SUBSCRIPTION PREDICTION FOR TERM DEPOSIT MARKETING

Richard Han
rickyhan24@gmail.com

ABSTRACT

In this project, I perform exploratory data analysis on a dataset consisting of customer marketing data to identify the most significant attributes of the data. Secondly, I apply a few classification models and identify the best-performing models in regard to accuracy, precision, and recall. The European banking institution, for which we are doing the analysis, can identify which features of their customers are most significant for subscription in order to improve marketing efforts.

1 Problem Statement

The task is to build a model to predict customer subscription to a term deposit given call center data about the customer such as their age, balance, the last contact month, etc. I need to identify which attributes are significant for prediction and to use those attributes to build a customer profile.

2 Data

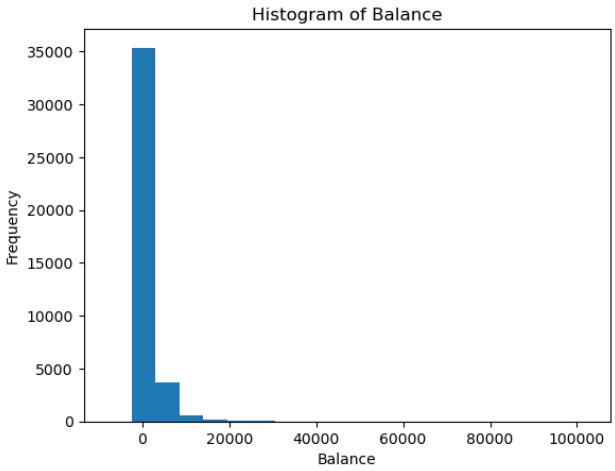
The data consists of 40000 rows of customers and 14 columns. The column 'y' indicates whether the customer subscribed or not to a term deposit. There are 5 numeric attributes 'age', 'balance', 'day', 'duration', and 'campaign'. 'day' is the last contact day of the month, 'duration' is last contact duration in seconds, and 'campaign' is number of contacts performed during the campaign and for the client. There are 4 binary attributes 'default', 'housing', 'loan', and 'y'. 'default' tells whether the customer has credit in default, 'housing' tells whether the customer has a housing loan, 'loan' tells whether the customer has a personal loan, and 'y' tells whether the customer subscribed or not to a term deposit. Finally, there are 5 categorical attributes 'job', 'marital', 'education', 'contact', and 'month'. 'job' is the type of job customer has, 'marital' gives marital status, 'education' gives education status, 'contact' gives the type of contact communication, and 'month' gives last contact month of the year.

There are 37104 customers who did not subscribe and only 2896 customers who did subscribe, making the dataset imbalanced.

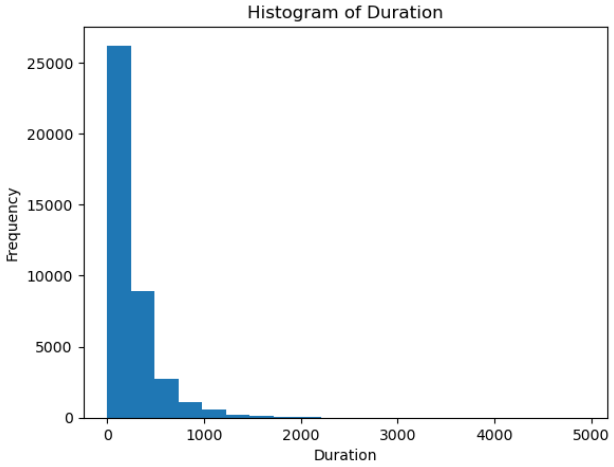
3 Exploratory Data Analysis

3.1 Results and Evaluation

To get a sense of what the customer looks like, we can plot histograms for each attribute. For example, here is the histogram for 'balance':

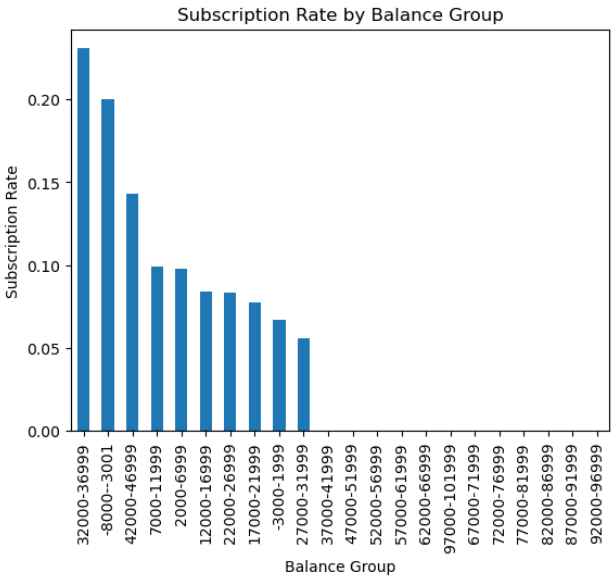


Most of the balances are between -2500 and 3000 euros. Here is the histogram for the duration of calls:



Many of the call durations were less than 250 seconds with the second largest bin of call durations between 250 and 500 seconds.

To get a sense of how each attribute might affect the subscription of a customer, we can look at bar plots for each attribute. For instance, here's the bar plot for balance:

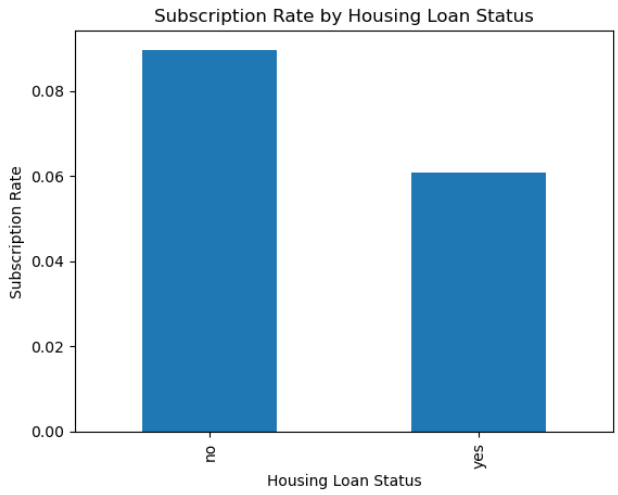


It appears that the best balance groups are 32000-37000, -8000-3000, and 42000-47000. However, their counts are not that high, as you can see in this table:

y balance	no	yes	subscription_rate
32000-36999	10	3	0.230769
-8000--3001	4	1	0.200000
42000-46999	6	1	0.142857
7000-11999	775	85	0.098837
2000-6999	5115	556	0.098043
12000-16999	228	21	0.084337
22000-26999	55	5	0.083333
17000-21999	95	8	0.077670
-3000-1999	30779	2215	0.067133
27000-31999	17	1	0.055556
37000-41999	7	0	0.000000
47000-51999	1	0	0.000000
52000-56999	3	0	0.000000
57000-61999	3	0	0.000000
62000-66999	3	0	0.000000
97000-101999	1	0	0.000000
67000-71999	0	0	NaN

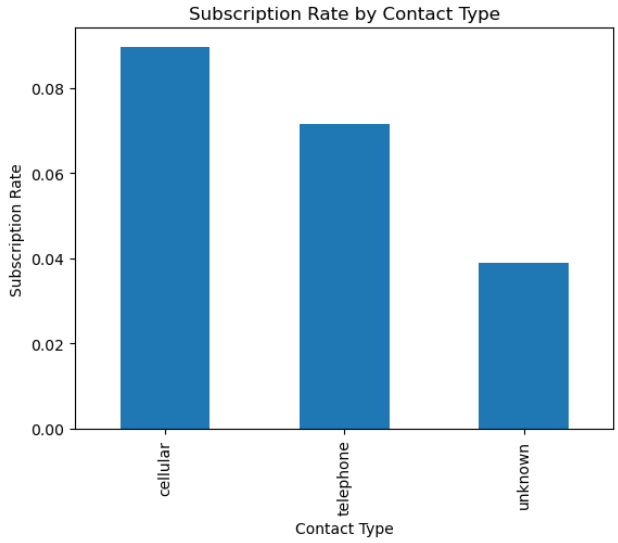
So, if we only consider balance groups with significant counts, then the best balance groups are 7000-12000, 2000-7000, and -3000-2000.

Here is the bar plot for ‘housing’:



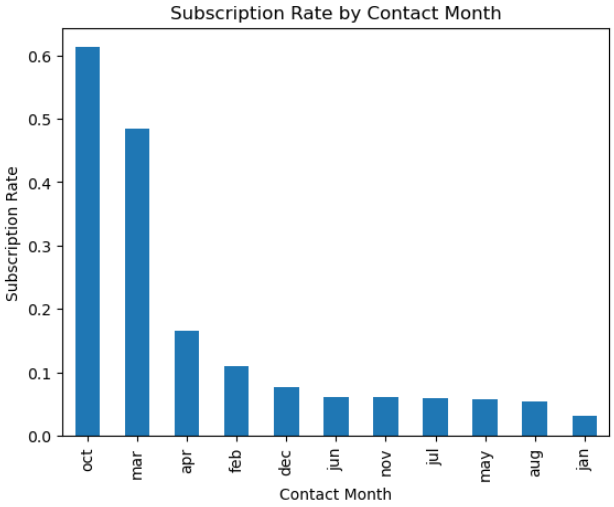
The customers without a housing loan tend to have a higher subscription rate.

Here is the bar plot for ‘contact’:



The best contact types are cellular, then telephone.

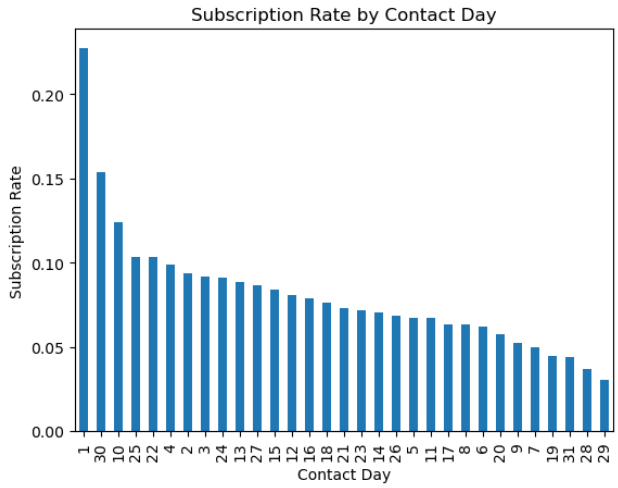
Here is the bar plot for ‘month’:



It appears that Oct., Mar., and Apr. are the best months. However, if we take a look at their counts, we can see that Oct's counts are low:

y	no	yes	subscription_rate
month			
oct	31	49	0.612500
mar	133	125	0.484496
apr	2267	451	0.165931
feb	2041	255	0.111063
dec	12	1	0.076923
jun	4440	294	0.062104
nov	3378	220	0.061145
jul	5996	384	0.060188
may	12741	791	0.058454
aug	4927	288	0.055225
jan	1138	38	0.032313

The best months are Mar, Apr, and Feb if we take into consideration the counts. Also, depending on whether we don't mind having lots of 'no' counts, the months of Jun, Nov, Jul, May, and Aug look promising as well. Here's the bar plot for 'day':



The best days to contact are on the 1st, the 30th, and the 10th. We should keep in mind that there are many other days that are promising.

Finally, for 'duration', if we take a look at only the cases where there are at least 100 subscriptions, the best duration ranges between 0 to 1500 seconds with higher conversion rates for 500-1250 seconds.

4 Classification

4.1 Data Preprocessing

In order to prepare the data for classification models, I converted the binary features 'default', 'housing', and 'loan' to 0's and 1's, and I one-hot encoded the categorical variables 'job', 'marital', 'education', and 'contact'. For the 'month' feature, I converted these to numbers between 1 and 12. Before applying the models, I also scaled the features 'age', 'balance', and 'duration'. To handle the imbalanced dataset, I will use oversampling and undersampling.

4.2 Methodology

First, I will apply logistic regression and random forest using all the features, paying attention to mean cv accuracy, test accuracy, precision for positive subscribers, and recall for positive subscribers; the precision and recall for negative subscribers remain high for all models. I will do this using both oversampling and undersampling. The reason for focusing on precision is that we don't want to count someone as a subscriber when they're not; this could be a waste of effort to try and convert them. The reason for focusing on recall is that we don't want to skip someone if they are a subscriber since those are the people we want to target; if customers who are potential subscribers are rare or difficult to find, we want to be able to find as many of them as possible.

Secondly, I will also use SHAP values to determine which features are most significant.

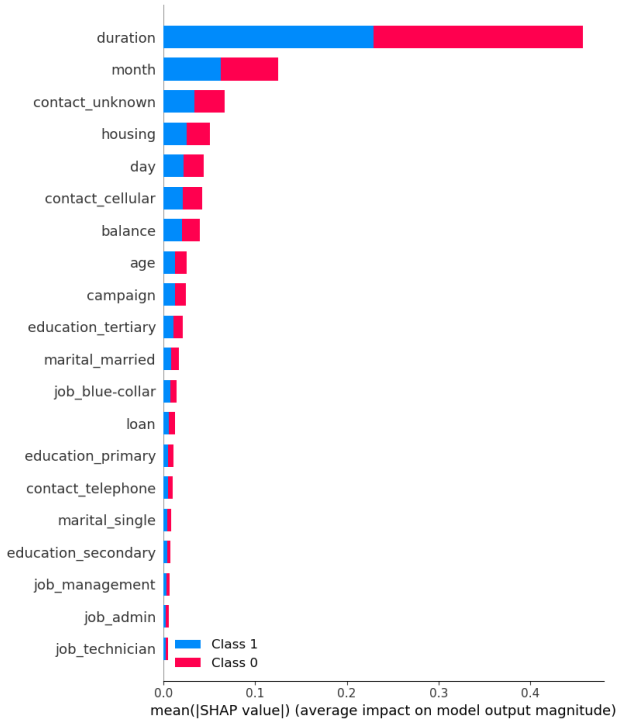
4.3 Results and Evaluation

Here is a summary table of the performance of all the models I considered. For each model, it gives the mean cv accuracy, test accuracy, and, for positive subscribers, precision, recall, and f1 score:

Model	Mean CV Acc	Test acc	Precision	Recall	f1 Score
Logistic regression w/oversampling	83.18%	85.28%	0.29	0.79	0.42
Logistic regression w/undersampling	82.72%	84.94%	0.28	0.79	0.42
Random Forest w/oversampling	98.18%	93.54%	0.53	0.44	0.48
Random Forest w/undersampling	88.07%	86.5%	0.32	0.88	0.47

The model with best recall is random forest with undersampling.

After training the random forest with undersampling, I used shap values to identify the most significant features:



As you can see, the most significant features are duration, month, contact, housing, day, and balance.

5 Conclusion

Exploratory data analysis and observing shap values determined that the most significant features are duration, month, contact, housing, day, and balance. The best balance groups were 7000-12000 euros, 2000-7000 euros, and -3000-2000 euros. The customers without a housing loan tended to have higher subscription rates, and the best contact types are cellular, then telephone. Thus, the customer profile the banking institution should target are customers without housing loans, who use cellular or telephone for contact, and who have balances 7000-12000 euros, 2000-7000 euros, or -3000-2000 euros.

The features duration, month, and day tell us how best to reach the target customers. The best months to contact them are Mar, Apr, and Feb; and, the best days to contact them are on the 1st, the 30th, and the 10th. If we don't mind contacting customers who reject in order to gain the ones that subscribe, the months of Jun, Nov, Jul, May, and Aug look promising as well. The calls should last no more than 1500 seconds (25 minutes) and preferably between 500-1250 seconds (8 minutes to 21 minutes).

Finally, random forest with undersampling had a mean cv accuracy of 88.07% and a high recall of 88%; if we mainly care about recall, then this would be a satisfactory model.