AllLife Bank
Personal Loan Campaign
PGP AIML - Ashley Campbell

Due 1/5/24

Contents / Agenda

Great Learning
POWER AHEAD

Executive Summary

- The model built can be used to predict if a customer will take out a loan and can correctly predict this 91.3% of the time.
- Income, family size and undergraduate education (in that order) are the most important variables in determining if a customer will take out a personal loan
- All members with an **income > $98,500 per year** should receive advertising.
- Members with an income **< $98,500 per year** will not be advertised to unless they have:
  - either **3 or more credit cards** AND a **CD account**
  - OR **4 or more credit cards** and an **income > $81,000** OR age **< 36 years**
- Consider looking further into customer occupations to determine likelihood of accepting loan with a specific type of occupation (i.e. entrepreneur).
  - Collect data on customer occupations during interactions through surveys, application forms, or during account setup.
  - Analyze data to determine if certain occupations are associated with increased likelihood of taking out a loan.

Business Problem Overview and Solution Approach

- **Objective**
  - **To predict whether a liability customer will buy personal loans,** to understand which customer attributes are most significant in driving purchases, and identify which segment of customers to target more.

- **Methodology**
  - Extract insights using Exploratory Data Analysis.
  - Determine which customer attributes are most associated with purchasing a personal loan
  - Create a model to predict which customers would be best to target for a personal loan offer
  - Test and revise model to minimize the risk of missing customers that would potentially accept the loan offer by **focusing on recall**

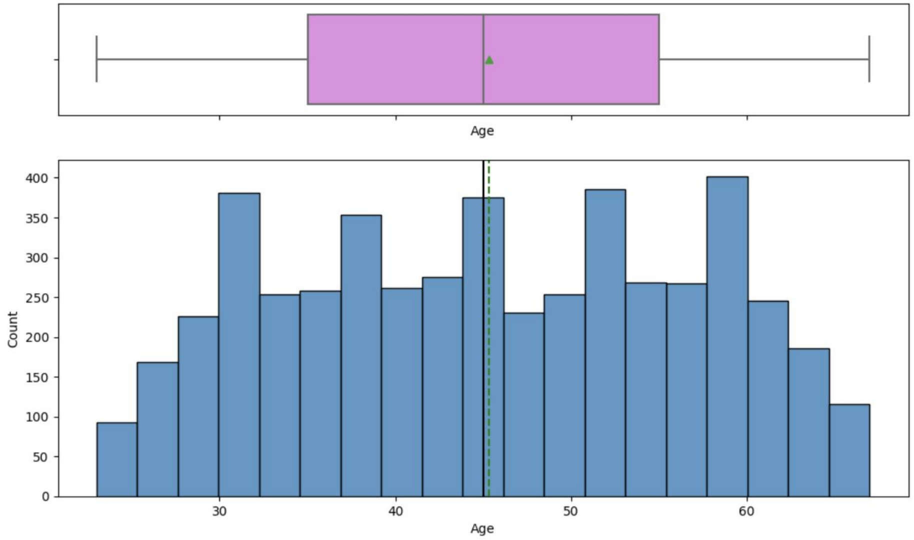
EDA Results

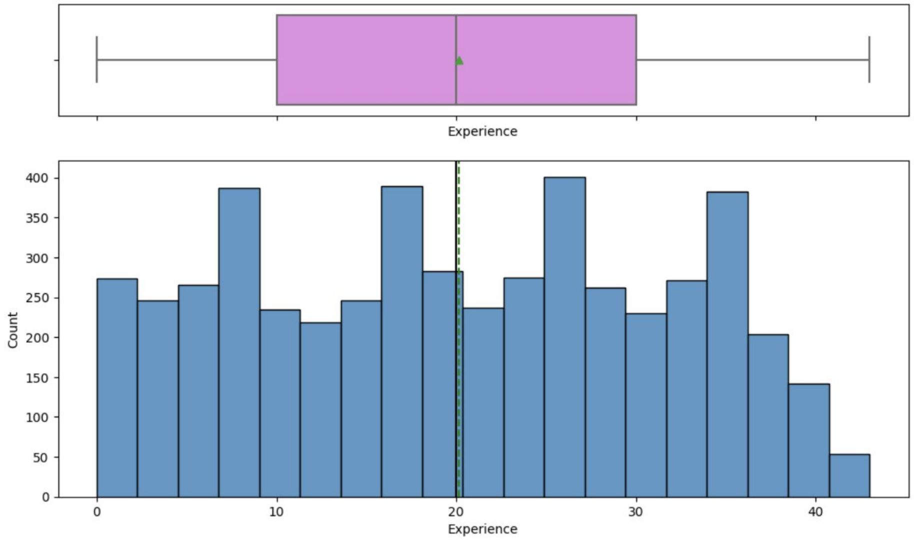
- There are **5000 rows** and **14 columns**
- The data contains several categorical features listed as integers, including:
  - whether or not members have a Securities Account or a CD Account,
  - the member's zip code
  - whether or not members have a credit card with a bank other than AllLife
  - whether they use online banking
- The **mean member age is 45**, with a min of 23 and max of 67
- **Mean years of experience is 20**, with a min of 0 and a max of 43
- **Mean yearly income is $74,000**, with a min of $8,000 & max of $224,000
- **Mean mortgage value is about $56,500**, with a min of $0 & max of $635,000
- **Avg monthly credit card spend is < $2000,** with a min of $0 & max of $10,000

EDA Results

- **Age** and **experience** provide very **similar data.**
- **Income, monthly credit card spend** and **mortgage** are right skewed, however, these outliers may provide good information for **target clients.**
- Most education data points are for members with an undergraduate education.
- The ratio of undergraduate education vs personal loan is lower than the ratio of either graduate or professional education to personal loan, **outliers** in the **undergraduate category** appear to be associated with the likelihood of taking a personal loan.
- Families with **more than 2 children** are more likely to take a personal loan.
- There are **slight differences in zip code** areas and likelihood of taking a loan.

Exploratory Data Analysis

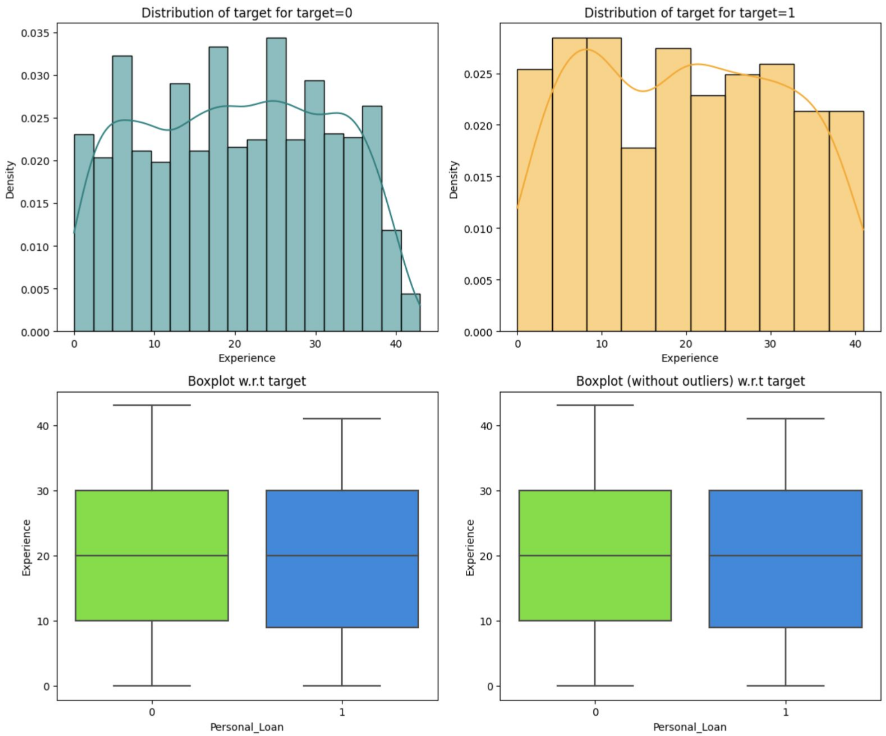


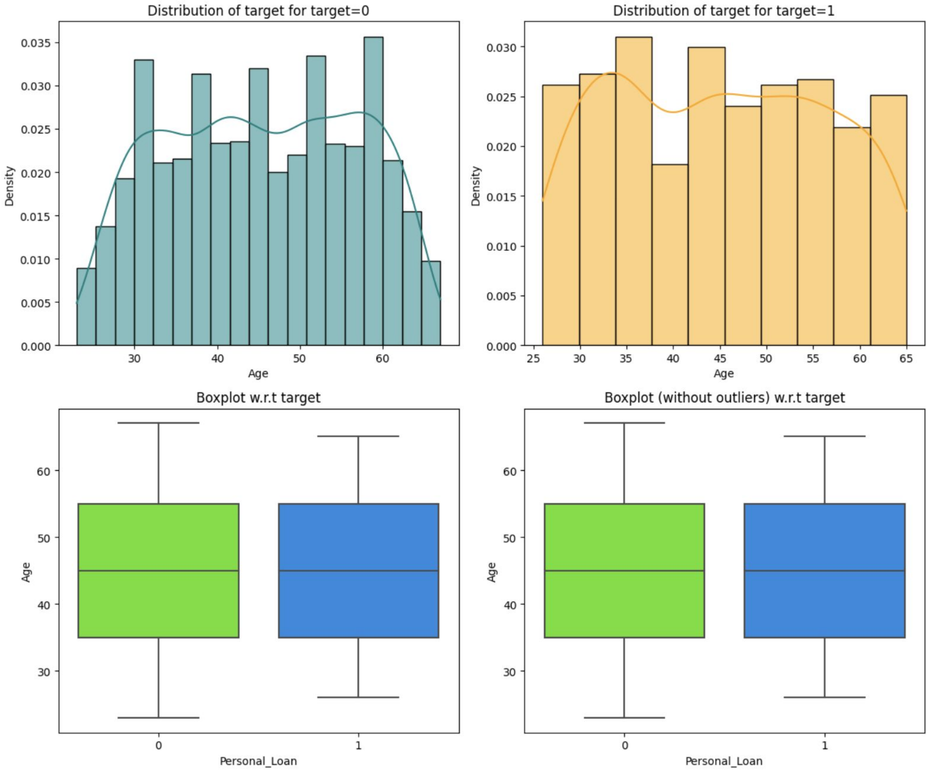


- Age and Experience are both uniformly distributed
- There are no outliers
- Average Age is about 45 years
- Average Experience is just under 20 years

Exploratory Data Analysis - Age and Experience

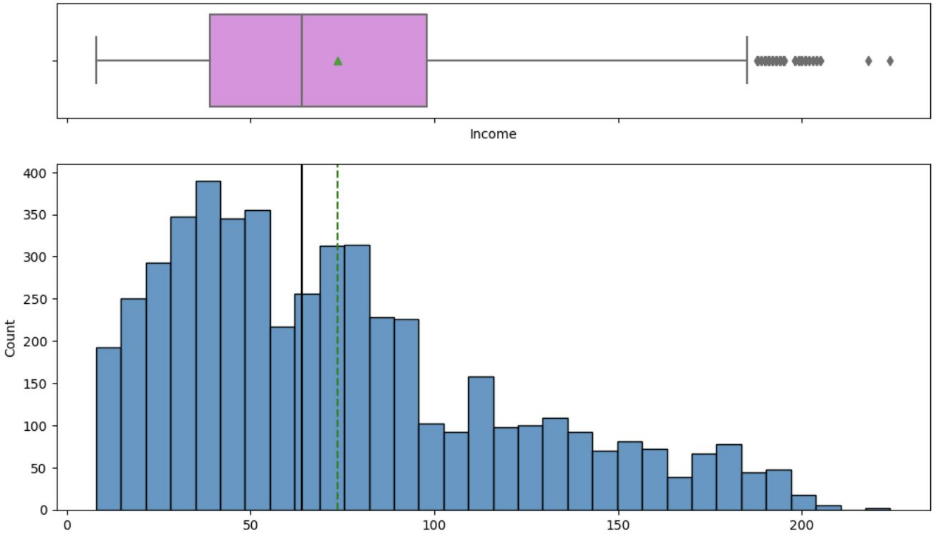
- There does not appear to be a difference between likelihood of taking out versus not taking out a loan for either age or experience features

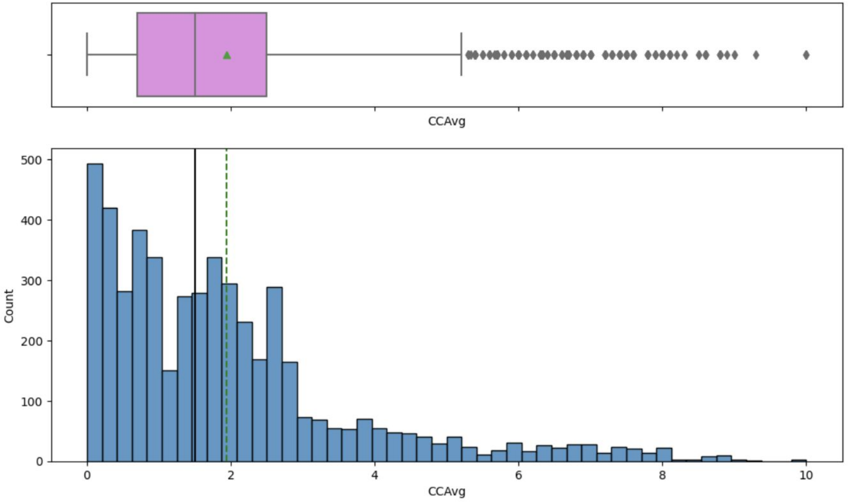


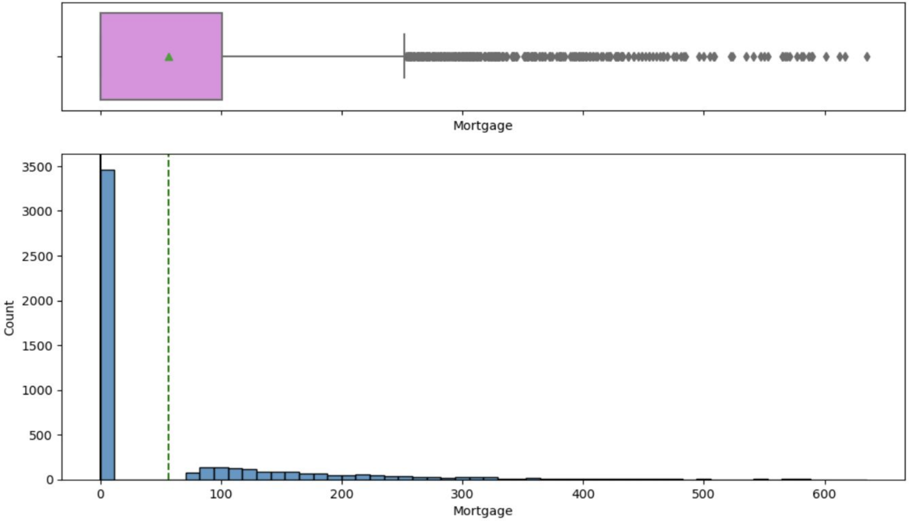


Exploratory Data Analysis

- Income, Mortgage and CCAvg are all right skewed, however, these are target features that may lead to loans

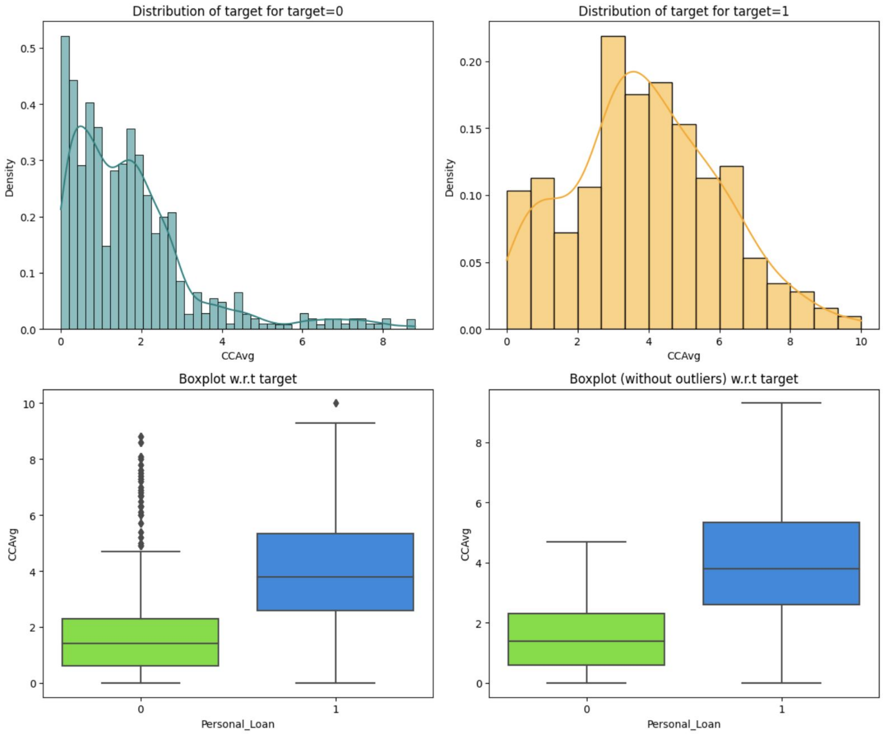


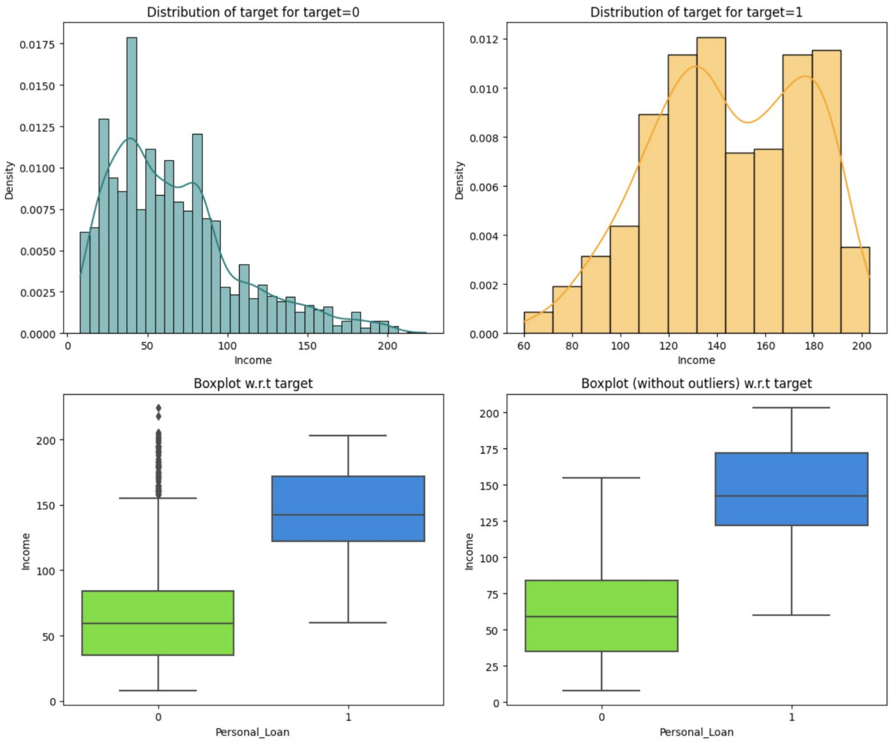




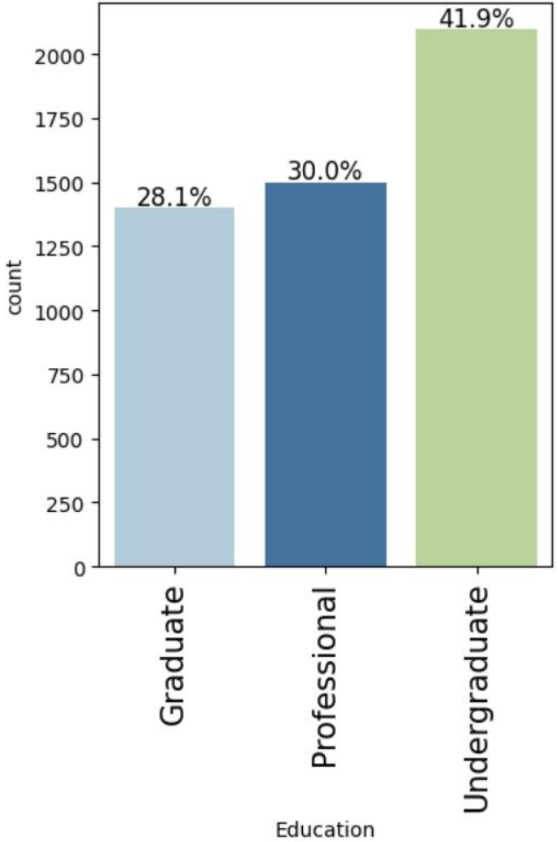
EDA - Avg Monthly Credit Card Spending & Income

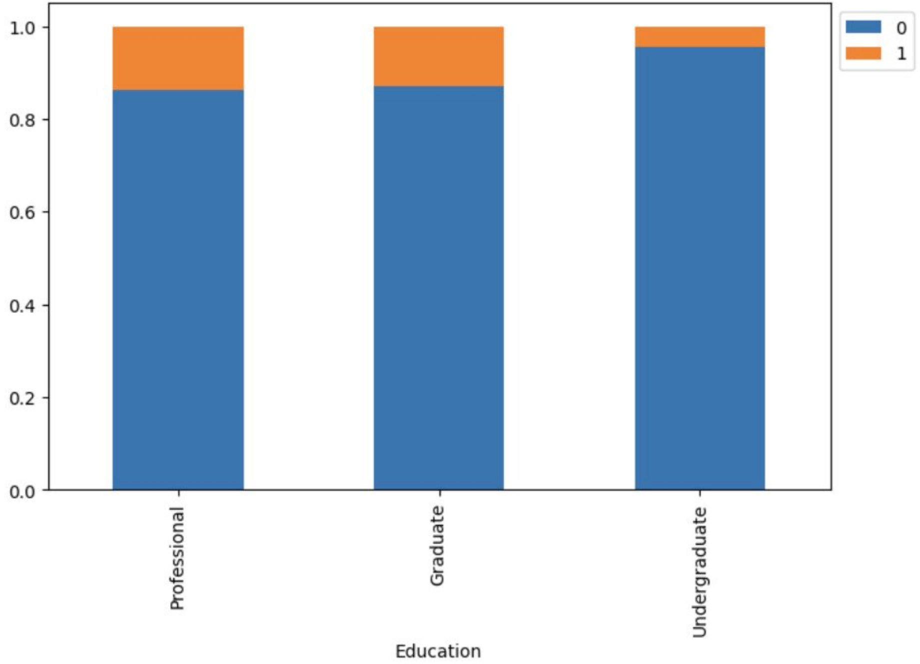
- Higher average monthly credit card spend and higher yearly income both associated with the likelihood of taking a personal loan





EDA - Education Level

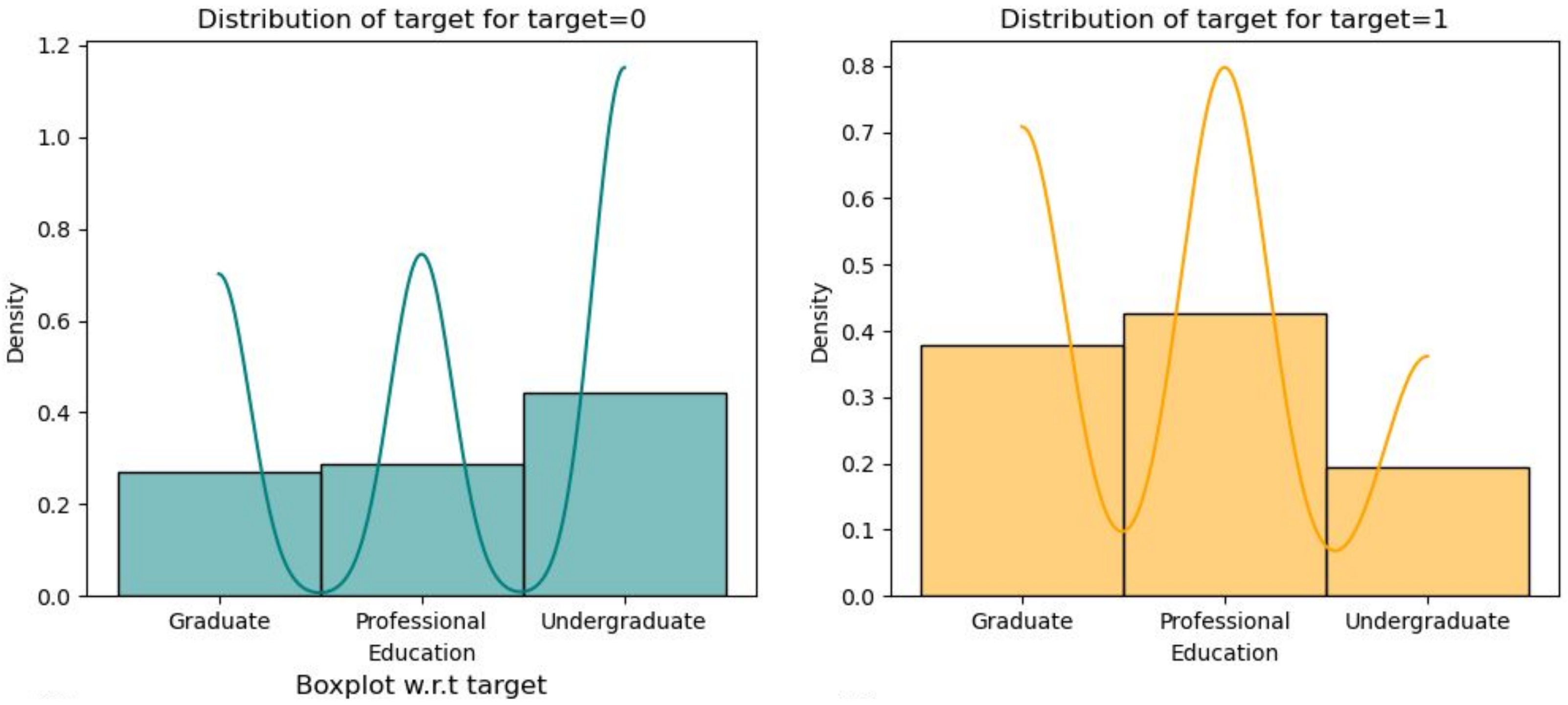




- Only 4% of those with an undergraduate level of education have a personal loan as opposed to 13% of those with a graduate degree and 14% of those with a professional degree

EDA - Education Level

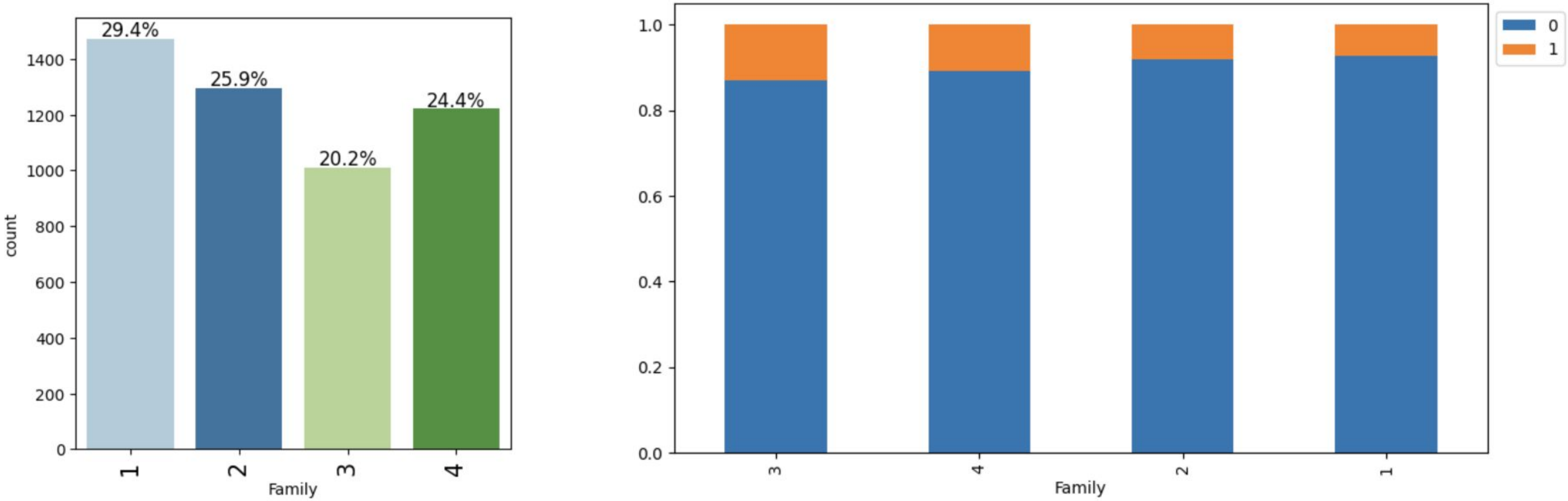
- Appear to be some outliers in Undergraduate Education category that are highly associated with taking a personal loan



Exploratory Data Analysis - Family Size

Number of children in families with a personal loan

- 11% of families with 4 children
- 13% of families with 3 children
- 7% of families with 2 children
- 8% of families with 1 child



Exploratory Data Analysis - Correlation Table

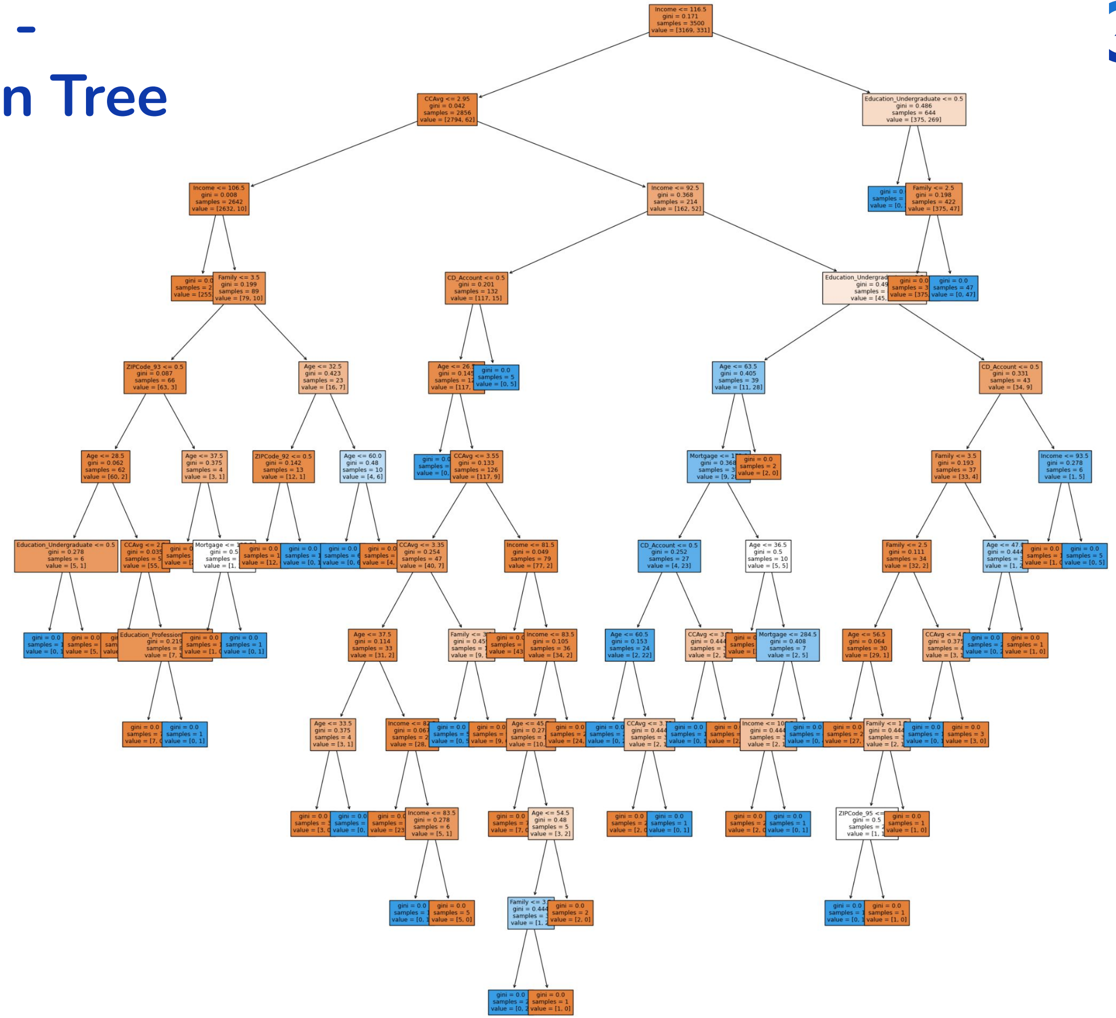
- Correlation between Age & Experience means the features are too similar
- Avg monthly credit card spend associated with income
- Mortgage also associated with income, but to a smaller degree

Data Preprocessing

- There are **no duplicate values** in the data set
- There are **no missing values** in the data set
- Outliers exist for income, mortgage and average monthly spending on credit cards
    - May represent potential customers → no treatment
- Decision Tree Models are not susceptible to outliers, so scaling is not necessary
- Age and experience highly correlated, too similar → **remove Experience** feature
- Encode categorical features

Model Building - Default Decision Tree

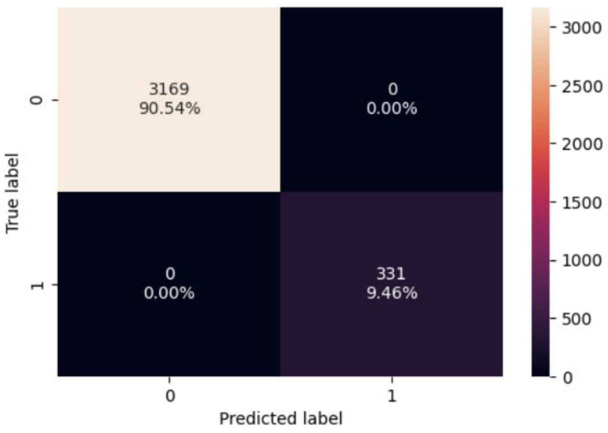
- It’s a mess.



Model Evaluation - Default Decision Tree

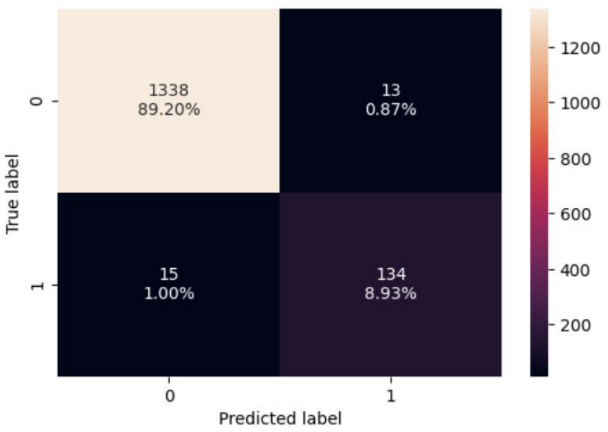
Focus on recall to minimize risk of missing personal loan opportunities

Model 1 Train

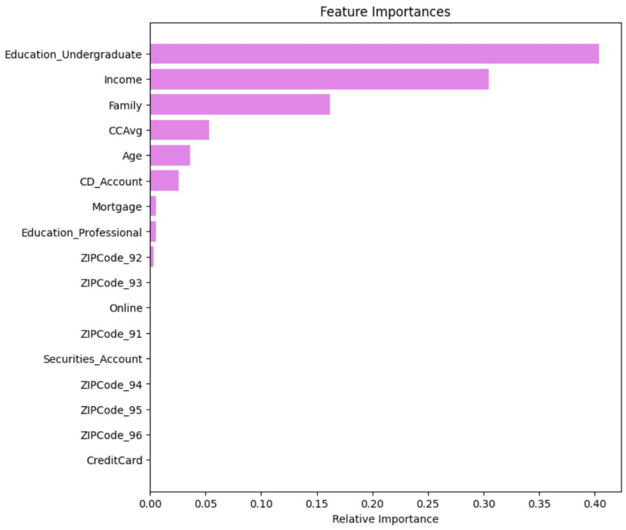


	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Model 1 Test



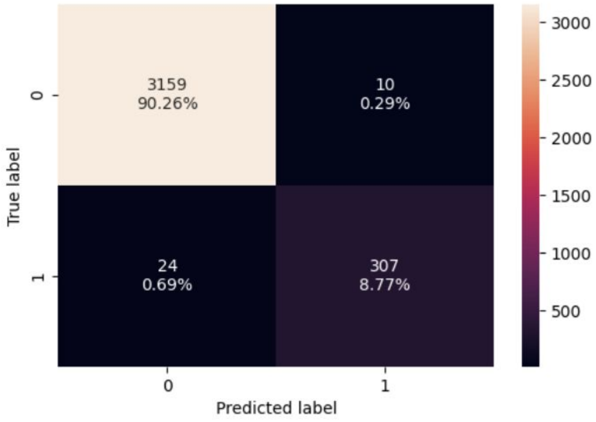
	Accuracy	Recall	Precision	F1
0	0.981333	0.899329	0.911565	0.905405



- Difference between recall in train and test scores indicates overfitting
- Undergrad Education, Income, and Family are most important features

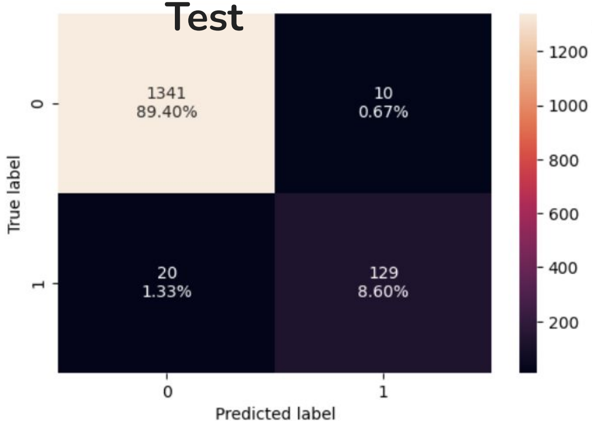
Model Evaluation - Model 2 Pre-Pruning - Max Depth 6

Model 2 Train

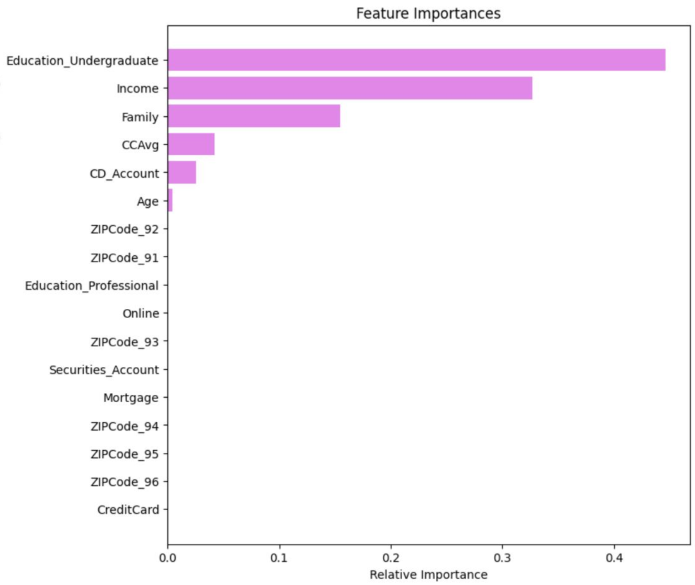


	Accuracy	Recall	Precision	F1
0	0.990286	0.927492	0.968454	0.947531

Model 2 Test



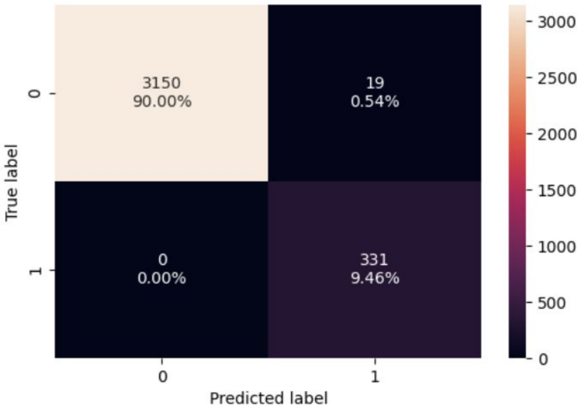
	Accuracy	Recall	Precision	F1
0	0.98	0.865772	0.928058	0.895833



- Recall dropped in both train and test models
- Still a significant difference in recall scores between train and test models
- Top feature importances are Undergraduate Education, Income, and Family

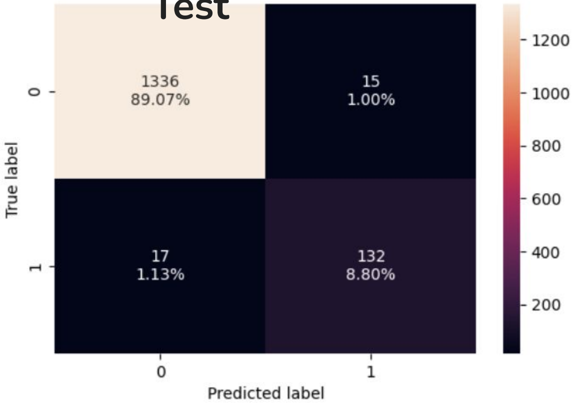
Model Evaluation - Model 3 - Post Pruning

Model 3 Train

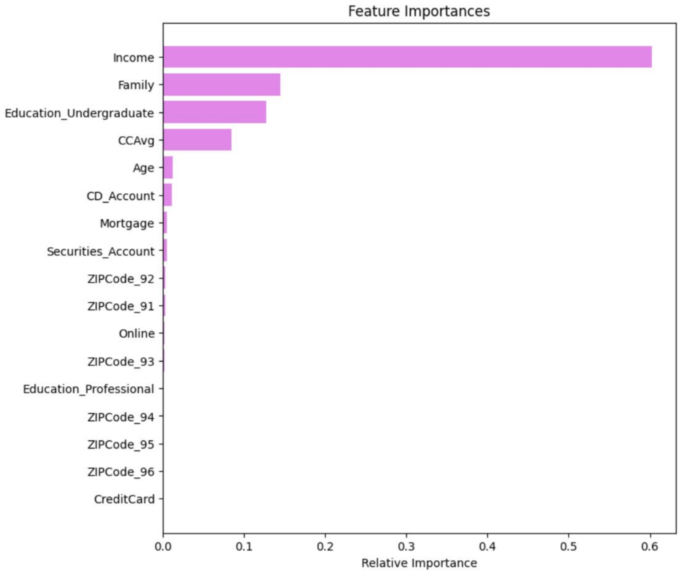


	Accuracy	Recall	Precision	F1
0	0.994571	1.0	0.945714	0.9721

Model 3 Test



	Accuracy	Recall	Precision	F1
0	0.978667	0.885906	0.897959	0.891892



- Still a significant difference between train and test recall
- Feature importances are still Income, Family, and Undergrad Education
- **Credit Card** (whether a member has a credit card with an outside company not related to feature importance - **remove from features and repeat the models**

Model Performance Improvement

- **Repeat Model with variations:**
  - **Change max depth** of tree
  - **Change CCP_alpha** values to optimize relationship between train and test recall values
  - **Remove features** with 0 importance to trees
    - **Credit Card** → has credit card with another bank
    - **Online** → uses online banking

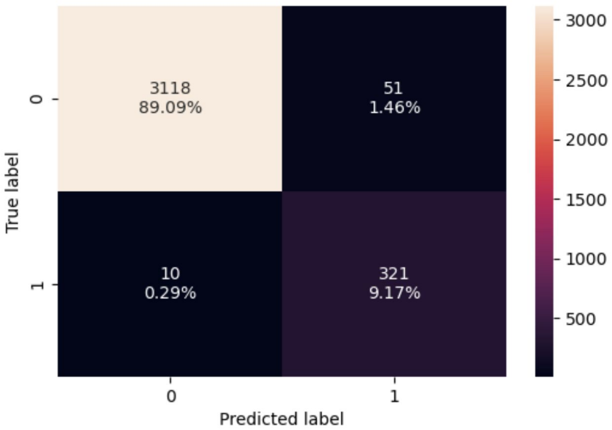
Model Comparisons

Model	Accuracy Train/Test	Recall Train/Test	Precision Train/Test	F1 Train/Test
1 - Default Tree	1.0 / 0.981333	1.0 / 0.899329	1.0 / 0.911565	1.0 / 0.905405
2 - Max Depth 6	0.990286 / 0.98	0.927492 / 0.865772	0.968454 / 0.928058	0.947531 / 0.895833
3 - Max Depth 4	0.987143 / 0.98	0.897281 / 0.845638	0.964286 / 0.947368	0.929577 / 0.893617
4 - No CC, online	1.0 / 0.980667	1.0 / 0.899329	1.0 / 0.905405	1.0 / 0.902357
5 - No CC, online, Max Depth 5	0.990286 / 0.98	0.927492 / 0.865772	0.968454 / 0.928058	0.947531 / 0.895833
6 -Best Accuracy ccp_a = 0.00062	0.994571 / 0.978667	1.0 / 0.885906	0.945714 / 0.897959	0.9721 / 0.891892
7 - ccp_a = 0.001	0.994571 / 0.972	1.0 / 0.926174	0.945714 / 0.816568	0.9721 / 0.867925
8 - ccp_a = 0.0015	0.982571 / 0.971333	0.969789 / 0.912752	0.862903 / 0.819277	0.913229 / 0.863492
9 - ccp_a=0.0016	0.981429 / 0.97	0.966767 / 0.90604	0.855615 / 0.813253	0.907801 / 0.857143

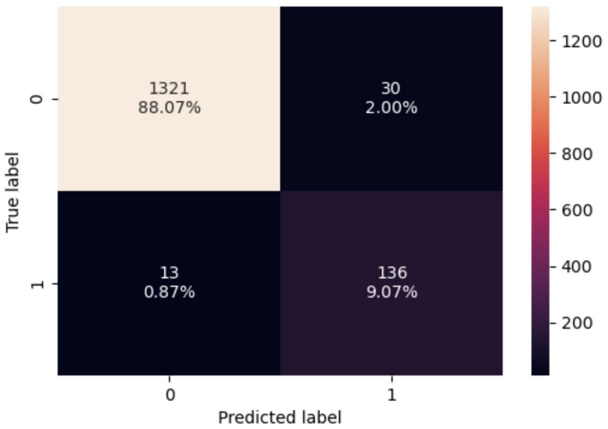
Model Performance Summary

- **Model evaluation criterion -> Model 8 ccp_alpha = 0.0015**
  - **Best recall score**, with **difference between test and train data minimized** to ensure the model will work well on a new data set
  - The cost of a missed loan is higher than the cost of extra advertising
  - **88-89%** of predictions were true positives **(loan advertised, loan taken)**
  - **9%** of predictions in both models true negatives **(no loan offered, none taken)**
  - **2%** false negatives **(no loan offered, opportunity missed)**
  - **< 1%** false positives **(loan offered, no loan taken, unnecessary advertising)**

Train Confusion Matrix



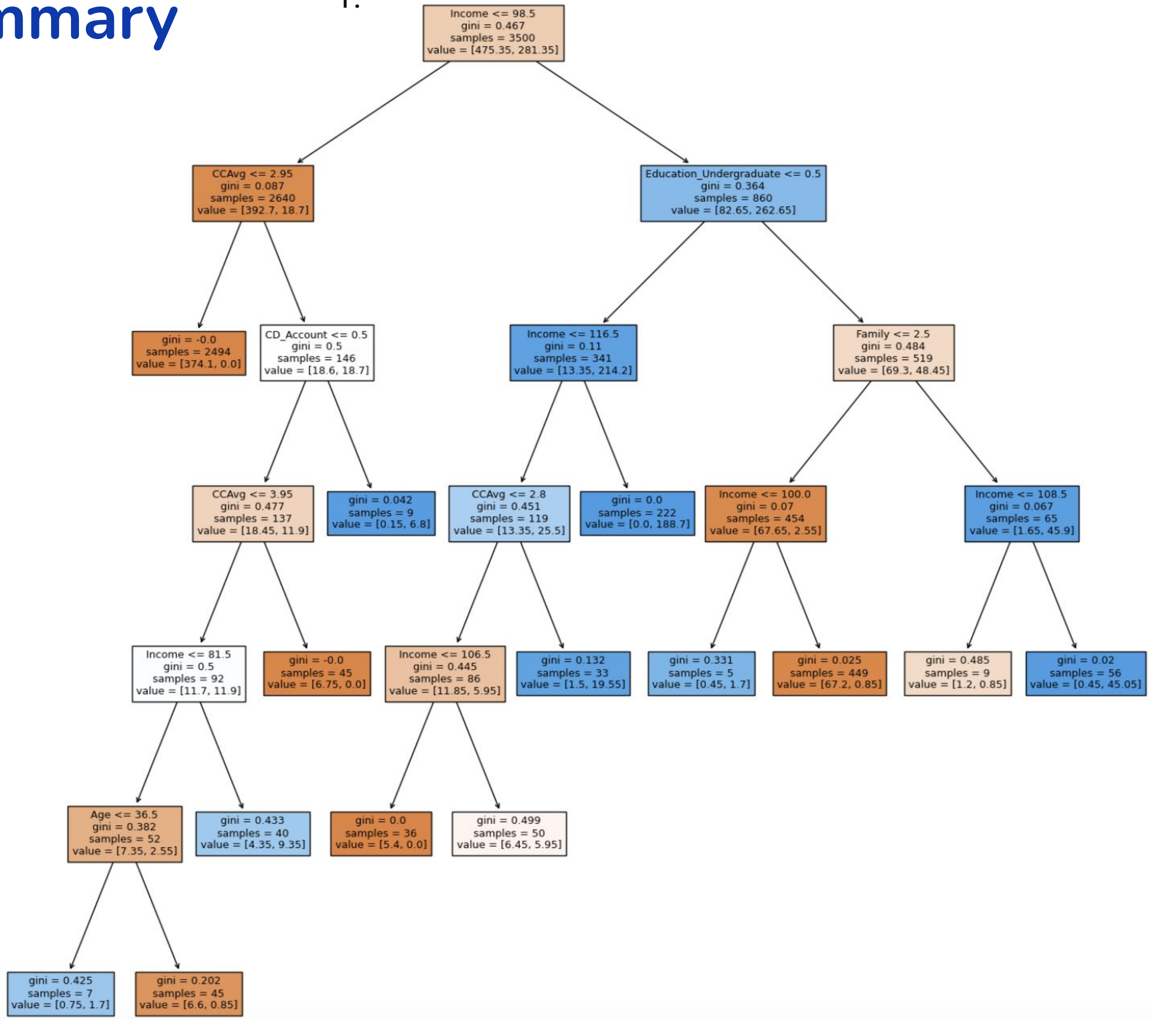
Test Confusion Matrix



Model Performance Summary

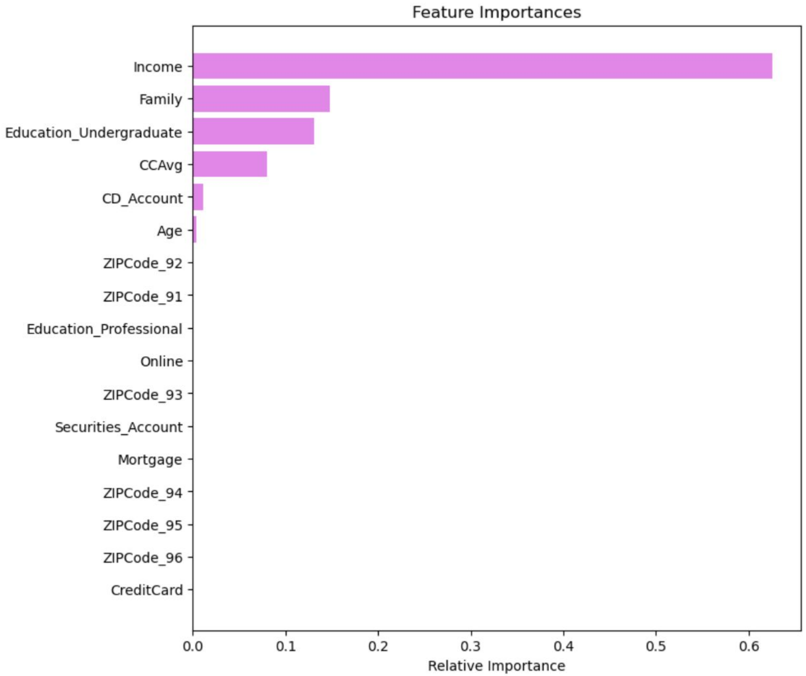
	Imp
Income	0.625392
Family	0.147620
Education_Undergraduate	0.131457
CCAvg	0.080060
CD_Account	0.011731
Age	0.003740
Securities_Account	0.000000
Online	0.000000
Mortgage	0.000000
ZIPCode_91	0.000000
ZIPCode_92	0.000000
ZIPCode_93	0.000000
ZIPCode_94	0.000000
ZIPCode_95	0.000000
ZIPCode_96	0.000000
Education_Professional	0.000000
CreditCard	0.000000

1.

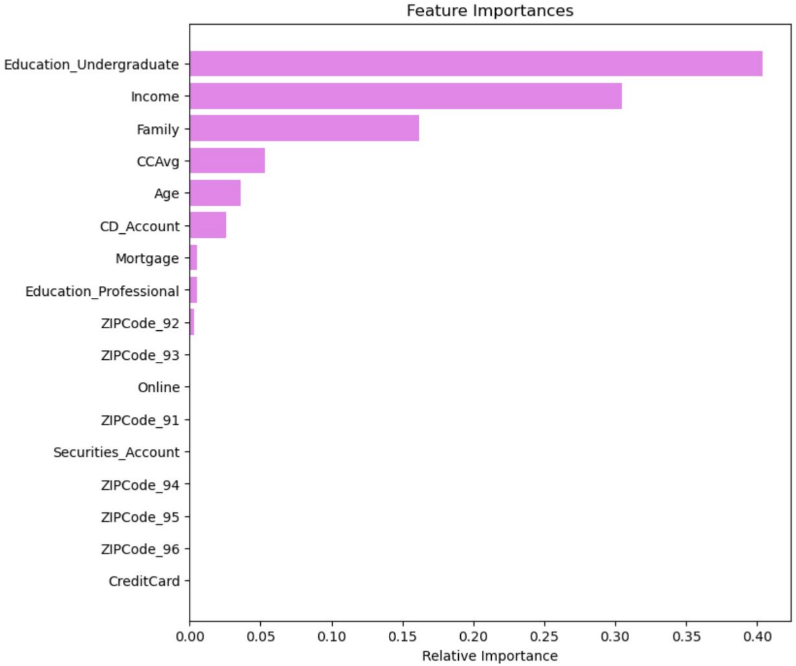


Model Performance Improvement

Final Tree



Default Tree



- The top 3 importance features from the default tree remain the same in the final tree but in a different order.
- The most importance features of the final decision tree in order are **Income**, **Family,** and **Undergraduate Education**

Decision Tree Flow Chart

APPENDIX

Data Background and Contents

- `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? (0: No, 1: Yes)
- `Securities_Account`: Does the customer have securities account with the bank? (0: No, 1: Yes)
- `CD_Account`: Does the customer have a certificate of deposit (CD) account with the bank? (0: No, 1: Yes)
- `Online`: Do customers use internet banking facilities? (0: No, 1: Yes)
- `CreditCard`: Does the customer use a credit card issued by any other Bank (excluding All life Bank)? (0: No, 1: Yes)

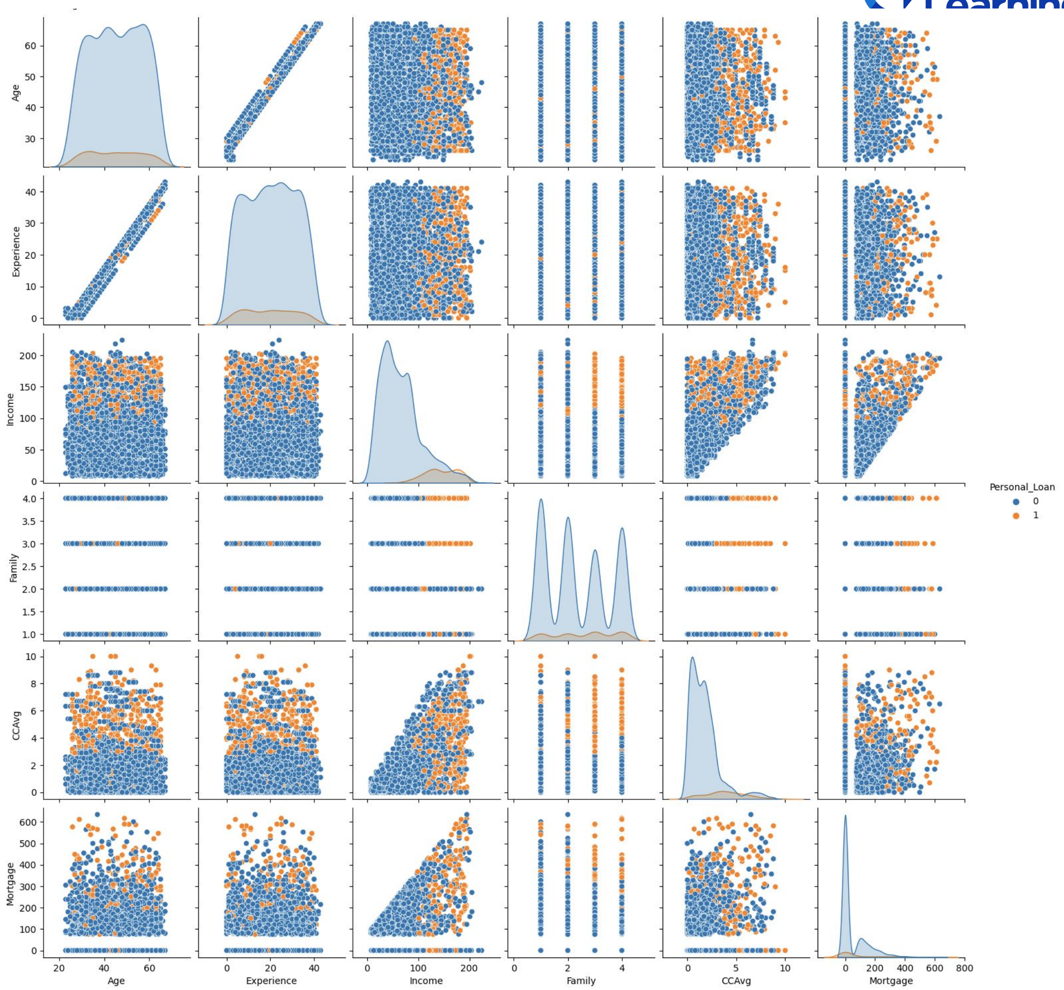
Data Background and Contents

- Size of the data: 5000 rows, 14 columns
- Type of data: 13 integers and 1 float, with 7 categorical variables converted to category
- Target variable: **Personal_Loan**
- Initial observations:
  - Age and experience features are very similar
  - Income, mortgage, and monthly average spending on credit cards all have outliers in the data
- **Summary statistics**:

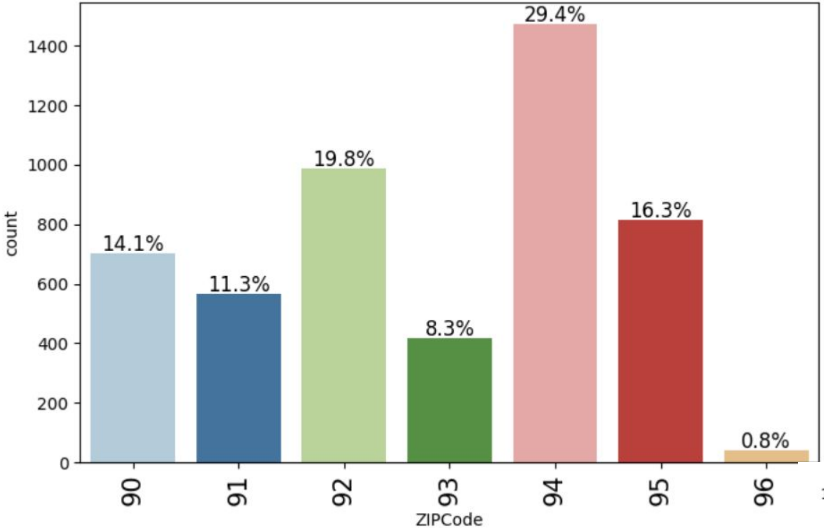
	count	mean	std	min	25%	50%	75%	max
Age	5000.0	45.338400	11.463166	23.0	35.0	45.0	55.0	67.0
Experience	5000.0	20.134600	11.415189	0.0	10.0	20.0	30.0	43.0
Income	5000.0	73.774200	46.033729	8.0	39.0	64.0	98.0	224.0
Family	5000.0	2.396400	1.147663	1.0	1.0	2.0	3.0	4.0
CCAvg	5000.0	1.937938	1.747659	0.0	0.7	1.5	2.5	10.0
Mortgage	5000.0	56.498800	101.713802	0.0	0.0	0.0	101.0	635.0

Exploratory Data Analysis

- We can see a clear link to the likelihood of having a personal loan and higher levels of income, higher mortgage, monthly credit card spend,m and more children

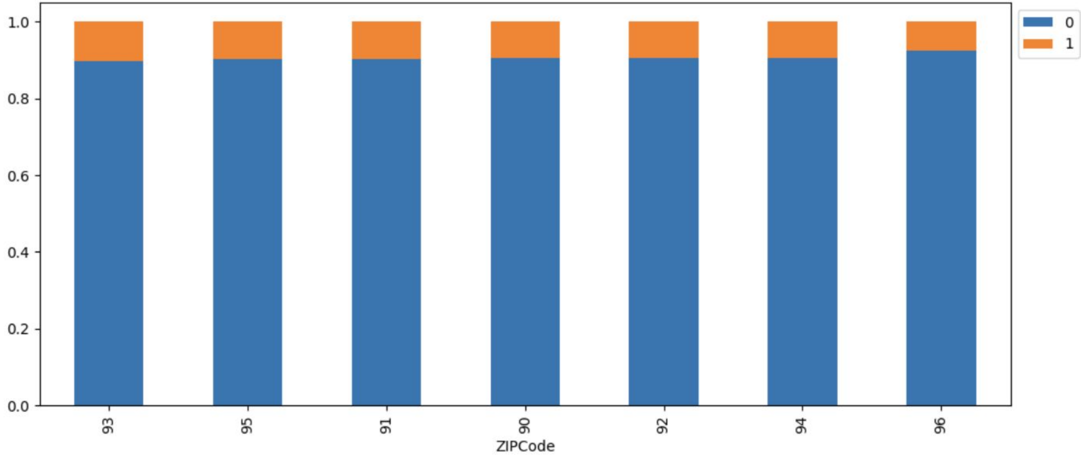


EDA - Zip Code



- The majority of data points come from zip code areas starting with 94

- There do appear to be slight differences in zip code and likelihood of taking a personal loan, but not significant differences



Great Learning

Happy Learning !