Sales Forecasting for Coporacion Favorita

SPRINGBOARD—CAPSTONE PROJECT 2

BY GEORGE TANG

DECEMBER, 2019

Importance of Sales Forecasting

- Ensure stores have the right items in stock at the right time
- Create positive shopping experience; ensure customers satisfaction and retention
- Optimize item inventory and shelf space usage
- Reduces perishable goods wastage, labor and transportation costs
- Increase revenue and profit

Coporacion Favorita

- Ecuador based grocery retailer
- Operators hundreds of supermarkets and carries over 200,000 products
- At the time (prior to 2017), company relied on subjective forecasting methods backed by little data analytics
- Hosted Kaggle Competition “Corporation Favorita Grocery Sales Forecasting” to challenge participants to develop innovative, data-driven solutions that improve sales forecasts

Project Tasks

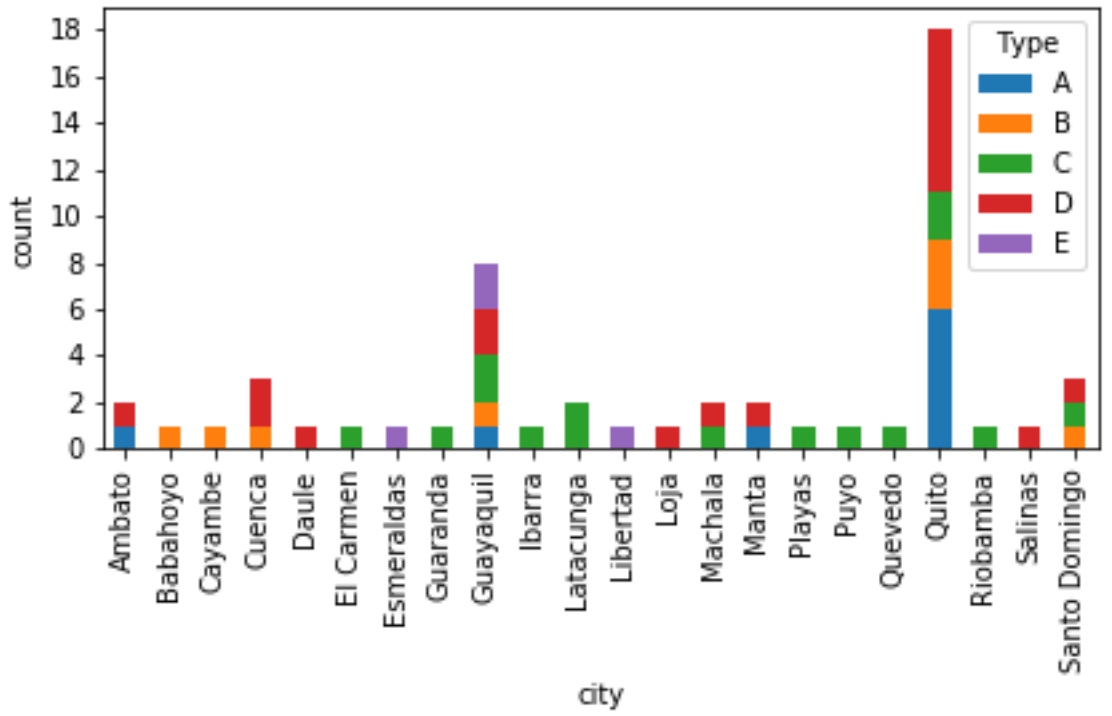
1. Data cleaning, wrangling, quality checking and feature engineering (covered in report)
2. Explore dataset and identify features that affect item sales
3. Develop and evaluate machine learning models for sales forecasting
4. Provide recommendations to management, and suggest future works

Dataset

- Sales data
  - From 2013/1/1 to 2017/8/15
  - Date, store number, item number, promotion status, unit sales
- Store information
  - Store number, city, state, type, cluster
- Item information
  - Item number, family, class, perishable (yes/no)
- Holidays and special events
- Oil price

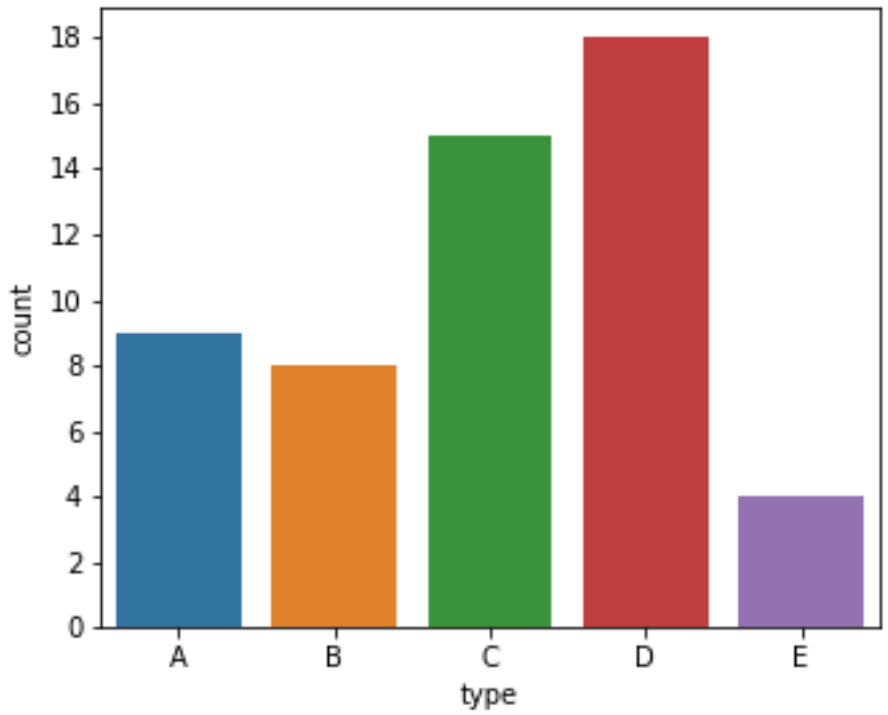
Store Information

- 52 stores (a subset of all stores) in 22 cities
- Quito has the most store (22), followed by Guayaquil (8)



Store Information

- 5 types: Megamaxi, Supermaxi, Gran Akí, Super Akí and Akí but labeling unknown
- Type D store is the most populous, followed by Type C

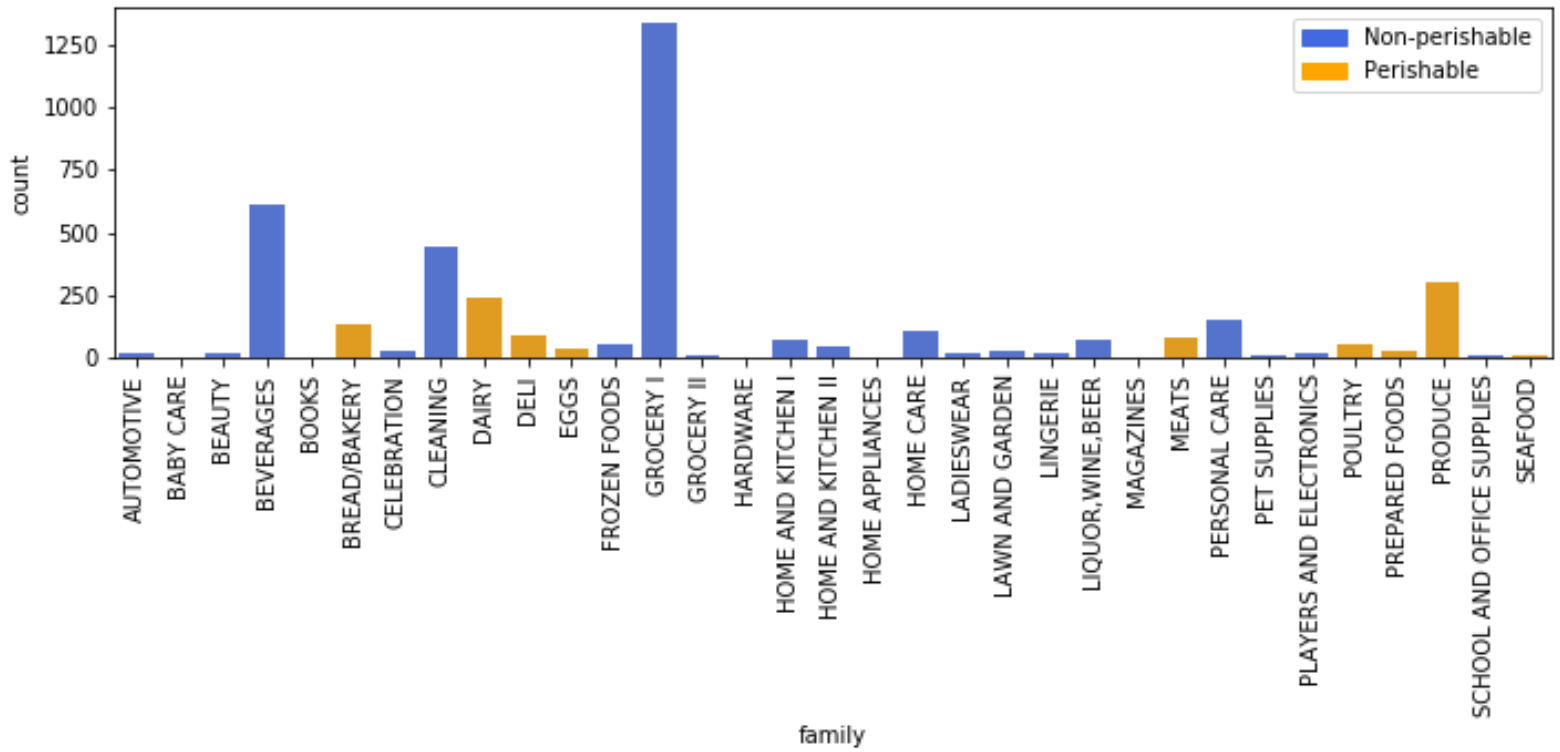


Item Information

- Total no. of items 4,400

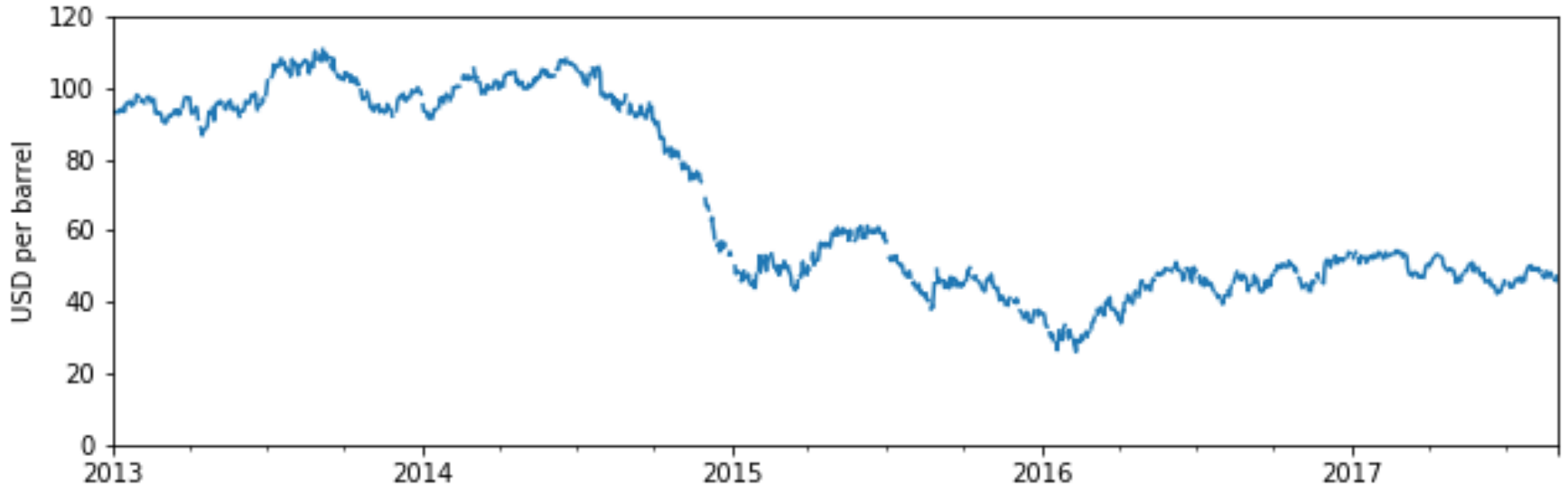
(subset of 200,000+)

- 986 items (24%) perishable



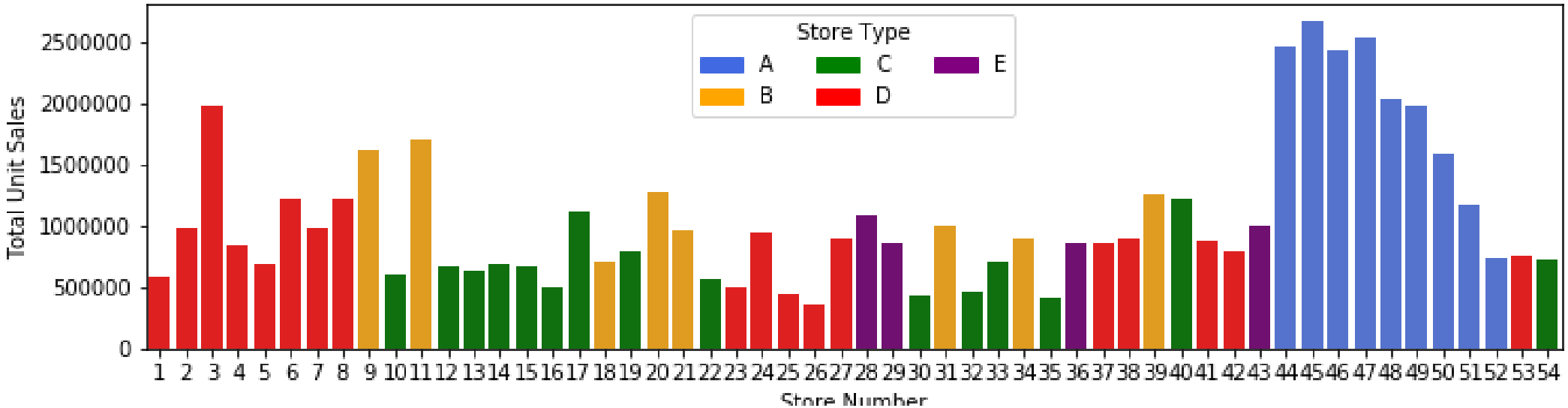
Oil Price

- Ecuador is an oil dependent country; consumer behavior may depend on oil price
- Significant drop in oil price between at the end of 2014



Sales Information – Store Type

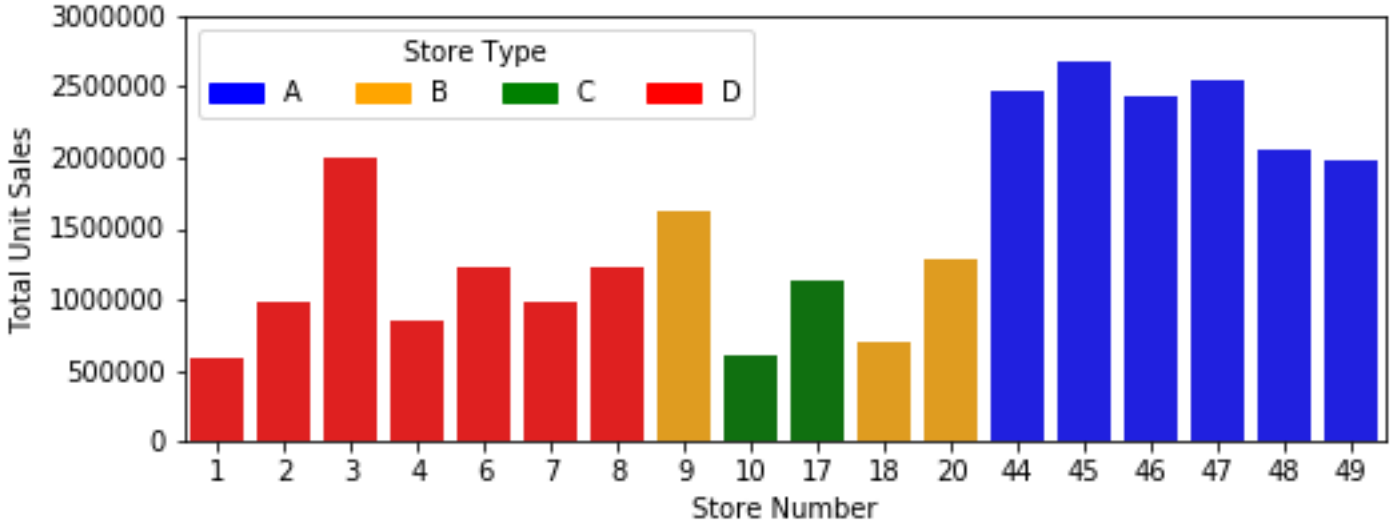
- Sum of GROCERY I units sales (across all items) in 2017
- Sum of unit sales are highest among Type D stores (most located in Quito), followed by Type A



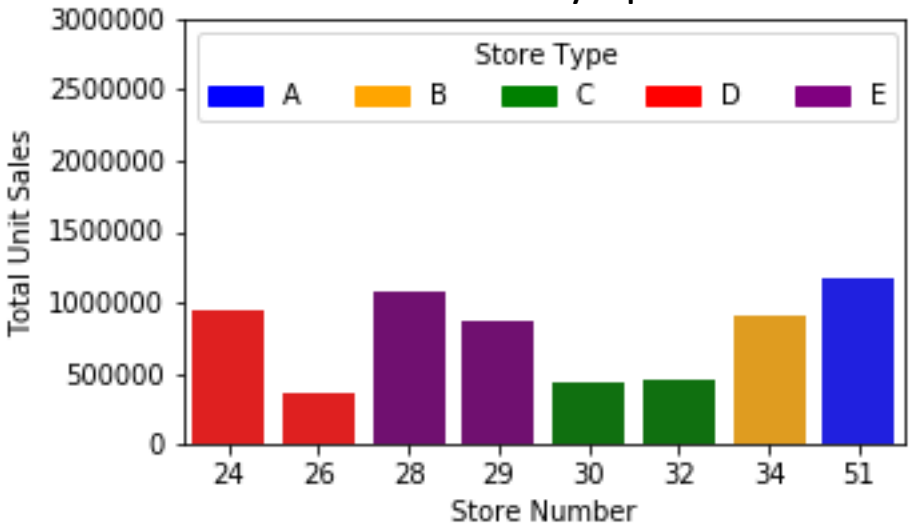
Sales Information – Store Type by City

- Sum of unit sales are typically higher in Quito than in Guayaquil

Quito



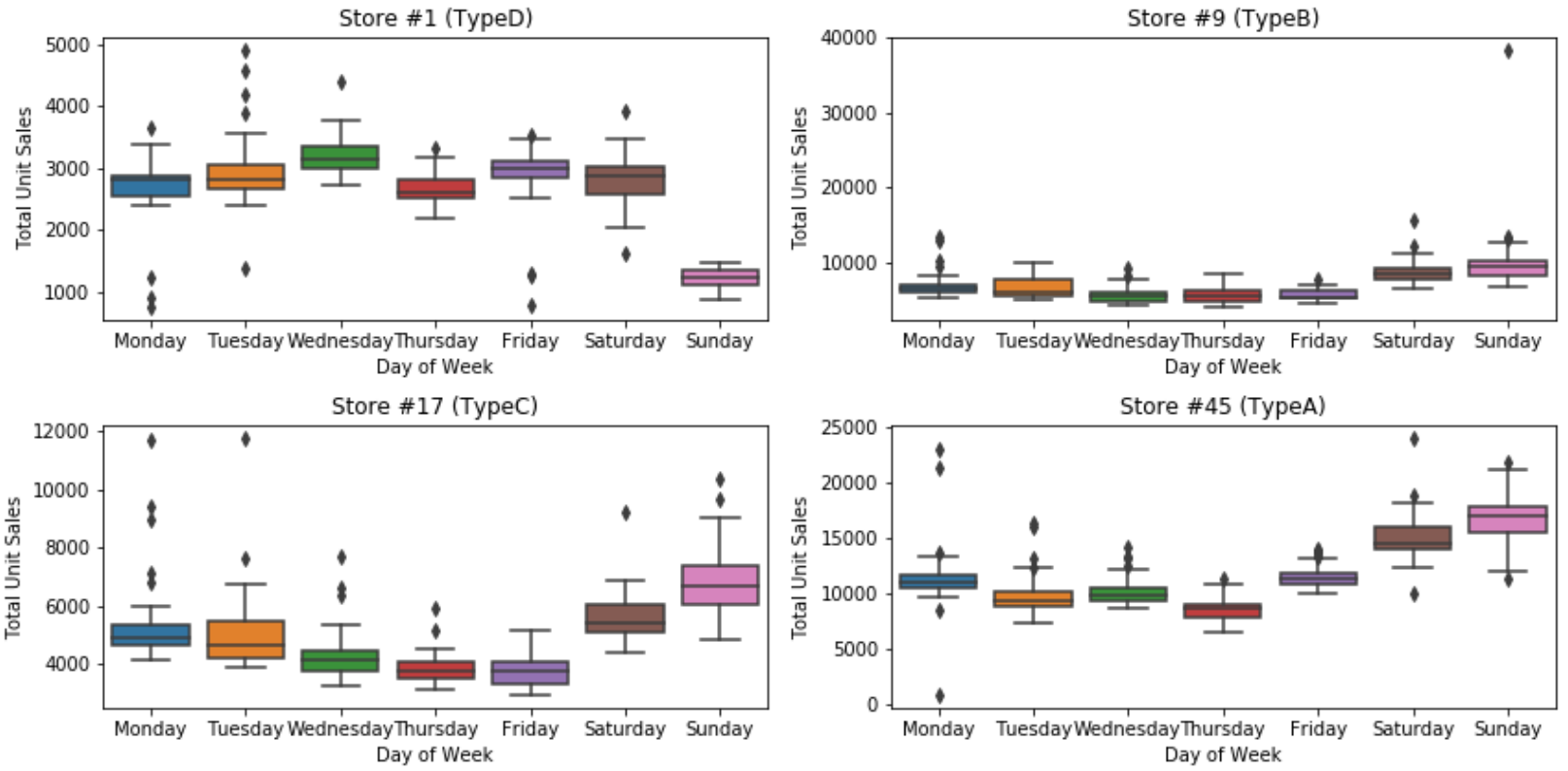
Guayaquil



Effect of Day of the Week on Sales

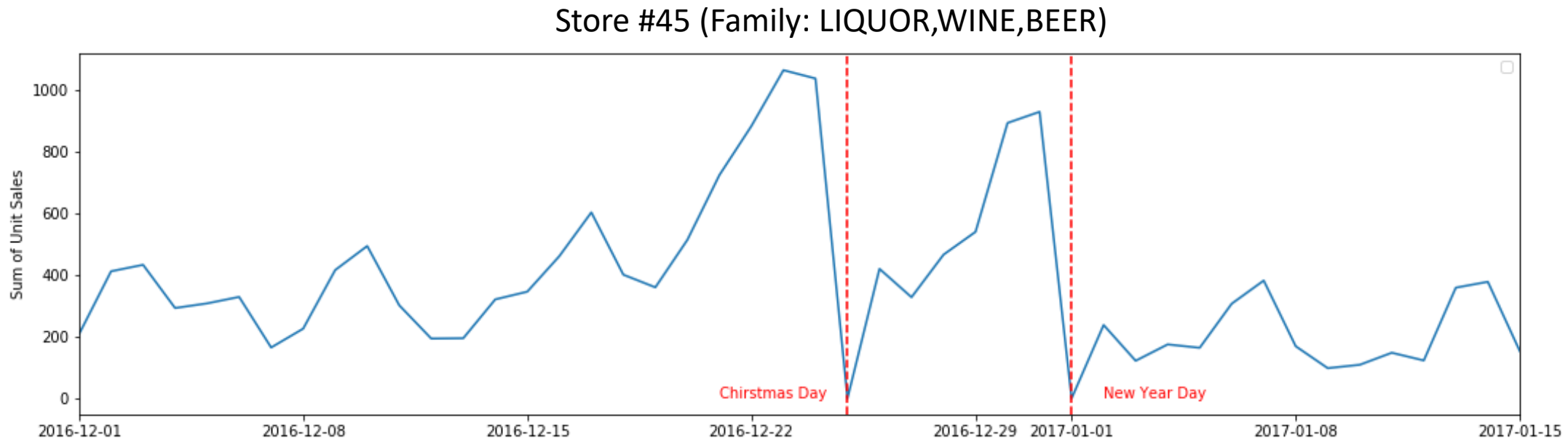
- Most stores have increased sales on weekends, but there are exceptions
- Depends on Store, Item, and Promotion Status

Family: Grocery I



Effect of Holidays on Sales

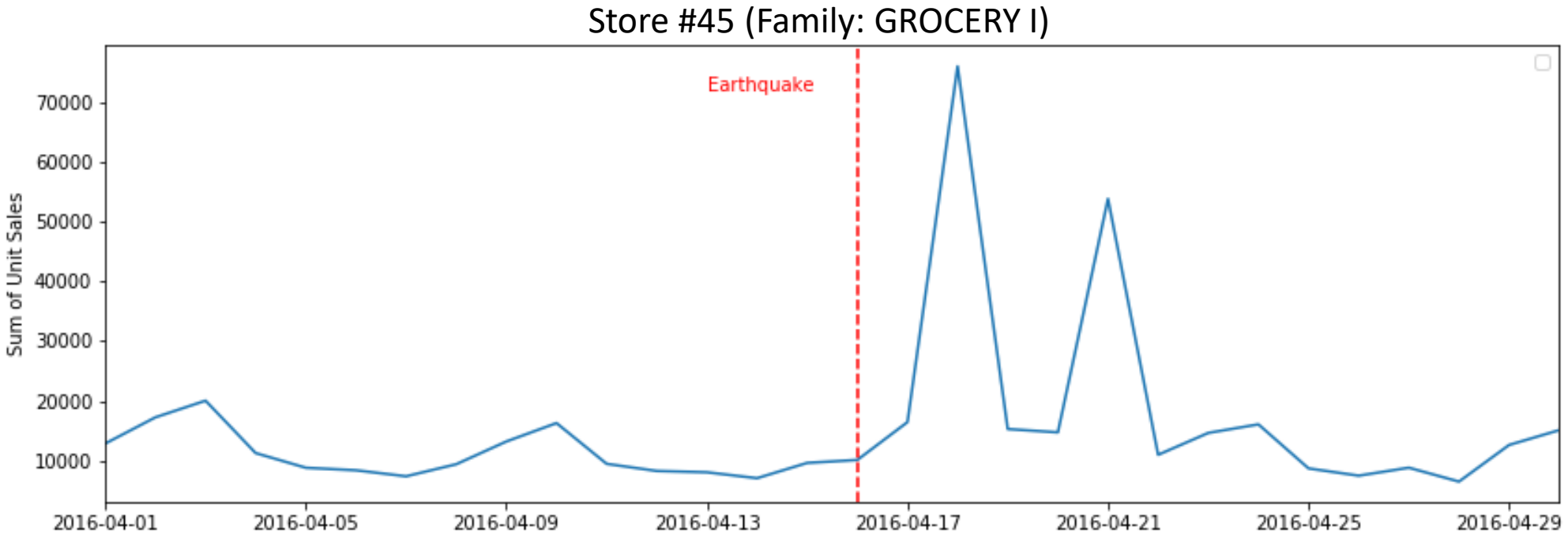
- Increased LIQOUR,WINE,BEER sales before Christmas day and New Year day



Effect of Special Events on Sales

Earthquake occurred on 2016/04/16

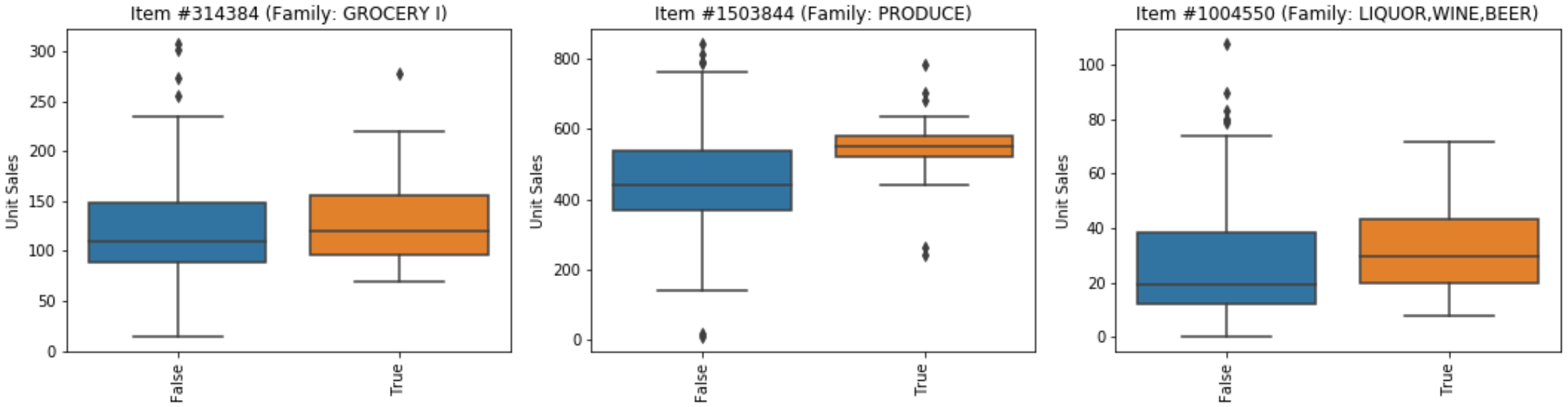
Increased in GROCERY I sales in response to nationwide charity drive



Effect of Promotion of Sales

- Promotion leads to increase in item sales, some more than others

Store #45



Sales Forecasting with Machine Learning

- Test data: Sales data from 2017/08/01 to 2017/08/15 (last 15 days of sales data)
- Evaluation metric: Normalized Weighted Root Mean Squared Log Error (NWRMSLE)
  - Based on Kaggle competition

$$NWRMSLE = \sqrt{\frac{\sum_{i=1}^{n} w_i [\ln(\hat{y}_i + 1) - \ln(y_i + 1)]^2}{\sum_{i=1}^{n} w_i}}$$

n: Total number of item sales across all stores each day in the 15-day period

w_i: Item weight (perishable: 1.25; non-perishable: 1)

$\hat{y}_i$: Predicted unit sales

y_i: Actual unit sales

Machine Learning Approach

- Based on Kaggle's winning solution with modification

1. Create a training, validation and test dataset
2. Use the training dataset to fit machine learning models
   - Create 15 separate models, one for each day of the 15-day period
   - Sales data of each day as target variables
3. Use the validation dataset to determine optimal hyperparameters
4. Make predictions with the test dataset and the best model for the 15-day sales forecasting period

Feature Engineering

Create new features based on:

- Historical sales data, e.g.
  - Unit sales 1 to 15 days before the reference date (1st day of forecasting period)
  - Mean, median, min and max sales and std deviation of sales 3,7, 14 and 30 days
  - Mean sales on days with promotion 3, 7, 14 and 30 days before the reference date
  - Mean daily sales for each date of week (i.e. Sunday, Monday etc.) 4 and 20 weeks before the reference date
- Historical promotion data, e.g.
  - Promotion status from 14 days before to 14 days after the reference date
  - Sum of promotion days with 7, 14 and 30 days before the reference date
  - First and last promotion days 14 days before the reference date
- Stores Information, e.g.
  - City, state, cluster
- Item Information, e.g.
  - Family, class

Machine Learning Models

- Use Microsoft LightGBM (gradient boosting) to establish a baseline performance
  - Grid search to determine optimal hyperparameters
  - Validation data to determine early stopping, i.e. model building will discontinue if there is no improvement to validation score
- Explore neural network (NN) for additional gain in accuracy
  - Evaluate 3 NN architectures
  - Validation data to determine early stopping

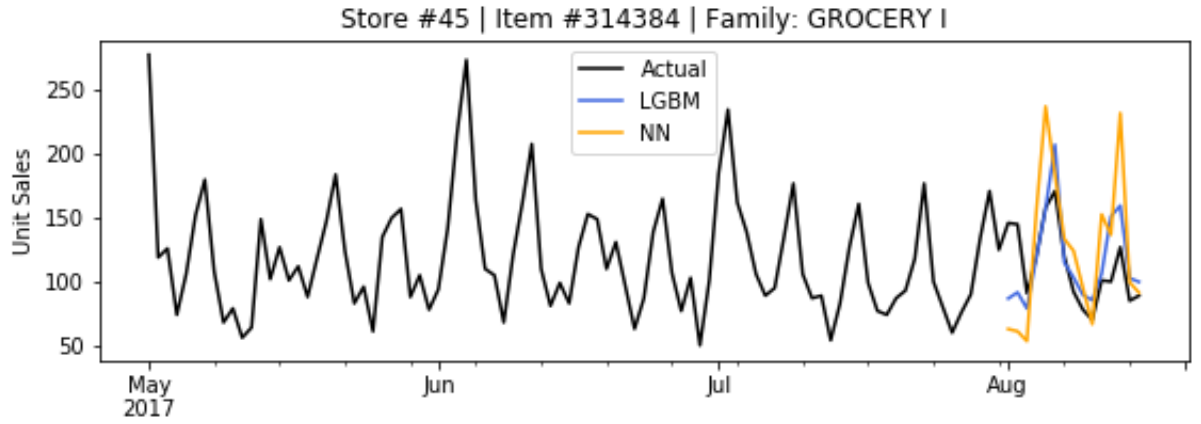
Model Performance

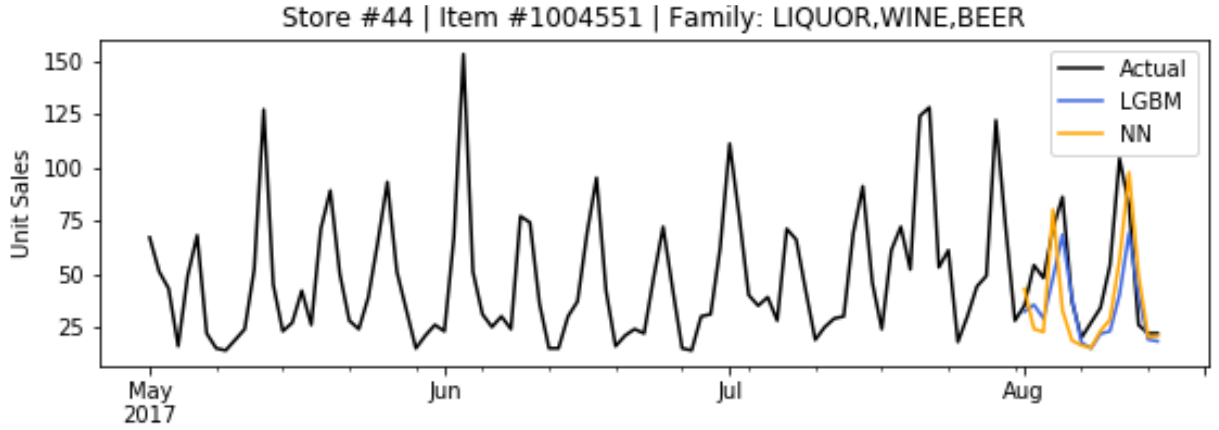
- Overall, the best LGBM model performs better than the best NN model based on NWRMSLE (lower is better)

Model	Validation NWLRMSLE	Test NWLRMSLE
LGBM	0.5866	0.5966
NN	0.6099	0.6222

Performance on Selected Items

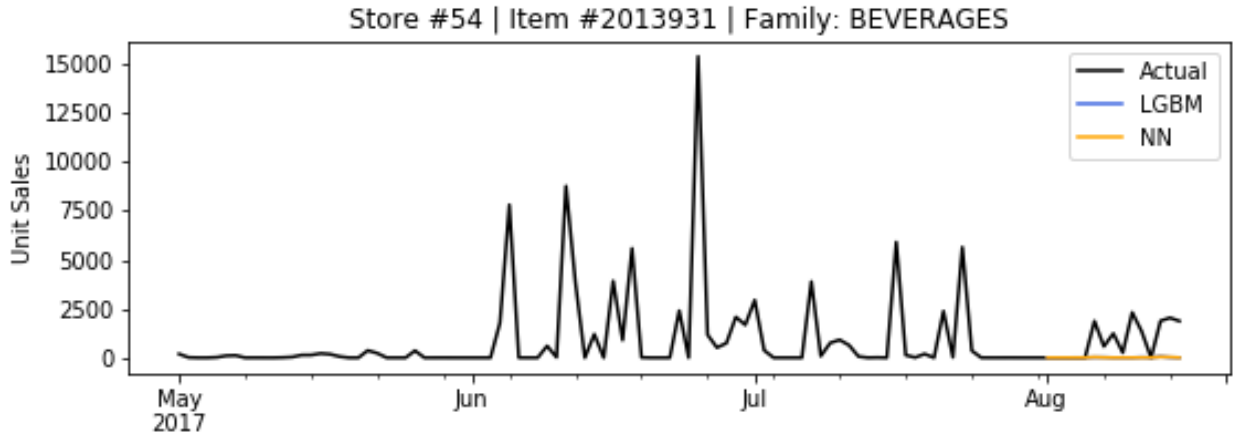
- Models performed reasonably well for items with sales that follow a seasonal pattern

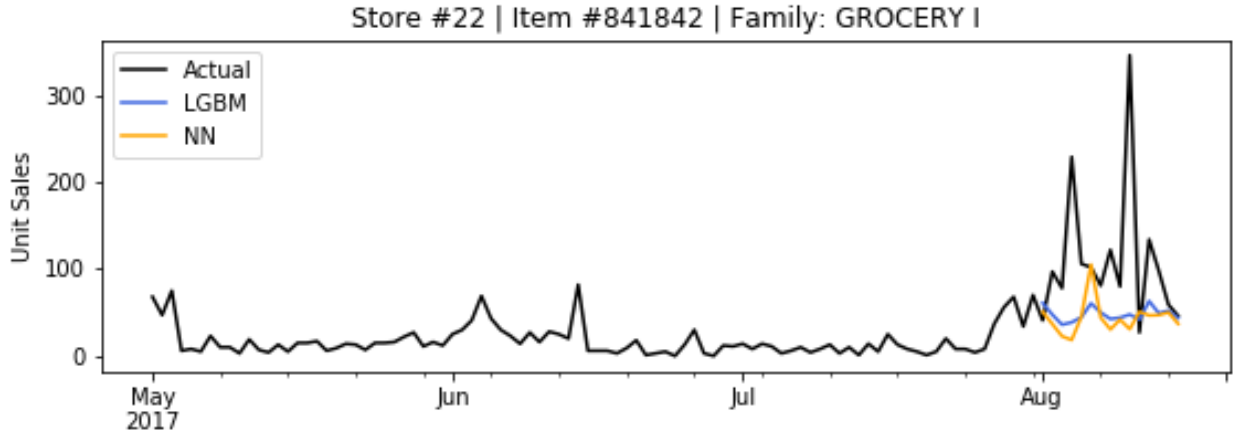




Performance on Selected items

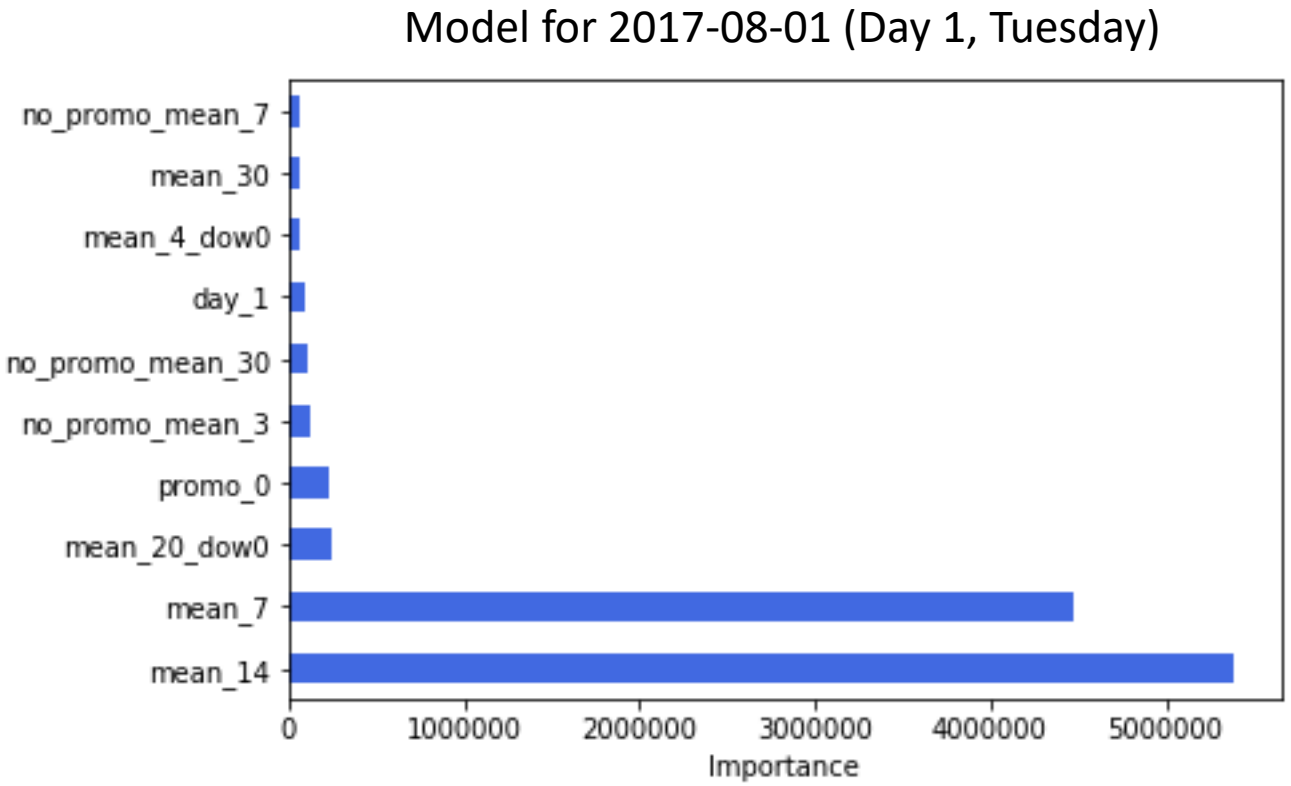
- Models performed poorly for items sales significantly from historical sales
- Model trained on historical data that does not project well into the future





Feature Importance

- Importance features include:
  - Mean sales in the last week or month
  - Mean sales in that day of the week
  - Promotion status on that day



Recommendations and Future Works

Recommendations:

- Use past sales data (e.g. mean sales in last 30 days, mean sales in the same day of week for the last 4 weeks) to predict sales in the coming weeks
- Predict sales during holidays based on historical data on similar events (e.g. last Christmas) and recent sales pattern
- Conduct emergency response planning for sales nationwide during extreme events that include risk management due to logistics interruption

Future Works:

- Improve prediction accuracy with models specific to data clusters (e.g. store, city, item family)
- Understand population demographics and dynamics
- Explore the use of commercial forecast softwares

Thank you!

Contact Information

- George Tang, aspiring data scientist
- Email: georgecctang@gmail.com
- Linkedin: https://www.linkedin.com/in/george-tang-b2b8005/