

A Credit Risk Machine Learning Classification Model



Problem

- ▶ How to minimize default Rates on personal loans?



The image shows a 'NOTICE OF LOAN DEFAULT' form. The form is titled 'NOTICE OF LOAN DEFAULT' in large, bold, black letters. Below the title, there is a section for 'Personal Information' which includes fields for 'Name (Last)', 'Name (First)', 'Name (Middle Initial)', 'Address (Mailing Address)', 'City', 'State', 'Zip', 'Home Telephone', and 'Other Telephone'. There are also fields for 'E-Mail Address' and 'Degree Year'. A blue and gold pen is resting on the form, pointing towards the 'City' field. The form is placed on a wooden surface.

The Dataset

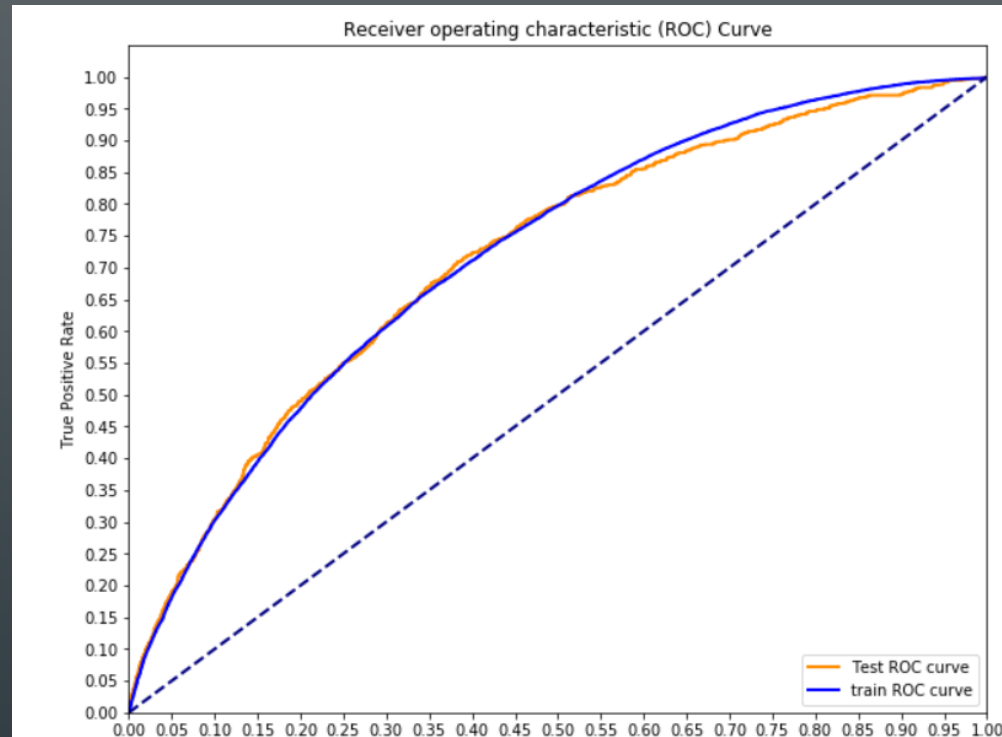
- ▶ Lending Tree borrowers data
- ▶ Conducted between 2007 and 2011
- ▶ 42500 rows x 52 columns
- ▶ Available on lendingtree.com

Data Preparation Steps

- ▶ Handling missing Values
- ▶ Eliminating columns leaking information from the future
- ▶ Synthetic Minority Oversampling (SMOT) for class imbalance

Evaluation Metric

- ▶ Accuracy not a good indicator
- ▶ True positive rate (recall)
- ▶ False positive rate

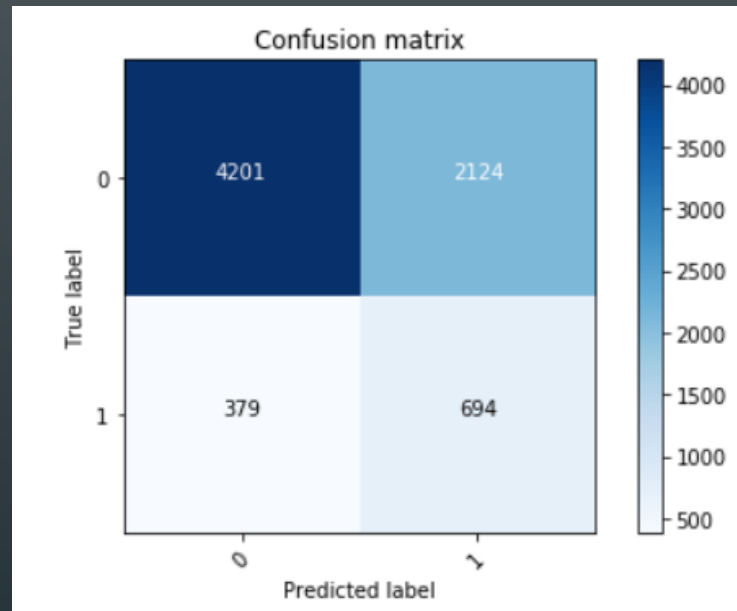


Different Classification Models

- ▶ Logistic Regression
- ▶ Support Vector Machine
- ▶ Random Forest
- ▶ XGBoost

Which Model Did Best ?

- ▶ Logistic regression provided the best results



Conclusion

- ▶ We can predict 65% of the defaults that were initially approved by Lending Tree screening process
- ▶ Drawback : the model still rejects a significant portion of the applicants who were not going to default

Next Steps

- ▶ Improve gridsearch to optimize random forest and XGBoost model performance
- ▶ Try different approaches to handling class imbalance
- ▶ Determine levels of interest rates insuring profitability despite undetected defaults