International Journal of Engineering & Technology, 7 (4) (2018) 4388-4392

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 15246
Research paper

Analysis of customer data using hybridized machine learning technique along with data exploration methods

G.S. Ramesh [1] *, Dr T V Rajini Kanth [2], Dr D Vasumaathi [3]

[1] Assistant Professor, Department of CSE, VNRVJIET
[2] Dean R&D, Professor in CSE, SNIST
[3] Professor department of CSE, JNTUH College of Engineering ,JNTU-H
**Corresponding author E-mail: ramesh_gs@vnrvjiet.ac.in*

Abstract

The introduction of Information Communication Technologies (ICT) like POS and sensors into retail and marketing the Sales Data and Customer data are increasing day by day seamlessly without any limitation and boundaries in an exponential growth. This huge voluminous data is threatening research community to develop suitable models for the identifying Target Customers, enhancement of particular Products sales etc. The need of Business Intelligence techniques is very much required in this scenario to address the Entrepreneurs, Business community. The Machine learning algorithms are also useful to Analyze Sales Volume or to Discover Most Likely to Buy Products or to Provide Price Recommendations. This paper addresses these problems with the help of data exploration using Visual Analytics techniques apart from predictive analytics. The Machine learning algorithms like K-mean Clustering, Logistic Model tree together as a hybridized Clustered based Logistic Model tree algorithm was applied apart from machine learning Data exploration techniques. Visual Comparisons were also made along with advanced statistical techniques and summarized the results for better conclusions.

Keywords: *Business Intelligence; Data Exploration; Logistic Model Tree; Machine Learning; Visual Analytics.*

1. Introduction

Over the past 10-20 years, estimating costumer buying pattern has been considered by organizations, and it is considered as most important research topic in estimating consumer behavior. Business Intelligence is the upcoming research area used by many business organizations like Amazon, IBM, Microsoft, Google etc for the enhancement of profits in business by processing huge business data generated by E-commerce applications. There are various tools available in this domain like Microsoft BI, IBM Watson Analysis etc. In computer science Machine Learning is a domain which allows the systems to analyze data using algorithms. The data may be in large volume and diverse in nature. Machine learning automatically detects interesting and useful patterns existing in the data. The Predictive Machine Learning techniques are applied on Historical or legacy sales transactional data to forecast sales volume of new and existing products. These results are useful to business community on the trend before introduction of new products or to discontinue old products. The recommender systems will analyze customer behaviour [1] and to predict mostly to buy products. These results will help the business community i.e. wholesalers or retailers can offer right products and prices to retain their loyal customers or to make new customers to become loyal customers. The major Potential Benefits are to explore deeper insights into sales and product data and also Insight to sales opportunities apart from improved planning and inventory management. Machine learning algorithms will help us to model and predict customer buying patterns. Data Mining using Machine Learning (ML) algorithms will make the companies to explore hidden insights existing in the data sets. The Machine learning algorithms are able to provide clear approvals for which movies to watch, books to read, and products to buy. These ML Algorithms shall be one of the influential classes of technologies [2] in the years to come. Most of the business organizations found that the ML algorithms made significant impact in their KPI's (Key Performance Indicators) for sales - such as new leads, up-sales, and sales cycle times etc. Statistical techniques, algorithms are used in Predictive analytics [10] along with machine learning techniques on the provided data, to find out the future outcomes based on previous data.

2. Related work

Rajan Gupta et al [3] in their paper stated that online purchases are taking place mainly due to transparency of product prices. The dynamic pricing is used by many retailers has increased their sales and margins. They proposed to develop a generic model by using machine learning techniques to improve best price procurement by clients on e-commerce platforms. Their major focus was on register-based e-commerce trades; but, the framework can be adjusted even for online shopping. They developed statistical and machine learning models to forecast the procurement decisions built on adaptive or intent valuing of a product based on various attributes of customer purchase history, etc. They focused on client sections for forecasting procurement rather than on specific purchasers.
Yuta Kaneko et al [4] in their paper stated that their aim is to build a model for predicting retail stores transactions using Deep learning algorithm. The sales prediction model was designed using 3 years POS data from a retail store which estimates the variations in sales on next day based on the details of particular day. The deep learning method which considered the L1 regularization has attained a sales prediction correctness rate of 86%. In contrast, the

correctness reduced by around 13% when the logistic regression method was used. These outcomes specify that deep learning is extremely suitable for building models that contain multi-attribute variables.

Yi Zuo et al [5] in their paper stated that most of the researchers used linear models to measure customer response of procurement intention to estimate the weight of the features that required such as age, income and gender, product value and sales elevation. The linear methods, especially linear discriminant analysis and logistic regression analysis are mostly used as the estimation models for procurement behavior. Demand of Machine Learning algorithms significantly increased as the investigation based on linear methods is inadequate to satisfy the needs of academics and experts method for knowledge discovery and data mining. They have employed two demonstrative machine learning methods namely Baye's classifier and support vector machine (SVM) and also compared their performances when applied on the real world data.

Stefan Meinzer et al [6] in their paper stated that the automotive industry is generating around one Terabyte of objective information every hour currently. This size will expressively grow by the number of linked amenities within the business. But, client gratification identification is mainly built on subjective surveys. They provided an industrial application that offers an answer with a large practical influence to stay in the hard competition. They addressed the fundamental questions like Can unhappy clients be grouped based on data that is given during each service call? Can the displeasure factors be concluded from service procedure data? The finest result for customer unhappiness classification was 88.8% achieved using the SVM classifier (RBF kernel). Furthermore, the 46 most potential pointers for unhappiness were recognized by the evolutionary feature selection. Their scheme was capable of categorizing customer unhappiness exclusively based on the objective data that is produced by virtually every service call.

S. HanumanthSastry et al [7] in their paper stated that they have applied clustering methods for detecting deviation in product trades and also used to recognize and equate trades over a specific period of time. They have used annual trade's data of a steel firm to examine trades Volume & Value based on dependent attributes namely products, quantities sold and customers. The demand for steel goods largely depends on attributes like price, customer profile, tax issues and Discounts. They analyzed trades data using K-Means & EM methods. Their study confirmed that partition methods like K-Means & EM algorithms are well suited to analyze trades data in contrast to the methods like DBSCAN & OPTICS or COBWEB.

Vaughn Aust [8] stated that ML allows sales teams to be more efficient, better targeted and can drastically reduce the amount of time that sales reps spend on presales tasks like prospecting and communicating. Sales Managers can increase rep efficiency using ML and also allows makes them to become better coaches. ML can even be utilized before hiring reps to determine what type of candidate will likely be a good sales rep down the road. With all of these newly discovered correlations and efficiencies—like the best time to call a certain prospect, the appropriate channel to reach them on, and the right things to say ML has shorten sales cycles, allowing reps to close more deals.

Cristina STOICESCU [9] stated in her article that the study of Customer behavior interdisciplinary and evolving science domain. Due to the faster development of technology and the radical modification of life style, customers begin to have gradually varied needs. The ICT tools were used to gather and study the consumer behavior. Initially companies have done analysis of consumer behavior data and later prediction was made to understand their behavior more effectively by the application of correlation techniques.

Daniel Faggella [10] addressed in his article about Present applications of predictive analytics for promotion and publicity, role of data and machine learning, Current market investigation on predictive analytics, Prominent merchants and service providers in predictive analytics and Connected discussions and articles.

Bernard Marr [11] stated in his article that how ML can be used in multiple ways by the companies. The possibilities are to interpret customer data, Improves Sales Forecasting by comparing with the legacy data and do sales predictions for optimal utilization of resources, Predict Customer Needs i.e. Anticipation of needs of customers and Efficient Transactional Sales i.e. machineries step in to handle certain trades efforts rapidly and efficiently can release up the persons in trades force to focus on the association. The other possibilities are Sales Communication i.e. Machineries can rapidly and effortlessly answer queries about price, product landscapes or contract terms. He concluded that Machines can take care of transactional sales to release up the persons in sales force to build associations and promote their leads in ways only humans can. By looking after the regular tasks for trade staff, machineries clear the way for the sales procedure to be better and more operative.

Reid Pryzant et al [12] stated in their paper that knowing consumer responses to divergence of data is significant as business intelligence and customer attitudes. They have considered textual product descriptions as important determinants of consumer choice.

Kris Johnson Ferreira et al [13] stated in their paper that they have used methods least squares regression, partial least squares regression, principal components regression, multiplicative (power) regression, regression trees and semi logarithmic regression for prediction. They also have used LP bound algorithm for further analysis.

Robert Siwerz et al [14] stated in his project that there is a statistically important variance between the SVM, MLP and RFBN when predicting the sales in a food store department. The SVM performed lower error measures than the other two methods. Since this study used on limited data, thus, one could hardly draw the conclusion that the SVM is always the most accurate method to use for sales prediction in a food store department. However, the result of this study can indicate what methods to look at when implementing machine learning methods to predict sales in the food industry.

3. Proposed system

The original raw data set i.e. Wholesale- Customer data [15] was considered and pre-processed for required format and then subjected to Data exploration techniques like Scatter 3D plot, Boxplot, Heat map etc. and identified the nature of attributes and analysed further for potential results. Then k- Means Clustering was applied. Logistic Model Tree was applied on the resulting Clustered data set and found that this Hybridization Machine learning technique yielded better results in analyzing the results.

4. Experimental results

The data Exploration Techniques were applied on the whole sale data. The following Fig.1 shows correlation diagram using Pearson coefficient and the inference is that the attributes are correlated to themselves particularly Detergent Paper, Grocery and Milk are correlated each other.

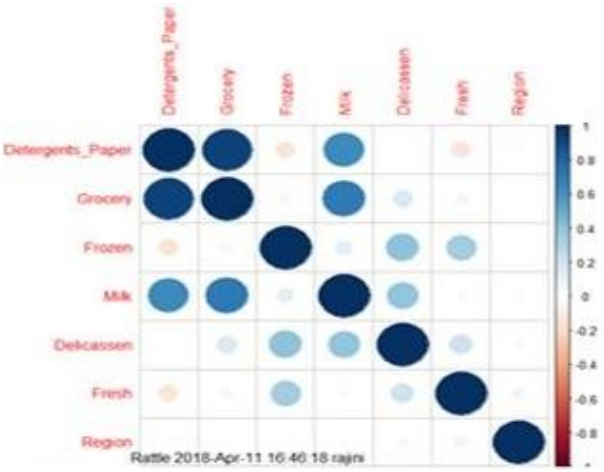

Fig. 1: The Correlation Plot Using Pearson.

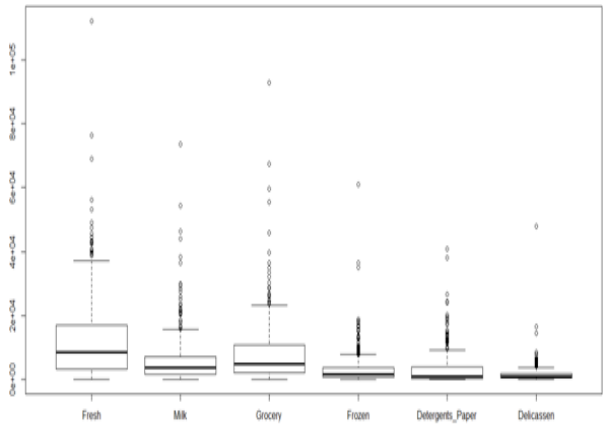

Fig. 2: Box Plot.

The outliers of the attribute were studied using Box plot shown in Fig.2 that the attributes Fresh, Milk, Grocery, Frozen, Detergents Paper has outlier data and this kind of diagram is used to show the outline of the distribution, its central value, and its variability. The Fig.2 shows that the plot of attributes Grocery vs Milk data was shown across 3 regions.

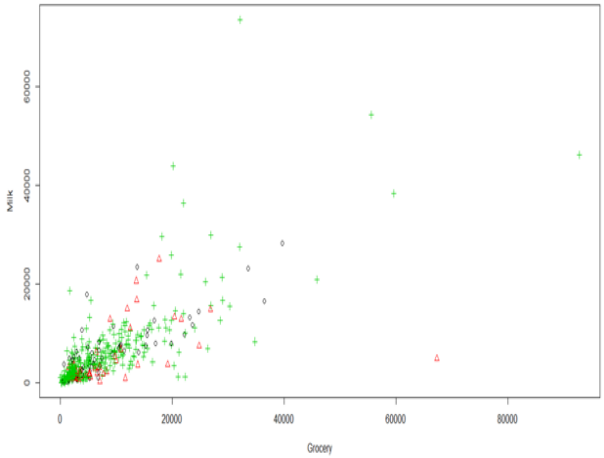

Fig. 3: Scatter Plot of Grocery vs. Milk.

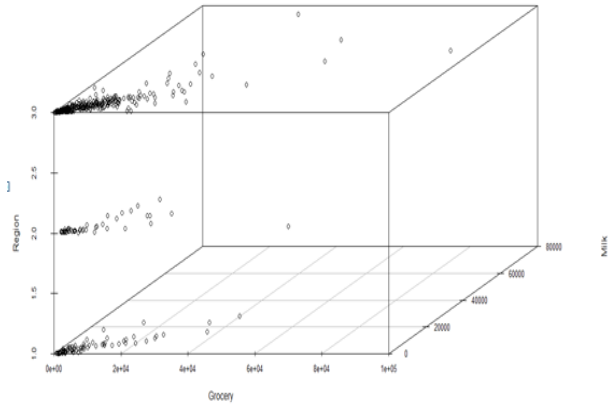

Fig. 4: 3D Scatter Plot.

The Scattered plot between Grocery vs Milk across 3 regions was shown in Fig.3. A 3D Scatter plot is shown below by Fig.4 in this three attributes considered were Grocery, Milk along three Regions. It is indicated that more data points were plotted in Region 3 followed by Region 1 and Region2. Region 3 has more sales of Groceries and Milk followed by Region1 and Region2 respectively.

The following Fig.5 represents a heat map i.e. a 2D display of a data matrix, calculate the similarity between Regions was calculated using distance function and then plotted it with a heat map. A heat map is a graphical picture of data where the distinct values contained in a matrix are denoted as colors. Heat maps assist us to get a prompt feel for an area by grouping places into classes and showing their density visually. The density is high where the color is dark.

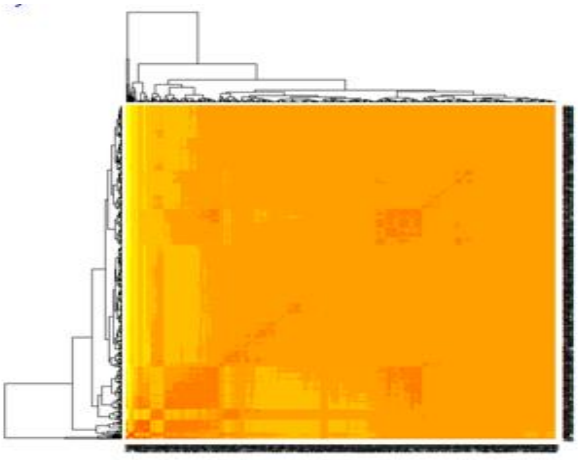

Fig. 5: Heat Map.

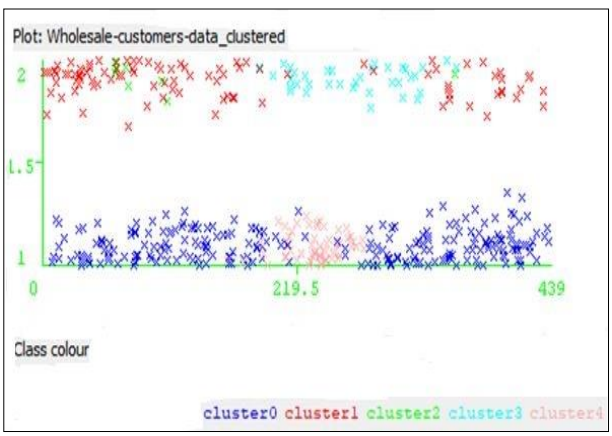

Fig. 6: Clusters Were Shown Channel Wise the Linear Regression Model Equation for Region vs. Grocery and Milk Is Given by Region = (-2.745e-06) * Grocery + (5.962e-06) * Milk + 2.530e+00 -- (I).

The Clustered diagram was shown under in Fig.6 with Instance Number on X-axis and Channel number on Y-axis reveals that Cluster0 and Clster4 used Channel-1 where as Cluster1, Cluster2, and Cluster3 used Channel2.

The Data set was subjected to K-Means Clustering algorithm with k = 5 numbers of clusters is shown below in Table-1. Euclidean distance was applied for making clusters. The 5 clusters are namely Cluster 0, Cluster1, Cluster2, Cluster3 and Cluster4. There are two channels of distribution and are Horeca and Retail. The Three regions are Lisbon, Oporto and Other Region. The two clusters Cluter0 and cluster4 used channel-1 for their distribution where as the other clusters Cluster1, Cluster2 and Cluster3 used Channel 2 for distribution. Maximum number of customers used Channel-1 for distribution. Most of the customers belong to other region than Lisbon, Oporto regions. Customers belong to Cluster2, Cluster0 and Cluster4 have spent more money for Fresh Products on annual basis. Lowest annual amount spent for fresh products was customers belong to Cluster3. Highest Annual Amount spent on milk was by customers of Cluster2 followed by Cluster3. The lowest was spent was by customers of Cluster0. The highest annual spending on Groceries was made by customers of Cluster2 followed by cluster3. The lowest annual spending was made by customers of Cluster0. The highest annual amount spent on frozen products was made by customers of Cluster0 followed by Cluster 4. The lowest annual amount was spent by customers of Cluster1. The highest annual amount spent on Detergents _Paper products was made by customers of Cluster2 followed by Cluster 3. The lowest annual amount was spent by customers of Cluster0. The highest annual amount spent on Delicatessen products was made by customers of Cluster2 followed by Cluster 1. The lowest annual amount was spent by customers of Cluster4.

Table 1: K-Means Clustered Data Set with 5 Clusters

Attribute	Full Data	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Instances	440	239	98	8	36	59
Channel	1.3227	1	2	2	2	1
Region	2.5432	2.8828	3	2.875	1.5	1
Fresh	12000.2977	13617.0879	9178.367	17674.5	6209.4722	12902.2542
Milk	5796.2659	3348.4184	8799.5918	36953.25	10104.3611	3870.2034
Grocery	7951.2773	3946.3389	13575.0918	51511.125	15983.25	4026.1356
Frozen	3071.9318	3901.5356	1428.0918	2383	2101.5	3127.322
Detergents _Paper	2881.4932	751.0711	5630.3469	26343.5	7493	950.5254
Delicatessen	1524.8705	1469.9707	1750.3061	2679.625	1556.1389	1197.1525

The customers whose annual spending is more on Frozen Products belongs to other Region and using Channel 1 are spending less annual amount on Products like Fresh, Milk, Grocery, Detergents Paper and Delicatessen. The customers belongs to Other Region and using channel 2 spends highest annual amount on the products namely Fresh, Milk, Grocery, Detergents Paper and Delicatessen are spending less annual amount on Frozen products. The highest number of Cluster0 customer's annual spending amount was more on Frozen Products than on other products like Milk, Grocery, Detergents Paper and Delicatessen belongs to other Region and used Channel1 for distribution. Cluster 2 has lowest number of Customers. Cluster2 customers annual spending amount was more on Products like Fresh, Milk, Grocery, Detergents Paper and Delicatessen belongs to Other Region used Channel2 and spend less on frozen products.

The grouped data set was used in Classification techniques through hybridization called Logistic Model Tree (LMT) [16, 17] was applied on subsequent clustered data. This LMT Classifier algorithm for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves. The procedure can deal with binary and multi-class target variables, numeric and nominal attributes and missing values.

Table 2: Performance Parameter Table

S. No	Name of the Parameter	value
1	Correctly Classified Cases	100%
2	Incorrectly Classified Cases	0%
3	Kappa statistic	1
4	Mean absolute error	0.0048%
5	Root mean squared error	0.0181%
6	Relative absolute error	1.8894%
7	Root relative squared error	5.0942%
8	Time taken to build model	0.56 sec
9	Accuracy	1
10	Precision	1

Table 3: Confusion Matrix

	Cluster #				
	0	1	2	3	4
cluster0	239	0	0	0	0
cluster1	0	98	0	0	0
cluster2	0	0	8	0	0
cluster3	0	0	0	36	0
cluster4	0	0	0	0	59

The TABLE-2 shows the performance parameters like Kappa statistic, accuracy and Precision of the algorithm indicates that the hybridized Logistic Model Tree proved to be good in terms of results as the values of Kappa statistic is1, Precision is 1and Accuracy is 1. Kappa is a chance-corrected measure of agreement between the classifications and the true classes. The Confusion Matrix represents that accuracy of classification in terms of performance of the hybridized algorithm namely Logistic Model Tree (LMT) is very good and is given in Table-3.

5. Conclusions

It was found that by accuracy parameters namely Confusion Matrix, Accuracy, Precision and Kappa Statistic showed that the hybridized technique called Clustered based Logistic Model Tree Classifier has yielded very good results when compared to individual application of algorithms. The kappa statistic measure the agreement of prediction with the true class and is 1 signifies there is complete agreement. The confusion matrix showed that this hybridized Clustered based Logistic Model Tree was more accurate in performance wise. It was found that The customers whose annual spending is more on Frozen Products belongs to other Region and using Channel 1 are spending less annual amount on Products like Fresh, Milk, Grocery, Detergents Paper and Delicatessen. The customers who spends highest annual amount on the products namely Fresh, Milk, Grocery, Detergents Paper and Delicatessen are spending less annual amount on frozen products. The highest number of Cluster0 customer's annual spending amount was more on Frozen Products than on other products like Milk, Grocery, Detergents Paper and Delicatessen belongs to other Region and used Channel1 for distribution. Cluster2 customer's annual spending amount was more on Products like Fresh, Milk, Grocery, Detergents Paper and Delicatessen and spends less on frozen products.

References

[1] Rudradeb Mitra, "*Predicting buying behavior using Machine Learning: A case study on Sales Prospecting (Part I)",* https://becominghuman.ai/predicting-buying-behavior-using-machine-learning-a-case-study-on-sales-prospecting-part-i-3bf455486e5d.

[2] Chris Glass, "*7 Ways Machine Learning Boosts Sales Performance and Drives Revenue Growth",* https://optymyze.com/blog/7-ways-machine-learning-boosts-sales-performance-and-drives-revenue-growth/.

[3] Rajan Gupta, Chaitanya Pathak, "*A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing"*, Elsevier Procedia Computer Science 36(2014) 599 – 605, www.sciencedirect.com.

[4] Yuta Kaneko, Katsutoshi Yada, "*A Deep Learning Approach for the Prediction of Retail Store Sales",* IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, Electronic ISBN: 978-1-5090-5910-2, Electronic ISSN: 2375-9259, DOI: 10.1109/ICDMW. 2016.0082.

[5] Yi Zuo, Katsutoshi Yada, A.B.M. Shawkat Ali, "*Prediction of Consumer Purchasing in a Grocery Store Using Machine Learning Techniques",* 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2016, Electronic ISBN: 978-1-5090-5753-5, https://doi.org/10.1109/APWC-on-CSE.2016.015.

[6] Stefan Meinzer, Ulf Jensen, Alexander Thamm, Joachim Hornegger, Björn M. Eskofier, *"Can machine learning techniques predict customer dissatisfaction? A feasibility study for the automotive industry",* 2017, Vol. 6, No. 1, Artificial Intelligence Research, URL: https://doi.org/10.5430/air.v6n1p80.

[7] S.HanumanthSastry and Prof.M.S.PrasadaBabu, *"Analysis & Prediction Of Sales Data In Saperp System Using Clustering Algorithms",* International Journal of Computational Science and Information Technology (IJCSITY) Vol.1, No.4, November 2013. https://doi.org/10.5121/ijcsity.

[8] Vaughn Aust, Machine Learning Can Turn Your Sales Team into Closers By drawing correlations out of massive amounts of customer data, machine learning can better inform the sales process—and enable your reps to close more deals, https://destinationCRM.com.

[9] Cristina STOICESCU, "*Big Data, the perfect instrument to study today's consumer behavior",* Big Data, the perfect instrument to study today's consumer behavior. Database Systems Journal vol. VI, no. 3/2015.

[10] Daniel Faggella, *"Predictive Analytics for Marketing – What's Possible and How it Works", https://www.techemergence.com/predictive-analytics-for-marketing-whats-possible-and-how-it-works/*.

[11] Bernard Marr, How Machine Learning Will Transform the Sales Function, https://www.forbes.com/sites/bernardmarr/2017/07/06/how-machine-learning-will-transform-the-sales-function/#3672c05f23d7.

[12] Reid Pryzant, Young-joo Chung, Dan Jurafsky, Predicting Sales from the Language of Product Descriptions, Proceedings of SIGIR, Tokyo, Japan, August 2017 (SIGIR 2017 eCom), 10 pages.

[13] Kris Johnson Ferreira, Bin Hong Alex Lee, David Simchi-Lev, Analytics for an Online Retailer: Demand Forecasting and Price Optimization, https://www.hbs.edu/faculty/Publication%20Files/kris%20Analytics%20for%20an%20Online%20Retailer_6ef5f3e6-48e7-4923-a2d4-607d3a3d943c.pdf.

[14] Robert Siwerz, Christopher Dahlén, Predicting sales in a food store department using machine learning, Degree Project In Computer Engineering, First Cycle, 15 Credits Stockholm, Sweden 2017.

[15] Abreu, N. (2011). Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. Mestrado em Marketing, ISCTE-IUL, Lisbon.

[16] Niels Landwehr, Mark Hall, Eibe Frank (2005). Logistic Model Trees. Machine Learning. 95(1-2):161-205. https://doi.org/10.1007/s10994-005-0466-3.

[17] Marc Sumner, Eibe Frank, Mark Hall: Speeding up Logistic Model Tree Induction. In: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, 675-683, 2005. https://doi.org/10.1007/11564126_72.