CORPORACIÓN FAVORITA

Grocery Sales Forecasting

Matt Barrett, Clarissa Franklin, Tim Lai, Brett Scroggins, Meyappan Subbaiah

kaggle

Project Outline

- Brick-and-mortar grocery stores are always in a delicate dance with purchasing and sales forecasting.
- Corporación Favorita is a large Ecuadorian-based grocery retailer that operate hundreds of supermarkets, with over 200,000 different products on their shelves.
- Corporación Favorita has challenged the Kaggle community to build a model that more accurately forecasts product sales.

CORPORACIÓN FAVORITA

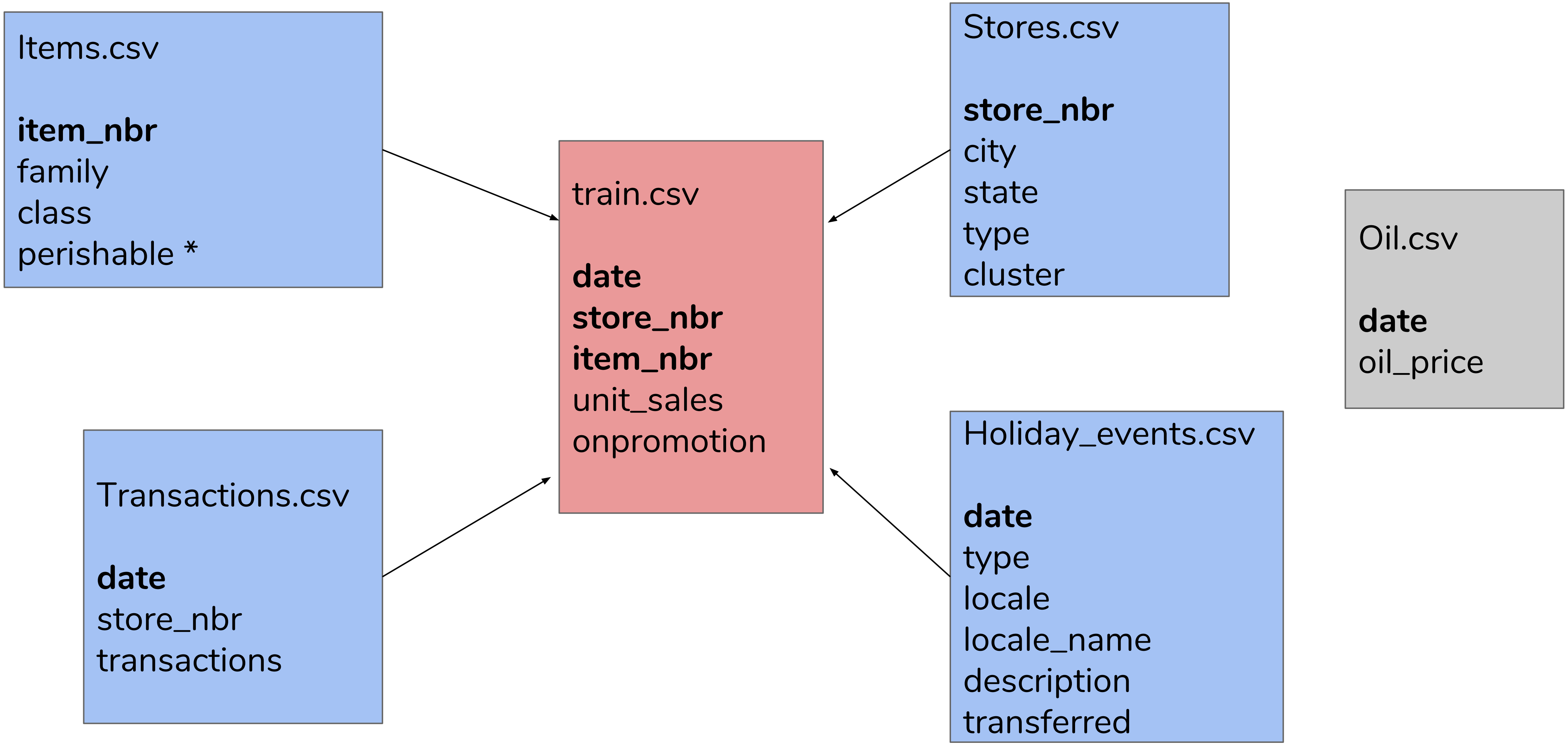
Data Complexity…

- Training Set (4.88 GB) - About 125 million rows
- Lots of categorical errors which will yield LOTS of dummies (ie, 4096 items)
- Leveraged cloud computing for data wrangling
- 10% sample for models

"Here's a list of 100,000 warehouses full of data. I'd like you to condense them down to one meaningful warehouse."

CORPORACIÓN FAVORITA

Data Explained



Project Goals

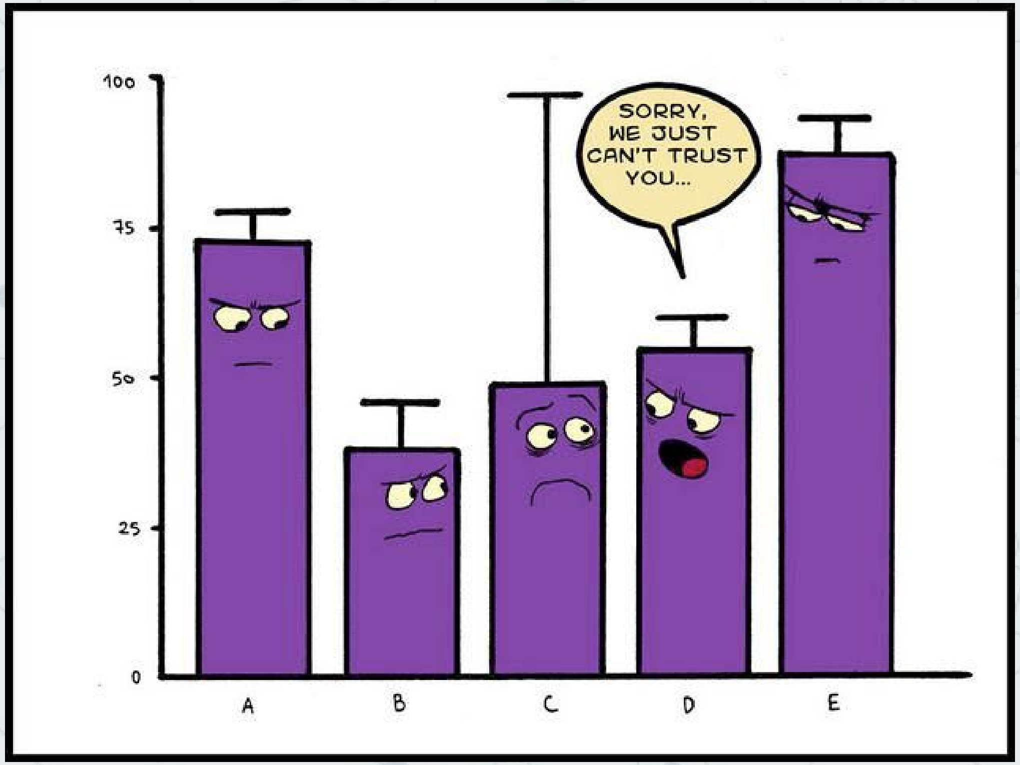
- Identify key factors
- Select an appropriate model minimizing loss metric
    - Loss metric: Normalized Weighted Root Mean Squared Logarithmic Error

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

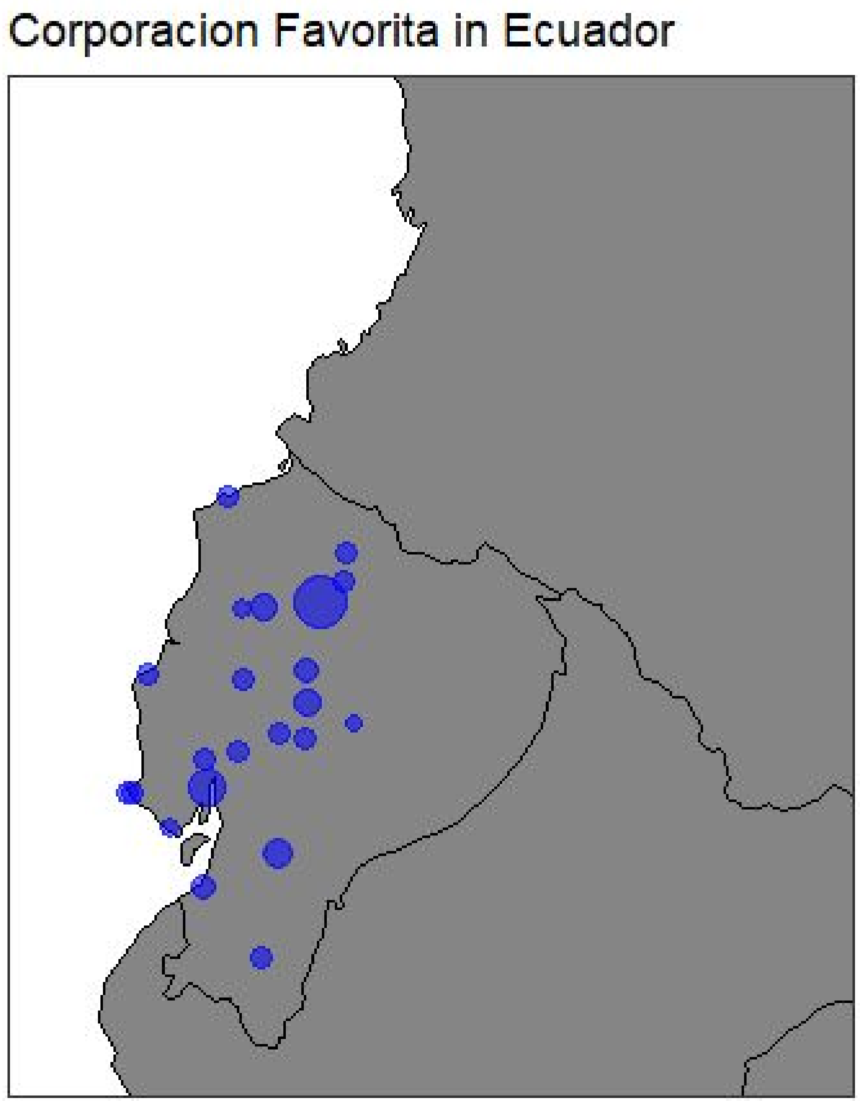
- Optimize inventory of individual products at each location to maximize profit

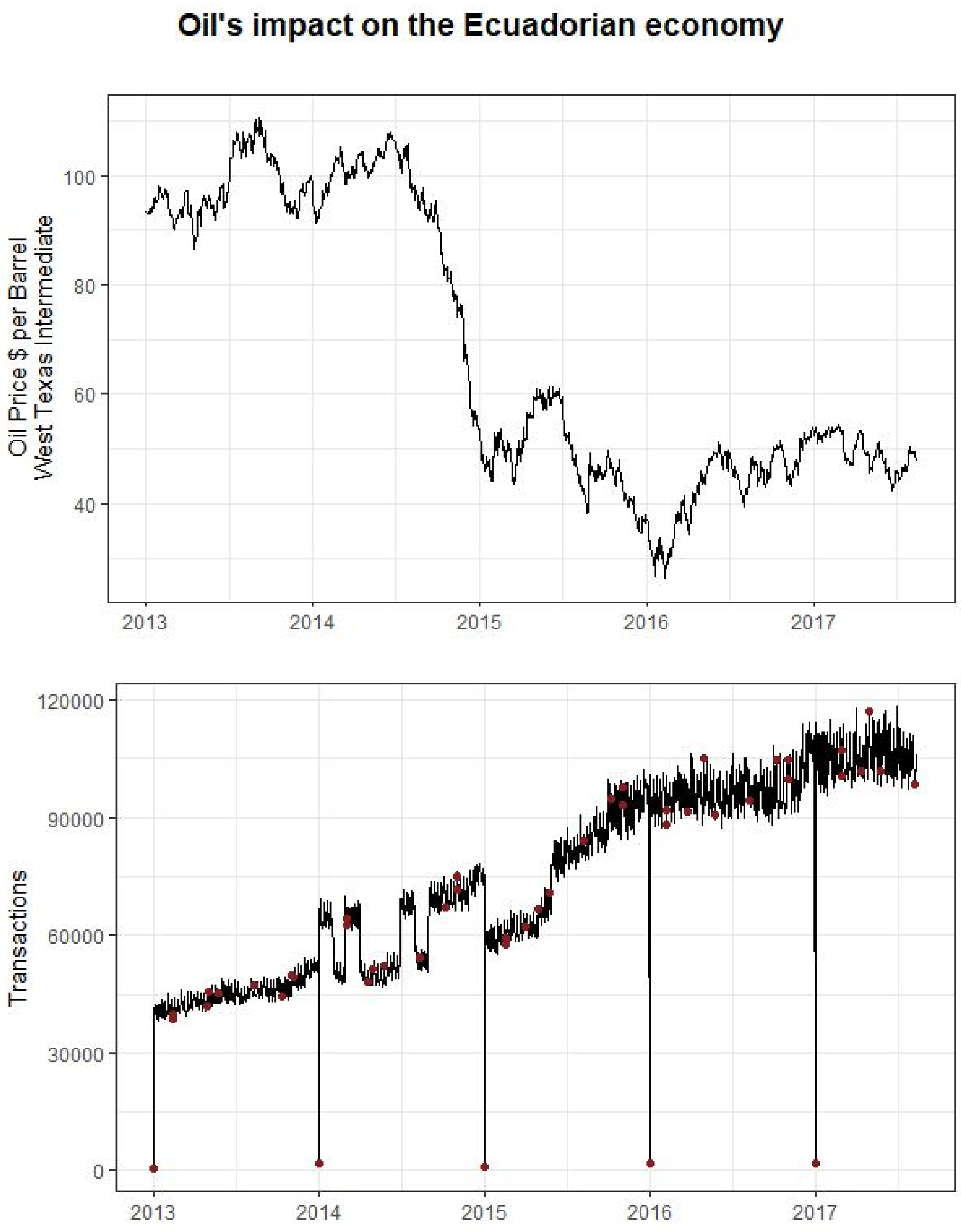
CORPORACIÓN FAVORITA

Exploratory Data Analysis

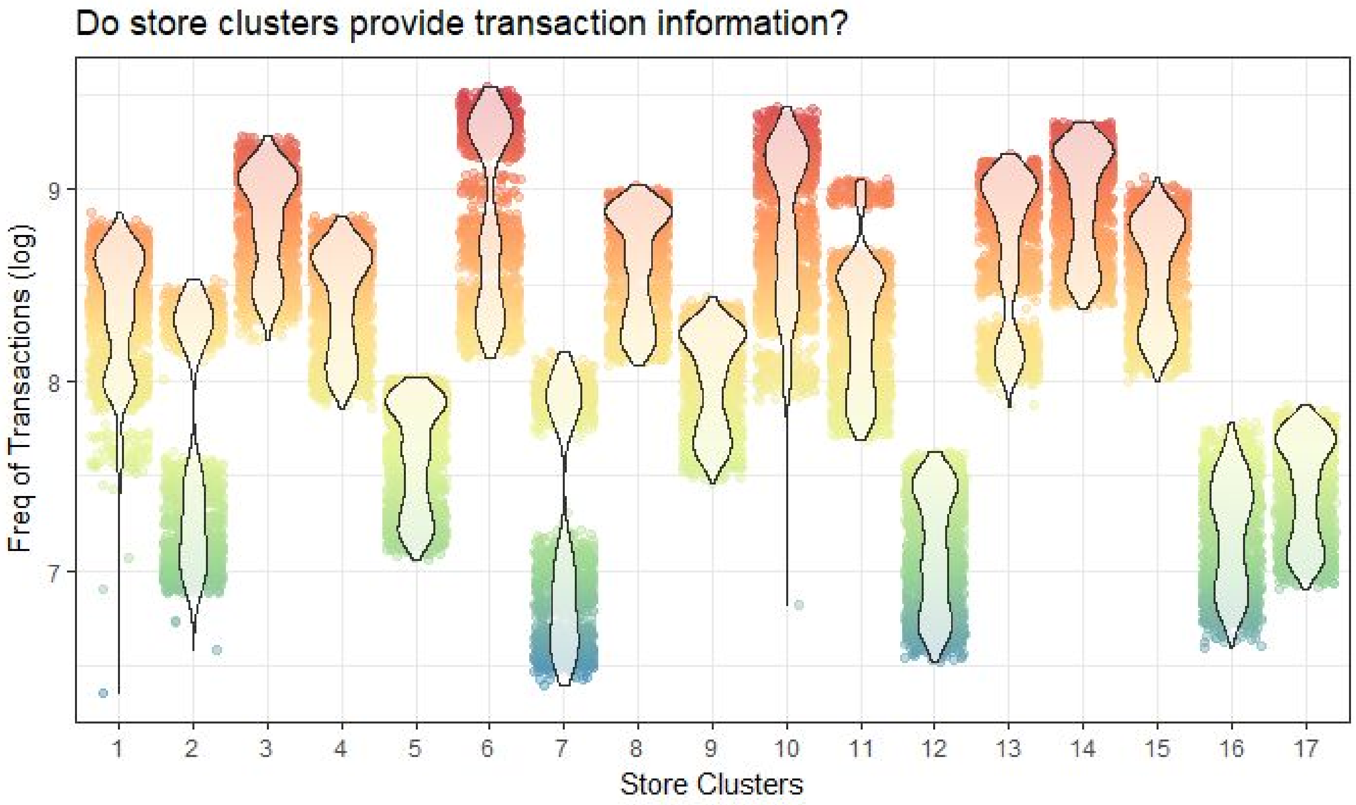


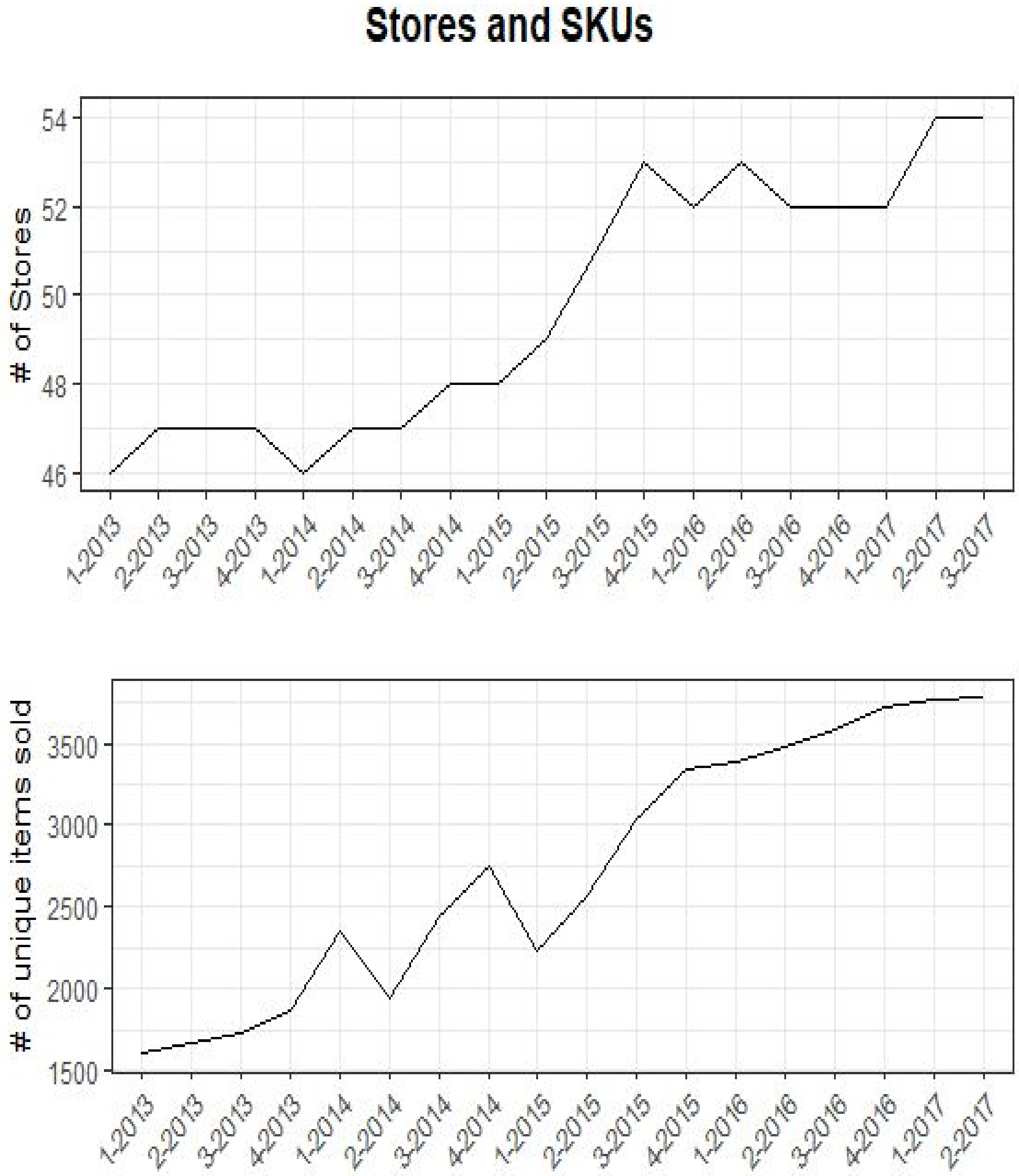
Economic Influences



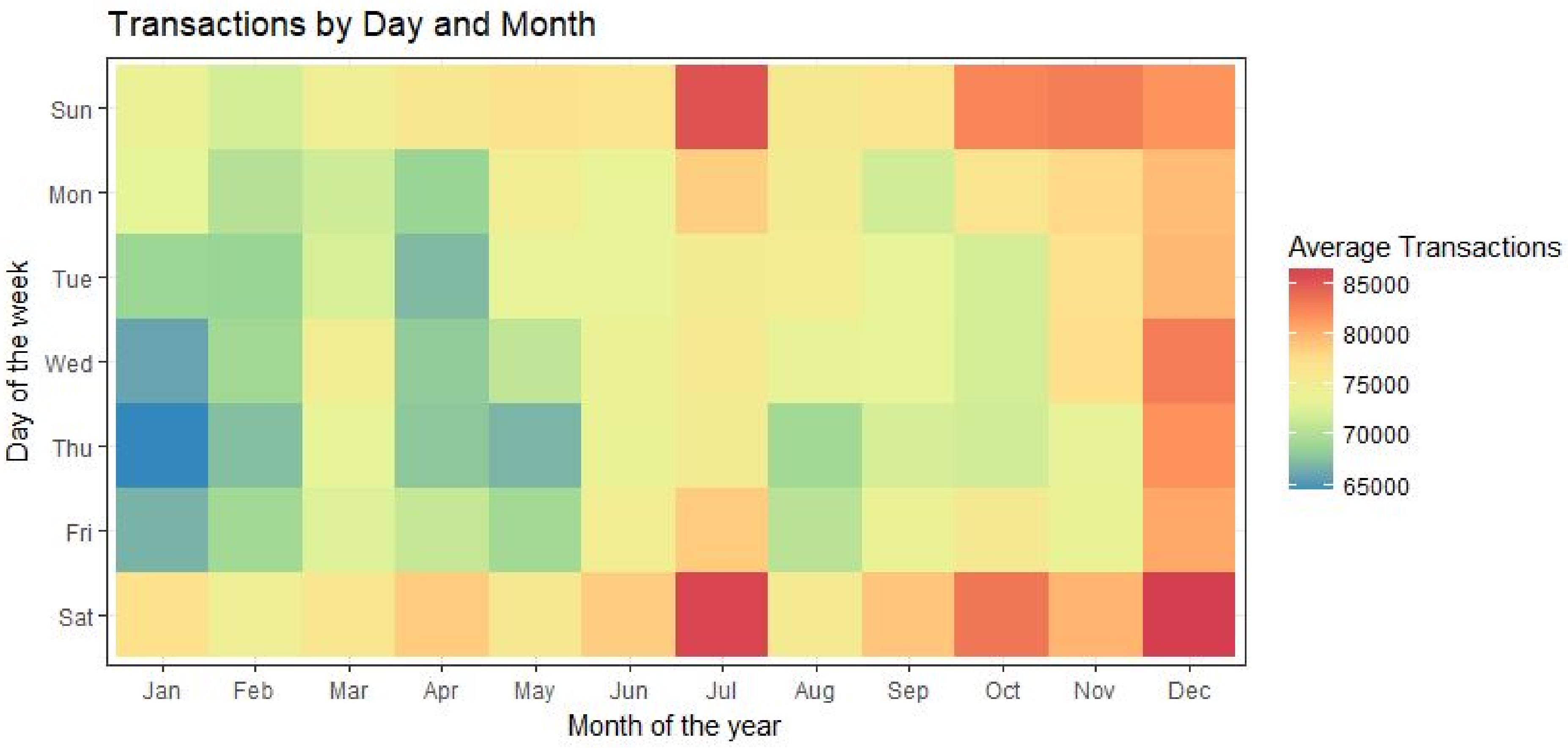


Stores and Products





Transactional Trends



Transactional Contents

Transactions by Group/Item (Log scale)

CORPORACIÓN FAVORITA

Feature Selection and Engineering

- Separated Full Date into day, month and year
- Dropped Store ID and Cluster - focused on store type (which grouped similar stores together)
- Holiday Importance
- Store Transactions - Created daily transactions per store

Model Fitting

Kaggle Scoring System

Normalized Weighted Root Mean Squared Logarithmic Error

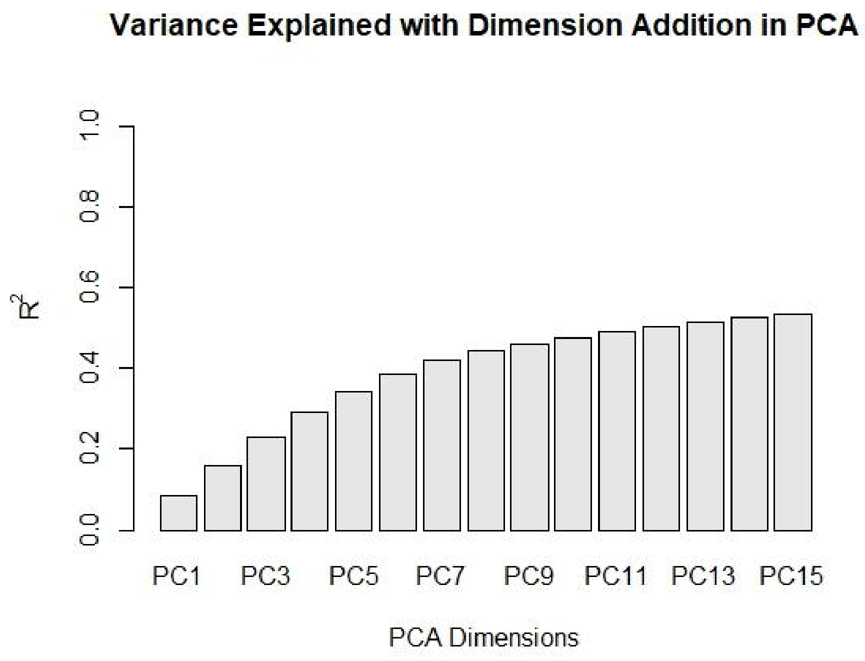
$$NWRMSLE = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(\ln(\hat{y_i} + 1) - \ln(y_i + 1)\right)^2}{\sum_{i=1}^{n} w_i}}$$

For reference, the naive forecast = 0.911 (Kaggle's benchmark score)

CORPORACIÓN FAVORITA

PCA

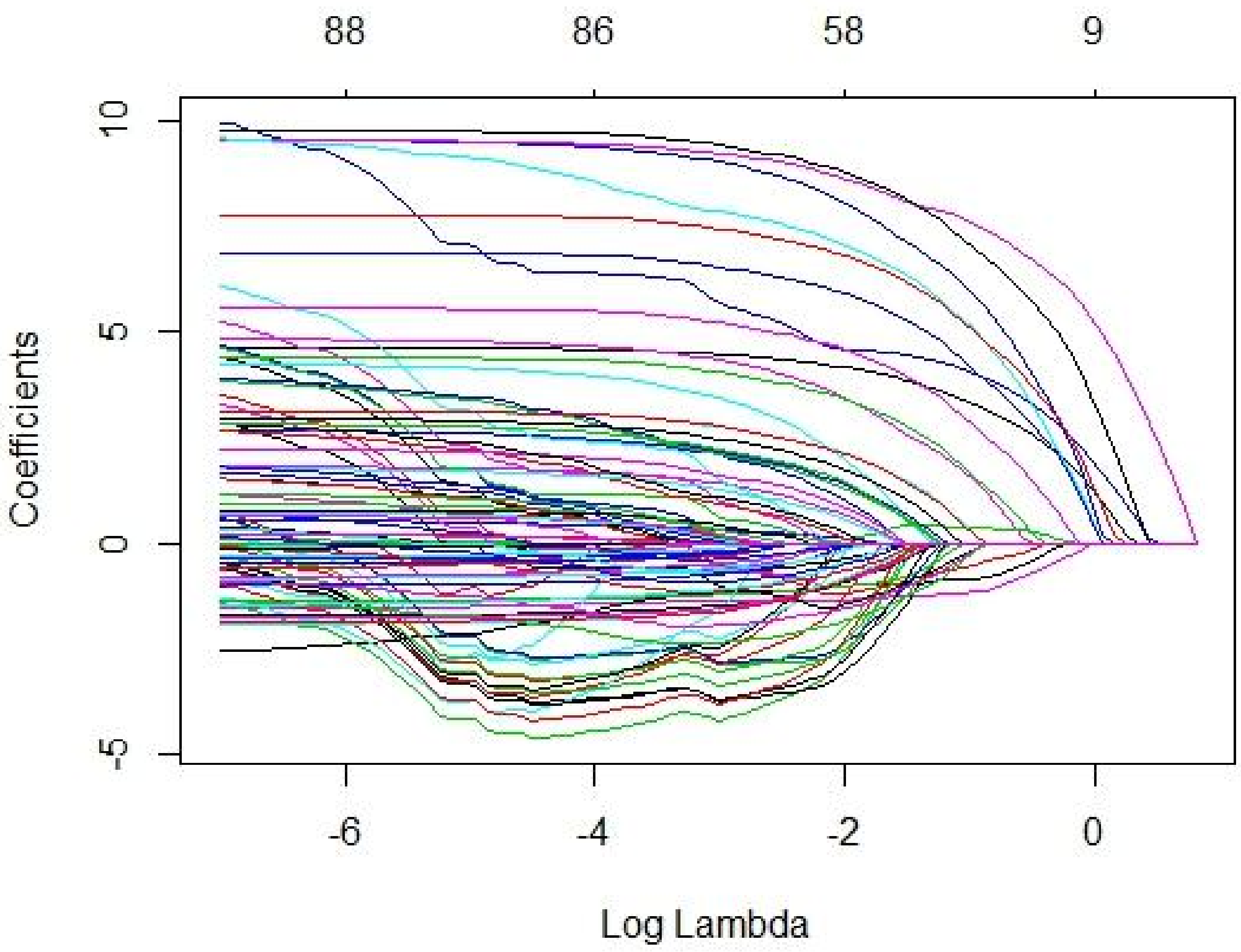
- Implemented to determine importance of variables
- Ultimately showed no potential
  - Limited variance explained even at high dimensions of PC's



Feature Reduction – LASSO

- Effect of factorizatior
  categorical variables

- Best lambda value: N
  (regular linear regres



RACIÓN FAVORITA

Experimental Results – Multilayered Perceptron

- Parameters
    - hidden layers
    - activation function
    - learn rate
    - batch size

Hidden Layers	Activation Function	Learning Rate	Batch Size	Score
2	tanh	0.005	500	0.916
4	relu	adaptive	200	0.903
5	relu	adaptive	1000	0.909

- Error Metric = 0.903

CORPORACIÓN FAVORITA

Experimental Results – Gradient Boosted Machine

- Parameters
  - number of trees
  - max depth
  - learn rate
  - min rows
- Predictors
  - 5 predictor variable subsets

# trees	max depth	learn rate	min_rows	nwrmsle
5	5	0.1	10	0.906
5	5	0.2	10	0.881
5	15	0.2	2	0.882
5	20	0.2	2	0.883
5	5	0.2	5	0.933

- Error Metric = 0.881

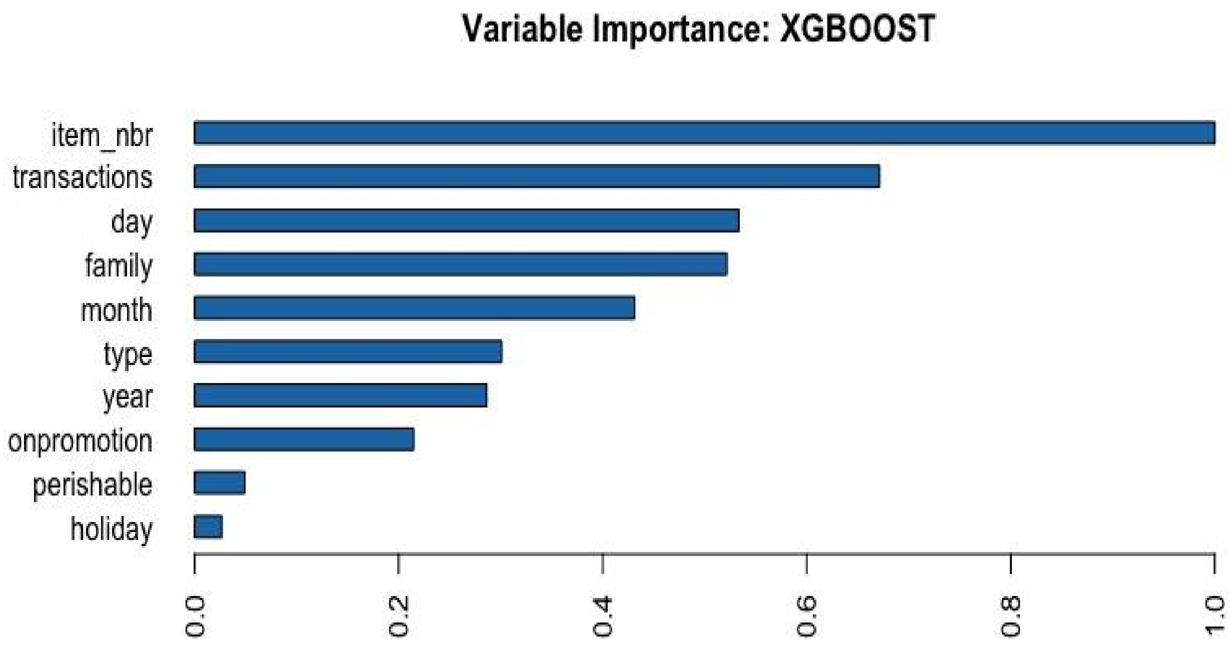
CORPORACIÓN FAVORITA

Experimental Results – XG Boost

- Number of Trees had a significant impact as did learn rate and max depth

- Time fitting was excessive for computing resources

- Error Metric = 0.793



CORPORACIÓN FAVORITA

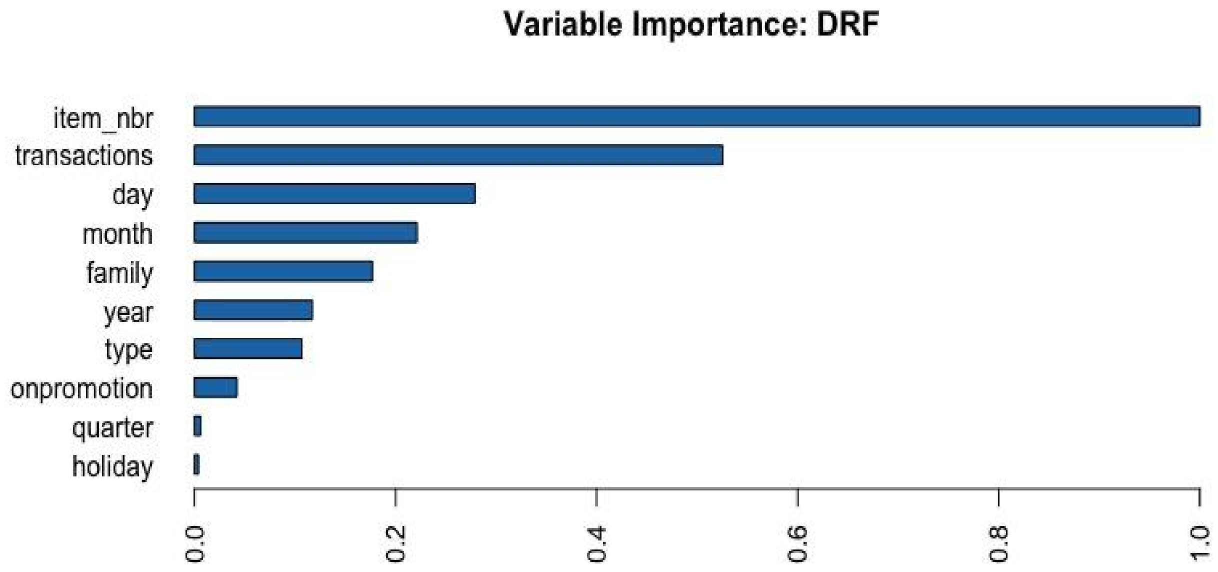
Experimental Results – Random Forest

- Output our Best Score

- Number of Trees did not impact much - depth was more important

- Again, time fitting was high for computing resources

- Error Metric = 0.728



CORPORACIÓN FAVORITA

Discussion of Models

Method	Scoring Metric (NWRMSLE)
Naive Forecast	0.911
MLP Regressor	0.903
GBM	0.881
XGBoost	0.793
Random Forest	**0.728**

Mean Item Sales Forecast	**0.726**

CORPORACIÓN FAVORITA

Modified Approach

- Came to the realization that our models were overwhelmed by the training set's size and complexity
    - Item Number was #1 on variable importance plots

- Models needed to focus on a more granular level

CORPORACIÓN FAVORITA

Revised Experimental Results

<u>Random Forest</u>

50 Trees
Max Depth of 10

Error Metric = 0.542

<u>XGBoost</u>

100 Estimators
Max Depth of 15
Learn Rate is 0.1

Error Metric = 0.639**

CORPORACIÓN FAVORITA

Revised Experimental Results

<u>GBM</u>

0.1 Learning Rate
Max Depth of 25
40 Estimators

Error Metric = 0.542

<u>MLP</u>

2 hidden layers
tanh activation function
learning rate of 0.005

Error Metric = 0.831

CORPORACIÓN FAVORITA

Method	Initial Approach Score	Revised Approach Score	With Log Sales
Naive Forecast	0.911		
Mean Item Sales Forecast	0.726		
Random Forest	0.728	0.542	0.513
XGBoost	0.793	0.553	n/a
GBM	0.881	0.542	0.507
MLP Regressor	0.903	0.831	0.795
Tree-Based Ensemble	n/a	n/a	0.511

Revised Experimental Results

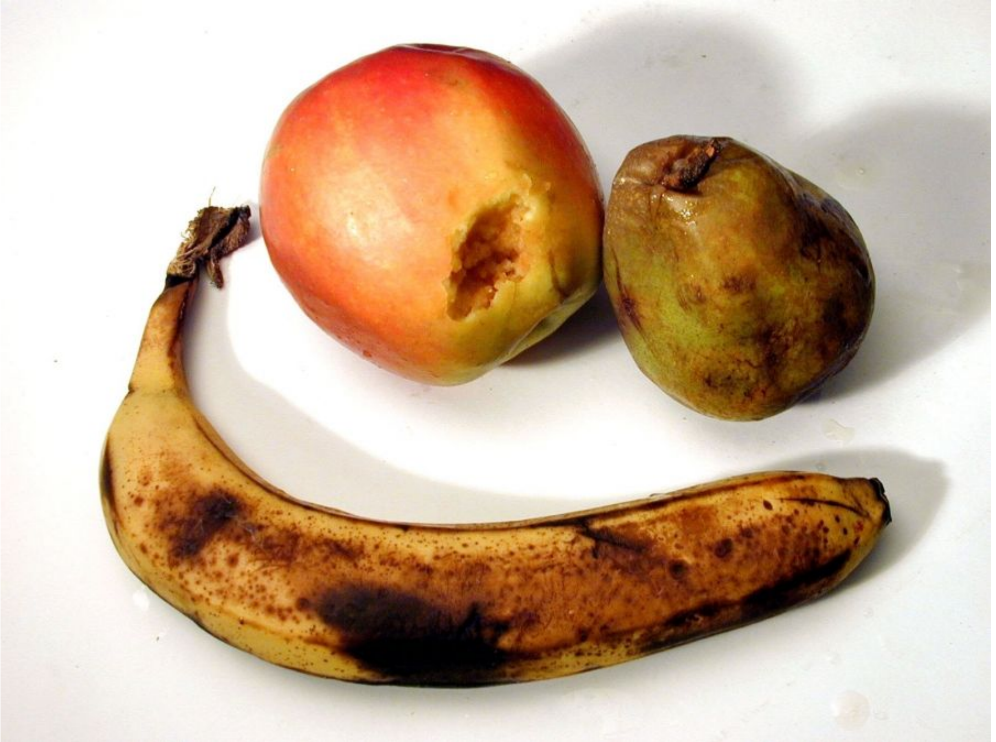
- Computational efficiency of revised approach
- What's next for our models?
    - By Store
    - Log Sales
    - Incorporate effects of oil

CORPORACIÓN FAVORITA

Key Takeaways

Business Value

- Why are we solving this problem?
  - Perishables
  - Struggling growing economy
  - Profitability



CORPORACIÓN FAVORITA

Next Steps

- Account further for the weight of perishable items due to a more limited shelf life
- Incorporate oil forecast to determine prediction impact
- Complexity and overfitting is still a challenge that we must account for

Any Questions?

Appendix

Set 1: onpromotion, year, month, day, quarter, type, cluster, item_nbr, family, class, perishable, store_nbr, transactions, holiday, day_of_week
Set 2: on promotion, year, month, store_nbr, transactions, holiday, day_of_week, avg. unit sales
Set 3: keep onpromotion, year, month, holiday, day_of_week, avg. unit sales
Set 4: keep onpromotion, year, holiday, day_of_week, avg. unit sales
Set 5: all basic predictors+average, using log(sales+1)