MSBD 5013 Project 2: M5 Uncertainty

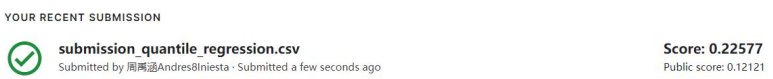
Hao Wu, Yuhan Zhou and Lingjun Guo {hwucc, yzhoudv, lguoal}@ust.hk
Department of Computer Science and Engineering, HKUST

1. Introduction

In the M5 Uncertainty project, the main problem is to estimate the uncertainty distribution of Walmart unit sales. We need to predict 9 quartile of item sales in different stores across various locations for two 28-day periods of uncertainty forecasting with the corresponding median and 50%, 67%, 95%, and 99% prediction intervals.

We implemented Quantile Regression with Keras, and tried normal distribution method to convert the accuracy results from the previous M5 Accuracy competition to uncertainty. And considering the Algorithmic logic and overall performance, we chose **Quantile Regression with Keras** as the winning model.

Private and Public Scores



2. Dataset

The datasets consist of previous 1941 days (starting from 2011-1-29) sales information of 3049 items in 10 stores of 3 states in US(CA, TX and WI). We also have a calendar dataset that gives information about dates and events.

- **Data Pre-processing**

The given dataset is extremely wide which has 1941 day features. So we reshaped the data, preprocessed the calendar, kept the features and dates we need, then melted it to a long data frame. Besides, we applied memory reduction which reduced 78.4% of the memory usage for all the given data sets

.

- **Exploratory Data Analysis**

We did EDA to find out the logic between each feature and sales.

3. Quantile Regression with Keras

Quantile regression allows us to estimate percentiles of the underlying conditional data distribution even in cases where they are asymmetric, giving us insight on the relationship of the variability between predictors and responses.

In this project, with Keras API, we use the pinball loss function to build our neural network and then make our prediction.

4. Evaluation

According to the rules of the Kaggle competition, they apply **Scaled Pinball Loss (SPL)** function for each time series and for each quantile to evaluate precision of the probabilistic, using following formula:

$$\mathbf{SPL(u)} = \frac{\frac{1}{h}\sum_{t=n+1}^{n+h}(Y_t - Q_t(u))u\mathbf{1}\{Q_t(u) \le Y_t\} + (Q_t(u) - Y_t)(1-u)\mathbf{1}\{Q_t(u) > Y_t\}}{\frac{1}{n-1}\sum_{t=2}^{n}|Y_t - Y_{t-1}|},$$

where Y_t is the actual future value of the tested time series at point t, Q_t (u) is the generated forecast for quantile u, h is the forecasting horizon, n is the length of the training data, and 1 is the indicator function that indicates whether Y is within the postulated interval. To be more specific, we will be asked to provide the **median**, and the **50%, 67%, 95%, and 99% Prediction Intervals(PI)**, where u is set to u_1=0.005, u_2=0.025, u_3=0.165, u_4=0.25, u_5=0.5, u_6=0.75, u_7=0.835, u_8=0.975, and u_9=0.995.

After calculating the SPL for all the 42,840 time series and all the requested quantiles, the final scores in the competition are ranked using the **Weighted SPL** (**WSPL**):

$$\boldsymbol{WSPL} = \sum_{i=1}^{42,840} w_i * \frac{1}{9}\sum_{j=1}^{9} SPL(u_j),$$

where w_i is the weight of the i_th series, and u_j represents the j-th quantile. The smaller WSPL score indicates the estimation is better.

5. Compare to other methods

	Quantile Regression	LGBM	LSTM	SES
Public Score	0.12121	0.15795	0.28558	0.31652
Private Score	0.22577	0.22739	0.32756	0.30010

In addition to the Quantile Regression Model, we also try some other models including LGBM, LSTM and SES. As can be seen in the table, Quantile Regression Model has a better performance on WPLS.

6. Conclusion

This project is another research aspect of Walmart sales prediction. Since we have completed the project of prediction accuracy, the submission file of the uncertainty project could also be transformed form the results of accuracy project if we consider the distribution of quantiles as the cumulative Normal Distribution. The outcomes using this method also perform well but not as good as the Quantile Regression Model made by Keras.

This project give us another standard to determine weather a prediction is good. For the future work, we will concentrate on optimizing the WPLS by applying CNN and RNN to improve the stability of prediction.

7. References

https://www.kaggle.com/code/ulrich07/quantile-regression-with-keras
https://www.kaggle.com/code/szmnkrisz97/point-to-uncertainty-different-ranges-per-level
https://github.com/ramuburla7250/M5---Forecasting-Uncertainty/blob/master/EDA.ipynb

8. Contribution

Model and Parameter Selection	- Hao Wu, Lingjun Guo
Coding and Evaluation	- Lingjun Guo, Yuhan Zhou
Poster writing	- Yuhan Zhou, Hao Wu