Business Presentation

Personal Loan Campaign Case Study

Background & Context

- **AllLife Bank** is a US bank that **has a growing customer base**.
- The **majority** of the Bank 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.
- A **campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success**. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio.
- **The Bank management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).**

Objective

- **Explore and visualize data**
- **Build a model**

  Model should be able to

  - 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.
- **Draw Conclusions and Business Recommendations**

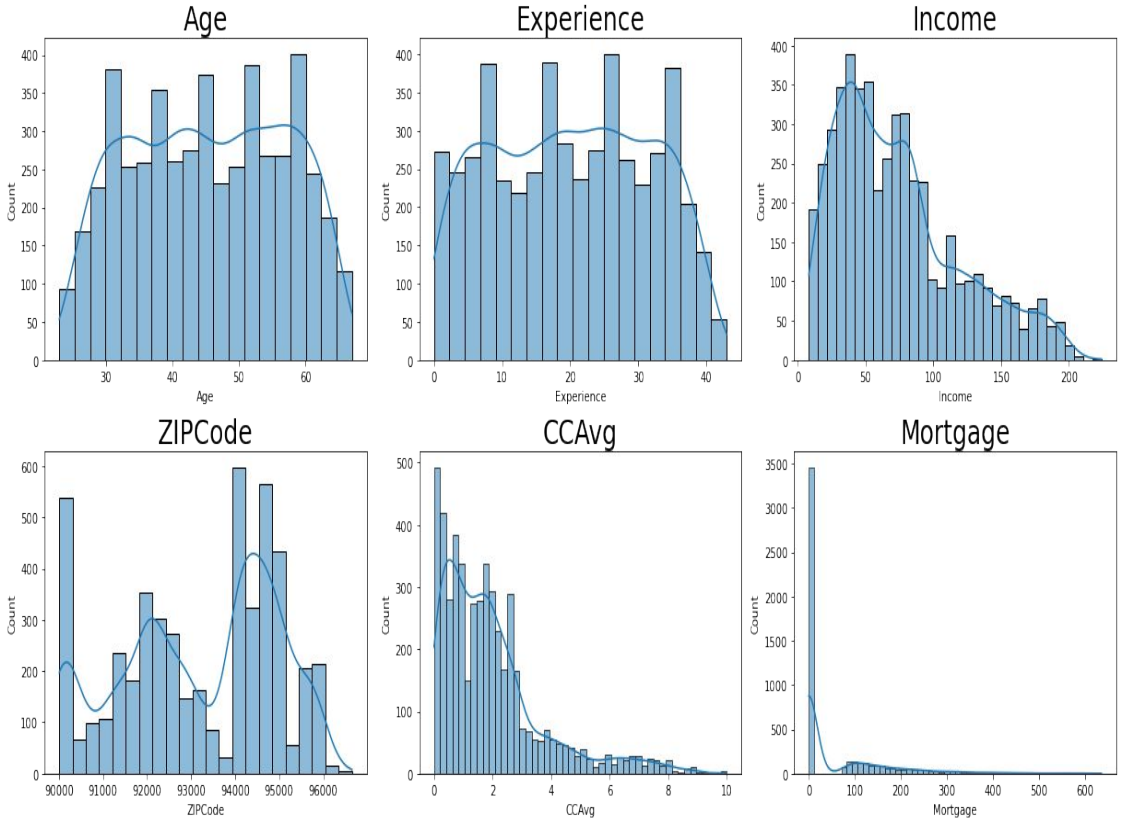
Data Overview

- Data contains 5000 (rows) customer data with 14 (columns) characteristics.
- There are no null values or duplicates in dataset.
- The ID column is of no significance and will be removed for analysis.
- The ZIP Code will be mapped to respective counties for analysis.
- Age, Experience, Income, Mortgage, CCAvg will be binned and mapped to new range columns respectively.
- There are negative values in Experience column which we will be imputing for analysis.

Variable	Description
ID	Customer ID
Age	Customer's age in completed years
Experience	#years of professional experience
Income	Annual income of the customer (in thousand dollars)
ZIP Code	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: Undergrad; 2: Graduate;3: Advanced/Professional
Mortgage	Value of house mortgage if any. (in thousand dollars)
Personal_Loan	Did this customer accept the personal loan offered in the last campaign?
Securities_Account	Does the customer have securities account with the bank?
CD_Account	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Do customers use internet banking facilities?
CreditCard	Does the customer use a credit card issued by any other Bank (excluding All life Bank)?

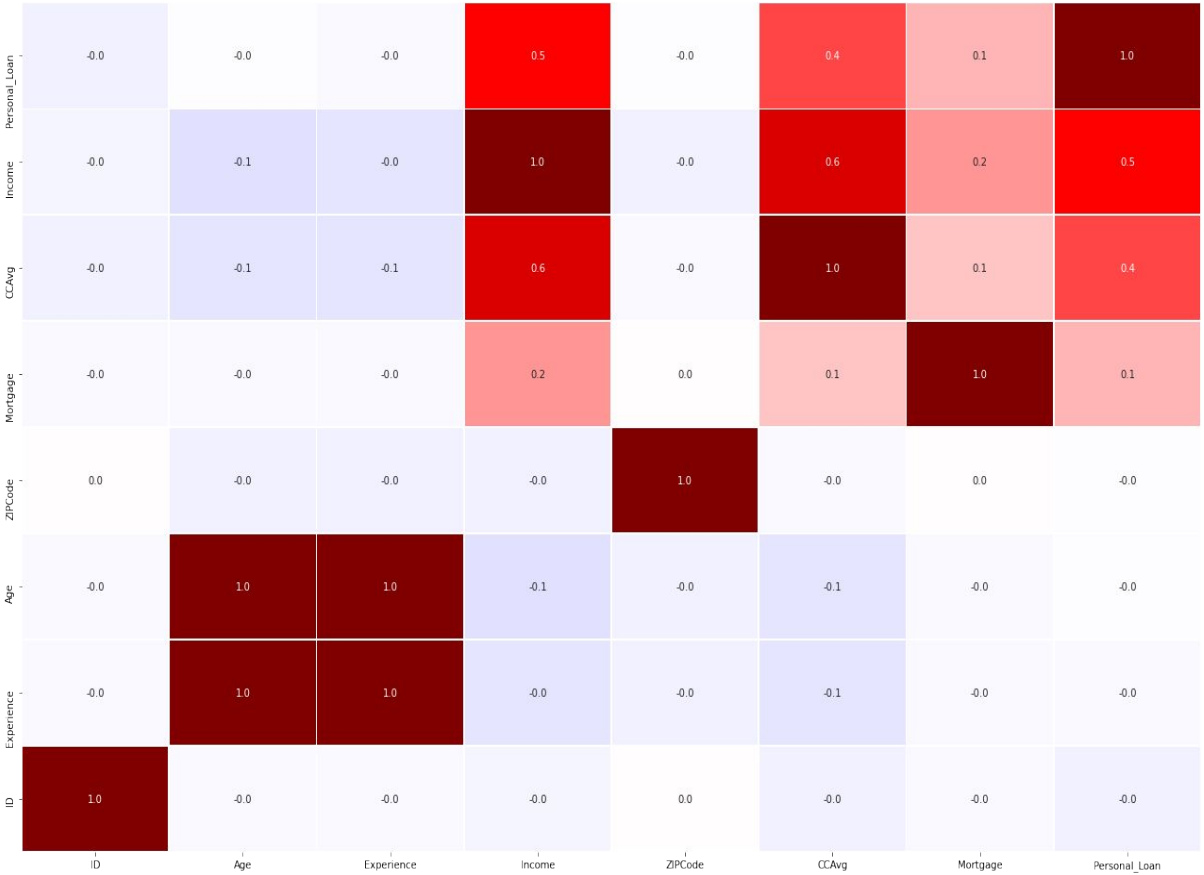
Exploratory Data Analysis - Univariate

- Age and Experience almost have non skewed distribution.
- CCAvg is heavily right skewed indicating outliers that needs to be capped and treated.Around 81% of the customers average spending is less than 3k a month.
- Income is right skewed indicating outliers.Around 75% of the customers have an Income below 100K and remaining 25% has greater than 100K
- Nearly 80% of the customers do not have any Mortgage.Because of 80% zero value, the mean of the distribution is skewed to 60K.



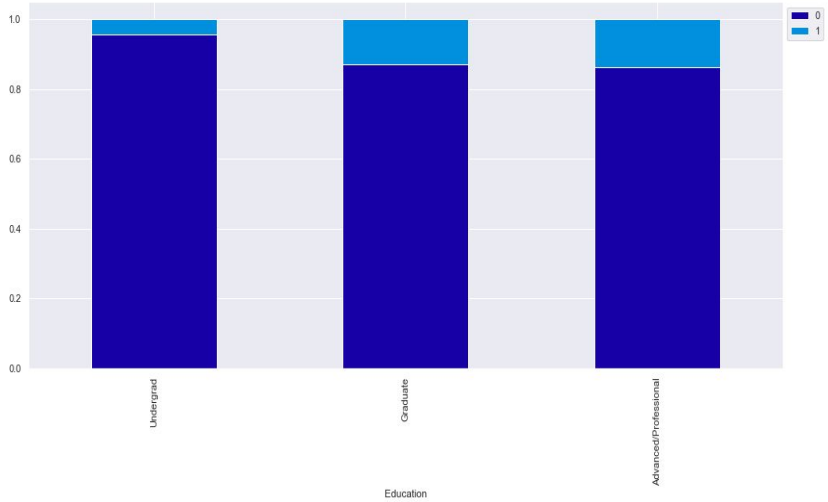
Exploratory Data Analysis - Bivariate

- Age and Experience are strongly correlated and we can drop one of these columns during modelling.
- CCAvg and Income is also 0.6 correlated.
- Personal Loan and Income have a correlation of 0.5.



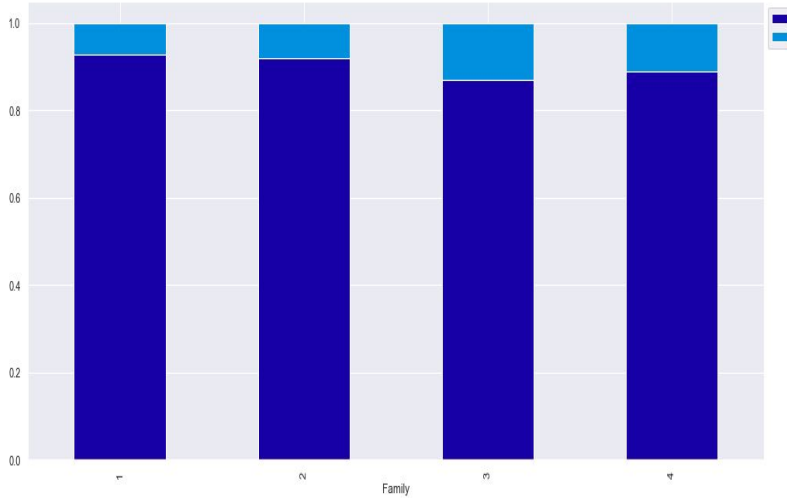
Exploratory Data Analysis - Bivariate

Education Level vs Personal Loan



- Customers with Advanced/Professional Education are the most(13.65%) who has got a Personal Loan very closely followed by Graduates category(12.97%)

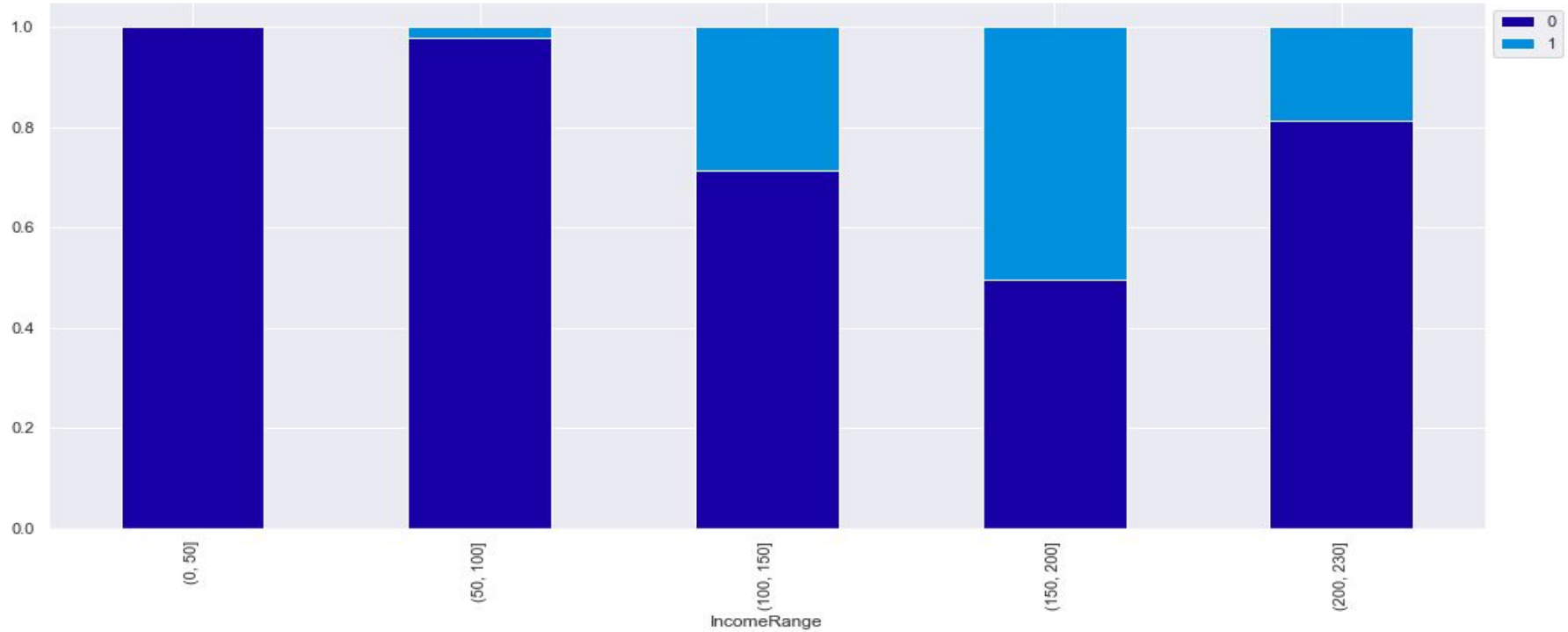
Family Size vs Personal Loan



- Customer family size of 3 are the most to buy the personal loan closely followed by family of 4 members

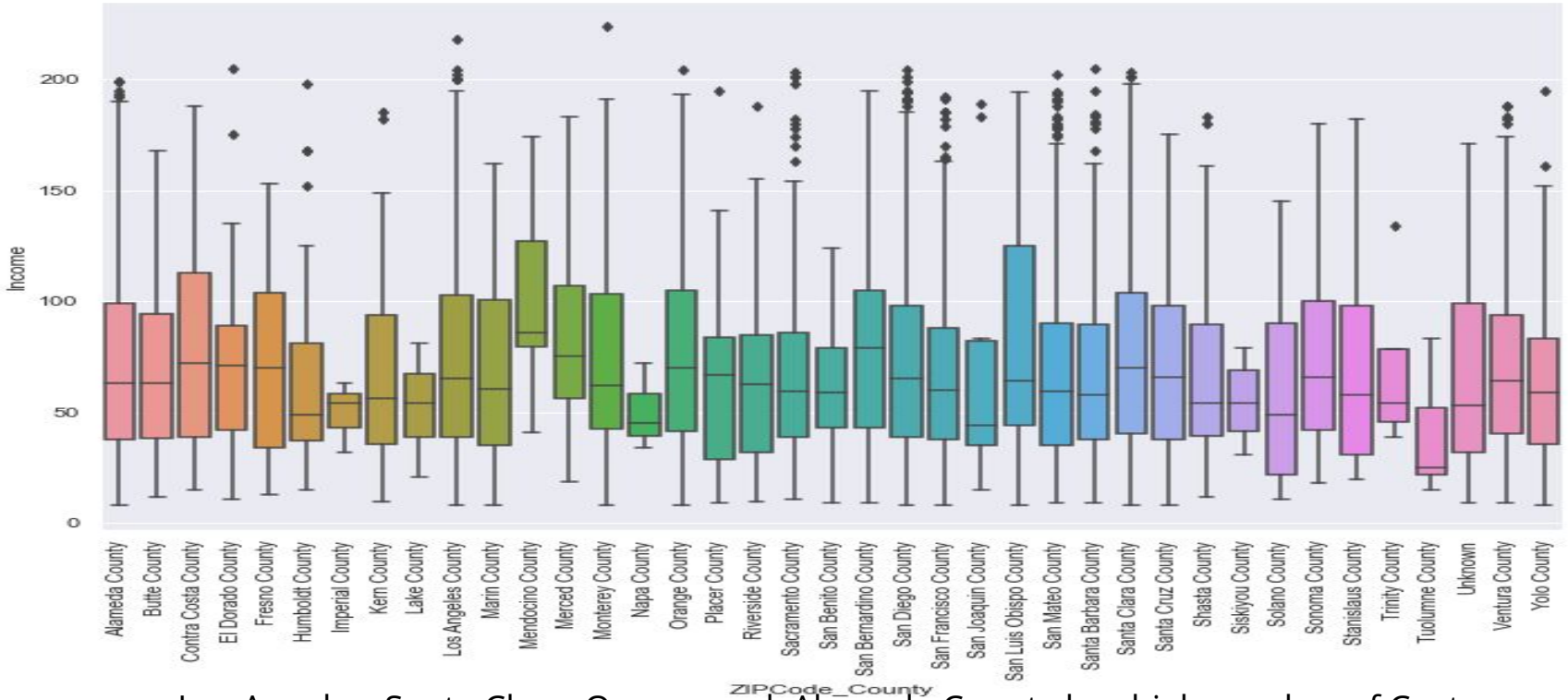
Exploratory Data Analysis - Bivariate

Income vs Personal Loan



- Most of customers who has bought a personal loan have an Income greater than 100k.

Exploratory Data Analysis - Bivariate



- Los Angeles, Santa Clara, Orange and ,Alameda County has high number of Customers as well as high income customers.

Assumptions - Logistic Regression

1. **Multicollinearity :**
   - Age and Experience were strongly correlated and removing them helped bringing down the VIF scores
   - ZIPCode_County also had multicollinearity issues and dropping that helped the model performance
2. The **outliers** in the numeric variables Income , CCAvg , Mortgage were treated using capping method before modelling.
3. **Dependent variable** is Personal Loan and is in **Binary form 0 and 1** satisfying the assumption
4. Logistic regression requires the **observations** to be **independent of each other** and we have confirmed that there is no data dependency.Unique data no duplicates.
5. **Larger Sample size** of 3500 train data vs 1500 test data was sufficient to build our model.

Model Performance Summary - Logistic Regression

	Model	Train_Accuracy	Test_Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
0	Logistic Regression Model - Statsmodels	0.969143	0.960	0.740181	0.657718	0.917603	0.915888	0.819398	0.765625
1	Logistic Regression - Optimal threshold = 0 .09	0.918000	0.920	0.897281	0.859060	0.540000	0.563877	0.674234	0.680851
2	Logistic Regression - Optimal threshold = 0 .35	0.962857	0.958	0.794562	0.724832	0.809231	0.830769	0.801829	0.774194

- **Our data has imbalanced class distribution and hence F1 score will be the right metric to use for Logistic Regression.**
- **We have been able to build a predictive model that can be used by the bank to find the customers who will buy a Personal Loan with an F1_score of 0.80 on the training set and 0.77 on test set (Logistic Regression - Precision-Recall Optimal threshold = 0.35 - with significant predictors).**
- **We tried to do Model Improvement using ROC-AUC threshold(0.09) and Precision-Recall curve threshold methods and Precision-Recall threshold of 0.35 gave the highest F1 score.**

Model Performance Summary - Logistic Regression

Dependent variable = Personal_Loan

	odds
const	4.717156e-07
Income	1.068642e+00
CCAvg	1.682016e+00
Family_3	1.519469e+01
Family_4	5.705258e+00
Education_Graduate	6.732155e+01
Education_Advanced/Professional	8.733981e+01
Securities_Account_1	3.641269e-01
CD_Account_1	3.916144e+01
Online_1	5.557446e-01
CreditCard_1	3.792363e-01

The coefficients of the logistic regression model are in terms of log(odd), to find the odds we have to take the exponential of the coefficients.

Therefore, odds = exp(b)

Income , Family size of 3 and 4, CCAvg, Education level of Graduates and Advanced/Professional(2&3),CD account Securities account,Online and CreditCard are the important predictor variables for this model .

Co-efficient Interpretation:

***Income:** Holding all other features constant a 1 unit change in Income will increase the odds of a customer buying a Personal Loan by 1.06 times.

* **CCAvg:** Holding all other features constant a 1 unit change in the CCAvg will increase the odds of a customer buying a Personal Loan by 1.68 times

* **Securities_Account_1:** Holding all other features constant a 1 unit change in the Securities_Account_1 will decrease the odds of a customer buying a Personal Loan by 0.36 times .

* Similarly negative(decrease) and positive(increase) co-efficients for other variables can be interpreted.

Model Performance Summary - Decision Tree

	Model	Train_Recall	Test_Recall
0	Initial decision tree model	1.00	0.85
1	Decision tree with hyperparameter tuning	0.87	0.73
2	Decision tree with post-pruning	0.98	0.97

- **In our case of predicting Personal Loan Buyers ,not being able to identify a potential customer is the biggest loss Business can face.**
- **Recall is the right metric to check the performance of the model.**
- **We tried hyperparameter tuning technique as well but the recall scores did not improve on the test data.**
- **Using the Decision tree Classifier with post-pruning technique with cc_alpha of 0.0067, we got the highest recall results and we were able to predict the False Negatives at 0.2% and False positives at 6% which gives us a very good reliable model with low error rates.**

Logistic Regression vs Decision Tree Classifier- Confusion Matrix Comparison

Model Name	True Positives	True Negatives	False Positives	False Negatives
Logistic Regression	7.20%	88.60%	1.47%	2.73%
Decision Tree Classifier	9.73%	84.07%	6%	0.20%

- **As False Negatives are opportunity cost, it is more expensive in our case than False positives which can be better planned and handled.**
- **Hence Decision Tree Classifier gives the best results with low False negative rates and highest recall on test data.**

**** True Positives:***
Reality: A customer buys a loan.
Model predicted: The liability customer will get converted to a loan customer buying a loan.
Outcome: The model is good.

**** True Negatives:***
Reality: A customer did NOT buy a loan.
Model predicted: The liability customer will NOT get converted to loan customer.
Outcome: The business is unaffected.

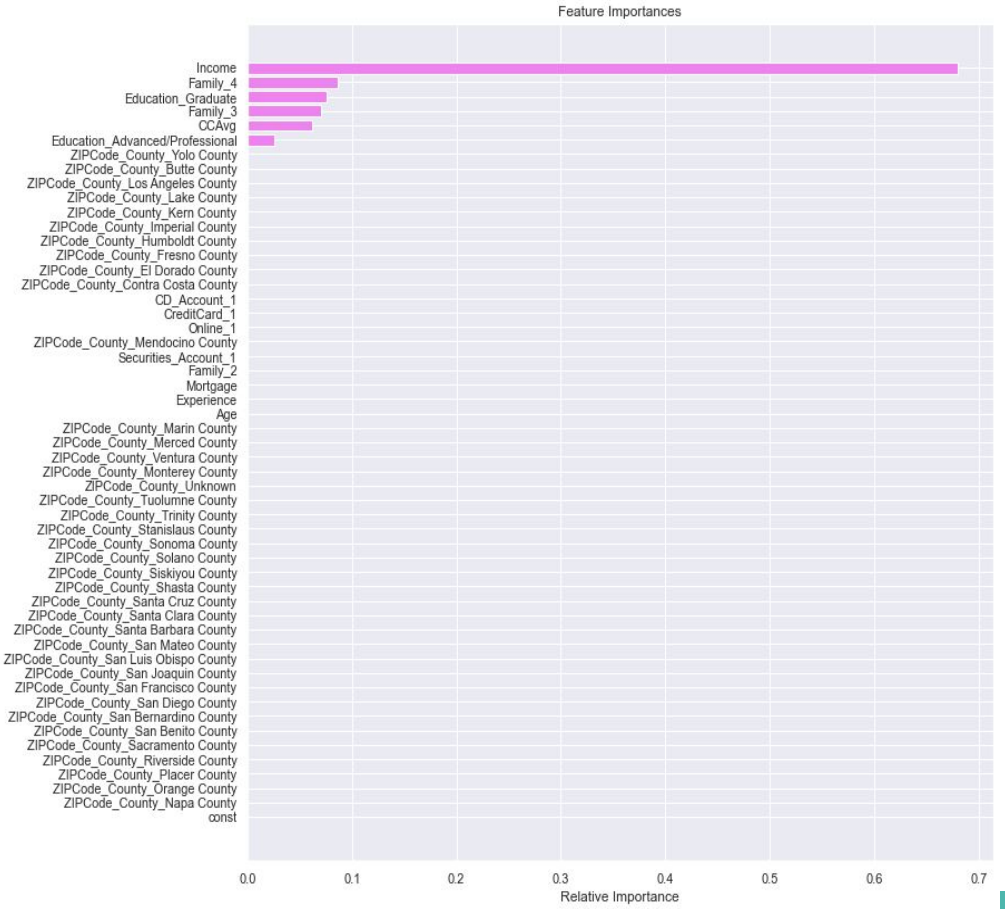
**** False Positives:***
Reality: A customer did NOT buy a loan.
Model predicted: The customer will get converted to a loan customer buying a loan.
Outcome: The team which is targeting the potential customers will be wasting their resources on the people/customers which will not be a very big loss compared to losing a customer who will buy a loan.

**** False Negatives:***
Reality: A customer buys a loan.
Model predicted: The customer will NOT buy a loan.
Outcome: The potential customer is missed by the sales/marketing team, the team could have offered the potential customer some discount or loyalty card to make the customer come again to purchase. (Customer retention will get affected.)

Feature Importance - Decision Tree

The model indicates that the most significant predictors of Potential Loan buyers are

1. **Income** greater than 92.5K dollars
2. **Family Size of 4 and 3** members
3. **Education level of 2 and 3** - Graduates and Advanced/Professionals
4. Average **Credit card usage** greater than **2.9k** a month



Conclusion

- **Decision Tree Classifier with post pruning technique** gave us the best model compared to Logistic regression with the **highest recall score of 98% on train and 97% on test data** and the least error of 0.2% False negatives and 6% False positives.
- **Income** being the top most feature to look at, **every 1 unit increase will increase the odds of customer buying a Personal loan by 1.06 times.**
- **Income, Family size 3 & 4, Education level 2 & 3(Graduates and Advanced Professionals), Credit card usage** greater than 2.9k are the key variables that has strong relationship with the dependent variable for the next campaign.
- The **bivariate results of EDA clearly matched with the the Decision Tree predicted important variables** and their parameters.

Recommendations

- If a customer's **Income** is greater than **92.5k** and his **Education level** is **Advanced/Professional/Graduate(level 3 or level 2)** then there is a very high chance that the **customer is going to buy a loan from the bank.**
- It is observed that the **family size of 4 and 3 members** has the likelihood of buying a loan.Those customers can be targeted by the marketing team as potential customers.
- The **Average Credit card usage of a customer is greater than 2900 USD** a month, those customers can also be targeted for loan.
- Employ the predictive model to predict potential customers (customers who can buy the product), and market Offers and deals on a real-time basis only to those customers.
- It is observed that **60% of the customers have online account**.Hence making attractive advertisements online with competitive offers/deals can attract more customers to buy the loan.
- 22% of Customers are from Los Angeles County that has the maximum number of customers.San Diego(11.4%) and Santa Clara(11.3%) is in second place for the count of customers.These have higher Income group customers as well, if we could device good marketing strategies, we can get a lot of conversions here too.