SNS GROUP (DATA SCIENCE ASSIGNMENT)

INTRODUCTION

I am **Ramya Krishnan**, a Data Scientist with a background in Mathematics and certifications in Data Science and Machine Learning. I specialize in predictive modeling, data analysis, and building machine learning models to derive valuable insights. This

project focuses on developing a machine learning model to predict **mobile phone sales prices**using data from **Flipkart's Mobile Products Dataset**. The project aims to enhance the accuracy of price predictions based on various phone attributes, such as brand, processor, RAM, and ratings.

SECTION 1: DATA ANALYSIS AND PREPROCESSING

TASK 1: DATA CLEARNING AND TRANSFORMATION

- **Loading the Dataset**:
  - The Flipkart Mobile Data is loaded into a pandas DataFrame for exploration and processing.
- **Handling Missing Data**:
  - Checked for missing values using df.isnull().sum().
  - Applied strategies like mean/median imputation for missing numerical data and mode imputation for missing categorical values.
- **Statistical Summary**:
  - Used df.describe() to get a summary of the dataset's numerical columns, such as sales price, ratings, and battery capacity.
- **Encoding Categorical Features**:
  - Applied **LabelEncoder** to categorical variables (brand, model, base_color, processor) to convert them into numerical form.

TASK 2: EXPLORATORY DATA ANALYSIS (EDA)

UNIVARIATE ANALYSIS

- **Sales Price Distribution**: Plotted the distribution of sales prices using a histogram to understand price ranges and detect outliers.
- **Brand Count**: Visualized the frequency distribution of various brands using a bar plot, showing the popularity of different mobile brands.

BIVARIATE ANALYSIS

- **Sales Price vs. Ratings**: Created scatter plots to show the relationship between sales_price and ratings, exploring how customer ratings affect pricing.
- **RAM vs. Sales Price**: Used box plots to examine the correlation between RAM sizes and sales price, analyzing price trends for different memory capacities.

MULTIVARIATE ANALYSIS

- **Pairplot**: Visualized relationships among features such as sales_price, RAM, ROM, battery_capacity, and ratings.
- **Correlation Matrix**: Generated a heatmap to visualize correlations between variables like battery_capacity, ratings, and sales_price, identifying which features strongly influence the target variable.

SECTION 2: MACHINE LEARNING MODEL DEVELOPMENT

TASK 1: MODEL DEVELOPMENT

DATA SPLITTING

- Split the data into training (80%) and testing (20%) sets using train_test_split() from scikit-learn.

MODELING TECHNIQUES

- **Linear Regression**: Built a baseline model to predict sales prices and evaluated it using metrics like **Mean Squared Error (MSE)** and **R^2**.
- **K-Nearest Neighbors (KNN)**: Implemented KNN to predict prices, tuning the hyperparameter k using the **Elbow Method** to minimize error.
- **Decision Tree Regressor**: Applied a Decision Tree model and tuned hyperparameters (max_depth, min_samples_split, min_samples_leaf) using **GridSearchCV** to prevent overfitting.
- **Random Forest Regressor**: Developed a Random Forest model, tuning n_estimators, max_depth, and min_samples_split for optimal results.
- **XGBoost Regressor**: Trained an XGBoost model, optimizing n_estimators, learning_rate, and max_depth using a grid search for enhanced performance.

EVALUATION MATRICS

- Evaluated the models based on **MSE**, **MAE**, **R^2 Score**, and **RMSE** on both the training and test data.

TASK 2: MODEL OPTIMIZATION

HYPERPARAMETER TUNING

- Performed hyperparameter tuning using **GridSearchCV** for each model to identify the best parameter settings.
- Focused on maximizing R^2 and minimizing error metrics to ensure robust performance on unseen data.

MODEL COMPARISON

Compared the performance of the models:

- **Linear Regression**: Showed baseline performance with high bias but was less complex.
- **KNN**: Performed moderately but struggled with higher-dimensional data.
- **Decision Tree**: Worked well but was prone to overfitting.
- **Random Forest**: Demonstrated better performance due to ensemble learning, handling non-linear relationships effectively.
- **XGBoost**: Achieved the best performance with the highest R^2 score and the lowest MSE, making it the optimal model for predicting sales prices.

SECTION 3: DATA VISUALIZATION AND COMMUNICATION

TASK 1: VISUALISATION DASHBOARD

CREATING THE DASHBOARD

- **Tool Used**: Tableau

- **Visualizations Included**:
    - **Bar Chart for Color-Based Discounts**: Shows discount percentages by color category, helping understand which colors are offered with higher discounts.

- **Donut Chart for Brand-Wise Sales**: Visualizes the distribution of sales across different brands, highlighting the most popular brands.
- **Bar Chart for Model-Wise Sales Price**: Displays sales prices for different mobile models, providing insight into pricing strategies for each model.
- **Tree Map for Model Details**: Illustrates overall details of various mobile models, such as sales and ratings, in a hierarchical format.
- **Side-by-Side Bars for Screen Size vs. RAM and ROM**: Compares screen sizes with RAM and ROM for different mobile phones, showing how these features vary together.
- **Pie Chart for Brand-Wise Ratings**: Represents the distribution of ratings across different brands, indicating customer satisfaction levels.

DELIVERABLE

Key Insights:

- The **bar chart for color-based discounts** reveals that certain colors receive more substantial discounts, potentially influencing consumer preferences.
- The **donut chart for brand-wise sales** indicates which brands have the highest sales, providing insights into market trends and brand performance.
- The **bar chart for model-wise sales price** highlights pricing strategies, showing which models are priced higher and potentially attracting different customer segments.
- The **tree map for model details** provides a comprehensive view of each model's performance, including sales and ratings, aiding in the identification of top-performing models.
- The **side-by-side bars for screen size vs. RAM and ROM** offer a comparison of how screen size correlates with other features, revealing trends in feature selection.
- The **pie chart for brand-wise ratings** helps assess customer satisfaction and brand reputation, guiding future product development and marketing strategies.

LINKS

DASHBOARD LINK:

https://public.tableau.com/app/profile/ramya.krishnan.a8410/viz/TechnologyProductsSalesDataSet/Dashboard1#2

GITHUB LINK:

https://github.com/Ramya19rk/SNS-Data-Science-Assignment

CONCLUSION

The **XGBoost Regressor** emerged as the most effective model for predicting mobile phone sales prices, outperforming other models in terms of accuracy and error metrics. By leveraging detailed data preprocessing, feature engineering, and hyperparameter tuning, this project highlights the power of machine learning in providing accurate and actionable insights into pricing strategies for retailers.

The interactive Tableau dashboard further enhances the project by presenting key insights in a user-friendly format. This visualization helps stakeholders understand complex data and make informed decisions based on sales trends, pricing strategies, and customer preferences.

Moving forward, additional improvements such as incorporating more features, refining the model with larger datasets, and further enhancing model interpretability will contribute to even more robust predictions. This project demonstrates the ability to effectively analyze mobile phone data and apply cutting-edge machine learning techniques to solve real-world pricing challenges.