Date: 21/08/2020

Major Project (Group - 26) – Initial Approach

Problem Statement - Using Exploratory Data Analytics and Other Trends in Data to Suggest the Appropriate Machine Learning Models for a User Given Dataset.

Initial Approach –

Due to lack of published material for the above problem statement we have formulated a simpler approach to start with this project. Once we get some substantial results from the same, we would eventually extend the project to its full requirements considering practical situations of use.

- Initially we will consider that we have to suggest the best algorithm from 5 basic machine algorithms and afterwards we would extend our work for as much algorithms as possible. The 5 ML algorithms we would be considering initially would be,
    1. Linear Regression
    2. Logistic Regression
    3. Naïve Bayes Classifier
    4. Decision Trees
    5. Support Vector Machines

- We would pick around 10 to 15 common datasets that have well published work that could help us to know what are the best algorithms that work with the respective dataset. For example, we would be picking a few datasets from the Scikit-Learn datasets, as they are very common and have well documented papers with enough information on which algorithms work best with them. Going further we would extend and test our work for more complex and different types of dataset.

- After this we would try out various statistical data analysis methods, few of which are mentioned below. We would also extend the number and quality of techniques used as the works proceeds.

    1. Mean, Median & Mode.
    2. Standard Derivation
    3. Anova
    4. T-test
    5. Hypothesis Testing
    6. Chi-Square Test
    7. Dimensionality Reduction Techniques Like PCA.

- We would also perform extensive visualisation-based analysis using various sorts of techniques available. We would also provide the user with an interactive dashboard where he/she can view and observe the trends & other statistical information in the through interactive plots. We would be using libraries like matplotlib, seaborn & plotly.

Resources Taken in Consideration

1. **(Research Paper) Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, Sebastian Raschka (**Link - https://arxiv.org/pdf/1811.12808.pdf).

2. **(Blog) -** https://medium.com/@statswork/statistical-data-analysis-statswork-c050101db816

3. **(Blog) -** https://www.bigskyassociates.com/blog/bid/356764/5-Most-Important-Methods-For-Statistical-Data-Analysis

4. **(Blog) -** https://medium.com/cracking-the-data-science-interview/the-10-statistical-techniques-data-scientists-need-to-master-1ef6dbd531f7

These is our initial working plan; we expect your opinion and inputs regarding how could we improve on this and carry out our work in a better way.

Sahil Gupta (9917103163)

Shivam Aggarwal (9917103169)