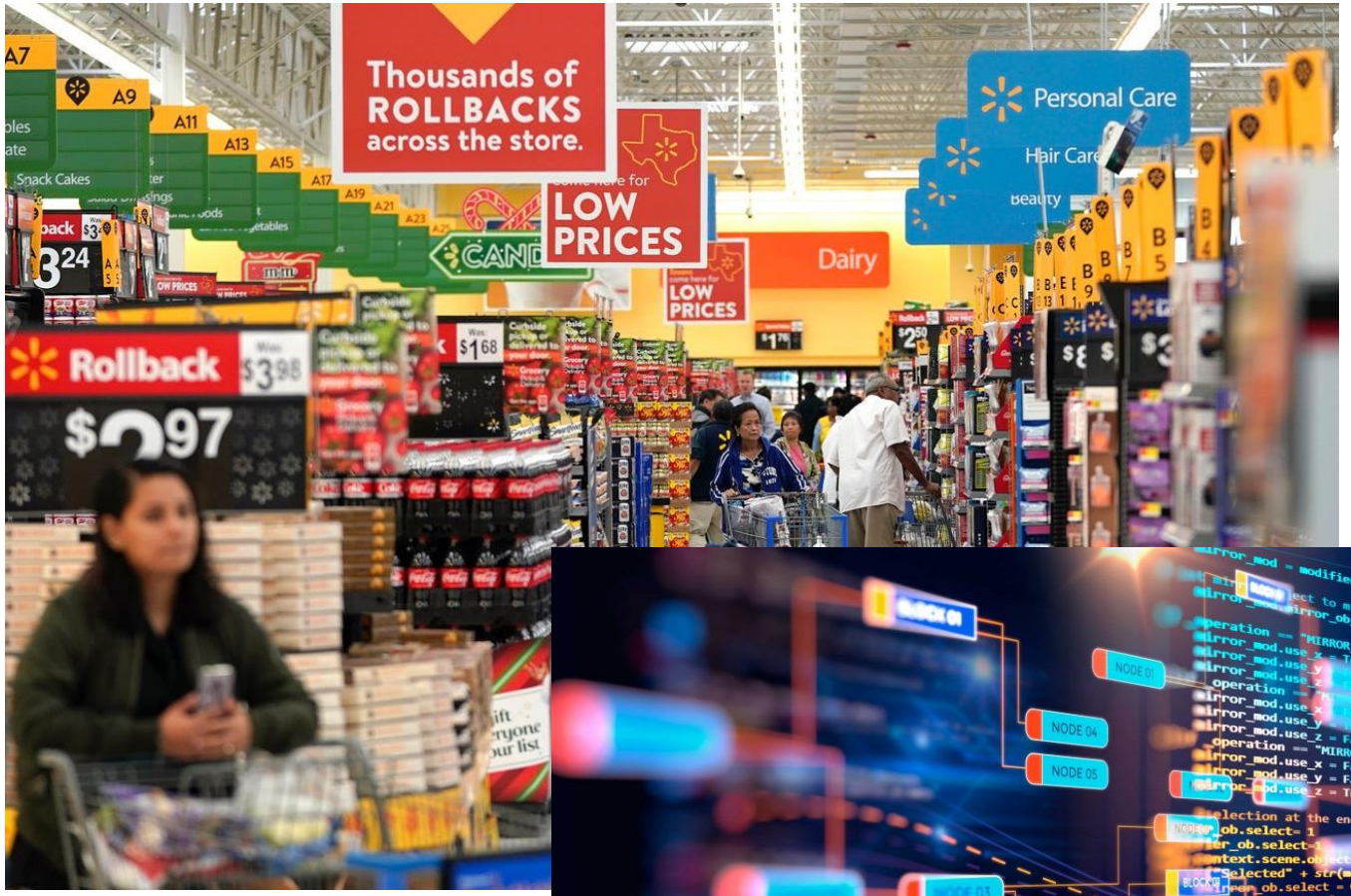
Forecasting Retail Store Sales Using Key Predictors

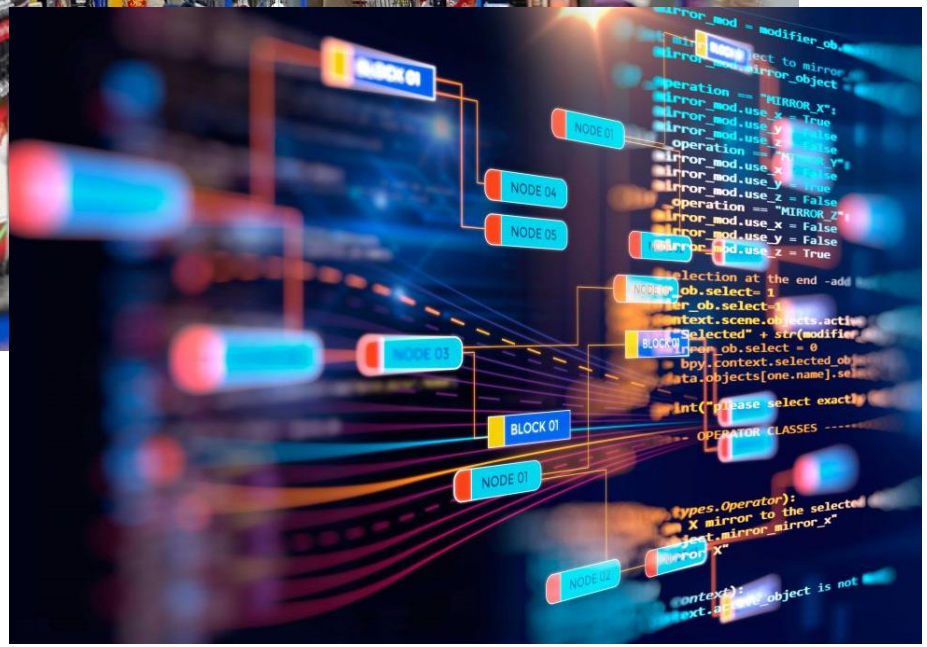
Thomas Jung

Ryerson University

Introduction

- Retail generates Big Data
- Past:
  - Simple linear regressions
  - Humans choose factors
- Present:
  - Complex machine learning algorithms
  - Machines choose most relevant factors

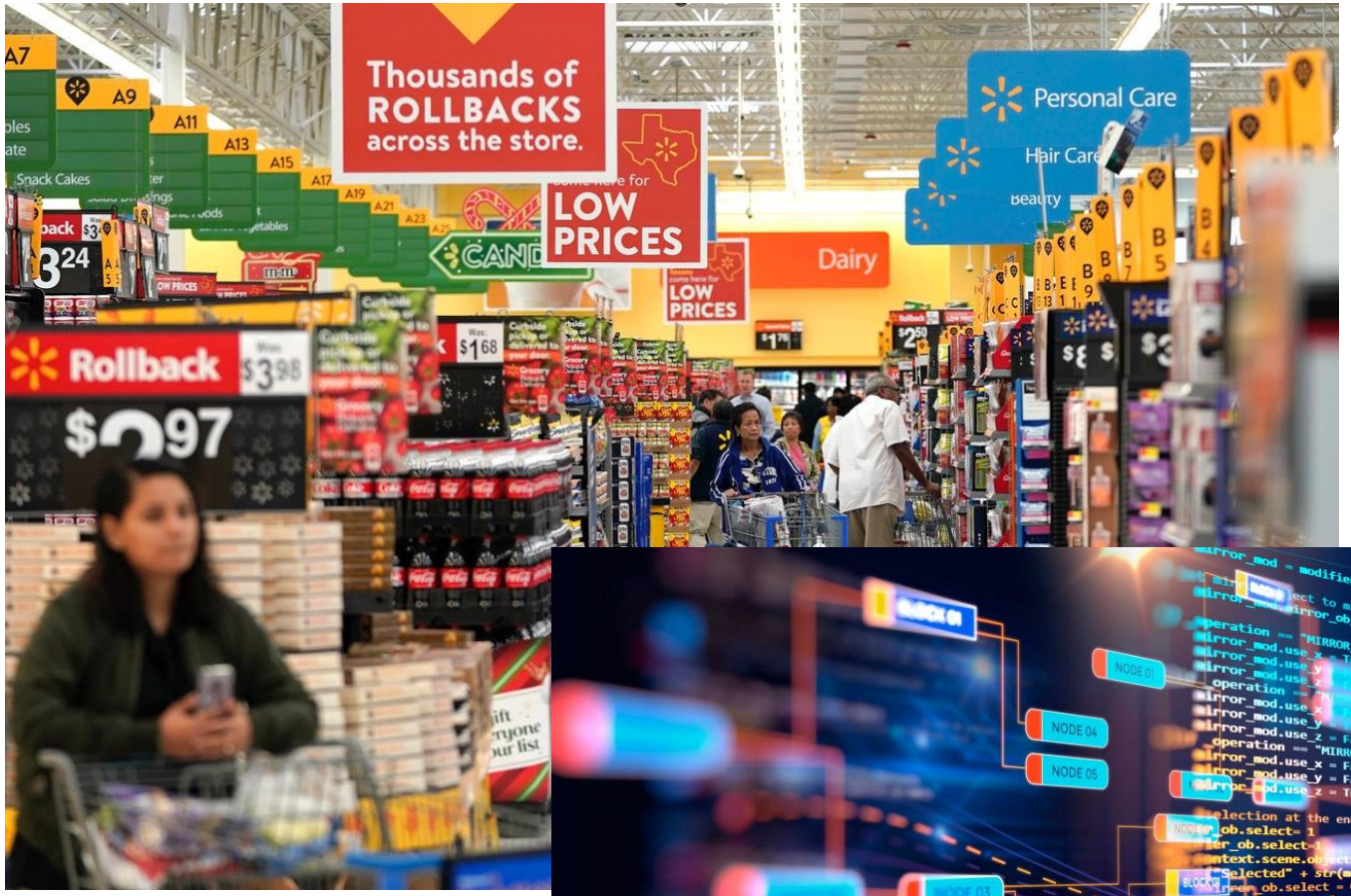


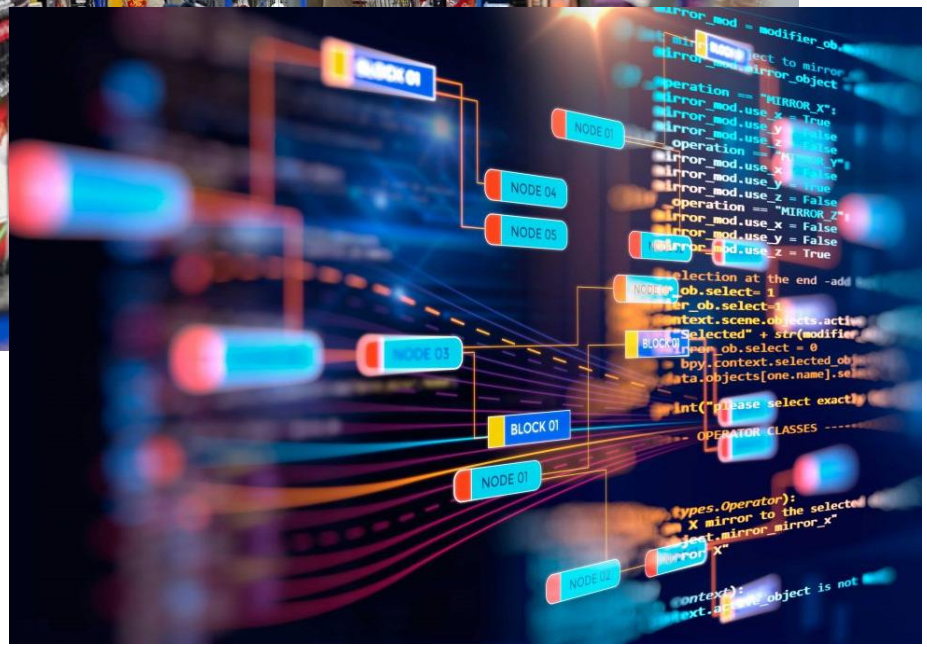


Ryerson University

Objective

- Predict retail sales volumes
- Predictive factors?
  - Historical sales
  - Date, holidays
  - Weather
  - Fuel prices, unemployment rates
- Machine learning algorithms
  - Multiple regression
  - XGBoost
  - Facebook Prophet

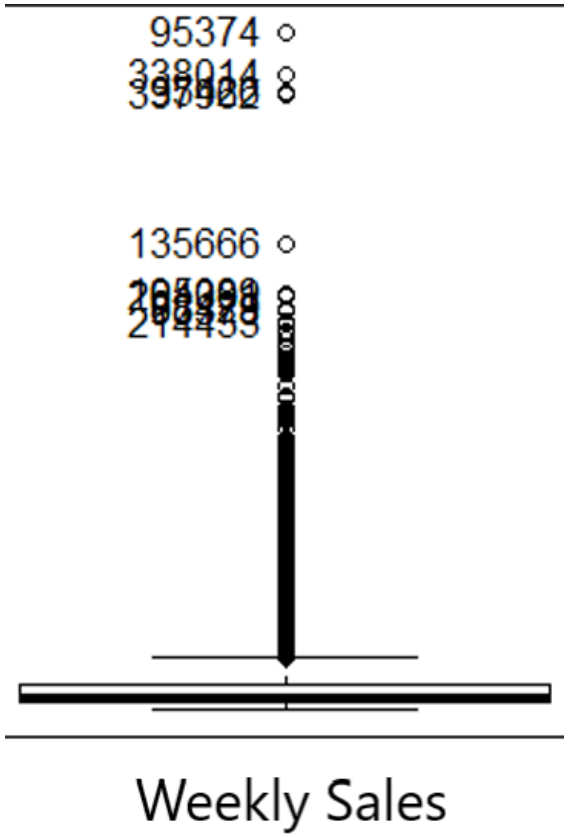




Ryerson University

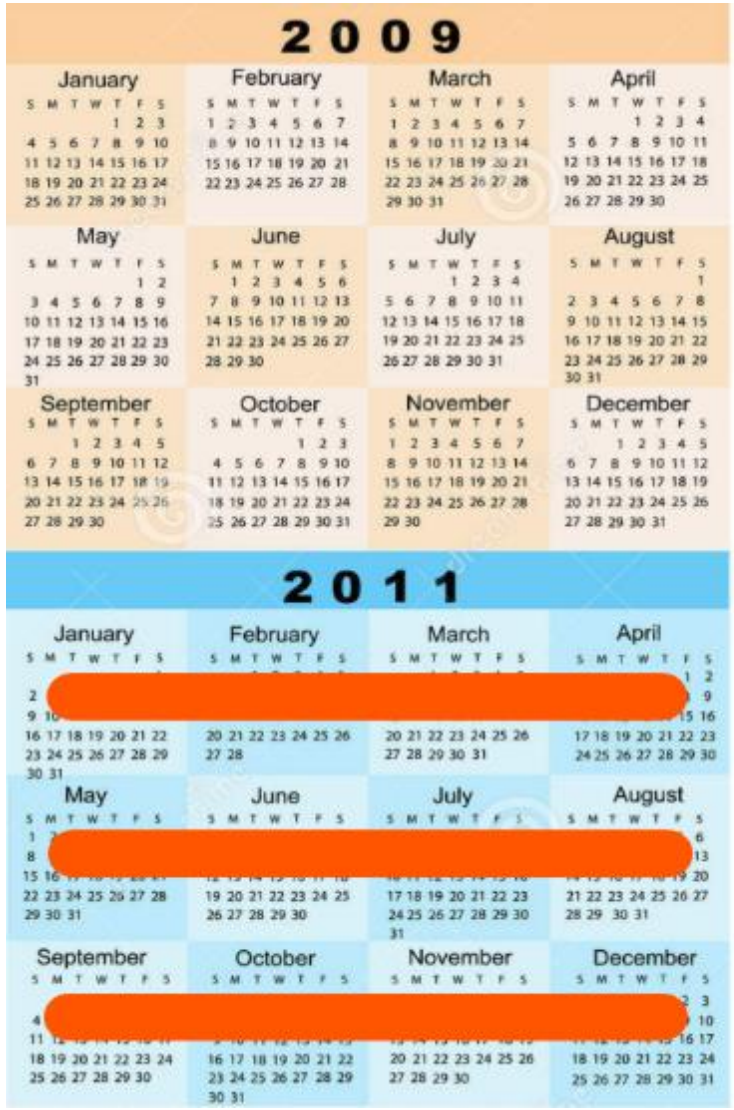
Dataset

File Name	Description	# of Attributes	# of Records
stores.csv	types and sizes of 45 Walmart stores	3	45
train.csv	weekly sales volumes between 2010-02-05 and 2012-11-01	5	421570
test.csv	test set for predicting sales volumes between 2012-11-02 and 2013-07-26	4	115064
features.csv	additional data related to store, department, and regional activity for the given dates including temperature, fuel price, markdowns, consumer price index, unemployment rate, and whether it is a holiday	12	8190



Ryerson University

Dataset



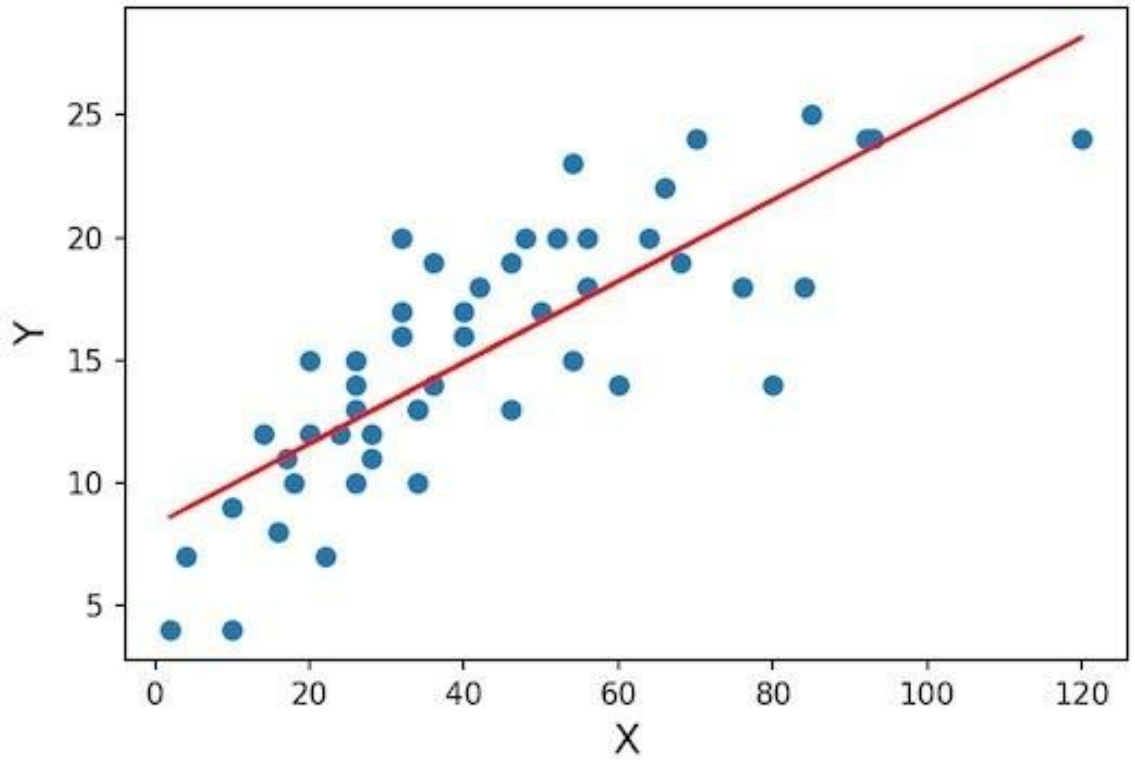
Training set

Test set

Ryerson University

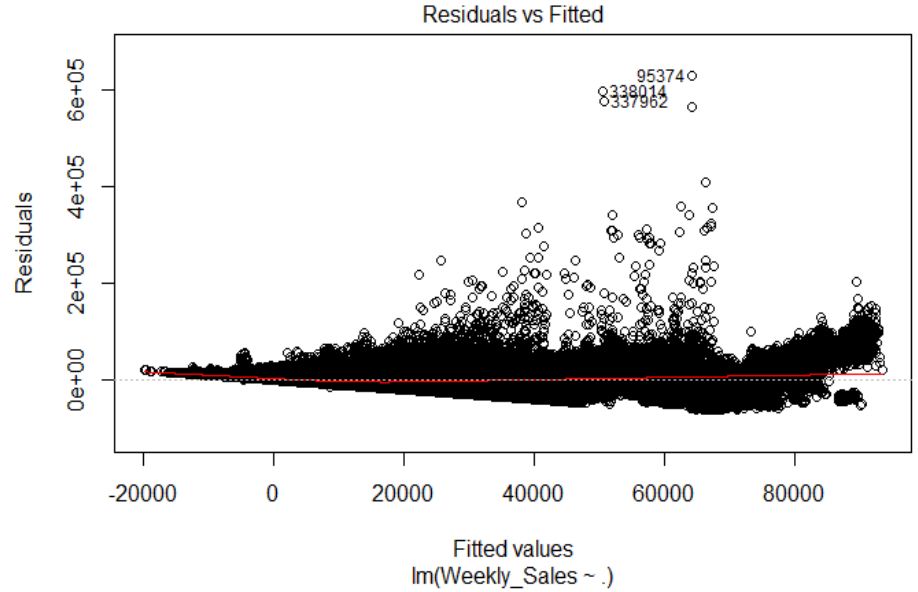
Multiple Regression

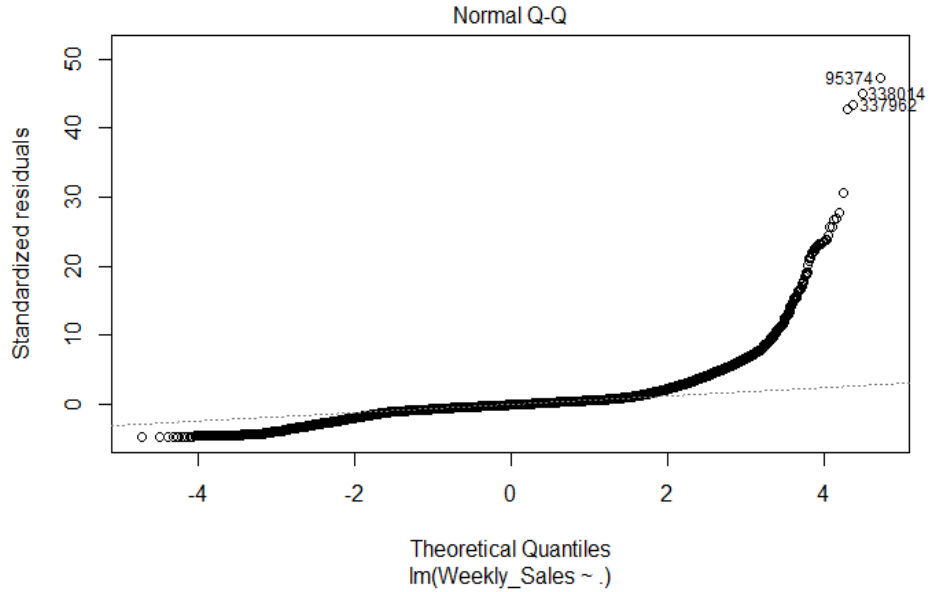
- Fastest to implement
- Parametric; most restrictive
- Assumptions/requirements
  - linear relationships between DV & IVs
  - Residual errors must be normally distributed & homoscedastic
  - No multicollinearity



Ryerson University

Multiple Regression

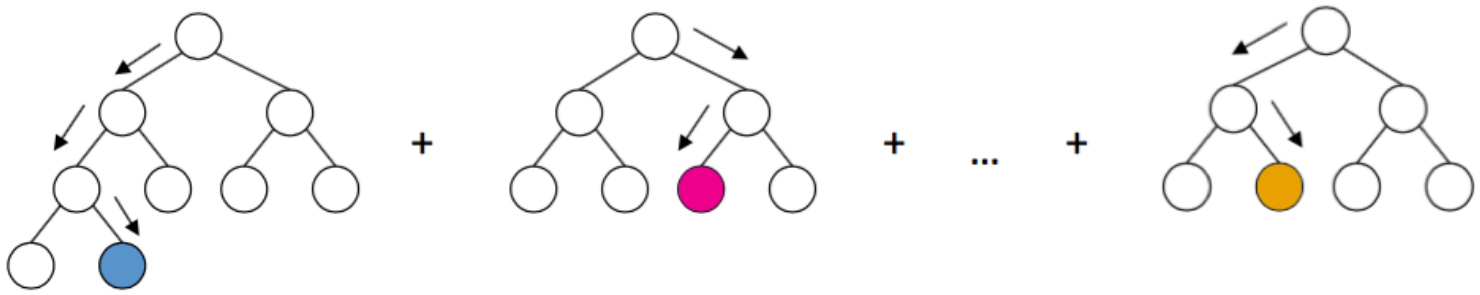




Ryerson University

XGBoost

- Ensemble learning algorithm
- Less restrictive:
  - Non-parametric
  - Robust to outliers & missing values
  - Built-in cross validation & regularization
- Required most time to implement



Ryerson University

Prophet

- Developed by Facebook
- Time series predictions
  - Non-linear trends
  - Seasonality & holiday effects
- Non-parametric
- Ability to add predictive regressors

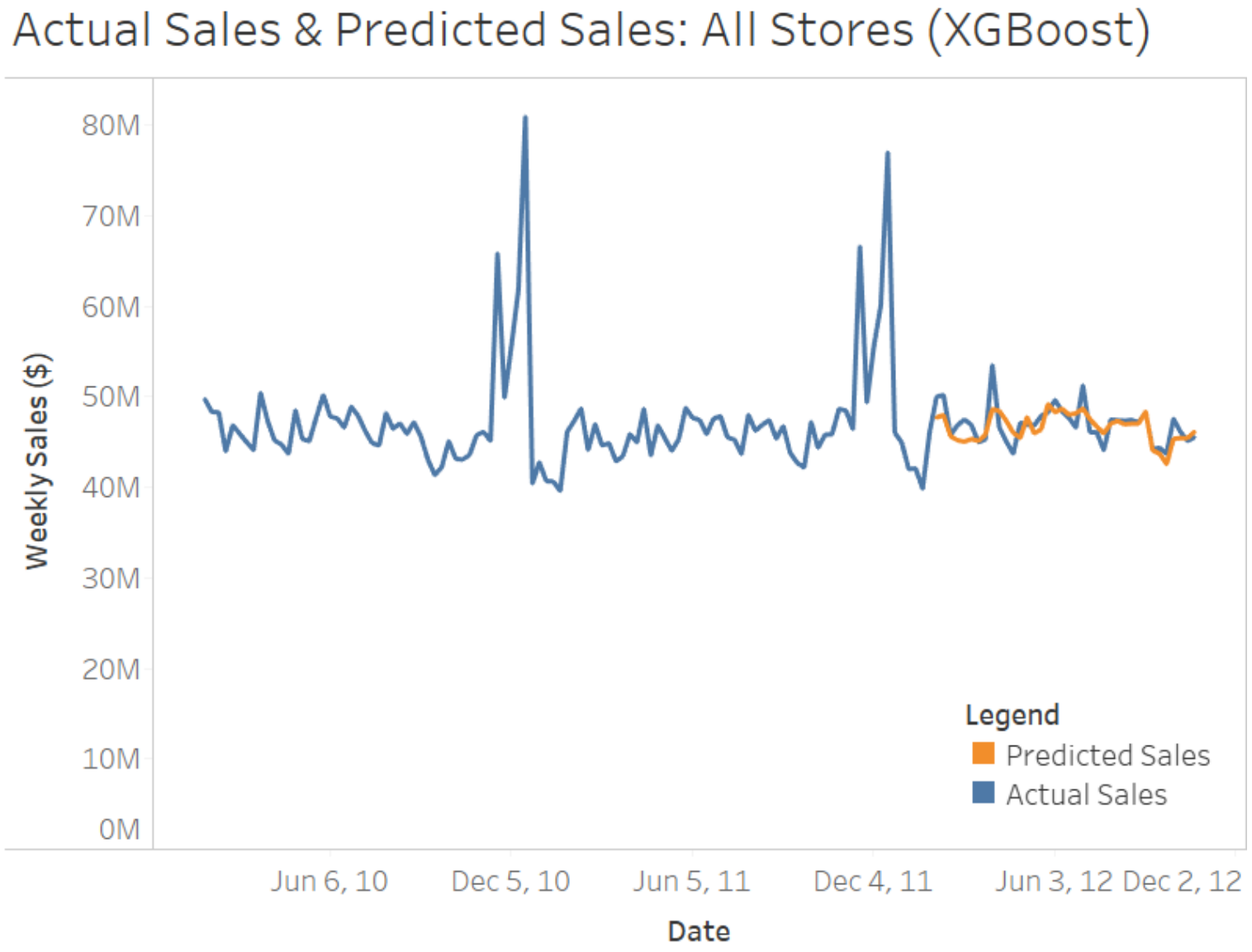
Ryerson University

Results

	RMSE ($)	PRED(30) (% of cases with less than 30% error)
Multiple Regression	12,229	30.7%
XGBoost	4,086	65.9%
Prophet	507,580	42.1%

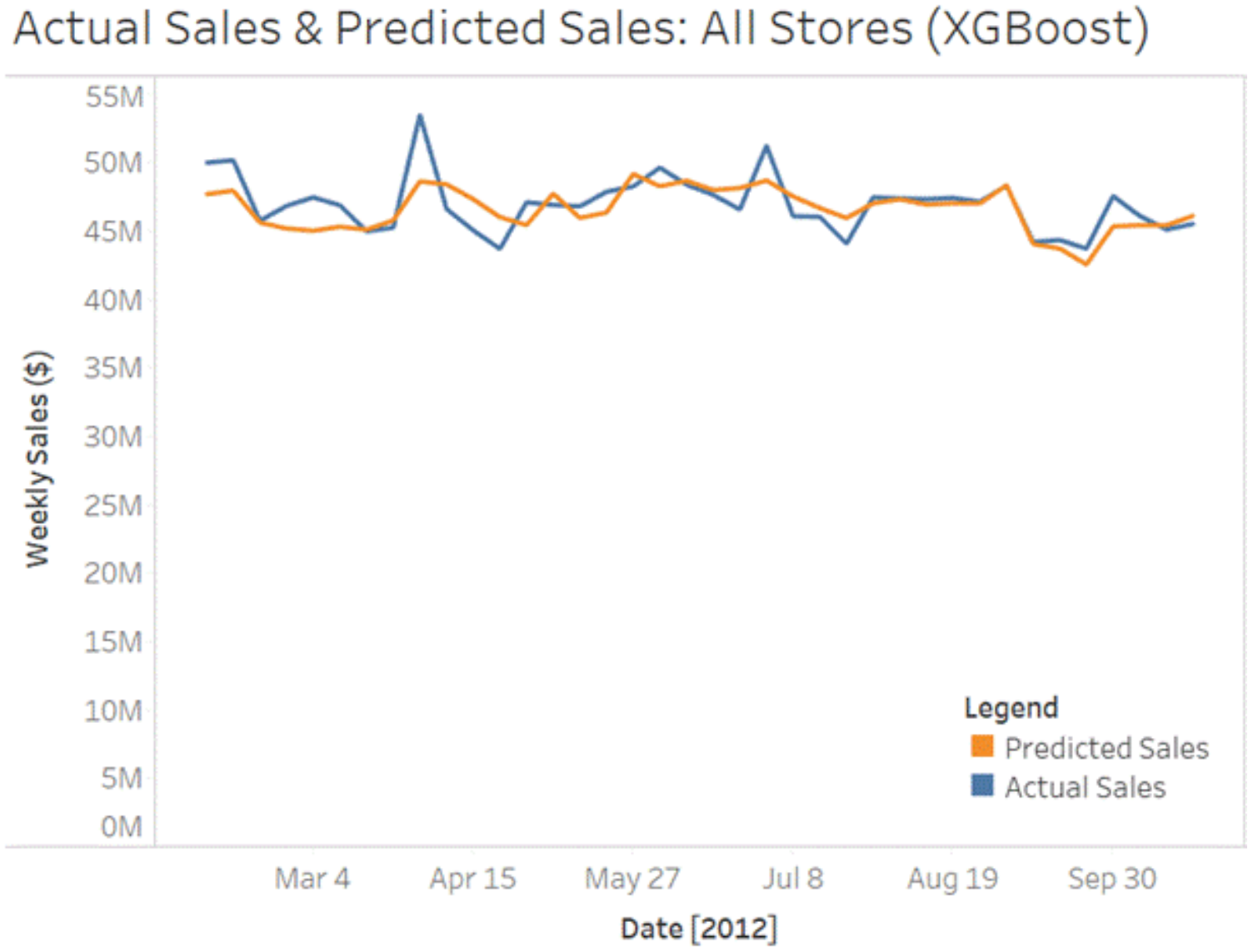
Ryerson University

Results: XGBoost



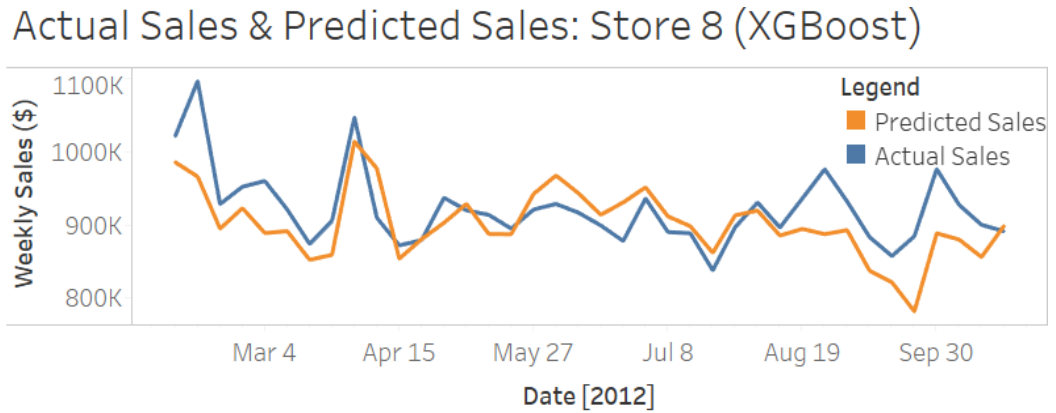
Ryerson University

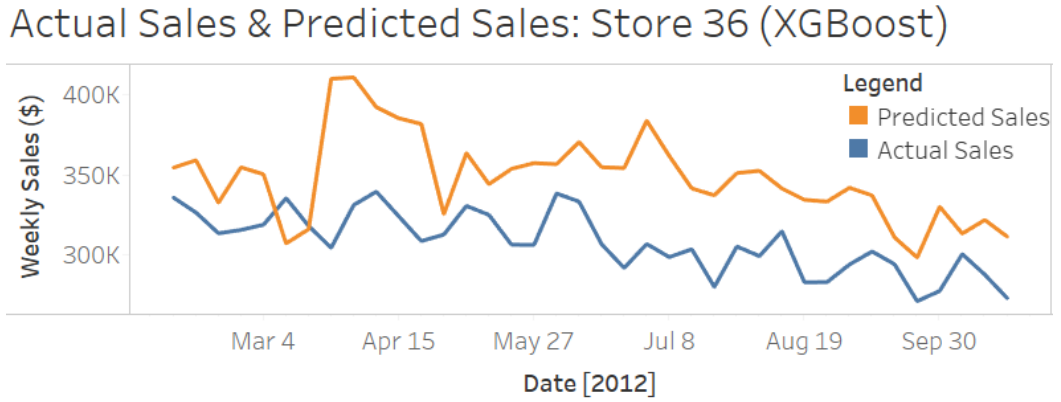
Results: XGBoost

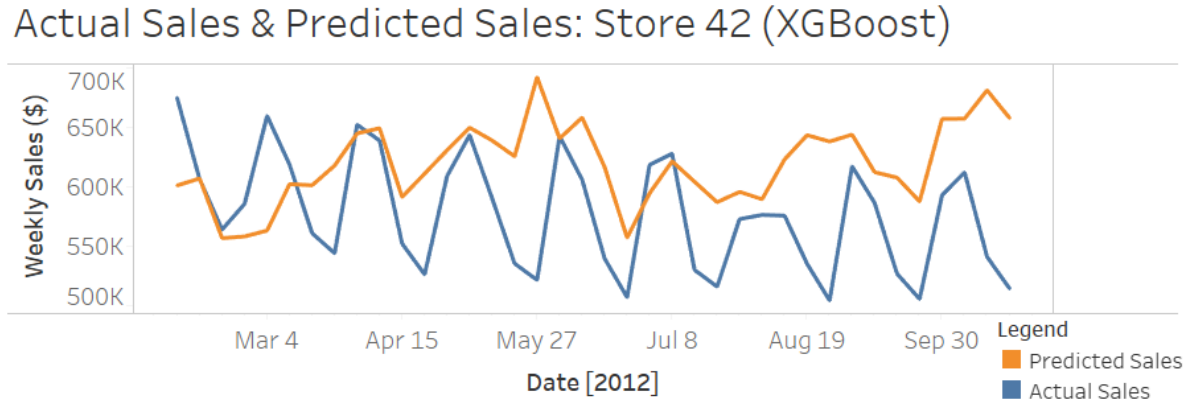


Ryerson University

Results: XGBoost

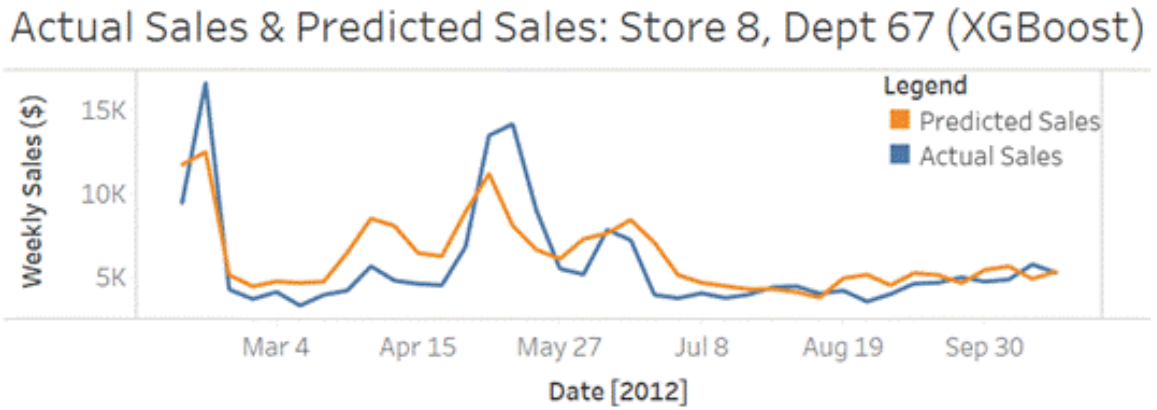






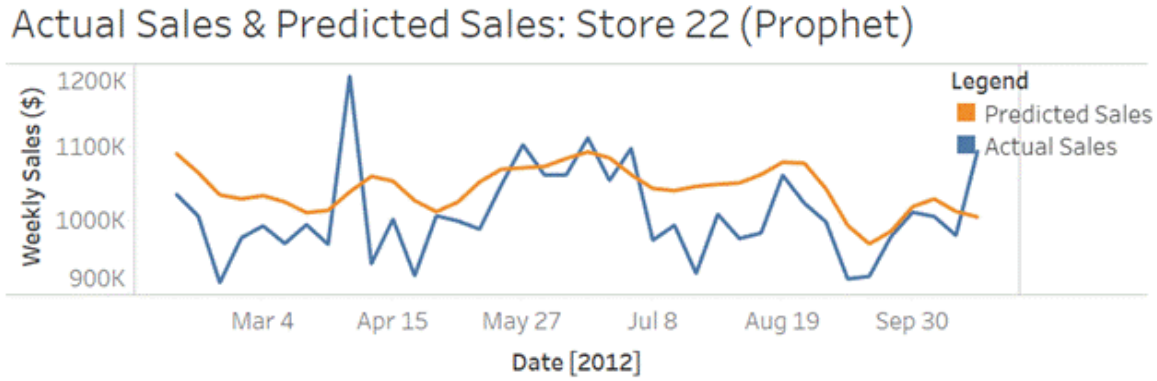
Ryerson University

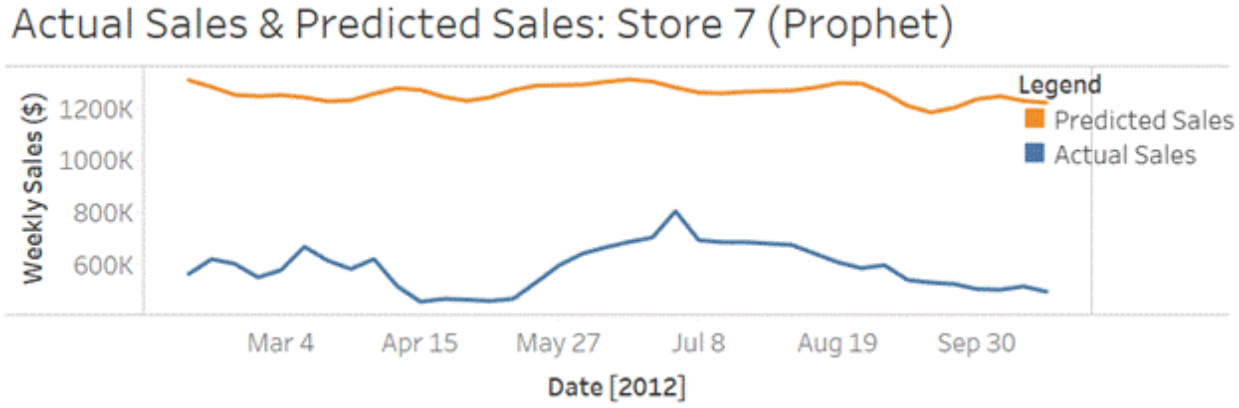
Results: XGBoost



Ryerson University

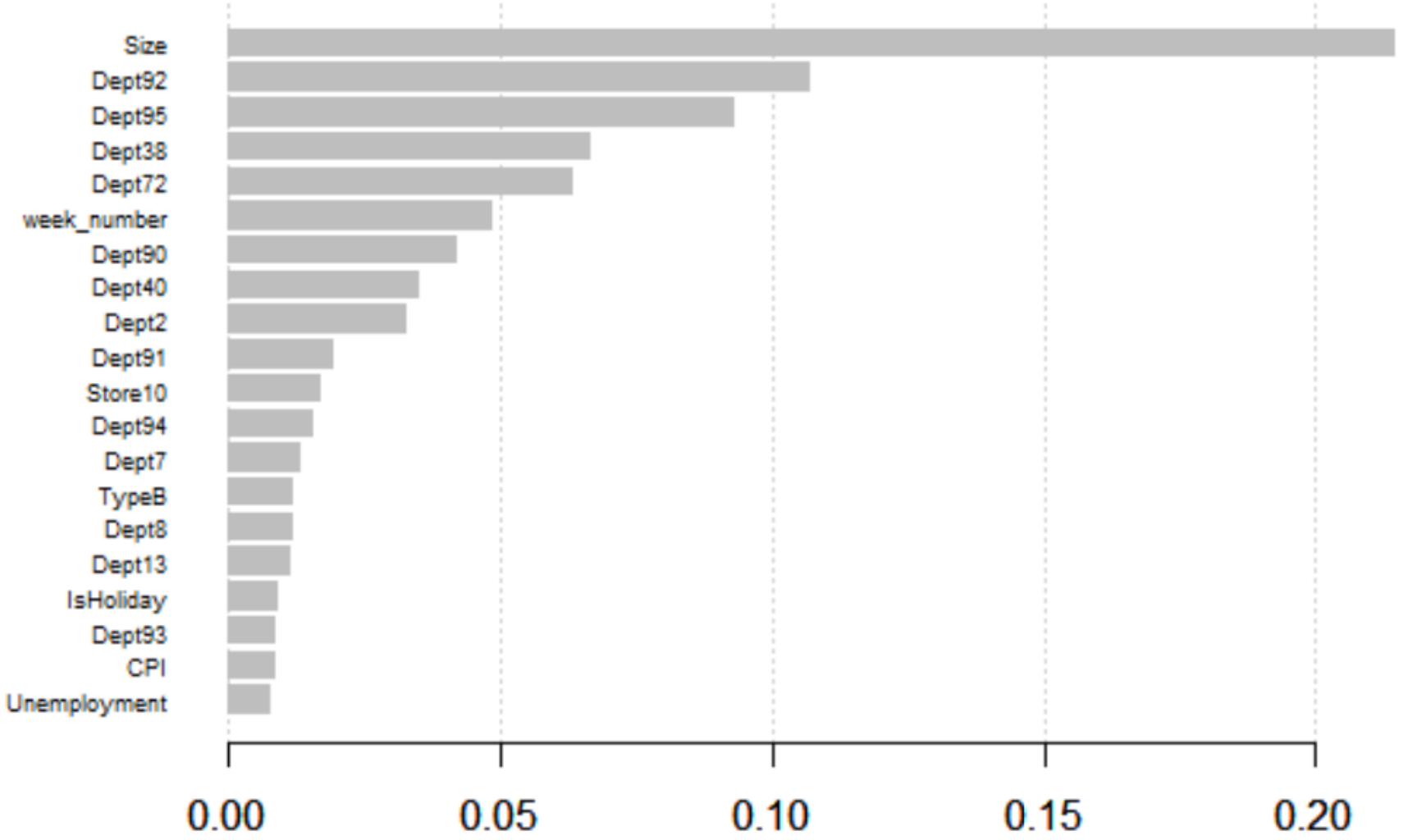
Results: Prophet





Ryerson University

Importance Matrix (XGBoost)



Ryerson University

Conclusion

- XGBoost was the best algorithm for predictions
- Enterprise-wide level:
  - 65.9% of predictions had less than 30% error
- Store-wide level:
  - Larger store predictions more accurate than smaller stores
- Department-wide level:
  - Small- to medium-sized departments more accurate
- Large spikes in sales
  - Captured most of the time in terms of date
  - Not fully captured in terms of magnitude

Ryerson University

Conclusion

- Hypothesis: weather, holidays, fuel prices, and unemployment rates can be used as predictors
- Results: these factors play a minor role
- Most predictive power
  - Size of the store
  - Store number
  - Department number
  - Week number

Ryerson University

Thank you.

Questions?

Ryerson University