MAFS 6010Z Project 3: M5 Forecasting - Accuracy

XIA Yiqiao yxiaat@ust.hk, Department of Mathematics, HKUST Video: https://youtu.be/RbGCxgbfmoM

1. Introduction

This competition is to estimate the point forecasts of the unit sales of various products sold in the USA by Walmart. I established two models, Light GBM and LSTM. The later presented much better prediction power and the model scored **1.15317** in Kaggle with team name **mafs6010Z_XIA**.

2. Data Pre-processing

2.1 Data Aggregation

- Sales data is converted from long to wide format for further merge.
- Categorical data, including calendar data and some price data, are transformed with Label Encoder to save memory.
- Also, there is a downcast process to save memory.

2.2 Fill NA

- Backward filled to maintain time series consistency.

2.3 Normalization

- Min-max normalization

3. Feature Engineering (for light GBM only)

LSTM is capable of generating features, so feature engineering process is for light GBM only.

Time series

- Lag for 1, 2, 3, 6, 12, 24, 36 data points.

Data mean

- Mean for sales data by item, state, store, category, department and cross terms were calculated separately.

Rolling Mean

- Sales data with rolling window 7 days.

Expanding Mean

- Expanding every 2 data points.

Trends

- Defined as (mean by date – mean by item, state, store, category and department.

4. Model Construction

Light GBM

- Models were trained by stores.
- Configurations

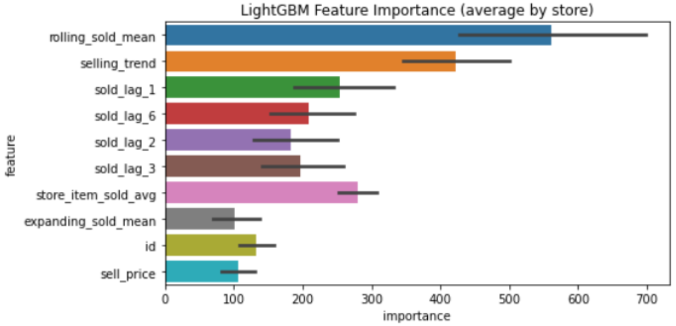
n_estimators = 1000,
learning_rate = 0.3,
Subsample = 0.8,
colsample_bytree = 0.8,
max_depth = 8,
num_leaves = 50,
min_child_weight = 300
Loss: mean squared error

LSTM Timestep = 28

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 100)               44000
_________________________________________________________________
dropout (Dropout)            (None, 100)               0
_________________________________________________________________
repeat_vector (RepeatVector) (None, 28, 100)           0
_________________________________________________________________
lstm_1 (LSTM)                (None, 28, 100)           80400
_________________________________________________________________
dropout_1 (Dropout)          (None, 28, 100)           0
_________________________________________________________________
time_distributed (TimeDistri (None, 28, 1)             101
=================================================================
Total params: 124,501
Trainable params: 124,501
Non-trainable params: 0
_________________________________________________________________
```

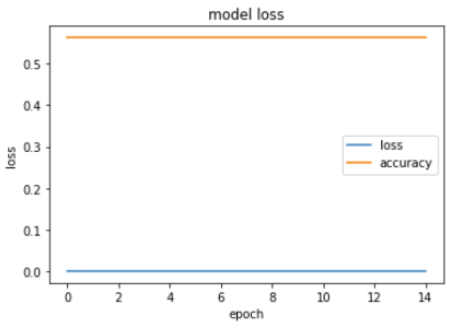
LSTM model constructon

5. Feature Importance – Light GBM Model



Top 10 feature importance

6. Model Performance – Accuracy (LSTM)



Accuracy and Loss for LSTM Model

7. Analysis and Conclusion

Light GBM did not perform well in the test data. It can be observed that sales data lag comprises 4 out of 10 top importance features, and trend data ranked 2nd. It's a pity that **calendar data and price data failed to contribute much** to the model. Further analysis **may include more time series techniques** like GARCH and ARIMA.

Loss for LSTM model is constant among epochs, which may due to the fact that a small list of data is used restricted by computation power. The model **assumed that it takes a while for event to have effect** on sales data, so only 14-54 days ago were used. It has been tested that 54-100 days data have marginal effect on accuracy, but smaller time range may be paid more importance in further analysis

8. References

[1] SHARMA, A. N. S. H. U. L. (2019). Time Series Forecasting-EDA, FE & Modelling. Kaggle. https://www.kaggle.com/anshuls235/time-series-forecasting-eda-fe-modelling
[2] PATEL, Y. A. S. H. V. I. (2021). Time Series Forecasting Using LSTM - M5. Kaggle. https://www.kaggle.com/yashvi/time-series-forecasting-using-lstm-m5/notebook