**Exploratory Data Analysis and Implementation of machine learning algorithm in mobile models’ dataset**

A project report for Internship Programme

Submitted by

Rishav Bairagya

Sarthak Dutta

Souma Das

BONAFIDE CERTIFICATE

Certified that this project work was carried out by the following members

***“Exploratory Data Analysis and Implementation of machine learning algorithm in mobile models’ dataset***“is the bonafide work of

|  |  |  |  |
| --- | --- | --- | --- |
| ***Name of student:*** | ***Rishav Bairagya*** | ***Signature:*** |  |
| ***Name of student:*** | ***Sarthak Dutta*** | ***Signature:*** |  |
| ***Name of student:*** | ***Souma Das*** | ***Signature:*** |  |

Acknowledgement

I take this opportunity to express my deep gratitude and sincerest thank to my project mentor, **Soumitra Biswas** giving most valuable suggestion, helpful guidance and encouragement in the execution of this project work.

I will like to give a special mention to my colleagues. Last but not the least I am grateful to all the faculty members of Globsyn Technologies or their support.

**Index**

1. **Introduction** 
   1. What is data analytics
   2. Libraries used in this project
   3. Implemented procedure
2. **Project Explanation**
   1. About dataset
   2. Explanation of each procedure with data
   3. Explanation of data
   4. Encoding data
3. **Machine learning**
   1. Libraries imported
   2. Terms used in machine learning
   3. Machine learning methods used summary
   4. Which model had the best accuracy
4. **Conclusion**
5. Introduction

In this Project, we have aimed to predict the selling price of mobile phones from different Brands using various machine learning algorithms, to find the best suited algorithm for our dataset, performance comparison, and an exploratory data analysis on the mobile sales dataset.

**1.1 WHAT IS DATA ANALYTICS**

Data analytics is the process of examining, cleaning, transforming, and modelling data to discover useful information, draw conclusions, and support decision-making. It involves using statistical techniques, algorithms, and software tools to analyse raw data and extract meaningful insights that can inform business strategies, improve processes, and predict future trends.

Key aspects of data analytics include:

* Data Collection: Gathering relevant data from various sources, such as databases, spreadsheets, or external sources like social media.
* Data Cleaning: Removing errors, duplicates, and inconsistencies from the data to ensure accuracy and reliability.
* Data Exploration: Using descriptive statistics and visualisation tools to understand the data, identify patterns, and detect anomalies.
* Data Modelling: Applying statistical models or machine learning algorithms to the data to identify relationships, predict outcomes, or classify information.
* Data Interpretation: Analysing the results to draw meaningful conclusions and provide actionable insights.
* Data Visualization: Presenting the findings in a clear and understandable way, often through charts, graphs, or dashboards.

Applications of Data Analytics:

* Business Intelligence: Improving decision-making by analysing sales data, customer behaviour, and market trends.
* Healthcare: Enhancing patient care through analysis of medical records and treatment outcomes.
* Finance: Detecting fraud, managing risk, and optimising investment strategies.
* Marketing: Personalising marketing campaigns by analysing consumer data and preferences.

**1.2 LIBRARIES USED IN THIS PROJECT**

Python is widely used in machine learning due to its simplicity, readability, and the extensive ecosystem of libraries an frameworks that facilitate the development and deployment of machine learning models.

Here are some Python Libraries that we used for our project:

1. Data Preprocessing:

Libraries: Pandas, NumPy, Sklearn-Ordinal Encoder.

Dataset cleaning, transforming, and organizing raw data into a suitable format for analysis. Libraries like Pandas and NumPy are commonly used for handling data structures and performing mathematical operations. Encoding and Scaling are done with the help of Sklearn Ordinal encoder.

1. Exploratory Data Analysis (EDA):

Libraries: Matplotlib, Seaborn.

Python's visualization libraries like Matplotlib and Seaborn allow for creating plots, charts, and graphs to understand the underlying patterns, trends, and relationships in the data. EDA is crucial for selecting features and understanding data distribution.

1. Model Training and Evaluation:

Libraries: Scikit-learn, XGBoost.

These libraries allow for the efficient training of models using various algorithms and evaluating their performance using metrics like accuracy, precision, recall, R2 score, etc. Cross-validation and hyperparameter tuning can also be done using these.

**1.3 PROCEDURE IMPLEMENTED**

Steps taken to complete this project:

1. Data Preprocessing: first we have begun by loading the dataset and performing data preprocessing tasks such as removing the duplicate values, null values, and cleaning the dataset by removing some unwanted white spaces in certain features, additionally we have converted the data irregularities under storage, and memory which will be very helpful while implementing ML algorithms. We have also performed Ordinal encoding for categorical values and then scaling them appropriately, this part will be explained in detail later on.
2. Exploratory Data Analysis (EDA): We conducted EDA to understand the distribution of variables, identify outliers, and gain insights into the relationships between features and the target variable.
3. Implementation of machine learning algorithms to our dataset  
   1. Regression Modelling: We have implemented various Machine Learning algorithms which include Linear Regression, Random Forest Regression, Support Vector Machine Regression, and K-Nearest Neighbours Regression, ADA-Boost regressor, Polynomial regressor.
   2. Model Evaluation: For each regression model, we evaluated its performance using Root mean square, Standard deviation, R² score, Absolute mean error to assess how well each model predicted smartphone prices.
4. Conclusion: After comparing the performance of various regression models, we have come up with a statement that the random forest regressor worked best for our dataset and was able to predict the selling price with accuracy score higher than other regressor models.

3. Machine Learning

**3.2 Terms used in machine learning implementation**

**Fitting:** It refers to the process of training a model by adjusting its parameters to match the patterns found in the training data. The goal is to make the model capable of accurately predicting or classifying new, unseen data based on what it has learned from the training set.

**Mean Squared Error (MSE)** is a common metric used to evaluate the performance of regression models in machine learning. It measures the average of the squares of the errors, where the error is the difference between the actual value and the predicted value by the model.

where is the actual value, is the predicted value, and n is the number of observations.

**Root Mean Square Error (RMSE):** is a measure of the differences between the predicted values generated by a model and the actual values observed. It is the square root of the average of the squared differences between predicted and actual values.

RMSE **=**

where is the actual value, is the predicted value, and n is the number of observations.

RMSE gives a sense of how far the predictions are from the actual values in the same units as the target variable. A lower RMSE indicates a better fit of the model to the data. RMSE is used since it is more comparable with another metric standard deviation.

**Standard Deviation**: It measures the amount of variation or dispersion in a set of values. In machine learning, it can describe the spread of the data points around the mean, or the variation of the errors (residuals) in predictions.

SD **=**

where represents individual data points and is the mean of the data points.

Since standard deviation is a metric which takes all values and find the net dispersion of the values with respect to the mean, if we compare it with RMSE of our machine learning model, we can find if our model prediction is more inclined based on features or it is simply returning comparing with the mean or its predictions are actually based on features.

**R2 (R-squared):** R2 is a statistical measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is also known as the coefficient of determination.

R2 **=**

where SSres is the sum of squares of residuals, and SStot is the total sum of squares (proportional to the variance of the data).

In simpler terms it is used to “measure the goodness-of-fit”

An R2 value of

* 0 indicates that the model does show any correlation with the given features vs label
* 1 indicates that the model explains high correlation with the given features vs label

A higher R2 value indicates a better fit of the model to the data.