Final Capstone Proposal - Predicting whether a client subscribes to Term Deposit after a direct marketing campaign**,** by *Harshini Gadige*

Business Problem

There are several strategies a marketing team focus in order to engage, influence and ultimately guiding the customers/clients to purchase a product. Direct marketing is still one of the most common strategies followed in this internet era where digital, social media marketing are leading the way. Banking domain has large scope in improvising traditional strategies for choosing customers to focus on or prioritizing them while direct marketing. The real problem lies when a marketing specialist is given a huge data and asked to start direct marketing with the customers to influence them to purchase a term deposit. This predictive model helps marketing team and the bank operations team to identify the clients who have higher probability to subscribe the term deposit and prioritize them first during direct marketing and thus it helps to increase the customers of a bank.

Client

The client for this project is any public/private sector bank, which provides various deposit schemes to increase the number of customers. The goal is to increase the high ensured capital for the banks to play around in the financial market.

Data

The data set used for this project is the Bank Marketing Data Set which contains information extracted from direct marketing campaigns of a Portuguese banking institution. The dataset has 17 columns with 11162 observations ordered by date. It includes both numerical and categorical variables with the output variable ‘y’ - which tells if a client subscribed to a term deposit or not. The data files will be uploaded in the github repository of the final capstone project.

High Level Approach and techniques used

I will be using data cleaning techniques to detect null values, rename the variable names to meaningful ones, verify the datatypes of each column and see if it is correct according to the context of the variable. Next, will be doing exploratory data analysis which includes univariate and multivariate analysis and calculate correlation factor to identify the variables which are highly correlated to the target variable. I’ll check the distribution of the target variable and convert it to a normal distribution using boxcox transformation technique(only if applicable).
Also will be doing feature engineering to convert the variables into features. Will use Pandas get_dummies to convert categorical variables to a form that could be used by ML algorithms.
Will be using ensemble techniques like Gradient Boosting, Random Forest in this project.
At the end I’ll be implementing the deep learning technique that I learnt from my specialization using Keras with Tensorflow backend to create a neural network. Finally I compare the models and perform evaluation metrics to identify the best one. I will work on hyperparameter tuning inorder to optimize the models.

Anticipated Challenges

The challenge that I anticipate is with the ‘duration’ variable. Below is the description given in the data set page about this variable.

duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

As it is mentioned in the dataset description, it highly affects the target variable and the real challenge lies in deciding whether to include this or discard this. I will be trying with both versions of the model one with including ‘duration’ and one without it to see the performance variation.

Deliverables

The deliverables for this capstone project include annotated source code, which includes data cleaning, EDA, Feature Engineering and predictive classification techniques using Machine Learning and Deep Learning. I’ll also include the data set files. Finally, this information will be visually presented in the form of a slide deck and presentation to stakeholders.