Caitriona O’Donohoe

Loan Approval Prediction System

Problem Statement

About 51% of Americans have taken out a personal loan in their lifetime. Processing these loan applications drive up labor costs for banks, whereas more technologically advanced banks can take advantage of automated processes, decreasing their operational costs. Using a well designed algorithm to predict loan approval can financially benefit the banks processing them, prevent red-lining, and protect banks from making *bad* loans, which have historically damaged the US economy.

Using loan approval data, we determine how can we predict the loan approval status for applicants using key applicant features?

Data Wrangling

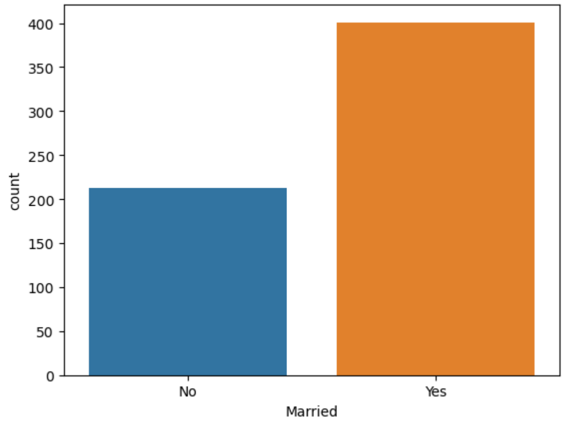
This Kaggle data set includes key features reviewed when one applies for a loan, including marital status, credit history, number of dependents, income, and more. Originally, the data was made up of 614 rows, including 12 features. Two of these features, applicant income and co-applicant income were consolidated into a new column: total income.

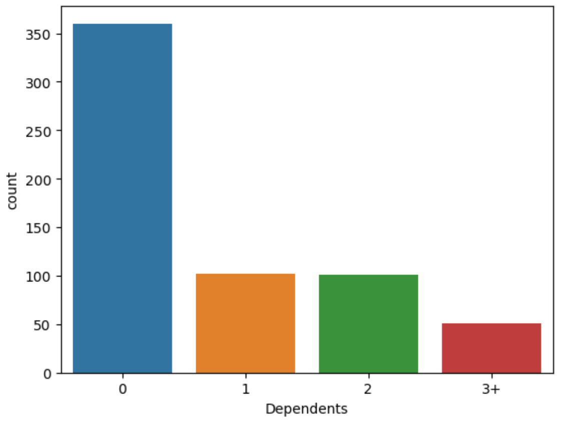
Several of our columns were missing a relatively small number of values, including gender, marital status, self-employment, number of dependents, term, and credit history. We rectified this by replacing these values using mode.

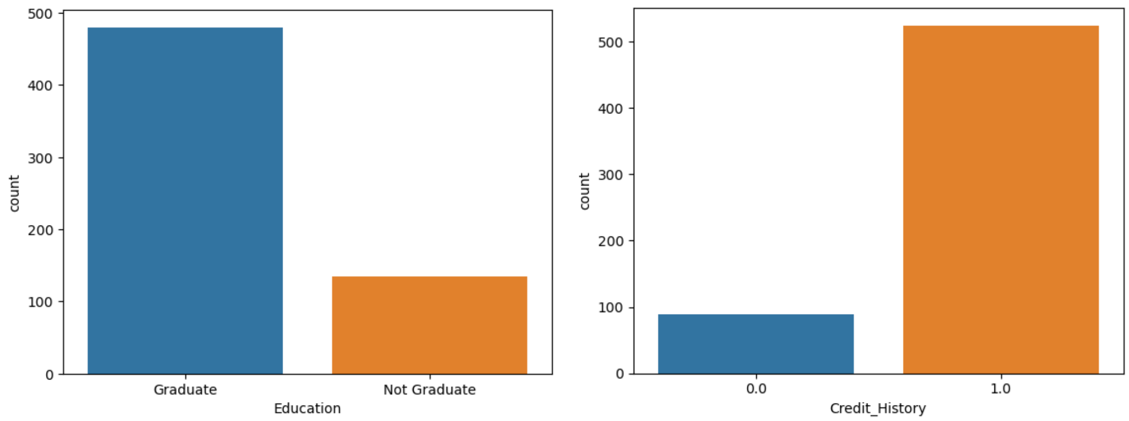
Additionally, all categorical values were converted to numerical values using Label Encoder.

Exploratory Data Analysis/Training & Pre-Processing

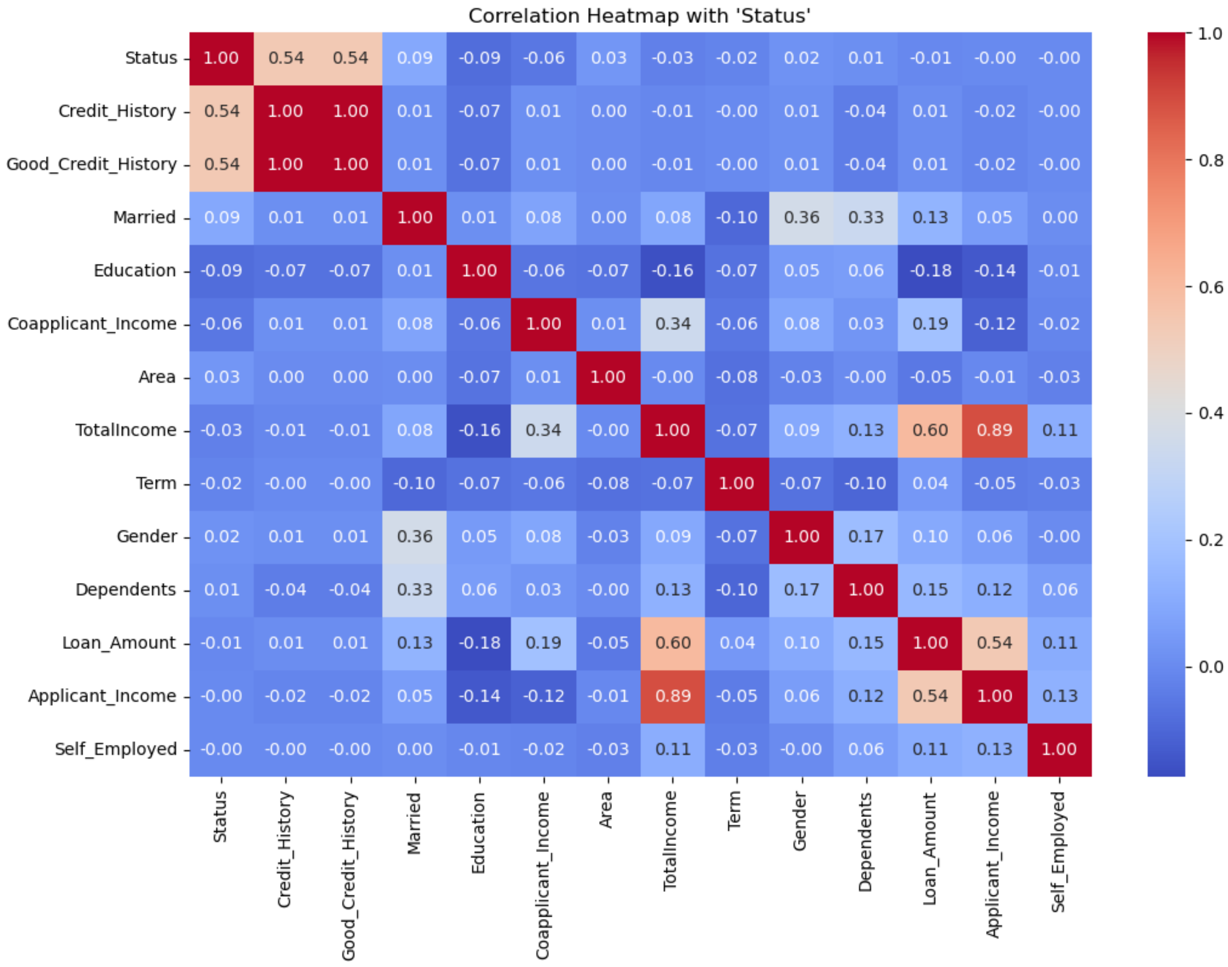
Using the correlation function in Python, we were able to determine how influential each of our features are relating to our target feature: Status. The most influential feature by far was credit history, with a correlation value of 0.540556. Other notable features include education, marital status, number of dependants, and income.







Using what we learned about credit history, we used manual feature engineering to create an additional column called Good Credit History; a binary column depicting whether or not an applicant has a good credit history.

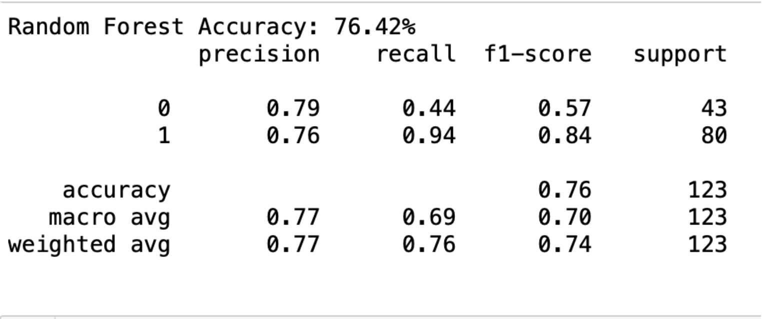


	Status	Credit_History	Good_Credit_History	Married	Education	Coapplicant_Income	Area	TotalIncome	Term	Gender	Dependents	Loan_Amount	Applicant_Income	Self_Employed
Status	1.00	0.54	0.54	0.09	-0.09	-0.06	0.03	-0.03	-0.02	0.02	0.01	-0.01	-0.00	-0.00
Credit_History	0.54	1.00	1.00	0.01	-0.07	0.01	0.00	-0.01	-0.00	0.01	-0.04	0.01	-0.02	-0.00
Good_Credit_History	0.54	1.00	1.00	0.01	-0.07	0.01	0.00	-0.01	-0.00	0.01	-0.04	0.01	-0.02	-0.00
Married	0.09	0.01	0.01	1.00	0.01	0.08	0.00	0.08	-0.10	0.36	0.33	0.13	0.05	0.00
Education	-0.09	-0.07	-0.07	0.01	1.00	-0.06	-0.07	-0.16	-0.07	0.05	0.06	-0.18	-0.14	-0.01
Coapplicant_Income	-0.06	0.01	0.01	0.08	-0.06	1.00	0.01	0.34	-0.06	0.08	0.03	0.19	-0.12	-0.02
Area	0.03	0.00	0.00	0.00	-0.07	0.01	1.00	-0.00	-0.08	-0.03	-0.00	-0.05	-0.01	-0.03
TotalIncome	-0.03	-0.01	-0.01	0.08	-0.16	0.34	-0.00	1.00	-0.07	0.09	0.13	0.60	0.89	0.11
Term	-0.02	-0.00	-0.00	-0.10	-0.07	-0.06	-0.08	-0.07	1.00	-0.07	-0.10	0.04	-0.05	-0.03
Gender	0.02	0.01	0.01	0.36	0.05	0.08	-0.03	0.09	-0.07	1.00	0.17	0.10	0.06	-0.00
Dependents	0.01	-0.04	-0.04	0.33	0.06	0.03	-0.00	0.13	-0.10	0.17	1.00	0.15	0.12	0.06
Loan_Amount	-0.01	0.01	0.01	0.13	-0.18	0.19	-0.05	0.60	0.04	0.10	0.15	1.00	0.54	0.11
Applicant_Income	-0.00	-0.02	-0.02	0.05	-0.14	-0.12	-0.01	0.89	-0.05	0.06	0.12	0.54	1.00	0.13
Self_Employed	-0.00	-0.00	-0.00	0.00	-0.01	-0.02	-0.03	0.11	-0.03	-0.00	0.06	0.11	0.13	1.00

Finally, we split our data using train/test split in sklearn, and normalized our numerical values using Min-Max Scalar, and the data was ready for modeling.

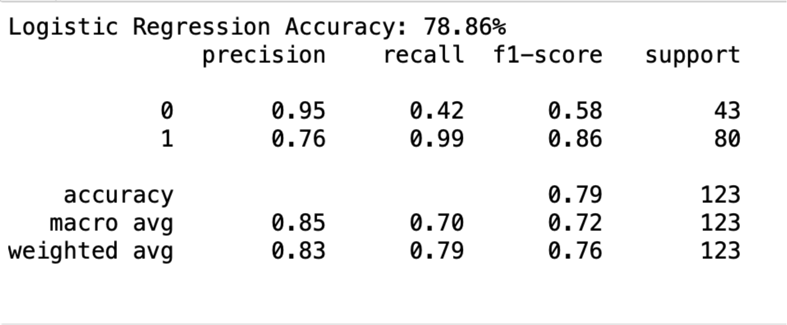
Modeling

I tested four different models for my project. The metric I used to evaluate each model was their accuracy score. Since this project is solving a classification problem, my first instinct was to use a random forest model, which returned an accuracy score of 76.42%.

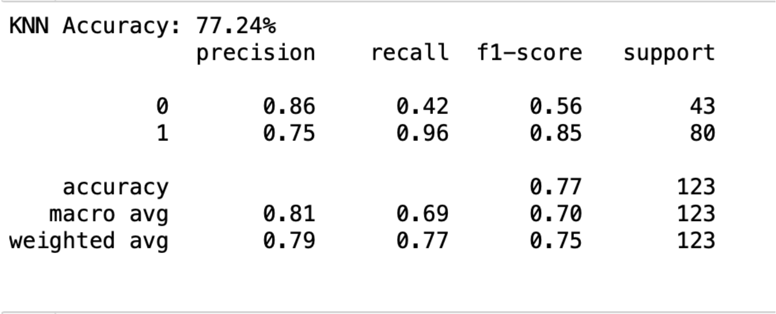


```
Random Forest Accuracy: 76.42%
              precision    recall  f1-score   support

           0       0.79      0.44      0.57        43
           1       0.76      0.94      0.84        80

    accuracy                           0.76       123
   macro avg       0.77      0.69      0.70       123
weighted avg       0.77      0.76      0.74       123
```

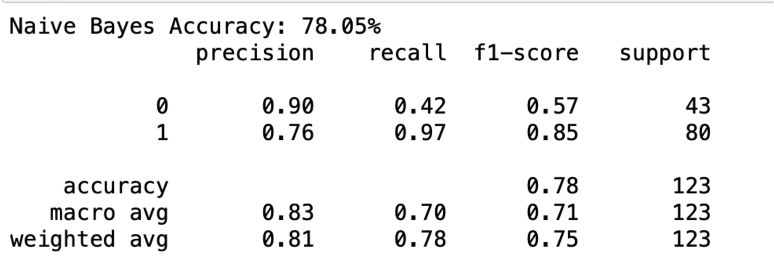
Not bad, but I was hopeful this could be improved. From there I explored other models; Naive Bayes, K-Nearest Neighbors, and Logistic Regression. While each performed better than the Random Forest model, the best was Logistic Regression, with an accuracy score of 78.86%.



```
Logistic Regression Accuracy: 78.86%
              precision    recall  f1-score   support

           0       0.95      0.42      0.58        43
           1       0.76      0.99      0.86        80

    accuracy                           0.79       123
   macro avg       0.85      0.70      0.72       123
weighted avg       0.83      0.79      0.76       123
```

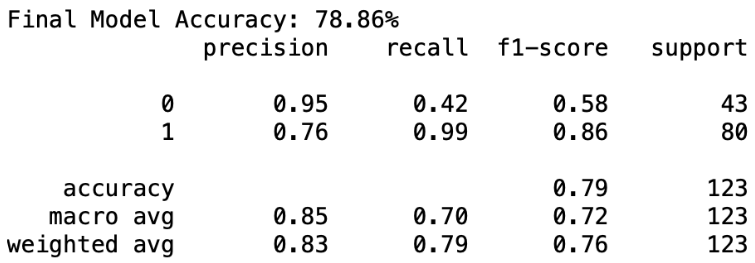


```
KNN Accuracy: 77.24%
              precision    recall  f1-score   support

           0       0.86      0.42      0.56        43
           1       0.75      0.96      0.85        80

    accuracy                           0.77       123
   macro avg       0.81      0.69      0.70       123
weighted avg       0.79      0.77      0.75       123
```



```
Naive Bayes Accuracy: 78.05%
              precision    recall  f1-score   support

           0       0.90      0.42      0.57        43
           1       0.76      0.97      0.85        80

    accuracy                           0.78       123
   macro avg       0.83      0.70      0.71       123
weighted avg       0.81      0.78      0.75       123
```

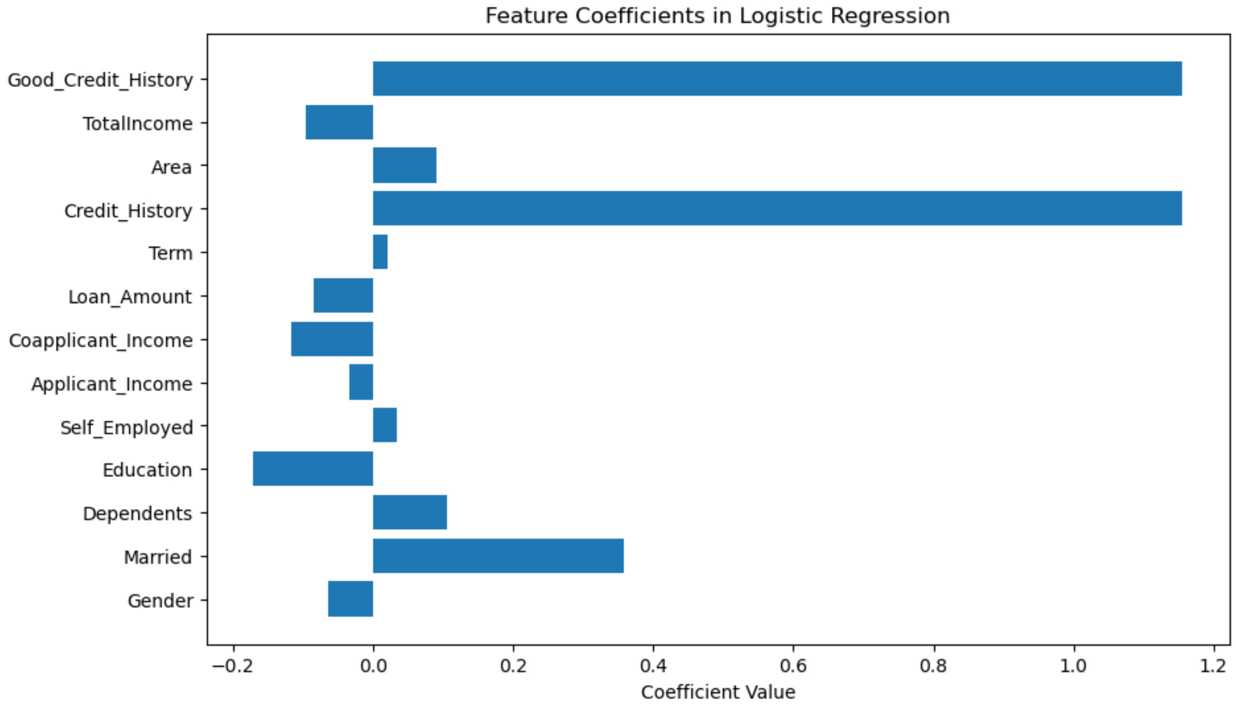
Using Grid Search Cross Validation, we found that the best value for C was 0.1. However, when incorporating this into our Logistic Regression model, there was seemingly no effect on the accuracy or classification report.



```
Final Model Accuracy: 78.86%
              precision    recall  f1-score   support

           0       0.95      0.42      0.58        43
           1       0.76      0.99      0.86        80

    accuracy                           0.79       123
   macro avg       0.85      0.70      0.72       123
weighted avg       0.83      0.79      0.76       123
```

Take-Aways/Final Conclusions

Finally, as a post-modeling analysis, I created a graph to understand the coefficients assigned to each feature in the Logistic Regression model, to see the impact of each feature of the model's predictions.



All in all, Logistic regression proves to be the most effective model for loan approval prediction. Credit History emerges as the most influential feature, which emphasizes the importance of maintaining a good credit history for loan applicants.

Using my findings a loan institution can enhance their business processes, reduce risks, and make more informed lending decisions. This could come in many forms. For example, implementing an automated loan approval system based on these insights from my predictive model, which could streamline the loan approval process and reduce manual effort and operating costs.

Another example, this model can be used to more effectively identify and assess the risk associated with each loan application. By better understanding the key features influencing loan approval, a loan institution can better identify high-risk applicants, proactively mitigating potential losses. Additionally, this can be a guide for customer segmentation. Using key features identified in this project, like credit history, education, etc. loan institutions can tailor loan products and terms to different customer segments, advancing their product offerings for various risk profiles.

I feel it's also important to highlight the societal and economic benefits to using such a model. Implementing a data-driven approach could help ensure fair lending practices. The loan institution can avoid discriminatory practices and provide equal opportunities to people from diverse backgrounds to access loans. Access to credit facilitates investment in education, businesses, and property ownership. By contributing to these areas, for a diverse clientele, the loan institution is helping foster economic growth at both individual and community levels.