AllLife Bank- Personal Loans

Contents

- Business Problem overview and Solution Approach
- Data Overview
- Exploratory Data Analysis – including key insights and observations
- Model Performance and Summaries
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

- AllLife Bank is based in the United States and has a growing customer base.
- The objective is to analyze information about current customers to determine what characteristics make an existing depository customer more likely to also accept a personal loan from AllLife Bank. This would generate additional revenue for the bank.
- The majority of AllLife Bank's customers only have depository or liability accounts, which do not generate revenue for the bank. Bank management wants to increase the number of current customers with depository accounts, to also having personal loans through AllLife Bank.
- Increasing the amount of personal loans issued to customers would increase revenue. Management wants to target the correct demographic for solicitation to increase the chances of selling personal loans. Targeting the wrong demographic would be a waste of advertisement dollars as well as management and staff's time.
- This model will help management identify the characteristics of existing customers who are most likely to accept a personal loan during a future lending campaign.

Data Overview

The data included variable such as:

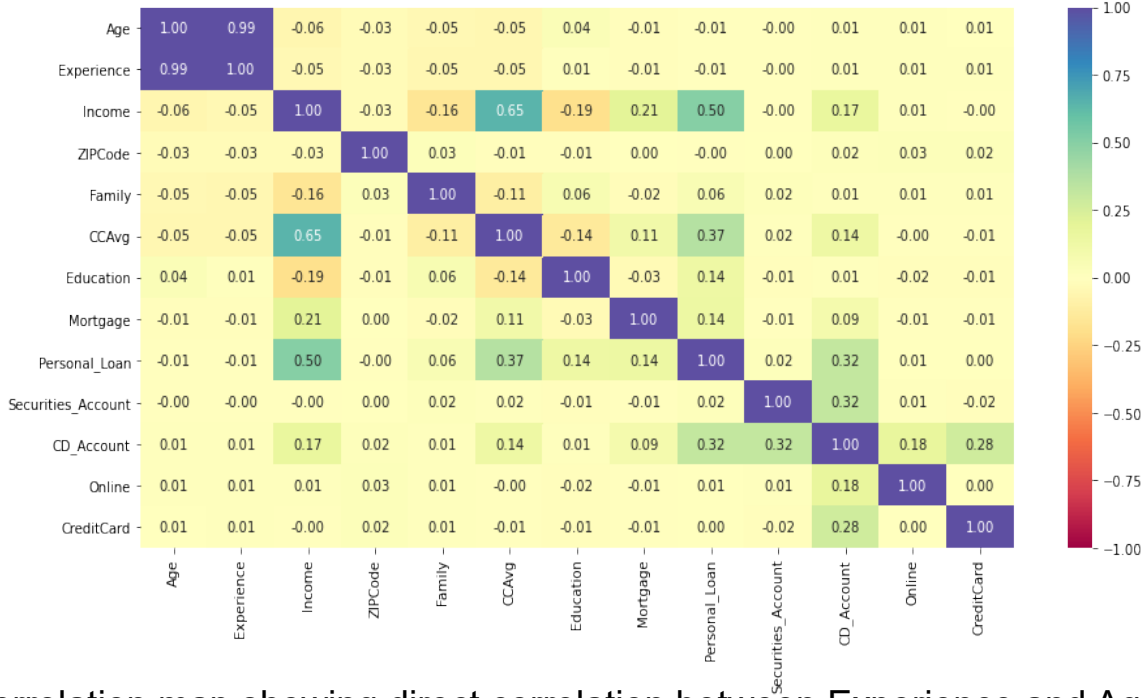
- Customer identification number
- Customer's age in years
- Number of years of professional experience
- Annual income of customer in thousands US dollars
- Home address zip code
- Family size
- Average credit card spending per month in thousands US dollars
- Level of education – Undergraduate, graduate or advanced/professional degrees
- Value of home mortgage (if any) in thousands US dollars
- Whether or not the customers accepted a personal loan on the previous campaign
- If the customer has a securities account with the bank
- If the customer has a certificate of deposit account with the bank
- Whether or not the customer uses the online banking services
- Whather or not the customer uses a credit card issued by a bank other than AllLife

Data Overview

During the data cleaning process the following steps were performed:

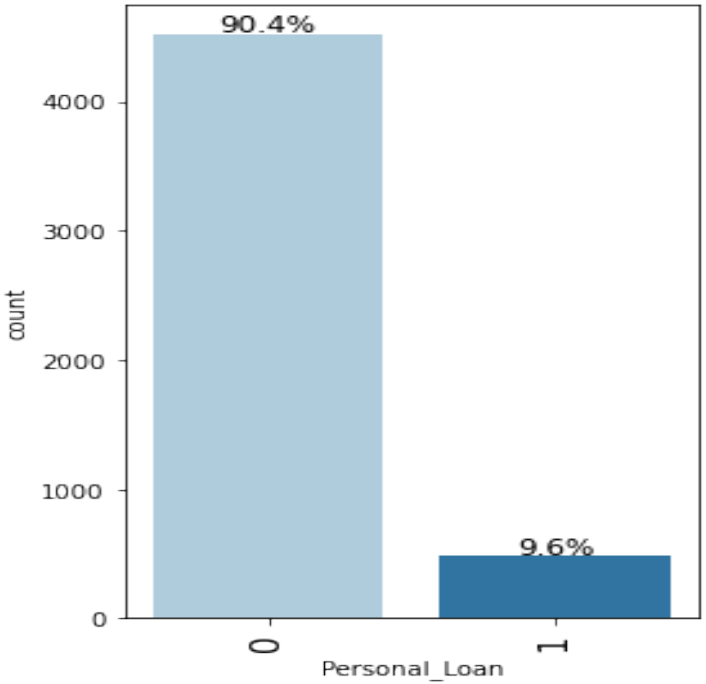
- 'ID' column was immediately drop as it was unnecessary
- 'Experience' column was dropped after generating a correlation heat map and determining that it was highly correlated with 'Age', with the correlation statistics of both those columns and other variables basically identical. 'Experience' was chosen over 'Age' because there were some negative values present in the column that would have required additional cleaning and correction.
- 'ZIPCode' was first converted into geographic major city area. This converted 467 unique characters to 244 unique characters. This was also excessive and the major cities were then bundled into counties, further reducing the unique character count to 38. Missing counties were filled in and then any county with a value count of 20 or less were bundled together into 'County-Other. This reduced the unique character count to a more manageable value of 25.
- The 'ZIPCode' and 'City' columns were dropped next.
- Outliers were left as is, see bivariate analysis – also decision trees are not sensitive to outliers

Exploratory Data Analysis

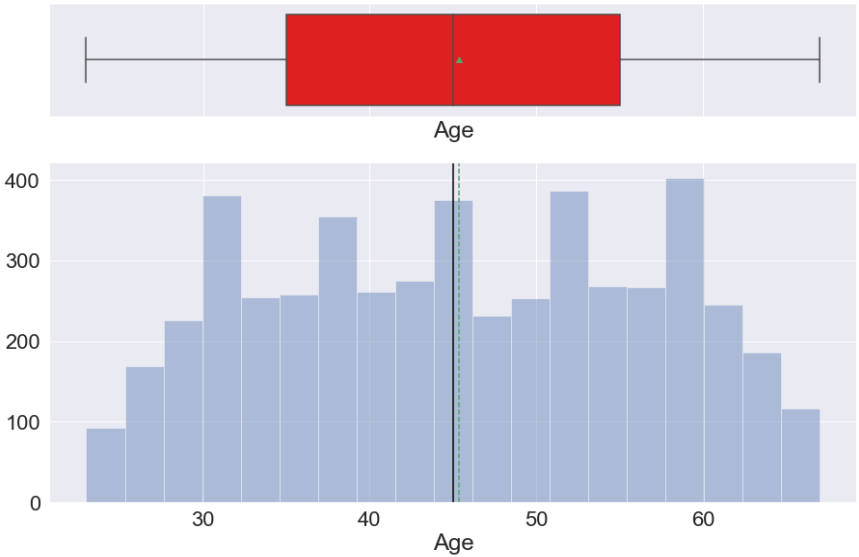


Heat correlation map showing direct correlation between Experience and Age. Notice identical correlation scores with the exception of Income, and those scores are very close.

Exploratory Data Analysis – Univariate Analysis

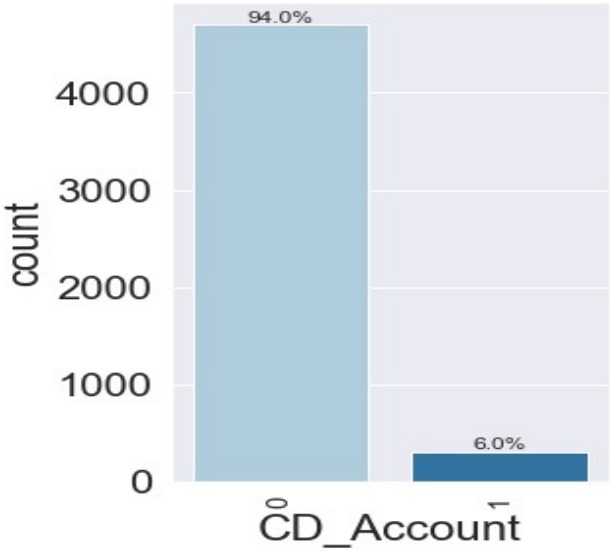


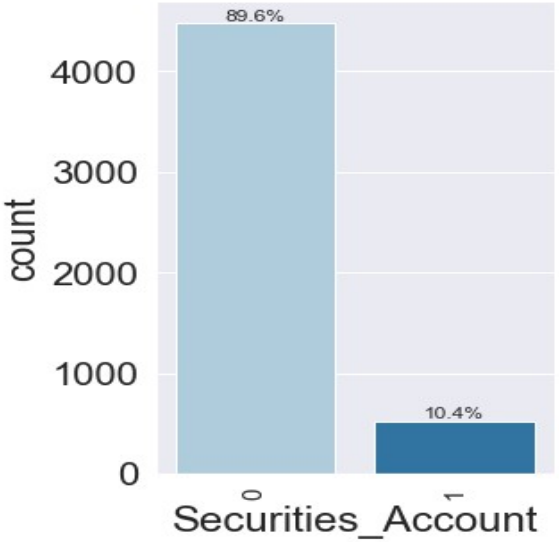
Only 9.6% of customers accepted a personal loan during the previous campaign.

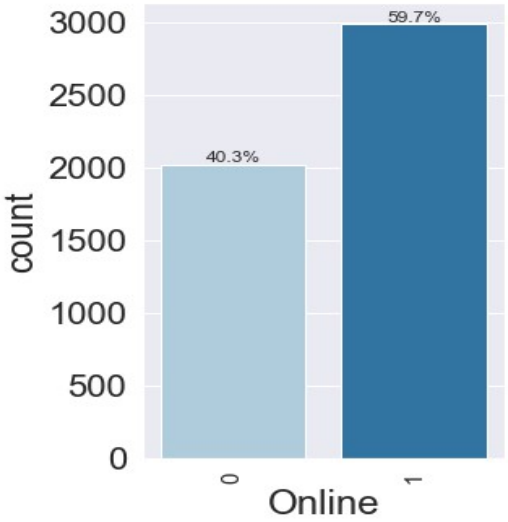


Age is reasonably normally distributed with almost equal mean and median

Exploratory Data Analysis – Univariate Analysis

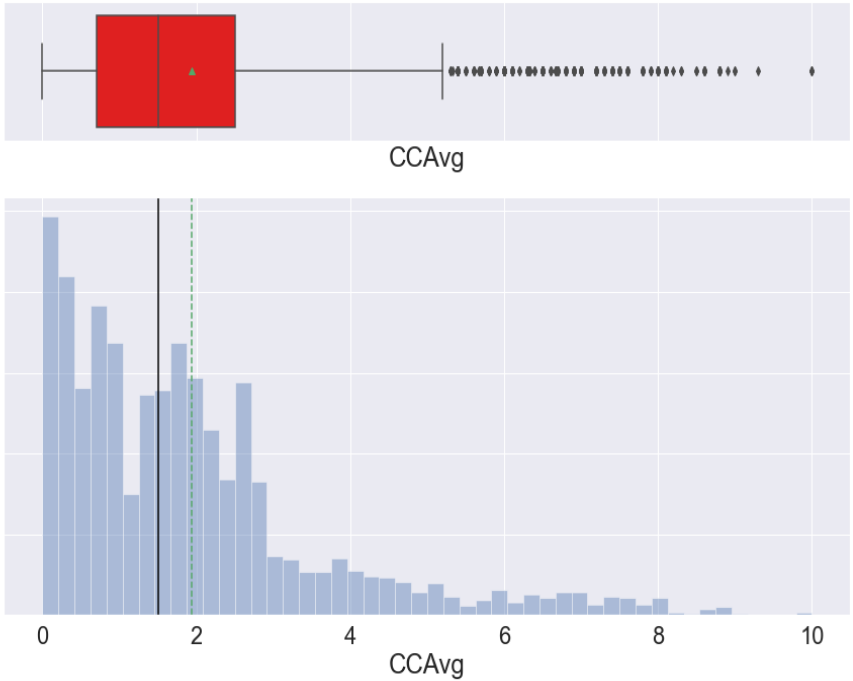


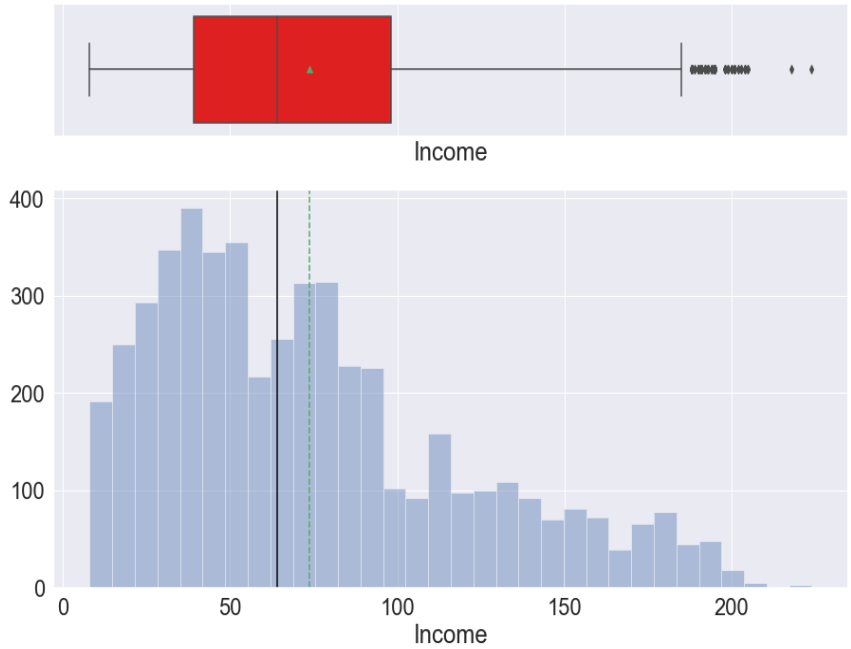




- 6% of customers have certificates of deposit accounts
- 10.4% of customers have a securities account
- 59.7% of customers use online banking services

Exploratory Data Analysis – Univariate Analysis

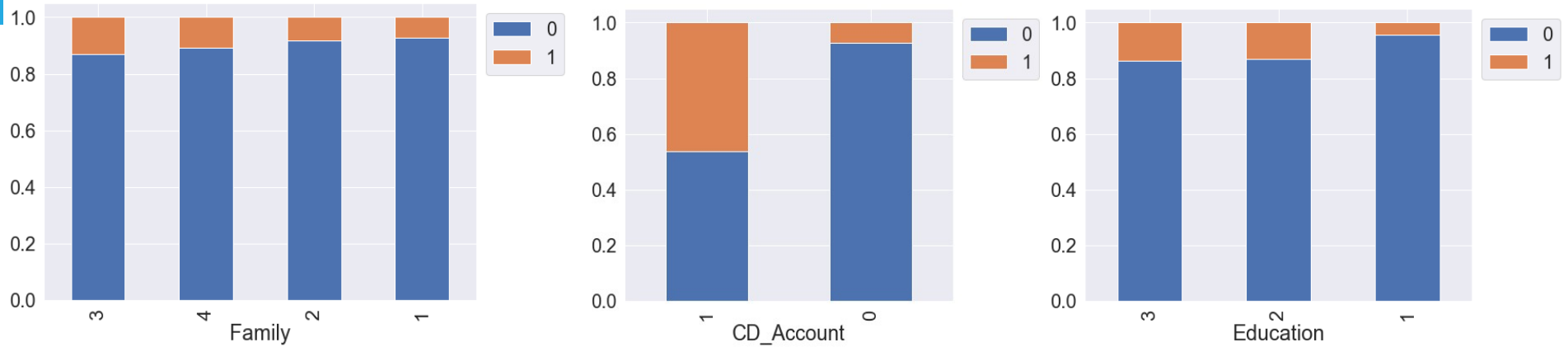




- Monthly credit card usage is right skewed with outliers reaching all the way up to 10k USD. It has a mean of 1.938k USD with a median of 1.500k USD

Income is slightly right skewed with s a few outliers above the 75% quartile. It has a mean of 73.774k USD and a median of 64.000k USD

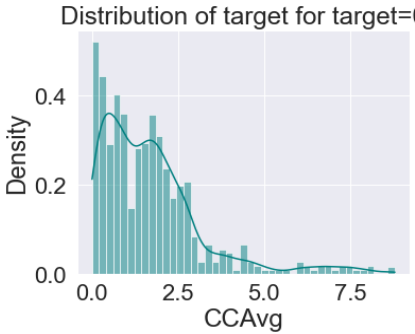
Exploratory Data Analysis – Bivariate Analysis

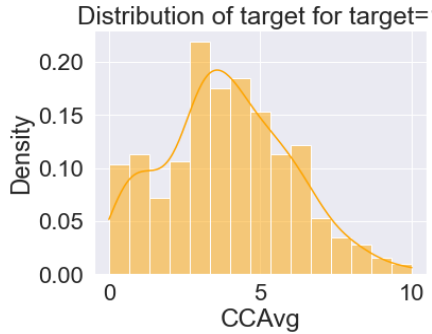


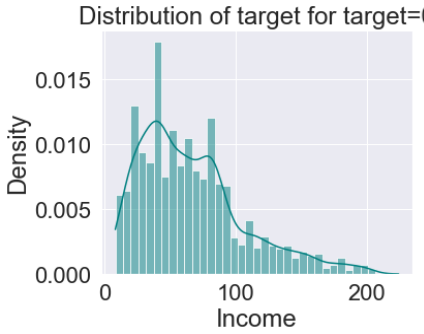
All bivariate analysis graphs are categories compared to Personal loans – 1 represents accepting a personal loan in the previous campaign whereas 0 represents not accepting a personal loan

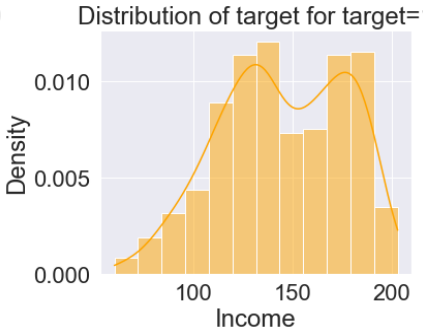
- Customers with 3 family members appear to have accepted more personal loans than those with 4, 2 or 1 member in the family
- Customers with certificates of deposit appeared to accept the a personal loan quite a bit more frequently than those without
- Customers with graduate degrees or advanced/professional degrees appear to accept personal loans more frequently than those with a bachelors degree or less

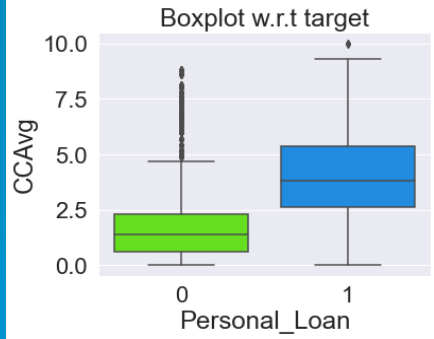
Exploratory Data Analysis – Bivariate Analysis

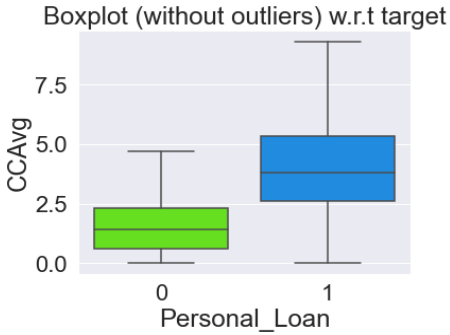


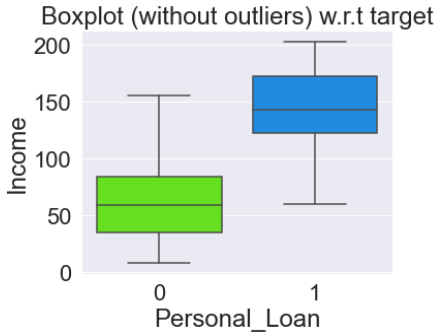












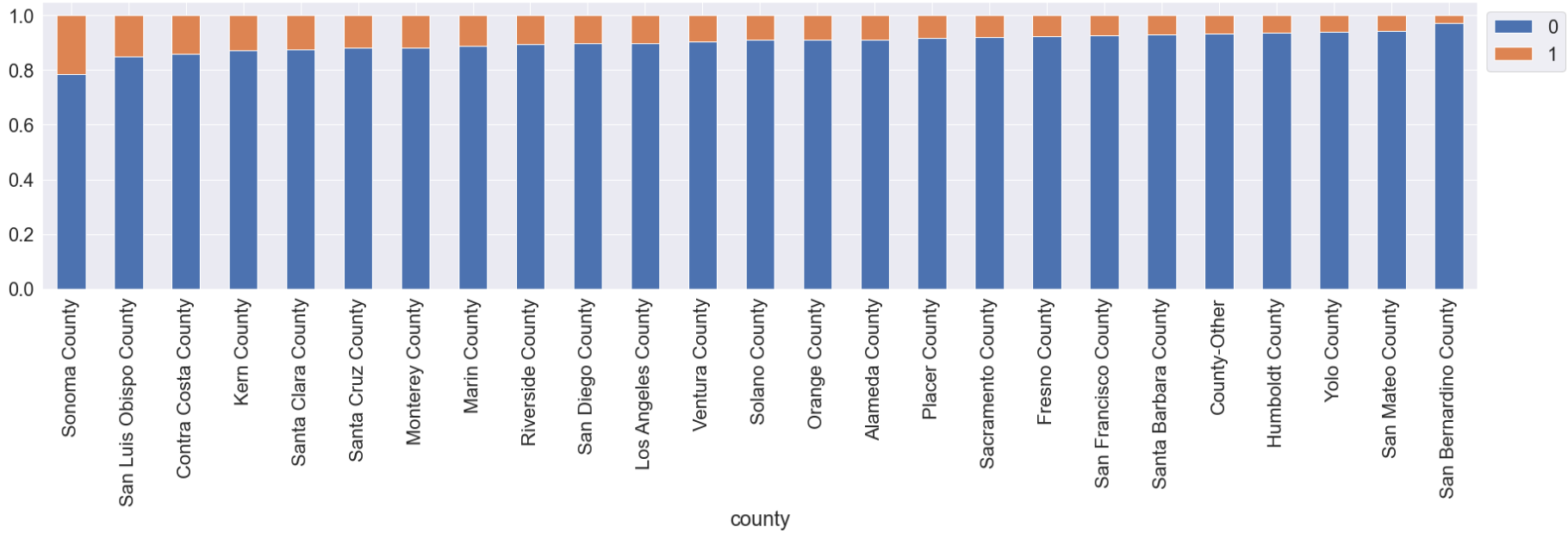
Customers with lower monthly credit card spending tended to not accept a personal loan whereas customers in the middle to upper regions were more likely to accept a personal loan.

Note: boxplot that adjusts for outliers without significant change to outcome, also there were much fewer for those accepting a loan as opposed to those who did not. Therefore outliers were left alone in model analysis

Customers with income less than approx. 100k USD were less likely to accept a personal loan.

Note: boxplot that adjusts for outliers without significant change to outcome, also there were much fewer for those accepting a loan as opposed to those who did not. Therefore outliers were left alone in model analysis

Exploratory Data Analysis – Bivariate Analysis



While Sonoma and San Luis Obispo counties only account for 28 and 33 customers out of the entire data set of 5000, those 2 counties have the highest percentages of acceptance of personal loans per capita in the previous campaign. Perhaps bank management should consider assessing customer acceptance criteria on a county by county basis.

Best Fitting Model(s) Performance Summary

Model Overview

- Model names: 2 models with identical testing performance metrics
  estimator
  bestmodel
- Target variable: Personal Loans
- Model: e*stimator:* Decision Tree Hyperparameter
  b*estmodel:* Decision Tree CCP-Alpha
- Method:e*stimator*: CVGridSearch- evaluates selection of pre-defined parameters and selects best fit based on array of aprameters
  bestmodel: np.argmax – generates maximum values along the axis and fits a model
- Number of observations: training data – 3500, testing data 1500
- Tree Depth: e*stimator*: 5 levels, 16 nodes
  bestmodel: 6 levels, 26 nodes
- Important Features: Education, Income, Family, CD Accounts for both (*bestmodel* also has Age)

Model Performance Summary – Performance Metrics

estimator - Training Data Set

Accuracy	Recall	Precision	F1
0.989714	0.927492	0.962382	0.944615

estimator - Testing Data Set

Accuracy	Recall	Precision	F1
0.981333	0.879195	0.929078	0.903448

bestmodel - Training Data Set

Accuracy	Recall	Precision	F1
0.992000	0.945619	0.969040	0.957187

bestmodel - Testing Data Set

Accuracy	Recall	Precision	F1
0.981333	0.879195	0.929078	.0903448

While the models have different, yet very similar training data metrics, their testing metrics were identical. Accuracy is very high on both models, but in the case, accuracy is not as meaningful as F1. F1 carries more weight because, as shown in the Univariate analysis, the class distribution of the variable, Personal Loans, is very uneven with only 9.6% accepting loans and 90.4% not accepting. If that distribution were more equal, Accuracy would hold more significance.

Accuracy refers to how frequently the model predicted correctly - did it correctly predict which customer would or would not accept a personal loan?

Recall (also called Sensitivity) also looks at the positive outcomes, but ALL customers who accepted a loan, whether predicted to or not

Precision looks only at predicted positive outcomes - regardless if correctly predicted to be positive or not - it is calculated by taking the total number of customers where were predicted to and did accept a loan and divide it by the total number of customers predicted to accept the loan whether accurate or not

F1 is the weighted average of Precision and Recall, it is another measure of accuracy and is more accurate when the data set has an uneven class distribution. In this data set, the distribution is uneven because there are many more customers who did not accept the loan as opposed to those that did. Therefore the F1 score is more meaningful than the Accuracy score.

Business Insights and Recommendations

- For AllLife Bank to run the most effective personal loan campaign, management needs to focus on customers with the following customer demographics:
  - an income of 116.5k USD or less
  - average spending of 2.95k USD per month on credit cards
  - an education of at least a Bachelors degree, plus addition studies
  - Persons with certificates of deposit accounts
- One area to consider addressing for future analysis is generating separate models based on geographic locations, such as counties. As noted on the bivariate analysis, Sonoma and San Luis Obispo counties accounted for only 33 and 28 of the total customer representation. However, those 2 counties had the highest percentage of acceptance per capita, than all the other counties.
- Another area to consider for analysis and subsequent campaign is the number of customers with AllLife Bank credit cards. The data provided asked if customers used credit cards from other financial institutions, but did not detail which ones also already had cards issued by AllLife. It would be interesting to assess whether customers already having that type of revolving credit would consider a personal loan.

greatlearning
Power Ahead

Happy Learning !