Predicting the laptop price in the world markets

1. Abstract

Machine learning is a subfield of artificial intelligence that deals with developing applications that can predict the future based on past data. If you are a data science enthusiast or practitioner, this article will assist you in creating your own end-to-end machine learning project from the ground up. We will create a project for predicting laptop prices. The problem statement is that if a user wants to buy a laptop, our application should be able to provide a rough price based on the user's configurations. Although it appears to be a simple project or simply developing a model, the dataset we have is noisy and requires a significant amount of feature engineering and preprocessing, which will pique your interest in developing this project. The purpose of this project was to use classification models to predicting laptop prices. This project uses machine learning methods to predict laptop price. Beside ML predictions, focus of this project is put on architecture and deployment. 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. Although it looks like a simple project or just developing a model, the dataset we have is noisy and needs lots of feature engineering, and preprocessing that will drive your interest in developing this project. The majority of the columns in a dataset used in our research is noisy and contains a lot of information. However, the more feature engineering we

do, the better the results will be. The only issue is that we have less data, but we will achieve a high level of accuracy. The only advantage is that having a large amount of data is preferable. We will create a project that will estimate the approximate cost of a laptop.

2. Design

This problem a design is Laptop Price, now days in markets it is real world problem, It's depended a lots of features Define Project Goal. ...

1. Determine Outcomes, Objectives, and/or Deliverables. ...
2. Identify Risks, Constraints, and Assumptions. ...
3. Prepare a Visual Aid. ...
4. Ballpark Your Budget. ...
5. Determine Approval and Monitoring Processes. ...
6. Use Proper Project Design Documents.

3. Data

If you're working on a personal project or playing around with a dataset or an API, this step may seem irrelevant. It's not. Simply downloading a cool open dataset is not enough. In order to have motivation, direction, and purpose, you have to identify a clear objective of what you want to do with data: a concrete question to answer, a product to build,

- **Kaggle and UCL**

Which contain lots of data one is Laptop price prediction at csv file format?

Once you've gotten your goal figured out, it's time to start looking for your data, the second phase of a data analytics project. Mixing and merging data from as many data sources as possible is what makes a data project great, so look as far as possible.

- **Clean Data:**

The next step (and by far the most dreaded one) is cleaning your data. You've probably noticed that even though you have a country feature, for instance, you've got different spellings, or even missing data. It's time to look at every one of your columns to make sure your data is homogeneous and clean. Now that you have clean data, it's time to manipulate it in order to get the most value out of it. You should start the data enrichment phase of the project by joining all your different sources and group logs to narrow your data down to the essential features. One example of that is to enrich your data by creating time-based features, such as:

WWW.KAGGLE.COM

- Preprocessing the data
  - Cleaning the data
  - Encoding the data

- Extracting date components (month, hour, day of the week, week of the year, etc.)
- Calculating differences between date columns
- Flagging national holidays

The data can be downloaded from the Kaggle competition page. There are two files train.tsv and test. tsv and a Kaggle submission template sample_submission.csv. The total size of the data is 1.03 GB after decompression. The files consist of product listings. These files are tab-delimited. train. tsv has 1,482,535 rows and test .tsv has 3,460,725 rows. Both train and test files have the following data fields. Now let us start working on a dataset in our Jupyter Notebook. The first step is to import the libraries and load 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.
Exploratory analysis is a process to explore and understand the data and data relationship in a complete depth so that it makes feature engineering and machine learning modeling steps smooth and streamlined for prediction. EDA involves Univariate, Bivariate, or Multivariate analysis. EDA helps to prove our assumptions true or false. In other words, it helps to perform hypothesis testing. We will start from the first column and explore each column and understand what impact it creates on the target 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 modeling to achieve high performance and generalized models. so let's get started with analyzing and preparing the dataset for prediction. A row in a data table is called a data point and a column is called a feature/variable. Going forward in this blog, I will use the words row and data point interchangeably. Same follows with column and feature/variable

In this dataset we explore the effects of the

- Distribution of target column: Working with regression problem statement target column distribution is important to understand. 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.
- 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 you plot a count plot (frequency plot) of a company then the major categories present are Lenovo, Dell, HP, Asus, etc.
- Type of laptop: Which type of laptop you are looking for like a gaming laptop, workstation, or notebook. As major people prefer notebook because it is under budget range and the same can be concluded from our data.
- 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.

4. Algorithms

- Feature Engineering and Preprocessing of Laptop Price Prediction Model

Feature engineering is a process to convert raw data to meaningful information. There are many methods that come under feature engineering like transformation, categorical encoding, etc. Now the columns we have are noisy so we need to perform some feature engineering steps.

A. Screen Resolution

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

- Extract Touch screen information: It is a binary variable so we can encode it as 0 and 1.one means the laptop is a touch screen and zero indicates not a touch screen.
- Extract IPS Channel presence information: It is a binary variable and the code is the same we used above. The laptops with IPS channel are present less in our data but by observing relationship against the price of IPS channel laptops are high.
- Extract X-axis and Y-axis screen resolution dimensions: Now both the dimension is 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.

B. CPU column

If you observe the CPU column then it also contains lots of information. If you 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.

C. Price with Ram

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

D. Memory column

Memory column is again a noisy column that gives an understanding of hard drives. many laptops came with HHD and SSD both, as well in some there is an external slot present to insert after purchase. This column can disturb your analysis if not feature engineer it properly. So If you use value counts on a column then we are having 4 different categories of memory as HHD, SSD, Flash storage, and hybrid.

E. 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.

F. 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, you can try something else if you find more correlation with price.

- **Models**

Logistic regression, k-nearest neighbors, and random forest classifiers were used before settling on random forest as the model with strongest cross-validation performance. Random forest feature importance ranking was used directly to guide the choice and order of variables to be included as the model underwent refinement.

1. Classification Algorithms

a) Naive Bayes
b) Decision Tree
c) Random Forest
d) Support Vector Machines
e) K Nearest Neighbors

2. Regression Algorithms

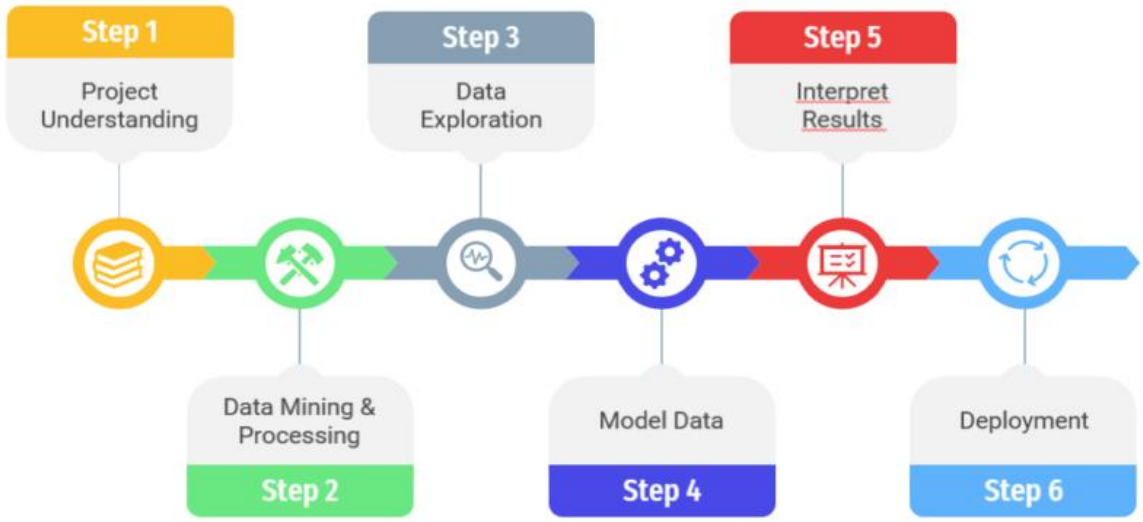
a) Linear regression

b) Lasso Regression

c) Logistic Regression

d) Multivariate Regression

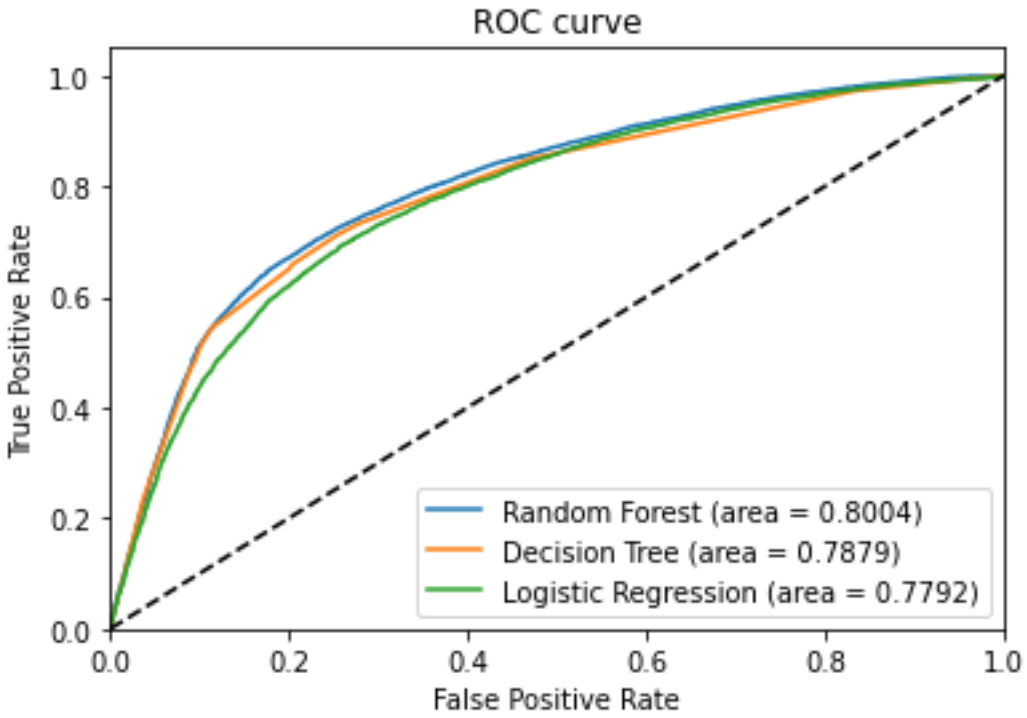
e) Multiple Regression Algorithms



Model Evaluation and Selection

The entire 130000-record training dataset was divided into 80/20 train vs. holdout, and all scores reported below were calculated using 5-fold cross validation on the training portion only. Because predictions on the 20% holdout were limited to the very end, this split was only used once, and the scores were only seen once. The official metric for Driven Data was classification rate (accuracy); however, class weights were included to improve performance against the F1

score and provide a more useful real-world application where classification of the minority class (functional needs repair) was required.

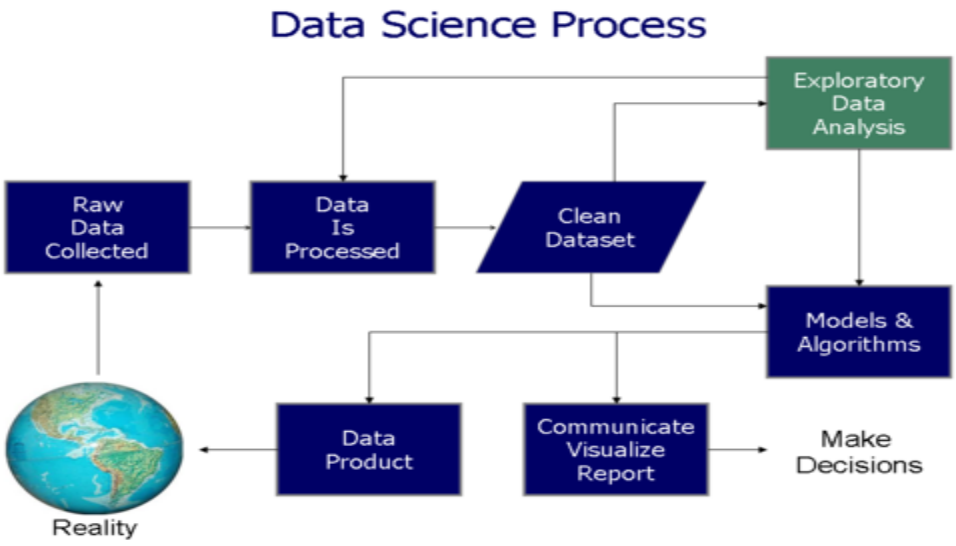


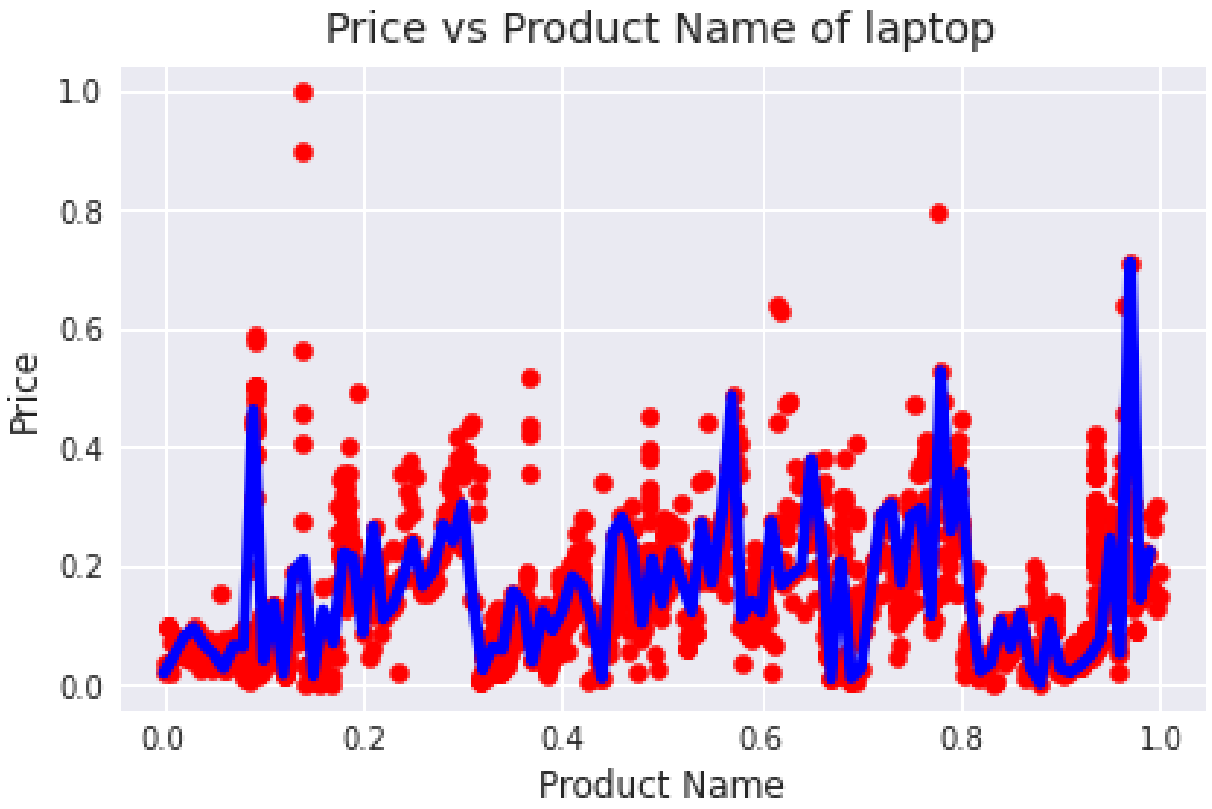
5. Tools used:

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- Tableau for interactive visualizations
- Jupiter notebook, Python 3.6 or 3.9
- Language: Python

6. Communication

In addition to the slides and visuals presented, Laptop price dataset will be embedded on my personal website and blog.


Data Science Process
Exploratory Data Analysis
Raw Data Collected
Data Is Processed
Clean Dataset
Models & Algorithms
Data Product
Communicate Visualize Report
Make Decisions
Reality


Price vs Product Name of laptop
Price
Product Name
0.0
0.2
0.4
0.6
0.8
1.0

Global Laptop Market

Market Share by Region (%)

Source: www.expertmarketresearch.com