IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Churn Analysis with Machine Learning Algorithms

Buket Onal[1], Metin Zontul[2]

Master Student, Department of Computer Engineering, Istanbul Aydin University, Istanbul, Turkey [1]

Associate Professor, Department of Computer Engineering, Istanbul Aydin University, Istanbul, Turkey [2]

Abstract: Competition conditions are increasing rapidly in almost every sector today. Along with the developments in the e-commerce sector, it has been seen that most of the developed countries are integrated and the development of logistics sector increases rapidly. Considering this increase in almost every sector, customer loyalty is of great importance for companies. By taking advantage of the data mining technology and taking into consideration the behavior exhibited customers, the data obtained can be modeled to determine the customers who have a tendency to leave the company. In this study, it was tried to reveal the lost customer behaviors by examining the shipping information of the customers working with a logistics company operating in Turkey. Data about 2.000 customers from the data received from the company were used in our application. Based on the customer shipment information, the input data were created by dividing into 5 classes. In the output data, the acquired and lost customers were taken into consideration. The information obtained by the data mining has been tested on the support vector machine algorithm. The data of these customers pertaining to past two years were divided into 3-month periods. Customer loss analysis was conducted for a total of 8 quarters including 7 sets of input data and 1 set of output data, and it is tried to make loss analysis estimation for the customers who have a tendency to leave the company in the next three months.

Keywords: Churn Analysis, Data Mining, Classification, Support Vector Machine.

I. INTRODUCTION

Today, many companies that serve in various sectors have to struggle to survive in the face of globalization and increasing competition. In order for companies to maintain their existence in the sector, it has become important to retain existing customers, minimize customer losses, and to identify customers who tend to prefer other companies.

Customer loss analysis has an important place in many sectors. The issue has been the subject of many scientific studies. Previous studies related to this subject have been examined. Tosun (2006) analyzed the customers who stopped using a private bank credit card. Using decision tree threshold values were determined in the direction of different rules and related examinations were carried out. The rules that negatively affect the results in the course of this study were subsequently discarded through post pruning method [1]. Akbulut (2006), has conducted his analysis based on the data of a cosmetics brand. The J.48 decision tree algorithm was chosen using the WEKA program. Using this algorithm, the customer behavior model was developed. In the direction of the results obtained, the customers are divided into two groups in the direction of amount of their purchases and the proposals have been made taking into consideration the reasons of preference of other companies by these customers [2]. Ozmen (2006) worked on calculating the likelihood of customers leaving a mobile telephone line operator based on termination of the contract of line holders, if there is no reattachment within 6 months after adding the line balance for a contingent line in the telecom sector. Decision trees and regression models were used. Complaints made to the call centers were observed as the most important factor [3]. Arifoğlu (2011) examined the data of GSM operator and compared Navie Bayes, Support Vector Machine, Probabilistic Neural Network and C-means algorithms. It has been shown that the C-Means algorithm gives the best result for the dataset used in conjunction with Adaptive Network-Based Fuzzy Inference System (ANFIS) [4]. Ercan (2015) selected data using correlation-based filtering and created an early prediction model by comparing the results of Bayesian Network, Logistic Regression, and SMO and Simple CART algorithms. Ensemble methods have been applied to increase the accuracy of the obtained model. The results show that the Simple CART algorithm is more successful in determining players who tend to leave the game. The developed model determined the players to leave the game with 68.20% accuracy [5]. Sarı (2016) estimated the sales demand of engine bearings using artificial neural networks. The results were compared with the results of the regression analysis (RA) and the time series. As a result, more realistic predictions were obtained with artificial neural networks [6].

Big volume data are stored in electronic environments, and accordingly need to reach the stored data in a crowded population is the main problem together with the technology that is improving rapidly compared to the past. In addition

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

to the purchasing behavior of customers, companies also keep various features belonging to the customers in the information storage media. Data mining has enabled the discovery of information that one can never imagine finding [7]. Thanks to the recorded data, useful and meaningful information can be obtained by detecting and revealing meaningful and confidential information in data stacks by using data mining techniques. According to the Gartner Group, data mining is the process by means of which a number of statistical and mathematical techniques are used to extract meaningful new data sets on stacked data, and where patterns and desired trends in these data sets are discovered. The acquired data mining results provide insight into how many customer-focused applications can be improved and measures can be taken to avoid of customer loss.

Companies take into account the data of past periods about customer loss analysis evaluations. One of the biggest difficulties in performing customer loss analysis when considering companies that have started to work with new companies is absence of sufficient information source, namely data set. The fact that the data set is limited causes incorrect and inaccurate customer trends. In the short term, it is necessary to determine the tendency to prefer another y. In the same way, to determine the tendency to lose long-term high-profit customers is a separate problem that stems from waiting for a long time, not expecting to lose a long-term employee, and being unable to get necessary measures to prevent loss of customer.

Considering the problems experienced in customer loss analysis, the aim of this study is to perform short-term customer loss analysis in the logistics sector. In doing so, first of all, data set in our hands has been cleared from meaningless and insignificant data. Later, using the data mining classification method, customers were classified based on their previous losing trends, taking into account the three-month shipments and total freight charges. 2000 customers form our data sets that demonstrate customer behaviors that have been acquired and lost in the classification process.
For the data that can be linearly separated in the classification by support vector machines, it is aimed to create two classes which are separated from each other by using the decision function obtained by training data. According to the functions in the dataset selected as DVM input, the output provided from this function is classified as belonging to two classes [9].

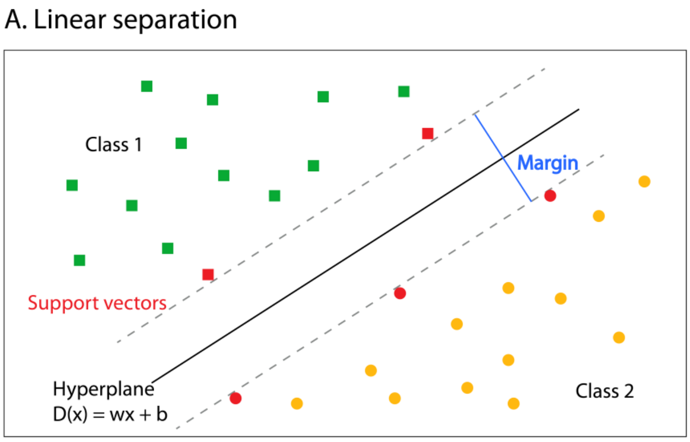

Fig 1. DVM model for linear separable data [10]

The DVM model for linearly separable data is shown in Figure 1.
In data mining, data may not always be linearly separated. In this case the problem arises about classification of data which cannot be separated linearly. In order to solve the problem arising from the fact that some of the training data belonging to dataset which cannot be separated linearly remain on the other side of the optimum hyper surface, a positive artificial variable (i ξ) needs to be defined. As in the case of the DVM method which can be linearly separated balance between maximizing the boundary that is the distance between vectors and minimizing false classification errors, , takes place by defining an equilibrium parameter ($0 < C < \infty$) denoted by C and taking positive values [11].
The DVM model for data which cannot be separated linearly is shown in Figure 2.

Data of which separation cannot be made linearly in input space, is transformed into a multi-dimensional space characterized as a space of features in order to provide a solution for the optimization problem, and thus a higher dimensional space representation is created. In the analyses made in this way, the optimum hyper surface where the classes can be separated from each other can be determined and the discrimination of the data can be done linearly.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

B. Non-linear separation

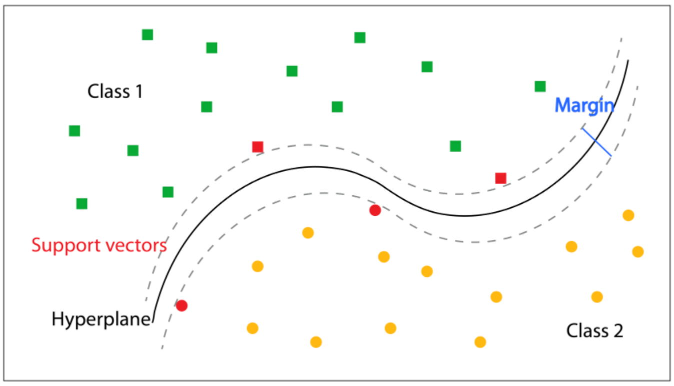

Fig 2. DVM model for data which cannot be separated linearly [10].

The DVM (Support Vector Machine) method is different from other methods because of the low complexity of algorithms and the low number of processes in learning process [12]. For this reason, the support vector machine algorithm is preferred for our nonlinear data source. With some of our data, algorithm training is provided first. The application was developed on the Spyder (Python 3.6) platform using the python language. The learning algorithm was tried on the test data to try to calculate the tendency of the customers to leave the company in the next three months.

II. MODEL

The raw data from the logistics company contains the shipping quantities and freight charges of the customers.

=IF(**Previous_period_volume**=0;"CONVERSİON";IF(**Current_period_volume**=0;"LOST";
IF(**Volume_difference**=0;"NOCHANGE";IF(**Volume_difference**<0;"DECLINER";"GAINER")))) (1)

Our data are divided into 5 different groups. The mathematical formula shown in Formula 1 is used to separate groups. The parameters used in this mathematical formula are described below:

Number of Packages in Previous Period: It indicates the total number of packages for the previous month period in which the analysis of the current customer is conducted.

Previous Period Freight Total: It indicates the total freight charge for the previous month period in which the analysis of the current customer is conducted.

Number of Working Days in the Previous Period: It indicates the total number of working days for the previous month in which the analysis of the current customer is conducted.

Number of Packages Period: It indicates the total number of packages for the current period of the month in which the analysis of the current customer is conducted.

Period Freight Total: It indicates the total freight charge for the current period of the month in which the analysis of the current customer is conducted.

Period Number of Working Days: It indicates the total number of working days for the current period in which the analysis of the current customer is conducted.

Previous Period Volume: It indicates the average package volume of the customer for the corresponding months in the previous period. It is obtained by dividing previous period number of packages parameter by the number of working days in the previous period.

Period Volume: It indicates the average package volume of the customer for the current month in the current period. The term is obtained by dividing the number of package parameter by the number of working days.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Previous Period Revenue: It indicates the average net revenue from customer for related months in the previous period. It is obtained by dividing the previous period freight total parameter by the number of working days in the previous period.

Period Revenue: It indicates the average net revenue from customer for related months in the current period. It is obtained by dividing the previous period freight total parameter by the number of working days in the previous period.

Volume Difference: It indicates the volume difference between the previous period and the current period.

The group definition obtained as a result of application of the mathematical formula used in grouping of customers.The groups obtained by applying the formulas using the values of the relevant parameters and the calculation details are given below:

No Change: If the volume difference parameter takes the value of 0, the group is determined as "no change." That is; if the package volume of the related period in the current period and the previous period has not been changed, the existing customer is not included in the lost customer list.

Conversion: If the customer's package volume for the previous period is 0 and an increase in the package volume for the current period is observed, then this is the acquired customer and included in the conversion group.

Gainer: If the current customer's volume difference value for the relevant period is greater than 0, it means this customer has increased the shipment volume and is included in the gainer group.

Decliner: If the current customer's volume difference for the relevant period is less than 0, this customer decreased the shipment volume and is included in the decliner group. Customers in this group are candidates to be lost.

Lost: If the customer has a package volume for previous period and package volume value is 0 in the related month for the current period, this is a lost customer and is included in the lost group.

Our data is subjected to normalization because it can be used in support vector machine algorithm. In this case, the data take numeric values.

TABLE 1CLASS NAMES AND VALUES OF THE DATA

Class Name	Value
No Change	0
Conversion	1
Gainer	2
Decliner	3
Lost	4

Table 1 indicated the new values we have received by normalizing our data. Data have been prepared in a total of 8 quarters as 4 quarters in the year 2015 and 4 quarters in the year 2016. The input value of the first 7 quarters in the generated data set represents the output value of the last quarter of 2016.

TABLE 2 SAMPLE DATA SET

Input 1	Input 2	Input 3	Input 4	Input 5	Input 6	Input 7	Output	DVM
1	3	2	3	3	2	2	2	0
1	4	1	2	3	3	2	2	0
1	2	3	3	4	1	3	2	0
1	1	1	1	1	1	1	2	0
1	2	2	2	3	3	2	2	0
3	3	2	3	2	4	1	4	1
1	1	1	1	1	3	2	4	1
1	2	2	3	3	2	3	4	1
1	3	4	1	1	1	1	4	1
1	2	2	3	4	1	3	4	1

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

In Table 2 an example from our data set is given. As you can see on the table, our output values represent the acquired lost customers. Since the learning of algorithm can be made more robust and at the same time the output values represent only 2 classes, changing output values in the DVM algorithm as 0 and 1 ensured achieving higher results.
Various parameters are used in our application to train the preferred DVM algorithm. These parameters are C, kernel, gamma, probability and random-state. The parameter, denoted by C, represents the error rate. The kernel parameter specifies the kernel type to be used in the algorithm. The kernel type can take linear, poly, rbf, sigmoid, precomputed values. The gamma parameter specifies the seed coefficient for the rbf, poly, and sigmoid kernel types. Probability parameter specifies whether the probability estimate is used in the algorithm. Since our data are linearly non-separable data, the participation of probability formula in the account will increase our accuracy. The random-state parameter represents the pseudo-random number generator that will use the data for the probability estimate.

By testing parameter values used for DVM algorithm it is aimed to achieve the best classification results in customer loss analysis. C, kernel and gamma parameters have been subjected to separate tests in order for customer loss analysis to be performed considering the behaviors of 2000 customers in our application. In our application, test combinations of the specified parameters have been made in order to determine the customers who tend to prefer another company at a high level. In the direction of the parameters specified as success criterion, the values that make the customer classification high have been determined.

The values of the C parameter used for support vector machines have a direct effect on the classification accuracy. Classification cannot be done correctly if the values are chosen too high or too low. The values of 0.1, 0.3, 0.5, 0.7, 1, 2.0, 3.0 and 4.0 were given for the C parameter and the value of 0.74 which is the best result according to these values was acquired when C parameter had values of 0.1, 0.3 and 0.5. It was also found that the estimated value decreased to 0.73 at 0.3 and 0.5 values from time to time. For this reason, 0.1-value was preferred in our algorithm.

The kernel parameter refers to the kernel type. The kernel type can take linear, poly, rbf, sigmoid, precomputed values [13]. When the linear kernel type is tested, the lowest result is 0.55. Our algorithm has an estimated value of 0.63 for the sigmoid kernel type and 0.69 for the poly kernel type. In the case of Rbf kernel type, a maximum value of 0.74 has been obtained and it has been decided to use this kernel type in our algorithm.

The gamma parameter specifies the kernel coefficient for the rbf, poly and sigmoid kernel types. Values of 0.3, 0.5, 0.7, 1, 2.0, 3.0, 4.0 and 5.0 are given for the gamma parameter. Gamma parameters were given different values while other parameters were kept constant. According to the results obtained, the more the kernel coefficient is increased, the lower the estimate value is. The appropriate result can be achieved when the parameter has values of 0.3 and 0.5. However, the more the kernel coefficient increases, the longer the algorithm learning time becomes and thus the 0.3 value is chosen as the correct value for our algorithm.

III.APPLYING THE MODEL

Our model covers 2000 customer behaviors that have been acquired and lost. Our application has been developed in Python language by using Spyder desktop platform. In order to develop our application ready libraries that the platform has provided us have been added as a reference to our project.

```
churn=np.array(pd.read_csv('C:\\Users\\Buket\\.spyder-py3\\Projects\\churn8.csv',sep=';'))
```

Fig 3. Taking the dataset in series

As shown in Figure 3, the data that is primarily needed for our program are put in project as series.

```
clf = ExtraTreesClassifier()
clf = clf.fit(X, y)
sum=0
j=0
for i in clf.feature_importances_:
    print(str(j)+':'+str(i))
    sum+=i
    j+=1
```

Fig 4. Variable importance analysis

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Variable significance analysis is performed for our input data with the ExtraTreesClassifier () function shown in Figure 4. Variable significance analysis is necessary to extract data that are not subject to modeling within the dataset. The data needed for the decision tree algorithm used in the analysis of variable significance takes input data expressed by x and output values expressed by y as parameters.

```
0:0.115567994733
1:0.122133447689
2:0.0935844192411
3:0.10682893515
4:0.116744562308
5:0.261789846375
6:0.183350794505
```

Fig 5. Result of variable significance analysis

As a result of the applied decision tree algorithms, the significance scores of our input values in our data set are shown in Figure 5. In our application, the model was chosen so that data having value lower than 0.11 are not considered. The test data to be used for learning algorithms covers 0.33 of the total model.

```
svm=OneVsRestClassifier(SVC(C=0.1,gamma=0.3,kernel='rbf', probability=True,
                            random_state=random_state))
```

Fig 6. Application of the support vector machine algorithm

The OneVsRestClassifier () function shown in Figure 6 applies a multi-class, multi-label strategy. In this strategy there is only one classifier for each class. Learning is carried out with the parameters we have already tested, which will be used in the DVM machine algorithm. After learning the algorithm, the test data other than the learning data were included in the application and the classification result was tried to be obtained.

	precision	recall	f1-score	support
0	0.72	0.78	0.75	330
1	0.76	0.69	0.73	330
avg / total	0.74	0.74	0.74	660

Fig 7. Classification result values

In the direction of the results obtained, the probability of a datum to belong to 0 class can be estimated with a 0.75 rate correctly. Probability of belonging to the 1 class can be estimated with a 0.73 rate correctly. In this case, class of a datum can be determined with a 0.74 rate correctly.

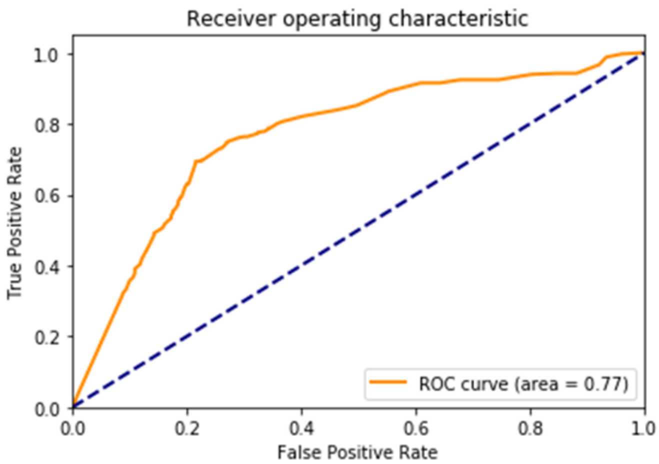

Fig 8. Display of results on ROC graph

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

As shown in Figure 8, the output quality of our data shown by the ROC graph has a curve ratio of 0.77. In the ROC graph, the Y-axis represents the true positive rate, the X-axis represents the false positive rate. The closer the curve is to the upper left corner, the better the evaluation is. In this case, the 1.0 ROC curve is considered ideal [14]. The larger the area under the curve, the better the quality is obtained. In this case, it is seen that we obtained a result close to the ideal curve ratio. This shows that we have achieved high success in our application.

IV. CONCLUSION AND DISCUSSION

Due to the low population in the past, important data could be stored through printed documents and it became harder today because of the increased population. With the fast-paced technology, it is now possible to keep and store personal data in electronic form, just like in any other fields.

As the data stored in the electronic environment provide great convenience, the increase of the volume of these data has caused some problems. With these stored data becoming an electronic mass, we have sought ways to make raw data useful. At this point, many studies have been carried out to classify the raw data with some algorithms and to make it more meaningful and quickly accessible when needed.

As in every sector, a large number of customers are available in the logistics sector and the data belonging to these customers are of great significance. Given the increasing competition, companies need to keep existing customers. Acquiring new customers requires more time and more expenses. For this reason, companies have been forced to take some precautions by trying to identify customers who tend to choose another company among existing customers by focusing on customer loss analysis.

In this study, data received from a logistics company operating in the World and Turkey were classified using data mining techniques and the support vector machine algorithm was utilized. It was ensured to identify customers who were likely s a result of the tests carried out in our work, 74% success was achieved. Determination of customers who have a tendency to leave the company within the next 3 months with high-rate estimation allowed company to take necessary measures and to keep existing customers in its portfolio within a short period of time. The data used in the study represent 2000 customers.

The current implementation can be further enhanced with more customers and better quality. In addition, conducting annual loss of customer analysis in terms of months instead of quarters will be more beneficial in future studies.

REFERENCES

[1] Tosun T., Churn Analysis on Credit Cards by Data Mining Techniques, Istanbul Technical University, Institute of Science and Technology, Master Thesis, Istanbul, 2006.
[2] Akbulut S., Customer Analysis and Customer Segmentation of a Cosmetic Brand with Data Mining Techniques, Gazi University, Institute of Science and Technology, Master Thesis, Ankara, 2006.
[3] Ozmen M., Churn Modeling In Telecommunications Sector, Istanbul Technical University, Institute of Science and Technology, Master Thesis, Istanbul , 2006.
[4] Arifoglu E., Churn Management By Using Fuzzy C-Means, Bahcesehir University, Institute of Science and Technology, Master Thesis, Istanbul, 2011.
[5] Ercan P., Detection of Churners in Internet Games Using Crm Approach: A Case Study on Pishti Plus, Middle East Technical University, Institute of Science and Technology, Master Thesis, Ankara, 2015.
[6] Sarı M., Artificial Neural Networks and Sales Demand Forecasting in an Automotive Industry, Sakarya University, Institute of Science and Technology, Master Thesis, Sakarya 2016.
[7] Bransten L., "Technology – power tools – looking for patterns: data mining enables companies to better manage the ream of statistics they collect; the goal: spot the unexpected", Wall Street Journal, 27 (12): 16- 20 (1999).
[8] Larose, D. T., 2005: Discovering Knowledge in Data: An Introduction to Data Mining, John and Wiley Sons Incorporated, USA.
[9] Namane, A., Guessoum, A., Soubarı, E. H. ve Meyrueıs, P., CSM neural network for degraded printed character optical recognition, Journal of Visual Communication and Image Representation, 25, 1171-1186, 2014.
[10] http://www.coxdocs.org/doku.php?id=perseus:user:activities:matrixprocessing:learning:classificationparameteroptimization
[11] Cortes, C., Vapnik, V., 1995, Support-Vector Network, Machine Learning, 20(3): 273–297.
[12] Osowski, S., Siwekand, K., and Markiewicz, T. (2004). MLP and SVM Networks. In: Proceedings of the 6th Nordic Signal Processing Symposium. (pp.37-40).
[13] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[14] Dirican A., R.O.C. Curriculum analysis and evaluation of diagnostic tests and computer application, İstanbul University, Health Sciences Institute, Doctoral Thesis,Istanbul,1991.