Corporación Favorita Grocery Sales Forecasting

A casestudy in modelling grocery sales in Ecuadorian supermarkets

David Scroggs

4-12-2022

Introduction

Accurate forecasting of product sales is important to retail stores, as stated in the competition description:

> *"Predict a little over, and grocers are stuck with overstocked, perishable goods. Guess a little under, and popular items quickly sell out, leaving money on the table and customers fuming."*

Competition is to forecast item sales in:

- 54 different grocery stores in 22 cities
- 4,100 sale items
- 4.5 years, 125 million rows of training data
- Forecast period of 2 weeks past final training data

Data

Competition model data is 7 tables, test/train containing the predictor variable and 5 other tables containing additional information.

Table	Rows	Cols	Column names
train	125,497,040	6	id, date, store_nbr, item_nbr, unit_sales, onpromotion
test	3,370,464	5	id, date, store_nbr, item_nbr, onpromotion
transactions	83,488	3	date, store_nbr, transactions
items	4,100	4	item_nbr, family, class,

EDA results

Evaluation metric

Normalized Weighted Root Mean Squared Logarithmic Error (NWRMSLE)

- Weight (wi) are for perishable items (weight = 1.5), other items weight = 1
- Metric accounts for predicting results of a varying order of magnitude

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

Model approach

1. Exploration of the training and auxiliary data sets
   - Explore the predictor variable
   - Explore the predictor relationship with other variables
2. Build a simple model on one item
   - Trial model algorithms
   - Trial data pre-processing
   - Test evaluation metric
   - Feature engineering
   - Model evaluation
3. Expand simple model
   - Model family of items (bread/bakery)
   - Expand model features
   - Tune hyper-parameters
   - Detailed model evaluation

Process

- Early EDA memory limit issues
- Proof of concept model
- Model fit times were long for the simple model
+ Predictor

Results

Final model - Tuned XGBoost

- Features modelled - item number, store number, temporal information (weekday, month, year), store type, store cluster, location (state)
- hyper-parameters tuned: tree depth, min data points in a node, randomly sampled predictors
- Model evaluation: nwrmsle = 0.696 (good kaggle results ~ 0.50-0.53)

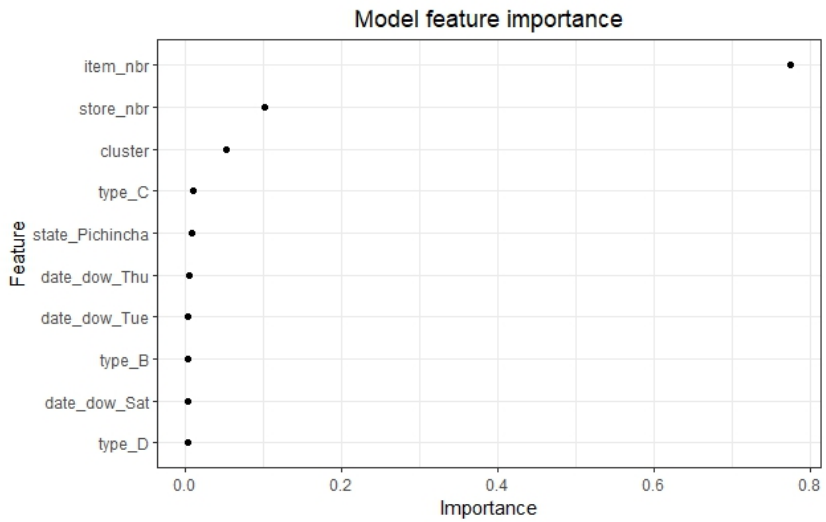

Figure 1: Feature importance

Residuals

- Model tends to under-estimate the result
- Stratification of residuals, particularly in higher unit sales

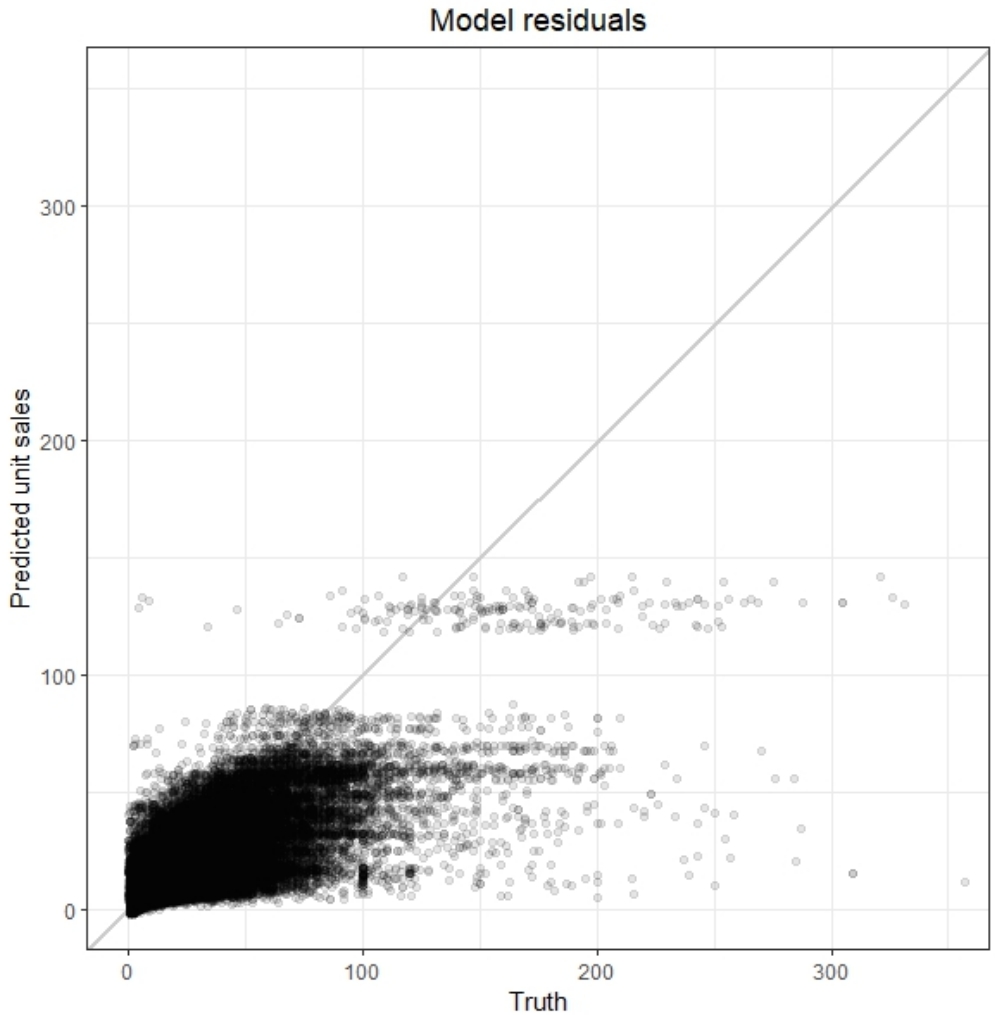

Figure 2: Residuals

Residual outliers

- Holidays not included in model features
  - Potential reason for some extreme outliers
- Store results insensitive to increase in sales
- Stores in 42, 49 large overestimates
  - Only stores in Quito city, and Type C
  - Both dummy features had high importance

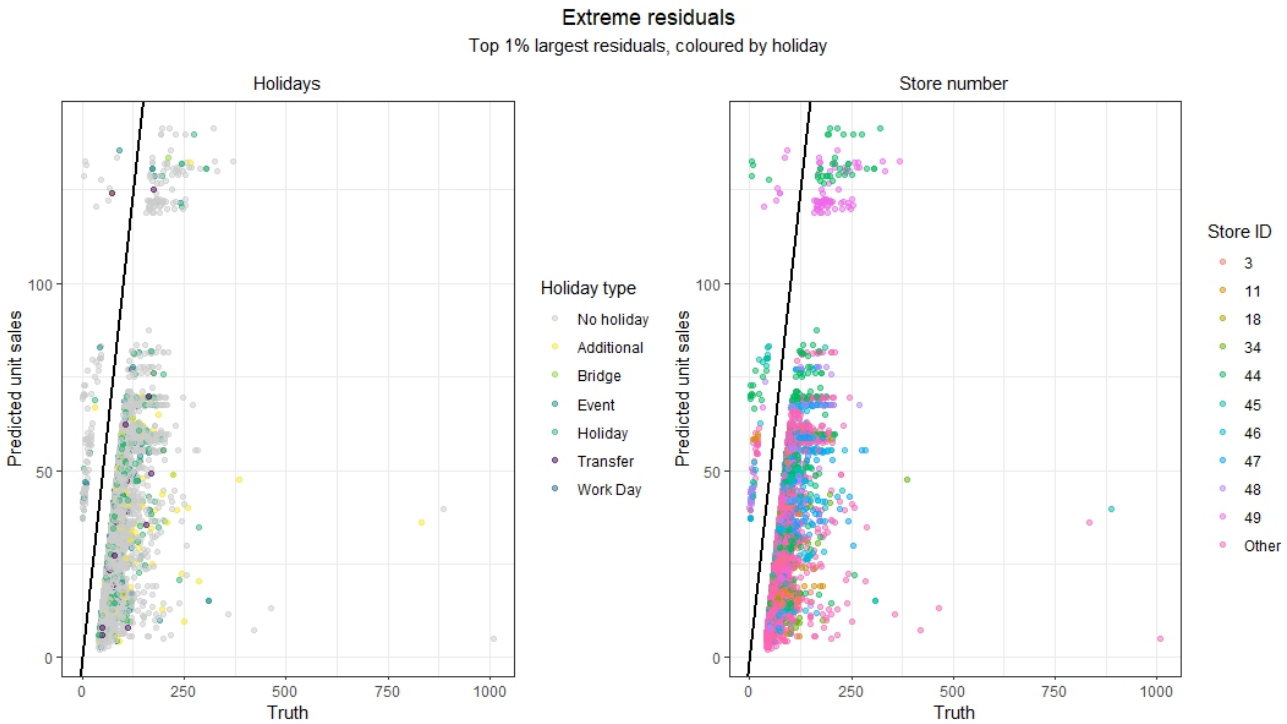

Figure 3: Extreme residuals

Limitations

- Large data set meant:
  - detailed EDA was difficult
  - modelling times were long
- Holiday data detailed and complex (holiday transfers)
- Bakery/bread family didn't have new items - issue not accounted for

Improvements/Future work

- Explore poor prediction performance of stores 42, 49
- Explore further feature engineering of bakery family model
  - Holiday information
  - Regional information
- Further EDA on other item families
- Further model hyper-parameter tuning
- EDA of effect of oil price, promotions, and re-investigate pay-day impact

Production deployment considerations

- New items/stores without history need coding
  - Work required to develop an approach to estimating these events
- Corporación Favorita work in 2 week horizons
  - Regular model retraining (daily)