LAB4

GROUP DISCUSION

We noticed that our datasets contain a lot of categorical variables. !0 of the 17 columns have the data type “object”, 3 features are Boolean and 4 are either float/int. During feature selection we tried to perform the initial Pearson’s correlation technique to select the most correlated features like it was advised in the lab tutorial. However, this method only applies to numeric variables. We opted to a Chi squared method of feature selection. After converting all our categorical to numeric using one-hot-Encoding from SkLearn, we noticed that features that have a lot of categories (city, state, threat level) lead to a lot of columns being created in our data Frame. This may lead to issues of multicollinearity among various variables, lowering model accuracy.

The flowing shows the first 10 results after performing the CHI SQURED method of variables selection. The features are listed in order of score and decreasing P-values.

Table

Description automatically generated

We went ahead a fitted a Logistic regression using all the features in our database. The following is the confusion matrix after fitting the model.

ACCURACY BASE MODEL : 0.7555205047318612

Classification Report :

precision recall f1-score support

False 0.77 0.97 0.86 954

True 0.54 0.09 0.16 314

accuracy 0.76 1268

macro avg 0.65 0.53 0.51 1268

weighted avg 0.71 0.76 0.68 1268

The aforementioned issues of multicollinearity complicated our process of feature selection so after consulting with our tutor he advised we pick and work on a different dataset that matched our level of python machine learning experience.

CONCLUSION

We picked a deferent dataset (Market Data) with a good mixture of categorical and numeric features to work with. So far, we have managed to load the dataset and perform Initial Data Exploration and feature selection using forward and backwards methods as well and Person’s correlation. We are planning to complete fitting Logistic regression as well as KNN models this coming weekend. This also include making a comparison between the two in order to pick the best model.