REPORT OF MINI PROJECT

On

Laptop Price Prediction

Department of Computer Engineering and Application

GLA University,

Mathura - 281406

Submitted by:

Animesh Raghuvanshi(171500047)

Kartik Agarwal(171500155)

Submitted To:

Mr.Pankaj Sharma
Assistant Professor
Dept. of Computer Engineering and Application

Branch/Sec:

B.tech-CSE/Sec-B

ABSTRACT

The aim of this project was to automate the process of predicting the cost of laptop automatically. In this system we take the big data set of different types of laptop according to their specification then and use it for predicting the cost.

It refers to various tasks and chores associated with the organization. By using python as programming language and dataiku Workbench as platform for training the data set, we completed the project. In the end we were able to provide fully implemented project to user.

To start off with the project you first need to have a clear definition of what you are about to do and how you are going to implement it. So far you have known what stock price prediction means and why is it important to have one such system to predict its value. Now you need to know how you will implement this ideology. This chapter will provide you with all those specific details that would definitely require to make such a system.

Python is a general-purpose programming language. Hence, you can use the programming language for developing both desktop and web applications. Also, you can use Python for developing complex scientific and numeric applications. Python is designed with features to facilitate data analysis and visualization.

Dataiku is a computer software company having it headquarter in New York City. The company develops collaborative data science software marketed for big data. Dataiku offers a free edition and enterprise versions with additional features, such as multi-user collaboration or real-time scoring. Dataiku is the collaborative data science software platform for teams of data scientists, data analysts, and engineers to explore, prototype, build, and deliver their own data products more efficiently.

INTRODUCTION

1.1-Genaral Introduction:

The main objective of the project is to analyse and observe which features are most helpful in predicting the price of laptop. To achieve this, we used machine learning classification methods to fit a function that can predict the best result.

Our goal is to develop a model that has the capacity of predicting the cost of laptop, we will split the dataset into features and the target variable. And store them in features and prices variables, respectively

Once a model has been trained on a given set of data, it can now be used to make predictions on new sets of input data.

The main problem is to build a model that will predict cost of laptop with a high degree of predictive accuracy given the available data. With 8 explanatory variables describing (almost), this competition challenges you to predict the result of estimated price of laptop in euros."

1.2-MODULES OF PROJECT

The entire project is broken down into four major modules which describe the step by step building and functioning of the project. The modules are as described as follows:

1. Choosing the dataset:

For this project, we choose the dataset from kaggle.com and start implemented it by analysing the data.

2. Selecting a suitable algorithm:

Now that the dataset has been chosen the second major task is to select an algorithm that would prove to be the best fit for the dataset chosen. There are a number of algorithms available that all would be able to help in predicting the price of the stocks. A few of those algorithms are –Ordinary Least Squares , Ridge regression, Lasso Regression, Logistic Regression, Random Forests, Gradient Boosted Trees, XG Boost, Decision Tree, Support Vector Machine, Stochastic Gradient Descent, K Nearest Neighbors, Extra Random Trees, Artificial Neural Network, Lasso Path and custom models. Out of all these algorithms we need to have one with the maximum result in minimum time.

Linear regression algorithm is used in this project to predict the prices of the project. It is the simplest algorithm for linear regression. The target variable is computed as the sum of weighted input variables. OLS finds the appropriate weights by minimizing the cost function (i.e., how'wrong' the algorithm is).

3. Training the data:

The chosen algorithm is then applied on the dataset and the results are verified before uploading the associated model onto the server. The code associated is tested with some quick examples so as to make sure that it is going to fulfil the objective. This will show some results in form of certain graphs which will depict the changes that will come into picture when the code is run and executed as per desire.

4. Connecting to Server:

After the data has been trained and fed the code the model is then connected to server to make it accessible globally. After executing several commands, the model is connected to the server. This connection makes the model to be globally accessible. An IP address is allotted using Kali Linux server from AWS Workplace. This IP address will redirect the browser to the model. But before that a website layout is designed so that the user becomes user friendly. a creative website layout is designed and uploaded to which the IP address so allotted will be redirecting. This module is quite important as it is that side of the module through which the user will interact and its layout has to be pretty simple and elegant for the users use it comfortably.

5. Predicting the price:

After the various executions on the dataset and connectivity to the server the model is finally made ready to predict the cost of laptop. The user finally gets what he wants. After entering all the inputs, the user, the predicted price of laptop will be shown.

1.3-Hardware and Software Requirement:

Software Specification:

- **Technology Implemented** : Machine Learning
- **Language Used** : CSS, Python, PHP
- **Database** : MySQL, PHP
- **User Interface Design** : Dataiku
- **Web Browser** : Chrome, Explorer, Firefox etc.
-

Hardware Requirements:

- **Processor** : Core i7
- **Operating System** : Windows, Linux
- **RAM** : 8 GB
- **Hard disk** : 1 TB, Graphic Card
- **Display** : Laptop Screen

Problem Definition:

The purpose of project deals with determining a cost of laptop and avoids the problems which occur when carried manually. Laptop cost predection as it is very clear from the name of project that this will be helping the unawared people about the actual cost of that laptop so that he might be aware of the actual worth of the laptop as this will help them to get a clear look about that project with a just little effort of placing some details to web site and they will be getting the answer to lot of their craving question

Now as we have taken all the general knowledge required to start this project now, we need to know how to implement all the tools along with the languages stated in the above context in a proper sequence so as to get the desired results.

A proper step by step algorithm needs to be followed so that we don't miss a step that would create some kind of error in our results. The step by step procedure is described as follows:

- Collect Dataset
- Pre-process the dataset
- Select Algorithm
- Design Front End
- Deploy Model to Server

Objectives

Collecting dataset and pre-processing it is the very first step towards the implementation of this project. In this step what we need to do is to collect the dataset that would be the best suitable for our work (project). The dataset used in this project is the dataset of laptop. This dataset is been chosen because of its accountability and promising nature. This dataset is also quite simple to understand.

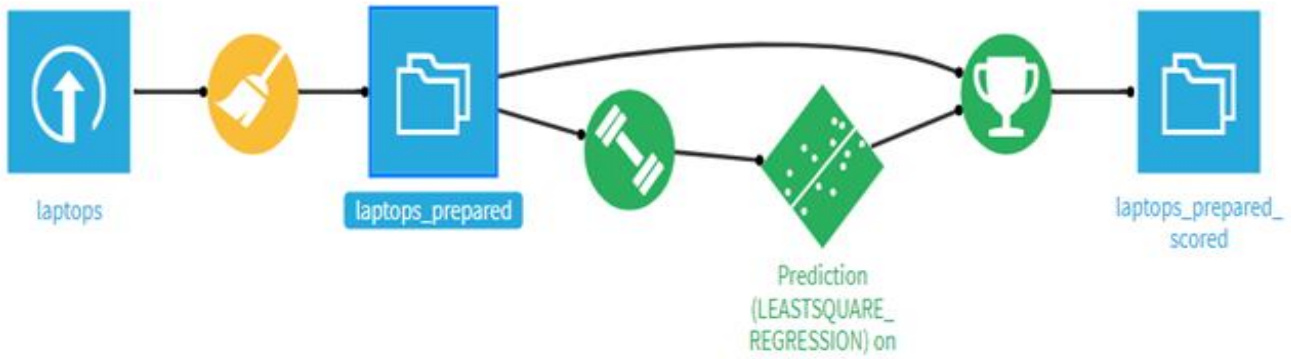
Before starting off the implementation lets first get through the tools which will be required for the implementation. There are a number of tools available for us to use but we need to be specific with our choice so as to pick the correct tool that working with them would be easy for the beginners too.

Dataiku

This platform is being used to the run the code and associate a model with it. This is the backbone of the system that we are trying to create here. This platform will be executing all the algorithms on the dataset that we will be providing and create a suitable model for the code to ru

The dataset is fed to Dataiku. The dataset is then trained. It is uploaded to it at the first instant. After uploading the data pre-processing is done. This is done to have the data in a proper format or the format that can be easing and properly used to serve our purpose.

Flow of data after processing at DATAIKU



In pre-processing the data, the columns are set to what we desire and we can remove some unwanted columns and rows such as columns having the value that will not affect the result. The rows and columns having null values should be removed from the dataset because when it will come to prediction of the price at some point where the value is fed as null an error might be generated that might crash the entire model. So it is better to remove such vulnerable and irrelevant data from the dataset.

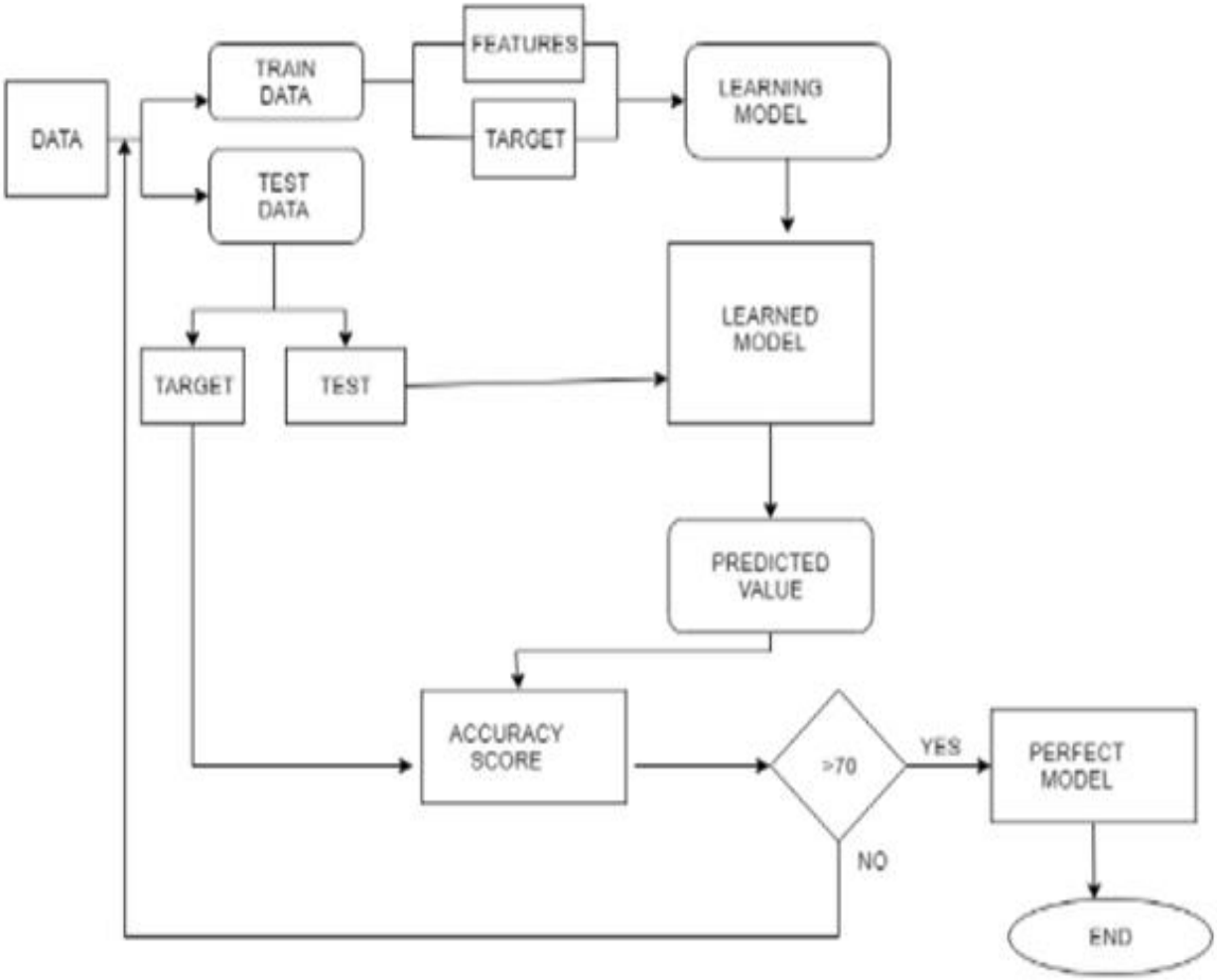
For example:

(i) In this dataset the date column has been removed as it is serving no purpose and more so ever our result doesn't depend on the date. Some values that are not in the desired datatype format are also changed.

(ii) Columns having values like 'yes' and 'no ' cannot be fed to the model so they need to be changed to integers so here in this case yes is changed to 1 and no is changes to 0. These numbers i.e. 0 and 1 are not compulsorily used any other number can also be used provided it is integer. After all this we get the prepared dataset and remember all this is done after the dataset is uploaded to Dataiku.

Methodology

The process of developing a complex product that tightly couples hardware devices with high-level software services requires an additional level of planning. For this project, we will exercise a proper product development approach to help you get familiar with the process of creating real-world hardware projects. This method can then be used to plan your own projects and take them to the next level. The following diagram describes a typical prototype development process, which always begins by defining the major goals that you want to achieve with your product:



Implementation Details

After the processing on the dataset has been done and dataset has been tested now, we need to make the dataset go through a algorithm test. But before doing this we need to first decide the prediction style that we would like to use. Dataiku provides two types of prediction styles:

(i) Automated Machine Learning

(ii) Expert Mode

The automated machine learning is the model which allows the user to get highly optimized model with minimal intervention. It will analyse your dataset, and depending upon preferences, select the best features handling, algorithms and hyper parameters. The expert mode will let you create your own model and let you have all the access to the model so that you can modify it according to your needs.

In this test a number of algorithms are implemented on the dataset. This gives us the idea as of which algorithm will provide the best result will maximum accuracy and in the least time possible. A number of algorithms are available in Dataiku like: Ordinary Least Squares, Ridge regression, Lasso Regression, Logistic Regression, Random Forests, Gradient Boosted Trees, XG Boost, Decision Tree, Support Vector Machine, Stochastic Gradient Descent, K Nearest Neighbors, Extra Random Trees, Artificial Neural Network, Lasso Path and custom models. Out of all these algorithms we need to have one with the maximum result in minimum time. One of the major factors helping us to decide which algorithm to select is the R2 score. Before proceeding further first let us know what R2 score is.

R2 score, also called "the coefficient of determination", is the proportion of variance in the dependent variable that is predictable from the independent variables. It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

The algorithm that will be having the maximum R2 score will be used for the prediction. The algorithm that proved to here be more promising than the others is.

After the code has been got ready, we have to design the front end. This is the user end. Through this the user will be interacting with the code and making use of it. This front end is designed using Slides. Slides is a website which helps you to get an adaptive layout for your model's front end. Select a suitable layout as per your desire and download its zip file. Now the format that has been downloaded might not be the perfect one to be implemented so we need to make some changes. Here Brackets help us to do make these changes easily.

After all the editing, we come to a major step of the project i.e. connecting it to the server. To initiate with this step, we need to have a server allocated to us so that we can use it and import our model onto it. This is facilitated by Amazon Web Services (AWS) workspace.

AWS Workspace

- EC2
- Instances
- AWS Marketplace
- Kali Linux Server

Contribution Summary

The team is doing a fantastic job for making this project. The credit of the work completed till now goes to each and every member of the team. Their commitment and devotion towards the work will definitely make this project a great one.

The work done by each member is as followed:

Animesh Raghuvanshi: is responsible for making all the necessary arrangements for this project. This include the work like making arrangement like collecting datasets, pre-process the dataset and selecting the algorithm.

Kartik Agrawal: is responsible for all the programming aspects of the project which includes making changes in the code as per the conditions or situations provided to us. He is responsible for designing frontend and deploying model to server.

Progress till date and Remaining work

Till now we have worked on the following aspects of the project which has been completed till date.

This includes:

1. For this project, we choose the dataset from kaggle.com and start implemented it by analysing the data.

2. Now that the dataset has been chosen the second major task is to select an algorithm that would prove to be the best fit for the dataset chosen.

Linear regression algorithm is used in this project to predict the prices of the project. It is the simplest algorithm for linear regression. The target variable is computed as the sum of weighted input variables. OLS finds the appropriate weights by minimizing the cost function (i.e., how‘wrong’ the algorithm is).

Remaining work:

1. The chosen algorithm is then applied on the dataset and the results are verified before uploading the associated model onto the server.

2. After the data has been trained and fed the code the model is then connected to server to make it accessible globally.

3.After the various executions on the dataset and connectivity to the server the model is finally made ready to predict the cost of laptop.

References

1. http://scikit-learn.org/stable/index.html
2. http://medium.com
3. http://anaconda.org
4. http :// pythonprogramming .net / http :// pythonprogramming .net / introduction-to-python-programming