The Erdős Institute Data Science Bootcamp
Team 11 project executive summary

Team members: Drew D. Ash, Laura Brade, Jamie Kimble

Data: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

GitHub: https://github.com/DrewAsh13/Erdos-2023_Project

Introduction: Banks and other financial institutions require new assets under management to function properly. To this end, acquiring new deposits is a fundamental goal for banks and other financial institutions. Direct marketing campaigns are a tool banks use to acquire new deposits. Our objective is to construct a predictive model which classifies whether or not an individual will deposit assets at a bank given certain demographic characteristics as well as economic indicators. Our data comes from a direct marketing campaign of a Portuguese banking institution which comprises 20 feature variables and 1 target variable.

Stakeholders: Bank management, director of marketing, director of asset management. More broadly, this kind of classification would be of interest to those interested in converting new customers or modeling such as default risk, such as on a bank loan.

KPI's: We have three metrics we compare our models by: precision, recall, true positive rate, and false negative rate. Additionally, we aspire to create a robust pipeline to train future models in collaboration with the marketing team.

Methods: We explored five different classification models: logistic regression, KNN, decision tree, and a random forest. While preprocessing the data we made two critical decisions. First, we dropped the feature "duration" which records how long, in seconds, the phone call lasted with the individual; This feature is not relevant for predicting who to call as we cannot predict in advance if someone will pick up the phone or not. Second, the variable "pdays", recording the number of days that passed by after the individual was last contacted from a previous campaign, specially encoded those individuals not previously contacted as 999. We discretized "pdays" into a categorical variable. We then compared our models using the metrics of precision, recall, and their true positive rate- as only 4,640 of the 41,188 contacted individuals in our data set (11.2%) deposited money as a result of the direct marketing campaign.

Results: From our analysis, the decision tree performed the best across all metrics- the results are presented in the table below. The Random Forest model seems unreasonably accurate and we would need to spend more time fine tuning this model before feeling confident in our model. Furthermore, our decision tree identified the Euro-libor 3 month rate, the consumer price index (CPI, cons.price.idx) and calling on Thursdays to be the most significant contributing factors.

Model	**Precision**	**Recall**	**True Positive Rate (TPR)**	**False Negative Rate (FNR)**
Logistic	**.6442**	**0.2166**	**.2166**	**.7834**
KNN	**.5274**	**0.2802**	**.2802**	**.7198**
Decision Tree	**.6745**	**0.1853**	**.6745**	**.3255**
Random Forest	**1**	**0.0011**	**1**	**0**

Future Work: We have four avenues of extending our work: increasing scalability and creating pipelines; Spending more time tuning hyperparameters and using boosting methods such as XGBoost or Adaboost, to increase performance on our KPI's; Trying a neural network, and spending more time on feature selection. In conjunction with feature selection, we would couple a more detailed analysis of not only the who the direct marketing campaign contacted, and ultimately converted, but also an analysis of our current depositors. We would use this information paired with the feature selection to inform how any additional dollars spent on the next direct marketing campaign could most effectively be used to increase the conversion yield.