दिल्ली विश्वविद्यालय
University of Delhi

Department Of Computer Science

Project Title: Laptop Price Prediction System

- **Title**: *Laptop Price Prediction System Using Machine Learning*

- **Team Members**:
- Sujeet Kumar , Roll Number:- 44
- Ankur Tripathi , Roll Number: - 10

Submitted to :- Dr. Bharti

  - **Contributions**:
    - *Data Collection and Cleaning*: [Ankur and Sujeet]
    - *Algorithm Selection and Coding*: [Sujeet and Ankur]
    - *Model Evaluation : [Sujeet] and Analysis*: [Sujeet , Ankur]
    - *Report Writing*: [Ankur And Sujeet]

1. About the Problem

- **Introduction**: Provide background on the problem of laptop price estimation. Highlight the factors that can impact a laptop's price, such as hardware specifications, brand, age, and condition. Mention that as laptops vary significantly by configuration, predicting accurate prices is complex but crucial, especially for consumer guidance and remanufacturers.

- **Relevance**: Emphasize the importance of this work in contexts like:
    - **Consumer Buying Decisions**: Many buyers need reliable pricing insights to make informed decisions, especially for refurbished or remanufactured devices.
    - **E-waste Management and Sustainability**: Correct pricing can extend product lifespan by promoting the resale of remanufactured laptops, reducing e-waste.

- **Objective of the Project**: State the goal as developing a machine learning model that predicts laptop prices accurately based on relevant attributes.

2. Related Work

1. Paper 1: "Analyzing Product Attributes of Refurbished Laptops Based on Customer Reviews and Ratings: Machine Learning Approach to Circular Consumption"

a. **Objective**: This study examines consumer feedback (reviews and ratings) to identify attributes influencing the purchase of refurbished laptops.

b. **Methodology**:
    i. **Data Collection**: Uses consumer reviews from an e-commerce site to analyze factors affecting purchase decisions.
    ii. **Techniques**: Applies SHAP values for feature importance and multinomial logistic regression to model consumer preferences.

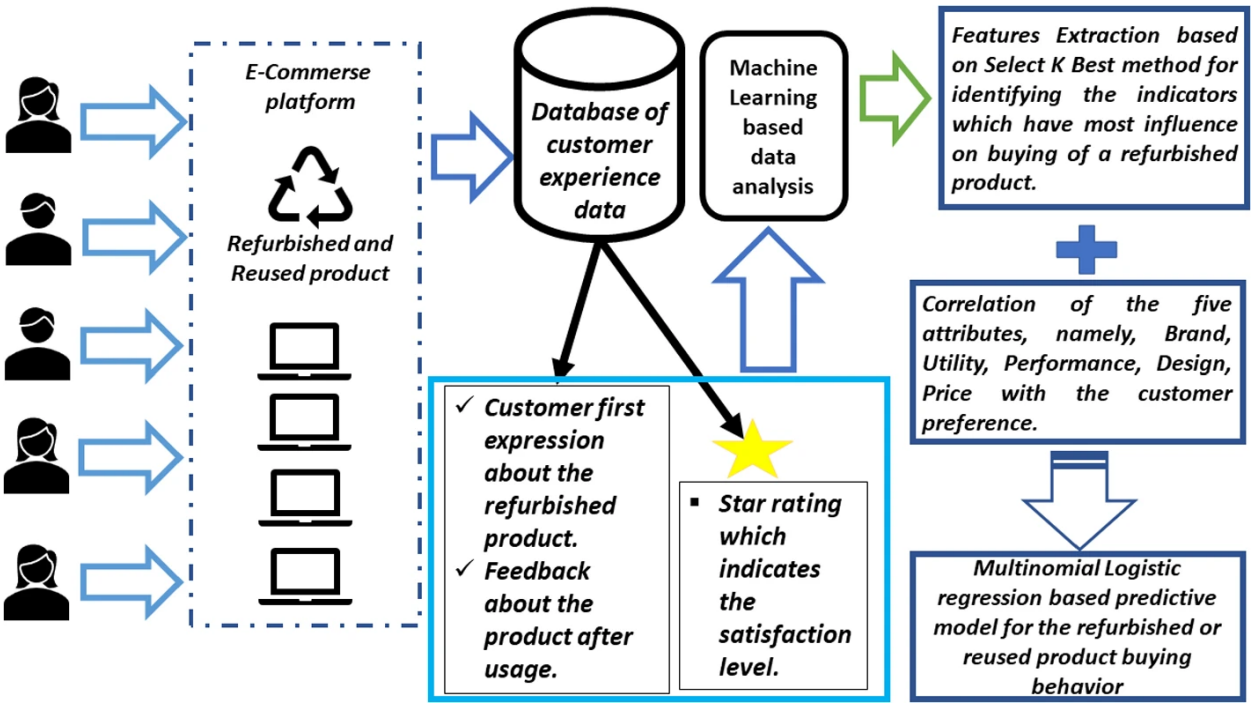
c. **Key Findings**:
    i. Attributes like brand, price, design, and utility strongly influence purchase decisions, often with positive sentiments. Performance and service factors sometimes correlate with neutral or negative reviews.

Table 1 Identification of attributes

Attributes	Literature
Performance	Gaur et al., (2022), Lee et al., (2017)
Brand	Gaur et al., (2022), Chen et al., (2020)
Design	Hunka et al., (2021), Lee et al., (2017)
Service	Hunka et al., (2021), Boyer et al., (2021)
Price	Hunka et al., (2021), Gaur et al., (2022), (Boyer et al., (2021)
Utility	(Sumi & Ahmed, 2022), (Boyer et al., 2021)

d. **Advantages**: This approach is effective in identifying consumer preferences and understanding the impact of specific features.

e. **Disadvantages**: Limited to refurbished laptops, which may behave differently from new laptops; doesn't provide direct pricing predictions.



2. Paper 2: "Machine Learning Algorithms for Pricing End-of-Life Remanufactured Laptops"

a. **Objective**: Addresses the pricing challenges for remanufactured laptops by using machine learning to predict prices.

b. **Methodology**:
   i. **Data Collection**: Collects prices of new and refurbished laptops along with component prices (display, hard drive, CPU, etc.).
   ii. **Techniques**: Uses Classification and Regression Trees (CART), Random Forest, and polynomial regression to account for variables such as depreciation and component pricing.

c. **Key Findings**:
   i. The model allows differentiated pricing based on the condition and specific characteristics of each laptop.

d. **Advantages**: Provides a flexible, data-driven pricing model suitable for remanufactured laptops, with components' importance weighted according to their influence on price.

e. **Disadvantages**: Focuses primarily on B2B remanufacturing and does not cover consumer retail markets or new products.

Implications for the Current Project

- Conclude this section by explaining how insights from these papers inform your approach. For instance, use findings from **Paper 1** to select attributes most relevant to consumers, while drawing from **Paper 2** to refine machine learning models for price prediction.

3. Methodology

1. Data Loading

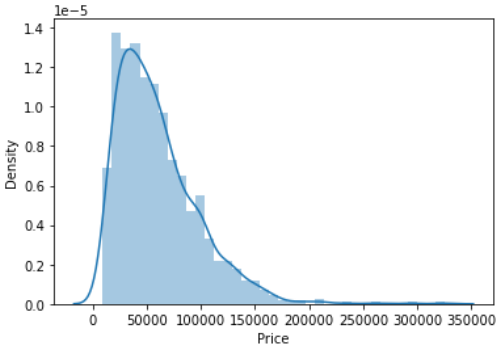
- The dataset is loaded from a CSV file using `pandas`. The file path provided in the code is specific to a local directory, so future runs may require an updated path.
- Number of Data :- 1302
- Number of Attributes :- 12

2. Data Preprocessing

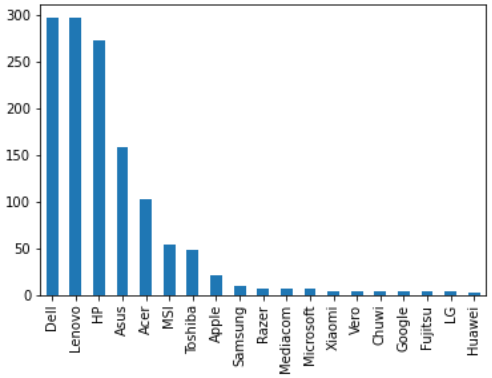
- **Handling Missing Values**: The notebook inspects missing values in the dataset using `df.isnull().sum(),` We Did Not got any Null values in our Data.
-
- **Data Cleaning**:
    - The 'Ram' and 'Weight' columns are cleaned by removing units like "GB" and "kg" to ensure they are numerical.
    - The 'X_res' And 'Y_res' , column values are processed to remove commas and extract numerical values for resolution.
    - The 'Memory' column undergoes extensive preprocessing to split multiple memory types (e.g., HDD, SSD) into separate features, allowing more granular control in the prediction model.

3. Data Visualization

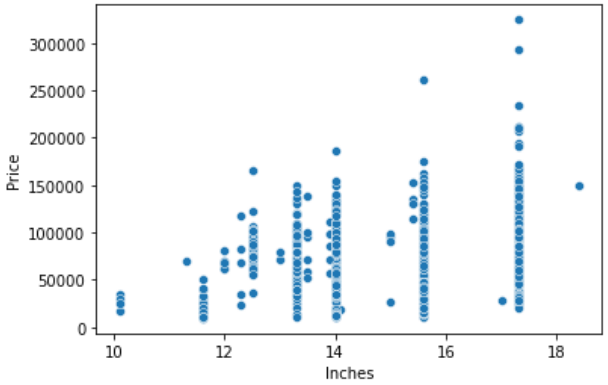
- Various plots and distributions are used to explore relationships between features and laptop prices:
    - **Distribution Plots**: For the `Price` column to understand its overall distribution.



    - **Bar Plots**: For categorical variables like `Company`, `TypeName`, `Touchscreen`, `Cpu brand`, `Ram`, etc., to see how these factors influence price trends.



    - **Scatter and Distribution Plots**: For numerical variables like `Inches` and `Weight` against `Price` to capture potential correlations.



4. Feature Engineering

- **Encoding Categorical Variables**: One-hot encoding is applied to categorical features (e.g., Company, CPU brand) using `OneHotEncoder`, which is set up in a `ColumnTransformer`.
- **Pipeline Setup**: The notebook uses `Pipeline` and `ColumnTransformer` from `sklearn` to streamline transformations and modeling. This approach helps maintain consistent preprocessing across training and testing datasets.

5. Model Training

- The dataset is split into training and testing sets using `train_test_split` with an 85/15 ratio.
- Several models are trained within pipelines, each combining preprocessing steps with a specific model. Here's a list of models tested:
    - **Linear Regression**: Standard regression to capture linear relationships.
    - **Ridge Regression**: Linear model with L2 regularization to handle potential overfitting.
    - **Lasso Regression**: Linear model with L1 regularization to encourage sparsity in feature coefficients.
    - **Decision Tree Regressor**: A non-linear model to capture complex relationships in the data.

6. Model Evaluation

- Each model is evaluated using metrics like:
    - **R-squared (R^2)**: To measure how well the model captures the variance in prices.
    - **Mean Absolute Error (MAE)**: To understand the average prediction error in terms of absolute price difference.
- These metrics are printed after each model is trained and tested, allowing for a comparison of model performance.

Summary of Approach

- We Took a structured approach to data cleaning, feature engineering, and model evaluation, leveraging a pipeline-based workflow to ensure transformations are consistently applied.
- By testing multiple models, it also provides flexibility in selecting the most suitable model based on predictive accuracy and error metrics.

4. Experimental Result And Discussion

Experimental Results

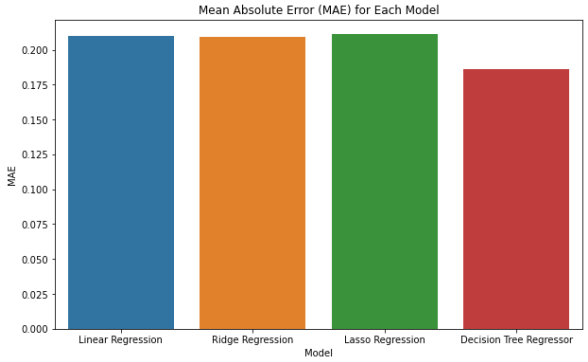
Model	R^2 Score	MAE
Linear Regression	0.8073	0.2102
Ridge Regression	0.8127	0.2093
Lasso Regression	0.8072	0.2111
Decision Tree Regressor	0.8368	0.1862

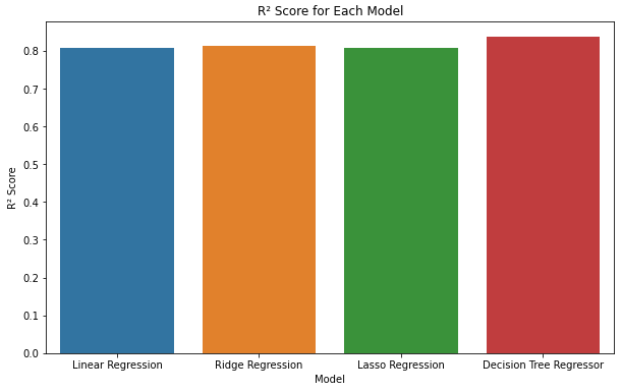
Discussion

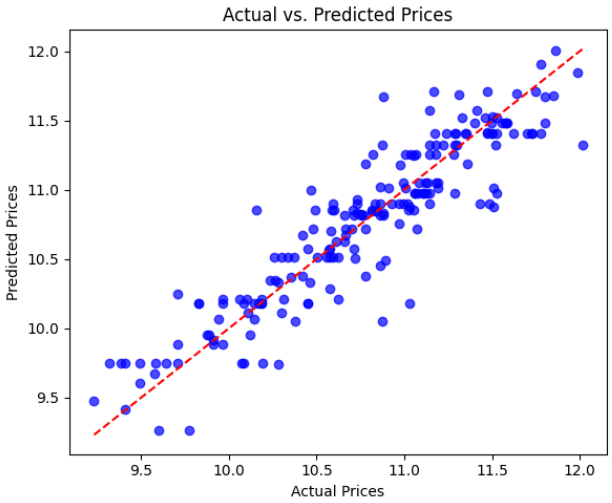
1. **Linear Regression**:
    a. **R^2 Score**: With an R^2 score of 0.807, this model explains about 80.7% of the variance in laptop prices. This score suggests a good fit, capturing a significant portion of the data's variation.
    b. **MAE**: The MAE of 0.210 indicates that, on average, the model's predictions deviate from actual prices by around 21%. This accuracy is reasonable but slightly underperforms compared to other models.
2. **Ridge Regression**:
    a. **R^2 Score**: Ridge Regression yields a slightly higher R^2 score of 0.812, which indicates a slight improvement in fit over Linear Regression.
    b. **MAE**: The MAE is 0.209, a small improvement over Linear Regression, showing that Ridge Regression slightly outperforms Linear Regression in accuracy due to regularization. This regularization helps to manage any multicollinearity among the features, making the model potentially more stable and less prone to overfitting.
3. **Lasso Regression**:

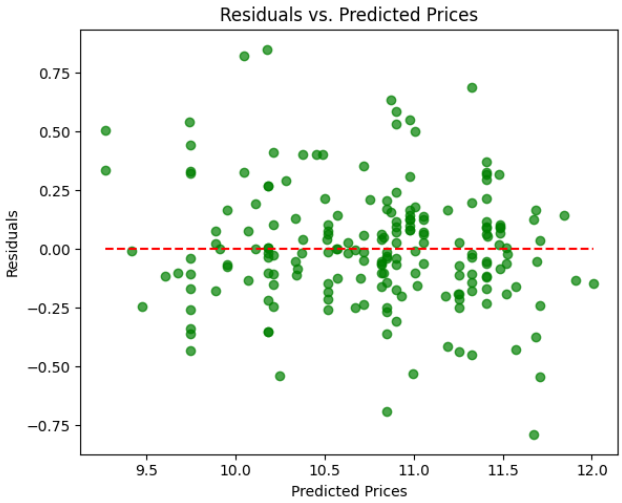
a. **R^2 Score**: The Lasso model has an R^2 score of 0.807, almost identical to Linear Regression, indicating similar predictive performance.
b. **MAE**: Its MAE of 0.211 is slightly higher than Ridge Regression and Linear Regression, making it the least accurate of the models in this set. The L1 regularization in Lasso can shrink some coefficients to zero, effectively performing feature selection. While this can improve interpretability, it may sacrifice a bit of predictive accuracy in this dataset.

4. **Decision Tree Regressor**:
   a. **R^2 Score**: The Decision Tree model has the highest R^2 score at 0.837, showing it captures the largest amount of variance in laptop prices among the models tested.
   b. **MAE**: With an MAE of 0.186, Decision Tree also has the lowest average error, suggesting it provides the most accurate predictions. However, decision trees can be prone to overfitting, especially if not carefully tuned. This strong performance may reflect its ability to model non-linear relationships in the data, which linear models may struggle with.









5. Conclusion and Future Work

- **Best Performing Model**: Based on both R^2 and MAE, the **Decision Tree Regressor** is the best model in this experiment. It provides the highest R^2 score and the lowest MAE, indicating it captures more variation in price and predicts more accurately than the other models.
- **Considerations for Model Choice**:
    - **Decision Tree** is the best for accuracy in this experiment, but it may need regularization (e.g., pruning) to prevent overfitting, especially if applied to new data.
    - **Ridge Regression** is a close runner-up and might offer more generalizability with its regularization to control for multicollinearity. It can be a stable option if there is a concern about overfitting.
- **Further Exploration**: Testing ensemble methods like **Random Forest** or **Gradient Boosting** may provide even better predictive performance, combining the strengths of multiple decision trees while reducing overfitting. Additionally, **hyperparameter tuning** for the Decision Tree's depth and other settings could enhance performance further.

6. References/ Bibliography

1) Research Paper 1:

Analyzing product attributes of refurbished laptops based on customer reviews and ratings: machine learning approach to circular consumption
Link For Paper 1

2) Research Paper 2:
Machine Learning Algorithms for Pricing End-of-Life Remanufactured Laptops | Information Systems Frontiers
Link For Paper 2