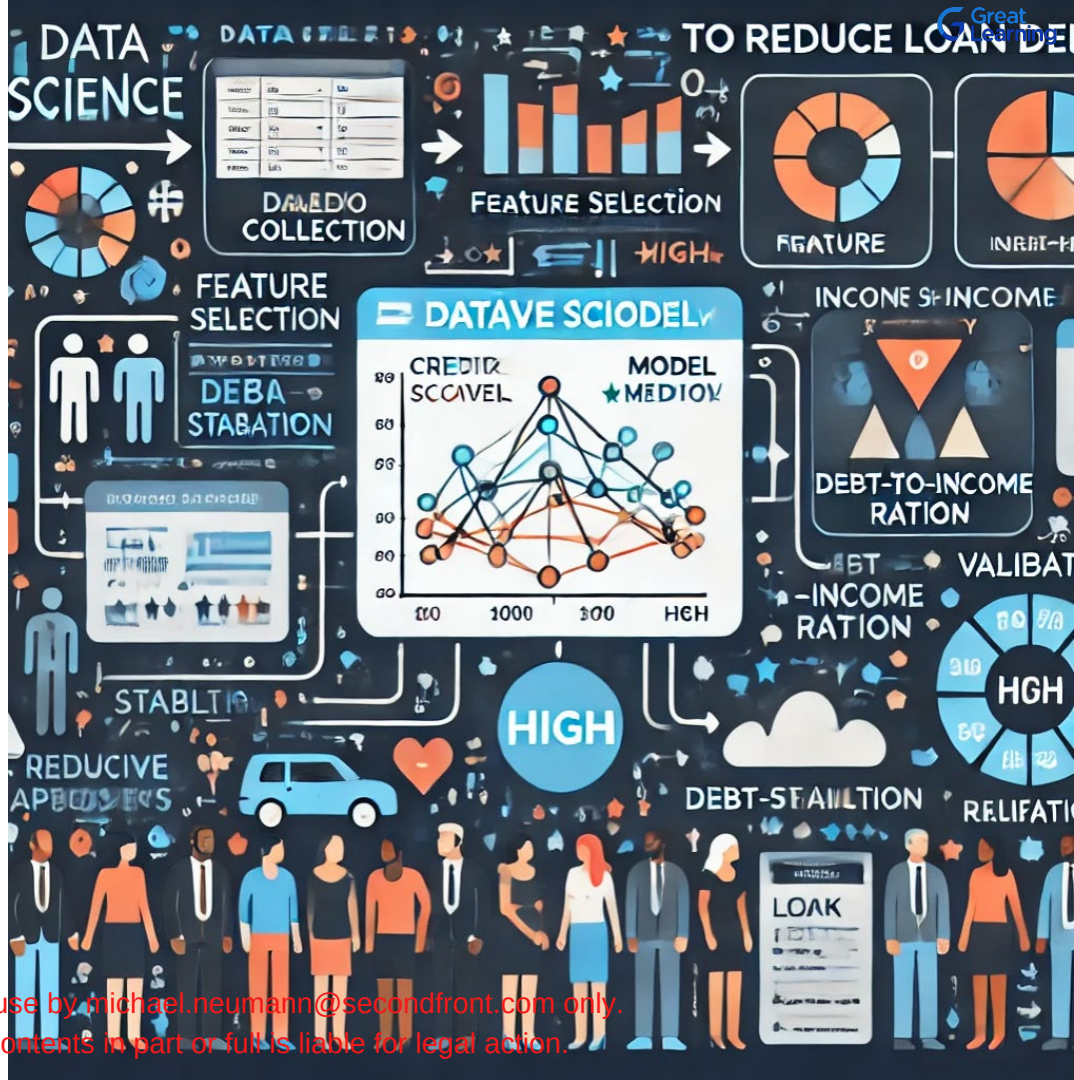
Requirement

- Develop a predictive model and identify key factors that influence the likelihood of loan default.
- Design a robust loan approval framework and reduce the risk of loans turning into defaults.

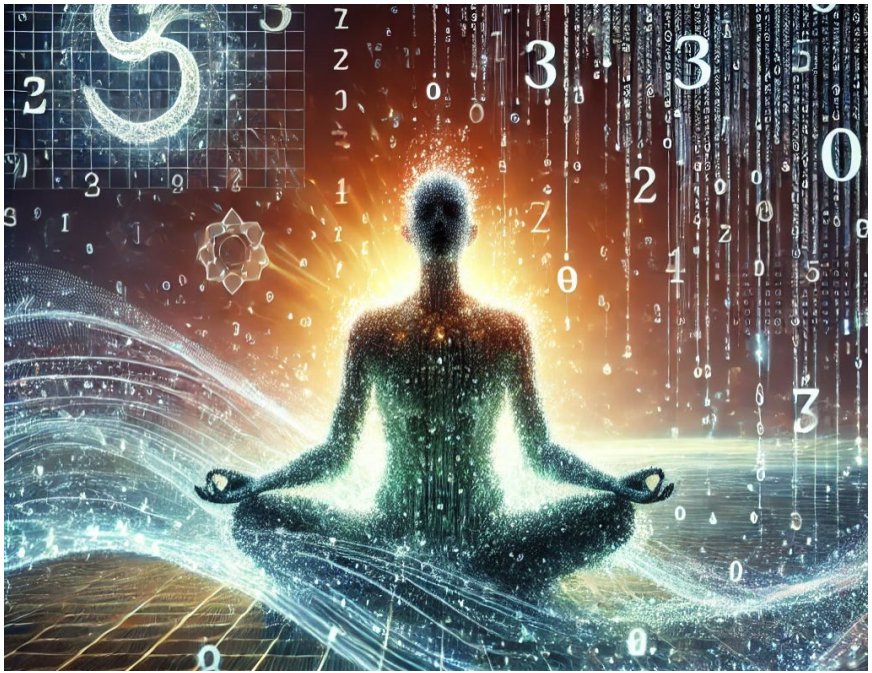


Operationalize Data

* ID: Unique identifier for the borrower.
* Delinquency_Status: Binary indicator whether borrower is delinquent (1/Y, 0/N) Loan_Term: Duration of the loan in months.
* Borrower_Gender: Gender of the borrower.
* Age_Group: Age group of the borrower.
* Loan_Purpose: Reason for taking the loan.
* Home_Status: Borrower's home ownership status.
* Credit_Score_Range: Range of the borrower's credit score.
* Income: Annual income of the borrower (in $1000s).
* Loan_Amount: Loan amount approved for the borrower.

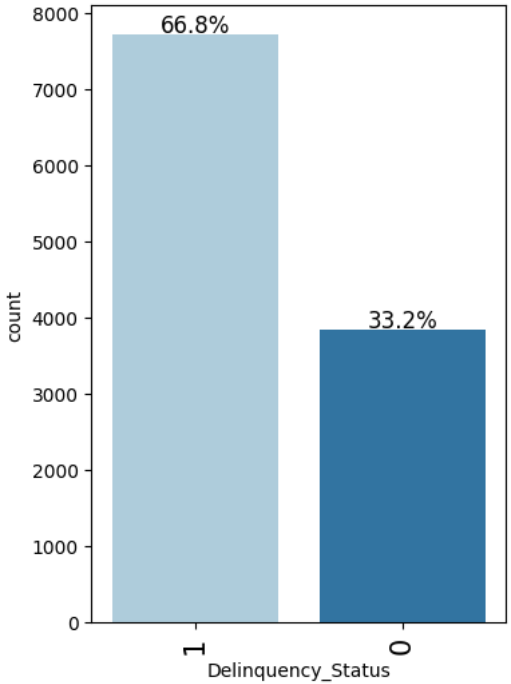
What Other Data Might We Want?

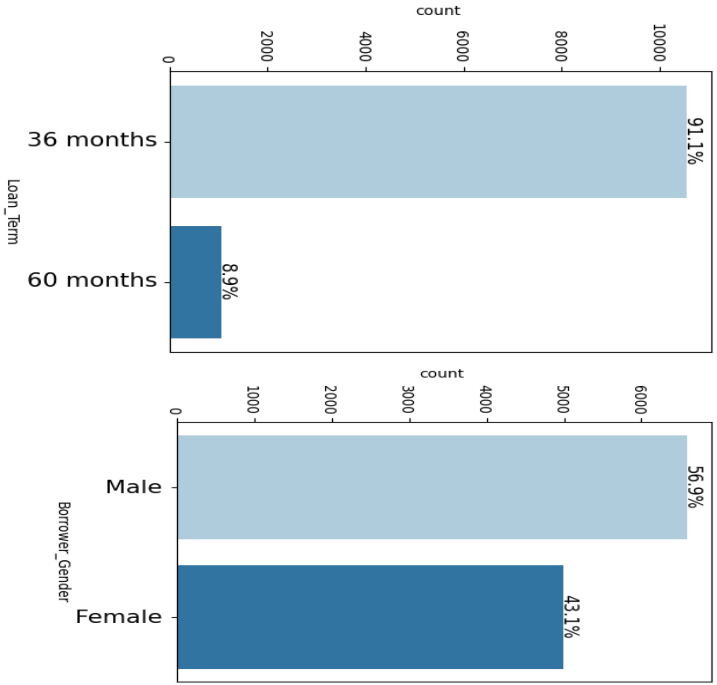
Becoming "One" with the Data

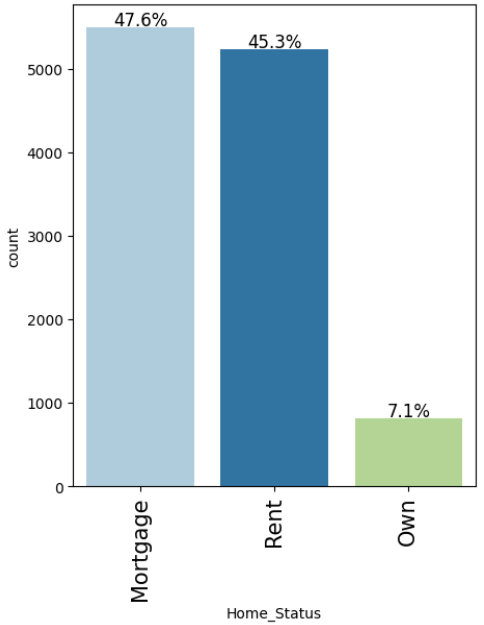


- Look at data headers `loan.head()`
- Shape of the data `loan.shape()`
- Data types `loan.info()`
- Duplicate IDs `loan["ID"].nunique()`
- Missing values `loan.isnull().sum()`
- Exploratory Data Analysis (EDA) - Visualize

Some Observations

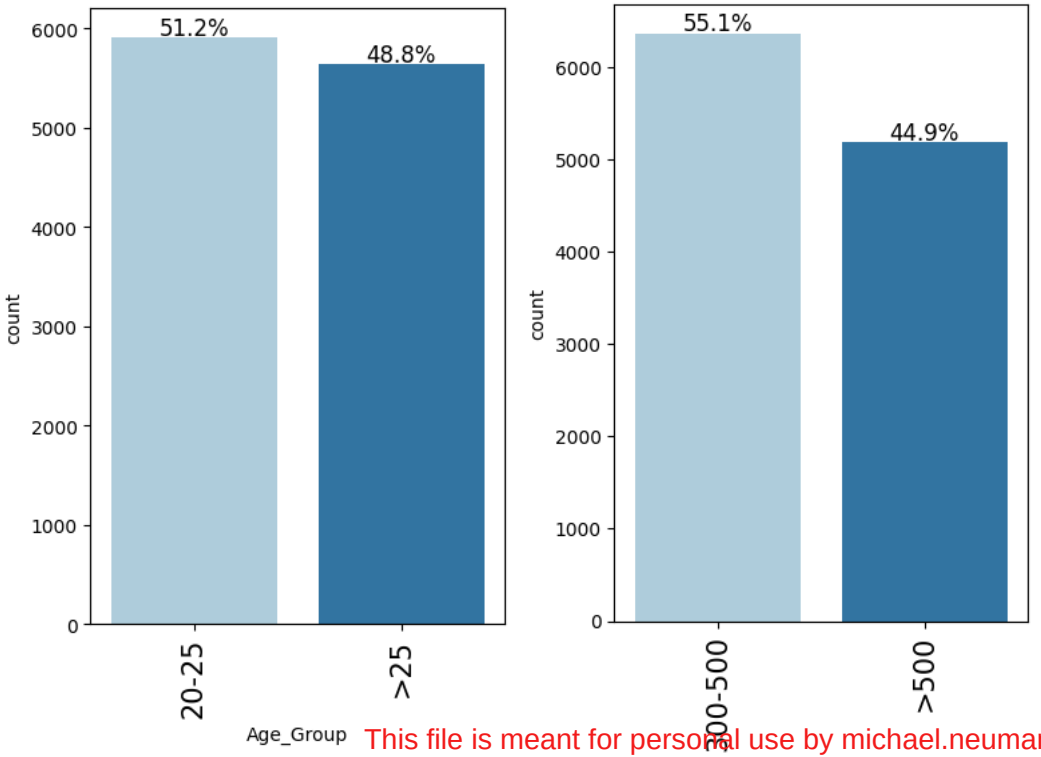






66.8% of Customers Delinquent

Some Observations



Purpose of EDA

1. Data Structure and Content
2. Data Quality
3. Summary Statistics
4. Relationships Btwn Variables
5. Distribution and Patterns
6. Target Variable Analysis (for Supervised Learning)
7. Feature Engineering Opportunities
8. Anomalies and Potential Bias
9. Contextual Insights

Data Cleaning

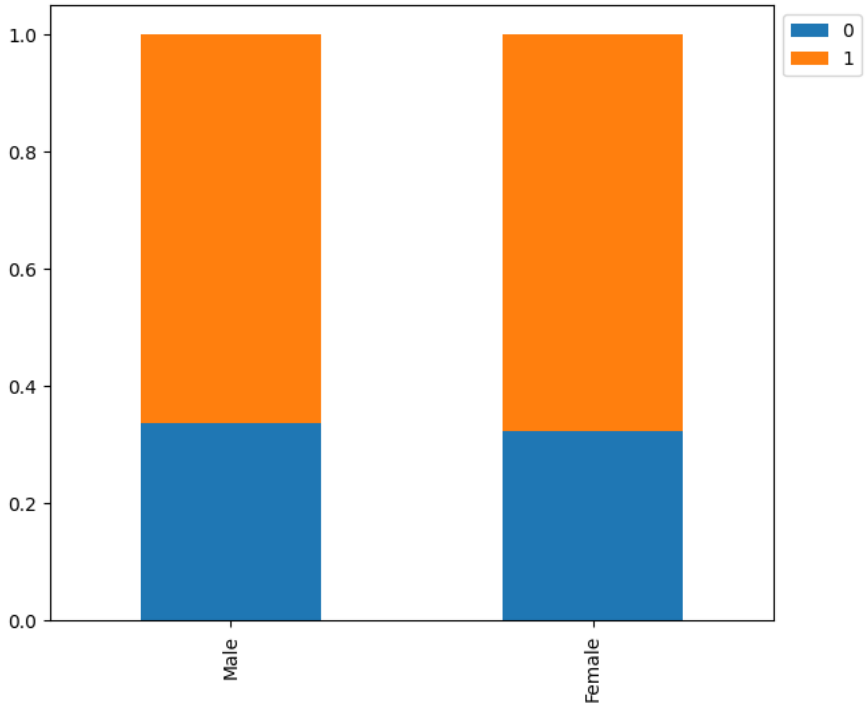
Example: Loan Purpose

```
'House', 'Car',
'Personal', 'Other',
'Wedding', 'Medical',
'other'
```

Identifying and correcting errors, inconsistencies, and inaccuracies in a dataset to ensure it is complete, accurate, and ready for analysis or modeling.

```
loan["Loan_Purpose"].replace("other", "Other", inplace=True)
```

Bivariate Relationships



We examine bivariate relationships to understand dependencies, correlations, or interactions between variables, which helps in feature selection, engineering, and identifying patterns to improve model performance/interpretability.

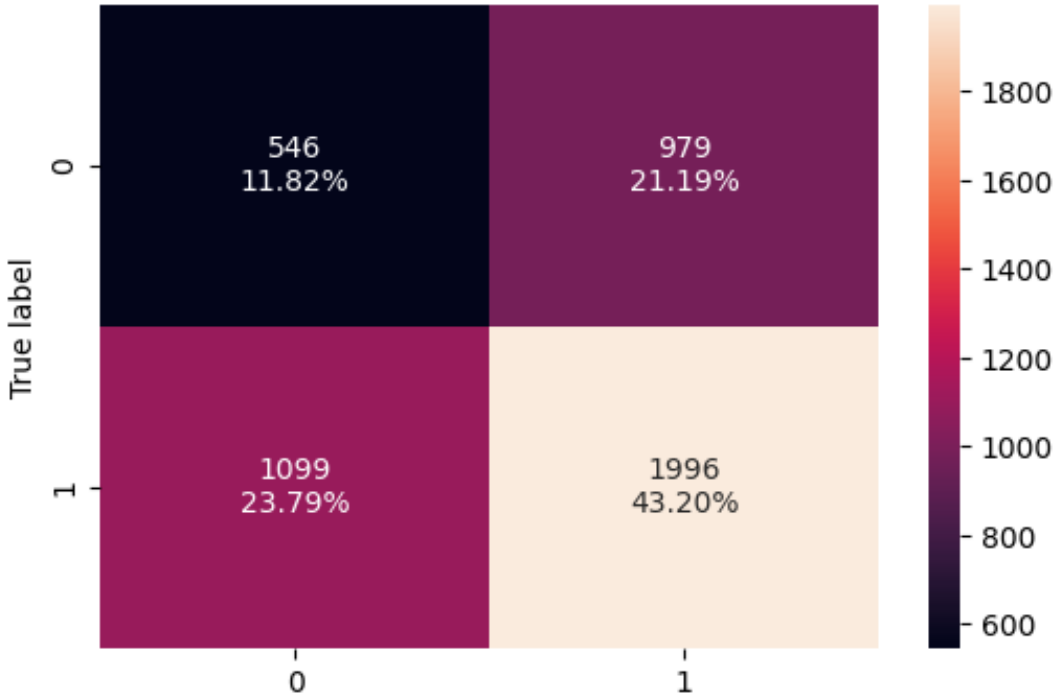
Analytics

Develop mathematical and computational models that analyze data, identify patterns, and make predictions or decisions, as the core means for deriving insights and enabling intelligent automation.

Analytic Decisions

1. Randomly split data
2. What type of error do we minimize?
   a) False Positive = Award delinquent loan (predict good, reality bad)
   b) False Negative = Fail to award good loan (predict bad, reality good)
3. It is worse to award a delinquent loan than failing to award a good loan
4. We therefore want to prioritize *RECALL* over *PRECISION*

The Confusion Matrix



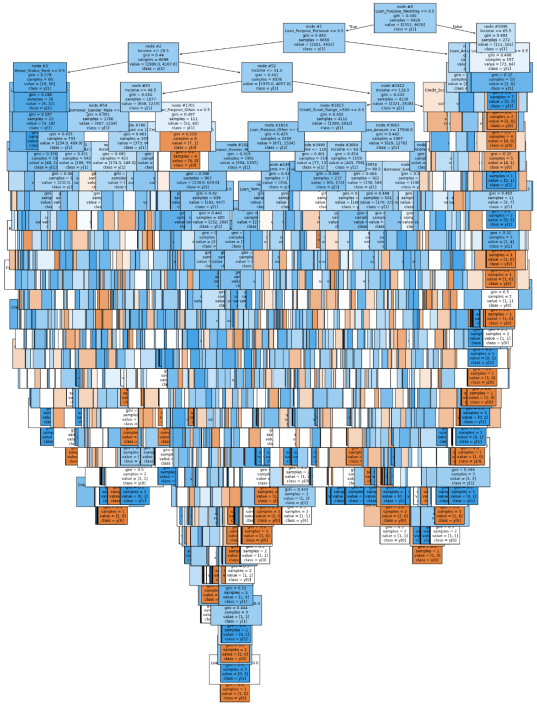
Training Performance

	Accuracy	Recall	Precision	F1
0	0.95612	0.942158	0.991564	0.96623

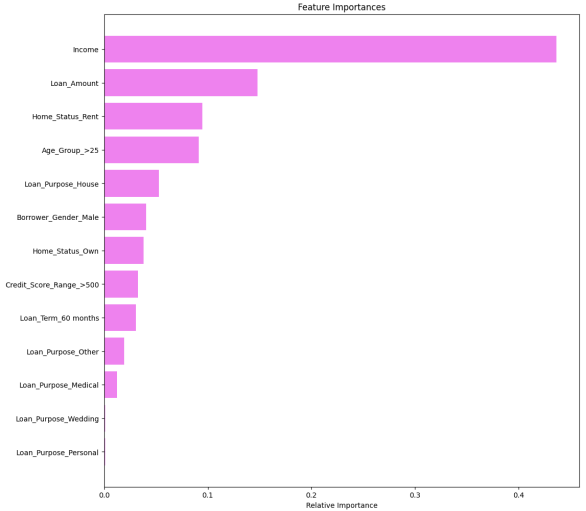
Test Performance

	Accuracy	Recall	Precision	F1
0	0.550216	0.644911	0.670924	0.657661

Visualize the Decision Tree



```
|--- Loan_Purpose_Wedding <= 0.50
|   |--- Loan_Purpose_Personal <= 0.50
|   |   |--- Income <= 20.50
|   |   |   |--- Home_Status_Rent <= 0.50
|   |   |   |   |--- Loan_Purpose_Medical <= 0.50
|   |   |   |   |   |--- Loan_Term_60 months <= 0.50
|   |   |   |   |   |   |--- Loan_Amount <= 22500.00
|   |   |   |   |   |   |   |--- Loan_Amount <= 17500.00
|   |   |   |   |   |   |   |   |--- Loan_Amount <= 12500.00
|   |   |   |   |   |   |   |   |   |--- Age_Group_>25 <= 0.50
|   |   |   |   |   |   |   |   |   |   |--- Home_Status_Own <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 4
|   |   |   |   |   |   |   |   |   |   |--- Home_Status_Own >  0.50
|   |   |   |   |   |   |   |   |   |   |   |--- weights: [0.00, 1.00] class: 1
|   |   |   |   |   |   |   |   |   |--- Age_Group_>25 >  0.50
|   |   |   |   |   |   |   |   |   |   |--- weights: [0.00, 1.00] class: 1
|   |   |   |   |   |   |   |   |--- Loan_Amount >  12500.00
|   |   |   |   |   |   |   |   |   |--- weights: [0.00, 3.00] class: 1
|   |   |   |   |   |   |   |--- Loan_Amount >  17500.00
|   |   |   |   |   |   |   |   |--- Borrower_Gender_Male <= 0.50
|   |   |   |   |   |   |   |   |   |--- Loan_Purpose_House <= 0.50
|   |   |   |   |   |   |   |   |   |   |--- weights: [1.00, 0.00] class: 0
|   |   |   |   |   |   |   |   |   |--- Loan_Purpose_House >  0.50
|   |   |   |   |   |   |   |   |   |   |--- Age_Group_>25 <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- weights: [0.00, 3.00] class: 1
...
|   |   |   |   |   |   |--- weights: [2.00, 0.00] class: 0
|   |   |   |   |--- Borrower_Gender_Male >  0.50
|   |   |   |   |   |--- weights: [0.00, 7.00] class: 1
```



Hyper Parameter Tuning

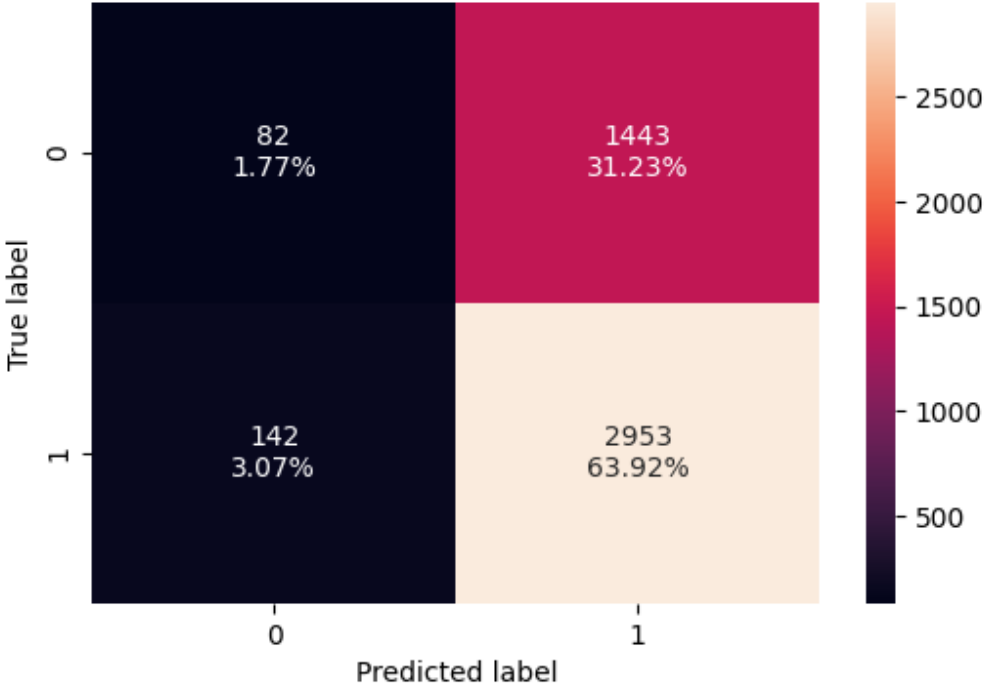
The **maximum depth** of the decision tree, or the number of levels from the root to the deepest leaf. Adjusts over/under-fitting

The **criterion** is the function used to measure the quality of a split. "*gini*" is faster. "*entropy*" provides better understanding/splits

The **splitter** strategy used to select the split at each node. "*best*" is more accurate. "*random*" reduces overfitting noisy data

The **minimum impurity decrease** required for a split to be performed. prunes unnecessary splits to simplify, reduce overfitting, & explainability

Optimize Hyper-Parameters



Initial Test Performance

	Accuracy	Recall	Precision	F1
0	0.550216	0.644911	0.670924	0.657661

Tuned Test Performance

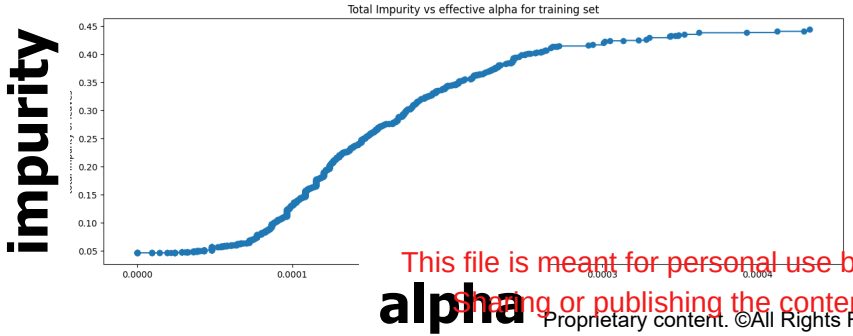
	Accuracy	Recall	Precision	F1
0	0.656926	0.95412	0.671747	0.788413

Complexity Reduction

We reduce **complexity**, evaluating whether a tree branch (cost) is higher than its benefit (impurity reduction / accuracy).

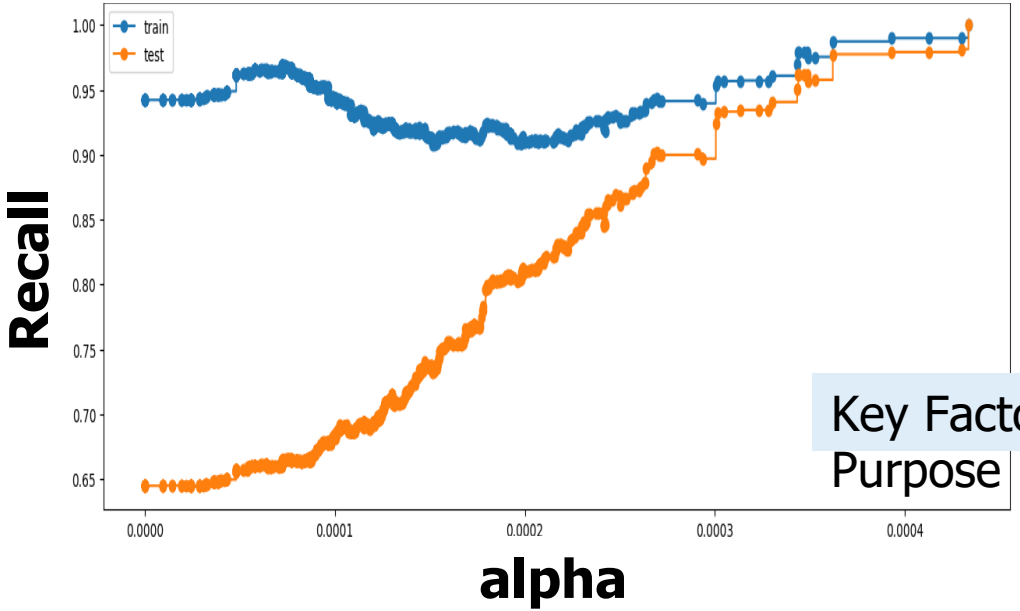
Pruning data controls for overfitting and generalizes better to new data.

Cost Complexity Pruning (CCP) computes a sequence of possible pruning options, calculating the **alpha** values that represent a penalty for tree complexity. Low alpha is high complexity. "**Impurity**" is error.



Recall vs alpha

Recall vs alpha for training & test data



Initial Test Performance

	Accuracy	Recall	Precision	F1
0	0.550216	0.644911	0.670924	0.657661

Tuned Test Performance

	Accuracy	Recall	Precision	F1
0	0.656926	0.95412	0.671747	0.788413

Post-Pruning Test

	Accuracy	Recall	Precision	F1
0	0.669913	1.0	0.669913	0.802333

Key Factors: FICO, Duration, Gender, Purpose