Grocery Sales Forecasting for Corporación Favorita

Yifei Zhang, Dongyuan Mao, Jing Zhao
Department of Electronic Engineering & Department of Materials Science Engineering

Motivation

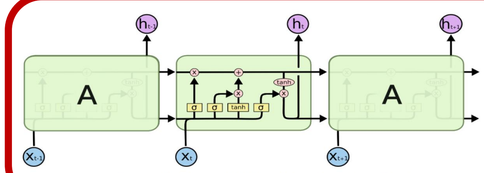
It is always in a delicate dance for grocery stores to decide future purchasing and do sales forecasting. We explored several methods to create a robust algorithm to make precise sales predictions for grocery stores given information about items, stores and history sales record.

Models & Methods

Linear Regression

- Extract features, create training pairs
- Train weights with loss minimization
- Apply weights to predict future sales.

Long Short-Term Memory

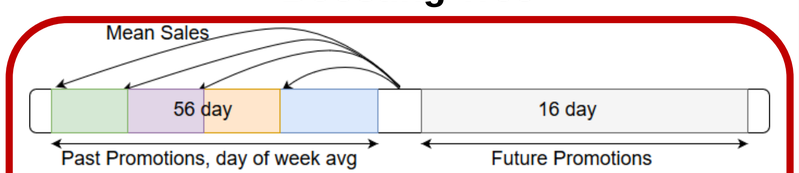


X_t: sales on day t
h_t: prediction of day t+1
Predict consecutive days: Set $X_t=h_{t-1}$

Moving Average

- Periodical weekly sales patterns in data
- Average sales number over long period
- Extract factors affecting daily sales
- Combine them to predict future sales

Boosting Tree



Features: mean of previous sales, mean of previous sales on the same weekday, class of the item to be predicted, sum of previous promotions, promotion on the day to be predicted, sum of future promotions

Results & Discussion

Linear Regression: 0.8031

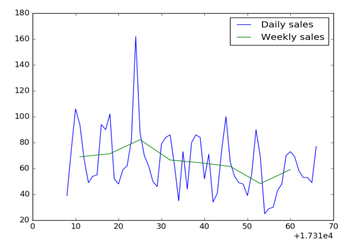
- Predict future sales with 30 day historical data
- Poor performance: Simple & Naive
- Time consuming, impossible to add other features

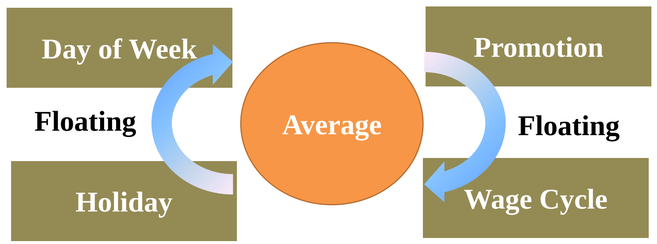
LSTM: unacceptable runtime

- Model from Keras
- Need training for each (store, item) pair
- Input: 7 previous sales
- Time consuming, hard to add other features

Moving Average: 0.535

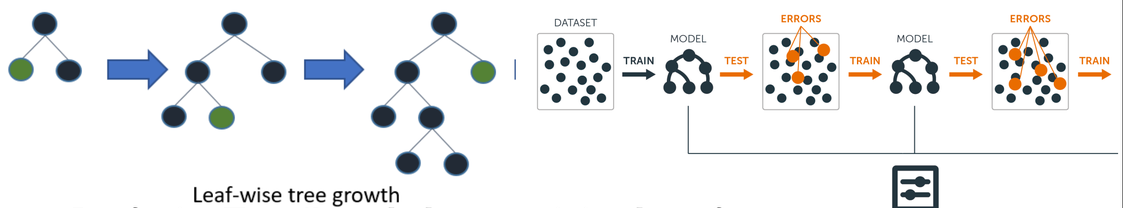
- Average part: Average sales number over a relatively long period
- Floating part: Factors that have impact on item's sales





$$\hat{y} = avg_{median} \cdot avg_{dow}/avg_{week} \cdot promo$$

Boosting Tree



Leaf-wise tree growth

- Leaf-wise Tree Growth: lower training loss, faster
- Prevent overfitting: max_depth, num_leaves, min_data_in_leaf, early_stopping
- Boosting: feature_fraction, bagging_fraction
- Categorical features: partition into 2 subsets, reorder the categories according to training target

Current score at Kaggle is 0.517, ranking 16/964

Conclusion

- Linear regression is not a promising method because of long run time and poor performance.
- Moving average is better because it fits the periodic sales pattern.
- Boosting tree is the best method. It is very flexible with adding new features. Therefore it has the ability to take more information into consideration and make most of them.

Future Work

Boosting tree is a high-potential method which we can dive deep into in future. Adding more features like oil prices, holiday events, store information may reduce error and make our algorithm more robust.

References & Acknowledgements

1. Data provided by Corporación Favorita: https://www.kaggle.com/c/favoritagrocery-sales-forecasting/data
2. Boozed'd Trees—Beer Sales Forecasting: http://web.stanford.edu/class/cs221/2017/restricted/pfinal/dzylber/final.pdf.
3. Predicting House Sales by Linear Regression: https://www.kaggle.com/rahulin05/predicting-house-sales-by-linear-regression
4. Deep Neural Networks for Sales Forecasting: https://dspace.cuni.cz/handle/20.500.11956/83139

We would like to thank Professor Percy Liang and Professor Stefano Ermon for teaching us the foundational material as well as our TA, Anna Wang, for giving us feedback and tips for improving our model/project.