**Volv AI Test**

* You are given a train data set having 1000 columns and 1 million rows.The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)

**Solution:**

Primary goal of any data entusiust after observing such data viz. 1000 colums (observations) and 1 million rows (Parameter / Variables) , Such Data might cost more compution and consume more time for model training, which can be cumbersome while taking inference or model deployment. The best solution is to reduce the dimentionality of Dataset with can be acheived my various mean. Most commonly used practice is generating Heat map of co-relation between variables and eliminating the variables with co-relation coeficient near to 1. Principle Component Analysis is one of the key Dimensionality reduction techniques used for Converting the dataset with n variable to n or less then n variable. This approach is very easy and might be most efficient as well. Feature engineering is one of the best approch out there to reduce the number of variables by means of feature extraction. PCA is one of the feature engineering technique aswell. In, simpler term feature are derived from existing variable, by combing most or perform in eigen vector operations. Libaries such as sklearn , pywavelet etc offer feature extraction at granular level. According to the dataset and domian of data, one or more approches of above can be implemented for Dimentionality Reduction.

* You are given a data set on cancer detection. You’ve build a classification model and achieved an accuracy of 96%. Why shouldn’t you be happy with your model performance? What can you do about it?

**Solution:**

In the case of Cancer dectection for any paitent using machine learning technique’s has its ups and downs. The ups being that the predictions are made instanly after data aquired by patient. While training such model an enthusist has to think like the doctor himself, one wrong step and the patient won’t even know he had cancer untill its to late. The major issue with the accuracy is, that even with 4% error their is the burden of predicting wrong for an certain group of indiviual. Classificated predicted valve thresshold can help a bit to easy mind but still won’t be enough to satisfy. Meaning as for eaach indiviual the prediction will return decimal valve, which denote probability of an even happening, usually the clasification rounds off the valve with 0.5 as thresshold i.e f the prediction is below 0.5 than EVENT DOESN”T OCCUR else EVENT OCCUR’S. We can experiment with his thresshold to prectict EVENT OCCUR when its lower which might help in increasing the overall accuracy and precision of predictions. Ideally the 100% accuracy is the only optimal solution for problem that involve, life threating outcomes.

* After spending several hours, you are now anxious to build a high accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of models could perform better than benchmark score. Finally, you decided to combine those models. Though, ensembled models are known to return high accuracy, but you are unfortunate. Where did you miss?

**Solution:**

Basically, an ensemble is a supervised learning technique for combining multiple weak learners models to produce a strong learner. Ensemble model works better, when we ensemble models with low correlation. Boosting is one of the ensembling techniques, while using 5 GBM the featues and learner of the model remain intant while combining them, they also have high correalation between them this lead to the unfortunate event. As Boosting itself is ensemble technique, to use ensembling agian is pointless.

* You are assigned a new project, which involves helping a food delivery company save more money. The problem is, company’s delivery teams aren’t able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine-learning algorithm can save them?

**Solution:**

The First and most important step while doing such project will be to find the reason why the food is delivered late. There can be number of factors that might affect the delivery. With the machine learning point the might lie with increase in demant, Vechicle issues such as traffic or lesser man power. The best fit that could save the business and deliver food on time would be forcasting. The most common termenolgy used in Data Science and Statistics is Time Series Analysis. With the help of algoriths like ARIMA,naiive or HOLT-WINTERS the problem can be tacked. With the help of short-term and long-term forcasting along exponential smoothing the best results can be achieved.

* Write code to train, test and deploy a model for sentiment analysis of Twitter’s tweets

**SOLUTION IN ATTACHMENT**

* Using any programming language, solve this:

a. You’ll be given a string such as “aaabbcc” and a number n (integer). These are the parameters to your function

b. Your function needs to remove any Alphabetic character that appear exactly n times

c. Your output should be “aaa”

**SOLUTION IN ATTACHMENT**

**def maxRepeating(str,num):**

**temp = str**

**n = len( str )**

**#a = []**

**cur\_count = 1**

**for i in range( n ):**

**if (i < n - 1 and str[i] == str[i + 1]):**

**cur\_count += 1**

**else:**

**if num == cur\_count:**

**res = str[i]**

**#a.append(res) #to get striped once**

**maxima = cur\_count**

**res = res\*maxima**

**cur\_count = 1**

**temp = temp.strip( res )**

**return temp**

**ent = input('Enter the String:')**

**number = int(input('Enter Number:'))**

**print(maxRepeating(ent,number))**

**OUTPUT:**

**/home/yog/PycharmProjects/String\_find/venv/bin/python /home/yog/PycharmProjects/String\_find/Problem\_6.py**

**Enter the String:aabbbcccnn**

**Enter Number:2**

**bbbccc**

**Process finished with exit code 0**