AllLife Bank Loan Marketing Model

Project #2- Machine Learning

Wednesday, August 7, 2024

Presented by: Sarah Lasater

Contents / Agenda

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

Executive Summary

OBSERVATIONS

Customer Features:

- Age: 50% of customers are between 35 and 54 years old, with a median age of 45.
- Professional Experience: 50% of customers have between 10 and 30 years of professional experience, averaging 20 years.
- Income: 50% of customers earn between $30,000 and $95,000, with a median income of $64,000 and an average of $73,700.
- Credit Card Spending: Customers with a personal loan spend between $2,500 and $5,500 monthly on average, whereas those without a loan spend between $500 and $2,250.
- Mortgage Values: 50% of mortgages range from $0 to $100,000, with an average mortgage of $56,500.
- Family Size: Family sizes of 1 (29.4%) and 2 (25.9%) are the most common among customers.
- Education Levels: Highest number of customers have education level 1 (2,096), followed by level 3 (1,501) and level 2 (1,403).
- Securities Account: 522 customers have a securities account, while 4,478 do not.
- CD Accounts: 140 personal loan holders have CD accounts (29%), compared to 162 of 4,520 non-loan holders (3%).
- Online Banking: 60% of personal loan holders use online banking, compared to 40.4% of non-loan holders.
- Credit Cards from Other Banks: 29.7% of personal loan holders have credit cards issued by other banks.
- ZIP Codes: Distribution of ZIP codes is similar between loan holders and non-holders.

CONTINUED

Executive Summary

DECISION TREE MODEL ANALYSIS

Model Performance:

- ***Sklearn Decision Tree:*** Shows an 8% better performance on training data compared to test data, indicating higher overfitting. Utilizes more features, contributing to increased complexity and performance disparity.

- ***Pre-Pruned Model:*** Demonstrates a 5.1% better performance on training data compared to test data. Achieves balanced F1 scores on both training (.93) and test sets (.879). Uses fewer features (six) which aids in reducing complexity and potentially enhances generalization.

- ***Post-Pruned Model:*** Underperforms with only 1.4% better performance on training data compared to test data. Utilizes only one feature (income), which is insufficient for effective decision-making.

CONTINUED

Executive Summary

RECOMMENDATIONS

1. Adopt the Pre-Pruned Decision Tree Model:
    - The pre-pruned model, despite signs of overfitting, demonstrates a more balanced performance and better generalization than the sklearn decision tree. Its use of fewer features reduces complexity, potentially leading to improved performance on unseen data and shorter prediction times.
2. Feature Selection:
    - Focus on the six key features used in the pre-pruned model: Income, Family, Education, CCAvg, and Age. This streamlined feature set is efficient and effective, balancing complexity with performance.
3. Improve Data Utilization:
    - Given the positive correlations between income and personal loans, as well as credit card spending, prioritize these factors in marketing strategies. *Higher-income customers and those with higher credit card expenditures are more likely to engage with personal loan offers.*
4. Targeted Marketing Strategies:
    - Tailor marketing efforts to *customers with higher incomes and those who actively use online banking services. These segments show a higher likelihood of taking personal loans and engaging with the bank's services.*
5. Continuous Model Monitoring:
    - Regularly evaluate and update the model to adapt to changing customer behaviors and improve its predictive accuracy.

The pre-pruned decision tree model is recommended for its balance of performance and efficiency, making it suitable for optimizing AllLife Bank's loan marketing efforts.

Business Problem Overview and Solution Approach

PROBLEM DEFINITION

AllLife Bank, a US bank that has a growing customer base that are mostly liability customers (depositors) with varying sizes of deposits, is interested in expanding their borrowers (asset customers) base to bring in more loan business and in the process, earn more through the interest on loans. Management wants to explore ways of converting its current liability customers to personal loan customers.

Business Problem Overview and Solution Approach

SOLUTION APPROACH / METHODOLOGY

The objective is to build a model that will help the marketing department to identify the potential customers who have a higher probability of purchasing the loan to then devise campaigns with better target marketing to increase the previous success ratio of 9%. We will utilize the Decision Tree methodology to develop, train, and test models for their efficacy in predicting which current AllLife Bank liability customers would be most likely to engage in a targeted marketing campaign to convert them into personal loan holders.

EDA Results- OBSERVATIONS (Univariate)

Age (SEE U1)	• 50% of customers fall between about 35 and about 54 years of age, with the median age being about 45 and the average being slightly higher at around 45.3 years of age.
Experience (SEE U2)	• 50% of customers have between 10 and 30 years of professional experience with the average falling right at 20 years of professional experience.
Income (SEE U3)	• 50% of customers' income falls between $39,000 and $98,000 with the median being about $64,000 and average income being around $73,700.
Monthly Credit Card Usage (SEE U4)	• 50% of customers spend an average of $700 to $2,500 per month on credit card usage with the average expenditure being $1,930 and median falling at $1,500. However, there are also many outlier customers spending between $5,000 to $10,000 per month.
Mortgage (SEE U5)	• 50% of customer mortgages fall between $0 - 101,000 with the average mortgage falling around $56,500. There are however also many customer outliers with mortgages falling between $250,000 and the maximum of $635,000.

CONTINUED

EDA Results- OBSERVATIONS (Univariate)

Family Size (SEE U6)	• With only options 1 through 4 available, family sizes of 1 rank highest at 29.4%, 2 in second at 25.9%, 4 in third at 24.4%, and 3 in fourth at 20.2%.
Education (SEE U7)	• It can be observed that the highest number of customers at 2,096 fall into the “Undergrad” education category, the second highest number at 1,501 fall into the “Advanced/ Professional” category, and the lowest number at 1,403 have an education rank of “Graduate”.
Securities Account (SEE U8)	• 4,478 bank customers do not have a securities account and 522 customers do have one.
CD Account (SEE U9)	• 4,698 bank customers do not have a securities account and 302 customers do have one.

CONTINUED

EDA Results- OBSERVATIONS (Univariate)

Online Banking (SEE U10)	• 2,016 customers do not use online banking and 2,984 customers do use it.
Credit Card (SEE U11)	• 3,530 customers do not have another credit card issued by any other bank and 1,470 customers do have one.

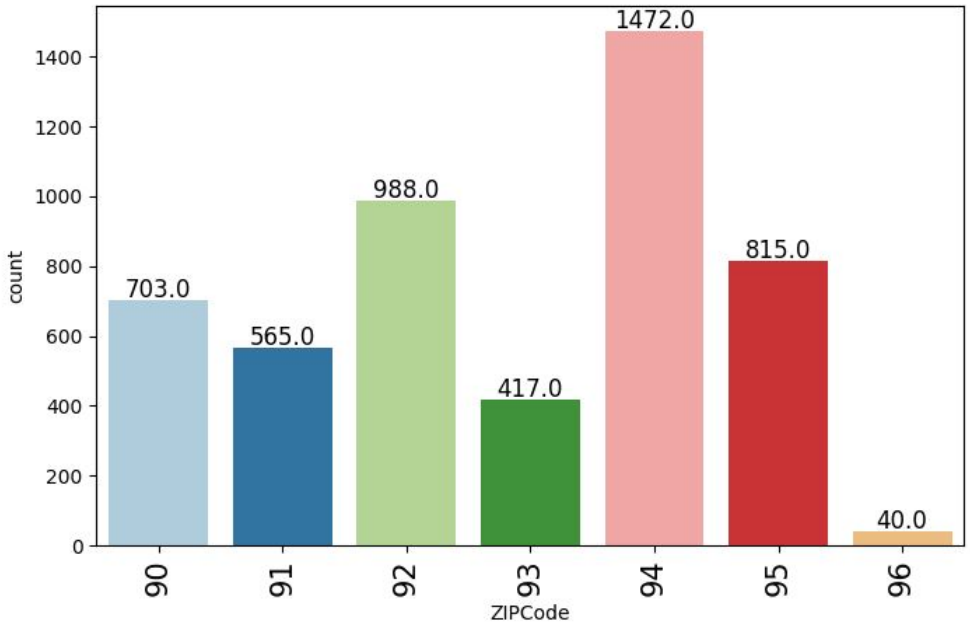
CONTINUED

EDA Results- OBSERVATIONS (Univariate)

ZIP Code

The barplot data indicates that the 7 unique zip codes rank from highest number of customers to lowest number of customers as follows:

1. 94: 1,472 customers
2. 92: 988 customers
3. 95: 815 customers
4. 90: 703 customers
5. 91: 565 customers
6. 93: 417 customers
7. 96: 40 customers



CONTINUED

EDA Results- OBSERVATIONS

CORRELATION

Features with highest positive correlation:

1. Age, Experience: 0.99
2. Income, CCAvg: 0.65
3. Income, Personal_Loan: 0.50

Features with highest negative correlation:

1. Income, Education: -0.19
2. Income, Family: -0.16
3. Education, CCAvg: -0.14

EDA Results- OBSERVATIONS (Bivariate)

Personal Loan vs Education (SEE B1)	• Customers with an education level “Undergrad” (2,096 customers): No loan= 2,003 / Yes loan= 93 • Customers with an education level “Graduate” (1,403 customers): No loan= 1,221 / Yes loan= 182 • Customers with an education level “Advanced/ Professional” (1,501 customers): No loan= 1,296 / Yes loan= 205
Personal Loan vs Family *(SEE B2)*	• The category of personal loan holders that has the highest count of loans is the family category of 4 with 134, with category 3 having just one less loan holder at 133. • The family category that has the highest instance of not taking out a personal loan is category 1 with a count of 1,365 customers, with category 2 having a count of 1,190.
Personal Loan vs Securities Account *(SEE B3)*	• The barplot indicates that of the 480 bank customers with a personal loan, 60 of them also have a securities account. This is 12.5% of the personal loan holders. • The barplot also indicates that of the 4,520 customers that do not have a personal loan, 462 of them do have a securities account. This is 10.2% of non-loan-holders or a total of 9.2% of the total number of 5,000 bank customers.

CONTINUED

EDA Results- OBSERVATIONS (Bivariate)

Personal Loan vs CD Account *(SEE B4)*	• Of the bank's 480 personal loan holders, 140 of those have a CD account as well. This is a rate of 29%. • Of the bank's 4,520 customers that do not have a personal loan, 162 of them do have a CD account. This is a rate of 3%. • The total number of bank customers with a CD account is 302, which is right at 6%.
Personal Loan vs Online Banking *(SEE B5)*	• Of the bank's 480 personal loan holders, 291 bank online while 189 do not. This is 60% of personal loan holders. • Of the bank's 4,520 customers that do not have a personal loan, 2,693 bank online while 1,827 do not. This is 40.4% of customers that do not have a personal loan. • The total number of 2,016 customers that do not utilize the bank's online platform is 40% of total banking customers.
Personal Loan vs Other Credit Card *(SEE B6)*	• Of the 480 bank customers that have a personal loan, 143 (29.7%) have a credit card issued by another bank. • Of the 4,520 customer base that does not have a loan, 1,327 (29.3%) do have a credit card issued by another bank. • A total of 1,470 customers have a credit card issued by another bank, leaving 3,530 customers that do not have one. 29.4% of all the bank's customers have a credit card issued by another bank.

CONTINUED

EDA Results- OBSERVATIONS (Bivariate)

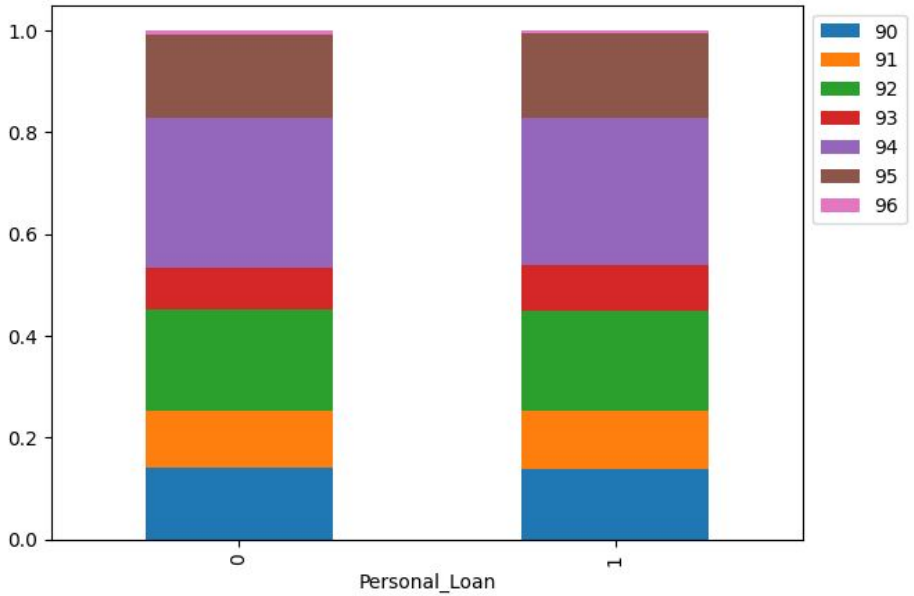
Personal Loan vs ZIP Code

Of the 480 customers that have a personal loan:

1. 94- 138 (28.7%)
2. 92- 94 (19.5%)
3. 95- 80 (16.6%)
4. 90- 67 (13.9%)
5. 91- 55 (11.4%)
6. 93- 43 (8.9%)
7. 96- 3 (0.6%)

Of the 4,520 customers that don't have a personal loan:

1. 94- 1334 (29.5%)
2. 92- 894 (19.7%)
3. 95- 735 (16.2%)
4. 90- 636 (14%)
5. 91- 510 (11.2%)
6. 93- 374 (8.2%)
7. 96- 37 (0.8%)



ZIP Code 94 is the most common in both groups, representing 28.7% of personal loan holders and 29.5% of non-loan holders. ZIP Code 92 and ZIP Code 95 also have a significant presence in both groups, with slightly higher proportions among non-loan holders compared to loan holders. ZIP Code 96 is the least represented in both groups, with only 0.6% of loan holders and 0.8% of non-loan holders.

CONTINUED

EDA Results- OBSERVATIONS (Bivariate)

Personal Loan vs Age (SEE B10)

- Ages 26 - 65 represented on both targets 0 and 1
- Personal loan: 50% of customers between 35-55 with a median age of 45.
- No personal loan: 50% of customers between 35-55 with a median age of 45.

(Without outliers) *(SEE B11)*

- Personal loan: 50% of customers between 35-55 with a median age of 45.
- No personal loan: 50% of customers between 35-55 with a median age of 45.

Personal Loan vs Work Experience (SEE B12)

- Work experience years 0 - 45 represented on distribution plot target 0 and 0-41 or 42 years on distribution plot target 1
- Personal loan: 50% of customers between 9 - 30 years work experience shown on w.r.t target boxplot.
- No personal loan: 50% of customers between between 10 - 30 years work experience shown on w.r.t target boxplot.

(Without outliers) *(SEE B13)*

- Personal loan: 50% of customers between 9 - 30 years work experience shown on w.r.t target boxplot without outliers.
- No personal loan: 50% of customers between between 10 - 30 years work experience shown on w.r.t target boxplot without outliers.

CONTINUED

EDA Results- OBSERVATIONS (Bivariate)

Personal Loan vs Income (SEE B16)

- Personal loan: 50% of customers between approximately $125,000 - approximately $175,000 income level shown on w.r.t target boxplot with a median income level around approximately $140,000.
- No personal loan: 50% of customers between approximately $30,000 - approximately $75,000 income level shown on w.r.t target boxplot with a median at around approximately $55,000 with many outliers above the approximately $155,000 income level.

(Without outliers) *(SEE B17)*

- Personal loan: 50% of customers between approximately $120,000 - approximately $170,000 shown on w.r.t target boxplot with a median income level around approximately $140,000 without outliers.
- No personal loan: 50% of customers between approximately $37,500 - approximately $80,000 shown on w.r.t target boxplot with a median income level around approximately $60,000 without outliers.

CONTINUED

EDA Results- OBSERVATIONS (Bivariate)

Personal Loan vs Average Monthly Credit Card Expense (SEE B18)

- Personal loan: 50% of customers show between approximately 2,500 - 5,500 average spending on credit cards per month shown on w.r.t target boxplot with a median average spending level around approximately 3,750.
- No personal loan: 50% of customers between approximately 500 - 2,250 average spending on credit cards per month shown on w.r.t target boxplot with a median average monthly spending level around approximately 1,500 with many outliers spending as much as approximately 9,000.

(Without outliers) *(SEE B19)*

- Personal loan: 50% of customers show between approximately 2,500 - 5,500 average spending on credit cards per month shown on w.r.t target boxplot with a median average spending level around approximately 3,750.
- No personal loan: 50% of customers between approximately 500 - 2,250 average spending on credit cards per month shown on w.r.t target boxplot with a median average monthly spending level around approximately 1,500.

CONTINUED

EDA Results- CONCLUSIONS (Univariate)

- Age and experience are nearly perfectly positively correlated- the older one is, the more work experience they will have gained.
- Income and monthly credit card spending are relatively highly positively correlated- the level of income one earns strongly influences their average spending on credit cards per month.
- Income and personal loans are also relatively highly positively correlated- the higher income one earns, the higher the instance the customer accepted the personal loan offered in the last campaign.
- Income and education are slightly negatively correlated- the less education one has had, the lower their earned income level.
- Income and family size are also slightly negatively correlated- the lower the number of family members in the household, the lower the level of income being earned.
- Lastly, education and monthly credit card spending have a slight negative correlation- the education level one has has a slight negative affect on their average monthly credit card expenses.

EDA Results- CONCLUSIONS (Bivariate)

Personal Loan vs Education

- Customers with education levels of “Graduate” and “Advanced/ Professional” have the highest rate of taking personal loans at a total of 387 loans vs 93 total loans taken for education level “Undergrad”, which is the education level with the highest number of customers at 2,096.

Personal Loan vs Family

- The barplot demonstrates that families with higher numbers of members are more likely to take out personal loans, indicating the need for additional income to supplement the difference needed to support a household with more individuals present.
- The barplot also demonstrates that families with a lower number of members are less likely to take out a personal loan, indicating the lack of need to supplement any missing income due to the fact that the household includes less individuals to support.

EDA Results- CONCLUSIONS (Bivariate)

Personal Loan vs Securities Accounts

- Customers with a personal loan have a higher proportion (12.5%) of having a securities account compared to those without a personal loan (10.2%). This suggests a potential association where personal loan holders are somewhat more likely to have a securities account. Of the total customer base, 9.2% have a securities account, indicating that a significant portion of customers, regardless of their loan status, hold securities accounts.

Personal Loan vs CD Accounts

- It can be concluded that the rate of instance of personal loan holders also having a CD account is relatively high with nearly one in 3 having both. In comparison to the overall number of 302 for the total general bank customer population of 5,000 standing at 6%, this indicates that almost half of the total CD account holders are also personal loan holders.

Personal Loan vs Online

- At a rate of 60%, a significant proportion of personal loan holders use the bank's online platform, which is notably higher than the 40.4% of non-personal loan holders who use the online banking service. This suggests that personal loan holders are more inclined towards using online banking. The percentage of personal loan holders who bank online (60%) is higher compared to the overall online banking usage rate (60% vs. 40%). This indicates that personal loan holders are more likely to utilize online banking services compared to the average customer.

EDA Results- CONCLUSIONS (Bivariate)

Personal Loan vs Credit Cards

- The percentage of personal loan holders with a credit card issued by another bank (29.7%) is very close to the percentage of non-personal loan holders with such a credit card (29.3%). This indicates that having a personal loan does not significantly impact the likelihood of holding a credit card from another bank compared to not having a personal loan. The overall percentage of customers with a credit card issued by another bank (29.4%) is consistent with the percentages observed among both personal loan holders and non-loan holders. This suggests that credit card ownership is fairly uniform across different customer segments within the bank. The data reveals that personal loan status does not have a significant impact on the likelihood of holding a credit card from another bank. Credit card ownership is relatively uniform across different segments of the bank's customer base.

Personal Loan vs Zip Codes

- The distribution of ZIP codes among personal loan holders and non-loan holders is nearly identical across the categories. For each ZIP code range, the percentages of customers with and without personal loans are closely aligned. Since ZIP code distribution does not show significant variation between loan holders and non-holders, targeting customers based on their ZIP code alone may not be effective for increasing personal loan uptake. Other factors beyond geographic location might be more influential in determining personal loan eligibility or interest.

EDA Results- CONCLUSIONS (Bivariate)

Personal Loan vs Age

- For both customers with personal loans and those without, the age profile is consistent, indicating that age does not significantly differentiate between these two groups. The central tendency and the interquartile range are identical for both categories, reinforcing that age alone does not determine personal loan ownership. Other variables should be examined to better understand personal loan ownership patterns.

Personal Loan vs Work Experience

- The work experience distribution for customers with personal loans is slightly skewed towards those with less experience (minimum 0 years and up to 41 or 42 years), while non-loan holders have a slightly broader experience range. The median experience years for both groups are relatively close at around 20 years. While there is considerable overlap in work experience between customers with and without personal loans, the personal loan holders generally have slightly less work experience.

EDA Results- CONCLUSIONS (Bivariate)

Personal Loan vs Income

- Customers with a personal loan generally have a higher income compared to those without a personal loan. The 50th percentile (median) income for personal loan holders is around $140,000 while for non-loan holders it is around $55,000. This suggests that customers with personal loans tend to have higher incomes. There is a clear trend where higher-income individuals are more likely to hold personal loans. These conclusions suggest that income is a significant factor in determining whether a customer has a personal loan, with higher incomes being associated with a higher likelihood of holding a personal loan.

Personal Loan vs Monthly Credit Card Usage

- Customers with a personal loan generally spend more on their credit cards each month compared to those without a personal loan. For personal loan holders, the 50th percentile (median) average spending is around $3,750, whereas for non-loan holders, it is around $1,500. This suggests that personal loan holders tend to have higher monthly credit card expenditures. Non-loan holders include some high spenders, with outliers spending as much as $9,000 per month. This suggests that while the majority of non-loan holders spend less, there are a few individuals with exceptionally high spending, which influences the overall distribution. The data suggests that higher credit card spending is associated with having a personal loan, potentially indicating a higher level of financial engagement or borrowing behavior among these customers.

Data Preprocessing

- Duplicate value check

  There were no duplicate values in the data.info dataframe.

- Missing value treatment

  There were no non-null values in the data.info dataframe therefore no need to do any missing value treatment.

Data Preprocessing

Feature Engineering

1. There were 467 unique numbers in the “ZIPCode” category. We found that the number of unique values reduced down to only 7 unique numbers when only taking into account the first two digits.
2. We then converted the data type of categorical features to “Category”.

Data preprocessing for modeling

1. When checking for unique values in the “Experience” category, there were values of -1, -2, and -3 in the data. A value correction to replace the numbers with 1,2, and 3 was performed.
2. It was also determined when checking for unique values in the “Education” category that only three unique values of 1, 2, and 3 were found.
3. The category “ID” is not relevant to this model therefore it was dropped from the dataframe.
4. The category “Experience” was found to be nearly perfectly correlated with “Age” therefore it was also dropped from the dataframe.
5. The dataset was split into two separate sets- train (70% = 3,500) and test (30% = 1,500)

Model Building

Model building steps of Decision Tree

1. Define functions to compute different metrics to check performance of classification model
    a. model_performance_classification_sklearn
    b. confusion_matrix_sklearn
2. Choose type of classifier (SEE SK1, PRE1, POST1)

Model Building

Model performance steps of Decision Tree

1. Check model performance on training data (SEE SK2, PRE2, POST2)
2. Review Accuracy, Recall, Precision, F1 scores (SEE SK3, PRE3, POST3)
3. Visualizing the Decision Tree
    a. Review Decision Tree & Decision Tree Rules
    b. Review Gini importance of tree features (SEE SK4A & B, PRE4A &B, POST4A &B)
4. Check model performance on test data (SEE SK5, PRE5, POST5)
5. Review Accuracy, Recall, Precision, F1 scores (SEE SK6, PRE6, POST6)

Model Building

Model performance

- All three of the decision trees- "Decision Tree sklearn", "Decision Tree (Pre-Pruning)" and "Decision Tree (Post-Pruning)" have a better performance on training data than testing data, classifying them all as overfitted although the pre-pruned tree has a 5.1% better performance on training set than test set while the sklearn tree has an 8% better performance on training set than test set. The post-pruned tree performed the worst at 1.4% better on training set than test set.

Model Performance Summary

Model evaluation criteria

- The pre-pruned decision tree has a similar F1 performance scores on training (.93) and test sets (.879).
- This model uses less features for decision-making than the sklearn decision tree which utilized them all and more features than the post-pruned decision tree, which only used one feature (income), which is far too few features to make sound recommendations. We will eliminate the post-pruned tree at this time and proceed with comparing the performance of the sklearn tree vs. the pre-pruned tree.
- Evaluating only six features will result in a shorter prediction time than the inclusion of all original features in the sklearn tree and it is likely to be less complex in comparison and ultimately yield better results on unseen data.

Model Performance Summary

Great Learning
POWER AHEAD

Overview of the final decision tree model and its parameters

We'll move ahead with the pre-pruned decision tree as our final model.

PARAMETERS

Choose the type of classifier- estimator = DecisionTreeClassifier(random_state=1)
Grid of parameters to choose from-

```
parameters = {
    "max_depth": np.arange(6, 15),
    "min_samples_leaf": [1, 2, 5, 7, 10],
    "max_leaf_nodes": [2, 3, 5, 10],
}
```

Type of scoring used to compare parameter combinations- acc_scorer = make_scorer(recall_score)
Run the grid search
Set the clf to the best combination of parameters
Fit the best algorithm to the data

Model Performance Summary

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

The pre-pruned model uses the following six features for decision-making:

- Income
- Family
- Education_2 (“Graduate”)
- Education_3 (“Advanced/Professional”)
- CCAvg
- Age

(SEE PRE4A & B)

Model Performance Summary

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

Training performance metrics

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	1.0	0.987714	0.836286
Recall	1.0	0.873112	0.933535
Precision	1.0	0.996552	0.359302
F1	1.0	0.930757	0.518892

Test performance metrics

Test performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.986000	0.978667	0.823333
Recall	0.932886	0.785235	0.906040
Precision	0.926667	1.000000	0.349741
F1	0.929766	0.879699	0.504673

Model Performance Improvement

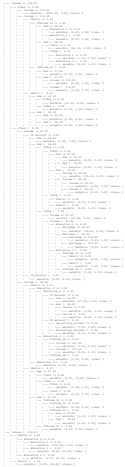
1. **Sklearn Decision Tree**: The model's rules are overly inclusive and complex, resulting in a large, convoluted pool of potential customers. This complexity can hinder its effectiveness as an accurate predictor for personal loan candidates. The model's extensive feature set and intricate rules make it less effective in identifying truly suitable candidates for the personal loan campaign.
2. **Pre-Pruned Decision Tree**: This model stands out as the most effective for predicting suitable candidates for the personal loan campaign. By considering six well-chosen features with appropriate weights, the pre-pruned tree strikes a balance between complexity and accuracy. Its nuanced decision-making capability ensures a more precise and actionable list of potential loan customers, making it well-suited for the campaign.
3. **Post-Pruned Decision Tree**: The post-pruned model, which relies on a single feature (Income), demonstrates too much simplification. Its limited feature set results in a model that is ineffective for predicting personal loan candidates. The lack of sufficient features impairs its ability to accurately assess customer suitability for the loan campaign.

Conclusion: While the sklearn decision tree is too complex and the post-pruned tree is too simplistic, the pre-pruned decision tree offers an optimal balance. It provides a concise, accurate, and actionable model for identifying potential loan candidates, making it the preferred choice for the personal loan campaign

Model Performance Improvement

Decision rules

Sklearn tree



The sklearn tree rules are far too inclusive and complex. This model will produce a large, complicated pool of customers that will not result in proving to be an accurate predictor of likely personal loan customers.

Pre-Pruned tree

```
|--- Income <= 116.50
|   |--- CCAvg <= 2.95
|   |   |--- Income <= 106.50
|   |   |   |--- weights: [2553.00, 0.00] class: 0
|   |   |--- Income >  106.50
|   |   |   |--- weights: [79.00, 10.00] class: 0
|   |--- CCAvg >  2.95
|   |   |--- Income <= 92.50
|   |   |   |--- weights: [117.00, 15.00] class: 0
|   |   |--- Income >  92.50
|   |   |   |--- Family <= 2.50
|   |   |   |   |--- weights: [37.00, 14.00] class: 0
|   |   |   |--- Family >  2.50
|   |   |   |   |--- Age <= 57.50
|   |   |   |   |   |--- weights: [1.00, 20.00] class: 1
|   |   |   |   |--- Age >  57.50
|   |   |   |   |   |--- weights: [7.00, 3.00] class: 0
|--- Income >  116.50
|   |--- Family <= 2.50
|   |   |--- Education_3 <= 0.50
|   |   |   |--- Education_2 <= 0.50
|   |   |   |   |--- weights: [375.00, 0.00] class: 0
|   |   |   |--- Education_2 >  0.50
|   |   |   |   |--- weights: [0.00, 53.00] class: 1
|   |   |--- Education_3 >  0.50
|   |   |   |--- weights: [0.00, 62.00] class: 1
|   |--- Family >  2.50
|   |   |--- weights: [0.00, 154.00] class: 1
```

The pre-pruned tree is well-suited to deliver accurate decisions for which customers would be suitable candidates for the personal loan campaign. It considers 6 features with appropriate weights assigned to each to get a very nuanced list of potential loan candidates.

Post-Pruned tree

```
|--- Income <= 98.50
|   |--- weights: [392.70, 18.70] class: 0
|--- Income >  98.50
|   |--- weights: [82.65, 262.65] class: 1
```

The rules of the post-pruned tree model show the tree is built with entirely too few features (Income) and its weights will not produce an effective model to predict appropriate customers for this personal loan campaign.

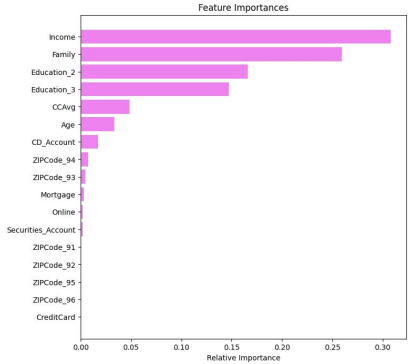
Model Performance Improvement

Decision feature importance

Sklearn tree

	Imp
Income	0.308098
Family	0.259255
Education_2	0.166192
Education_3	0.147127
CCAvg	0.048798
Age	0.033150
CD_Account	0.017273
ZIPCode_94	0.007183
ZIPCode_93	0.004682
Mortgage	0.003236
Online	0.002224
Securities_Account	0.002224
ZIPCode_91	0.000556
ZIPCode_92	0.000000
ZIPCode_95	0.000000
ZIPCode_96	0.000000
CreditCard	0.000000

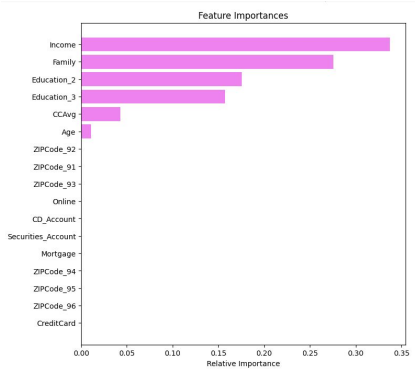
The sklearn tree is entirely too large and complex with too many features included, leading to a very complicated model.



Pre-Pruned tree

	Imp
Income	0.337681
Family	0.275581
Education_2	0.175687
Education_3	0.157286
CCAvg	0.042856
Age	0.010908
CD_Account	0.000000
Online	0.000000
Securities_Account	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
Mortgage	0.000000
CreditCard	0.000000

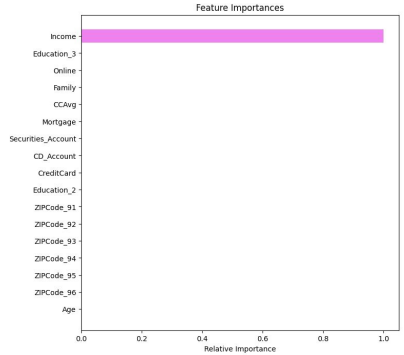
The pre-pruned tree is a smaller, much more manageable size. It prioritizes only the top 6 features for inclusion in the model, building a more reasonable and reliable model with a medium processing time.



Post-Pruned tree

	Imp
Income	1.0
Age	0.0
ZIPCode_91	0.0
Education_2	0.0
ZIPCode_96	0.0
ZIPCode_95	0.0
ZIPCode_94	0.0
ZIPCode_93	0.0
ZIPCode_92	0.0
CreditCard	0.0
Online	0.0
CD_Account	0.0
Securities_Account	0.0
Mortgage	0.0
CCAvg	0.0
Family	0.0
Education_3	0.0

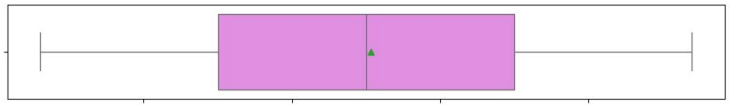
The post-pruned tree is extremely small with only one single feature (Income) used to determine a customer's likelihood of candidacy for taking personal loan.

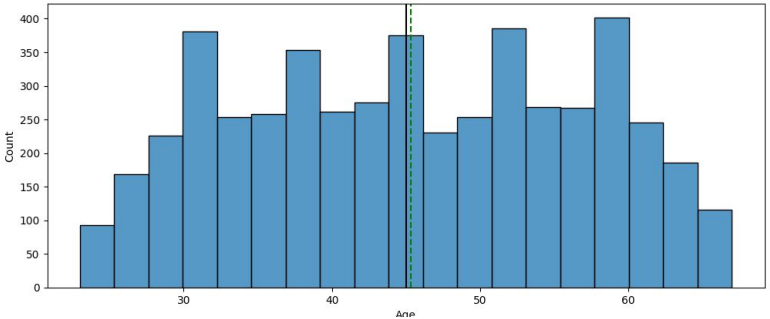


APPENDIX

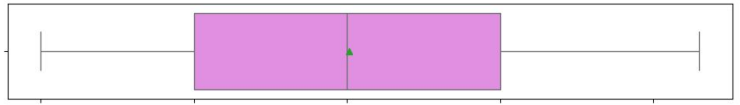
Data Background and Contents- (Univariate Data)

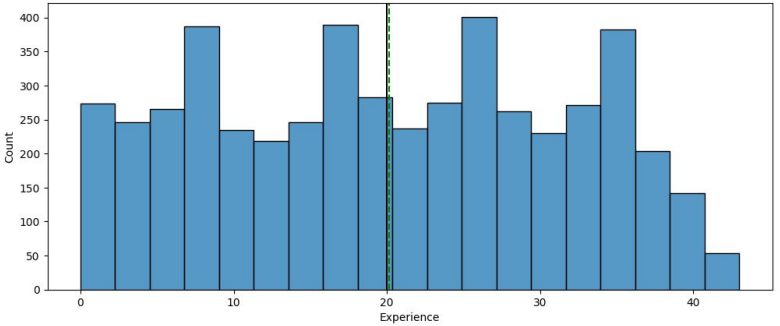
U1



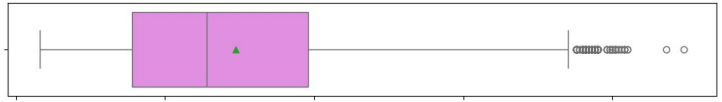


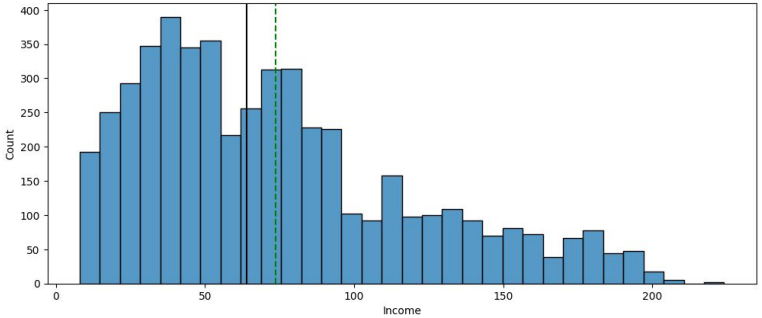
U2



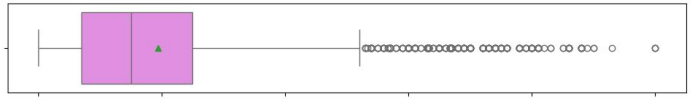


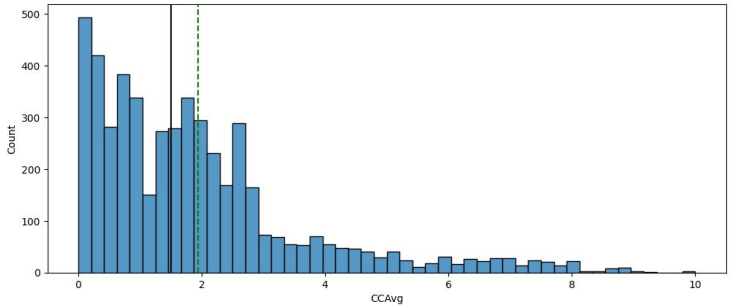
U3



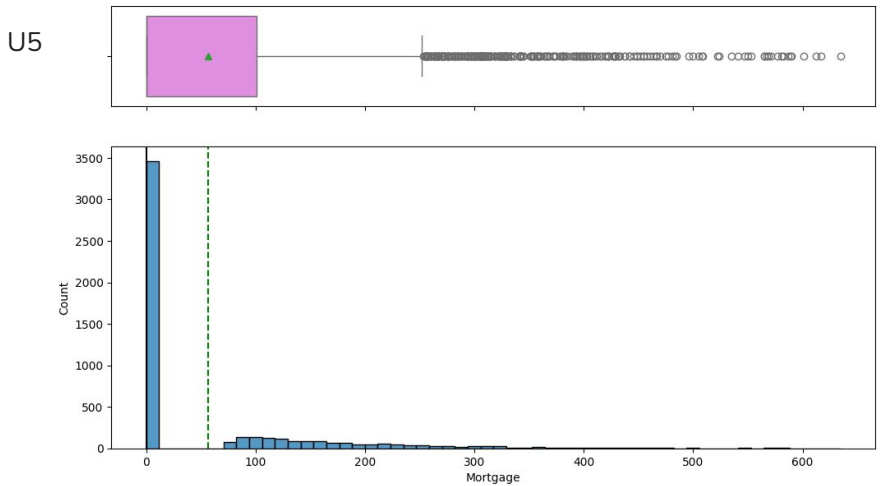


U4



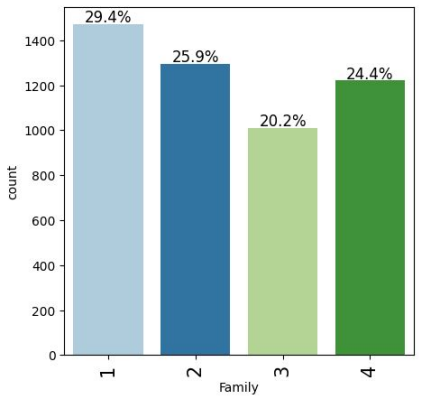


Data Background and Contents- (Univariate Data)

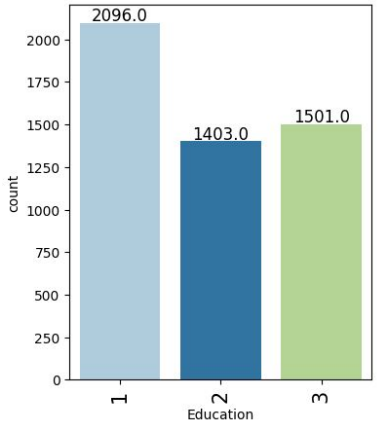


Data Background and Contents- (Univariate Data)

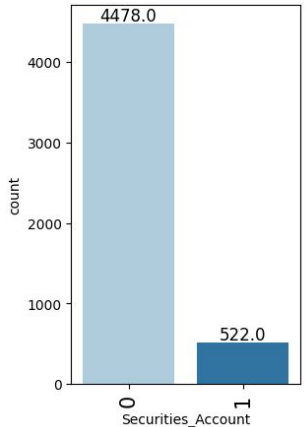
U6



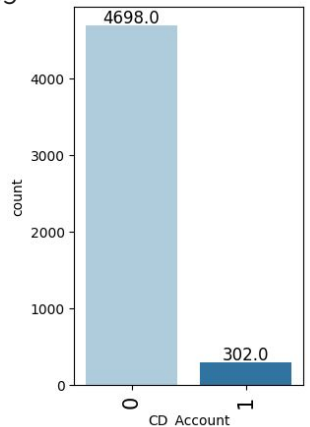
U7



U8

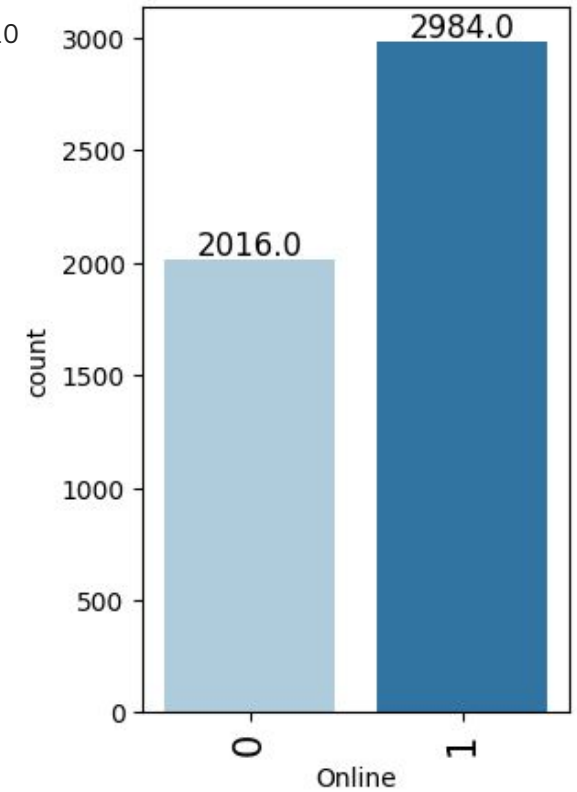


U9

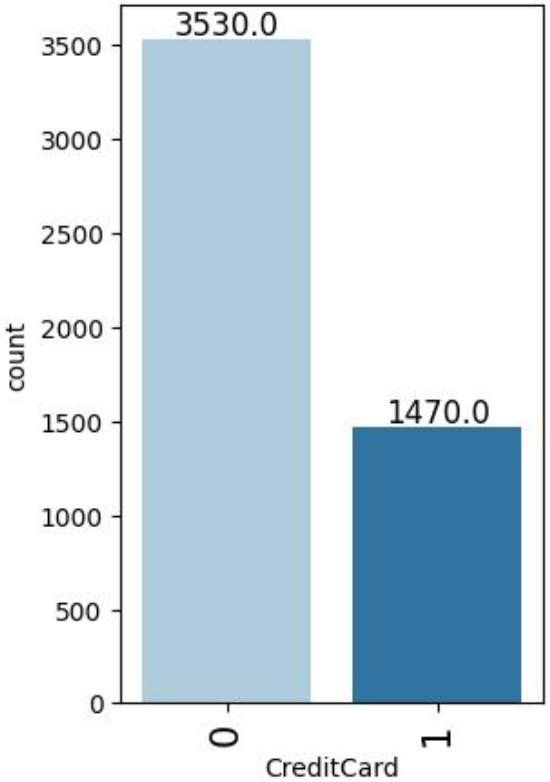


Data Background and Contents- (Univariate Data)

U10

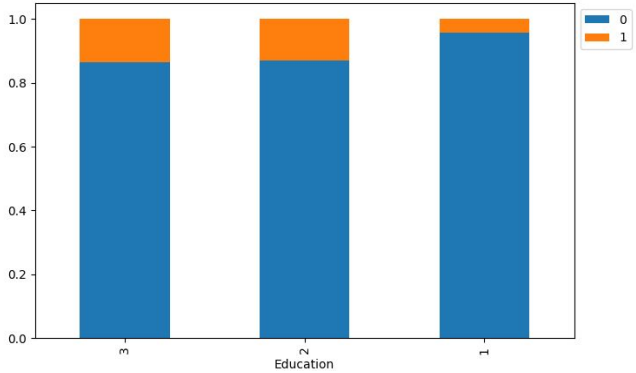


U11

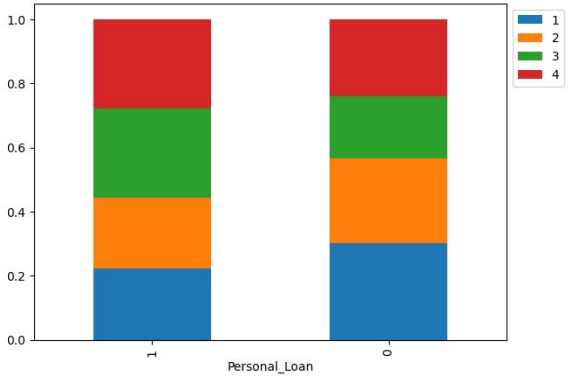


Data Background and Contents- (Bivariate Data)

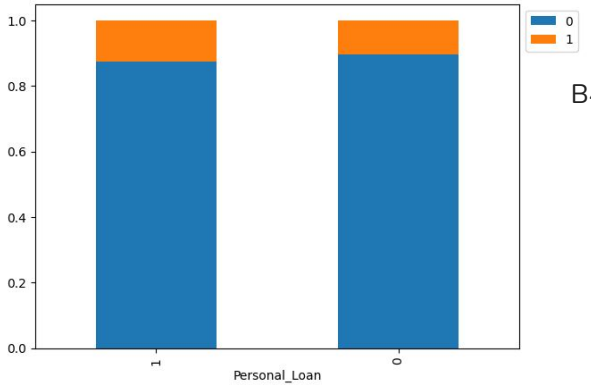
B1



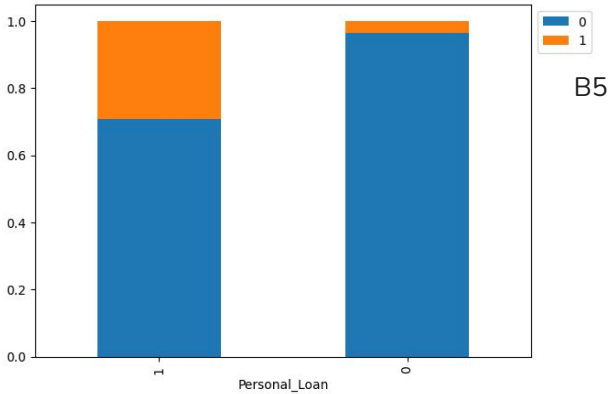
B2



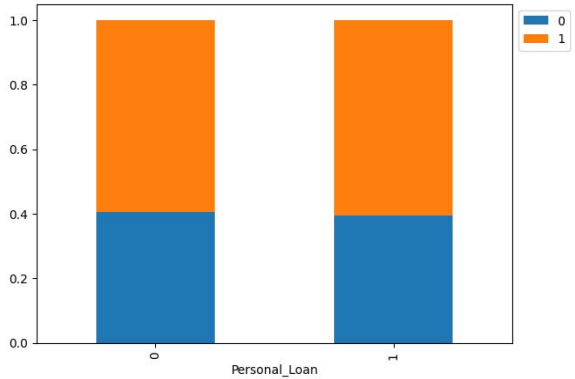
B3



B4

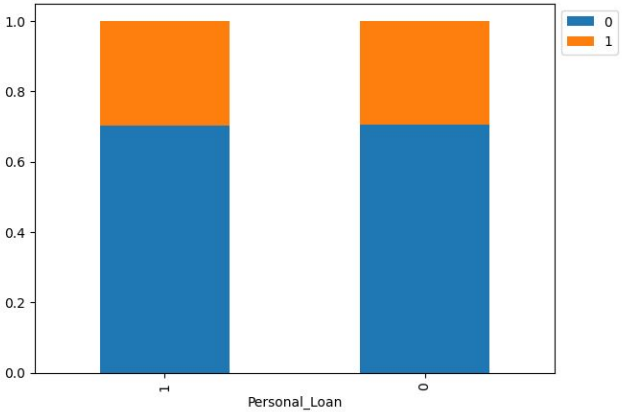


B5

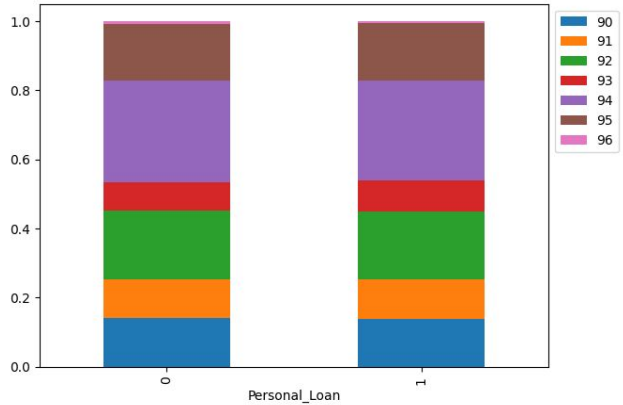


Data Background and Contents- (Bivariate Data)

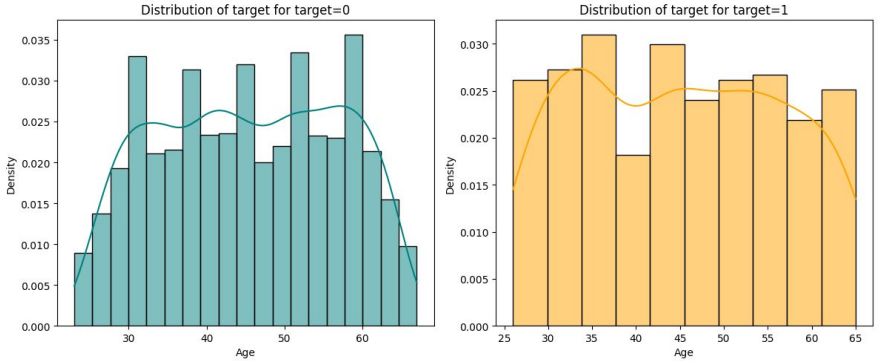
B6



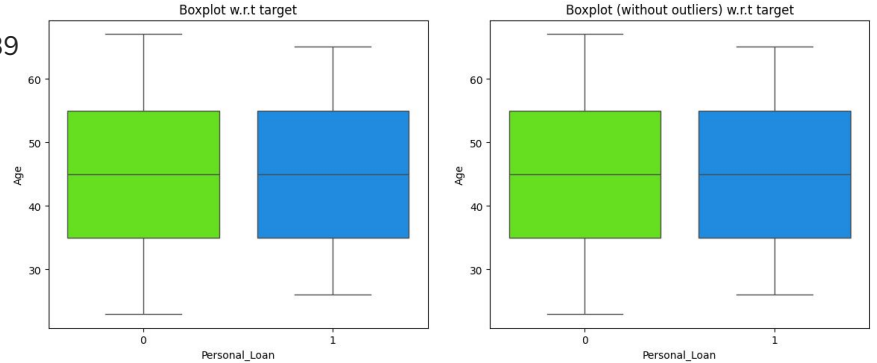
B7



B8

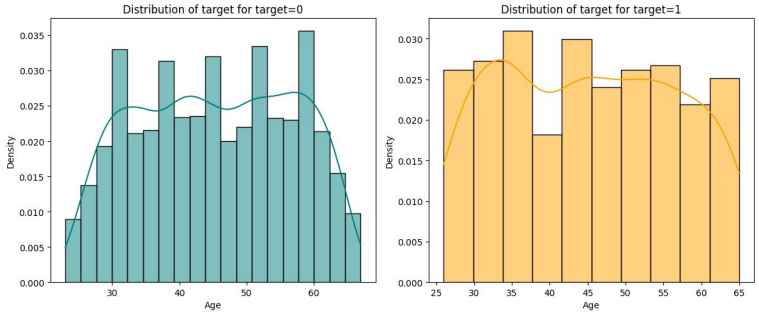


B9

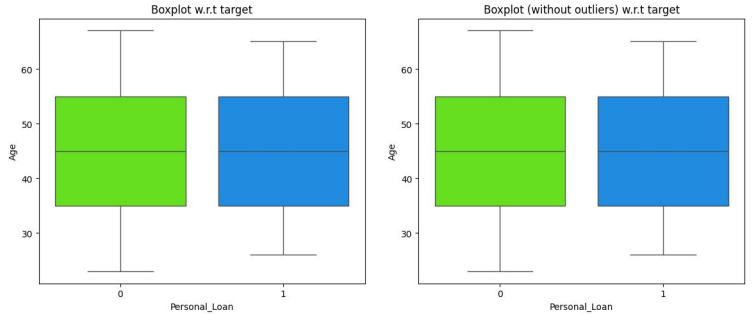


Data Background and Contents- (Bivariate Data)

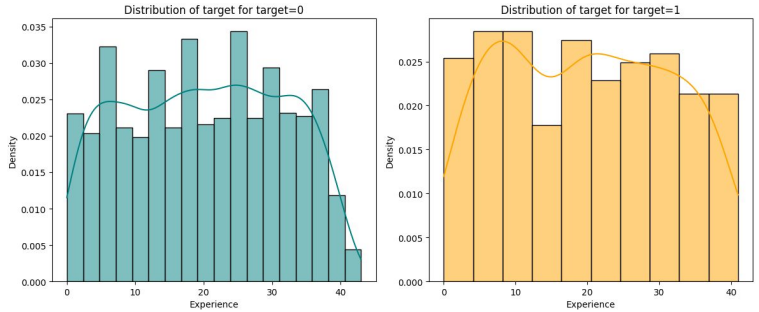
B10



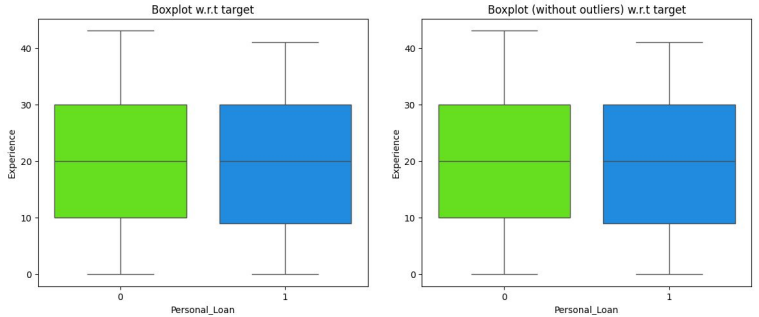
B11



B12

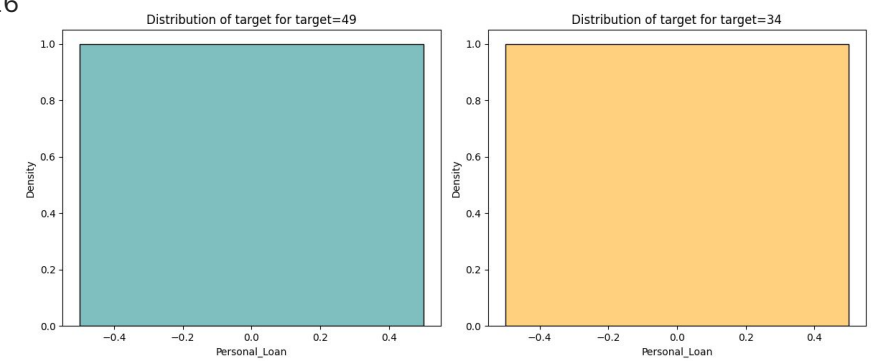


B13

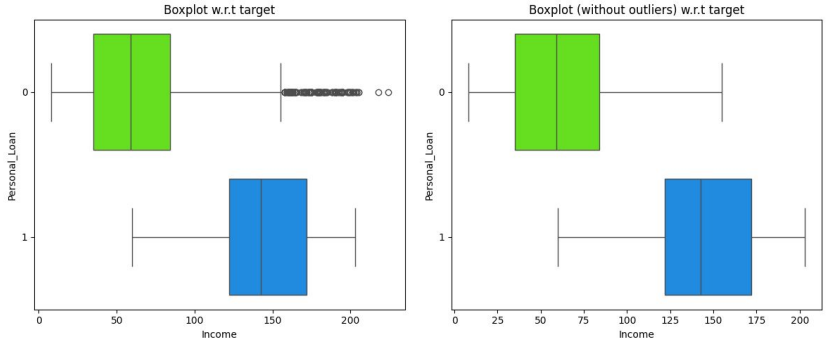


Data Background and Contents- (Bivariate Data)

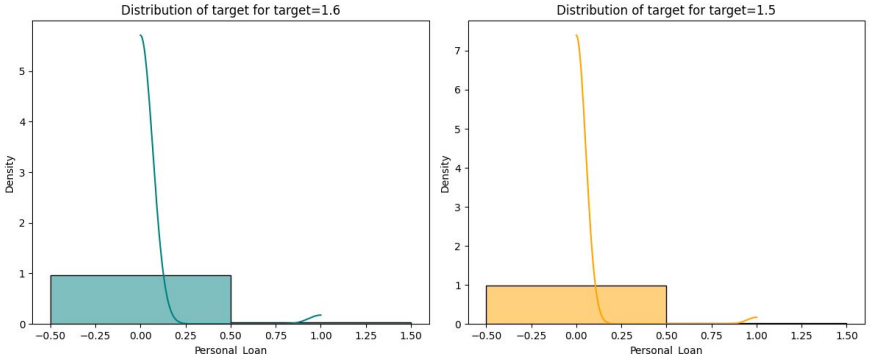
B16



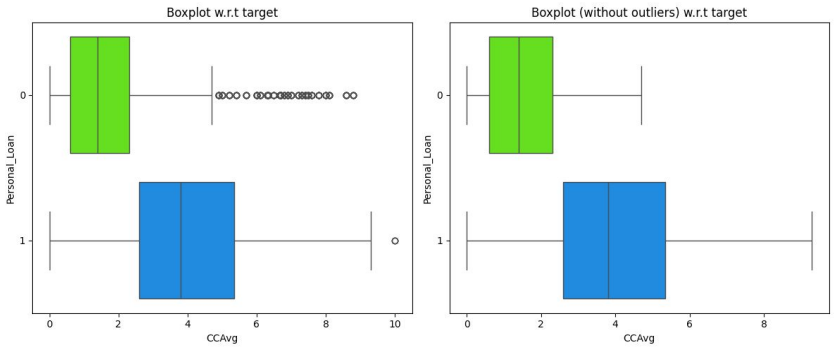
B17



B18

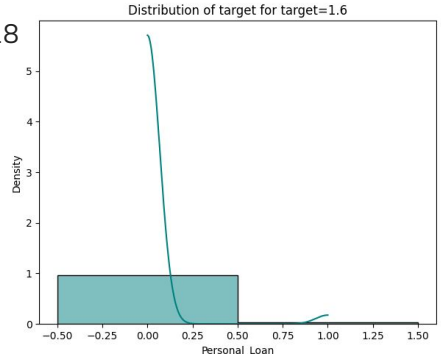


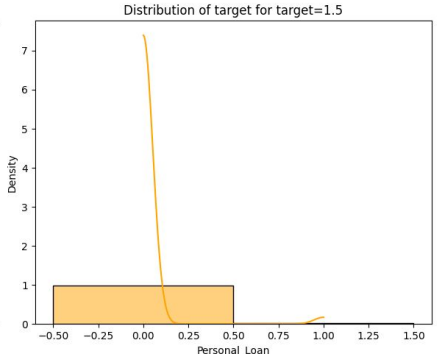
B19



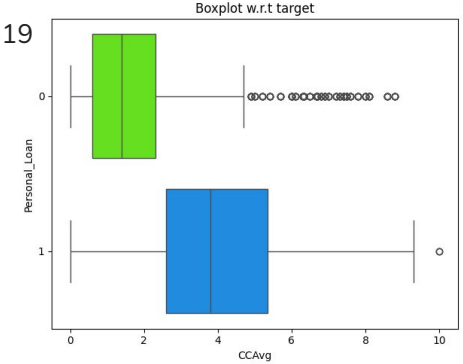
Data Background and Contents- (Bivariate Data)

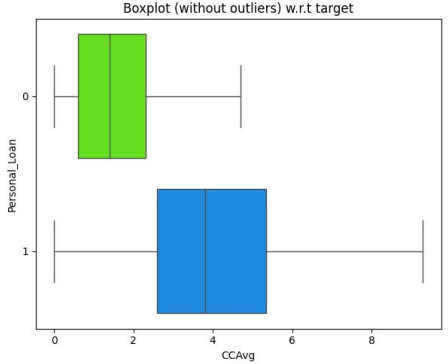
B18
c





B19





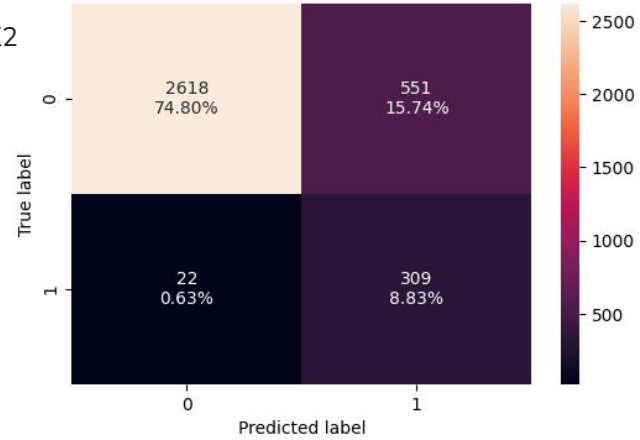
Data Background and Contents (Sklearn)

SK1

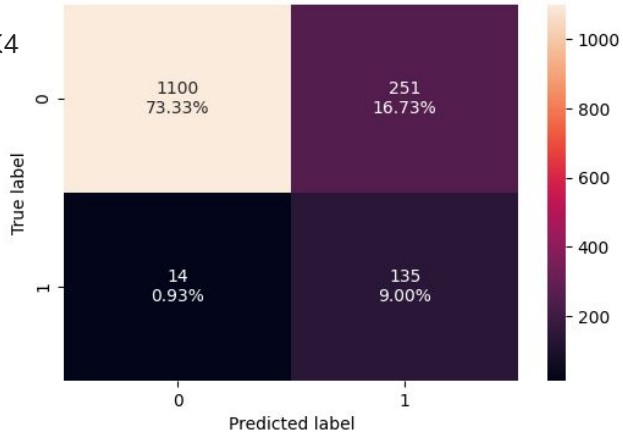
DecisionTreeClassifier

DecisionTreeClassifier(random_state=1)

SK2



SK4



SK3

decision_tree_tune_post_train

	Accuracy	Recall	Precision	F1
0	0.836286	0.933535	0.359302	0.518892

SK5

decision_tree_tune_post_test

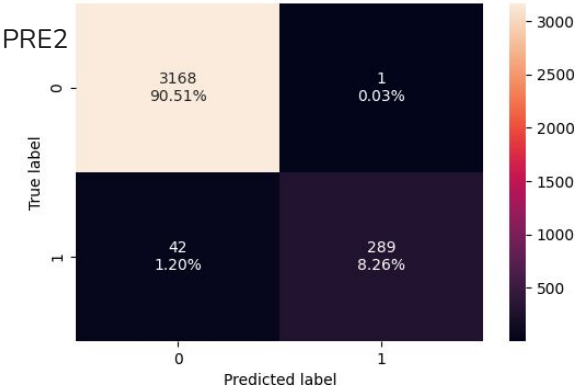
	Accuracy	Recall	Precision	F1
0	0.823333	0.90604	0.349741	0.504673

Data Background and Contents (Pre-Prune)

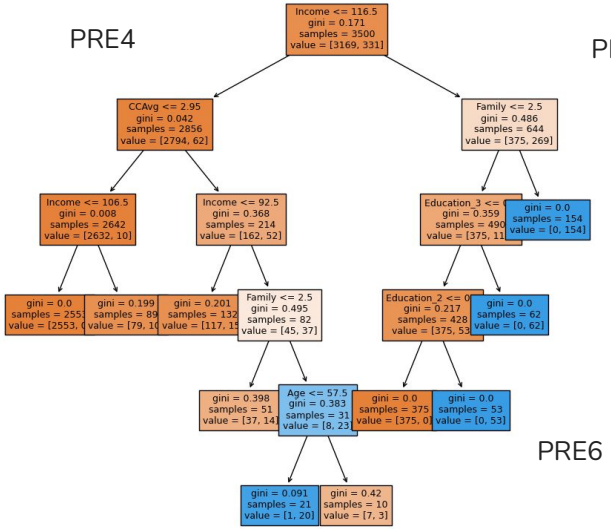
PRE1

```
DecisionTreeClassifier(max_depth=6, max_leaf_nodes=10, min_samples_leaf=10,
                       random_state=1)
```

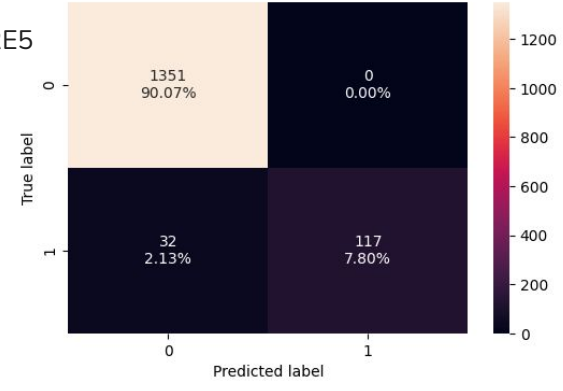
PRE2



PRE4



PRE5



PRE3

```
decision_tree_tune_perf_train
```

	Accuracy	Recall	Precision	F1
0	0.987714	0.873112	0.996552	0.930757

PRE6

```
decision_tree_tune_perf_test
```

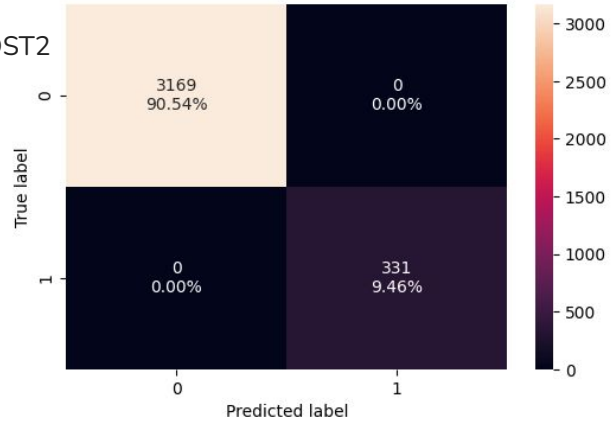
	Accuracy	Recall	Precision	F1
0	0.978667	0.785235	1.0	0.879699

Data Background and Contents (Post-Prune)

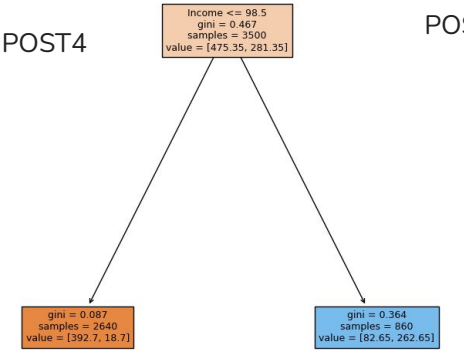
POST1

```
DecisionTreeClassifier(ccp_alpha=0.04708834100596766,
                       class_weight={0: 0.15, 1: 0.85}, random_state=1)
```

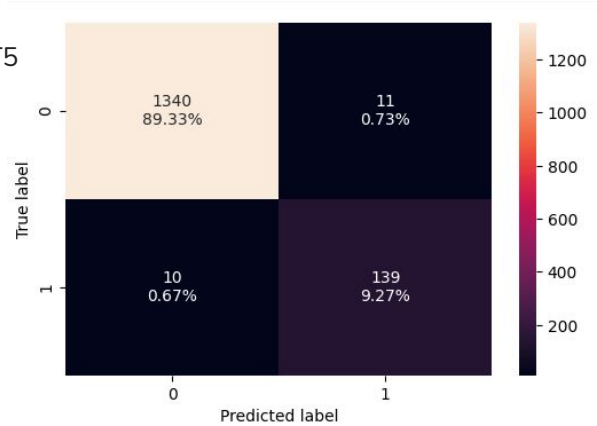
POST2



POST4



POST5



POST3

```
decision_tree_perf_train
```

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

POST6

```
decision_tree_perf_test
```

	Accuracy	Recall	Precision	F1
0	0.986	0.932886	0.926667	0.929766

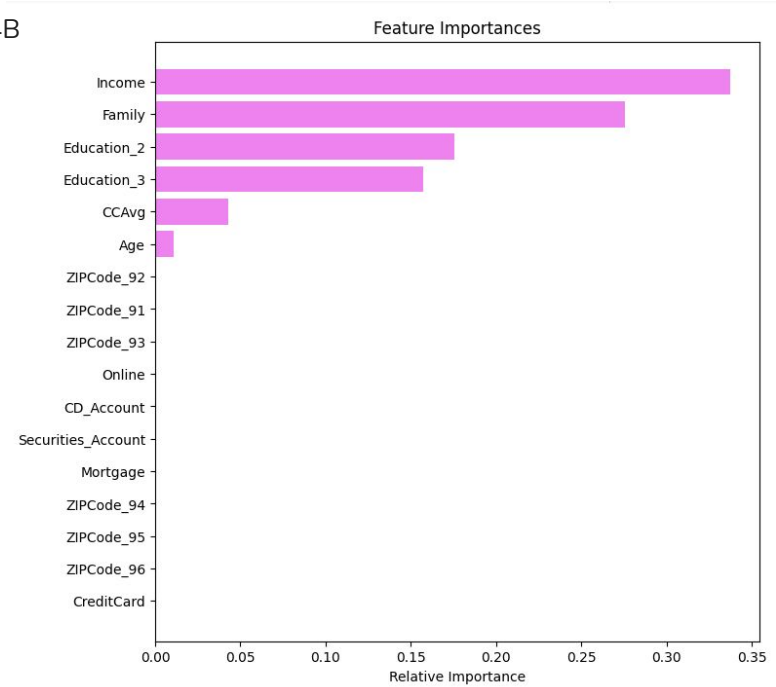
Slide Header

Final Decision Tree Model: Pre-Pruned Tree

PRE4A

	Imp
Income	0.337681
Family	0.275581
Education_2	0.175687
Education_3	0.157286
CCAvg	0.042856
Age	0.010908
CD_Account	0.000000
Online	0.000000
Securities_Account	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
Mortgage	0.000000
CreditCard	0.000000

PRE4B



Great Learning

Happy Learning !