ALL LIFE BANK PROJECT

2021

OBJECTIVES

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

BUSINESS PROBLEM OVERVIEW

BACKGROUND

- AllLife Bank is a US Bank that seeks to grow is Assets(Loan) base by capitalizing on several factors that could influence same.
- As a senior data scientist at AllLife Bank, I am tasked with coming up with a model which the marketing team would adopt as a tool to design the marketing campaigns and target potential customers for the purchase of Personal loans

SOLUTION APPROACH (MACHINE LEARNING)

- Define the problem and perform an Exploratory Data Analysis
- Illustrate the insights based on EDA
- Data pre-processing
- Model building - Logistics Regression and CART
- Establish Model Parameters
- Model performance evaluation and Improvement
- Return the Most Import Features/Variables influencing Personal Loan Purchase
- Actionable Insights & Recommendations

BUSINESS IMPLICATIONS

- Maximize Processing and Management Fees which impacts on AllLife's bottom-line
- Increase Customer retention and traffic and ultimately unprecedented deposits
- Targeted Marketing and Media Campaigns
- Minimize the cost of funds via increased personal loan asset offer
- Optimize Sales and Profit Margin
- Effective Allocation and redistribution of resources especially for product design, pricing and Ad-campaigns
- Developing potential markets for more revenue

DATA MANIPULATION

- Identification of Missing Values
- Fixing Columns: (ID and ZIPCode) were dropped ultimately prior to Modelling
- Fixing the ordered variable ,Education and converting ultimately to Categorical data type
- Conversion of Data types
- Treating Negative Values in the ’Experience’ variable and dropping the rows containing them prior to Modelling as well
- Prior to modelling, Year was dropped in place of a new variable, Age created for a cleaner prediction

DATA INFORMATION

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)
ZIPCode	Home Address ZIP codeImportant features by models
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)

Observations	Variables
5000	14

Float64	Int64
1 CCAvg	13 Name Location Fuel_Type Transmission Owner_Type Mileage Engine Power New_Price

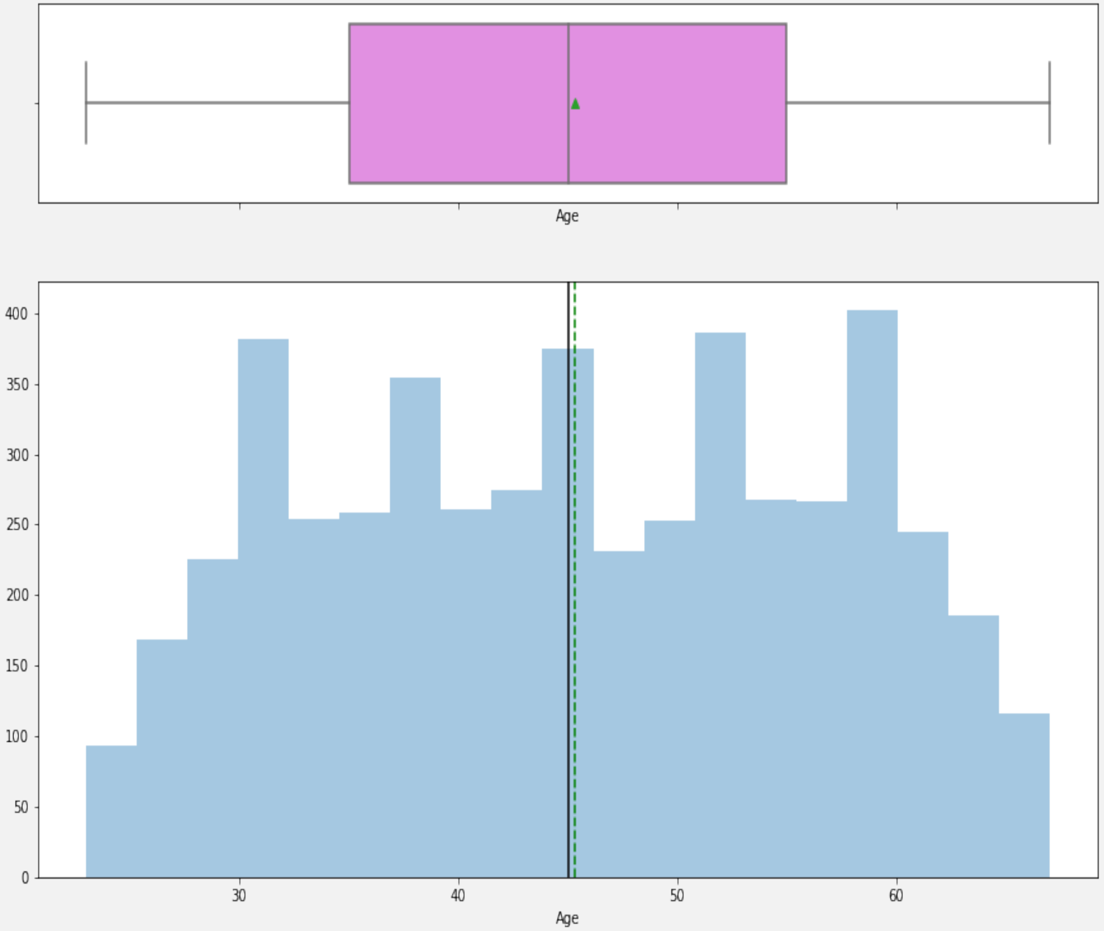
Missing Values in Data
None

Arbitrary Values in Data
Experience (52 negative values)

Columns Dropped
ID
ZIPCode

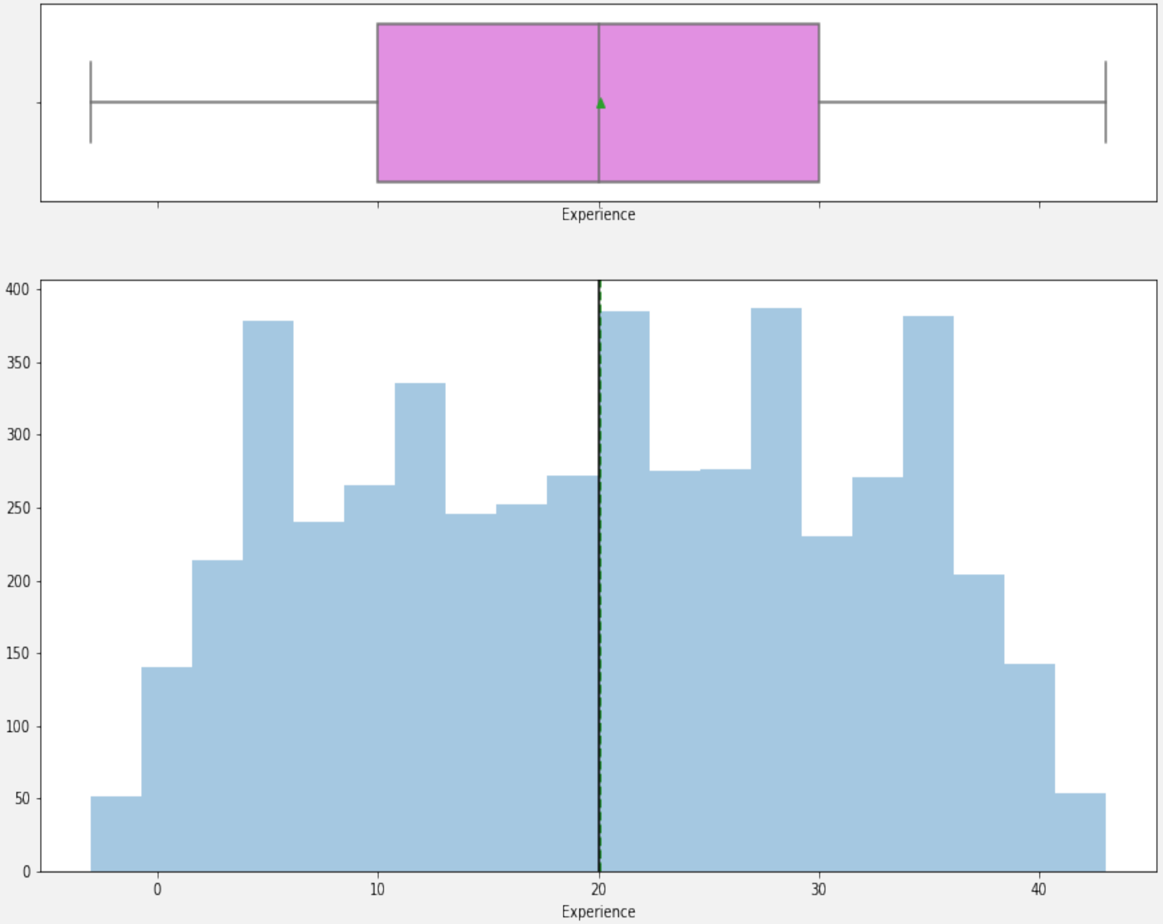
EXPLORATORY DATA ANALYSIS

AGE



- The distribution of Age approximately normal
- There are no outliers in this variable.
- 3rd quartile(Q3) is equal to 55 which means 75% of customers are aged 55years and below

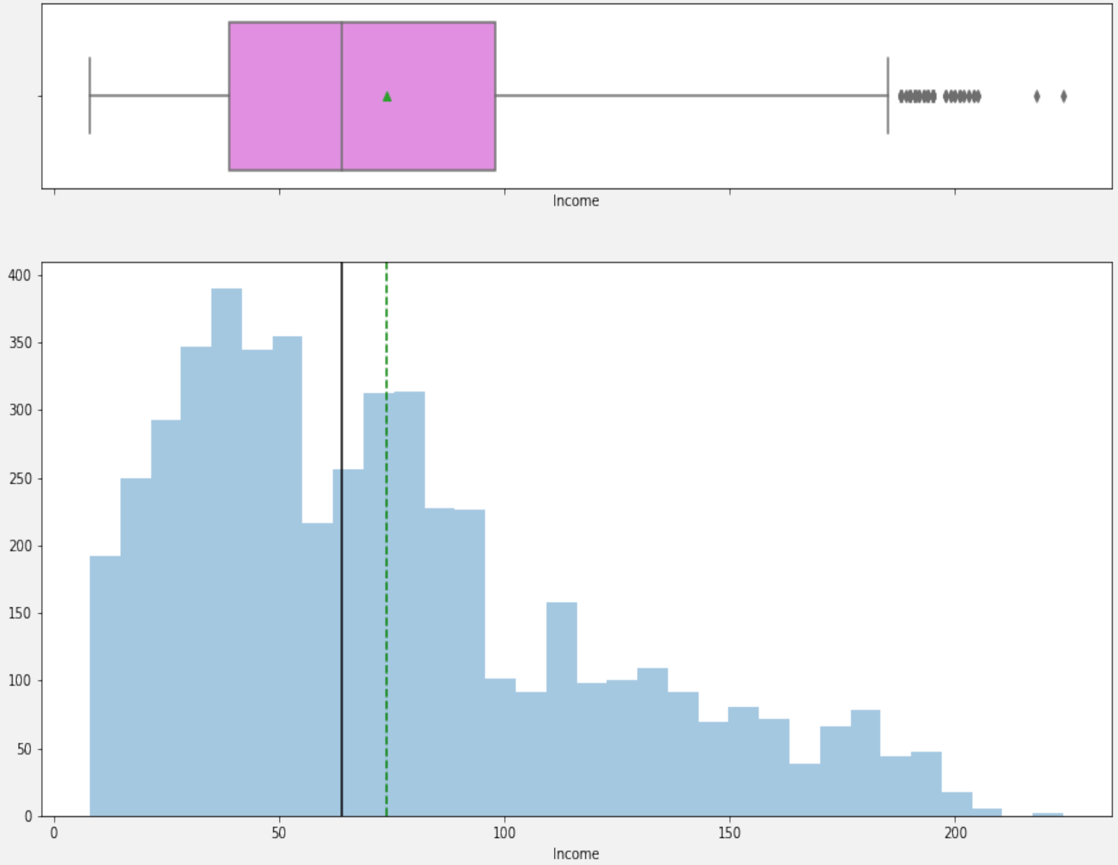
EXPERIENCE



- Normal distribution
- No outliers
- 75% has a professional experience of 50 years or less and an average of 20 years.
- It’s a highly educated population

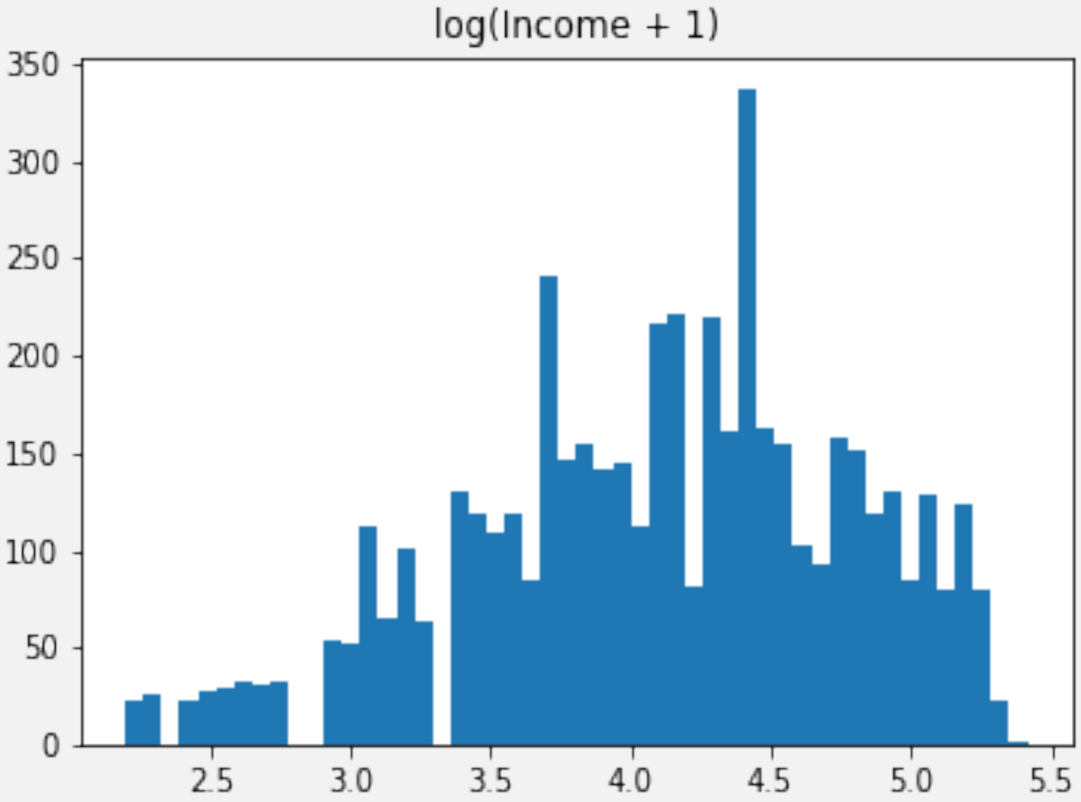
EXPLORATORY DATA ANALYSIS

INCOME



- Income is Right_skewed
- Presence of Outliers
- Wide range with Maximum at 224,000 and least at 8,000

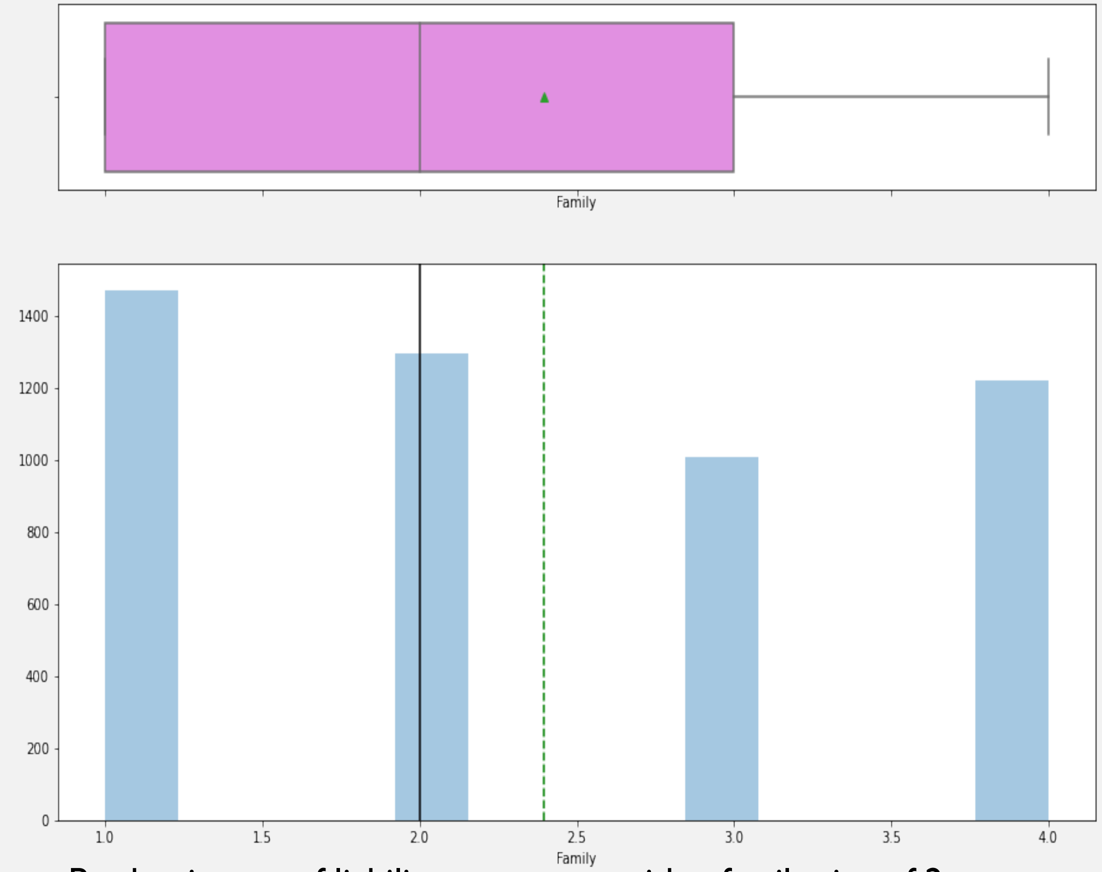
LOGPLOT-INCOME



- Income behaves better on a Log scale

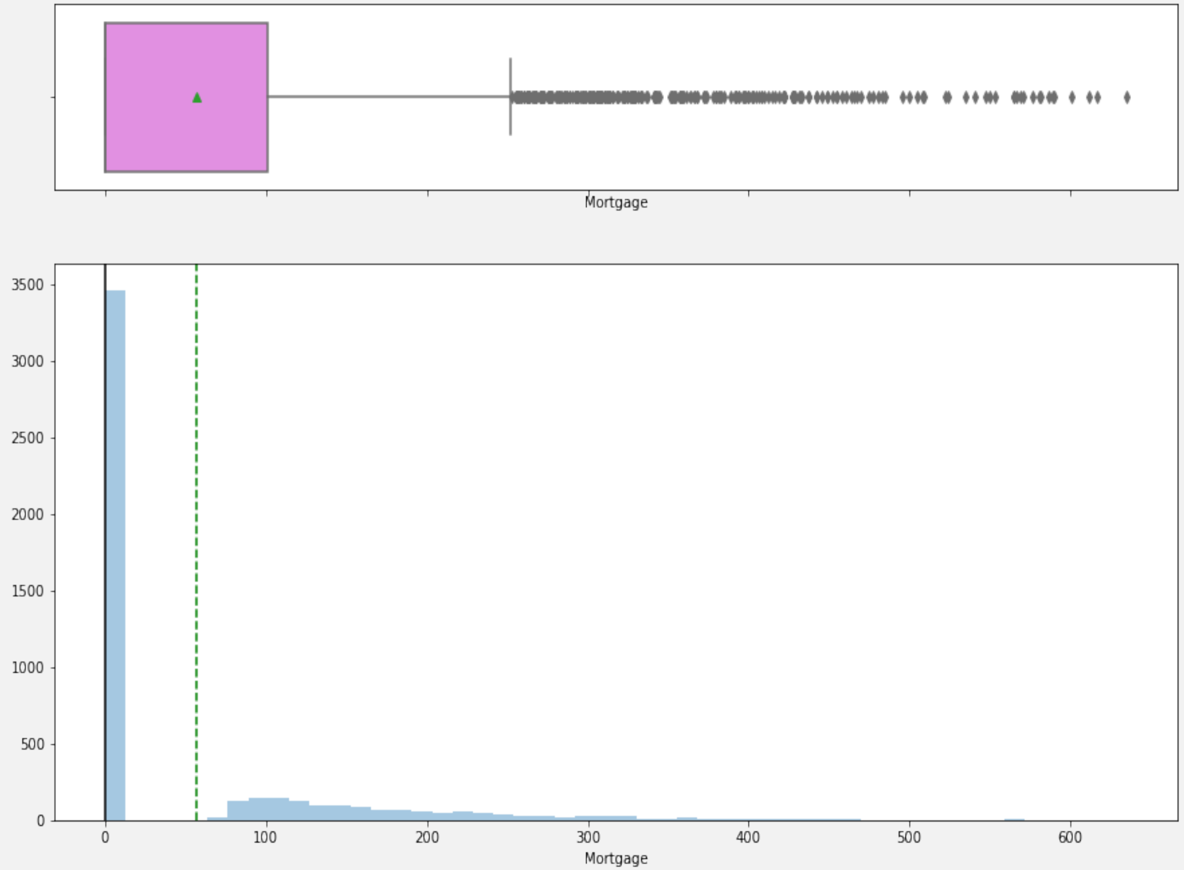
EXPLORATORY DATA ANALYSIS

FAMILY



- Predominance of liability customers with a family size of 2
- It is a uniform distribution
- 75% have a family size of 3 and below
- No outliers

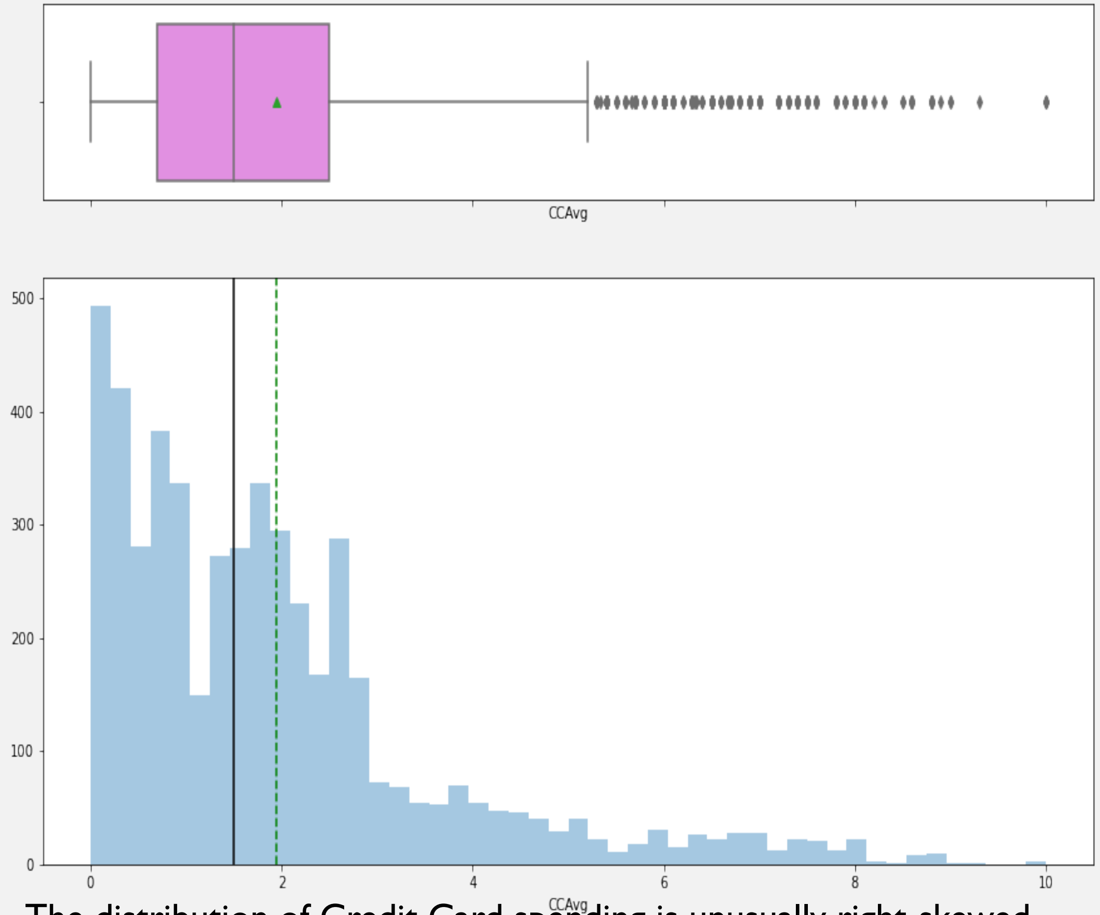
MORTGAGE



- Right skewed
- Presence of outliers
- 50% have no mortgage
- This distribution is less geared
- There was no effect of the Logplot on Mortgage

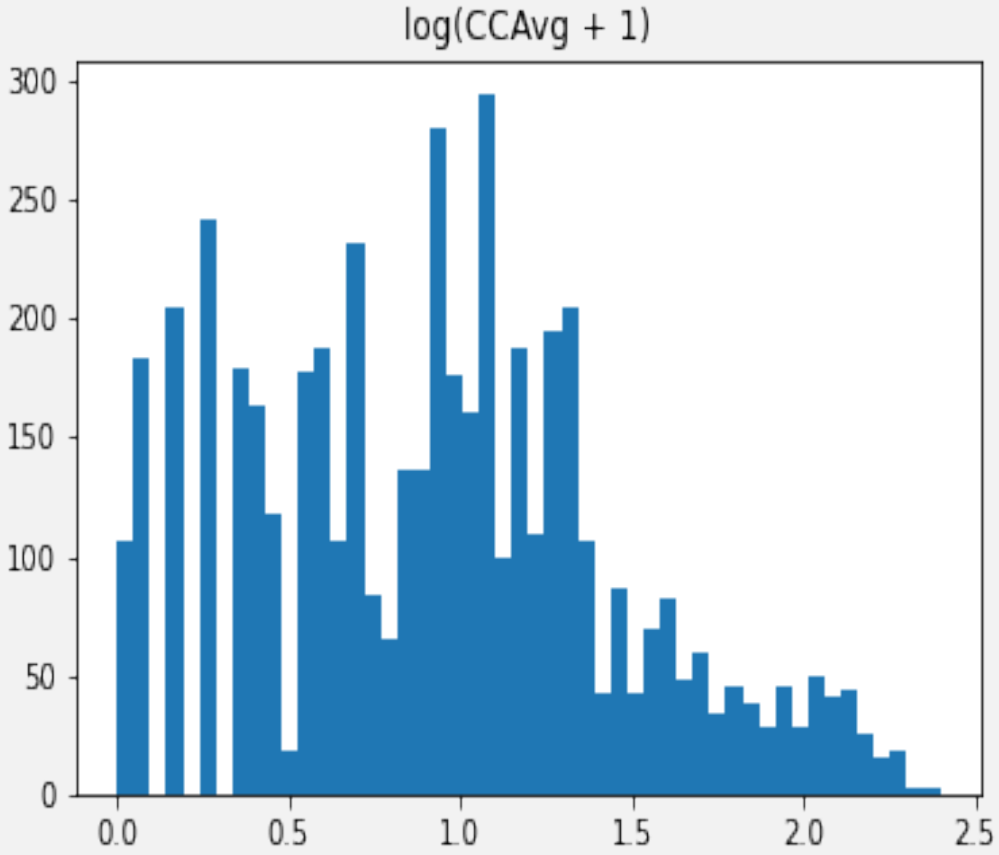
EXPLORATORY DATA ANALYSIS

CCAvg



- The distribution of Credit Card spending is unusually right-skewed.
- There are outliers in this variable.
- 75% of the liability customers in the distribution spent 2,500 or less from their credit cards with 1,930 dollars spent on the average

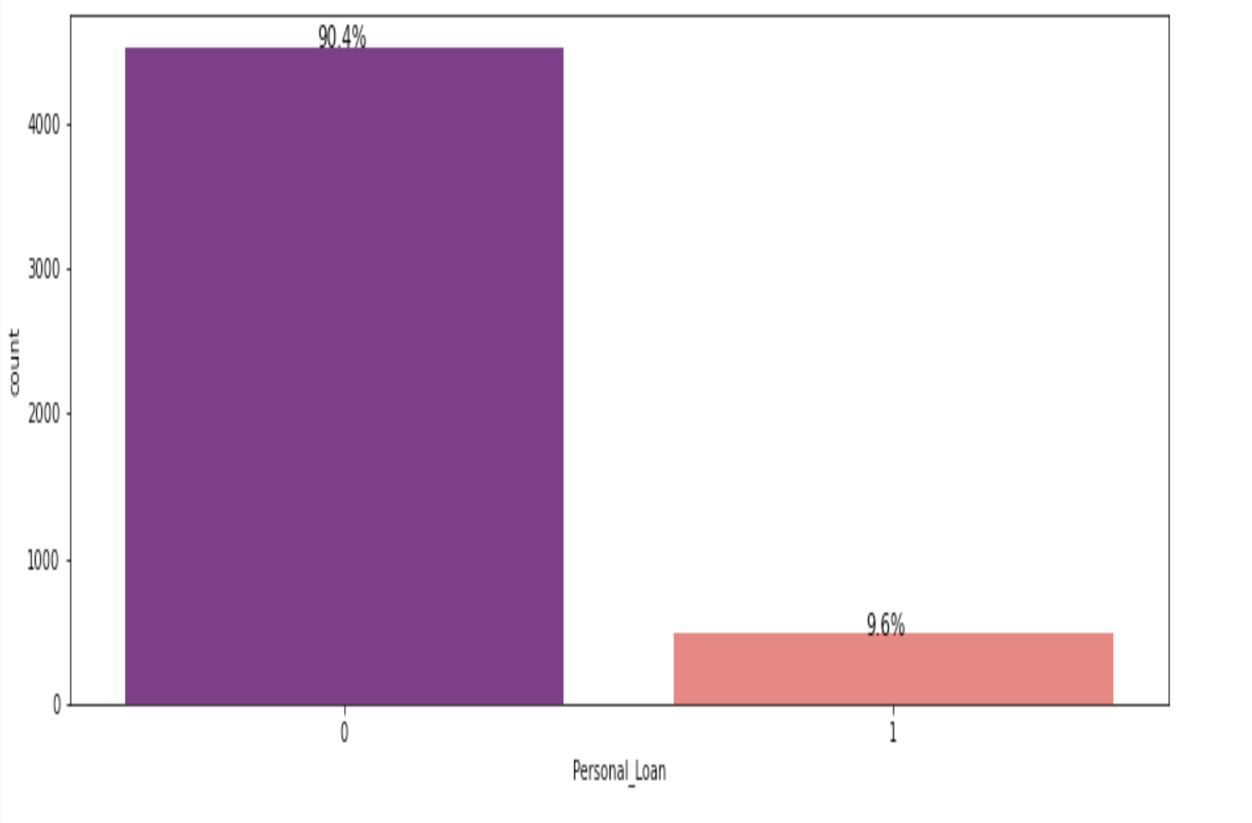
LOG TRANSFORMATION-CCAvg



- The Power of Log Transformation and its effect on kurtosis. We can see an apparent effect on CCAvg behaviour

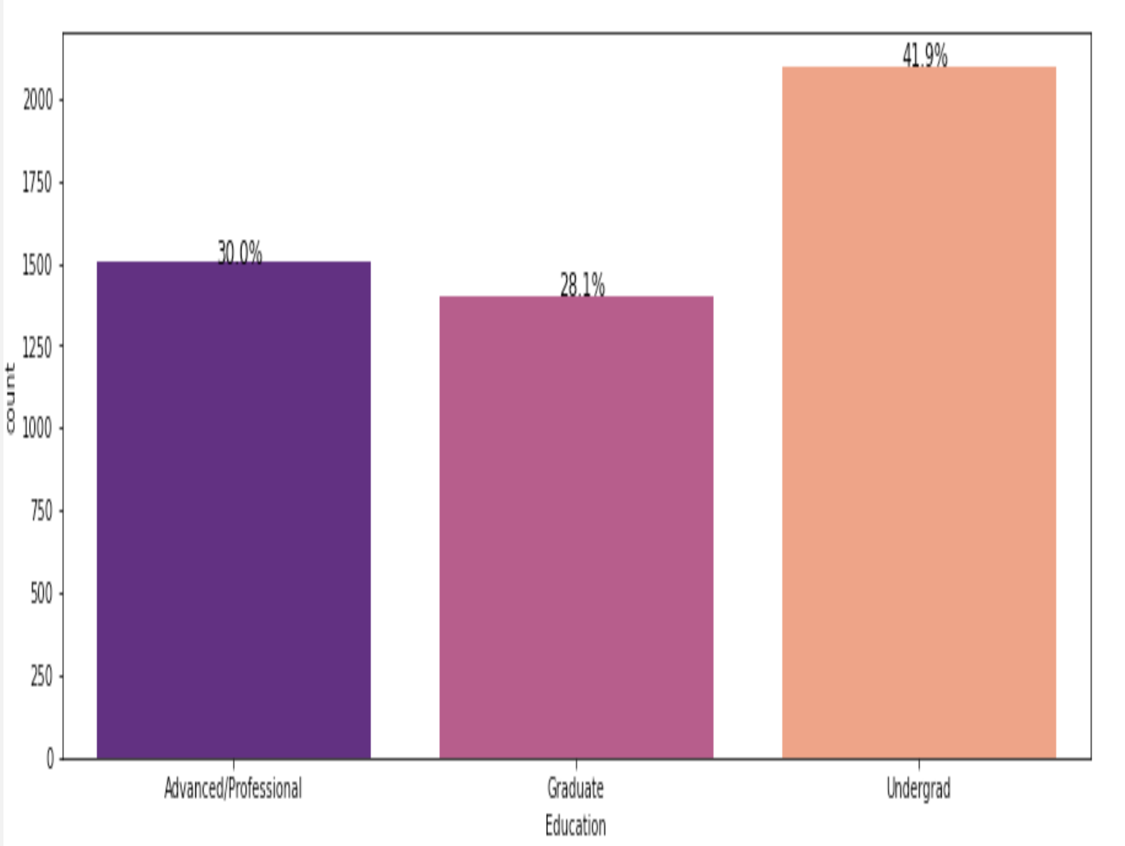
EXPLORATORY DATA ANALYSIS

PERSONAL LOAN



- 90.4% of the customers did not accept the personal loans in the last campaign.
- 9.6% of the customers accepted personal loans previously

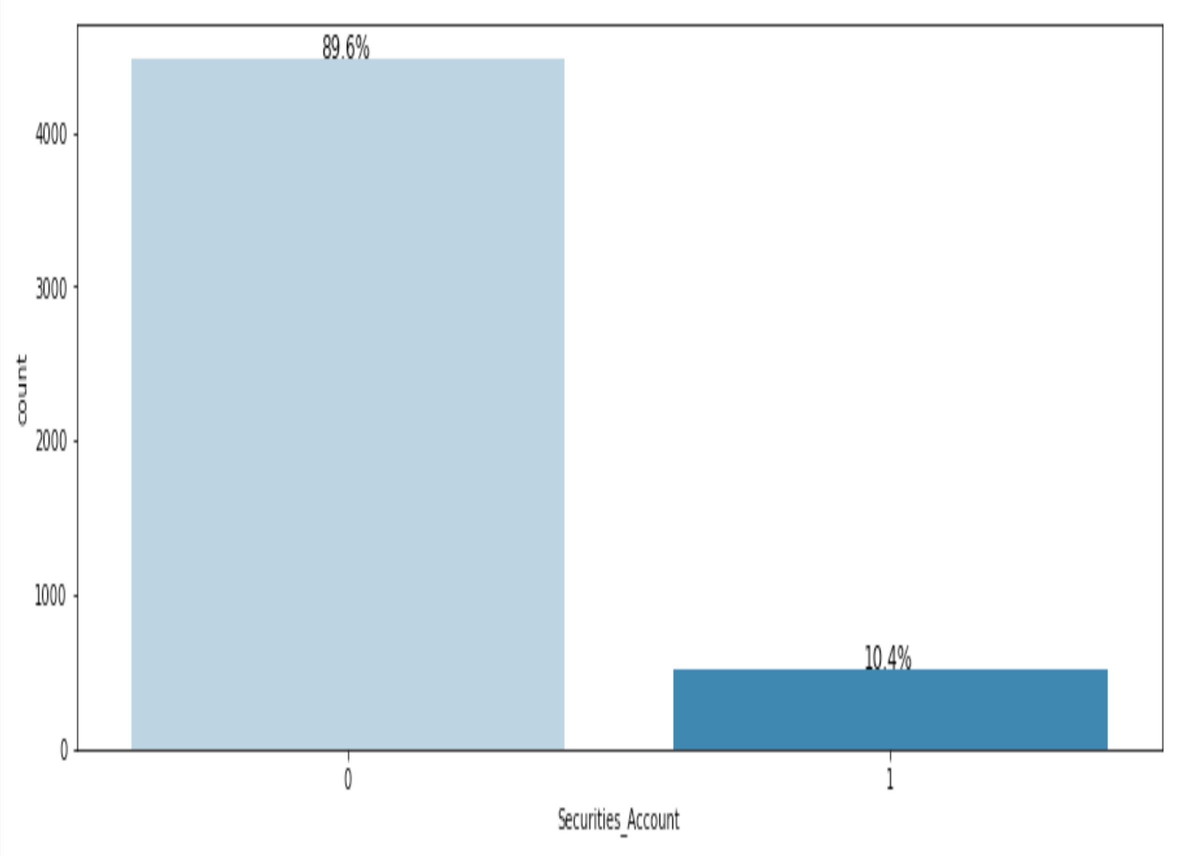
EDUCATION



- 41.9% of the customers are Undergrads with Advanced/Professional following closely
- Graduate Education accounts for 28.1% of the distribution of customers

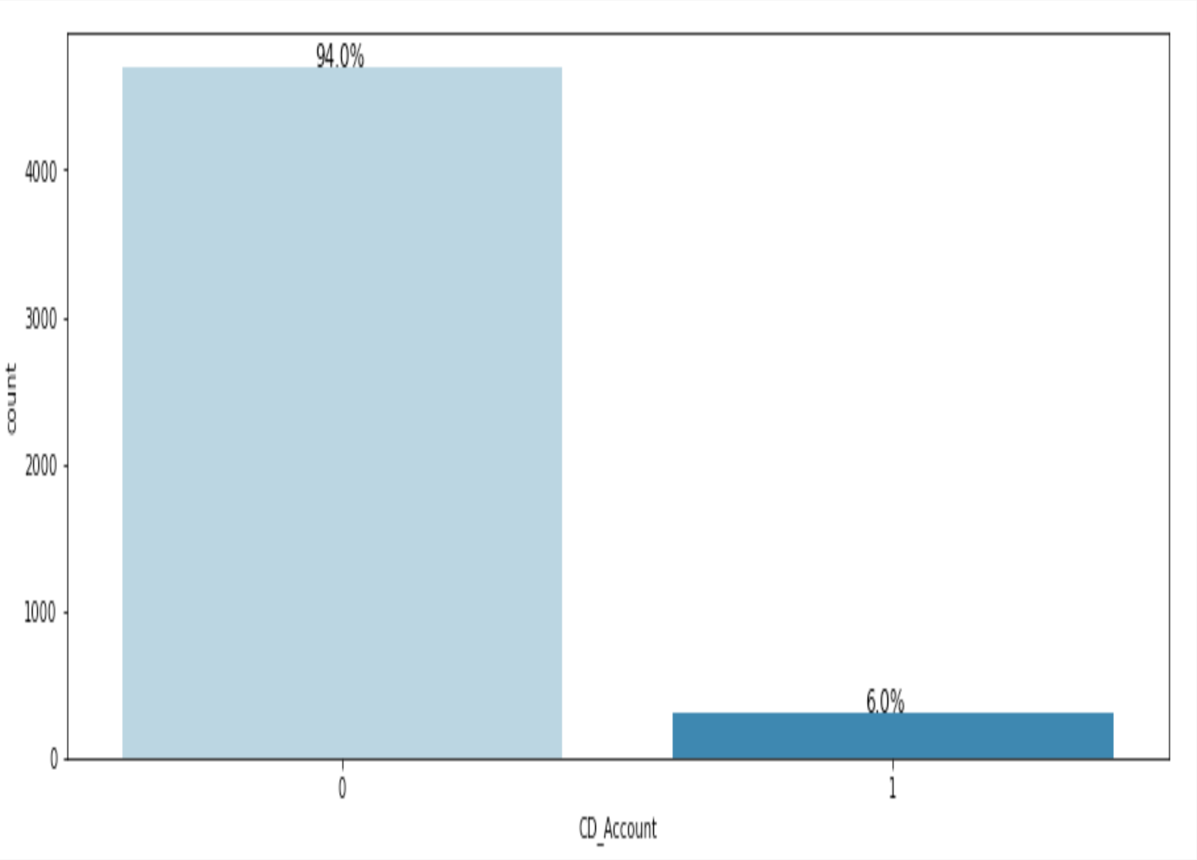
EXPLORATORY DATA ANALYSIS

SECURITIES ACCOUNT



- 89.6% of the customers do not maintain a security account with AllLife Bank while 10.4% does.

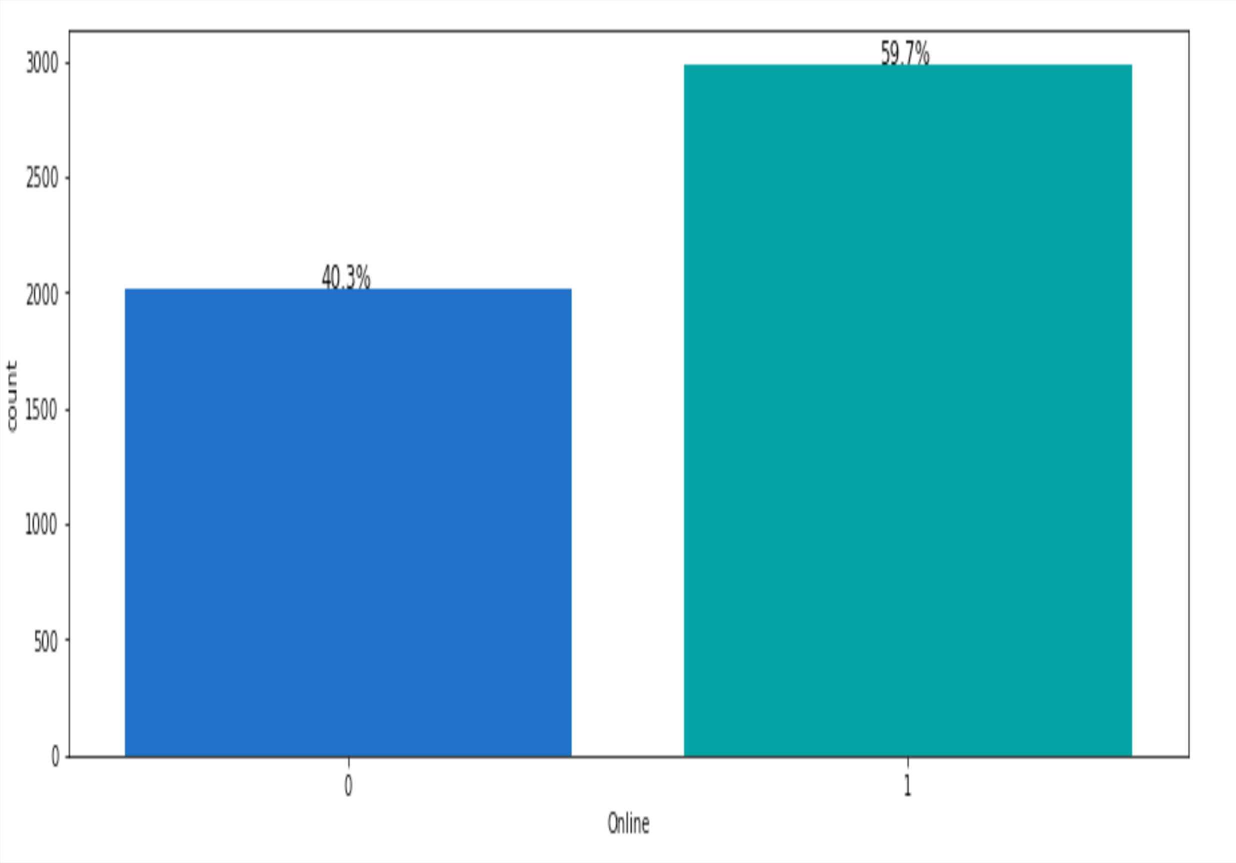
CD ACCOUNT



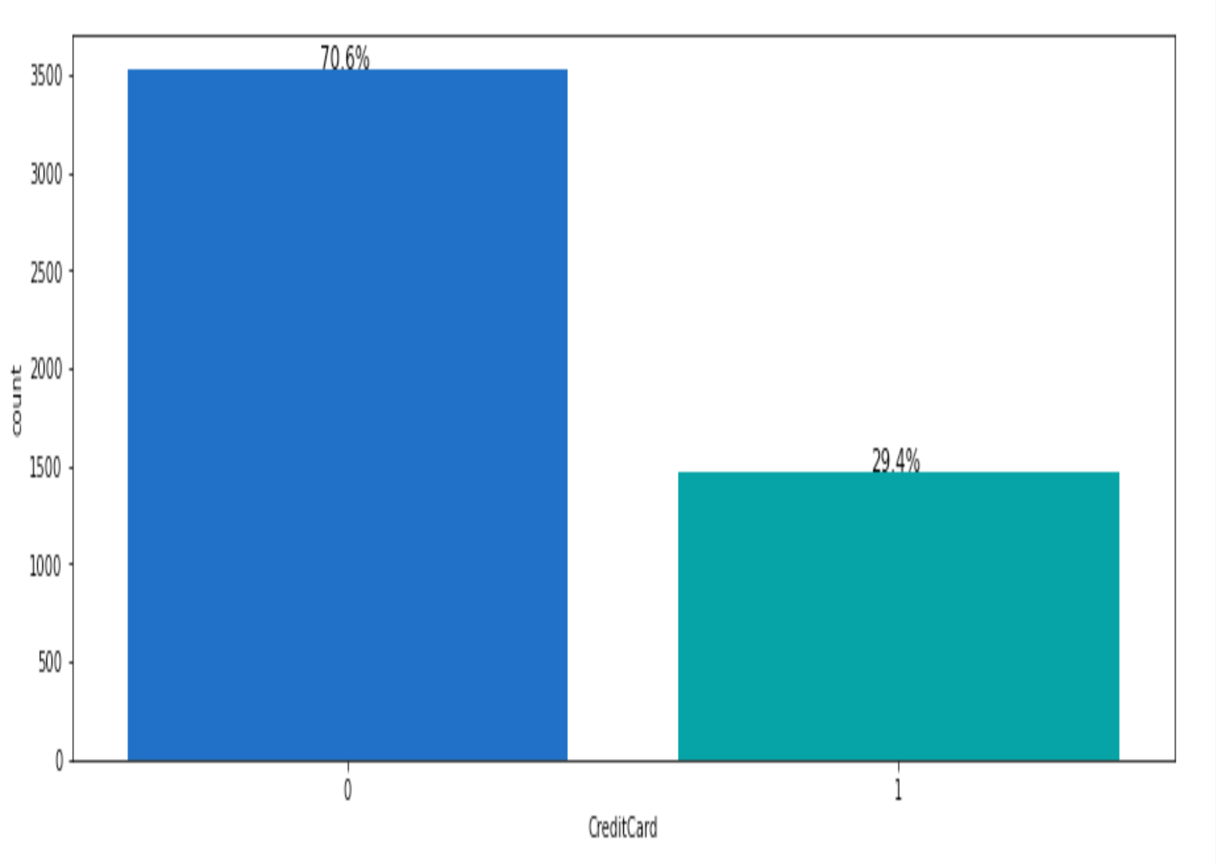
- 94.0% do not maintain a Certificate of Deposit account with AllLife Bank while 6.0% does.

EXPLORATORY DATA ANALYSIS

ONLINE



CREDIT CARD

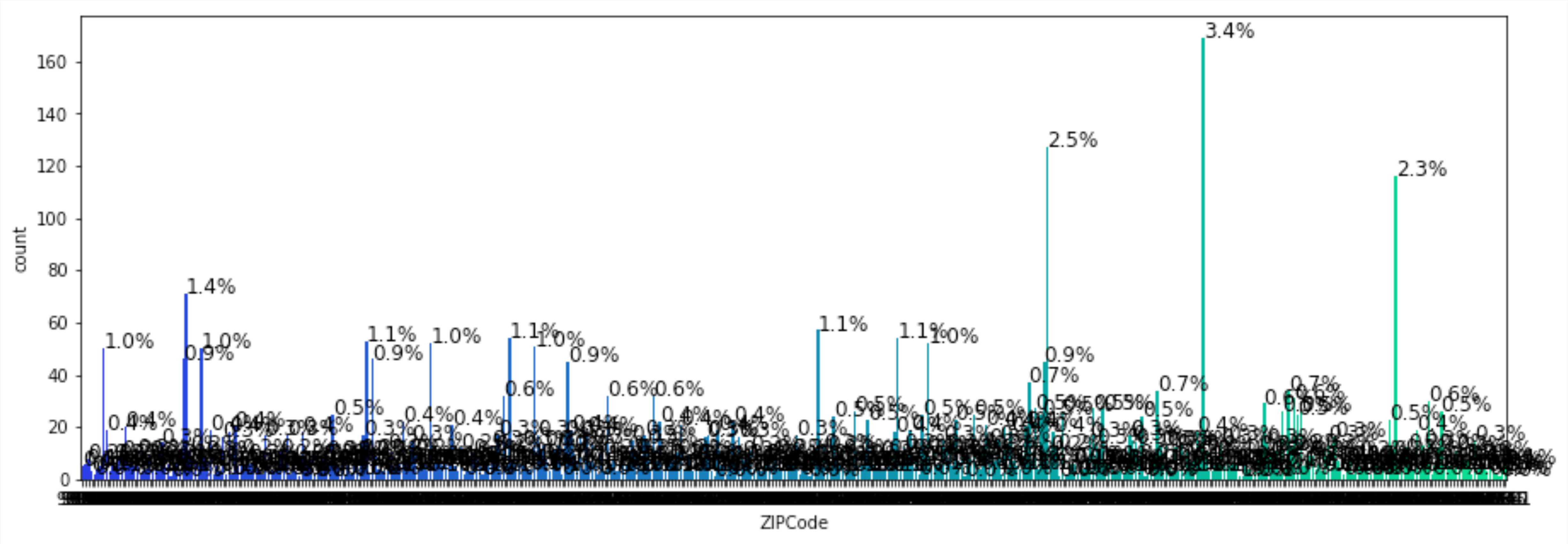


- 40.3% of the customers of AllLife Bank do not use online banking facilities while 59.7% engage online banking facilities.

- 70.6% of the customers do not use credit cards issued by other banks while 29.4% do

EXPLORATORY DATA ANALYSIS

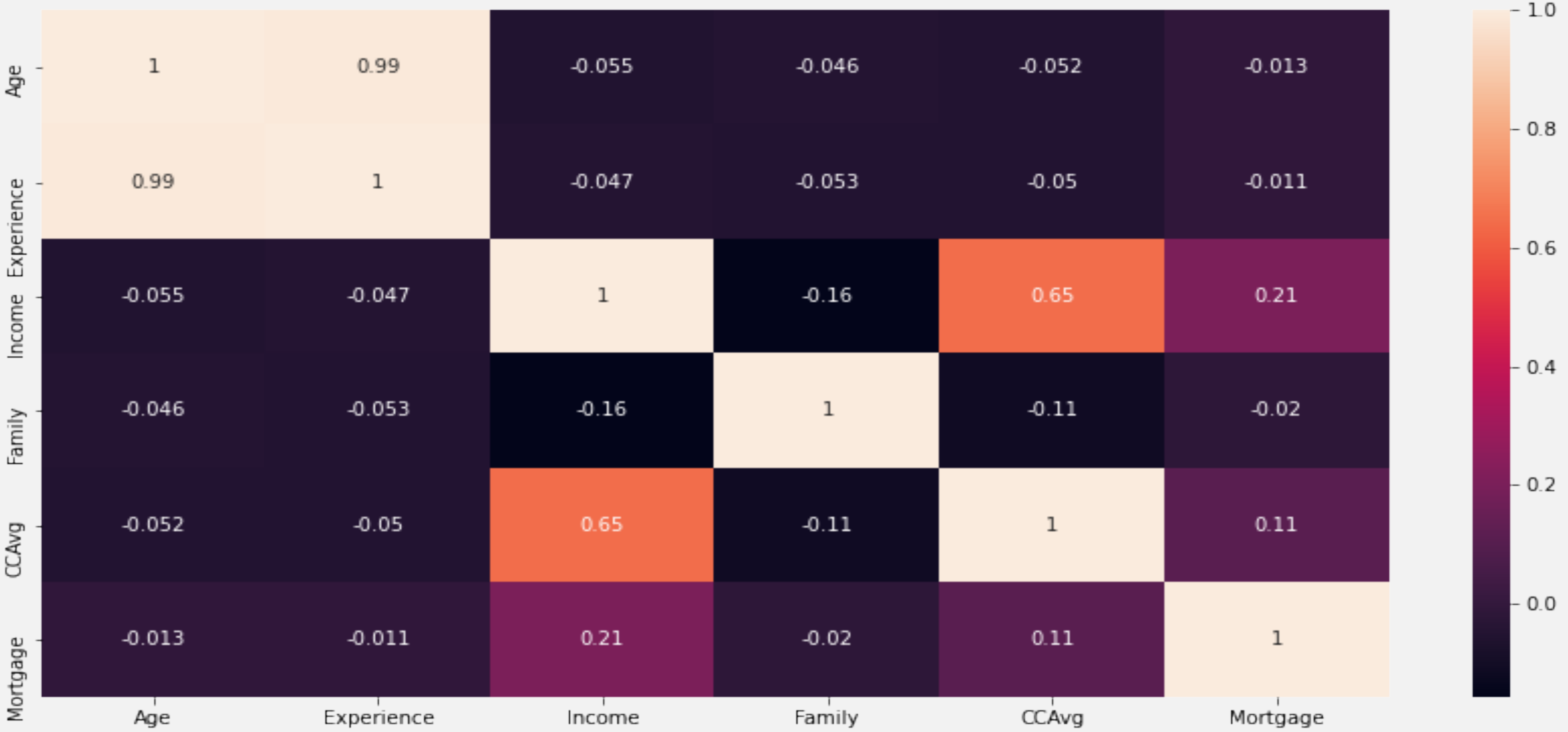
ZIPCode



- Reconciling the plot with the value_count info obtained earlier, we see that the top 4 locations of interest by Percentage customer base and patronage have ZIPs 94720,94305, 95616,90095,93106 respecively
- When looked up via a Google search, these are Berkeley, CA; Stanford CA; Davis ,CA; Los Angeles,CA and Isla Vista,CA respectively

EXPLORATORY DATA ANALYSIS

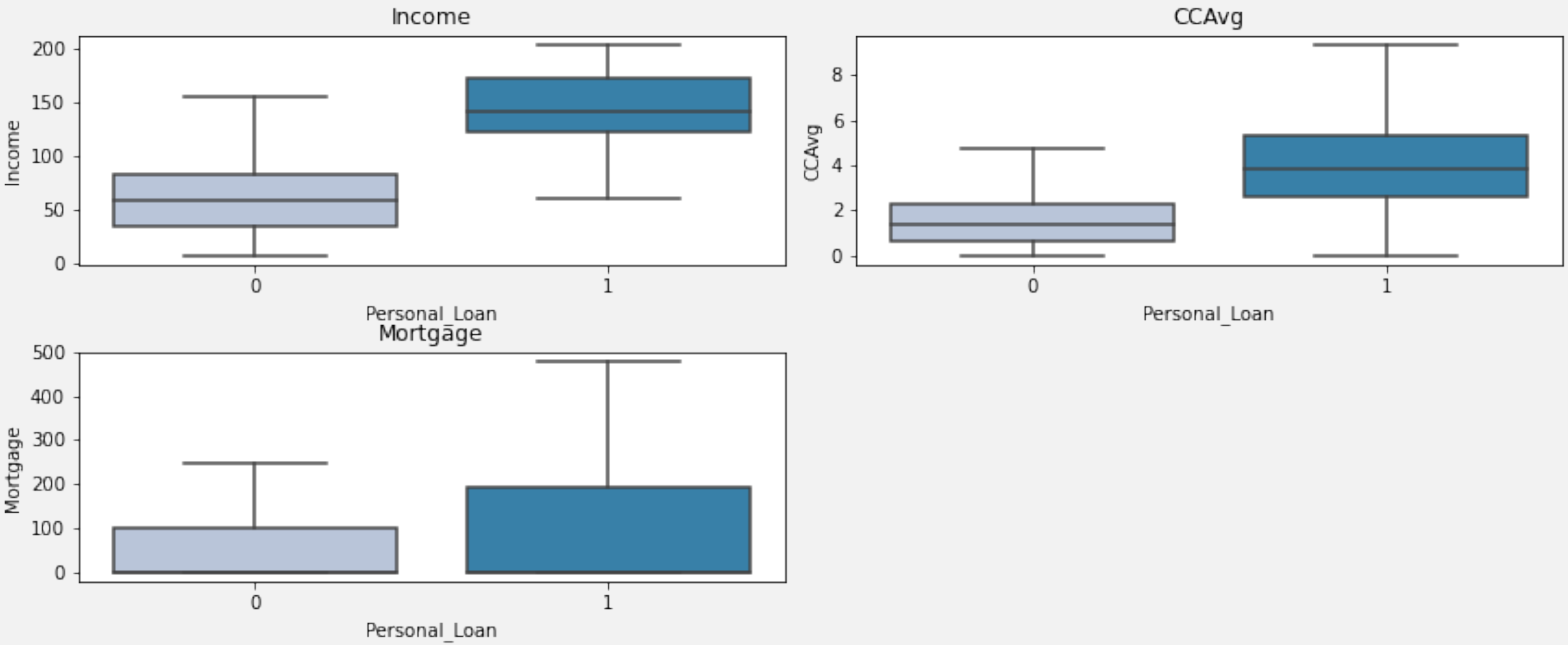
Heat Map



- Age and Experience are highly correlated
- Same can be concluded for CCAvg and Income. This clearly alludes to the fact that the higher the income, the higher the propensity for credit card spending
- There is a relatively fair correlation between Mortgage and Income. This establishes that a higher income would ideally endear the management of AllLife Bank to finance a mortgage facility
- There is clearly little or no correlation between the other variables

EXPLORATORY DATA ANALYSIS

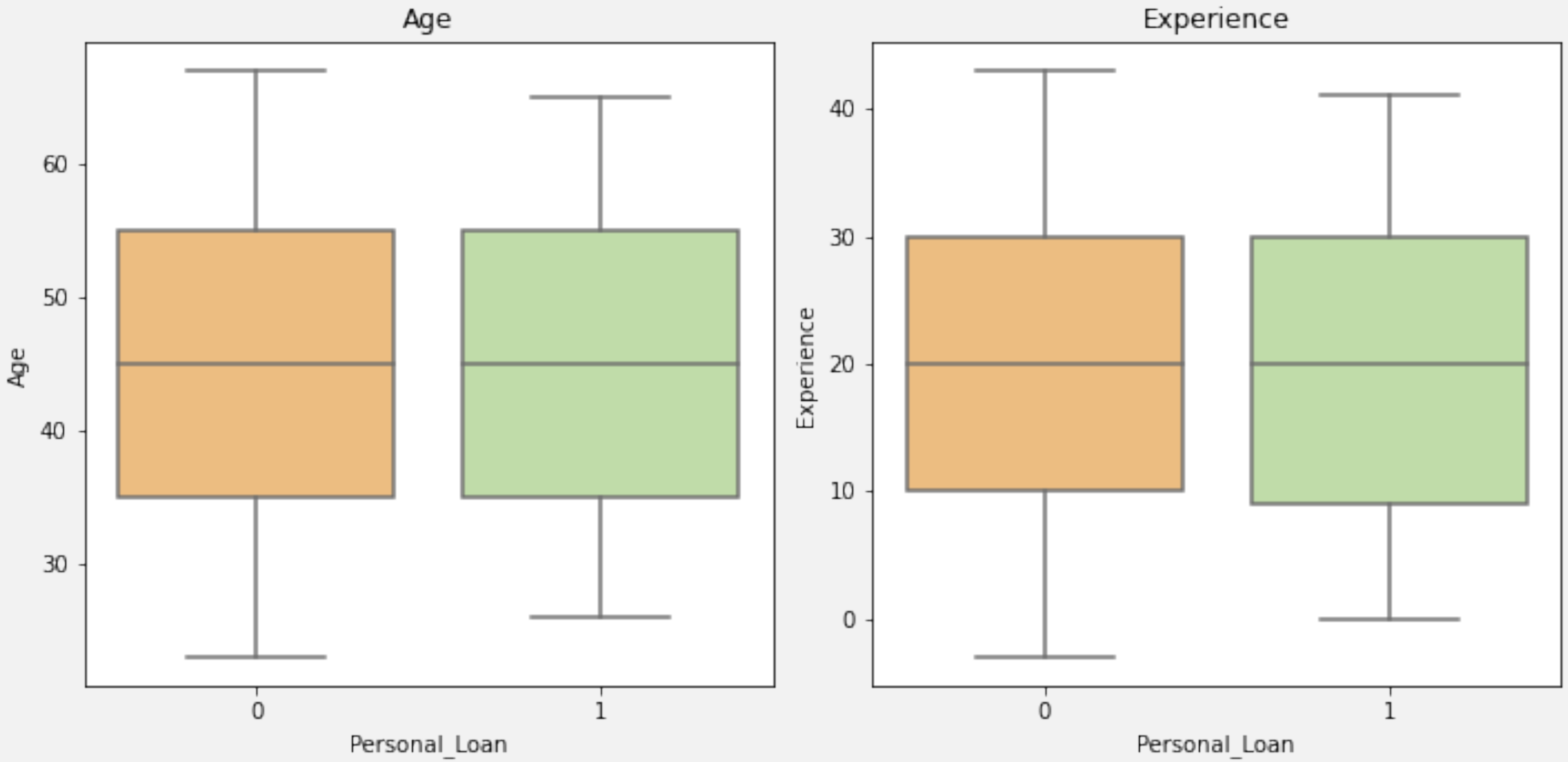
PERSONAL LOAN VS INCOME ,CCAVG , MORTGAGE (WITHOUT OUTLIERS)



- Higher incomes, higher credit card spending and higher mortgages are huge potentials for a personal loan consideration
- Validates the correlation between Income, CCAvg and Mortgage
- Customers with higher credit card spending and mortgage with attract more fees

EXPLORATORY DATA ANALYSIS

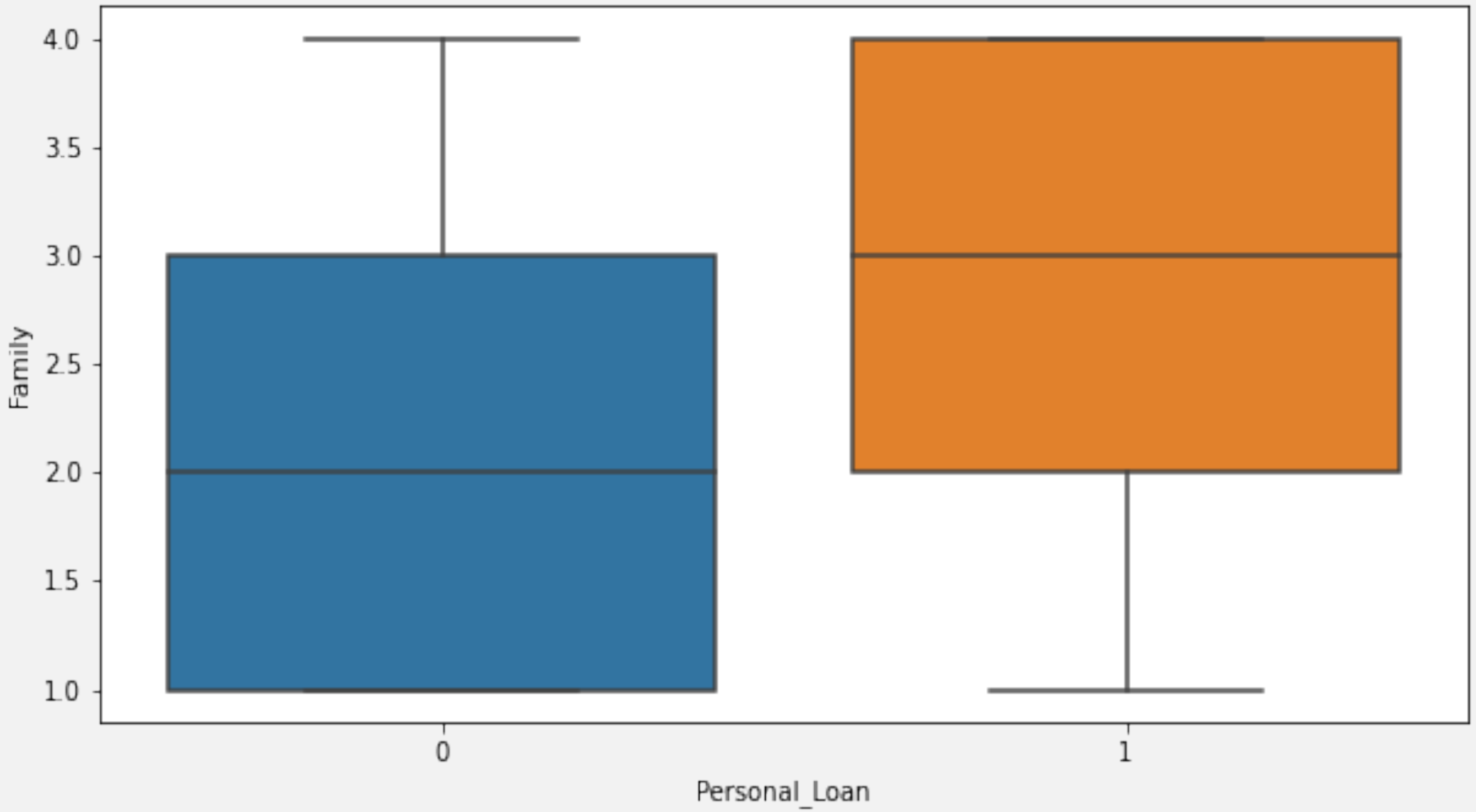
PERSONAL LOAN VS AGE AND EXPERIENCE



- Strong correlation between Age and Experience based on similar plots
- There is an equal drive on Personal Loan potentials for both classes who did not accept loan offers at the previous campaign
- The affinity of AllLife Bank management to sell a personal loan to customers with a high professional experience
- Same applies to one with an age of 35 and above who did not accept a personal loan offer

EXPLORATORY DATA ANALYSIS

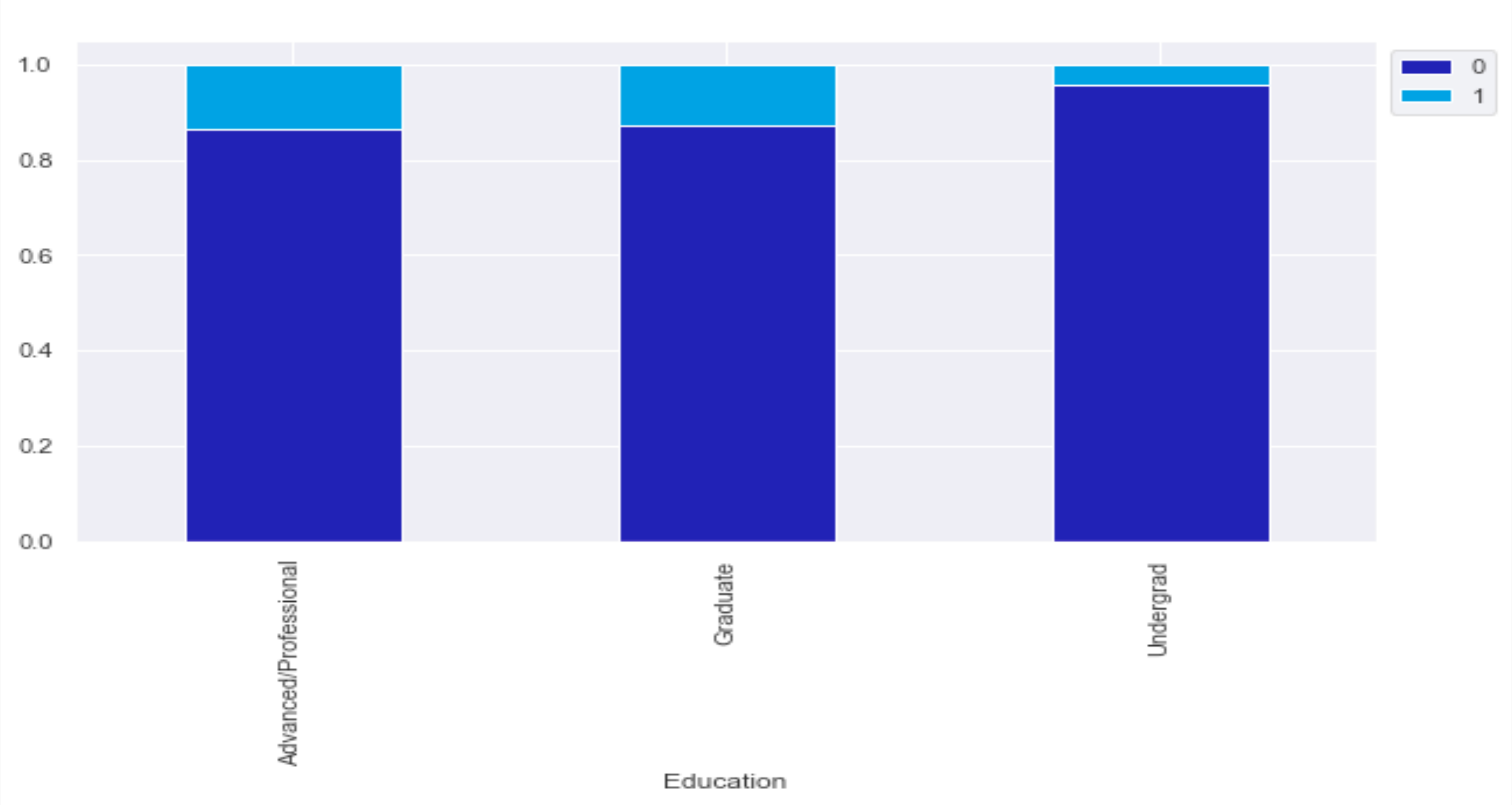
PERSONAL LOAN VS FAMILY



- Liability customers with family sizes of 2 or more accessed personal loans in the last campaign
- While 50% at least of those who did not have a family size of 2 or less

EXPLORATORY DATA ANALYSIS

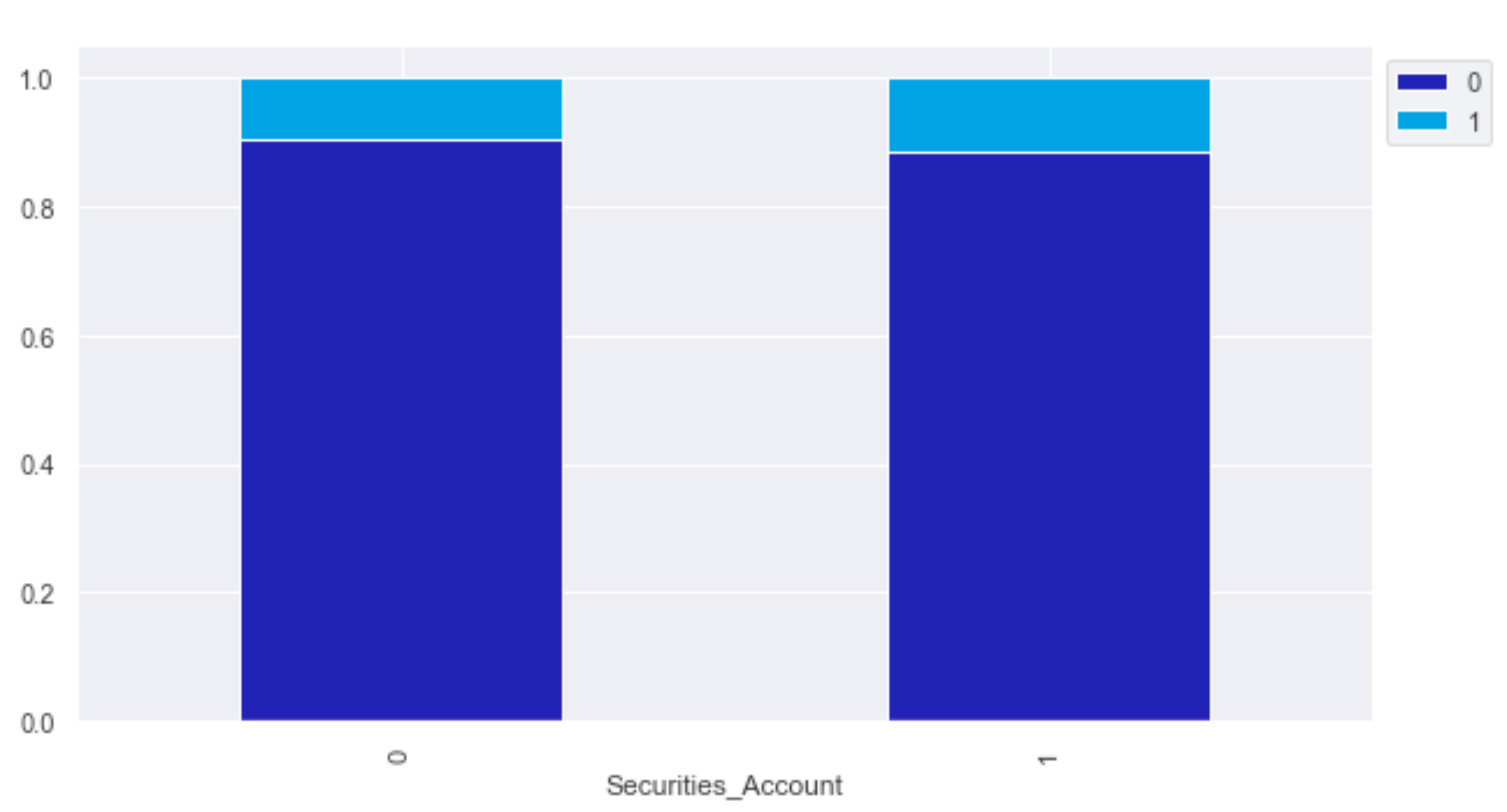
PERSONAL LOAN VS EDUCATION)



- Advanced professionals leads the pack with roughly 15% chance of being granted loan offers compared to Graduates with a 13%
- Undergrads have a hugely untapped market

EXPLORATORY DATA ANALYSIS

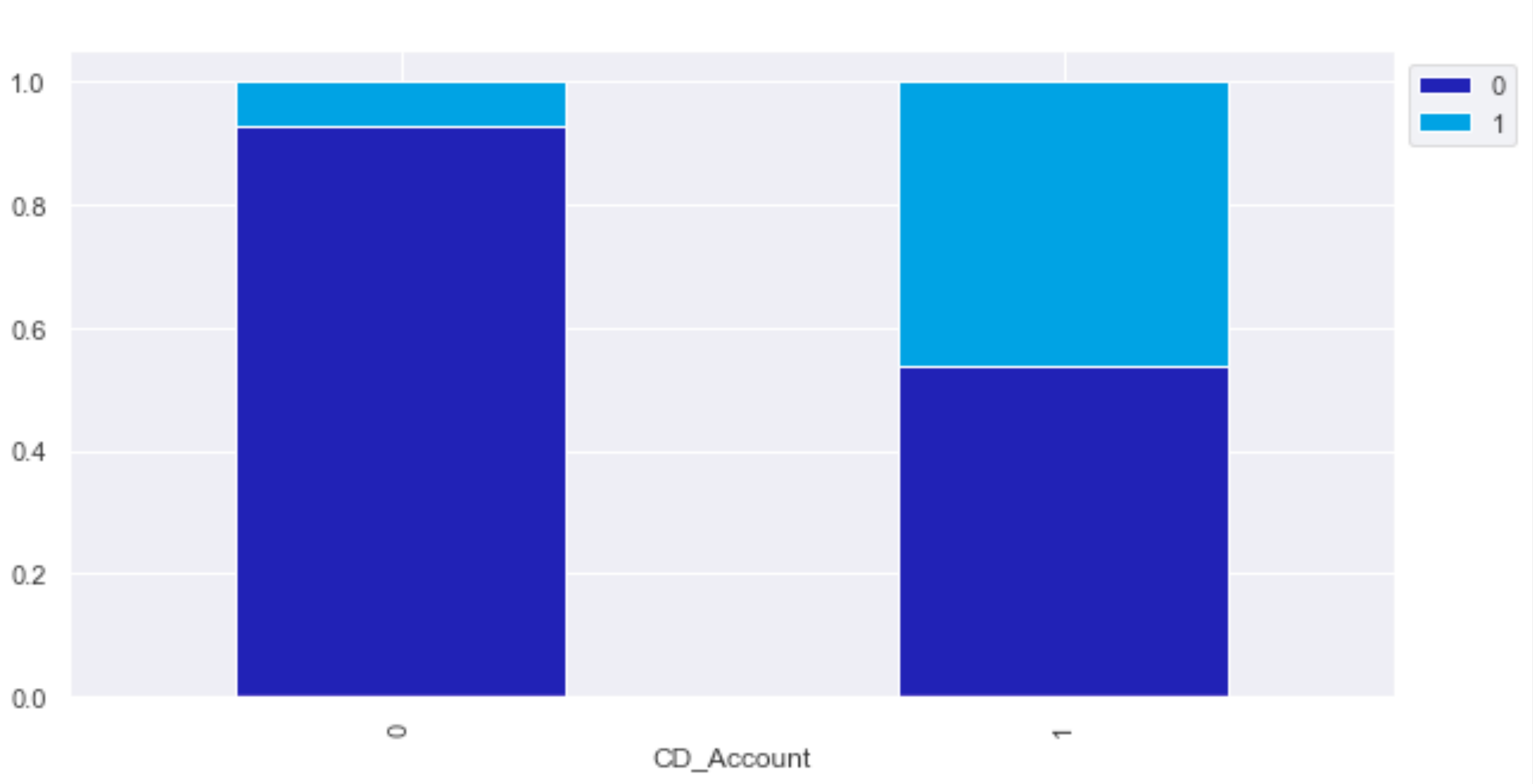
PERSONAL LOAN VS SECURITIES ACCOUNT



- Customers with Securities account have a 16% chance of being granted a personal loan
- Customers without a Securities account have a 10% likelihood of being considered for an offer

EXPLORATORY DATA ANALYSIS

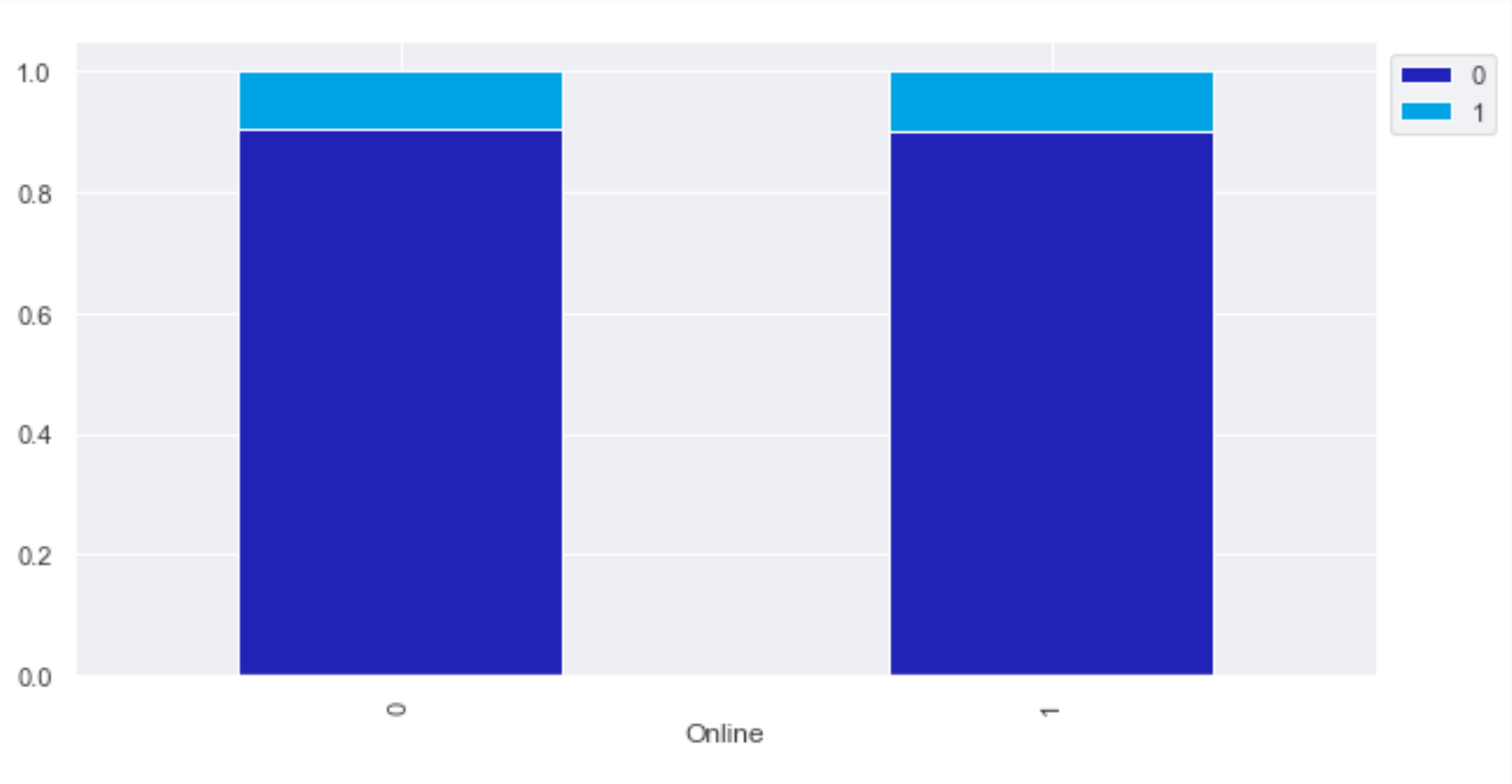
PERSONAL LOAN VS CD ACCOUNT



- There is 50% chance for customers with Certificate of Deposits to be considered for a loan
- While the likelihood to be considered without one stands at approx. 7%

EXPLORATORY DATA ANALYSIS

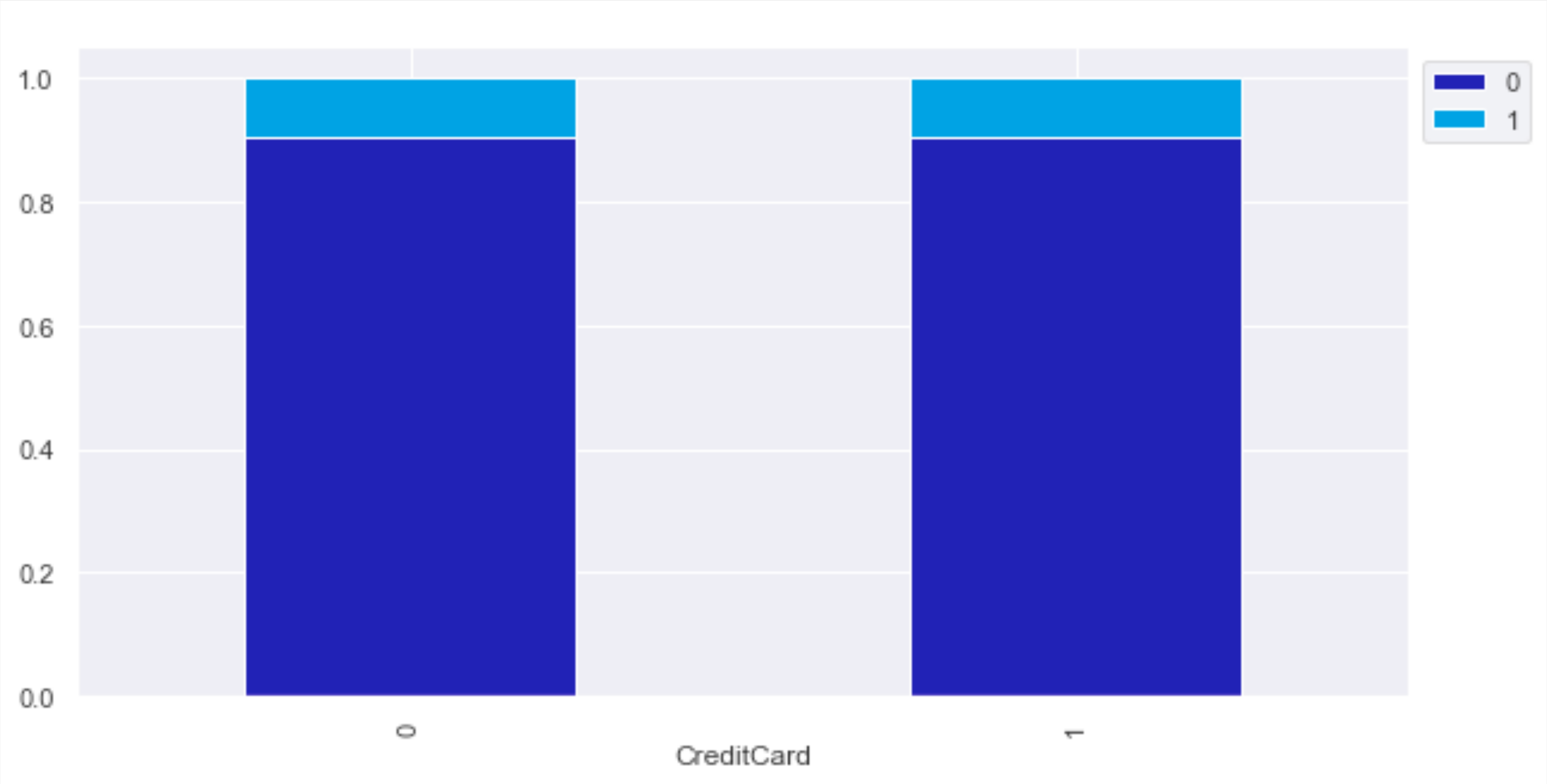
PERSONAL LOAN VS ONLINE



- Customers who engage in online banking activities have an almost at per chances of being considered for a loan purchase when compared to those without
- This stands at approx. 10% and 9% respectively

EXPLORATORY DATA ANALYSIS

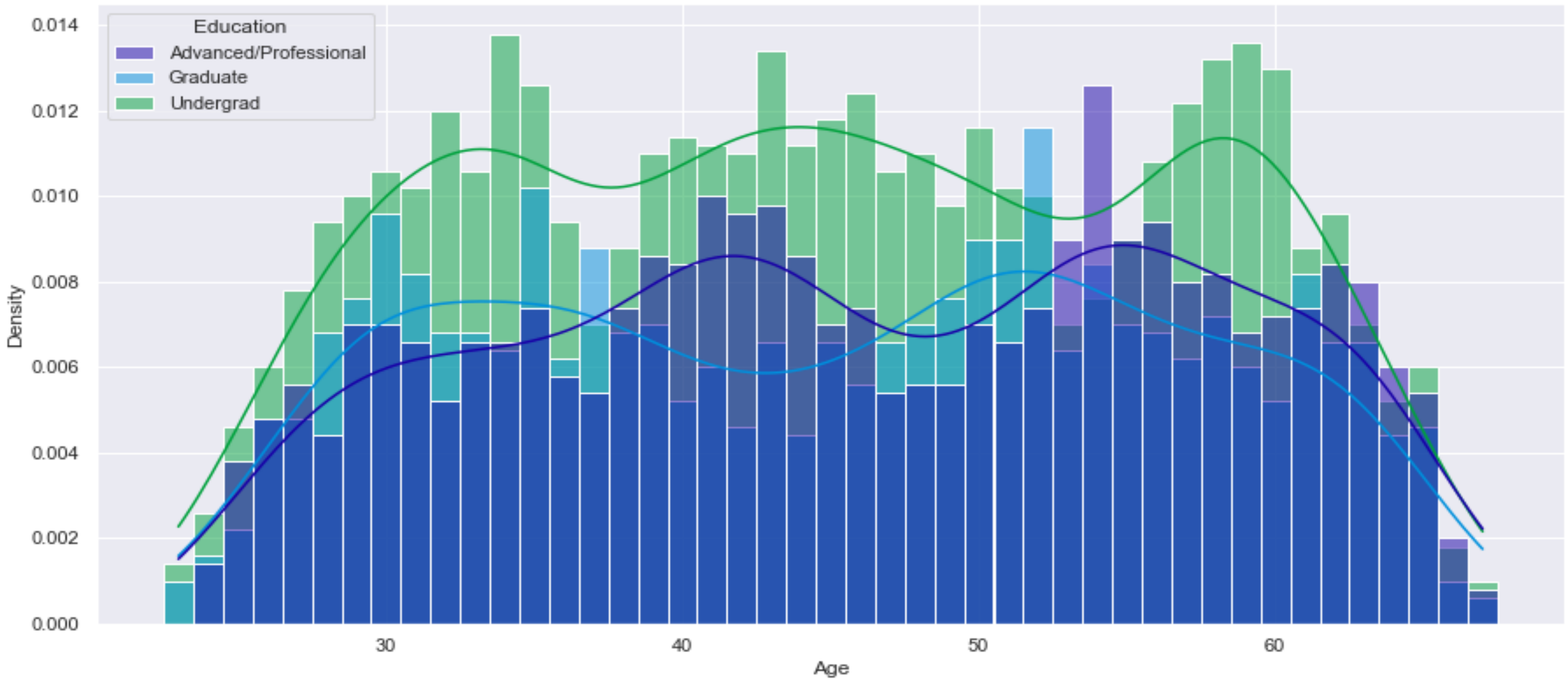
PERSONAL LOAN VS CREDIT CARD



- This is seen as a relatively important feature in loan consideration
- This also represents a huge market for AllLIfe Bank

EXPLORATORY DATA ANALYSIS

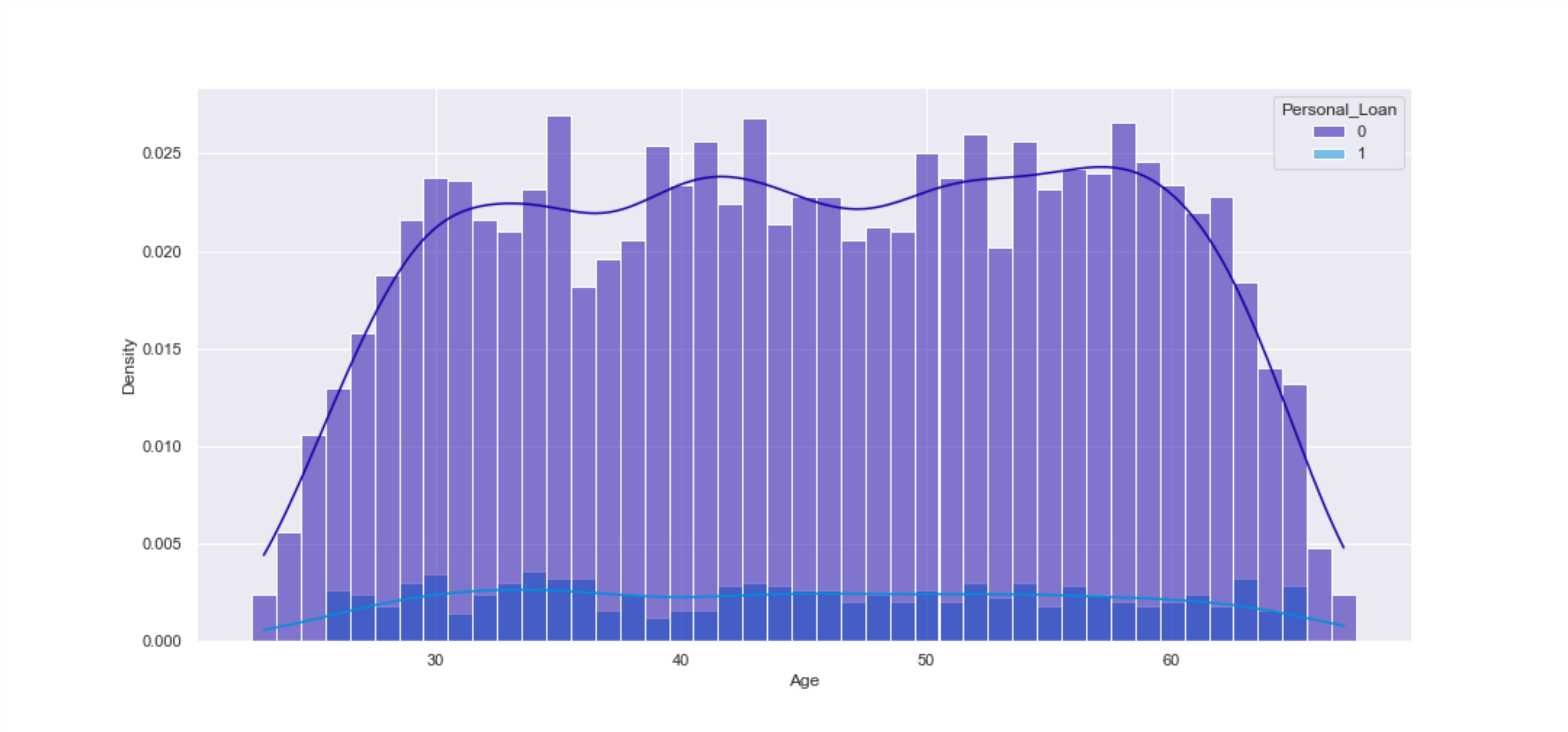
AGE DISTRIBUTION BY EDUCATION



- Advanced Professionals have much higher ages
- This is closely followed by Graduates

EXPLORATORY DATA ANALYSIS

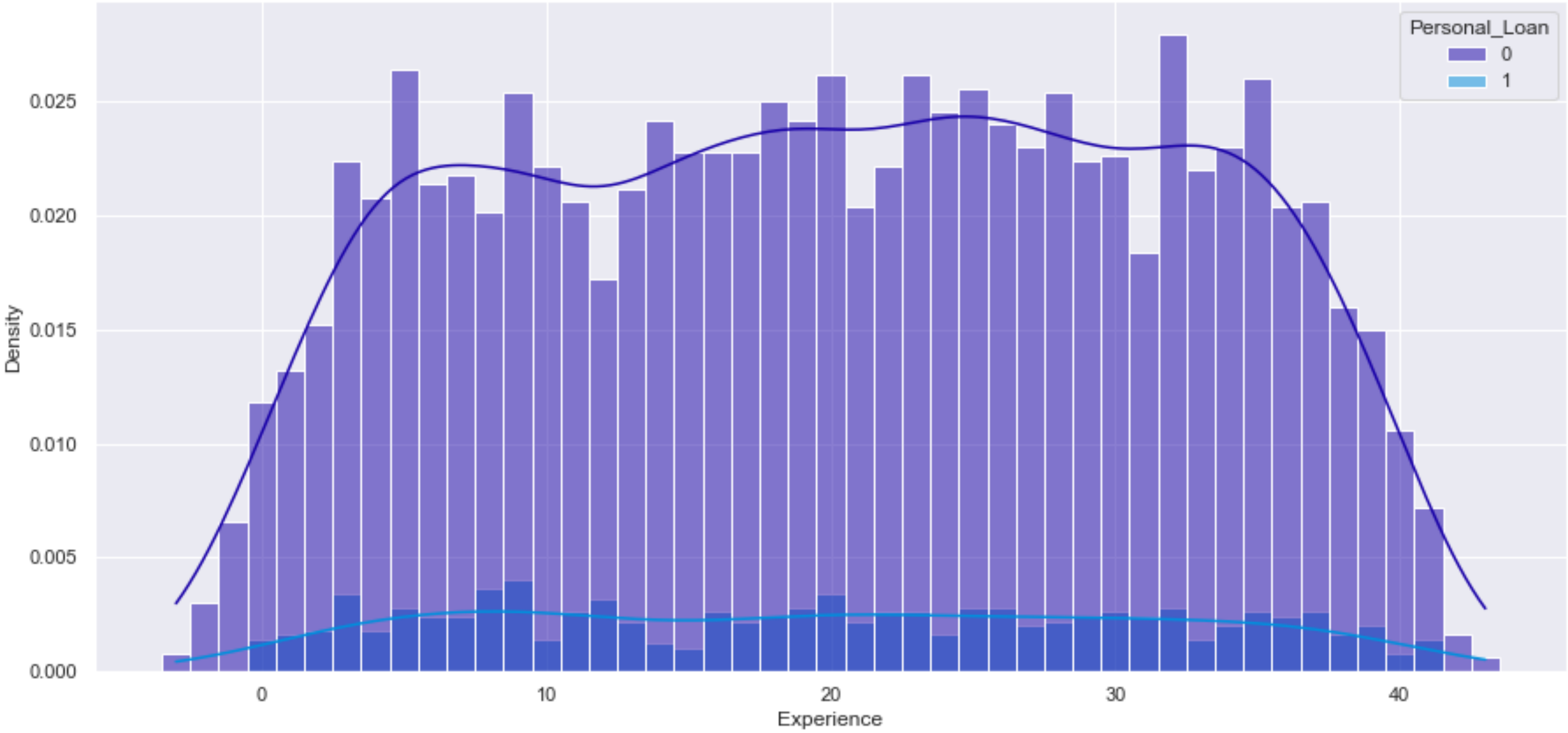
AGE DISTRIBUTION BY PERSONAL LOAN



- It is clear that those who took loans are lesser in number compared to those who did not considering the age factor

EXPLORATORY DATA ANALYSIS

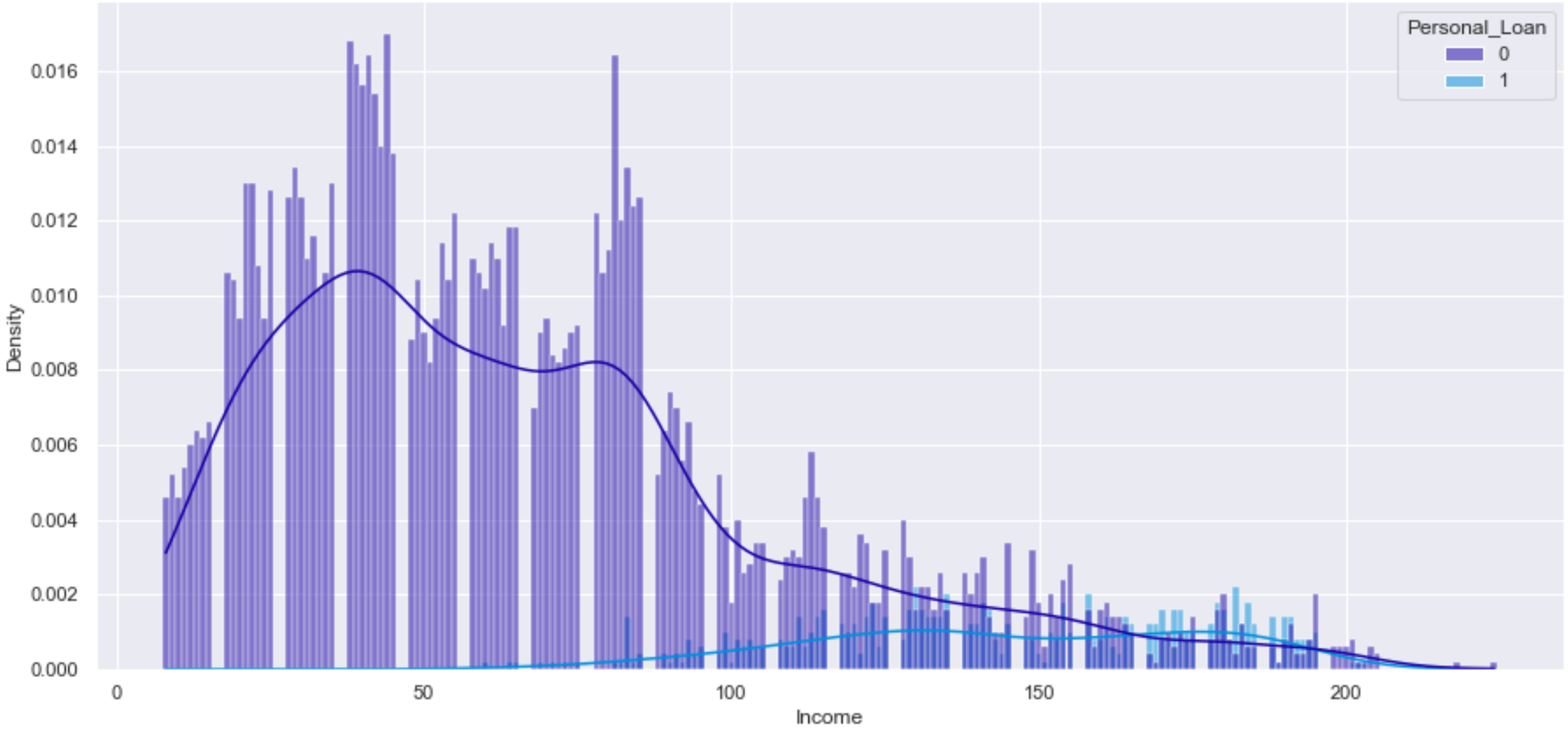
PERSONAL LOAN VS EXPERIENCE



- This is a clear correlation between Age and Experience with similar plots
- Those who did accept the loan have a higher experience than those who did not

EXPLORATORY DATA ANALYSIS

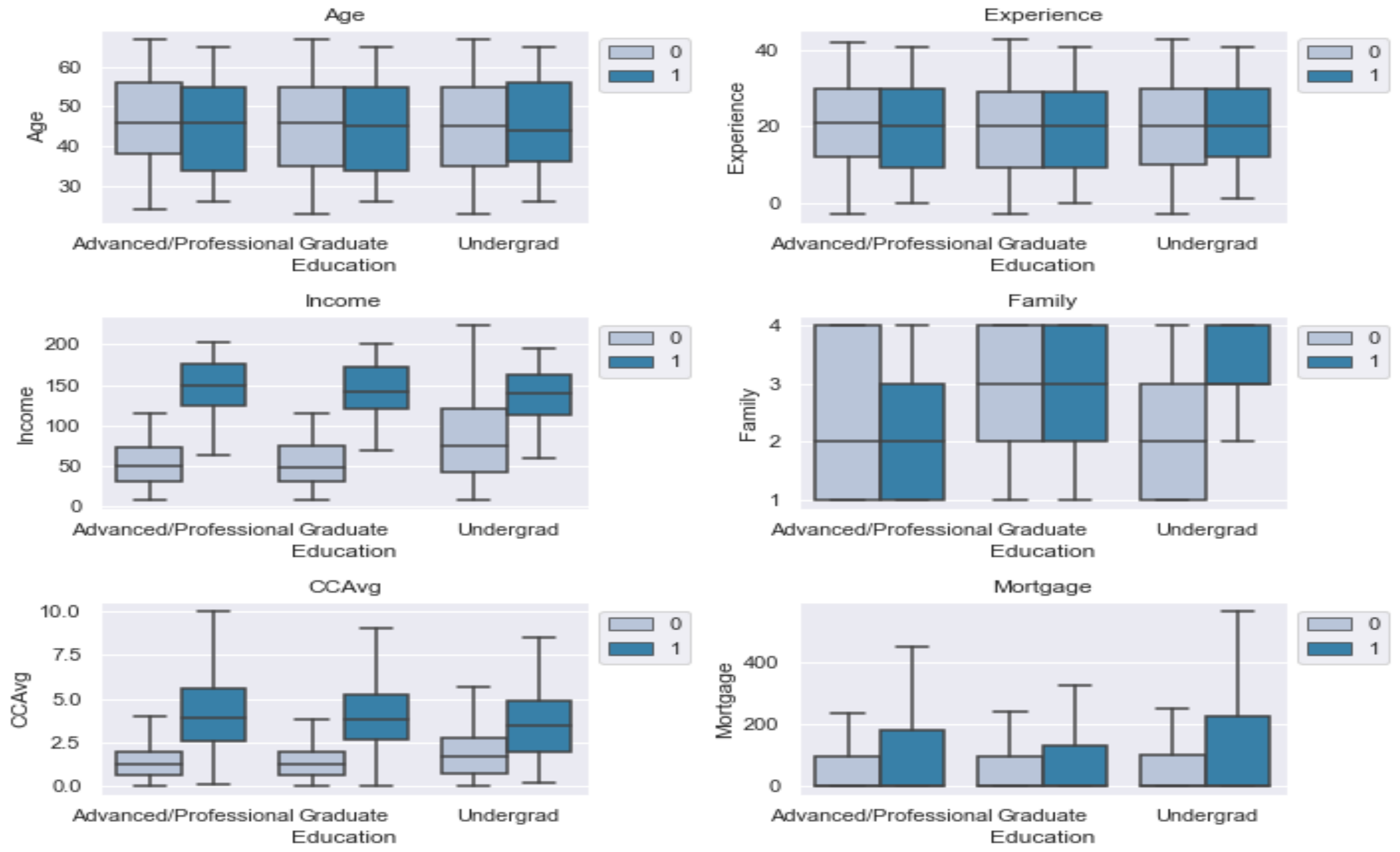
INCOME VS PERSONAL LOAN



- It appears quite obvious that those with a higher income were considered for a loan offer in the previous campaign
- Income clearly is a major feature in loan consideration

EXPLORATORY DATA ANALYSIS

CUSTOMER LEVEL ANALYSIS



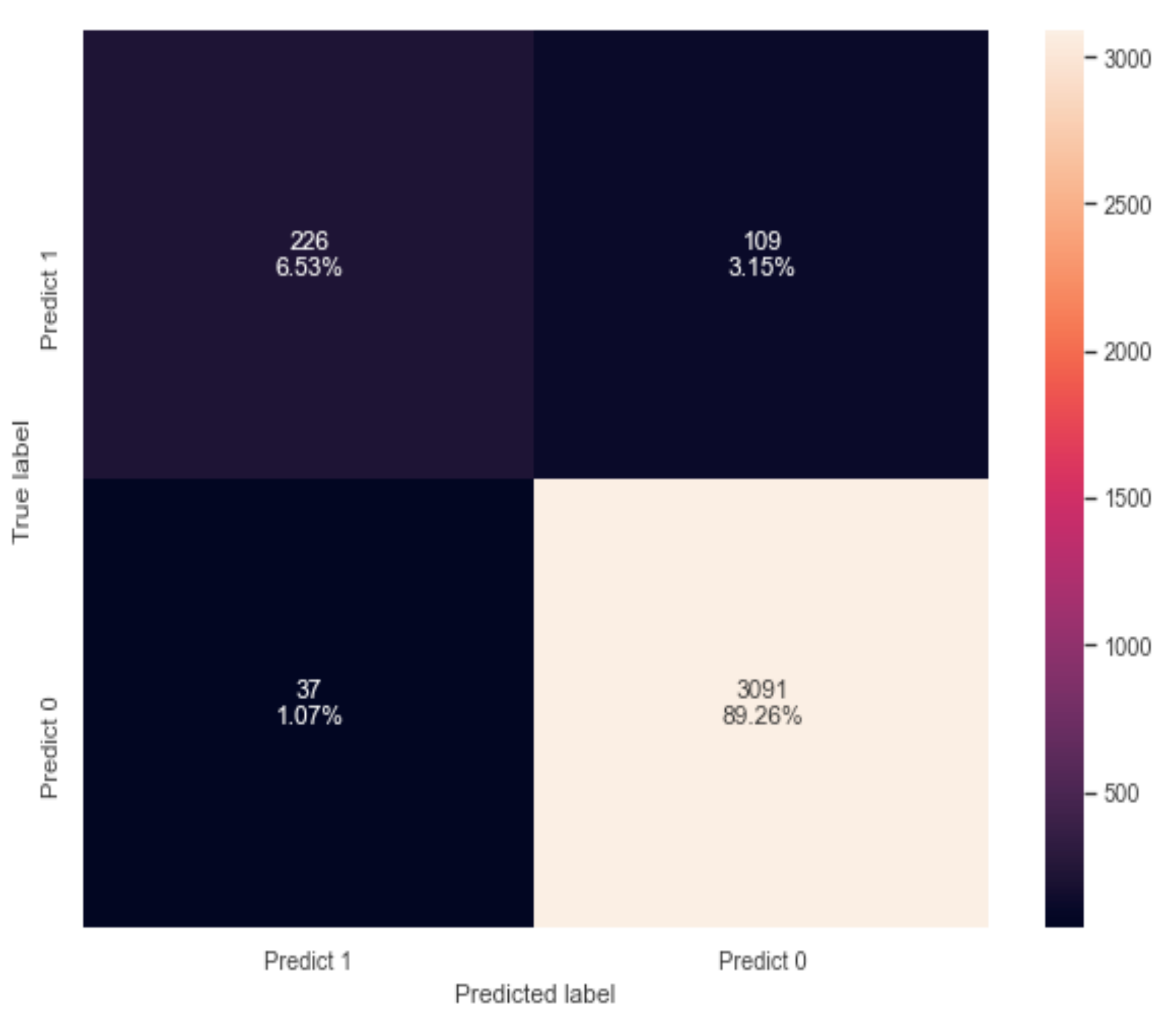
- Advanced Professional customers is a very huge market with potentials as seen
- A significant chunk of customers did not accept personal loans in the last campaign. AllLife has a huge prospect base to grow their bottomline
- Undergrads trumped the other two classes in access to loans but yet maintains the highest number of untapped market without loans. This is a core target for the marketing team
- Foe a wealthy and credit worthy customer base, considerations for loan purchase should target more customers with higher experience and Income
- Clearly Income and Credit card spending are key factors that determines loan purchase
- Advanced Professionals with higher family sizes accessed more loans than their counterparts. Experience counts greatly as well

MODEL PERFORMANCE SUMMARY

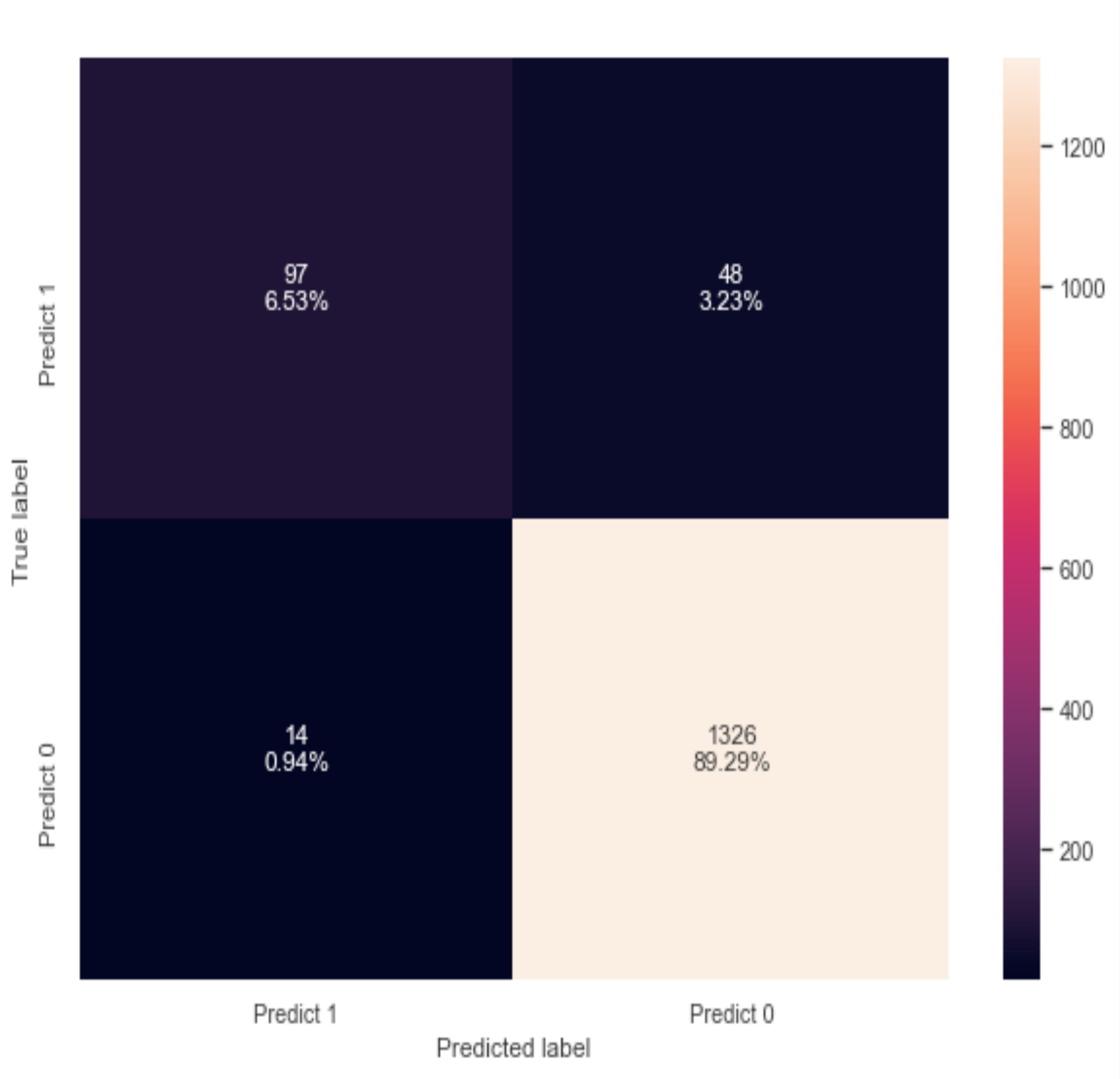
OVERVIEW OF ML MODELS AND PARAMETERS

- Prior to Modeling, Outliers were treated and further Data-Preprocessing was done to identify independent variables viable for the prediction process
- Next step was to create dummy variables using a function that automates On-Hot encoding for a more surgical approach. Education was hot encoded
- Data was Preprocessed and Split into Train and Test sets at a 70:30 Ratio
- Modelling was executed using the Logistic Regression (Sklearn) and Logistics Regression(Statsmodels) and results compared
- Multicollinearity was addressed and a model chosen with improved model performance parameters (Accuracy, Precision, Recall, RUC and f1scores
- As an alternative, the CART model was equally employed for predicting our models where Hyper-Parameter Tuning model parameters were compared to Post Pruning Results after a care iteration of models using the Decision Tree Classifier
- Independent variables (X) against a target or dependent variable(Y=Personal Loan) in this case.

CONFUSION MATRIX FOR PRIOR LR PREDICTION

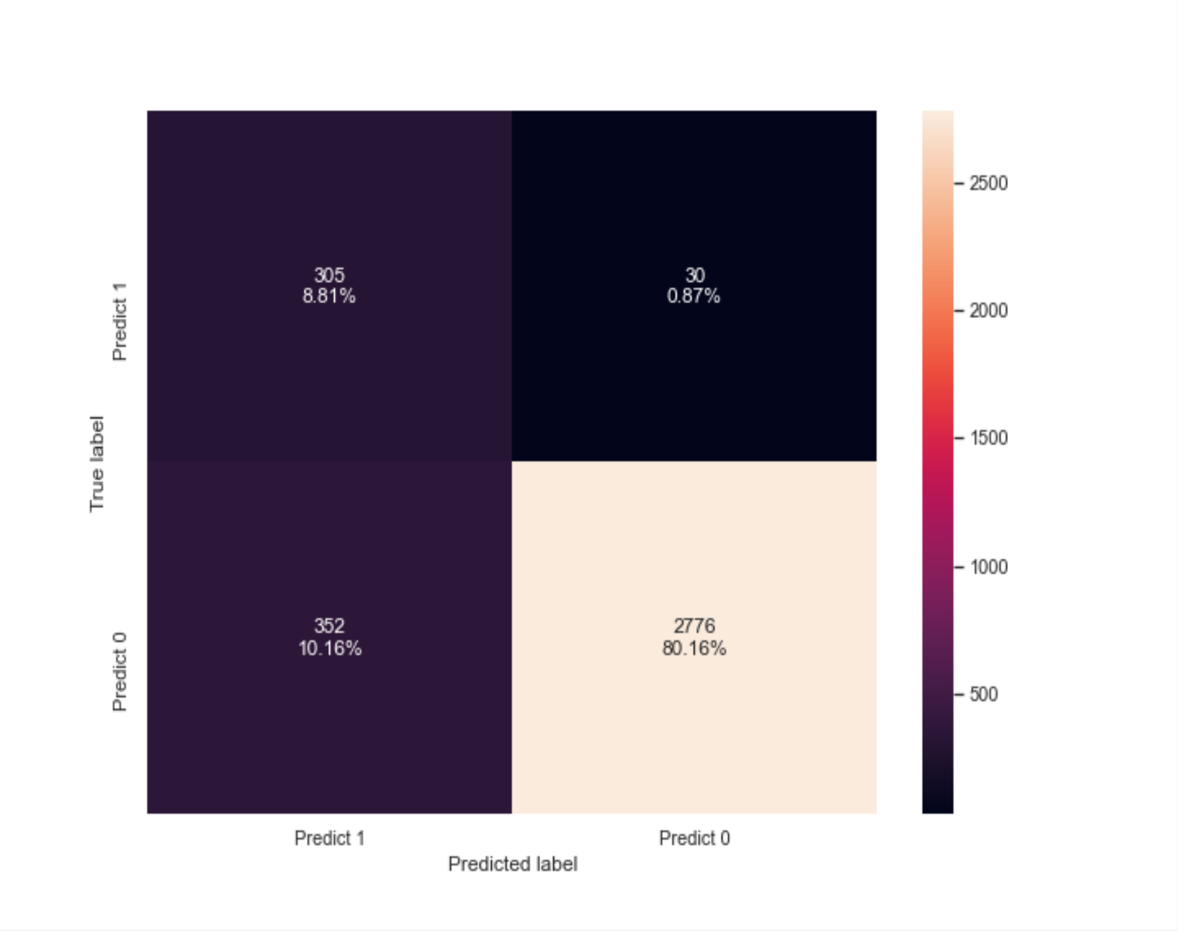


TRAIN SET

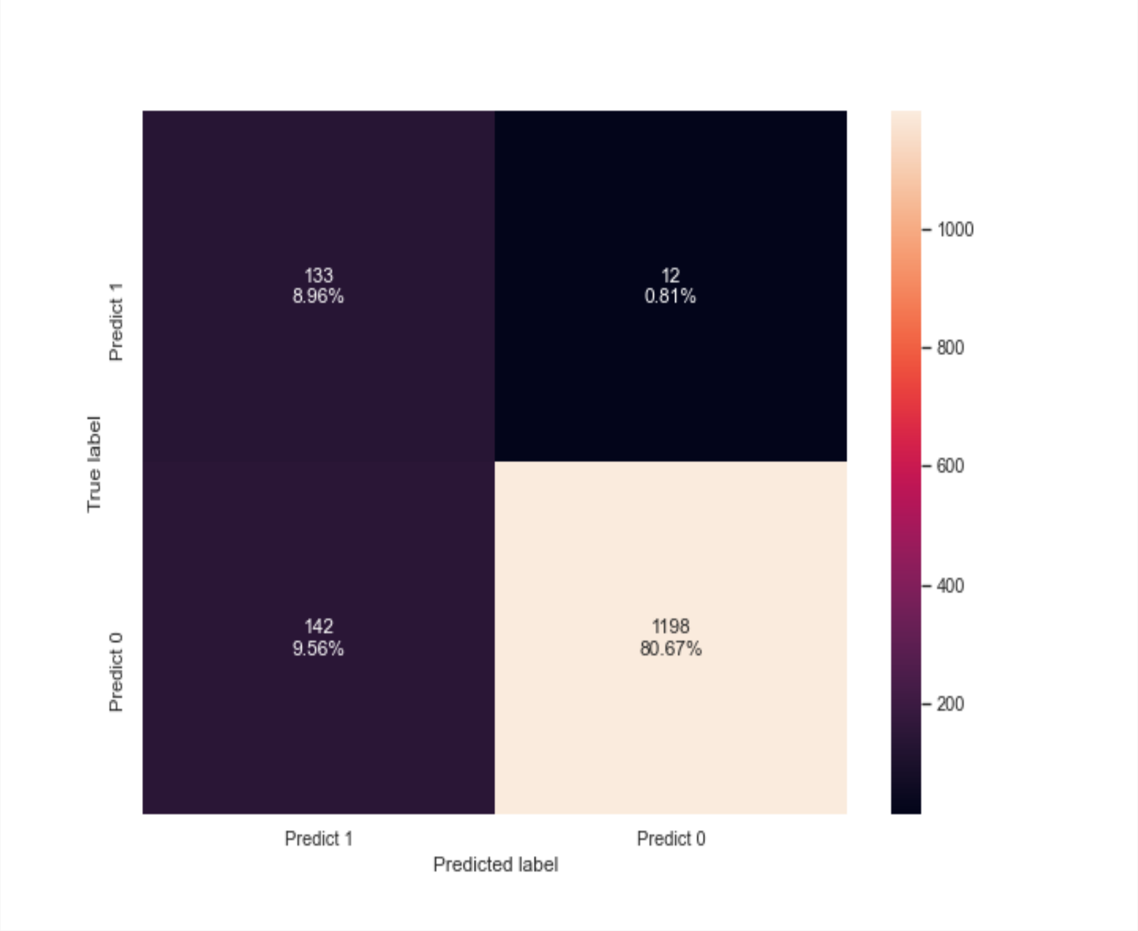


TEST SET

CONFUSION MATRIX FOR POST LR PREDICTION



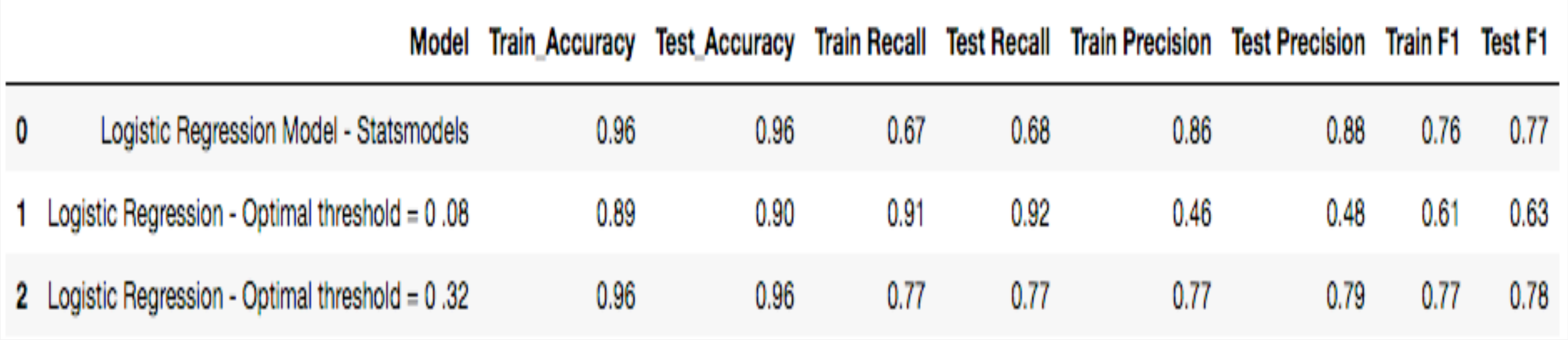
TRAIN SET



TEST SET

After choosing optimal threshold , true positives has increased from 6.5% to 8.96% while false positive has decreased from 3.1% to 0.81%

MODEL PERFORMANCE INSIGHTS- LOGISTICS REGRESSION



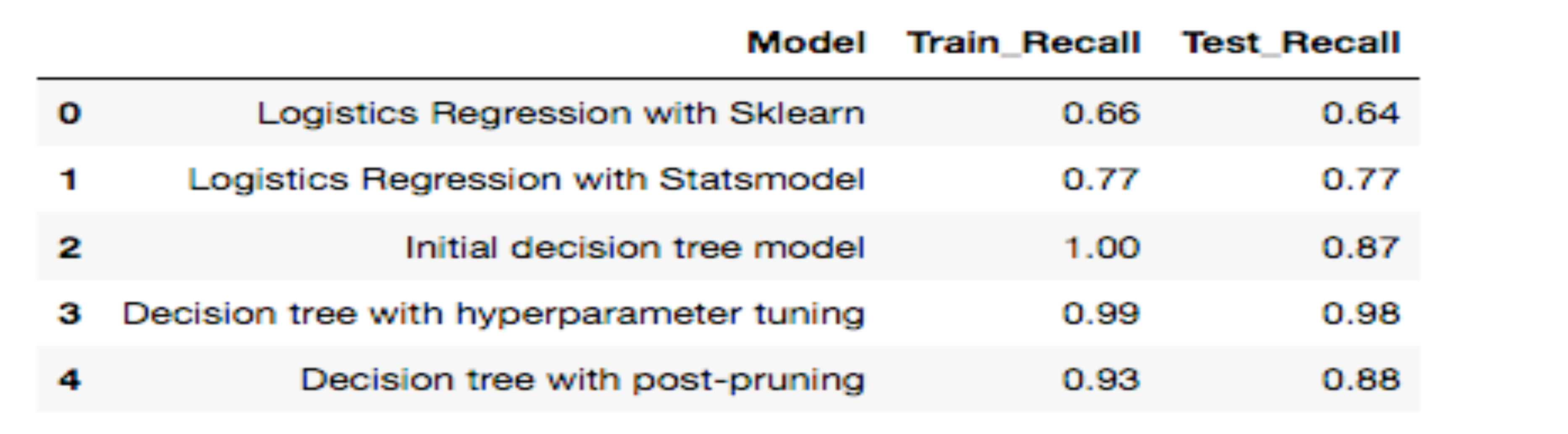
	Model	Train_Accuracy	Test_Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
0	Logistic Regression Model - Statsmodels	0.96	0.96	0.67	0.68	0.86	0.88	0.76	0.77
1	Logistic Regression - Optimal threshold = 0 .08	0.89	0.90	0.91	0.92	0.46	0.48	0.61	0.63
2	Logistic Regression - Optimal threshold = 0 .32	0.96	0.96	0.77	0.77	0.77	0.79	0.77	0.78

- A predictive model that can be used by the AllLife Bank to target potential personal loan purchasers with an f1_score of 0.77 on the training set and formulate marketing strategies accordingly as well as allocate scarce resources toward maximising profit and turnover via fees and good customer traffic accordingly. (Statsmodels - Logistic Regression - with significant predictors).

- Coefficients of some levels of Income, CCAvg, CD_Account, Family are positive for which an increase in these will lead to a proportional increase in chances of being offered a personal loan

- Coefficients of Undergraduate Education,Online Banking, Credit card, are negative increase in these will lead to adecrease in chances of a liability customer getting a personal loan

MODEL PERFORMANCE INSIGHTS-CART(DECISION TREE)



	Model	Train_Recall	Test_Recall
0	Logistics Regression with Sklearn	0.66	0.64
1	Logistics Regression with Statsmodel	0.77	0.77
2	Initial decision tree model	1.00	0.87
3	Decision tree with hyperparameter tuning	0.99	0.98
4	Decision tree with post-pruning	0.93	0.88

- From the table above, it confirms that the Decision Tree model is by far trumps the Logistics regression models
- Easy interpretation is one of the key benefits of Decision Trees.
- Personal Loan Purchase was analyzed using Logistics Regression and Decision Tree Classifier to build a predictive model for the same.
- The models built can be used to predict if a customer is going to contribute to qualify for a loan and ultimately contribute to profit and Revenue generation (by purchasing) or not.
- Different trees and their confusion matrix were analysed to get a better understanding of the model.
- The robustness of Decision Trees to outliers as well as how much less necessary Data manipulation is using CART models is unbeatable.They are also imbalance-proof during analysis.
- Income, CD_Account , Family, Credit card spending(CCAvg) and Creditcard of other Banks owned by liability customers are the most significant variables influencing the dcision to offer Personal loan to prospects .
- We established the importance of hyper-parameters/ pruning to reduce overfitting

IMPORTANT FEATURES BY MODELS

Logistics Regression	Initial Decision Tree	HyperParameter Tuning	Post Pruning
'CD Account	Income	CCAvg	Income
Family	Family	Income	CD_ Account
Income	CCAvg	Credit Card	Family
CreditCard	Experience	Family	CCAvg
Education	Credit card	Education_Undergraduate	Credit card
Education_Undergraduate	Mortgage	Education_Graduate	Educated_Undergraduate
	Education_Undergraduate	Online	Online

- Reference can be made to the Jupyter notebook for Graphical Visualizations of Features

CONCLUSIONS-KEY INSIGHTS FROM EDA

- 75% of the liability customers are aged 55 yrs an below with Advanced Professionals having much higher ages
- Advanced Professionals have a minimum of 50 years experience and average 25years across the distribution
- There is a mean income of approx. $74,000 with the lowest earning $8,000 and the maximum at $224,000. Hence the presence of outliers. It is equally a high income earning distribution.
- 75% of the customers across the distribution maintain a family size of 3 or less, even though advanced professionals with higher family sizes enjoy more benefits
- At least 50% of the customers do not have a mortgage. The distribution is not a highly geared one. This speaks to a healthy credit portfolio held by most customers of the bank
- 75% of the liability customers in the distribution spent 2,500 or less from their credit cards with 1,930 dollars spent on the average. This is a core target.
- Average Credit card spending was at $1,900 with 25% of the customers spending about $700 or less and about 50% spending $1,500 or less. This is indeed a very good credit rating for these customers and can be favored during consideration for loans
- 41.9% of the customers are Undergrads with Advanced/Professional following closely at roughly 37%
- 90.4% of customers did not accept personal loans at the last campaign and 89.6% do not maintain a securities account with AllLife Bank
- 94% do not maintain a Certificate of Deposit with the bank. This clearly is a bracket the marketing team should focus on to design juicy asset products
- 70.6% of the customers do not use credit cards issued by other banks while 29.4% do. This speaks to an untapped opportunity to sell personal loans
- Berkeley, CA; Stanford CA; Davis ,CA; Los Angeles, CA and Isla Vista, CA locations have the highest customer base by distribution in the state of California
- Customers with higher income tend to engage spend more from their credit cards. This is a core determinant for loan qualification
- Advanced professionals have a 15% chance of being granted loans compared to Grads and Undergrads based on Experience
- There is 50% chance for customers with Certificate of Deposits to be considered for a loan compared to those without

BUSINESS RECOMMENDATIONS

Based on the key insights regarding the granting personal loans to liability customers in AllLife Bank, the following are ideal recommendations to the board;

- These insights can be inferred to form the basis of a formidable marketing vis-à-vis advert campaigns to gain a competitive edge, more market share and ultimately increased revenue
- The huge bracket of undergrads may present a high risk but the marketing team could design low priced assets to attract this group which could ultimately generate the much needed critical mass for AllLife Bank and growing retained balances for the Bank in the near furture
- As evidenced in the analysis, an extensive and aggressing marketing initiative ought to be considered to tap into the significant chunk of customers with potentials for a huge certificates of deposit which obviously serves as a security for any personal loan granted as well as a retained deposit which offers AllLife Bank the value added opportunity to make a convenient spread as income as well as cash management and handling fees.
- The present campaign should embrace considerations should be purely based on factors that will favor an asset portfolio design with value added incentives for huge credit card spenders as this oils the wheels of more turnover and income for the bank
- Assets(Loans) should be priced competitively with value added services to enable customer retention and referral that will ultimate spike an attendant traffic in patronage and ultimately revenue
- A sizeable chunk of customers without credit cards of other banks presents a viable opportunity to drive the growth in the sales of personal loans tied to AllLife's Credit card products which in turn drives the asset base and income of AllLife Bank
- It is quite apparent the training data is not exhaustive. More effort should be geared toward leveraging on a wider observation sample to draw more inclusive, extensive and impactful insights
- There are instances of customers without mortgage ,maintaining huge deposits and are high income earners. As well as Advanced professionals who clearly exceptional and personalised service solutions would endear to take up personal loans. As such the marketing team should develop Premium asset-class products for High-Networths as they are known

THANK YOU