# Review of Data Science Technique Applied to Online Food Delivery Services

Volkan Mazlum, Mert Alp Kuvandık

201811045 - 201911038

Department of Computer Engineering, University of Cankaya

CENG-480

Term-Project
Dec 12, 2022

## 1 Introduction

I will not re-explain this dataset obtained from the literature. In general, the codes are based on the M6Class column. Different algorithms are already used by everyone. Our aim in this study is to develop a basic process to reach a result, and the result we get at the end of this process has a visible value. The process we apply covers the basic processes in the field of Data Science. This includes processes in the below:

- DATA CLEANING,
- DATA INTEGRATION,
- DATA TRANSFORMATION,
- DATA REDUCTION,
- DATA DISCRETIZATION,
- FEATURE ENGINEERING,
- MODEL and EVALUATION

These processes will be discussed in detail in the Experimental Setup section. To briefly talk about the problems, first, it is necessary to get rid of the missing values and outliers' data in the dataset. After these are cleared, dimensionally reduction operations can be

performed in cases where the data is large. In this section, the technique of deleting one of the ones with very high correlation values can be applied or the most important features can be obtained by using different algorithms such as SelectKBest. These processes can be performed with techniques such as PCA. There are examples of all of these in our project. Normally, categorical data is converted to binary data with techniques such as one-hot encoding, but in this dataset, these operations have been done before. Data transformation parts are ready in the dataset. We did not do any extra work on these parts. In addition, metrics obtained with different cross-validation techniques were tried to be compared. It is aimed to take the research one step further by using different algorithms, not just basic algorithms like KNN. We did all our estimation and work for the M6Class feature. Using other features, we tried to find the most suitable restaurant for the customer to order food.

## 2 Experimental Setup

First of all, I would like to talk about the Data Cleaning phase. A missing value search was carried out for the non-categorical features of the extracted dataset. Afterward, the correlation map was obtained. From here, those with a high correlation relationship were excluded. (Dimensionality Reduction) Max, and min values were checked, and outlier detection was tried (using the z score). The detected outliers were filled with the median values of the features. Afterward, visualization techniques were used to determine whether there was a normal distribution or not.

In the Data Reduction phase, it was first aimed to show, step by step, how the PCA algorithm works. Later, dimensionality reduction was achieved with graphics by using different Sparse PCA, Kernel PCA, Incremental PCA, Locally Linear, and many other PCA derivatives. After these processes, accuracy values were obtained thanks to the train and test setters, and comparisons were made. In general, accuracy values of around 0.96 were obtained.

In the Data Discretization phase, the available features were tried to be divided into intervals. Using these intervals, estimation was done with the decision tree algorithm. During the Feature Engineering phase, 5 highest value features ('DistanceKm', 'TimeMinutes', 'M1DeliveryCost', 'M5DeliveryTimeFulfillment', 'M6DeliveryCostPerKm') were selected using the RFE algorithm. After selection, learning and estimation processes were carried out with the Random Forest algorithm. An unrealistic accuracy of 1 was obtained. So, we decided to use the Monte Carlo Cross-Validation technique, but it did not solve the problem. We decided to use different algorithms instead of these algorithms. Later, Polynomial Feature and Linear Regression were used together. A very high accuracy value of 0.98 was obtained. Then, the order of importance of the features was prepared with the ExtraTreesClassifier. These data were graphed. Later, the KNN algorithm was tested on this data. An accuracy of 0.97 and a ROC of 0.90 was obtained.

In the Model and Evaluation phase, it was decided to use dataset 2 instead of dataset 1. The same operations were performed again. Dimensions were reduced by selecting certain features. The XGBoost algorithm, which was intended to use data from the very beginning, was used at this stage. 0.99 accuracy, 0.97 ROC-AUC value was obtained. Later, the ANN algorithm was also applied. 0.97 accuracy, 0.98 AUC value was obtained. But a result as good as XGBoost has not been achieved so far. Although more than one algorithm has been tried in the project, XGBoost seems to be the most suitable algorithm for this dataset. It is not expected that the dataset, which has gone through operations such as cleaning and feature selection in general, will give a bad result in any algorithm. Because this was our data purpose from the beginning of the project. As a result, we have not seen any value below 0.90 so far. In a subsection, the results obtained in tabular form will be given. There is an opportunity to examine this in more detail. Next, we wanted to take this research one step further and examine how much the results changed using oversampling. Through a detailed examination of the problems in the Imbalanced datasets, we can see the differences.

# 3 Results

| Algorithms | Additional algorithm | Accuracy | ROC |
|---|---|---|---|
| LinearRegression | PolynomialFeatures | 0.98 | - |
| KNN | - | 0.97 | 0.90 |
| **XGBoost** | **K-Fold** | **0.99** | **0.99** |
| ANN | - | 0.98 | 0.99 |
| KNN | Kernel PCA | 0.95 | 0.90 |
| Random Forest | RFE + Monte Carlo Cross Validation | 0.98 | 0.98 |
| Decision Tree | Discretization + KFold | 0.99 | - |
| XGBoost | - | 0.99 | 0.97 |

```
######### Testing Accuracy Results #########
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        52
           1       1.00      1.00      1.00      1803

    accuracy                           1.00      1855
   macro avg       1.00      0.99      1.00      1855
weighted avg       1.00      1.00      1.00      1855

[[  51    1]
 [   0 1803]]
```

Figure-1 for XGBoost

```
######### Testing Accuracy Results #########
              precision    recall  f1-score   support

           0       0.75      0.69      0.72        52
           1       0.99      0.99      0.99      1803

    accuracy                           0.98      1855
   macro avg       0.87      0.84      0.86      1855
weighted avg       0.98      0.98      0.98      1855

[[  36   16]
 [  12 1791]]
```

Figure-2 for ANN

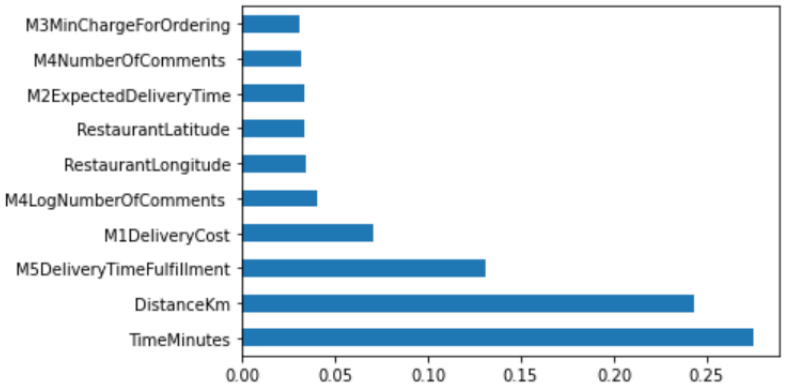

Figure-3 for Feature Importance

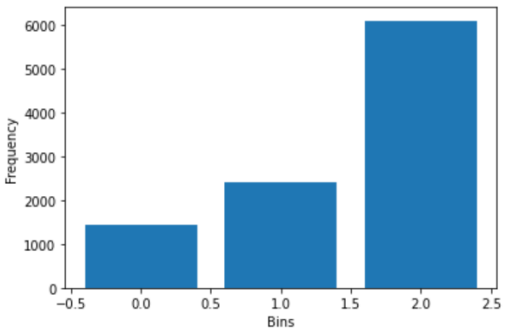


Figure-4 for Discretization for M1DeliveryCost

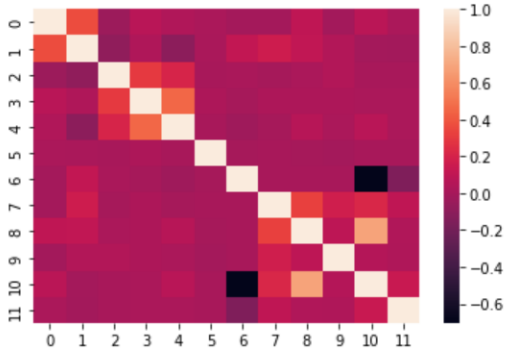


Figure-5 for Correlation

| Algorithms | Additional algorithm | Accuracy | ROC |
|---|---|---|---|
| KNN | - | 0.98 | 0.98 |
| **XGBoost** | **K-Fold** | **0.99** | **0.99** |
| Random Forest | RFE + Monte Carlo Cross Validation | 0.99 | 0.99 |
| Decision Tree | Discretization + KFold | 0.99 | - |
| XGBoost | - | 0.99 | 0.99 |

Table for Solution and Metrics to Imbalanced Dataset

```
######### Testing Accuracy Results #########
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00       906
         1.0       1.00      1.00      1.00       901

    accuracy                           1.00      1807
   macro avg       1.00      1.00      1.00      1807
weighted avg       1.00      1.00      1.00      1807

[[904   2]
 [  1 900]]
```

Figure-6 for XGBoost with Oversampling

```
######### Testing Accuracy Results #########
              precision    recall  f1-score   suppo

         0.0       0.98      0.98      0.98      9
         1.0       0.98      0.98      0.98      9

    accuracy                           0.98     18
   macro avg       0.98      0.98      0.98     18
weighted avg       0.98      0.98      0.98     18

[[892  14]
 [ 17 884]]
```

Figure-7 for KNN with Oversampling

# 4 Conclusion

Different algorithms were tried on the dataset and different results were obtained. As can be seen in the upper part, there are noticeable high values in all of them. We would like to point out that these processes, which we apply to us, work for us and can be easily applied in future studies. Based on these processes, which everyone knows, the research can be done quickly, and the results can be reached easily. When we look at the values, XGBoost seems to be the most usable algorithm. It has already been seen to be very successful in most studies in the literature. That's why we wanted to include it in our project.