Benjy Katz

Week 11

Nov 27, 2023

Colab: https://colab.research.google.com/drive/1OcZuTlFjYlp11ByWdxA7QPuhz8J3hHWC?usp=sharing

After separating out future data and old data and removing customer and sales data, the rmse sharply diminished. Obviously, this is more in line with real world conditions since you can not know how many customers will walk in on any given day. I based my model off previous sales and added a column which includes last year's sales and customer data. In order to encode the store id data, I used a target encoder. This translates the storeid to how it relates to general sales data. This might have been redundant but it was a nice encoded number to maintain the identity of the store. Most other data was one hot encoded like day of week, year, and store type. Some data was left out due to redundancies with the two tables.

When the zeros are taken out of consideration my model predicts as follows

MSE: 1580265.08

RMSE: 1257.09

RMSPE: 1.3041149443807064e+17

I am not sure where the RMSPE comes from and it seems too far off.

When the zeros are left in

MSE: 1226190.89

RMSE: 1107.34

RMSPE: 9.234931598858386e+17

I am not sure why RMSPE is so out of whack, but it is likely that I am not implementing the function correctly.

Leaving the zeros in improves RMSE because it is very easy to predict sales when the sales will be zero. If the store is closed, there will be $0 in sales. It does not take an advanced model to figure that out. This pads the stats for all other predictions and improves the RMSE.