Identifying Machine Learning Models to Predict Serious Credit Delinquency

Noah Keogh

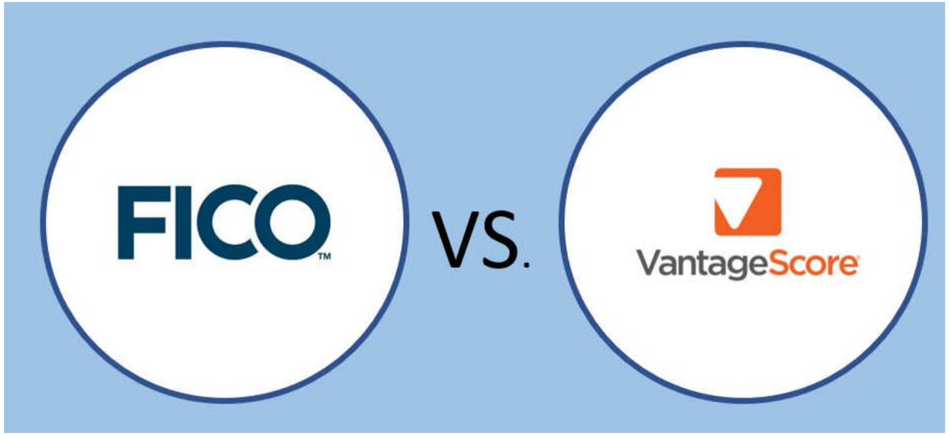
What is a Credit Score?

- Created in 1989 with Introduction of FICO score [1].
- Gives quantitative measure to a borrower’s potential riskiness [2].
- Scores range from 300 to 850 [3].

How do Models Calculate Credit Score?

Common Credit Score Factors:

1) Payment history
2) Credit utilization ratio
3) Total debt
4) Credit mix
5) Account age/depth of credit
6) Hard inquiries

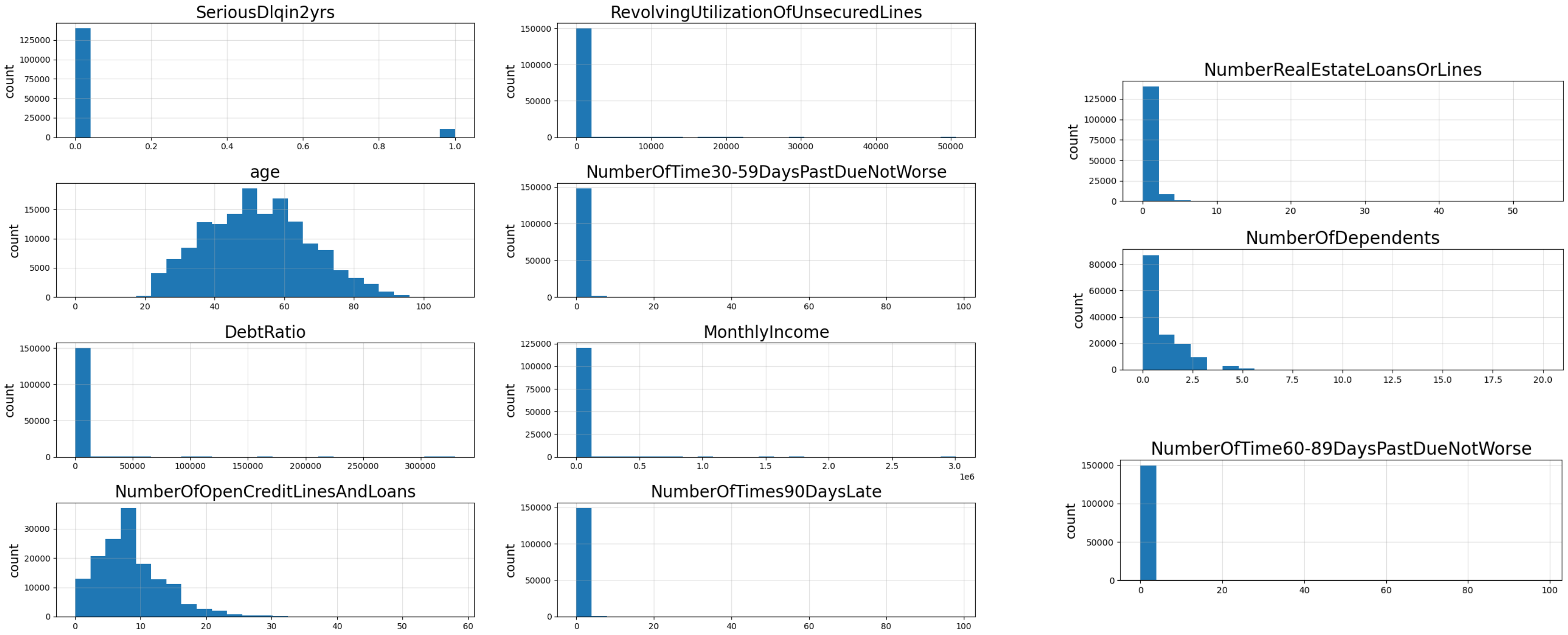


Business Objective

Supervised Model: To classify a given person as being likely or not likely to experience serious delinquency.

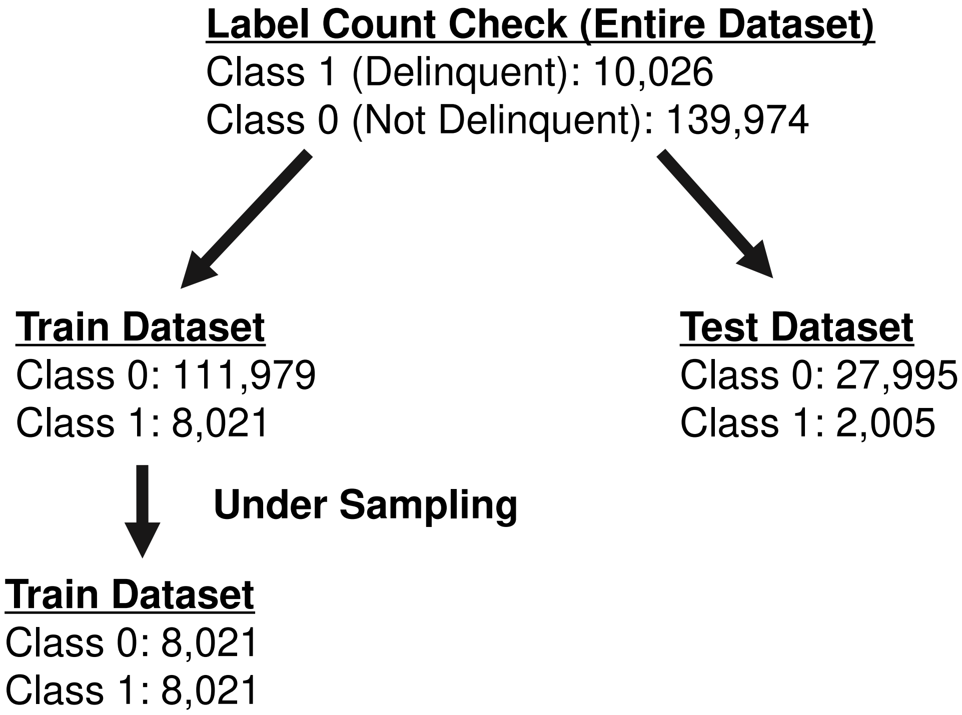
Unsupervised Model: Any financial tendencies or behaviors of people in the dataset. Do these tendencies have any correlation to risk of serious delinquency?

Dataset



- Publicly Available on Kaggle
- 11 Features (one considered label)

Data Wrangling



Label Count Check (Entire Dataset)
Class 1 (Delinquent): 10,026
Class 0 (Not Delinquent): 139,974

Train Dataset
Class 0: 111,979
Class 1: 8,021

Under Sampling

Train Dataset
Class 0: 8,021
Class 1: 8,021

Test Dataset
Class 0: 27,995
Class 1: 2,005

Data Wrangling

1) Missing Values (Entire Dataset):

Monthly Income: 29,731

Number of Dependents: 3,924

2) Label Balancing (under sampling)

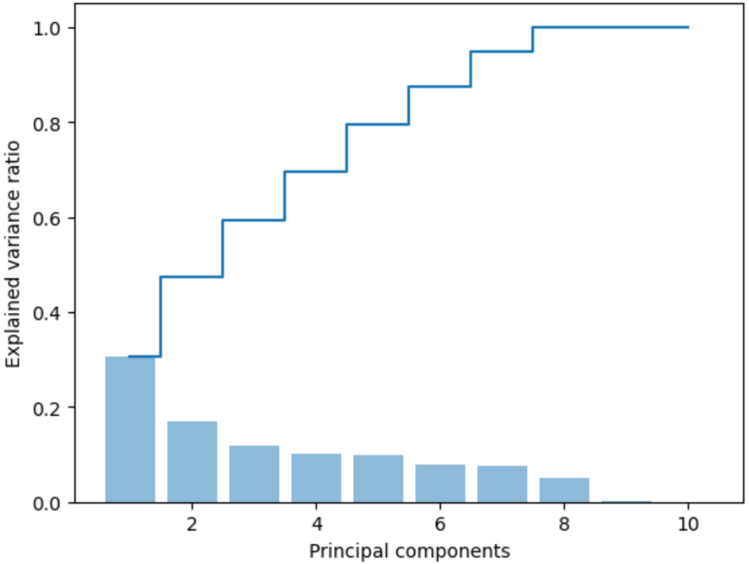
3) Feature Standardization

4) PCA Dimensionality Reduction

10 dimensions 7 dimensions

(94.8% explained variance)

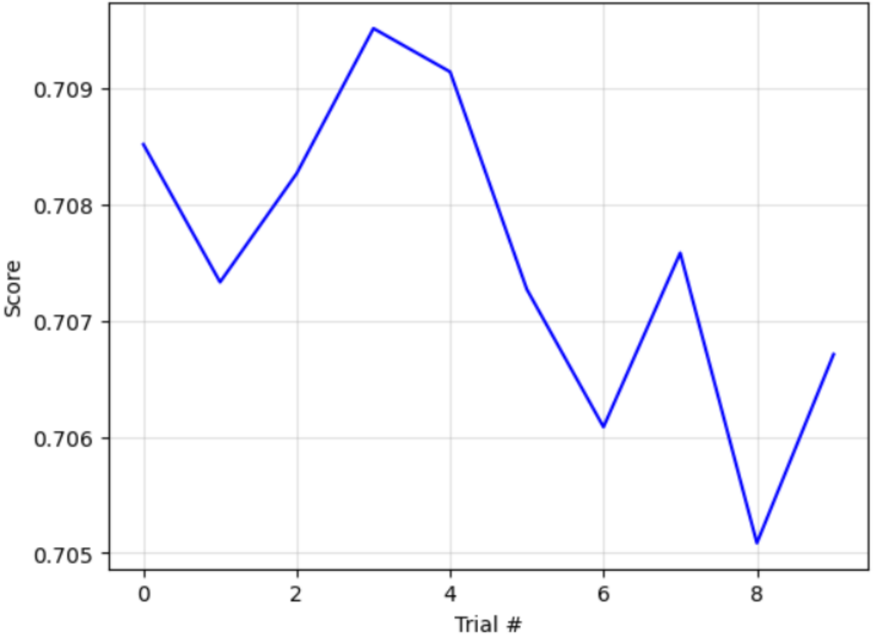
```
Explained Variance Ratio:
[0.30587879 0.16867738 0.11931453 0.10115368 0.09935303 0.07923475
 0.07456405 0.05026824 0.00102548 0.00053007]
Cumulative Sum of EVR:
 [0.30587879 0.47455617 0.5938707  0.69502438 0.79437741 0.87361216
 0.94817621 0.99844445 0.99946993 1.        ]
```



Model Testing

Data Modeling: Logistic Regression

```
Average CV Score: 0.70754917897229
```



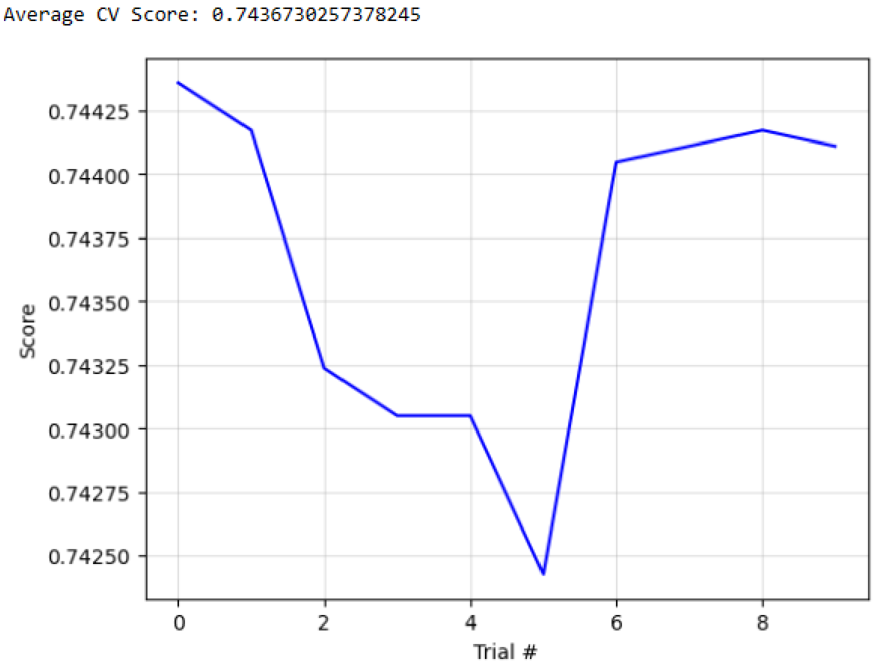
Average Accuracy: 70.75%

Nested Cross Validation:
10 trials conducted

Parameters Tested

1) C: 1, 10, 100
2) Solver: lbfgs, liblinear

Data Modeling: Support Vector Machine (SVM)



```
Average CV Score: 0.7436730257378245
```

Average Accuracy: 74.37%

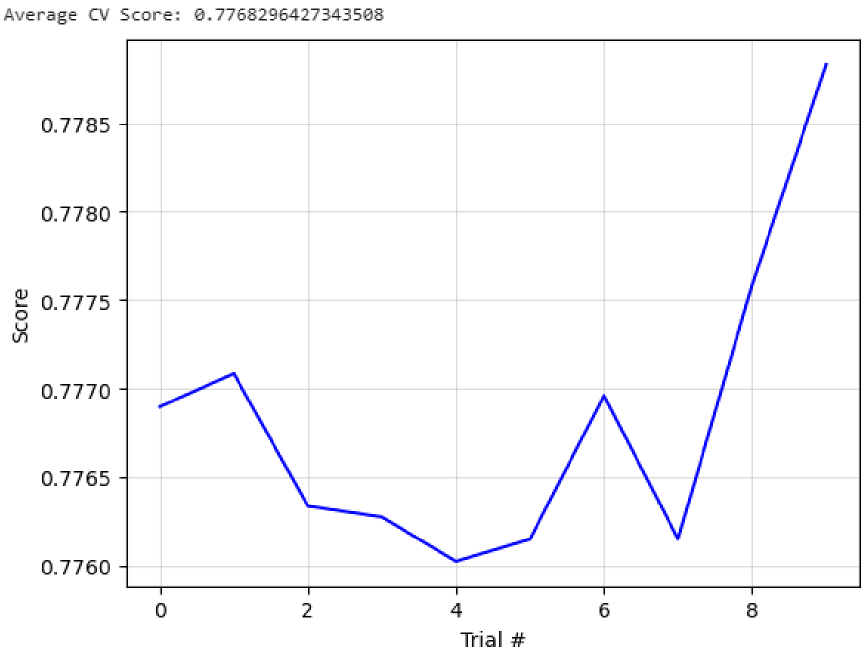
Nested Cross Validation:

10 trials conducted

Parameters Tested

1) C: 1, 10
2) Gamma: 0.01, 0.1
3) Solver: rbf

Data Modeling: Random Forest Classifier



```
Average CV Score: 0.7768296427343508
```

<u>**Average Accuracy: 77.68%**</u>

<u>**Nested Cross Validation:**</u>

10 trials conducted

Parameters Tested

1) Max Depth: 4, 6, 8
2) Min Samples Leaf: 30, 100

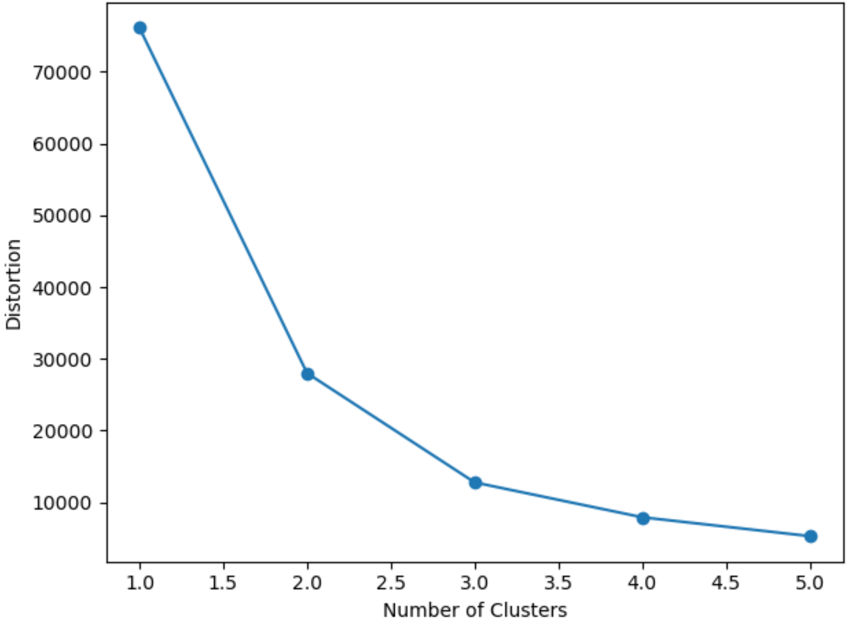
Data Modeling: K-Means-Clustering

1) Dimensionality Reduction

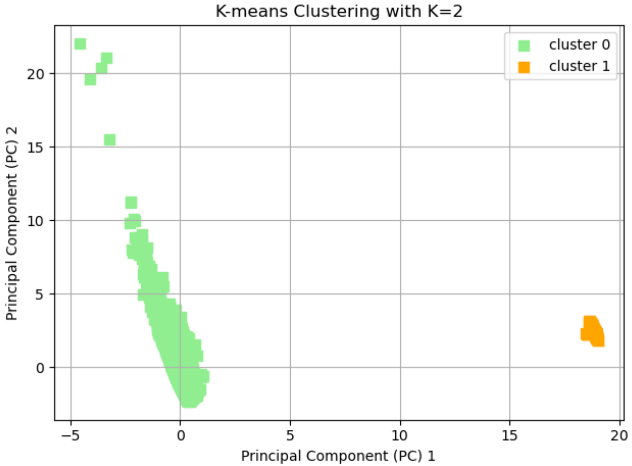
10 dimensions □ 2 dimensions (47.46% variance)

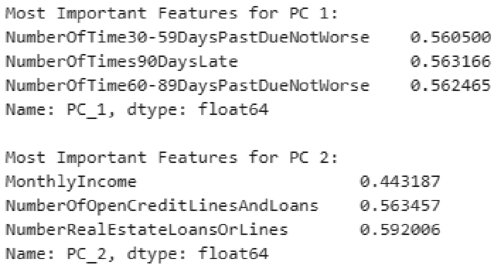
2) Clustering

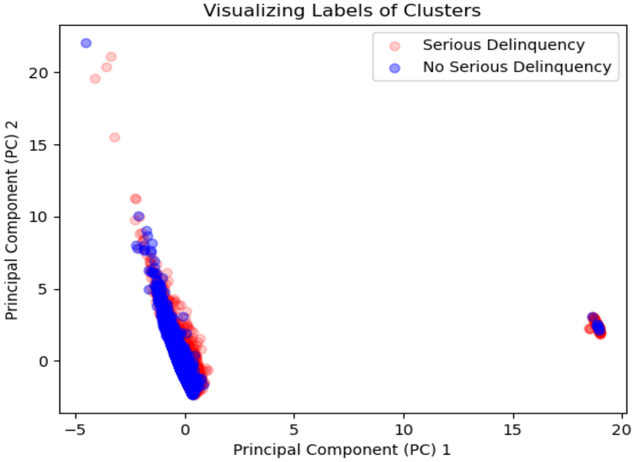
Use K=2 to perform clustering



Data Modeling: K-Means-Clustering

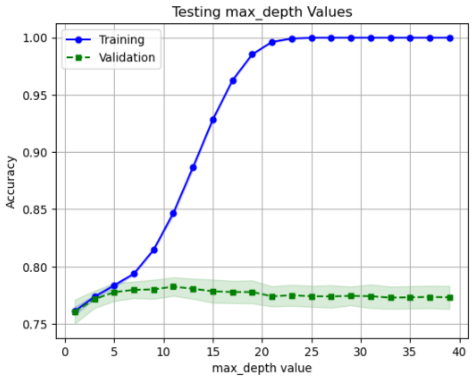

K-means Clustering with K=2

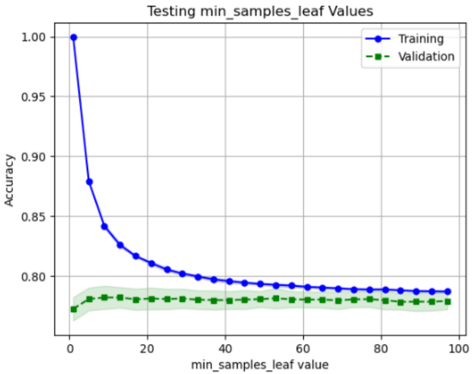

```
Most Important Features for PC 1:
NumberOfTime30-59DaysPastDueNotWorse    0.560500
NumberOfTimes90DaysLate                 0.563166
NumberOfTime60-89DaysPastDueNotWorse    0.562465
Name: PC_1, dtype: float64

Most Important Features for PC 2:
MonthlyIncome                         0.443187
NumberOfOpenCreditLinesAndLoans       0.563457
NumberRealEstateLoansOrLines          0.592006
Name: PC_2, dtype: float64
```

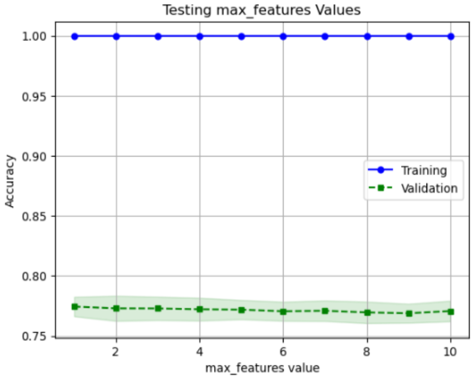

Visualizing Labels of Clusters

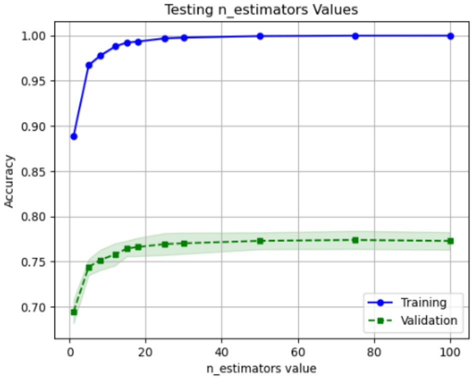
Random Forest Classifier Optimization

Hyperparameter Testing: Validation Curves





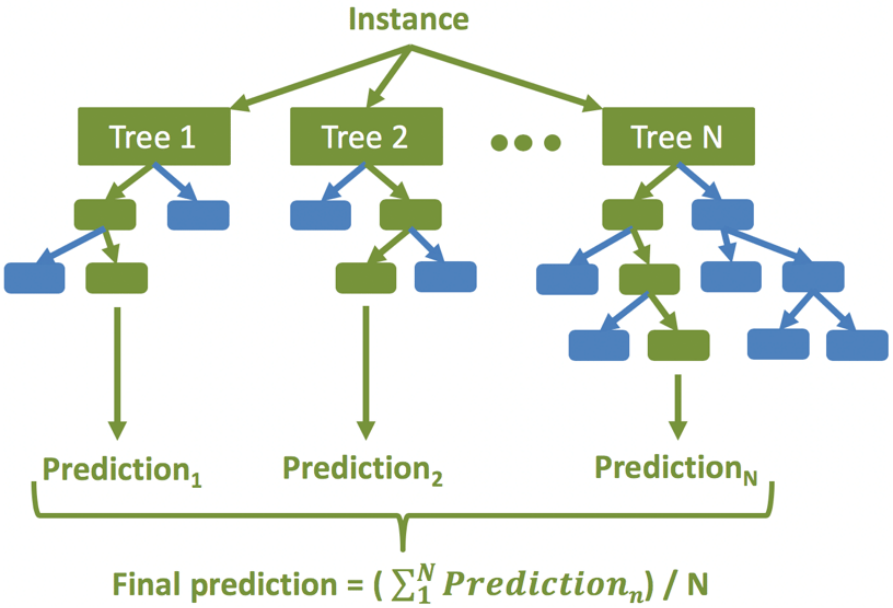




Optimal Values for Each Hyperparameter to Test in Randomized Search

max_depth: 3 to 8
min_samples_leaf: 45 to 100
max_features: 1 to 10
n_estimators: 15 to 25

Randomized CV Search for Final Model



Accuracy in CV: 78.12%

Model Hyperparameters

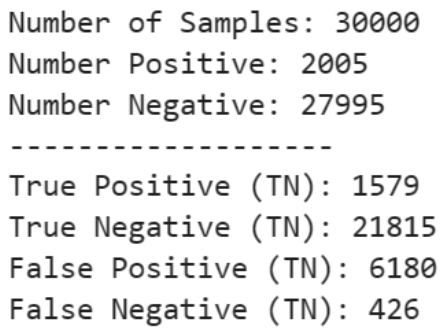
1) max_depth: 7
2) max_features: 5
3) min_samples_leaf: 48
4) n_estimators: 19

Assessing Accuracy

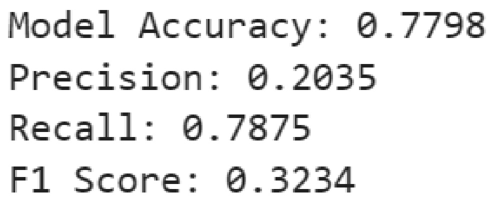
<u>Test Dataset</u>

Class 0: 27,995

Class 1: 2,005



```
Number of Samples: 30000
Number Positive: 2005
Number Negative: 27995
--------------------
True Positive (TN): 1579
True Negative (TN): 21815
False Positive (TN): 6180
False Negative (TN): 426
```



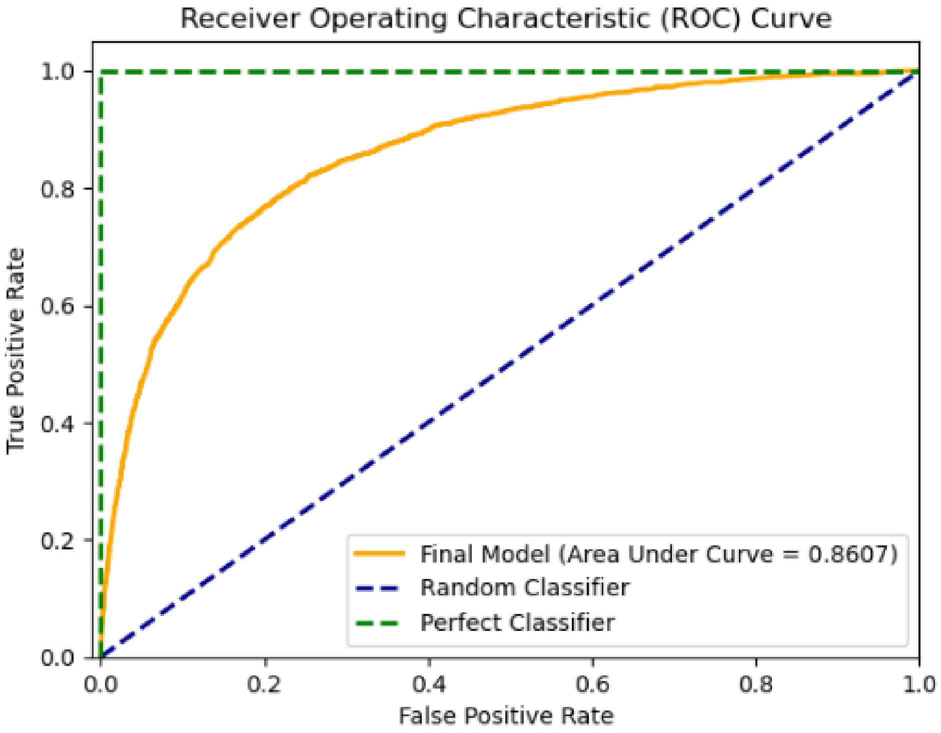
```
Model Accuracy: 0.7798
Precision: 0.2035
Recall: 0.7875
F1 Score: 0.3234
```

Assessing Accuracy

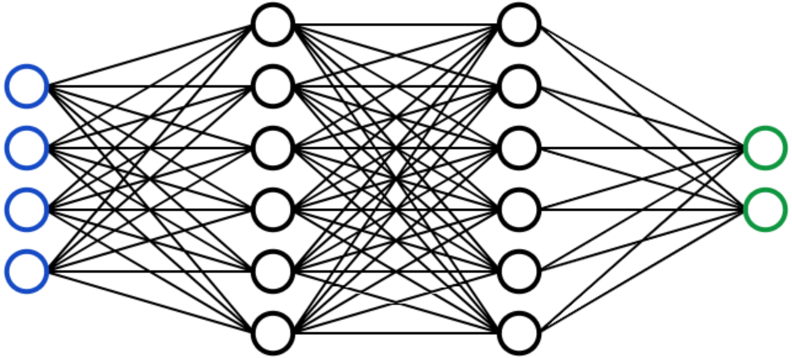
		Predicted 0	Predicted 1
Actual	0	TN	FP
Actual	1	FN	TP

False Positive Rate (FPR) = $\frac{FP}{(FP + TN)}$

True Positive Rate(TPR) = $\frac{TP}{(TP + FN)}$



Future Explorations



- Exploring different models such as neural networks.

- Further optimizing tested models such as logistic regression and SVMs.

- Collecting more data and training examples to improve performance.

Questions?

Sources

[1] https://www.creditrepair.com/blog/education/when-were-credit-scores-invented/

[2] https://www.onemainfinancial.com/resources/credit/credit-scoring-models#:~:text=A%20credit%20scoring%20model%20is%20an%20algorithm%20used,helps%20lenders%20make%20informed%20decisions%20when%20approving%20loans.

[3] https://www.bankrate.com/personal-finance/credit/no-credit-score-zero-credit/