BLG 454E Learning From Data
Term Project Report

Ismail Salih Namdar, Kadir Emre Oto, Suheyl Emre Karabela

***Abstract*— Returns of items in online shopping affects both company and store profit. In order to prevent and reduce the number of returns, machine learning techniques are used. The historical data, which will be used as a training data set, from the company has some attributes. However, in order to improve prediction of returns data set is cleaned and restructured. Some features are extracted and some features are added to dataset to increase prediction. After that, tensorflow, randomforest, sklearn and keras machine learning techniques are applied to that data set.**

I. INTRODUCTION

Online shopping companies are worried about increasing number of returning items from customers. For that reason, both companies and store's cost are increasing rapidly and this causes dissatisfaction to everyone because even customer do not want to deal with returning items. So, in order to improve user experience and reduce cost, machine learning algorithms will be used to prevent returns while keeping user satisfaction high as possible. Historical data set that contains orders from customer will be used as a training data set. Our Kaggle team name is: 150140032_150140109_150140055. We scored 6283 at the Kaggle class competition.

II. DATA SET USED

In this study, historical shopping dataset is used. Data set contains several attributes about the sale customer and store. Firstly, data must be cleaned because it has some missing and corrupted values. Our approuch to data cleaning is similar to Prediction of Return in Online Shooping article [1]. Delivery date attribute is corrupted in some of the orders. So, we calculated this by adding average delivery time to order date. Color attribute is missing in some places. This will corrected by adding value of black to missing orders. We calculated age of the customer and the membership time and added these as features. We checked the closeness of the order date to some special days like valentine's day. Again this data cleaning approaches is identical to Prediction of Return in Online Shooping article [1]. Lastly, orders which has corrupted birthday date are removed from the date set in order to prevent some misprediction that can cause.

III. METHODS USED

We used python programming language because it has wide variaty of machine learning libraries. It is a great language for data analyze and aplling machine learning algorithms. We added and substracted some features in data set. We converted delivery date and order date to delivery time and we calculated customer's age. Also, special dates are added. In addition to that, we added risk factor to item, customer and manufacturer. Risk factor calculation is done similar to the referenced paper [1]. We used tensorflow, randomforest, sklearn and keras. In tensorflow, FtrlOptimizer is used for optimizing weights. We obtained %46.2 accuracy in tensorflow. After that, we used random forest algorithm in scikit-learn. In random forest, we calculated importance of each feature in data set. Importance of features can be seen in Figure 1.

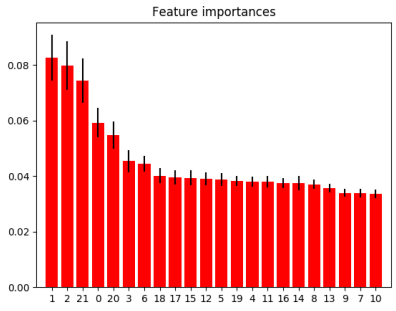

Fig. 1. Feature Importances

However, random forest algorithm reached maximum %66 accuracy. We also implemented a neural network in keras with adam optimizer which reaches %66 accuracy.

IV. RESULTS

We used the splitted test data accuracy and Kaggle competition loss to evaluate the model. Both keras and sklearn provides the loss and accuracy information after each iteration of the algorithm. We trained the models with different hyperparameters and different sized models. Our accuracy didn't improve after %66.

V. CONCLUSIONS

In conclusion, we first have done feature extraction. After that, we converted and dropped some features from historical data set in order to supply the machine learning algorithm with more meaningful data.

REFERENCES

[1] İ. Bilgen and Ö. S. Saraç, "Prediction of return in online shopping," in *Signal Processing and Communications Applications Conference (SIU), 2015 23th*, pp. 2577–2580, IEEE, 2015.