Ilari Pajula

Time Series Forecasting Model for Orders

Data Set: Wolt Takeaway Orders 2020 (orders_autumn_2020.csv)

FEATURE ENGINEERING

The model will use calendar days as data points, there are 10 features in data after feature engineering. Each order is grouped according to its timestamp day, and the average values of a calendar day's orders are used as one data point's features. The label of the dataset is the frequency of orders per day.

<u>FEATURES</u>**:**

'ACTUAL_DELIVERY_MINUTES - ESTIMATED_DELIVERY_MINUTES', 'ITEM_COUNT', 'ESTIMATED_DELIVERY_MINUTES', 'ACTUAL_DELIVERY_MINUTES', 'CLOUD_COVERAGE', 'TEMPERATURE', 'WIND_SPEED', 'PRECIPITATION', 'Year sin', 'Year cos'

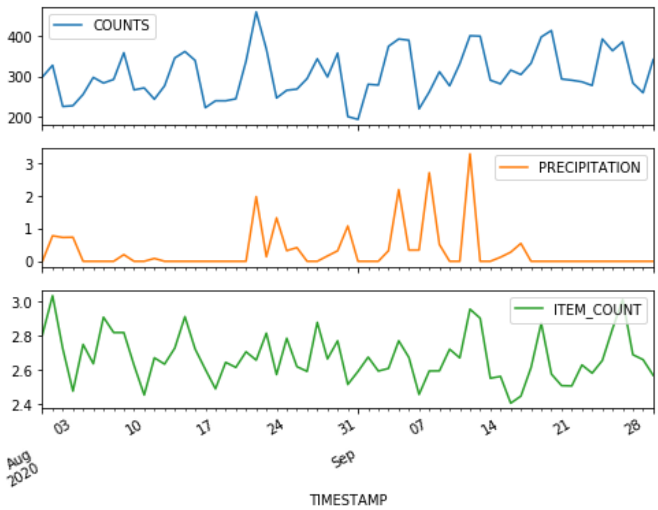
<u>LABEL</u>**:**

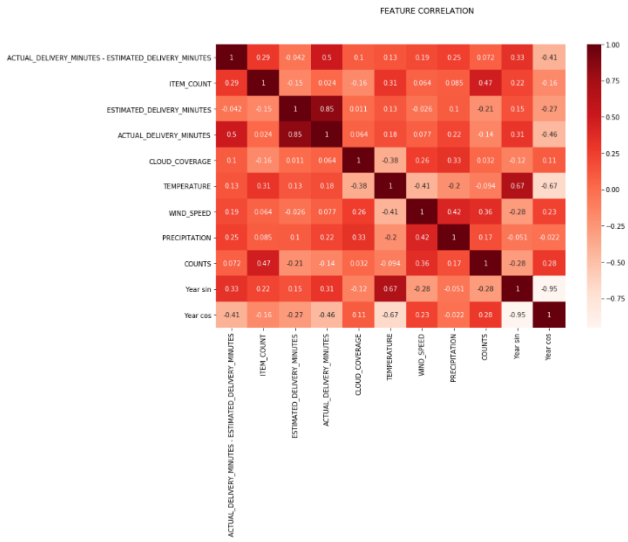
' COUNTS'

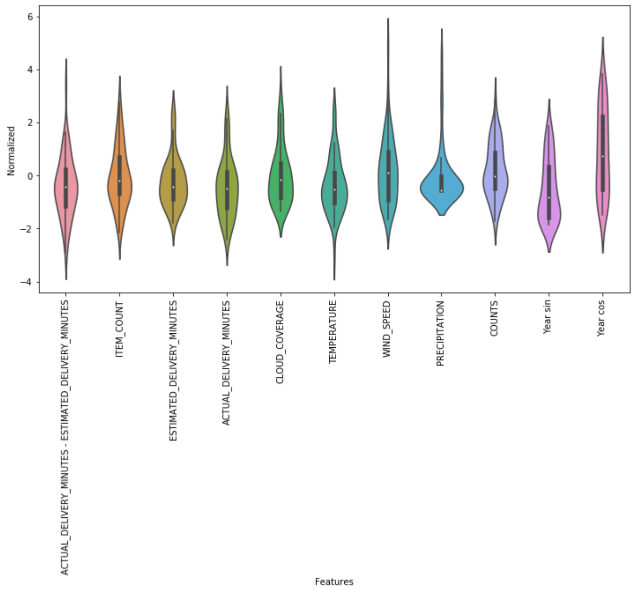
Timestamps are manipulated into sine and cosine signal data during feature engineering. This model will not incorporate location data so I have removed features that correspond to venue and user location.

<u>DATA EXPLORATION:</u>

Visualizing some of the features after feature engineering.



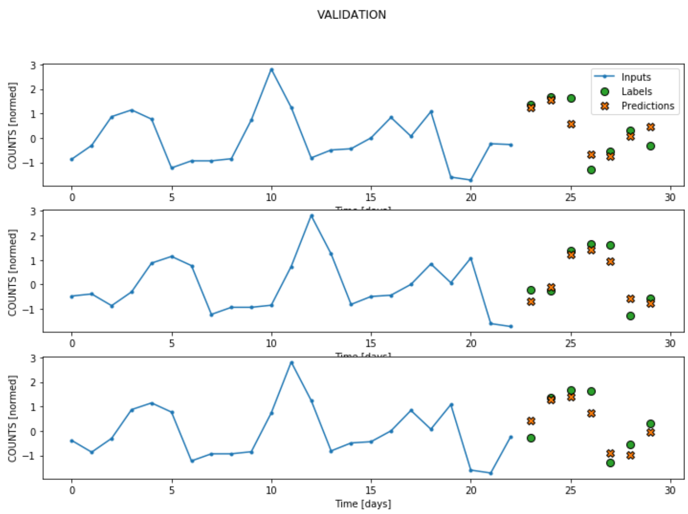


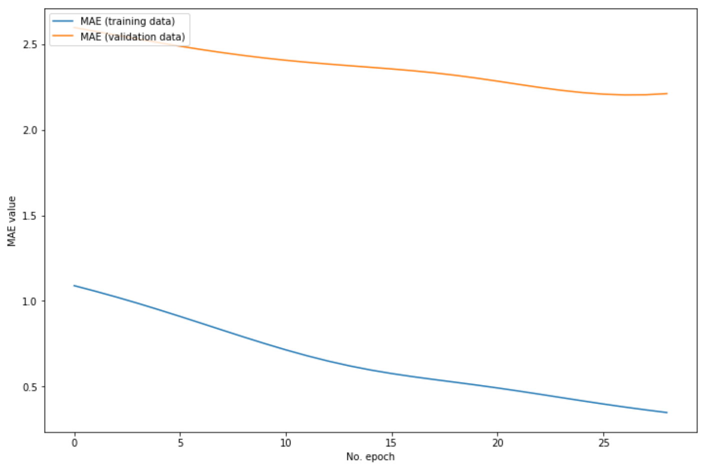


The model first builds a window using the WindowGenerator class, which will pack the training, validation, and testing data into a comfortable format for Tensorflow to read.

The model uses the first 70% of the data for training and the latest 50% for validation. The training and validation data have 20% overlap, because the daily training data is limited after the orders are grouped. However, the predictions made by the validation data will not have any overlap with training data.

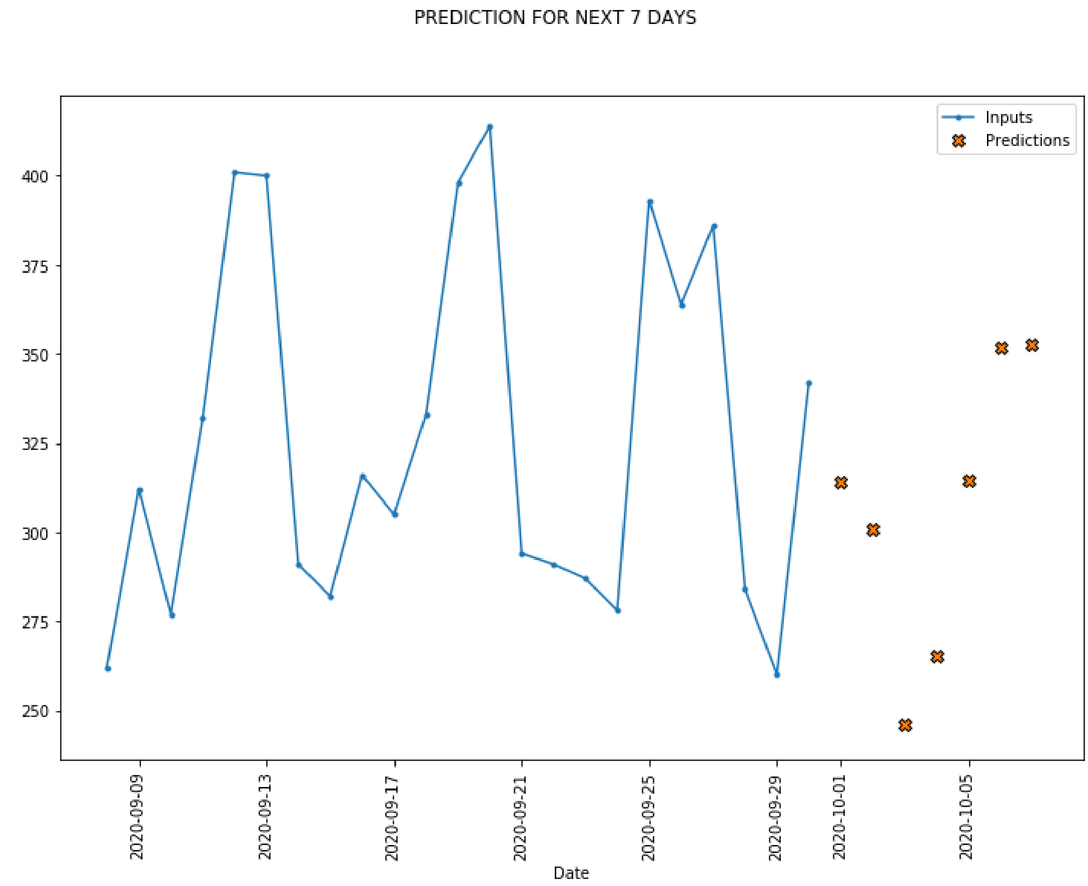
The model uses 23 days as input data, and then tries to predict the next 7 days. After training the model on a convolutional neural network (CNN) the validation results are as shows. The training and validation loss per training epoch are also shown below. The metric used to evaluate the model is mean average error.





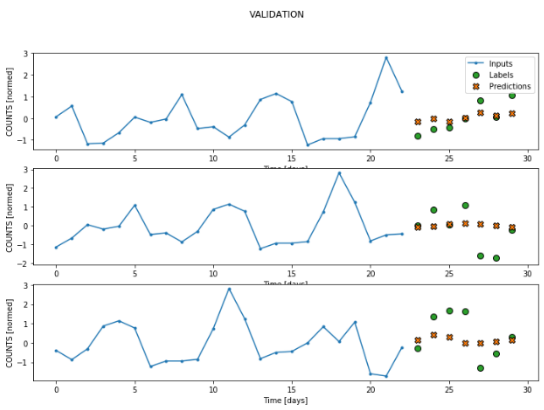
The model can now be used to predict the latest 7 days, which the data does not have validation data for.

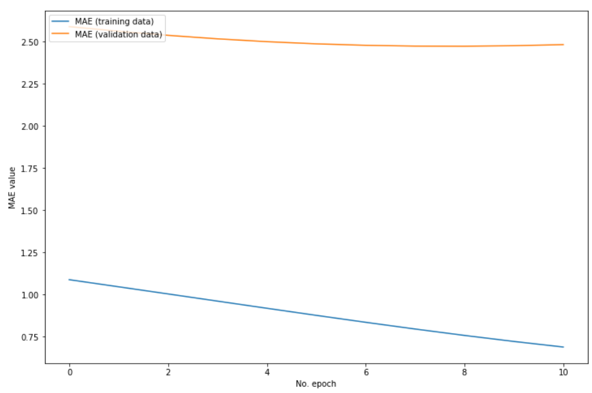
CNN MODEL PREDICITION



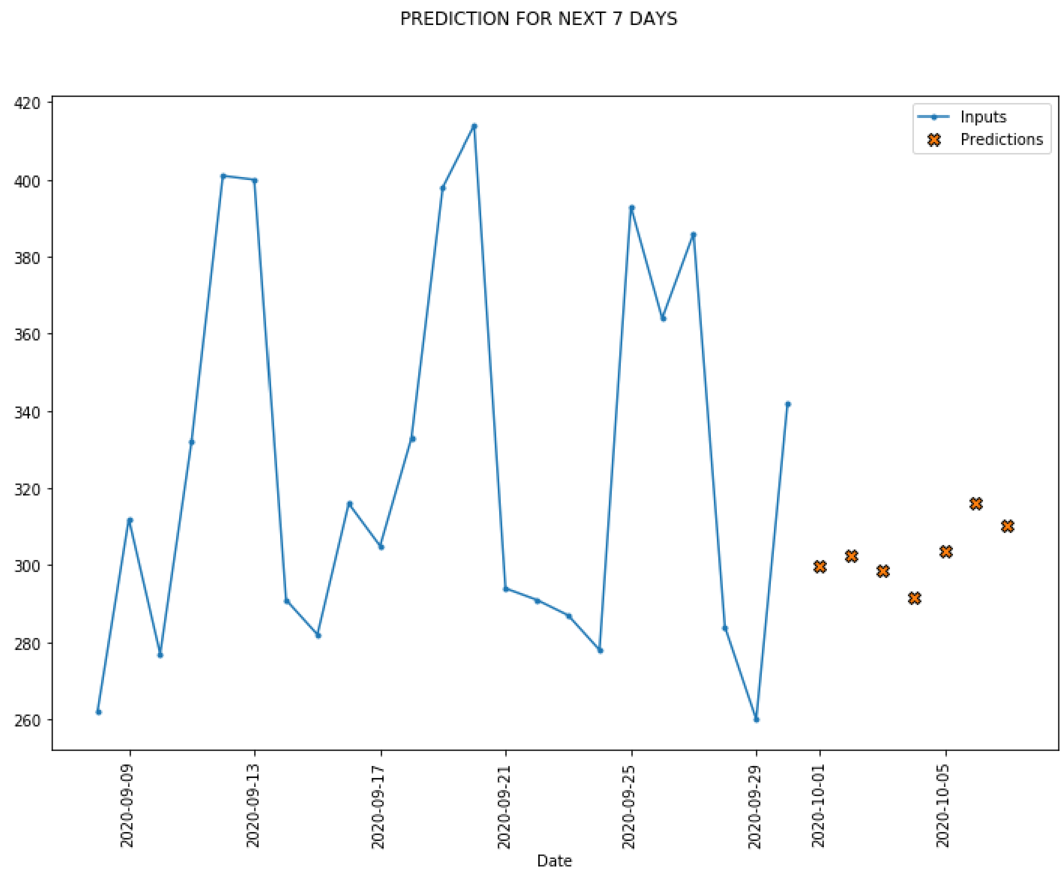
Further Development / Alternative Model

An alternative dense model could be used to make forecasts for this data as well. A dense model is a bit more powerful than a linear model but will likely produce less accurate forecasts than the CNN model, which is well suited for this type of time-series data. Below are the results from the dense model training and forecast.





DENSE MODEL PREDICTION



CONCLUSION

The CNN model gets better training results than the Dense model, and its evolution can be observed when visualized with the training epochs. However, both models are likely to suffer from overfitting because there is some overlap between the training and validation data.

While I am relatively happy with how the models turned out, I would have some improvement points if I were to continue with this project. The most important improvement would be to gain access to more data, so there would be more reliable training and validation.

An improvement on this model would include a more refined feature engineering stage. I was not sure how to implement better use of order delivery times, location data, and weather data. The model as it is now will probably not benefit from these features that much, as indicated in the correlation heatmap of these features where many values are close to zero.

CNN models are black boxes, which means that we do not know what weights the features are assigned. Getting better feature engineering with the inclusion of the location data would allow the DenseNet model, which can be supervised, to become more reliable and allow forecasts to become more versatile. These models could potentially rely on weather forecasts, and location trends to make a better forecasts of order frequency per day.