**Summer 2021 Project 2**

**Rashmi Patel, Taylor Bonar, Michael Burgess**

**Introduction:**

The goal of the following analysis seeks to identify and describe the key factors and interactions which affect client decision to subscribe for term deposit and predict if the customer will make a subscription to term deposit or not. The dataset is obtained from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) which related to the direct marketing campaign of a Portuguese banking institution.

**Data Description:**

The direct marketing campaign data includes of a multitude of variables including, age, job type, marital status, education, etc. and a categorical response variable. Several data points in multiple variables which was unknown were set to NA. This dataset is collected over time from May 2008 to November 2011. The dataset is delivered as delimiter separated values with 41,188 observations, 20 inputs and one response variable. Of the inputs, there were 10 continuous variables and 10 categorical string variables.

**Exploratory Data Analysis:**

**Chart, surface chart

Description automatically generated with medium confidence**When the data was examined with some basic plots, we found that some categorical variables have unknown group (1.5% of the data) in them which was contributing nothing to the analysis being done. So, moved forward replaced those unknown values with NA. After the replacement, we removed those NA values from the data. After getting the cleaned data, we started doing exploratory data analysis.

A picture containing graphical user interface

Description automatically generatedThe following results were found:

1. Out of 30488, 26629 clients (87.34%) did not choose subscription for term deposit. Only 3859 clients (12.66%) chose to subscribe for term deposit. Keeping this in mind, we’ll need to ensure either our sampling is weighted, or consider algorithms that can oversample or under sample when fitting our logistic regression model, otherwise we may introduce bias and lower our accuracy on predicting when a client successfully subscribes to a bank term deposit.
2. While examining the normality of numeric variables using histograms, we found that there are some potential normality issues. Age, Duration, Campaign, Previous campaigns are the variables that are highly right-skewed, and others seems to be distorted across different with some visual skewness.
3. While examining the categorical variables, we found the following results:
4. In job variable, the highest 3 categories were- admin with 25.3%, blue-collar with 22.47% and technician with 16.37%. Other job categories contribute only to 35.85% in which 0.8% were categorized as “unknown”.
5. In marital status, 60.52% of the clients were married with 28.09% single, 11.2% divorced and 0.19% were categorized as “unknown”.
6. Looking at the education of clients, the top 3 category have 29.54% of university graduates, 23.1% of high school education, and 14.68% of basic 9 year. Other education categories contribute up to 32.68% in which 4.2% were “unknown” and 0.04% were illiterate.
7. In housing variable which describing whether clients have housing loan or not, 52.38% of the clients have housing loan, 45.21% of the clients did not have housing loan and remaining 2.4% was categorized as “unknown”.
8. Looking at the loan variable which describe whether the client have personal loan or not, 82.43% of the clients have personal loan, 15.17% did not have personal loan and remaining 2.4% is categorized as “unknown”.
9. In preferred method of contact, 63.47% of the clients chose cellular communication type and 36.53% of the clients chose telephone to be preferred method of contact.
10. In last contact month of the year, 33.43% of the clients were last contacted in the month of May, 17.42% in the month of July and 15% in the month of August. No clients were contacted in the month of January and February.
11. Talking about the day of the week for last contact, 20.94% of the clients were last contacted on Thursday, 20.67% on Monday and the rest of them were last contacted in other 3 business days.
12. In poutcome variable which describes the outcome of previous marketing campaign, we found that only 3.33% of the marketing campaign were successful, 10.32% were failure and 86.34% were non-existent.
13. In default variable which describes whether a client has credit in default or not, we found that only 0.01% of the clients had credit in default. 79% of the client did not have credit in default and 20.87% were categorized as “unknown”.
14. A picture containing chart

    Description automatically generatedWhen we analyzed the multicollinearity of variables (correlation between numeric variables) using corrplot () function, we found the following results:
    1. emp.var.rate and euribor3m are 97% positively correlated.
    2. nr.employed and euribor3m are 94% positively correlated.
    3. nr.employed and emp.var.rate are 90% positively correlated.
    4. cons.price.idx and emp.var.rate are 77% positively correlated.
    5. cons.price.idx and euribor3m are 67% positively correlated.
    6. cons.price.idx and nr.employed are 49% positively correlated.
    7. pdays and previous are 59% negatively correlated.
    8. Nr.employed and previous are 49% negatively correlated.
    9. euribor3m and previous are 44% negatively correlated.
    10. emp.var.rate and previous were 40% negatively correlated.

**Linear Discriminant Analysis:**

In exploring possible models for predicting whether a client would subscribe to a term deposit or not, we applied a linear discriminant model to the data. The data were reduced to include only continuous predictors and were examined to see if the data met the assumptions of the linear discriminant model.

A picture containing chart

Description automatically generated

Although the data meets the assumptions of independence and no multicollinearity, it lacks any clear linearity or consistency in its variance. For these reasons, a linear discriminant model would not be recommended for making accurate predictions.

Although the model isn’t recommended, we applied the linear discriminant model for the sake of comparison. The data were split into an 80/20 randomly sampled train/test set. As expected, the accuracy is lacking at only 66.89% as shown in our confusion matrix or 68.4% as shown in the ROC curve. Because the data set is heavily unbalanced the linear discriminant model only achieved a sensitivity of 29.36% but had a specificity of 88.08%. A technique like under sampling may help to balance the sensitivity and specificity, but as stated previously the linear discriminant model will never provide great accuracy for the given data.

|  |  |
| --- | --- |
| A picture containing text, receipt  Description automatically generated | Chart, line chart, scatter chart  Description automatically generated |

**LASSO Logistic Regression Model**

When utilizing a dataset with many predictors, it can be preferable to use an automatic variable selection method to help reduce unnecessary variables. In doing so for logistic regression, we opted to use LASSO as our method. Before creating our LASSO regression model, we needed to determine our ideal value for controlling the shrinkage of coefficients, lambda.

To pick our ideal lambda, we need to perform k-fold cross-validation to reduce our cross-validation prediction error rate. As can be observed from our k-fold cross validation plot, the vertical dashed line indicates that the log of our optimal value of lambda is approximately -5 (exact lambda is) to minimize our prediction error. Additionally, we can infer what coefficients are non-zero from our coefficient plot of our cross-validation method for our minimal lambda, in which we can observe that we have two coefficients exactly equal to zero (marital status of single and education level of high school).

|  |  |
| --- | --- |
| Graphical user interface  Description automatically generated | Graphical user interface, application, table  Description automatically generated |

However, we do have an additional option in our regularization method to balance accuracy and simplicity. As we elected to use a cross-validation technique, we also can determine a lambda to give us our most simple model that lies within one standard error of our optimal value of lambda. The following coefficient plot is much simpler and only 32 variables have a coefficient exactly equal to zero:

Graphical user interface

Description automatically generated with low confidence

When choosing which lambda value to use when fitting our model, we generally want a balance between accuracy but also simplicity. That way, we can easily interpret the model if need be. As we seek to understand how coefficients interact within the model and understanding how our data comes into play, we shall resort to the simpler model and use a lambda within one standard error of our optimal lambda for our LASSO model instead of focusing on accuracy.

Once we fit our model, we determined that our test data was roughly 90% accurate in predicting whether a client would or would not subscribe to a bank term deposit. With this in mind, we can conclude our coefficients from our simpler model may be most relevant and significant for this direct marketing campaign.