9/12/2024

**Sehrish Mubarik**

***Mobile Products Data Analysis and Rating Prediction Using Machine Learning***

**1. Project Overview**

This project involves analyzing and processing a dataset (mobile\_data.csv) with the goal of understanding the relationships between product features and building a machine learning model to make predictions, such as estimating mobile ratings or predicting prices.

**2. Tools and Libraries Used**

* **Pandas**: Data manipulation and preprocessing.
* **Matplotlib & Seaborn**: Visualizing data relationships and distributions.
* **Scikit-learn**: Implementing machine learning models for predictive analysis.
* **Python**: General-purpose programming for data analysis and modeling.

**3. Data Summary**

The dataset contains 828 entries with the following columns:

* **Title**: Names of mobile products.
* **Rating**: Customer ratings (target variable for modeling).
* **Reviews**: Number of customer reviews.
* **Discount Price**: Discounted price.
* **Original Price**: Original price of the product.
* **Off**: Percentage discount (categorical).

**Key Observations:**

* Missing Values:
  + Reviews: 306 missing entries.
  + Discount Price: 549 missing entries.
  + Off: 549 missing entries.
  + Original Price: 2 missing entries.

**4. Steps Performed**

**4.1 Data Loading**

* The dataset was loaded into a Pandas DataFrame, and basic exploratory commands were used to understand its structure and missing values.

**4.2 Data Cleaning**

* Handled missing values using:
  + Statistical imputation for numeric columns (mean or median).
  + Dropped non-essential columns (Discount Price and Off) with high proportions of missing data.

**4.3 Exploratory Data Analysis (EDA)**

* Visualizations of relationships between key variables, such as:
  + Ratings and Prices.
  + Original vs. Discount prices.
* Identification of trends and patterns using Seaborn plots (e.g., boxplots, scatterplots).

**4.4 Machine Learning Model**

A machine learning model was implemented to predict mobile product ratings based on the following features:

* **Features**:
  + Reviews
  + Discount Price
  + Original Price
* **Target Variable**:
  + Rating

**Model Details:**

* **Algorithm Used**: Support Vector Machine (SVM).
* **Preprocessing**:
  + Normalization or scaling of numeric features to improve model performance.
  + Encoding categorical variables if necessary.
* **Performance Metric**: Accuracy score or Mean Squared Error (MSE).

**Results:**

* The initial model achieved an accuracy score of **0.015**, which suggests poor performance.
* Further steps such as hyperparameter tuning (using GridSearchCV) and feature engineering are recommended to improve the model.

**5. Outcomes**

* The dataset was successfully cleaned and prepared for analysis.
* Exploratory visualizations provided insights into trends and distributions.
* An initial machine learning model was implemented, but its accuracy indicates a need for improvement.

**6. Recommendations**

1. **Model Optimization**:
   * Perform hyperparameter tuning for the SVM model (e.g., adjust kernel type, C, and gamma).
   * Experiment with alternative models, such as Random Forests or Gradient Boosting, for better performance.
2. **Feature Engineering**:
   * Create new features, such as price-to-rating ratio or discount percentage, to enhance predictive power.
   * Address class imbalance in the target variable, if any.
3. **Evaluation Metrics**:
   * Explore additional metrics like R² or RMSE for regression problems, as the accuracy might not fully capture model quality.
4. **Further Analysis**:
   * Investigate the influence of categorical features (e.g., Off) by one-hot encoding or ordinal encoding.