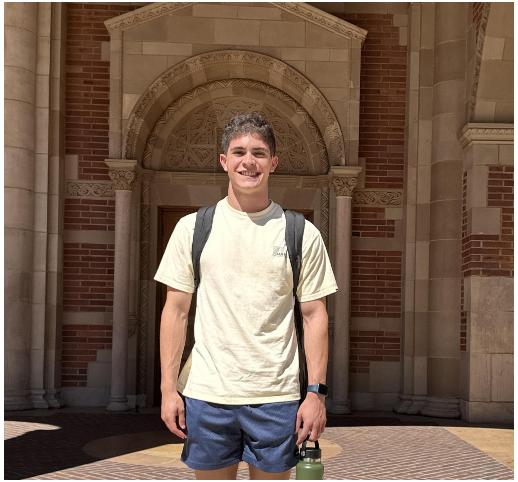
Stats M148 Capstone

Vanessa Chan & Jason Clark

UCLA College | Physical Sciences Statistics & Data Science

fingerhut

Who are we?



Jason Clark (he/him/his)

Major: Data Theory

Graduation: Fall 2024

Vanessa Chan (she/her/hers)

Major: Data Theory

Graduation: Fall 2024

Overview

1. Predictive Modeling for Incomplete Journeys
2. Forecasting the Number of Orders Shipped
3. Forecasting the Number of Started Journeys
4. Promotion Analysis
5. Customer Survival Analysis
6. Customer Segmentation Beyond Success vs. Failure
7. Wrap-Up

Predictive Modeling for Incomplete Journeys

Creating our Training Dataset

To create our training dataset we functionalized a way to generate a random time for each completed journey and then made the cut at that time. We then calculated the number of each incident type that occurred in the cut journey and pivoted wider.

Pros: Allows us to use all of the data, none goes to waste

Cons: Does not oversample originally incomplete journeys, which would not be accurate to the actual dataset's behavior when cut-off at one point in time

Feature Engineering

In the process of generating our training dataset, we engineering a few features as to improve the predictive accuracy of our model:

activation_month: the month the user opened an account with Fingerhut and began making actions on the site

time_since: the time (in days) since an action was recorded on the site for a given user (this includes actions performed by the user as well as system generated actions)

journey_time: the time (in seconds) between the latest action on the site for a given user and the first action on the site for the user

Random Forest

The model that we chose for our predictive modeling was a Random Forest with the following hyperparameters:

mtry: 2

# of trees: 1000

cutoff: 80% fail & 20% success

In addition to the features we engineered, we used all of the counts of the different event types except order_shipped.

Performance:
(Brier Score)

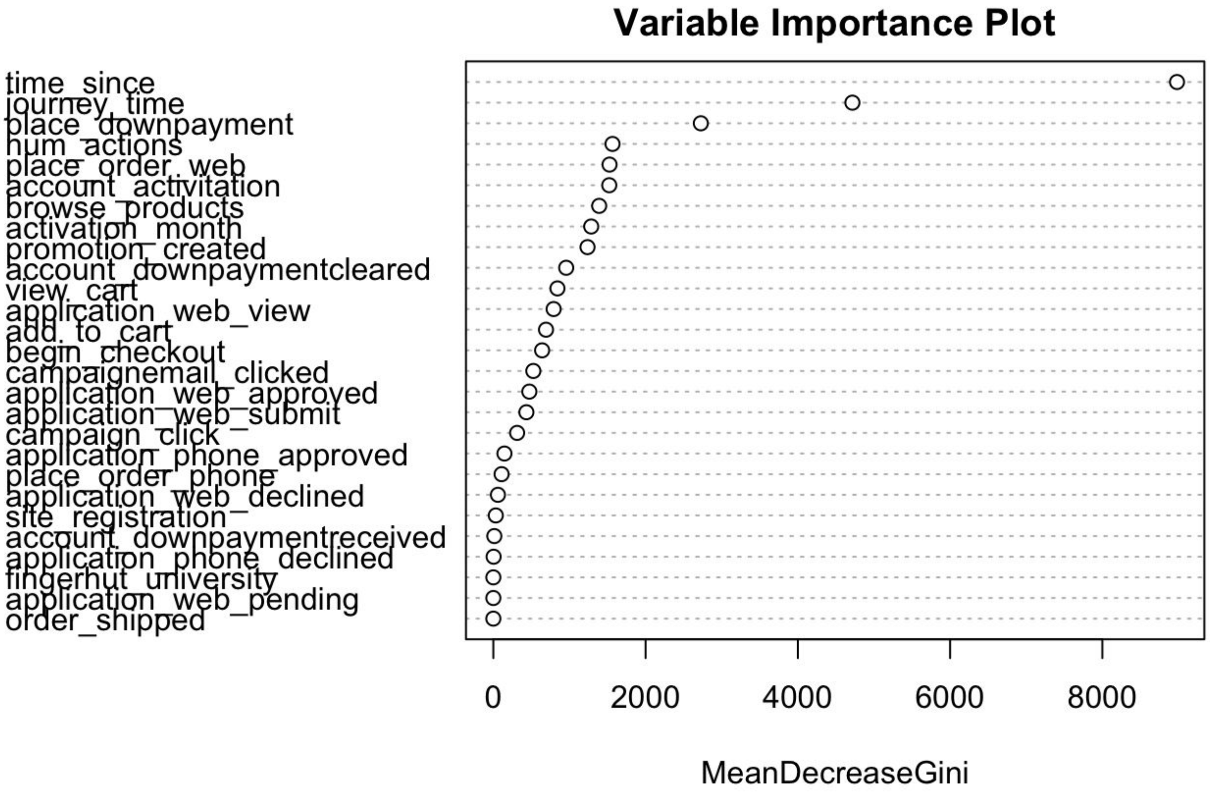
Reference #1 (predict fail):
0.04260856

Reference #2 (predict the proportion of success):
0.06556514

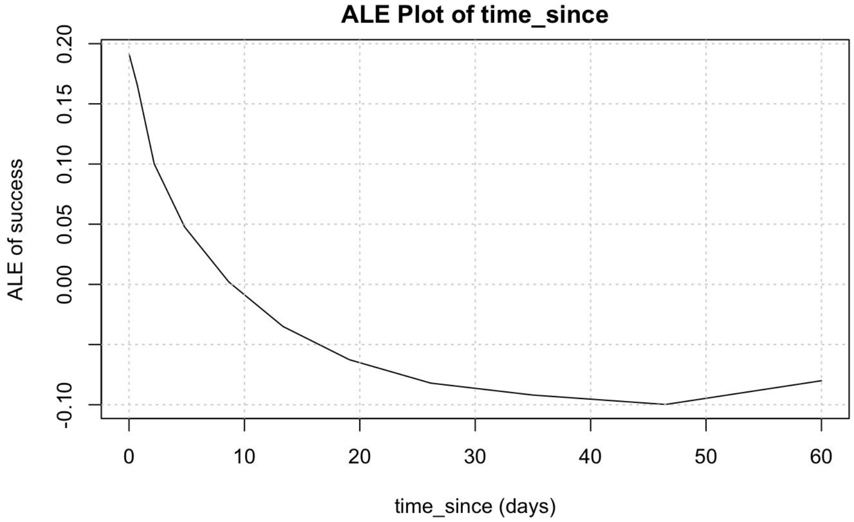
Validation Set #1: 0.04005497

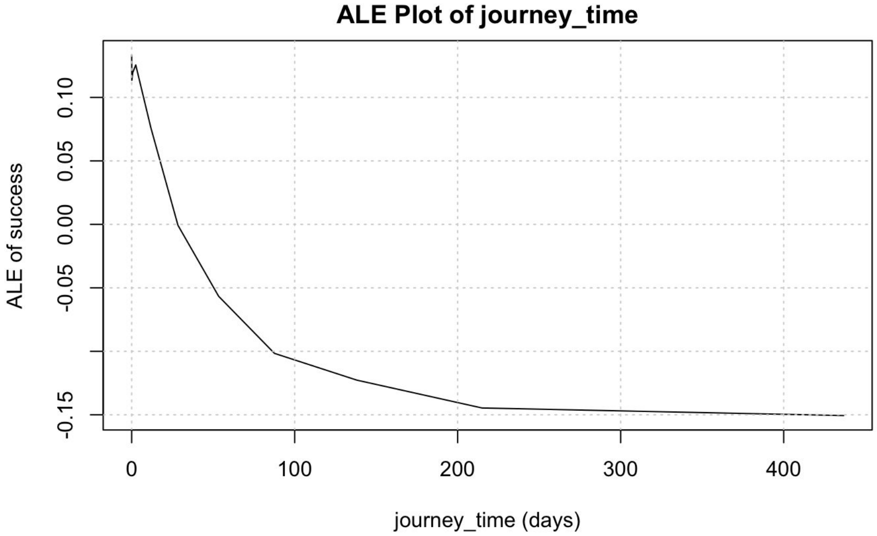
Validation Set # 2: 0.03962162

Interpretation of Model



Interpretation of Model (cont.)

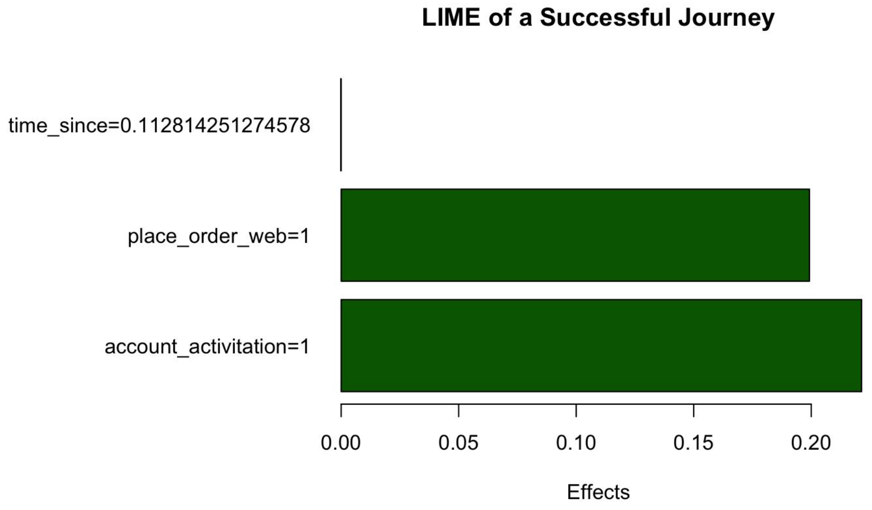


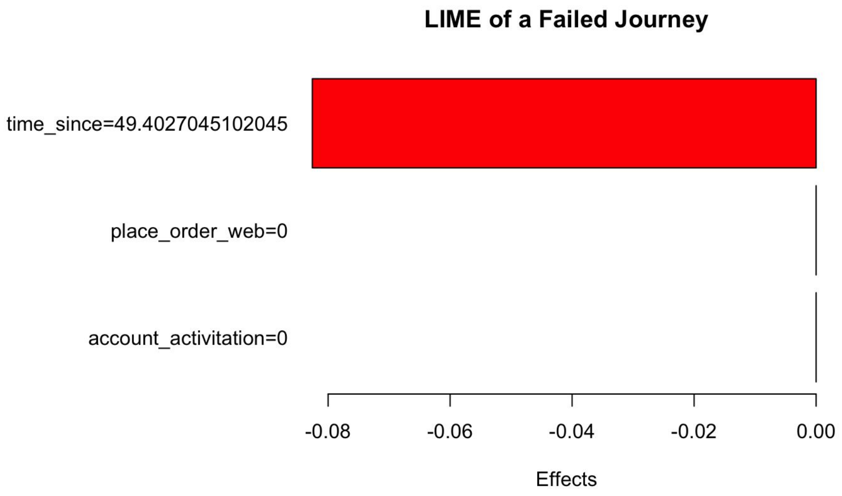


time_since: time since last action on the site (user and system)

journey_time: the amount of time from the first action to the last

Interpretation of Model (cont.)





Forecasting the Number of Orders Shipped

Aggregating Technique & Model Selection

In order to create a Time Series forecast for future orders shipped, we aggregated by the number of orders shipped by month.

The model that we selected for our time series was Generalized Additive Model (GAM) using the **prophet** package.

$$y(t) = g(t) + s(t) + \epsilon_t.$$

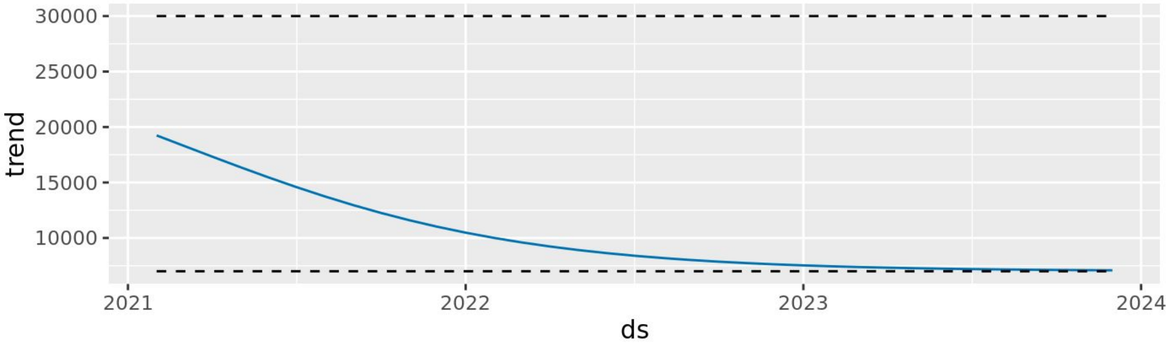
- $g(t)$: trend (non-periodic changes)
- $s(t)$: seasonality (periodic changes)
- ϵt: error term, default prior $\epsilon \sim N(0,0.5)$

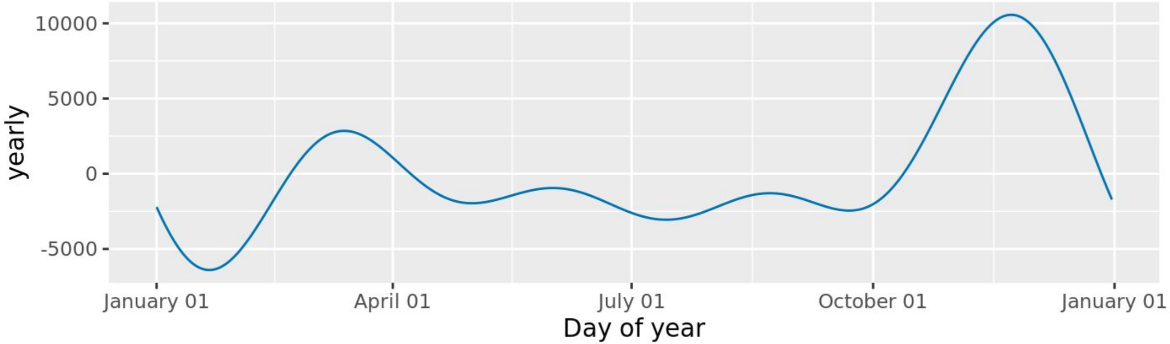
From "Time series analysis using Prophet in Python — Part 1: Math explained" by: Sophia Yang, Ph.D.

This model allows us to better capture trends and seasonality in our data.

For our model we chose a logistic trend and seasonality with a 4th order Fourier Series.

Coefficients & Seasonality

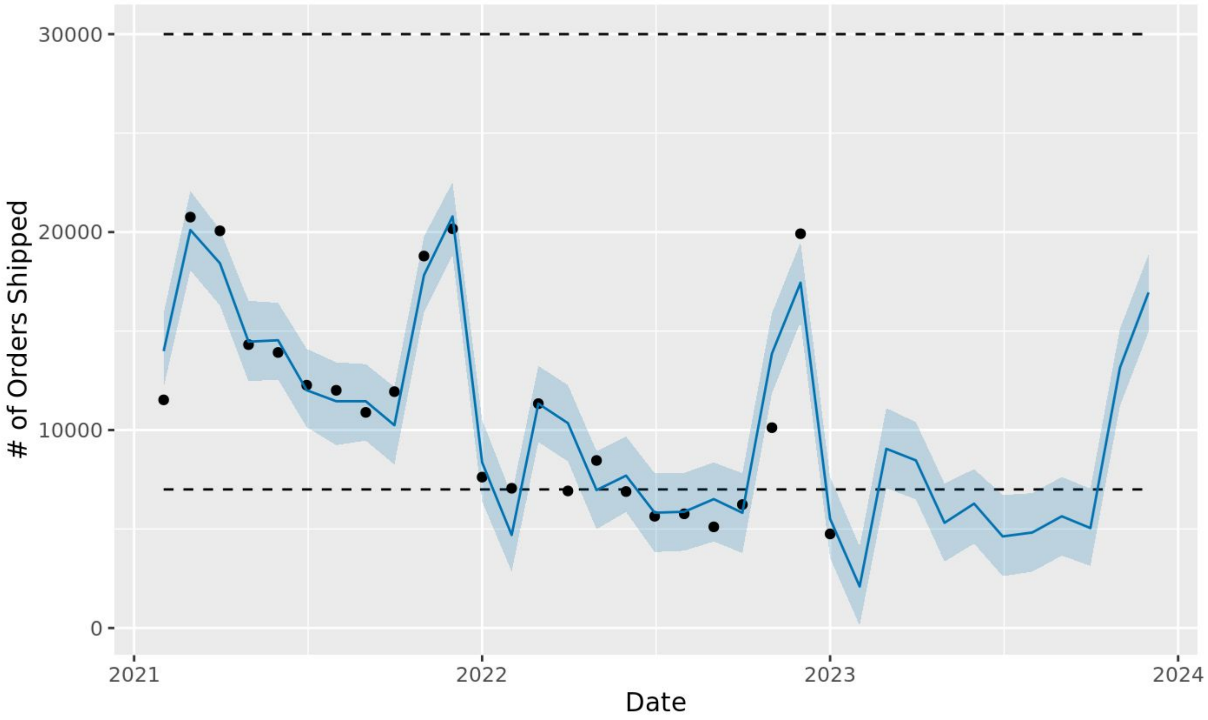




The trend shows the logistic nature with a floor of 7000 and a cap of 30000.

The seasonality shows a spike during the US holiday season and the US tax season.

Forecast



Using the data we make a forecast of the number of order shipped through 2023.

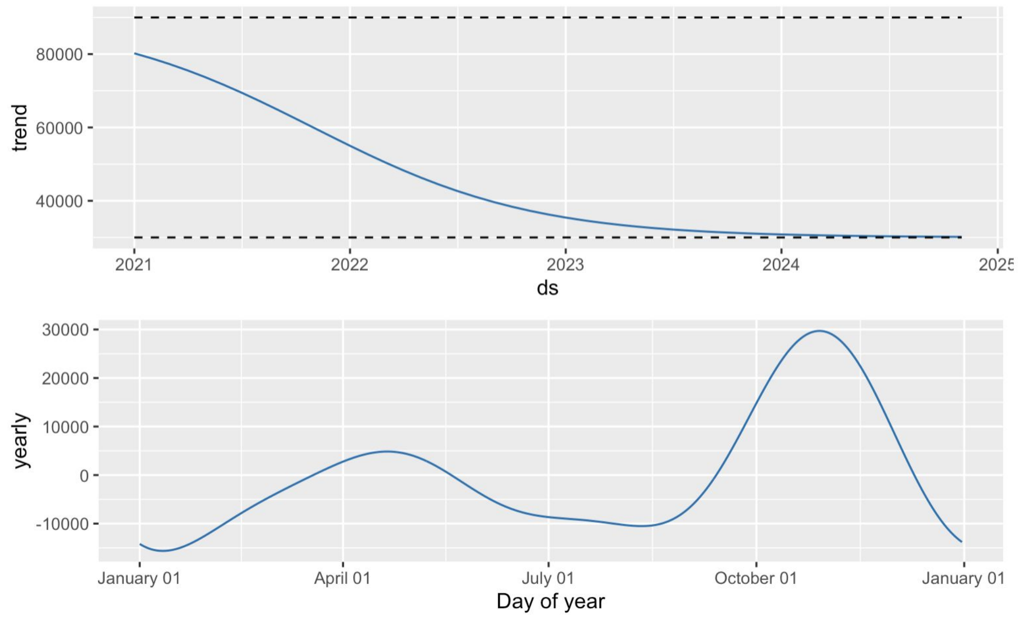
Performance:

Reference (based on only last year's numbers): 2059.66

Root Mean Square Error (RMSE): 1995.133

Forecasting the Number of Started Journeys

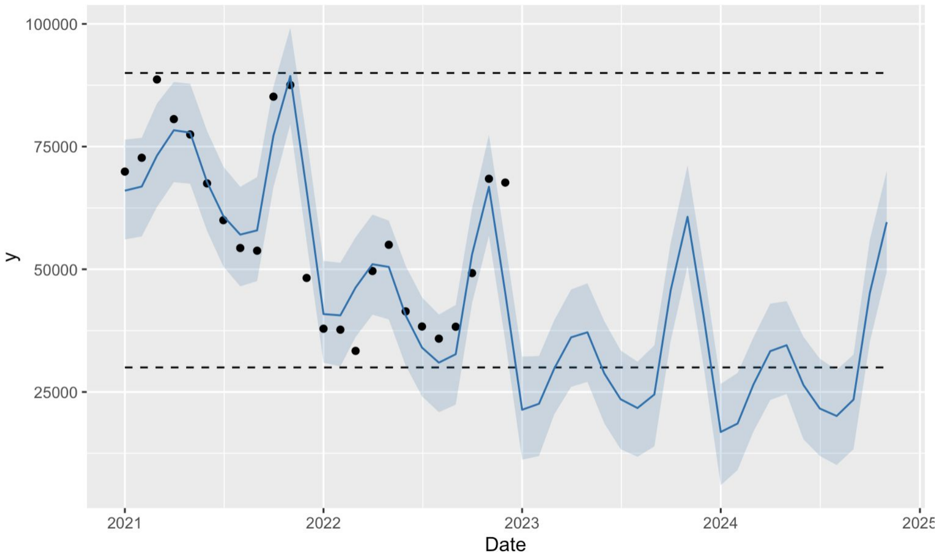
Coefficients & Seasonality



The trend shows the logistic nature with a floor of 30000 and a cap of 90000.

The seasonality shows a large spike during the US tax season and a smaller spike in the US holiday season.

Forecast & Context



Social Context

- The time series model captures the decrease in users from 2021 to 2023. However, that might be due to the increase in online shopping due to the pandemic in 2021, followed by the subsequent economic downturn

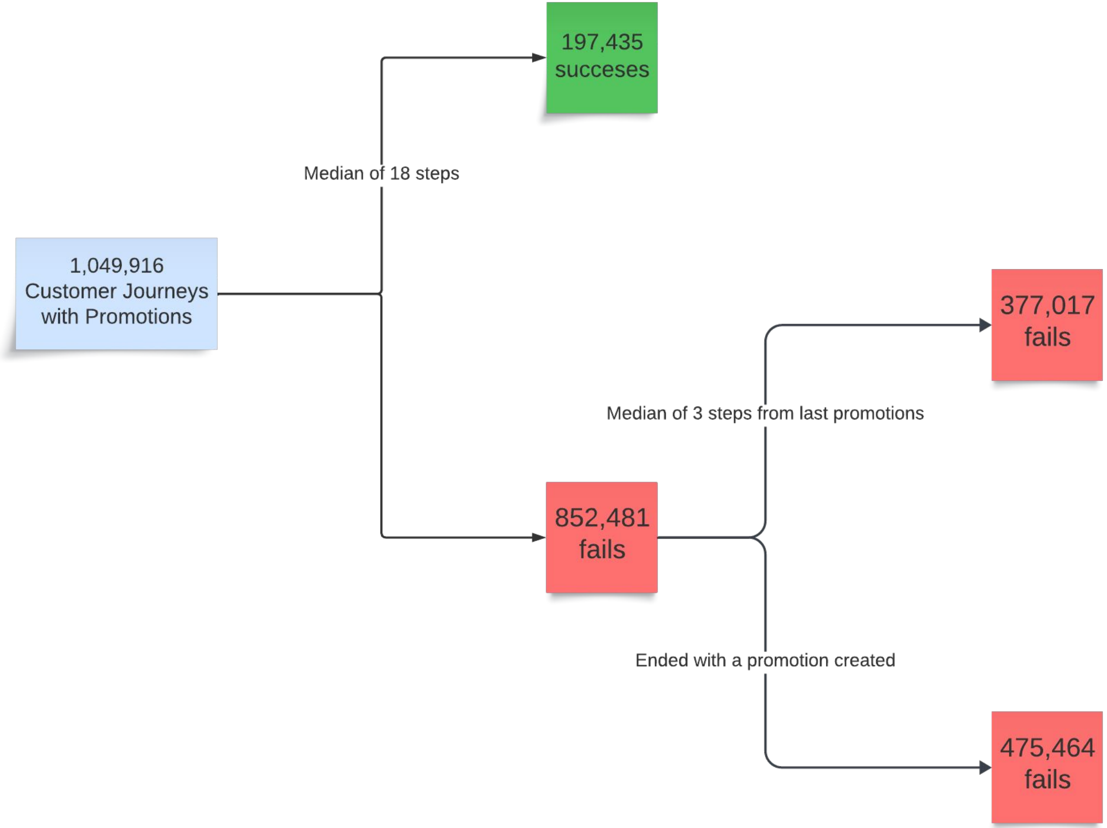
Logarithmic Flooring

- Applied when there is a “saturation level”, or a natural upper and lower limit to the number present
- Using a logarithmic model also emphasizes local trends over global trends, meaning that seasonality will be more prominent

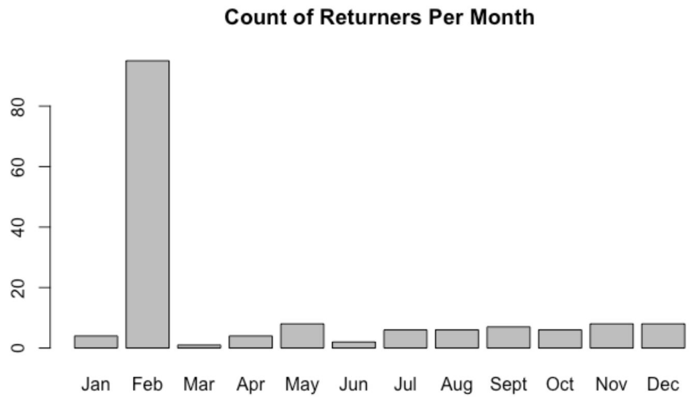
https://www.jstor.org/stable/2582785?seq=6

Promotion Analysis

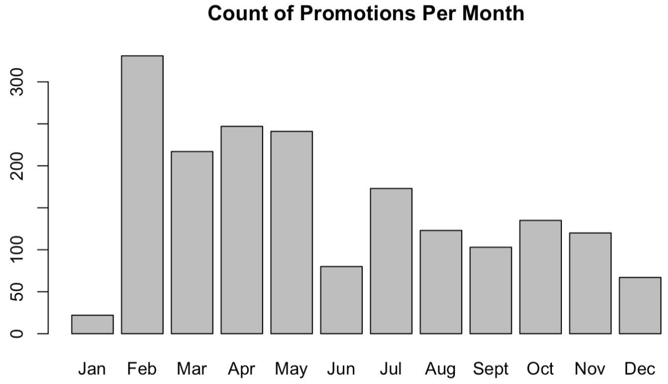
Promotions Flowchart



Returner’s Perspective



Returner:
- Defined as a user who left the application for longer than 60 days but ultimately returned
- ie. there is a 60 day gap somewhere in their user journey



Analysis:
- Corroborates with graph that shows most promotions occur in February

Customer Survival Analysis

Methodology

- A survival analysis curve is a statistical tool that plots the probability that an event hasn't occurred at each point in time.
- In this case, the “death event” is order_shipped, which means we are directly observing the customers that have **not yet** placed an order.
- Made with the **survival** package on R Studio

Kaplan-Meier Curve:

- Mathematically: 1 - the CDF (cumulative distribution function) of the event occurring.
- The graph looks like a "step ladder" as it is calculating the probability at every point in time

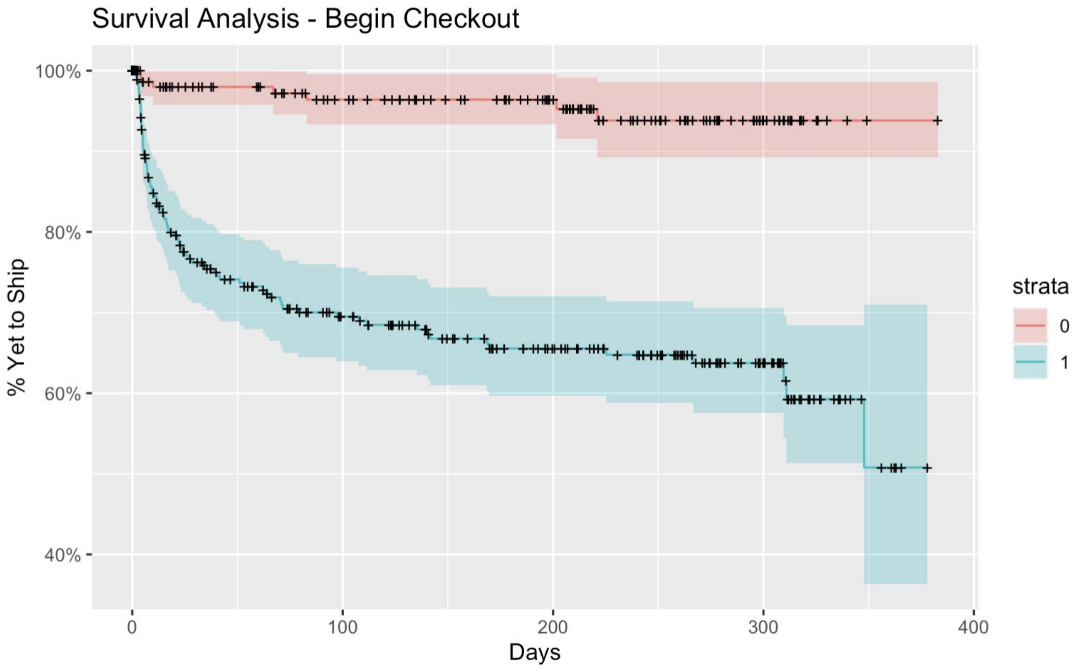
Benefits:

- Intuitive and Interpretable
- Parametric (requiring less assumptions).

https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/
https://www.graphpad.com/guides/survival-analysis

“Check-Out” Retention

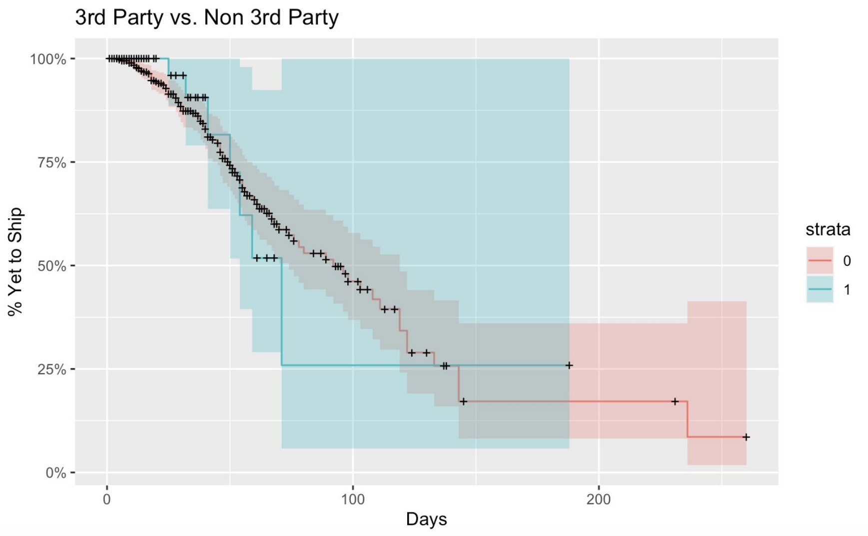
- How many users who initiate the check-out process successfully purchase an item?

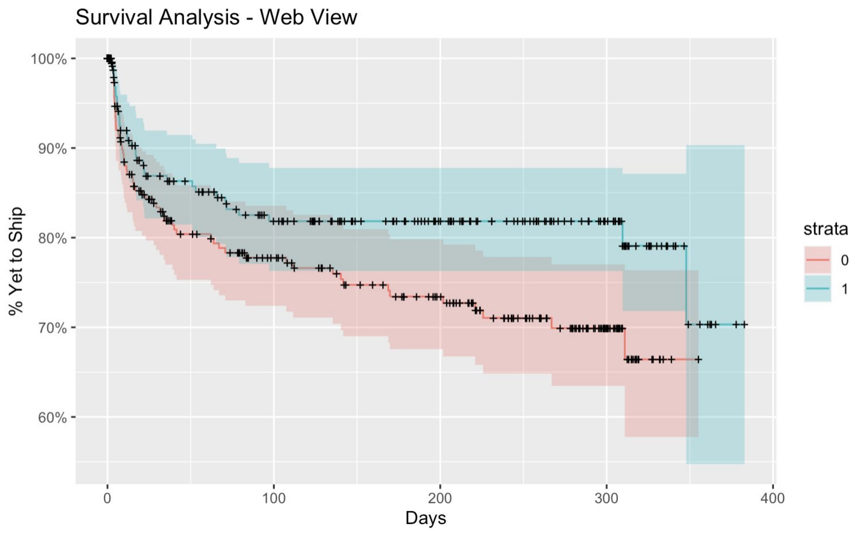


- In general, the results above show that almost 30% of people who browse products, view cart, and begin checkout successfully purchase an item in the first 50 days.

- It also shows that about 50% of people who Begin Checkout do not result in an order_shipped

Analysis of Other Variables





There are significantly more customers who did not come from third-party affiliates.

Of those who did, most placed their orders within the first 50 days.

Those who viewed the platform on web were about 10-20% less likely to place an order than those who viewed the platform on mobile devices.

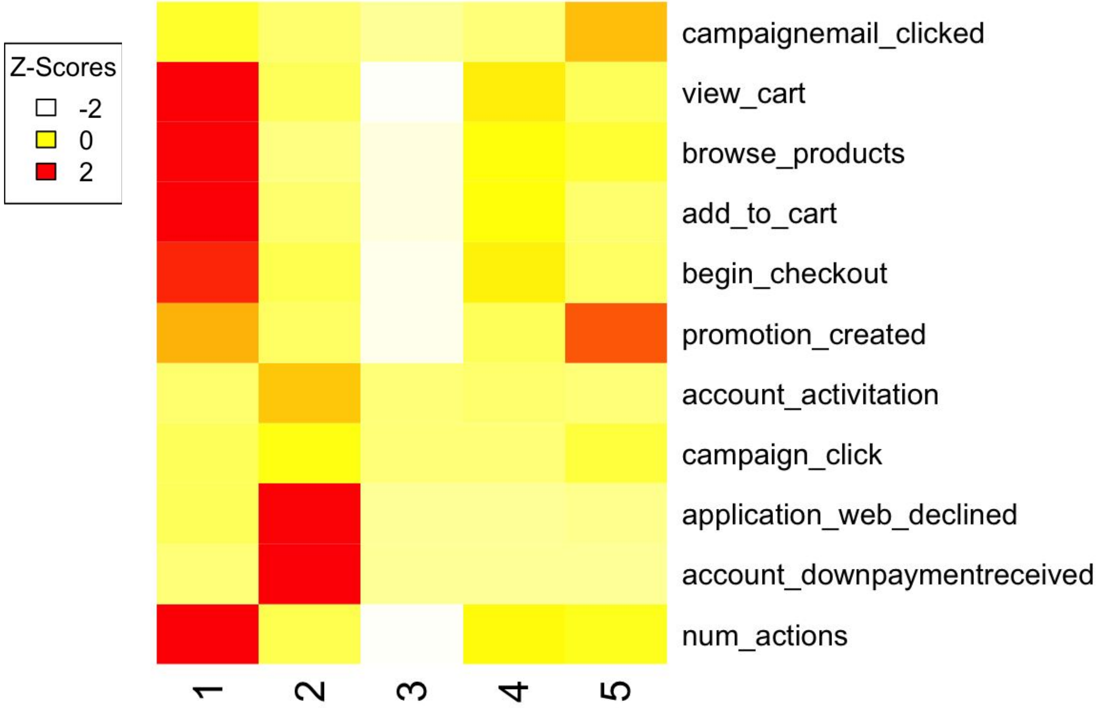
Customer Segmentation Beyond Success vs. Failure

Problem Definition & Methodology

Problem: Can we define segments of customers beyond just whether their journey will be successful or a failure?

Methodology: After flattening the data so that for each user we have the number of times they performed each action, use K-Means Clustering to “over-cluster” users that had successful journeys and users that had a failed journey. By inspection, then manually combine clusters so that a few major clusters can be identified.

Clusters: Successful Journeys



12,578	12,374	170,766	56,896	26,749

Cluster 1: high number of actions, browse and add many things to cart

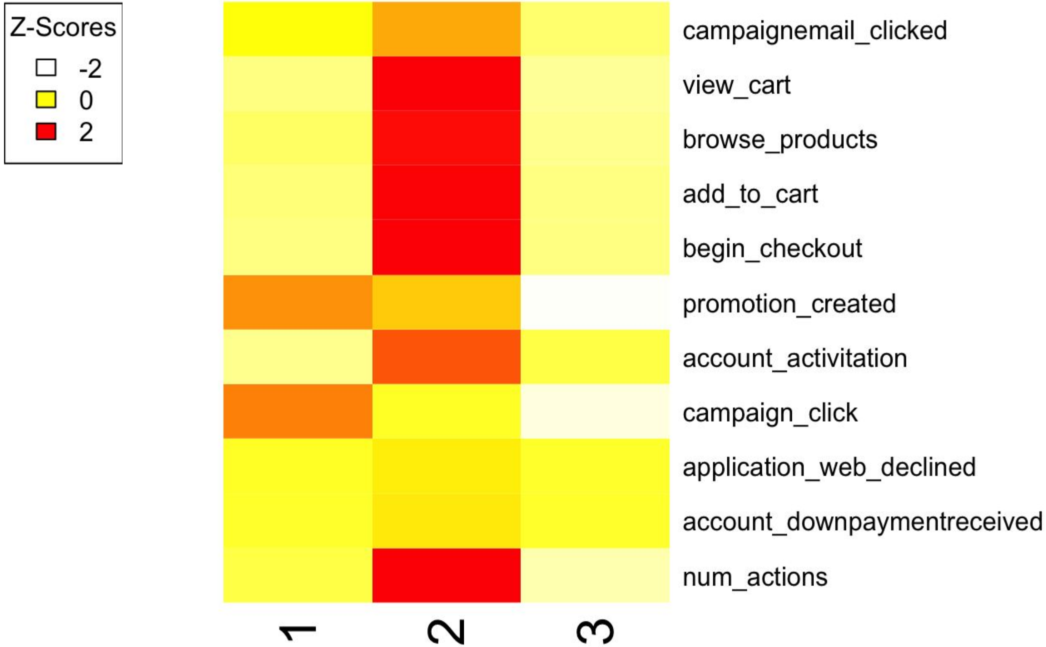
Cluster 2: determined, application declined many times before being accepted

Cluster 3: quick and know exactly what they want

Cluster 4: average customers

Cluster 5: most receptive to promotions

Clusters: Failed Journeys



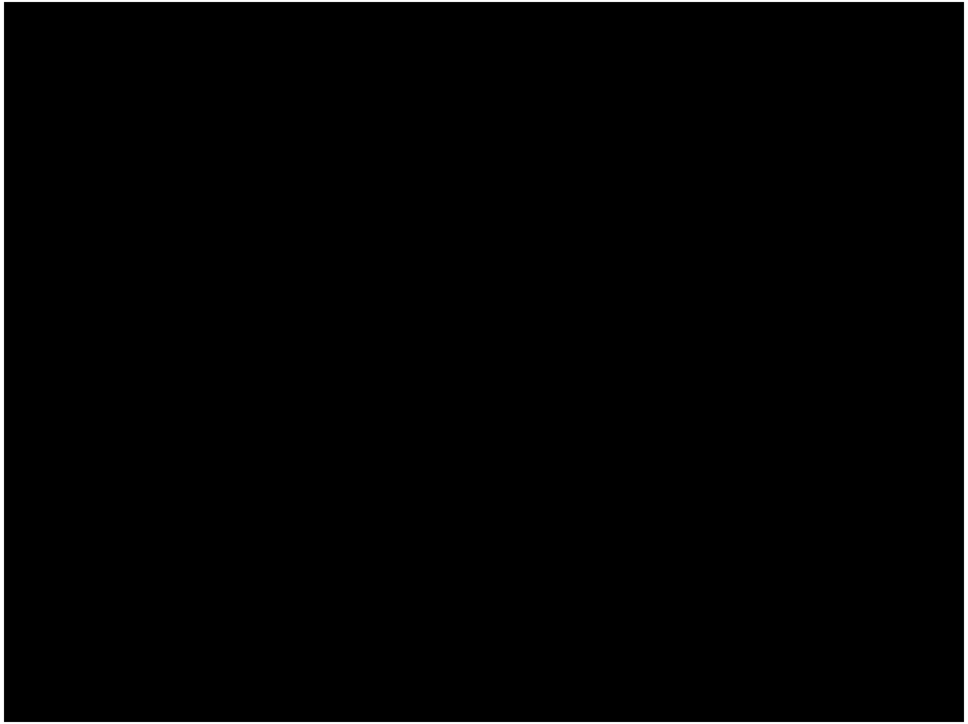
301,622	127,939	564,050

Cluster 1: targeted by promotions

Cluster 2: mostly likely to purchase something

Cluster 3: perform a few actions then never come back to the site

How Journeys Fall into Clusters



Wrap-Up

Recommendations

Seasonality

- Market to prospective customers leading up to tax season (March-April) and the US holiday season (Oct-Nov).

Check-Out Retention

- Impose a “check-out timer” or shopping cart email reminders to push people to completing their check-out journey after 50 days

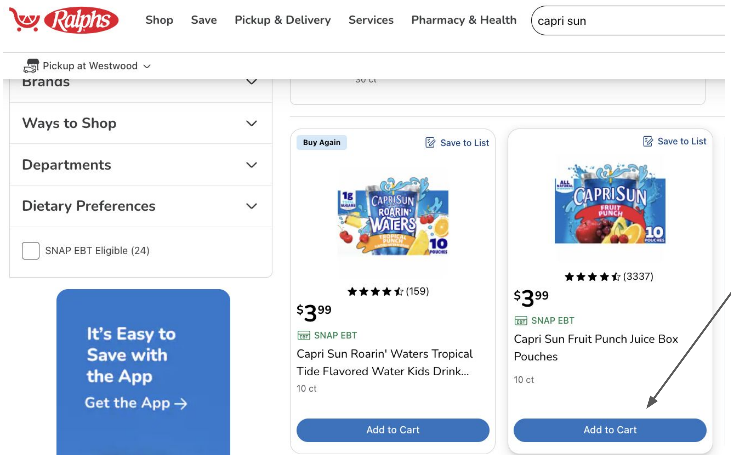
Web View vs Mobile View

- There could be some interesting UI/UX research on how to increase purchases on the website.

Recommendations Cont.

Clustering

- To convert more users from failed cluster 1 to failed cluster 2, promotions should lead to a product browse page once clicked
- “Add to cart” function can be added to the browse products page to kickstart someone’s checkout journey



These add to cart buttons on the Ralphs website allows users to add to cart while browsing, which are the two most significant actions from failed cluster 2

Special Thanks to…

Ben Thompson (BlueStem Brands) for providing the data and serving as a liaison between FingerHut and UCLA

Professor Maierhofer for his mentorship and support this quarter as we brought what we have learned about in class to application

Joseph Resch for providing guidance and feedback to our work

UCLA Department of Statistics & Data Science for developing Stats M148 and providing us the opportunity to showcase our skills

Questions? Comments?