Final Project Presentation

Maria DeCaro, Erin Lloyd, Courtney Gerard, Gabriel Lucey, Aiden Robinson, and Connor Cluett

Data Mining Problem

- Banks want to predict likelihood of “Defaulting” on loan/credit payments
- What factors increase likelihood a customer defaults?
- What risk is tied to a customer?
- Can the Bank reliably predict the customer will repay their loan?
- **Goal: Make a comprehensive machine learning model that can predict who will default on their loan**

Dataset Overview

- Overview:
  - Rows: 255, 347
  - Columns: 18
- Data Columns
  - LoanID: Unique identifier for each loan application
  - Age: Applicant’s age in years
  - Income: Applicant’s income
  - LoanAmount: Amount of the loan requested
  - Credit Score: Applicant’s credit score
  - MonthsEmployed: # of months employed
  - NumCreditLines: # of credit lines
  - InterestRate: Interest rate on the loan
  - LoanTerm: Term of the loan in months
  - DTIRatio: Debt-to-income ratio
  - Education: Applicant’s education level
  - EmploymentType: Applicant's employment type.
  - MaritalStatus: Applicant's marital status.
  - HasMortgage: Whether the applicant has a mortgage.
  - HasDependents: Whether the applicant has dependents.
  - LoanPurpose: Purpose of the loan.
  - HasCoSigner: Whether the loan has a co-signer.
  - Default: Binary indicator of loan default (1 for default, 0 for non-default).

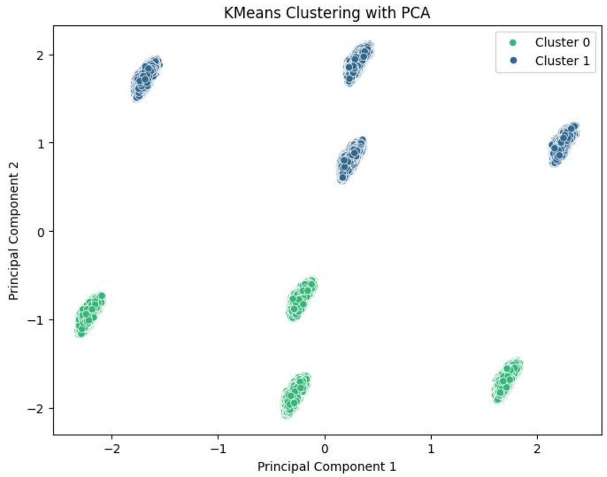
General Preprocessing Steps

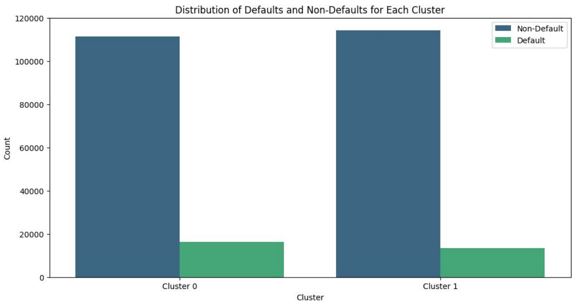
- Check for missing values
  - Nothing missing from our data
- Examined what data type each column was and converted it if needed
  - Categorical data into numeric
- Normalized numeric features using min-max scaling
  - This method scales it to a fixed range between 1 and 0
- Dropped unnecessary columns
  - Only 1 dropped - LoanID
- Saved all of this into a new CSV file

```
Number of Default instances: 29653
Number of Non-Default instances: 225694
Percentage of Default instances: 11.61%
Percentage of Non-Default instances: 88.39%
```

Dataset Imbalance

- Our dataset has a severe class imbalance
- Why?
  - This dataset reflects the natural distribution of loan outcomes
  - In real world scenarios, defaults are less common compared to successful repayments
- SMOTE
  - Oversampling the minority class





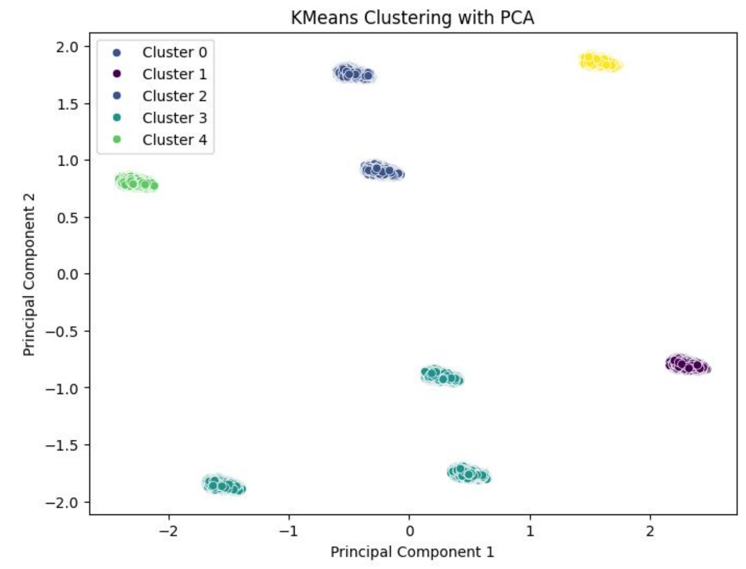
Clustering

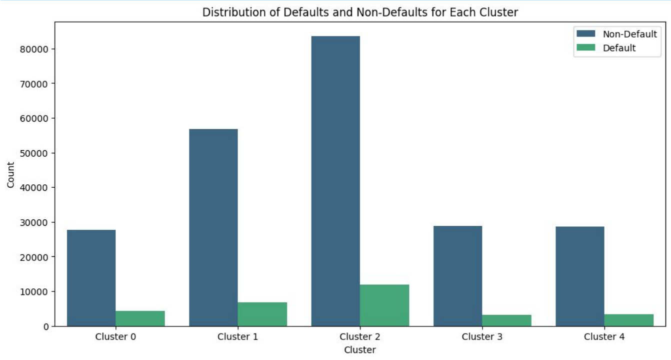
Tried without PCA first
- Uninterpretable Results

Performed K Means clustering with PCA
- 2 clusters
- 5 clusters

Silhouette score for 2 clusters: 0.0317320336106
Silhouette score for 5 clusters: 0.0539852511404

Fails to identify any real patterns due to dataset





```
Empty DataFrame
Columns: [antecedents, consequents, antecedent support, consequent support, support, confidence, lift, leverage, conviction, zhangs_metric]
Index: []
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(LoanAmount)	(Income)	0.999996	0.999991	0.999987	0.999991	1.000000	-3.786904e-11	0.999996	-0.000009
1	(Income)	(LoanAmount)	0.999991	0.999996	0.999987	0.999996	1.000000	-3.786904e-11	0.999991	-0.000004
2	(LoanAmount)	(InterestRate)	0.999996	0.999813	0.999809	0.999813	1.000000	-8.141849e-10	0.999996	-0.000187
3	(InterestRate)	(LoanAmount)	0.999813	0.999996	0.999809	0.999996	1.000000	-8.141849e-10	0.999813	-0.000004
4	(InterestRate)	(Income)	0.999813	0.999991	0.999804	0.999991	1.000000	-1.628370e-09	0.999813	-0.000009
...	...	...	...	...	...	...	...	...	...	...
95	(LoanAmount)	(DTIRatio, CreditScore)	0.999996	0.991837	0.991832	0.991837	1.000000	-3.552118e-08	0.999996	-0.008163
96	(DTIRatio, CreditScore)	(Income)	0.991837	0.999991	0.991828	0.999991	1.000000	-7.104237e-08	0.991837	-0.000009
97	(DTIRatio, Income)	(CreditScore)	0.993721	0.998098	0.991828	0.998095	0.999997	-3.237180e-06	0.998290	-0.000520
98	(CreditScore, Income)	(DTIRatio)	0.998090	0.993730	0.991828	0.993726	0.999997	-3.275201e-06	0.999477	-0.001726
99	(DTIRatio)	(CreditScore, Income)	0.993730	0.998090	0.991828	0.998086	0.999997	-3.275201e-06	0.998278	-0.000526

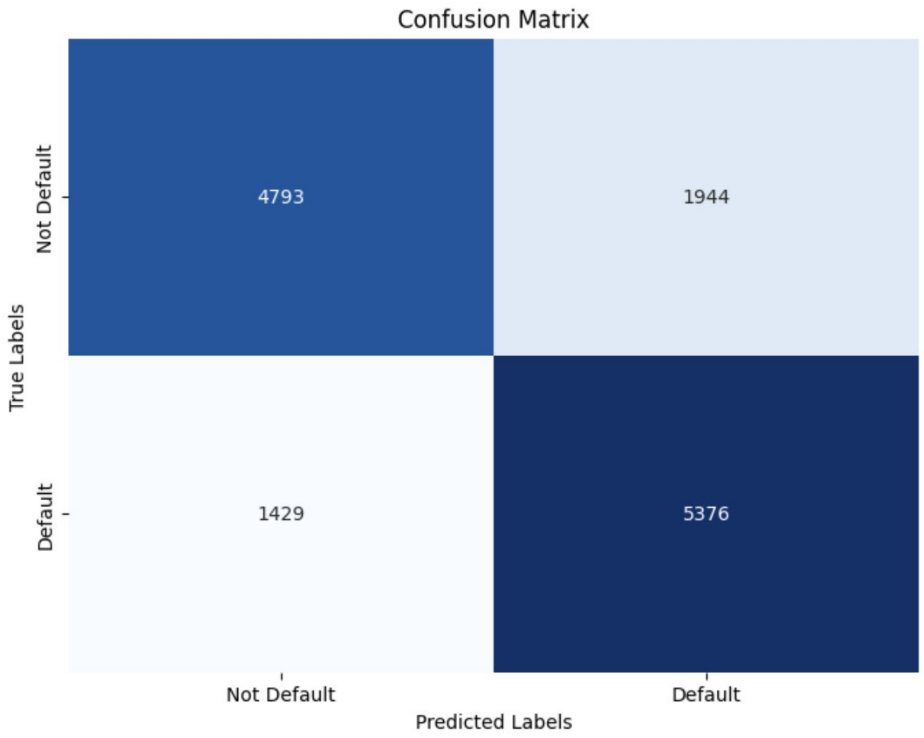
Association Rule Mining

- Can generate rules
- No rules are generating with the consequent of a person defaulting or not defaulting

- Data exploded when using One-Hot encoding
- Subsetted the data originally because of run time, but now running with 100% of the data there are still no rules being generated

- Could be helpful to understand other relationships in our data but not our business problem

```
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.71      0.74      6737
           1       0.73      0.79      0.76      6805

    accuracy                           0.75     13542
   macro avg       0.75      0.75      0.75     13542
weighted avg       0.75      0.75      0.75     13542
```

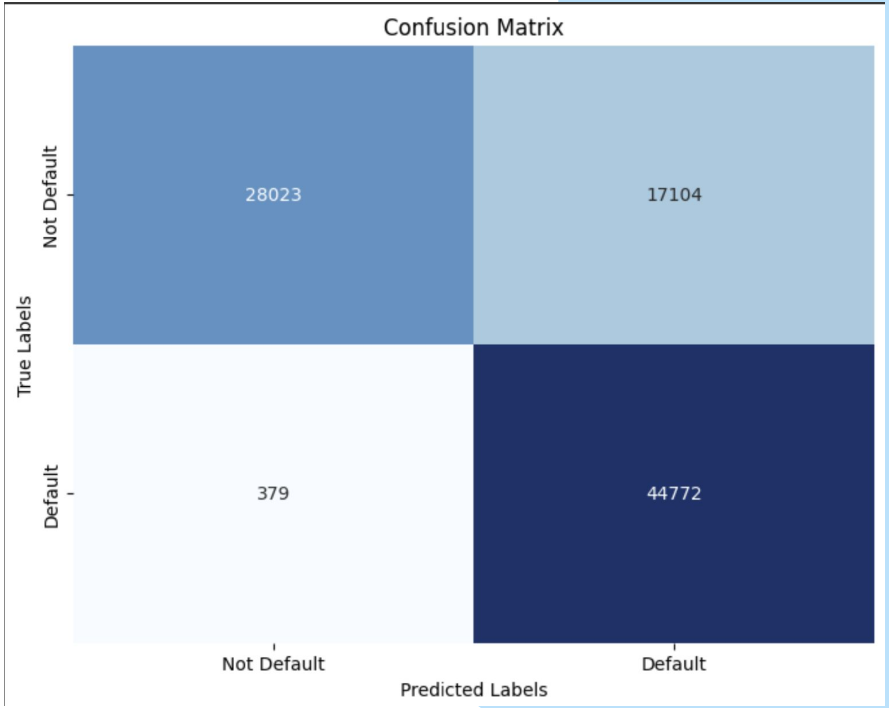


Standard Vector Machine (SVM)

- Data Preprocessing
  - SMOTE
- Model Performance
  - Accuracy of 75% (with smote)
  - Using a subset of 15% of the overall data
  - Long Runtime
- Model Hyperparameters
  - Kernel: RBF
  - C = 1.0
- Without smote, got a 88% accuracy, but super imbalanced, especially when already using a subset of the data

K-Nearest Neighbors (KNN)

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.62      0.76     45127
           1       0.72      0.99      0.84     45151

    accuracy                           0.81     90278
   macro avg       0.86      0.81      0.80     90278
weighted avg       0.86      0.81      0.80     90278
```

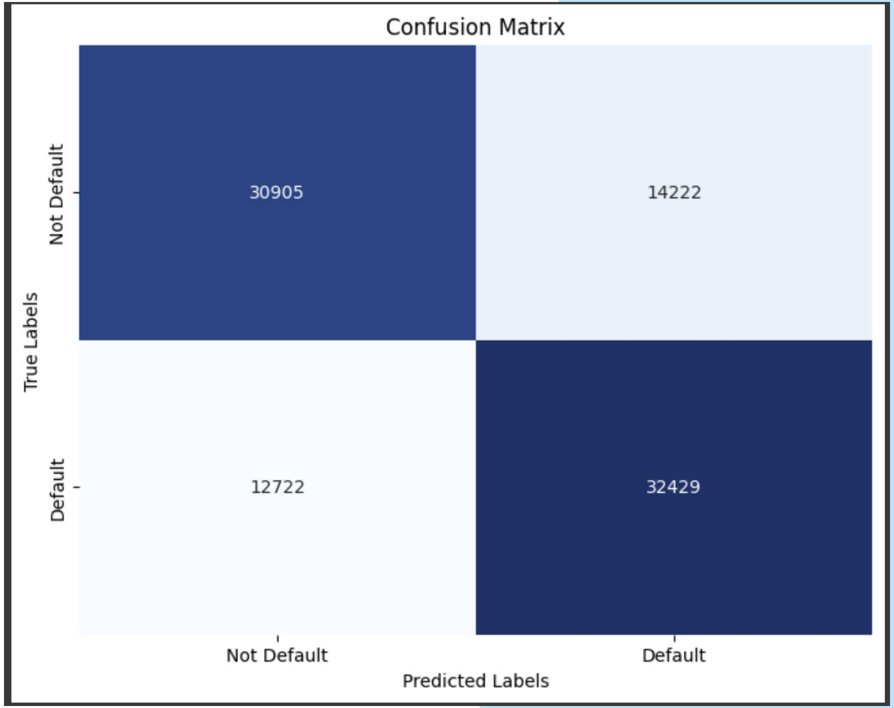


- Without SMOTE
  - ACC: 88%, F1: .07
- With SMOTE
  - ACC: 81%. F1: .80
- SMOTE helps to increase F1 score. KNN here also did an exceptional job at predicting if someone will default.

```
Selected Features and their Importance Scores:
          Feature       Score
0             Age  991.147593
1          Income  343.608125
2      LoanAmount  249.812510
3  MonthsEmployed  315.319187
4    InterestRate  574.348579
```

Classification Report:

	precision	recall	f1-score	support
0	0.71	0.68	0.70	45127
1	0.70	0.72	0.71	45151
accuracy			0.70	90278
macro avg	0.70	0.70	0.70	90278
weighted avg	0.70	0.70	0.70	90278

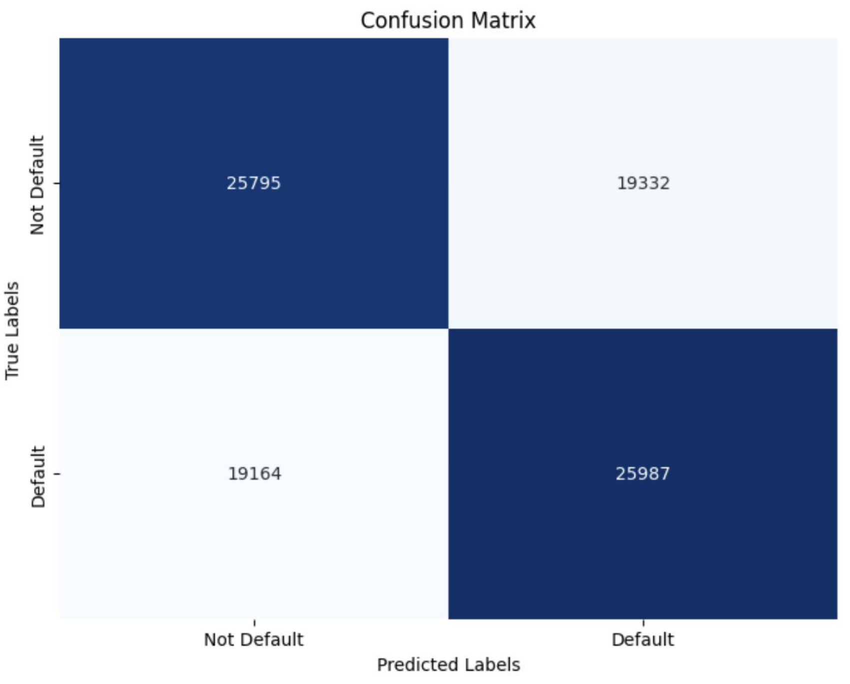


Logistic Regression

- Without SMOTE
  - ACC: 89%, F1: .06
- With SMOTE
  - ACC: 70%. F1: .70
- SMOTE helps to increase F1 score, but Logistic Regression struggle because relationships are not defined by a boundary

Multinomial Naive Bayes

```
Classification Report:
              precision    recall  f1-score   support

           0       0.57      0.57      0.57     45127
           1       0.57      0.58      0.57     45151

    accuracy                           0.57     90278
   macro avg       0.57      0.57      0.57     90278
weighted avg       0.57      0.57      0.57     90278
```

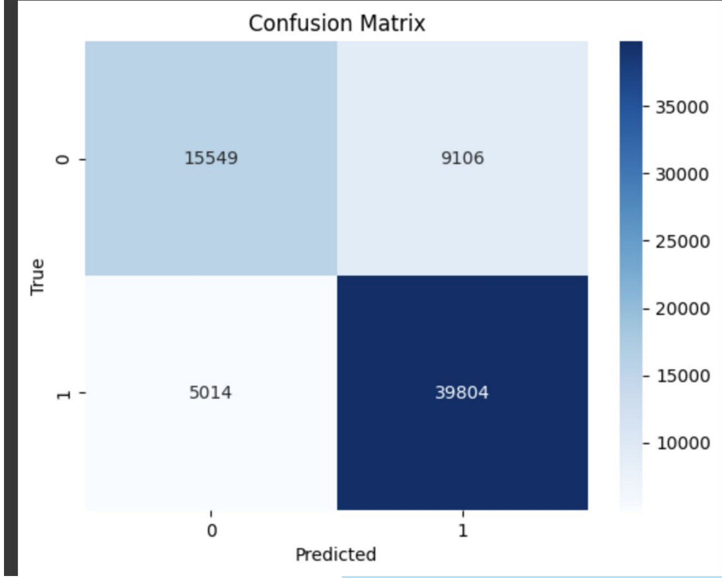


- Data Preprocessing
  - SMOTE
- Model Performance
  - Used GridSearch and still not getting any higher of an accuracy
  - Overall not great accuracy at 57%
- Overall, I think this is just not the best model for this data
- Guassian Naive Bayes may be a better approach!

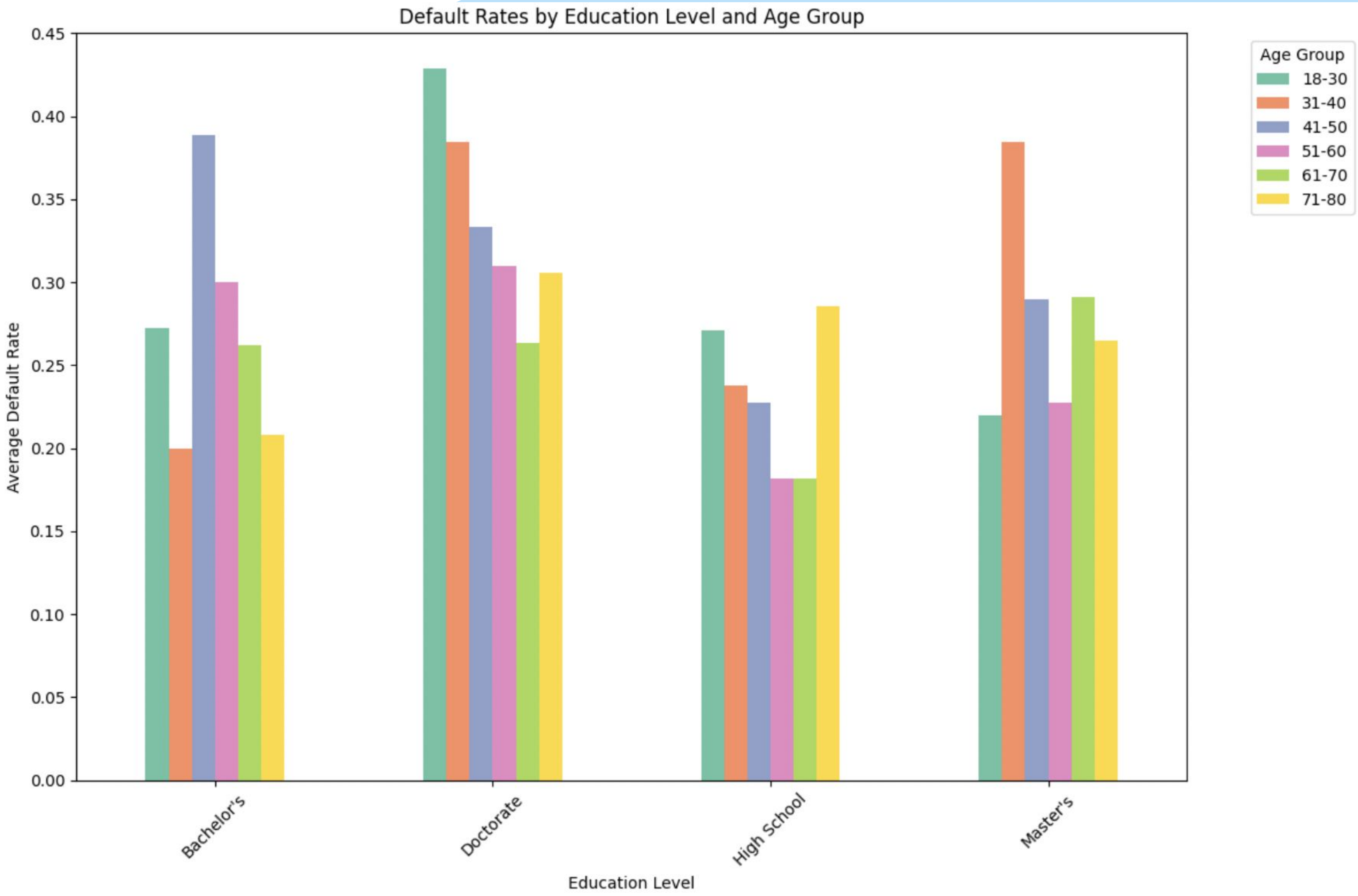
Gaussian Naive Bayes

Classification Report:

	precision	recall	f1-score	support
0	0.76	0.63	0.69	24655
1	0.81	0.89	0.85	44818
accuracy			0.80	69473
macro avg	0.78	0.76	0.77	69473
weighted avg	0.79	0.80	0.79	69473



- Accuracy was originally 68.08%
- SMOTE helped address the class imbalance and improved the Naive Bayes classifier's performance in predicting loan defaults
- After using Borderline - SMOTE the accuracy increased from 72.63% to 74.03%
- Used SMOTE-ENN and got an accuracy around 80%
  - ENN- Edited Nearest Neighbors

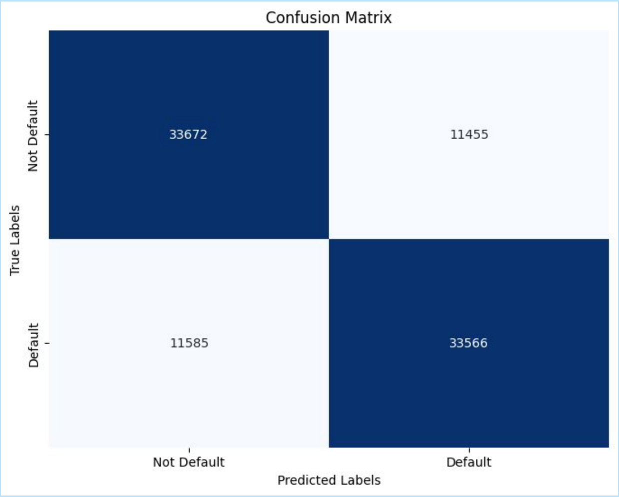

Default Rates by Education Level and Age Group
Average Default Rate
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Bachelor's
Doctorate
High School
Master's
Education Level
Age Group
18-30
31-40
41-50
51-60
61-70
71-80

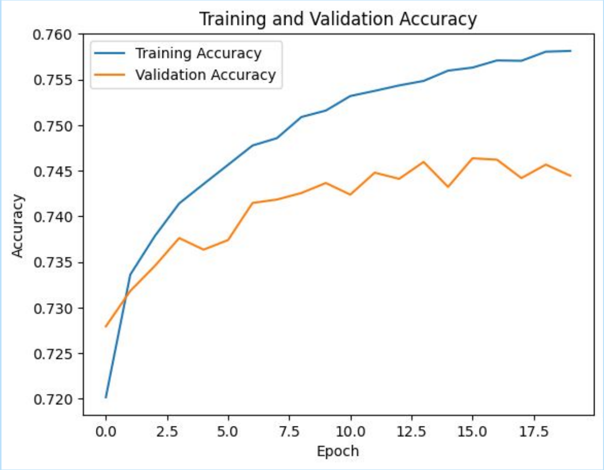
Deep Learning

- SMOTE for dataset imbalance
- Attempted RUS and class weights

- ANN Model
- Input Layer with 64 Neurons
- Hidden layer with 32 Neurons
- Output Layer with 1 Neuron
- Adam optimizer to adjust learning rates

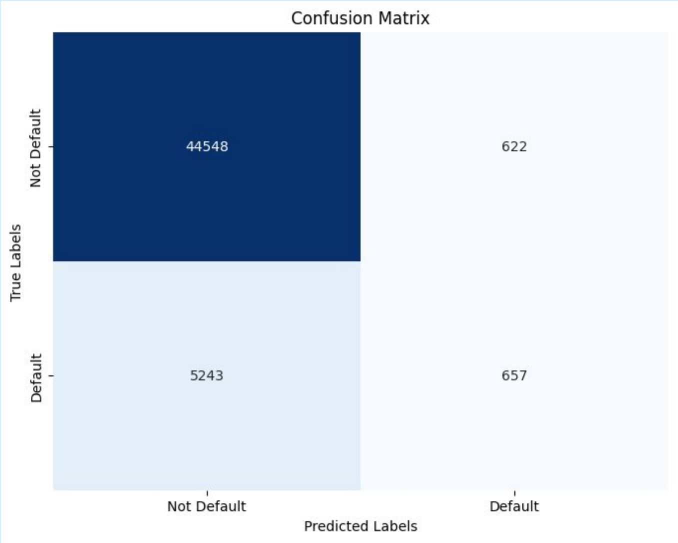
- 20 Epochs





Classification Report:

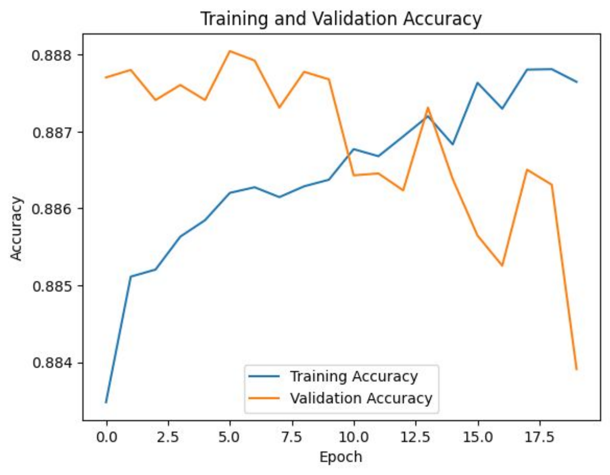
	precision	recall	f1-score	support
0	0.74	0.75	0.75	45127
1	0.75	0.74	0.74	45151
accuracy			0.74	90278
macro avg	0.74	0.74	0.74	90278
weighted avg	0.74	0.74	0.74	90278



Deep Learning

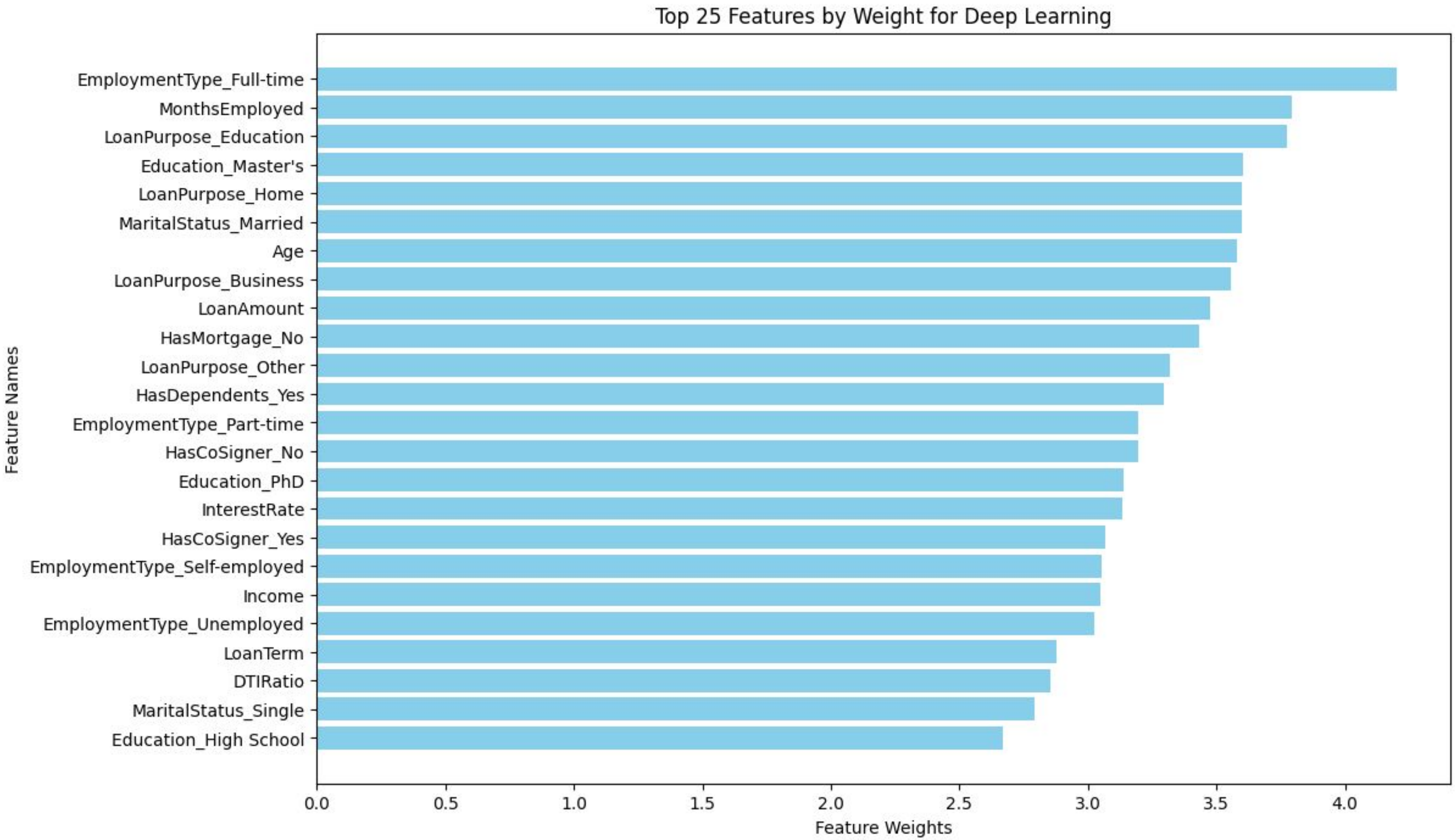
Without SMOTE

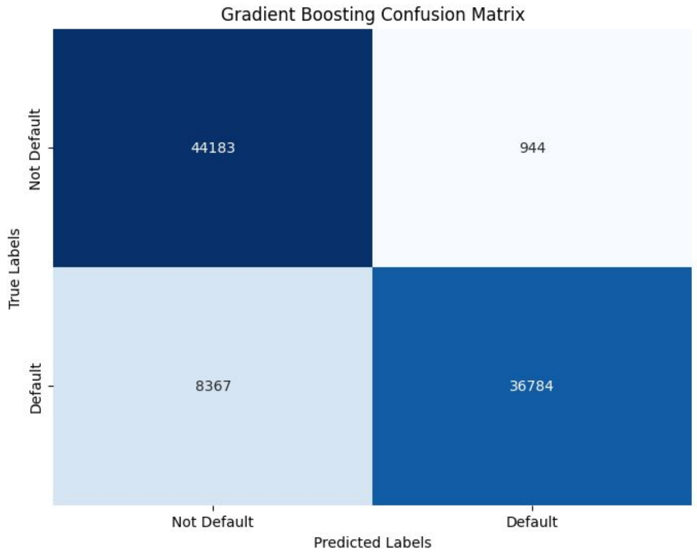
- Higher Accuracy - 0.89
- F1 score for default instances - 0.18
- Major Decrease in Validation Accuracy



Classification Report:

	precision	recall	f1-score	support
0	0.89	0.99	0.94	45170
1	0.51	0.11	0.18	5900
accuracy			0.89	51070
macro avg	0.70	0.55	0.56	51070
weighted avg	0.85	0.89	0.85	51070


Top 25 Features by Weight for Deep Learning
EmploymentType_Full-time
MonthsEmployed
LoanPurpose_Education
Education_Master's
LoanPurpose_Home
MaritalStatus_Married
Age
LoanPurpose_Business
LoanAmount
HasMortgage_No
LoanPurpose_Other
HasDependents_Yes
EmploymentType_Part-time
HasCoSigner_No
Education_PhD
InterestRate
HasCoSigner_Yes
EmploymentType_Self-employed
Income
EmploymentType_Unemployed
LoanTerm
DTIRatio
MaritalStatus_Single
Education_High School
Feature Names
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Feature Weights

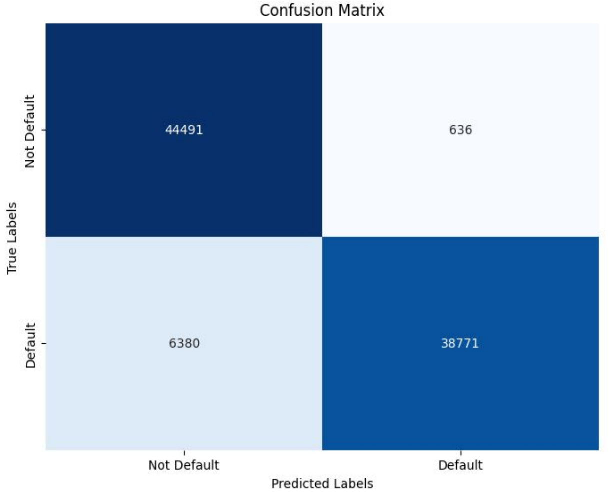


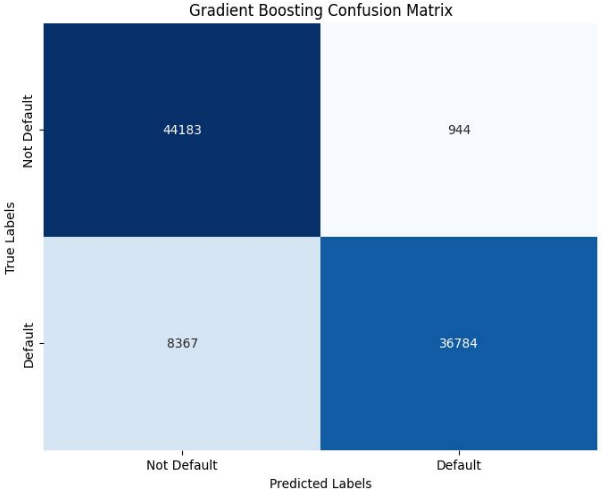

```
Gradient Boosting Accuracy: 0.8968630231064046

Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.98      0.90     45127
           1       0.97      0.81      0.89     45151

    accuracy                           0.90     90278
   macro avg       0.91      0.90      0.90     90278
weighted avg       0.91      0.90      0.90     90278

Frequency of Predictions for Gradient Boosting:
0    52550
1    37728
```

Gradient Boosting Machine

- Data Preprocessing
  - One-hot encoding
  - SMOTE
- Model Performance
  - Overall High with 89% Accuracy
  - Initial was 8% F1 Score

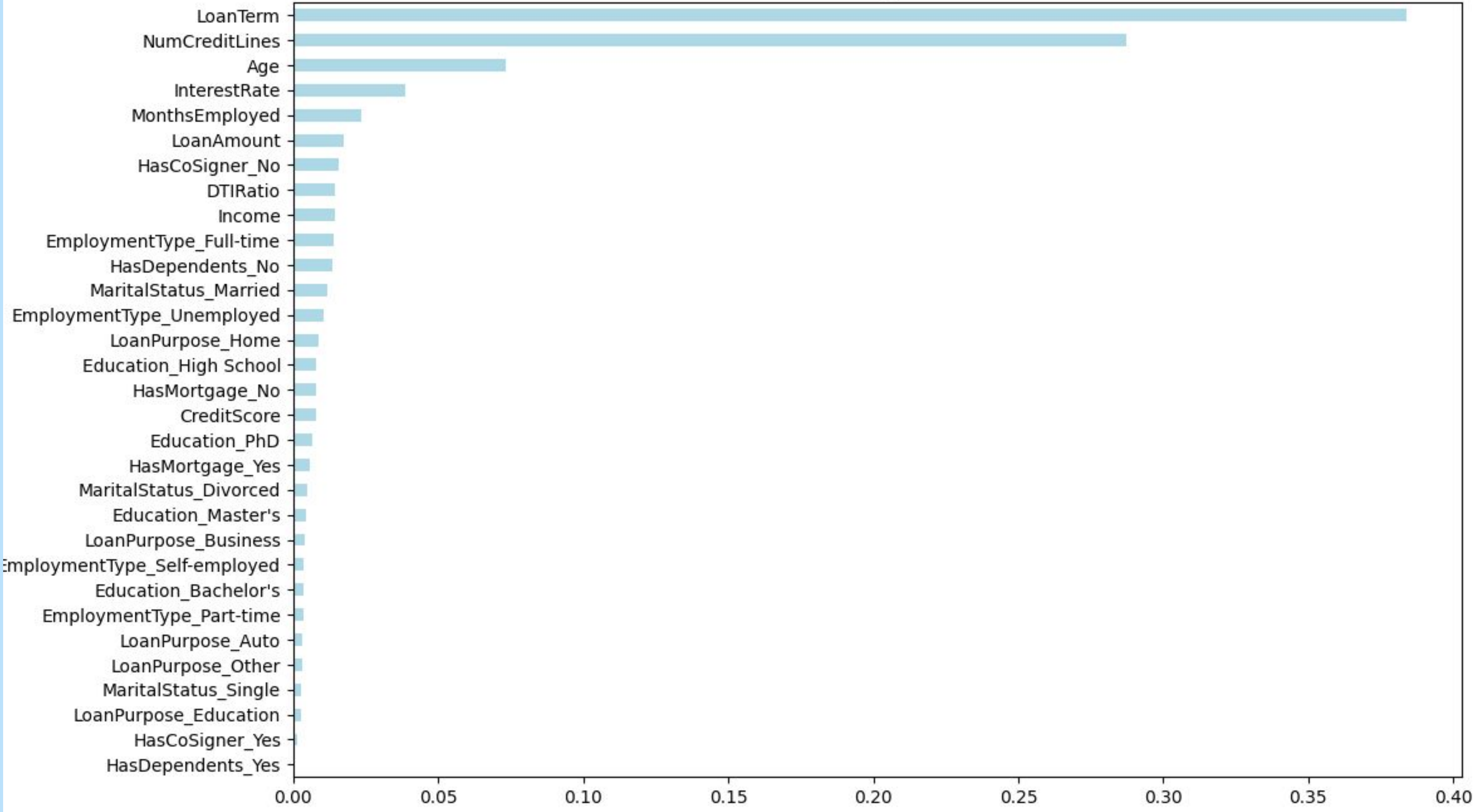
```
Accuracy: 0.9222844989920025

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.99      0.93     45127
           1       0.98      0.86      0.92     45151

    accuracy                           0.92     90278
   macro avg       0.93      0.92      0.92     90278
weighted avg       0.93      0.92      0.92     90278

Frequency of Predictions:
0    50871
1    39407
```



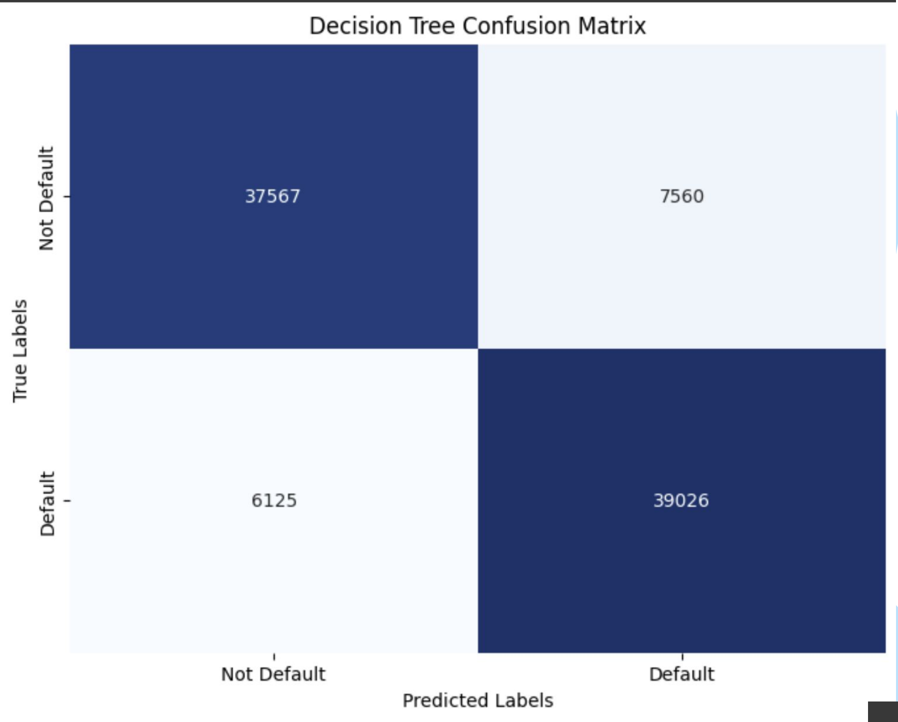


XGBoost

- Model Training
  - 100 estimators
  - Log loss evaluation metric
- Accuracy
  - 92% - high overall performance
- High Recall Ensure Reliable Identification



LoanTerm
NumCreditLines
Age
InterestRate
MonthsEmployed
LoanAmount
HasCoSigner_No
DTIRatio
Income
EmploymentType_Full-time
HasDependents_No
MaritalStatus_Married
EmploymentType_Unemployed
LoanPurpose_Home
Education_High School
HasMortgage_No
CreditScore
Education_PhD
HasMortgage_Yes
MaritalStatus_Divorced
Education_Master's
LoanPurpose_Business
mploymentType_Self-employed
Education_Bachelor's
EmploymentType_Part-time
LoanPurpose_Auto
LoanPurpose_Other
MaritalStatus_Single
LoanPurpose_Education
HasCoSigner_Yes
HasDependents_Yes
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40



Decision Tree

- Without SMOTE
  - ACC: 80%, F1: 55%
- With SMOTE
  - ACC: 85%, F1: 85%
- SMOTE helped to correct the imbalance of Defaulting payments

Classification Report:

	precision	recall	f1-score	support
0	0.86	0.83	0.85	45127
1	0.84	0.86	0.85	45151
accuracy			0.85	90278
macro avg	0.85	0.85	0.85	90278
weighted avg	0.85	0.85	0.85	90278

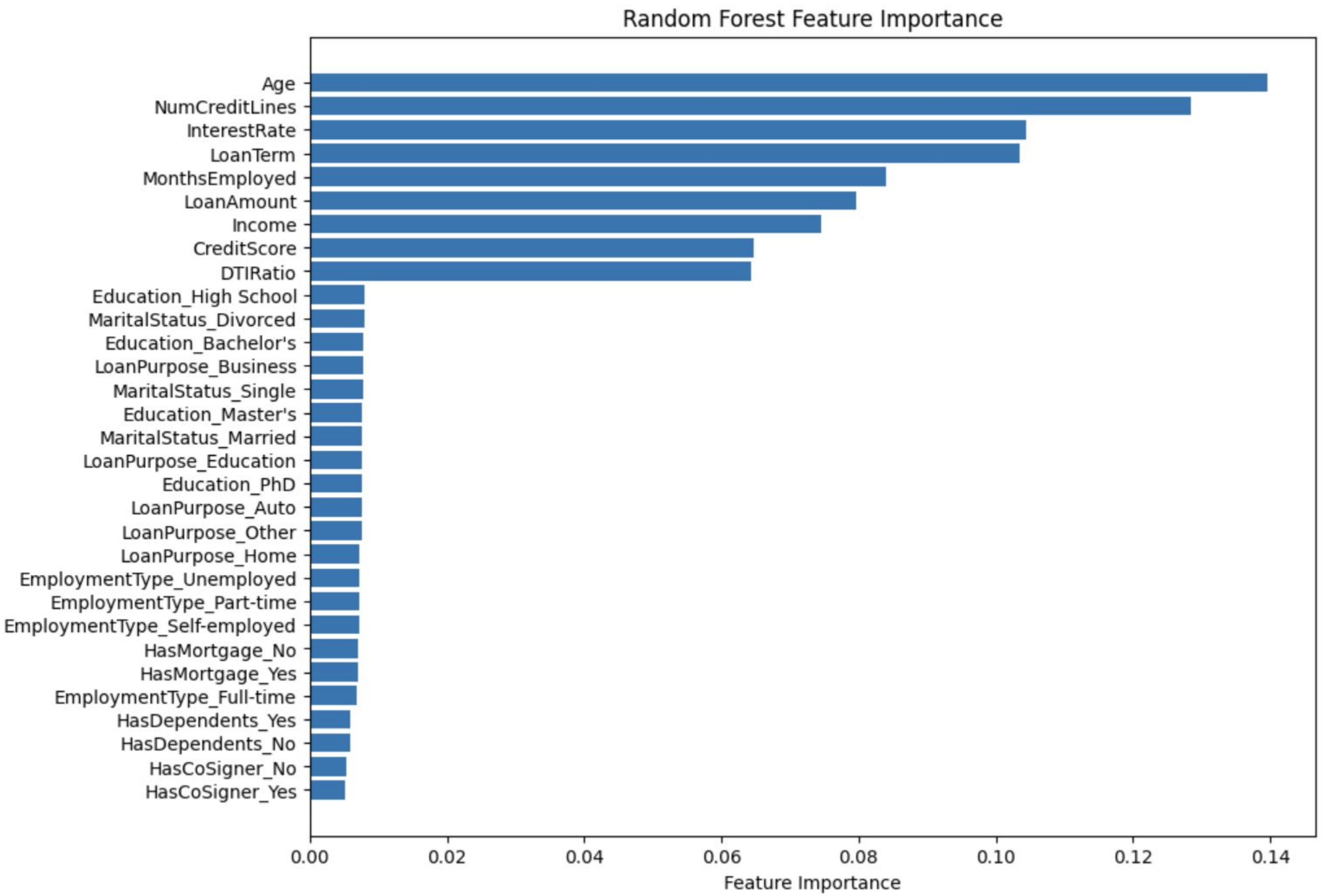
Accuracy: 0.918219278229469

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.93	0.92	45127
1	0.93	0.90	0.92	45151
accuracy			0.92	90278
macro avg	0.92	0.92	0.92	90278
weighted avg	0.92	0.92	0.92	90278

Random Forest

- Data Preprocessing
  - SMOTE
  - One-hot encoding
- Parameters
  - 100 trees
- Model Performance
  - 91.8% accuracy



Random Forest Feature Importance
Age
NumCreditLines
InterestRate
LoanTerm
MonthsEmployed
LoanAmount
Income
CreditScore
DTIRatio
Education_High School
MaritalStatus_Divorced
Education_Bachelor's
LoanPurpose_Business
MaritalStatus_Single
Education_Master's
MaritalStatus_Married
LoanPurpose_Education
Education_PhD
LoanPurpose_Auto
LoanPurpose_Other
LoanPurpose_Home
EmploymentType_Unemployed
EmploymentType_Part-time
EmploymentType_Self-employed
HasMortgage_No
HasMortgage_Yes
EmploymentType_Full-time
HasDependents_Yes
HasDependents_No
HasCoSigner_No
HasCoSigner_Yes
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Feature Importance

Models Accuracy

With SMOTE

- Support Vector Machines - 75%
- Logistic Regression - 70%
- Multinomial Naive Bayes - 57%
- Gaussian Naive Bayes - 80%
- ANN Deep Learning - 74%
- Gradient Boosting Machine - 90%
- **XGBoost - 92%**
- Decision Tree - 85%
- **Random Forest - 92%**

Final Business Insight

- What makes a person default on a loan?
  - Loan Terms
  - Credit Lines
  - Age
  - Employment
    - Type
    - Length of

Thank You!

Any Questions?