Personal Loan Campaign

Alexander Stevenson

Contents

Executive Summary

My primary objective was to develop a predictive model for AllLife Bank to identify liability customers who are most likely to accept a personal loan offer. Through rigorous analysis of customer data, I uncovered key attributes that significantly influence purchasing decisions, enabling the bank to refine its marketing strategies and maximize loan conversions.

Key Findings & Recommendations

My analysis revealed that income, average credit card spending (CCAvg), and holding a CD account (CD_Account) are the strongest predictors of loan acceptance. To enhance conversion rates, I recommend prioritizing marketing efforts toward customers exhibiting these characteristics, as they represent the highest likelihood of engagement.

To mitigate the class imbalance in loan acceptance data, I suggest using oversampling techniques or cost-sensitive learning to prevent the model from being biased toward non-loan takers, thereby improving predictive accuracy.

Model Development & Performance

After testing multiple Decision Tree models, I selected a post-pruned model as the most effective, demonstrating strong performance on unseen data with a recall of 0.925, ensuring more eligible customers are approved, while maintaining a respectable accuracy of 0.962. I prioritized recall over precision because AllLife is focused on rapid expansion, aiming to approve more qualified applicants and increase loan volume. If the primary concern were minimizing loan defaults, precision would have been the priority.

The model achieves an F1-score of 0.823, indicating a strong balance between recall and precision. The model maintains a reasonable precision of 0.741, preventing excessive loan approvals for unqualified applicants.

Business Impact

By implementing this model, AllLife Bank can:

- <u>Enhance loan targeting</u> – Reduce marketing inefficiencies and increase engagement with high-potential customers.
- <u>Increase revenue</u> – Expand its loan customer base while optimizing risk exposure.

- Leverage data-driven strategies – Improve long-term profitability through informed decision-making.

Deploying this model will allow AllLife Bank to confidently identify and target potential loan customers, significantly improving the efficiency and success rate of its marketing campaigns. By leveraging these insights, the bank can drive sustainable growth and competitive advantage in the personal loan market.

Business Problem Overview and Solution Approach

- **Context**

AllLife Bank is a US bank that has a growing customer base. The majority of these customers are liability customers (depositors) with varying sizes of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio.

As a Data Scientist at AllLife Bank, I must build a model that will help the marketing department to identify the potential customers who have a higher probability of purchasing the loan.

- **Objective**

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

- **Solution Approach / Methodology**

1. Problem Definition and Exploration:

    - Clearly define the problem and understand the objectives, such as identifying potential customers interested in personal loans and understanding significant customer attributes.

    - Perform Exploratory Data Analysis (EDA) to gain insights into the data, understand the distribution, and identify patterns among variables.

2. Data Pre-processing:

    - Handle missing values, if any, to ensure completeness of the data.

    - Detect and treat outliers, as they can significantly affect the model's performance.

    - Conduct feature engineering where necessary to create new features that may contribute to better model predictions.

    - Encode categorical variables and normalize numerical features if required.

3. Model Building:

    - Choose Decision Tree as the primary model due to its interpretability and ability to capture non-linear relationships.

    - Define the evaluation criteria, such as accuracy, precision, recall, and the F1-score, to gauge the model's performance.

4. Model Evaluation and Optimization:

    - Evaluate the initial model performance and iterate to improve it through techniques such as pruning (both pre and post-pruning) to prevent overfitting and enhance generalization.

    - Compare different models based on the defined evaluation criteria to select the best-performing model.
    - Extract decision rules and determine feature importance to understand which features play crucial roles in prediction.

5. Insights and Recommendations:

    - Summarize key insights from the model, such as which customer attributes are most indicative of purchasing a loan.

    - Provide actionable recommendations for the marketing team on targeting specific customer segments to increase campaign success.

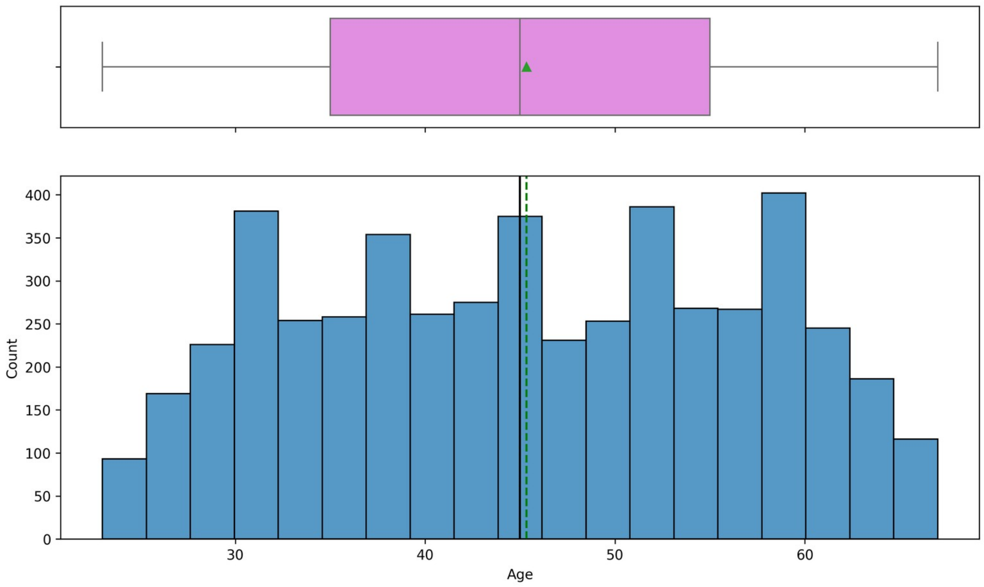
EDA Results

- **Data points**

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?
Securities_Account: Does the customer have a 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)?

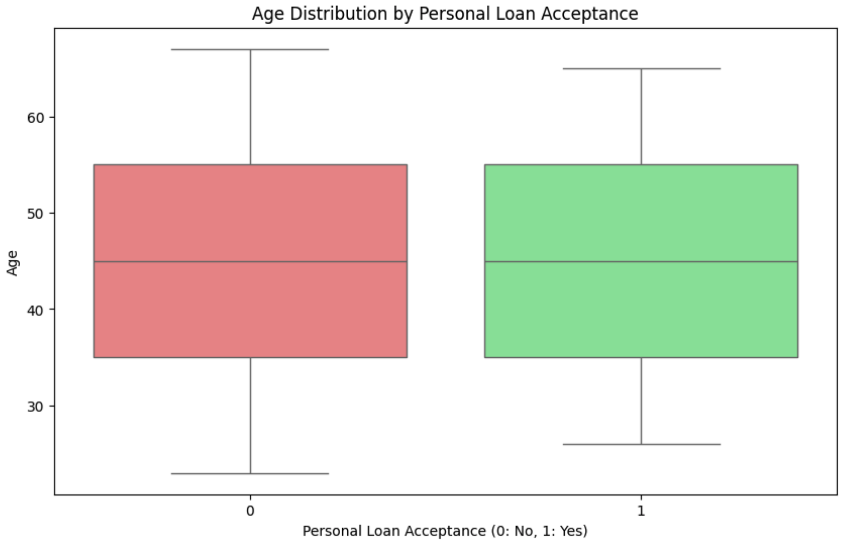
- **Key results from EDA**

1. Customer Profile Analysis:
    - Age:



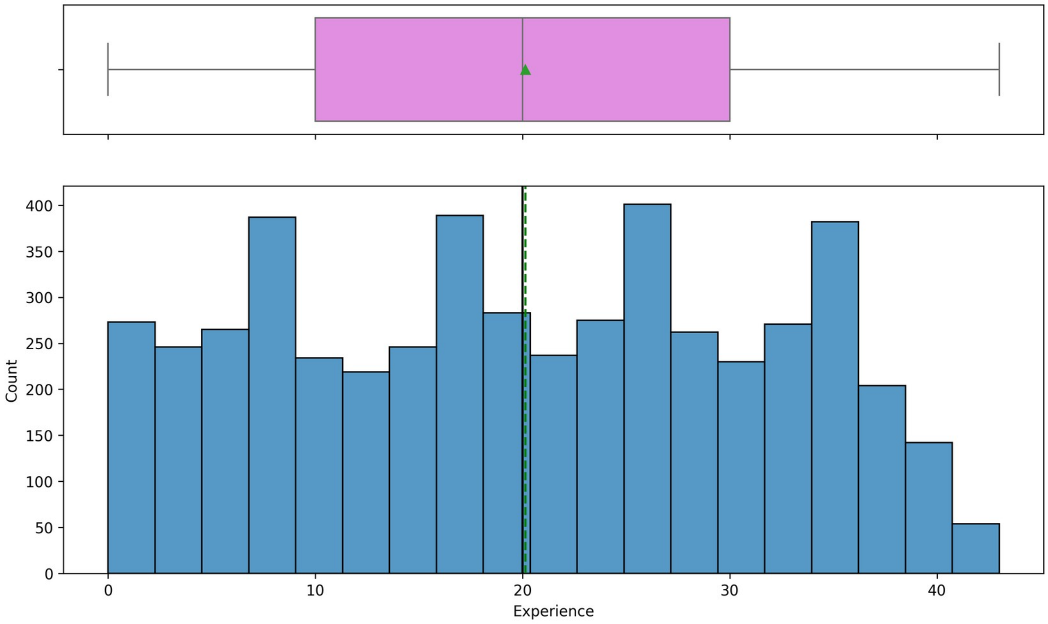
Minimum age: 23
Maximum age: 67
Mean age: 45.3384
Median age: 45.0

The correlation coefficient between age and personal loan acceptance is -0.0077. This value indicates a very weak, almost negligible, negative correlation. This suggests that a customer's education level has little to no influence on their decision to accept a loan.



A box plot was used to visualize the relationship between age and personal loan acceptance (above). The plot suggests a potential trend where customers who accepted personal loans tend to be only very slightly older on average compared to those who did not. However, this is negligible, and there is significant overlap in age ranges between the two groups, indicating that age alone is not a definitive predictor of loan acceptance.

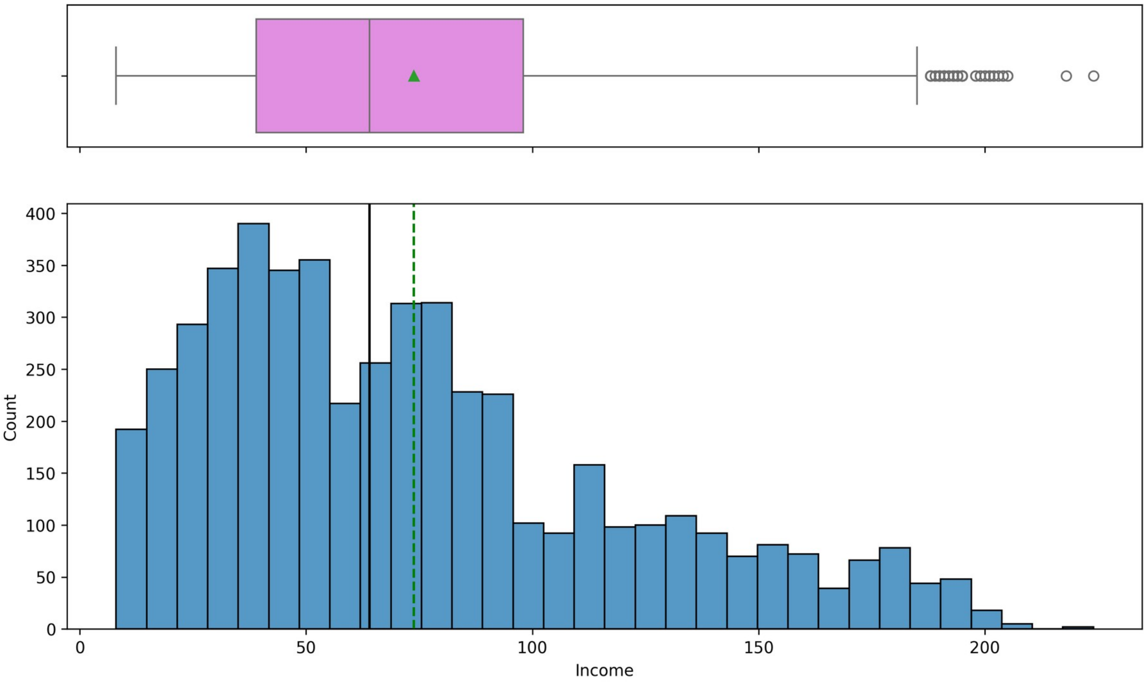
- <u>Experience</u>:



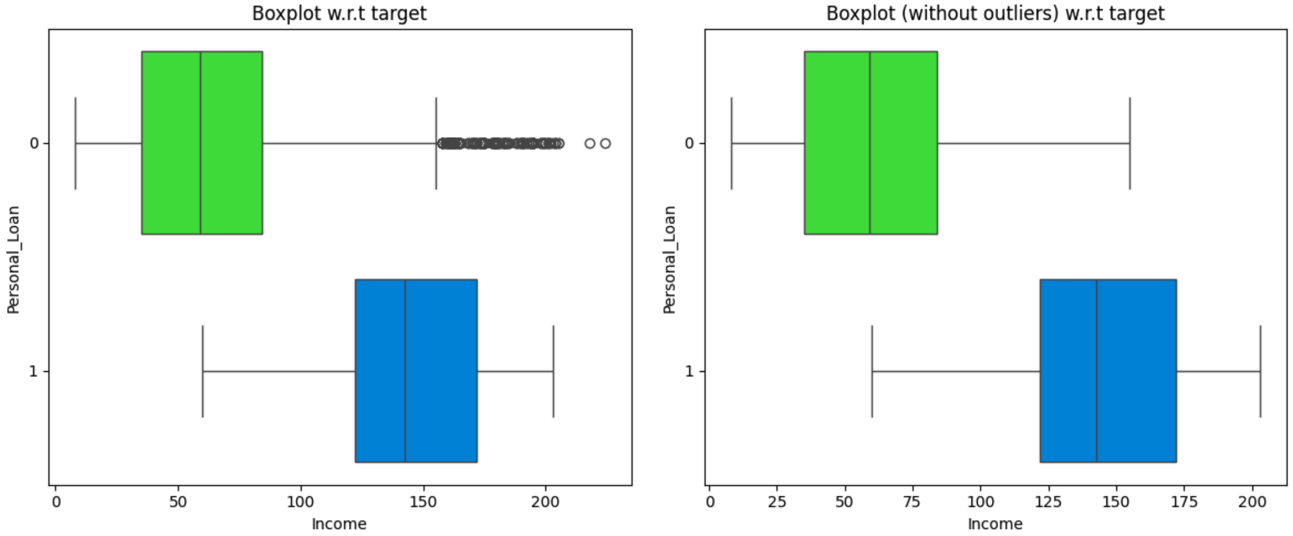
Minimum experience: 0
Maximum experience: 43
Mean experience: 20.1346
Median experience: 20.0

The correlation analysis revealed a very weak negative correlation (-0.0083) between professional experience and personal loan acceptance. This indicates that there is practically no linear relationship between these two variables. A customer's years of experience appears to have little to no influence on their likelihood of accepting a personal loan offer. Other factors, such as income, credit history, or demographic attributes, likely play a more dominant role in driving loan acceptance decisions.

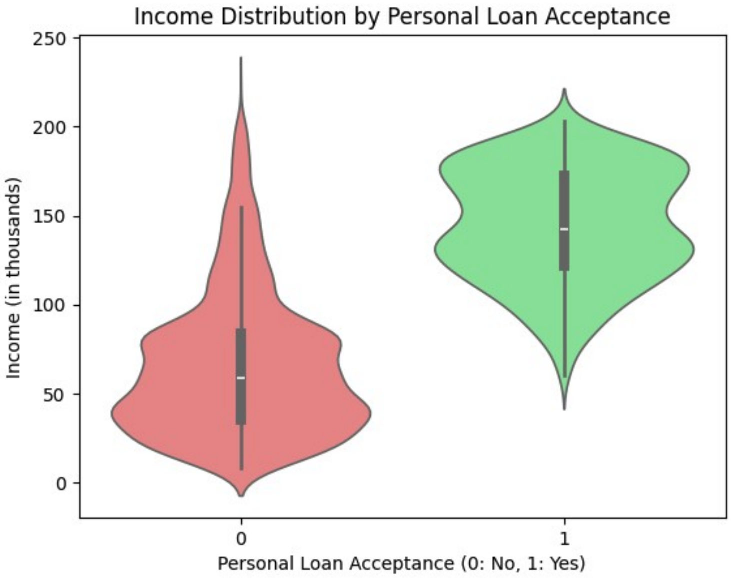
- Income:



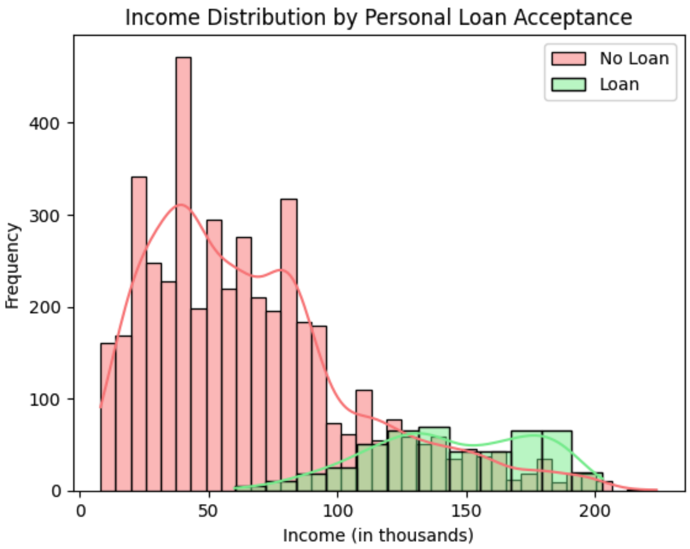
Minimum income: 8
Maximum income: 224
Mean income: 73.7742
Median income: 64.0



We can observe (above) the correlation between income and acceptance of the loan from the previous campaign. As we can see, the income for those who did not accept the personal loan (0) was much less than those who did accept the personal loan (1). This is true irrespective of outliers being present.



A



B

The analysis revealed a moderate positive correlation (0.5025) between income and personal loan acceptance. This indicates that as a customer's income increases, their likelihood of accepting a personal loan also tends to increase.

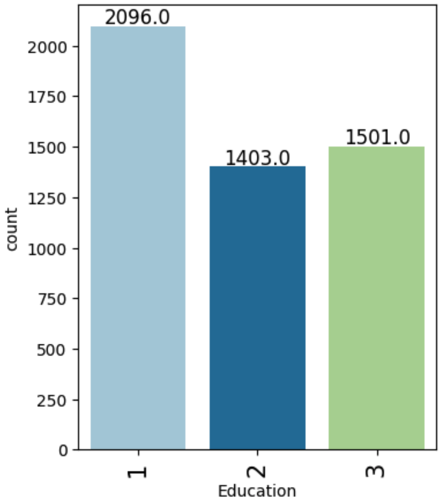
This relationship is further illustrated by the violin plot (Figure A) and overlaid histograms (Figure B). The violin plot shows a clear shift in the distribution of income for customers who accepted personal loans, with a higher density of individuals in the upper income ranges compared to those who did not. The overlaid histograms provide a similar perspective, revealing that a larger proportion of loan acceptors are concentrated in the higher income brackets.

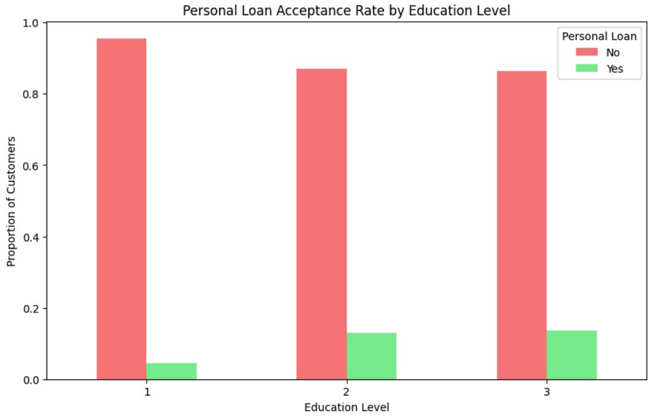
These findings suggest that income is a significant factor in influencing personal loan acceptance decisions. Customers with higher incomes appear to be more inclined towards taking out personal loans, potentially due to greater financial capacity or a higher propensity for borrowing.

- <u>Education Level</u>:

Education

1	42 %
3	30 %
2	28 %



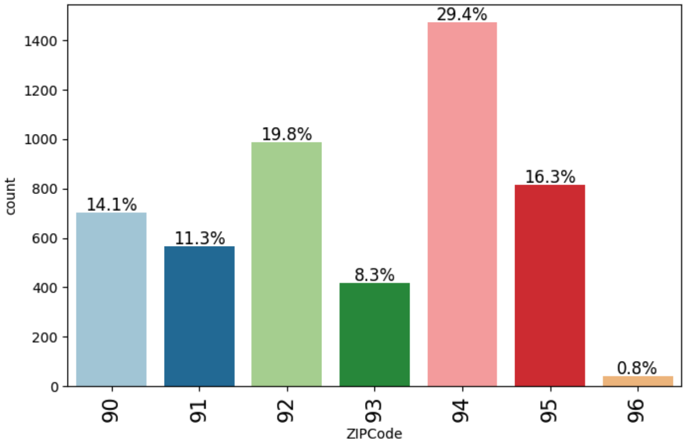


Analysis reveals a weak positive correlation (0.1367) between a customer's education level and their likelihood of accepting a personal loan. This suggests that, while there is a tendency for customers with higher education to be slightly more inclined towards personal loans, the relationship is not strong enough to be a primary driving factor.

- Zip Code:

Chi-Square Statistic: 0.6 p-value: 0.9964

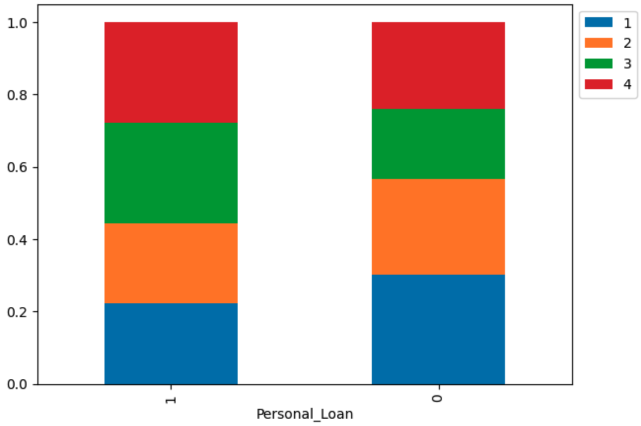
The results of the Chi-Square test indicate that there is no statistically significant relationship between ZIP code and personal loan acceptance. The very small Chi-Square statistic and the high p-value suggest that any observed differences in personal loan acceptance rates across various ZIP codes are likely due to random variation, rather than indicating a true underlying association between these two variables.



- Family:

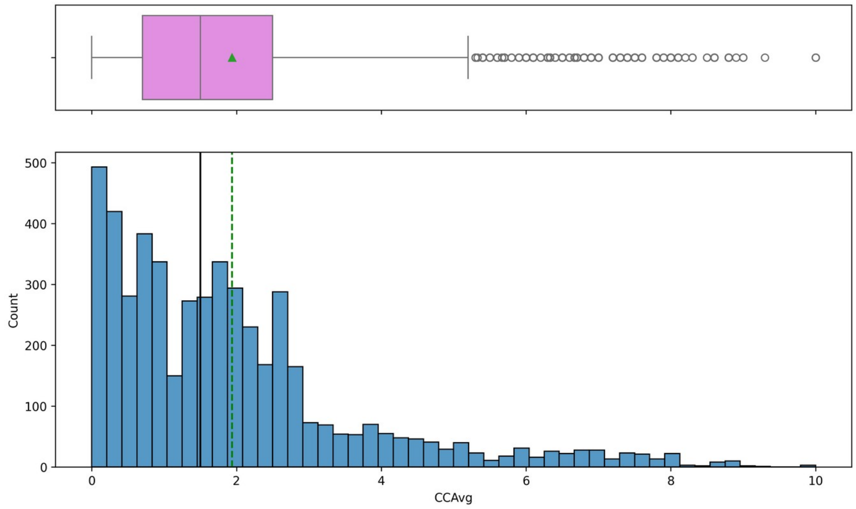
The correlation between Family size and Personal Loan acceptance is 0.061, indicating a negligible correlation between these two variables. This suggests that family size has little to no impact on whether a customer is likely to accept a personal loan. Based on this information, the bank's marketing department would be better off focusing its efforts on other variables with stronger predictive relationships, rather than allocating resources to target customers based on family size when promoting loan products.

1	0.29 %
2	0.26 %
4	0.24 %
3	0.20 %

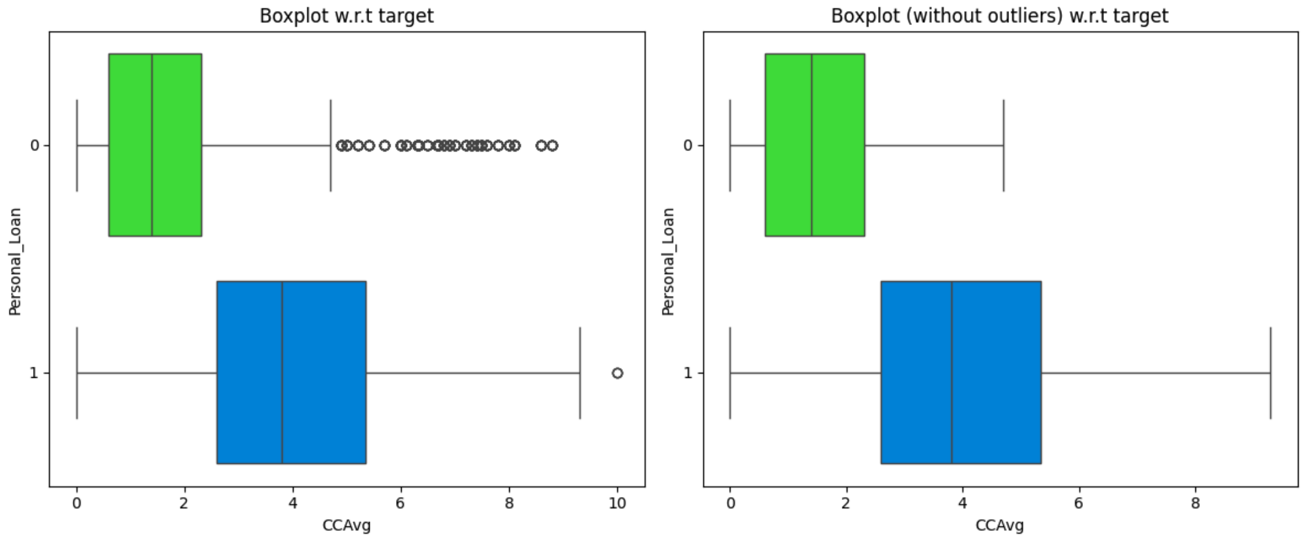


2. Spending and Financial Behavior:

- CCAvg (Credit Card Average Spend):



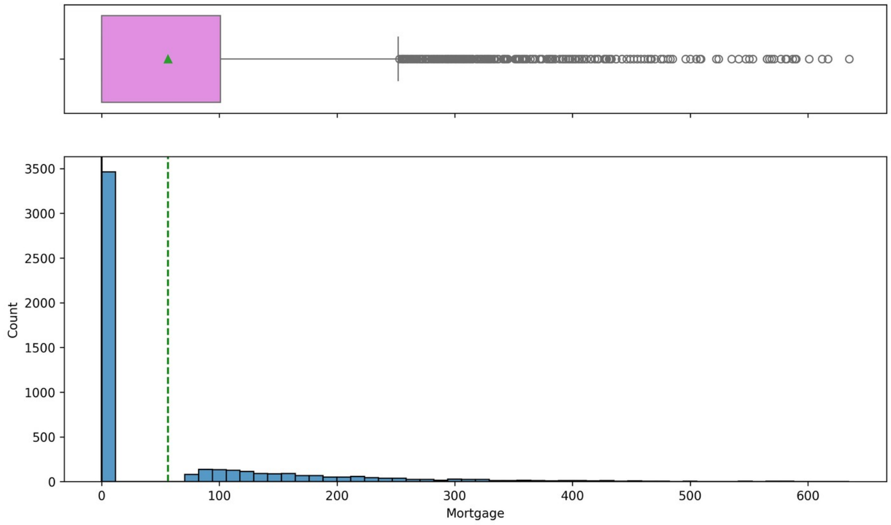
Minimum CCAvg: 0.0
Maximum CCAvg: 10.0
Mean CCAvg: 1.94
Median CCAvg: 1.5



A correlation of 0.37 between average credit card spending (CCAvg) and personal loan acceptance suggests a moderate positive relationship. Customers who tend to spend more on their credit cards are also somewhat more likely to be interested in personal loans. While not a very strong relationship, it's still a noticeable trend that the marketing department could leverage for targeted campaigns. Focusing on customers with higher CCAvg *could* increase the success rate of personal loan offers, although we must remember that correlation does not equal causation and other factors might be at play.

- Mortgage:

Minimum mortgage: 0
Maximum mortgage: 635
Mean mortgage: 56.5
Median mortgage: 0.0



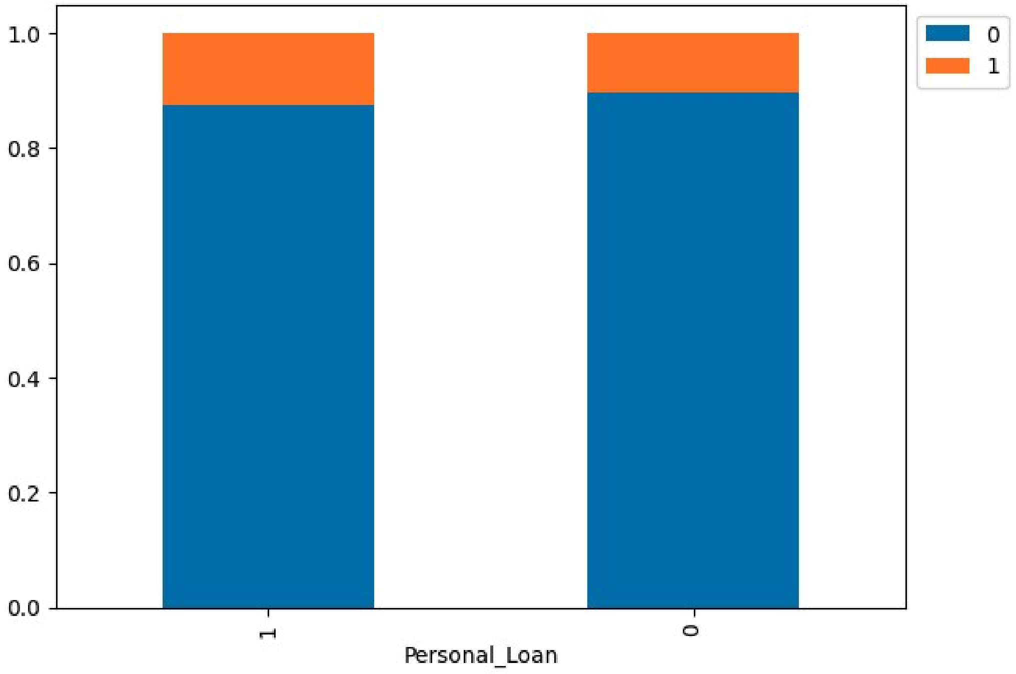
The correlation of 0.14 between mortgage value and personal loan acceptance indicates a weak positive relationship, suggesting that customers with higher mortgages are only *slightly* more likely to accept personal loans. This weak correlation implies that mortgage value is not a strong predictor of loan acceptance and should not be a primary factor for targeted marketing campaigns.

3. Banking Behavior:

   - Securities and CD Accounts:

Personal_Loan vs Securities_Account

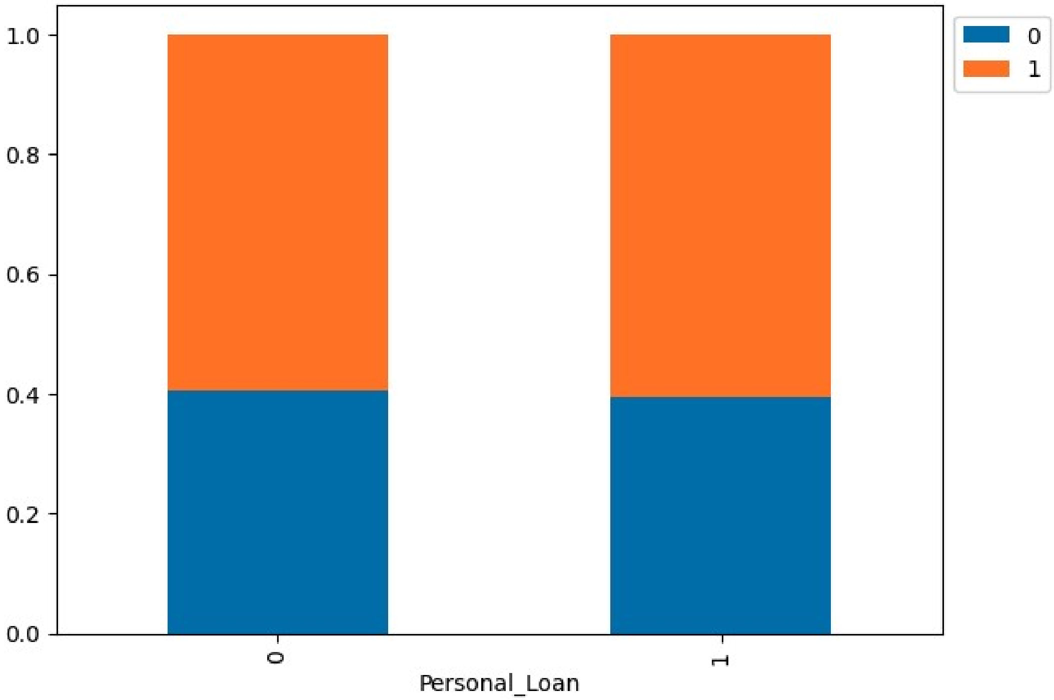
```
Securities_Account

0    4478
1     522

Securities_Account

0    89.56 %
1    10.44 %
```



The point-biserial correlation between Personal_Loan and Securities_Account is 0.02, indicating a very weak positive relationship. This suggests that customers with a securities account are only *slightly* more likely to accept a personal loan, but this relationship is almost negligible. The p-value of 0.12 is greater than the common significance level of 0.05, meaning this correlation is not statistically significant. In practical terms, having a securities account doesn't appear to be a strong predictor of whether a customer will accept a personal loan offer.

Personal_Loan vs CD_Account

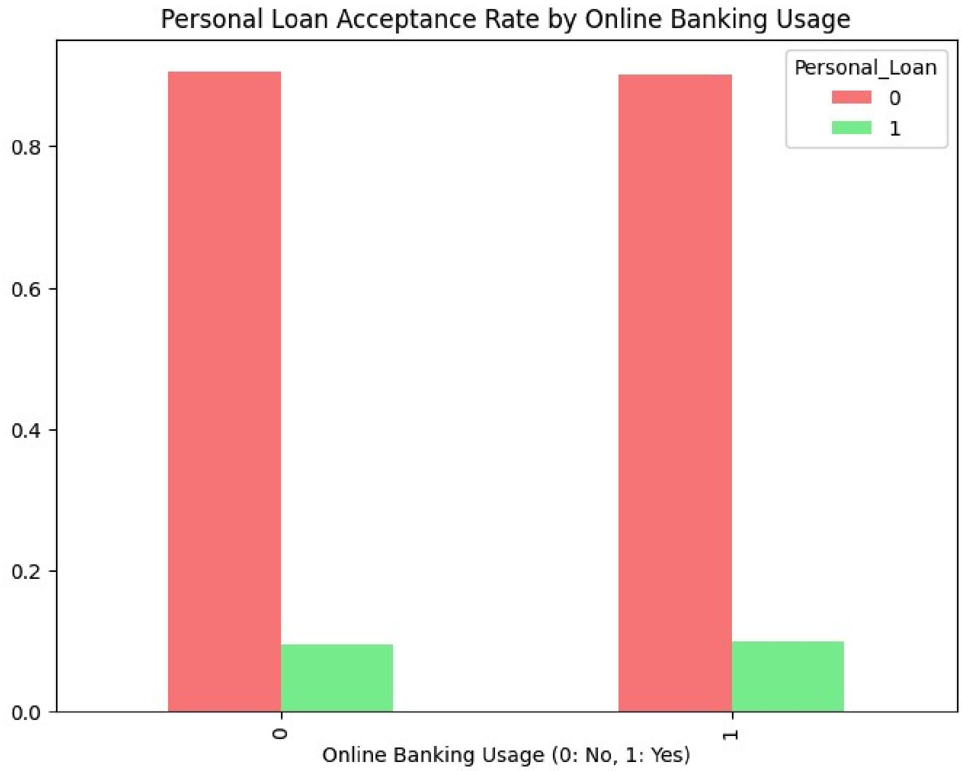
```
CD_Account
0    4698
1     302

CD_Account
0    94 %
1    6 %
```



The point-biserial correlation between Personal_Loan and CD_Account is 0.32, indicating a moderate positive relationship. This suggests that customers with a certificate of deposit (CD) account are more likely to accept a personal loan compared to those without a CD account. The extremely small p-value (1.27e-116) strongly indicates that this correlation is statistically significant. This means that the relationship between having a CD account and accepting a personal loan is likely not due to random chance. For the bank, this suggests that customers with CD accounts are a promising target group for personal loan campaigns, as they have a higher probability of accepting such offers.

- Online Banking and Credit Card Ownership:

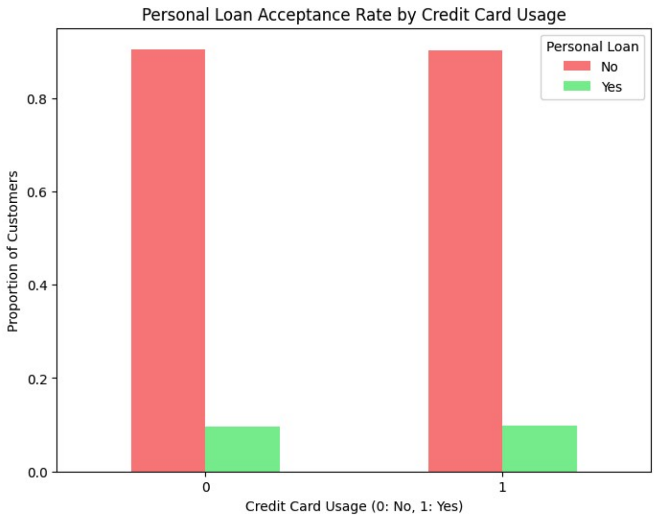
Online Banking Usage



The analysis indicates a weak and statistically insignificant correlation between online banking usage (Online) and personal loan acceptance (Personal_Loan). The calculated point-biserial correlation coefficient is 0.00627, suggesting a negligible linear relationship. The high p-value of 0.657 further supports this conclusion.

The accompanying bar plot (above) visually reinforces this lack of strong association, showing similar proportions of personal loan acceptance among online banking users and non-users. Therefore, while online banking is a popular service, it does not appear to be a significant factor in predicting a customer's likelihood to accept a personal loan offer from AllLife Bank.

Credit Card



The analysis reveals a very weak and statistically insignificant correlation between credit card usage (CreditCard) and personal loan acceptance (Personal_Loan). The calculated point-biserial correlation coefficient is only 0.0028, indicating a negligible linear relationship between these two variables. This finding is further supported by the high p-value of 0.843, which is much greater than the typical significance level of 0.05.

The accompanying bar plot (above) visually confirms this lack of association. It shows that the proportion of customers accepting or declining personal loans is nearly identical for both credit card users and non-users. Therefore, we can conclude that a customer's use of a credit card issued by another bank does not appear to be a strong predictor of their likelihood to accept a personal loan offer from AllLife Bank.

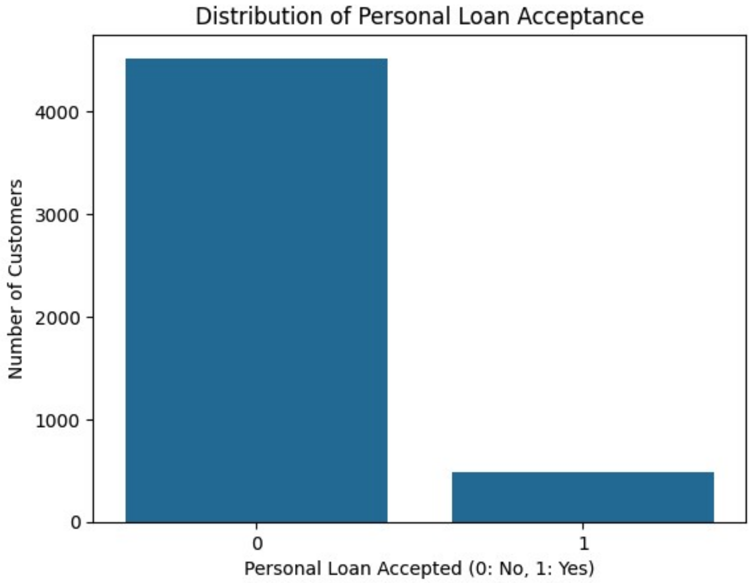
Further Considerations

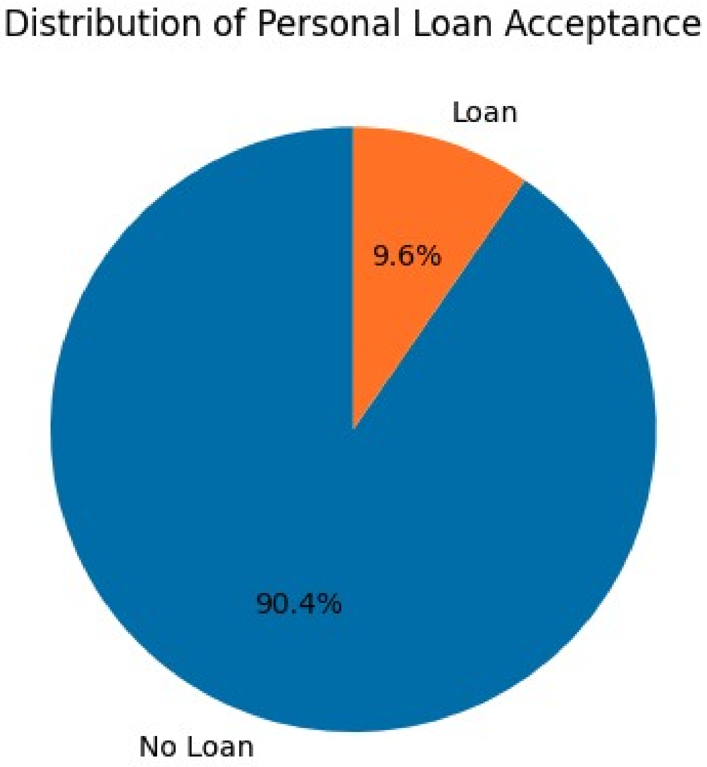
Correlation vs. Causation: While the crosstab (above) shows a relationship, it doesn't necessarily mean family size *causes* loan acceptance. Other factors could be influencing both family size and loan decisions.

Statistical Significance: To determine if these observed differences are statistically significant, you could perform a chi-squared test of independence. This test would assess whether the relationship between family size and loan acceptance is likely due to chance or a real association.

4. Target Variable Distribution:

- Personal Loan Acceptance:





The target variable, Personal_Loan, exhibits a significant class imbalance. Out of 5000 total customers, a large majority (4520 or 90.4%) did not accept the personal loan offer, while a relatively small portion (480 or 9.6%) accepted it. This imbalance suggests that during model building, techniques like oversampling, undersampling, or cost-sensitive learning might be necessary to prevent the model from being biased towards the majority class (loan declined) and to improve its ability to predict loan acceptance accurately.

5. Correlations and Relationships:

- Correlations between various features and the target variable used to identify potential predictors of loan acceptance.

During my exploratory data analysis, I investigated correlations between different features and the target variable, Personal_Loan, to identify potential predictors of loan acceptance. I observed the strongest positive correlations with Income, CCAvg (average credit card spending), and CD_Account (certificate of deposit account). This suggests that customers with higher incomes, higher credit card spending, and those who already hold a CD account with the bank are more likely to accept a personal loan offer. Conversely, features like Securities_Account, Online, and CreditCard showed weak and statistically insignificant correlations with Personal_Loan, indicating they might not be strong predictors.

- Interesting bivariate relationships that could suggest influential factors for the model.

Further investigation into bivariate relationships revealed interesting patterns. For instance, the distribution of income for loan acceptors was notably skewed towards higher values compared to those who declined the loan, supporting the strong positive correlation. Similarly, customers with higher education levels exhibited a greater tendency to accept personal loans. Interestingly, while family size showed a slight positive correlation, its impact might not be as prominent as income or credit card spending. Overall, my analysis suggests that financial factors, particularly income and credit card spending, play a crucial role in influencing loan acceptance, while demographic factors like education level might have a secondary influence. Features related to other product holdings or online banking usage seem to have a minimal impact on loan acceptance decisions.

Data Preprocessing

- Duplicate value check
  - Zero duplicates found

- Missing value treatment
  - No missing values observed, so no treatment necessary

- Outlier check (treatment if needed)
  - Number of outliers in each column:

Age	0
Experience	0
Income	96
Family	0
CCAvg	324
Mortgage	291

  - Outlier percentage in each column:

Age	0.00
Experience	0.00
Income	1.92
Family	0.00
CCAvg	6.48
Mortgage	5.82

  - No treatment of outliers is necessary. Since the outliers are within a reasonable range and likely represent real high-income or high-spending individuals, who are relevant to loan acceptance, they do not need to be removed or adjusted. We evaluated the data in its original state, other than the feature engineering describe next.

- Feature Engineering
    - I employ several feature engineering techniques to enhance the dataset for modeling. Firstly, I transform the ZIP Code column by extracting the first two digits, effectively grouping similar zip codes into broader geographical regions and reducing dimensionality. Secondly, I convert several columns, including the modified ZIPCode, to the 'category' data type, ensuring Pandas and other libraries handle them appropriately as categorical variables, which could improve model performance. Lastly, I address data quality by replacing negative values in the Experience column with more plausible values, correcting potential data errors.
    - These feature engineering steps aim to improve the dataset for modeling by reducing dimensionality, enhancing model compatibility with categorical features, and improving data quality. These techniques aim to enhance the dataset's suitability for model training and ultimately improve the predictive power of the model being built.

- Data preprocessing for modeling
    - Beyond feature engineering, I undertake crucial data preprocessing steps to prepare the dataset for machine learning. I apply one-hot encoding to convert categorical features (ZIPCode, Education) into numerical representations, making them suitable for algorithms. The entire feature set is then converted to a consistent 'float' data type to facilitate better training.
    - The data is split into training and testing sets using train_test_split, allowing for model evaluation on unseen data. To prevent redundancy and potential overfitting, the highly correlated Experience column is dropped. Finally, the Personal_Loan column is designated as the target variable, defining the model's prediction objective.
    - These preprocessing steps are crucial for ensuring data compatibility with algorithms, enabling robust model evaluation, and preventing issues like overfitting. By carefully preparing the data in this way, the notebook establishes a solid foundation for building effective and reliable machine learning models to predict personal loan acceptance.

Model Building

- Model building steps of Decision Tree

I followed a structured approach to build a Decision Tree model. First, I carefully prepared the data through feature engineering and preprocessing, ensuring it is suitable for the algorithm. Next, a Decision Tree classifier is initialized, with specific parameters for splitting criteria and reproducibility (criterion="gini", random_state=1).

The model is then trained using the training data, where it learns patterns between features and the target variable. Initial performance is evaluated using metrics like accuracy and a confusion matrix on the training data. Additionally, the tree structure and confusion matrix were visualized to understand the decision-making process and performance on training and data. Crucially, the model's generalization ability is assessed by evaluating it on unseen test data, using the same performance metrics.

To enhance performance and prevent overfitting, hyperparameter optimization (HPO) / tuning was employed through techniques like pre-pruning or post-pruning. Finally, the best-performing model was selected based on the evaluation results, ready for deployment and prediction on new data. This structured approach ensures a robust and reliable Decision Tree model for predicting personal loan acceptance.

Model Performance Summary

- Model evaluation criterion

My primary focus during model evaluation was recall. This is because the business goal was to identify as many potential loan customers as possible, even if it meant a slightly higher risk of false positives. Maximizing the number of truly interested customers targeted was the priority.

- Overview of the final decision tree model and its parameters

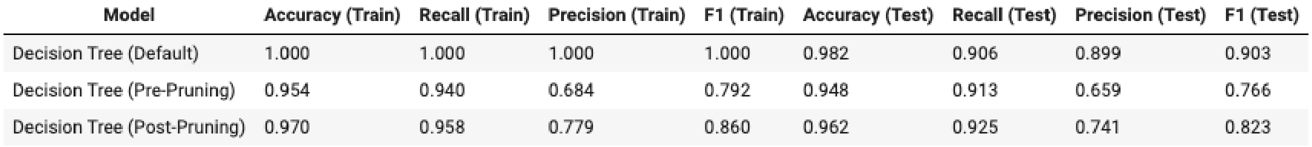
After building and evaluating different Decision Tree models, I selected the post-pruned Decision Tree as the final model. This model used a max_depth of 6, max_leaf_nodes of

50, and min_samples_split of 10. I believe this configuration provides a good balance between model complexity and generalization performance.

- Summary of most important features used by the decision tree model for prediction

The model highlighted Income, CCAvg (average credit card spending), and CD_Account (certificate of deposit account) as the most influential features for predicting loan acceptance. These features appear to be key factors in customer decision-making regarding personal loans.

- Summary of key performance metrics for training and test data of all the models in tabular format for comparison



Model	Accuracy (Train)	Recall (Train)	Precision (Train)	F1 (Train)	Accuracy (Test)	Recall (Test)	Precision (Test)	F1 (Test)
Decision Tree (Default)	1.000	1.000	1.000	1.000	0.982	0.906	0.899	0.903
Decision Tree (Pre-Pruning)	0.954	0.940	0.684	0.792	0.948	0.913	0.659	0.766
Decision Tree (Post-Pruning)	0.970	0.958	0.779	0.860	0.962	0.925	0.741	0.823

The above table compares the performance of three Decision Tree models: Default, Pre-pruned, and Post-pruned. The Default model, with no parameter tuning, achieved perfect accuracy and recall on the training data but slightly lower scores on the test data (0.982 accuracy, 0.906 recall). This suggests that the Default model is likely overfitting

The Pre-pruned model showed a reduced performance on both training and test sets, particularly in precision, indicating a potential for overfitting mitigation. The Post-pruned model demonstrated a balance between training and test performance, with a slight decrease in training accuracy and recall but an improved recall on the test set (0.925), aligning with the focus on maximizing true positive predictions. Overall, the Post-pruned model appears to offer the best generalization and alignment with the business objective.

Model Performance Improvement

Improvement in model performance with pruning: By experimenting with different pruning techniques, I was able to significantly improve the performance of the Decision Tree model. Pre-pruning helped in reducing overfitting, but it came at a cost – a slight dip in recall on the test set. This trade-off led me to explore post-pruning, which proved more effective. By carefully adjusting the cost-complexity parameter, I fine-tuned the model to achieve a better balance between training and test performance. Importantly, post-pruning resulted in a higher recall on the test set (0.925) while maintaining a respectable accuracy of 0.962.

Decision rules and feature importance: A closer look at the decision rules of the post-pruned model highlighted Income, CCAvg, and CD_Account as the most influential factors for predicting loan acceptance. Customers with high income and credit card spending were more likely to accept personal loan offers, which makes intuitive sense. The feature importance analysis confirmed this observation, with Income being the most crucial predictor, followed by CCAvg and CD_Account. This analysis offered valuable insights into the key drivers of loan acceptance among AllLife Bank's customer base.

Data Background and Contents

The dataset used for this project is provided by AllLife Bank, a US-based financial institution. Their goal is to expand their personal loan customer base by converting existing liability customers (depositors) into borrowers. The data represents a sample of the bank's customers and includes information about their demographics, financial behavior, and product holdings.

Contents

The dataset contains the following variables:

Customer Demographics: Age, Experience, Income, ZIP Code, Family
Financial Behavior: CCAvg (average credit card spending), Mortgage
Product Holdings: Securities_Account, CD_Account, Online, CreditCard
Target Variable: Personal_Loan (indicates whether the customer accepted a personal loan offer)

Types of Variables

The dataset comprises a mix of numerical and categorical variables:

Numerical: Age, Experience, Income, CCAvg, Mortgage
Categorical: ZIP Code, Family, Education, Securities_Account, CD_Account, Online, CreditCard, Personal_Loan

Data Structure and Organization

The data is organized in a tabular format, with each row representing a customer and each column representing a specific attribute. The dataset initially included a unique customer ID, which was removed for modeling purposes.