Project 4

Group 3: Jorge Chavez, Alejandra Gomez, Brittany Wright

Project Proposal

The aim of our project is to uncover patterns between loan information and loan applicant information.

Our Data

- We are using a bank loan status dataset
  - 34,469 records
- Source: https://www.kaggle.com
- Limitations
  - Income & Loan quantities are in the millions
  - Most credit scores are high; lowest numbers are average scores
  - Lack of Data

Cleaning the Data

- Merged the training and testing datasets from Kaggle
- Dropped all null values
- Dropped outlier credit scores
- Recoded the “Loan Status” variable

```python
#import data to clean and check null values for train data
credit_train = pd.read_csv("credit_train.csv")
credit_train.isnull().sum()
```

[3] ✓ 0.2s

```
Loan ID                          514
Customer ID                      514
Loan Status                      514
Current Loan Amount              514
Term                             514
Credit Score                   19668
Annual Income                  19668
Years in current job            4736
Home Ownership                   514
Purpose                          514
Monthly Debt                     514
Years of Credit History          514
Months since last delinquent   53655
Number of Open Accounts          514
Number of Credit Problems        514
Current Credit Balance           514
Maximum Open Credit              516
Bankruptcies                     718
Tax Liens                        524
dtype: int64
```

```python
#check null values
credit_train.isnull().sum()
```

[6] ✓ 0.0s

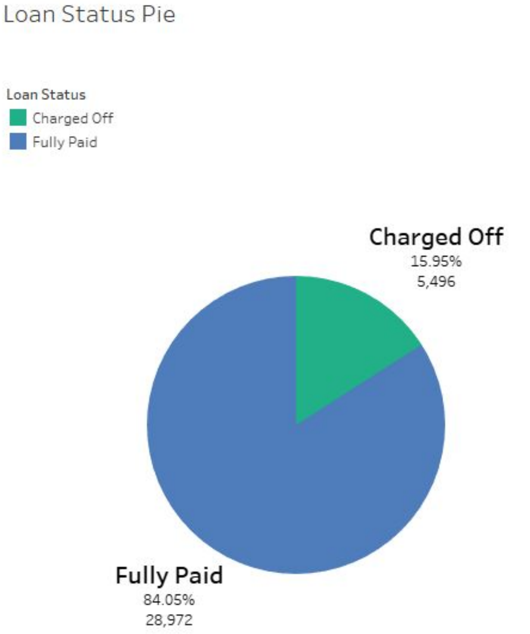
```
Loan ID                         0
Customer ID                     0
Loan Status                     0
Current Loan Amount             0
Term                            0
Credit Score                    0
Annual Income                   0
Years in current job            0
Home Ownership                  0
Purpose                         0
Monthly Debt                    0
Years of Credit History         0
Months since last delinquent    0
Number of Open Accounts         0
Number of Credit Problems       0
Current Credit Balance          0
Maximum Open Credit             0
Bankruptcies                    0
Tax Liens                       0
dtype: int64
```

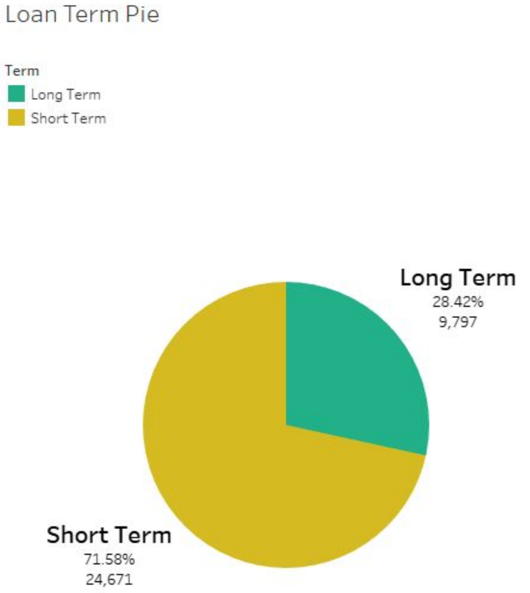
Questions

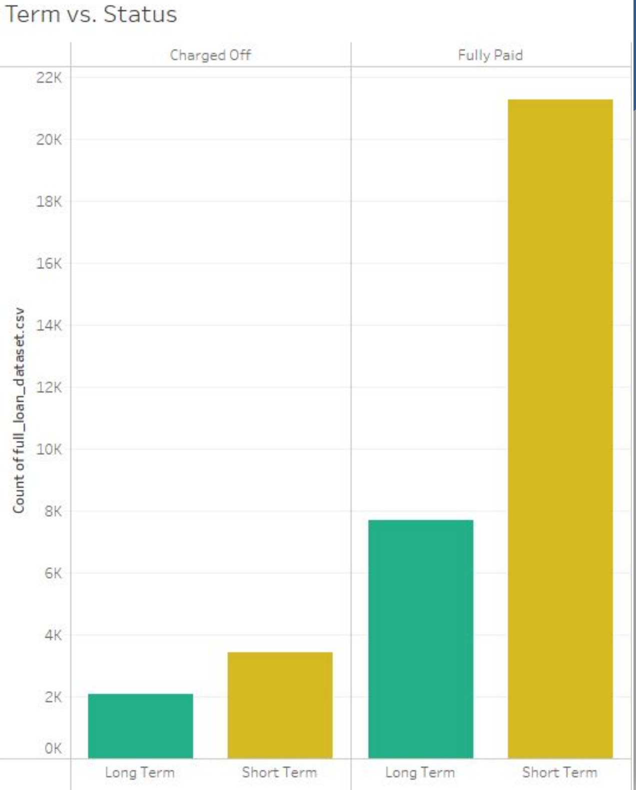
- What are the relationships between the loan status & credit scores to the other variables?
- Does the relationship between income and credit score have sufficient strength to generate an accurate credit score?
- Can we predict loan status based on current loan amount, monthly debt, maximum open credit, current credit balance, years of credit history, months since last delinquent, number of open accounts, credit score, and annual income?

Question 1

Relationship between Loan Status and Loan Term





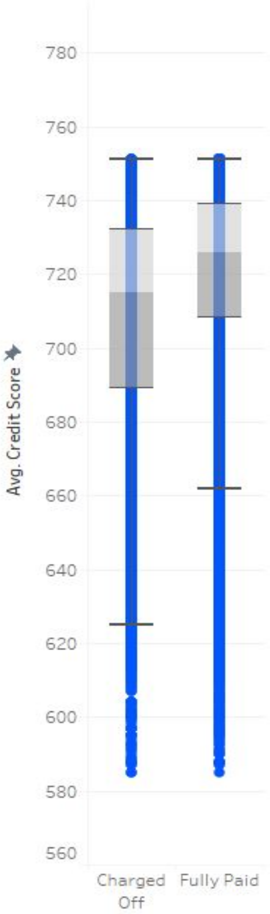


Relationship between Loan Status, Credit Score and Annual Income

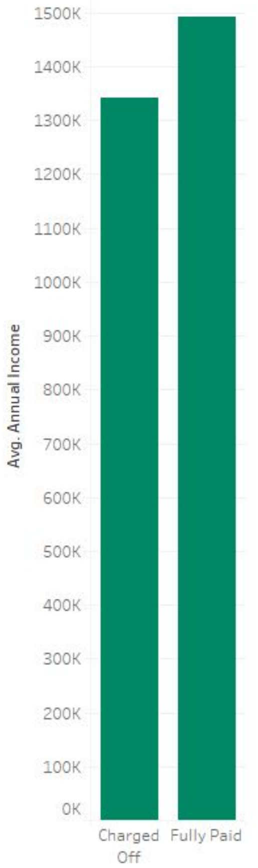
Credit Scores Categories

Credit Score Category	Charged Off % of Total Count of full_loan_dataset.c..	Charged Off Count of full_loan_dataset.csv	Fully Paid % of Total Count of full_loan_dataset.c..	Fully Paid Count of full_loan_dataset.csv
Average	13.57%	746	6.92%	2,004
Good	19.29%	1,060	15.40%	4,463
Very Good	55.22%	3,035	58.80%	17,036
Exceptional	11.92%	655	18.88%	5,469

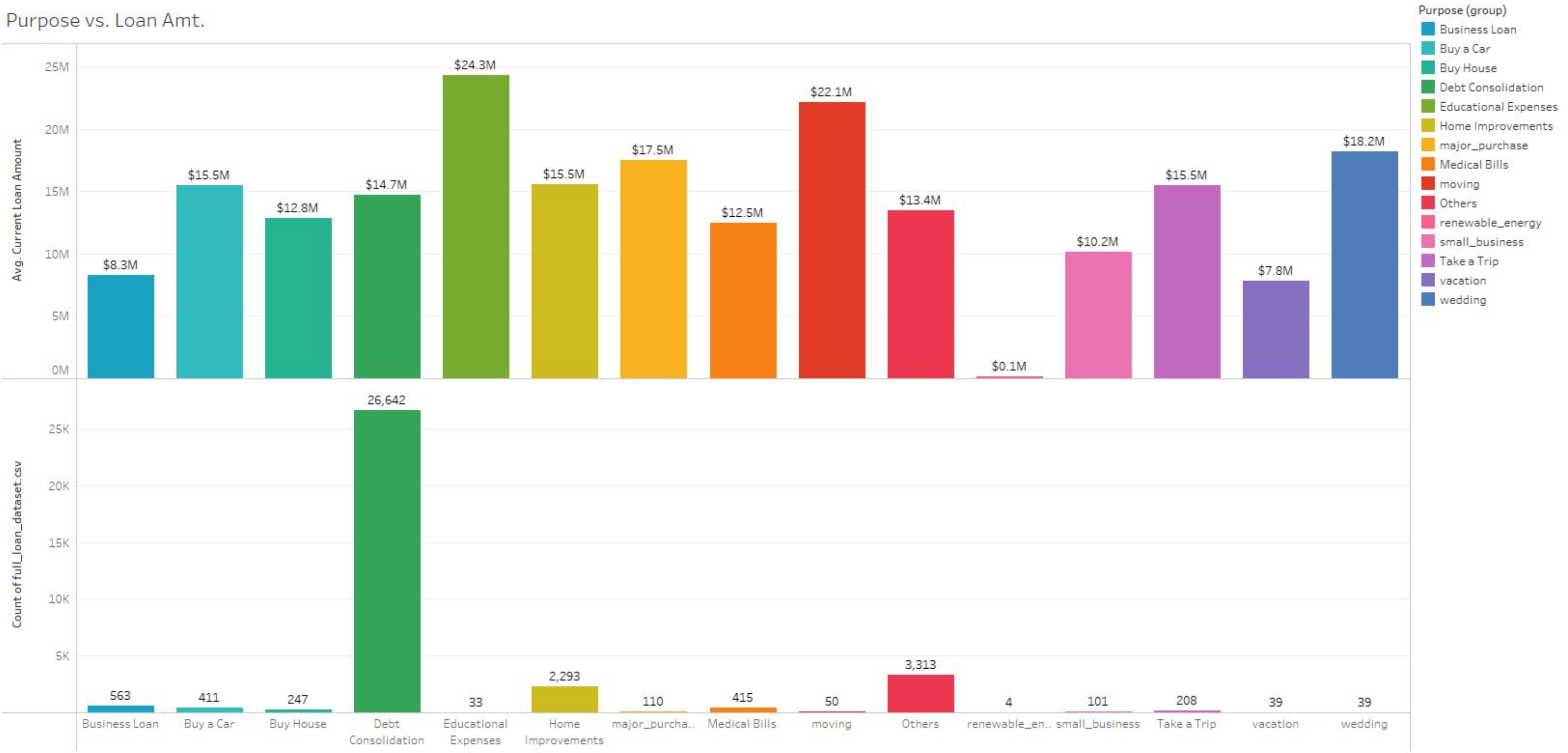
Status vs. Credit Score



Status vs. Annual Inc.



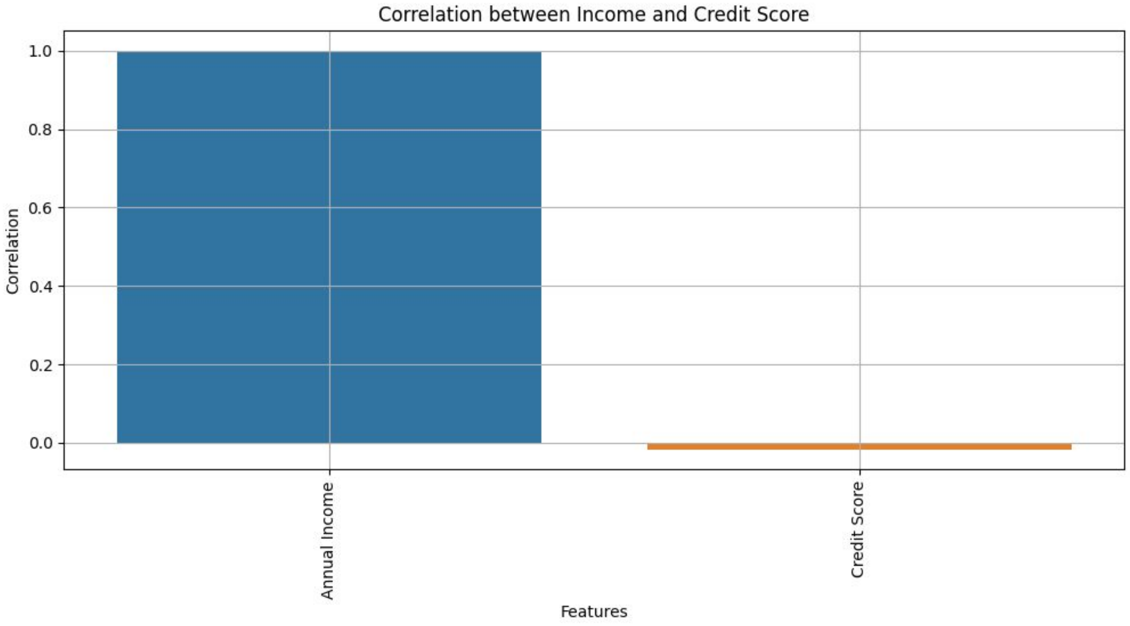
Relationship between Loan Purpose and Loan Amount



Question 2

Random Forest Regression:

- Used a random forest regressor on just Income/ Credit Score.(R2 score came to be 5%)
- Meaning not enough data/features were used to enable predicting credit score based off income alone accurately



New Regression Tree w All Features

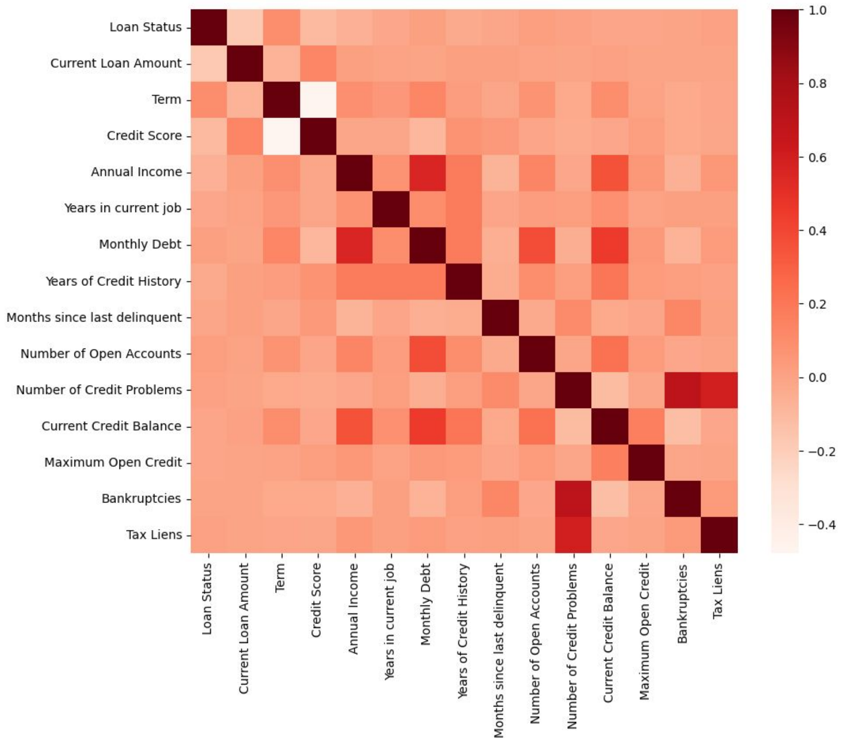
- Decided to use all features to allow model to ingest as much data to see if this will improve score
- Heat map details relationship between features

Feature Importance

```
# Random Forests in sklearn will automatically calculate
importances = rf_model.feature_importances_
# We can sort the features by their importance
sorted(zip(rf_model.feature_importances_, X.columns), re
```

203] ✓ 0.0s

```
[(0.6976201713759477, 'Term'),
 (0.11332549917673167, 'Current Loan Amount'),
 (0.05701004720834927, 'Maximum Open Credit'),
 (0.027657907883268593, 'Monthly Debt'),
 (0.024126555912081963, 'Loan Status'),
 (0.018753434201665746, 'Years of Credit History'),
 (0.018391064430298976, 'Current Credit Balance'),
 (0.013518817652461212, 'Annual Income'),
 (0.010910370254427223, 'Number of Open Accounts'),
 (0.010151636369485772, 'Months since last delinquent'),
 (0.003978408988601662, 'Number of Credit Problems'),
 (0.002593647348389447, 'Bankruptcies'),
 (0.0014250689367483354, 'Years in current job'),
 (0.0005373702615424883, 'Tax Liens')]
```



Final Score for Model

- Insufficient Features
- Weak Feature Importance
- Insufficient Data
- Data Imbalance
- Other Factors: Credit score prediction is a complex task influenced by various factors beyond income alone. Consider incorporating additional features or exploring alternative machine learning algorithms that may better capture the underlying patterns and relationships in the data.

random forest regressor

```
#create regressor model and train
rf = RandomForestRegressor(n_estimators = 300, max_features
[207] ✓ 3.9s

#make predicts on test data
prediction = rf.predict(X_test_scaled)

#R2 score
score5 = rf.score(X_test_scaled, y_test)
percentage = score5 * 100
print('R2: %.2f%%' % percentage)
[208] ✓ 0.1s
... R2: 26.58%
```

Question 3

Random Forest

- Top 10 most important features
- Accuracy Score: 0.84
- Classifying Fully Paid
  - Precision: 0.84
  - Recall: 1.00
- Classifying Charged Off
  - Precision: 0.78
  - Recall: 0.01

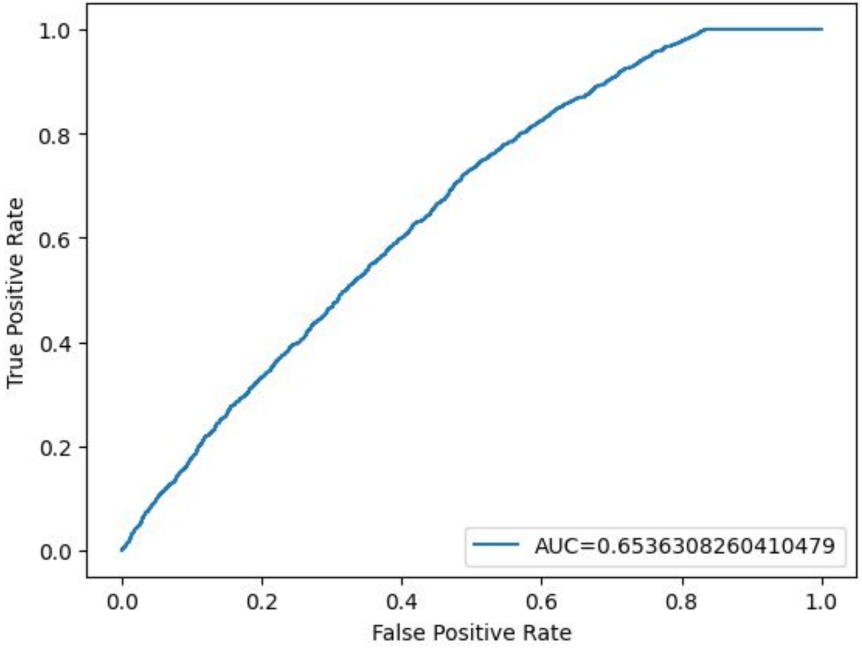
```
# Get the feature importance array
importances = rf_model.feature_importances_
# List the top 10 most important features
importances_sorted = sorted(zip(rf_model.feature_importances_, X.columns), reverse=True)
importances_sorted[:10]
```

[82] ✓ 0.0s

```
[(0.12002576564297994, 'Current Loan Amount'),
 (0.10083233921719331, 'Annual Income'),
 (0.09533412077978105, 'Monthly Debt'),
 (0.09248753732657634, 'Maximum Open Credit'),
 (0.09208273811707002, 'Current Credit Balance'),
 (0.09092201495901993, 'Credit Score'),
 (0.09060063448872557, 'Years of Credit History'),
 (0.08315615417435834, 'Months since last delinquent'),
 (0.06555833938865463, 'Number of Open Accounts'),
 (0.011934134046475923, 'Number of Credit Problems')]
```

Logistic Regression

- Decided to use top 9 from random forest originally since the last feature only explained roughly 1% of the variance.
- Accuracy: 0.81
- Classifying Fully Paid
  - Precision: 0.82
  - Recall: 1.00
- Classifying Charged Off
  - Precision: 0.00
  - Recall: 0.00



Over/UnderSampling

Oversampling

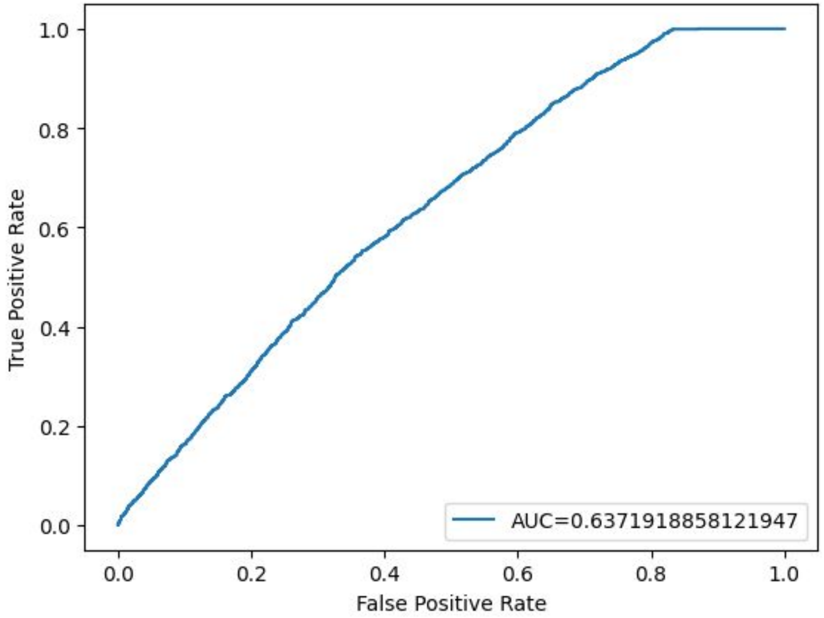
- Accuracy: 0.55
- Classifying Fully Paid
  - Precision: 0.86
  - Recall: 0.54
- Classifying Charged Off
  - Precision: 0.23
  - Recall: 0.62

Undersampling

- Accuracy: 0.55
- Classifying Fully Paid
  - Precision: 0.86
  - Recall: 0.54
- Classifying Charged Off
  - Precision: 0.23
  - Recall: 0.62

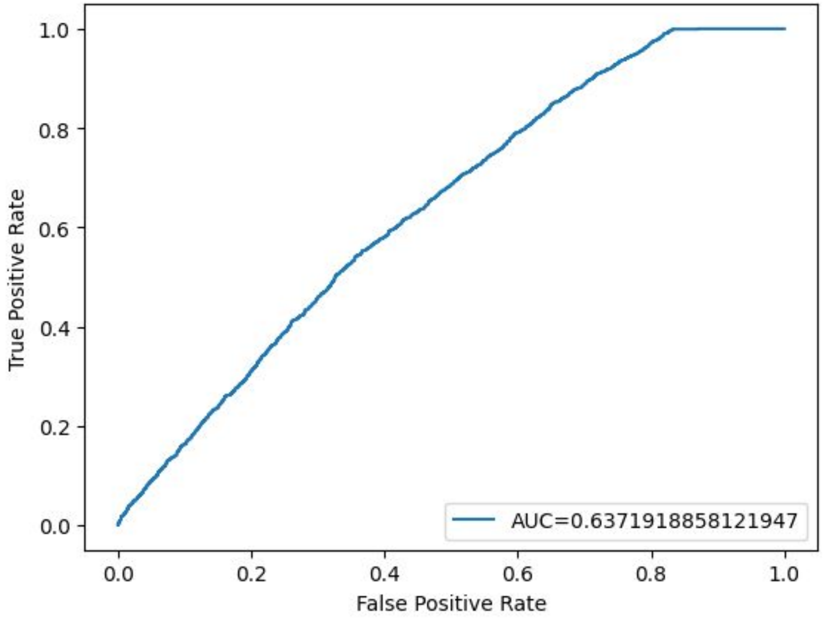
Changing the Number of Predictors to 5

- Changed the number of predictors to the top 5 as each explained roughly 10% of the variance.
- Predictors included current loan amount, monthly debt, maximum open credit, current credit balance, and annual income.



5 Predictors

- Accuracy: 0.83
- Classifying Fully Paid
  - Precision: 0.84
  - Recall: 1.00
- Classifying Charged Off
  - Precision: 0.00
  - Recall: 0.00



Future Considerations

- Obtaining more data
- Getting more balanced data
- Getting data sources with more information
  - Information about income
  - Coding of missing values
  - Accurate credit score data

The End

Questions?