BLG 456E Learning From Data Term Project Report

Buse Sibel Korkmaz

Gülnur Kaya

***Abstract*—The most important enemy of e-commerce companies is return. Most of them give free-return guarantee to customers in order to increase customer happiness. But it causes trouble in nowadays. New goal for firms, finding solution without damaging customer happiness. The aim of this research is predicting that the product will return or not with machine learning methods according to historical data. The data includes several information about customers and products. The importance of this study is that generating new features which increases accuracy of prediction.**

I. INTRODUCTION

Online shopping is popular but dangerous business. Online shopping gives easiness to retailers about some issues such as rent of office or salaries of workers but it has some disadvantages. Retailers give free-return guarantee to increase customer happiness but it causes high cost of returns. According to Forbes, even well-known retail brands faced with return rates over 30% from divisions of their catalog [1]. The aim of this study forecasting that the product will return or not with machine learning methods according to historical data. The dataset consists information about customers and products. The importance of this study is that generating new features with feature engineering which increases accuracy of prediction.

The dataset which is used in this research was represented to "Data Mining Cup Competition" (DMC-2014) [2]. In the competition, it is desirable to train artificial learning models that are trained through training data sets and classified data that are completely independent of the test set and that do not include class knowledge.

II. DATA SET USED

The data set which is studying in this research belongs to commercial organization. The data sources from shopping records of the organization. The data set consists customer and product record and return status for 481092 shopping. The variables in dataset are shown in Table I. They used with original names for the sake of intelligibility.

Table I: Original Data Set

Features	Class of Feature	Values of Feature	Number of missing data
orderItemID	number	1 2 3 4 5 6 7 8 9..	0
orderDate	date	"2012-04-01"…	0
deliveryDate	date	"1990-12-31",…	39419
itemID	number	186 71 71 22 151 …	0
size	category	"1","10","10+",..	0
color	category	"amethyst",..	143
manufacturerID	number	25 21 21 14 53 87..	0
price	number	69.9 70 70 39.9 …	0
customerID	category	794 794 794 808..	0
salutation	category	Company,"Family",..	0
dateOfBirth	date	"1655-04-19",…	48889
state	category	"Baden-Wuerttemberg",..	0
returnShipment	label of class	0,1	0

In data preparation part, missing values were replaced with most repetitive values. In colors, symbols of '?' were replaced with black. In deliveryDate symbols of '?' were replaced approximate date according to mean of deliverytime. In size feature there are both categorical and numerical data. For conversion numerical to categorical of size feature we used Table II for women and Table III for men [3]. Categorical values were converted to numerical values. We generated deliveryTime feature according to deliveryDate and orderDate, age feature according to dateOfBirth and membershipTime according to creationDate from given features in order to increase accuracy.

Table II: Size conversion for women

International Size	US Size	UK Size	French Size	Italian Size	German Size
XS	4	6	34	36	32
XS	6	8	36	38	34
S	8	10	38	40	36
S	10	12	40	42	38
M	12	14	42	44	40
M	16	16	44	46	42
L	18	18	46	48	44
L	20	20	50	50	46
XL	22	22	52	52	48
XL	24	24	54	54	50
XXL	26	26	56	56	52
XXL	28	28	58	58	54

Table III. Size conversion for men

US/ International Size	UK Size	French Size	Italian Size	Euro Size
XS	13,5	35	38	35
XS	14,0	36	39	36
S	14,5	36	40	37
S	15,0	37	41	38
M	15,5	37	42	39
M	15,5	38	43	40
L	16,0	38	44	41
L	16,5	39	45	42
XL	17,0	39	46	43
XL	17,5	40	47	44
XXL	18,0	40	49	45
XXL	18,5	41	52	46

New features were created

deliveryTime: According to orderDate and deliveryDate deliveryTime was calculated. Then, mean of deliveryTime was found and it was used in order to fill '?' values in deliveryDate.

age: According to dateOfBirth feature age of customers were calculated and splitted three category: Customer who older than 55, who is between 55 and 35 and who is younger than 35. dateOfBirth feature is replaced with age feature. In training the instances which has '?' value in dateOfBirth was dropped.

importantDate: If ordered item is a special day gift and it is delivered after than special day, return probability of item increases. In this study valentine's day and new year is investigated.

membershipTime: With using creationDate customer's membershipTime was calculated and replaced this with creationDate in data set.

itemID: Each item and returnShipment value of item was grouped. After this, return probability of each item is calculated and replaced with itemID.

manufacturerID: Each manufacturer and returnShipment value of manufacturer was grouped. After this, return probability of each item which is produced by the manufacturer is calculated and replaced with manufacturerID.

customerID: Each customer and returnShipment value of customer was grouped. After this, return probability of each customer is calculated and replaced with customerID.

size: Each size and returnShipment value of size was grouped. After this, return probability of each size is calculated and replaced with sizeID.

In training, %25 of instances were excluded to use in testing our model. In testdata return probabilities of customerID, itemID and manufacturerID is calculated according to trainingdata information because returnShipment information does not exist in testdata.

III. METHODS USED

In this project, scikit-learn Gaussian Naïve Bayes libraray was used. State, color, salutation and orderItemID were excluded feature set because they decrease the accuracy. Bagging and boosting algorithms were also tried but they did not improve the accuracy so they were not used. Neural network, Logistic, Decision Tree and Random Forest were also tried but same reason they were not used. Support Vector Machine classifier also tried but it is time consumer algorithm so it was not selected.

IV. RESULTS

As a result, according to error scores we decided to used Naïve Bayes. Some error score on test data which is separated from given training data.

Table IV. Error Scores

Classifier	Error
Decision Tree	40.033
Bagging	52.102
Naive Bayes	37.145
AdaBoost	37.940
Logistic	38.766
Neural Network	38.766

V. CONCLUSIONS

In conclusion, according to Kaggle results our team placed 18th . Our error score is 7901. This project give us to experience challenging competition ambiance and we had to chance to use every method which we learned in class. It is a good implementation practice.

REFERENCES

[1] "The ticking time bomb of e-commerce returns", https://www.forbes.com/sites/stevendennis/2018/02/14/the-ticking-time-bomb-of-e-commerce-returns/#10b1ffeb4c7f

[2] "Data-mining-cup (DMC) 2014 task", http://www.data-miningcup.de/en/service/download-center/.

[3] "European sizes conversion", https://www.blitzresults.com/en/european-sizes/