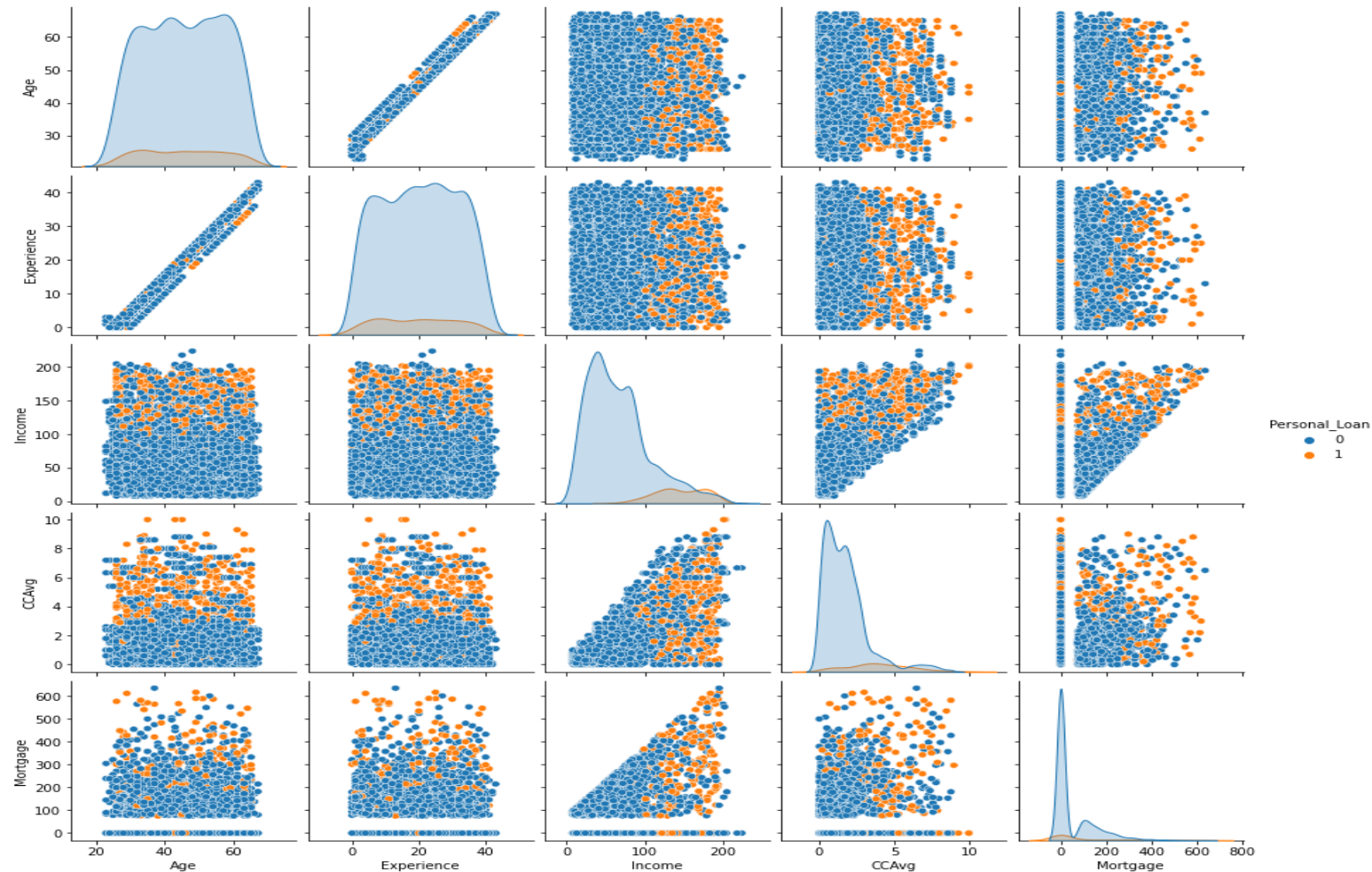
Exploratory Data Analysis : Correlation Matrix



•**Observations:**

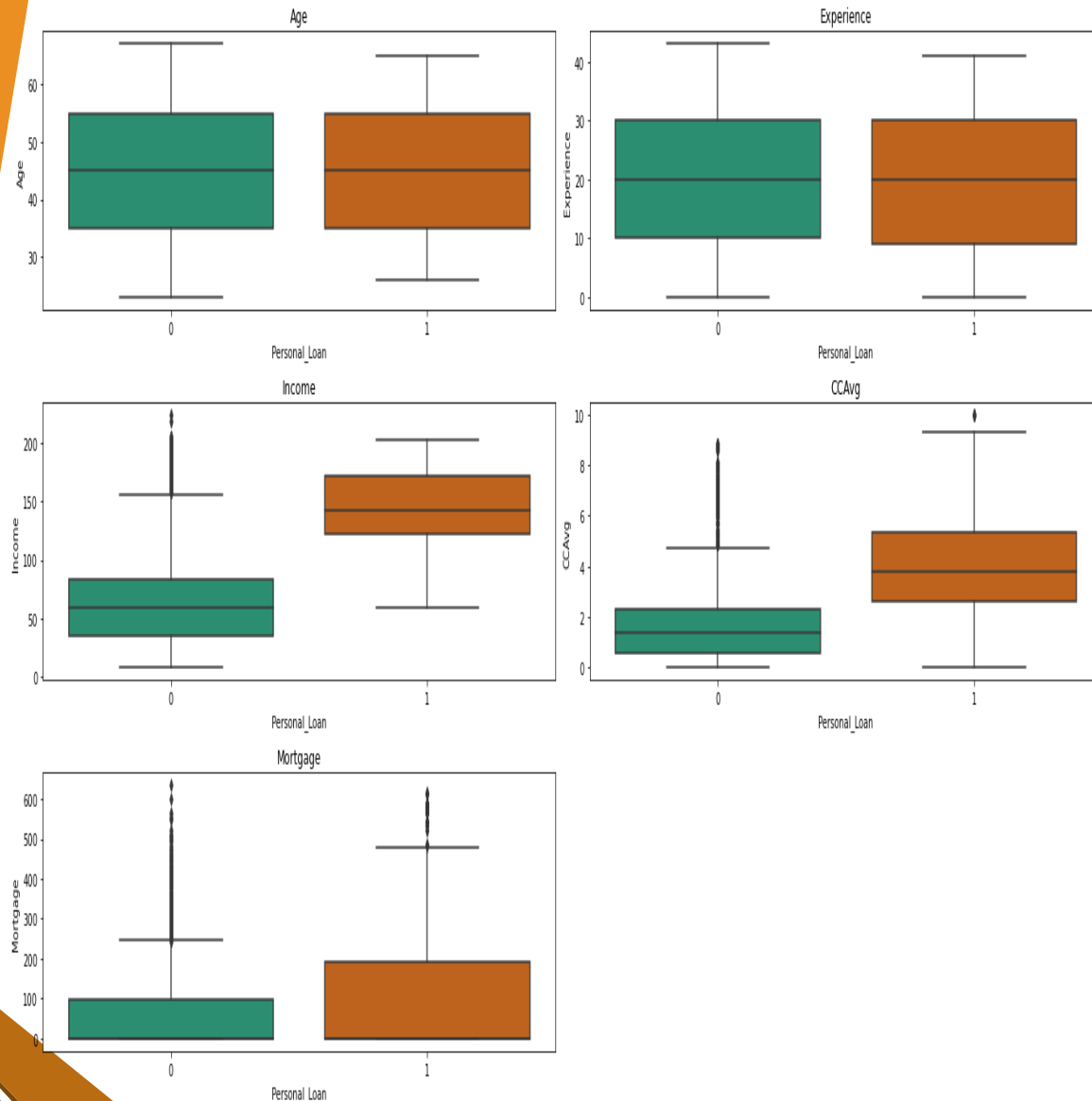
- Age and Experience have the highest correlation at 0.99. We suspect multi-collinearity between these variables
- Income and CC_Avg have the next highest positive correlation at 0.65. This suggests that customers with higher income have higher Credit card charges.
- Income and CCAvg have a positive correlation with Mortgage.

Exploratory Data Analysis : Pair Plot with dependent variable Personal_Loan



- The pair plot shows a more varying distribution in the variables between customers who took a loan and those who didnt.
- We see again that the distribution for Age and Experience is very similar. This could suggest possible multicollinearity
- There are overlaps that make it difficult to interpret who has personal loans and who doesnt, hence we will analyze further with other plots

Bivariate and Multivariate Analysis Variables:



**Numerical Variables with
Personal_Loan**

Observations:

- The mean values for Age is the same for both categories of Personal Loans
- Similarly the mean values for Experience is also almost equal for both categories of Personal Loan. Both these variables don't have any outliers
- Customers who have Personal Loans also have high Mean **Income and Credit Card Average** compared to customers who don't have a loan. Interesting we see several outliers in the higher end for both these variables in Class 0.
- The mean value for Mortgage at both levels is 0.0 (in dollars). This is because the majority of the customers don't have Mortgages. However, we see that customers with higher mortgages have Personal loans. But, we also see that there are several outliers in the high end again for customers who don't have a loan.
- The above plot suggests a correlation between Income, CCAvg, and Mortgage. Customers with high values for these variables have taken loans. This could suggest them as possible features of customers that can be targeted.

Bivariate and Multivariate Analysis Variables:

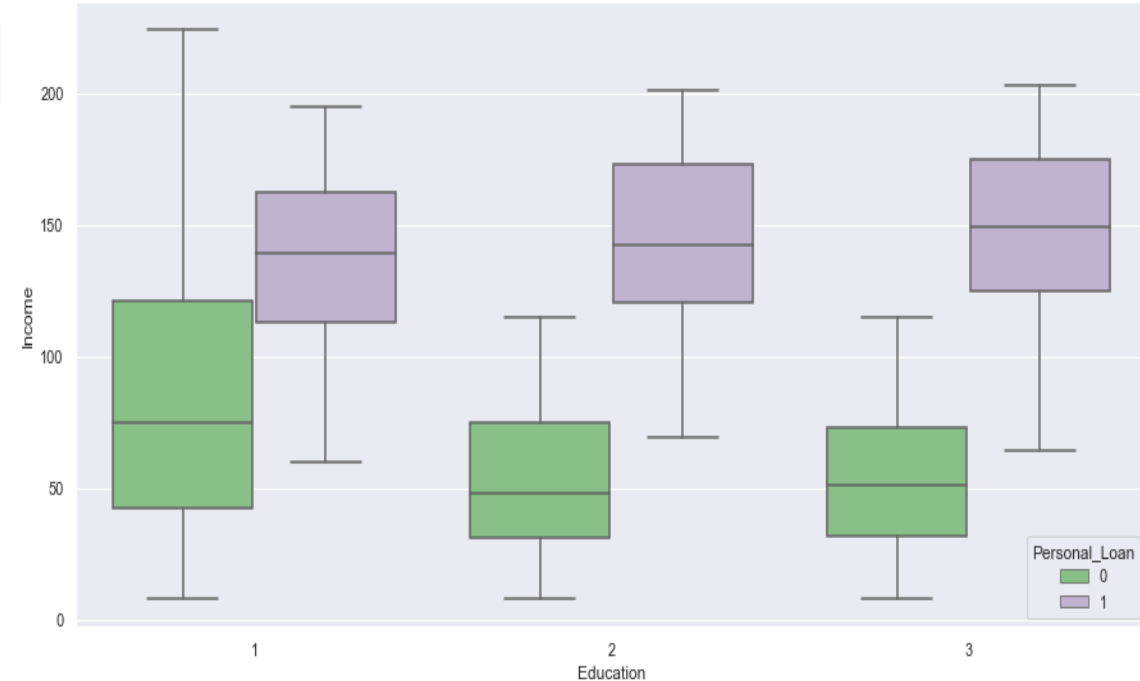
CD_Account Vs
Personal_Loan



•**Observations:**

- In the 302 customers have a CD_account, almost 50% have a Personal Loan
- This suggests that customers who have a CD_account are likely to buy loans and can be a possible target feature.

Education Vs Income Vs Personal Loan

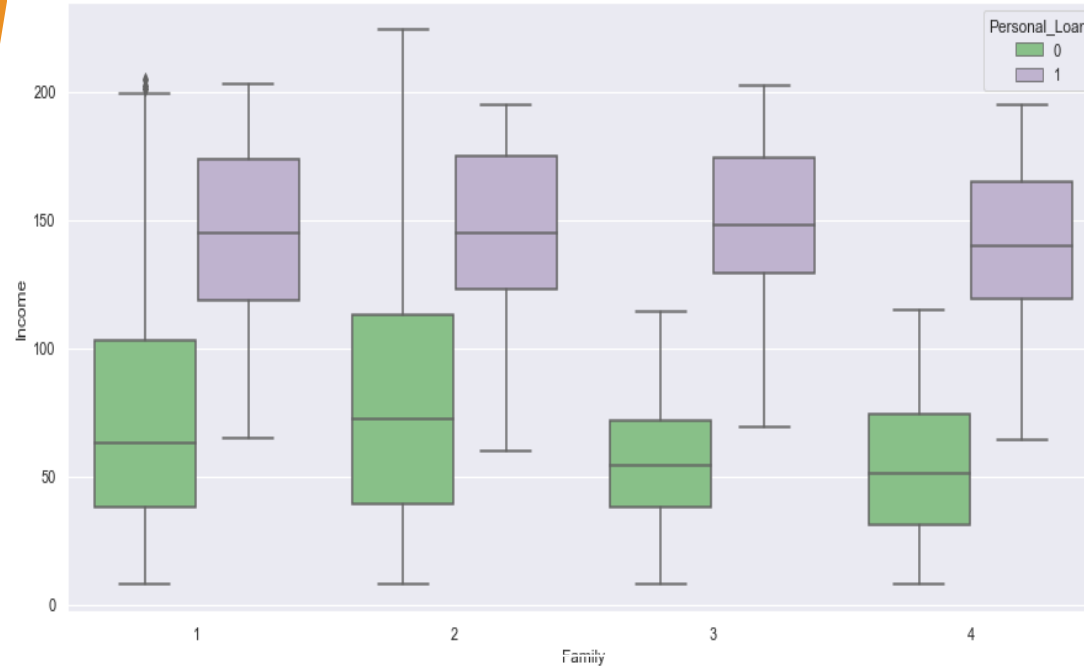


Observations:

- As Education level increases, Mean Income also increases.
- Customers with Education level 2 and 3 who have personal loans have a much higher mean income than Education level 1 customers

Bivariate and Multivariate Analysis Variables:

Price Vs Manufacture Year



Observations:

- Income level among all Family groups is significantly higher for customers who have a Personal Loan

Year Vs Kilometers_Driven



Observations:

More Customers with higher income and CCAvg > 2.5 (in thousand dollars) have personal loans.

Model Building - Overview

Model Evaluation Criteria:

Model can make two kinds of wrong predictions:

- Wrongly Identify customers as loan borrowers but they are not - False Positive
- Wrongly identifying customers as not borrowers but they actually buy loans - False Negative
- Since the Banks wants to identify all potential customers who will purchase a loan, the False Negative value must be less.

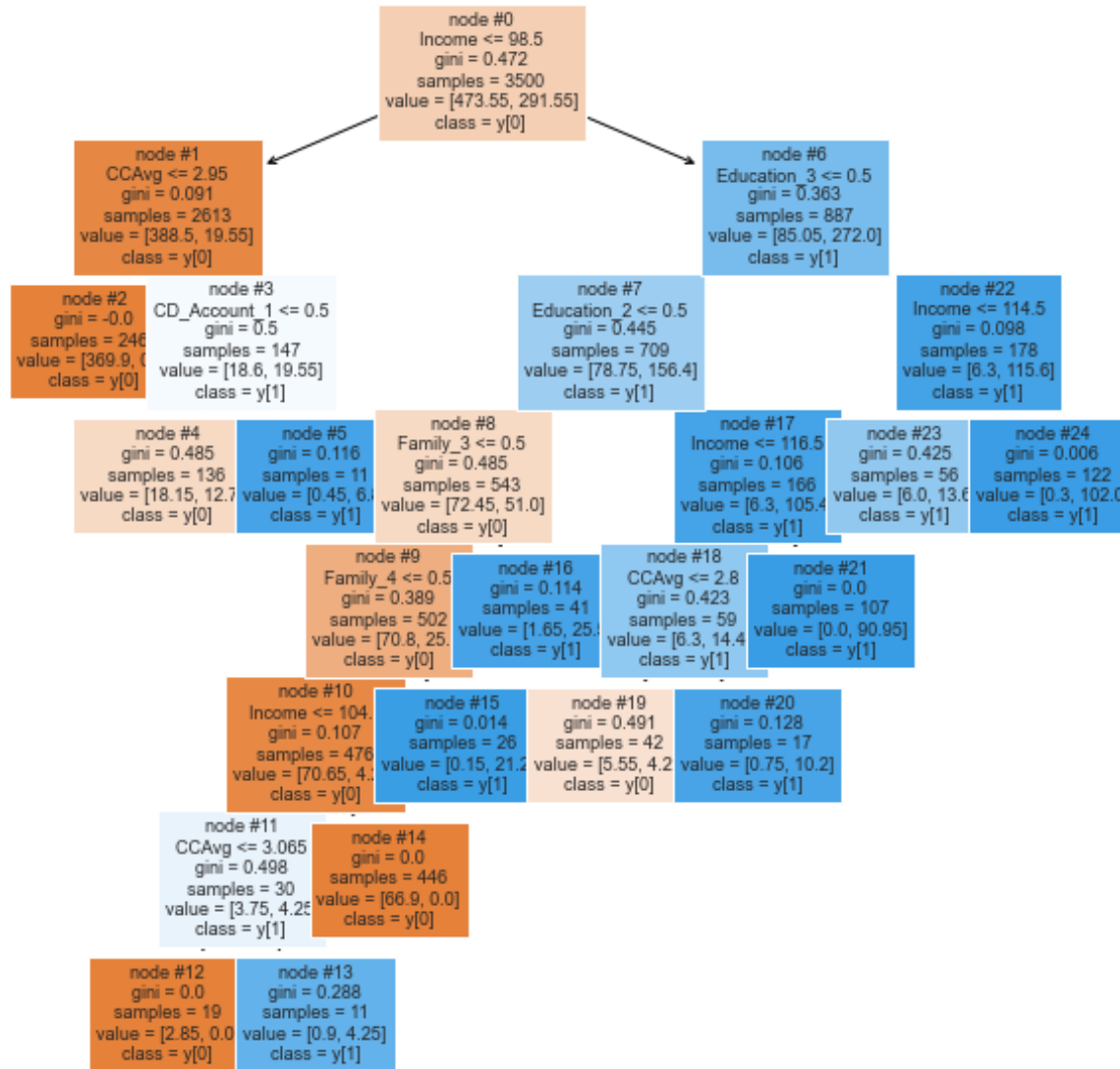
How to reduce losses:

- Recall is the Performance metric that must be improved.
- The Recall score must be maximized and greater the score the less the chance of missing potential customers.

Comparison of all Models for Personal_Loan Prediction

Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall
Logistic Regression Model-sklearn	0.94	0.94	0.52	0.46
Logistic Regression-Statsmodel-mutlicollineari...	0.96	0.97	0.73	0.68
Logistic Regression-Optimal Threshold =0.2017	0.95	0.94	0.85	0.83
Logistic Regression-Optimal Threshold =0.25	0.96	0.95	0.84	0.81
Sequential Feature Selction Method	0.94	0.94	0.48	0.51
Initial Decision Tree	1.00	0.99	1.00	0.91
Decision tree- hyperparameter tuning(pre-prun...	0.81	0.80	0.99	0.93
Decision tree- Cost Complexity post-pruning	0.97	0.97	0.94	0.93

Visualizing the Tree Model



Conclusion:

- From the table we see that the Decision Tree - Cost Complexity(Post-Pruning) model with impurity alpha value = 0.03 ;has the best Accuracy of 97% and Recall of 93.4%. This is the best model for this dataset.
- The most important features from this model are:
 - Income
 - Education_2 (Graduate)
 - Family_4 (family of 4)
 - CCAvg (Credit Card Average)
 - CD_Account1 (Customers with a CD account)
- This model is also not affected by outliers or extreme values

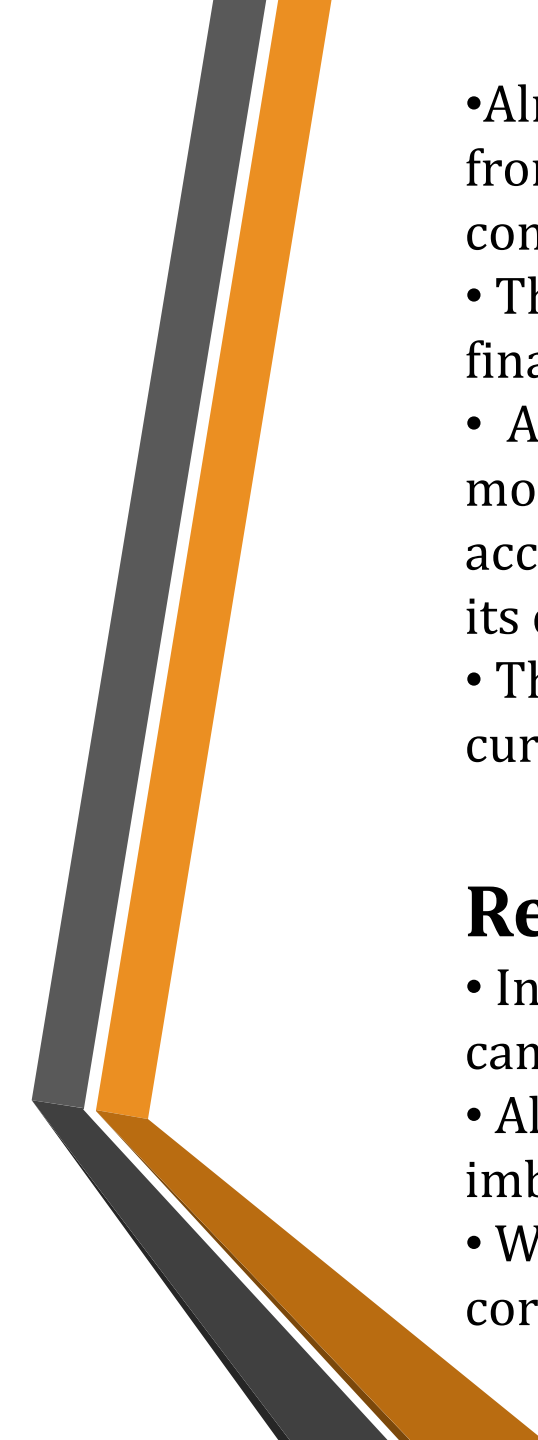


Business Insights

- Income has been the consistent variable across all the models built. This shows that it the most important variable to consider.
- From the final model, we see the following segments/variable combinations of customers who have a strong probability of borrowing loans.
 - Higher Income and Higher Credit Card Average
 - Higher Education levels (2 &3) and Higher Income
 - Larger Family size and high income
 - Customers with a CDAccount
- These potential customer segments can be targeted by the marketing team to increase the conversion rate.

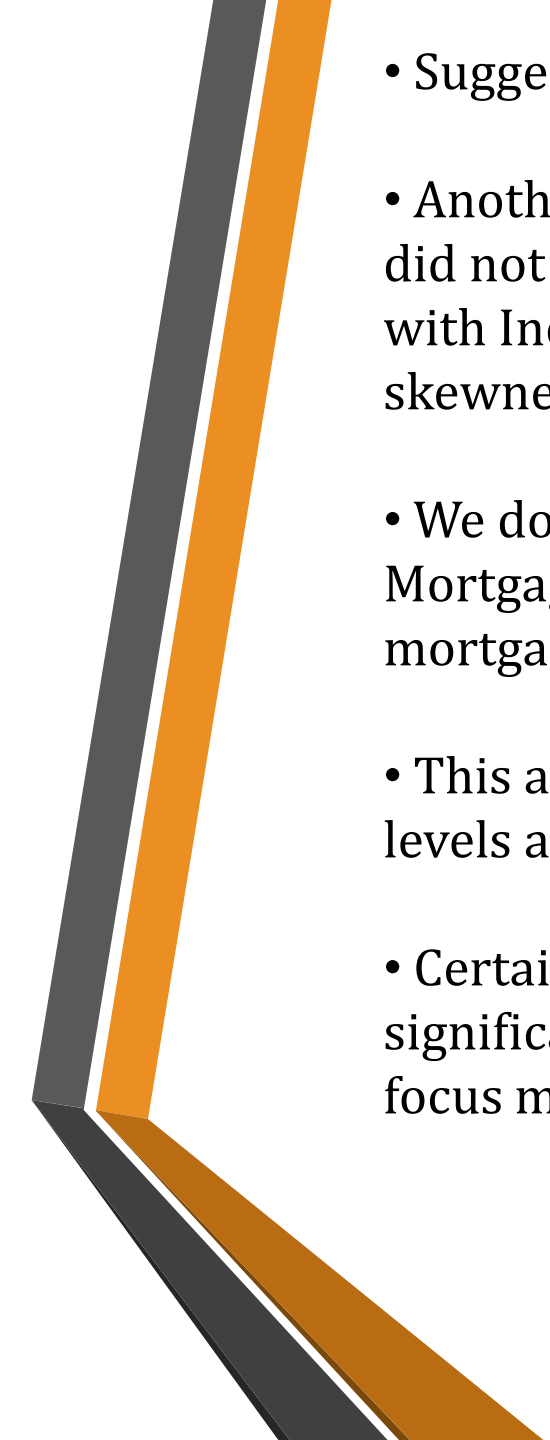
Possible recommendations to the campaign and business:

- A customer with higher income has a better rate of paying back the loans, hence such customers should be treated with more incentives in the campaign.
- Target customers with high education levels as we see that such customer tend to have high income paying jobs.
- Customers with high credit card average are also potential target. Hence the bank can offer suitable combination offers to the bank's credit card along with purchasing a loan. These decision require careful consideration of the bank's lending policies.
- Offering Customers with a larger family and high income flow, incentives beneficial to the whole family rather than just the customer.

- 
- Almost 50% of customers who have a CD account purchased a loan. Any material from the previous campaign that centered on this area can be improved to increase conversion rate.
 - The bank may offer lower interest rates at larger payment periods with possible financial investment suggestions to improve customer retention.
 - A key idea would be to help educate/provide insights to the customers on better money management, saving and refinancing solutions, to their existing savings account. By ensuring the customer engagement, the bank can increase its loyalty with its customers.
 - The Bank and its marketing campaign must ensure that they are updated on all the current social and economic conditions of the market while promoting.

Recommendations:

- In this dataset, only 9.6% of total customers purchased a loan after the last campaign.
- Although we were able to build a 97% accuracy model with a high recall, the imbalance in the data is very high.
- We do not know how the model would predict if the imbalance were to be corrected.

- 
- Suggested recommendations would be to decrease the imbalance in the dataset.
 - Another imbalance in feature was on Mortgage. More than 50% of customers in this dataset did not have a house/mortgage. Even though we saw that mortgage had a positive correlation with Income and customers with higher mortgages also purchased loans, there was heavy skewness in this variable and it was not considered an important feature in the final model
 - We do not know if this is due to its skewness or due to high zero mortgage values. Since Mortgage is also a type of loan, a better approach would be to separate customers at various mortgage levels and then build a predicting model.
 - This approach may also bring the location / ZIPCode variable into consideration as Mortgage levels are related to the location of the house.
 - Certain variables(i.e Online, Securities Account and CreditCard) in the dataset had no significance to the dependent variable. Future data collected can ignore these variables and focus more on correcting the imbalance.



Thank you