ACTA INFOLOGICA, 2018; 2(1): 45-51
Research Article

ACTA INFOLOGICA
ISSN: 2602-3563

A Real Life Web Based Marketing Optimization Framework With External Data

Şadi Evren Şeker*

ABSTRACT

Big data and data science studies in recent years are booming exponentially, parallel to the data collected and increased processing speeds. As an inevitable consequence, most of the web-based companies are migrating their business models to novel technologies based on the big data and data science research. This paper is based on a real life experience based on one of the web stores with highest volume sales in Turkey. The project was building a data science model on big data technologies to make estimations based on the external data, such as weather conditions, customer demography, news at newspapers, current product alternatives, financial facts (like currency exchange rate or stock market values) and most importantly the sentimental analysis and opinion mining on social network, blogs and news. In the paper, details of problems and possible solution alternatives and methodology for problem solving and solutions and outcomes of the study are explained in the given order.

Keywords: Marketing, opinion mining, targeted marketing, big data, data science, customer behavior.

Information of Author(s):

Şadi Evren ŞEKER
ORCID: 0000-0002-7323-3695
sadievrenseker@sehir.edu.tr
İstanbul Şehir Üniversitesi, İşletme ve Yönetim Bilimleri Fakültesi, Yönetim Bilişim Sistemleri Bölümü

DOI: 10.30801/acin.356344

Submit Date:	20.11.2017
Accept Date:	03.01.2018
Publish Date:	26.06.2018

(*) Contact Author
Address: İstanbul Şehir Üniversitesi, İşletme ve Yönetim Bilimleri Fakültesi, Yönetim Bilişim Sistemleri Bölümü, İstanbul, Türkiye ● **Telephone Number:** +90 0216 4444 034

ACTA INFOLOGICA
ISSN: 2602-3563

1. INTRODUCTION AND PROBLEM DEFINITION

E-marketing is an increasing trend and the advances in e-marketing is directly affecting the web-based sales [1]. An Internet company in Turkey, has invested for a research project to collect data and develop a data science project to increase marketing success. By definition of the project, marketing success criteria is the percentage of sales for ads displayed. Also the data is grouped into two categories, internal and external as below:

- internal data is defined as the data in the local databases of the company, like customer information, product information, sales information
- external data is defined as any data source, which might affect the customer behavior about the advertisement, such as the weather conditions, financial data (like currency rates or stock market values), opinions and sentimental effects of daily news or blogs or social networks.

The project has three major questions:

1. What are the correlated parameters and which parameters have the highest affect on marketing success?
2. What is the best technology for implementing the project?
3. What is the best data science solution for the decision-making and customer / advertisement matching optimization?

During the project all three questions above are solved and a working real life project is implemented [2]. For the first question, some temporary research oriented systems are installed and some prototype coding is developed on the temporary systems. Although the system was not built on a full scaled technology in this step, data is collected by using some sampling algorithms and some discovery correlation algorithms such as Kendall's-Tau, Spearman's-Rho or Goldman-Kruskal's-Gamma [3] is executed to understand the correlation between the parameters and the marketing success.

For the second major question, the prototype implementation is executed on some technologies like Apache Hadoop, Microsoft Azure, Apache Spark and for the data science layer, MLLib, AzureML and Mahout libraries are implemented for test purposes.

Finally for the third question, several data science solutions are tested. Because the data was collected from man different unstructured sources, during the preprocessing, some imputation, data cleaning and noise reduction algorithms are implemented. Although some of the dimension reduction algorithms tested, the success of these algorithms was not high enough to take into consideration for the case. For example, Principal Component Analysis (PCA) and Latent Dirichlet Allocation (LDA) algorithms are tested an the contribution of the algorithms was less than 1% for most cases and had negative affect in some cases. After the data preprocessing and extracting the feature vectors, a limited set of machine learning algorithms are tested. The reason for limitation is the number of suitable algorithms in big data world, which means they can run on map-reduce based, is limited within the libraries.

As a summary, the problem is defined on the selection of best technology, best algorithm on most important parameters based on both external and internal parameters. The project outputs the best alternative advertisement for any customer.

2. METHODS

The project can be divided into three sub projects as defined in the introduction and problem definition section. For all subparts of the project different methods are tested and details of these methods are provided within this section. So the section is organized in three subsections, which are the preprocessing and parameter selection with correlation factors, technology decision and data science problems and solutions.

ACTA INFOLOGICA
ISSN: 2602-3563

2.1. Preprocessing and Parameter Selection with Correlation Factors

Before dealing with the data science layer problems, some preprocessing algorithms applied and feature vectors are extracted.

During the preprocessing phase, the noisy data is cleaned and some of the missing data is imputed. The data cleaning is only applied on the external data, since the data source is unstructured and data collection is implemented by some background processing running on servers and the data source is located under some other organizations. For the internal data, where it is structured and gathered via database queries, there is no such problem as noisy data but there are some missing values for some of the data sets. For example, database might not hold the whole information about customer demographics like age or birthplace of the person. This information is collected from some of the customers by their own will. So, there is no imputation for the external data and no data cleaning for the internal data because of their structures.

For the imputation phase, k-nearest-neighborhood (k-nn) algorithm is deployed with k=3, and the missing values are predicted by the rest of the data set with using 3 most common samples. For the data cleaning, the row-wise deletion technique is implemented for higher confidence.

Just after the preprocessing phase, the correlation factors are calculated over the feature vectors extracted.

Pearson's Rho function is simply division of covariance of two parameters to the multiplication of standard deviation of each parameter as demonstrated in equation (1).

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \quad (1)$$

Pearson's Rho function yields a value between -1 and +1 and 0 means there is no correlation while +1 means they are highly correlated and -1 means there is a high correlation in the negative direction.

Kendall's Tau function is based on concordant and non-concordant pairs and concordant pairs are considered as the agreement between two parameter features. In below equation the count of non-concordant is subtracted from the number of concordant samples and the value is divided to a normalization factor, which is the total number of pairs within the given data set. The formulation of Kendall's Tau is demonstrated in equation (2).

$$\tau = \frac{\sum_{i=0}^{n} concurdant_i - \sum_{i=0}^{n} nonconcurdant_i}{n(n-1)/2} \quad (2)$$

Similar to Pearson's Rho, Kendall's Tau also yields a value between -1 and +1 and again the 0 output means no correlation, +1 means high correlation in same direction and -1 means high correlation in negative direction.

Finally, Goldman Kruskal's Gamma is based on the probability of randomly selected pair of observation in the same order (Ps) or opposite order (Pd) and can be easily calculated by the division of difference to summation between these two parameters. This formula is given in equation (3).

$$\gamma = \frac{P_s - P_d}{P_s + P_d} \quad (3)$$

Similar to the Pearson's Rho and Kendall's Tau correlations, the Goldman Kruskal's Gamma is also between -1 and +1 and again the 0 output means no correlation, +1 means high correlation in same direction and -1 means high correlation in negative direction.

After calculating all three correlation coefficient [4] [5], we go for the arithmetic average of methods and try to rank the correlation between each parameter and the output parameter.

ACTA INFOLOGICA
ISSN: 2602-3563

2.2. Deciding Technology

Decision on the technology for the project depends on many different criteria, such as the current knowledge and experience of employees on a certain technology, current software license agreements, current technologies already installed on the servers and so on. The fortunate part of the project was, the company was investing its first big data / data science project and they were open for any technology decision. The technology decision is divided into three layers during the technology management of the project:

- Decision on data science technologies
- Decision on big data technologies
- Decision on the server technologies depending on the above decisions

So the first step of technology decision was deciding on the final step of the project. During the project, the data science layer was determinant for the rest of the technology decisions. The reason was, for any alternative technology in the data science layer with less success would create a negative motivation for the project, so data science layer is accepted as the determiner of the technology for the project.

The data science level algorithms are decided at the first step and details of the algorithms are explained in subsection 2.3 of this paper. After the decision of the algorithms, performance tests are executed and the performance of technologies are compared on the private cloud servers. Company was already holding a private cloud with Linux and Windows operating systems and core database of the company was built on MS SQL. For the apache technologies, Hadoop and spark, the tests ran on the Linux servers and for azure, the tests ran on the Microsoft cloud. After the benchmark test comparison, company decided to go on the spark technology and searched for a suitable server alternative. The best alternative with the cost/performance and easy maintenance was amazon web services (AWS), that the company would reach. Although AWS offers some plans and technology alternatives, another decision is done for the simple storage service (S3) and running spark in the AWS cloud [6].

2.3. Data Science Problems and Solutions

The first data science experiments are executed on sampled data with Rapid Miner software and the results achieved during the experiments are carried on to the production phase. During the experiments, five basic problems are researched:

- Feature extraction technique for the opinion mining
- Machine Learning Algorithm for the opinion mining
- Sales Forecasting
- Ad Matching
- Customer Segmentation

During the feature extraction for opinion mining phase, three major techniques are tested, term frequency – inverse document frequency (TF-IDF), Word-to-vector and Bag-of-Words [7]. During the machine learning for opinion mining, the classification algorithms, such as, k-nn, decision tree, random forest and naïve bayes algorithms are tested. For the forecasting, linear regression with, least squares, Lasso and ridge alternatives are tested, also isotonic regression and random forest algorithms are tested at this step too. Finally for the ad matching problem, recommender algorithms are tested based on content based filtering and collaborative filtering.

As a conclusion of the section 2, the technology decision can be demonstrated as in Figure 1.

ACTA INFOLOGICA
ISSN: 2602-3563

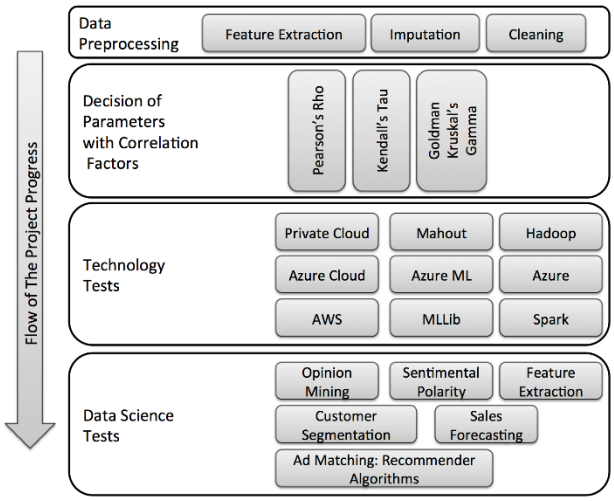

Figure 1. Overview of the Project Progress

The best results achieved on TF-IDF during the feature extraction, random forest for the opinion mining, logistic regression for the customer segmentation and k-nn algorithm based on the user/product similarity. Finally for the customer segmentation problem, a decision tree algorithm is implemented over the x-means algorithm. X-means algorithm ran with unknown k parameter for k-means, bounded between 2 to 120 and best solution achieved for the 38 segments of customers. Again, for the decision tree on the customer segmentation, best results achieved by the random forest algorithm. Details of only best practice algorithms are provided below.

TF-IDF is one of the text mining methods used for feature extraction from natural language data sources [8], [9] [10] [11].

The TF-IDF calculation is provided in equation (1).

$$TF - IDF(t, d, D) = tf(t, d) x\ idf(t, D) \qquad (4)$$

Where t is the selected term, d is the selected document and D is all documents in the corpus. Also TF-IDF calculation in above formula is built over term frequency (TF) and inverse document frequency (IDF), which can be rewritten as in equation (5).

$$tf(t, d) = \frac{f(t, d)}{\max\{ f(w, d) : w \in d \}} \qquad (5)$$

where f is the frequency function and w is the word with maximum occurrence. Also the formulation of IDF is given in equation (6).

$$d(t, D) = log \frac{|D|}{|\{d \in D : t \in d\}|} \qquad (6)$$

where |D| indicates the cardinality of D, which is the total number of documents in the corpus.

Another crucial algorithm mostly applied on similar data sets is the random forest (RF) algorithm [12]. Random forest is based on the decision tree approach and the method applies bagging [13] [14], random feature selection [15], and shape quantization [16] methods together.

Random forest algorithm simply creates random decision trees and combines the decision trees depending on their success rates. The similar approach can be easily applied on the other algorithms as a meta-classifier. In this approach the decision tree divides the problem space and the majority voting is applied for each portion of data. Although some techniques like boosting may show better success rates, bagging has an advantage of avoiding the over-fitting problem. By a general definition, bagging reduces the variance and avoids the over-fitting [17].

ACTA INFOLOGICA
ISSN: 2602-3563

In this study, we have applied bagging both directly by using random forest and applying techniques as an ensemble classifier over other algorithms.

The second attempt was implementing the random forest algorithm and the most important parameter for the algorithm is the number of trees. This parameter indicates the number of randomly generated decision trees and the random forest algorithm will work as a meta-decision tree over these random decision trees. We have defined the number of random trees as 10, which means random forest algorithm will generate 10 random decision trees and another decision tree over these 10 decision trees as a meta-decision tree.

3. FINDINGS

As an output of the research for different alternatives in section 2, for the problem statement in section 1, the best solution and the overview of the project can be demonstrated as in Figure 2.

The project has two major data source as input. The first group of data source is the internal data source with customer and product statistics. The second group data source is the external data source, which includes the weathers, social networks [18], news and blogs. In order to collect external data, some resident processes are deployed and a temporary database is implemented. All the information collected from both internal and external data is gathered in S3 server in an AWS service for long time storage and online processing as a big data solution. Spark server is running on top of simple storage service (S3) and machine-learning algorithms are running on top of spark server with full map-reduce advantages. So the spark server can load balance and use scaling and cost reduction advantages.

Finally, project provides the required reports and the optimized matching algorithm between the customer and the products and/or advertisements.

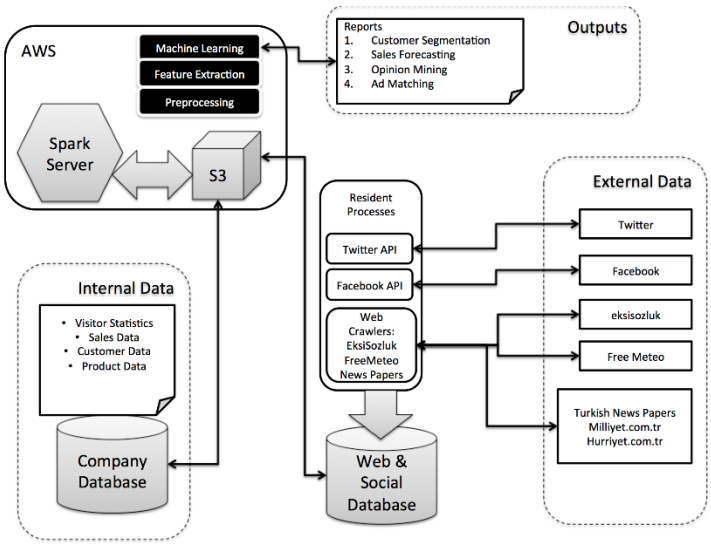

Figure 2. Overview of the Project and Problems Deployment

4. DISCUSSION AND CONCLUSION

The project was about deciding the data science solution alternatives and technology investment for a big data project of a web based company in Turkey. During the project, the determinant factor was the data science policy and technology for the rest of the project investment. After achieving a certain success improvement in the data science prototyping, company researched for the best big data investment. Also from the project it was certain that, the external data sources are affecting the recommender system together with the internal data sources. Project has some unique properties like the first time application of recommender system based on multiple external data sources, working on big data platform and specialized on web marketing in Turkey. Of course, the project will be

ACTA INFOLOGICA
ISSN: 2602-3563

improved by time considering different feature extraction methods like time series analysis or implementing new machine learning algorithms on the big data world but it can be considered as one of the first steps in the area.

REFERENCES

[1] Sadi Evren Seker, "Real Life Machine Learning Case on Mobile Advertisement: A Set of Real-Life Machine Learning Problems and Solutions for Mobile Advertisement," in *Computational Science and Computational Intelligence (CSCI), 2016 International Conference on*, 2016.

[2] Mehmet Lutfi Arslan, Sadi Evren Seker, and Cevdet Kizil, "Innovation driven emerging technology from two contrary perspectives: A case study of Internet," *Emerging Markets Journal*, vol. 3, no. 3, p. 87, 2014.

[3] Sadi Evren Seker, *Weka ile Veri Madenciliği*.: Draft2Digital, 2015.

[4] Sadi Evren Seker, Cihan Mert, Nuri Ozalp, and Ugur Ayan, "Time series analysis on stock market for text mining correlation of economy news," *Int. J. Soc. Sci. Humanity Stud*, vol. 6, no. 1, pp. 66-91, 2013.

[5] Sadi Evren Seker, Yavuz Unal, Erdinc H Kocer, and Zeki Erdem, "Ensembled correlation between liver analysis outputs," *International Journal of Biology and Biomedical Engineering*, vol. 8, pp. 1-5, 2014.

[6] Abhishek Gupta and Dejan Milojicic, "Evaluation of HPC Applications on Cloud," in *Open Cirrus Summit (OCS) 2011*, vol. 6, 2011, pp. 22-26.

[7] Sadi Evren Seker and Cihan Mert, "A Novel Feature Hashing for Text Mining," *Journal of Technical Science and Technologies*, vol. 2, no. 1, pp. 37-40, 2013.

[8] Marc-André Mittermayer, "Forecasting intraday stock price trends with text mining techniques," in *HICSS '04 Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)*, vol. 3, 2004, pp. 64-73.

[9] Robert P. Schumaker and Hsinchun Chen, "Textual analysis of stock market prediction using breaking financial news:The AZFin Text system," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 2, pp. 1-19, 2009.

[10] Saman Halgamuge, Y Zhai, and Arthur Hsu, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction," in *Advances in Neural Networks - ISNN 2007 (Lecture Notes in Computer Science)*, vol. 4493, 2007, pp. 1087-1096.

[11] Gabriel P. C Fung, Jeffrey X Yu, and Wai Lam, "News sensitive stock trend prediction," *Lecture Notes in Computer Science*, vol. 233, pp. 481– 493, 2002.

[12] Breiman L, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[13] Breiman L, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49-84, 1996.

[14] Breiman L, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123-140, 1996.

[15] Ho TK, "Random Decision Forests," *Proceedings of the 3rd International Conference on Document Analysis and Recognitio*, pp. 278-282, 1995.

[16] Amit Y and Geman D, "Shape quantization and recognition with randomized trees," *Neural Computing*, vol. 9, no. 7, pp. 1545-1588 , 1997.

[17] Watanachaturaporn P and Varshney PK, Arora MK Xu M, "Decision tree regression for soft classification of remote sensing data:," *Remote Sensing of Environment* , vol. 9, no. 3, pp. 322-336 , 2005.

[18] Sadi Evren Seker and Atik Kulakli, "Macroeconomic ICT Facts and Mobile Telecom Operators via Social Networks and Web Pages," *Journal of Business Economics and Management*, vol. 4, no. 2, pp. 99 - 104, 2016.

This work was presented at the 4th International Management Information Systems Conference and published in the conference abstract book.