Hackathon Context :

The data is related to direct marketing campaigns of a financial institution.The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed. You will have to analyze the dataset in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank.

Input :

Task-1

1. You will be given a data.csv file. Use this dataset for training your model.
2. You will be given a test_data.csv file. Give predictions on this dataset.
3. We have provided the attribute information.

Task-2

4. We have given 5 ML Questions from the concepts we learned on Day 1 of ML Marathon.

Process :

What are we expecting from your submission ?

1. You have to upload the jupyter notebook with your solution.
2. You have to upload the result.csv file **(with 1 target column “deposit”)** that you predicted for test_data.csv.
3. Precision, Recall and F1 Scores for the final model you have selected, you have to fill in the google form.
4. Submit your answers to 5 ML Questions in the PDF format and upload it.

Output / Goal :

The classification goal is to predict if the client will subscribe (yes/no) to a term deposit (variable y).

Note : Don’t cheat as we will be assessing the notebooks very carefully on the below mentioned 5 points. We wish you all the best. Happy Learning !!

Selection Criteria :

1. **How well did you understand the dataset ?**
   Try giving explanations in the jupyter notebook.

2. **What information did you collect after doing Data Visualization?**
   We are not interested in the number of graphs, but in the information you gathered after you plotted a graph. Try giving the insight you got from a particular graph and use that information forward.

3. **Which model have you chosen and why ?**
   We are interested in knowing which different models you tried and why you moved to a particular model. Try giving proper explanations for each transition you are making.

4. **Finally the metrics !!**
   We are interested in knowing your (Precision, Recall and the F1 Score) for the final model you have chosen.

5. **Your final thoughts on some of the ML concepts that we discussed on day 1.**
   You will be given 5 Questions that you have to answer with proper explanations.

Attribute Information:

Input variables:

1 - **age (numeric)**

2 - **job** : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

3 - **marital** : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - **education** (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - **default**: has credit in default? (categorical: 'no','yes','unknown')

6 - **balance:**

7 - **housing**: has housing loan? (categorical: 'no','yes','unknown')

8 - **loan:** has personal loan? (categorical: 'no','yes','unknown')

related with the last contact of the current campaign:

9 - **contact**: contact communication type (categorical: 'cellular','telephone')

10 - **month**: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

11 - **day**: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

12 - **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.

other attributes:

13 - **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

15 - **previous**: number of contacts performed before this campaign and for this client (numeric)

16 - **poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Output variable (desired target):

17 - **deposit** - has the client subscribed a term deposit? (binary: 'yes','no')

ML Questions

1. While attempting Problem 1 in the hackathon, which feature/features according to you played the most important role in prediction of the deposit? Can you visualize the same and which algorithm you used for determining that and why?

2. The data science team in your company is trying to build a classifier with a target having two possibilities (yes/no) and what they are interested in is that the solution must be fast enough to be used by the deployment team to predict for the new dataset. Which algorithm do you suggest and why?

3. Suppose the data you used in building a model has a low variance and low bias. Can you guess how the data will look like in the predicted outcome? Give a proper explanation.

4. Suppose you have a dataset that has missing values, and you tried multiple classification algorithms but you are not getting good results, which algorithm do you recommend to the team and why?

5. Can we use the same data for testing and training, if yes then why and if no, why not? Please provide 1 example that you used to come up with your findings. (**Note**: You can try this in the jupyter notebook and paste the screenshots in the PDF).

I wish you all the Best. Accept the challenge and Happy Learning !!

Remember Participation is more important than winning.

Thanks & Regards,
Pallavi Pannu