SDS 323: Final Project

Kyle Carter, Crystal Tse, Jinfang Yan

Professor James Scott

5/11/2020

Bank Marketing

Abstract

In this analysis, we seek to discover what would make marketing campaigns for a Portuguese bank's term deposit offering more effective during the 2008 - 2013 recession. We use logistic regression, decision tree, random forest, naive Bayes, and hierarchical clustering machine learning models to determine the most useful variables in predicting if a contact will subscribe to a term deposit. We analyzed a large dataset of 41,188 observations and 21 features that included demographic information on the bank client, the history of contact with the client, and macroeconomic indicators. Due to an imbalanced data set, we rely on AUC (area under the curve) as our evaluating metric and find that the logistic regression model was the most effective when cross validated on a test set. The random forest, when validated, was close to the logistic model. Our most important recommendations include: contact people who were called during the previous campaign, limit the number of calls during a campaign, contact retirees and students, spread the volume of calls over the full year, and be vigilant about macroeconomic trends such as the Euribor rate and consumer confidence index.

Introduction

We want to predict not only if a customer will deposit at the bank, but what potential actionable strategies the bank can undertake to attract a larger number of customers in its marketing campaigns. Often, multiple phone calls to the same client were required to assess if the client would subscribe to the product of a bank term deposit or not. The goal is to conserve resources by preventing calling people who are not likely to be interested in term deposits, and instead find a more receptive audience.

The data contains 41,188 observations from telemarketing campaigns of a Portuguese banking institution promoting term deposits in the period 2008 - 2013, as Portugal was experiencing a financial crisis. In 2008, Portugal plunged into the international Great Recession, and 2010 - 2014 was the most challenging part of the financial crisis, characterized by an international bailout and austerity by the government. Thus, it is worth noting that this data does not reflect an economy in steady state and is more reflective of saving habits in times of economic hardship.

A term deposit, or time deposit, is an interest-bearing bank account with a predetermined date of maturity that generally offers a greater rate of return than savings accounts. The dataset includes 21 attributes, including a binary variable y that indicates whether the client subscribed to the deposit or not; the contact communication type; various traits about the potential customer such as age, job, education level, engagement in a housing or personal loan, and

history of contact with the client; and various measures of the health of the economy such as the consumer confidence index.

This analysis has key implications for understanding factors that affect individual decision-making at both the personal and macro level. Understanding market segmentation and socioeconomic background indicators of what makes certain people more likely to become customers has tangible benefits for the bank, but it can also have higher-level implications for macroeconomists seeking to understand the impact their policies may have on aggregate saving, especially in times of economic hardship like Portugal was experiencing at the time.

This data set presents several problems, which we have tried our best to deal with reasonably. The first is that it is imbalanced; only about 11% of the observations accepted a term deposit. We have tried to alleviate this issue primarily by choosing tree-based methods, which implicitly look at both classes via its splitting rules. Also, when we evaluated our models on test sets, we largely ignored accuracy, since it scores the overall class distribution, and focused on sensitivity and specificity. Therefore, we valued the ROC and AUC for validation since they incorporate both metrics. The second is that several variables have many unknown values. We purged unknown values from the data set after thoroughly examining each variable, reducing our data set to about 38,200 observations. However, in the future it would be worth revisiting with advanced methods of imputation, which come with their own drawbacks.

To understand what might cause a person to be more likely to subscribe to a bank term deposit, several methods were considered after data preprocessing. A few duplicate rows were removed, and while there were no missing values, a large fraction of observations had “unknown” values. It was also necessary to remove the “duration” variable, or the measure of the length of the last phone call in seconds. This is in contrast with the original research paper, which preserved the duration variable (Moro et al., 2014). However, the duration of the last phone call is highly correlated with the dependent variable of subscription to a term deposit. Clearly, if the customer was completely unreceptive to telemarketing and a deposit subscription, then the duration of the phone call would be 0. Furthermore, understanding duration does not yield actionable insights, since the bank cannot target people if duration is an unknown before contacting them. For this reason, the data preprocessing diverges from previous literature.

Other variables were also determined to have high multicollinearity with each other, leading to an unstable model if kept in the analysis. The variables “pdays,” “nr.employed,” and “loan” were all highly correlated with various other indicators of the history of contact with the customer, macroeconomic indicators, and personal attributes of the client (Fig. 1). In addition, observations with unknown marital status and job type were removed, and the variable “default” added to noise since only 3 observations had credit in default. By removing these highly correlated or noisy predictors, the model should be able to more closely find the relationships between the feature variables and obtain more reliable results.

Figure 1. Correlation Matrix Showing Multicollinearity

Methods

A main factor in determining what analysis could be performed was the fact that the dataset contained a mix of both quantitative and qualitative variables. Given our goal of predicting a binary outcome of subscription, we want to determine the likelihood that a contact will subscribe and create a threshold that gives a higher overall success rate.

In previous literature (Moro et al., 2014), logistic regression, decision trees, neural networks, and support vector machine methods were used. However, neural networks and support vector machine results are difficult to interpret and were not included in this analysis. Instead, we expanded upon the decision tree's susceptibility to noise through usage of random forests and tried to replicate the findings from the logistic regression, while also exploring other methods of data analysis.

We first considered simple models for initial prototyping and understanding of the data. Logistic regression is useful for predictive modeling and classification, which aligns with the goal of predicting whether a customer is likely to subscribe to a term deposit or not.

Naive Bayes is another potentially good model in terms of its relative simplicity and interpretability. It is strong in classification and working with categorical variables, which compose about half of this dataset. A particularly useful aspect is that it returns probabilities, which is directly actionable once a threshold is determined for how much the bank is willing to take a chance that the contact may potentially not become a subscriber. However, naive Bayes might oversimplify with its assumptions that each variable is completely independent of the other. We noticed that some of the variables are closely correlated with each other, such as

previous, pdays, and poutcome, which all are measures of the history of contact with the potential client, so this is a weakness to keep in mind.

Hierarchical clustering was also considered, as it works well with categorical variables. Although clustering may not be ideal for prediction (classification is better), it is worth noting the importance of variables and how they might be related to each other. It also offers benefits for market segmentation and interpretability, since grouping variables together can help provide insight into higher-level patterns and help segment the market.

Decision trees are useful in that they provide class probabilities which are directly interpretable by the bank in determining potential people to contact. They are also a helpful visual guide. However, trees tend to include noise, which blurs the broader insights that the bank wants to discover about its ideal market for term deposits. This can be resolved by using random forests, which average bootstrapped samples of decision trees to filter out the noise and find the main trend.

One conventional way of navigating the qualitative variables is to convert them to binary values that indicate the existence of an attribute or not, a process known as one-hot encoding. However, this would not be appropriate for certain methods such as K-means clustering, which uses Euclidean distance to measure the relative closeness of different observations. By converting all the categorical variables such as job type to dummy variables, this would make the Euclidean distance a somewhat meaningless measure of similarity since there is no reasonable or intuitive mean for dummy variables. The model would run into the “curse of dimensionality”, where observations become quite similar to each other, which is not ideal for classifying what individuals to market towards. Thus, we decided not to use K-means clustering on this dataset.

To summarize, logistic regression, naive Bayes, hierarchical clustering, decision trees, and random forests were used to analyze the data. K-means clustering was ultimately not used due to the difficulty of handling categorical variables.

Results

The random forest model predictably had a high accuracy, due to the imbalanced data set described above. The model produced a high specificity but had a low sensitivity; the model predicts a lot of false negatives but few false positives. However, it should be noted that the random forest model had a higher sensitivity than the logistic regression and decision tree models. This was to be expected but yields tangible results for the bank. This implies that the bank can accurately determine whether a person will subscribe to a term deposit, since there are few false positives, but may also accidentally rule out candidates that are actually likely to subscribe, since the model has a high false negative rate. The bank should not hastily dismiss a potential customer base, so it is worth considering how to reduce this false negative rate.

The variable importance plot for random forests demonstrates that euribor3 is the most important variable, followed closely by age (Fig. 2).

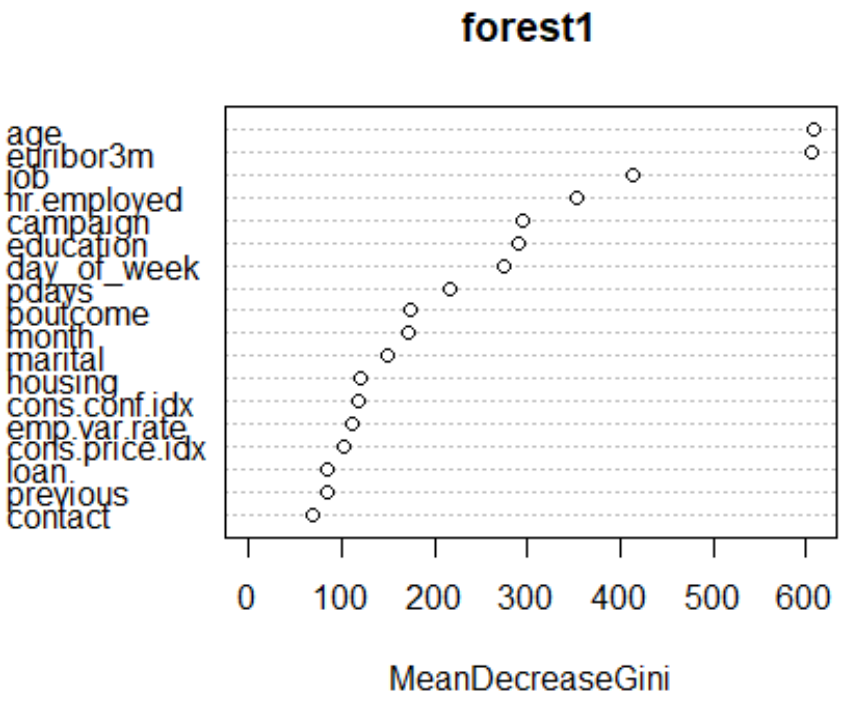

Figure 2. Variable Importance Plot for Random Forests

Hierarchical clustering shows the relationships between the different variables. The y-axis, or height of the tree, measures distance between the clusters. For example, job, education, age, and marital status are all equally equidistant from the other variables, which could mean that those personal traits are not as important in determining subscription outcomes since they are in a different branch than the desired “y” (Fig. 3). However, the result that euribor3m, or the Euribor three-month rate, is closely related to the dependent variable is consistent with the result from random forests.

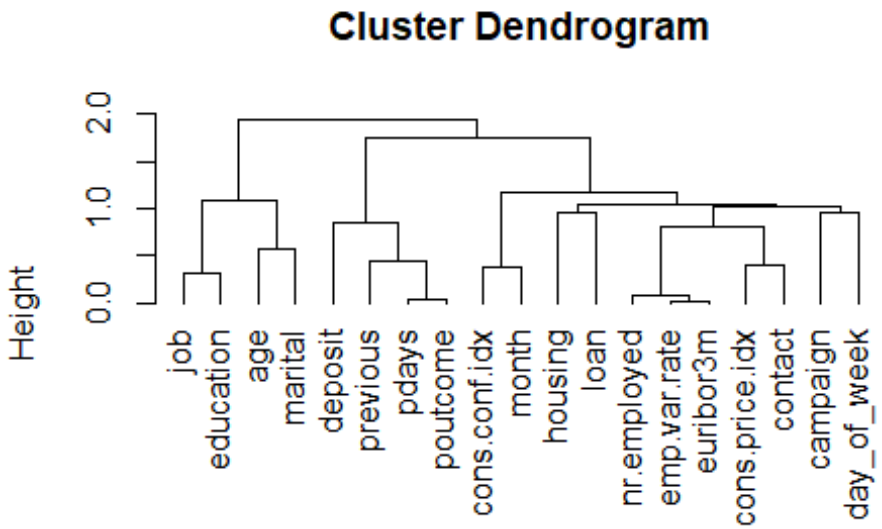

Figure 3. Hierarchical Clustering Dendrogram

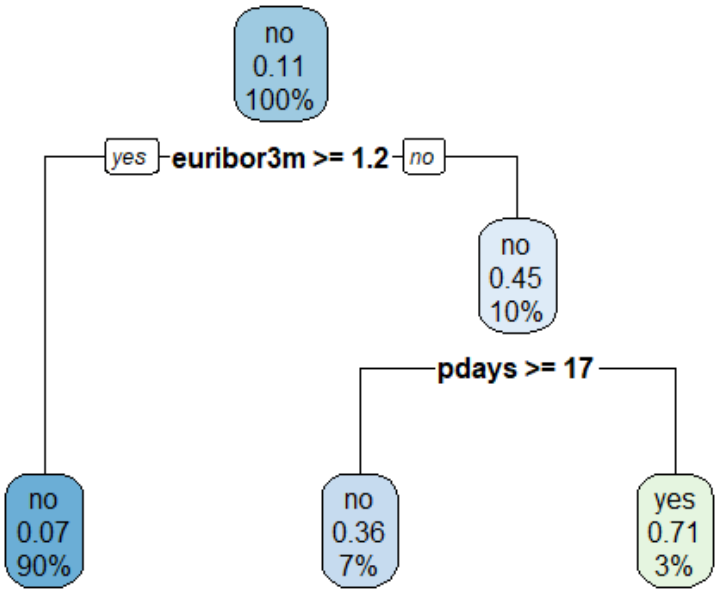

Figure 4. Decision Tree for Term Deposit Subscription

In Figure 4, the Euribor 3-month rate is the predictor variable used for the primary split. When the Euribor 3-month rate is greater than or equal to 1.2, people subscribe to a deposit with predicted probability of 0.07; 90% of people end up not subscribing to a deposit. When the Euribor 3-month rate is lower than 1.2, 10% of people choose to not subscribe to a deposit with a probability of 0.46.

The second split separates the outcome of the previous marketing campaign. When the previous marketing campaign was a failure or nonexistent, then the predicted probability is 0.37 of subscribing, leading to no deposit. Conversely, a node holding a total 3% of the observations exists when the previous outcome was a success.

Conclusion

We found that the logistic regression model performed the best when validated on a test set, followed closely by the random forest model, with AUC scores of 0.79 and 0.78 respectively (Fig. 5). Although every model struggled on this data set, we believe that the flexibility of the logistic model was enough to beat other methods. The random forest is a tree-based method and should be more robust to imbalance; however, it is likely that it overfitted to the training data.

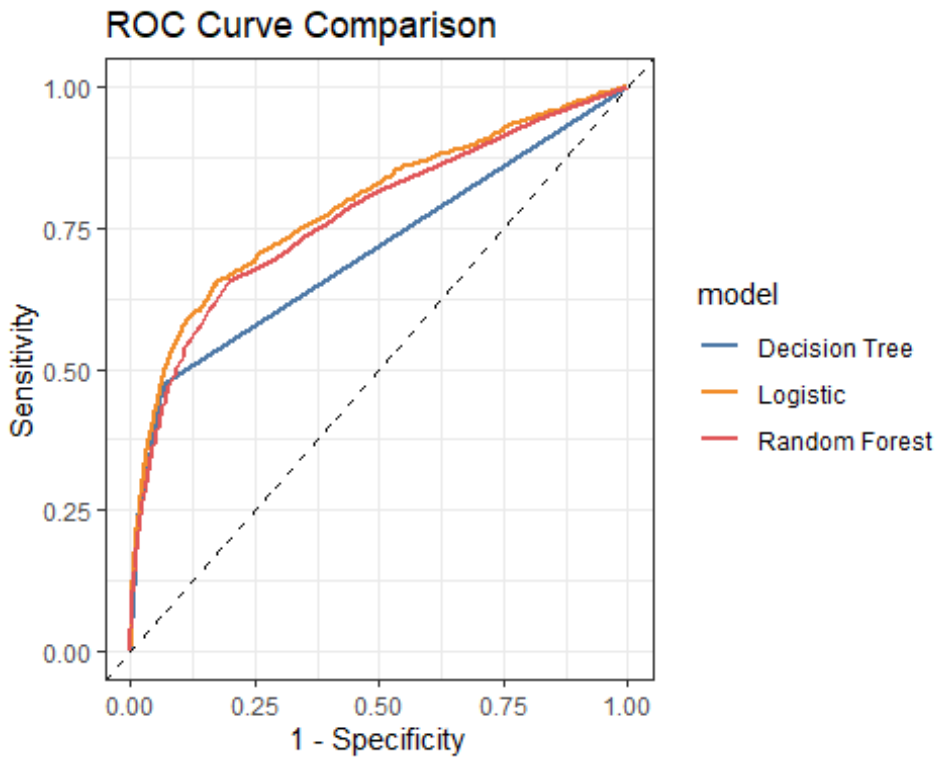

Figure 5. ROC Plot Shows Model Performance

A concern with the random forests model is that the results are not immediately obvious. Faceting on “euribor3m,” it is evident that those that subscribe to a term deposit actually did so when the interest rate was lower, which is counterintuitive (Fig. 6). One would usually expect a higher savings rate to attract more subscriptions to term deposits, since a higher interest rate is meant to incentivize saving. However, the mean interest rate for instances when there was a successful subscription was 2.12%, as opposed to 3.8% for when the contact did not subscribe. The median had an even more pronounced difference in interest rates, with 1.2% and 4.8% for subscribers and non-subscribers respectively. According to the logistic regression, people are

more likely to accept a deposit when the Euribor 3-month rate increases, whereas the decision tree shows the opposite result. This could be inconsequential, as the rates were in flux after the 2008 financial crisis, and the economic state was atypical. Regardless, this information should be taken lightly and data from a more stable time frame should be collected. This finding opposes the obvious incentive of a higher risk-free rate of return. This counterintuitive result coincides with the findings from the original research paper, which explained that decreases in interest rates during and after a downturn are because the government is trying to encourage spending to spur the economy (Moro et al., 2014). However, this has the reverse effect of causing people to want to save even more.

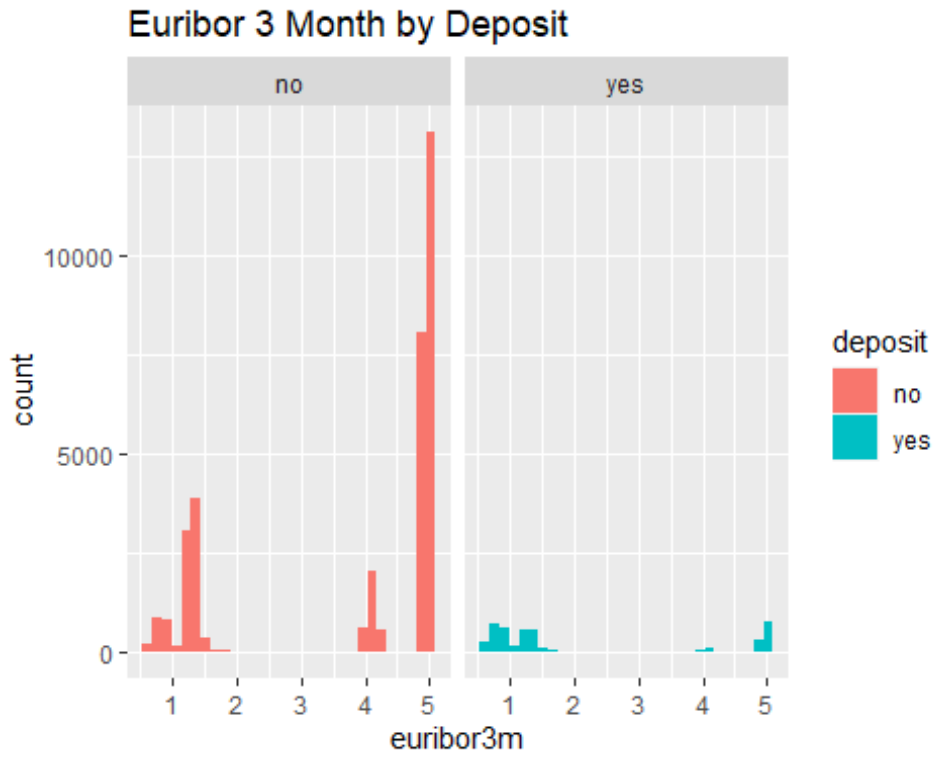

Figure 6. Counterintuitive Findings in the Data

One other way of interpreting this could be that contacts could be more willing to subscribe at a lower rate since it is during a time when the regular savings account rates are correspondingly high (not as high as a term deposit but higher than the inflation rate). The bank could influence more contacts to subscribe when the savings account rate drops too low (e.g., below the inflation rate), because the term deposit rate will be more enticing.

The other economic indicator variables were also affected by the circumstances, so not much stock can be placed upon them. For instance, deposits decrease when the employment variation rate increases, however this is likely confounded by the financial crisis.

By looking at the variable importance plots and prior analysis, we find a select few variables that determine the likelihood of a potential customer subscribing to a term deposit. Age is a primary indicator, especially those over the age of 60 and under the age of 30. Job type was also important, although it was correlated with age. Still, retirees and students had the highest subscription rates, so perhaps offering a student or retiree bonus to attract these customers would increase the number of deposits. Retirees will have a higher demand for term deposits in order to gain interest through risk-free payments, since they tend to have low risk tolerance. It is the same with students, who face uncertainty and cannot usually afford to have money locked up for long periods.

There was no significant difference among days of the week (although only weekdays were recorded), a higher proportion of deposits was recorded in the fall. Whether this was done in anticipation of holiday shopping or just a mere coincidence would require more data. May was the month with the most calls, but also the lowest rate of success. The campaign should be spread over the full year to capture any seasonal trends.

The customers who were more educated were more likely to subscribe to a term deposit, however this could be the result of affluence increasing with education level (richer customers need a place to store their money), so any speculation here would be dubious. Customers were slightly more likely to accept a deposit if they already have a housing loan, however this relationship was not statistically significant.

For the campaign itself, previously contacted people were more likely to accept a term deposit, even if they had rejected the deposit. However, people should not be harassed, as the proportion of accepted deposits drops precipitously after 4 calls; the callers should gently remind potential customers, not chase them.

The most pressing issue is the lack of information; there are thousands of unknowns in the data set. It is imperative that the bank collects data reliably so that a more effective model can be constructed. Whether this issue is from lack of survey engagement or from loss of bank data, the quality of data should be prioritized.

By capturing, leveraging, and analyzing massive volumes of data, bank and financial services companies can capitalize on new data-driven business opportunities. Bank and financial services companies will be able to generate insights that create better customer experiences, improve operational efficiency, and drive sales.