Forecasting Grocery Sales with Regression Models and Random Forests

Victor de Fontnouvelle
Justin Weltz
Thomas Thornton

Introduction

- Finite supply
- Perishable

How to predict demand?

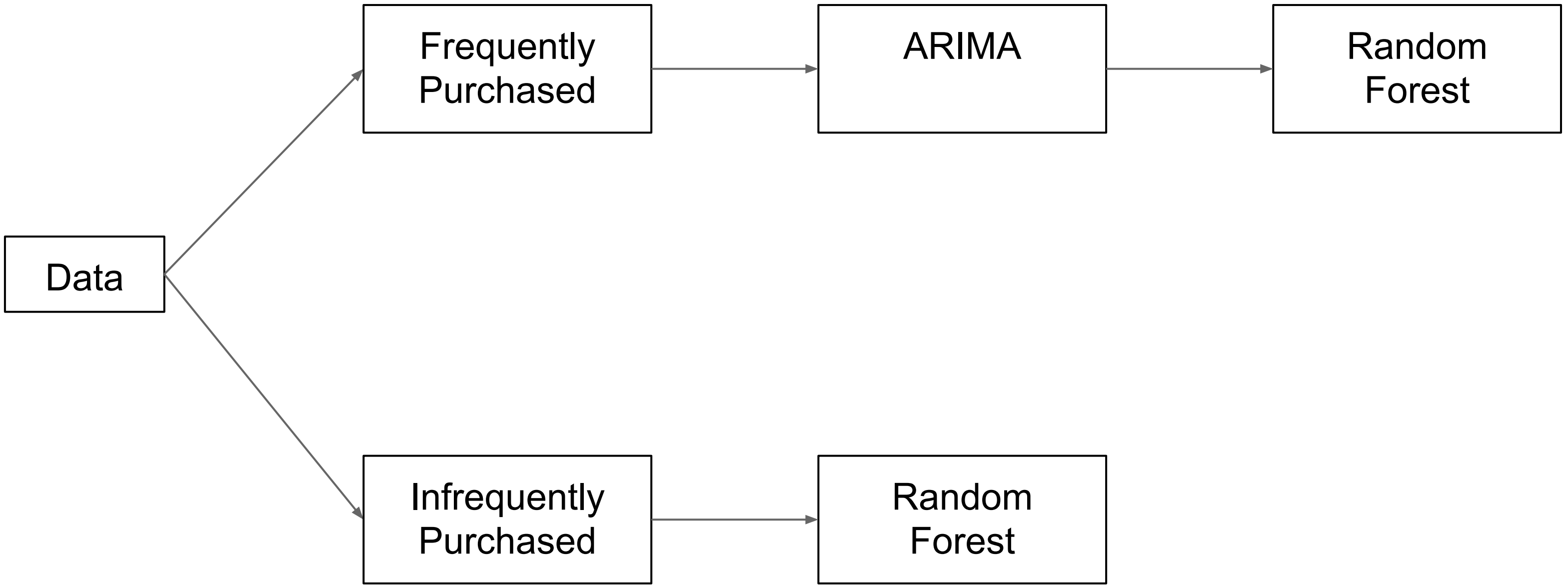
Corporacion Favorita

- 54 stores
- 200,000 items
- 125M transactions

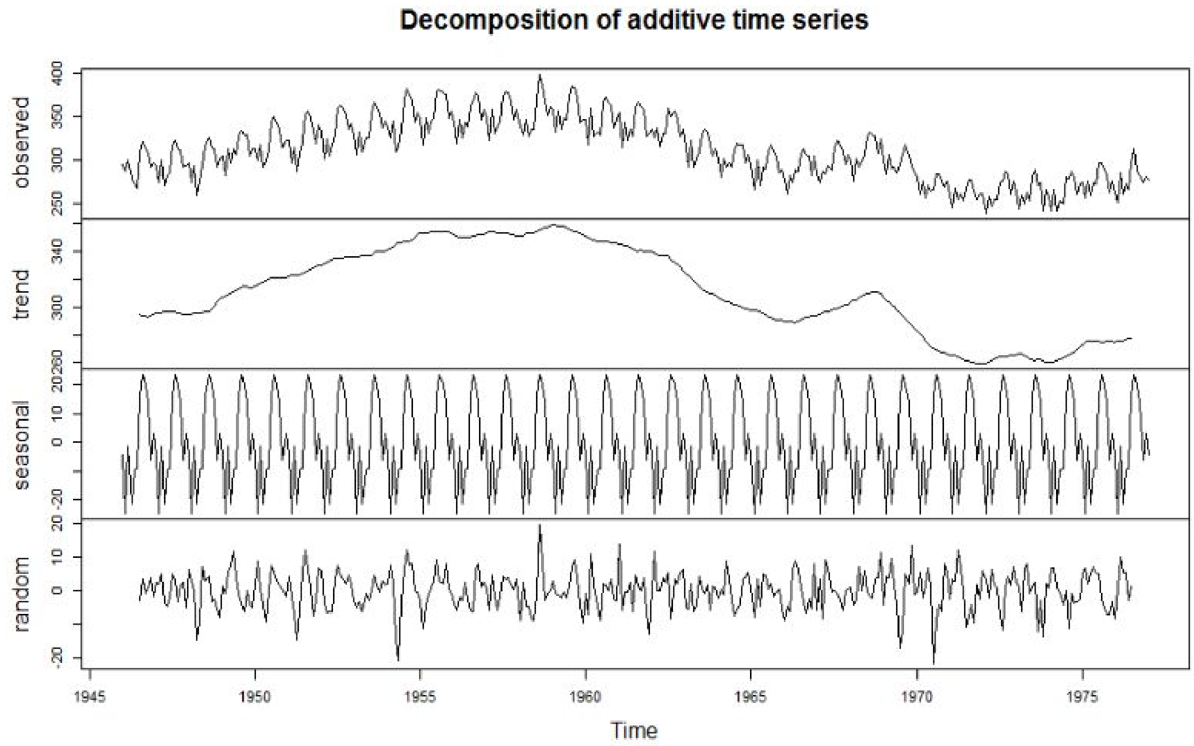
Datasets & Wrangling

- Items - type
- Stores - location, cluster
- Oil - date, price
- Train/Test Data - item number, store number, date ➜ **sales**
  - Joined tables
  - Converted categorical to binary
  - Added numerical variable: days since payday

Dual Model Pipeline



Time Series Forecasting

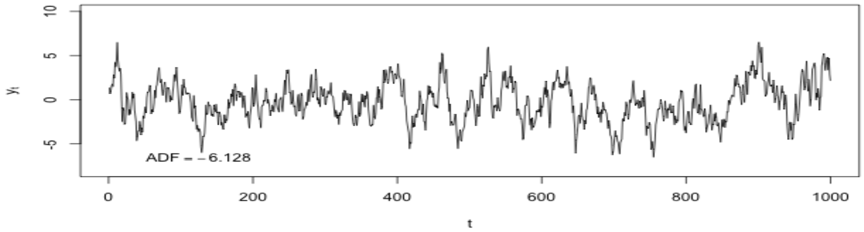


The ARIMA Model Equations

AR

$$Y_t = \alpha + \rho Y_{t-1} + e_t$$

I



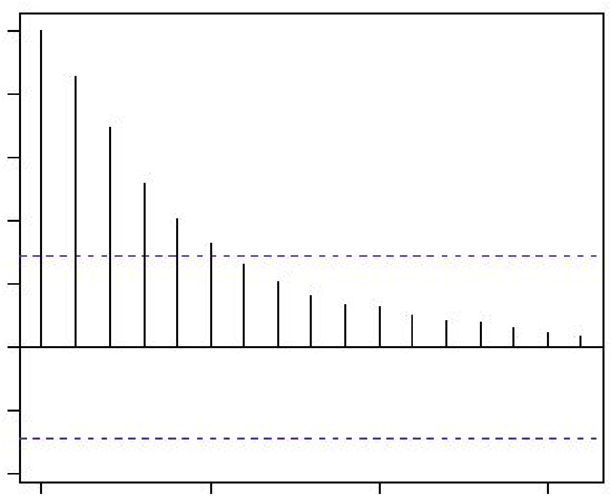
MA

$$Y_t = \alpha + \rho e_{t-1} + e_t$$

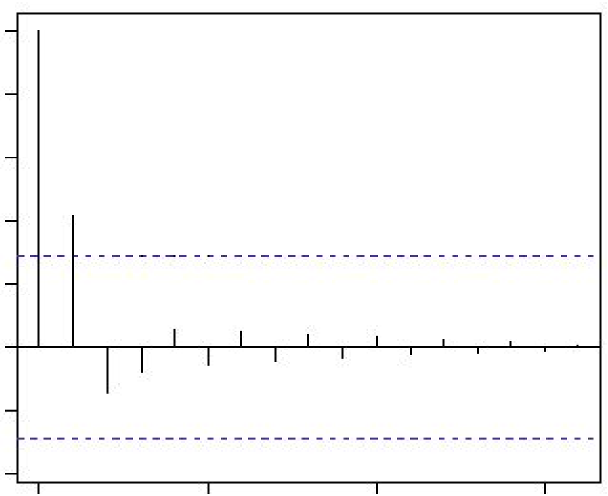
The ARIMA Model Choice

AR

MA



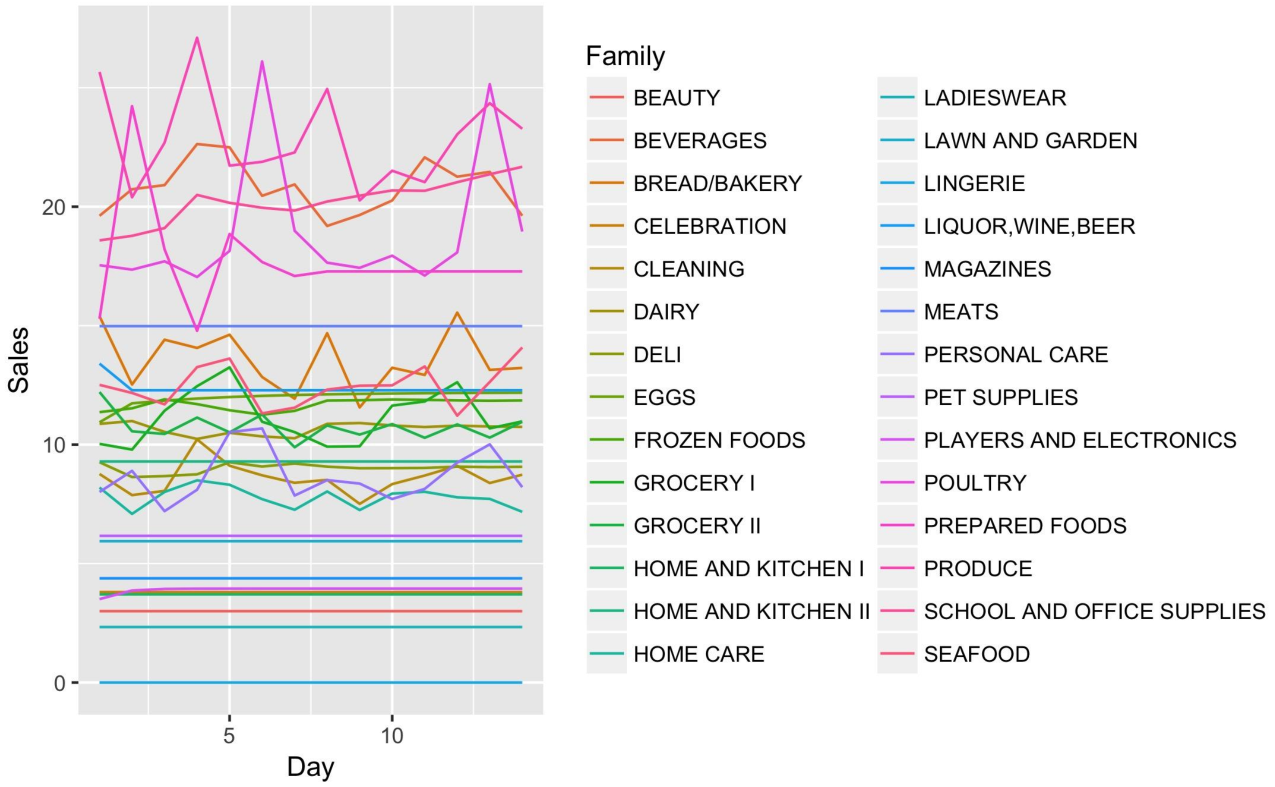
ACF Plots



Popularity


Silk
SOYMILK
Kids Fortified
8th continent
soymilk
Dean's
Easy
FAT FREE MILK
HEAVY WHIPPING CREAM
milk chug
3.49
1.99
3.29
Enhanced

ARIMA Forecast

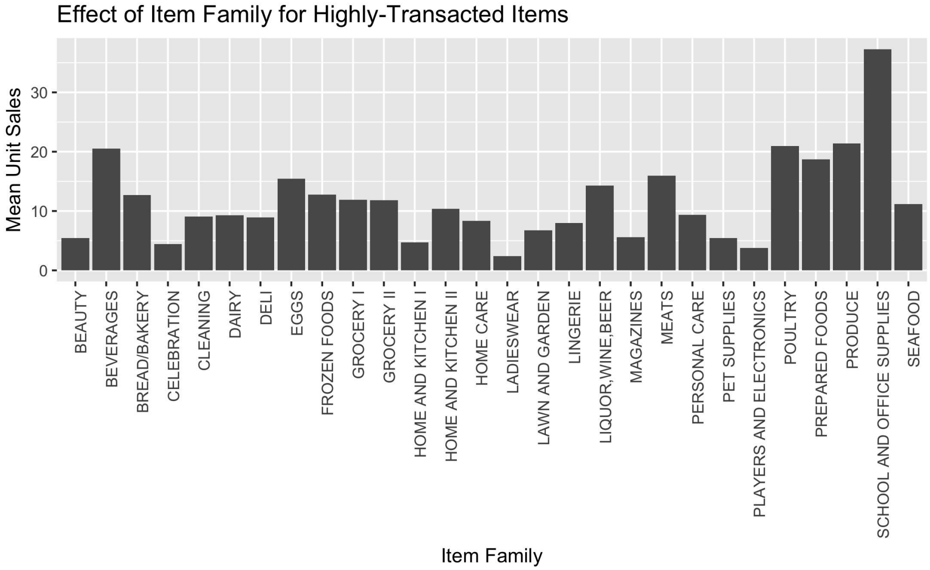


Random Forest Creation

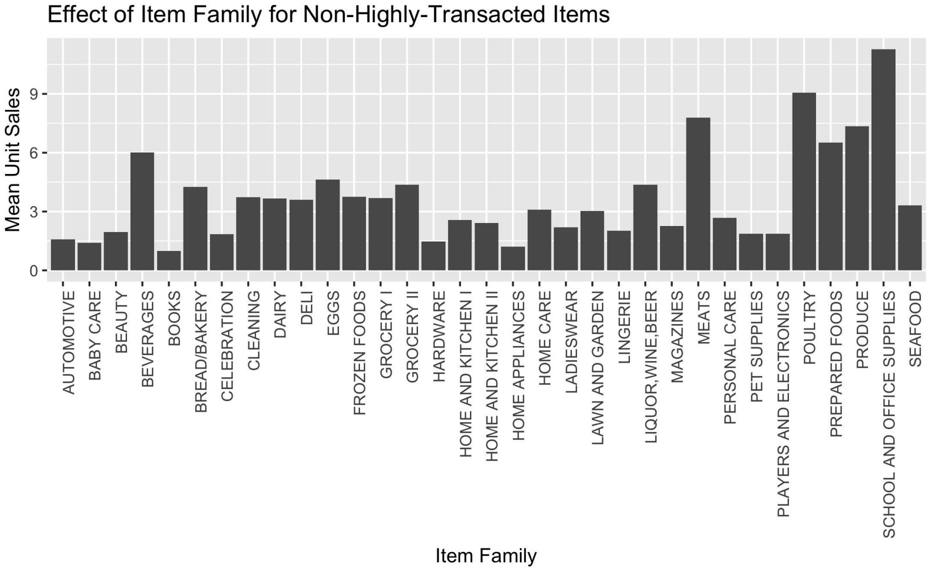
Variables Included:

- Item Family
- Store City
- Store Cluster
- Store Type
- Day of Week
- Days Since Payday
- Popularity

Item Family

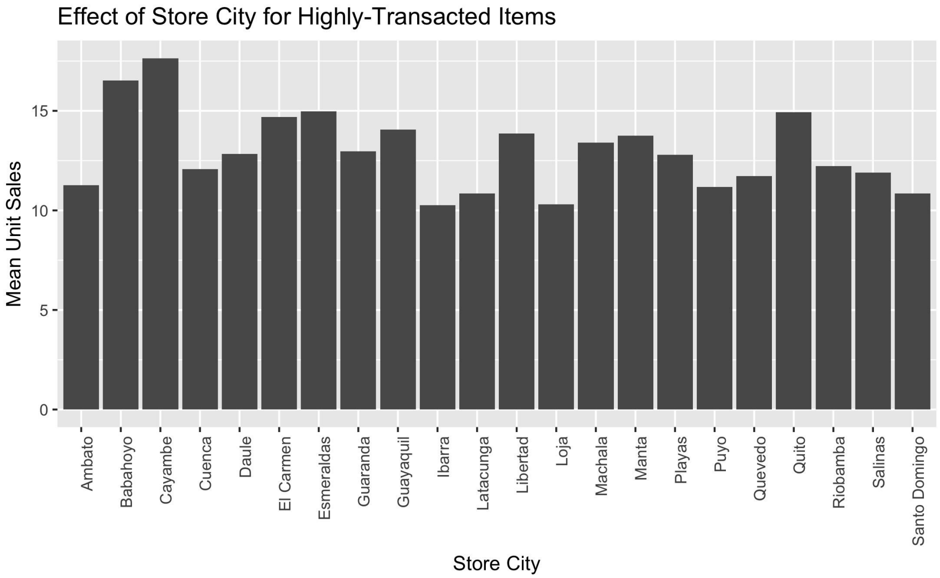


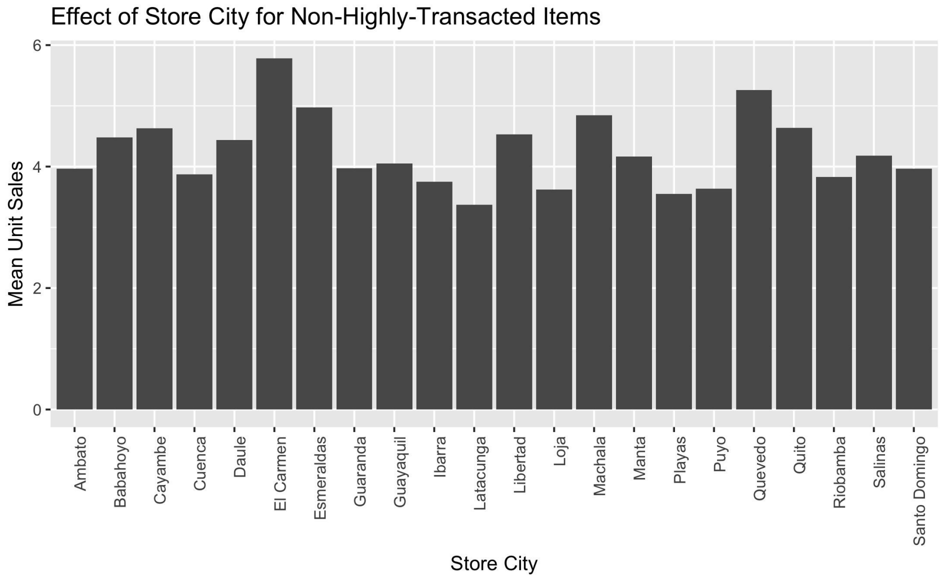
Effect of Item Family for Highly-Transacted Items
Mean Unit Sales
0
10
20
30
BEAUTY
BEVERAGES
BREAD/BAKERY
CELEBRATION
CLEANING
DAIRY
DELI
EGGS
FROZEN FOODS
GROCERY I
GROCERY II
HOME AND KITCHEN I
HOME AND KITCHEN II
HOME CARE
LADIESWEAR
LAWN AND GARDEN
LINGERIE
LIQUOR,WINE,BEER
MAGAZINES
MEATS
PERSONAL CARE
PET SUPPLIES
PLAYERS AND ELECTRONICS
POULTRY
PREPARED FOODS
PRODUCE
SCHOOL AND OFFICE SUPPLIES
SEAFOOD
Item Family



Effect of Item Family for Non-Highly-Transacted Items
Mean Unit Sales
0
3
6
9
AUTOMOTIVE
BABY CARE
BEAUTY
BEVERAGES
BOOKS
BREAD/BAKERY
CELEBRATION
CLEANING
DAIRY
DELI
EGGS
FROZEN FOODS
GROCERY I
GROCERY II
HARDWARE
HOME AND KITCHEN I
HOME AND KITCHEN II
HOME APPLIANCES
HOME CARE
LADIESWEAR
LAWN AND GARDEN
LINGERIE
LIQUOR,WINE,BEER
MAGAZINES
MEATS
PERSONAL CARE
PET SUPPLIES
PLAYERS AND ELECTRONICS
POULTRY
PREPARED FOODS
PRODUCE
SCHOOL AND OFFICE SUPPLIES
SEAFOOD
Item Family

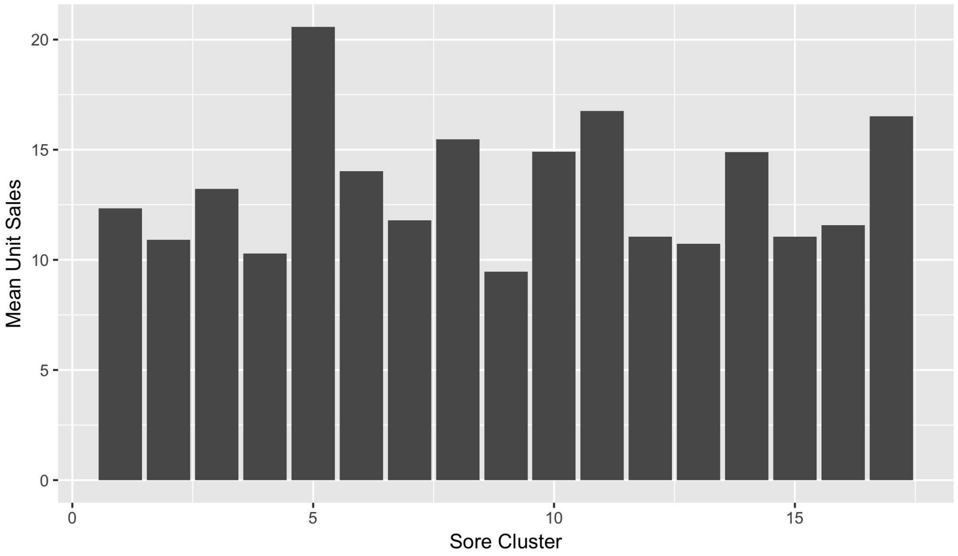
Store City


Effect of Store City for Highly-Transacted Items
Mean Unit Sales
0
5
10
15
Ambato
Babahoyo
Cayambe
Cuenca
Daule
El Carmen
Esmeraldas
Guaranda
Guayaquil
Ibarra
Latacunga
Libertad
Loja
Machala
Manta
Playas
Puyo
Quevedo
Quito
Riobamba
Salinas
Santo Domingo
Store City

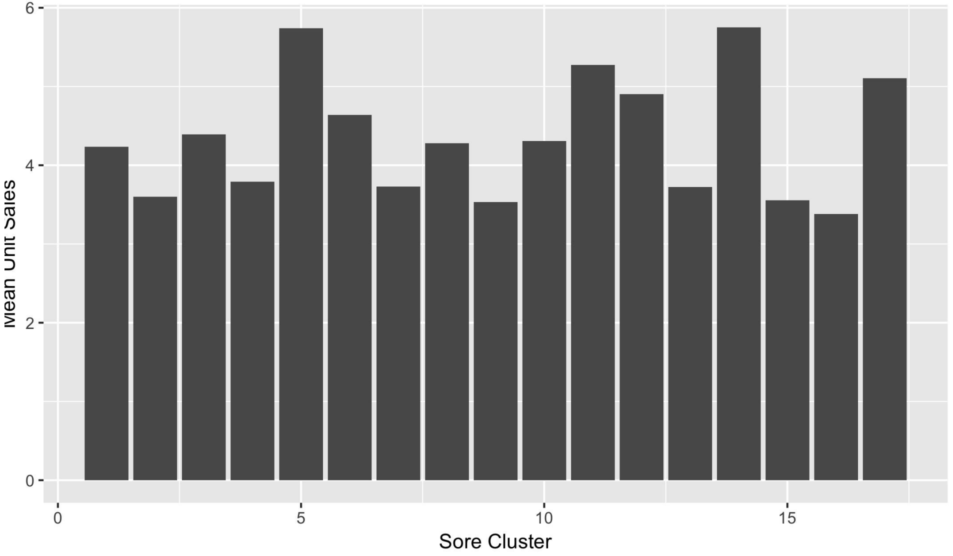

Effect of Store City for Non-Highly-Transacted Items
Mean Unit Sales
0
2
4
6
Ambato
Babahoyo
Cayambe
Cuenca
Daule
El Carmen
Esmeraldas
Guaranda
Guayaquil
Ibarra
Latacunga
Libertad
Loja
Machala
Manta
Playas
Puyo
Quevedo
Quito
Riobamba
Salinas
Santo Domingo
Store City

Store Cluster

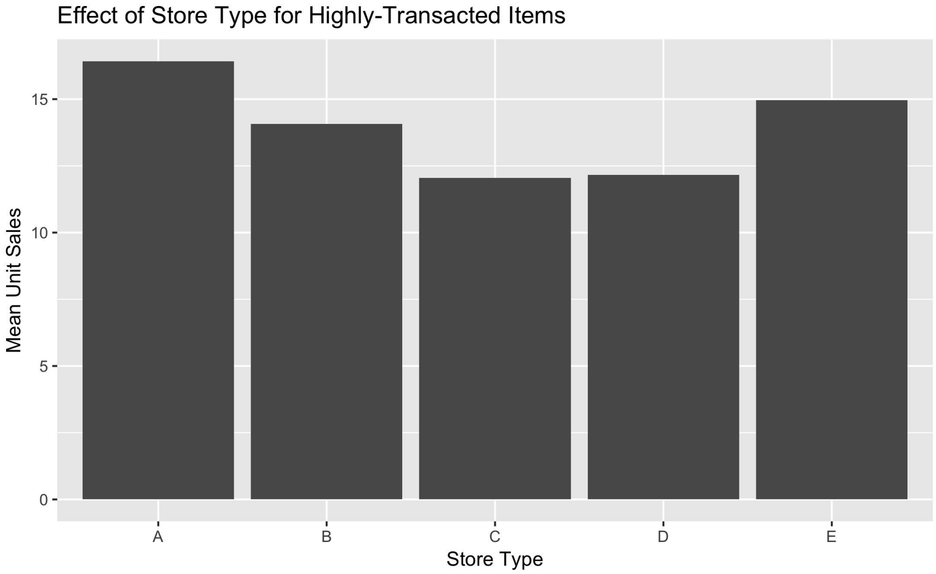
Effect of Store Cluster for Highly-Transacted Items



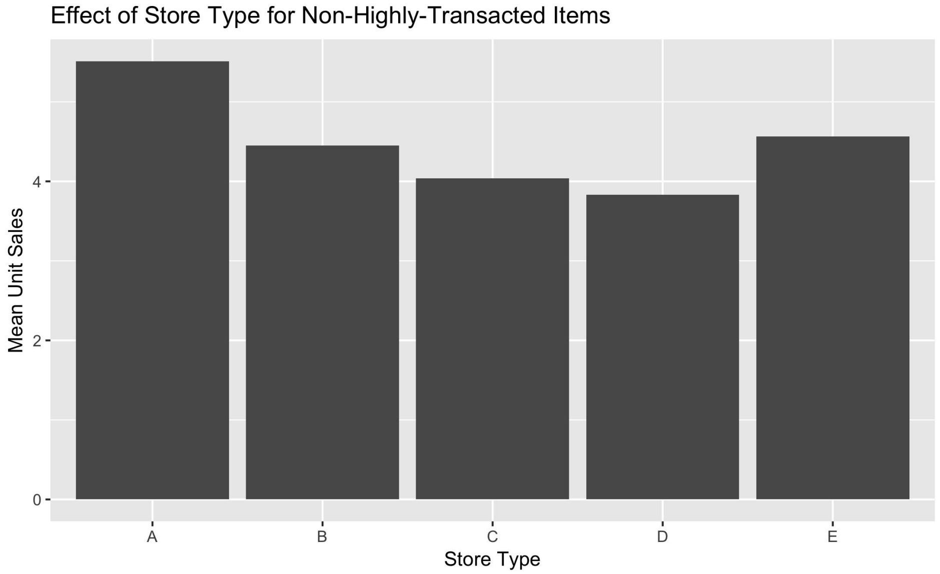
Effect of Store Cluster for Non-Highly-Transacted Items



Store Type

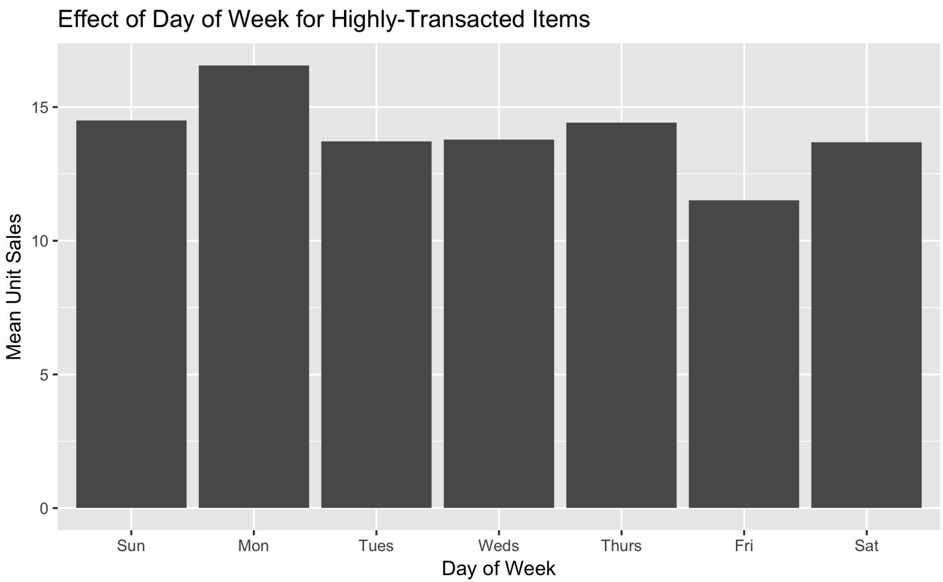


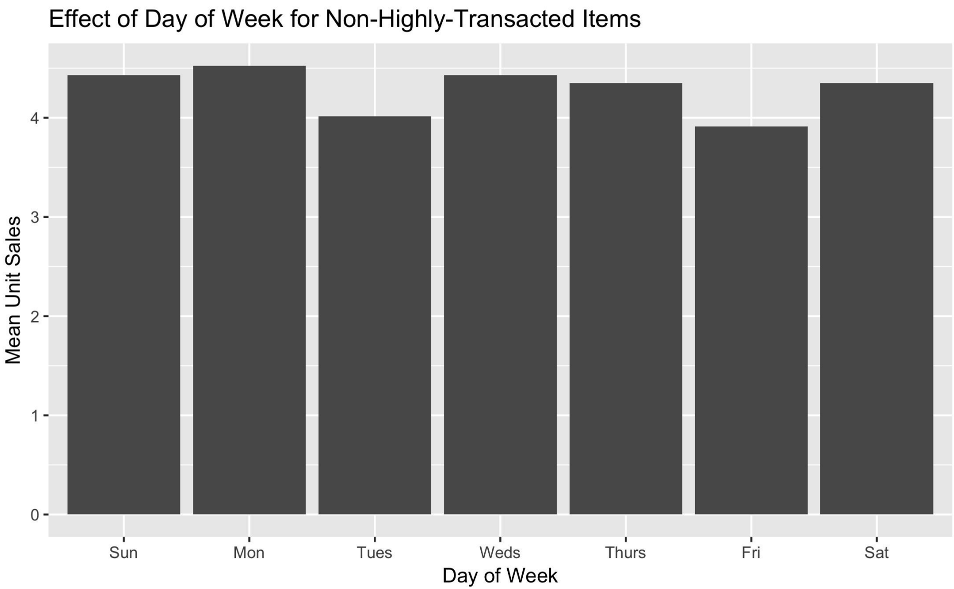
Effect of Store Type for Highly-Transacted Items
Mean Unit Sales
15
10
5
0
A
B
C
D
E
Store Type



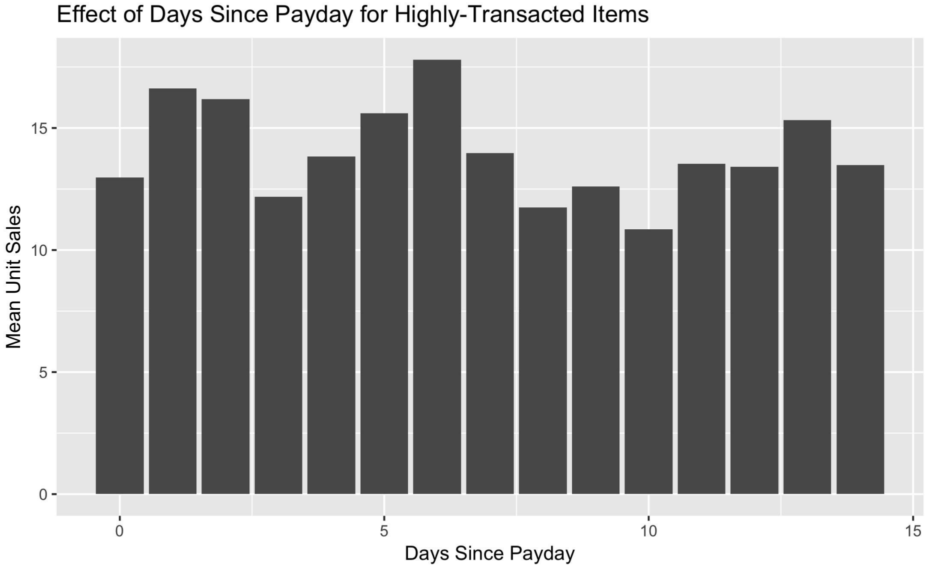
Effect of Store Type for Non-Highly-Transacted Items
Mean Unit Sales
4
2
0
A
B
C
D
E
Store Type

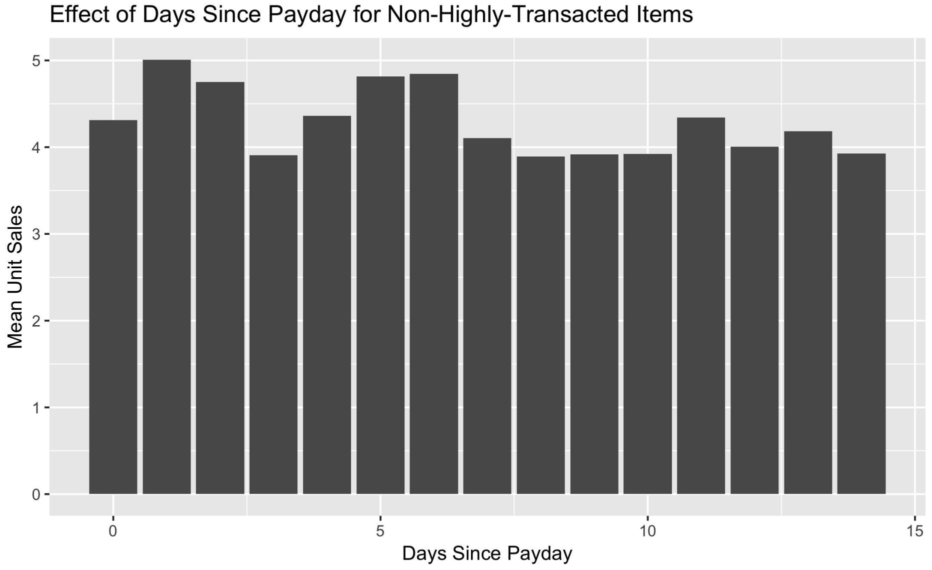
Day of Week


Effect of Day of Week for Highly-Transacted Items
Mean Unit Sales
15
10
5
0
Sun
Mon
Tues
Weds
Thurs
Fri
Sat
Day of Week

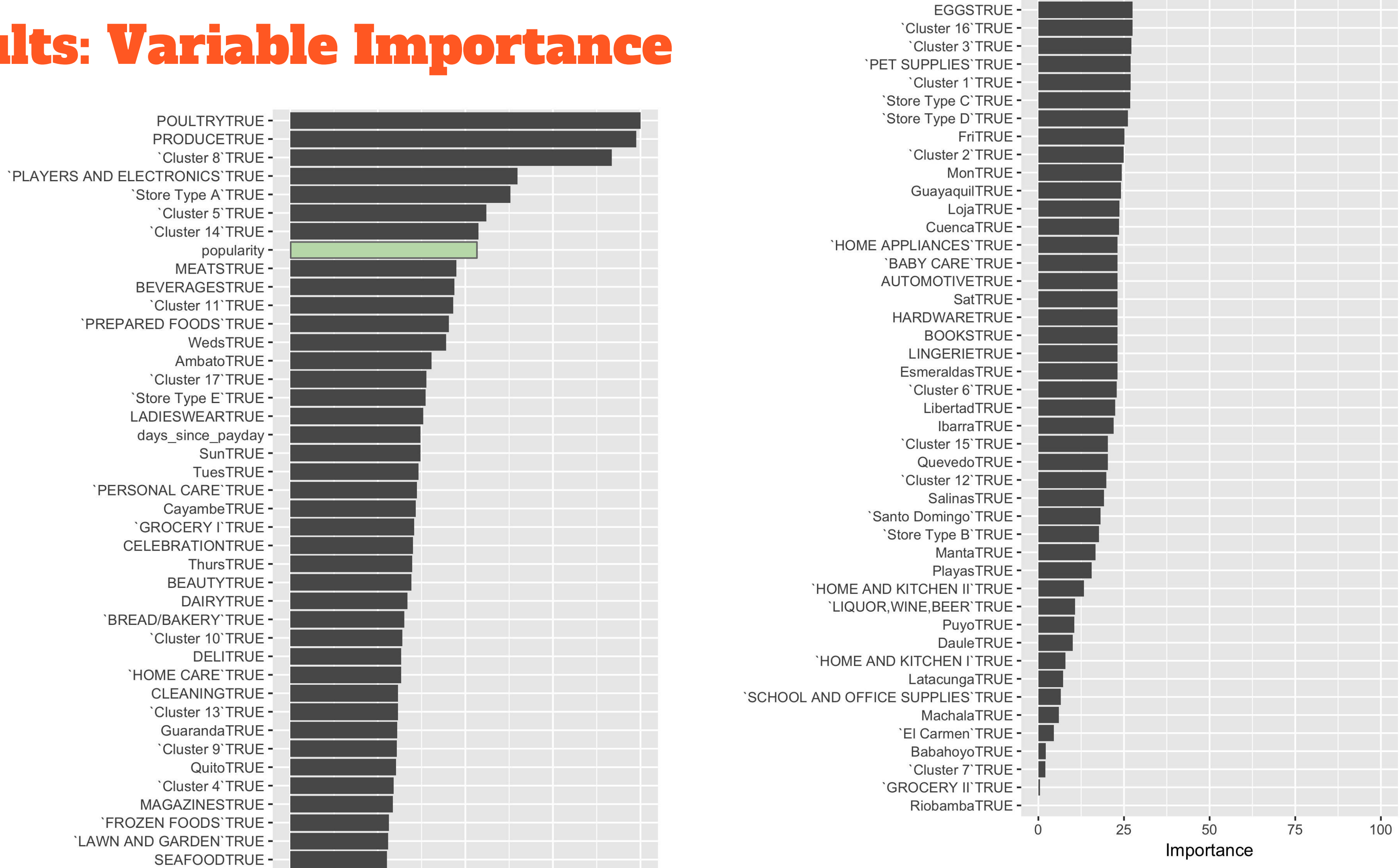

Effect of Day of Week for Non-Highly-Transacted Items
Mean Unit Sales
4
3
2
1
0
Sun
Mon
Tues
Weds
Thurs
Fri
Sat
Day of Week

Days Since Payday

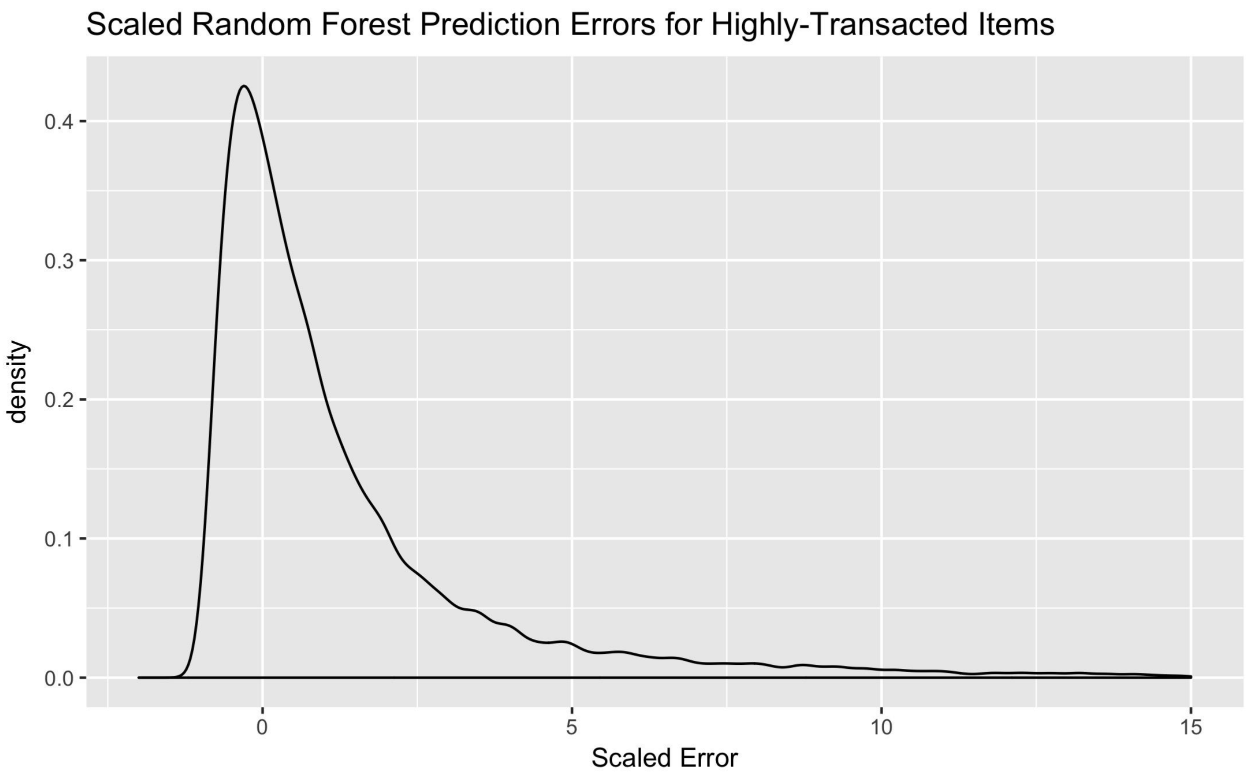




Results: Variable Importance



Results - Prediction Errors



Results: Kaggle

933	▾153	**Matheus Facure**	1.294
934	▾153	**Yosuke Abe**	1.295
935	▾153	**rjuer**	1.295
936	▾153	**Jeffrie**	1.295
937	▾153	**Anjukan Kathirgamanathan**	1.299
938	▾153	**mhaulrich**	1.303
939	new	**Victor de Fefontnouvelle**	1.307
940	▾106	**Magic Logic**	1.309
941	▾154	**tomgrek**	1.310

1036 Teams Total

Potential Improvements

- **0.25% of training set used**
- Add data for no sales
- Incorporate other datasets
  - Holiday
  - Weather
  - Economy