

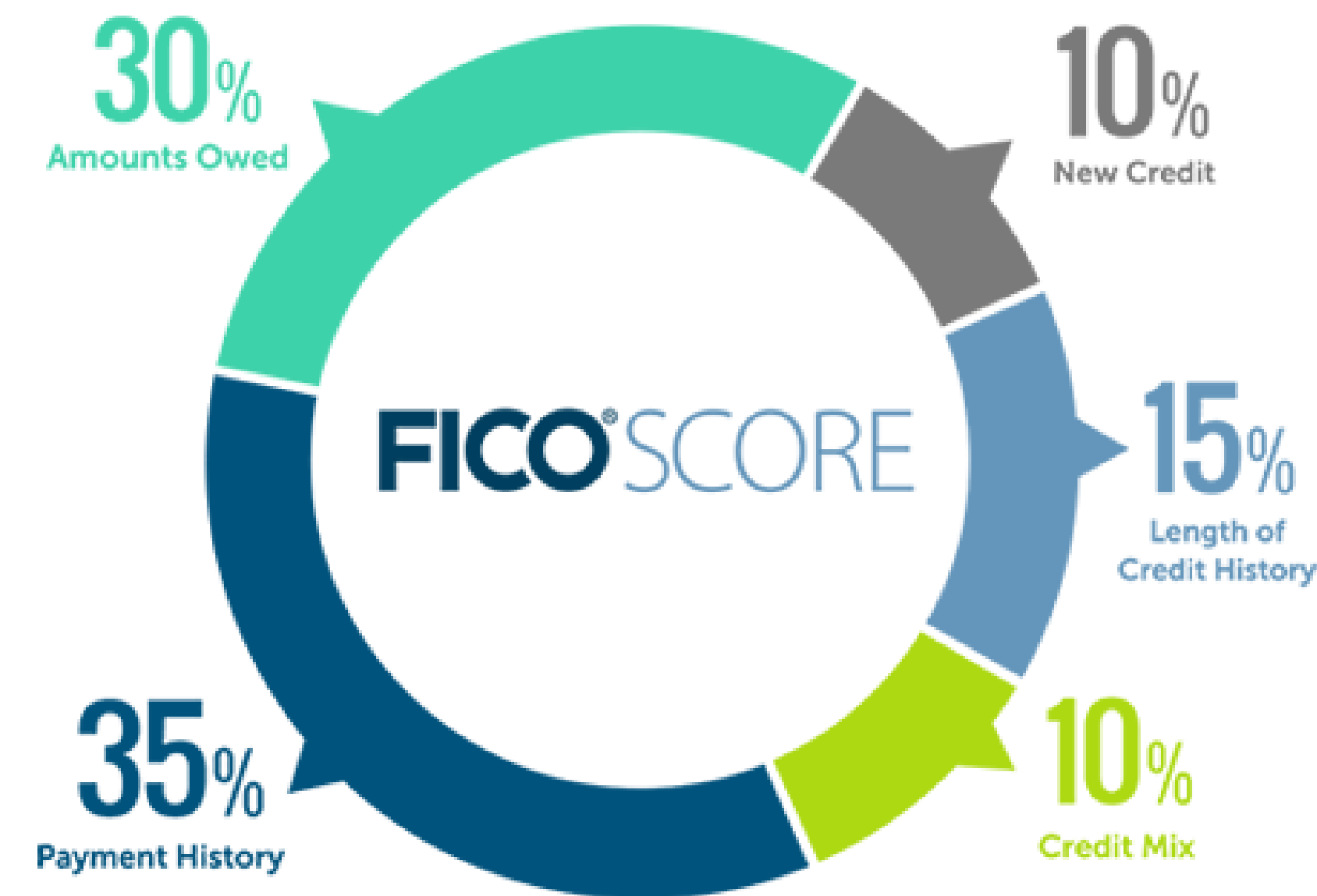
Predicting Credit Risk

Caleb Puckett, Dr. Robert Kelley

Bellarmino University Data Science Program

ABSTRACT

A credit score communicates the likelihood of defaulting on a loan; this project will use predictive analytics to analyze factors not used as typical determinants of this score. The current factors of credit score have been chosen for good reason, but in this project, an attempt will be made to link credit risk with more general indicators of one's situation. This can help remove the high bar to entry to obtain a good credit score.

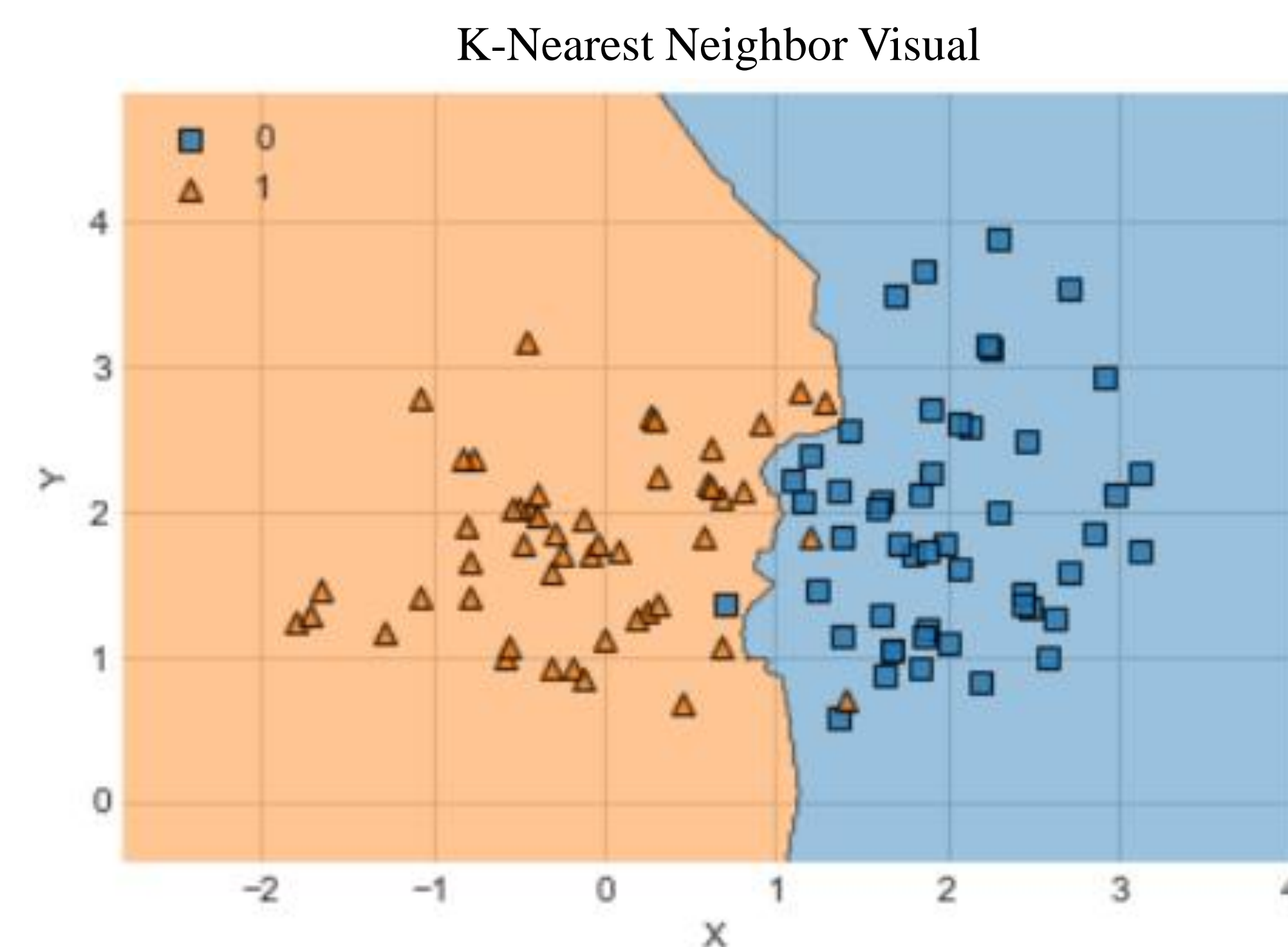


OBJECTIVES

- Using more general circumstances to predict likelihood of credit default
- The current measure is effective at predicting this risk, but every factor is dependent upon previous borrowing activity
- This creates difficulty when individuals without previous borrowing experience attempt to get a good credit score
- Catch-22 scenario which could be handled by understanding other indicators of credit risk
- This would not replace current credit scoring system – it would complement it for those with no borrowing history

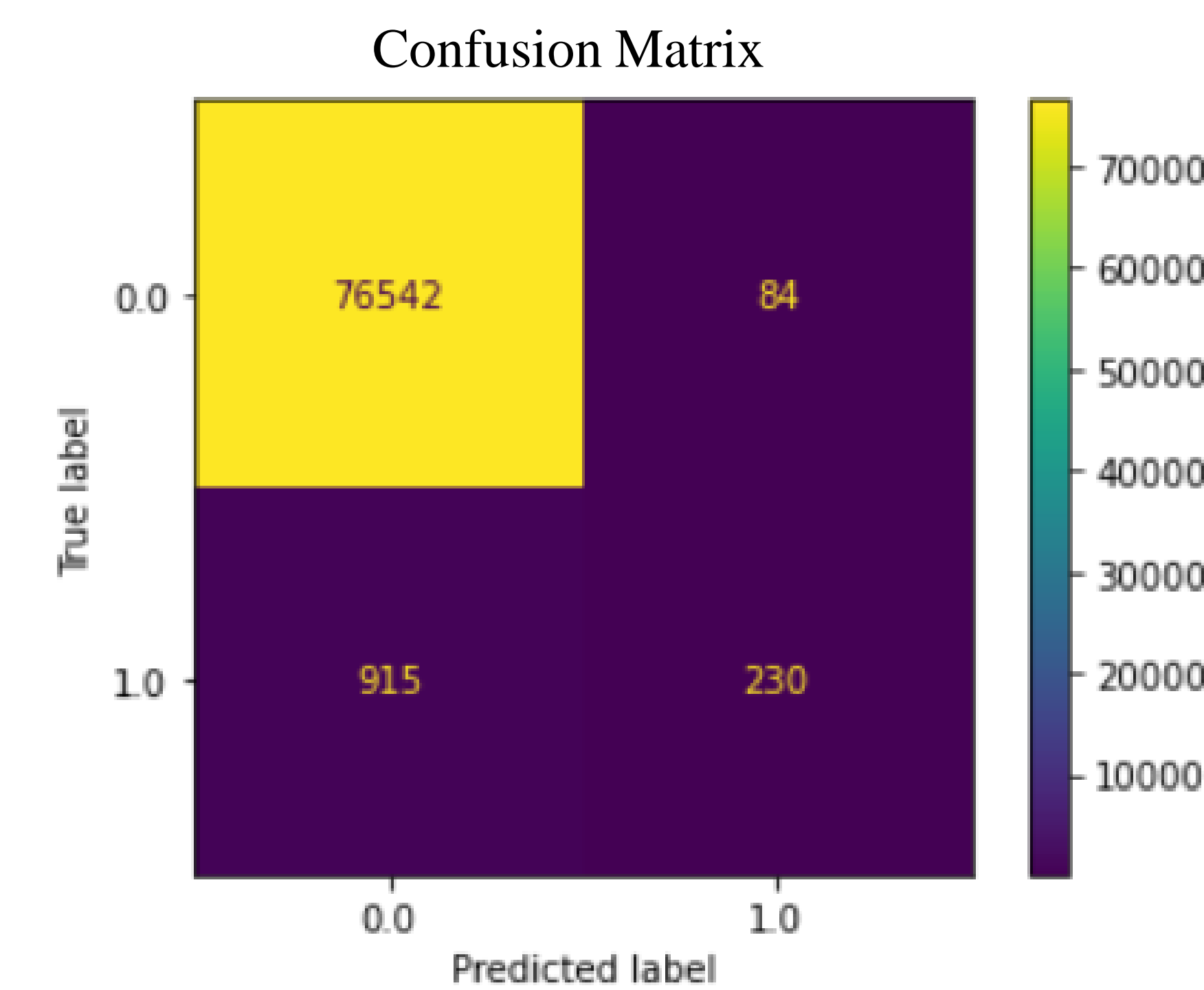
METHODS

- Data was gathered online from Kaggle
- Data cleaning and manipulation were performed
- Exploratory data analysis was performed
- Two machine learning models were implemented for predictive purposes: Decision Tree and K-Nearest Neighbor
- Work was all done in Python besides the occasional use of Tableau for visualizations



RESULTS

- Models used in this project were not especially accurate at predicting inability to repay loans
- Both models were ~98.5% accurate – insufficient because ~98% of the payments in the data are not late
- Approximately the same accuracy could be obtained if our model predicted on-time payments for every individual
- Below is a confusion matrix showing the prediction results



CONCLUSIONS/FUTURE WORK

- The results were not robust and did not accurately predict credit risk as hoped
- Decision Tree and K-Nearest Neighbor machine learning models were used, but there are more models worth trying
- These include Random Forest (many Decision Trees) and Kernel Support Vector Machine (more complex analysis)
- These will likely provide more accurate results based on their ability to analyze more rigorously
- After running these additional models, it is worth considering the addition of other variables not included in these data

REFERENCES

Data obtained from Kaggle at the following link:

https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction?select=application_record.csv

CONTACT

Puckettcaleb7@gmail.com

(502)-608-6668