Cake Price Prediction

Ibrahim Tarek Mohamed | Inter2Grow | 23/11/2023 |Github

Data Preparation

- After loading the dataset and check it's shape (4000,8) started to check the numerical and objects columns in the dataset to solve the problem of object columns later.
- Checked the dataset statistic properties STD, Mean, Median, min, max, quartiles.
- Checked if there is any massing value to handle it.
- Checked the outliers in numerical columns.

Handel Categorical columns

There is many ways to handle the categorical values I used two of them one-hot encoder and label encoding.

LABEL ENCODING AND ONE HOT

Drop sold_on column as this might cause a problem while training the model as we have 7 values from 0 : 6 the distance between 6 to 0 is very large so the model can be bias.

Convert the object columns into categorical than use cat.codes function to label encode it.

Calculate the correlation between data's features.

Normalize the dataset columns I used two normalize techniques min max scaler and standard scaler cause min max scaler didn't perform well as the distance between numerical feature and label encoding feature is very large so stander scaler solved this problem.

The same steps done using one-hot encoding.

MODEL SLECTION

After many visualizations of the data, I found the relation between the price and the other feature that strongly effect price is a linear relation that can be represented by linear regression model but it will be good if I tried many others models like XGBoost, SVM and others.

After training the different models, as expected linear regression is the best model with then `Random Forest regression.`

`The evaluation matrix was MSE and accuracy score.`

`LR = 96.67  MSE = 13.4`

`RF = 96.3  MSE = 13.9`