BOGAZICI UNIVERSITY

INDUSTRIAL ENGINEERING

IE 582 PROJECT REPORT

2021-2022 FALL

GROUP 5

2021702093 Mert Güneş

2021702135 Ayşe Nur Sülün

CONTENTS

I. INTRODUCTION

The project is based on solving a data problem based on e-commerce, which is one of the most important big data sources of today. Unlike previous assignments, this time we need to focus on working with more raw data. While it is aimed to examine the customers separately and to make a gender estimation for each customer throughout the problem, all the actions of a customer are included in the data at hand. Considering that a train data of more than 5 million lines belongs to approximately 5600 customers in total, feature engineering becomes very important for this problem.

While Python was used as the programming language in the project, the main libraries were determined as "Pandas", "Numpy", "Seaborn" and "Sklearn". While Pandas and Numpy are mostly used in data processing, Sklearn library is sufficient for data of this size in modeling. Apart from this, Seaborn was used for the necessary graphics.

If we look at the features in the raw data that we can use, user_action, sellingprice and categories come to the fore. Apart from this, features such as the time of action and product gender can also be seen as suitable for use in the model. Since most of the available data is categorical data, while encoding is needed, numerical feature “price” can be taught directly. Finally, since the feature to be estimated is also categorical, it can be said that there is a classification problem.

	time_stamp	contentid	user_action	sellingprice	product_name	brand_id	brand_name	businessunit	product_gender	category_id	Level1_Cate
0	2020-12-02T22:26:14.023Z	39918893.0	favorite	3099.00	PerfectCare 600 EW6F449ST A+++ 9 KG 1400 Devir...	8511.0	Electrolux	Beyaz Eşya	Unisex	1272.0	
1	2020-12-08T23:15:04.603Z	3558544.0	favorite	3079.00	WW90J5475FW A+++ 1400 Devir 9 kg Çamaşır Makinesi	3228.0	Samsung	Beyaz Eşya	NaN	1272.0	
2	2020-12-05T16:19:01.157Z	31292729.0	favorite	3999.00	KM 9711 A++ 9 kg Çamaşır Kurutma Makinesi	10989.0	Vestel	Beyaz Eşya	Unisex	1276.0	
3	2020-12-05T16:28:00Z	6363103.0	visit	2544.00	CMI 9710 A+++ 1000 Devir 9 kg Çamaşır Makinesi	10989.0	Vestel	Beyaz Eşya	NaN	1272.0	
4	2020-12-02T22:26:59Z	39918893.0	visit	3099.00	PerfectCare 600 EW6F449ST A+++ 9 KG 1400 Devir...	8511.0	Electrolux	Beyaz Eşya	Unisex	1272.0	

Figure I.1 Head of Train Data

II. APPROACH

Regarding the approach to the solution of the problem, first feature engineering was carried out on the available data. To elaborate on this part, a total of 10 different features have been created. To explain these features in order:

-Selling Price: It is the average of the prices of the products investigated by the customer.
-Basket: It shows the number of products that the customer puts in the basket.
-Favorite: Shows the number of products that the customer has added to their favorites.
-Order: It shows the number of products ordered by the customer.
-Search: It shows the number of products the customer is looking for.
-Level1: It is the Level 1 Category which the customer takes the most action.
-Level2: It is the Level 2 Category which the customer takes the most action.
-Level3: It is the Level 3 Category which the customer takes the most action.
-Product Gender: It is the product gender that the customer takes the most action on.

These features are created for both test and train data. Null data are filled with "Unisex" for product gender and "0" for the rest. All features were combined by accepting the customer ID as an index and the necessary data for the model was created.

In addition, although standardization is not considered necessary in the model, categorical features have been transformed by considering the number of categories. For example, since Level 1 feature contains too many categories, its encoding is done with Label Encoding, while One Hot Encoding is used for Product Gender.

The feature to be predicted "Gender" is divided into two features as Male and Female, Female feature is separated from train data for prediction, Male feature is dropped.

Finally, IsolationForest was tried to preprocess the data. Although it is thought that it would be correct to discard the outlier data, not discarding this data affected the results better. It can be said that the diversity in the data is more beneficial.

Since the labels of the test data at hand are missing, the train data is divided into validation data to test the model. In doing so, two different methods were followed. First of all, since the first lines of the data are more regular, a code was written that aims to use them as validation, and tests were carried out on it. In the second, the data was randomly divided so that the gender distribution was compatible with the train data. Close results were obtained in both methods. The former was more overfitting, while the latter lowered the scores a little more.

```
X_train_last2 = X_train_last[:1000]
X_train_last3 = X_train_last[1000:]
y_train_last2 = y_train_last[:1000]
y_train_last3 = y_train_last[1000:]
```

```
st2,y_train_last3,y_train_last2 = train_test_split(X_train_last,y_train_last,test_size=0.3,random_state=42,stratify=y_train_last)
```

Figure II.1 Preparing of Validation Data

In the last part, estimation was made with four different models:

First, a basic classification model, Logistic Regression, was used. It can be said that it is an ideal model for less complex data. While there is no need for an extra parameter tuning, the model is sufficient by simply increasing the number of iterations.

As the second model, Random Forests algorithm, estimator number and max feature number were used by tuning with 5 cross validations. It has been seen that the model created with Max 3 features, 200 estimators is one of the best models.

As the third model, Extra Gradient Boosting, which can be considered as one of the deepest algorithms, was used, learning rate, max depth and estimator number were tuned with 5 cross validations. 0.1 learning rate, 10 max depth and 50 estimator were seen as the most suitable parameters.

Finally, the Support Vector Machines algorithm is used as a different perspective. C and model degrees were tuned, but sufficient results could not be achieved in this model.

III. RESULTS

In the results of the model, the accuracy scores, Roc Auc scores and balanced accuracy of the 4 models were compared. The results are given below, respectively.

Logistic Regression: Accuracy 0.834, Roc_Auc_Score 0.8, Balanced_Accuracy_Score 0.8
Random Forests: Accuracy 0.872, Roc_Auc_Score 0.8, Balanced_Accuracy_Score 0.8
Extra Gradient Boosting: Accuracy 0.84, Roc_Auc_Score 0.81, Balanced_Accuracy_Score 0.81
Support Vector Machines: Accuracy 0.65, Roc_Auc_Score 0.5, Balanced_Accuracy_Score 0.5

Although the best model looks like Random Forests when all scores are examined, the best model in the leaderboard scores is XGBoost because of its Balanced Accuracy Performance, so the final submission was made on its. Confusion Matrix is below. True Positive Ratio is enough, but cannot be said same for True Negatives.

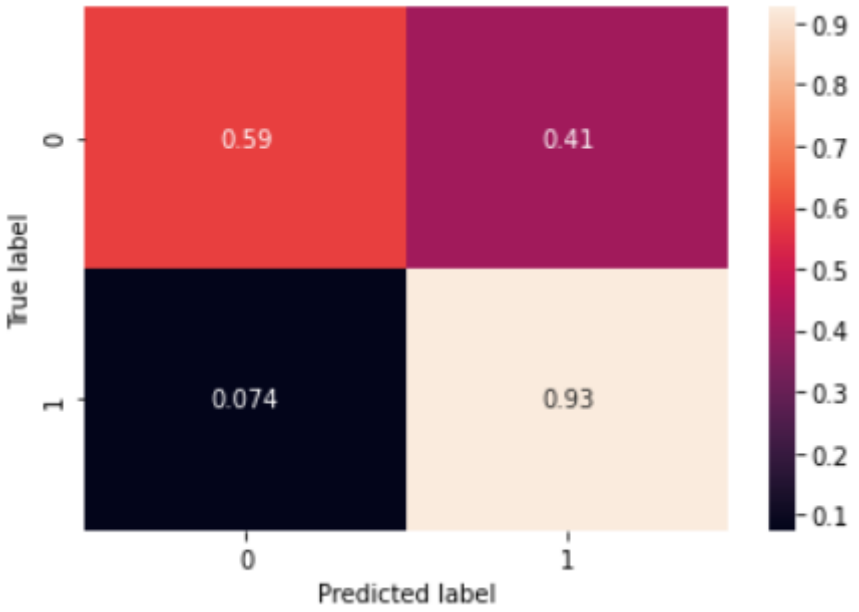

Figure III.1 Confusion Matrix

IV. CONCLUSION AND FUTURE WORK

As a result, when the best model was taught, 0.8356 Auc Score and 0.71 Balanced Accuracy Score were obtained according to the leaderboard. In terms of Feature Engineering, it can be said that it is a project that improves people in many respects, considering that it is worked with e-commerce data and estimation is made on raw data. Accurate estimation of customer gender above 0.8 is also important in terms of examining big data. The low Balanced Score may also be associated with the fact that the data has more female customers than male customers.

In terms of future work, 10 features can be seen as insufficient. When models are taught with more features and more RAM, better results can be obtained when all parameters are tuned for each model. Unfortunately, my friend Ayşe Nur decided to leave her master's degree after the project groups became clear, so it became a project that I (Mert Güneş) carried out alone. It can also be said that it is a project that can be solved better with group work. However, I wanted to write his name to the project because of his initial help, for your information.

Code Link:

https://github.com/BU-IE-582/fall21-mrtgnsboun/blob/gh-pages/Project%20IE%2058 2%20Mert%20G%C3%BCne%C5%9F%202021702093.ipynb