Forecasting Demand for Divvy – *Chicago's bike share system*

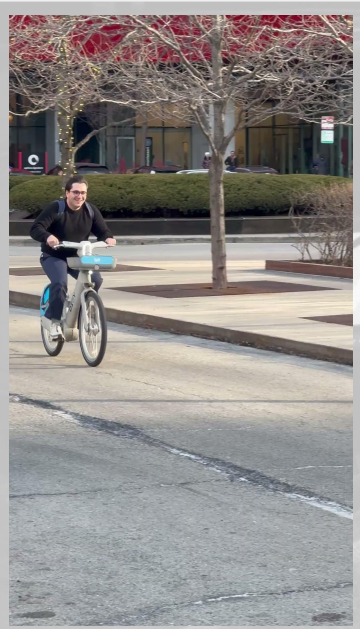
March 12th, 2025

Andrew Castillo, Vincent Chiro
Zimeng Huang, Alex Lee, Manas Vemuri

DIVVY

Introducing the team…

Andrew Castillo



Zimeng Huang

Alex Lee

the team…

Vincent Chirio

Manas Vemuri

Agenda

- Understanding Divvy's bike share system
- Data-Driven Decision Framework & Data Overview
- Exploratory Data Analysis & Feature Engineering
- City-wide Modeling
- Station-level Modeling
- Insights & Recommendations
- Implementation & Future Work

Understanding Divvy – Chicago's bike share system

Divvy Overview

Divvy is Chicago's bike-sharing system launched in 2013 – providing convenient, sustainable, and affordable transportation **integrated into Chicago's L (train) and bus network.**

- **5.8M rides in 2024**

Despite success, **Divvy faces challenges with bike availability**

- **High-demand stations** frequently experience **bike shortages**
- Solutions require efficient bike redistribution

Forecasting Demand & Solving Bike Redistribution

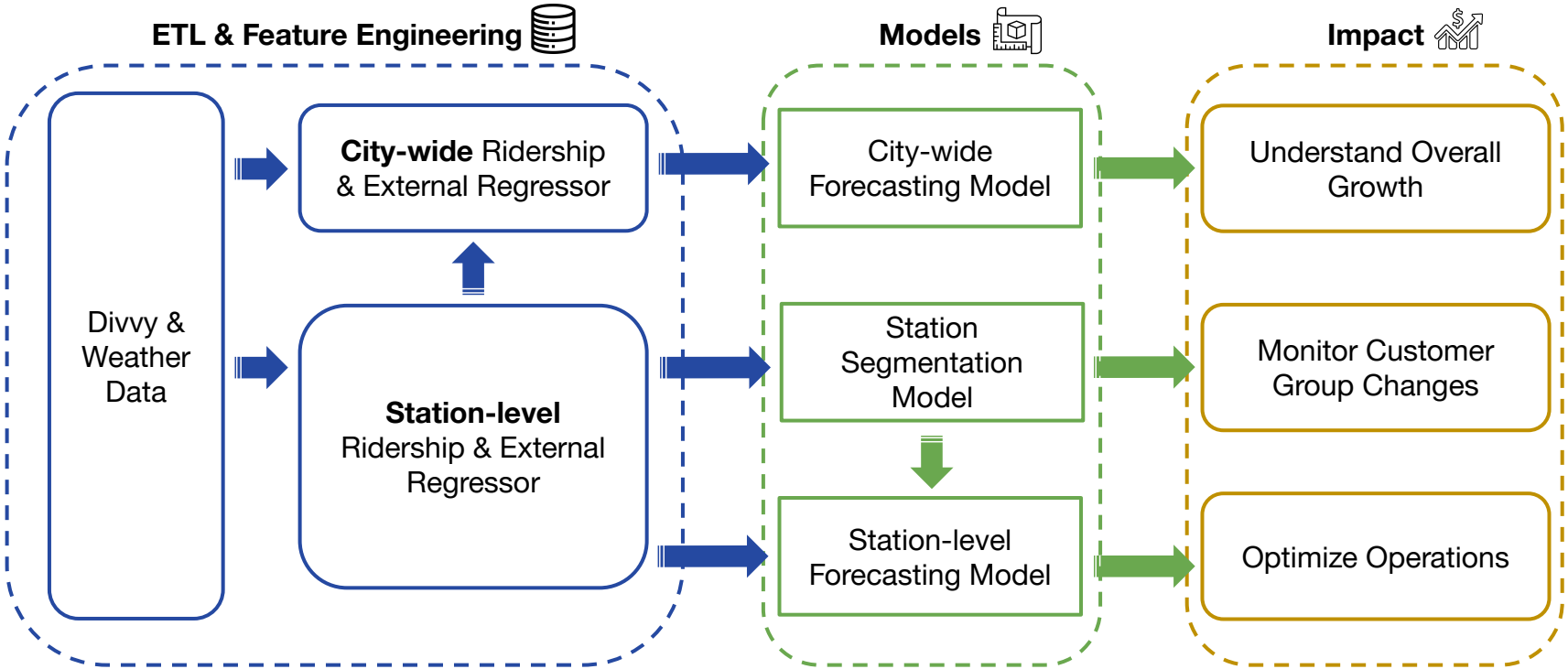
Demand Forecasting

Generate accurate bike demand forecasts, enabling better resource allocation

Optimize Bike Redistribution

- **Data-driven allocation strategies** based on predicted demand patterns
- Evaluate **expansion potential for '*Bike Angels*' program** to incentivize customer-assisted bike redistribution

Data-Driven Decision Framework: Forecasting & Business Impact



Data Overview – What does the ridership data look like?

Daily Ridership – Divvy Data:

Publicly available data from Divvy's website, including **daily ridership information aggregated** into monthly and quarterly files.

Key Features:

- **Ride Identification:** Unique ride IDs for each journey
- **Bike Types:** Classification of bikes used (traditional vs. electric options)
- **Trip Timing:** Ride start and end timestamps

ride_id	rideable_type	started_at	start_station_name	start_station_id	end_station_name
5DEFF0229C83F70F	electric_bike	Tuesday, October 22, 2024	N Paulina St & Lincoln Av	20253	Franklin St & Lake St
A97E23B7238AF887	electric_bike	Sunday, October 13, 2024	N Paulina St & Lincoln Av	20253	Southport Ave & Wrightw
616D981E9B1749A7	electric_bike	Friday, October 11, 2024	N Paulina St & Lincoln Av	20253	Montrose Harbor

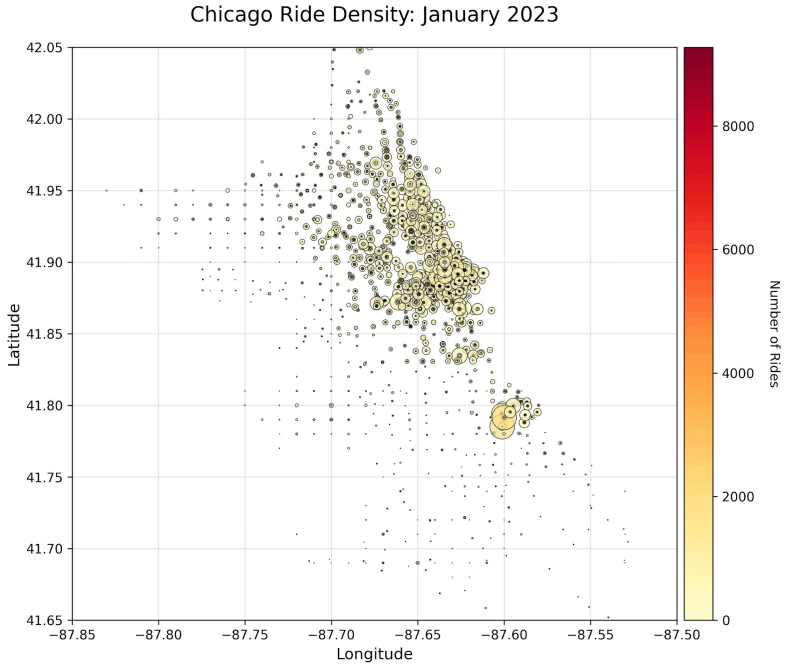
Incorporating Weather Data:

- **Temperature** – Daily high/low temperatures are likely to significantly impact ridership
- **Precipitation** – Rainfall and snowfall amounts could discourage ridership

Exploring Seasonal Patterns – Higher Ridership Outside the Loop in Off-Peak Months

Confirming suspicions around seasonal patterns:

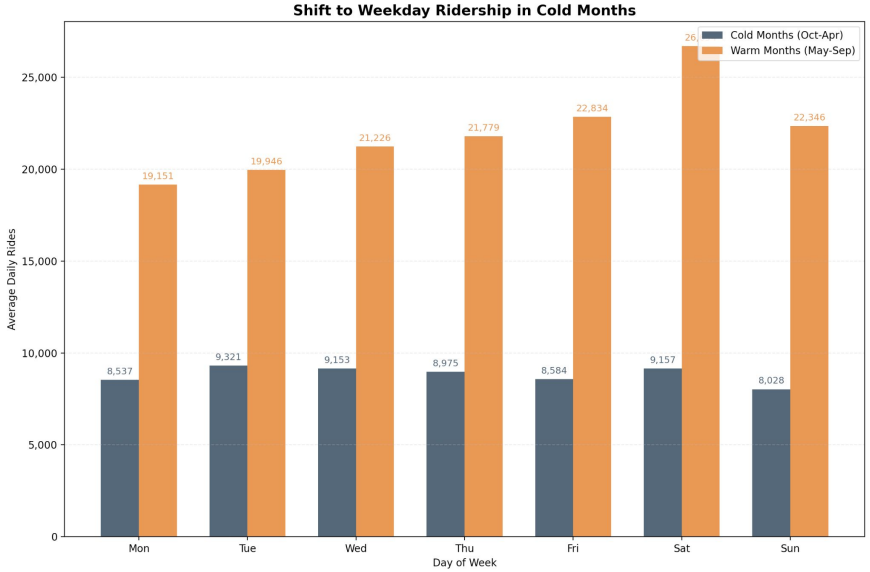
- **Seasonal Impacts:** Distinct patterns emerge around warmer months
- **Ridership shifts to colleges and commutes** during off-peak months (e.g. University of Chicago)



Weather conditions influence both ride volume and patterns

How weekly patterns shift with seasons:

- Cold months show relatively consistent ridership across weekdays (Mon-Fri) – suggests riders use Divvy for commutes when it's colder
- Warm months show higher weekend ridership, with Saturday having the highest ridership of all days, suggesting during the summer riders are using Divvy for leisure



Feature Engineering – Exposing Lags & Seasonalities

Data	Feature Engineering
E-bike Availability	Moving averages of e-bike proportions
Stations	Moving averages of # of unique stations
Date	Month, Year, Day , Week of Year, Day of Month, Weekday, Weekend
Weather	Temperature, Rain Intensity, Snow
Ridership Lags & Moving Averages (XGBoost)	Lags and moving-averages for ridership at different intervals (e.g. 1, 7, 30, 365)

City-wide Forecasting: Prophet Delivers Best Performance in Baseline Models

Model	MAE	MSE	MAPE	sMAPE	MASE	R^2
ARIMA (6,1,2)	9096	127,253,356	66%	97%	3.30	-1.623
SARIMA ((0,0,0)(0,1,0,365))	3203	18,693,134	44%	33%	1.16	0.615
Prophet	2414	9,781,558	35%	28%	0.88	0.798

AutoARIMA and GridSearchCV used for model tuning

- Prophet delivered the best performance and fastest training
- Based on these initial results, we prioritized Prophet and other advanced models with external data integration, excluding classical ARIMA/SARIMA approaches

City-wide Forecasting: Incorporating External Regressors

Model	MAE	MSE	MAPE	sMAPE	R^2
Prophet with stations	2,775	13,688,579	24.5%	**23.9%**	0.82
Prophet with temperature	2,005	6,717,550	29.0%	**25.7%**	0.86
XGBoost with temperature	2,295	8,815,499	26.5%	**23.1%**	0.82
Linear Regression with ARIMA errors	2,589	10,423,372	42.9%	**30.7%**	0.79

- Adding external regressors of stations and temperature does further improve Prophet's predictive power
- Using XGBoost with temperature gives us the lowest error metrics
- Linear Regression with ARIMA errors is a big improvement on ARIMA but is not able to keep up with the more advanced models

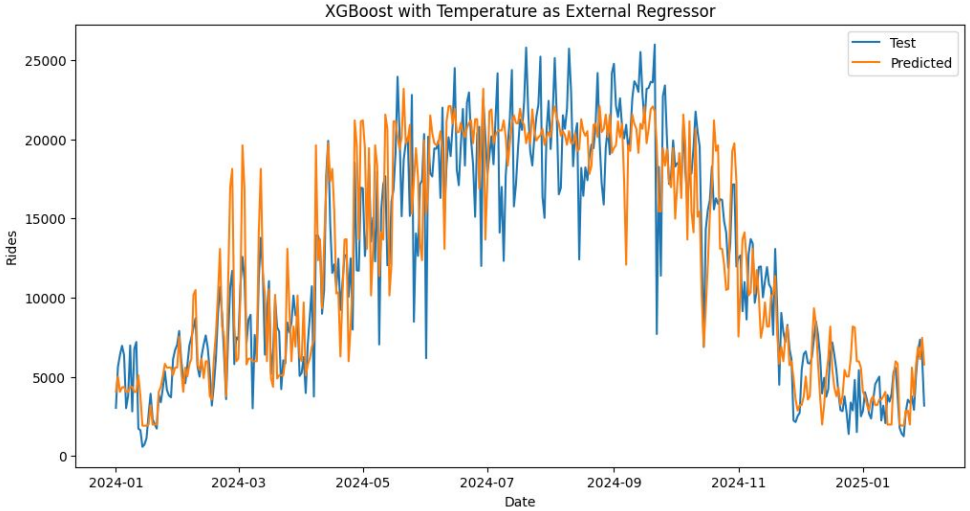
XGBoost with Temperature Excels in Forecasting

Preferred Model

XGBoost + Weather

XGBoost (& decision trees) do well with capturing non-linear relationships
→ Temperature explains a large part of variance in ridership

Prophet assumes a more linear effect of external regressors whereas XGBoost can see threshold effects



City-wide Forecasting: Insights on Overall Growth

Citywide Growth (2023 -> 2024): -1.5%

- The projected negative growth suggests a decrease in Divvy bike demand from 2023 to 2024.

Citywide Forecasting Can help Divvy:

Spot Long-term Trends	• Citywide ridership forecasts helps Divvy identify growth patterns and seasonal shifts, guiding high-level strategic planning.
Prepare for Changing Demand	• Divvy can adjust the total number of bikes, optimize station placements, and allocate maintenance resources based on fluctuations in demand.
Evaluate Past Strategies	• Comparing forecasted growth with historical performance helps Divvy assess how well past initiatives have worked and informs future decisions.

Using Dynamic Time Warping to Cluster Stations

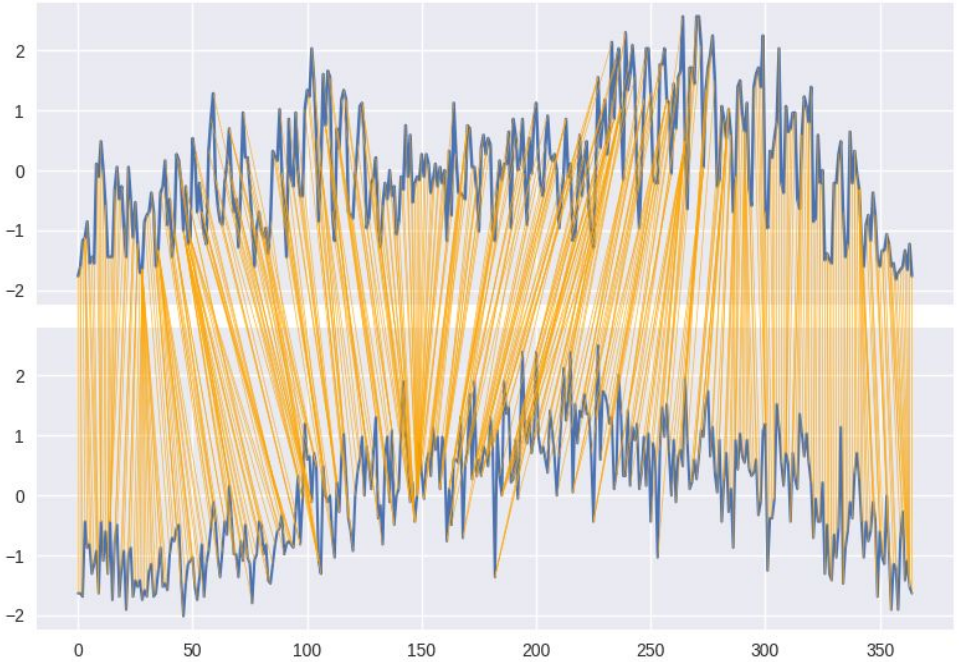
Dynamic Time Warping

- Capture time series *patterns*.
- Robust to noise.
- Normalized for volume discrepancies.

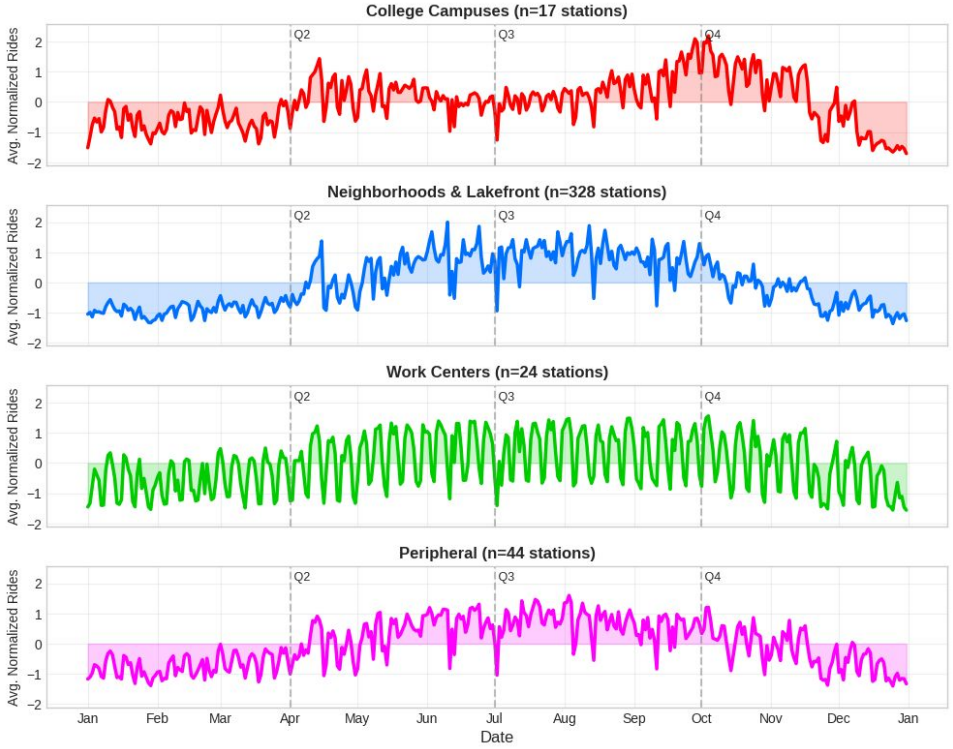
Hierarchical Clustering

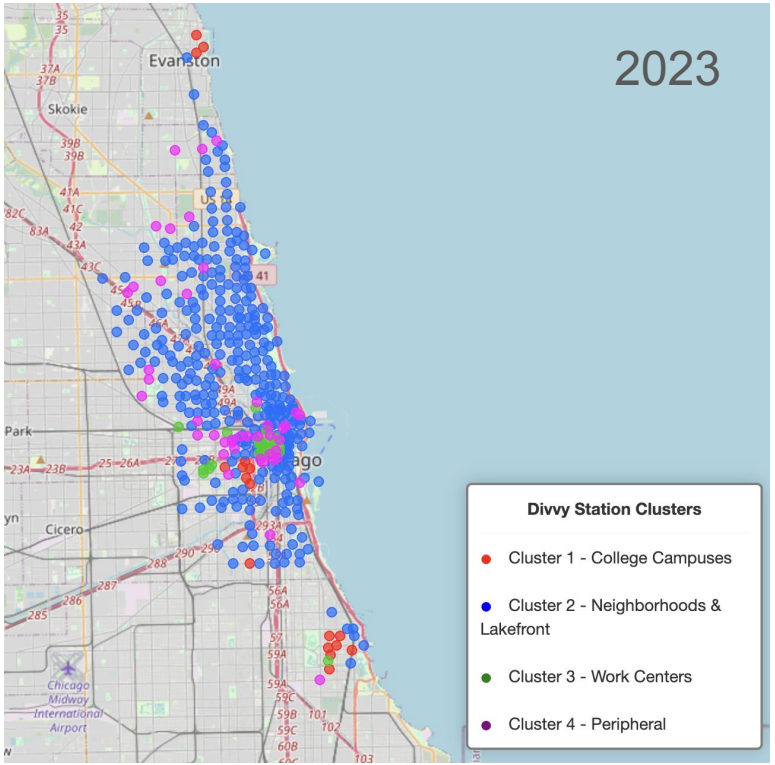
- Easy integration with DTW.
- Optimal for cut-point tuning and evaluating inter-cluster similarity.

900 W Harrison St. vs. Aberdeen St & Jackson Blvd

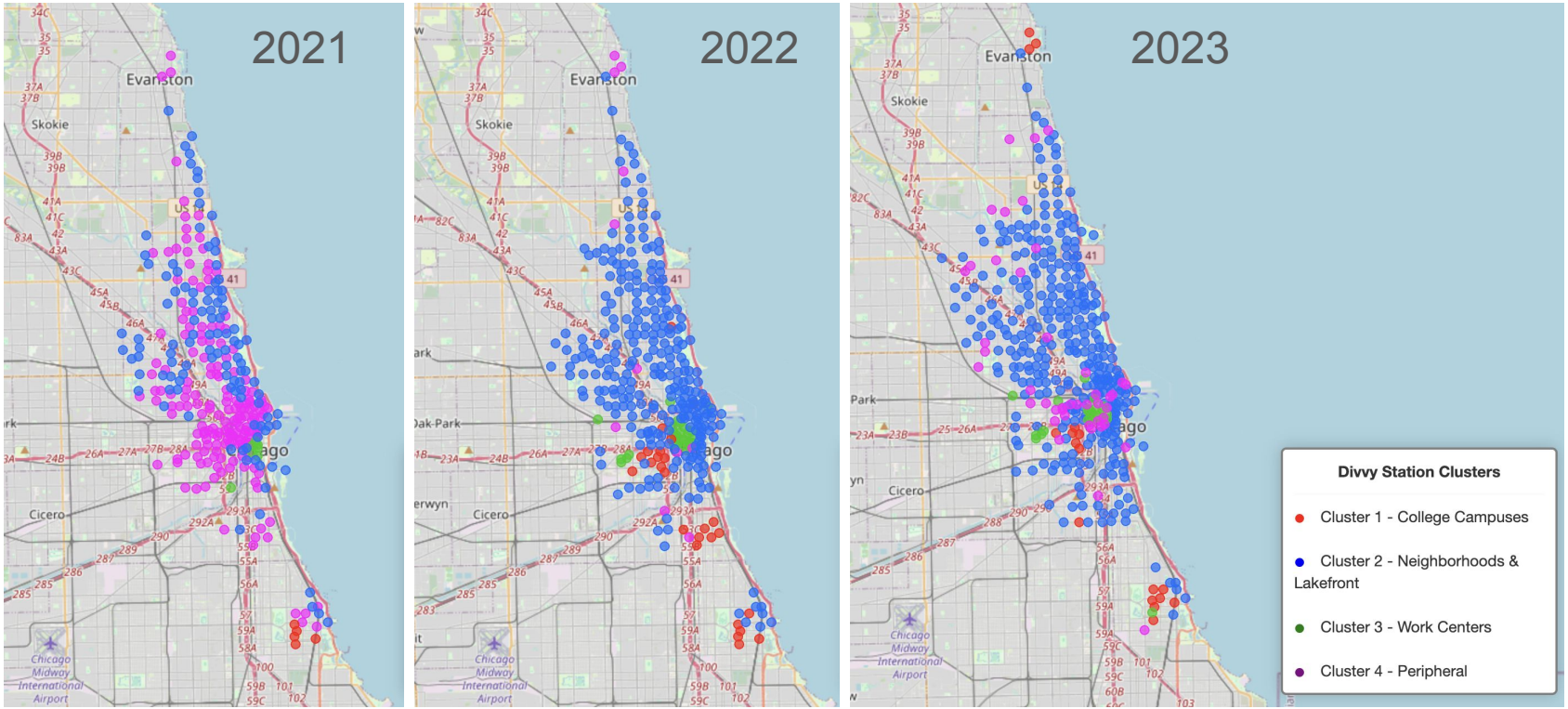


Clustering: Results & Qualitative Interpretation





Clustering: Station Clusters Over Time



Clustering: Why the Mild Cluster Shift Over Time?

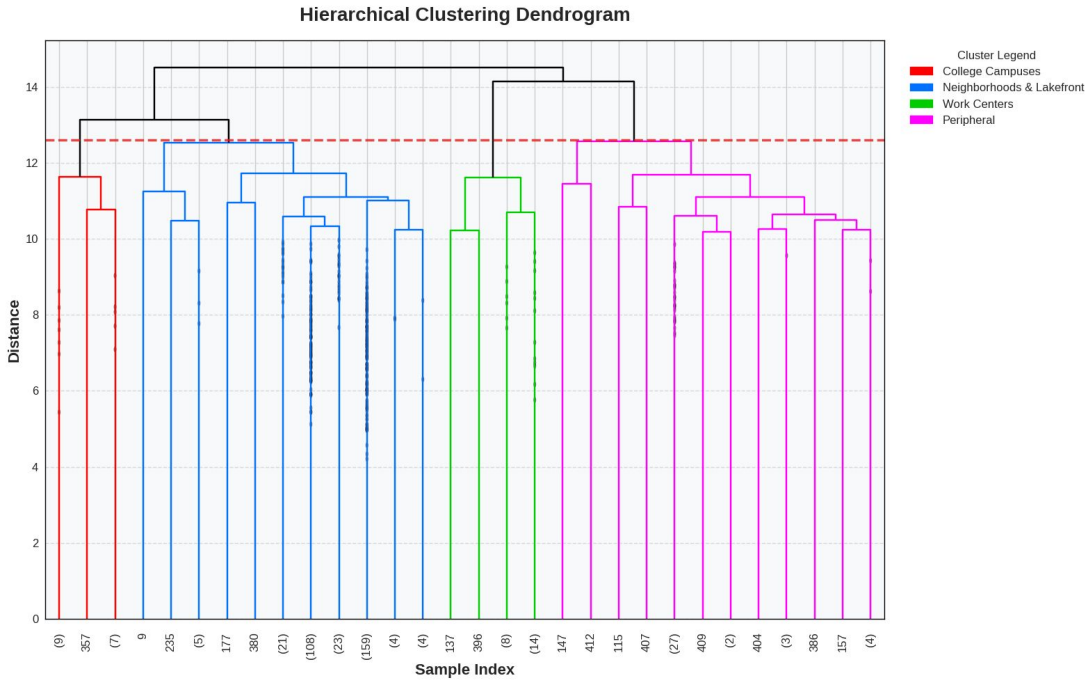
Arriving at our cluster levels

Optimal Cut point at 4 clusters based on elbow method with in-cluster SSE.

Anywhere from 3 to 6 clusters are reasonable choices based on dendrogram.

Requires maintenance

- Range in cluster viability suggests instability in clustering as station patterns and demographics shift overtime



Station-level Forecasting: One Model vs. Individual Models

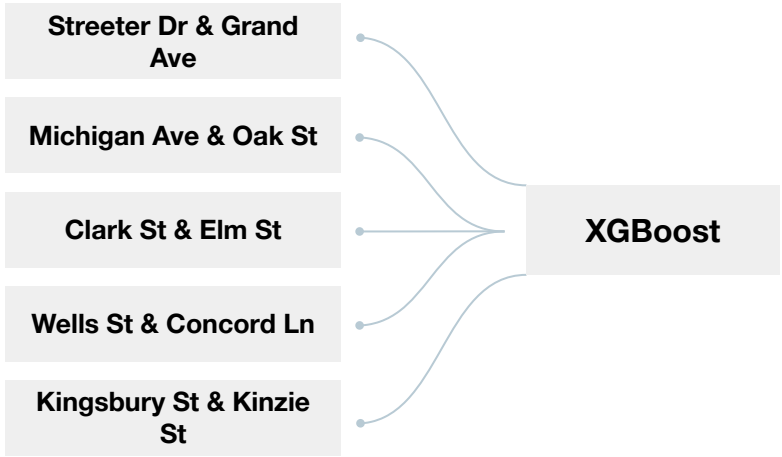
Why build and maintain individual models when you can have one?

Efficiency & Scalability

- One XGBoost model handles all stations simultaneously vs. maintaining 1,000+ individual models
- Dramatically reduces training and deployment time (hours vs. days)
- Easier to update as new stations are added to the system

Station Incorporated Through One-hot Encoding as features

Incorporates Clusters to Capture Further Segmentation by Location



Mixed Results for XGBoost on One-Day Ahead Forecasting

Model Performance

Model	MAE	MSE	MAPE	sMAPE	R^2
XGBoost	4.79	89.97	48.67%	35.38%	89.3%
XGBoost + Clustering	4.73	86.41	48.63%	35.32%	90.0%

→ Impressive R^2

→ Poor sMAPE – on average we're off by 35% per station per day

❌ ***falls short of industry standards of 5-15%***

Refining Our Approach for Real-Time Decision System for Redistribution

Enhanced Model Development

- Build on the existing XGBoost framework that efficiently handles all stations
- Incorporate supply data as additional features to improve prediction accuracy
- Add temporal features to capture patterns by time of day, day of week, etc.

Error Reduction Strategy

- Target significant improvement in sMAPE metrics (from 35% to 10%)
- Implement station-specific error weighting

Station-level Modeling: Real-time Operation Decision System

Demo usage in real-time decision: alerts on significant demand change in 2024

Demand Change in Stations	Case Definition	Alerts Standard	Alerts Sent	Precision	Recall
Significant Increase	**Increase > 50 rides the next day**	**Predicted increase > 200**	1409	92%	2%
Significant Decrease	**Decrease > 50 rides the next day**	**Predicted decrease > 200**	1394	100%	4%

- While the system may not capture all demand fluctuations across stations, it serves as an extreme-case alarm system to inform operational strategies.
- For instance, bikes can be redistributed at midnight based on alerts signaling significant demand shifts at different stations.

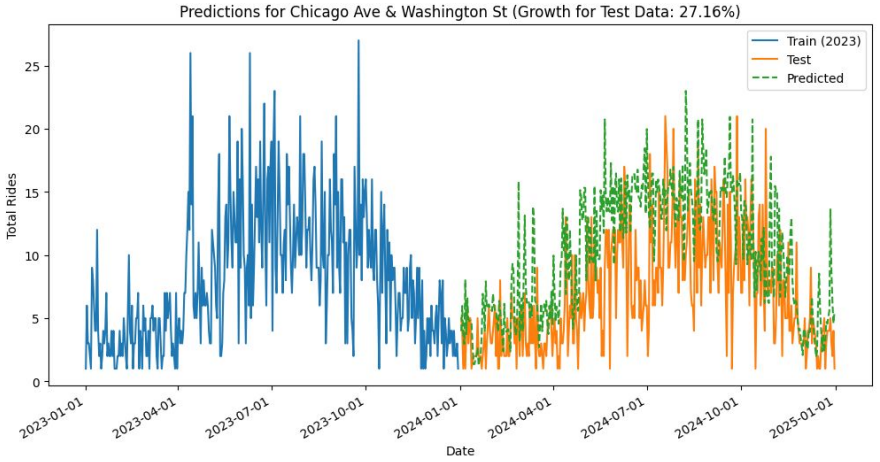
Station-level Long-term Forecasting: Growth Potential (without lags)

Stations rated for highest growth (2023 → 2024):

- Station: Chicago Ave & Washington St, Growth: 27.16%
- Station: Marine Dr & Ainslie St, Growth: 23.77%
- Station: Fairbanks Ct & Grand Ave, Growth: 23.16%

How Divvy can leverage this information:

- Expand bike docks & bike availability at high-growth stations to meet demand
- Plan for expansion efforts to add new stations to demand areas
- Scale back or relocate underused stations to optimize resources



Incorporating Lagged Features on: Considerations for Forecasting

Data Leakage

- Lagged values depend on recent actual ride data
- These values aren't available when making true future forecasts

Takeaways

Incorporating lagged features limits your forecast window to next-day

Incorporating Recursive Forecasting Methods

- Recursive forecasting uses predictions from previous time steps as inputs for future predictions
- Creates a chain of predictions where each forecast builds upon previous forecasts

Limitations: Errors compound over time as predictions feed into future predictions, your largest effective forecasting window might be only 14-days!

Refining Our Approach for Real-Time Decision System for Redistribution

Adding External Factors	• Factors like city events, gas prices, and public transit disruptions could help improve forecast accuracy
Advanced Pricing Model	• Analyze Divvy's pricing data to develop dynamic pricing • This could help optimize their strategy and maximize profits
Real Time Data Integration	• Real time data like current weather and traffic reports could help make more granular forecasts to make adjustments within the day
Incorporating Supply Data	• Demand forecasts are only part of the equation, with supply side data we can better help prioritize restocking and where to concentrate efforts

Thank You!

Thank you to Sam Tagle (@tagle.foto) for our first and last slide pictures. Like street photography? Check out his work!

Appendix 1: Variables for City-wide xgboost with lags

Variables

Influential Factor	Information Utilized	Variables	Variable Type	Transformation	Notes
Seasonality	**Historical Ridership**	Yesterday	Continuous		
		Last Week	Continuous		
		Last Month	Continuous		
		Last Quarter	Continuous		
		Last Year	Continuous		
Demand	**Time**	Year	Continuous	X = X - 2020	Yearly difference
		Whether 2020	Dummy		Pandemic
		Month	Dummy		Monthly difference
		Week of year	Continuous		Weekly difference in a year
		Day of week	Dummy		Daily difference in a week
		Is weekend	Dummy		Weekend vs Weekday
	Holiday	Whether it's holiday	Dummy		
	Weather	Temperature	Continuous		
		Rain levels	Dummy	cut by median rain	levels = no, moderate, heavvy
		Snowfall flag	Dummy		rain vs no rain
Supply	**Station**	Yearly Moving Average of Counts	Continuous		
	E-bike	Yearly Moving Average of proportions	Continuous		
Price	**Price plans**	Annual Membership	Continuous		
		Day Pass Price	Continuous		
		Unlock Fee exists	Dummy		
		Member ebike unlock fee exists	Dummy		