Project Report

on

Laptop Price Prediction using various regression models

Submitted for Internship/ Industrial training requirements

of

B.Tech and M.Tech

in

Mathematics and Data Science

Submitted by

Arindam Phatowali

Scholar No. 214104021

Under the guidance of

Dr. Amit Bhagat

Department of Mathematics, Bioinformatics and Computer Applications

Maulana Azad

National Institute of Technology Bhopal - 462003 (India)

MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY, BHOPAL (M.P)-462003

CERTIFICATE

This is to certify that Arindam Phatowali (Sch no- 214104021), a student of Dual Degree (B.Tech+M.Tech) of batch 2021 -2026 has completed the project titled "Laptop Price Predictor" being submitted in the partial fulfilment of the requirement of the completion of Dual Degree in Mathematics and Data Science to Maulana Azad National Institute of technology Bhopal under my supervision.

Date: 10th November 2023
Place: Bhopal

Dr. Amit Bhagat
Assistant Professor
(Supervisor)

INDEX

Abstract

The "Laptop Price Prediction" project aims to predict the price of laptops using regression analysis. The project involves the use of machine learning techniques to analyse a dataset of laptop specifications and their corresponding prices, then a model is built that can estimate the price of a laptop based on its specifications. The process involves data collection, preprocessing, exploratory data analysis, feature selection, model building, and evaluation. The final model can be used by both buyers and sellers to estimate laptop prices, providing a valuable tool in the marketplace. Future work could involve collecting more data to improve the accuracy of the model or exploring other machine learning techniques for price prediction.

Acknowledgement

I take this opportunity to express my sincere gratitude and respect to MANIT Bhopal for providing me a platform to pursue my studies and carry out 3rd year summer project.

I have a great pleasure in expressing my deep sense of gratitude to Dr Amit Bhagat Sir for their constant support and encouragement throughout the course of this project.

I also extend my thanks to all the faculty of MDS who directly or indirectly encouraged me.

Finally, I would like to thank my friends for all their moral support they have given me during the completion of this work.

INTRODUCTION

1.Relevance of the Project:

The “Laptop Price Prediction” project is crucial as it aids in market analysis, benefits consumers, predicts trends, assists in inventory management, and facilitates competitive analysis. It enables businesses to understand pricing influences and make informed decisions. Consumers can estimate costs for specific laptop features, aiding in purchase decisions. The model can identify pricing trends from historical data, useful for consumers and businesses. Retailers can predict laptop prices for inventory management and sales forecasting. Lastly, businesses can compare their pricing with the market for strategic adjustments.

2. Problem Statement:

Our project aims to predict laptop prices based on user input. The problem we are addressing is that there are many different combinations of configurations that can be done, so if people want to buy a new laptop, then our model should have all the prices sorted by their configuration. While it may seem like a straightforward model development project, the complexity lies in the noisy dataset that requires extensive feature engineering and pre-processing. This complexity adds an intriguing layer to the project, making it a more engaging and challenging endeavour.

3.Objective of the Project:

- Data Analysis: To analyse a dataset of laptop specifications and their corresponding prices, understanding the relationship between different features and how they influence the price.

- Model Development: To develop a regression model that can accurately predict the price of a laptop based on its specifications.

- Model Evaluation: To evaluate the performance of the model using appropriate metrics and ensure it can make accurate predictions.

- Business Application: To provide a tool that can be used by businesses for pricing their products, and by consumers to make informed purchasing decisions.

4.Scope of the Project

The aim of this project is to predict the price of laptops using regression analysis. The project involves the use of machine learning techniques to analyse a dataset of laptop specifications and their corresponding prices, and then build a model that can predict the price of a laptop based on its specifications. Further, future enhancement can be done such as incorporating more data or exploring other machine learning techniques to improve the accuracy of the price prediction.

Process

The project aims to predict laptop prices based on various features using regression models. The problem statement is that if any user wants to buy a laptop, then our application should be compatible to provide a tentative price of laptop according to the user configurations.

The process of the project involves 6 major steps:

1. Data Collection:

The data for this project is collected from Kaggle (link- https://www.kaggle.com/datasets/mohidabdulrehman/laptop-price-dataset). The dataset includes features such as the brand, screen size, processor type, RAM, storage type and size, graphics card, and more. Most of the columns in a dataset are noisy and contain lots of information. But with feature engineering, we will get more good results. The only problem is we are having less data, but we will obtain a good accuracy over it.

2. Data Preprocessing:

Import the necessary libraries like NumPy, pandas, matplotlib and seaborn and load the data. After that we will take a basic understanding of data like its shape, sample, is there are any NULL values present in the dataset. Understanding the data is an important step for prediction or any machine learning project. The collected data is pre-processed to handle missing values, outliers, and categorical variables. For instance, changes were made in weight and Ram column to convert them to numeric by removing the unit written after value. We need a few changes in weight and Ram column to convert them to numeric by removing the unit written after value.

3. Exploratory Data Analysis (EDA) and Feature Engineering: EDA is performed to understand the data better. This includes visualizing the distribution of different features and their relationship with the laptop price. In other words, it

helps to perform hypothesis testing. Now, Feature engineering is a process to convert raw data to meaningful information. Feature engineering involves creation of new features based on existing ones to improve the model's performance. Feature selection techniques are used to select the most relevant features.

We will start from the first column and explore each column and understand what impact it creates on the price column. At the required step, we will also perform preprocessing and feature engineering tasks. Our aim in performing in-depth EDA is to prepare and clean data for better machine learning modelling to achieve high performance and generalized models. So, let's get started with analysing and preparing the dataset for prediction.

1. Distribution of target column:

   Working with regression problem statement, the target column distribution is important to understand. Here, our target column is the price column. By our observation we see that the distribution of the target variable is skewed, and it is obvious that commodities with low prices are sold and purchased more than the branded ones.

2. Company column:

   We want to understand how does brand name impacts the laptop price or what is the average price of each laptop brand? If we plot a frequency plot of a company then the major categories present are Lenovo, Dell, HP, Asus, etc. Now if we plot the company relationship with price then we can observe that how price varies with different brands. Razer, Apple, LG, Microsoft, Google, MSI laptops are expensive, and others are in the budget range.

3. Type of laptop:

Which type of laptop we are looking for like a gaming laptop, Ultrabook, or 2 in 1 convertible. As major people prefer notebook because it is under budget range and the same can be concluded from our data.

4. Does the price vary with laptop size in inches?

⇨ A Scatter plot is used when both the columns are numerical, and it answers our question in a better way. From the below plot we can conclude that there is a relationship but not a strong relationship between the price and size column.

5. Screen Resolution:

Screen resolution contains lots of information. Before any analysis, first we need to perform feature engineering over it. If we observe unique values of the column then we can see that all value gives information related to the presence of an IPS panel, touch screen is present or not, and the X-axis and Y-axis screen resolution. So, we will extract the column into 3 new columns in the dataset.

a. Extract Touch screen information:

   It is a binary variable so we can encode it as 0 and 1. 1 means the laptop is a touch screen and 0 indicates not a touch screen. If we plot the touch screen column against price, then laptops with touch screens are expensive which is true in real life.

b. Extract IPS Channel presence information:

   It is a binary variable, so it can also be encoded as 0 and 1. The laptops with IPS channel are present less in our data but by observing relationship against the price of IPS channel laptops are high.

c. Extract X-axis and Y-axis screen resolution dimensions:

   Now both the dimensions are present at end of a string and separated with a cross sign. So first we will split the string with space and access

the last string from the list. then split the string with a cross sign and access the zero and first index for X and Y-axis dimensions.

d. Replacing inches, X and Y resolution to PPI:

If we find the correlation of columns with price using the corr function, then we can see that inches do not have a strong correlation but X and Y-axis resolution have a very strong resolution so we can take advantage of it and convert these three columns to a single column that is known as Pixels Per Inch (PPI). In the end, our goal is to improve the performance by having fewer features. Now when we will see the correlation of price then PPI is having a strong correlation. So now we can drop the extra columns which are not of use. At this point, we have started keeping the important columns in our dataset.

6. CPU column:

If we observe the CPU column then it also contains lots of information. If we again use a unique function or value counts function on the CPU column, then we have 118 different categories. The information it gives is about preprocessors in laptops and speed.

To extract the preprocessor, we need to extract the first three words from the string. We are having an Intel preprocessor and AMD preprocessor, so we are keeping 5 categories in our dataset as i3, i5, i7, other intel processors, and AMD processors.

How does the price vary with processors? => We can again use our bar plot property to answer this question. And as obvious the price of i7 processor is high, then of i5 processor, i3 and AMD processor lies at the almost the same range. Hence price will depend on the preprocessor.

7. Price with Ram:

Again, bivariate analysis of price with Ram. If we observe the plot, then Price is having a very strong positive correlation with Ram or we can say a linear relationship.

8. Memory column:

Memory column is again a noisy column that gives an understanding of hard drives. Many laptops come with HHD and SSD both. In some there is an external slot present to insert after purchase. This column can disturb our analysis if not feature engineered properly. So, if we use value counts on a column then we are having 4 different categories of memory as HHD, SSD, Flash storage, and hybrid.

First, we have cleaned the memory column and then made 4 new columns which are a binary column where each column contains 1 and 0 indicate that amount four is present and which is not present. Any laptop has a single type of memory or a combination of two. So, in the first column, it consists of the first memory size and if the second slot is present in the laptop then the second column contains it else, we fill the null values with zero. After that in a particular column, we have multiplied the values by their binary value. It means that if in any laptop particular memory is present then it contains binary value as one and the first value will be multiplied by it, and same with the second combination. For the laptop which does have a second slot, the value will be zero multiplied by zero is zero.

Now when we see the correlation of price then Hybrid and flash storage have very less or no correlation with a price. We will drop this column with CPU and memory which is no longer required.

9. GPU Variable:

GPU (Graphical Processing Unit) has many categories in data. We are having which brand graphic card is there on a laptop. We are not having how many

capacities like (6Gb, 12 Gb) graphic card is present. So, we will simply extract the name of the brand.

If we use the value count function then there is a row with GPU of ARM so we have removed that row and after extracting the brand GPU column is no longer needed.

10. Operating System Column:

There are many categories of operating systems. We will keep all windows categories in one, Mac in one, and remaining in others. This is a simple and most used feature engineering method, we can even try something else if we find more correlation with price.

When we plot price against operating system then as usual Mac is most expensive.

11.Log-Normal Transformation:

We saw the distribution of the target variable above which was right-skewed. By transforming it to normal distribution performance of the algorithm will increase. We take the log of values that transform to the normal distribution which we can observe below. So, while separating dependent and independent variables we will take a log of price, and in displaying the result perform exponent of it.

4. Model Building:

Now we have prepared our data and hold a better understanding of the dataset. So let's get started with machine learning modelling and find the best algorithm by training various regression models like linear regression, random forest, xgboost etc on the pre-processed data to achieve maximum accuracy.

We will import all the necessary libraries first before proceeding any further. Our libraries include train_test_split, ColumnTransformer, Pipeline, OneHotEncoder, r2_score,mean_absolute_error, LinearRegression, KNeighborsRegressor,

DecisionTreeRegressor, RandomForestRegressor and XGBRegressor all from sklearn.

Split in train and test:

X = data.drop(columns=['Price'])

y = np.log(data['Price'])

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.15, random_state = 2)

Implement Pipeline for training and testing: Now we will implement a pipeline to streamline the training and testing process. First, we use a column transformer to encode categorical variables which is step one. After that, we create an object of our algorithm and pass both steps to the pipeline. Using pipeline objects, we predict the score on new data and display the accuracy. For example-

```
# Preprocessing
step1 = ColumnTransformer(transformers=[
    ('col_tnf',OneHotEncoder(sparse=False,drop='first'),[0,1,7,10,11])
],remainder='passthrough')
# Xgboost regressor
step2 = XGBRegressor(n_estimators=45,max_depth=5,learning_rate=0.5)
# Create a pipeline
xgb = Pipeline([
    ('step1',step1),
    ('step2',step2)
])
# Fit and predict
xgb.fit(X_train,y_train)

y_pred = xgb.predict(X_test)
# Print scores
print('R2 score',r2_score(y_test,y_pred))
print('MAE',mean_absolute_error(y_test,y_pred))
```

```
R2 score 0.8890315532690259
MAE 0.16094995656431285
```

5. Model Evaluation: The performance of the models is evaluated using metrics such as Mean Absolute Error (MAE) and R-squared.

6. Model Selection: We got the highest accuracy of 88.9% with XGBoost regressor. Hence, we select XGBoost regressor as our final model for prediction.

Result:

```
# Initialize an empty list to store user input
user_input = []
# Ask the user for each feature
user_input.append(input("Company (e.g., Asus/Lenovo/Dell/HP etc): "))
user_input.append(input("TypeName (e.g., Ultrabook/Gaming/Notebook/2 in 1 Convertible): "))
user_input.append(float(input("RAM (e.g., 8): ")))
user_input.append(float(input("Weight (e.g., 1.37): ")))
user_input.append(int(input("Touchscreen (0 for No, 1 for Yes): ")))
user_input.append(int(input("IPS (0 for No, 1 for Yes): ")))
user_input.append(float(input("PPI (e.g., 226.983): ")))
user_input.append(input("CPU (e.g., Intel Core i3/i5/i7): "))
user_input.append(int(input("HDD (e.g., 0/500/1000): ")))
user_input.append(int(input("SSD (e.g., 128/256/500): ")))
user_input.append(input("GPU (e.g., Intel/AMD/Nvidia): "))
user_input.append(input("Operating System (e.g., Windows/Linux/Mac etc): "))

# Convert the user input list to a numpy array
user_input_array = np.array([user_input], dtype=object)

# Use the fitted voting_regressor pipeline to predict the price for the user's laptop features
predicted_price = xgb.predict(user_input_array)

# Print the predicted price
print("Predicted Price:", (np.e)**predicted_price[0])
```

```
Company (e.g., Asus/Lenovo/Dell/HP etc): Asus
TypeName (e.g., Ultrabook/Gaming/Notebook/2 in 1 Convertible): Ultrabook
RAM (e.g., 8): 16
Weight (e.g., 1.37): 1.4
Touchscreen (0 for No, 1 for Yes): 0
IPS (0 for No, 1 for Yes): 1
PPI (e.g., 226.983): 160
CPU (e.g., Intel Core i3/i5/i7): Intel Core i5
HDD (e.g., 0/500/1000): 0
SSD (e.g., 128/256/500): 512
GPU (e.g., Intel/AMD/Nvidia): Nvidia
Operating System (e.g., Windows/Linux/Mac etc): Windows
Predicted Price: 69349.71687203702
```

Project Link- https://github.com/arinp95/Laptop_Price_Prediction

Conclusion

The project used a dataset of laptop features and prices collected from Kaggle.com and pre-processed it to handle missing values, outliers, and categorical variables. Exploratory data analysis (EDA) is performed to understand the relationship between different features and the target variable and used feature selection techniques to identify the most relevant features for price prediction. Then we trained and evaluated several regression models, and then used the voting regressor which offered 88% prediction precision. Thus, the project provides valuable insights into the factors that influence laptop prices and can be used to make informed decisions when buying a laptop.

References:

1. Laptop Price Dataset | Kaggle
2. https://numpy.org/doc/stable/user/index.html#user
3. https://pandas.pydata.org/docs/user_guide/index.html
4. 3.12.0 Documentation (python.org)
5. scikit-learn: machine learning in Python — scikit-learn 1.3.1 documentation