Identifying At-Risk Customers

Before Approving Credit Applications.

Milestone 5

Michelle Helfman

Bellevue University

DSC630-T302 Predictive Analytics (2241-1)

Professor Andrew Hua

October 28, 2023

Term Project Summary

Introduction:

As the US economy is emerging from the COVID-19 pandemic, more people are falling behind on their credit card and car loan payments. Rents have increased dramatically to offset the loss of rental income during the pandemic. Student loan payments resumed in October. Overpriced homes purchased during the bidding wars of 2021 and 2022 have lost considerable value, and with home prices dropping and interest rates at record-high levels, selling these homes will be challenging. There are no lower interest rates for refinancing, so mortgages will begin to default.

The Federal Reserve (Fed) continues to raise interest rates to control inflation, making purchasing more challenging. Also, banks and other lenders had already begun tightening credit for months, which went into overdrive after the bank failures in the spring of 2023. These factors place additional importance on identifying customers at risk of defaulting before approving credit applications.

Data Selection:

A credit risk dataset from Kaggle was used to simulate the credit bureau information used on loan applications and other credit approval forms. The dataset consists of 32581 rows and 12 columns of data to analyze customer credit risk using current loan information to train our model. The information is a combination of 4 categorical (Home Ownership, Intent, Loan Grade, and Prior Defaults) and 8 numerical (Age, Income, Employment Length, Loan Amount, Interest Rate, Current Loan Status, Loan Percent Income, and Credit History Length) columns.

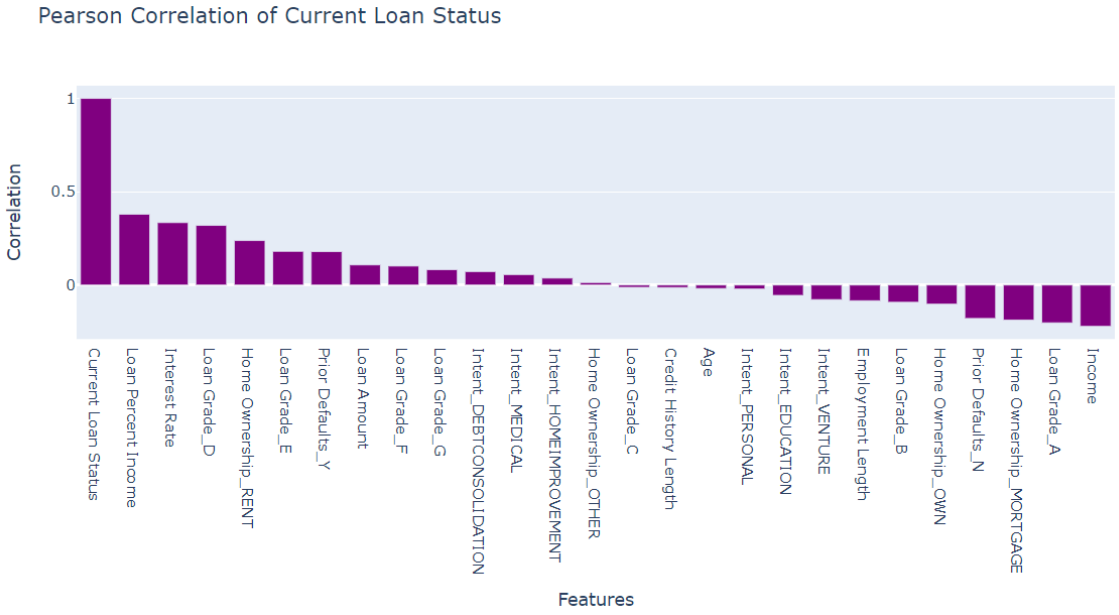
The categorical columns will be divided into separate columns based on the unique information and then converted to numerical values.

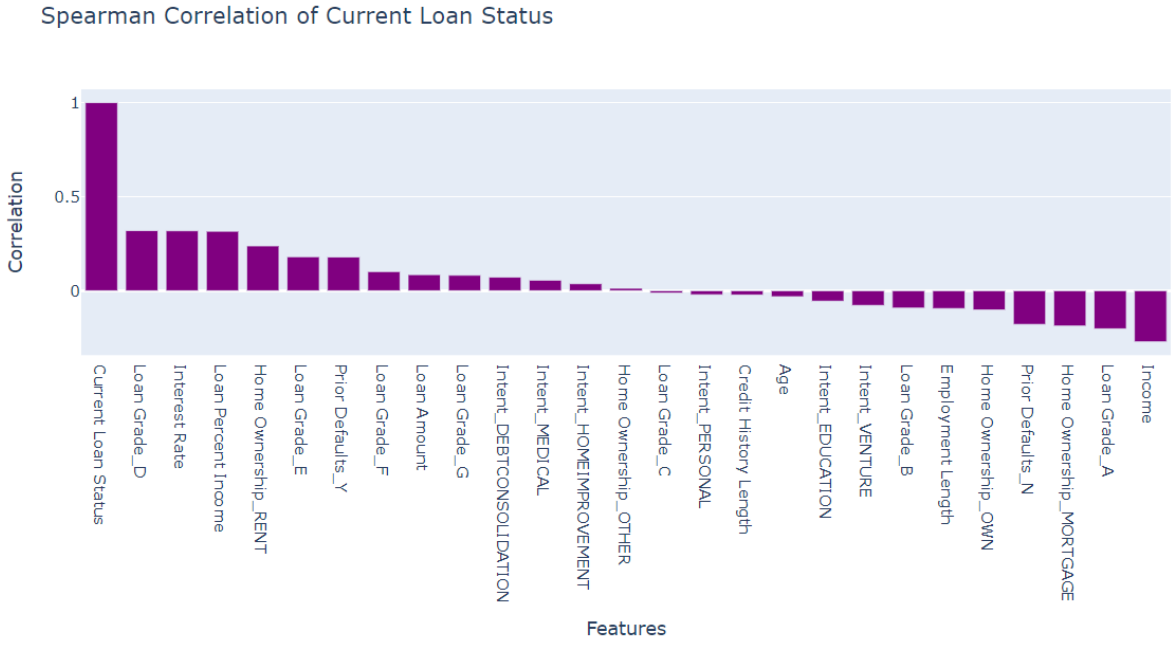
- Age – The age of the customer
- Income – Annual income in USD
- Home Ownership – Rent, Mortgage, Own, or Other
- Employment Length – Length of current employment in years
- Intent – Debt Consolidation, Education, Home Improvement, Medical, Personal, or Venture
- Loan Amount – Current loan amount
- Loan Grade – A through G, A is excellent, G is considered very poor
- Interest Rate – Current interest rate
- Current Loan Status – Good standing or default
- Loan Percent Income - Percent of income of the current loan
- Prior Default – Prior loan that is in default
- Credit History Length – Years of reporting credit

Methodology:

Exploratory Data Analysis (EDA) was performed to better understand the Credit Risk dataset. For each numeric column, the minimum, maximum, mean, and percentile statistics were given. Details of the categorical, non-numeric columns, consisting of the various elements, were described. Columns with outliers and missing information were identified and corrected by filling in the missing information or deleting the entire row. The categorical information is separated by unique values into individual columns and converted to numeric dummy variables.

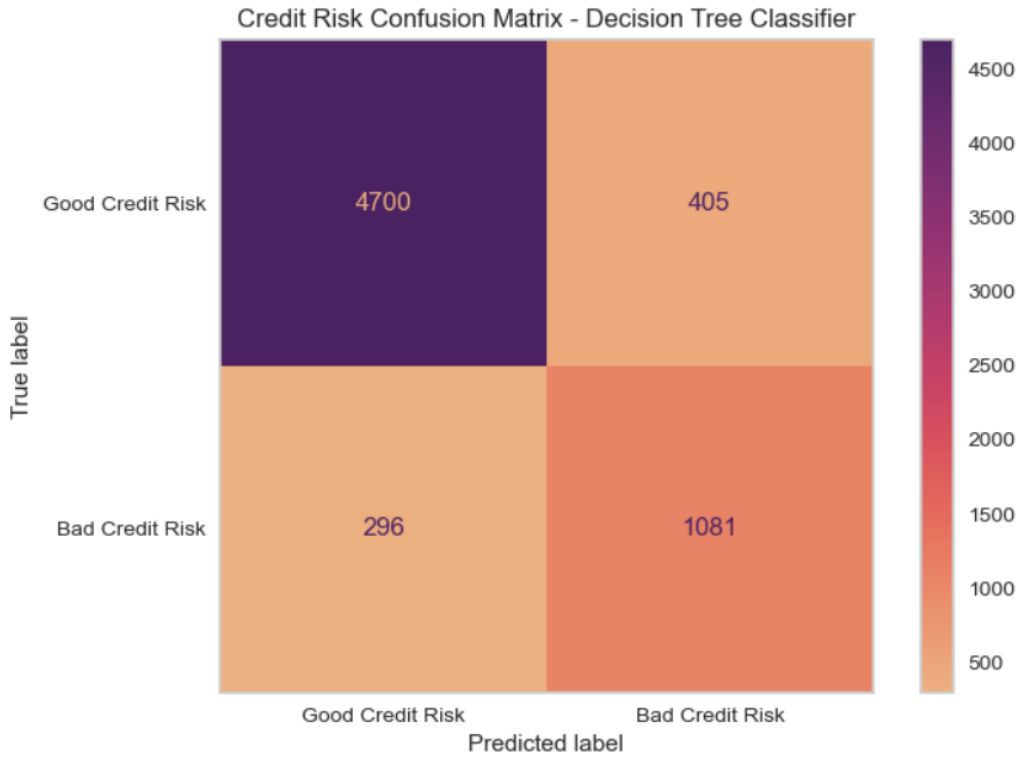
The Decision Tree Classifier and Logistic Regression models were used to create a binary outcome: Good Credit Risk or Bad Credit Risk. The Pearson and Spearman Correlation methods compare all the columns, including the dummy variables, to the Current Loan Status (Target Variable) and evaluate the level of correlation.





After running both models none of the columns were considered highly correlated to the Current Loan Status.

The Decision Tree Classifier and Logistic Regression models with 80% train and 20% test sets were used to determine the models' results and accuracy.



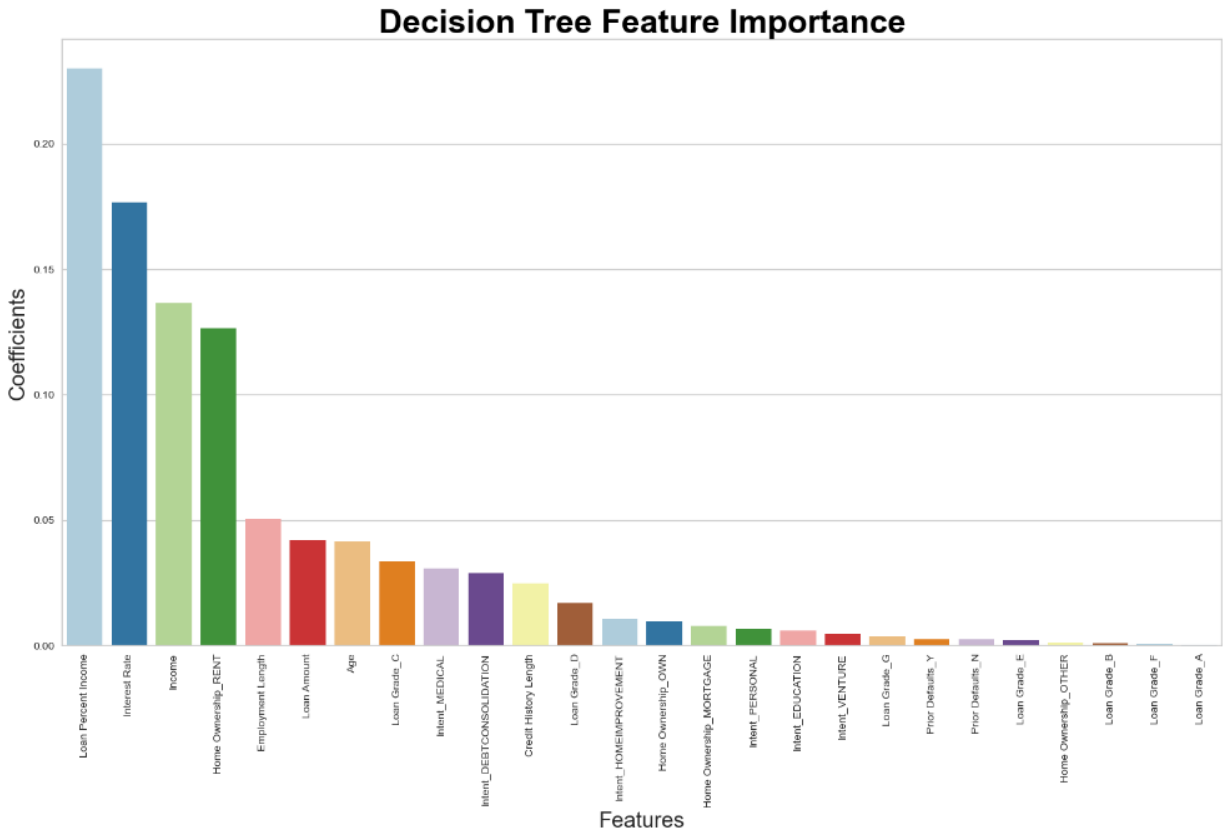
Classification Report For Decision Tree Classifier

	precision	recall	f1-score	support
Good Credit Risk	0.94	0.92	0.93	5105
Bad Credit Risk	0.73	0.79	0.76	1377
accuracy			0.89	6482
macro avg	0.84	0.85	0.84	6482
weighted avg	0.90	0.89	0.89	6482

The Decision Tree Classifier predicted Good Credit Risk with 94% accuracy and Bad Credit Risk at 73% with an overall accuracy score of 89%. The Recall scores of all the records with

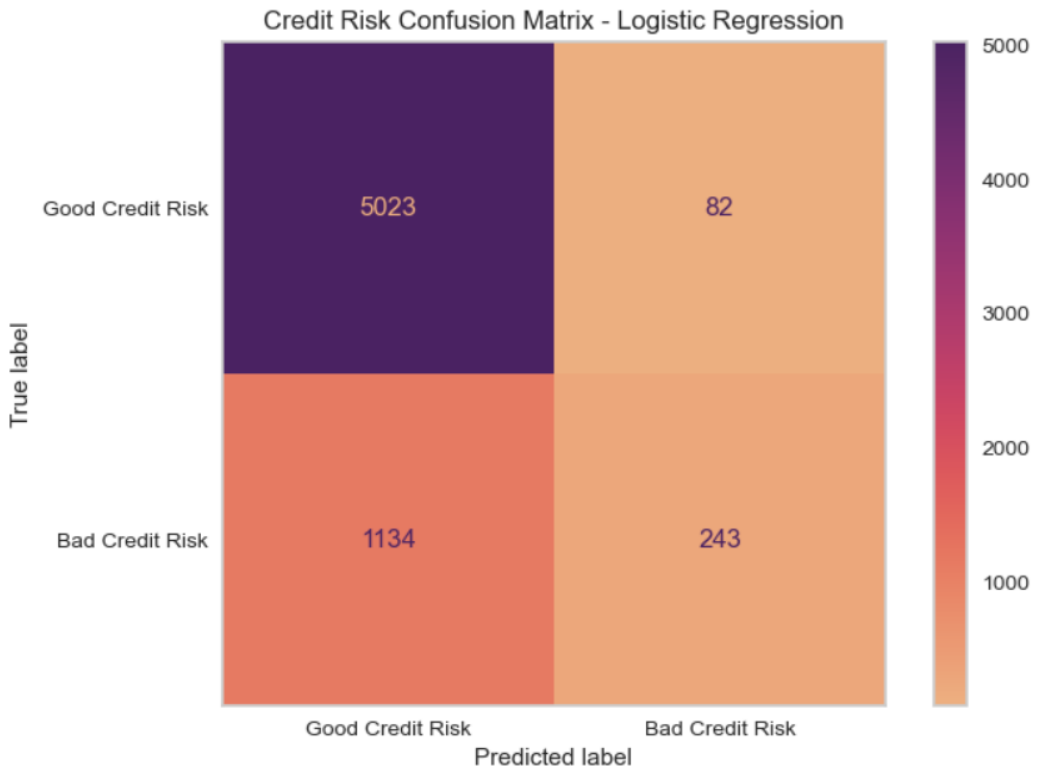
Good and Bad Credit Risk predicted correctly are 92% and 79%, and the combination of both Recall and Precision scores yielded scores of 93% and 76%, respectively.

The Decision Tree Classifier also ranked how the features contributed to the accuracy of this model.



The top 5 features that contribute the most to predicting the accuracy of the Decision Tree Classification model are Loan Percent Income, Interest Rate, Income, HomeOwnership_Rent, and Employment Length. The five features that contribute the least are Lone Grade_E, Home Ownership_Other, Lone Grade_B, Lone Grade_F, and Lone Grade_A.

The Logistic Regression Model was run and produced similar results.

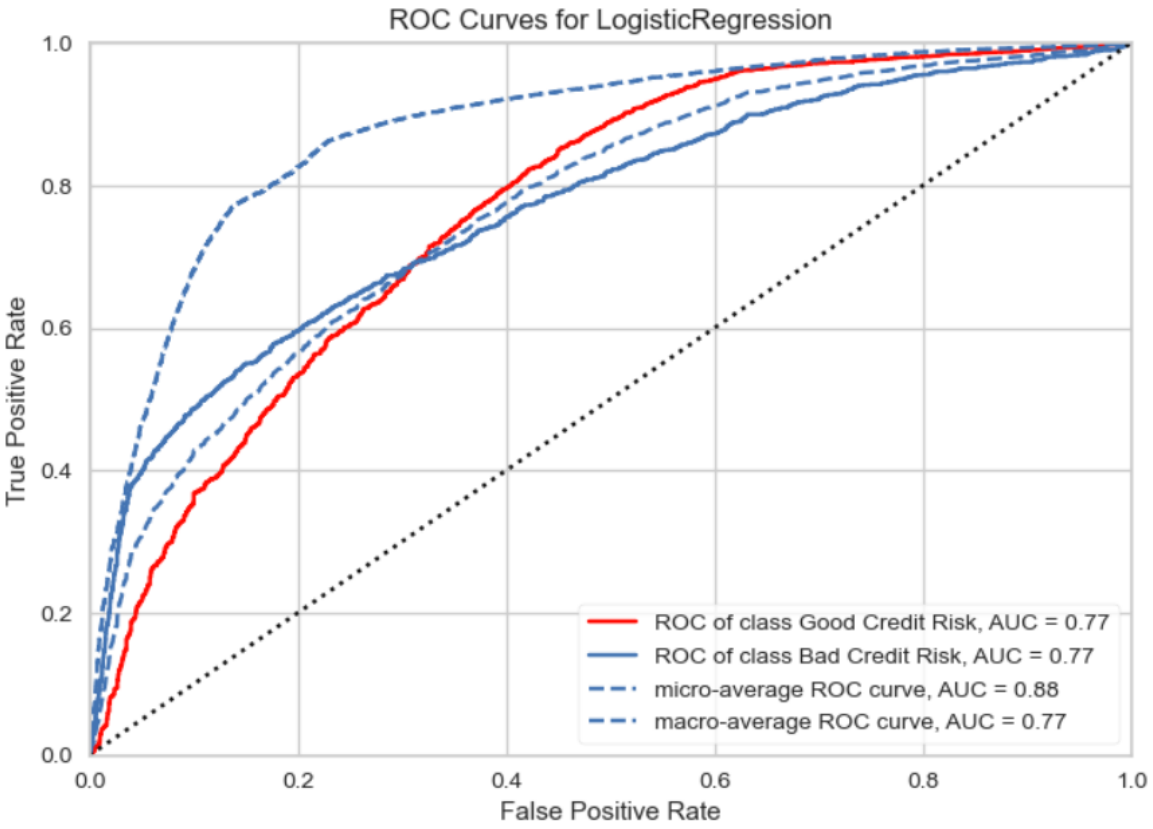


Classification Report For Logistic Regression

	precision	recall	f1-score	support
Good Credit Risk	0.82	0.98	0.89	5105
Bad Credit Risk	0.75	0.18	0.29	1377
accuracy			0.81	6482
macro avg	0.78	0.58	0.59	6482
weighted avg	0.80	0.81	0.76	6482

The Logistic Regression Model predicted Good Credit Risk with 82% accuracy and Bad Credit Risk at 75% with an overall accuracy score of 81%. The Recall scores of all the records with Good and Bad Credit Risk predicted correctly are 98% and 18%, and the combination of both Recall and Precision scores yielded scores of 89% and 29%, respectively.

A second accuracy test was performed on this model using the ROC (Receiver Operating Characteristics) Curve.



The ROC Curve shows that the Logistic Regression Model is a good predictor of Credit Risk, with an overall accuracy score of 77%.

Results:

Both the Decision Tree Classifier and Logistic Regression models give results on whether an applicant is a good or bad credit risk based on the Current Loan Status. While both models return binary results, the accuracy of the models is very different.

- The Accuracy Score for the Decision Tree Classifier was 89%, while the accuracy of the Logistics Regression model was 81%. The results are the same as part of the Classification Report or when calculated separately.
- Both models have false Good and Bad Credit Risk rates above 10%. The Decision Tree Classifier was 11%, and the Logistics Regression model was 19%.

- The Precision score on the Classification Report reflects a correct percentage of Good Credit Risk predictions. The Decision Tree Classifier performed better than the Logistic Regression model at 94% to 82%, respectively.
- The Classification Report also gives a Recall score on the proportion of actual Good Credit Risk records that were identified correctly. The Logistics Regression model had a better Recall score of 98% over the Decision Tree Classifier, with 92%

Conclusion:

Credit, either by loan or credit card, is essential in everyday life and for the future. Credit is needed for mortgages, transportation, education, business starting, and health and well-being. Banks and lenders take on risk when extending credit, and it is essential to minimize that risk.

I recommend using the Decision Tree Classifier model, which can become an even better model with some additional credit-related information: credit bureau score, rent or mortgage payment, and other information from a standard credit application. This model is already very good at 89% accuracy, but with credit being so important, an accuracy rate of over 93% (outstanding) would make one more comfortable with the results.

Because having credit is so important, relying solely on the outcome of a statistical model would create a significant risk of a wrong decision. An actual person should also review the individual parts of the credit profile before making a final determination.

References

Tory Newmyer, Aaron Gregg, and Jaclyn Peiser (August 30, 2023)

Delinquencies rise for credit cards and auto loans, and it could get worse

https://www.washingtonpost.com/business/2023/08/30/delinquencies-credit-auto-loans/

Diana Olick (November 7, 2022)

Here's how much equity U.S. homeowners have lost since May

https://www.cnbc.com/2022/11/07/homeowners-lost-1point5-trillion-in-equity-since-may-as-home-prices-drop.html

Lao Tse (2020). Credit Risk dataset. (Retrieved, September 4, 2023)

https://www.kaggle.com/datasets/laotse/credit-risk-dataset