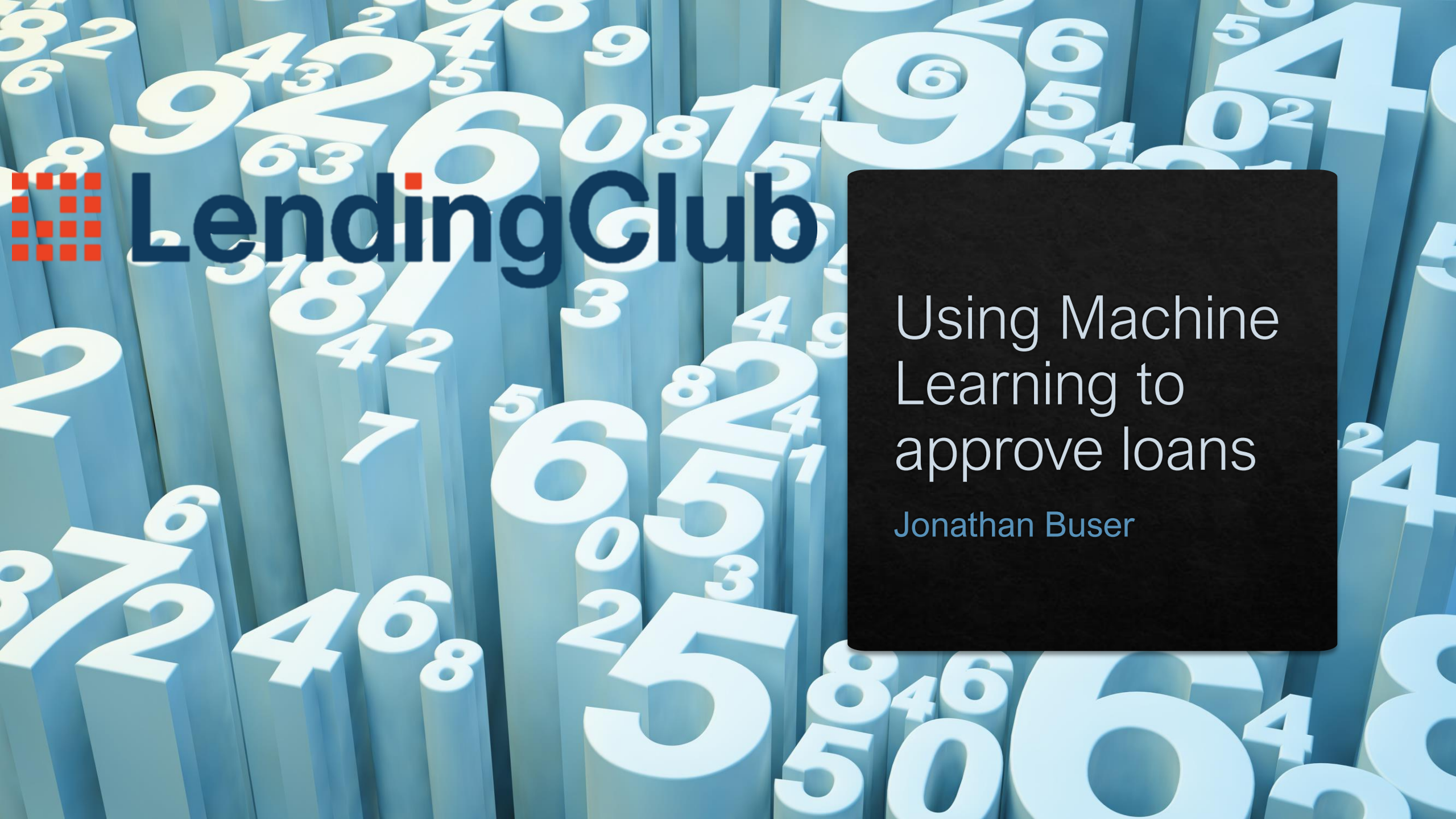


Using Machine Learning to approve loans

Jonathan Buser

Introduction

The Lending Club

Lending Club is a peer-to-peer lending platform that connects borrowers and investors.

One of the biggest challenges for Lending Club is identifying high-risk borrowers who are likely to default on their loans.

The goal of the model is to help Lending Club reduce the number of defaults and save money in the long run by increasing the recall of our model.

Dataset

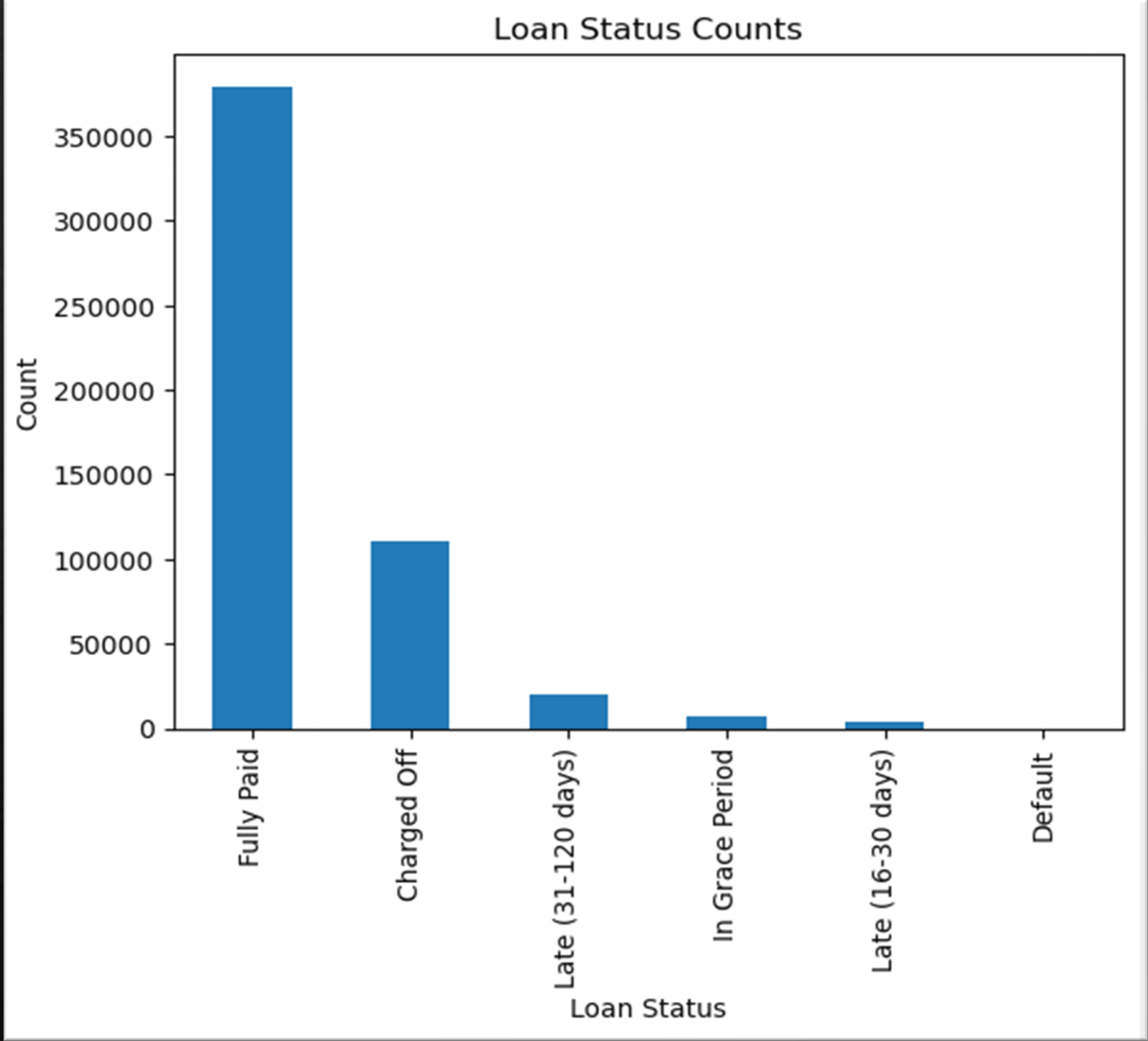
The Data

The dataset used in this project consists of Lending Club's loan data from 2016 to 2018, including 521,435 records and 95 features or noncurrent loans.

The target variable is the loan status, which indicates whether the loan was paid off fully or defaulted.

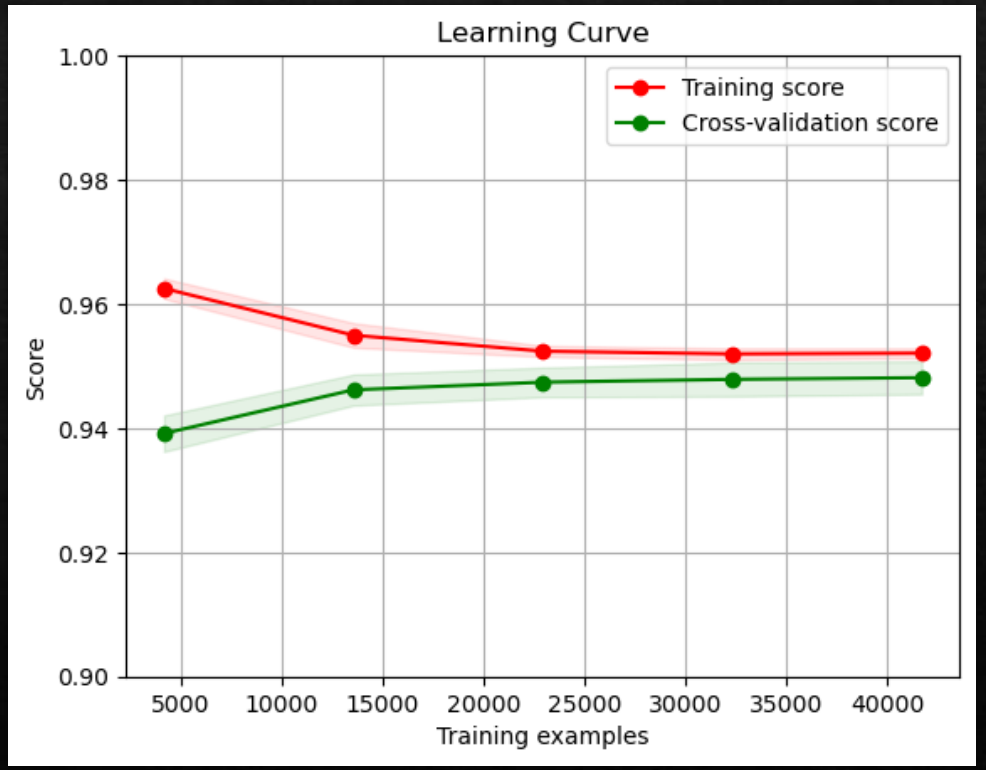
The dataset is highly imbalanced, with 72% of loans paid off fully and 27% defaulting.

Removed all features that are not available at loan application.



Modeling

XGBoost and Smote-EEN	We used a technique called SMOTE-ENN to create combination of over-sampling and under-sampling to balance the class distribution in the dataset and improve the model.
	We also applied feature engineering techniques, such as one-hot encoding and feature scaling, to prepare the data for modeling.
	Finally, we trained the model using an algorithm called XGBoost, which helps the model learn from the data and make predictions based on what it has learned.
	I also tried a KNN, Logistical Regression and Random Forest but the best and fastest model was XGBoost.



Results

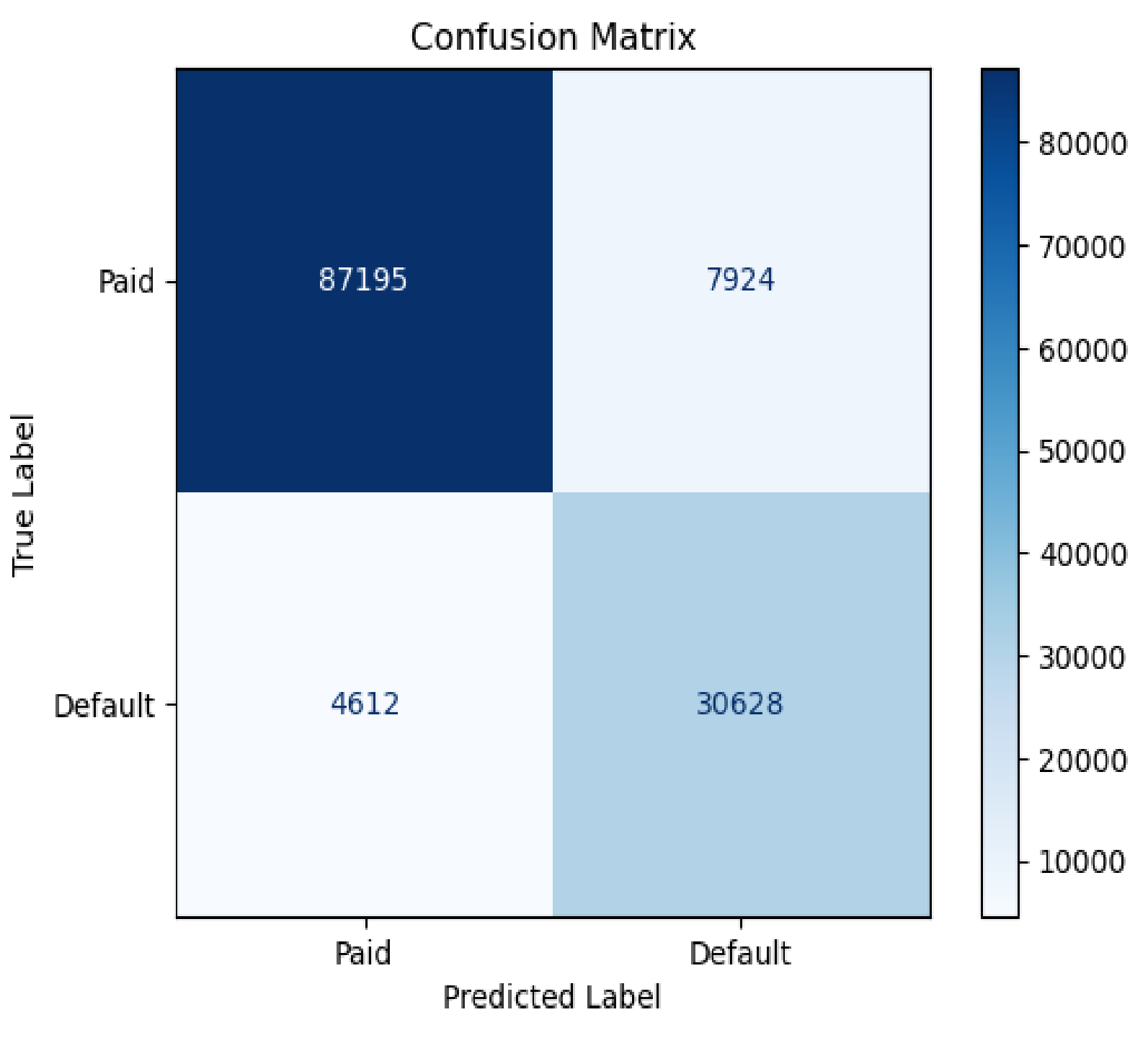
Accuracy of 90.38%

Recall of 86.91%

Precision of 79.45%

F1 score of 83.01%

Overall, our model performed well on the task of predicting loan default with high accuracy, recall, and F1 score.



Conclusion

What we Learned

Our XGBoost model trained on Lending Club's loan data with SMOTE-ENN oversampling and feature engineering techniques can potentially reduce the number of defaulting loans from 27,000 to 5,000 per 100,000 applications, bringing the default rate down to 5%.

Assuming people who default don't pay off on average 75% of their loan, our model can potentially save Lending Club around $228,839,500 per year (assuming 100,00 applications per yet).

We acknowledge that our analysis ignores the potential profits from loans that our model would have denied even though they paid off their loans.

Our model can help Lending Club identify high-risk borrowers and reduce the number of defaults, ultimately saving them a significant amount of money in the long run.

Next Steps

Firstly, we recommend further testing and validation of our model using more recent data, as this can ensure its continued effectiveness in predicting loan default in the ever-changing landscape of the lending industry.

Collaboration with Lending Club's data and risk management teams can provide valuable insights and ensure the model's compatibility with their business operations. I’m not 100% on how to fill all the null values for the credit checks.

In conclusion, while our model shows great potential for Lending Club to reduce the number of defaults and save significant amounts of money in the long run, there is still room for improvement and refinement.

Thank You!

Jonathan Buser

https://www.linkedin.com/in/jonathan-buser/

https://github.com/SpaceMonkey0453