**Mobile Price Range Prediction**

**Name: Mohd Aun Farooqui**

**INDEX**

**SL.NO Topics**

1. **Abstract**
2. **Introduction**
3. **Project Goal**
4. **Attributes**
5. **Exploratory Data Analysis**
6. **Data Visualization**
7. **Algorithms**
8. **Conclusion**
9. **Challenges**
10. **Abstract**

The mobile phone industry has been rapidly growing and evolving over the years. With the increasing number of mobile phone options available in the market, it can be challenging for customers to choose the right phone at the right price. In this project, we aim to predict the price range of mobile phones based on various features like battery power, RAM, internal storage, etc. We have used a dataset consisting of 2,000 mobile phones and their features to train several machine learning models and select the best performing model. We have also used SHAP (Shapley Additive explanations), a model explainability tool, to understand the importance of different features in predicting the price range. The goal of this project is to provide customers with a better understanding of mobile phone pricing and help them make informed purchase decisions.

1. **Introduction**

The mobile phone has become an integral part of our daily lives, and with the emergence of new technology, the market for mobile phones has become extremely competitive. With this competition, mobile phone manufacturers are continuously working on improving their products and adding new features, which in turn is leading to an increase in the number of mobile phones available in the market.

To aid buyers in selecting the best mobile phone to suit their needs, we can use Machine Learning algorithms to predict the price range of a mobile phone based on its features. In this project, we will use a mobile phone dataset to develop a machine learning model that can accurately predict the price range of a mobile phone.

1. **Project Goal**

The goal of this project is to predict the price range of a mobile phone based on various features such as battery power, RAM, internal memory, and others. The project aims to build a machine learning model that can accurately classify the price range of a mobile phone and provide insights into which features are most important in determining the price range.

1. **Attributes**

* Battery\_power : Battery capacity in mAh
* Blue: Has bluetooth or not
* Clock\_speed: speed at which microprocessor executes instructions
* Dual\_sim: Has dual sim support or not
* Fc: Front Camera mega pixels
* Four\_g: Has 4G or not
* Int\_memory: Internal Memory capacity
* M\_dep: Mobile Depth in cm
* Mobile\_wt: Weight of mobile phone
* N\_cores: Number of cores of processor
* Pc: Primary Camera mega pixels
* Px\_height: Pixel Resolution Height
* Px\_width: Pixel Resolution Width
* Ram: Random Access Memory in MB
* Touch\_screen: Has touch screen or not
* Wifi: Has wifi or not
* Sc\_h: Screen Height of mobile in cm
* Sc\_w: Screen Width of mobile in cm
* Talk\_time: longest time that a single battery can last over a call
* Three\_g: Has 3G or not
* Wifi: Has wifi or not
* Price\_range: This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost), 3(very high cost)

1. **Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a critical first step in any data analysis project. It helps to understand the data and identify patterns, relationships, and outliers in the dataset. For the bike sharing demand prediction project, EDA was performed on the dataset to gain insights into the data and understand its structure.

1. **Data cleaning**

Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in a dataset. This process involves handling missing or incorrect data, removing duplicates, converting data types, and dealing with outliers. Data cleaning is a crucial step in any data analysis or machine learning project as it ensures that the data is accurate and reliable for use in modeling and analysis. Failure to clean the data can result in incorrect conclusions, biased models, and poor business decisions.

Our dataset has no missing and duplicate values.

1. **Data manipulation**

Data manipulation is the process of transforming and cleaning the data to make it more structured and organized for further analysis. In this project, data manipulation was performed using various techniques such as merging, grouping, filtering, and sorting to prepare the data for modeling.

1. **Data Visualization**

Data visualization is the graphical representation of data and information using visual elements like charts, graphs, and maps. It helps to communicate complex data and information in an easily understandable and accessible format.

* Pie chart
* Box plot
* Heatmap
* Pair plot
* Count plot

1. **Algorithms**

* **Decision Tree**

A decision tree is a machine learning algorithm used for both regression and classification tasks. It is a tree-like model where the internal nodes represent features, the branches represent the decision rules, and the leaf nodes represent the outcomes or predictions.

* **Logistic Regression**

Logistic Regression is a statistical algorithm used for binary and multi-class classification problems. It is a supervised learning algorithm that predicts the probability of a dependent variable that can take one of two values (0 or 1) based on one or more independent variables. It estimates the probability of an event occurring based on previous data observations. Logistic regression assumes that there is a linear relationship between the independent variables and the log odds of the dependent variable. It uses a logistic function, also known as a sigmoid function, to transform the linear equation into a probability score that lies between 0 and 1.

* **XgBoost Classifier**

XgBoost (Extreme Gradient Boosting) is a popular and powerful open-source gradient boosting library that can be used for classification, regression, and ranking problems. It uses a tree-based ensemble model that sequentially trains a series of weak models (typically decision trees) and combines their predictions to form a final model prediction. XgBoost is known for its speed and performance, as well as its ability to handle large datasets with high dimensionality. It also has several built-in regularization techniques to prevent overfitting and improve generalization performance.

* **Support Vector Classifier**

Support vector classifier (SVC) is a type of binary classification algorithm that uses a hyperplane to separate data points into classes. The algorithm tries to find a hyperplane that maximizes the margin between the two classes, which are defined by their closest data points, known as support vectors. The goal of SVC is to find the decision boundary that best separates the two classes, while also minimizing the number of misclassifications.

SVC is often used in machine learning for classification tasks where the data is not linearly separable. In these cases, the SVC uses a kernel function to transform the data into a higher-dimensional space, where it may be linearly separable. Common kernel functions include the radial basis function (RBF), polynomial, and sigmoid kernels.

1. **Conclusion**

* Based on the analysis and modeling performed in this project, we can conclude that the mobile price range can be predicted with a good accuracy using machine learning algorithms.
* Ram has continuous increase with price range while moving from Low cost to Very high cost.
* The mobile price range has almost constant battery, internal memory, and primary camera.
* We started by exploring the data and cleaning it by handling missing values, feature scaling, and one-hot encoding.
* After that, we performed feature selection and feature engineering to extract new features from existing ones.
* We then trained and evaluated four machine learning models: Decision Tree Classifier, Logistic Regression, XGBoost Classifier, and Support Vector Classifier.
* We used cross-validation and grid search to optimize the hyperparameters of each model.
* Our best performing model was the Logistic Regression, which achieved an accuracy of 95% on the train set and 92% on the test set.
* We used SHAP values to understand the feature importance and explain the model predictions.
* The feature RAM, battery power, pixel size played more important role in deciding the price range of mobile phone.
* Overall, our project demonstrates the effectiveness of machine learning algorithms in predicting mobile price range and can be used by mobile manufacturers to estimate the price range of their products based on the features. However, there is still room for improvement in terms of feature engineering and exploring other machine learning algorithms.

1. **Challenges**
2. **The amount of data collected:** There are some unnecessary raw data collected which do not contribute much to the study. Thus, to identify them and eliminate them will be a challenge.
3. **Handling Null Values and outliers:** - Identifying Null values and outliers and handling them. Handling them is different in different cases. So, we need to analyse it and to process.
4. **Processing raw data into meaningful data:** Sometimes the columns can’t be used as to understand how apply functions on these columns to get more relevant results.
5. **Visual representation:** One size doesn’t fit all. Thus, finding exact graphs to represent the data is challenging.
6. **References**
7. <https://www.almabetter.com/notes>
8. <https://www..python.org.com>
9. <https://www.stackoverflow.com>