Predictive Framework for E-Commerce Product Categorization on Azure Cloud

Nisha[1], Amit saxena[2]

[1]Mtech scholar computer science department, Truba institute of engineering & information Technology, Bhopal, M.P, India
[2]Associate Professor (H.O.D) Computer Science Department, Truba institute of engineering & information Technology, Bhopal, M.P, India

---***---

Abstract - E-commerce refers to the Electronic commerce and defined as buying and selling of products over electronic systems such as the Internet. With the widespread use of Internet, the trade conducted electronically (online) has grown extraordinarily. The E-commerce companies have a large database of products and a number of consumers that use these data. To address this data and information explosion, e-commerce stores are applying machine learning to identify and customize the product category information. Data scientists in this field are utilizing machine learning potential to build unmatched competitiveness in the market by finding purchase preferences, customer churn and product suggestions etc. Applying popular Machine Learning algorithms to huge datasets brought new challenges for the ML practitioners as traditional ML libraries do not support well processing of large datasets. So to address the issue, data and computation can be distributed to any Cloud Computing environment with minimal effort. Cloud computing paradigm turned out to be valuable alternatives to speed-up machine learning platforms.

The work, first discusses the machine learning and its importance in predictive analytics. Introduction to multiclass classification is presented. Few E-commerce classification frameworks and need of cloud platforms to analyze ever growing E-commerce data is briefly surveyed then. Finally, this work proposes the predictive framework for E-commerce Product Classification which is developed over Microsoft Azure Cloud. The proposed framework predicts the Product Category in a large E-commerce dataset released by a famous e-commerce company for a competition. The proposed classifier is build using 'Multiclass Logistic Regression' by choosing the optimal parameters. The results obtained by proposed model are evaluated and presented in terms of accuracy. The work also demonstrates the use of leading cloud environment for machine learning. The results obtained in this research are promising and the dissertation also directs the future research work in the field.

***Key Words*: E-commerce, Machine Learning, Product Classification, Microsoft Azure, Cloud Computing**

1. INTRODUCTION

Background – Classification of E-commerce Data

In the present era of communication web is the best medium for doing business. The barrier of time and space has been break down by online businesses comparing to physical shops or office. Big companies around the world are realizing that e-commerce is not just a way of buying and selling over Internet rather it is helpful in improving the competence than other giants in the market. To raise their prospective markets, the big companies like Amazon, are mounting and starting new ventures in very short span of time. So, now each of these companies wants to utilize Machine Learning (ML) potential to fabricate unmatched competitiveness in the market.

That is why, Data scientists have been in huge demand in E-Commerce segments. To better understand this evolution, it is vital to make acquainted oneself with all the probable applications of ML in the E-Commerce business scenario. ML has empowered businesses to analyze all queries, whether searched or abandoned, from all the users. Predictive analytics which is based on machine learning can improve sale probability and find customer churn, by analyzing customer's past click-through actions, purchases, preferences and history in real-time.

The advancement toward E-Commerce in the Internet has produced business policies like e-wallets, which is still not available at physical level. Businesses offer more choices to consumers by means of E-commerce. Increasing choice, on the other hand, has also amplified the amount of data and information that consumers must process before they are able to choose which objects meet their needs.

In past decade, E-commerce stores employed following advertising and marketing strategies:

- ✓ Remarketing: E-commerce firms follow individuals through the use of website cookies or other mechanisms and then present product advertisements significant to them while visiting other websites.

- ✓ Website Personalization: E-commerce websites are commonly able to change the appearance of their pages and emphasize certain products depending on the visitor profile.
- ✓ Deal Customization: Product Deals are offered to certain customers on the website if it is expected that they may get customer's increased purchase response in the future. Also, items that are bought frequently are suggested to new or potential customers.
- ✓ Email personalization: Personalized emails could be sent to potential customers based on their earlier visits to the website or partner websites.

Summarizing below few reasons that justify/establish the need of Cloud platforms to analyze E-commerce data:

A. Fast Analysis

Very large amounts of training data need a great deal of computer memory and processing power for classification and regression analysis. Particularly with data representing complex non-linear behaviors, as with text, speech, handwriting, face recognition, stock price prediction, and financial forecasting, the computing bill, requirement can be quite large. However, the emergence in recent years of cloud computing changes everything. IaaS providers like Amazon Web Services (AWS) and 'Google Cloud' platform now offer access to virtually unlimited computing power on demand, in the form of clustered parallel servers that can be used for an hourly fee.

B. Machine Learning on Cloud environment for Fast Prediction in Big Data

As the data is growing at faster rate and becoming "Big Data", the computation speed for prediction and other operations is predictable. For solving complex ML problems across a wide range of industries and applications, cloud environment is most appropriate which is virtually available with unlimited computational and storage capacity on pay-as-use model.

C. Balanced and Imbalanced Datasets

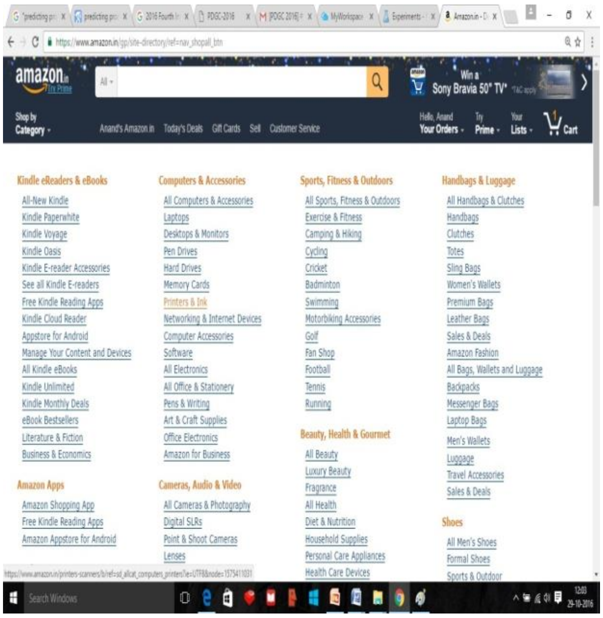
For classification analysis, real-world data from structured, semi-structured and unstructured databases used is often imbalanced. So, to predict something significant from such datasets, machine learning researchers often use artificially balanced data in increasing new techniques and algorithms.

D. Cost Effectiveness

The major costs of data analysis includes: Processing and Memory cost for learning and testing predictive models; Additional processing and analysis for optimizing parameters. Cloud computing eradicate the need for dedicated high-power computers / servers by making it possible to purchase memory, processing power, etc. only as needed, and quite in cost-effective manner. In the next section cloud computing is briefly introduced.

2. LITERATURE REVIEW

From past decades E-commerce is growing rapidly as it enables consumers to acquire any product within few clicks. The key component required for success of online shopping platform is their ability to retrieve appropriate products for the consumers very quickly. E-commerce firms got enhanced rapidly with the application of machine learning techniques such as classification and association rule mining. Machine learning tasks became a challenging job because of few factors such as ever growing amount of data for classification and constraints on response time. A snapshot of a famous shopping website [www.amazon.in], signifying an option of shop by category, in Figure 2.1. The website has apply association principle and signify the frequently bought items like mobile covers and protective glass etc., along with a mobile.



Kindle eReaders & eBooks
All-New Kindle
Kindle Paperwhite
Kindle Voyage
Kindle Oasis
Kindle E-reader Accessories
See all Kindle E-readers
Free Kindle Reading Apps
Kindle Cloud Reader
Appstore for Android
Manage Your Content and Devices
All Kindle eBooks
Kindle Unlimited
Kindle Monthly Deals
eBook Bestsellers
Literature & Fiction
Business & Economics

Amazon Apps
Amazon Shopping App
Free Kindle Reading Apps
Amazon Appstore for Android

Computers & Accessories
All Computers & Accessories
Laptops
Desktops & Monitors
Pen Drives
Hard Drives
Memory Cards
Printers & Ink
Networking & Internet Devices
Computer Accessories
Software
All Electronics
All Office & Stationery
Pens & Writing
Art & Craft Supplies
Office Electronics
Amazon for Business

Cameras, Audio & Video
All Cameras & Photography
Digital SLRs
Point & Shoot Cameras
Lenses

Sports, Fitness & Outdoors
All Sports, Fitness & Outdoors
Exercise & Fitness
Camping & Hiking
Cycling
Cricket
Badminton
Swimming
Motorbiking Accessories
Golf
Fan Shop
Football
Tennis
Running

Beauty, Health & Gourmet
All Beauty
Luxury Beauty
Fragrance
All Health
Diet & Nutrition
Household Supplies
Personal Care Appliances
Health Care Devices

Handbags & Luggage
All Handbags & Clutches
Handbags
Clutches
Totes
Sling Bags
Women's Wallets
Premium Bags
Leather Bags
Sales & Deals
Amazon Fashion
All Bags, Wallets and Luggage
Backpacks
Messenger Bags
Laptop Bags
Men's Wallets
Luggage
Travel Accessories
Sales & Deals

Shoes
All Men's Shoes
Formal Shoes
Sports & Outdoor

https://www.amazon.in/printers-scanners/b/ref=sd_allcat_computers_printers?ie=UTF8&node=1575411031

Figure 2.1 Screenshot of Shop by Category option at www.amazon.in

The important findings of work [23] signify that the area of customer withholding received most research attention. The most popular research areas among these are one-to-one marketing and loyalty programs. The most commonly used models for data mining in Customer Relationship Management (CRM) are Classification and Association Rule Mining based models.

Very large amounts of training data can require a great deal of computer memory and processing power for classification and regression analysis. Particularly with data representing complex non-linear behaviours, such as text, speech, handwriting, face recognition, stock price prediction, and financial forecasting, the computing bill can be quite bulky. As Machine Learning is a time consuming task, so Cloud computing paradigm can be an important alternatives to speed-up machine learning platforms [23]. The review provides a roadmap for future investigation in the field of application of data mining techniques in CRM.

The shopping platforms like Amazon, e-Bay, Walmart and Yahoo etc. categorize products into different product taxonomies which make it hard for sellers to classify. The major challenge acknowledged in E-Commerce is categorization of goods.

Zornitsa Kozareva [24] of Yahoo labs, concluded that the different taxonomies of arranging products are in use at various famous E-commerce shopping firms. Different taxonomies organize products making it tough and labor-intensive for sellers to classify the products. An automatic product categorization mechanism is proposed to address the challenge, which assigns the correct product category from a taxonomy for a given product title. In work [24], 319 categories organized into 6 levels and performance evaluation is done for 445 product titles using multiple algorithms. The best f-score of 0.88 is obtained.

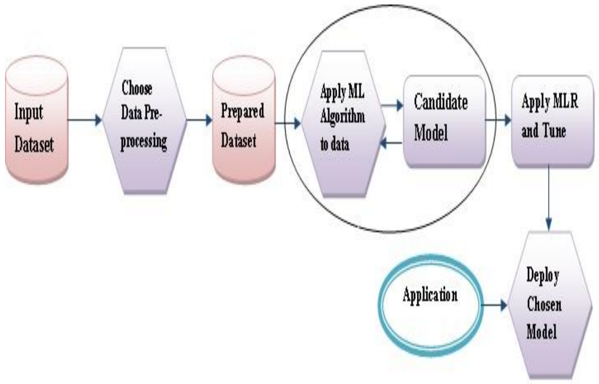
3. PROPOSED WORK

Proposed Predictive Framework for E-Commerce Product Categorization on Azure

The Proposed Framework is presented in Figure 3.1, which actually employs a Multiclass ML model with tuning of parameters for more accuracy. The input dataset is first processed and converted into a format which is most suitable for giving more accurate results. The pre-processing methods also affect the results of machine learning.

It is an iterative process as different data pre-processing techniques are available to apply on raw data. The machine learning algorithms are iteratively applied in the next step, and candidate model is determined. For Prediction, the ML algorithms typically apply some statistical analysis like regression or more complex approaches like decision forest to the data. The data scientist chooses the best ML algorithms and decides which aspects of the prepared data should be used to generate more useful results.

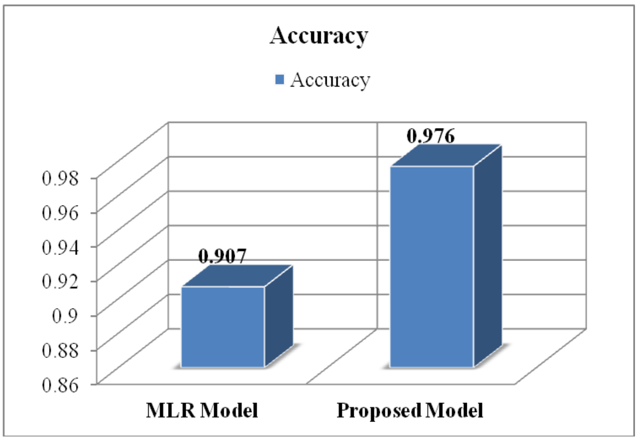
Here in the proposed framework, the hyper parameter tuning is also applied to the model for better predictive accuracy. Application of tuning method provides better classification of products or simply better predictive accuracy than using the model alone. At last the classifier (model) is deployed and tested on test dataset. To tune the ML model, a validation dataset is required.



The steps that are followed are given below:

1. Create New Resource within Machine Learning Analytics solution.
2. Import/Upload the dataset.
3. Pre-process the dataset. The step is optional in case the data is already preprocessed.
4. Split and partition the dataset into training, validation and test set. In our experiment Training, Validation and Testing are taken 25%, 25% and 50% respectively.
5. Identify categorical attributes from among given features except target feature and cast them into categorical features.
6. Apply Machine Learning Algorithm to Train the model. Here we applied Multiclass Logistic Regression (Machine Learning) Algorithm to Train the model.
7. Tune the model by adjusting the algorithm hyper parameters.
8. Apply "Score Model" to Score both the Models (Classifiers) i.e. Simple Multicast Logistic Regression (MLR) and Tuned MLR with standard metrics.
9. Apply "Evaluate Model" Compare both the models.
10. Run the experiment steps 2 through 9 to run the experiment.

From the above experiment we get the result as shown in figure:



4. CONCLUSIONS

The companies doing online business wants to utilize machine learning potential to build unmatched competitiveness in the market. In this paper, we proposed an Azure ML framework for E-commerce product categorization. The model used optimized Multicast Logistic Regression algorithm to train the classifier. The evaluation results show that the proposed classifier performs better in terms of accuracy. We have performed experiment by tuning the original LR based model parameters. In this work, we demonstrated the performance of a Cloud based classifier framework that maximizes overall classification accuracies independent of computational resource limitations.
Most of the datasets such as human genome, social networks can be classified as big data. The proposed research can provide potential approach for training and testing of big data for addressing multi-class classification problems. So, further research will repeatedly evaluate the framework with different ML algorithms, optimization parameters, ensemble methods and e-commerce databases. In future the model can be optimized to handle imbalanced datasets from various sources and domains. Also, the model can be modified for applying on Hadoop MapReduce [34] platform. Further research can be done for refining the model, for classification of data with varied inputs like images, tech specification etc. Other learning approaches that are based on incremental learning can be applied. In addition, a formal approach needs to be discovered to suggest sub/hierarchical categories to optimize the classifier performance.

ACKNOWLEDGEMENT

Our thanks to the expert Professor Amit saxena, who have contributed towards development of this paper.

REFERENCES

[1] Pine II, B.J. and Gilmore, J.H. “The Experience Economy” Boston: Harvard Business School Press, 1999.

[2] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems 25 (6) (2009) pp. 599–616.

[3] Wentao Liu, “Research on Cloud Computing Security Problem and Strategy”, 978-1-4577- 1415-3/12, IEEE 2012

[4] Rodrigo N. Calheiros, Adel Nadjaran Toosi, Christian Vecchiola, Rajkumar Buyya, “A coordinator for scaling elastic applications across multiple clouds”, Elsevier - Future Generation Computer Systems 28 (2012) 1350–1362

[5] National Institute of Standards and Technology (2011) NIST cloud computing reference architecture: Version 1. NIST Meeting Report

[6] P. Mell, T. Grance, “The NIST Definition of Cloud Computing, National Institute of Standards and Technology”, ver. 15, 9 July 2010.

[7] J. K. Han, Micheline, Data mining: concepts and techniques: Morgan Kaufmann, 2001.

[8] P. Simon, Too Big to Ignore: The Business Case for Big Data: John Wiley & Sons, 2013.

[9] I. H. a. F. Witten, Eibe Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann 2005.

[10] C. M. Bishop, Pattern recognition and machine learning: Springer, 2006.

[11] J. a. H. Friedman, Trevor and Tibshirani, Robert, The elements of statistical learning vol.1: Springer series in statistics Springer, Berlin, 2001.

[12] Andy Liaw and Matthew Wiener, “Classification and Regression by random Forest”, R News, ISSN 1609-3631, Vol. 2/3, December 2002

[13] Belady C. “In the data center, power and cooling costs more than the it equipment it supports” 2007. URL:http://www.electronics-cooling.com/articles/2007/feb/a3/.

[14] Clark C, Fraser K, Hand S, Hansen JG, Jul E, Limpach C, Pratt I, Warfield A. Live migration of virtual machines. Proceedings of the 2nd Symposium on Networked Systems Design and Implementation (NSDI 2005), USENIX, Boston, MA, USA, 2005.

[15] Google, Google Apps. http://www.google.com/apps/.

[16] Salesforce. Salesforce CRM applications and software solutions. http://www.salesforce.com/eu/crm/products.jsp.