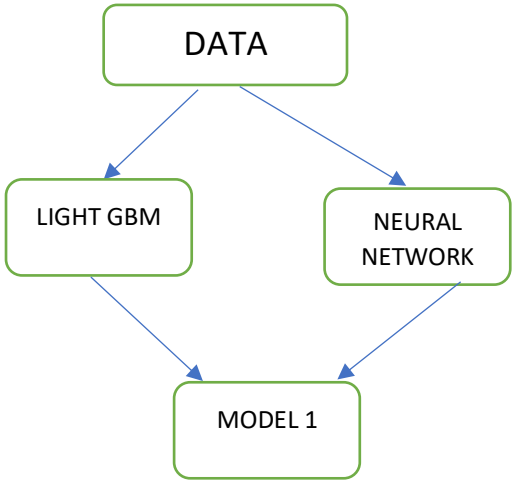
JOBATHON MAY28-MAY30

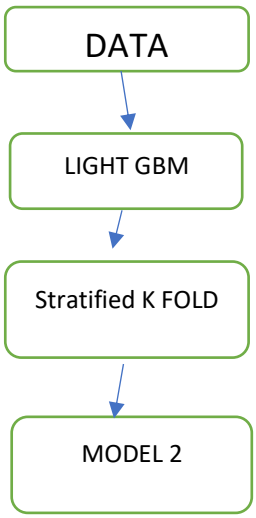
This dataset aims to solve the problem of finding the customer who would like to take interest in taking the credit. The data contains the details of customers who all are eligible for taking credit cards

INITIAL ATTEMPT



This was my first Approach in Solving the Problem, after visualising the dataset using EDA and other transformations, the first approach which clicked in my head is to use Ensemble Techniques and then If that did not work the use Neural Network and if that did not give me the Required solution then use a combination of both, This Approach after all the Pre-processing and EDA gave me a result of 85.4% ROCAUC on Submission, viz I had used AUTO VML and Randomized Search CV for getting the best parameters for LIGHT GBM and Neural Network, Here is the final Approach which Worked for me.

FINAL ATTEMPT



In the previous Attempt i.e. where my model was a combination of Neural Network and LightGBM, I got a descent score, but then I started exploring LightGBM more ,I used different kinds of hyperparameter techniques to get the best Parameters like Randomized Search CV, Bayesian Optimization and all, after finding the best parameters I used Stratified KFOLD CROSS VALIDATION on my model and tested on the testing dataset, it gave me a descent score of 87.1% which was more than the previous one.

2.Approach For the Problem

My initial approach for the problem was simple, First I build the baseline model with logistic regression and submitted it. After then I started exploring the ensemble techniques like LightGBM, Random Forest and many more, After that I started exploring models and checked my score for the validation set. After giving a dry run for these models, I checked out for possible transformations which I can make in my data, like can I make a transformation such that I get the Data in Gaussian Distribution and then try out Neural Network and other techniques, so I explored on EDA part more and I found that the Avg_balance column was left skewed and then I applied box-cox transformation to make it centrally-skewed, then after came the categorical features, first I used get dummies method, which did not give a good score ,then I moved on with mean encoding which gave a descent score not that great as well, then I used Label Encoder on all columns except for the Region Column ,on Region Col I used mean encoding, After all Encodings I decided to check out for Outliers, I also did the outlier removal also but that did not work and model performance decreased then I realized sometimes in the finance sector domain outliers helps us to capture the underlying patterns in the data which rarely occurs. I applied 10 Fold Cross validation to train all my models and reached to final model.

3.Feature Engineering Ideas

I tried many variants here, first I started of with categorical features, I noticed that there is one feature which had Null values, I fixed it by filling in with some Random Value, like 'DK', then I used the Dummies method first and checked with all my ensemble and Neural Nets, It did not give that good score, then I switched on to Label Encoding except for the Region Col where I applied Mean Encoding as it's count for more, after then I had a look at the distributions, I noticed few transformations may help us, so I made few Quantile and Log, Box-Cox Transformations on the data, and noticed that there are few outliers, I removed the outliers and noticed that this dropped my performance, outliers may help sometimes, so I filled those outliers with the most frequent value of that col, then after I moved on for Scaling and Transforming the Data.