greatlearning
Power Ahead

Personal Loan Campaign Modelling

By Jacob Siegel

Contents

- Introduction and Business Context
- Data Overview & Single Variable Analysis
- EDA & Multi Variable Analysis
- Logistics Regression Model
- Decision Tree Model
- Conclusion and Recommendations

Business Problem Overview and Solution Approach

- **Context:** AllLife Bank is interested in expanding its borrower base (personal loans) to bring in more loan business and revenue.
- **Goal:** Understand if a customer will buy a personal loan based on the data set provided.
- **Implications:** Use the insights from the analysis to create a competitive advantage and market to customers that are more likely to buy a personal loan.

Data Overview

- The raw data set contains 5000 rows and 14 column.
- There were no missing data or duplicate rows.

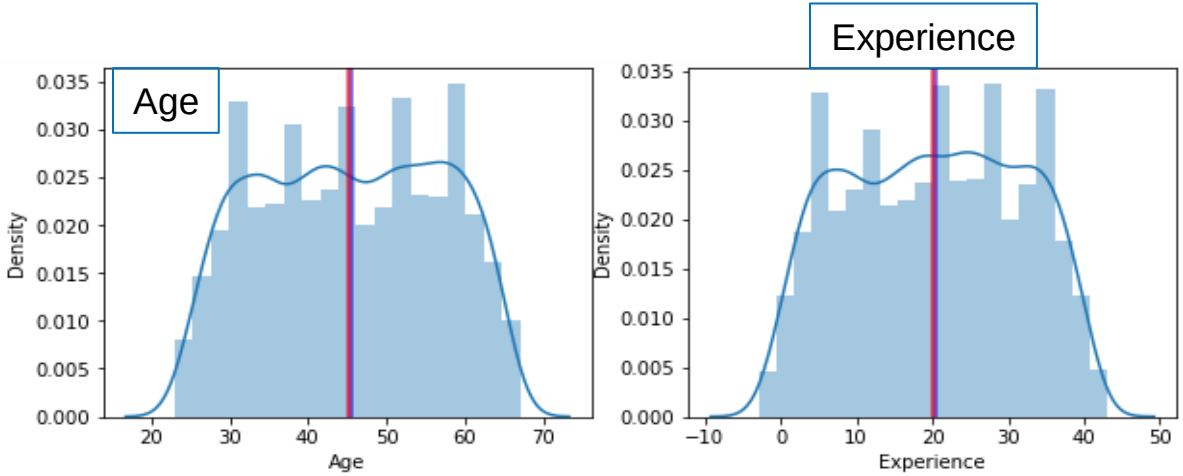
Personal Loan Breakdown:
480 people took a personal loan
4520 did not take a personal loan

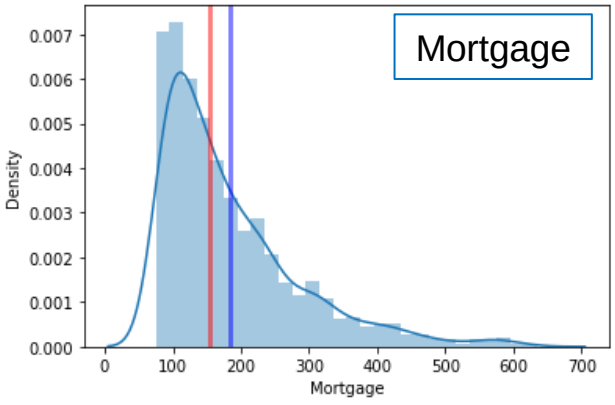
Column Description

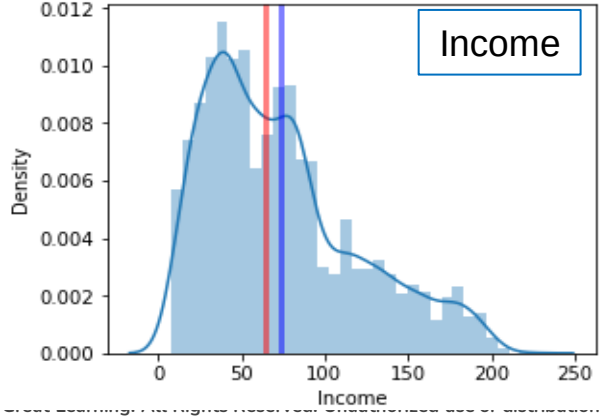
Age: Customer's age in completed years
Experience: #years of professional experience
Income: Annual income of the customer (in thousand dollars)
ZIP Code: Home Address ZIP code.
Family: the Family size of the customer
CCAvg: Average spending on credit cards per month (in thousand dollars)
Education: 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)?

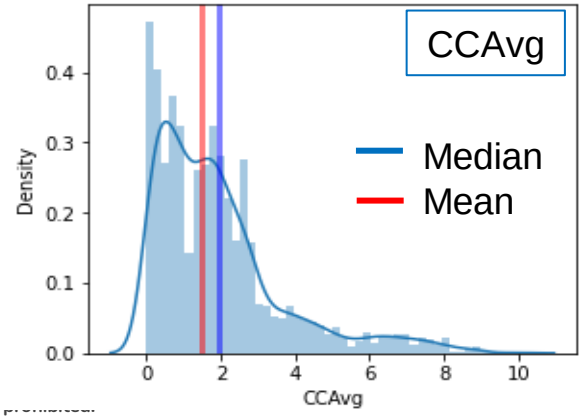
Data Overview: Numerical Variables

- Age and experience have a similar normal distribution.
- The majority of people (3462) do not have mortgages. This data will be converted to a categorical variable to represent having a Mortgage (yes or no).







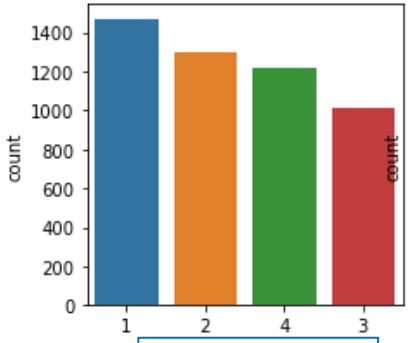


Data Overview: Categorical Variables

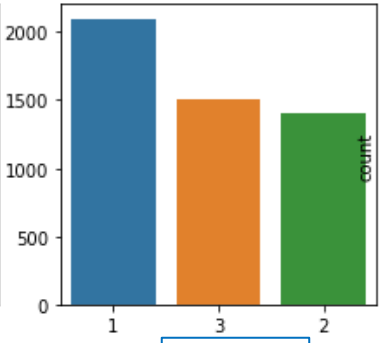
greatlearning
Power Ahead

- The majority of the people in the data set are from single family and have an education of “1”
- The majority do not have: a personal loan, a security account, a CD account, or a Credit Card.

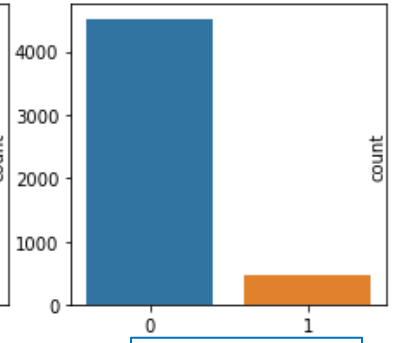
Family



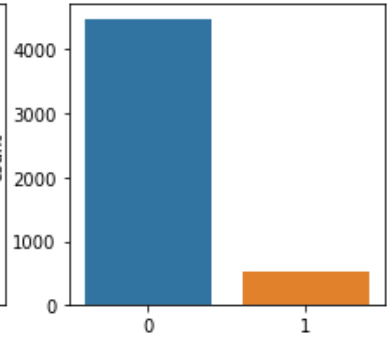
Education



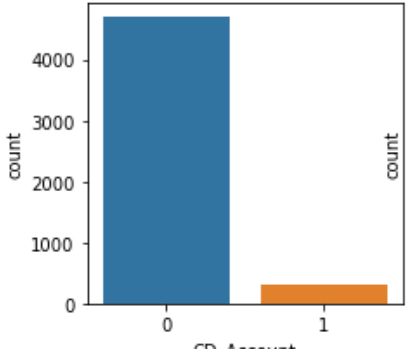
Loan



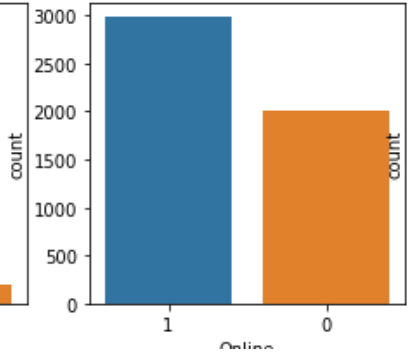
Securities



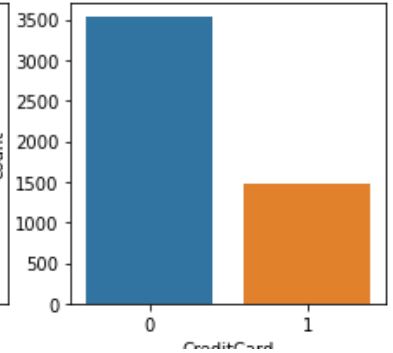
CC Account



Online



Credit Card



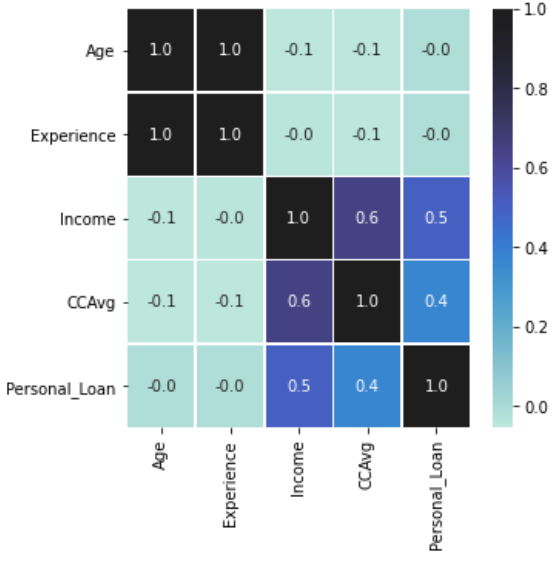
ZIPCode	
94720	169
94305	127
95616	116
90095	71
93106	57

There are 467 unique zip coded in the data. This table shows the most common zip coded

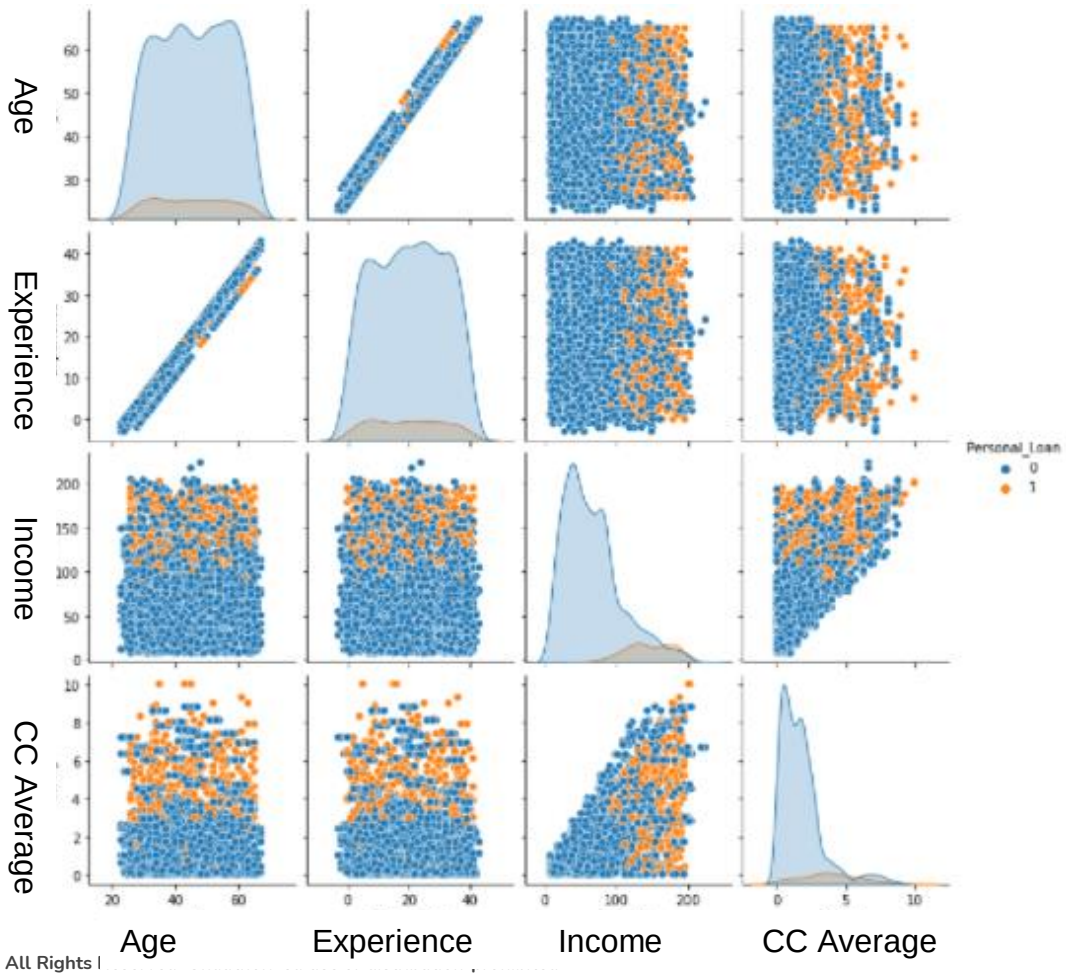
EDA: Continuous variables

- Experience and age are exactly linearly related. Most other data is not strongly correlated.
- The strong linear relationship may be due to previous data manipulation.

Correlation Values



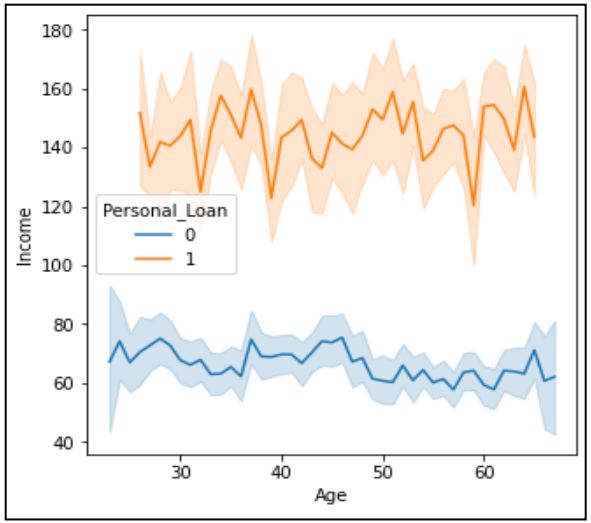
Cross Plots



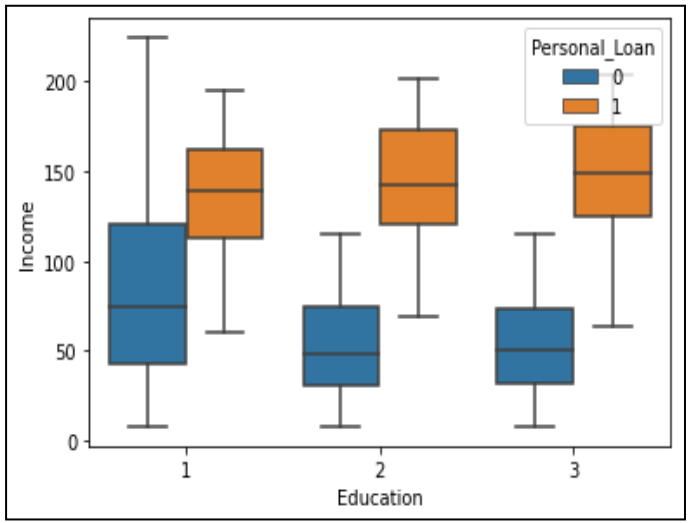
EDA: Closer Examination of Income and Personal Loan

- Income is not correlated with age, however, people who took a personal loan had higher incomes on average. This can also be seen with Education, in each group people who took a personal loan had high higher income on average.
- Credit card average is slightly correlated with income, however, there appears to be an arbitrary cut off of minimum income per credit card that may be artificial.

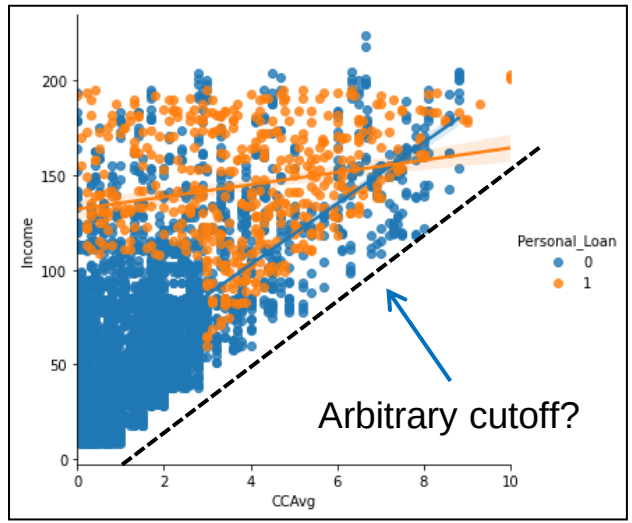
Income vs. Age



Income vs. Education



Income vs. CC Average



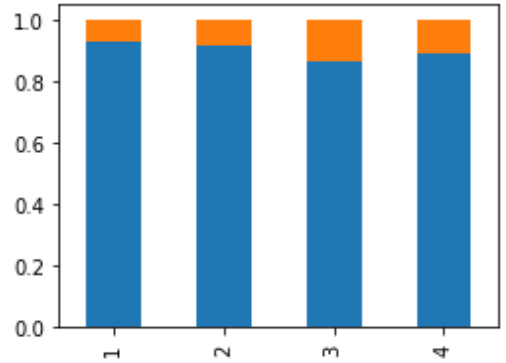
EDA: Personal Loan in Categorical Variables

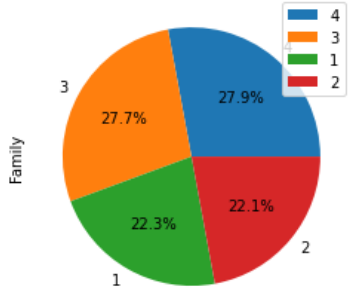
- Families of 3, Education levels of 2 and 3, and people with a CD Account have higher proportions of Personal Loans.
- Securities Account, Online, Credit Card, and Mortgage do not have significant differences in people with a personal loan and are not shown here.

Ratio of No personal Loan to Personal Loan by category

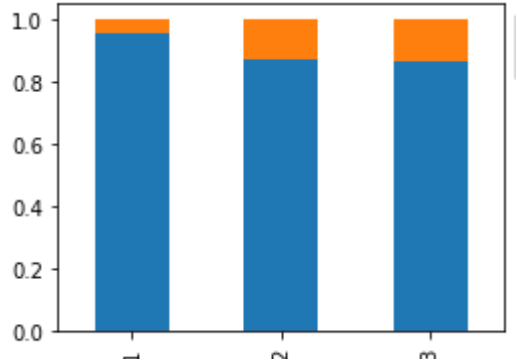
Breakdown of Personal Loans by Category

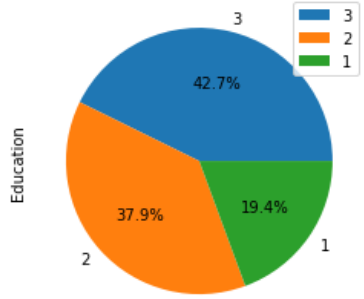
Family



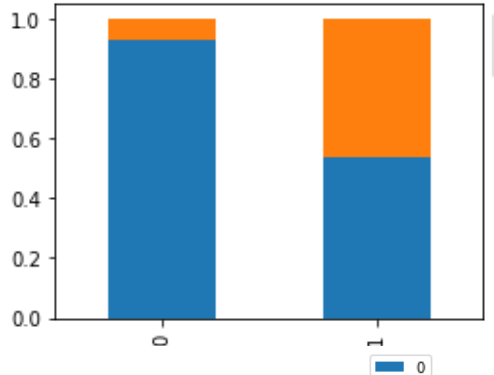


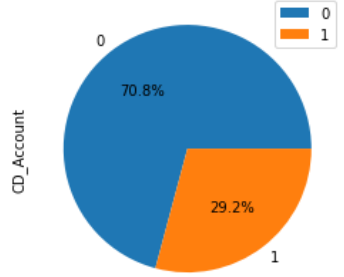
Education





CD_Account



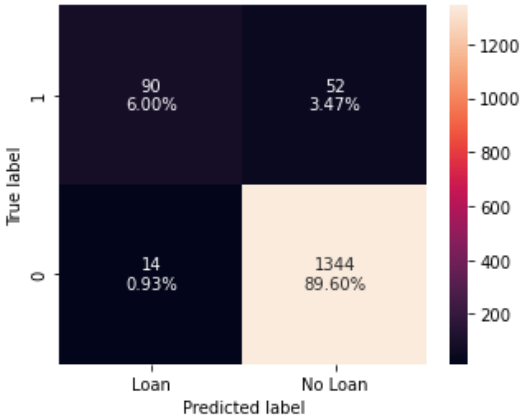


Logistic Regression Models

- Logistics regression models were run with both sklearn and statsmodels. Training and Test data were generated with a 70/30 split.
- The sklearn initial model had an accuracy of 96% on the test data. After apply the optimal threshold determined by the ROC curve the new model has an accuracy of 92%, however there are fewer false positives. This model may be preferred to reduce marketing costs.

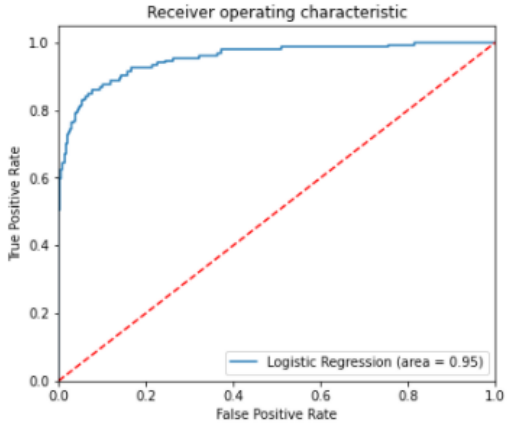
First Test

Accuracy %96
False Positives 52 (%3.47)



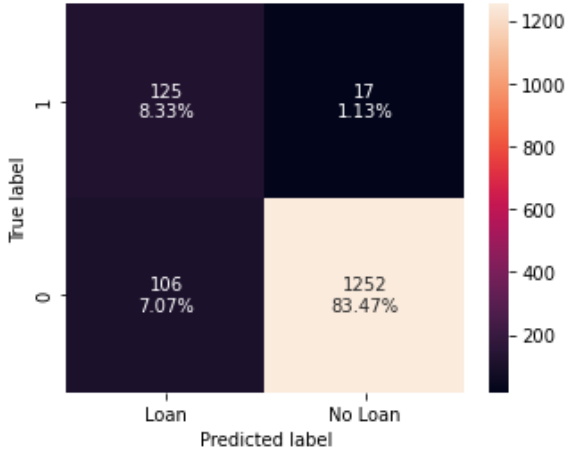
ROC Curve

Optimal threshold: 0.13



Test Optimized

Accuracy %92
False Positives 17 (%1.13)



Logistics Regression Models: Variable Coefficients

- Having a CD account or Education level 2 or 3 are the highest coefficients. However, these are categorical variables.

- Income also has a strong influence, though smaller coefficient, because it is a continuous variable with a wide range of values.

Table of Variable Coefficients from Statsmodel

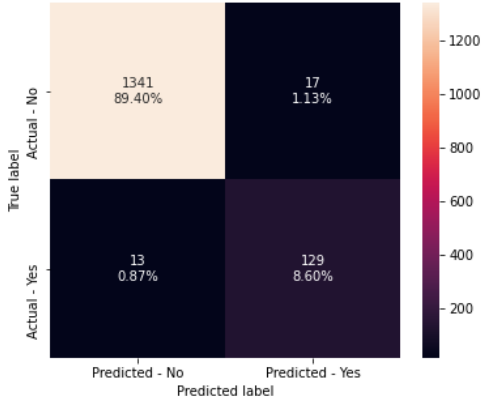
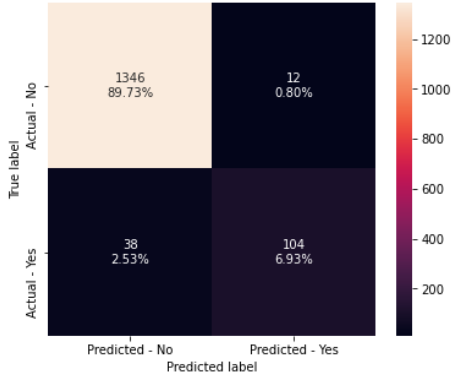
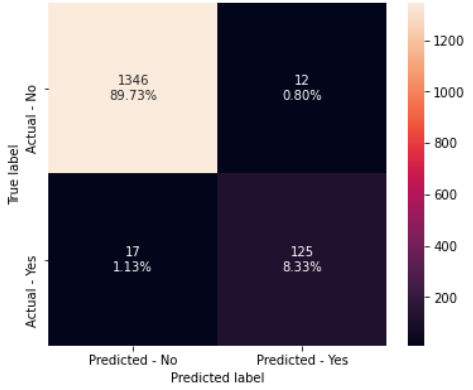
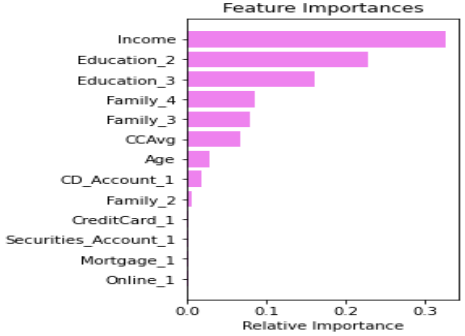
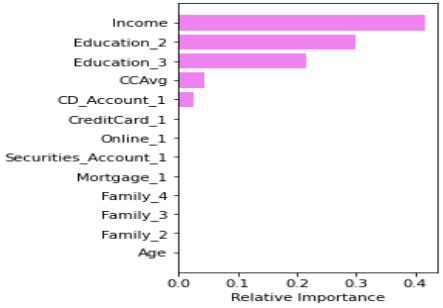
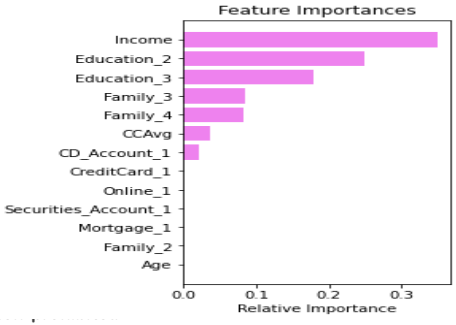
	coef	Odds_ratio	probability	pval
CD_Account_1	4.69	109.35	0.99	0.00
Education_3	1.45	4.26	0.81	0.00
Education_2	1.45	4.24	0.81	0.00
Income	0.03	1.03	0.51	0.00
Age	-0.11	0.89	0.47	0.00
Online_1	-1.24	0.29	0.22	0.00
Family_2	-1.36	0.26	0.20	0.00
Securities_Account_1	-1.55	0.21	0.17	0.00
CreditCard_1	-1.88	0.15	0.13	0.00

Decision Tree Models: Overview

- The following slide outlines the result of three decision tree models
  - **Model 1:** an unrestricted decision tree.
  - **Model 2:** a depth limited decision tree (depth = 3).
  - **Model 3:** a grid search decision tree.
- All three models offer high accuracy and recall. However, model 1 or 3 my be proffered as they are fewer false negatives.

Decision Tree Models: Results

greatlearning
Power Ahead

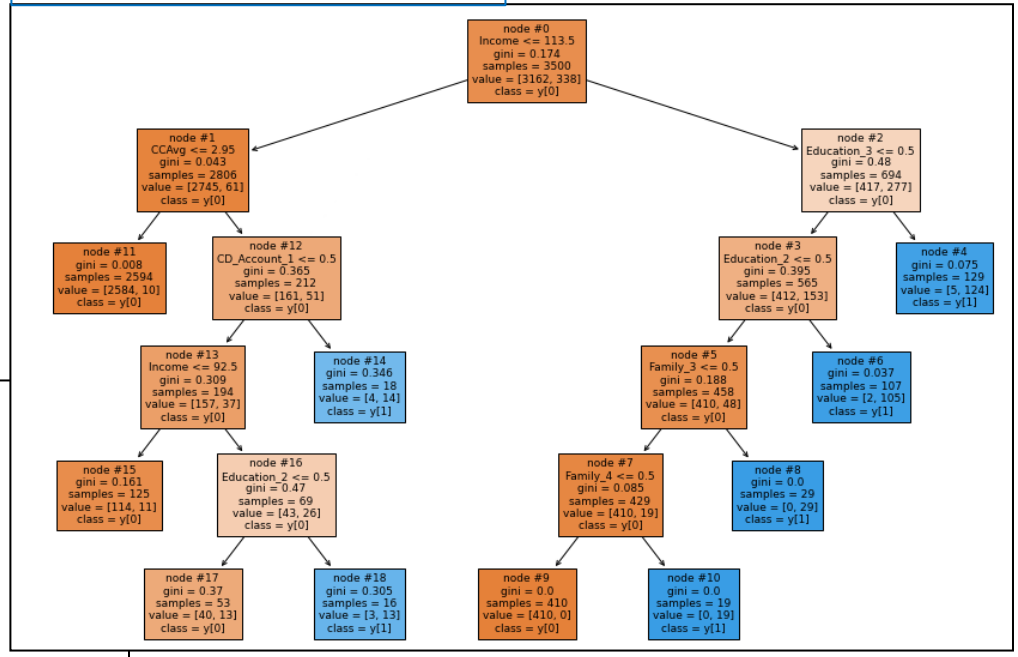
	Unrefined Model	Depth Restricted Model	Grid Search Model
	Test Accuracy: 98 Test Recall: 91	Test Accuracy: 97 Test Recall: 73	Test Accuracy: 99 Test Recall: 88
Confusion Matrix			
Feature Importance			

Conclusion and Recommendations

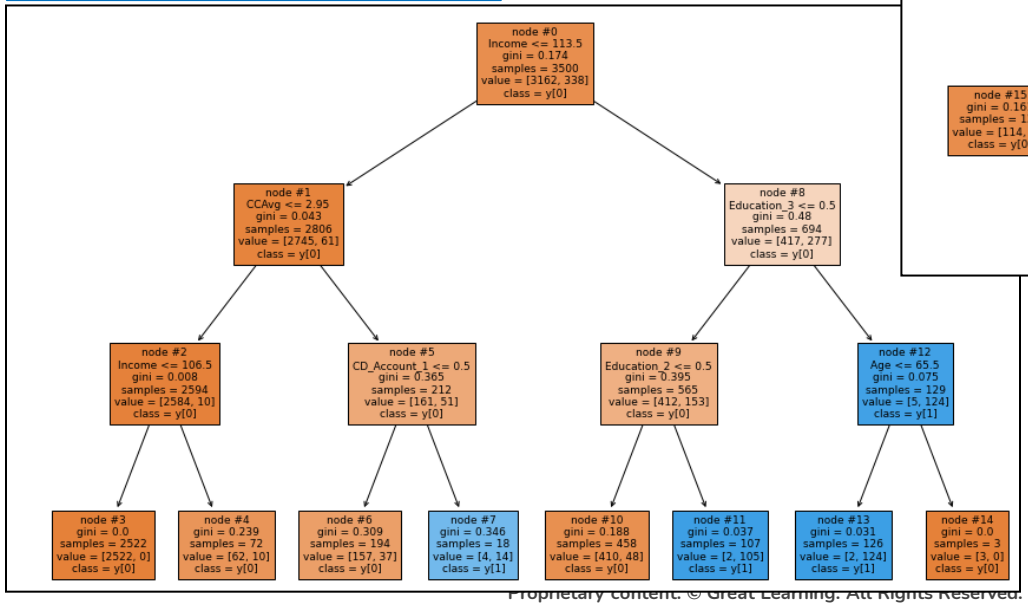
- **Data Pre Processing:**
  - Zip Code was removed from the data as there were more than 400 unique zip codes.
  - Experience was not included in the models as it was 100% linearly correlated with age.
  - Mortgage was converted to a categorical variable (yes/no) for the models.
- **EDA**
  - Income seems to be one of the strongest differentiators of people who took a loan.
  - Families of 3, Education levels of 2 and 3, and people with a CD Account have higher proportions of Personal Loans.
- **Model Results**
  - Both the logistics regression and decision tree models provide a strong fit to the data.
  - *Strong indicators for a loan:* Income, Education level, CC average, CD Account, age
  - *Intermediate indicators for a loan:* Family size, credit card account
  - *Weak indicators for a loan*: online, securities account, having a mortgage.
- **Recommendations**
  - The proffered model is the decision tree grid search as it has the fewest false negatives, which means the most amount of people will be successfully marketed to apply for the loan.
  - The bank should also focus on bring in more customers with higher income and higher education as they are more likely to get a loan.

Apendix 1: Decision Trees

Grid Search Model



Depth Restricted Model



Appendix 2: Age vs Experience

- There is no variability in the data and the cross correlation is nearly 1 which suggest the data may have been modified before we received it.
- Because of this relationship Experience was not included in the models.