Predictive Analysis of E-Commerce data from multi-category store

Shailja Pandit
Anish Omprakash Pandey
Surya Kiran Golagani
Department of Information Systems, California State University
Los Angeles
e-mail: spandit3@calstatela.edu, apandey9@calstatela.edu , sgolaga@calstatela.edu

Abstract: This project will illustrate the usage of Machine Learning algorithms on E-Commerce datasets. For this project, we will use our knowledge, research, and development of a predictive model to compare the results of all the models and predict the probability of the item being sold. This paper focuses on analyzing eCommerce data from a multi-category store. It presents how the products and the users are connected to each other and how the analysis affects the business.
Our goal is to use five Machine learning models to predict, in this paper. Based on the accuracy of the regression models, the multi-category store can implement the strategy in order to manage new items for the store. We have used four Regression algorithms for predicting the Item Price ad One Multiclass Classification model to predict the Event type.

Keywords: Regression, Classification, MultiClass Classification, Hadoop, Price, EventType, R2 , RMSE, Precision, Recall, Zepplin, Train Split Validator, Cross Validator

1. Introduction

We live in the world of e-commerce. There are many shops on the internet. The Internet has made it possible to trade with anyone, anywhere. We can buy products without leaving the house and compare prices in different stores within seconds. We don't just find what we really want and accept the first offer that is appropriate. And we think it's interesting to see the world through the data it produces. So, we decided to play with the e-commerce number to deepen our understanding.
The objective of the four Regression Algorithms includes building a model that predicts the optimal price of the item considering the features of the item and the users. This analysis is useful so that the Company does not overcharge or undercharge for the Item. If the outlet charges more than the market price, then it might miss on the potential customers. On the other hand, if the outlet charges less for a particular item, then it might miss out on the potential revenue. So we use the machine learning model to predict the optimal price of the Item. Also, the latter part of the project focuses on predicting the event type, by this prediction we can predict if a particular item is being purchased or is only being viewed or added to the cart.

2. Data Set Used

The dataset used in this paper is taken from an open-source platform which is Kaggle.com. The data is for the month of October 2019. There are 9 columns and about 4 M rows. Each row in the data signifies an even which are related to the users and the products.
URL: https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store?select=2019-Oct.csv
Dataset Size: 5 GB, Format: CSV

The different columns present in the dataset are listed below:

- Event Time: This column describes when the event took place.
- Event Type: It takes one entry -View, Cart, Purchase
- Product Id: It is the ID of the product.
- Category Id: It is the category of the product.
- Category Code: It is the code assigned to each category
- Brand: It represents the Brand name of the product.
- Price: It is the price of the product
- User Id: It is the Id of the user which is generated at the login. It is a unique value.

3. Technical Specifications

Cluster Version	Hadoop 3.2.1-amzn-3.1
No of CPUs	8 OCPUs
Pyspark Version	3.0
Number of Nodes	3
CPU speed	2.20 GHz
Total Storage	481 GB

4. Related Work

[1] A paper published in Springer link focuses on the price forecasting model for e-commerce products which is based on time series and sentiment analysis. The prediction model they have used in their project is Autoregression, Time Series, SVM, SSA-ARMA prediction model, and neural network. This paper similar to ours is also related to price prediction, but the algorithms and the approach used by them differ from what we used in our paper.

[2] A paper on the '*Analysis of recommendation algorithms for e-commerce*' evaluated various algorithmic choices for CF-based recommender systems. The recommender systems apply data analysis techniques to the problem of helping the customer find which products they would like to purchase. This algorithm is basically used in all e-commerce websites and differs from the algorithms which we have used. Our algorithms are strictly used to only predict price.

[3] A paper on '*Analysis of e-commerce behavior in Multi-Category Store*' comes up with various visualizations based on human behavior towards the purchase, this analysis would help in growth and earning a profit which is very crucial for any manufacturer. This paper differs a lot when compared to ours. Our

paper focuses more on predictive analysis and their paper focuses more on descriptive analysis.

5. Background/Existing work

In our project, we have used 4 regression algorithms for Item Price Prediction and 1 Classification Algorithm for Event Type(View/Cart/Purchase) Prediction.

5.1 Regression

Regression is a supervised machine learning technique that is used to predict continuous values. In part one of our project involving Item price prediction, we make use of Regression models, as price, the target variable is a continuous numeric variable. We used four Regression algorithms -Decision Forest Regression, Gradient Boost Tree Regression, Linear Regression, and Random Forest Regression. We used the Tune Model Hyperparameters module for improved model performance, Train Split Validator module, and Cross Validate module to ensure if the model generalizes well and the Permutation feature importance module to eliminate less important features iteratively. For price prediction using SparkML we used the algorithms -Gradient Boosted Tree Regression, Decision Tree Regression, and Random Forest Regression. The regression models used are based on the lab work involving predicting the 'arrival delay' in the 'flights' data. A similar process of creating a pipeline for feature transformation and training a regression model was performed. We also used a Cross Validator to find the best-performing parameters. The evaluation metrics used are Root Mean Square Error (RMSE) and Coefficient of Determination(R2).

5.2 Classification

Classification is a supervised ML technique that categorizes a set of data into classes. We used multi-class classification models in part two of our project involving event-type prediction. We build a model to classify the Event-Type as View/Cart/Purchase. Since our Event-Type contains three classes instead of two, therefore, a multi-class classification is used instead of a Binary Classification. Binary class classification is used when there are only two categories, we use multi-class classification when there are more than two categorical values. In this project, we used Decision Tree Multi-Class Classification model. The evaluation metrics used are Accuracy, Test error, Recall, and Precision. Accuracy is a fraction of Properly predicted cases. Test Error Implies error on a test set. Precision quantifies the number of positive class predictions that actually belong to the positive class. Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

5.3 Workflow Architecture

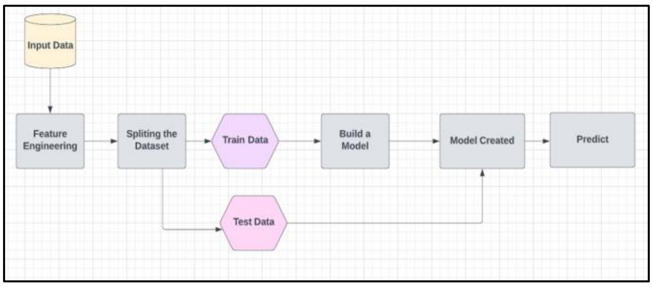


Fig 4.2 Work Architecture

Input Data: Input the data. That data can be structured or unstructured.

Feature Engineering: frame a machine learning problem in terms of what we need to foresee and what sort of observation data we have to make those predictions.

Splitting the Dataset: Splitting the dataset into train and test data.

Build a Model: This is a significant step to picking a legitimate model to implement and predict the output. The model is inbuilt on the Train data set. Once the model is fit. We run it using the Test Data set.

Predict: This is the step where a solution is obtained. Also, further, we can evaluate and compare the results.

6. Predictive Analysis

Predictive Analysis is a branch of analytics that focuses on making predictions about future outcomes using historical data. Predictive models help businesses retain customers and grow their business, such as in our case. Predictive models can also be used by businesses to forecast inventory and manage resources.

In this paper, we will discuss a few of the machine learning models mentioned in the paper later. The predictive analytics models are designed to access historical data, discover patterns, observe trends, and then use the information to draw up predictions about future trends.

7. Our Work

We implemented various Machine learning models to predict price and the EventType. We made use of the whole dataset to build predictive models.

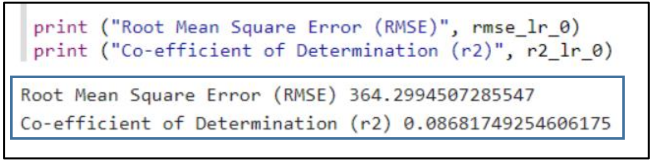
Big Data is a collection of enormous datasets of aggregate volume, velocity, and variety such a large amount of data is difficult to manage and process. To deal with such an enormous load of data methods like Hadoop, Data Mining, Machine learning, etc. are used. Machine learning is a various and energizing field to deal with big data and there are numerous approaches to define it. It transforms data into a program and automates the automation system, Following are a few Machine learning algorithms:

7.1 Linear Regression

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.'

We used Cross-Validation and Train Split Validation Model for Linear regression. We split the data into 70% train and 30% test. Tune Model Hyperparameters were used to find the best-performing model. The Cross-Validation helped generalize the model. Parameters were defined using Param Grid. Fig 7.1a and 7.1b below shows the evaluation results for Linear Regression.

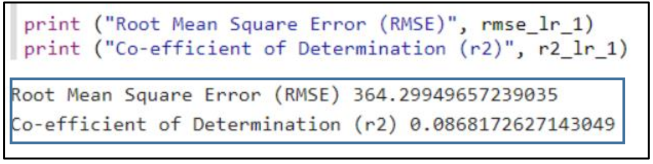
Fig 7.1a- Linear regression Using Train split Validator



```
print ("Root Mean Square Error (RMSE)", rmse_lr_0)
print ("Co-efficient of Determination (r2)", r2_lr_0)
```

```
Root Mean Square Error (RMSE) 364.2994507285547
Co-efficient of Determination (r2) 0.08681749254606175
```

Fig 7.1 b – Linear regression using Cross Validator



```
print ("Root Mean Square Error (RMSE)", rmse_lr_1)
print ("Co-efficient of Determination (r2)", r2_lr_1)
```

```
Root Mean Square Error (RMSE) 364.29949657239035
Co-efficient of Determination (r2) 0.0868172627143049
```

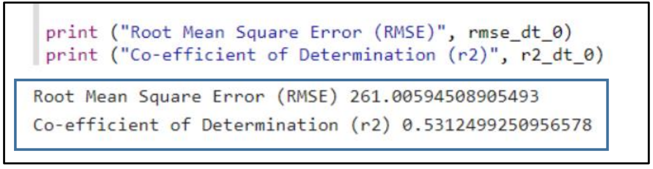
We got approximately same RMSE and R2 values when Linear Regression Algorithm was run using Train Split and Cross Validator. It is clear from the results that the RMSE value is too high which is 364.299 and the Coefficient of Determination which is R2 is very low which is 0.086 . Hence, It is not a good model for prediction.

7.2 Decision Tree Regression

As we could see from Section 7.1, the Linear Regression is not the best model for Price prediction due to high RMSE values. Therefore we used Decision Tree Regression Algorithm for Price Prediction.

We used Cross-Validation and Train Split Validation Model for Decision Tree regression. We split the data into 70% train and 30% test. Tune Model Hyperparameters were used to find the best-performing model. The Cross-Validation helped generalize the model. Parameters were defined using Param Grid. Fig 7.2a and 7.2b below shows the evaluation results for Decision Tree Regression.

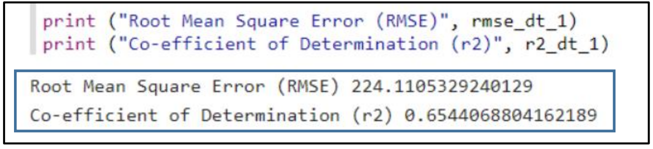
Fig 7.2a- Decision Tree regression Using Train split Validator



```
print ("Root Mean Square Error (RMSE)", rmse_dt_0)
print ("Co-efficient of Determination (r2)", r2_dt_0)
```

```
Root Mean Square Error (RMSE) 261.00594508905493
Co-efficient of Determination (r2) 0.5312499250956578
```

Fig 7.2b- Decision Tree regression Using Cross Validator



```
print ("Root Mean Square Error (RMSE)", rmse_dt_1)
print ("Co-efficient of Determination (r2)", r2_dt_1)
```

```
Root Mean Square Error (RMSE) 224.1105329240129
Co-efficient of Determination (r2) 0.6544068804162189
```

As we can see from the figures above, the R2 value for the cross validator model 0.654 is better than that of the Train split Validator model 0.532. Also, the CV model has lower RMSE than the TSV model. But overall we can see that Although the Coefficient of Determination for the Decision Tree Algorithm model is better than that of the Linear Regression Model, but is not the best.

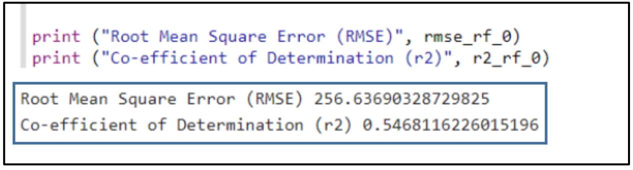
7.3 Random Forest Regression

Random forests are the most adaptable and simple to utilize supervised learning algorithms. It can be utilized both for classification and regression. However, it is generally utilized for classification. A forest consists of trees. Random forests have a range of uses, for example, recommendation engines, image classification, and feature selection.

As we could see from Section 7.1and Section 7.2, the Linear Regression and Decision tree did not work out well for Price prediction due to high RMSE values. Therefore we used Random Forest Regression Algorithm for Price Prediction.

We used Cross-Validation and Train Split Validation Model for Random Forest regression. We split the data into 70% train and 30% test. Tune Model Hyperparameters were used to find the best-performing model. The Cross-Validation helped generalize the model. Parameters were defined using Param Grid. Fig 7.3a and 7.3b below shows the evaluation results for Random Forest Regression.

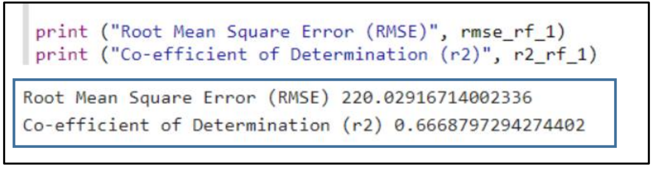
Fig 7.3a- Random Forest regression Using Train split Validator



```
print ("Root Mean Square Error (RMSE)", rmse_rf_0)
print ("Co-efficient of Determination (r2)", r2_rf_0)
```

```
Root Mean Square Error (RMSE) 256.63690328729825
Co-efficient of Determination (r2) 0.5468116226015196
```

Fig 7.3b- Random Forest regression Using Cross Validator



```
print ("Root Mean Square Error (RMSE)", rmse_rf_1)
print ("Co-efficient of Determination (r2)", r2_rf_1)
```

```
Root Mean Square Error (RMSE) 220.02916714002336
Co-efficient of Determination (r2) 0.6668797294274402
```

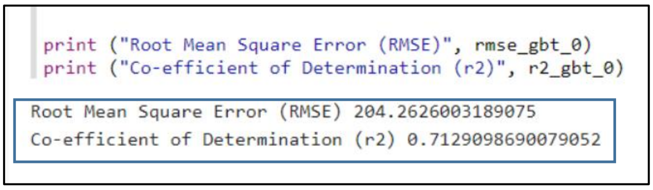
As we can see from the figures above, the R2 value for the cross validator model 0.667 is better than that of the Train split Validator model 0.546. Also, the CV model has lower RMSE than the TSV model. But overall we can see that Although the Coefficient of Determination for the Random forest Algorithm model is better than that of the Decision Tree Regression Model, but is not the best.

7.4 Gradient Boost Tree

Gradient Boost Tree (GBT) is a machine learning technique used in the regression. It gives a prediction model in the form of an ensemble of weak prediction models, which is typically a decision tree. When a decision tree is a weak learner, the resulting algorithm is called a gradient boost tree, it usually is better than the random forest algorithm.

All the above-listed models are not the best model for Price prediction. Now we will try the GBT regression model. We used Cross-Validation and Train Split Validation Model for Decision Tree regression. We split the data into 70% train and 30% test. Tune Model Hyperparameters were used to find the best-performing model. The feature Importance module was used to determine the best features to use in the model and Cross-Validation helped generalize the model. Parameters were defined using Param Grid. Fig 7.4a and 7.4b below shows the evaluation results for GBT Regression.

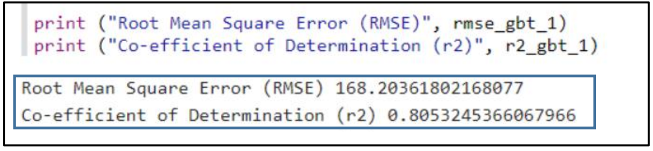
Fig 7.4a- GBT regression Using Train split Validator



```
print ("Root Mean Square Error (RMSE)", rmse_gbt_0)
print ("Co-efficient of Determination (r2)", r2_gbt_0)

Root Mean Square Error (RMSE) 204.2626003189075
Co-efficient of Determination (r2) 0.7129098690079052
```

Fig 7.4b- GBT regression Using Cross Validator



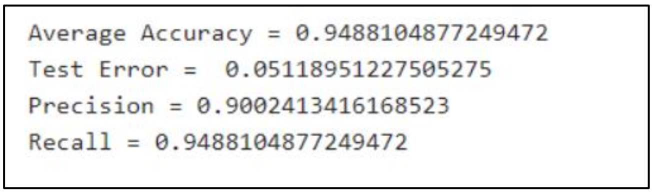
```
print ("Root Mean Square Error (RMSE)", rmse_gbt_1)
print ("Co-efficient of Determination (r2)", r2_gbt_1)

Root Mean Square Error (RMSE) 168.20361802168077
Co-efficient of Determination (r2) 0.8053245366067966
```

As we can see from the figures above-Fig 7.4 and Fig 7.4b, the R2 value for cross validator model 0.805 is better than that of the Train split Validator model 0.712. Also, the CV model has lower RMSE 168.20 than the TSV model 204.26. But overall we can see that based on the RMSE and R2 value , the GBT regression model is the best fit for Item Price Prediction.

7.5 Decision Tree Classifier

Decision tree classifiers are considered to be the most famous of the best-known methods of data classification representation of classifiers. Researchers from different disciplines and backgrounds have tackled the problem of extending a decision tree from available data, such as machine study, pattern recognition, and statistics. Decision tree classifiers created the classification model by building a decision tree.



```
Average Accuracy = 0.9488104877249472
Test Error =  0.05118951227505275
Precision = 0.9002413416168523
Recall = 0.9488104877249472
```

We have used multiclass classification in our case, we chose event type as our target variable, which has 3 classes: purchase, view, and cart renamed as 0,1 and 2. So here we are predicting whether a customer is purchasing the device, viewing the device, or adding the device to the cart. we have performed decision tree classification on our data and the results tabulated are average accuracy of 0.95 and a test error is 0.05. Where, Accuracy is just the percentage of predictions that were made correctly, and the Test error is the performance of the model on unseen or test data. The Precision is 0.9002, Precision Indicates how many positive predictions were actually true.

8. Comparison Table

According to the Comparison table below.

Algorithms	**Results for CV**	**Time Taken to fit the model**
Linear Regression	R2: 0.086 RMSE: 364.299	4 min 54 sec
Random Forest Regression	R2: 0.67 RMSE: 220.029	14 min 15 sec
Decision Tree Regression	R2: 0.65 RMSE: 224.11	9 min 11 sec
GradientBoost Tree Regression	R2: 0.81 RMSE: 168.20	35 min 58 sec
Decision Tree Classifier	Precision: 0.90024 Recall: 0.94881	10 min 13 sec

We can arrange various Regression Algorithms in the below order.

On the basis of time :

GBT> RF >DT> LR (GBT taking the most time and LR taking the least)

On the basis of accuracy :

GBT>RF>DT> LR (GBT having the best accuracy and LR having the least).

Thus, we can conclude that even though the GBT model takes the maximum time to run. It is the best model, with R2 closer to 1 and the least RMSE value.

9. Conclusion

This paper attempts to come up with the best-performing models for price prediction and Event Type Prediction. We performed a predictive analysis in Hadoop-Spark Cluster using various Regression and Classification algorithms. We used four regression algorithms for price prediction and one multi-class classifier for Event Type Prediction.

When referring to the table above, for price prediction, we have achieved an RMSE of 168.20 for Gradient Boost Tree, and an R2 value of 0.81. Which makes this algorithm the best fit for price prediction When compared to other regression algorithms. For event-type prediction, the algorithm used is a decision tree multi-classifier. The precision achieved for this algorithm is 0.9002 and the value of recall is 0.948, which makes it a good model for event type classification.

Predicting Item Price helps the company to set the optimal price for their Items and it also helps them understand how different features of the listing can be used to accurately predict the price. With better price estimates, an eCommerce website can reach an equilibrium price that optimizes profit and affordability. Predicting if the Item is being purchased or just viewed or added to the cart, will also help compare the Items and help the retailers know which Item is in high demand.

References

[1] Tseng, KK., Lin, RY., Zhou, H. “ Price prediction of e-commerce products through Internet sentiment analysis.” Electron Commer Res 18, 65–88 (2018). https://doi.org/10.1007/s10660-017-9272-9, October 2017

[2] Sarwar, Badrul, George Karypis, Joseph Konstan, and John Riedl. "Analysis of recommendation algorithms for e-commerce." In *Proceedings of the 2nd ACM Conference on Electronic Commerce*, pp. 158-167. 2000.

[3]Sachdeva, Sanya, and Supriya Raheja. "Analysis of e-commerce behavior in Multi-Category Store." 2013

[4]Github Link- **https://github.com/shailjapandit05/CIS-5560-e-Commerce-Prediction-Project**

[5] Dataset Source- **https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store?select=2019-Oct.csv**