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 and 125 million rows of training data
- Forecast period of 2 weeks past final training data

Teams given 3 months to produce their best result, uploaded via Kaggle

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, perishable
oil	1,218	2	date, dcoilwtico
holidays	350	6	date, type, locale, locale_name, description, transferred
stores	54	5	store_nbr, city, state, type, cluster

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

Exploratory Data Analysis

Results

- Clear day-of-week trends (autocorrelation)
- Significant range in store sales (1-15 million units/year)
- Item family types
  - A few types had a large proportion of all items

Issues

- Early EDA memory limit issues
- Number and range of unit sales made visualisation challenging

First simple model

- A single common bread/bakery item
- 83,500 rows of data

Results

- Trialled 3 model recipes
  - Recipe 1: store number and temporal data
  - Recipe 2: store information (number, location, type, cluster and daily transactions) and temporal data
  - Recipe 3: As above and includes pay-day information
- Trialled 2 model algorithms
  - Random forest
  - XGBoost
- Model algorithms had similar accuracy
- Recipe 2 had best accuracy

Issues

- Random forest model fit times were long for the simple model

First model results

Recipe	model	.metric	mean	n
Recipe 1	boost_tree	rmsle	0.3808435	5
Recipe 1	rand_forest	rmsle	0.5025597	5
Recipe 2	boost_tree	rmsle	0.3148345	5
Recipe 2	rand_forest	rmsle	0.3095069	5
Recipe 3	boost_tree	rmsle	0.3150564	5
Recipe 3	rand_forest	rmsle	0.3092414	5

Bakery/bread family model

- A model fitting the bread/bakery item family

Results

- XGBoost model and Full recipe from simple model used
- Hyper-parameters tuned: tree depth, min data points in a node, randomly sampled predictors
- Model evaluation: nwrmsle = 0.696 (good kaggle results $\sim$ 0.50-0.53)

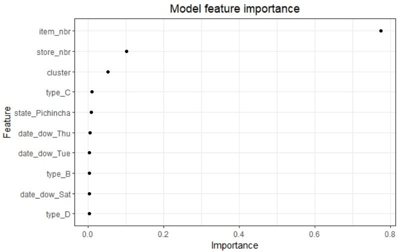

Figure 1: Feature importance

Hyper-parameter tuning results

mtry	min_n	tree_depth	trees	learn_rate	.metric	mean	n
13	2	8	1000	0.02	rmse	8.672	4
13	21	15	1000	0.02	rmse	8.692	4
13	21	8	1000	0.02	rmse	8.795	4
13	40	15	1000	0.02	rmse	8.820	4

Note: Metric is rmse

Residuals

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

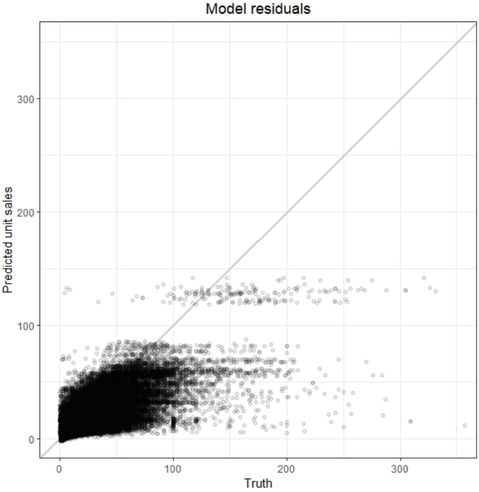

Figure 2: Residuals

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

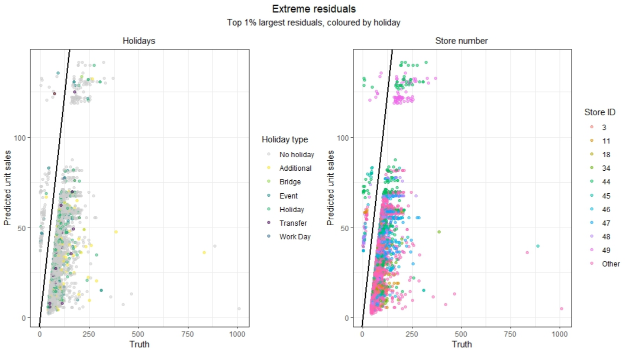

Figure 3: Extreme residuals

Limitations

- Detailed EDA was difficult due to:
  a) size of dataset
  b) diversity of product items
  c) time constraints
- Model parameter (feature and hyper parameters) exploration limited
- Holiday data detailed and complex (holiday transfers)
- Items modelled were common - no new items/stores in test set
  - Method required for full data set

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
- Investigate ARIMA model for temporal effects

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/weekly)
  - Develop item forecast performance metric
    - Continuous monitoring of results
  - Monitor and log large model errors (new holidays)