**CLASSIFYING THE PRICE OF MOBILE PHONES BY APPLYING ML MODELS WITH A MULTINOMIAL APPROACH USING PYTHON LANGUAGE**

**Statement of Work SOW (V1)**

**Business Understanding & Problem Discovery**

**NAME: HEET SONI**

**STUDENT NUMBER: 100768241**

**PROFESSOR: MARCOS BITTENCOURT**

**EXECUTIVE SUMMARY**

Mr.Bush started his new mobile company. For better profit, he wanted to check which mobile feature is affecting the price of the mobile phone. Mobile features like battery, Wi-Fi, 3g/4g support, sim support, camera, clock speed, etc are changing the price of the phone. With the help of various machine learning models, we will identify which mobile feature is more important. Moreover, in the end, we will classify the mobile price range.

**BUSINESS PROBLEM STATEMENT**

Classify the mobile pricing in comparison with mobile set features. Additionally, identify which mobile features affecting the price most for a better mobile selling point.

**RATIONAL STATEMENT**

Mr.Bush wanted to see how each mobile set feature is changing the price of the mobile phone. He is looking for the relationship between the phone's price and its features. With the help of that relation, he wanted to build one multinomial model for the classification of the mobile price range.

**REFINED PROBLEM STATEMEN**

The key problem is to classify the mobile price range as compared to its feature selections. Here, in this problem, every mobile feature is affecting the mobile price. So, the subproblem is to illustrate the relationship between price and the features.

**DATA REQUIREMENTS**

* The first data requirement will be to find out all the valuable mobile features set that affect our pricing range.
* For deploying a more accurate and trustable model, we need diverse and large data.
* The data should be in excel or .csv format for execution.
* The data must be in the proper size for the final selected model.
* The dataset is expected to not have any kinds of special characters.
* The dataset is required to have all the values in numerical format.

**DATA SOURCE**

Mobile price classification:

<https://www.kaggle.com/iabhishekofficial/mobile-price-classification>

**DATA DESCRIPTION**

The dataset contains various types of independent variables and one dependent variable. Dataset has 2001 rows and 21 columns. All the variables are in numerical formats.

Independent variables:

Battery power, Blue, Clock speed, Dual sim, Fc, Four g, Int memory, M deep, Mobile w, N cores, Pc, Px height, Px width, Ram, Sc h, Sc w, Talk time, Three g, Touch screen, wi-fi.

Dependent variables:

Price range

**ASSUMPTION**

* First, we assume that our data is coming from a genuine source.
* Additionally, we are believing that there was no data manipulation happened before getting the data source from the source.
* We assume that the dataset has all the valuable features to conduct this pricing model.
* For better accuracy, we assume that all the independent variables are highly correlated to our dependent variable.
* Our dataset has all the correct values and no null or unwanted values.
* We believe that we have enough data to train and test our ML model.

**LIMITATIONS**

* For accurate training and testing sets, we only limited datasets.
* We do not know which company's phones we are putting them for the test.
* The dataset is categorical, so to run the linear regression model is hard.
* This dataset might be biased if the dataset is collected via only one user.

**CONSTRAINS**

* The main constrain is that some of the independent variables are not well defined.
* The final count for the dependent variables is an imbalance.
* We do not have the proper information about which variables are more suitable for the model building.
* We have no information about the outliers available in the given dataset. They may occur in the EDA process.
* Few of the independent variables have some value as "0" and we do not know whether these values are null values or real measured values.

**TEST PROCESS**

* In the beginning, we will perform the EDA process on the dataset.
* Will identify null values, outliers, and unwanted data columns.
* Remove all the null values from the dataset and with the help of the Tukey method, we will remove all the outliers.
* To analyze the relationships with the mobile's price range, we will plot some histograms and density graphs.
* Hit map will illustrate the correlation between the independent variables.
* With that, we can have a rough idea about which variables are useful for the model.
* To ensure the correlation between variables, we will plot a few scatter plots.
* Find the VIF factor score to see the multicollinearity in the dataset.
* If the dataset is imbalanced, we will use the SMOTE technique.
* In the end, we will deploy one suitable model for the modified dataset.
* We will focus on the accuracy score of the mode.
* To ensure a better accuracy score, we will use cross-validation and hyper parameters.
* In the end, we will focus on our accuracy score as our ground truth. This will give us the measurement of our test process.