International Journal of Electrical, Electronics and Computer Engineering **6**(1): 45-50(2017)

ISSN No. (Online): 2277-2626

Predictive Modelling for E-Commerce Data Classification Tasks: An Azure Machine Learning Approach

***Kajal Govinda Fegade*[*], *Dr. Varsha Namdeo*[**] *and Prof. Ravindra Gupta*[**]**

**Research Scholar, Department of Computer Science & Engineering,*
RKDF Institute of Science & Technology, Bhopal, Bhopal, (Madhya Pradesh), INDIA
[**]Associate Professor, Department of Computer Science & Engineering,
RKDF Institute of Science & Technology, Bhopal, Bhopal, (Madhya Pradesh), INDIA

(Corresponding author: Kajal Govinda Fegade, kajalfegade28@gmail.com)
(Received 27 November, 2016 Accepted 08 January, 2017)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The predictive analytics is intensely integrated into the current society. From spam email filtering, to predicting movies you like based on reviews, to categorize product within e-commerce data, to predict purchase behavior, the result will be determined by the use of this technology. But the computational challenges posed by growing datasets in real-world, can't be handled by a single computer. Cloud computing platforms, on the other hand, are able to handle large volumes of data for information retrieval and prediction tasks. When making scientific predictions, Machine Learning (ML) has unique ability to evaluate large number of variables than a human possibly could do. Again ML is a time consuming task, so Cloud computing paradigm proved to be an important alternatives to speed-up machine learning tasks. Combining the advantages like handling large volume of data, speed of execution, scalability and use of exciting new technologies like Azure Machine Learning Studio help to solve critical business problems like categorization of products in E-Commerce domain. In this paper, we will explore the necessity of Machine Learning in E-commerce domain and use of Cloud platform for performing predictive analysis. This work demonstrates predictive analysis for classifying E-commerce product data in real and leading cloud computing platform: Microsoft Azure ML Studio. Two predictive paradigms are built using popular ML classification algorithms on real cloud platform and evaluated on basis of Accuracy. The algorithm used for classification purpose is Multiclass Decision Forest and Multiclass Logistic Regression. The performance of both the models on small dataset is evaluated on basis of Classification Accuracy. For small dataset Multiclass Decision Forest is winner. Developing and building a predictive model with ML classification on real cloud platform and measuring its effectiveness will be the key focus of this work.

Keywords: Microsoft Azure, Machine Learning, E-commerce, Predictive Analytics, Classification, Cloud Computing, Data Mining

I. INTRODUCTION

Data mining tasks is already becoming vital in E-Commerce (EC) domain because it is recognized as important enablers of EC. The major reasons that flourished the data mining in this domain are, growing data that should be handled or analyzed by organizations and companies. Another reason is to gain knowledge about their customers and trends in marketing. Data Mining (DM) can be applied in many ways to grow the business of EC companies. Identifying the tendency of a specific customer to buy a product of specific category is a type of the knowledge and it is also regarded as an important basis for one-to-one marketing. As the quantity of consumer products is increasing the ability of classifying these is crucial. For the success of such web platforms the key component is their ability to retrieve desired products for the consumers very quickly.

From the data mining perspective, prediction of product category in EC data is a classification problem. Predictive modelling involves application of Machine Learning (ML) techniques for Classification and other DM tasks. A snapshot of a famous shopping website [1] suggesting an option of shop by category, in Fig. 1. The website is applying classification principle and suggesting the consumer products that are grouped in various categories like: Books, Sports, Fitness & Outdoors, Handbags & Luggage, and Beauty, Health & Gourmet etc.

The factor such as ever growing amount of data for classification and constraints on response time, have made DM tasks a challenging job in the EC domain. So, to override the constraints on size of data to be classified and computational performance, the choices of cloud computing platforms is available.

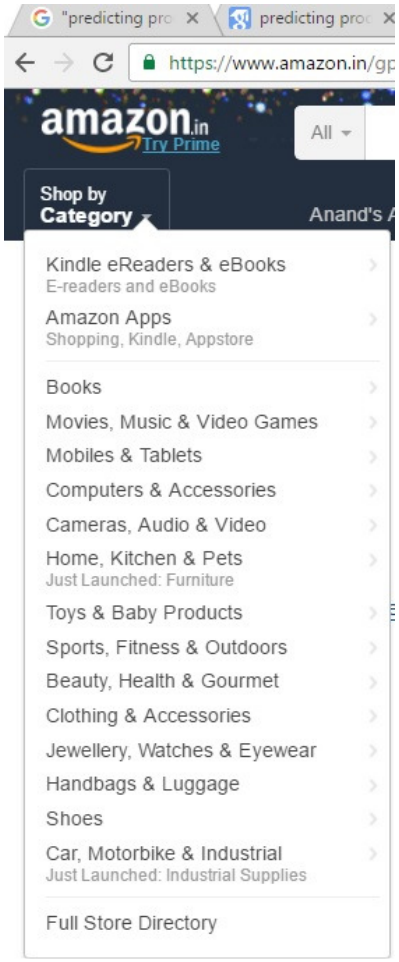

Fig. 1. Screenshot of Shop by Category at EC Website www.amazon.in.

Classification and regression analysis on very large amounts of training data can require a great deal of computer memory and processing power. Especially with data representing composite non-linear behaviours, as with text, speech, handwriting, face recognition, stock price prediction, and financial forecasting, the computing bill can be quite large. However, the emergence in recent years of cloud computing changes everything. The shopping websites like Flipkart, Amazon, Myntra e-Bay and Yahoo etc. systematize products into diverse product catalogue which make it difficult for sellers to categorize.

The categorization of goods is major challenge identified in E-Commerce.

There are numerous classifiers based on various algorithms are available for product classification. Sometimes, a classifier might override the others in classification performance for a specific set of data and sub-methods involved. In general, it is can't be said that one method always outperforms all the other methods for every possible situation. Here in this work two predictive models are proposed and evaluated and compared on basis of accuracy.

Section-2 presents few papers that are reviewed to investigate the application of machine learning in E-commerce domain, need of Cloud platforms for analyzing E-commerce data and importance of ML in predictive analysis. The work also explores the research areas in e-commerce domain. Azure Cloud based Predictive modelling for predicting product category within E-commerce product dataset using two algorithms is proposed and demonstrated in section – 3. The performance evaluation of models is presented in section – 4. Finally the paper is concluded in section – 5.

II. LITERATURE REVIEW

It is essential that the goods are listed in accurate catalogues so that users find their products in appropriate categories. For flourishing business and product sales the product categorization at E-commerce sites is a necessity. The paper [3] explores the experimental outcomes that were conducted with various feature classification methods in combination with three main classifiers SVM, Naïve Bayes, K-Nearest Neighbors, along with LDA which is an unsupervised topic classifier in documents.

Kozareva [4] studied various different taxonomies of systematizing products that are in use at those well-known shopping platforms. Diverse taxonomies of products are making it difficult and laborious for sellers to group the products. To deal with confront an automatic product categorization means is proposed, which allocate the correct product class from a catalogue for a given product title. The evaluation for performance is done for 445 product titles in 319 categories and 6 levels using several algorithms. The best f-score of 0.88 is obtained.

The work [5] focuses the troubles and challenges in handling Big Data classification and prediction of attack in network traffic data. The prediction framework requires implementation of ML approaches to capture global knowledge of the traffic patterns. The author also discussed that Big Data properties put significant challenges to implement ML frameworks.

The authors [6] analyzed the recommender systems at six leading websites and examined how recommender systems help E-commerce sites increase sales. The authors tried to explain that how recommender systems are related to some traditional database analysis techniques. The catalogue of recommender systems is created by studying the consumer inputs, the necessary knowledge from the database, the ways the recommendations are offered to consumers, recommendation creation technologies, and personalization level of the recommendations. The privacy implications of such recommender systems are examined.

The work [7] indicates that the area of customer retention attracted most research thought. One-to-one marketing and loyalty programs are considered most popular research areas. In Customer Relationship Management (CRM) Classification and Association Rule Mining based models are most common in DM tasks. Authors have considered that cloud computing platforms are a valuable alternative to speed-up the ML tasks. The review provides a roadmap for future research in the field of application of DM techniques in CRM.

The authors in their work [8] presented a supervised learning method for classifying products into a set of known categories. For building classifier features the product catalog information from different distributors on Amazon.com is used. The purpose is to show the improvement in automation of product categorization.

By exploiting the existing category of a product and ML techniques the authors [17] provided a semi-automatic solution to the re-classification problem, i.e., re classifying an already categorised product in to another taxonomy. They employed a huge number of classifiers and filters to figure out which category is best suited for a product. 76% accuracy is achieved in their classification attempts and final detection of category involves manual step.

The authors [2] have investigated how CC model influenced the field of ML. The environments like R, Octave and Python along with well-liked statistical tools are now fixed in the cloud as well. IaaS providers like Microsoft Azure [16], Amazon Web Services (AWS) [9] and 'Google Cloud' [10] platform now offer access to virtually unlimited computing power on demand, in the form of parallel server clusters that can be used for an hourly fee. The popular mathematical tools and libraries are already deployed in cloud. The existing platforms also allow users to create a Hadoop [11] cluster in the cloud and run jobs on it. Large number of libraries is now available for distributed implementations of ML algorithms for complex systems, data mining and data analytics.

III. PROPOSED PREDICTIVE MODELS ON AZURE ML PLATFORM

Framework Used

The framework used for development of Predictive models is simply based on Data Mining approach. The framework shown in figure 2 showing only major steps and, forms the basis of predictive or classification analysis that has been done on E-commerce data. The major steps are defined below:

I. Select / Load Input data set

II. Pre-processing

-Apply Feature extraction for finding relevant features

-Received prepared dataset as output

-Generate Training and Testing data set

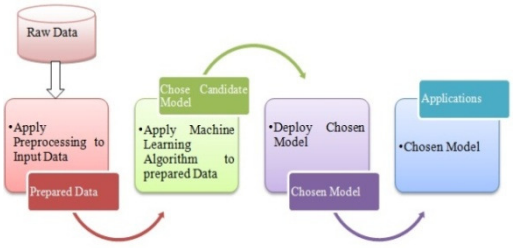

Fig. 2. Flow diagram.

III. Build the Classifier / model by applying Machine learning algorithm to the "training" dataset

IV. Apply Multiple Classifiers on testing data set. Perform / Obtain Prediction (classification) of the testing set.

V. Utilize the "test" set predictions to calculate all the performance metrics (Measure Accuracy and other parameters)

VI. Chose Candidate model (best classifier) on the basis of evaluation.

All the steps done on data mining tools those are available on Azure.

Deployment and Execution of Implemented Classifiers

Based on the generalized framework a predictive model is built over Microsoft Azure ML platform using two ML algorithms: Multiclass Decision Forest [13] and Multiclass Logistic Regression [14]. Here the 1% of dataset [15] provided, is uploaded. The dataset contains total of 618 instances/rows and nine categories of products i.e. Class – 1 to 9. The models build are shown in Fig. 3.

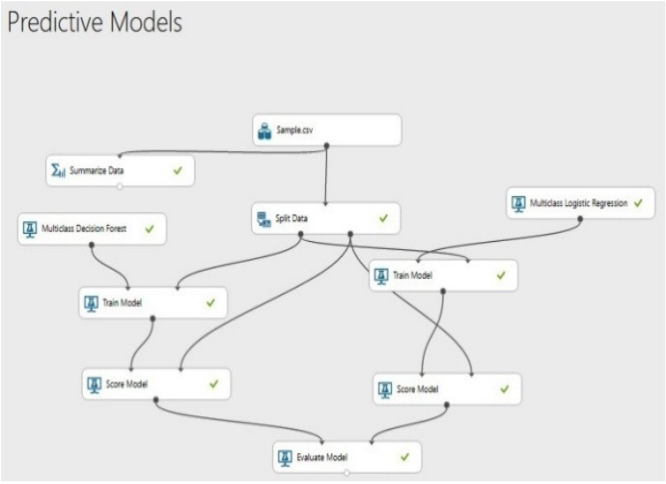

Fig. 3. Actual Predictive Classifiers built at Azure.

Following are the chief experimental steps for deployment and execution of Predictive Classifiers:
(i) Create New Resource: Machine Learning Analytics solution.
(ii) Importing Data: Import/Upload the dataset.
(iii) Preprocessing: Pre-process step is omitted as the dataset is already processed.
(iv) Split and Partition Dataset: Randomly split and partition the data into 30% training, and 70% testing, using the 'Split Data' module.
(v) Apply Machine Learning: Apply (1) Multicast Decision Forest, and (2) Multiclass Logistic Regression, (Machine Learning) Algorithms to Train the models. Here metric for measuring performance is also configured.
(vi) Score the Model: Apply "Score Model" to Score both the Models (Classifiers) with standard metrics.
(vii) Evaluate Model: Apply "Evaluate Model" Compare both the models. The 'Evaluate model' scores the classification model with standard metrics. It also visualizes the results through confusion matrix (see Figure 4 & 5) after running the experiment.

IV. PERFORMANCE EVALUATION

Performance metric determines how well the DM algorithm is achieving accuracy on a given dataset. For example, if we apply a classification algorithm on a dataset, we first check to see how many of the data points were classified correctly. This is a performance metric and the formal name for it is "Accuracy." Accuracy is defined as the quantity of correctly classified instances divided by the total number of instances:

$$Accuracy = \frac{Number\ of\ correct\ Predictions}{Number\ of\ Instances} \quad (1)$$

Table: 1 showing the classifier Accuracy achieved by two classifiers. Also the evaluation results for predicted class of known instances from test set are determined. The confusion matrix for evaluating results of both classifiers: Multiclass Decision Forest (MDF) and Multicast Logistic Regression (MLR) are shown in figure 4 and 5 respectively.

Table 1: Classifier Results: Accuracy.

Classifier Results: Accuracy		
Classifiers	***Overall Accuracy***	***Accuracy (%)***
Multicast Decision Forest based Classifier	0.891	89.1
Multiclass Logistic Regression based Classifier	0.665	66.5

The confusion matrix represents actual and predicted class on vertical and horizontal scale respectively. The results depicts that classifier based on Decision Forest identifies those classes well which are not even recognised by Logistic Regression based classifier. Class 1 and 4 is not correctly classified by LR based classifier while other classifier classified it up to certain level. Here Multiclass Decision Forest algorithm performed well for small set of instances. Except for two classes i.e. 1 and 4, MDF achieved classification accuracy of more than 75%.

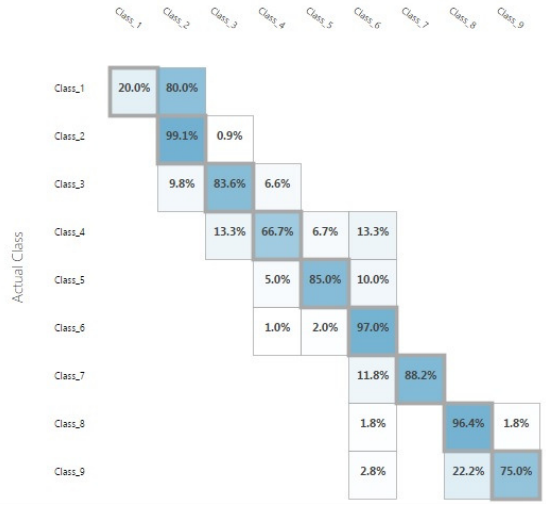

Fig. 4. Confusion Matrix for Multicast Decision Forest Algorithm.

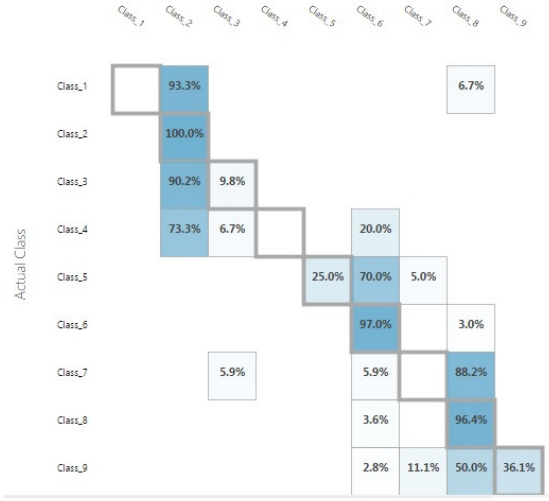

Fig. 5. Confusion Matrix for Multicast Logistic Regression.

V. CONCLUSION, CONTRIBUTIONS AND FUTURE DIRECTIONS

The machine learning process isn't especially simple. To make life easier for people doing machine learning, Cloud provides several different components. Several Cloud platforms like Azure [12] Machine Learning was designed for applied Machine Learning and enables the process of creating ML models. Cloud platform uses best-in class algorithms and a simple drag-and-drop interface along with easy deployment service.
Our main contributions are synthesized below:

1) The importance of predictive analysis in E-commerce domain is studied along with need of cloud platforms for analyzing such data is established.
2) The work also proposed and demonstrated generalized predictive modelling. Two predictive models for predicting product category within E-commerce product dataset are demonstrated. The models are deployed on Azure Cloud. Predictive models are built using popular ML classification algorithms and evaluated on basis of Accuracy.
3) The classification accuracy of two popular ML algorithms is presented in work.
In general, it can't be said that one method always outperforms all the other methods for every possible situation. Here Multiclass Decision Forest algorithm performed well for small set of instances. As a future work the same classifiers will be tested and evaluated for large datasets. Also, the improvements through various pre-processing and ensemble methods would be suggested. In future we propose a dynamic ML model at cloud platform, for product categorization of e-commerce website.

ACKNOWLEDGMENT

This research was guided by Prof. Ravindra Gupta Associate Professor, at Department of Computer Science & Engineering, RKDF Institute of Science & Technology, Bhopal. I would like to thank him for providing insight and expertise that greatly assisted the research. I would also like to show my gratitude to the Dr. Varsha Namdeo, Associate Professor & Head, Department of Computer Science & Engineering, RKDF Institute of Science & Technology, Bhopal for sharing their pearls of wisdom with us during the course of this research, and we thank Prof. Anand Motwani, Associate Professor & Head, Computer Science & Engineering Department, at Sagar Institute of Science, Technology & Research (SISTec-R), Bhopal, India for providing knowledge assistance about real Cloud Computing technologies which greatly improved the manuscript. I am also thankful to my family and colleagues for support.

REFERENCES

[1]. www.amazon.in
[2]. Daniel Pop, "Machine Learning and Cloud Computing: Survey of Distributed and SaaS Solutions", https://www.researchgate.net/publication/257068169.
[3]. Srinivasu Gottipati and Mumtaz Vauhkonen, "E-Commerce Product Categorization",

[4]. Zornitsa Kozareva, "Everyone Likes Shopping! Multi-class Product Categorization for e-Commerce", Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1329–1333, Denver, Colorado, May 31 – June 5, 2015, Association for Computational Linguistics.

[5]. Suthaharan, S., "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning. Performance Evaluation Review", 41(4), ACM 2014

[6]. J. Ben Schafer, Joseph A. Konstan, John Riedl, "E-Commerce Recommendation Applications", *Data Mining and Knowledge Discovery*, **5**, 115–153, 2001, Kluwer Academic Publishers, Netherlands

[7]. E.W.T. Ngai, Li Xiu, D.C.K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", *Expert Systems with Applications* **36** (2009) 2592–2602, Elsevier.

[8]. Sushant Shankar and Irving Lin, "Applying Machine Learning to Product Categorization",

[9]. https://aws.amazon.com/machine-learning/

[10]. https://cloud.google.com/ml/

[11]. Apache Hadoop Website http://hadoop.apache.org/

[12]. David Chappell, "Introducing Azure Machine Learning", A Guide For Technical Professionals, Sponsored by Microsoft Corporation, 2015.

[13]. https://msdn.microsoft.com/en-us/library/azure/dn906015.aspx

[14]. https://msdn.microsoft.com/en-us/library/azure/dn905853.aspx

[15]. Otto Group Product Classification Challenge. Available: https://http://www.kaggle.com/c/otto-group-product-classification-challenge

[16]. www.portal.azure.com

[17]. Abels S. and Hahn A., "Reclassification of electronic product catalogs: The"apricot" approach and its evaluation results. InformingScience",9:31–47 (2006).