Machine Learning Nanodegree

Capstone Proposal

Product Supplier Search Process

Anirban Paul

July 23, 2018

1 Domain Background

The average customer of today owns and uses at least four devices – mobile, tablet, laptop and a desktop on a daily basis through various ecommerce websites like Flipkart, Amazon, Aliexpress etc. At least 50-60% of customers use their smartphones and tablets to make online purchases. Numerous amounts of data are generated every day in various ecommerce websites. Lots of important and valuable information is stored in these data which can be used to discover various patterns and which could give some great insights. The benefits of using machine learning in ecommerce can be found in this link: https://www.loop54.com/how-machine-learning-can-benefit-your-e-commerce-company

With the help of data various customer buying patterns can be discovered which could help retailers in selling their products at a good profit. In other ways these information can also be used to help customers in choosing the best product and recommend them to select the best sender or the top ten vendors for the particular product as in this case is ASUS Laptops. I am going to develop recommender systems which will help the customer to know the top ten vendors for a product. are very powerful tools that are user friendly and provide germane suggestions to users that are quite accurate, and can help create a positive impact.

I would be using three machine learning approaches for building the model for the prediction of the vendors. I will be using AdaBoost, Gradient Boosting and Decision Tree Regressor(if needed) to predict the top ten vendors at for ASUS laptops. Amongst these three models the ADA Boost will serve as a benchmark model and an attempt will be made to build at least one model which outperforms or is at par with the benchmark model, which I am looking forward to achieve with the other two model (Decision Tree Regressor and Gradient Boosting)

Gradient boosting is a machine learning method created by Leo Breiman for regression and classification related problems, which produce mostly decision tree in the form of an ensemble of weak models. . It builds stage-wise models like other boosting approach and it allows optimization of differentiable loss function.

AdaBoost, short for Adaptive Boosting, is a machine learning algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Gödel Prize for their work. It can be used in combination with many other various types of algorithms to improve performance and get better result. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier.

Some of the previous work of seller selection proces can be found in the below link:

https://www.deltabid.com/vendor-selection-process/

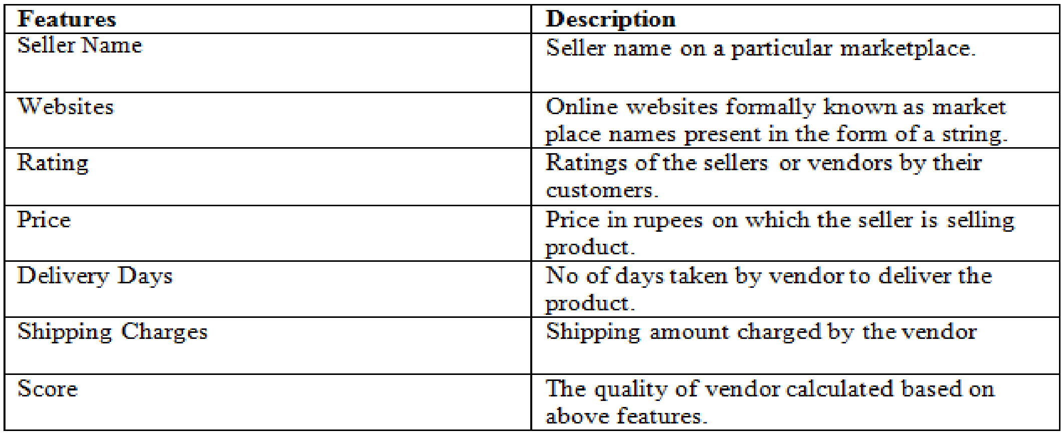
2 Problem Statement

In this project I will be using three machine learning technique to develop a recommender system that gives top ten vendors for ASUS laptop. There will various sellers and their features depending on those features a scoring function will be calculated which will serve as a target label, more the scores better will be the vendor. The three machine learning techniques are ADA Boost, Decision Tree Classifier and Gradient Boosting. These methods will be used to predict the scores of different vendors on the test set. I will be using AdaBoost, Decision Tree Regressor(if needed) and Gradient Boosting to predict the top ten vendors for ASUS laptops across various market place. Amongst these three models the ADA Boost will serve as a benchmark model and an attempt will be made to build at least one model which outperforms or is at par with the benchmark model, which I am looking forward to achieve with the other two model (Decision Tree Regressor and Gradient Boosting)

The problem is a regression problem, in which the main task is to predict the scores of the vendors and that score will be used to determine the top ten vendors from the test set. There will be various input features like Seller Name, Price, Ratings, Delivery Days, Websites, Shipping Charge and Score.

3 Datasets and Inputs

The dataset that I have used for this project is manually collected from various websites like Flipkart, Amazon, Aliexpress, Snapdeal etc. containing 6 features such as Seller Name, Price, Ratings, Delivery Days, Websites, Shipping Charge and 1 label which is known as Score which is basically a scoring function calculated depending different features. The dataset has 50 different sellers. Description of different features is given below:



Features	Description
Seller Name	Seller name on a particular marketplace.
Websites	Online websites formally known as market place names present in the form of a string.
Rating	Ratings of the sellers or vendors by their customers.
Price	Price in rupees on which the seller is selling product.
Delivery Days	No of days taken by vendor to deliver the product.
Shipping Charges	Shipping amount charged by the vendor
Score	The quality of vendor calculated based on above features.

The mathematical equation for scoring function is Score=(ratings*100000)/(delivery days*price*shipping charges). I have multiplied ratings with 100000 because the score was coming very small.

4 Solution Statement

The goal of this project is to build a recommender system model that can predict the top ten vendors for ASUS laptop. There will various sellers and their features depending on those features a scoring function will be calculated which will serve as a target label, more the scores better will be the vendor. The three machine learning techniques are ADA Boost, Decision Tree Classifier and Gradient Boosting. These methods will be used to predict the scores of different vendors on the test set. I will be using AdaBoost, Decision Tree Regressor(if needed) and Gradient Boosting to predict the top ten vendors for ASUS laptops across various market place. Amongst these three models the ADA Boost will serve as a benchmark model and an attempt will be made to build at least one model which outperforms or is at par with the benchmark model, which I am looking forward to achieve with the other two model (Decision Tree Regressor and Gradient Boosting)

5 Benchmark Model

The Benchmark model for this type of problem is decided as ADA Boost Regressor with the prediction accuracy of 80% of top 10 sellers in test set and other model like Decision Tree Regressor(if needed) and Gradient Boosting will be tried on the test set to obtain better or at par performance with Benchmark Model. The final accuracy of different with the benchmark model will be produced in excel file(CSV) to show the whether top ten vendors produced by the algorithms is present in the original top ten vendors in the test set list or not.

6 Evaluation Metrics

This project will be evaluated with regards to the model's ability to predict the top ten vendors. I expect that the two model that I am going to use other than ADA Boost model which will be my benchmark model among them at least one model amongst them to perform better than the benchmark model. The project should be able to train successfully on random training data and then test the model and produce an accurate score based on features. The Evaluation metrics for this case is Mean Squared Error and also predicted top 10 sellers in test set by the various model will be compared with the original and based on no of wrong vendors present in top 10 will be used to say accuracy.

7 Project Design

This project comprises of three different machine learning algorithms they are ADA Boosting, Decision Tree and Gradient Boosting to predict the scores for given features of the sellers. Each of these algorithms will be composed of these steps

Data Exploration: Visualizing the dataset, dropping the string format columns, cleaning the dataset, check relevance of every column to the target column for example the dataset has some values in string format. Some string format column has to be dropped. Then the data is split into training and testing data. I will be looking into different classic machine learning models, and their discoveries in diseases risks. I will be developing three classifications using AdaBoost, Decision Tree Regressor(if needed) and Gradient Boosting to predict the top ten vendors for ASUS laptops across various market place. Amongst these three models the ADA Boost will serve as a benchmark model and an attempt will be made to build at least one model which outperforms or is at par with the benchmark model, which I am looking forward to achieve with the other two model (Decision Tree Regressor and Gradient Boosting) After data preprocessing I will be splitting the training dataset into training(70 percent) and testing(30 percent) sets using train test split etc.

Training: After randomly splitting the dataset into training and testing sets I will be training the every models on the samples.

Testing: After successful training of the model I will be testing with other models to predict the scores which should be close to the original scoring function and tried to get less Mean Squared Error and also try to improve the accuracy of predicting the top ten vendors.

The main aim of this project is to build a recommender system model that can predict the top ten vendors for ASUS laptop As discussed I am using the dataset which is collected manually from various datasets . I am using Jupyter Notebook for the development of the project. The code has been written in Python language. The data has to be preprocessed. Then we have to apply regression algorithms to the data according to the three algorithms discussed earlier. The Mean Squared Error and the accuracy in producing the top 10 sellers in checked each algorithm is then observed and then the top ten vendors by each algorithms is compared with the original top ten sellers from the test set using excel sheet.