Dalya Adams
MSDS 422 section 59
Bank Marketing Classification

In order to execute an effective marketing campaign, ensuring that the right client sees the right ad, or comes in contact with the marketing medium, is essential. While marketing the product to every single customer of the company, or every person in an area, could be done, it is neither practical nor the best way to spend marketing money. In order to ensure the right customer sees the intended message, clusters of customers most likely to buy the product need to be identified. By targeting the customers in these identified clusters, companies are better able to utilize their limited resources.

Throughout this project, the goal was to identify which customers would respond to the bank's most recent marketing call. Specifically, has the client subscribed to a term deposit? While the dataset had many different variables, from age of the client to the date they were contacted, this study focused on the impact of three specific variables on the client subscribing to a term deposit. These three variables were: default (Does the client have credit in default?) , housing (Does the client have a housing loan?) and loan (Does the client have a personal loan?). By identifying the impact of these three client details on the likelihood of them subscribing to a term deposit, the bank is better able to target these customers, increasing the efficiency of the marketing campaign while also decreasing the cost associated with the marketing campaign.

In attempting to increase the efficiency of the bank's marketing campaign, machine learning classification models were implemented to classify which customers would subscribe to the term deposit. The models were then evaluated for their accuracy utilizing the area under the receiver operating characteristic (ROC) curve. The data utilized for this analysis is a subset of the bank's clients who were approached with the term deposit option. In the dataset, 3,705 clients were approached with only 71 answering in the affirmative. The dataset consist of the demographic details of the clients, as well as the 'yes' or 'no' answer provided when presented with the term deposit opportunity. Two models were utilized in this analysis: a Logistic Regression and a Naïve Bayes model. These two models, while both well-known classification models, operate differently. The logistic regression determines the probability of an event occurring by minimizing the error from the model, whereas the Naïve Bayesian model estimates the joint probability of an event occurring, based on the assumption that all features are conditionally independent of each other.

Utilizing Scikit Learn in Python, both the Logistic Regression and the Naïve Bayesian models were implemented. To ensure that the accuracy of the model would hold when presented with new data, 10-fold cross validation was utilized on both models. The average area under the ROC curve for the logistic regression was 0.6117 whereas the area under the ROC curve for the Naïve Bayesian model was 0.6110. Neither of the models are extremely accurate but provide us more detail than random chance would. Since the logistic regression is the more accurate model we will look into this model in more detail. Exporting the coefficients from this model, we learn that defaulting on a loan increases the probability a client will sign up for a term deposit, whereas having a home loan or personal loan decreases the probability of signing up for a term deposit.

As a recommendation for management, pitching term deposits to banking clients who have previously defaulted on a loan will increase the probability that a client will purchase a term deposit. Clients with housing loans and personal loans are not ideal to target, as these types of loans decrease the probability that a client will purchase a term deposit. An additional segment of the population that has an increased probability of purchasing a term deposit are the subset of clients without housing or personal loans who have not defaulted on a loan. In short, focusing marketing campaigns on these two subsets of clients, will increase the term deposit adoption rate with in the bank.

Appendix

Coefficients of Logistic Regression

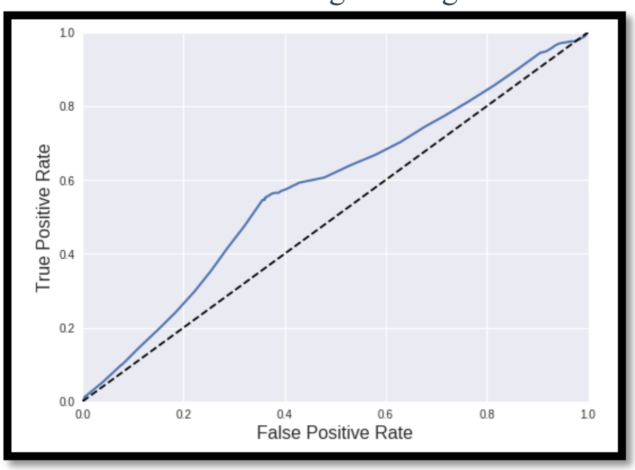
Default	Housing	Loan
0.299	-0.672	-0.540

Logistic Regression model predictions for test cases

Logistic regression model predictions for test cases:

	default	housing	loan	predict_NO	predict_YES
0	1.0	1.0	1.0	0.945729	0.054271
1	1.0	1.0	0.0	0.892349	0.107651
2	1.0	0.0	1.0	0.900786	0.099214
3	1.0	0.0	0.0	0.811988	0.188012
4	0.0	1.0	1.0	0.953277	0.046723
5	0.0	1.0	0.0	0.906588	0.093412
6	0.0	0.0	1.0	0.914016	0.085984
7	0.0	0.0	0.0	0.834890	0.165110

ROC Curve for Logistic Regression



Precision versus Recall for Logistic Regression