# Personal Loan Campaign

## Project 2 – Machine Learning

October 17, 2024

# Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

# Business Problem Overview and Solution Approach

- Objective: We need to predict whether a liability customer will buy personal loans, understand which customer attributes are most significant in diving purchases, and identify exactly which segment of customers to target more.
- Problem: When doing exploratory data analysis we noticed that more than 90% of customers did not take a personal loan – class imbalance in this variable.

# Executive Summary - Call to Action

Key Predictors from the Decision Tree (in hierarchical order):

1.Income: Higher income often indicates a higher likelihood of loan acceptance. Customers with more disposable income may be targeted as they are more likely to qualify and show interest in loans.

2.CCAvg (Credit Card Spending): Customers who spend more on credit cards per month may be more likely to accept personal loans. They may already demonstrate a tendency to use credit, indicating potential demand for personal loans.

3.Education Level: Higher education levels (e.g., Graduate or Advanced degrees) often correlate with a higher likelihood of accepting loans, as educated customers might be more comfortable with financial products.

4.Ownership of Multiple Accounts: Customers with multiple accounts (e.g., securities or CD accounts) may be more engaged with the bank's financial products and are thus more likely to take out personal loans.

5.Mortgage: Customers with mortgages might need personal loans to finance home-related expenses or pay off higher-interest debts.

# Executive Summary - Call to Action

Example of a Likely Customer Profile:

- Age: Middle-aged (around 35–55 years old)
- Income: High income (e.g., more than $100,000 per year)
- CCAvg: Higher credit card spending (e.g., more than $2,500 per month)
- Education: Graduate or Advanced degree
- Account Ownership: Likely to have a securities account, CD account, or both
- Mortgage: May have a mortgage, but not necessarily a very large one compared to income.

# Executive Summary - Call to Action

Actionable Insights:

1. Target high-income customers with significant credit card spending.
2. Focus on customers with higher education levels as they may be more open to exploring financial products.
3. Engage customers who already have multiple accounts with the bank, as they may trust the bank's offerings more.
4. Tailor campaigns to middle-aged customers (aged 30-50), who are often in their prime borrowing years.

This profile should guide future marketing teams to identify the best customers for targeted loan campaigns

# Business Problem Overview and Solution Approach

- The dataset has the following structure based on the first few rows:
- Customer ID
- Age (Years)
- Experience
- Income
- ZIP Code
- Family size of customer
- Average spending on credit cards per month (in thousands of dollars)
- Education level (given values of 1 for Undergrad, 2 for Graduate and 3 for Advanced/Professional degree)
- Mortgage value of the customer’s house (Thousands)
- Personal Loan – Binary value if accepted (0 for “No”, 1 for “Yes)
- Securities Account (also a binary 0/1 value)
- Certificate of deposit account (also binary 0/1 value)
- Does customer use online banking facilities(also binary 0/1 value)
- Does customer use credit card from another bank (also binary 0/1 value)

# EDA Results

- There are no missing values in the data set – all info is fully populated
- Age: The mean/median average age of all customers is 45 years old, from a range of 23 to 67 years old
- Experience: There is an outlier minimum of -3 years in the “Experience” column suggesting a data entry error. Otherwise, the mean value of workforce expereice is 20 years, from a range of 43 years for all
- Income of customers ranges from $8000 - $224000 (in thousands) with a median income of $64,000 (in thousand dollars)
- The median family has about 2 people with 4 being highest recorded maximum
- Most customers have no mortgages (less than 50%) and did not take the loan (only 9.6% have loans)
- Almost 90% have no securities account and almost 95% lack a CD account
- Only 60% of customers seemingly use online banking and and about 30% of customers have credit cards with another bank

# Data Preprocessing

- Duplicate value check – No duplicates were found in the data - ran data.duplicated().sum() to check rows that were exact duplicates of others
- Missing value treatment – No missing values in data set (this was verified previously)
- Outlier check – performed an IQR analysis: 1) There were negative values in the “Experience” column – someone typed negative values as “experience” – this can be corrected. 2) Income and Mortgage both show high variability but this depended on whether extreme values are actually valid. We tested for Age, Income, Experience, Credit Card Spending and Mortgage outliers.
- Feature Engineering
- Data preprocessing for modeling

# Data Preprocessing – Outlier Check

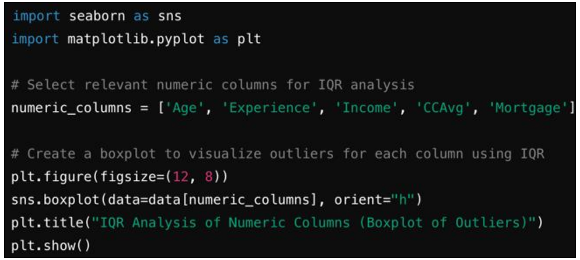
- Outlier check – performed an IQR analysis: 1) There were negative values in the “Experience” column – someone typed negative values as “experience” – this can be corrected. 2) Income and Mortgage both show high variability but this dependsd on whether extreme values are actually valid. Outliers should be considered valid.

Testing for Age:

•Q1 (25th percentile): 35 years
•Q3 (75th percentile): 55 years
•IQR (Interquartile Range): 20 years
•Lower Bound: 5 years (no actual outliers below this in the dataset)
•Upper Bound: 85 years (no actual outliers above this in the dataset)

•**Number of Outliers**: 0

Code Snippet:



```
import seaborn as sns
import matplotlib.pyplot as plt

# Select relevant numeric columns for IQR analysis
numeric_columns = ['Age', 'Experience', 'Income', 'CCAvg', 'Mortgage']

# Create a boxplot to visualize outliers for each column using IQR
plt.figure(figsize=(12, 8))
sns.boxplot(data=data[numeric_columns], orient="h")
plt.title("IQR Analysis of Numeric Columns (Boxplot of Outliers)")
plt.show()
```

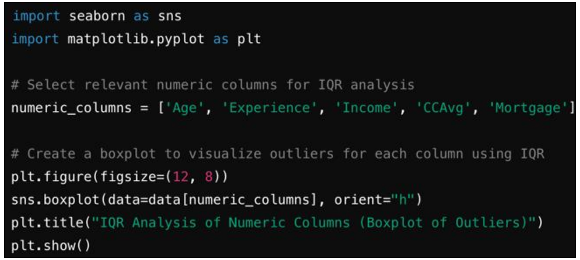
# Data Preprocessing – Outlier Check

- Outlier check – performed an IQR analysis: 1) There were negative values in the “Experience” column – someone typed negative values as “experience” – this can be corrected. 2) Income and Mortgage both show high variability but this dependsd on whether extreme values are actually valid. Outliers should be considered valid.

Experience:

•Q1 (25th percentile): 10 years
•Q3 (75th percentile): 30 years
•IQR (Interquartile Range): 20 years
•Lower Bound: -20 years (some corrections were done for negative experience values)
•Upper Bound: 60 years (no actual outliers above this in the dataset)
•**Number of Outliers: 0 after fixing negative experience values**

Code Snippet:



```
import seaborn as sns
import matplotlib.pyplot as plt

# Select relevant numeric columns for IQR analysis
numeric_columns = ['Age', 'Experience', 'Income', 'CCAvg', 'Mortgage']

# Create a boxplot to visualize outliers for each column using IQR
plt.figure(figsize=(12, 8))
sns.boxplot(data=data[numeric_columns], orient="h")
plt.title("IQR Analysis of Numeric Columns (Boxplot of Outliers)")
plt.show()
```

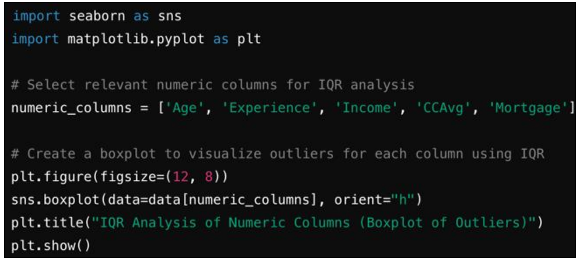
# Data Preprocessing – Outlier Check

- Outlier check – performed an IQR analysis: 1) There were negative values in the “Experience” column – someone typed negative values as “experience” – this can be corrected. 2) Income and Mortgage both show high variability but this dependsd on whether extreme values are actually valid. Outliers should be considered valid.

Income:

- Q1 (25th percentile): $39,000
- Q3 (75th percentile): $98,000
- IQR (Interquartile Range): $59,000
- Lower Bound: -$49,500 (no outliers below this, since income can't be negative)
- Upper Bound: $186,500 (there are some outliers above this limit)
- **Number of Outliers: A few customers have incomes exceeding $186,500**

Code Snippet:



```python
import seaborn as sns
import matplotlib.pyplot as plt

# Select relevant numeric columns for IQR analysis
numeric_columns = ['Age', 'Experience', 'Income', 'CCAvg', 'Mortgage']

# Create a boxplot to visualize outliers for each column using IQR
plt.figure(figsize=(12, 8))
sns.boxplot(data=data[numeric_columns], orient="h")
plt.title("IQR Analysis of Numeric Columns (Boxplot of Outliers)")
plt.show()
```

# Data Preprocessing – Outlier Check

- Outlier check – performed an IQR analysis: 1) There were negative values in the "Experience" column – someone typed negative values as "experience" – this can be corrected. 2) Income and Mortgage both show high variability but this dependsd on whether extreme values are actually valid. Outliers should be considered valid.

CCAvg (Credit Card Average Spending):

•Q1 (25th percentile): $700

•Q3 (75th percentile): $2,500

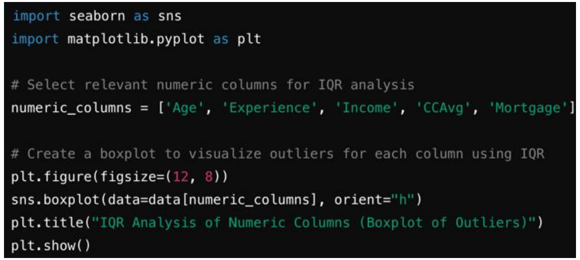
•IQR (Interquartile Range): $1,800

•Lower Bound: -$1,900 (no outliers below this limit)

•Upper Bound: $5,100 (some outliers above this limit)

•Number of Outliers: A few customers spend more than $5,100 on average

Code Snippet:



```
import seaborn as sns
import matplotlib.pyplot as plt

# Select relevant numeric columns for IQR analysis
numeric_columns = ['Age', 'Experience', 'Income', 'CCAvg', 'Mortgage']

# Create a boxplot to visualize outliers for each column using IQR
plt.figure(figsize=(12, 8))
sns.boxplot(data=data[numeric_columns], orient="h")
plt.title("IQR Analysis of Numeric Columns (Boxplot of Outliers)")
plt.show()
```

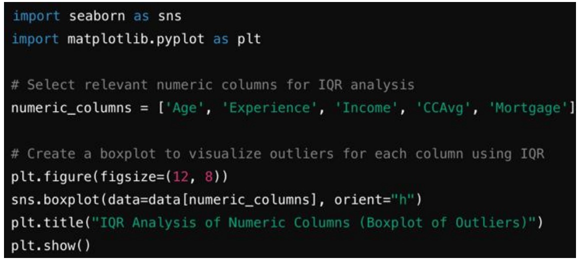
# Data Preprocessing – Outlier Check

- Outlier check – performed an IQR analysis: 1) There were negative values in the “Experience” column – someone typed negative values as “experience” – this can be corrected. 2) Income and Mortgage both show high variability but this dependsd on whether extreme values are actually valid. Outliers should be considered valid.

Mortgage:

- Q1 (25th percentile): $0 (many customers have no mortgage)
- Q3 (75th percentile): $101,000
- IQR (Interquartile Range): $101,000
- Lower Bound: -$151,500 (no outliers below this limit)
- Upper Bound: $252,500 (some customers have mortgages exceeding this)
- **Number of Outliers: A few customers have mortgages greater than $252,500**

Code Snippet:



```
import seaborn as sns
import matplotlib.pyplot as plt

# Select relevant numeric columns for IQR analysis
numeric_columns = ['Age', 'Experience', 'Income', 'CCAvg', 'Mortgage']

# Create a boxplot to visualize outliers for each column using IQR
plt.figure(figsize=(12, 8))
sns.boxplot(data=data[numeric_columns], orient="h")
plt.title("IQR Analysis of Numeric Columns (Boxplot of Outliers)")
plt.show()
```

# Data Preprocessing – Feature Engineering

- We added new features to better assist in model construction:
- 1. Income to Mortgage Ratio – Mortgage Divided by Income
- 2. Total Accounts – A sum of the securities account, CD Account and Credit Card Ownership
- 3. Spending to Income Ratio – Average Credit Card Spending divided by Income

- We then grouped (binning) customers into income brackets (low, medium or high) to simplify income analysis. This was done based on the Gaussian distribution of income levels
- We also tried grouping the Age groups into young, middle aged and senior by the same measure

# Data Preprocessing – Feature Engineering

- We will convert categorical features such as **Education**, **Income_Bin**, and **Age_Group** into one-hot encoded variables. This process will help integrate categorical variables into machine learning models.

- We will scale continuous features like **Income**, **CCAvg**, and **Mortgage** using the standard scaler from the scikit-learn library, which will normalize these values to ensure that they are on the same scale.

- Each feature is rescaled using formula: $z=(x-\mu)/sd$ so it has a mean of 0 and an sd of 1

# Model Building

- We can handle the class imbalance by using a decision tree classifier, which inherently handles some level of imbalance by focusing on information gain at each split. However, to further improve its performance on imbalanced data, we can adjust the class weights during training. This will give more importance to the minority class (customers who accepted the loan) while training the model.

To further improve this training model, we decided to split the dataset into 80% training data and 20% testing data

A decision tree classifier was trained on the aforementioned preprocessed data, with class balancing to handle class imbalance in the target variable (in this case Personal Loans)

Predictions were made on the test data sheet

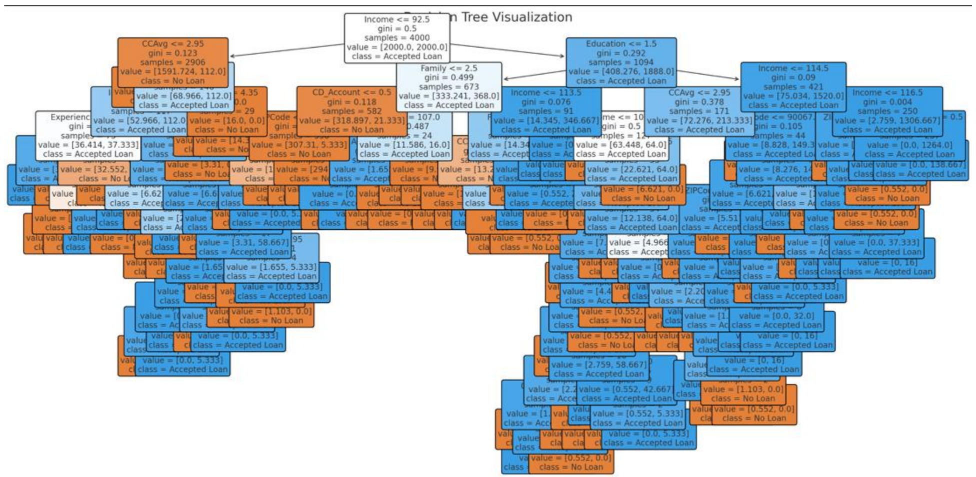
Model performance was evaluated on Accuracy Precision, Recall and F1-Score

# Business Problem Overview and Solution Approach

- We can handle the class imbalance by using a decision tree classifier, which inherently handles some level of imbalance by focusing on information gain at each split. However, to further improve its performance on imbalanced data, we can adjust the class weights during training. This will give more importance to the minority class (customers who accepted the loan) while training the model.

The decision tree model's performance on the test set is as follows:

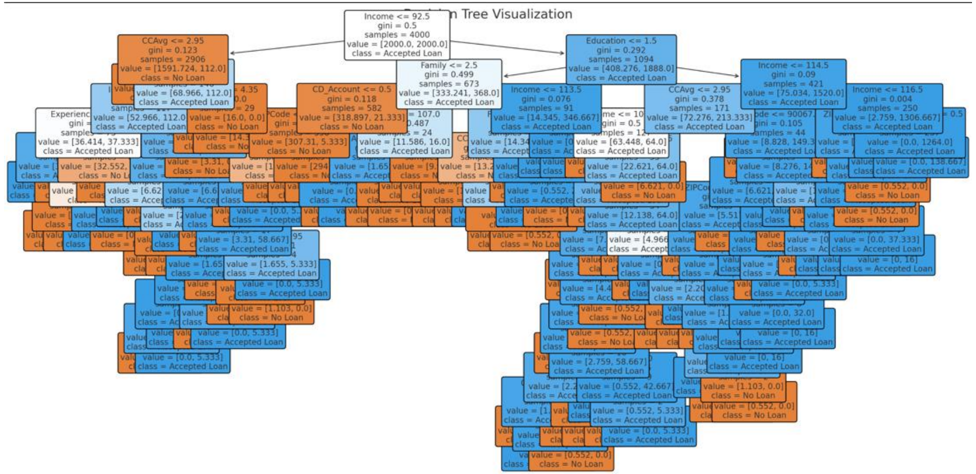
•Accuracy: 98%
•Precision for "No Loan": 99%
•Precision for "Accepted Loan": 92%
•Recall for "No Loan": 99%
•Recall for "Accepted Loan": 90%
•F1-score for "No Loan": 99%
•F1-score for "Accepted Loan": 91%



# Model Performance Summary

The decision tree model's performance on the test set is as follows:

- Accuracy: 98%
- Precision for "No Loan": 99%
- Precision for "Accepted Loan": 92%
- Recall for "No Loan": 99%
- Recall for "Accepted Loan": 90%
- F1-score for "No Loan": 99%
- F1-score for "Accepted Loan": 91%



# Model Performance Improvement

- Additional features can be removed such as Zip Code if they are not showing any correlation or relevance to the training dataset
- We focused narrowly on income and to a lesser extent age, as the key metric to judge a customer’s likeliness to take out a loan

# APPENDIX

Great Learning

# Happy Learning !