Predicting Term Deposit Subscription

Milestone 5- Final Project Submission

Moshe Burnstein

DSC630

Professor Andrew Hua

May 22, 2023

Introduction

Banks use term deposits to raise cash and in turn lend out that money at a higher interest rate than they are giving for the term deposit (ThinkBank). Term deposits, also known as time deposits, are offered both by banks and by credit unions. The three defining characteristics of a term deposit according to bankrate.com are a guaranteed interest rate, a fixed maturity date that interest accrues until, and a penalty for early withdrawal (https://www.bankrate.com/banking/term-deposit-vs-call-deposit/#term). Term deposits are referred to as certificates of deposit (CD) when issued by banks and as share certificates when issued by credit unions. It is commonly assumed that people put money in CDs when young (less than 20) and when old (older than 60). This is because parents instruct their children to put their money into CDs to protect the principal until the child matures and can invest wisely. The youngster often does not rely on these monies for his day-to-day living. When the youngster matures, he can no longer afford to keep his money locked up for a set period of time. He also sees much greater potential profits investing elsewhere. After he has accrued his money during his earning years and matriculates to "older" age he needs to maintain this money safely. He can once again afford to lock his money away for a period in a term deposit.

Does marketing change this paradigm? Would people during their earning years be uninterested in such a product when marketed to appropriately? Would marketing "turn off" people from investing in these term deposits? Would younger people or older people respond positively to a marketing campaign? How can one determine who might respond positively to a marketing campaign, while excluding those who will subscribe for the term deposit irrespective of the campaign? Siegel in Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or

Die (Siegel, 2013), describes how the Obama Campaign leveraged predictive analytics to learn who to reach out to. They needed to find swing voters who would respond positively to marketing. They needed to exclude voters who would anyway vote for Obama as well as voters who would have voted for him but will not now because of the outreach. This is a particularly vexing problem to quantify because one cannot produce samples of the same voter voting once without being marketed to and then again after being marketed to.

These questions are of paramount interest to banks, and to the world in general. Banks can efficiently market to prospective customers who would be swayed by the marketing to subscribe to term deposits. The Obama Campaign was able to predict on a granular basis who was appropriate to reach out to, and who not (Siegel, 2013). The world at large may embrace the concept that predictive modeling to choose appropriate candidates for outreach is a boon.

The goal of this project is to inform the bank of the most profitable ways to target prospective subscribers. The bank subject matter experts will have to collaborate with me both when building the model and then when deploying it. They will inform me of the costs of contacting potential customers vs the profitability of gaining a subscriber. This will dictate the balance between aiming for higher precision scores and higher recall scores. The project will potentially tell us how best to solicit customers through a phone campaign. I anticipate that profits gained by each successive subscriber will far outgain losses from adding a few false positives (potential subscribers who will not subscribe).

The data which I plan to use for this project is a .csv file with “;” separators. The default separator for .csv files is commas, but one can use 'read.csv' in R to load a file with ';' separators. There were no issues loading and visualizing the data. There are no null values in the dataset. There are “999” values in the pdays column. This column describes the number of days since the

client was last contacted in a previous campaign. The value “999” is used to represent a potential client who was not previously contacted. I will address this if I need to by coding “999” as “0”. Alternatively, I will remove all such rows if I need focus on days since contact because I would be studying only subjects who were previously contacted. The target variable is imbalanced... only 13% of the subjects subscribed. I will have to weigh the pros and cons of creating more samples of subscriptions as opposed to leaving the target imbalanced. If I leave it imbalanced, I will have to pay special attention to model scores by using confusion matrixes and other scoring metrics which score each class of the target. I plan to create two subset dataframes(df) from the original dataset, one containing all subscribers and the second containing all nonsubscribers. This will enable me to both visualize and wrangle each class separately. No single variable is highly correlated with the target, but there are potential groupings which promise to be informative.

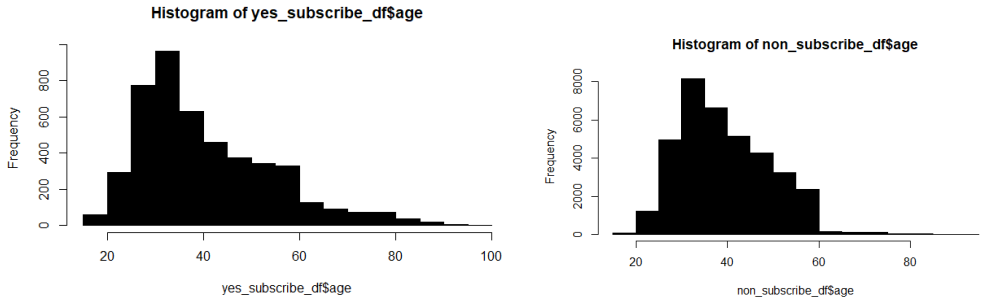
The highest correlation exists between duration of phone contact and subscription. There is a strong caveat given in the attribute information in the UCI Repository. It cautions against using duration in any realistic predictive model. It states that if duration=0, then there will most definitely be a non-subscription. Duration is not known before the contact call is made. After the contact is made, the target is perforce known. The duration variable is confounding because it is reasonable to consider a longer duration of contact to be a driver of subscriptions, and therefore it would be a good predictor. This requires much investigation. Is the positive correlation between duration and subscription causal? Or does a potential subscriber who plans to subscribe need to have a longer contact duration to learn the terms of the term deposit, and a third variable causes both the longer duration and the subscription.

Would directing bank employees to engage potential clients in longer conversations drive subscriptions? While one may assume that merely socializing with a prospective client does not

drive subscriptions, maybe engagement itself has an effect? Even if social engagement does not cause subscriptions, maybe engagement in pitching the term deposit does?

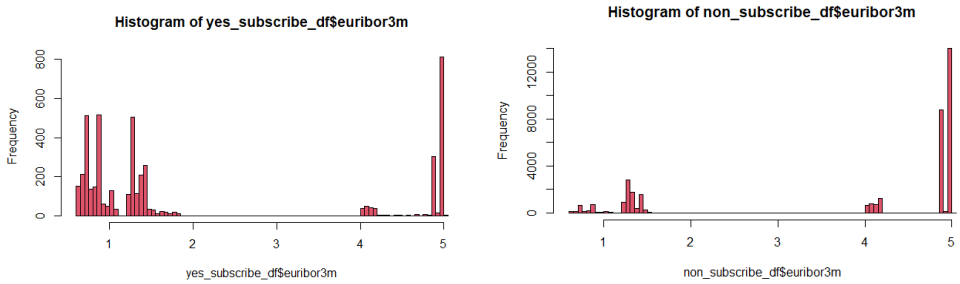
Methods/Results

The target variable, subscription (yes/no), is imbalanced. There are 36,548 who declined to subscribe, and 4,640 who subscribed. Only 12.7% subscribed. I created a second dataframe fro the original dataframe containing only subscribers, and a third dataframe containing only non-subscribers. I did this to investigate and compare these different populations. Histograms of the variables age, eurobir3m, and duration show distinct differences between subscribers and non-subscribers.



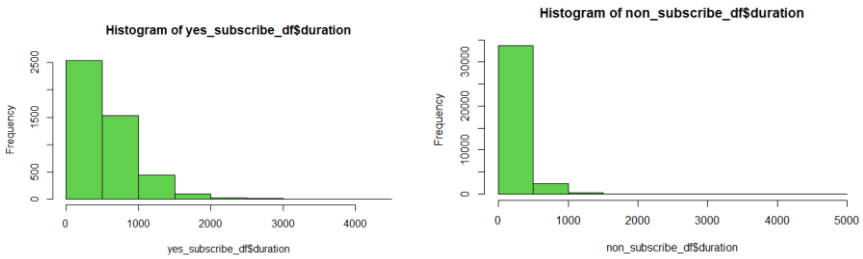
One must take note that while the x-axis is on the same scale throughout, aside from no 100 year-old outliers in the non-subscribe df, the scale of the y-axis frequency is up to 1,000, and not 8,000. This is because there are only a fraction of the number of observations in the subscribe df as there are in the entire df, or in the non-subscribe df. The shape of the distribution differs from subscribers to non-subscribers. The vast majority of non-subscribers are between 20 and 60 years of age, while subscribers are well-represented both in the under 20 age bin and in the over 60 bins.

Histograms of the eurobir3m rate in the two subset dataframes (yes- and non-subscribers) show different distributions.



Note that the scales of the plots are different ranges because of the differing magnitudes of counts. It is clear, however, that the yes subscibers are right-skewed as opposed to the non subscribers. The lower euribor3m rates are positively correlated with subscriptions. This confirms what I anticipated: people invest in fixed returns to protect themselves from lowering interest rates.

Histograms of call duration in seconds display different distributions between the two subset dataframes.

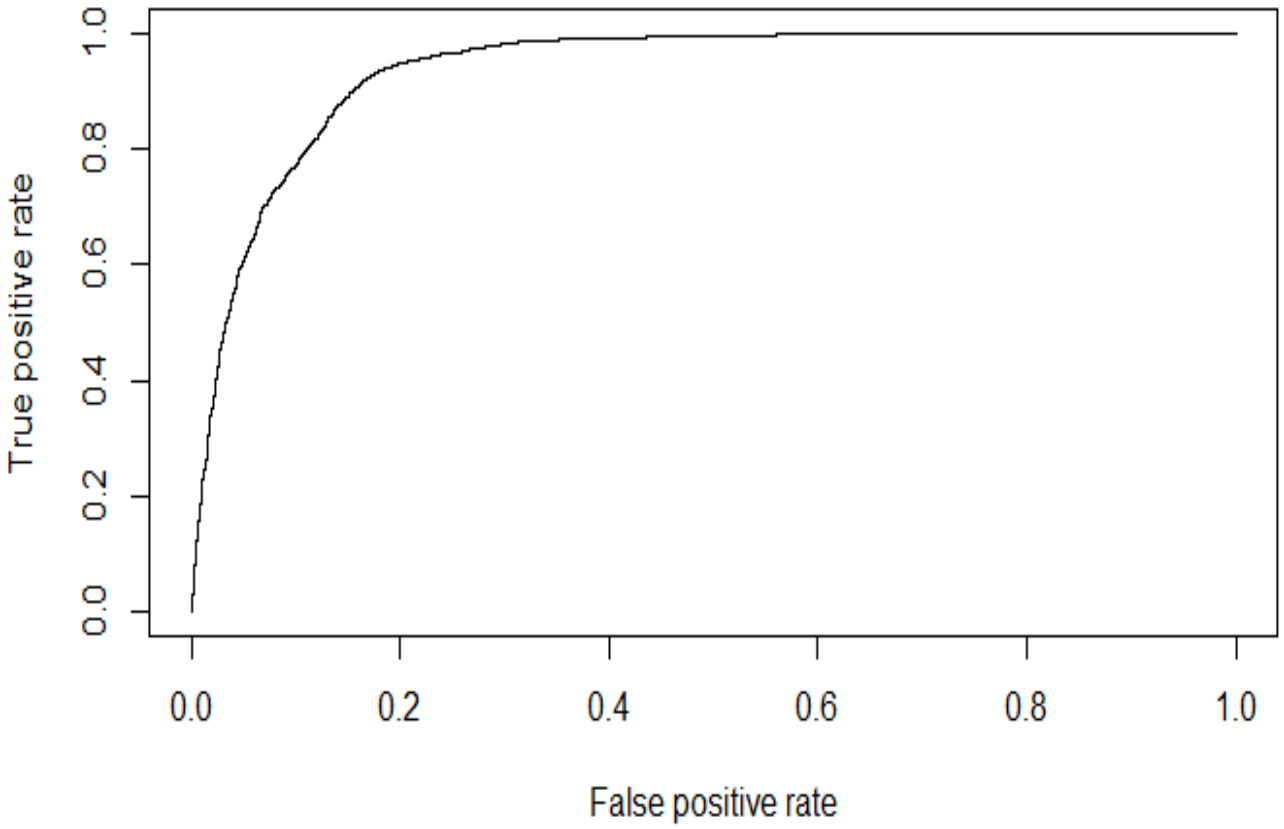


There are clearly more phone calls of longer duration in yes-subscribe. We must determine how to leverage this positive correlation.

One must consider how to model this dataset. All the twenty independent variables are poorly, if at all, correlated with the target feature: subscription. Also, the target classes are imbalanced. There are 36,548 observations of people who declined to subscribe, and 4,640 observations of

subscribers. Subscribers make up merely 12.7% of the target variable. This problem is accentuated by the fact that presumably stakeholders wish to make better predictions on subscribers than non-subscribers. Specificity is paramount because one does not wish to lose any potential subscribers, even if this means that one must entertain more false positives. The model should weed out definite non-subscribers to save that wasted expenditure on campaigning to them.

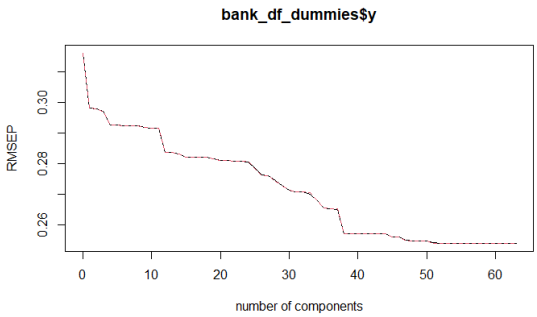
I first created a logistic regression model using all features of the dataset. This type of model does not perform well on non-linear data, so I did not expect a robust model. It will serve as my baseline model. I created dummy variables for my categorical non-numeric features. In r this leaves the original columns in the dataframe, so I had to remove the original columns which show NA when modeling. I then scaled the dataframe to ensure that different magnitudes will not bias the models. Scaling is not necessary for tree-based models because these models are not affected by magnitude of variables. While the accuracy of this model is 83%, the specificity is an abysmal 3%, an utter failure.

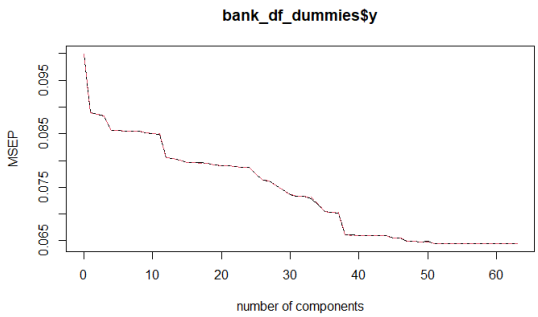


I then created a RandomForest model to visualize important features with its varImplot function, and to produce a predictive model. Random forest is a meta estimator which leverages multiple decision trees by averaging them to improve accuracy and prevent overfitting. It is not affected by magnitudes among variables and therefore does not need scaling.

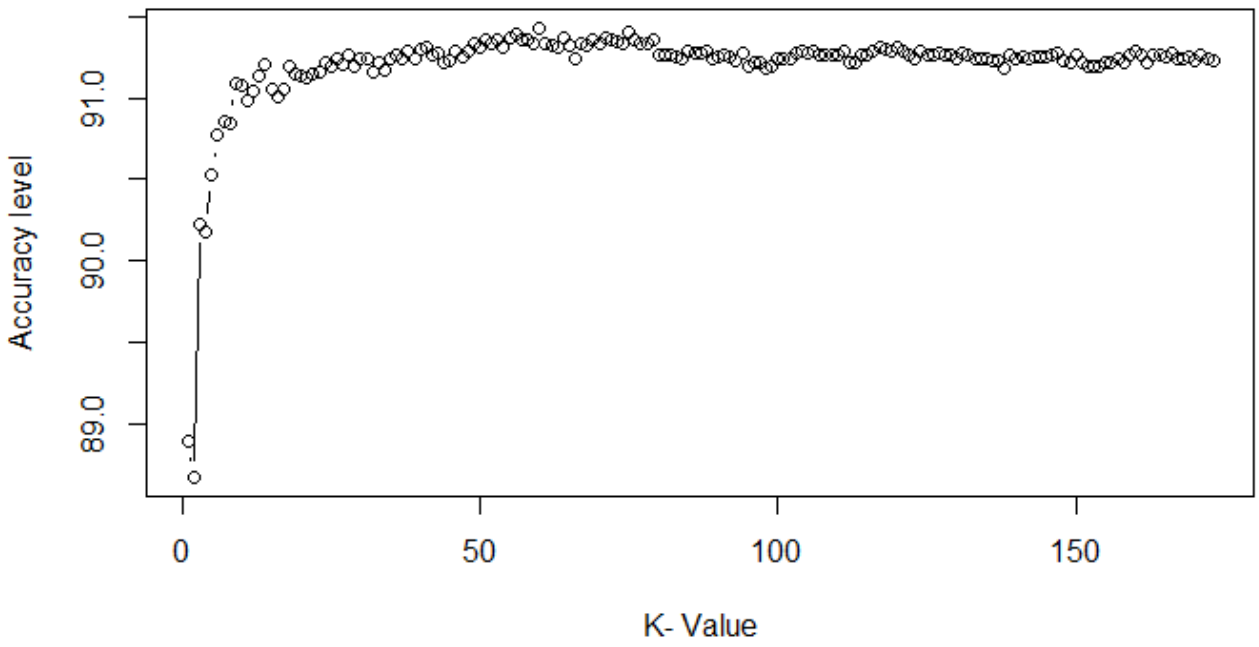
The graph presents duration as the greatest driver of subscriptions, followed distantly by euribor3m and age. The higher the mean decrease gini, the greater that variable affects the model. This calculates the splits in trees based upon the Gini Impurity Index ((n.d.). https://campus.datacamp.com/courses/introduction-to-machine-learning-in-r/how-much-will-i-earn?ex=6). While the sensitivity of this model was 97%, the specificity was only 43%.

I then created a Principal Components Regression (PCR) model to visualize the most important principal components. This type of model combines multiple linear regression with principal component analysis to reduce features. This in turn reduces degrees of freedom.





One must use upwards of 37 components to create a viable model. This is evident both in Root Mean Square Error of Prediction and the Mean Square Error of Prediction. These are the gold standard metrics for error predictions. The data is so poorly correlated with subscription that one cannot capture the data to effectively model it without using most features. I created a Support Vector Machine (SVM) with hyperparameter tuning. While it produced a sensitivity of 98%, it only produced a specificity of 31%.
I then investigated kNN models. These models model proximity from a point to a group k of points to classify a point by calculating the votes of k nearest neighbors on class. I first created models for k=168 and k=169 because these values sandwich the square root value of the number of observations. I then iterated from k=1 to k=173 to find the best performing model.



I visualized accuracies of all these ks on a graph and consulted the output list to find k=60 producing the best accuracy. More important to us than the 91.4% accuracy on the test set is the 50.0% specificity. It correctly predicted subscribers 714 times and incorrectly predicted subscribers only 726 times. This has proven to be the most robust model yet.

Ensemble models have been all the rage over the past several years because they have produced the most robust models. Ensemble learners address the main deficiencies of simple, weak learners- noise, bias, and variance (Aporras, 2020). Bagging and boosting improve the stability of models by combining multiple weak learners. These models generate more training data. Bagging allows any element in the data to be sampled randomly, while boosting weights the observations. Boosting often offers the most robust model. I built a bagged mode which produced a sensitivity of 96%, but more importantly produced a specificity of 56%. The XGBoost model which I built proved to be the most robust. While producing a lesser sensitivity

(92%) than the bagged model (96%), the boosted model produced a specificity of 68%. It is well worth the slight tradeoff in sensitivity for the excellent improvement in specificity.

Conclusions and Recommendations

Lower euribor3m rates are a driver of subscriptions. I therefore recommend an aggressive marketing campaign when euribor3m rates drop. This is sensible because people tend to invest in fixed rate returns when interest rates drop. Further investigation is required because the data suggests that the other index rates do not correlate with subscriptions. Consumer price index and consumer confidence index showed no such association with subscriptions. Furthermore, we must confirm our extrapolation from on Portuguese banking institution to American banking institutions.

Age most definitely plays a role in subscriptions. There is a much greater percentage of subscribers in the younger and older groups than in the main population. While there are only 804 observations in the group of younger/older, a whopping number of 529 people from this group subscribed. The 804 sample is too small to really study, so we should look for more data on younger and older people's subscriptions. Alternatively, it is most likely a worthwhile investment to conduct our own study on younger and older groups.

Longer contact duration is positively correlated with subscriptions. It would be appropriate to survey subscribers and ask them if they had previously decided to subscribe before they were engaged in a prolonged phone contact or did the engagement of longer duration influence them in deciding to subscribe. Phone contacts must be made cautiously to positively influence the on-the-fence clients, and not to turn off others.

Of the more explainable models, the kNN model is the best. However, the boosted model is the most robust. A specificity of 0.6796 on a subscriber class translates into immediate profits because a random sample would only produce 12.7% subscribers, leading to unnecessary

expenditures on campaign marketing to non-subscribers. This model is ready for immediate deployment.

Model Scores Table

Model Name	Accuracy	Kappa	Sensitivity	Specificity	Balanced Accuracy
GLM using all features	0.8335	-0.0514	0.92121	0.03304	0.47713
RandomForest	0.9115	0.4814	0.9737	0.4312	0.7024
kNN k=60	0.9138	0.516	0.9665	0.4958	0.7312
SVM with hyperparameter tuning	0.9051	0.3746	0.9802	0.3066	0.6434
Ensemble Bagging	0.9145	0.5441	0.9583	0.5613	0.7598
XGBoost	0.9084	0.3932	0.921	0.6796	0.8003

References

Aporras. (2020, November 3). What is the difference between bagging and boosting? ⋆ Quantdare. Quantdare. https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

Bennett, R. (n.d.). Term deposit vs. call deposit. Bankrate. Retrieved March 21, 2023, from https://www.bankrate.com/banking/term-deposit-vs-call-deposit/#term

Moro, S., Cortez, P., & Rita, P. (2014, June). A Data-Driven Approach to Predict the Success of Bank Telemarketing. UCI Machine Learning Repository: Bank Marketing Data Set. Retrieved March 21, 2023, from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Improving your model. R. (n.d.). https://campus.datacamp.com/courses/introduction-to-machine-learning-in-r/how-much-will-i-earn?ex=6

Siegel, E. (2013). Predictive analytics: The power to predict who will click, buy, lie or die. John Wiley and Sons. Pages 213-217

What does your bank do with your money? ThinkBank. (n.d.). Retrieved April 19, 2023, from https://www.tmbank.com.au/thinkbank/industry-news/what-does-your-bank-do-with-yourmoney

What is the K-nearest neighbors algorithm?. IBM. (n.d.-a). https://www.ibm.com/topics/knn