Name:- Swapna Vaidya ***Class:-*** MSDS - 422

Evaluating Classification Models

(1) *A summary and problem definition for management*

A Portuguese bank conducted seventeen telephone marketing campaigns between May 2008 and November 2010. The bank recorded client contact information for each telephone call. The bank wants its clients to invest in term deposits. The bank is interested in the following management problems:-

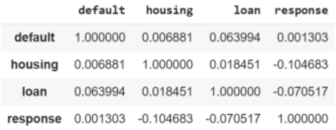
A. Which of the two modeling methods would you recommend and why?

B. And, given the results of your research, which group of banking clients appears to be the best target for direct marketing efforts (similar to those used with previous telephone campaigns)?

(2) *Discussion of the research design, measurement and statistical methods, traditional and machine learning methods employed*

Based on the exploratory data analysis that was done the survey data, below were some of the findings:-

- In all there are 4521 instances with 17 features in the data
- Of the 17 features, 11 features are non-numerical and 6 features that are numeric
- There are no missing values in any of the columns in the data set.
- Clients with a marital status of divorced or singles have the highest response rate. Similarly, clients who have not defaulted on the credit or have no housing loan or any personal loan also have the highest response rate. Next, clients who are retired or are students or clients who have tertiary education also have the highest response rate.
- As far as the numerical variables are concerned there wasn't any pattern to point out.
- Based on the correlation matrix of default, loan and housing to response variable, there seems to be negative correlation between housing and loan and a slight positive correlation with default.



	default	housing	loan	response
default	1.000000	0.006881	0.063994	0.001303
housing	0.006881	1.000000	0.018451	-0.104683
loan	0.063994	0.018451	1.000000	-0.070517
response	0.001303	-0.104683	-0.070517	1.000000

(3) *Overview of programming work*

This assignment was done in Google Colab, using 4 main libraries (Pandas, numpy, matpotlib, Seaborn and Skylearn).

- ***Ingestion:-*** Using pandas the data was first loaded from Google OnDrive. Then the data was subset into 3 sets:- 1) for all the explanatory variables to answer management question of which group of clients should they target. 2) for all the numerical variables to see if there were any insights. 3) for modeling purposes.
- ***Exploration:-*** of the data was done using methods like describe(), info() and corr(), crosstab().
- As part of exploration, ***visualization*** was done using histograms and scatterplots.
- ***Data Cleaning and Transformation:-*** The values of the subset created for the models with the 3 features default, loan and housing were transformed into binary values of 0 and 1. The data was then split into train and test data using the "train_test_split" method of the Skylearn package. The label "response" was separated from the feature values for the train and test data
- ***Modeling:-*** The first model selected was a logistic model with solver "liblinear". The cross validation method of cross_val_predict() with 3 folds was used with method "decision_function" to get the scores for all possible thresholds. These scores were used to calculate the ROC AUC of 59.83% and draw the ROC curve. The approximate model accuracy was 88.30%
  The second model used was a Bernoulli Naïve Bayesian model since the 3 features had binary values of 0 and 1. The cross validation method of cross_val_predict() with 3 folds was used with method "predict_proba" to get probabilities. The positive probabilities were then used to calculate the ROC AUC metric of 59.66% and plot the ROC curve. The approximate model accuracy was 63%
  Evaluation on the Test Data:- Next both the model were evaluated on the test data using the same process above to ensure there is no overfitting or underfitting of the model. The logistic model gave a ROC AUC of 56.72% while the Bernoulli Naïve Bayesian gave a ROC AUC of 56.36%

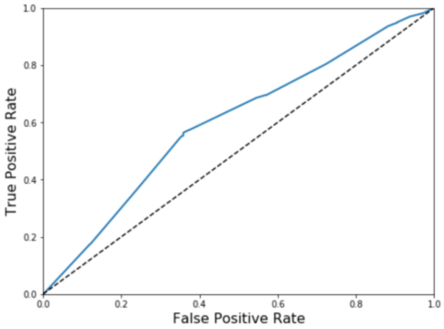
(4) *Review of results with recommendations for management*

After the exploratory data analysis and modeling exercise has been completed, following are the recommendations to the questions raised by the Portuguese Bank management:-

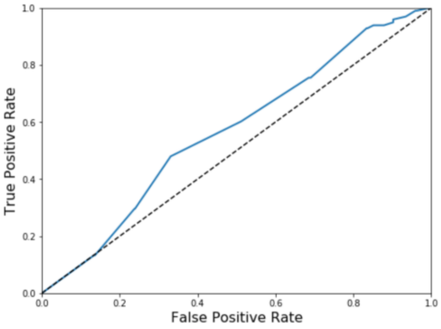
A. Which of the two modeling methods would you recommend and why?

Of the 2 modeling methods, I would recommend the logistic model as the ROC AUC is slightly better than the Bernoulli Naïve Bayesian model in both the training and the test data. ROC AUC for the “Logistic model” is 59.83% with a model accuracy of 88.30%. Having said that both the models give a pretty low AUC metric since the positives are very small in number. It would be better to evaluate the model with some more explanatory features than just “default”, “loan” and “housing” to make a strong recommendation to the management.

Training Data



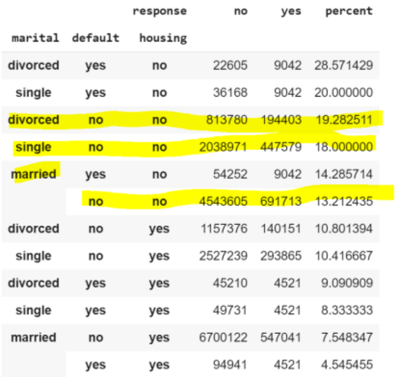
Test Data



B. Given the results of your research, which group of banking clients appears to be the best target for direct marketing efforts (similar to those used with previous telephone campaigns)?

Based on the analysis, various groups of clients should be targeted:-

- Clients with a marital status of “divorced” or “Single” or “Married” who have not defaulted or have no housing or personal loan have a high percentage of positive responses



		response	no	yes	percent
marital	default	housing			
divorced	yes	no	22605	9042	28.571429
single	yes	no	36168	9042	20.000000
divorced	no	no	813780	194403	19.282511
single	no	no	2038971	447579	18.000000
married	yes	no	54252	9042	14.285714
	no	no	4543605	691713	13.212435
divorced	no	yes	1157376	140151	10.801394
single	no	yes	2527239	293865	10.416667
divorced	yes	yes	45210	4521	9.090909
single	yes	yes	49731	4521	8.333333
married	no	yes	6700122	547041	7.548347
	yes	yes	94941	4521	4.545455

- Clients who are retired or students
- Clients who have tertiary education
- Finally, the bank should do marketing campaign of the term deposit as based on the last campaign it resulted in generating lot of positives.