MATH5470

Spring 2023-24

Kaggle - M5 Forecasting Accuracy

#	Topic	Time allocated (minutes)	Slide number
1	Project introduction	1	2
2	Data analysis using exploratory data analysis (EDA)	3	3-4
3	Preprocess of the data using feature engineering	2	5
4	Model training results with scores	2	6-10
5	Conclusion	1	11
	Total	9	

Group members :

GUPTA, Anchal

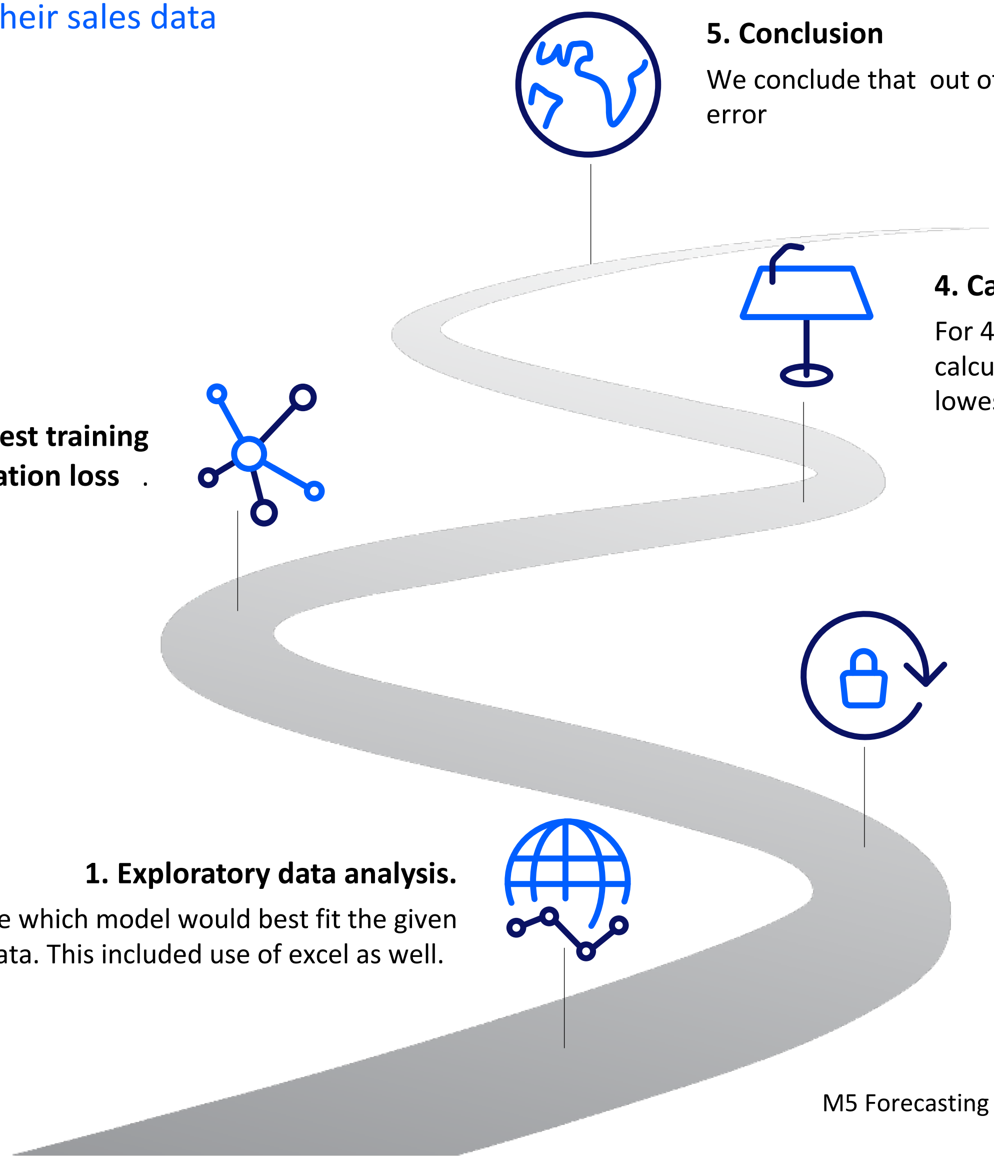
GAMAGE NANAYAKKARA, Dinusara Sasindu

SEO, Minji

ZHAO, Hang

Project Introduction and Workflow

The project entails estimation of the unit sales of Walmart retail goods for the next 28 days based on their sales data from January 2011 to June 2016



5. Conclusion

We conclude that out of all models selected LGBM model selected has lowest error

4. Calculate Private and public score

For 4 models we used as sample, programs were run to calculate private and public score. This gave us model with lowest error what is predicted

3. Ran various models to see which model has lowest training and validation loss .

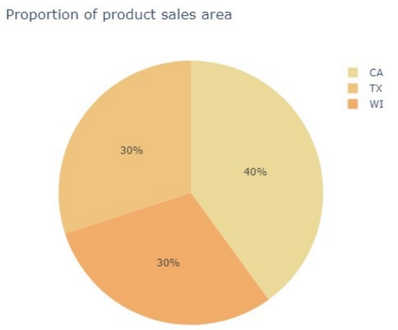
2. Feature processing

Perform feature processing to select what works better between time series model or tree model feature processing

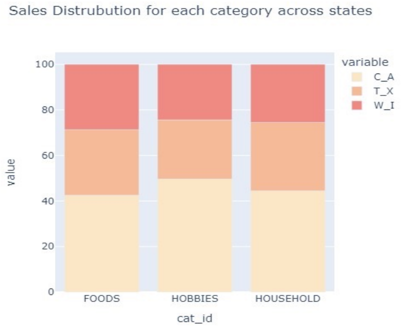
1. Exploratory data analysis.

We utlised various methods to analyse which model would best fit the given data. This included use of excel as well.

1. EXPLORATORY DATA ANALYSIS 1/2



Proportion of product sales area

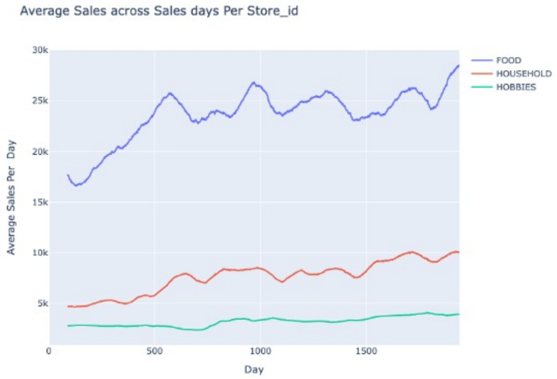


Sales Distrubution for each category across states

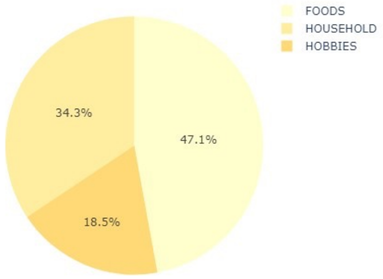
Calculate sales percentage across the store and product categories

Conclusion - Similar distribution of sales per stores

ANALYSIS 1



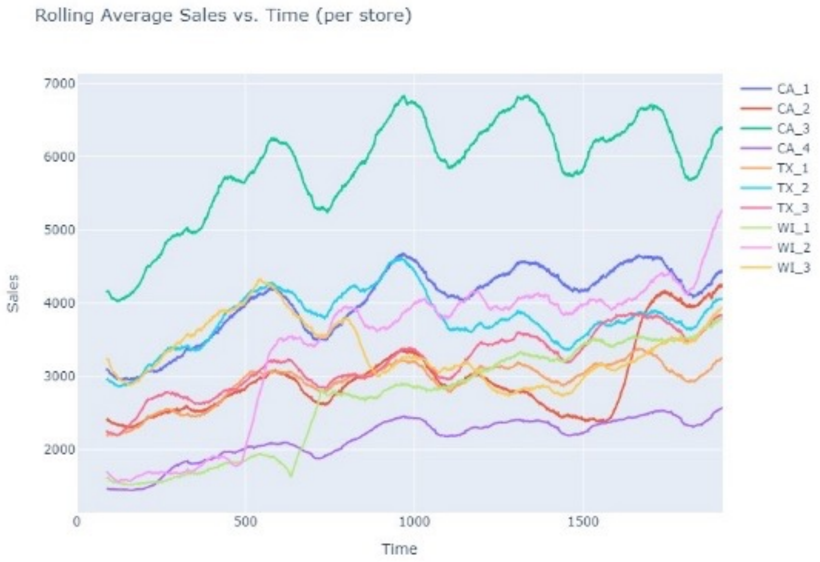
Average Sales across Sales days Per Store_id



Calculate sales percentage across product categories

Conclusion - “FOODS” category is expected to be dominant factor in the model performance

ANALYSIS 2



Rolling Average Sales vs. Time (per store)

Obtain the sales trend per store over 2000 days

Conclusion –

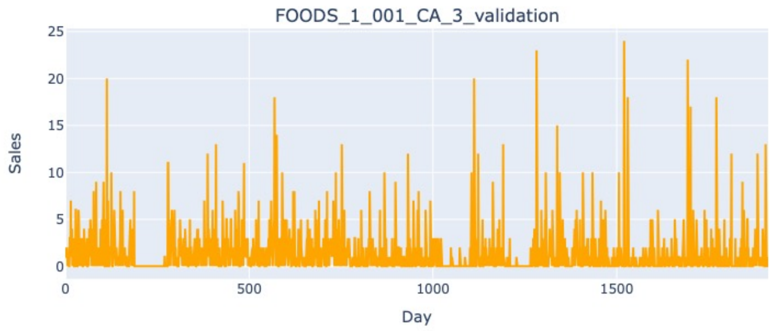
a) Each store has different sales trend. Store code CA_3 has fluctuating trend vs TX_2 has been steady

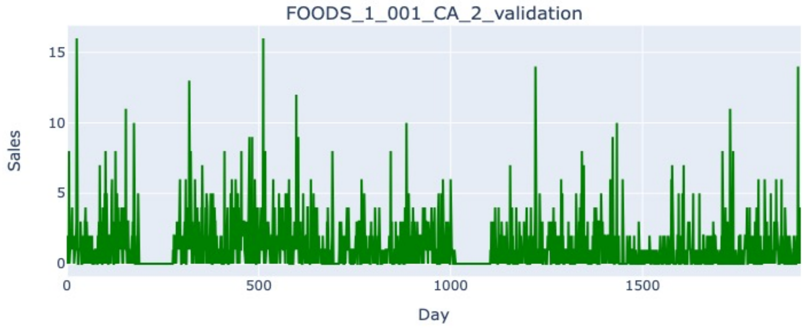
b) Different model per store may yield higher accuracy if a greater volume of data can be obtained.

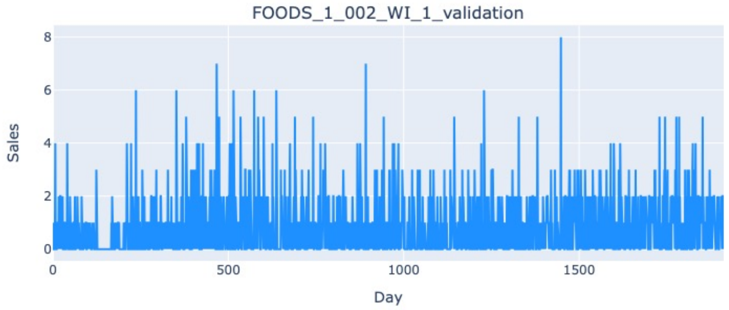
ANALYSIS 3

1. EXPLORATORY DATA ANALYSIS 2/2

Certain days have higher sales than others. Our analysis shows this can be attributed to one or two events on same day







2. Feature Processing

Integrate time information to the event information

Data intake

Date and time when each sales were made.

Any special events on the day (e.g., Superbowl, Valentines' day, etc.)

What was analysed

Implement recurrent neural network (RNN) and long short-term memory (LSTM) architectures.

Check if there is an enhance in the predictive performance of the model on providing more information on the sales.

Conclusion

Such processing improved the performance of RNN and LSTM.

Create new features from temporal intervals

Data intake

Percentage difference in sales price between consecutive weeks

Historical sales trend over 7-day and 28-day intervals

What was analysed

Tailor the features for light gradient boosting machine (LGBM) and extreme gradient boosting (XGB)

Check if there is an enhancement of the predictive performance of the model on providing more information on the sales.

Conclusion

Such processing improved the performance of LGBM and XGB.

3. MODELs TRAINING

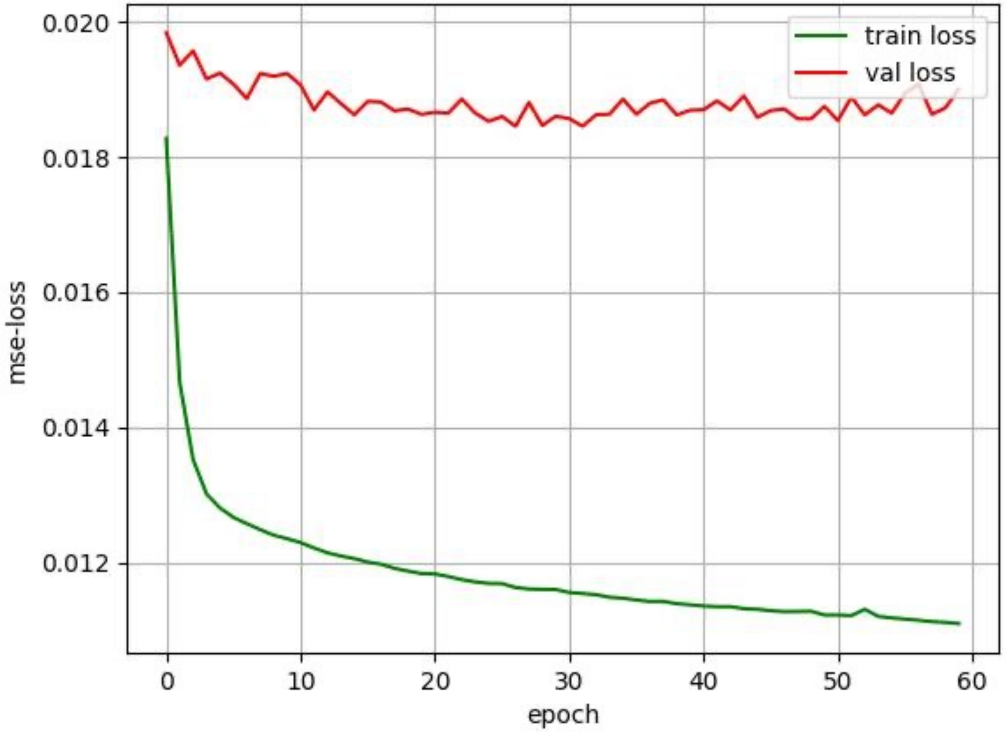
Long Short-Term Memory ("LSTM")

Process

1 **90% of the data was used for training and the rest 10% was for validation.**

2 **Mean square error (MSE) was calculated as a loss function.**

3 **A point with minimum train loss and minimum validation loss was selected.**



3. MODELs TRAINING 2/4

Recurrent neural network model (“RNN”)

Process

1 **The model included three recurrent layers to capture long-term dependencies.**

2 **Dropout layers were added in between RNN layers to prevent overfitting.**

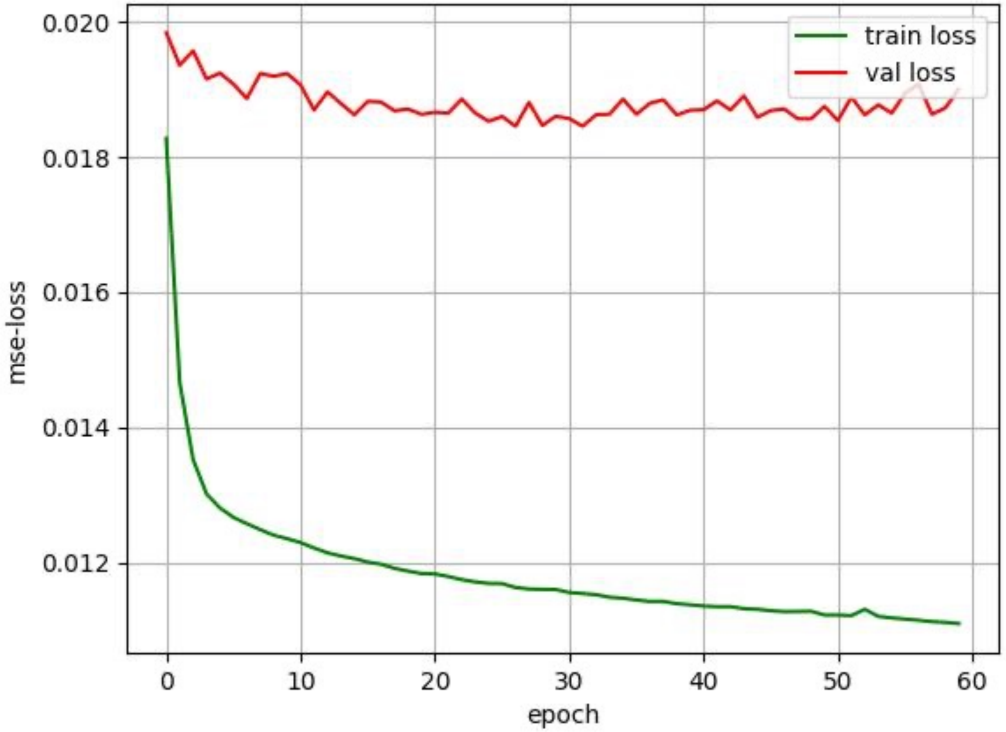
3 **Output of the RNN layers was passed to dense layer to make predictions.**

4 **Training parameters:**

a) Optimizer: Adam

b) Number of epochs: 60

c) Loss function: mean squared error



3. MODELs TRAINING 3/4

Light GBM model ("LGBM")

Process

1 **Training parameters:**

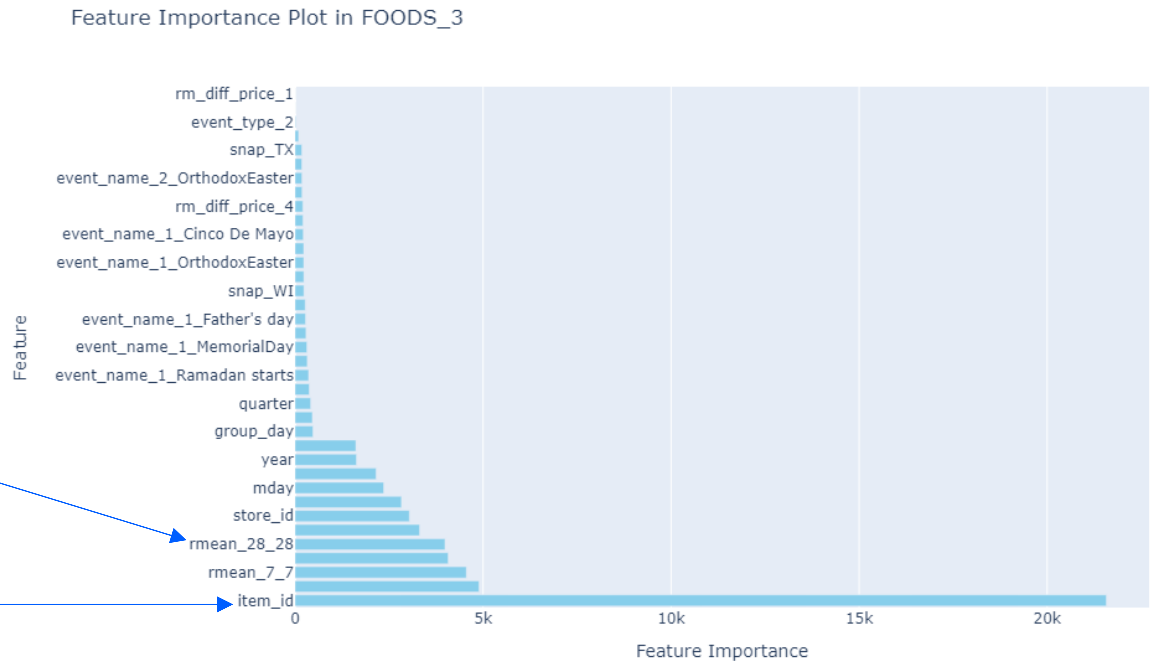
a) Learning rate: 0.09

b) Number of epochs: 2000

c) Loss function: negative log likelihood

2 **"Rolling mean of lagged values" Temporal information that was added during pre-processing**

3 **Product type being the most dominant factor**



3. MODELs TRAINING

4/4

Extreme Gradient Boosting (“XGBoost”)

Process

1 **Training parameters:**

a) Learning rate: 0.09

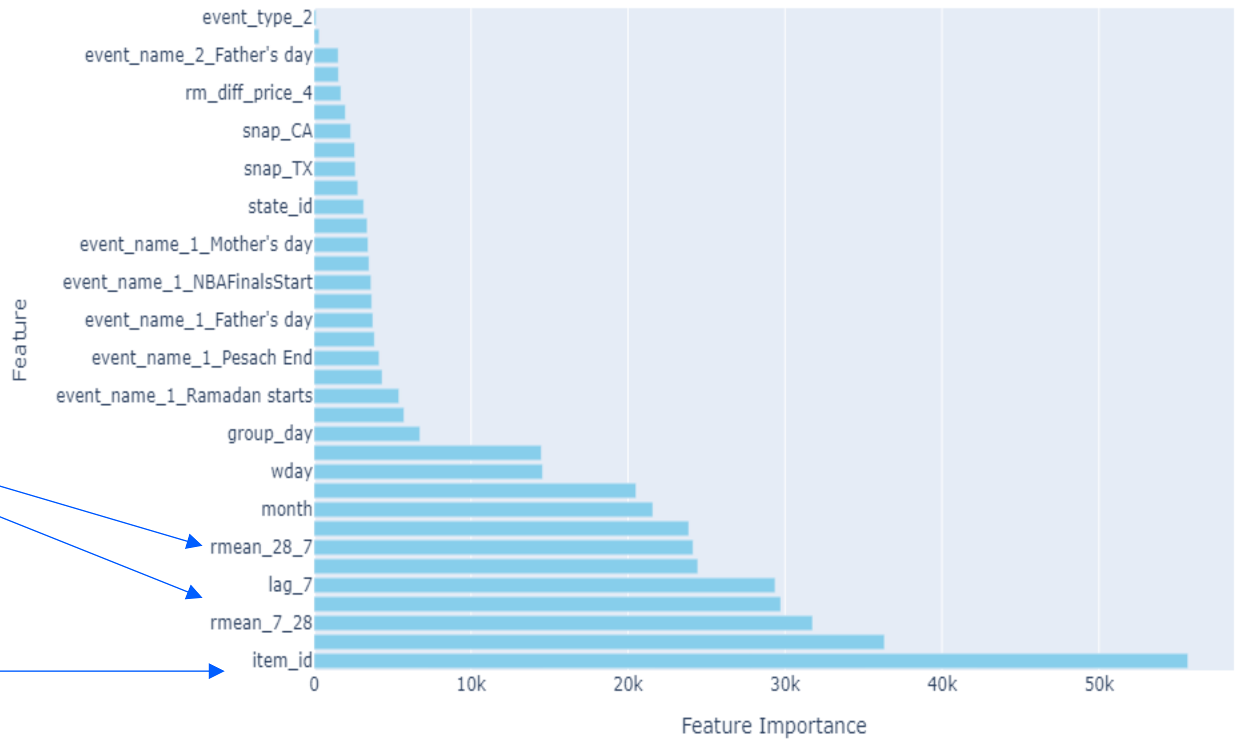
b) Number of epochs: 2000

c) Loss function: negative log likelihood

2 **Similar to LGBM, temporal information was ranked high in importance**

3 **Product type being the most dominant factor**

Feature Importance Plot in FOODS_3



4. CALCULATE PRIVATE AND PUBLIC SCORE

Each score represents the mean error from the test data per model. The LGBM model showed the lowest Kaggle in both private and public score. This means LGBM estimated the Walmart's sales most accurately out of four models trained.

MODELS	PRIVATE SCORE	PUBLIC SCORE	REMARKS
LGBM	0.58542	0.74072	Most accurate since it has the lowest error
XGBoost	0.58911	0.75612	
RNN	1.21168	1.25751	
LSTM	1.36598	1.44894	Least accurate with the highest error

5. CONCLUSION

1 One must experiment with implementation of various models as well as importance of model selection in optimizing predictive performance.
Based on comparative analysis, we saw that LGBM model outperforms the others with lowest error rate.

2 Deep learning models like LSTM and RNN may face challenges in accurately predicting integer-valued sales due to their existing inherent tendency to produce continuous outputs.

3 FUTURE RECOMMENDATIONS - Leveraging a combination of deep learning and GBM techniques can provide valuable insights into retail sales forecasting.