**Literature review:** There have been a few papers that have discussed this requirement. A common thread across all papers was the use of GLM based algorithms. Apart from that some other algorithms used were Neural Networks (in 1), Random Forests (in 1), KNN (in 1), CART (2), Naïve Bayes (in 3) and Support Vector Machines (SVM) (in 3). Out of these, Neural Networks and Random Forests seemed to stand out to giving better performances (as discussed in 1). We have not used KNN in our approach as we cannot interpret the effect of different predictors on our dependent variable (as discussed in 1). We have not used Neural Networks since it is a black box. Also, a Neural Network does not fit well to data that was not part of the original training dataset (as discussed in 1). In our approach, we did not use SVM as it tends to use a lot of processing power and can sometimes be non-responsive (as discussed in 3). Based on the literature review, we decided to apply GLM, CART and Random Forests for training the predictive models.

Data Imbalance (as discussed in 1) was another factor that was considered in one of the papers. This was addressed in that approach by using Over / Under sampling (or a mix of both) from the training dataset. However, the results from each of these approaches can vary considerably when applied in a real world situation. It will also differ based on the algorithms that will be applied. We have not addressed this in our approach since we believe that the data imbalance will be inherent in real data and the applied model should appropriately apply some bias based on this.

Duration was one of the variables highlighted in almost all papers. Some of the papers resorted to extensive feature engineering (1 and 3). However, the results in such papers showed that the basic variables like Duration were the ones that had higher predictive power as opposed to other exotic features. Again in our approach, we did not delve deep into feature engineering and stuck to the basic feature engineering. The advantages of extensive feature engineering seemed to be negligible.

References:

1. - Who Will Subscribe A Term Deposit? Jiong Chen (jc4133), Yucen Han (yh2645), Zhao Hu (zh2210), Yicheng Lu (yl3071), Mengni Sun (ms4783)
2. Predictive Modeling to Improve Success Rate of Bank Direct Marketing Campaign - Vaidehi R
3. A Data Mining Approach for Bank Telemarketing Using the rminer Package and R Tool - Sérgio Moro, Paulo Cortez , Raul M. S. Laureano