Demographics or Economics?

Approaching Class Imbalance Problems in a Sales Environment

Andrew Mooney

The Sales Problem

The big questions in any industry:

- **Who?**
- **Why?**
- **When?**

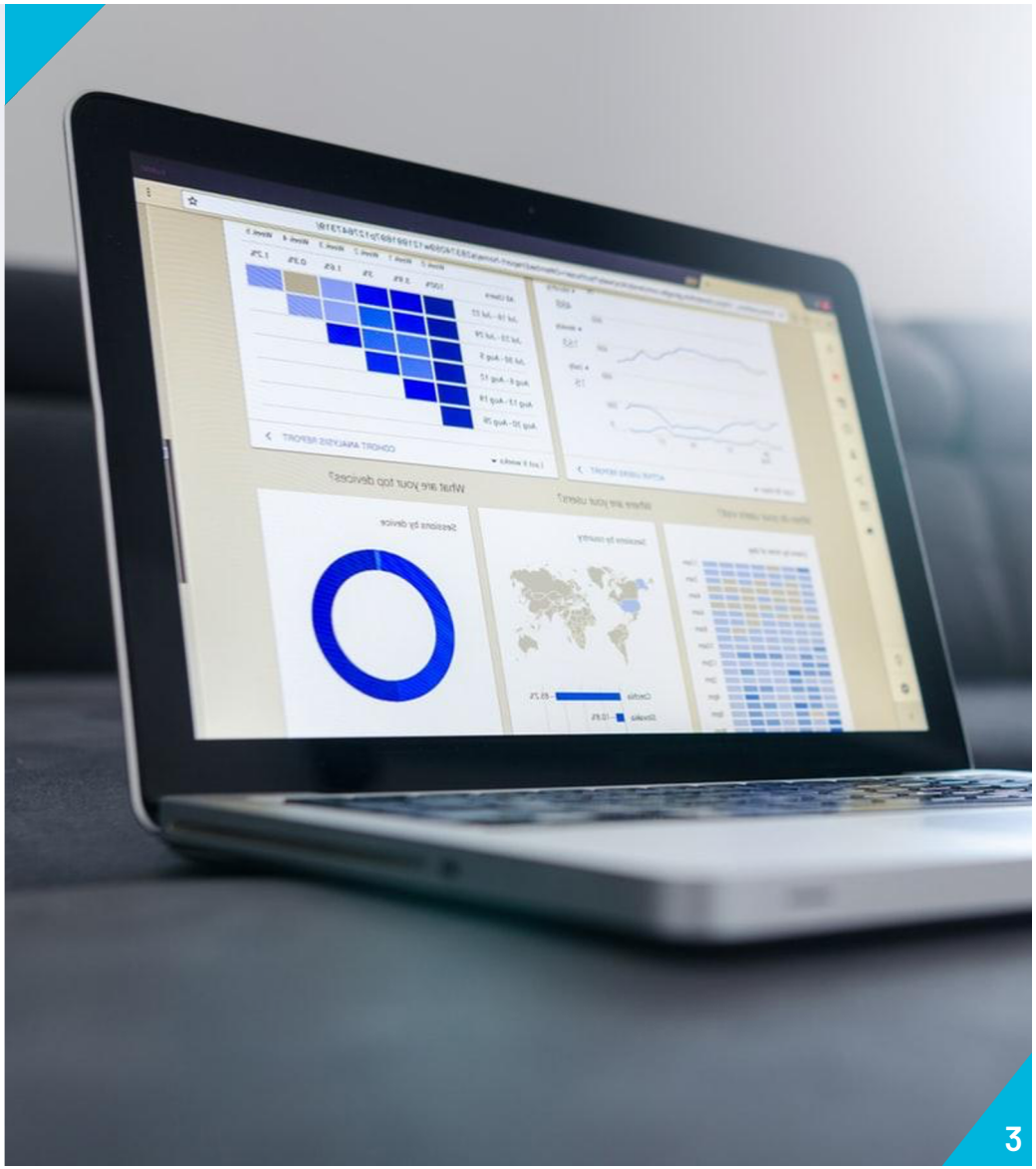
Indicators to explore:

- **Demographics**
- **Purchase History**
- **Macro Economic Factors**

What's worse? False Positive or False Negative?

The Data

- Bank product sold on subscription basis
- 41k rows of information
- Massive class imbalance problems (88.9% "no")
- Few nulls, but a number of 'unknown' categories



Methodology

Model Selection

- False Positives vs. False Negatives – maximize the F1 Score
- Accuracy must be greater than baseline 88.9%
- Explore the following models:
  - Logistic Regression
  - K Nearest Neighbors
  - Decision Tree
  - Random Forest
  - XGBoost

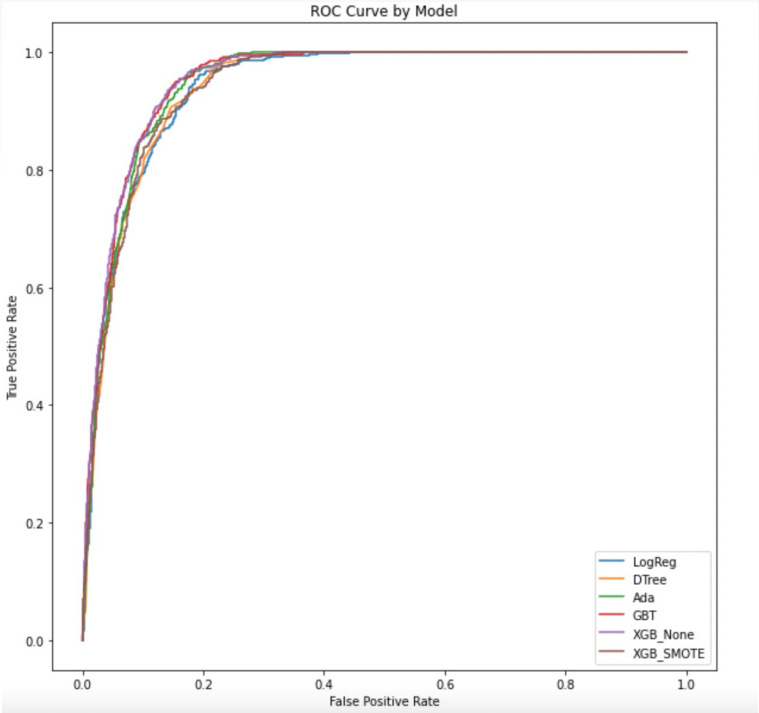
Imbalance Issues

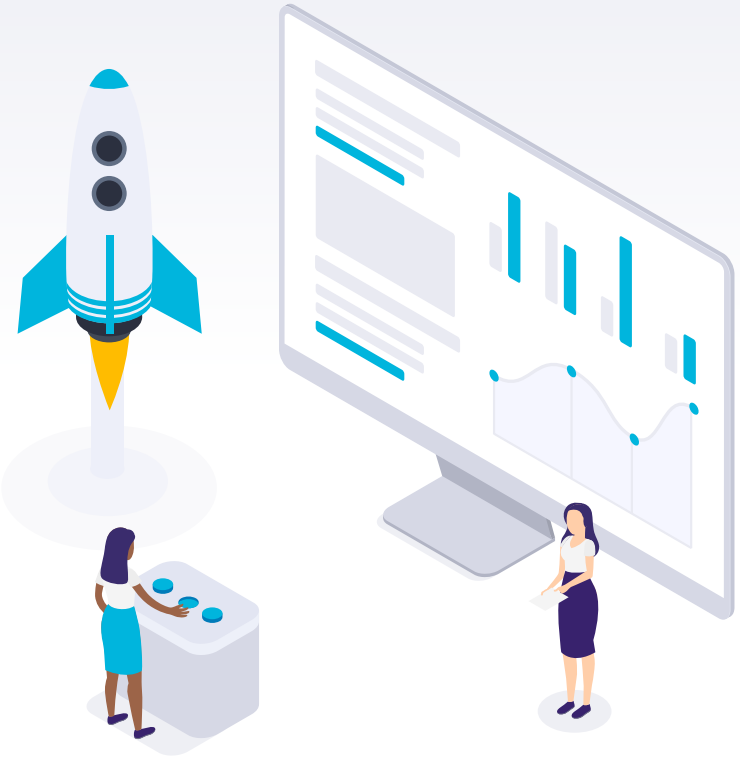
- No Transformation
- SMOTE
- Under-sample the majority class
- Compare model scores on all 3 datasets

Final Comparison

- Performed on holdout data
- Measured by accuracy and FP vs. FN rates

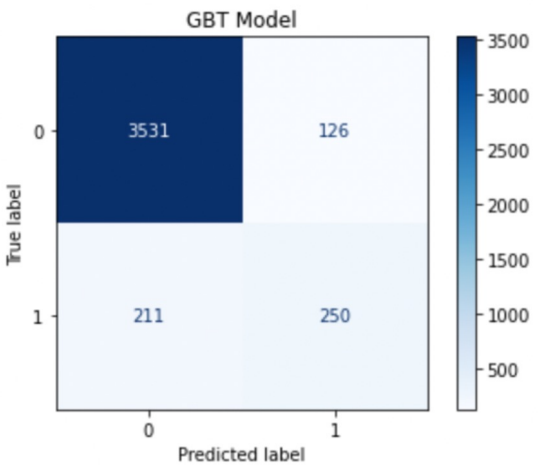
Results and Conclusions

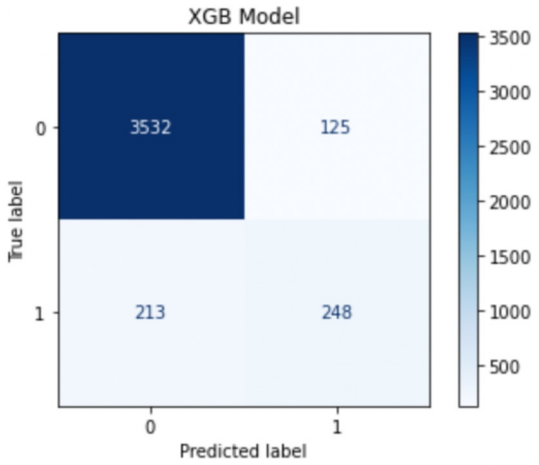




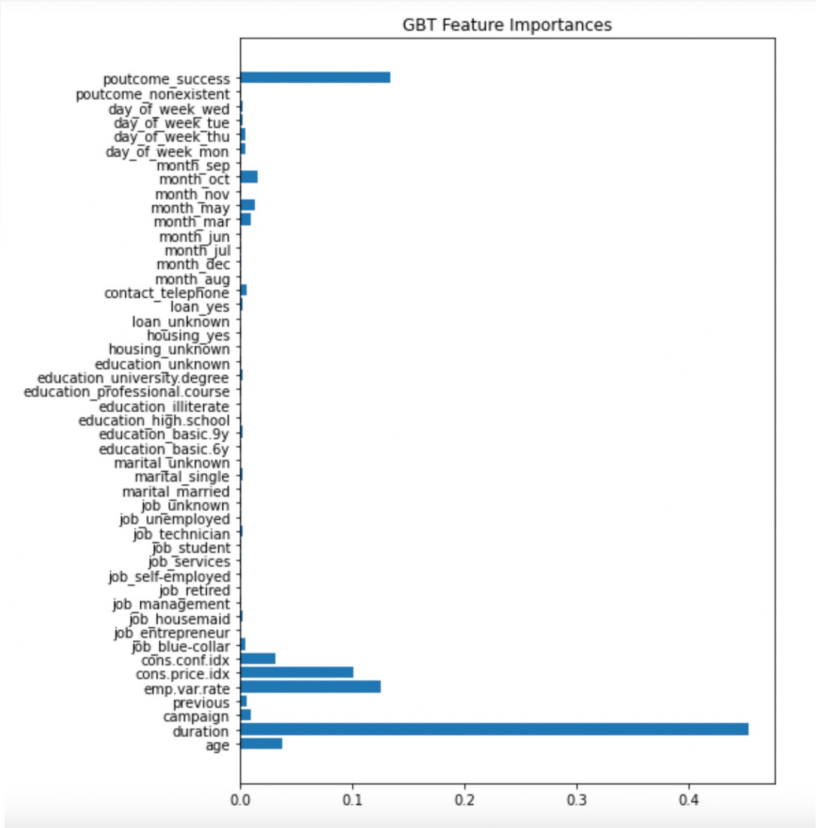
Final Comparison

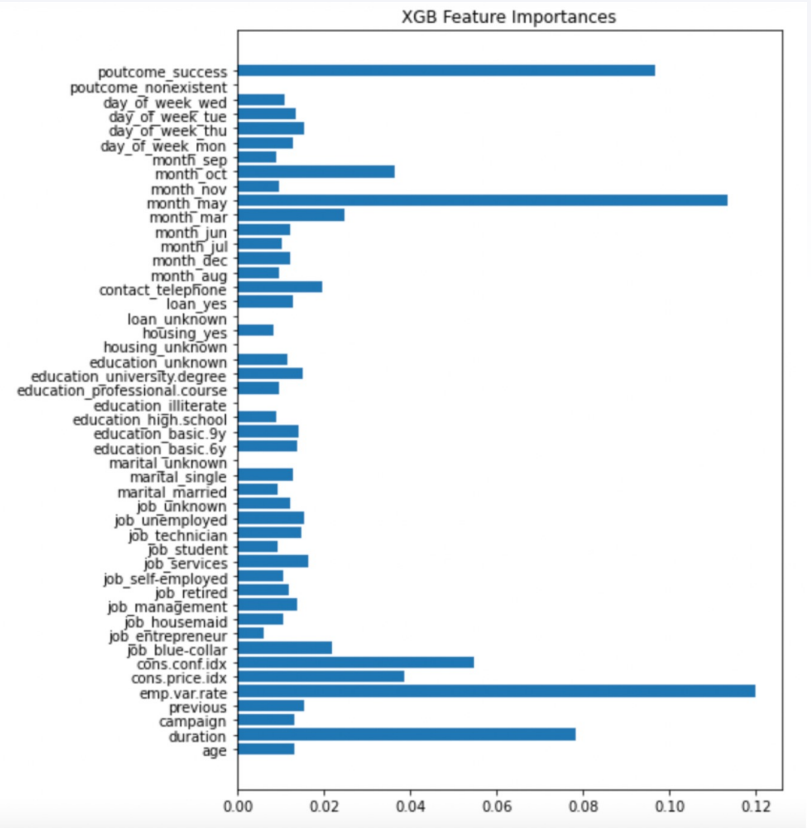
	LogReg	DTree	Ada	GBT	XGB_None	XGB_SMOTE
Accuracy	0.914279	0.913550	0.911608	0.918164	0.917921	0.902623
Precision	0.617391	0.624703	0.699588	0.664894	0.664879	0.555970
Recall	0.616052	0.570499	0.368764	0.542299	0.537961	0.646421
F1 Score	0.616721	0.596372	0.482955	0.597372	0.594724	0.597793
ROC-AUC Score	0.783963	0.763647	0.674401	0.753922	0.751890	0.790670





Most Important Predictors





Conclusions

- Macro-Economic Factors are useful
- Demographics can be important
- Past purchase history is a strong indicator
- The data with the least transformations offered the best results
- More data is always better than less

THANK YOU!

Any questions?

You can find me at:

- andrew.k.mooney@gmail.com
- github.com/andrewkmooney/bank_campaign