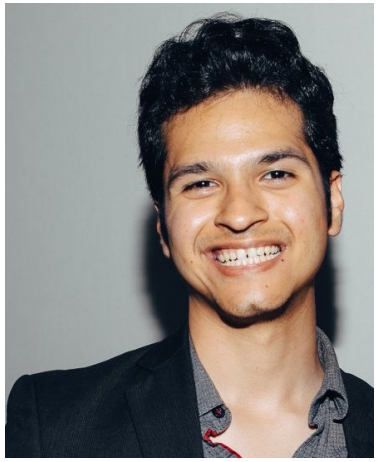
Wharton Customer Analytics Datathon

February 5, 2021

Omar Nunez (W'21), Rachel Levin (W'21), and Alana Levin (W'21)

The Team

Three Wharton seniors who love all things data



Omar Nunez

- Wharton ‘21
- Majors: Statistics & Finance
- Engagement Manager for MBAs & UGs @ Venture Lab

Rachel Levin

- Wharton ‘21
- Majors: Statistics & Finance
- Minor: Math
- President of Wharton UGR Data Analytics Club
- TA for STAT 471, OIDD 101

Alana Levin

- Wharton ‘21
- Majors: Finance and Public Policy
- President of Wharton Global Research & Consulting

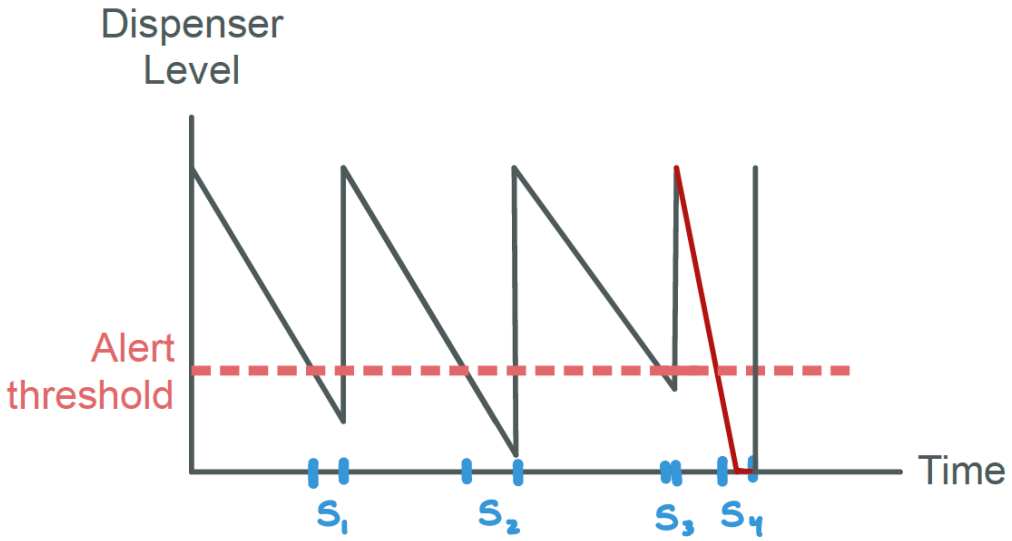
Executive Summary

We explored a variety of modeling techniques to extract insights from Essity's Tork data sets

- Theme: Claims and Product
- Data Exploration
- Prediction methodologies:
  - Stepwise / Multiple Linear Regression
  - Elastic Net
  - Classification Tree
- Recommendation

Project Motivation

Reduce number of times when dispenser is both out-of-stock and consumers expect it to be stocked

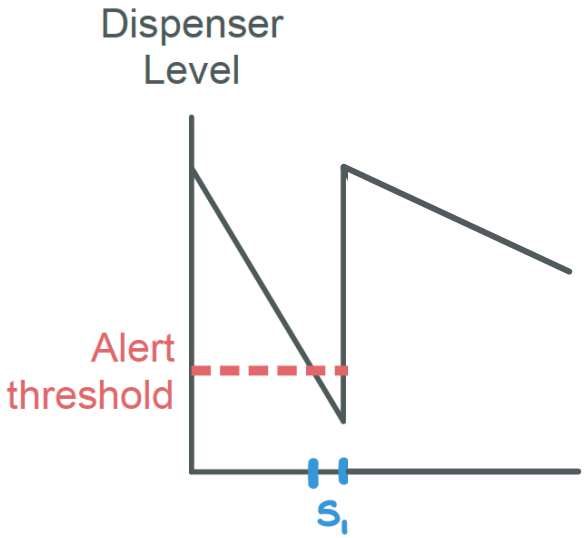


Service time s is a random variable
Assume s independent of consumption velocity

High-traffic days may cause dispensers to run out of stock when people need them the most

Project Motivation

Reduce number of times when dispenser is unnecessarily stocked



Service time s is a random variable
Assume s independent of consumption velocity

During low-traffic days, stocked dispensers are inefficient — they have low utilization

Evidence in the Data

To illustrate the data, we built what we call:

The Essity Utilization Matrix™

Traffic days	Status: No Soap	Status: Soap
Below-average	Sweet spot	Low util.
Above-average	Bad service	Sweet spot

Evidence in the Data

Case Study: Site 32

Traffic days	Status: No Soap	Status: Soap
Below-average	76%	24%
Above-average	52%	48%

Game plan

Reach for the sweet spots by dynamically shifting 'empty_level' on scu_df

Key Metric: Daily Traffic

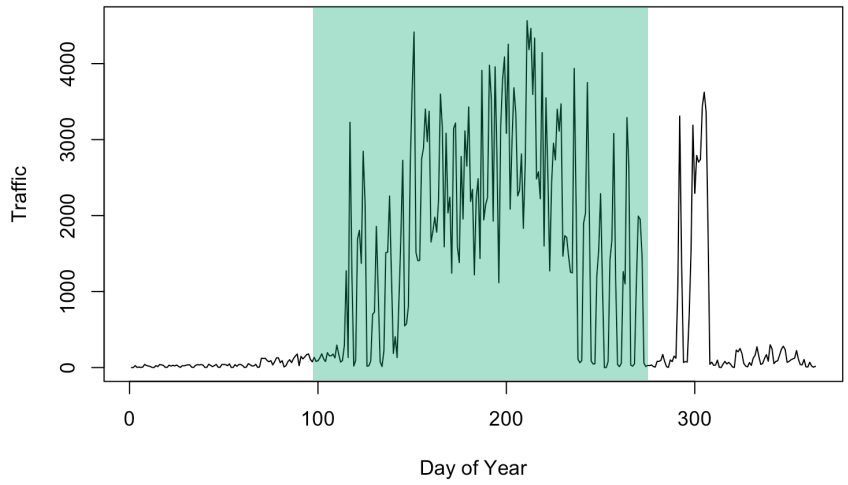
Data Preparation

```
dailyCountStart <- aggregate(x = loc_ppl_df[c("counter_in")],
                  by = loc_ppl_df[c("LocationId", "PeopleCounterId", "Year", "Month", "Day")],
                  FUN = min, na.rm = TRUE)

dailyCountEnd <- aggregate(x = loc_ppl_df[c("counter_in")],
                  by = loc_ppl_df[c("LocationId", "PeopleCounterId", "Year", "Month", "Day")],
                  FUN = max, na.rm = TRUE)

dailyCount_df["dailyTraffic"] <- dailyCount_df["endCount"] - dailyCount_df["startCount"]
```

Location 702 Daily Traffic



Prediction Methodology

Data Preparation

- Data set for analysis: dailyCount_df
  - 5,630 rows
  - Features of interest: LocationId, dailyTraffic, dailyTrafficLag, weekDay, and standardized versions of continuous variables
  - Excluded 7 outliers and rows where dailyTrafficLag = NA (i.e., no record of previous day)
- Split data into train and test sets (80:20)
- Conduct a range of regression and classification analyses

Prediction Method #1

Stepwise / Multiple Linear Regression

- Regression Framework
  - Response: dailyTraffic.scale (scaled)
  - Covariates: LocationId, time, time2, time3, dailyTrafficLag1.scale, weekDay, and interactions
- Conclusion: These predictors moderately explain variation in dailyTraffic. The linear model provides a starting point for future feature addition.

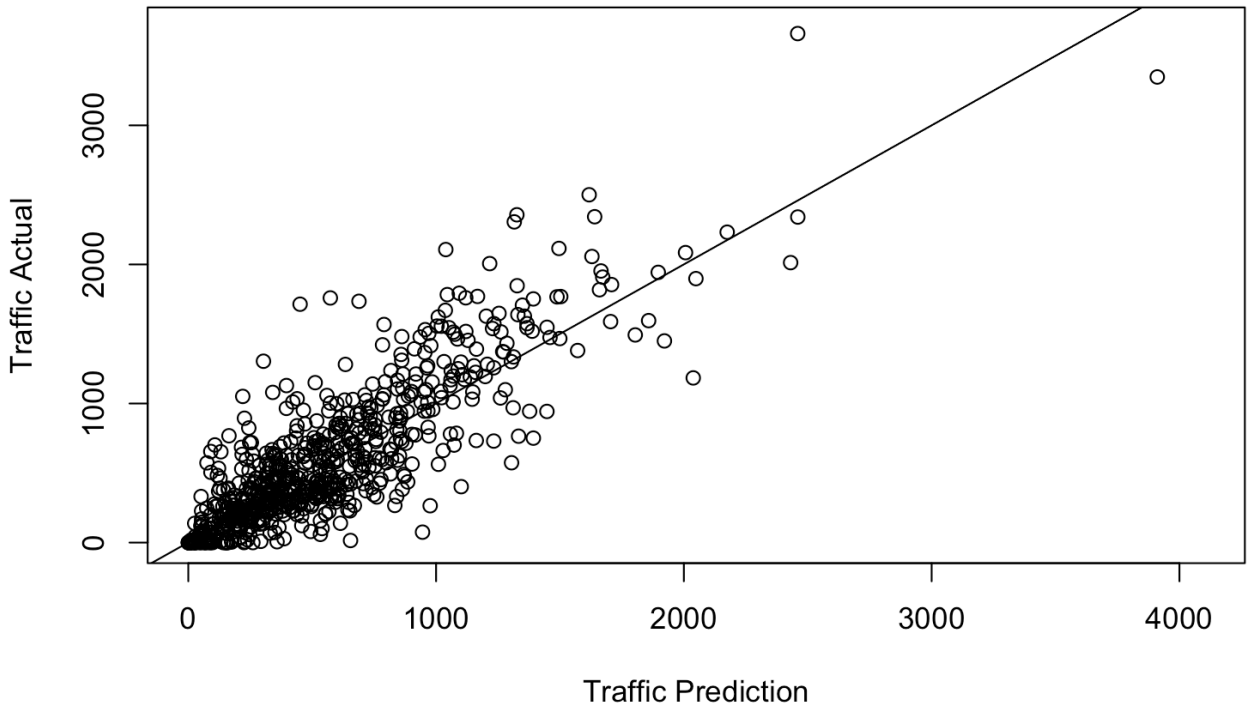
Predictive Performance

Out-of-sample R^2: 73%
Mean Abs. Error: 203

$$\sqrt{dailyTraffic} = \sum_{i=1}^{p} \beta_i \cdot LocationId_i + \left(\sum_{n=1}^{3} \beta_{p+n} \cdot time_n \right) + \beta_{p+4} \cdot \sqrt{dailyTraffic_l} + \cdots + \beta_{p+5} \cdot weekday + interactions + \varepsilon$$

Prediction Method #1

Multiple Linear Regression



Prediction Method #2

Penalized Regression: Elastic Net

- Regularization framework
  - Response: dailyTraffic (continuous)
  - Covariates: LocationId,

    dailyTrafficLag, and weekDay
- Cross-validated parameters: α = 0.1, λ = 0.35
- Poor out-of-sample performance
- Conclusion: dailyTraffic is not well-explained by the covariates in the model

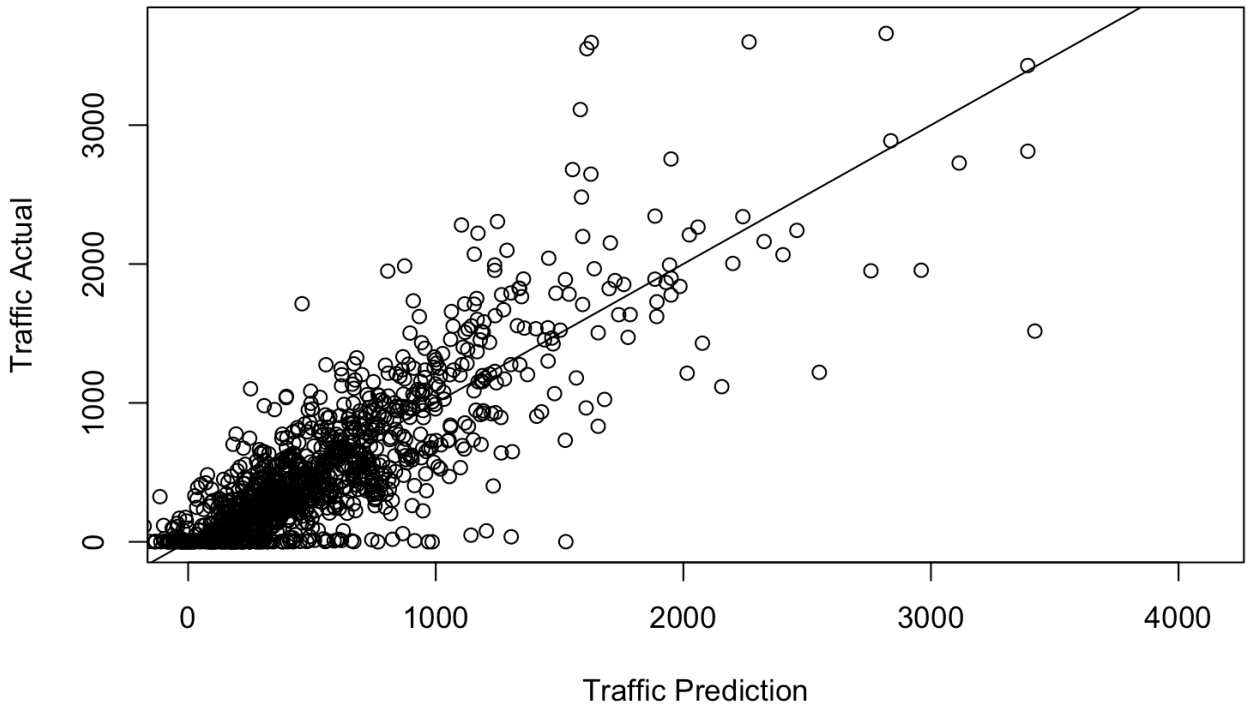
Predictive Performance

Out-of-sample R^2: 72%
Mean Abs. Error: 218

$$\frac{\sum_{i=1}^{n}(y_i - x_i^J \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{m} |\hat{\beta}_j| \right)$$

Prediction Method #2

Penalized Regression: Elastic Net



Prediction Method #3

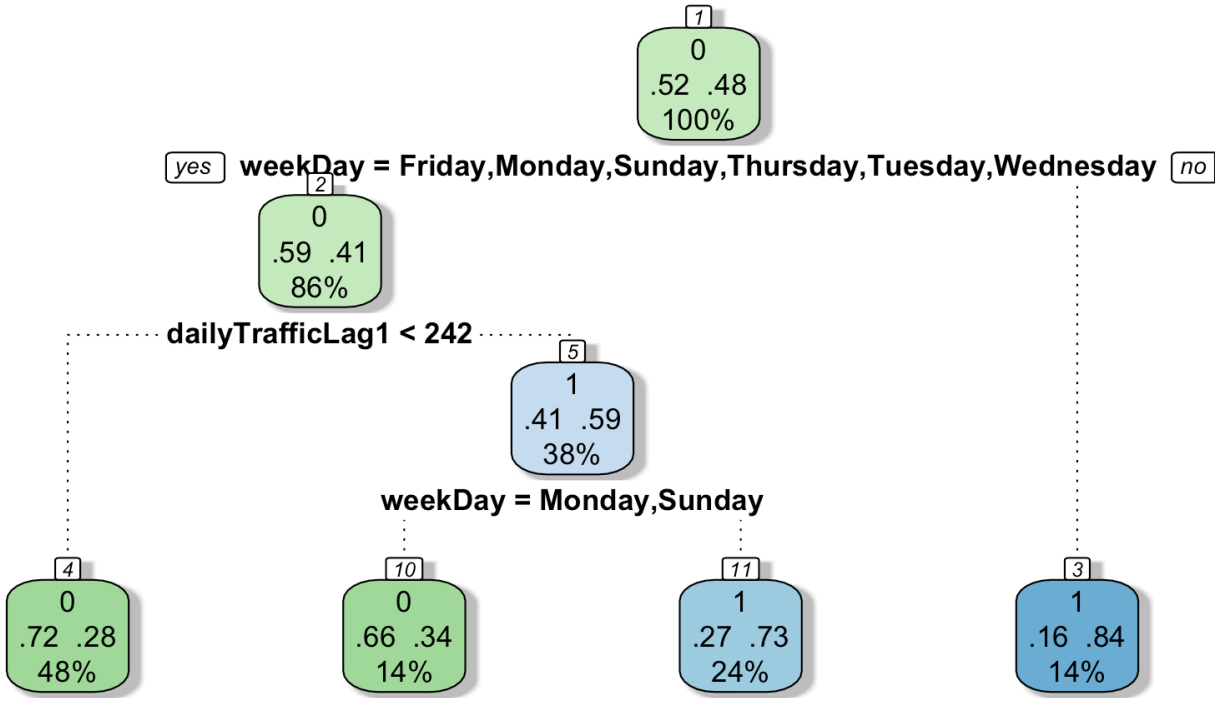
Classification Tree

- Feature engineering: TrafficAboveAvg (binary)
  - If daily traffic at a given location on a given day is above average (relative to that location's performance)
  - Covariates: LocationId, dailyTrafficLag, weekDay

Out-of-Sample AUC
0.7439

Prediction Method #3

Classification Tree



Out-of-Sample AUC

0.7439

Recommendation

- Classification tree advantages:
  - Easier to implement in the business setting
  - More interpretable
  - Flexible and conducive to tuning

Final Thoughts

- What's next? Our analyses enable Essity to…
  - Expand daily traffic modeling to all 15 major customers
  - Compare claims and wastage relationships among clients
  - Map associations between subscription type, sensor data, and traffic patterns
  - Present easily interpretable classification results to management
  - Use the Essity Utilization Matrix to minimize "Bad service" instances and maximize the "Sweet spots"

Thank You!