Mobile Price Range Classifier Documentation

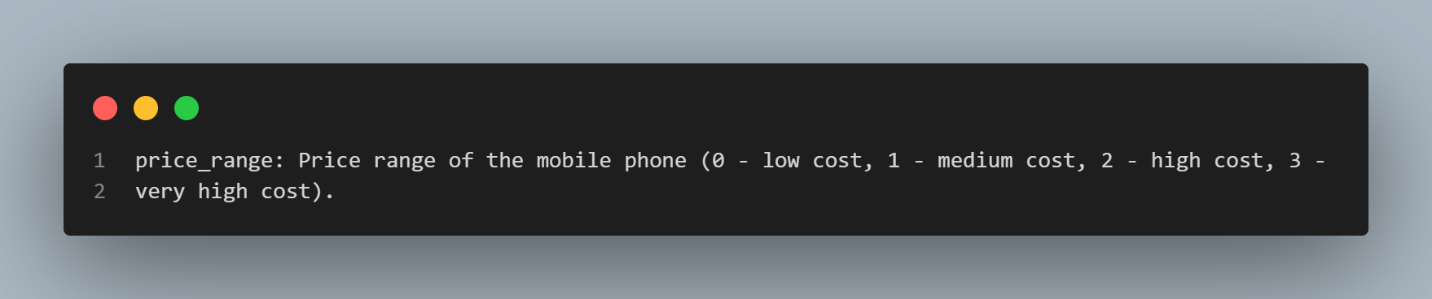
Section 1: Introduction

In the contemporary era, the mobile phone industry is at the forefront of consumer technology. This project, part of the Mentorness internship, aims to harness machine learning to predict mobile phone price ranges based on various features.

Section 2: Problem Statement

The objective is to create a classifier that can discern between different price brackets of mobile phones, helping consumers and businesses alike in decision-making processes.

Section 3: Dataset Description

The dataset consists of several mobile phone attributes, including battery power, camera specifications, memory, and connectivity options. It categorizes phones into four distinct price ranges which are to be predicted, providing a comprehensive view of the market landscape.

Section 4: Methodology

The methodology Section will include data exploration, data cleaning, feature analysis, model selection, and model validation.

Data Exploration and Cleaning: Loading Data: The dataset is loaded into a pandas DataFrame, allowing for manipulation and analysis.

A screen shot of a computer code

Description automatically generated

Identifying Data Types: Ensuring correct data types are assigned to each feature for accurate analysis.

A black rectangular with yellow green and blue text

Description automatically generated Handling Missing Values: Checking for missing values to maintain data integrity.

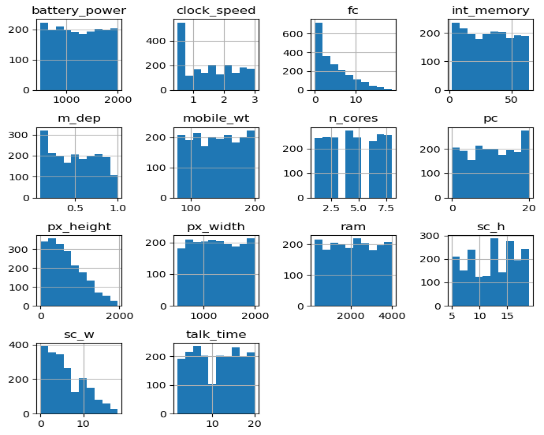
Exploratory Data Analysis (EDA):

Statistical Summary: Describing the dataset statistically to identify distributions and outliers.

A black rectangular with yellow green and blue dots and white text

Description automatically generated

A graph with lines and dots

Description automatically generated with medium confidenceData Visualization: Using histograms, boxplots, and heatmaps to visualize data and discover patterns and correlations.

A diagram of a number of numbers and a number of squares

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generated

Section 5: Preprocessing

Data preprocessing involves transforming raw data into an understandable format for machine learning models.

Feature Engineering: Categorical Encoding: Binary categorical variables such as 'blue', 'dual\_sim', etc., are encoded appropriately (0 or 1).

Feature Scaling: Numerical features are scaled using standardization to improve the performance of the algorithms, although this step is considered as a part of EDA but it is seen later on the model section as per the figure below.

Section 6: Model Selection, Evaluation, and Testing

Several machine learning models are explored for this classification task.

Model Selection:

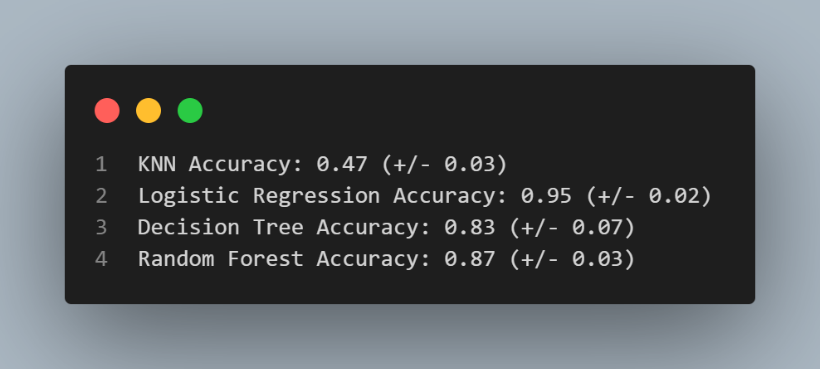
K-Nearest Neighbors (KNN): A model that classifies data based on the majority class of its nearest neighbors.

Logistic Regression: A model that predicts the probability of categorical outcomes.

Decision Trees and Random Forest: Models that predict the value of a target variable by learning decision rules inferred from the data features.

Model Evaluation:

Cross-validation: Using cross-validation to assess the performance of the models and prevent overfitting.

A screen shot of a computer code

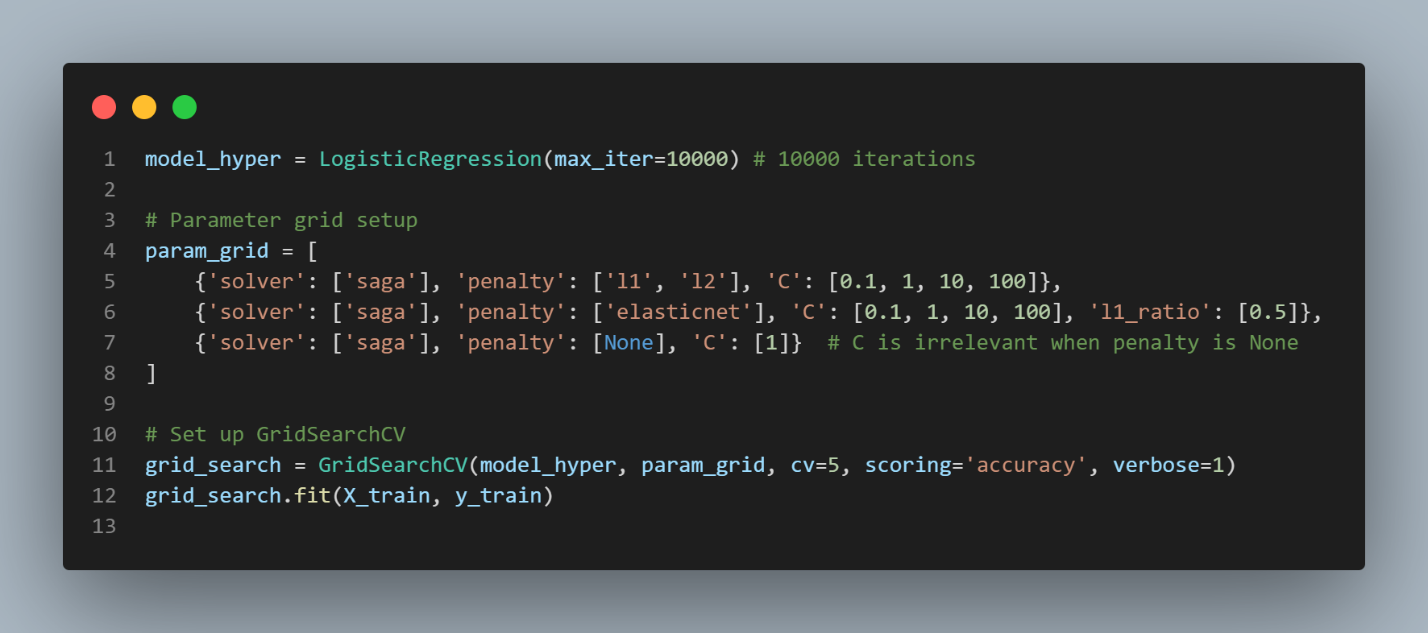
Description automatically generated

Model Performance:

Accuracy Score: The ratio of correctly predicted observations to the total observations.

Precision, Recall, and F1-Score: Metrics that provide a more detailed view of the model performance beyond accuracy.

Hyperparameter Tuning: Employing GridSearchCV to find the optimal parameters for the selected model (Logistic Regression).



Making predictions: Using the pre-tuned model (similar metrics to the tuned version) on the test set while measuring the frequency count of the neighbors’ price ranges.

A black screen with white text

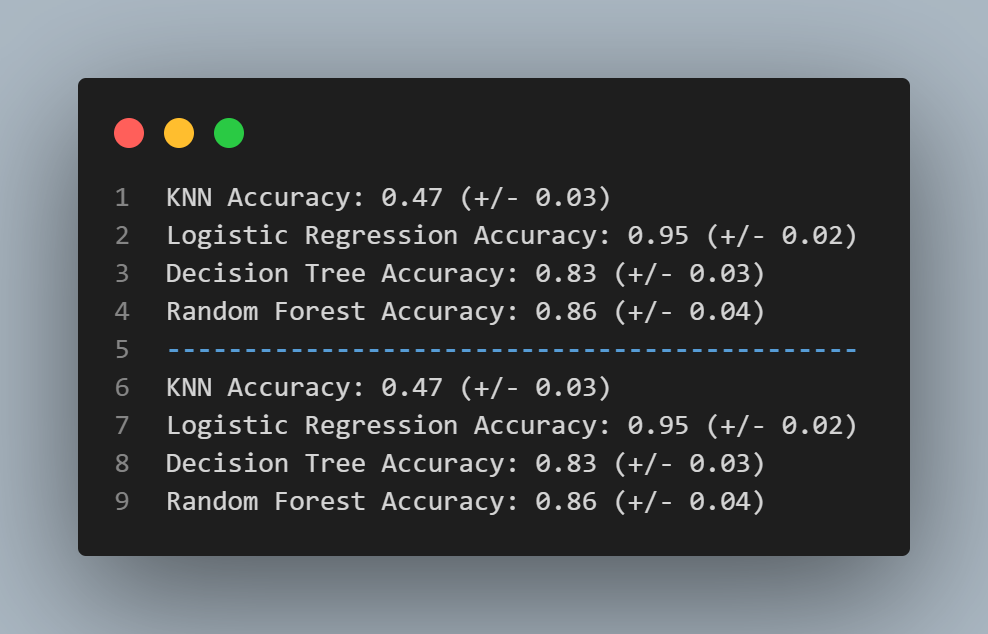
Description automatically generated

Section 7: 2 cases of feature selection

Case 1: Numerically (Statistically associated) and Categorical variables

Case 2: Highly correlated variables

Both Case 1 and Case 2 yielded similar results to the original model using all features.



Section 8: Conclusion

The conclusion synthesizes the findings, discusses the effectiveness of the models, and provides recommendations for practical applications and future research. Future Work: Suggests potential improvements, additional data collection, and exploration of more complex models.

Yaser Alshuaybat

24/4/2024