Data Mining

CSCI-6401-01

Phase 5 – Data Modelling

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.

Hardware used :

Tool : Jupyter Notebook
System : Apple M1 Chip Version 13.2
Memory : RAM 8GB
System-Type : 64-bit MAC OS

Outcomes

As we have chosen the regression modelling technique for the dataset and the research question we have, we have to use the decision tree regression in order to identify the delivery scams being going on. So, far before creating the decision tree we have to pre-process the data in the dataset first in order to remove the irrelevant data that is present in the dataset, So we have performed pre-processing but including only the columns that are required by the project to train the data. So here we have excluded irrelevant data like Restaurant_ID, Driver_at_restuarant_datetime, Driver_ID, Total_time elapsed which are less needed for our data modelling. After pre-processing, we have trained the set and added the features and target in order to train the dataset and then evaluate the prediction for the delivery scams from the generated decision tree regression.

For the performance metrics, we have used the mean square error and r square error for the regression model which is used to evaluate the performance metrics and here MSE gives the average of the regression data which is the difference between the actual data and the predicted data that we have acquired by training and predicting the data whereas r squared error gives us the variance of the regression model that ranges in between 0 and 1 that resembles the performance metrics of this model.

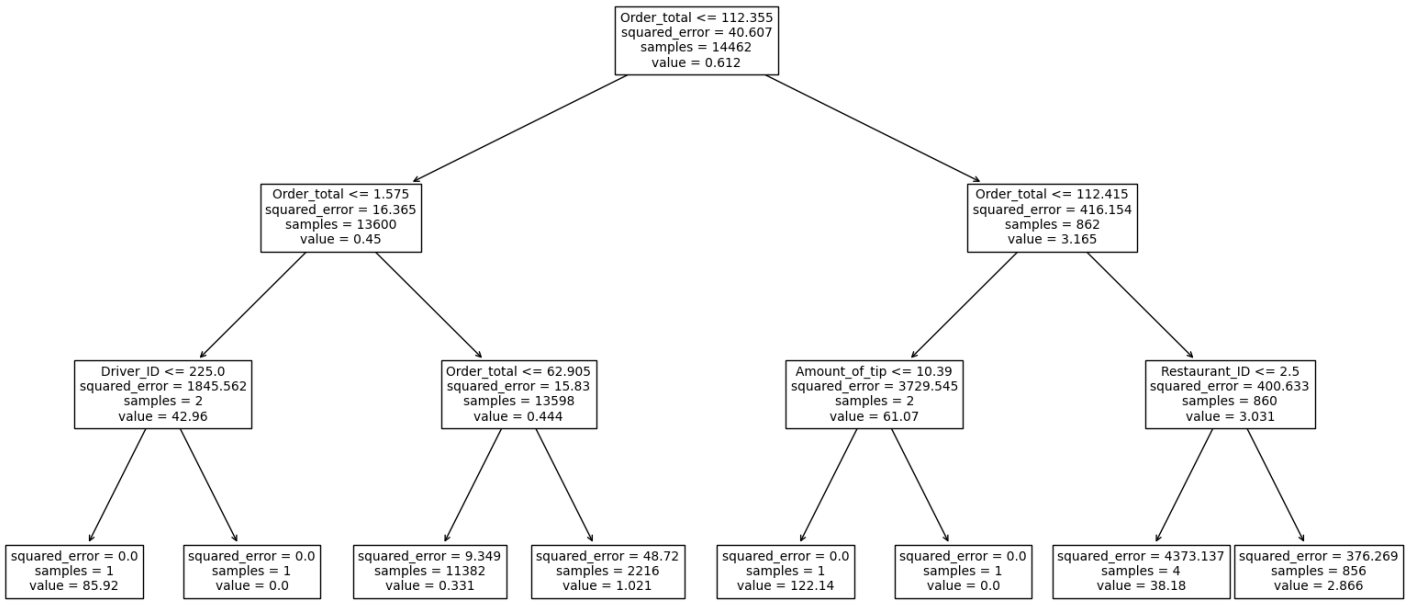
Mean squared error: 22.026627763339206

R-squared: 1.0023211323739158463

Visualisation

Here we have used the decision tree for visualisation in which it contains the subsets as the nodes that are split up into the internal nodes and leaf nodes. This is done after pre-processing, So we have built this regression model by using sklearn where we have set the maximum with to 3 so that the decision tree wont be over fitting for the data we require.

Here for the visualisation, we have used matplotlib that we have imported before the pre-processing to plot the decision tree with the branches.



Conclusion

From the data modelling we have done in this project till know we got know that it is needed to pre-process the data before creating the decision tree for the regression technique because not doing so might cause or lead to overfitting of the decision tree which results the unstable outcome and also we got to know we have to create multiple decision tree instead of creating a single decision tree to get more accurate outcome from different perspectives and we have to just play along with the max features and the target in order to get the best possible performance metrics.

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