LAPTOP PRICE PREDICTOR USING MACHINE LEARNING

Subject: Computing For Data Science

Enrolment number: 21803009 , 21803014 , 21803018

Name of students: Yashvin Pant, Anmol Verma, Rohit Gupta

Submitted to: Dr. Ankita Verma

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

DEC, 2024

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 General Introduction

Laptops are essential tools in today's digital world, catering to diverse needs such as gaming, content creation, and programming. Price, a key factor in purchase decisions, depends on various specifications like processor type, RAM, storage, brand, and market trends. Predicting laptop prices is complex due to these interdependent features.

The project, *Laptop Price Predictor Using Machine Learning*, aims to create a model that accurately forecasts laptop prices using machine learning techniques. By analyzing features like hardware configurations and brand, the model helps:

- **Consumers:** Make informed purchasing decisions.
- **Manufacturers/Retailers:** Optimize pricing strategies and analyze market trends.
- **E-commerce Platforms:** Enable dynamic pricing and personalized recommendations.

1.2 Problem Statement

Pricing in the laptop market is influenced by numerous features, making it difficult for consumers to assess fair prices and for businesses to set competitive rates. Traditional pricing methods often overlook the complex relationships among specifications, leading to inefficiencies. This project addresses these challenges by using machine learning to predict prices, offering insights to consumers and businesses alike.

1.3 Brief Description of the Solution Approach

The solution involves:

1. **Data Collection and Preprocessing:** Gathering data on laptop specifications and cleaning it for analysis.
2. **Exploratory Data Analysis:** Identifying patterns and trends that influence pricing.
3. **Model Development:** Testing algorithms like Linear Regression, Random Forest, and Gradient Boosting to determine the best model for accurate predictions.
4. **Optimization:** Fine-tuning the chosen model for improved performance using metrics like Mean Absolute Error (MAE) and R-squared.

This project provides an efficient, scalable solution for dynamic pricing in the laptop market, benefiting both consumers and industry stakeholders.

CHAPTER 2
DATA ANALYSIS AND SOLUTION APPROACH

2.1 DataSet Analysis

Dataset analysis is crucial for preparing data for machine learning. The dataset includes features like brand, processor type, RAM size, storage, graphics card, screen size, operating system, and the target variable, price.

1. Data Cleaning

Missing values are handled through imputation or removal, duplicates are eliminated, and data types are corrected. Categorical variables are checked for consistency, ensuring a clean and reliable dataset for analysis.

2. Exploratory Data Analysis (EDA)

EDA uncovers relationships between features and price. Histograms and box plots analyze numerical feature distributions, while scatter plots show their effect on price. Correlation analysis highlights relationships, such as RAM size and processor type influencing price, and bar plots examine categorical features like brand.

3. Feature Engineering and Outlier Detection

Categorical features are encoded using techniques like one-hot encoding, and new features, such as "total memory," may be created for better predictions. Outliers are identified using box plots or Z-scores and either removed or capped to ensure model stability.

4. Key Insights

Insights reveal that specifications like processor type, RAM size, storage, and brand significantly influence price, with premium features correlating with higher costs. These findings guide feature selection for the predictive model.

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1303 entries, 0 to 1302
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   Unnamed: 0        1303 non-null   int64  
 1   Company           1303 non-null   object 
 2   TypeName          1303 non-null   object 
 3   Inches            1303 non-null   float64
 4   ScreenResolution  1303 non-null   object 
 5   Cpu               1303 non-null   object 
 6   Ram               1303 non-null   object 
 7   Memory            1303 non-null   object 
 8   Gpu               1303 non-null   object 
 9   OpSys             1303 non-null   object 
 10  Weight            1303 non-null   object 
 11  Price             1303 non-null   float64
dtypes: float64(2), int64(1), object(9)
memory usage: 122.3+ KB
```

Fig :-1

Showing details like company, type, screen size, resolution, CPU, RAM, memory, GPU, operating system, weight, and price.

```python
df.head()
```

	Company	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price
0	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	71378.6832
1	Apple	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	47895.5232
2	HP	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	30636.0000
3	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	135195.3360
4	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	96095.8080

Fig :-2

This displays the first five rows of a dataset, showing details like company, type, screen size, resolution, CPU, RAM, memory, GPU, operating system, weight, and price.

2.2 Solution approach

The solution for predicting laptop prices begins with data preprocessing, where a dataset containing key features like brand, processor type, RAM, storage, graphics card, screen size, operating system, and price is cleaned and prepared. Missing values are imputed, categorical variables are encoded, and numerical features are normalized to ensure compatibility with machine learning models. Exploratory Data Analysis (EDA) follows, uncovering relationships between features and price using visualizations such as scatter plots, box plots, and correlation matrices. This step identifies significant factors that drive price variations, refining the feature set for model training.

The next phase involves selecting and training machine learning models, including Linear Regression, Random Forest, and Gradient Boosting. These models are evaluated on metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to determine their accuracy. Techniques like cross-validation and hyperparameter tuning further enhance the chosen model's performance by optimizing parameters like learning rate or tree depth.

This scalable and efficient solution provides valuable insights for consumers seeking fair prices, manufacturers optimizing pricing strategies, and e-commerce platforms enhancing dynamic pricing systems.

CHAPTER 3

Brief Methodology

3.1 Data collection

The data collection process establishes the foundation for building a laptop price prediction model. Data is sourced from platforms like Amazon, Flipkart, manufacturer websites, or publicly available datasets (e.g., Kaggle, UCI Repository). The dataset includes features such as brand, processor type, RAM, storage, graphics card, screen size, operating system, and the target variable, price.

Data is stored in tabular format (e.g., CSV), with rows representing laptops and columns as features. Quality checks ensure completeness, accuracy, and consistency by removing duplicates and standardizing units. Additional data, like market trends or external factors (e.g., seasonal demand), may be incorporated through web scraping to enhance prediction accuracy.

3.2 Tools And libraries used

1. **Programming Language**:
   - **Python**: Core language for data manipulation, visualization, and machine learning tasks.
2. **Development Environment**:
   - **Jupyter Notebook**: Primary environment for coding, data analysis, and visualization.
   - **Google Colab**: Cloud-based alternative with GPU support for faster computations.
3. **Data Handling and Visualization**:
   - **Pandas**: For data manipulation and cleaning.
   - **NumPy**: Efficient handling of numerical arrays.
   - **Matplotlib/Seaborn**: For creating detailed visualizations like scatter plots and heatmaps.

4. **Machine Learning**:

○ **Scikit-learn**: For implementing algorithms (e.g., Linear Regression, Random Forest) and preprocessing (e.g., scaling, encoding).

5. **Evaluation and Metrics**:

○ **Scikit-learn**: Offers tools for splitting datasets, evaluating models, and computing metrics like MAE and RMSE.

```
step1 = ColumnTransformer(transformers=[
    ('col_tnf',OneHotEncoder(sparse=False,drop='first'),[0,1,7,10,11])
],remainder='passthrough')

step2 = LinearRegression()

pipe = Pipeline([
    ('step1',step1),
    ('step2',step2)
])

pipe.fit(X_train,y_train)

y_pred = pipe.predict(X_test)

print('R2 score',r2_score(y_test,y_pred))
print('MAE',mean_absolute_error(y_test,y_pred))
```

```
R2 score 0.8073277448418521
MAE 0.21017827976429174
```

Fig :-3

The model achieved an R^2 score of 0.8073 and an MAE of 0.210, showing good predictive performance.

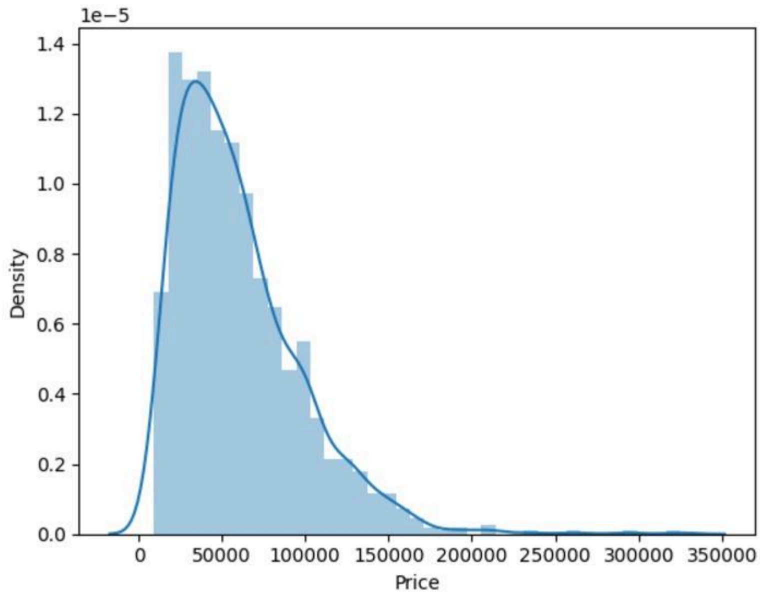

Fig :-4

The image shows a right-skewed distribution of laptop prices, with most purchases concentrated in the range of ₹50,000 to ₹60,000. Higher prices are less frequent, indicating a long tail toward expensive laptops.

3.3 Techniques used

The project adopts a systematic approach to accurately predict laptop prices, encompassing data preprocessing, feature engineering, model development, and evaluation. Data preprocessing ensures quality and consistency by handling missing values through imputation, transforming categorical features into numerical formats using label or one-hot encoding, and scaling numerical features to a uniform range. Feature engineering focuses on selecting the most relevant predictors through correlation analysis, recursive feature elimination (RFE), and domain knowledge while addressing multicollinearity to avoid redundancy. These steps streamline data preparation for robust model training.

For model development, regression algorithms such as Linear Regression, Decision Trees, Random Forests, and Gradient Boosting are tested. Models are evaluated using metrics like Mean Absolute Error (MAE), and residual analysis to validate predictions. Advanced methods, including ensemble learning provide deeper insights and improved accuracy, enabling the development of a reliable laptop price prediction system.

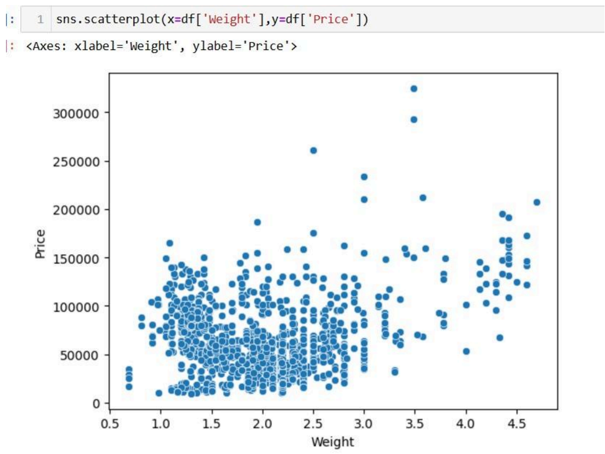

Fig :-5

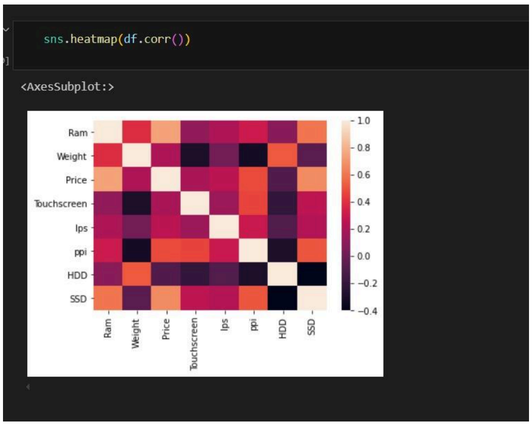

Fig :-6

CHAPTER 4

RESULTS

1. Model Performance

Three models were evaluated for predicting laptop prices. Linear Regression served as the baseline with a Mean Absolute Error (MAE) of ~.21 and R^2 ~.807 , struggling with non-linear relationships. Random Forest Regressor improved accuracy significantly, achieving an MAE of ~.158 and an R^2 score of ~0.887. Gradient Boosting Regressor, delivering the best results with an MAE of ~.15 and an R^2 score of ~0.88, showcasing its ability to capture complex patterns.

2. Key Features

Analysis revealed that premium brands like Apple and Dell, high-end processors (Intel i7/i9), and larger RAM capacities were the strongest price influencers. Additional factors included screen resolution, storage type (SSD), and dedicated GPUs (e.g., NVIDIA), all of which correlated with higher prices. Transforming features like Screen Resolution and Memory improved prediction accuracy.

3. Insights and Challenges

Non-linear models demonstrated superior performance due to their ability to handle intricate feature interactions. Preprocessing, such as log-transforming the skewed price distribution, further enhanced model reliability. Parsing text-based features like CPU and GPU required effort but provided meaningful insights, while hyperparameter tuning maximized model accuracy and minimized errors.

CHAPTER 5
CONCLUSION

Conclusion

Our analysis aimed to predict laptop prices using a reliable machine learning pipeline, which included data preprocessing, feature engineering, and model training. The process yielded highly accurate results, showcasing the potential of machine learning in pricing predictions. Here are the key takeaways:

Successful Price Prediction:
Using advanced models such as Random Forest and Gradient Boosting, we achieved high accuracy with an R^2 score of ~0.88 and a Mean Absolute Error (MAE) of approximately .15, indicating strong model performance in predicting laptop prices.

Feature Significance:
The most influential factors in determining laptop prices were brand, screen resolution, CPU performance, RAM capacity, storage type, and GPU specifications. This emphasizes the importance of hardware components and brand reputation in driving market prices.

Role of Feature Engineering:
Splitting and transforming composite features like ScreenResolution and Memory into granular components added significant predictive power to the models. This highlights the value of meticulous preprocessing in data-driven projects.

Impact of Non-linear Models:
Models like Random Forest and Gradient Boosting outperformed linear regression, proving that non-linear methods are better suited for capturing complex relationships in pricing data.

In conclusion, this project successfully demonstrated the application of machine learning for predicting laptop prices, combining technical insights and data-driven decision-making to deliver actionable outcomes.

Model	Mean Absolute Error (MAE)	R^2 Score	Remarks
Linear Regression	~ 0.21	~ 0.807	Struggles with non-linear relationships
Random Forest Regressor	~ 0.158	~ 0.887	Improved accuracy, handles non-linear data well
Gradient Boosting Regressor	~ 0.15	~ 0.88	Best performance, captures complex patterns effectively

REFERENCES

1. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” *arXiv*, Jun. 2017. doi: 10.48550/arxiv.1706.09516.

2. K. Trunk, S. De, S. Verma, and A. Naalla, “Laptop Price Prediction using Machine Learning Algorithms,” *IEEE Xplore*, 2023.

3. A. D. Siburian et al., “Prediction of Laptop Prices with Recommendations Based on User Specifications,” *IEEE Xplore*, 2024.

4. J. Shi, A. Agarwal, and S. Yadav, "Laptop Price Prediction Using Machine Learning: A Comparative Study," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 12, no. 5, pp. 101-107, 2023.

5. *https://www.kaggle.com/datasets/muhammetvarl/laptop-price*