Data Analysis - Low Code

Machine Learning Project
Personal Loan Campaign

November 2, 2025

Andres Dodero

Contents / Agenda

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

Business Problem Overview

- AllLife Bank wants to increase the number of personal loan customers among its existing liability customers (depositors). Currently, only a small fraction of customers take personal loans. The goal is to predict which liability customers are most likely to accept a personal loan offer, so the marketing team can target them effectively and improve campaign success rates.
- Objective: Predict which existing liability customers are most likely to accept a personal loan offer, enabling targeted marketing and higher conversion rates.

Executive Summary: Insights & Recommendations

Focus marketing on high-income customers

- Income is the most influential predictor of loan acceptance
- Target > $90K income customers

Segment by Education Level:

- Select customers with graduate or professional education; higher conversion chance.

Credit Card Average Spending:

- Use credit card spending as a secondary filter for campaigns.
- Target > $3K / month spenders

Optimize Marketing Spend:

- Post-pruned model hast best balance between Recall (93%) and Precision (79%).
- Reduces wasted targeting while capturing most potential loan customers

Executive Summary: Additional Recommendations

- Fine-tune the campaign if needed for families with 3-4 members
- Ignore ZIP codes, online banking, and mortgage data as they add no predictive value after pruning.
- Bundle personal loans with credit card offers for high CCAvg customers.

Solution Approach

Data Preparation:
- Load & clean data, handle anomalies, encode categorical variables, and drop irrelevant features.

Exploratory Data Analysis (EDA):
- Perform univariate and bivariate analysis to understand patterns and correlations.

Model Building
- Train a Decision Tree Classifier to predict loan acceptance.
- Split data into training and test sets.

Model Optimization
- Apply pre-pruning (limit depth, min samples, leaf nodes).
- Apply post-pruning using cost-complexity pruning (ccp_alpha).

Model Evaluation
- Compare models using Accuracy, Precision, Recall, F1-score, and confusion matrix.
- Select the best-performing model.

Insights & Recommendations
- Identify key drivers (e.g., income, credit card usage).
- Suggest targeted marketing strategies for high-probability segments.

Data Overview Summary

Dataset Size:

- Rows: 5,000 / Columns: 14 (after dropping ID, 13 remain)

Column Types:

- Numeric: Age, Experience, Income, Family, CCAvg, Mortgage
- Categorical: Education, Personal_Loan, Securities_Account, CD_Account, Online, CreditCard, ZIPCode

Target Variable:

- Personal_Loan (0 = No, 1 = Yes) / Acceptance rate: 9.6% (480 customers)

Key Statistics:

- Age: Mean = 45 years; range 23–67
- Experience: Mean = 20 years; range 0–43
- Income: Mean = $73.8K; range $8K–$224K
- CCAvg: Mean = $1.94K; range $0–$10K
- Mortgage: Median = $56K; highly skewed with outliers up to $635K
- Family Size: 1–4 members (most common = 1)Education: Levels 1–3 (Undergrad, Graduate, Professional)

Data Quality:No missing values.

- Negative experience values corrected.
- ZIPCode simplified to first two digits (7 unique categories).

EDA Results

These fields were analyzed using a boxplot and histogram:

- Age: Customer's age in completed years
- Experience: #years of professional experience
- Income: Annual income of the customer (in thousand dollars)
- ZIP Code: Home Address ZIP code.
- Family: the Family size of the customer
- CCAvg: Average spending on credit cards per month (in thousand dollars)
- Education: Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional
- Mortgage: Value of house mortgage if any. (in thousand dollars)
- Personal_Loan: Did this customer accept the personal loan offered in the last campaign? (0: No, 1: Yes)
- Securities_Account: Does the customer have securities account with the bank? (0: No, 1: Yes)
- CD_Account: Does the customer have a certificate of deposit (CD) account with the bank? (0: No, 1: Yes)
- Online: Do customers use internet banking facilities? (0: No, 1: Yes)
- CreditCard: Does the customer use a credit card issued by any other Bank (excluding All life Bank)? (0: No, 1: Yes)

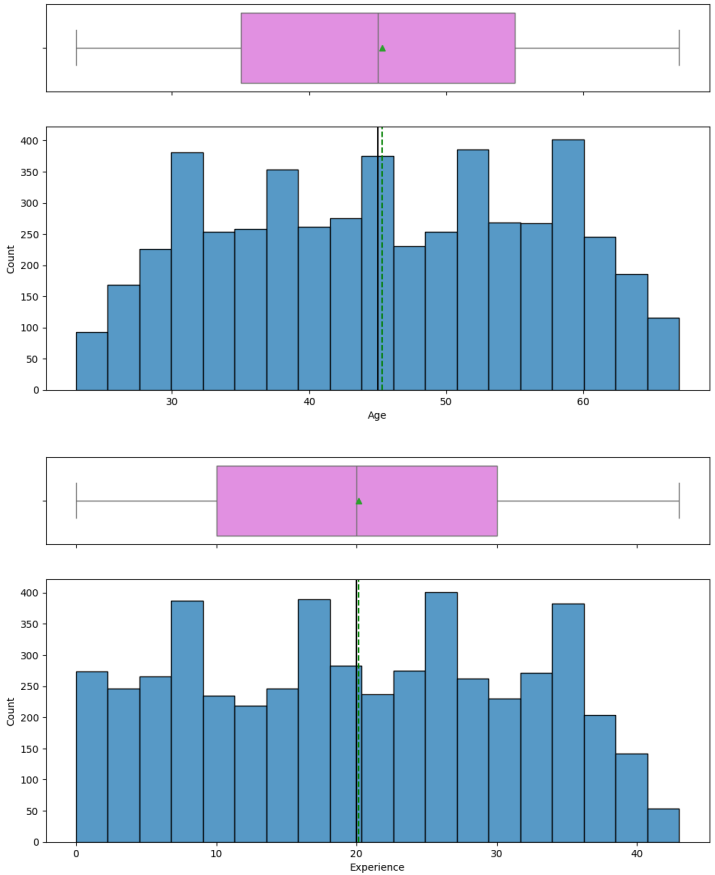
EDA Results

Age

- Ages span from about 23 to 67 years.
- Mean is 45 years.
- No significant outliers; follows a normal distribution.

Experience:

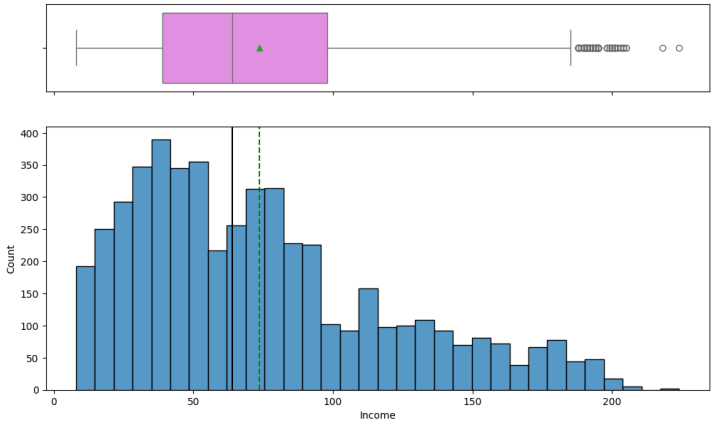
- Experience spans from 0 to about 43 years.
- Mean is 20 years.
- No significant outliers (negative values were fixed earlier in preprocessing).
- Histogram shows a relatively uniform distribution



EDA Results

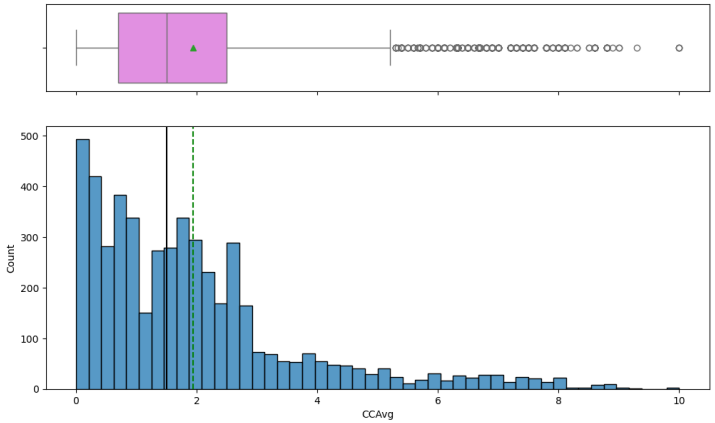
Income

- Wide range: $8K–$224K; median = $73.77K.
- Graph shows right-skey; most customers make < $100K
- Many outliers making ~ $190K or more.



CCAvg (Credit Card Average Spend):

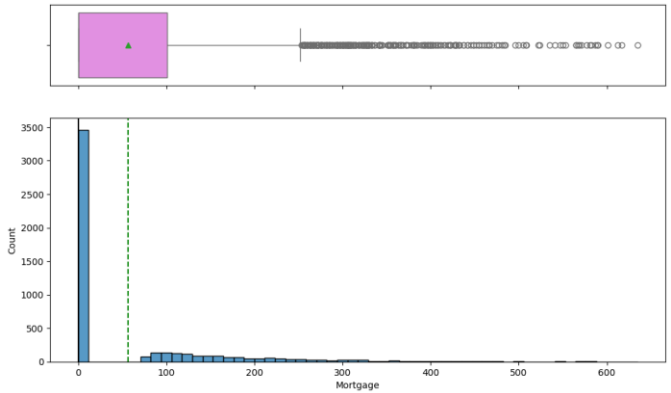
- Range: 0–10K; mean = 1.9K.
- Graph shows right-skew; most customers spend less than ~$3K per month
- Many outliers spending > $5K per month



EDA Results

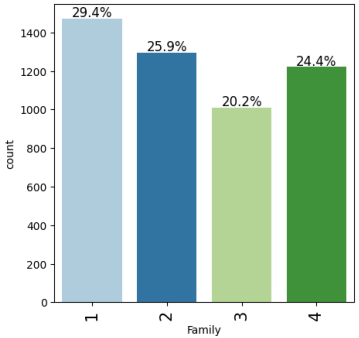
Mortgage:

- Extremely right-skewed: Most customers have zero mortgage.
- A small number of customers have very high mortgage values (up to $635K).
- Numerous outliers beyond $300K, visible in the boxplot.



Family Size:

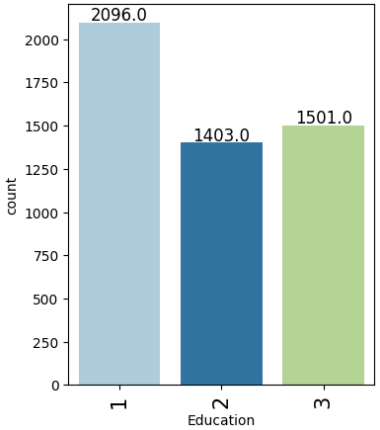
- Family size of 1 is the most common, but not by much.
- The differences are minimal between each size value.



EDA Results

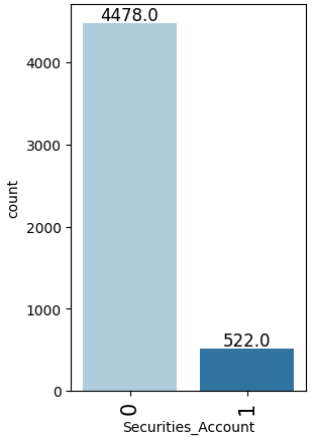
Education:

- Most customers are classified as “Undergrad” or 1
- While undergraduates dominate, graduate and professional levels together form a significant portion (~58% combined).



Securities:

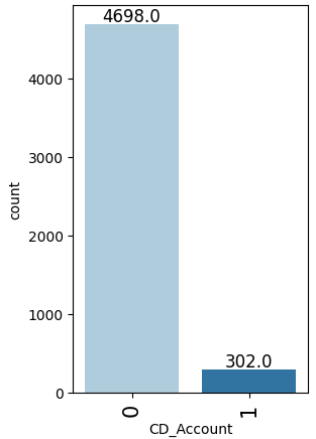
- Most customers do not have a Securities account (89%)



EDA Results

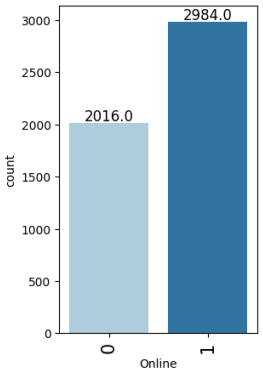
CD Account:

- Same as Securities, customers with a CD account are the majority.



Online Banking:

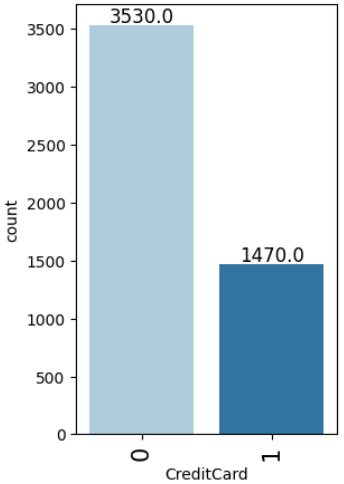
- Majority of customers use online banking (59.7%), but a significant portion still does not (40.3%).



EDA Results

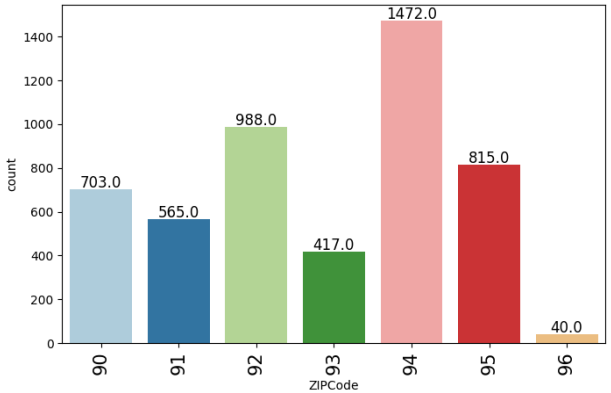
Have a Credit Card with the Bank:

- Customers without a credit card from another bank (0): 3,530 (≈ 70.6%).
- Customers with a credit card from another bank (1): 1,470 (≈ 29.4%).
- Majority of customers do not hold external credit cards.



Zip Code:

- Strong concentration on Zip Code 94 with 1,472 customers (≈ 29.4%).
- ZIPCodes 92 (988 customers) and 90 (703 customers) are next in size.
- ZIPCode 96 is extremely rare (only 40 customers).



Data Preprocessing

Duplicate value check

- Duplicate values were searched in the data frame. There are repeated values for several fields like zip code, years of experience, age, etc., but those were expected. However, no duplicate records were identified. Used data.duplicated().sum() to count duplicate records.

Missing value treatment

- No missing values were identified.

```
data.isnull().sum()
```

	0
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

Data Preprocessing

Outlier check: Calculated Interquartile Range (IQR) for numeric columns.

- Income: ~1.92% outliers.
- CCAvg: ~6.48% outliers.
- Mortgage: ~5.82% outliers.
- Age, Experience, Family: 0% outliers.

Analysis

- Outliers mainly occur in Income, CCAvg, and Mortgage, which are expected because these features have wide ranges.
- **Decision**: Outliers were not removed, likely because they represent real high-value customers, and are important to predict loan default.

Data Preprocessing

Feature Engineering.

- When accounting for the 1st 2 digits of the Zip Code, 7 unique values were identified
- Converted these categorical columns to the category data type: Education, Personal_Loan, Securities_Account, CD_Account, Online, CreditCard, ZIPCode

Data preprocessing for modeling

- Dropped “Experience” as it is perfectly correlated with Age
- Dropped “Personal Loan” since it is the target variable, and stored it separately in “Y” for modeling
- Kept all other customer details as input features.
- Converted ZIP code and Education into separate columns for each category
- Made sure all data is in a numeric format for the model.
- Split the data into: 1) Training set (70%) 2)Test set (30%).
- Checked the balance of loan approvals: 90% did not take a loan, 10% did.

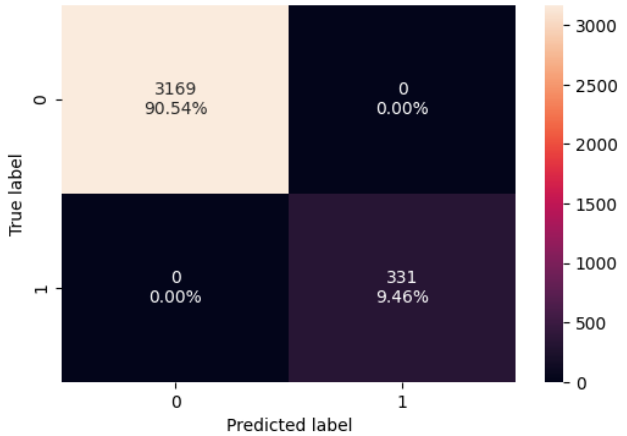
Model Building

Model building steps of Decision Tree

- After the Data preprocessing, created a Decision Tree Classifier
- Fit the model on the training data (70%)
- Computed metrics: Accuracy, Recall, Precision, F1 using a custom function. Performance was perfect due to the overfitting of the data with many leaf node values = 1

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

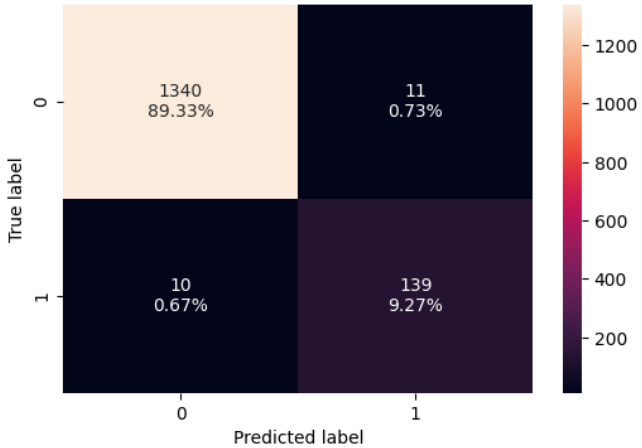
- Visualized confusion matrix.



Model Building

Tested the performance using the test data (30%)

- Confusion Matrix:



- Computed metrics: shows a decrease from the perfect 1.0 in the training data

	Accuracy	Recall	Precision	F1
0	0.986	0.932886	0.926667	0.929766

Model Performance Summary

- Model evaluation criterion
- Overview of the final decision tree model and its parameters
- Summary of most important features used by the decision tree model for prediction
- Summary of key performance metrics for training and test data of all the models in tabular format for comparison

Model Performance Summary

Model evaluation criterion

- The bank wants to maximize loan conversions while minimizing false positives (offering loans to who won't accept).
- Missing potential buyers (false Neg) is costly because it reduces campaign success.
- ~9.5% of the customers accepted the loan; data is skewed or unbalanced.
- High accuracy can be misleading since predicting "No Loan" for all gives ~90.5% accuracy but no business value.

Key Metrics

- Accuracy: Provides prediction correctness, but not ideal for this project since high accuracy can be misleading.
- **Recall: Measures how many actual loan buyers are correctly identified. Highly critical since the goal is to maximize conversions; missing potential customers hurts revenue.**
- Precision: Measures how many predicted buyers actually take the loan. It has importance since the bank wants to avoid spending on customers unlikely to accept loans.
- F1 Score: Harmonic mean of precision and recall. Useful because both recall and precision matter for campaign efficiency.

Recommended Criterion:

- Recall will capture most potential loan acceptances. Use F1 for balanced performance.

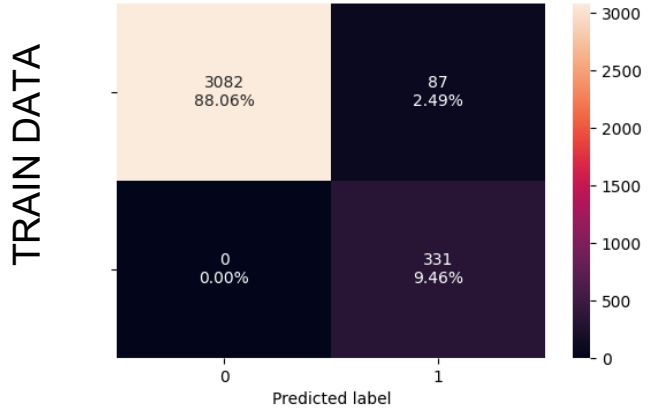
Model Performance Summary – PrePruned

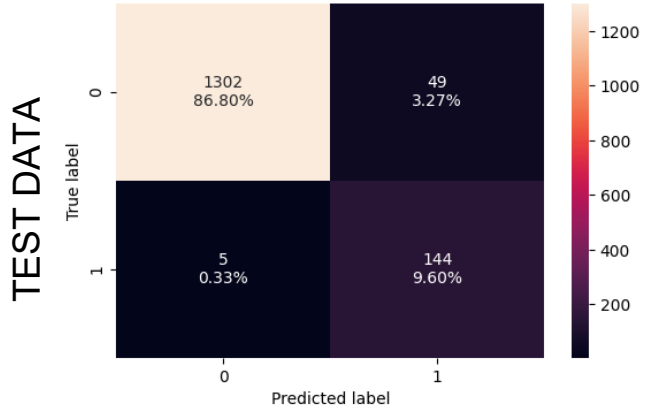
The algorithm that maximized RECALL tended to choose very small trees, which risked overpruning and losing important features that indicate customer risk. After testing multiple configurations, strong performance was achieved, retaining key features by selecting the parameters shown below.

- Max depth: 6
- Max leaf nodes: 25
- Min samples split: 10

Best test recall score: 0.966

Accuracy	Recall	Precision	F1
0.964	0.966443	0.746114	0.842105

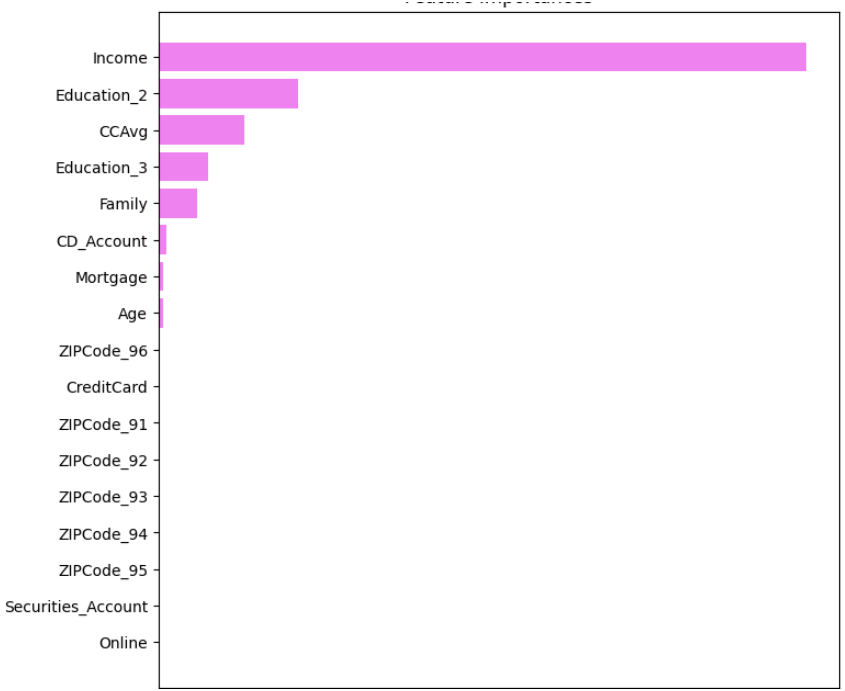




Model Performance Summary – PrePruned

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

- Income has the highest relative importance, and therefore is the Root Node
- Customers with Education = Graduate or Advanced show higher conversion rates.
- Customers with higher monthly credit card spending tend to accept loans more often.
- Family size plays a moderate role; larger families might have higher financial needs.
- CD Account, Mortgage and Age have very low importance; the rest have almost 0 impact.



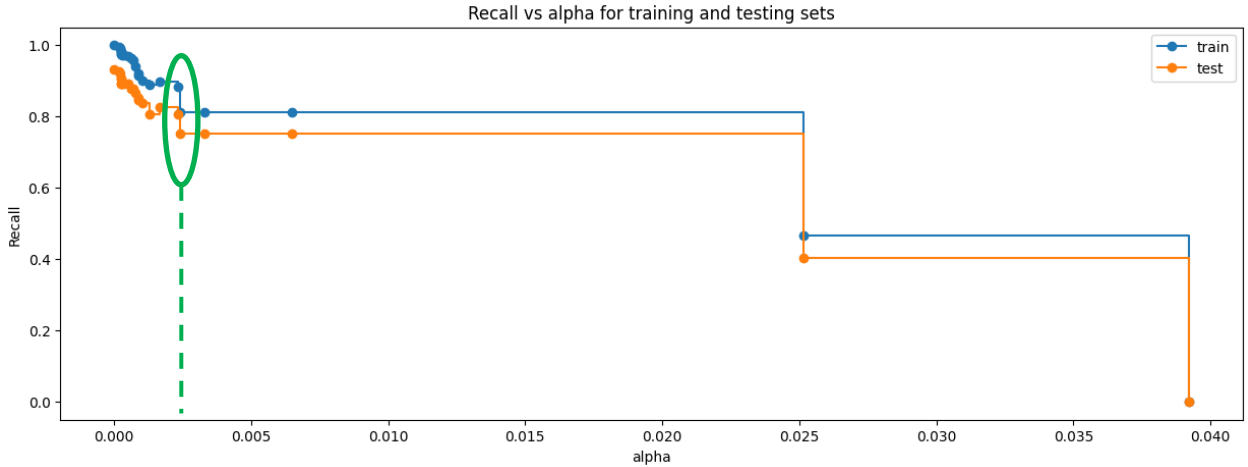
Model Performance Summary – PostPruning

Calculated the Cost Complexity Pruning Path for the Decision Tree; the goal is to reduce the tree size and see how that affected the model's impurity errors

- We want a model that finds nearly all potential loan customers (high Recall) while avoiding unnecessary complexity that could lead to overfitting or poor performance on new data.

Overview of the adjusted decision tree model and its parameters

- The tree has 5 levels that split on: 1) Income 2) CCAvg and Family 3) CD_Account, Education_3, Income, 4) CCAvg, Education_2, Income. 5) Income, CCAvg 6) Leaf nodes (final decisions).
- 31 nodes in total, with 10 Leaf Nodes (smallest is samples = 5)
- The best alpha was determined at 0.0025 based on this analysis and a significant drop in nodes in the first set



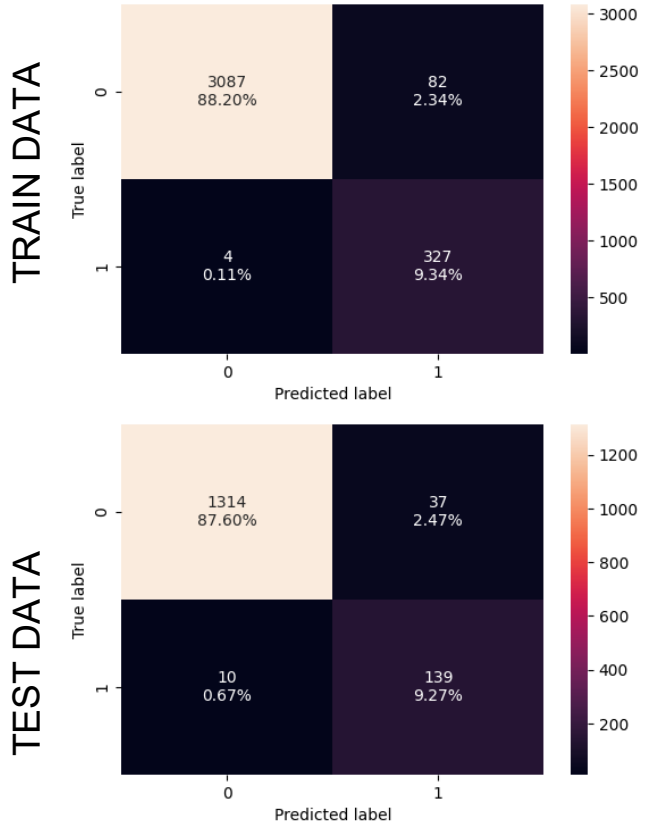
Model Performance Summary – PostPruned

Test Data:

- Recall decreased in PostPruning, while F1 score and Precision Increased compared to the PrePruning model.
- Recall continues to be strong (only 10 missed out of 149 actual loan takers)
- Overall, the model correctly predicts 96%+ of cases, with very few missed loan customers (0.67%) and limited false targeting (2.47%). This ensures efficient marketing and high conversion potential

Best test recall score: 0.933

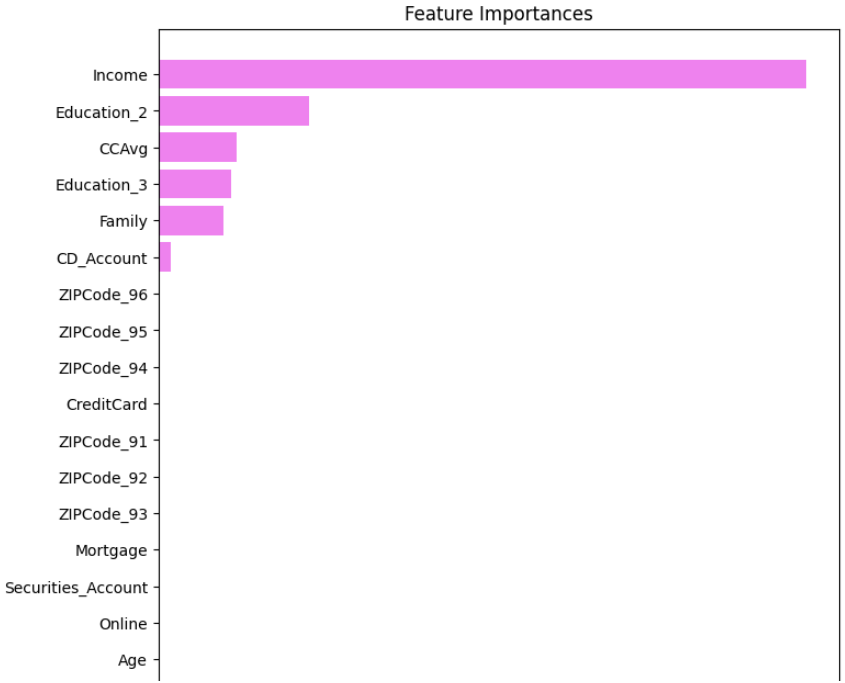
Accuracy	Recall	Precision	F1
0.968667	0.932886	0.789773	0.855385



Model Performance Summary – PostPruned

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

- Income remains the most influential feature and is still the root driver of predictions.
- Customers with Education = continues to play a significant role in loan acceptance.
- Customers with higher monthly credit card is an important secondary indicator of likelihood to accept a loan.
- Family: no change in importance (small); larger families might have higher financial needs.



Model Performance Summary

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

Training Data Performance

Metric	Default Model	Pre-Pruned	Post-Pruned
Accuracy	1.000	0.975143	0.975429
Recall	1.000	1.000	0.987915
Precision	1.000	0.791866	0.799511
F1 Score	1.000	0.883845	0.883784

- Pre-pruning improved generalization and boosted Recall (great for targeting loan customers), but Precision dropped significantly.

Test Data Performance

Metric	Default Model	Pre-Pruned	Post-Pruned
Accuracy	0.986	0.964	0.968
Recall	0.933	0.966	0.933
Precision	0.927	0.746	0.790
F1 Score	0.930	0.842	0.855

- Post-pruning slightly reduces Recall compared to pre-pruning but improves Precision; fewer wasted marketing efforts while still catching most positives.

Model Performance – Different Pruning Techniques

Default Tree:
- Very deep and complex, likely overfitted.
- Excellent performance on training sets, but high risk of overfitting.

Pre-Pruned Tree:
- Recall improved (93.3% to 96.6%) / fewer missed potential loan buyers.
- Precision dropped (92.7% to 74.6%) / more false positives thus higher marketing $$.
- Improvement: A simpler tree (smaller, fewer nodes), with a strong focus on recall.

Post-Pruned Tree:
- Recall improved (93.3% to 96.6%) / fewer missed potential loan buyers.
- Precision dropped (92.7% to 74.6%) / more false positives thus higher marketing $$.
- Improvement: similar performance vs. Pre-Pruned, but improves Precision, thus less wasted marketing resources.

Model Performance – Decision Rules & Key Features

Income is the root node

- Customers with higher income thresholds are more likely to accept a loan.
- For example: If Income > ~98K, the probability of loan acceptance increases significantly.

Education level matters

- Graduate (Education_2) and Advanced (Education_3) customers are strong indicators of acceptance.
- Combined with income, these splits refine targeting.

Credit Card Average Spending (CCAvg)

- Higher monthly credit card spending (> ~2.95K) often signals openness to loans.
- Family size and CD Account
- Family size plays a minor role; CD Account presence slightly increases likelihood.

Key Features:

- Income dominates as the most important feature
- Education 2 and 3 account for 22% of the importance
- Credit Card average spending contributes 7.5%

APPENDIX

Data Background and Contents

Great Learning
POWER AHEAD

Data Background:

- Source: AllLife Bank customer dataset from a personal loan campaign.
- Objective: Predict which liability customers are likely to accept a personal loan offer.
- Size: 5,000 records, 14 attributes.
- Target Variable: Personal_Loan (0 = No, 1 = Yes).

Data Contents:

- Demographics: Age, Experience, Family size.
- Financials: Income, Mortgage, CCAvg (monthly credit card spending).
- Education: Levels (Undergrad, Graduate, Advanced).
- Banking Behavior: Securities Account, CD Account, Online banking usage, CreditCard usage.
- Location: ZIP Code.
- Target: Loan acceptance indicator.

Additional Analytical Insights

- The original data showed that only 9.6% of customers accepted the loan, making the dataset highly skewed.
- Income and Education are key drivers; this aligns with the business logic that higher income and education correlate with loan eligibility and interest.
- Family size plays a minor role; targeting market efforts by family sizes of 3 and 4 could improve loan acceptance but not by much.
- CD Account presence adds slight influence.

Great Learning

Happy Learning !