Data Mining

CSCI-6401-01

Phase 6 – Optimization

Team Name

Technocrats

Team Members

Sravani Thoomuganti (Team Lead) – sthoo1@unh.newhaven.edu
Kamal Siddharth Teki – kteki1@unh.newhaven.edu
Sai Teja Gunda – sgund9@unh.newhaven.edu

Company Name

DoorDash

Datasets

https://www.kaggle.com/code/deyashinichakravorty/doordash-dataanalysis/notebook

https://platform.stratascratch.com/data-projects/delivery-durationprediction?tabname=solution

About the Datasets

The datasets we chose date back two years and come from a trusted domain. The door dash data analysis datasets will be used by us. Since we obtained the data from the official Kaggle website, which has historical stock data, we can guarantee that it is accurate to the nth degree. We also got dataset of delivery duration prediction which we also use in our research. We have taken this from strata scratch website.

Research Question

Ways to identify fraudulent activity on the DoorDash platform, such as delivery scams?

Data Modelling Techniques

If we go through out data presented in the datasets most of the data will be in the form of int and float values which are almost numbers rather than any other format. So, we have decided to use the regression data modelling technique where we will be using decision tree regression where the data is converted into subsets of data with a parent and child node format in order to detect the accuracy of the delivery scams being going on.

Parameters & Hyper Parameters

There are various hyper parameters used in this regression model like criterion, max_depth, min_samples_split, min_samples_leaf, max_features which are used to to train the model for the prediction of the delivery scams. Here, criterion is the mean squared error 'mse' which is set by default used for our purpose. max_depth can be said to be the maximum depth of the decision tree we have generated where the tree has various leaves which are nodes continuing till the end. Whereas min_samples_split can be defined as the minimum number of samples that are much needed by the decision tree to split the internal node with the default as 2. Similarly, min_samples_leaf is the minimum number of samples that are need at leaf node level in the decision tree. Here we have trained the required columns after pre-processing.

Other than this, we have also utilised parameters like random_state which is used to gain the control over generating the random number using which are done by an algorithm that is the primary reason in splitting the nodes at decision tree level in this regression modelling.

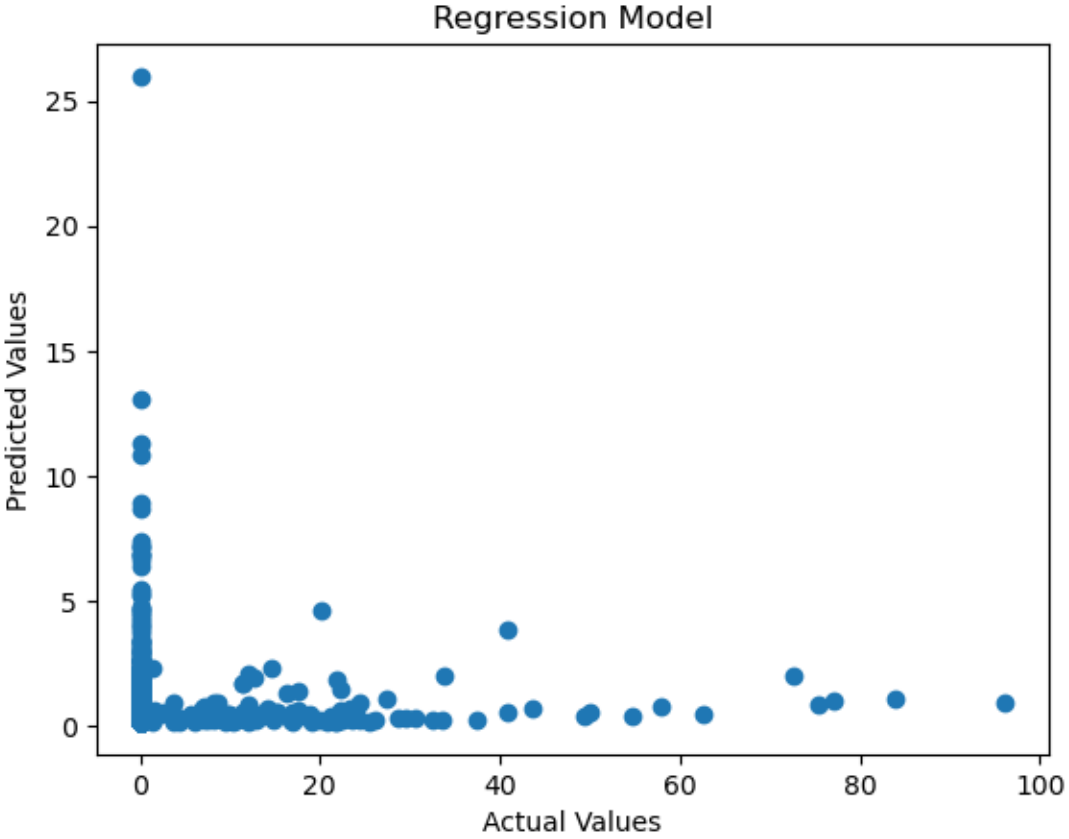
Optimization Techniques

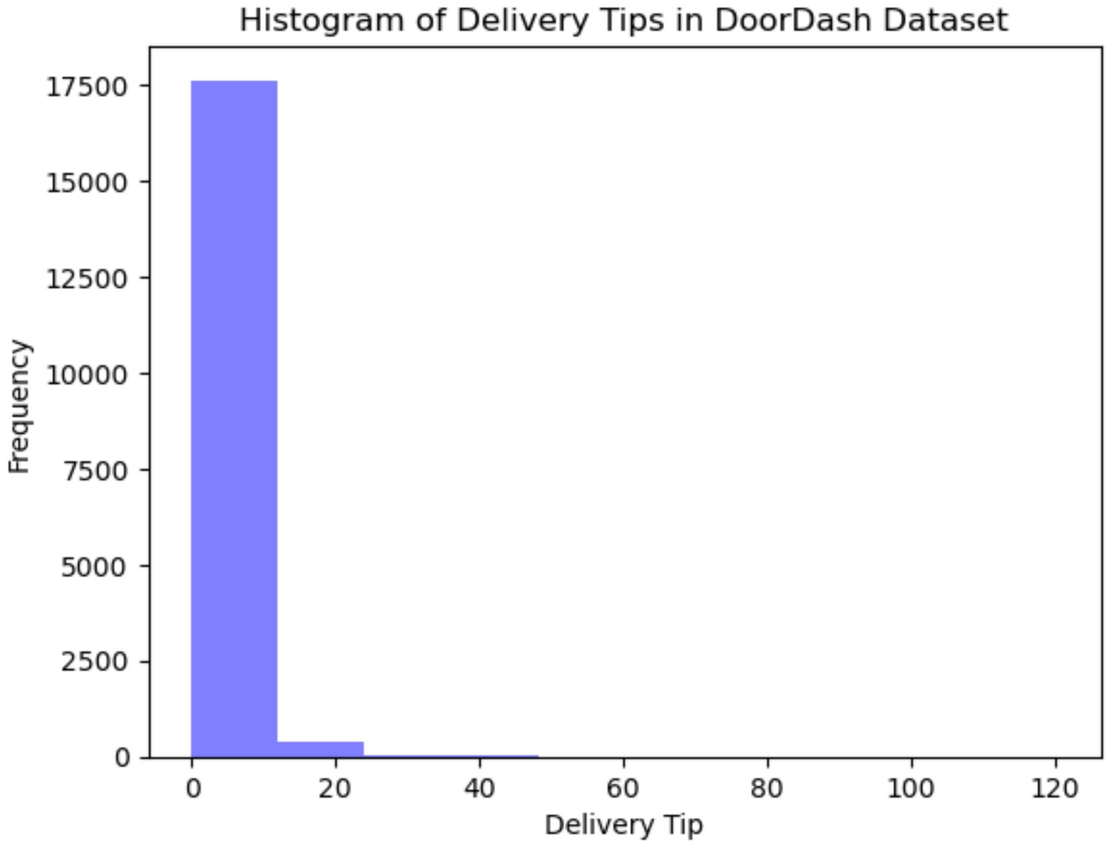
- Grid Search Cross Validation
- Randomized Search Cross Validation
- Feature Selection

For optimization we have used the random forest which is utilized to build various or multiple decision trees and then merges these trees in order to get the efficient accuracy for the problem or research question. So, for the optimisation we have split the data to train and test the data and we have chosen the decision tree regression for regression model, which indeed is used to train the model and then we have tune the model, as the results we acquired through data modelling is not enough or satisfied with the output, This is would increase the efficiency of the output. So, we have used grid search with cross-validation technique, randomized search with cross validation and feature selection from the random forest to acquire the output for hyperparameter tuning.

Visualization

For the visualization we have used matplotlib to visualise the trained model and especially matplotlib.pyplot has been used by importing it to the python kernel in the notebook to create the scatterplot and histograms in order to identify the fake reviews by analysing the delivery tips to check for the real accounts for the visualisation purpose.





```
MSE: 22.796789869701914
R-squared: 0.03736733930377434
```

Conclusion

So we here conclude that in order to proceed with the optimization we have utilized Random Forest model along with the hyperparameter tuning and also the feature selection. Along with these we have used the grid search cross validation in order to best hyperparameters that can be utilized for the model, where we used the feature selection to choose the most important ones from the random forest. By using these optimization techniques we are able to improve the accuracy of the trained model in order to identify the fake reviews and also we have used visualisation to show a analysis on the data validation we have made by improving and training the model

https://github.com/kamalsiddharthteki/Team-Technocarts