Predictive Model for Multiclass Classification of E-Commerce Data: An Azure Machine Learning Approach

Kajal Govinda Fegade
Research Scholar
Department of Computer Science & Engineering
RKDF Institute of Science & Technology, Bhopal

Ravindra Gupta
Associate Professor
Department of Computer Science & Engineering
RKDF Institute of Science & Technology, Bhopal

Varsha Namdeo, PhD
Associate Professor
Department of Computer Science & Engineering
RKDF Institute of Science & Technology, Bhopal

ABSTRACT
Electronics Commerce (E-Com) is one among the various business methodologies that addresses the growing requirement of business organizations, and customers. The E-commerce industry is one of the world's leading and growing industries with market worth around $22.1 trillion globally. Through E-Com, companies are developing the competence in business domain. The business giants like Amazon, Flipkart, etc. utilizing Machine Learning (ML) potential to make matchless competitiveness in the market through data analytics and business intelligence. ML has empowered businesses by analyze the data collected through various sources including social media reviews. Data scientists are in huge demand in E-Commerce market researches because predictive data analytics based on ML can enhance sale prospects and discover the reasons of customer churn, by analyzing customer's reviews and click-through actions, preferences and past purchase history, in real-time.

Massive increase in the volume, variety and velocity of data generated through various businesses or E-Commerce platforms pose a huge computational and storage challenge for data analysis and intelligence tasks. Addressing the computational and storage needs for business intelligence tasks, cloud computing paradigm is evolved. The data and computation can be distributed to any Cloud computing environment with minimal effort nowadays. Also, Cloud computing paradigm turned out to be valuable alternatives to speed-up machine learning platforms.

The work, first discusses the 'E-Commerce advantages', 'Importance of Machine Learning in E-Commerce Domain', 'Cloud Computing and Need of Cloud platforms for Machine Learning tasks'. Also, the background for 'E-Commerce Product Data Classification Task' 'is established. Introduction to multiclass classification and the literature survey for various classification tasks is presented. Finally, a Predictive Model for E-commerce Data Classification Task is proposed and deployed over Microsoft Azure Cloud. The proposed model predicts the Product Class from a large product dataset released by a well-known e-commerce company for a competition. The proposed model is build using 'Neural Network' (Multiclass) and R-Script module for better convergence. The obtained results are compared with benchmark model "Multiclass Logistic Regression" and evaluation is done on basis of prediction accuracy. Proposed work also demonstrates the use of one of the foremost cloud environments for machine learning. The results attained by the proposed model are promising and the paper also mentioned the future research work in the field.

Keywords
Predictive Modeling, Computer Vision, E-commerce, Data Classification, Machine Learning, Microsoft Azure, Cloud Computing.

1. INTRODUCTION

1.1 Introduction to E-Commerce
The E-Commerce (EC) industry is one of the world's leading and growing industries with market worth around $22.1 trillion globally, according to latest "Union Nations Conference on Trade and Development" (UNCTAD) [1] estimates. Electronic Commerce (EC) is one among the various business methodologies that addresses the growing requirement of business organizations, and customers. It is a way of improving services at reduced cost while ensuring the speedy delivery. E-Com offers various advantages like: Non-Cash Payment; Frequent Service availability; Customer Reach (Marketing); Improved Sales & Customer Communication; Support & Inventory Management. EC also helps the government to deliver public services such as healthcare, education, social services at a reduced cost and in an improved manner.

1.2 Importance of Machine Learning in E-Commerce Domain
Through EC companies are developing the competence in business domain. The business giants like Amazon, Flipkart, etc. utilizing Machine Learning (ML) potential to make matchless competitiveness in the market through data analytics and business intelligence. ML has empowered businesses by analyze the data collected through various sources including social media reviews. Predictive data analytics based on ML can enhance sale prospects and discover the reasons of customer churn, by analyzing customer's reviews and click-through actions, preferences and past purchase history, in real-time. So, Data scientists are in huge demand in E-Commerce market researches.

To address this data and information explosion, EC stores are applying machine learning to customization principles to their presentation in the on-line store [2]. The application of ML models to mine large datasets grows stronger with the extraordinary raise in the logs of personal data and consumer web browsing behavior. Such ML based models are generally designed to gain insights from large unstructured and structured datasets.

EC data classification and need of cloud platforms for such tasks is established and summarized here. Running Data Mining and performing Business data analysis for business

intelligence tasks like Classification. Prediction and Regression on huge volumes of data can require a great deal of memory and processing power. So, the providers like Amazon Web Services (AWS), IBM BlueMix and 'Google Cloud' platforms are now offering access to said services on demand, in the form of clustered parallel servers on the basis of an hourly fee. These cloud based ML platforms provides a wide range of suitable enterprise and business applications, with unlimited computational and storage capacity on pay-as-use model.

The languages like Python, R etc. along with various statistical tools are embedded in Cloud platforms to develop new techniques to handle such datasets. Cloud computing is briefly introduced below.

1.3 Cloud Computing

Cloud Computing shown significant move in modern 'Information and Communication Technology' (ICT) by providing enterprise applications and powerful architecture to perform scalable and complex computing. The benefits of cloud computing comprise of virtualized resources, distributed and parallel processing, security, and service integration with scalable infrastructure capabilities. CC must also ensure Quality of Service (QoS) requirements negotiated through Service Level Agreements (SLAs) agreed between interacting entities i.e. cloud providers, consumers and brokers [17]. So, Cloud computing paradigm turned out to be valuable alternatives to speed-up data mining and business intelligence tasks [3].

A number of choices are available for different workload management, performance and computational requirements and Cloud Computing (CC) grew out as requirement. The Cloud Computing [4, 5] is a novel model for computing that came from distributed, grid, and parallel computing along with a mix of virtualization technology. This paradigm of computing delivers services through a pool of highly virtualized resources and appears as a single large resource. The features like dynamic scaling of applications, software platforms, and infrastructures according to agreed Service Level Agreements (SLAs) [6] are allowable to customers. Cloud computing can not only lessen the cost and constraint for automation and computerization by people and enterprises but can also offered reduced infrastructure maintenance fee, user access and efficient management [7]. In view of said advantages, companies like Amazon, IBM, Microsoft etc. developed a large number of applications that supports varying cloud platforms thus resulted in a tremendous increase in the scale of data generated and consumed by such applications. The popular statistical tools and environments like Octave, R and Python are now embedded in the cloud as well.

1.4 E-Commerce Product Data Classification Task (Background)

Because of the growing attractiveness and acceptance of EC websites, users face an ever growing burden in actually selecting the right product from the large number products offered. So for a pleasant shopping experience, the product must be organized in categories where users can easily search and buy with high confidence. Similarly, such shopping platforms are facing difficulties in categorization of products into various categories. So, such platforms need efficient methods and system that predicts product's category. The application of machine learning techniques boosted the field. But the factors such as ever rising data for applying ML algorithms puts constraints on response time and made ML tasks a challenging job in the EC domain. A snapshot of a famous shopping website [8], suggesting an option of shop by category, in Figure 1. The website is applying classification principle and suggesting the consumer products various categories like: Books, Sports, Fitness & Outdoors, Handbags & Luggage, and Beauty, health & Gourmet etc.

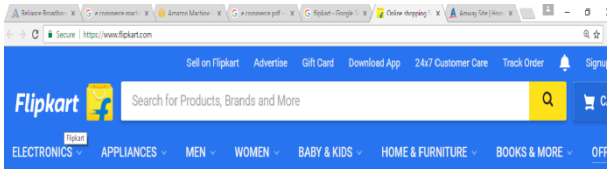

Fig 1: Screenshot showing various Product Categories at www.flipkart.com

2. LITERATURE REVIEW

The Comprehensive product data classification means fitting wide-range of product choices in a well-structured class of product catalog. This is necessary for easy navigation through the Products and good sales of products. This can be also achieved by hierarchical classification of products and through recommender systems, based on suitable ML techniques. In past decade, E-commerce firms employed following advertising and marketing strategies like: Remarketing; Website Personalization; Deal Customization; Email personalization. In this section we present a survey of relevant researches.

In our previous work [9], necessity of Machine Learning in E-Commerce domain and utilization of Cloud platform for performing predictive analysis is analyzed. The work demonstrates predictive analysis for classifying EC product data in real and leading cloud platform: Microsoft Azure. Two predictive paradigms are built on real cloud platform using popular ML classification algorithms: Multiclass Decision Forest and Multiclass Logistic Regression (MLR), and evaluated on basis of Accuracy. The performance of both the models is evaluated on basis of Classification Accuracy on one percent data provided at [10].

The authors of work [11], tested 11 data mining classification techniques, and found out that decision table is one of the best classifiers giving high accuracy. Also, the paper proposed a recommender system based on decision table classifier to help customers' find their products on EC websites. The dataset chosen is composed of ordering log file for 3 months. It consists of 304 instances and 26 attributes. The classifiers used are: K Star, Filtered Classifier, J48, Naïve Bayes, and Classification via Clustering, Ridor, JRIP, Decision Table, Bayes Net, END and Simple Cart.

Cloud-based platforms are capable of handling large volumes of information manipulation tasks, thereby necessitating their use for large real-world data set computations. The work [12] focuses on building a Generalized Flow within the Microsoft Azure Machine Learning Studio (MAMLS) (real cloud computing platform). It performs multi-class as well as binary classification to maximize the overall classification accuracy. The classification characteristics of the proposed flow are evaluated and compared on 3 public data sets with respect to existing state-of-the-art methods.

Product classification for EC sites is a necessity for successful business and product sales [13]. It is vital that the products are listed in accurate categories so that users find their products in appropriate categories. The paper [13] explores the experimental results that were conducted with various feature classification methods in combination with three main

classifiers Naïve Bayes, SVM, K-Nearest Neighbors, along with LDA an unsupervised document topic classifier.

Authors [14] presented a method for classifying products into a set of known categories by using supervised learning. For building features for classifier the product catalog information from different distributors on Amazon.com is used. The purpose is to show the improvement in automation of product categorization.

3. MULTICLASS ARTIFICIAL NEURAL NETWORK

Many ML algorithms are available in literature [15]. A more detailed description of the machine learning algorithms can be found for example in [15]. For this work we used Multiclass Neural Networks. For better understanding multiclass classification is and Multiclass Neural Networks is explained here.

In multiclass Classification, the classifier can be used to predict multiple outcomes. The following classification models can be applied to data sets that contain three or more unique class labels. In this work, we apply Multiclass Artificial Neural Network [16] for building proposed classification framework. The equation for predicting value of independent variable Y using Neural Network is represented below.

If we want to classify an instance belonging to one of K classes, then Multiclass classification with a linear neural network is a quite simple extension of the binary classification setup. We may think that our second layer node, could output 0, 1 ...K−1.

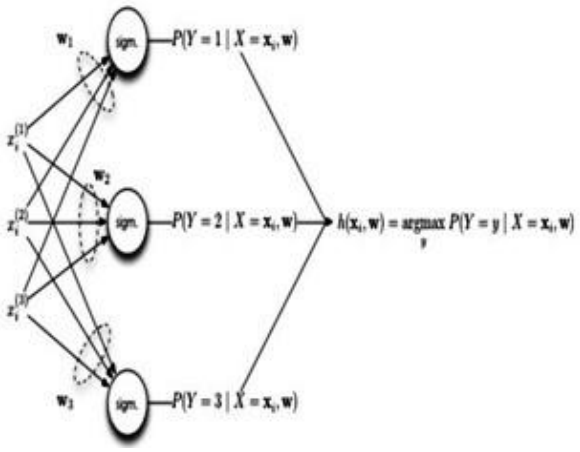

Fig 2: Multiclass Artificial Neural Network

Instead of **|w|=M** (where **M** is the number of features), here we have **|w|=MK**, **K** is number of classes. So in the example figure 2, we have **K** = 3, and 3 features, giving us 9 weights in total.

In fact, this is the neural network view of multinomial logistic regression. Recall the previous likelihood used in binary logistic regression:

$$P(Y = y \mid X = \mathbf{x}_i, \mathbf{w}) = \frac{1}{1 + \exp(-y\mathbf{w}^T\mathbf{x}_i)}$$

To extend this to K classes, we use the following likelihood:

$$P(Y = k \mid X = \mathbf{x}_i, \mathbf{w}) = \frac{\exp(\mathbf{w}_k^T\mathbf{x}_i)}{\sum_{k'}^{K} \exp(\mathbf{w}_{k'}^T\mathbf{x}_i)}$$

Dependent variable: Here the value (Y) representing the value or process, we are trying to predict or understand (e.g., product category).

Training Instances: Here the variables (Xi = 1 to n) are used to model or to predict the values of dependent variable.

Weights (w): Weights are computed by the likelihood. The weights are values, one for each data instance, that represent the type of relationship and strength with the particular data instance has to the Y. When the relationship is a strong one, the Weight is large. Weak relationships are associated with coefficients near zero.

4. PROPOSED WORK

The Proposed Model is presented in Figure 3, which actually employs a Multiclass Neural Network ML model with R Script Module for faster evaluation and lesser overall time. The dataset is already pre-processed and converted into a suitable format for obtaining more accurate results within lesser time. The pre-processing methods affect a lot in final evaluation results of ML model. It is a good practice to apply such processes on raw data. In the next step dataset is split into two subsets known as training and testing set. Generally a small fraction of dataset is taken to train the model/classifier. Here the dataset is divided into the ratio of 30:70 i.e. train: test. To build the model various ML algorithms are applied and tested iteratively in the next step and best model is determined. The purpose of Prediction is fulfilled by applying the ML algorithms involving mathematical models and statistical analysis like regression analysis or more complex approaches like neural networks and genetic algorithms to the data. The best model based on ML algorithm algorithms is chosen by data scientist to decide many aspects to generate more useful results. Application of Multiclass Neural Network provides better data classification better predictive accuracy than other models like Logistic Regression.

The actual model build using steps of proposed Predictive Model, at Microsoft Azure ML platform, is revealed in Figure 4. Snapshot of original model built over MAMLS is shown in Figure 4.In our earlier work [9] we proposed the two such models based on Decision forest and Logistic Regression algorithms. The models are applied on 1% of E-commerce Dataset. The results revealed that Multiclass Decision Forest is winner on small datasets, but on evaluating the models on large (full) dataset the Logistic Regression is performing well.

Furthermore, we compared the result of proposed model with MLR based model on the basis of Accuracy and Class based Accuracy. The class based accuracy is also obtained using confusion matrix. Here, proposed model is shown.

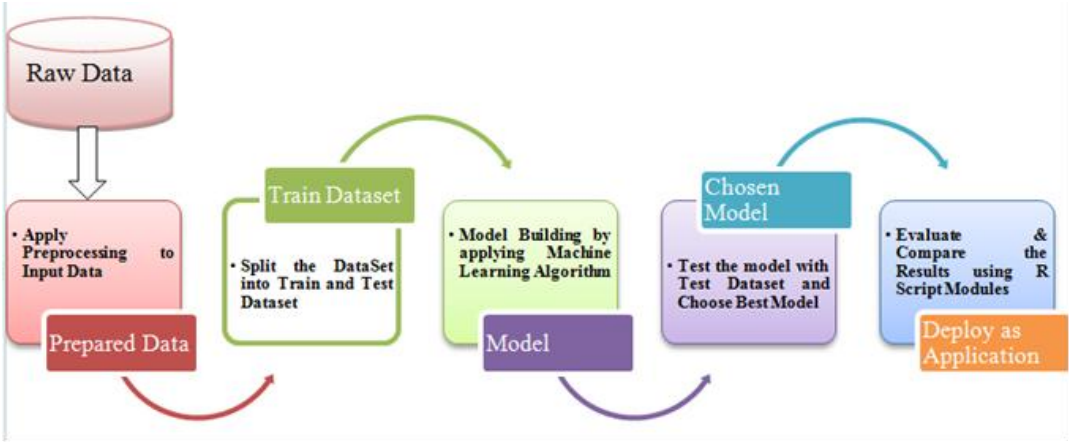

Fig 3: Proposed Predictive Model

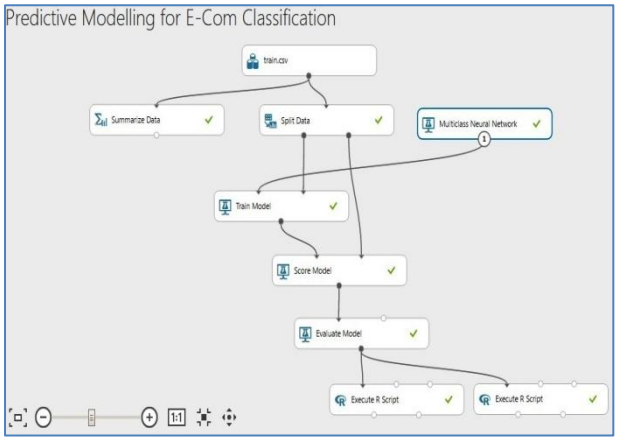

Fig 4: Model built using Azure ML

Table 1. Product Instances (Class wise)

Classes	Instances
Class 1	1929
Class 2	16122
Class 3	8004
Class 4	2691
Class 5	2739
Class 6	14135
Class 7	2839
Class 8	8464
Class 9	4955

5. EXPERIMENTAL SETUP AND RESULT ANALYSIS

A large E-commerce company: Otto Group [10] provided the dataset having 61,878 product instances in total. There are 93 features in total and each product possesses 1 or more features. The purpose of this work is to build a predictive model to classify a particular instance in correct class of product. For model building and evaluation the dataset is partition into 30% for Training, 70% for testing, hyper-tuning and evaluating the model.

The simulation parameters of Multiclass NN are shown in Figure 5.

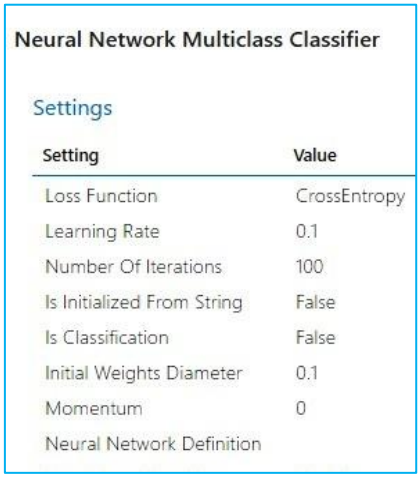

Fig 5: MNN Classifier

5.1 Parameters

The Few important parameters are:

5.1.1 Learning rate

It defines the size of the step taken at each iteration before correction.

5.1.2 Number of iterations

This specifies the maximum number of times the algorithm should process the training cases.

5.1.3 Initial weights diameter

It specifies the node weights at the start of the learning process.

6. EXPERIMENTAL RESULTS: EVALUATION PARAMETER, ANALYSIS AND DISCUSSION

In this work the evaluation metrics that taken into account are Accuracy and Class based Accuracy. Accuracy is defined as the number of correctly classified instances divided by the total number of instances:

$$Accuracy = \frac{Number\ of\ correct\ Predictions}{Number\ of\ Instances}$$

The results attained by MLR based predictive model is 91.5%, while the proposed model has attained accuracy of 98.5% as shown in Figure 6. The results also demonstrated that, with proposed model the capability of identifying individual classes is also enhanced. As it is evident from results (Table 2 & Figure 7):

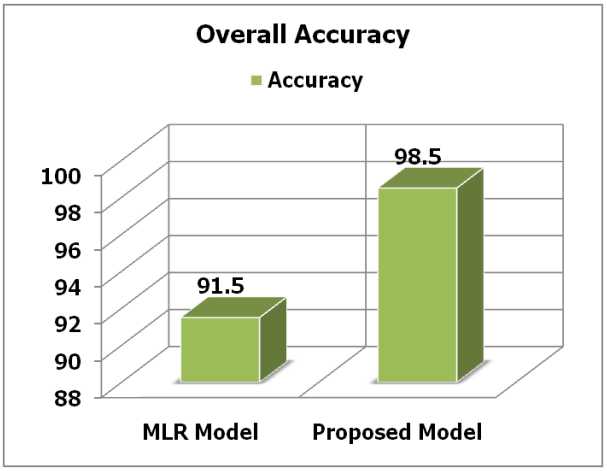

Fig 6: Comparison of Accuracy

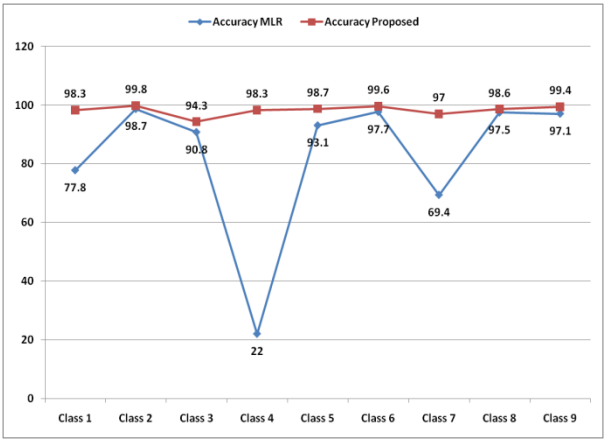

Fig 7: Comparison of Class based Accuracy

Table 2. Algorithms Results: Class based Accuracy

	Accuracy	
	MLR	**Proposed**
Class 1	77.8	98.3
Class 2	98.7	99.8
Class 3	90.8	94.3
Class 4	22	98.3
Class 5	93.1	98.7
Class 6	97.7	99.6
Class 7	69.4	97
Class 8	97.5	98.6
Class 9	97.1	99.4

7. CONCLUSION AND FUTURE WORK

In continuation to previous work [9] in which 2 predictive models for predicting product category within Ecommerce product dataset are demonstrated. In our previous work Multiclass Decision Forest algorithm performed well for small set of instances, over MLR. The application of machine learning techniques to predictive methods is not simple. Also, the computational power of single system has lot of limitations. Cloud provides ML platforms with several components for such computational intensive tasks. Cloud platforms like Microsoft Azure ML workspace is designed for applying ML i.e. for creation of ML models. Cloud platform provides best-in class implementation of techniques and algorithms with a simple drag-and-drop interface along with easy maintenance and operations service.

Most of the datasets such as text, images, human genome and social networks can now be categorized as big data. The work proposed here can provide potential approach for training and testing of big data as well as for addressing multi-class classification problems. So, further research will repeatedly evaluate the model with different optimization parameters, ensemble methods and other databases.

8. ACKNOWLEDGMENTS

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 and 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 and Head , Computer Science & Engineering Department, at Sagar Institute of Science , Technology and 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.

9. REFERENCES

[1] United Nations Conference on Trade and Development http://unctad.org/en/Pages/Home.aspx

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

[3] 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

[4] 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.

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

[6] 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

[7] L.Chih-Wei, H.Chih-Ming, C.Chih-Hung, Y.Chao-Tung, "An Improvement to Data Service in Cloud Computing with Content Sensitive Transaction Analysis and Adaptation, Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual, 2013, pp.463–468.

[8] Flipkart www.flipkart.in

[9] Kajal Govinda Fegade, Dr. Varsha Namdeo and Prof. Ravindra Gupta, "Predictive Modelling for E-Commerce Data Classification Tasks: An Azure Machine Learning Approach", International Journal of Electrical, Electronics and Computer Engineering, IJEE-ComE 6(1): 45-50(2017)

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

[11] R. A. E. D. Ahmeda, M. E. Shehaba, S. Morsya and N. Mekawiea, "Performance Study of Classification Algorithms for Consumer Online Shopping Attitudes and Behavior Using Data Mining," 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, 2015, pp. 1344-1349.

[12] M. Bihis and S. Roychowdhury, "A generalized flow for multi-class and binary classification tasks: An Azure ML approach," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 1728-1737.

[13] Srinivasu Gottipati and Mumtaz Vauhkonen, "E-Commerce Product Categorization"

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

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

[16] Multiclass Neural Network [Online] https://msdn.microsoft.com/en-us/library/azure/dn906030.aspx

[17] Heena Kaushar, Pankaj Ricchariya and Anand Motwani, "Comparison of SLA based Energy Efficient Dynamic Virtual Machine Consolidation Algorithms", International Journal of Computer Applications 102(16):31-36, September 2014.