Data Mining Classification Assignment

Oguz Birdal

April 26, 2019

1 Abstract

Explanation of the experimental setup and answers for the five questions in classification assignment can be found in this report. Identifier list of properties, which are labeled as true to take a contract, are can be found in "identifiers.txt". All the code implemented to make predictions are in the "classification.py".

2 Answers

2.1 Question 1

The existing company have far less true values in prediction columns than "high-Value" column. It can be seen that they were too fussy while taking on contracts. Taking on contract also requires some money and it is possible that they took on contracts until they run out of budget. Also the company can earn money if they can find a buyer for the property. Otherwise they will loose money even if the property is high valued so rather than taking on contracts for every high valued property, classifying the properties which buyers would be interested to buy and took on contracts for them is better for company's profit.

2.2 Question 2

In the "train.csv" file the prediction of the existing company and the true labels of high valued properties are available. I decided to use graphs to see if I can find any clue about which method is used for prediction. I used histograms below and focused on splits for binary classifying for each feature. True and false labels were separated strictly for some features like "district" as can be seen in figure. I tried to find where the tree branches out at top if a decision tree algorithm is used for prediction. The splits in the graph strengthen the probability of a decision tree algorithm be used. To understand which type of classification model was used by the other company I decided to train some models on "train.csv" according to "prediction" column in the file and compare the results. Then I split the "train.csv" file into two dataframes to train the model with %80 of the data and test it with the other %20. Because of the strong chance of a decision tree model was used, I trained the first model using random forest classifier which gave the best precision 0.982. I also trained models using k-nearest neighbor, naive bayes and logistic regression. These models gave precisions 0.901, 0.856 and 0.831 respectively. It can be seen with

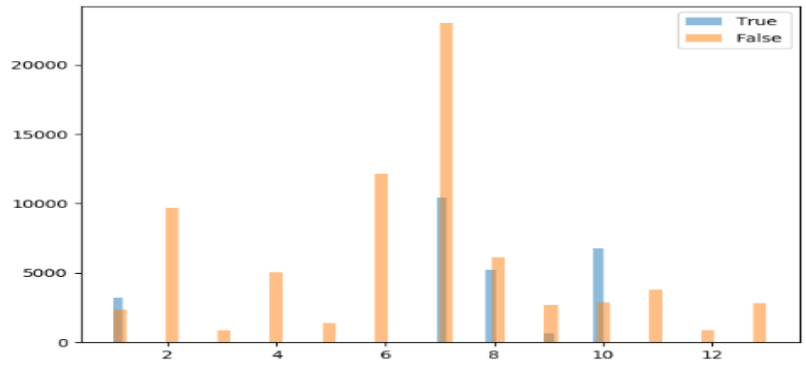

Figure 1: Distribution of district column

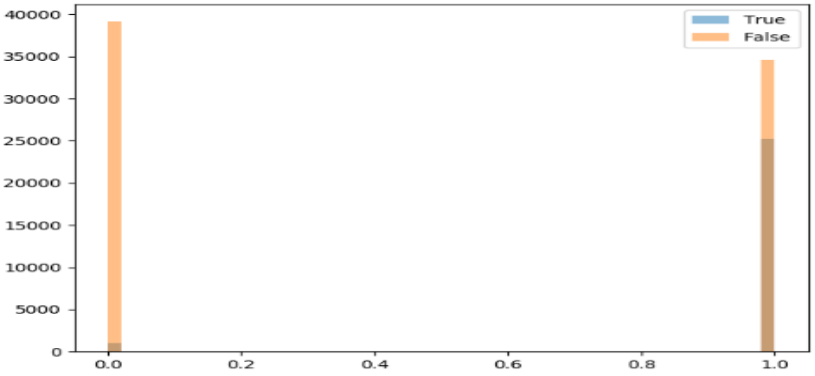

Figure 2: Distribution of subway column

all results that a decision tree model was used to make a prediction by existing company.

2.3 Question 3

I used the same Random Forest model which gave the best precision in the previous question for training set but this time I trained the model with whole "train.csv" and estimated the "highValue" and "prediction" columns separately for "test.csv". To calculate the estimated profit that existing company would make, I kept the rows with true labels on prediction column. If "prediction" and "highValue" columns are both true it means this rows are our true positives and the company will earn 600 EUR and spend 450 EUR for each . If "prediction" column is true and "highValue" column is false, it means that the company will earn 100 EUR and still spend 450 EUR for each property. The calculations can be seen in the code file. According to estimations of random forest model the true positive count is 4239 while false positive count is 860. Following the calculations profit assumption of the existing company is 334850 EUR.

2.4 Question 4

To maximize my company's profit the most important things to do are creating a model to predict high valued properties in "test.csv" and predicting how existing company made their predictions because we should not take same contracts with them to not lose money. I already created a %98 similar model to existing company using random forest classifier in the previous questions. Next step is predicting high valued properties in "test.csv" with a proper classifier. To find the best classifier I tried different classifiers to predict "highValue" column in "train.csv" such as random forest, k-neighbors, gaussian naive bayes, logistic regression, multilayer perceptron, stochastic gradient descent and linear support vector. Between these classifiers best results were belong to random forest classifier again with 0.76 precision. I have made a bit of normalization in the features and tried removing some features but the best results came out with using all of the features. The second best precision was belong to k-neighbor classifier. I tried to run the model with different number of neighbors and chosen odd numbers because the number of the classes are even. The best precision 0.759 was with 7 neighbors. Then I tried naive bayes and logistic regression which had the precisions 0.706 and 0.628 respectively. SGD, MLP and LinearSVC did not result very well. Especially MLP and LinearSVC predicted labels all true or all false that gave a precision around 0.50 which is the worst precision possible for a binary classifier. In the light of such information I decided to use random forest model to predict high valued properties in "test.csv". After predicting high valued properties in "test.csv" I dropped the rows that I estimated the existing company already took on contract. Finally I defined the properties to take contracts which has true labels on prediction column and not taken by the existing company. The identifiers can be found in "identifier.txt" file.

2.5 Question 5

I decided to take contracts to all high valued properties which the existing company did not take on contract. In the "identifier.txt" there are 4837 properties that are going to be taken by our company. The model that I used had 0.76 precision so when we take estimation errors into account we can say that number of true positive would be 3676 and false positives would be 1161 approximately. Following the calculations profit assumption of the company is 145110 EUR. I estimated the prediction of the existing company before with the model which has 0.98 precision. So if we calculate the chance of overlapping with existing company, they will earn 30000 EUR less, which is 304850 EUR, than in case of monopoly in the market.