

Enhancing Customer Acquisition at AllLife Bank

Jul 4, 2024

Hemant Sharma

Contents / Agenda

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

Executive Summary

- Target High-Income Customer:
 - *Insight:* income is the most significant predictor of loan acceptance, with higher income customers showing likelihood of accepting personal loans.
 - *Recommendation:* focus marketing efforts on customers with higher incomes, as they are more likely to respond positively to personal loan offers.
- Enhance Online Banking Experience:
 - *Insight:* over 50% of customers are active on online banking, indicating a large proportion of loan offer acceptance coming from customers using online banking services.
 - *Recommendation:* improve the online banking platform to make loan application processes more user-friendly, which will help to also promote personal loan offers through online channels.
- Leverage Family & Education Information:
 - *Insight:* customers with larger families & higher education levels are more likely to accept personal loan offers.
 - *Recommendation:* personalize marketing campaigns to emphasize benefits of personal loans for family expenses & educational advancements.
- Refine Credit Card User Targeting:
 - *Insight:* customers with moderate credit card spending (CCAvg) are primary targets for personal loans
 - *Recommendation:* identify customers with specific spending habits & target them with personal tailored loans to match their financial behaviors.

Executive Summary

- Focus on Key Demographics:
 - *Insight:* Age, Family, & Education levels significantly influence customers' loan acceptance capabilities.
 - *Recommendation:* segment customers by age, family size, & education levels to create targeted marketing strategies & improve loan acceptance rates.
- Adjust Pruning Techniques:
 - *Insight:* pre-pruning & post-pruning techniques helped to improve model generalization & performance. Pre-pruning with 'max_depth=6' helped balance the performance metrics.
 - *Recommendation:* review & adjust decision tree model parameters daily, to ensure optimal performance. Use pre-pruning parameters as the baseline, refining them based on updates from new data.

Business Problem Overview and Solution Approach

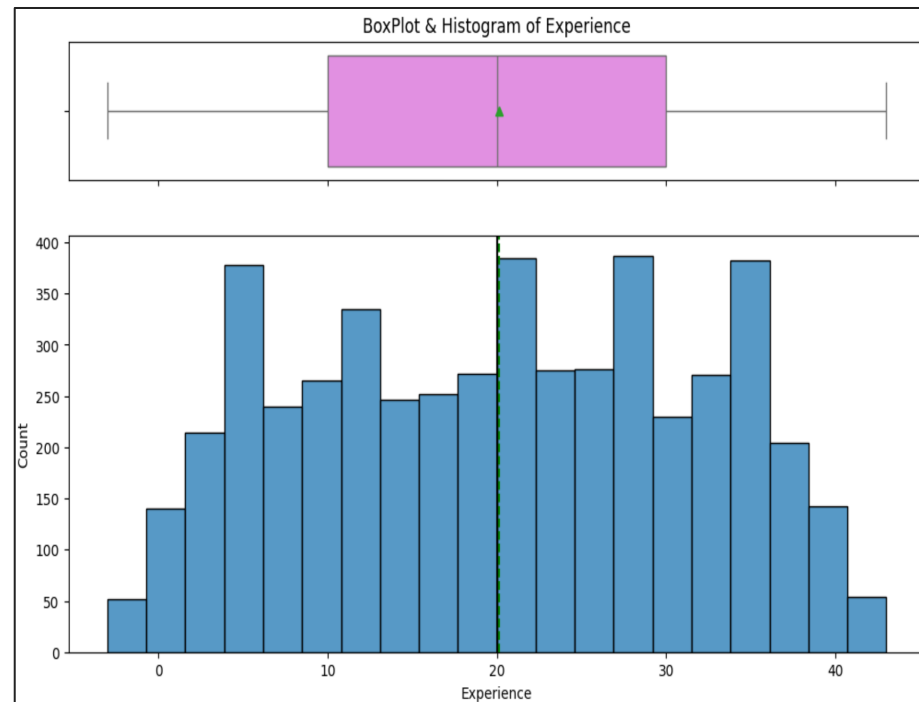
- AllLife Bank wants to increase the number of asset customers (borrowers) who will apply for personal loans, while maintaining those customers as liability customers (depositors) as well. To achieve this, the bank needs a data-driven model to identify and target customers with a higher probability of taking out loans, enhancing the effectiveness of their marketing campaigns.
- To achieve this analysis, we will develop a predictive model using machine learning techniques to analyze historical customer data and identify key attributes that influence loan applications. By segmenting customers based on these attributes, we can create targeted marketing campaigns to improve conversion rates.

EDA Results

- EDA Results – Experience
- EDA Results – Income
- EDA Results - Online
- EDA Results – Heatmap
- Age vs. Personal Loan
- Experience vs. Personal Loan
- Income vs. Personal Loan
- CCAvg. vs. Personal Loan

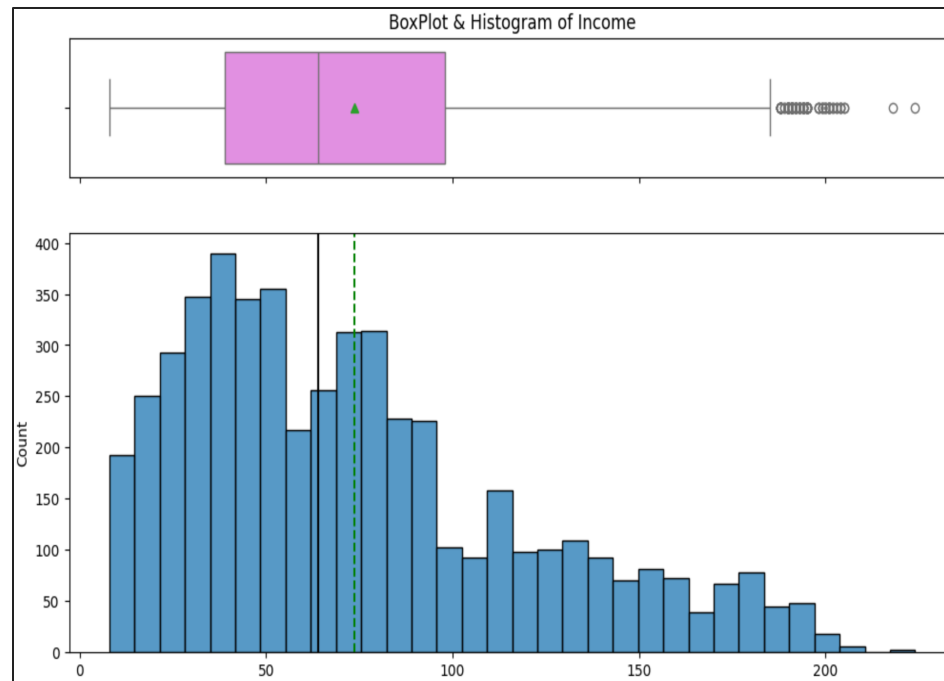
EDA Results – Experience

- Boxplot:
 - The median 'Experience' is prominently indicated by a line within the box, suggesting a central value around 20 years of work 'experience'. It is also safe to assume, since there are no outliers within the boxplot, that the data is relatively consistent, following a predictable pattern.
- Histogram:
 - Also illustrates a relatively uniform distribution with peaks occurring across several levels. This indicates that though customers' work 'Experience' varies widely, most have between 10 to 30 years of 'Experience'.



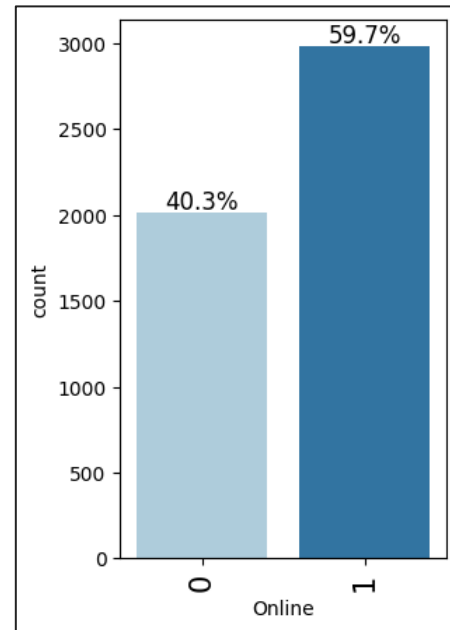
EDA Results- Income

- Boxplot:
 - Most customers' 'income' lies between \$20k-\$100k, concentrated around \$40k-\$60k, highlighting a middle-income majority. However, the box plot is also skewed towards the right (positive skewness) revealing several high-income outliers, which indicates a smaller segment of wealthier customers.
- Histogram:
 - Presents a steady decline in frequency as income increases past \$60k-\$70k, suggesting higher incomes are rarer. In contrast, the right-skewed box plot with many outliers highlights a significant presence of high earners, emphasizing income disparities among customers.



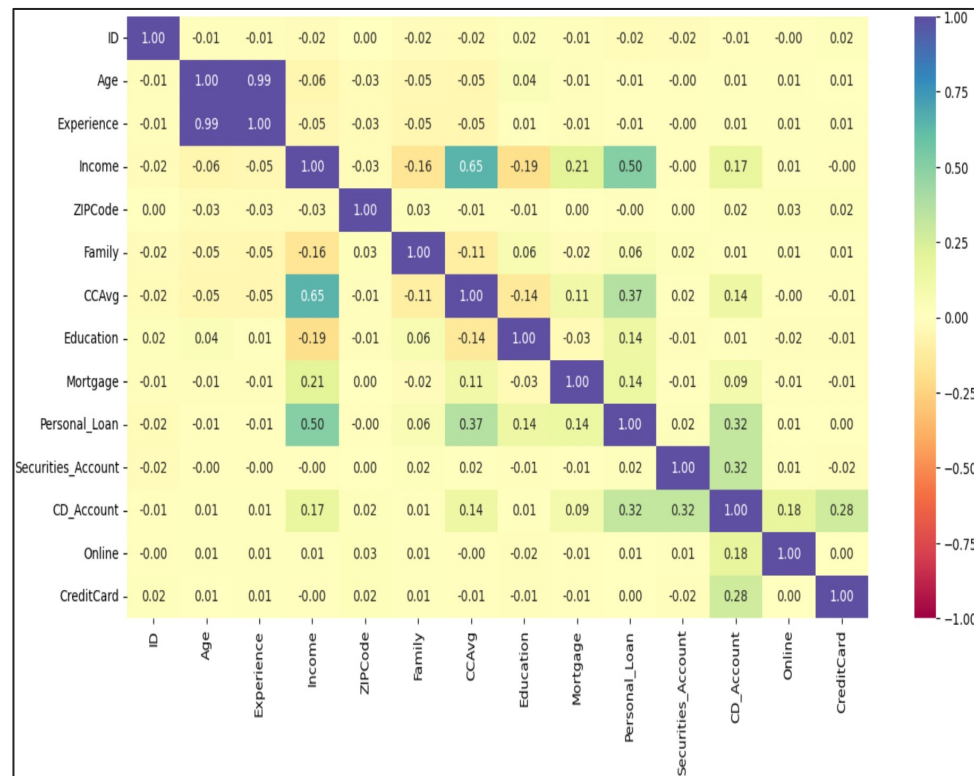
EDA Results- Online

- Barplot:
 - Indicates that over 50% of customers use online banking services, suggesting a significant majority of the customers are engaged with the bank's digital platforms.
 - This highlights the importance of maintaining & improving the online banking platform, as there is a substantial opportunity to leverage this for promoting personal loans to attract remaining 40.3% of customer base.



EDA Results - Heatmap

- *Age & Experience*: strong correlation (.99) between these two categories, which equates to older customers having more experience & younger customers having less experience.
- *Income & CCAvg*: strong correlation (.65) between income & CCAvg suggests that customers with higher income tend to have higher credit card usage.
- *Personal Loan & Income*: moderate correlation (.50) between these two categories suggests that customer who may have a higher income, are more likely to take a loan from AllLife Bank.
- *Education & Income*: negative correlation (-.19) between these two categories suggests that as the level of education increases, there is a possibility for income to decrease or that a higher level of education does not equate to higher income.



Age vs. Personal Loan

○ Histograms:

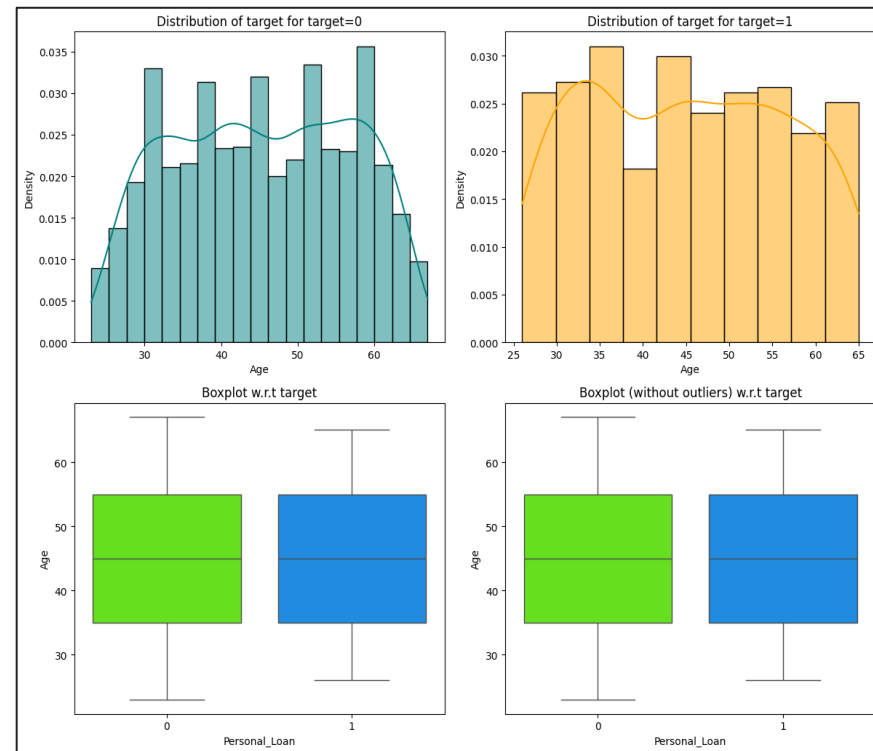
- Age distribution in *Teal histogram* (target = 0) shows a uniform spread with slight peaks, which suggests that a large range of customers have not taken loans.
- Age distribution in *Orange histogram* (target = 1) appears slightly skewed towards middle ages of 30-50yrs old, with a prominent peak around the mid-40. This suggests that customers within this range are more likely to take personal loans compared over other age groups.

○ Boxplots:

- *Green Boxplot's* (target = 0) median age is placed in center of the box, with the interquartile range (IQR) indicating that half of the customers fall within this middle age range.
- *Blue Boxplot's* (target = 1) median age appears slightly higher, and the IQR slightly narrower, which represents a higher concentration of age range among the loan applicants.

○ Overall Analysis:

- Middle-aged customers appear more likely to take personal loans, due to higher financial needs or their ability to be financially stable which allows them to qualify for loans.
- The visuals also present us with the fact that, the lack of a loan is common across all ages but with a slight preference for older (>50yrs) and younger (<30yrs) extremes.



Teal / Green (target = 0) – represents customers who have NOT taken personal loans.
Orange / Blue (target = 1) – represents customers who HAVE taken personal loans.

Experience vs. Personal Loan

○ Histograms:

- As per the distribution of experience of customers, the *Teal histogram* has a uniform distribution overall, but also shows a high peak around 10-20yrs of experience.
- Like the graph before, the *Orange histogram* also has a smooth, yet more prominent peaked distribution, exactly around 10-20yrs of experience, suggesting that as customer tend to get closer to this work experience range, there is a higher probability for them to take out loans.

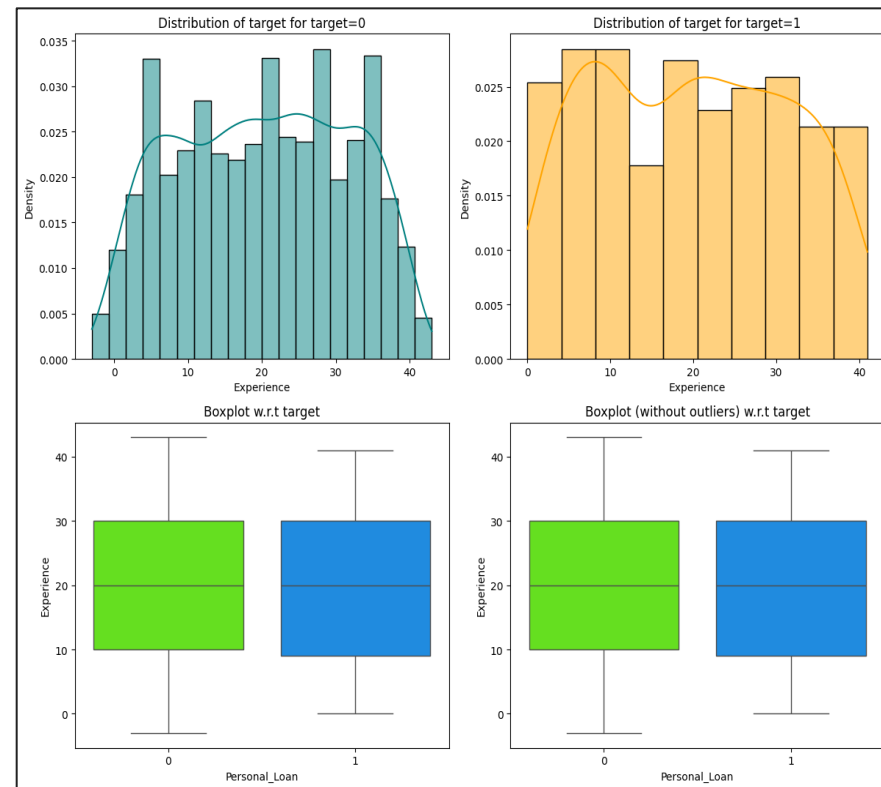
○ Boxplots:

- Green boxplot* shows the IQR with median roughly around 20yrs experience, while the whiskers extend outwards to cover most of the experience range, with no outliers.
- Blue boxplot* presents a similar median of around 20yrs experience, but with a slightly broader IQR, indicating a little more wide spread in years of experience.

○ Overall Analysis:

- Boxplots suggest that while median experience across both groups (green= 0 & blue =1) is somewhat similar, customers who have taken loans (blue =1) have a slightly higher level of experience in years, which can influence their decision to take out loans.

Teal / Green (target = 0) – represents customers who have NOT taken personal loans.
Orange / Blue (target = 1) – represents customers who HAVE taken personal loans.



Income vs Personal Loan

○ Histograms:

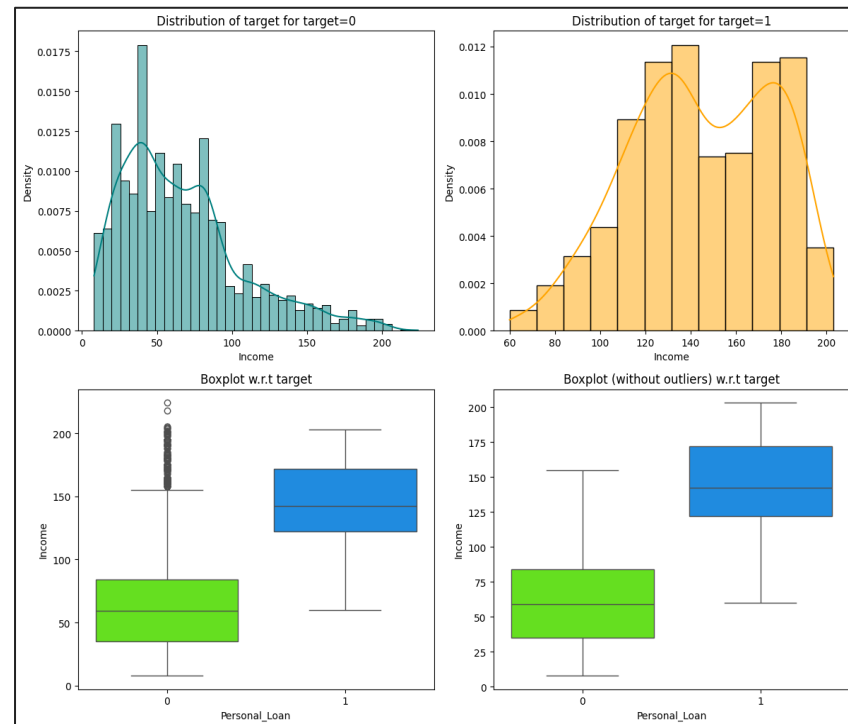
- The *Teal histogram* shows a high-density peak at lower income values, gradually tapering off as income increases. This suggests that most individuals who have not taken a personal loan, mostly have lower incomes
- The *Orange histogram* is more concentrated around the higher income values, with a significantly high peak around \$130k. This indicates that individuals with higher income levels are more likely to take a personal loan.

○ Boxplots:

- *Green boxplot* displays the median income is relatively lower, with a wider IQR. Boxplot also presents several outliers above upper whisker, indicating high number of individuals with much higher income than median who haven't taken a loan.
- *Blue boxplot* displays median income is higher & box is narrower, shifting upwards. This indicates that individuals who have taken a personal loan tend to have higher income.

○ Overall Analysis:

- Customers who have taken personal loans, generally fall into the higher income category, as opposed to those who have not taken personal loans or can't afford them due to lower income as seen by the presence of outliers in the non-loan group (*green boxplot*).



Teal / Green (target = 0) – represents customers who have NOT taken personal loans.

Orange / Blue (target = 1) – represents customers who HAVE taken personal loans.

CCAvg vs Personal Loan

○ Histograms:

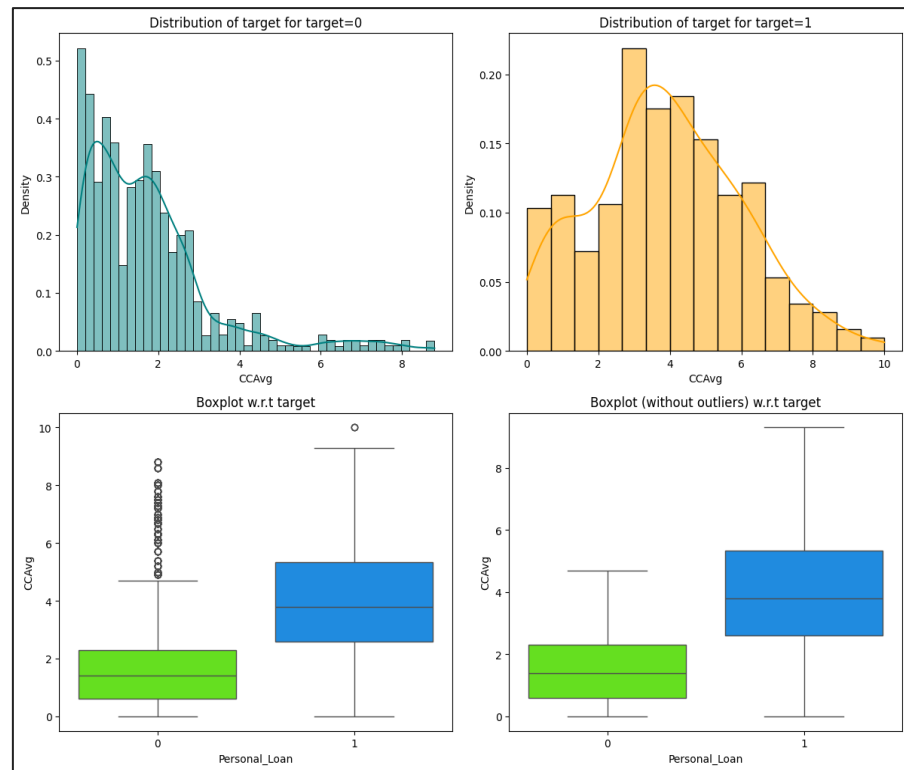
- *Teal histogram* presented is skewed rightwards, indicating that most customers who don't have a personal loan spent less each month on credit cards. This is shown by the significant peak at the lower spending levels.
- *Orange histogram* presented has an evenly spread distribution for customers who have taken a personal loan. The distribution peaks at higher values of spending, suggesting customers with personal loans generally have a higher credit card usage.

○ Boxplots:

- *Green boxplot* shows a lower median spending, with a broader IQR range & outliers, which indicates some customer have significantly higher credit card spending than the median, even though they have not yet signed up for any personal loans.
- *Blue boxplot* shows a higher median, with a lower IQR, which is consistent with the findings that customers with higher credit card spending are more open to accepting personal loans.

○ Overall Analysis:

- Customers who have had a higher CCAvg are more likely to accept personal loans & vice-versa, potentially reflecting the fact that higher overall financial activity equates to better creditworthiness for loans.



Teal / Green (target = 0) – represents customers who have NOT taken personal loans.

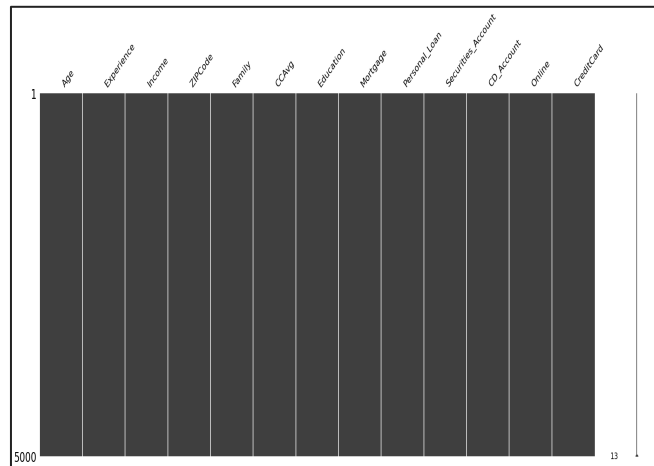
Orange / Blue (target = 1) – represents customers who HAVE taken personal loans.

Data Preprocessing

- Duplicate & Missing Values
- Outlier check (treatment if needed)
- Feature Engineering
- Data Preparation for Modeling

Duplicate Values & Missing Values Check

- As seen in the visuals to the right, there are no duplicate values, across the dataset.
- Additionally, the visual to the bottom right suggests there are no missing values within the dataset, indicating the dataset does not include any erroneous details (duplicate or missing values) & is highly pure.



```
Age 0
Experience 0
Income 0
ZIPCode 0
Family 0
CCAvg 0
Education 0
Mortgage 0
Personal_Loan 0
Securities_Account 0
CD_Account 0
Online 0
CreditCard 0
dtype: int64
```


Data Preprocessing: Outlier Check

- As visible in the to the right, attributes such as 'Mortgage' (5.82%), 'Securities_Account' (10.44%) 'CD_Account' (6.04%) suggest high variability in the financial behavior of AllLife Bank customers. These could indicate that few customers may have higher values, which could heavily skew analysis.
- Possible Treatment:
 - We could transform the data for 'Securities_Account' & 'CD_Account' by feature engineering to make them into "categorical" type, to normalize the data & reduce the impact of outliers.

```
ID          0.00
Age         0.00
Experience  0.00
Income      1.92
ZIPCode     0.00
Family      0.00
CAvg        6.48
Education   0.00
Mortgage    5.82
Personal_Loan 9.60
Securities_Account 10.44
CD_Account  6.04
Online      0.00
CreditCard  0.00
dtype: float64
```

Feature Engineering

- 'ZIPCode' column was transformed to only keep the first two digits. This feature engineering will potentially help capture regional trends without having too many unique categories for the 'ZIPCode' feature, as can be seen in snippets attached
- In addition to reducing the 'ZIPCode' feature to first two digits, we also changed it to the "category" type from "int" type. Same was done for features such as 'Education', 'Personal_Loan', 'Securities_Account', 'CD_Account', 'Online', & 'CreditCard', to improve accuracy of calculations.
- Overall Analysis:
 - Reducing the number of unique values in ZIPCode & transforming said features into categorical types, simplifies the input data, allowing the model to uncover patterns of meaning without overfitting of insignificant details in data.

```
Number of unique values in ZipCode: 467
```

```
Number of unique values if we take first two digits of ZIPCode: 7
```

Data Preparation for Modeling

- Feature Removal:
 - We chose to remove the 'Experience' feature due to its high correlation with 'Age' to prevent data redundancies.
- Dummy Variables:
 - Converted categorical variables 'ZIPCode' & 'Education' to dummy variables to reduce multicollinearity.
- Data Splitting:
 - Split the dataset into training (70%) & testing (30%) sets to evaluate the model's performance on new data.
- Loan Uptake Analysis:
 - Analysis given shows that both the training & testing models show a consistent distribution of individuals who have & have not taken personal loans.

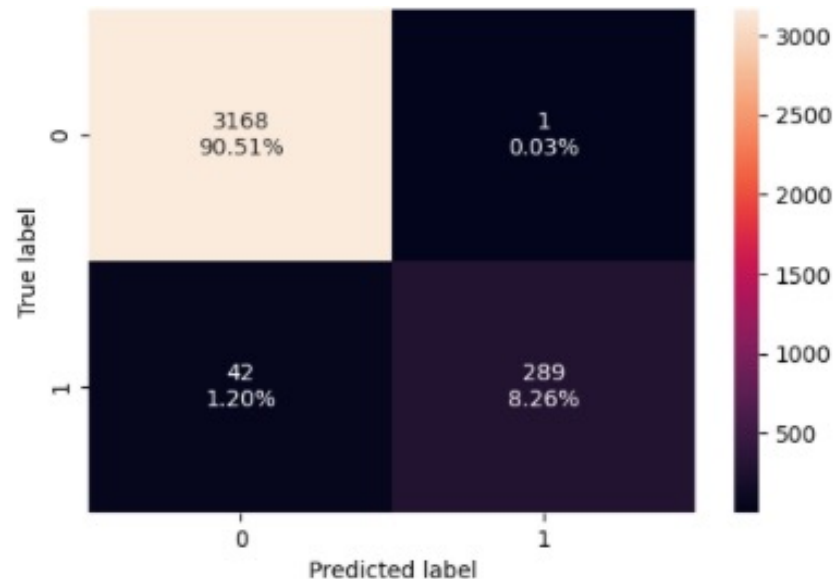
```
Shape of Training set : (3500, 478)
Shape of test set : (1500, 478)
Percentage of classes in training set:
Personal_Loan
0    0.905429
1    0.094571
Name: proportion, dtype: float64
Percentage of classes in test set:
Personal_Loan
0    0.900667
1    0.099333
Name: proportion, dtype: float64
```

Model Building

- Data Preprocessing:
 - Loading the dataset & understanding the structure (#of rows, columns, type of data, missing/duplicate values if any, etc..)
 - Handle any missing or duplicate values, & if required, transform any variable data types to different data type (categorical to numerical & so on).
 - Splitting data into features (X) & target variables (Y). In this scenario, Y being 'Personal_Loan'.
 - Upon splitting the data, we further divide it into training & testing sets through the 'train_test_split'.
- Training Models:
 - Implemented use of 'DecisionTreeClassifier' from 'sklearn.tree' library, to help setup decision tree modeling.
 - Set parameters for modeling such as 'gini criterion', 'min_sample_split' or 'max_depth' & trained the models on training set.
- Evaluating Models:
 - Evaluated model performances using metrics such as accuracy, precision, recalls, & F1 scores.
 - Used the 'confusion_matrix' to get a detailed breakdown about the predictions.
- Hyperparameter Tuning:
 - Implemented the use of 'GridSearchCV' on the training data to find the best parameters that would fit the model predictions.
 - Re-evaluated the model with the best parameters suggested to find any improvements in performance.

Model Performance- Training Data

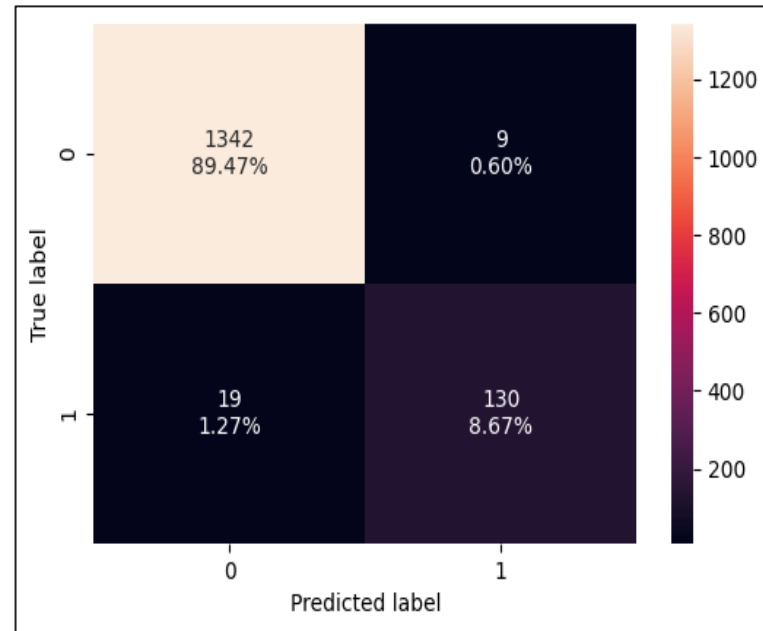
- Modeled using the training data, which includes features (x_{train}) & the target variable (y_{train}) indicating whether customer took a personal loan.
- Confusion matrix is displayed as a heatmap, presenting the number of of correct/incorrect predictions made by the model on training data.
 - True Negative 0 (no loan): model predicts 3168 cases correctly, equating to 90.51% of the training set.
 - True Positive 1 (loan): model correctly predicts 289 cases of actual loans, equating to 8.26% of the training set.
- Evaluation Criterion:
 - Accuracy: proportion of correct predictions among total cases examined. (98.77%)
 - Recall: proportion of actual positives correctly identified. (87.31%)
 - Precision: proportion of positive predictions that are correct. (99.65%)
 - F1-Score: balance between precision & recall. (93.07%)



	Accuracy	Recall	Precision	F1
0	0.987714	0.873112	0.996552	0.930757

Model Performance- Testing Data

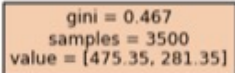
- Modeled using the testing data, which includes features (x_{test}) & the target variable (y_{test}) indicating whether customer took a personal loan.
- Confusion matrix is displayed as a heatmap, presenting the number of of correct/incorrect predictions made by the model on testing data.
 - *True Negative 0 (no loan)*: model predicts 1342 cases correctly, equating to 89.47% accuracy of the testing set.
 - *False Positive (no loan)*: model predicts 9 cases incorrectly, equating to .60% inaccuracy of customer who did not take loans but were identified to have taken loans.
 - *True Positive 1 (loan)*: model predicts 130 cases correctly, equating to 8.67% accuracy of the testing set for customers who took personal loans.
 - *False Negative (loan)*: model predicts 19 cases incorrectly, equating to 1.27% inaccuracy of customers who did take loans but were identified to have not taken loans.
- Evaluation Criterion:
 - Accuracy: proportion of correct predictions among total cases examined. (98.13%)
 - Recall: proportion of actual positives correctly identified. (87.25%)
 - Precision: proportion of positive predictions that are correct. (93.52%)
 - F1-Score: balance between precision & recall. (90.28%)



	Accuracy	Recall	Precision	F1
0	0.981333	0.872483	0.935252	0.902778

Model Performance- Final Decision Tree

- Parameters Used:
 - Max_depth = 6
 - Max_leaf_nodes = 10
 - Min_samples_leaf = 10
 - Random_state = 1
- Overall Analysis:
 - The final decision tree (pre-pruned model) with 'max_depth=6' was selected based on its balanced performance metric & generalization ability.
 - Model used 'Income' as the root node for splitting, indicating that 'Income' is a significant predictor for determining if customers will take a personal loan or not.



gini = 0.467
samples = 3500
value = [475.35, 281.35]

Model Performance Summary

- *Accuracy*: the decision tree model on the training set indicates it fits the data almost too well. Same is reflected on the classification of the performance metrics, with scores of 1.0 across all key measures (accuracy, recall, precision, F1 score).
- *Precision/Recall*: high precision & recall values on the training set indicate that the model is quite good at correctly identifying true positives & has fewer false positives.
- *F1-Score*: the training set model performed much better than the testing set model, most likely because we used parameters such as 'max_depth' & 'min_samples_split' to help prevent overfitting.
- *Feature Importance*: the use of 'Income' as the root node for splitting, signifies the importance of determining if customer would take loans. In addition, the early & frequent presence of features such as 'Education', 'Family' & "CCAvg" also played a key role to predict if customers were to take out loans or not.
 - Income = 33.77%
 - Family = 27.56%
 - Education (Level 2)= 17.57%
 - Education (Level 3)= 15.73%
 - CCAvg= 4.29%

Note: *You can use more than one slide if needed*

Model Performance Improvement

- Model Performance Improvement: Pruning Techniques
- Comparative Analysis: No Pruning vs Pre-Pruning vs Post-Pruning
- Summary of Decision Rules & Features

Model Performance Improvement: Pruning Techniques

- *No Pruning Performance:*
 - **Training set:** shows a clear sign of overfitting with such perfect percentages across analysis.
 - Accuracy = 100%
 - Recall = 100%
 - Precision = 100%
 - F1-Score = 100%
 - **Testing set:** performance metrics drop, indicating model is more precise, but the lower recall suggests model may have predicted the positive cases inaccurately.
 - Accuracy = 97.87%
 - Recall = 78.52%
 - Precision = 100%
 - F1-Score = 87.97%
- *Pre-Pruning Performance(max_depth=6):* max depth helped to control the complexity of the decision tree & prevent overfitting on the Training & Testing sets.
 - **Training Set:** with reduction in the recall value, the model is better at identifying the true positives in data. Additionally, the improvement in precision & f1-score indicated better positive predictions & balance between recall & precision.
 - Accuracy = 98.77%
 - Recall = 87.31%
 - Precision = 99.65%
 - F1-Score = 93.08%
 - **Testing Set:** improvement in accuracy indicates better generalization, along with recall improvement allowing model to capture more true positive cases.
 - Accuracy = 98.13%
 - Recall = 87.25%
 - Precision = 93.52%
 - F1-Score = 90.28%

Conclusion:

Implementing the use of (max_depth) allowed for a more balanced model, reducing the redundancies of the overfit training set to improve the testing set. The (max_depth) enhanced the analyses of Recall & F1-Scores, which are crucial to locate missing true positives (potential loan customers).

Model Performance Improvement: Pruning Techniques

- *Post-Pruning (Cost-Complexity)*: use of parameter (ccp_alpha) to evaluate complexity of tree & remove sections not contributing to reducing cost function to select the optimal pruned tree.
 - **Training Set**: recall value significantly improved, capturing more true positives, while precision remains high indicating most positive predictions are correct.
 - Accuracy = 98.67%
 - Recall = 86.54%
 - Precision = 99.23%
 - F1-Score = 92.35%
 - **Testing Set**: accuracy lower than pre-pruned model, but an improvement in recall value & precision value indicating most true positives captures are correct.
 - Accuracy = 98.05%
 - Recall = 86.72%
 - Precision = 93.12%
 - F1-Score = 89.74%

Comparative Analysis: No Pruning vs Pre-Pruning vs Post-Pruning

- Overfitting:
 - *No Pruning*: training set is well overfitted, showing signs of lower test set performance, especially in recall.
 - *Pre-Pruning*: avoid overfitting by controlling the 'max_depth' of the tree, resulting in better test set performance & good balance between precision & recall.
 - *Post-Pruning*: reduced overfitting by removing unnecessary branches through use of 'ccp_alpha'.
- Balance between Precision/Recall:
 - *No Pruning*: high precision, low recall indication the model could be missing several values.
 - *Pre-Pruning*: better balance & significant improvements in recall, while maintaining a high precision rate.
 - *Post-Pruning*: almost similar balance to pre-pruning, with a higher recall & precision rate bringing about a significantly improved F1 score.
- Generalization:
 - *No Pruning*: poor generalization due to overfitting, clearly indicated by a significant drop in performance metrics.
 - *Pre-Pruning*: improved & balanced generalization, indicated by a higher test accuracy, recall & F1-Score.
 - *Post-Pruning*: almost similar results of generalization to that of pre-pruning, indicating both techniques effectively reduced overfitting.

Summary of Decision Rules & Features

- Parameters:
 - 'max_depth'= 6
 - 'max_leaf_nodes'= 10
 - 'min_samples_leaf' = 10
 - 'random_state'= 1
- Most Important Features Used:
 - Income: 33.77%
 - Family: 27.56%
 - Education (level 2): 17.57%
 - Education (level 3): 15.73%
 - CCAvg: 4.29%

- Key Performance Metrics for Training & Testing Data Models

Data Tree Model	Accuracy	Recall	Precision	F1-Score
No Pruning (training)	1.0	1.0	1.0	1.0
Pre-Pruning (training)	0.9877	0.8731	0.9965	0.9308
Post-Pruning (training)	0.9867	0.8654	0.9923	0.9235
No Pruning (test)	0.9787	0.7852	1.0	0.8797
Pre-Pruning (test)	0.9813	0.8725	0.9352	0.9028
Post-Pruning (test)	0.9805	0.8672	0.9312	0.8974

APPENDIX

Data Background and Contents

○ Features in Data:

- *ID*: unique identifier of each customer (dropped in analyses due to redundancy)
- *Age*: customer age in years.
- *Experience*: years of professional experience.
- *Income*: annual income of customer (thousands of dollars)
- *ZIPCode*: customers' residential zip code.
- *Family*: family size of customers.
- *CCAvg*: customers' average credit card spending per month (thousands of dollars)
- *Education*: customers' education levels identified as listed below:
 - 1 = undergrad
 - 2 = graduate
 - 3 = Advanced/Professional
- *Mortgage*: value of customers' house mortgage
- *Personal Loan*: target variable in analyses, indicating if customer accepted loan offer.
 - 0 = No
 - 1 = Yes
- *Securities_Account*: if customer has a securities account with AllLife Bank.
 - 0 = No
 - 1 = Yes
- *CD_Account*: if customer has a CD account at AllLife Bank.
 - 0 = No
 - 1 = Yes
- *Online*: if customer uses internet banking
 - 0 = No
 - 1 = Yes
- *CreditCard*: if customer uses a credit card issued by AllLife Bank.
 - 0 = No
 - 1 = Yes

Additional Insights

- Age & Experience:
 - Customers in age range 35-50 with approx. 10-20 years of experience are more likely to take out personal loans
- Income & Spending:
 - Customers with moderate credit card spending are primary targets for loan offers.
- Family & Education:
 - Larger families & higher education levels seem to correlate with higher loan acceptance.
- Online Banking:
 - Over 50% of customers use online banking, indicating a tech-savvy customer base.