Joel Fry
Connor Allison
Nick Mata
RoyalPriesthood Olola

Case Study #1 Report

Data Mining Bank Information for Telemarketing

Executive Summary

We constructed two models, a Linear Discriminate Analysis (LDA) and a Logistic Regression Model (GLM) model. Both provided good results. We found the LDA model provided the highest prediction accuracy and highest positive predictive value when forecasting for the likelihood of a customer signing up for a term deposit. The GLM provided more balanced sensitivity and specificity with a lower accuracy and had significantly higher false positives. Both models had similar complexity, therefore, it did not play a role in choosing the model.

The Problem

A bank wants to increase its long-term deposits with a telemarketing campaign targeting existing current bank customers. The goal is to contact current bank customers who are most receptive to making term deposits while avoiding customers who are unlikely to make term deposits. Using a data set comprised of 150 variables, researchers found 22 key features that were most relevant to predicting success. Compiling almost 53,000 records between 2008 and 2013 they needed a way to sift through the total customer base to find receptive customers (Moro et al, 2014). Using these techniques, they wanted to know if they could better predict the likelihood of a customer making a term deposit by optimizing customer targeting. This would have a two-fold outcome by both reducing operational costs and enhancing customer experience. The root of the problem was they did not have a way to utilize data-driven insights to improve their ability to make good decisions to best help their customer base.

Review of Related Literature:

Banking institutions have used, and can use, all available techniques to include Principal Component Analysis with clustering algorithms, or customer segmentation based on shared characteristics, to improve customer product targeting (Moro 2015). We found for this case study, already having customer data, that we could exclude methods used to find new customers in the community and focus on existing customers with the bank. Machine Learning Algorithms and predictive modeling both provide usable results based on the existing data, however, for this case study we chose to focus on using the CRISP-DM (Cross-Industry Standard

Process for Data Mining) because of the distinct phases allowing us to step through the study and document finds (Shah 2024). Business Understanding, the bank identified a product they needed to sell and asked the key question, which of our customers will buy it. Data Understanding, bank members collected data about current customers, and organized it in a way to use in four models. Data Preparation, as addressed in the previous step, they cleaned and transformed any data to make sure it was usable. Modeling, in the process of understanding their using data, they identified four models to use, Neural Networks, Decision Trees, Logistic Regression, and Support Vector Machine (SVM) (Moro et al., 2014). Evaluation, after running the four models, they determined the Neural Network was the best model to use. Deployment, they took the outcome and started making contact with the bank's customers.

Existing Methodologies use in the area:

For our review of the data, we used a logistic regression model and LDA to identify who would likely want to make a term deposit. In both models, we split the data at 70/30 for training and testing. Additionally, we ran several versions of the models to try to find an optimal model that would be most applicable to the business problem.

Data

A pre populated data set from the current banking customers was provided which included information related to all aspects of their lives and banking history. Categories include age, marital status, education, which types of products they are currently using and whether they were contacted for this campaign. A total of 150 variables were identified, and narrowed to 22 key points that best predicted who would purchase a term deposit (Moro et al., 2014).

From our subset of the data, our logistic regression model identified nine variables that were significant in identifying the customers most likely to make the term deposit. A review of the subset of data found there were unknown values, but no missing data. We converted pdays into three dummy variables, if they had been contacted in the last week, two weeks, or three weeks, with the final outcome of never being contacted; these covered all records in the data. We converted categorical data to factors, then ran the logistic regression model. The outcome revealed age, contact, month, campaign, poutcome, pdays, emp.var.rate, cons.price.idx, cons.conf.idx were all significant. It is notable that the nine we determined are also in the 22 used in the study and we did not identify a variable that was not previously determined significant. Additionally, the nine we identified covered the various areas used in the study, covering specifics about the customer, when and how they were contacted, along with economic conditions.

Other cleaning included removing duration as a variable. It skewed the data because if it was zero then the record was treated as a no, however, if they picked up the phone and had any contact it could be either a yes or no. This takes into account acting on the predictions of the

model which does not help in making future predictions. It was useful in the study to find the accuracy of their model but does not play a role in ours since we cannot act on the results.

Findings

We found, of the 150 variables, there was a set of data that could mostly predict positive outcomes, while avoiding negative outcomes. By increasing the positive predictive value, it would increase the bank efficiency in targeting its customers. Most notably, it is possible to use a combination of user data with socio-economic data to positively predict results. To identify the variables, we used a stepwise fit and checked for multicollinearity using VIF. We did not have to drop any variables based on these checks as they did not have multicollinearity.

```
glm(formula = y_train ~ contact + month + campaign + emp.var.rate +
    cons.price.idx + cons.conf.idx + daysdummy1 + daysdummy2,
    family = binomial, data = df_train)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -110.11266   16.58613  -6.639 3.16e-11 ***
contacttelephone   -0.70453    0.24874  -2.832  0.00462 **
monthaug            0.10710    0.37154   0.288  0.77314
monthdec            0.48773    0.61471   0.793  0.42753
monthjul            0.08715    0.34269   0.254  0.79927
monthjun            0.21127    0.31765   0.665  0.50598
monthmar            1.77601    0.45090   3.939 8.19e-05 ***
monthmay           -0.38729    0.27657  -1.400  0.16141
monthnov           -0.47886    0.33475  -1.431  0.15257
monthoct           -0.33539    0.45595  -0.736  0.46199
monthsep           -0.10843    0.45988  -0.236  0.81361
campaign           -0.08585    0.04127  -2.080  0.03751 *
emp.var.rate       -0.69292    0.07006  -9.890  < 2e-16 ***
cons.price.idx      1.18070    0.17994   6.562 5.32e-11 ***
cons.conf.idx       0.05644    0.01813   3.112  0.00186 **
daysdummy1          1.58008    0.28081   5.627 1.84e-08 ***
daysdummy2          1.22216    0.51040   2.394  0.01664 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2014.1  on 2882  degrees of freedom
Residual deviance: 1580.3  on 2866  degrees of freedom
AIC: 1614.3
```

Our accuracy rate on the LDA was .897, sensitivity was .354, the specificity was .961 and a positive predictive value was .517. This was better than the logistic regression model accuracy was .789, sensitivity was .677, the specificity was .802 and a positive predictive value was .287. Ultimately, we used the logistic regression model for a stepwise fit to identify which variables were significant and used them in the LDA model. With a positive predictive value that was half

of the LDA model, we did not think the logistic regression model gave valuable output other than it was useful to identify significant variables. Our determination was based on the idea it was better to get more yes's per phone call than making more phone calls to get all the yes's.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 829  39
         1 277  91

               Accuracy : 0.7443
                 95% CI : (0.719, 0.7685)
    No Information Rate : 0.8948
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2487

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.70000
            Specificity : 0.74955
         Pos Pred Value : 0.24728
         Neg Pred Value : 0.95507
             Prevalence : 0.10518
         Detection Rate : 0.07362
   Detection Prevalence : 0.29773
      Balanced Accuracy : 0.72477

       'Positive' Class : 1
```

GLM CONFUSION MATRIX

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1063   84
         1   43   46

               Accuracy : 0.8972
                 95% CI : (0.879, 0.9136)
    No Information Rate : 0.8948
    P-Value [Acc > NIR] : 0.4128621

                  Kappa : 0.3659

 Mcnemar's Test P-Value : 0.0003861

            Sensitivity : 0.35385
            Specificity : 0.96112
         Pos Pred Value : 0.51685
         Neg Pred Value : 0.92677
             Prevalence : 0.10518
         Detection Rate : 0.03722
   Detection Prevalence : 0.07201
      Balanced Accuracy : 0.65748

       'Positive' Class : 1
```

LDA CONFUSION MATRIX

We started by cleaning the data, taking out records with unknown loans. We recognized we needed to remove duration as it would skew results and was based on the outcome of the original model. The goal was to predict who would say yes, and the duration measured the original study checking their results which would not be useful in ours. A duration with a time means someone from the bank contacted a customer and the customer responded in some way. The duration with no time means they did not pick up, which was counted towards the nos. We took what was left into a stepwise model to identify which variables were significant. Continuing with a confusion matrix, we wanted to understand how well the model worked with the results of the stepwise selection, and checked multicollinearity several times throughout when applicable. After having these results, we then checked for the optimum cut off point to factor in the unbalanced nature of the data. This gave us the outcome for the logistic regression model. Finding the optimal cut off leads the model to give less than optimal results. The adjusted data set led to better balanced sensitivity and specificity, however, that was valuable for a real-world outcome. Using the same variables and cleaned data, we used them for an LDA model and found better results.

The outcome of the models was the LDA had a higher accuracy and higher positive predictive value, which was the desired outcome component.

Conclusion and recommendation

While trying to understand the business problem, and the data, we had to take more time to understand what we really needed out of the information. We determined positive predictive value was most important over accuracy as we wanted to maximize positive outcomes per phone call rather than finding every single positive outcome. We found with the pdays variable, it provided insight on when someone from the bank should contact a customer for them to be most receptive to saying yes to the term deposit. Changing it to categorical showed it was better than keeping it continuous or changing it to a factor because then we could pinpoint when was the best time to contact someone.

In the end, there were several more things we could do to better fit the model. We noticed we could test for interactions between the variables to see if that played a role. It would benefit our study to find out why March was more productive than any other month. Lastly, we identified using several other models to increase accuracy while minimizing false results.

Shah, U. S. (n.d.). Bank Marketing Predictive Analysis: A Deep Dive Using CRISP-DM. Medium.

Moro, S., Cortez, P., & Rita, P. (2014). *A data-driven approach to predict the success of bank telemarketing.*

Moro, S. (2015). *Feature Selection Strategies for Improving Data-Driven Decision Support in Bank Telemarketing.*

Jin, W., & He, Y. (2019). Three data mining models to predict bank telemarketing.

Appendix:

Reference PDF Titled Case Study 1