Project Proposal
Topic: Term Deposit Subscription Prediction

Chintan Thakkar - cdt303
Kushal Shah - ks5574

Q1) What is the problem?

The term deposits are of prime interest for banks, since they can invest those funds in higher gain investments to make profits. Our dataset represents the data of clients of a bank from their marketing campaign, and based on client's background, the bank would like to know whether the client would subscribe to a term deposit plan. The fact that a client investing in one plan at a bank is likely to invest in other attractive plans offered by the same bank, the banks could then persuade these clients to purchase the term deposit first and then offer them other promotional plans, and thus invest their marketing resources wisely.

Q2) How will you learn the background?

We learn about the background by considering the Banking Term Deposit Dataset available in the UCI Dataset repository and the explanation given for all of it's features. We can gain some insight into the background of the problem by performing a simple exploratory analysis on some of the features to see how they are related to the outcome. Especially, we would like to explore details about economy metrics like employment variation rate, consumer price index, consumer confidence index and see how they impact the person's decision to invest in term plans.

Reference link:
https://www.fool.com/investing/2016/12/31/3-reasons-rising-consumer-confidence-is-good-for-b.aspx

Q3) What kind of data will you use?

We would be using data that describes the person's personal/occupation/monetary background and current economic conditions to determine if the person is likely to purchase a term plan from a bank.

Bank Client Data: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

VARIABLE	DESCRIPTION	EXAMPLES
age	age of client	numeric
job	type of job	categorical
marital	marital status	categorical

education	level of education	categorical
default	has credit in default?	categorical
housing	has housing loan?	categorical
loan	has personal loan?	categorical

Data related to last contact in the current campaign:

VARIABLE	DESCRIPTION	EXAMPLES
contact	contact communication type	categorical
month	last contact month of year	categorical'
day_of_week	last contact day of the week	categorical
duration	last contact duration, in seconds	numeric

Other Attributes:

VARIABLE	DESCRIPTION	EXAMPLES
campaign	number of contacts performed during this campaign and for this client	numeric
pdays	number of days that passed by after the client was last contacted from a previous campaign	numeric
previous:	number of contacts performed before this campaign and for this client	numeric
poutcome	outcome of the previous marketing campaign	categorical

VARIABLE	DESCRIPTION	EXAMPLES
emp.var.rate	employment variation rate - quarterly indicator	numeric
cons.price.idx	consumer price index - monthly indicator	numeric
cons.conf.idx	consumer confidence index - monthly indicator	numeric
euribor3m	euribor 3 month rate - daily indicator	numeric
nr.employed	number of employees - quarterly indicator	numeric
y - **TARGET**	has the client subscribed a term deposit?	(binary: 'yes','no')

Q4) What kind of model will you build? And how would you preprocess/evaluate?

Here is our plan to get started with data preparation before we can model the data.

Correlation analysis:

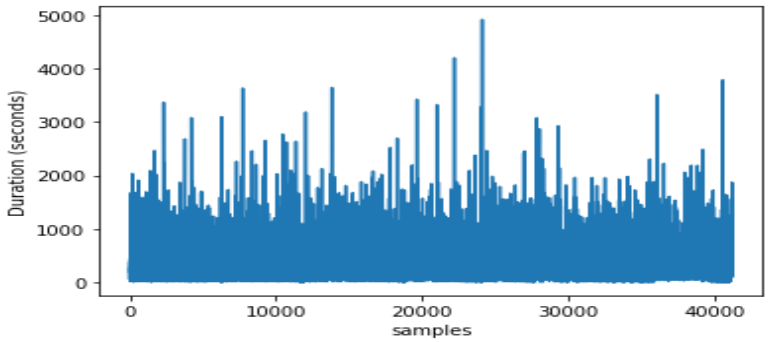
Highly correlated features are simply redundant for the model and we would prefer to keep only one feature out of all the high co-related ones, since this would help in feature set reduction for better performance.

One hot encoding:

Since our data contains many categorical variables like job status, marital status etc, we would convert these fields to numeric fields by using one hot encoding.

Feature scaling:

With preliminary analysis, we analyzed that the field duration has the maximum variance and contains values ranging from 0 to 4918, so we would be scaling down this feature. Meanwhile, we are also seeing if other features like age, price-index and confidence-index need to be normalized. Refer below figure to see high variance in duration feature.



Data Sampling:

Currently, the dataset has 90% of target values as 'NO'. Before training the model, we would have to sample the dataset such that the model is not biased due to highly one-sided outcome values. For example, if the model is biased so as to predict outcome as the majority categorical class from the training set, even then it will achieve an accuracy of 90% in this case, so one approach we have thought of is to downsample the majority class and upsample the minority class before we fit the model, hoping this would eliminate the bias.

Models:

We intend to test some of the following classification algorithms and choose the one which performs the best.

1. Logistic Regression
2. SVM
3. KNN
4. Random Forests

Prediction optimizations:

We will use cross-validation in order to evaluate the above algorithms. Also, we would compare different kernel functions for SVM, values of k for KNN, experiment tree depth and other parameters for Random forest to name a few and see what configurations work the best.

Evaluation:

We will create a confusion matrix based on our model. Based on this matrix, we can calculate the accuracy, precision, recall and F-score. Using these metrics, we can then plot the ROC and AUC curve to get more clarity on the performance.

Q5) What assumptions are safe to make?

We assume that there are no additional attributes affecting the outcome apart from the attributes present in the dataset. Also, at the moment, logically we assume that the "contact" feature does not affect the Target variable as the customer may be using telephone for such potential investments, so as to not share their personal cellular contact.