PERSONAL LOAN PREDICTION
ALL LIFE BANK

Project 4 - Supervised Learning - Classification

Authored by C. Kamakani Kahunahana

BUSINESS OVERVIEW
PROBLEM AND SOLUTION APPROACH

PROBLEM

Only a small percentage of All Life Bank's are taking advantage of personal loans. All Life is looking to increase revenues by increasing the number of customers who take out personal loans.

AllLife Bank is a US bank that has 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 rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio.

SOLUTION APPROACH

Build a predictive pricing model for to predict which customers would be good candidates to target for personal loans using Logistic Regression and Decision Tree Modelling.

Using data provided, we will identify key variables from their customer database to identify which characteristics are best at identifying which customers are more likely to take out a personal loan.

Based on this analysis I will build and test two types of models, logistic regression and decision tree, to help management make better business and marketing decisions.

Objectives:

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

BASE DATASET OVERVIEW

Details	Amount
Observations	5000
Variables	14
Null	None

This is a description of the base dataset however, much transformation will need to be done prior to performing any analysis or model construction

Variable	Type	Description
ID	Integer	Customer ID
Age	Integer	Customer's age in completed years
Experience	Integer	#years of professional experience
Income	Integer	Annual income of the customer (in thousand dollars)
ZIPCode	Integer	Home Address ZIP code
Family	Integer	Family size of the customer
CCAvg	Float	Average spending on credit cards per month (in thousand dollars)
Education	Integer	1: Undergrad; 2: Graduate;3: Advanced/Professional
Mortgage	Integer	Value of house mortgage if any. (in thousand dollars)
Personal_Loan	Integer	Did this customer accept the personal loan offered in the last campaign? (Dependent Variable)
Securities_Account	Integer	Does the customer have securities account with the bank?
CD_Account	Integer	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Integer	Do customers use internet banking facilities?
CreditCard	Integer	Does the customer use a credit card issued by any other Bank (excluding All life Bank)?

01
DATA PREPROCESSING

DATA PRE-PROCESSING

PROCESSING VARIABLES AND UNNECESSARY DATA

Variable	Pre-Processing Strategy
ID	Identical to index; will be dropped during model building
Experience	Perfect correlation (1.0) with Age will be dropped
ZIPCode	Using uszipcode database to identify City, County, and State and split into separate columns
Income	Wide variation and outliers in the upper range; the data was grouped into categories based on distribution [‘<25K','26-50K','51-75K','76-100K','101-125K','126-150','>151K']
CCAvg	Wide variation and outliers in the upper range; the data was grouped into categories based on distribution ['<1K','1<2K','2<3K','3<4K','4<5K','5<6K','>6K']
Mortgage	Wide variation and outliers in the upper range; the data was grouped into categories based on distribution ['none','<100K','101-200K','201-300K','301-400K','>401K']

- Dropped ID
- Experience dropped because correlation to Age
- Convert ZIPCode to City, County, and State (State dropped because 100% California)
- Create categorical variables for Income, CCAvg, and Mortgage

DATASET OVERVIEW
NUMERICAL VALUES

NUMERICAL VARIABLE DESCRIPTION

	Count	Mean	Standard Deviation	Minimum	25%	50%	75%	Maximum
Age	5000	45.35	11.46	23.00	35.00	45.00	55.00	67.00
Experience	5000	20.10	11.46	-3.00	10.00	20.00	30.00	43.00
Income	5000	73.77	46.03	8.00	39.00	64.00	98.00	224.00
ZIPCode	5000	NA	NA	NA	NA	NA	NA	NA
Family	5000	2.40	1.15	1.00	1.00	2.00	3.00	4.00
CCAvg	5000	1.98	1.75	0.00	0.70	1.50	2.50	10.00
Education	5000	1.88	0.84	1.00	1.00	2.00	3.00	3.00
Mortgage	5000	56.50	101.71	0.00	0.00	0.00	101.00	635.00
Personal_Loan	5000	0.10	0.29	NA	NA	NA	NA	NA
Securities_Account	5000	0.10	0.31	NA	NA	NA	NA	NA
CD_Account	5000	0.06	0.24	NA	NA	NA	NA	NA
CreditCard	5000	0.60	0.49	NA	NA	NA	NA	NA
Online	5000	0.49	0.46	NA	NA	NA	NA	NA

- Negative values in experience are replaced with 0 as someone cannot have negative years of professional experience

NEW DATASET OVERVIEW
CATEGORICAL VALUES

CATEGORICAL VARIABLE DESCRIPTION

	City	County	Income_Group	CCAvg_Group	Mortgage_Group
Count	5000	5000	5000	5000	5000
Unique	245	39	7	7	6
Top	Los_Angeles	Los_Angeles	26-50K	<1K	none
Frequency	375	1095	1242	1683	3462

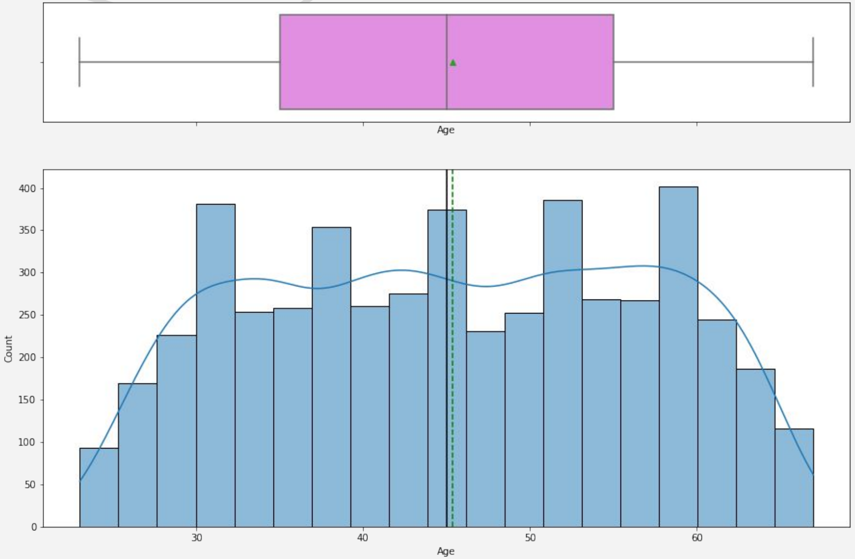
- Location of Los Angeles is the most common. There are 375 different cities identified.
- There are 39 different counties with Los Angeles being the most common
- Income was split into 7 sub-groups with annual salary of between $26K-50K being the most common
- Credit Card Average balance carried was split into 7 sub-groups with < $1K being the most common
- Mortgage balance carried was split into 6 sub-groups with no mortgage being the most common

02

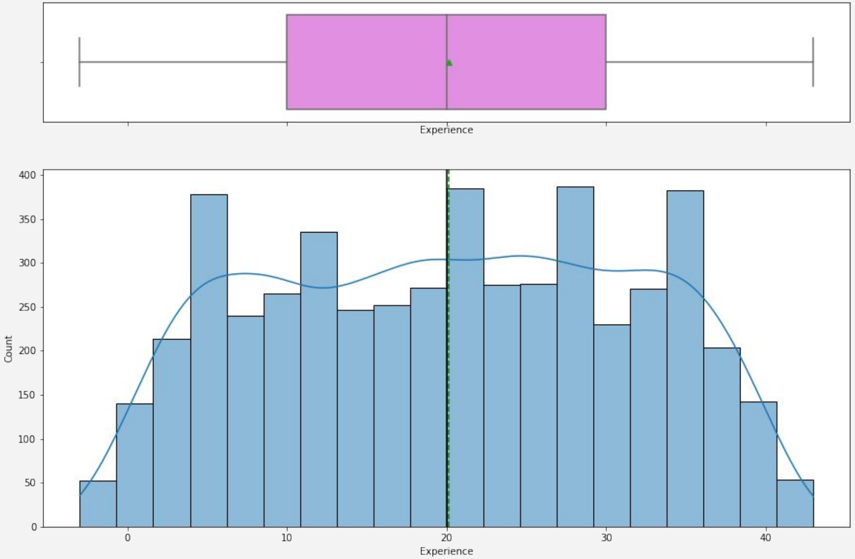
EXPLORATORY DATA ANALYSIS

CUSTOMER ANALYSIS
AGE AND EXPERIENCE

CUSTOMER AGE



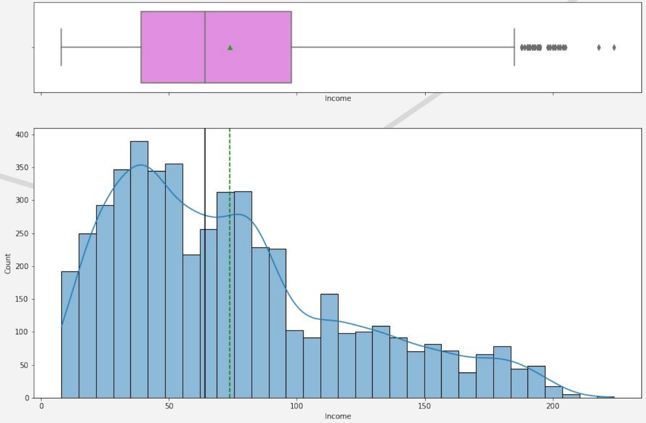
YEARS OF PROFESSIONAL EXPERIENCE



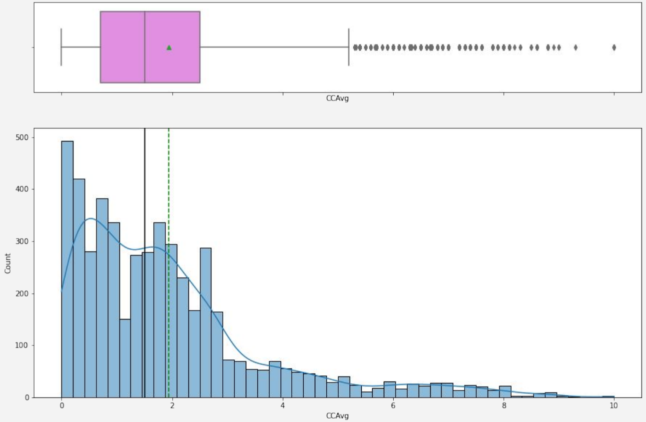
- Mean and Median age are both 45
- Mean and Median experience are both 20 years
- Correlation between the data sets is 1.0
- Experience is dropped from further analysis

CUSTOMER ANALYSIS
INCOME, CCAVG, AND MORTGAGE

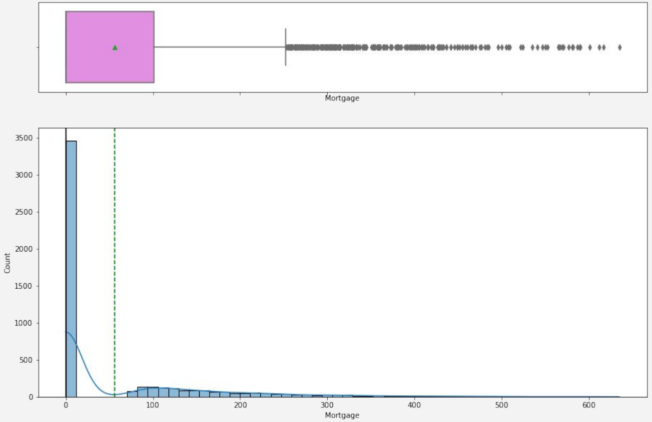
ANNUAL INCOME



AVERAGE CREDIT CARD BALANCE



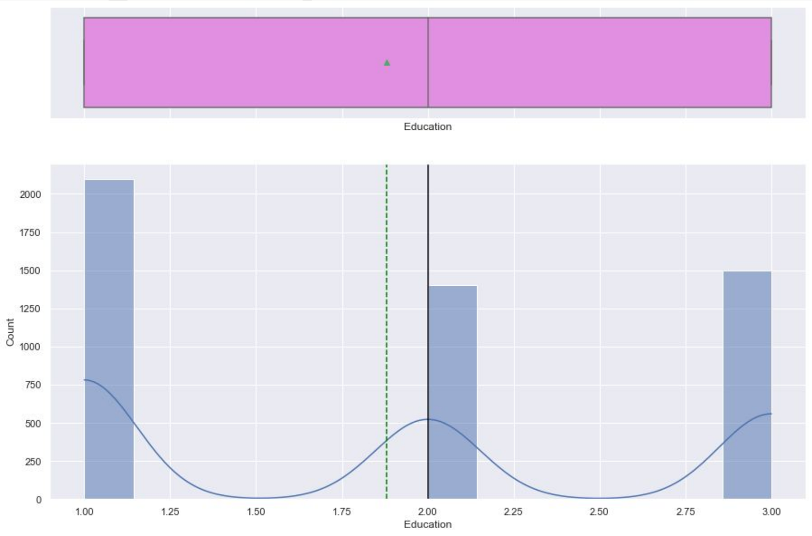
MORTGAGE BALANCE



- All three plots have a right skew with many outliers in the upper range
- May need to be trimmed or capped or grouped for further analysis
- Median Income is $64K with min of $8K and max of $224K
- Median CC Balance is $1.5K with min of $0 and max of $10K
- Median Mortgage is 0 hence most do not have mortgages, mean is $54.5K and max is $635K

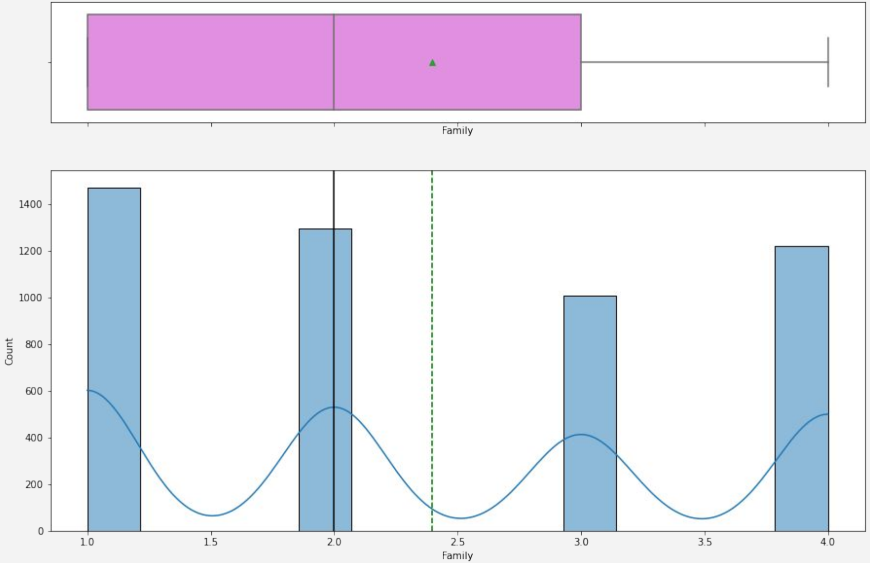
CUSTOMER ANALYSIS
EDUCATION AND FAMILY

YEARS OF EDUCATION



- Education has three values in the dataset 1, 2, and 3

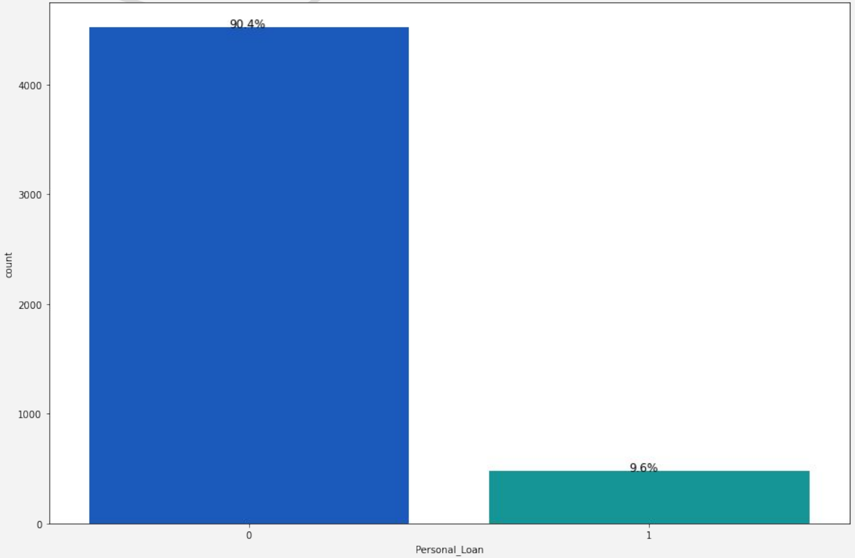
FAMILY SIZE



- Customer family size is fairly evenly distributed

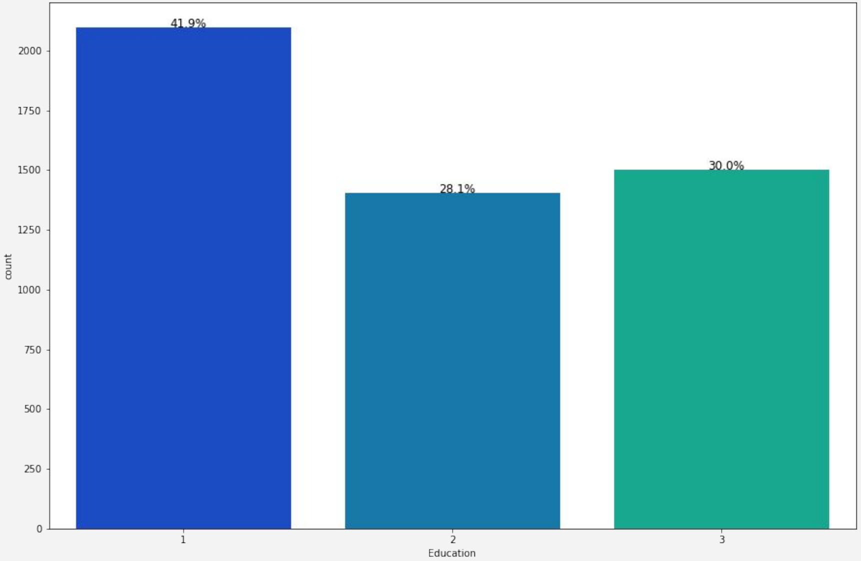
CUSTOMER ANALYSIS
PERSONAL LOAN AND EDUCATION

PERSONAL LOAN (Y/N)



- Personal loan is our dependent variable
- 90.4% if customer do not have a personal loan

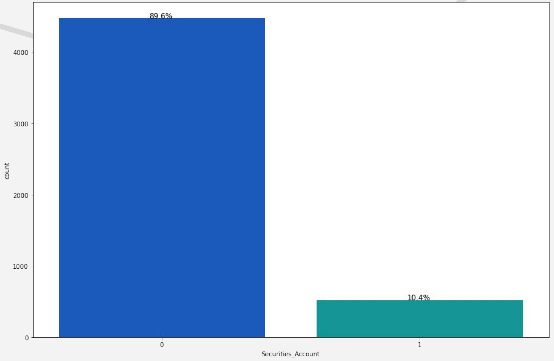
EDUCATION LEVEL



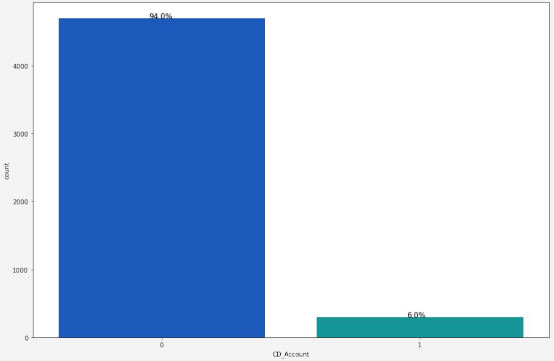
- 41.9% of customers report an education level of Undergrad
- Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional

CUSTOMER ANALYSIS
SECURITIES ACCOUNT, CD ACCOUNT, CREDIT CARD ACCOUNT, AND ONLINE BANKING

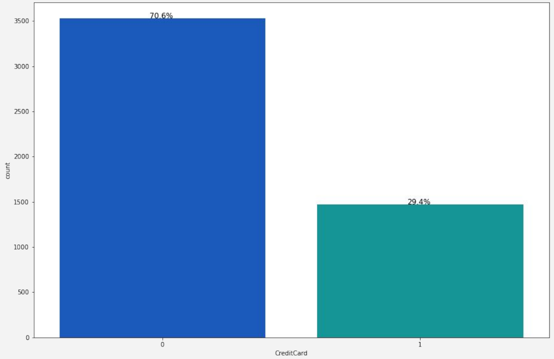
SECURITIES ACCOUNT (Y/N)



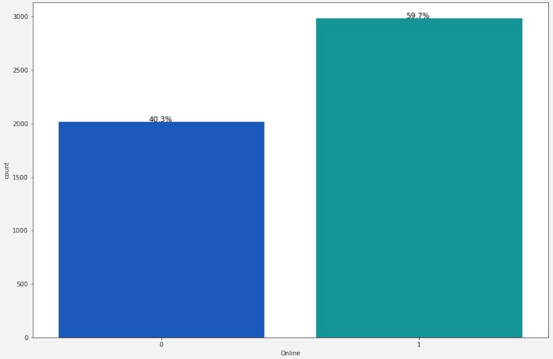
CD ACCOUNT (Y/N)



CREDIT CARD ACCOUNT (Y/N)



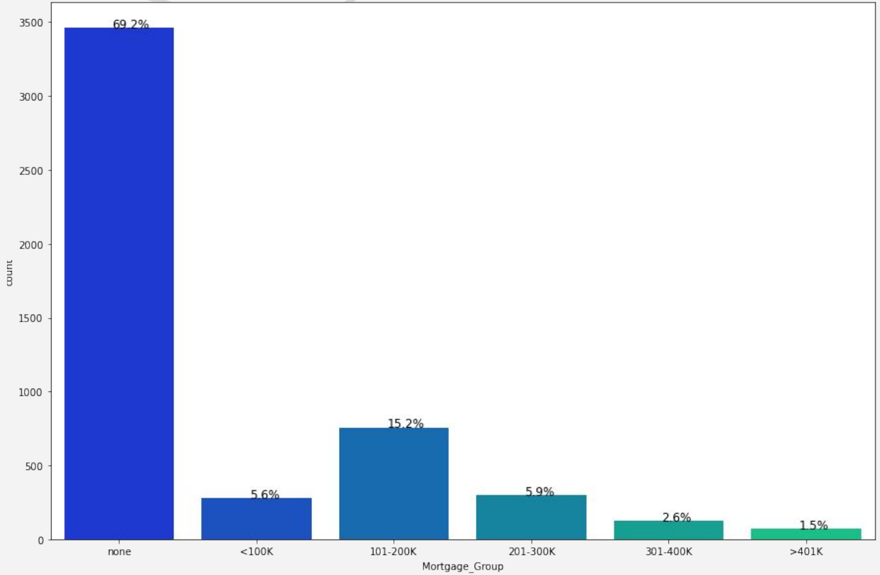
ONLINE BANKING (Y/N)



- 89.6% of customers do not have a securities account
- 94% customers do not have a CD account
- 70.6% of customers do not have a credit card with a bank other than All Life
- 40.3% of customers do not use internet banking

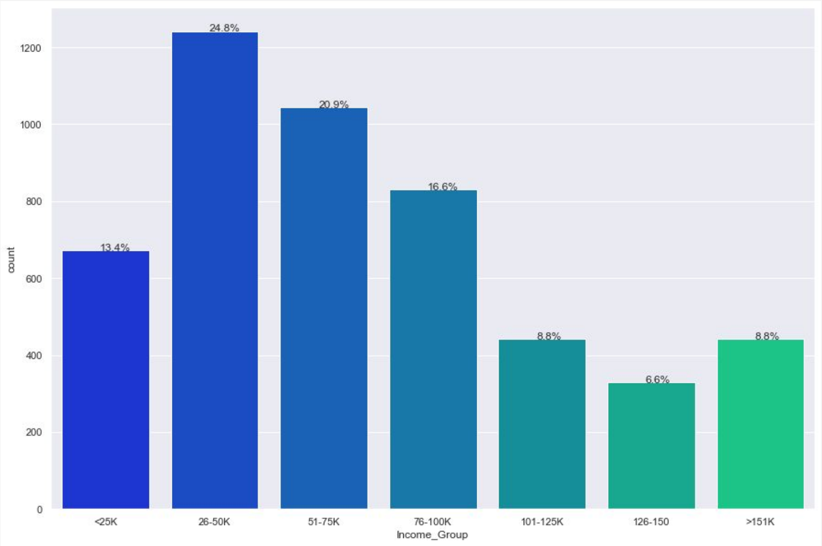
CUSTOMER ANALYSIS
MORTGAGE AND INCOME

MORTGAGE BALANCE



- 69.2% of customers do not carry a mortgage

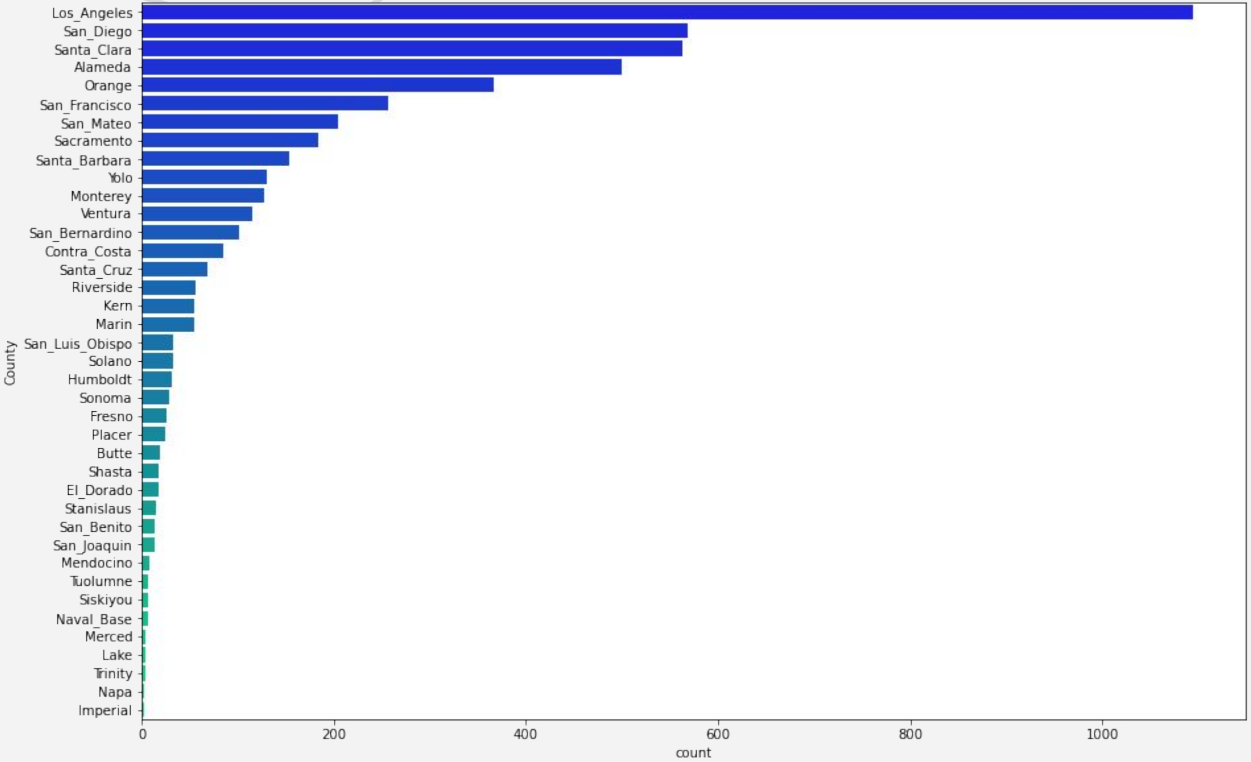
ANNUAL INCOME



- 24.8% earn an annual salary between $25-50K
- 8.8% earn more than $151K

CUSTOMER ANALYSIS
COUNTY OF HOME ADDRESS

HOME COUNTY

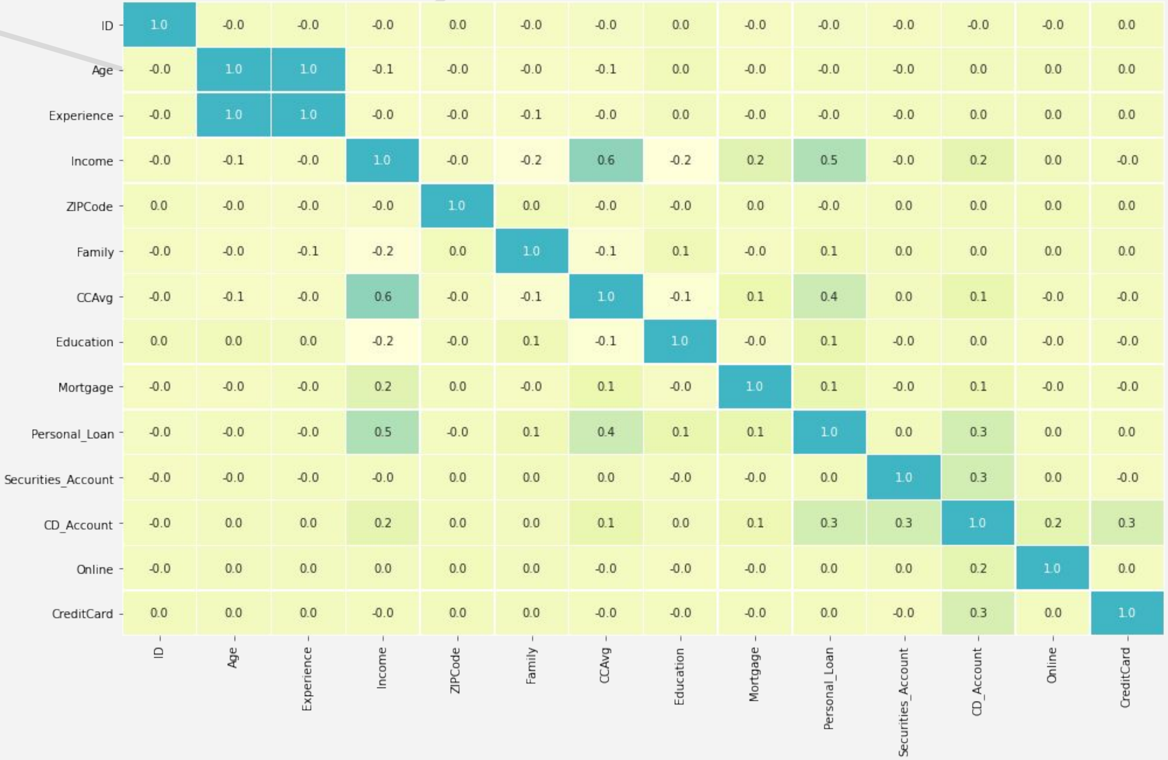


- 39 different counties reported
- Most customers live in Los Angeles and San Diego followed by Santa Clara and Alameda

03

BIVARIATE
& MULTIVARIATE
ANALYSIS

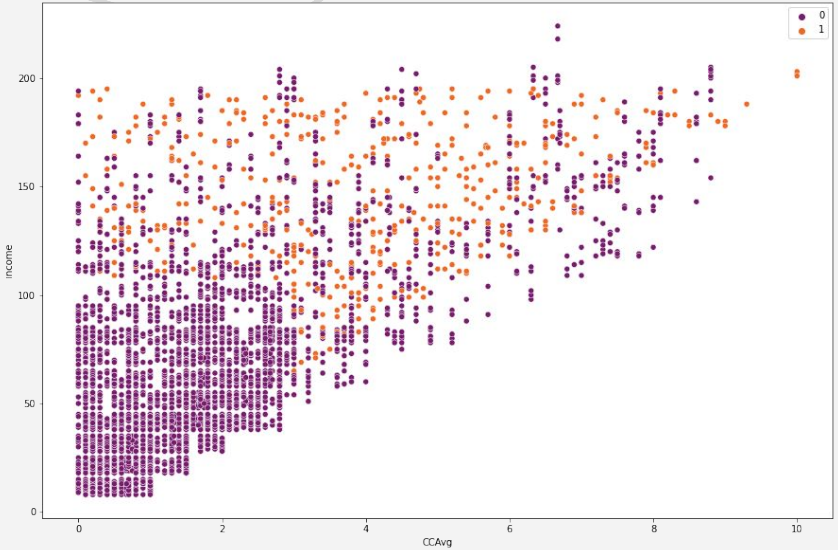
CUSTOMER ANALYSIS
CORRELATION MATRIX



- Age and Experience are perfectly correlated
- Income, CCAvg, and Mortgage have some relationship to each other
- CD Account has a 0.3 correlation with Personal Loan, Securities Account, and Credit Card
- Personal Loan has slight correlation to Income and CCAvg

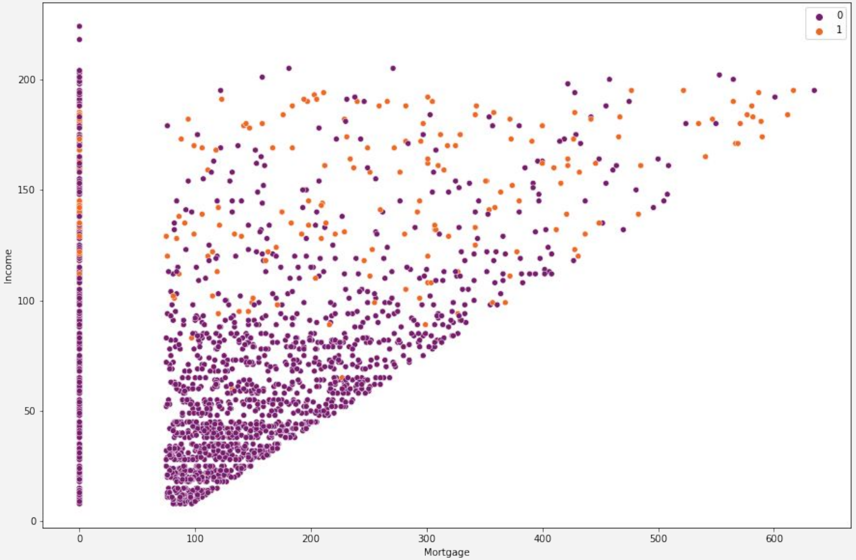
PERSONAL LOAN ANALYSIS
CREDIT CARD AND INCOME

CC AVG & INCOME VS PERSONAL LOAN



- Credit card average balance and income are good indicators of personal loans
- Seems to be a strong relationship between Credit card average and mortgage balance

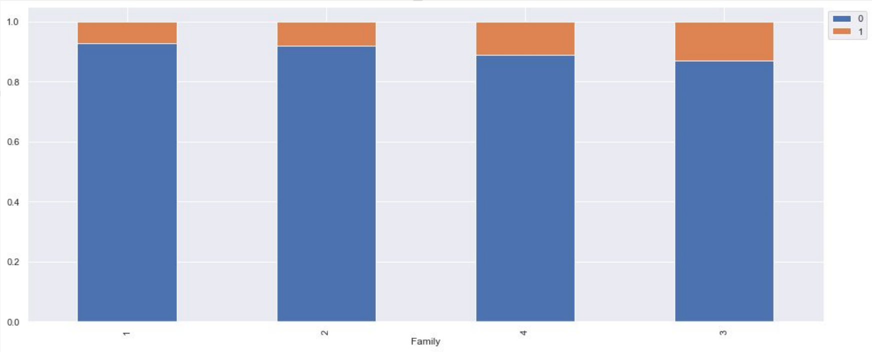
MORTGAGE & INCOME VS PERSONAL LOAN



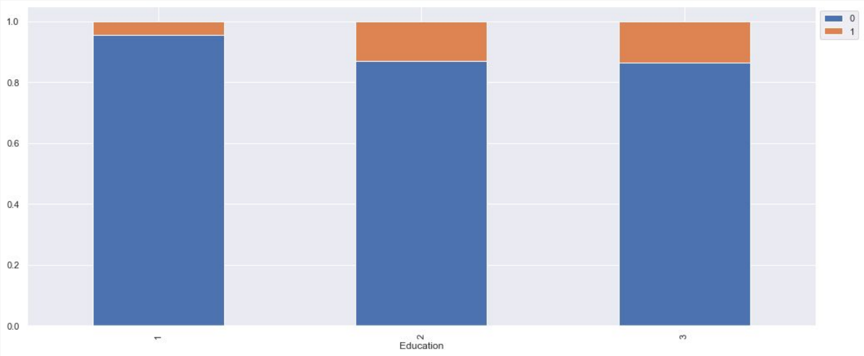
- Mortgage amount and Income seem to have strong relationships with whether or not a customer has a personal loan

PERSONAL LOAN ANALYSIS
FAMILY SIZE AND EDUCATION LEVEL

FAMILY SIZE VS PERSONAL LOAN



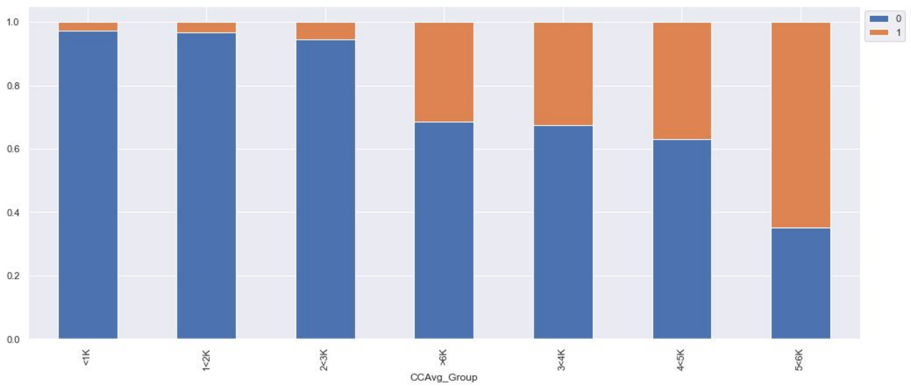
EDUCATION LEVEL VS PERSONAL LOAN



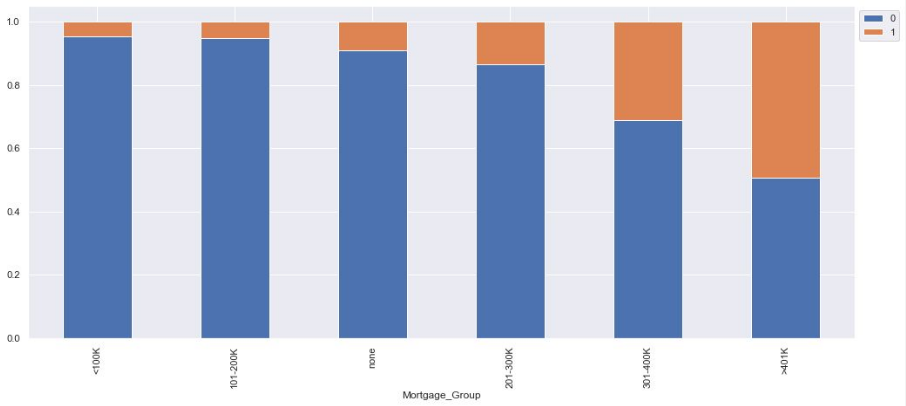
- Families of 3 and 4 persons have more personal loans
- Higher educated customers tend to take out more personal loans

PERSONAL LOAN ANALYSIS
CC AVERAGE AND MORTGAGE BALANCE

CREDIT CARD AVERAGE BALANCE VS PERSONAL LOAN



MORTGAGE BALANCE VS PERSONAL LOAN

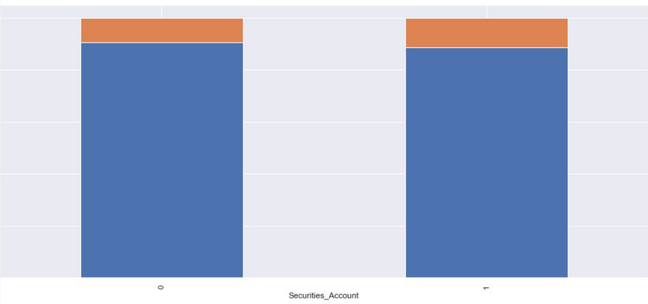


- Customers with higher credit card balances are more likely to take out a personal loan
- Customers with higher mortgages are more likely to take out a personal loan

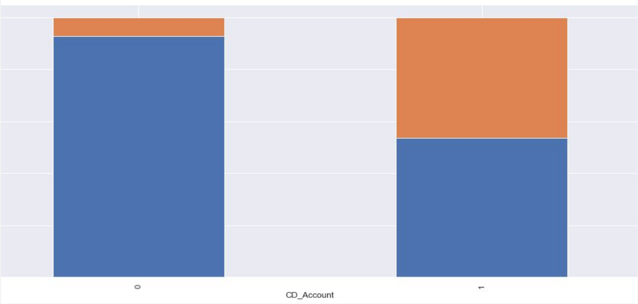
CUSTOMER ANALYSIS

SECURITIES ACCOUNT, CD ACCOUNT, CREDIT CARD ACCOUNT, AND ONLINE BANKING

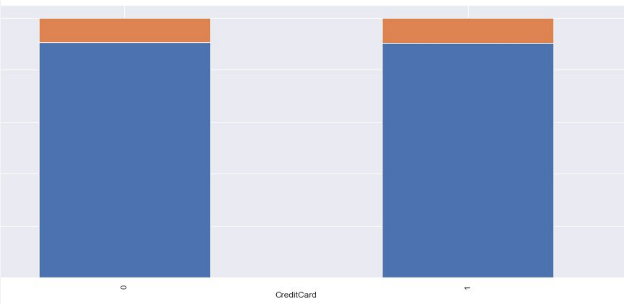
SECURITIES ACCOUNT



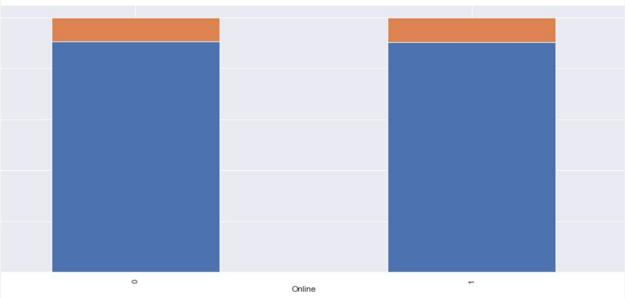
CD ACCOUNT



CREDIT CARD ACCOUNT



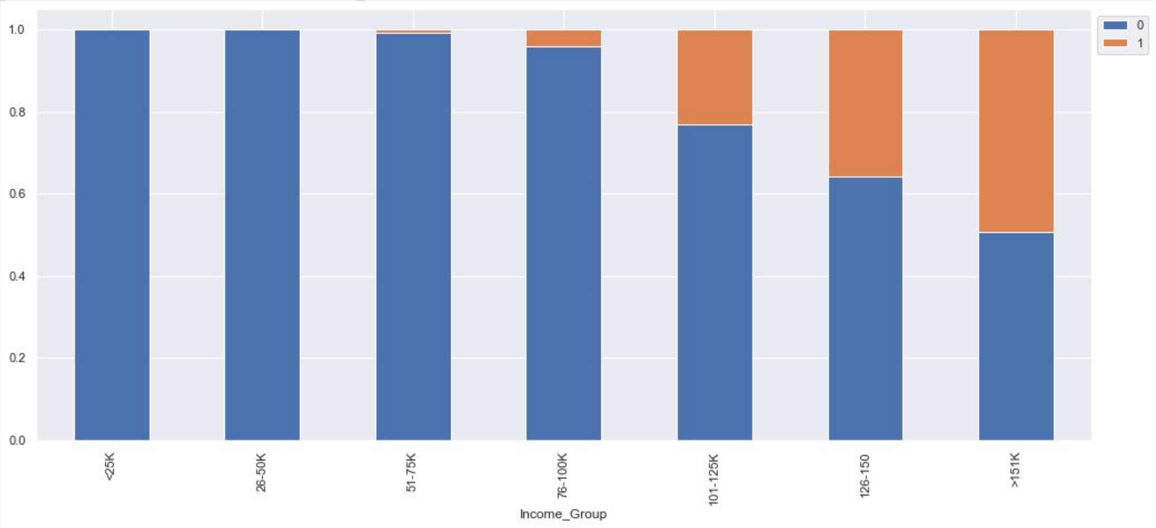
ONLINE BANKING



- Securities, credit card or online banking accounts are not good indicators of personal loans
- Customer with CD Accounts are more likely to take out a personal loan

PERSONAL LOAN ANALYSIS
ANNUAL INCOME

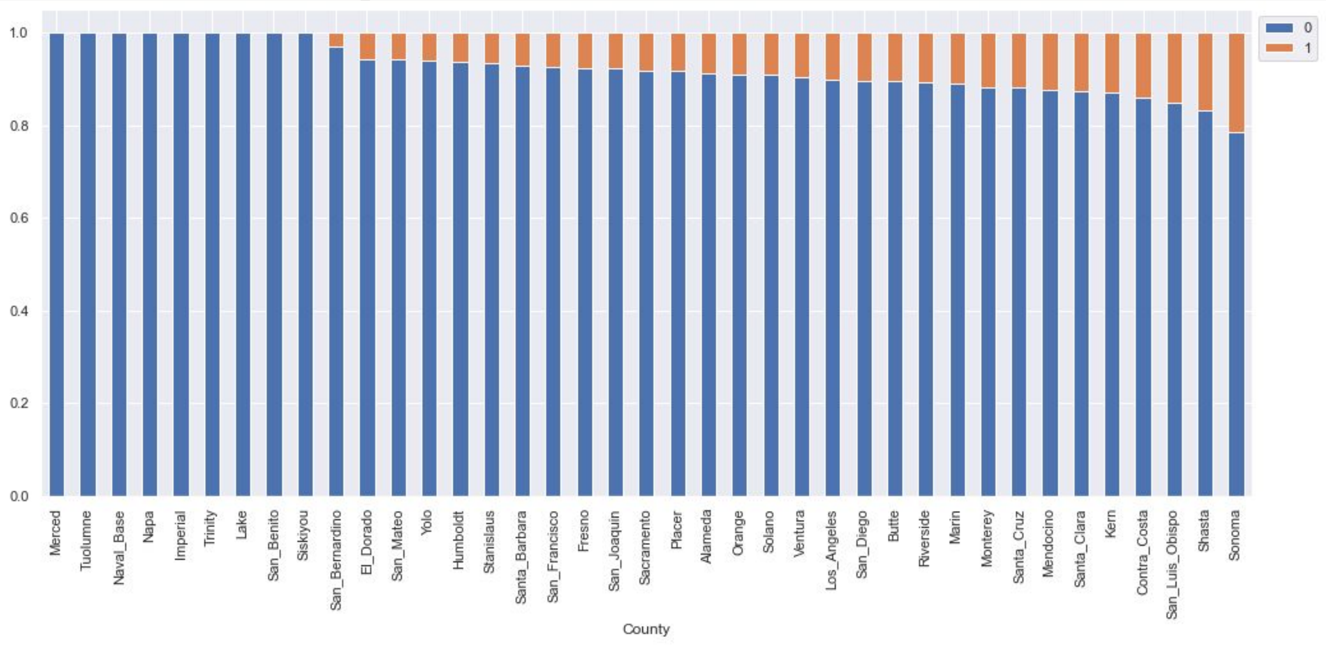
ANNUAL INCOME



- Higher income customers are more likely to take out a personal loan

PERSONAL LOAN ANALYSIS
HOME COUNTY

HOME COUNTY



- Customers in Northern California tend to take out more loans than those in Southern California

EXPLORATORY DATA ANALYSIS
SUMMARY CONCLUSION AND OBSERVATIONS

CUSTOMER DEMOGRAPHICS

- 45 years old with 20 years of professional experience
- $64K annual income
- $1.5K in credit card balances
- No mortgage
- Educated with 41.9% having at least some college
- 89.6% of customers do not have a securities account
- 94% customers do not have a CD account
- 70.6% of customers do not have a credit card with a bank other than All Life
- 40.3% of customers do not use internet banking
- Most live in Los Angeles

PERSONAL LOAN CUSTOMER ANALYSIS

- Families of 3 and 4 persons have more personal loans
- Higher educated customers tend to take out more personal loans
- Customers with higher credit card balances are more likely to take out a personal loan
- Customers with higher mortgages are more likely to take out a personal loan
- Securities, credit card or online banking accounts are not good indicators of personal loans
- Customer with CD Accounts are more likely to take out a personal loan
- Higher income customers are more likely to take out a personal loan
- Customers in Northern California tend to take out more loans than those in Southern California

RECOMMENDATIONS

- Al life should look to target upper income customers with higher education living in Northern California who hold higher mortgages, credit card balances and have larger families for personal loans.

- 69.2% of All Life Bank customers do not currently have a mortgage which could present an opportunity to target them with mortgage products and services

04

MODEL BUILDING

LOGISTIC REGRESSION

BUILDING THE LOGISTIC REGRESSION MODEL

Approach

1. Data Preparation
2. Partition into training (70%) and test data (30%)
3. Run Model
4. Remove Collinearity
5. Observe variable coefficients
6. Measure performance using
    a. Accuracy and Recall

Objectives

- 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.

Model can make wrong predictions as:

- Predicting a person has a Personal Loan but actually does not.
- Predicting a person doesn't have a Personal Loan but actually they do.

Which case is more important? Both are important.
Because this is only a target market model and not an underwriting model we can accept some false positives. However, too many false positives could lead to an inefficient allocation of marketing spend and man hours.

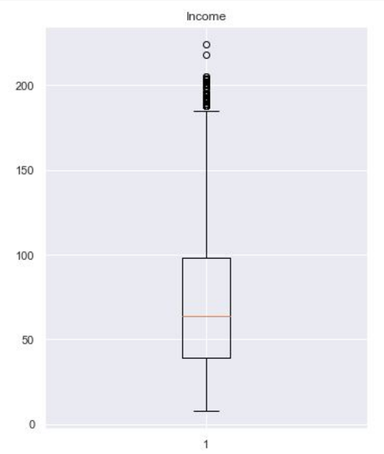
How to reduce losses?
We can use accuracy since the downside of a False Positive is minimized. We are only attempting to predict whether or not a customer will apply for a loan not whether or not they will default. For this application, we want to identify all customers likely to apply for a loan and can evaluate their application for approval at a later step in the process.

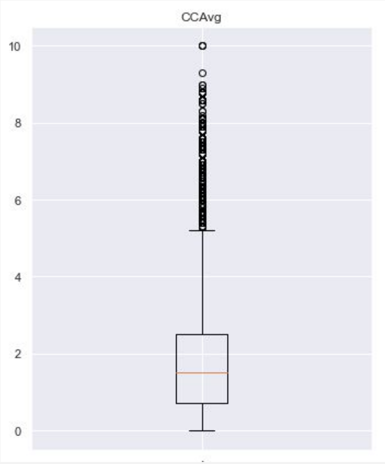
However, Recall and F1_score should also be maximized, the greater the f1_score higher maximizes both Accuracy and Recall. Recall can help us be more efficient with our marketing by reducing the number of customers we market to who are less likely to apply for a loan.

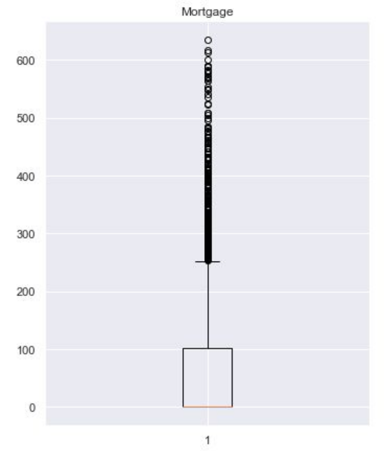
A good balance of both Accuracy and Recall are good metrics for this project.

DATA PREPARATION
OBSERVE OUTLIERS

OUTLIERS







- Income, CCAvg, and Mortgage have outliers in the upper range
- However, these values seem rationale and are likely to reoccur therefore, we will choose to bucket them into ranges rather than trim or cap
- Trimming and capping at the 1.5*IQR was modelled but did not produce results superior to bucketing while increasing model complexity

OUTLIERS TREATMENT
FINAL DATASET BEFORE DUMMY VARIABLES

VARIABLE DESCRIPTION

Variable	Type	Description
Personal_Loan	**Integer**	**This is our dependent variable**
Age	Integer	Customer's age in completed years
Family	Integer	Values are 1, 2, 3, or 4
Education	Integer	Values are 1, 2, or 3
Securities_Account	Integer	Values are 1 and 0
CD_Account	Integer	Values are 1 and 0
CreditCard	Integer	Values are 1 and 0
Online	Integer	Values are 1 and 0
Mortgage_Group	Category	OneHotEncode into dummy variables
Income_Group	Category	OneHotEncode into dummy variables
CCAvg_Group	Category	OneHotEncode into dummy variables

- CCAvg_Group was eventually dropped due to high collinearity with Income and Mortgage

FINAL MODEL PERFORMANCE
METRICS

MODEL PERFORMANCE SUMMARY

Metric	Train	Test
Accuracy	0.9606	0.9533
Recall	0.8429	0.8121
Precision	0.7643	0.7423
F1 Score	0.8017	0.7756
ROC-AUC Score	0.9079	0.8905

- Accuracy score of 95.3% exceeds the base accuracy of 90.4% indicating the model has some predictive value
- Using the Recall measure, of the customers with a personal loan the model can label them correctly 81.2% of the time.
- Precision tells us that of all the customers labeled as loan borrowers, 74.2% were actual borrowers
- F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.
- ROC-AUC Score close to 1.0 indicates a good classification model. A score of 0.89 is a good score.

FINAL MODEL PERFORMANCE
COEFFICIENTS AND RECOMMENDATION

COEFFICIENT SUMMARY

Variable	Coefficient
Constant	-31.4202
Family	0.7723
Education	1.7758
Securities_Account	-1.2856
CD_Account	4.0128
Online	-0.5375
CreditCard	-1.3424
Income_Group_26-50K	-15.1894
Income_Group_51-75K	19.4302
Income_Group_76-100K	22.2489
Income_Group_101-125K	25.1726
Income_Group_126-150K	27.0920
Income_Group_>151K	27.3051

- Family and Education are positive indicators for Personal Loans. An increase in the the size of the family shows a 0.77x increase and Education has a 1.76x increase
- A Securities Account, Online Account, and Credit Card from another bank are negative indicators for Personal Loans but having a CD Account increase the likelihood by 4.01x
- Income seems to be an important indicator with customers with incomes of 26-50K (-15.19x) less likely to take a out a personal loan but increases as incomes rise 51-75K (19.4x), 76-100K (22.2x), 101-125K (25.2x), 126-150K (27.1x), >151K (27.3x).

RECOMMENDATION

Based on these findings, All Life Bank should target their marketing efforts for Personal Loans to customers with large families, higher education, and incomes above $50,0000 per year. Customers with securities accounts, credit cards from other banks, and who bank online are less likely to take out a personal loan.

05

MODEL BUILDING

DECISION TREE

BUILDING THE DECISION TREE

Approach

1. Data Preparation
2. Partition into training (70%) and test data (30%)
3. Build a CART model on the training data
4. Tune the model and prune the tree, if required
5. Measure performance using
    a. Accuracy, Recall and F1

Objectives

- 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.

Model can make wrong predictions as:

- Predicting a person has a Personal Loan but actually does not.
- Predicting a person doesn't have a Personal Loan but actually they do.

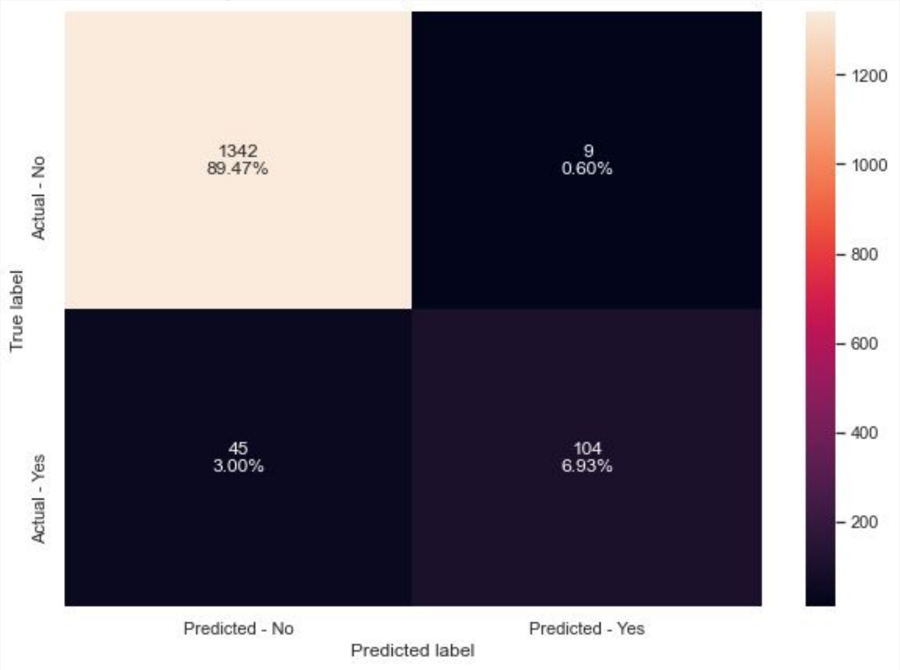
Which case is more important? Both are important.
Because this is only a target market model and not an underwriting model we can accept some false positives. However, too many false positives could lead to an inefficient allocation of marketing spend and man hours.

How to reduce losses?
We can use accuracy since the downside of a False Positive is minimized. We are only attempting to predict whether or not a customer will apply for a loan not whether or not they will default. For this application, we want to identify all customers likely to apply for a loan and can evaluate their application for approval at a later step in the process.

A good balance of both Accuracy and Recall are good metrics for this project.

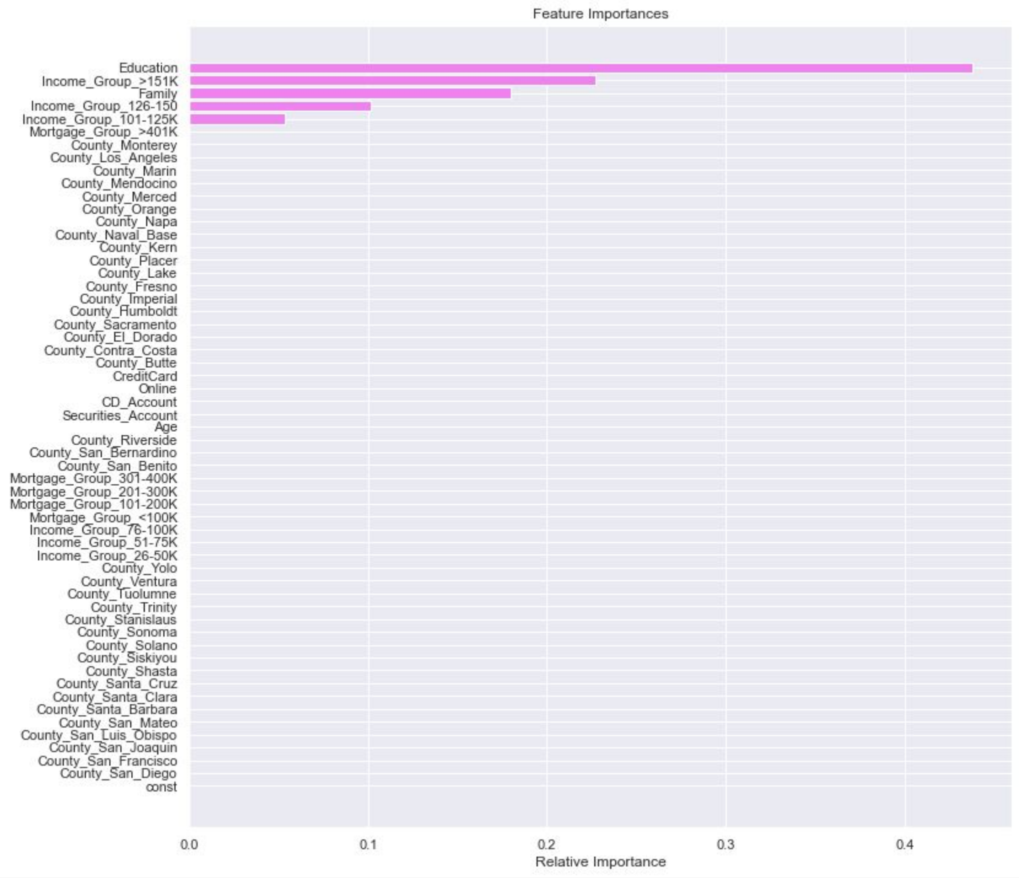
CONFUSION MATRIX

CONFUSION MATRIX



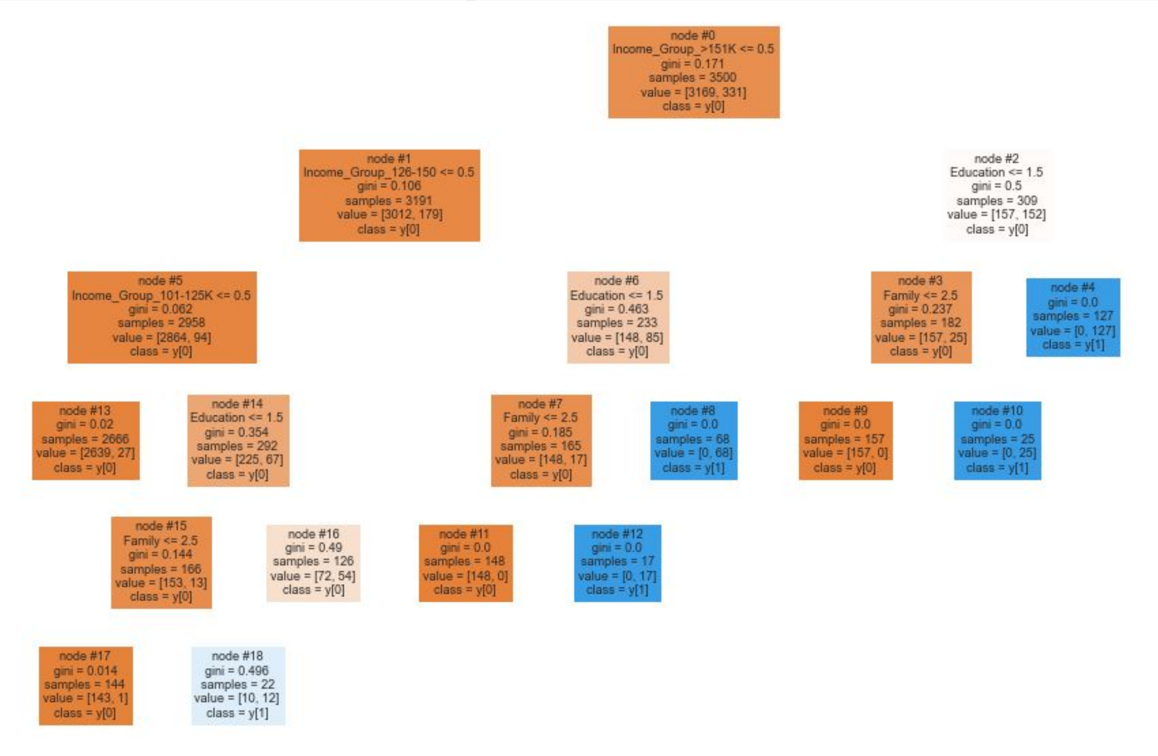
- The model predicted whether or not a customer took out a personal loan or not with 96.4% accuracy which exceeds the base accuracy of the dataset of 90.4%

FEATURE IMPORTANCE



- Education is the most important feature
- Income about $151K is a strong indicator
- Larger families tend to take out more personal loans
- County was not a good indicator
- Income was not a good indicator

VISUALIZING THE DECISION TREE



- Root node was Income_Group_>151K
- Education was the most powerful predictive factor

FINAL MODEL PERFORMANCE
METRICS

DECISION TREE PERFORMANCE SUMMARY

Model	Recall Test	Recall Train	Accuracy Test	Accuracy Train
Initial decision tree mode	0.997	0.832	0.999	0.9651
Decision tree with restricted maximum depth of 5	0.752	0.698	0.974	0.9642
Decision tree with hyperparameter tuning	0.752	0.698	0.974	0.9643
Decision tree with post-pruning	0.997	0.832	0.989	0.9734
Logistic Regression optimal threshold	0.843	0.813	0.961	0.0953

- Initial decision tree showed typical characteristics of overfitting
- Decision Tree with restricted maximum depth of 5 had identical metrics as hyperparameter tuning but the later produced a model with less complexity
- Post pruning using cc-alpha model produced the best results but model was very large and complex
- Logistic Regression model produced comparable accuracy scores but with the added benefit of coefficient ratios which help business managers more accurately target customers

RECOMMENDATION

Based on these findings, All Life Bank should target their marketing efforts for Personal Loans to customers with higher education, larger families and higher incomes above $150,0000 per year

06

KEY INSIGHTS & RECOMMENDATIONS

KEY INSIGHTS

OBJECTIVES

1. **To predict whether a liability customer will buy a personal loan or not.**

   We can fairly conclude that our models are generally good at predicting whether or not a customer is more likely than other customers to take our a personal loan

2. **Which variables are most significant.**

   Annual Income, Education, and Family Size are the most significant positive predictors of whether or not a customer will take out a personal loan. Securities accounts, online banking, or a credit card from another institution were negative indicators.

3. **Which segment of customers should be targeted more.**

   Somewhat counter intuitively, All Life Bank should target customers with higher incomes, more education and bigger families as they are the most likely to take out a personal loan. As importantly, they are also the most likely to pay it back.

BUSINESS RECOMMENDATIONS

1. **Key Insight: Customers with higher incomes are more likely to take out a personal loan**
   Recommendation: Target these customers with marketing around big ticket and luxury purchases such as vacations, jewelry, and other non-essential items.

2. **Key Insight - Customers with bigger families are more likely to take out a personal loan**
   Recommendation: Target these customers with family centric uses of funds such as family vacations, school trips, private education, sporting events, concerts, electronics and other big ticket purchases particularly during the holidays

3. **Key Insight - Higher Educated customers are more likely to take out a personal loan**
   Recommendation: The company should look to focus their advertising at recent graduate of Graduate or PHD level schools with enticing ads for celebrating their achievements or funds for celebration because "you've earned it."

4. **Key Insight - Customers with Securities Accounts, Credit Cards from other banks and who bank online are less likely to take out a personal loan**
   Recommendation: These may indicate more sophisticated customers with more reserved spending habits. Target them with other products such as home mortgages since 69.2% of the banks customers do not have a mortgage.

**BONUS - Target customers in Northern California versus Southern California as a area for growth.