Improving User Experience with Machine Learning

Case Study: 🚲-sharing

Marc Puls | Code Analytics | 03.04.2025

AGENDA

INTRODUCTION

Capital Bikeshare (CBS) system in Washington, DC (USA)

Time period: 2021 - 2023

Datasets:

1. Hourly, weather, casual/member | 1.095 rows
2. Daily, weather, casual/member | 26.280 rows
3. Daily, Time, Start & End stations, casual/member | 10.693.996 rows

EDA | OBSERVATIONS

Different coordinates for same station → Used **first sample** for aggregation

Haversine & geo.py calculate distance at birdflight (!) ≠ urban regions
→ Considered as **acceptable inaccuracy** as all records treated the same

Multi-day rides → **Excluded**! Focus on same-day rides.

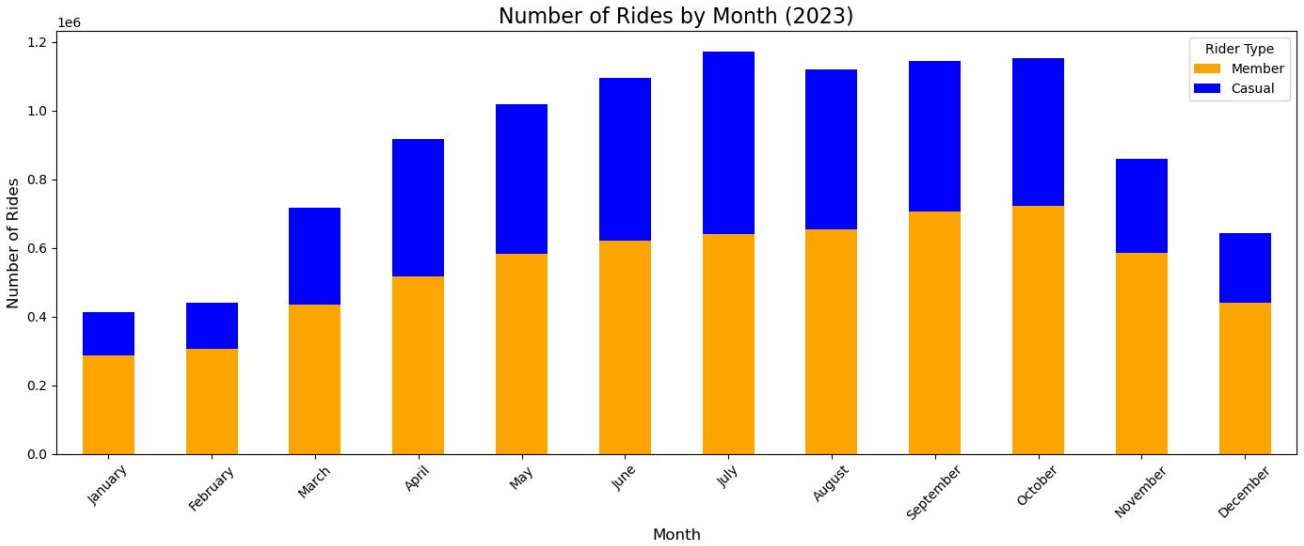
Growing trend Y-on-Y for total rides → Focus on **most recent** data (2023)

EDA | MEMBER VS. CASUAL

Seasonality!

Member ↗

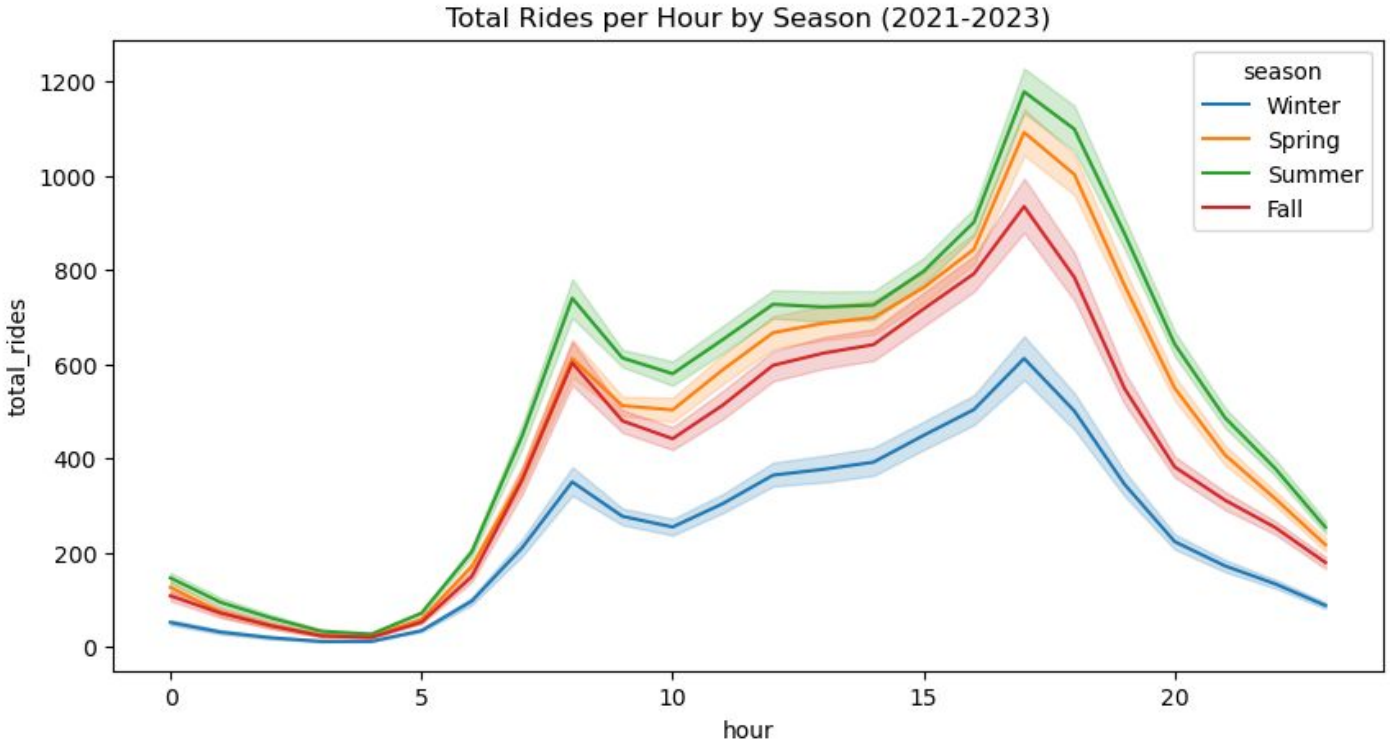
Casual ↑



EDA | TIME OF DAY & SEASON

2 Peaks ↑

Seasonality ↑



KPI | IDEA GENERATION

Customer satisfaction matters:

1. Bike availability at pickup
2. Empty docking unit at return

Reduce cost of **rebalancing stations** (daily, weekly)

Realign station **capacity** to actual demand

Systematic solution required: Analysis of **Turnover & Imbalance**

KPI | DEFINITION WITH S-M-A-R-T METHOD

Specific: Reduce number of stations with elevated imbalance levels for CBS

Measurable:

- Metric 1: Weekly (Im-)Balance at station level
- Metric 2: Weekly Turnover by station
- Metric 3: Responses to CBS's bi-annual questionnaire in section "Bike & Docking availability"

Attainable: Imbalanced Stations -10%, Customer Satisfaction +15%

Relevant: Rebalancing costs (driver, trucks, etc.), Customer Satisfaction

Time-bound: Within 12 months (Seasonality → Year-on-Year)

KPI | FORMULA | DEFINITION

*By **station** and **day**:*

BALANCE = COUNT of RETURNS - COUNT of PICKUPS

*By station and **aggregated** by **week** or **year**:*

CUMULATIVE BALANCE = COUNT of RETURNS - COUNT of PICKUPS

KPI | FORMULA | EXAMPLE

Selected Station: 11th & S St NW

	STATION	WEEK_NR	SEASON	LAT	LNG	COUNT_PICKUP		COUNT_RETURN		CUMULATIVE_BALANCE
1040	11th & S St NW	1	1	38.913761	-77.027025	832	-	848	=	16
1041	11th & S St NW	2	1	38.913761	-77.027025	864		792		-72
1042	11th & S St NW	3	1	38.913761	-77.027025	748		764		16
1043	11th & S St NW	4	1	38.913761	-77.027025	780		780		0
1044	11th & S St NW	5	1	38.913761	-77.027025	628		660		32

KPI | FORMULA | DEFINITION

*By station and **aggregated** by week or year:*

$$\text{IMBALANCE FACTOR} = \frac{\text{CUMULATIVE BALANCE}}{\text{CAPACITY*}}$$

IMBALANCE FACTOR	IMBALANCE CATEGORY
>= +0.5	TOO MANY (Bikes)
> -0.5 & <+0.5	BALANCED (Bikes)
<= -0.5	TOO FEW (Bikes)

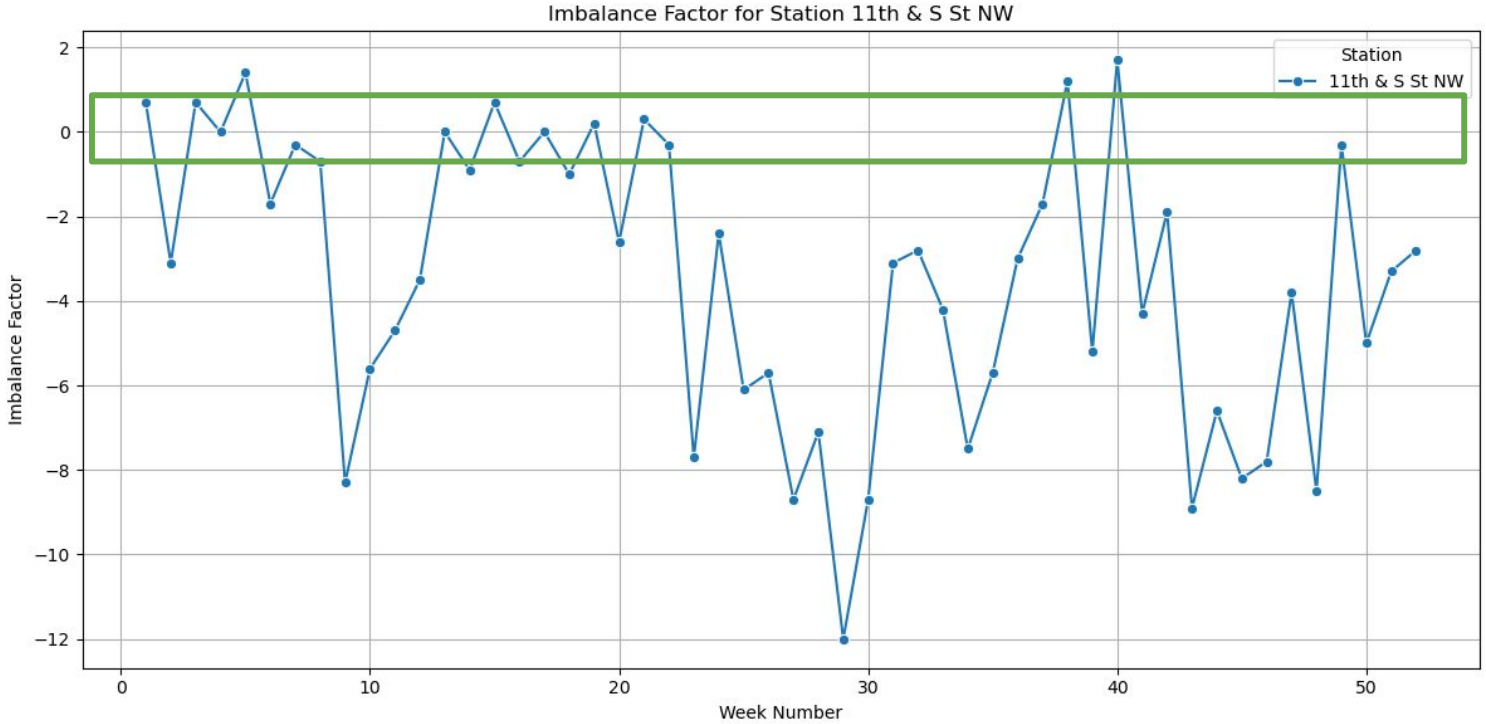
(*) CAPACITY = Number of docks per bike station [Range: from 9 to 55 docks]

KPI | FORMULA | EXAMPLE

Selected Station: 11th & S St NW

WEEK_NR	SEASON	LAT	LNG	COUNT_PICKUP	COUNT_RETURN	CUMULATIVE_BALANCE	CAPACITY	IMBALANCE_FACTOR	IMB_FAC_CATEGORY
1	1	38.913761	-77.027025	832	848	16 /	23 =	0.7 →	too_many
2	1	38.913761	-77.027025	864	792	-72	23	-3.1	too_few
3	1	38.913761	-77.027025	748	764	16	23	0.7	too_many
4	1	38.913761	-77.027025	780	780	0	23	0.0	balanced
5	1	38.913761	-77.027025	628	660	32	23	1.4	too_many

KPI | FORMULA | EXAMPLE



KPI | FORMULA | DEFINITION

By station and **aggregated** by week or year:

$$\text{TURNOVER} = \frac{\text{COUNT PICKUP}}{\text{CAPACITY}}$$

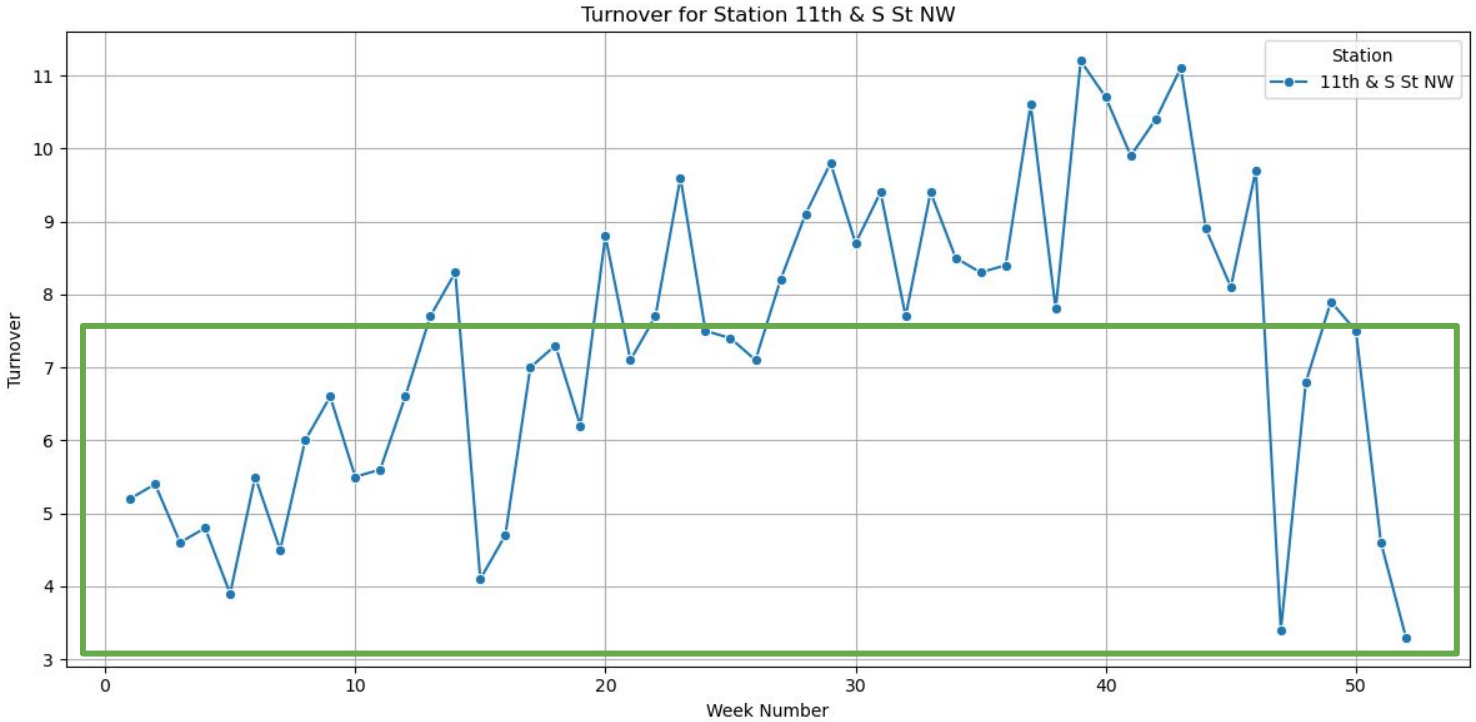
TURNOVER	TURNOVER CATEGORY
> 2	highly-utilized (station)
< 0.33 & <= 2	well-utilized
<= 0.33	under-utilized

KPI | FORMULAS | EXAMPLE

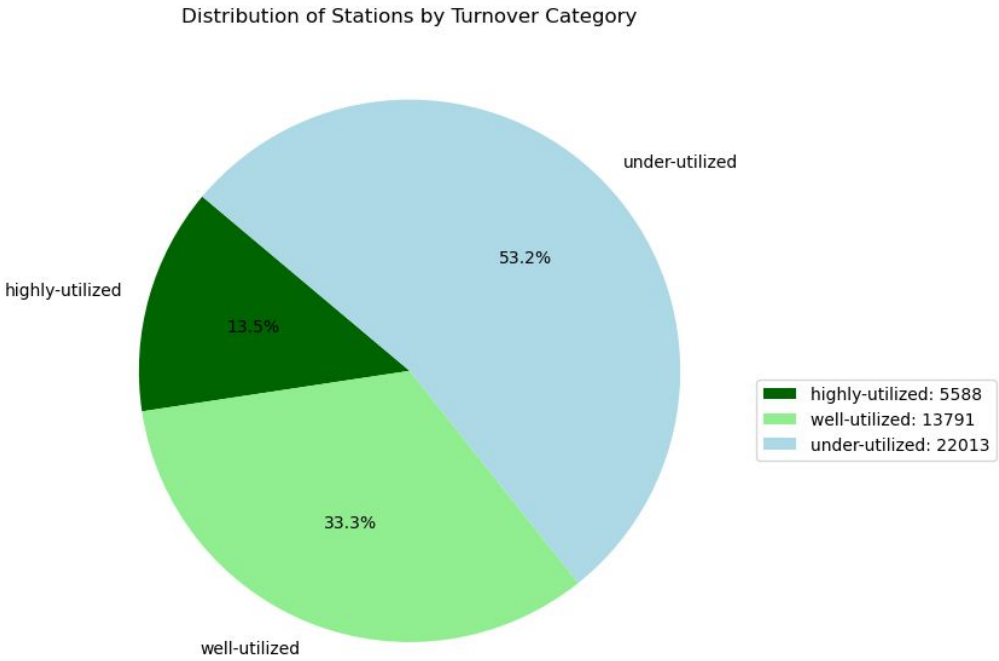
Selected Station: 11th & S St NW

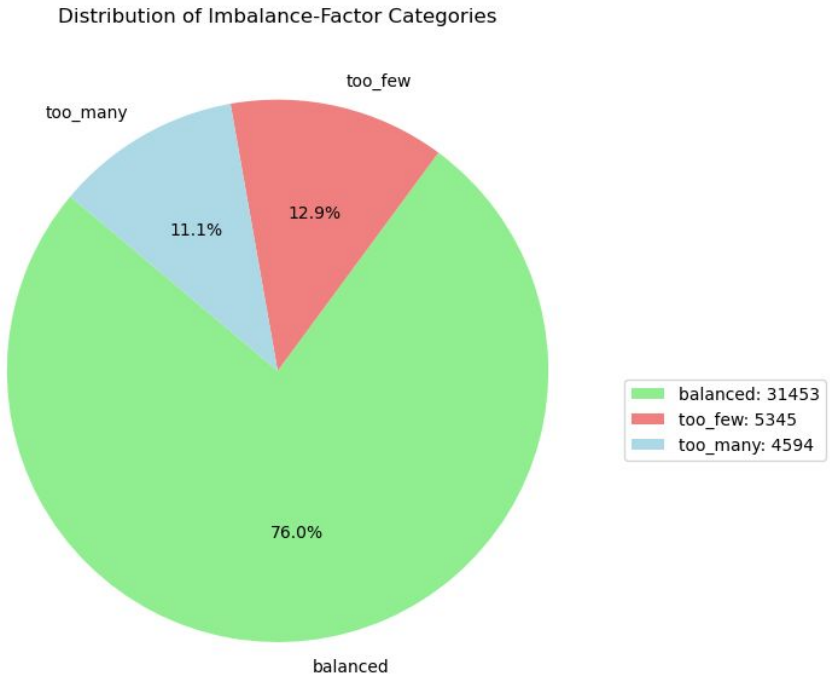
WEEK_NR	SEASON	LAT	LNG	COUNT_PICKUP	COUNT_RETURN	CUMULATIVE_BALANCE	CAPACITY	IMBALANCE_FACTOR	IMB_FAC_CATEGORY	TURNOVER	TO_CATEGORY
1	1	38.913761	-77.027025	832	/	7	/	23	=	5.2	highly-utilized
2	1	38.913761	-77.027025	864	792	-72	23	-3.1	too_few	5.4	highly-utilized
3	1	38.913761	-77.027025	748	764	16	23	0.7	too_many	4.6	highly-utilized
4	1	38.913761	-77.027025	780	780	0	23	0.0	balanced	4.8	highly-utilized
5	1	38.913761	-77.027025	628	660	32	23	1.4	too_many	3.9	highly-utilized

KPI | FORMULA | EXAMPLE



KPI | FIRST RESULTS *





(*) Based on weekly station data

KPI | CATEGORIES *

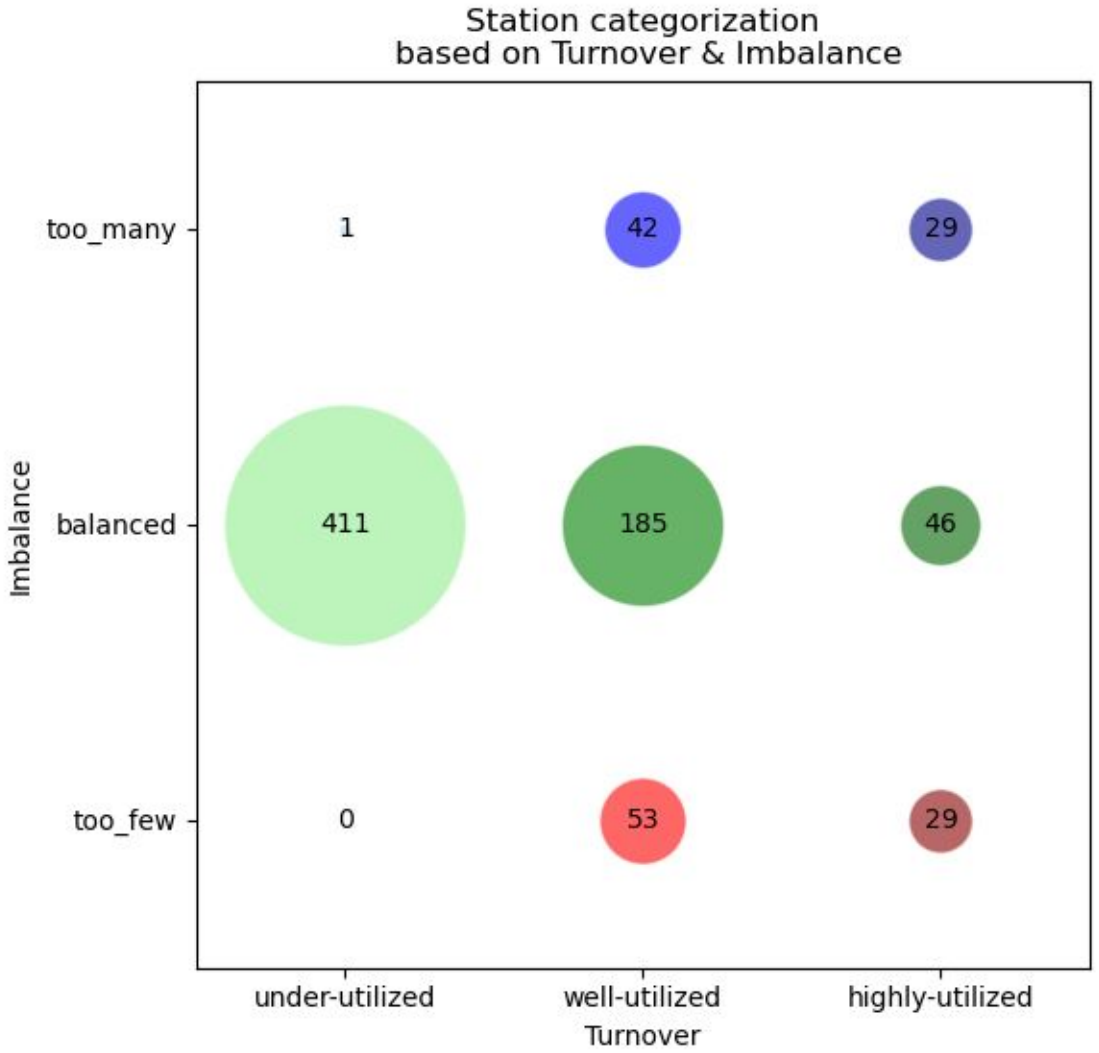
Legend:

- Too many bikes
- Balanced
- Too few bikes

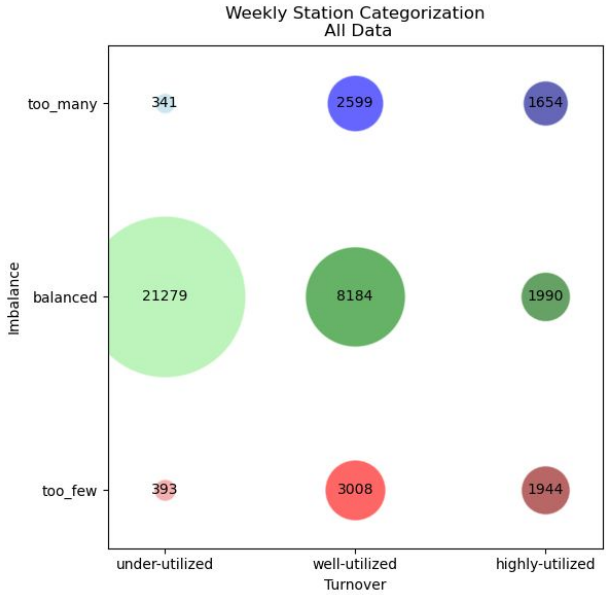
Priorities!

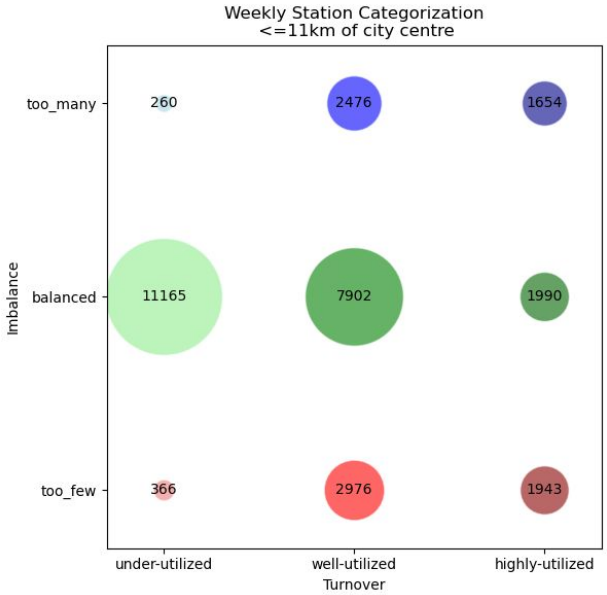
1. Too few & Highly-utilized
2. Balanced & Under-utilized

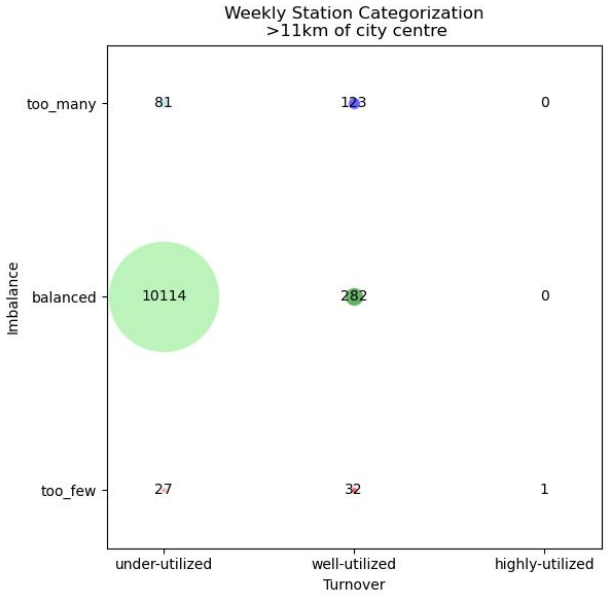
(*) Based on annualized station data



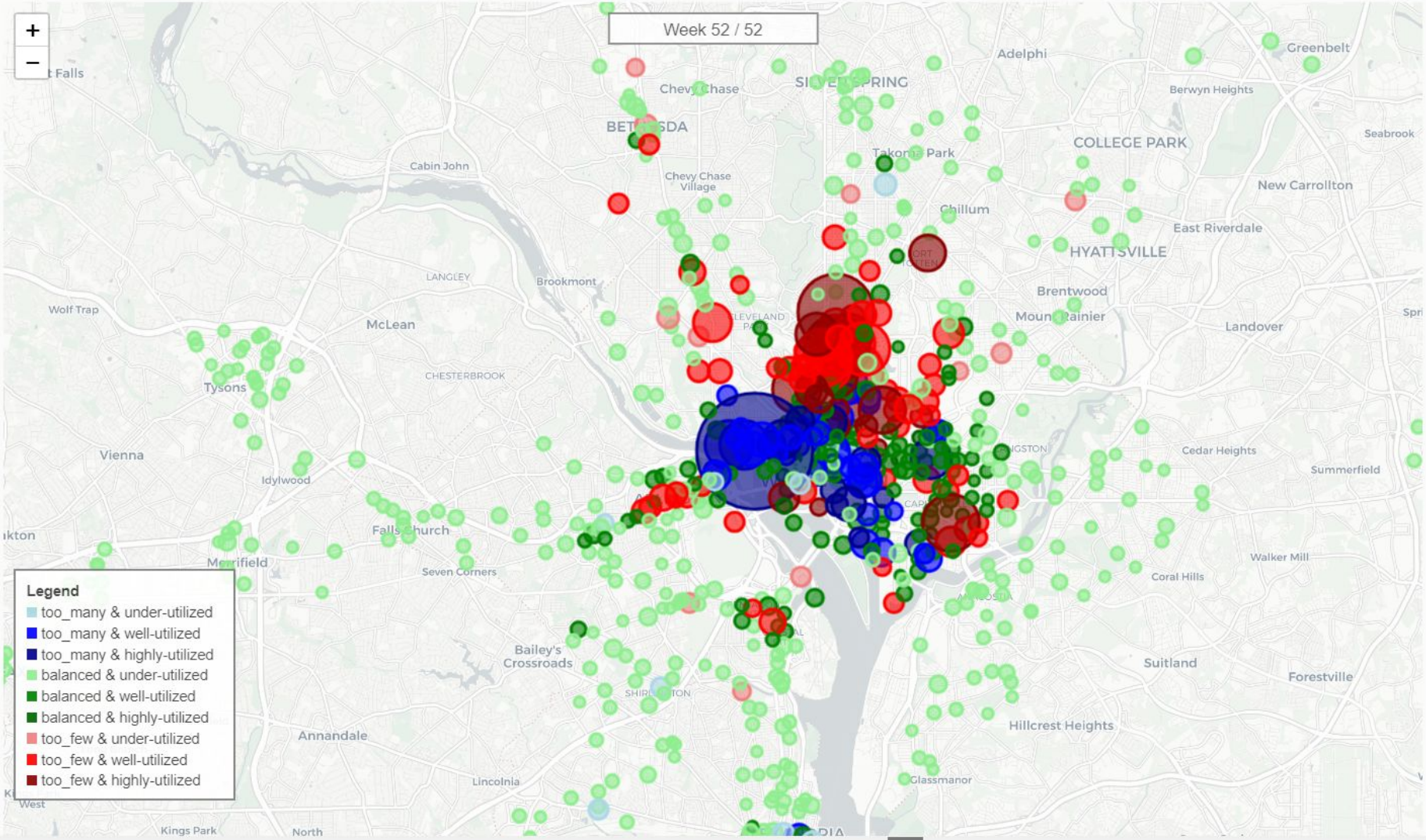
KPI | CATEGORIES *





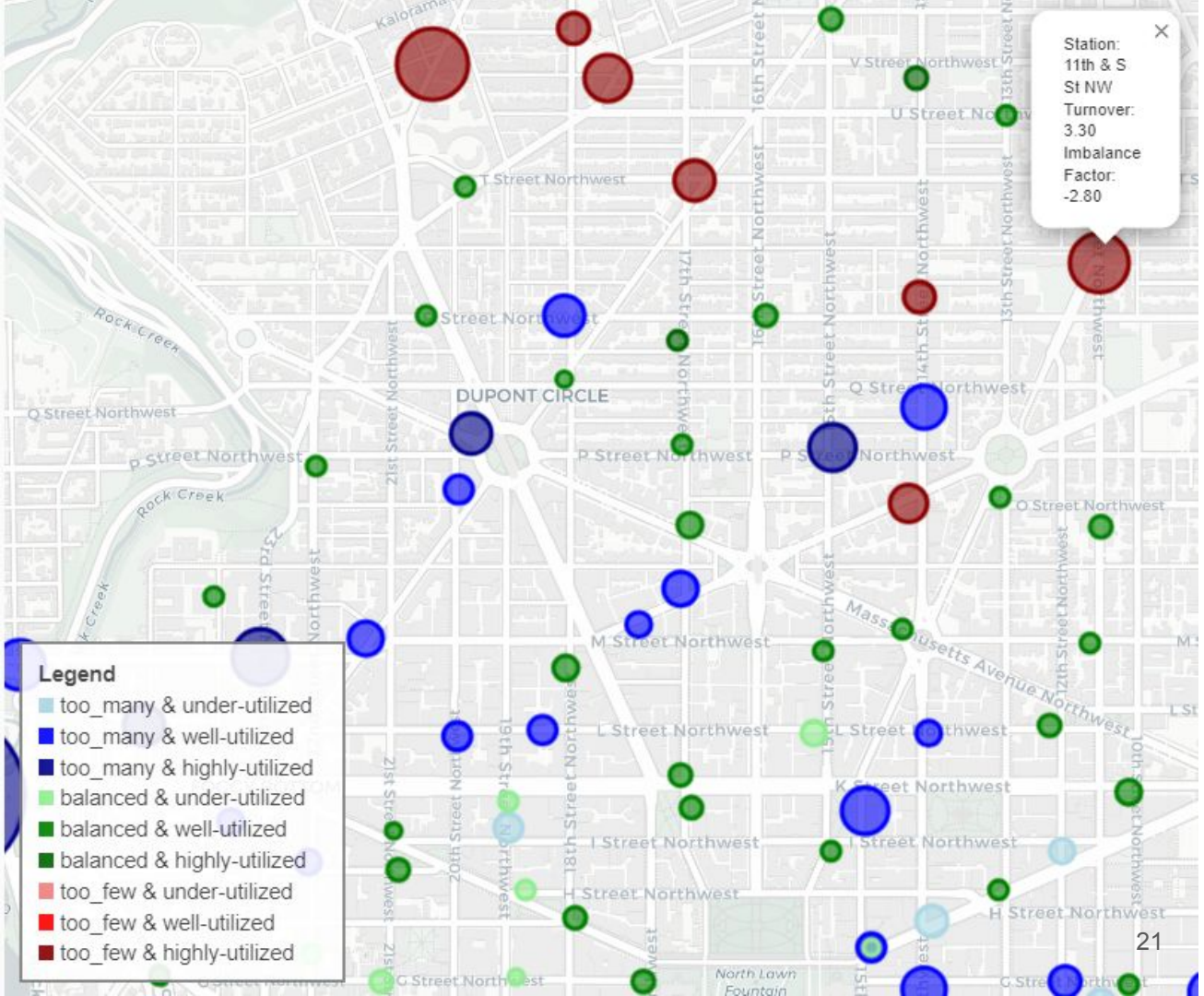


(*) Based on weekly station data


Week 52 / 52
Legend
too_many & under-utilized
too_many & well-utilized
too_many & highly-utilized
balanced & under-utilized
balanced & well-utilized
balanced & highly-utilized
too_few & under-utilized
too_few & well-utilized
too_few & highly-utilized
Adelphi
Greenbelt
Chevy Chase
Berwyn Heights
BETHESDA
COLLEGE PARK
Takoma Park
Seabrook
Cabin John
Chevy Chase Village
New Carrollton
Chillum
East Riverdale
HYATTSVILLE
LANGLEY
Brookmont
Brentwood
Wolf Trap
McLean
Landover
CHESTERBROOK
Tysons
Vienna
Idylwood
Cedar Heights
Summerfield
Falls Church
Merrifield
Seven Corners
Walker Mill
Coral Hills
Bailey's Crossroads
Suitland
Forestville
Annandale
Hillcrest Heights
Lincolnia
Glassmanor
Kings Park
North

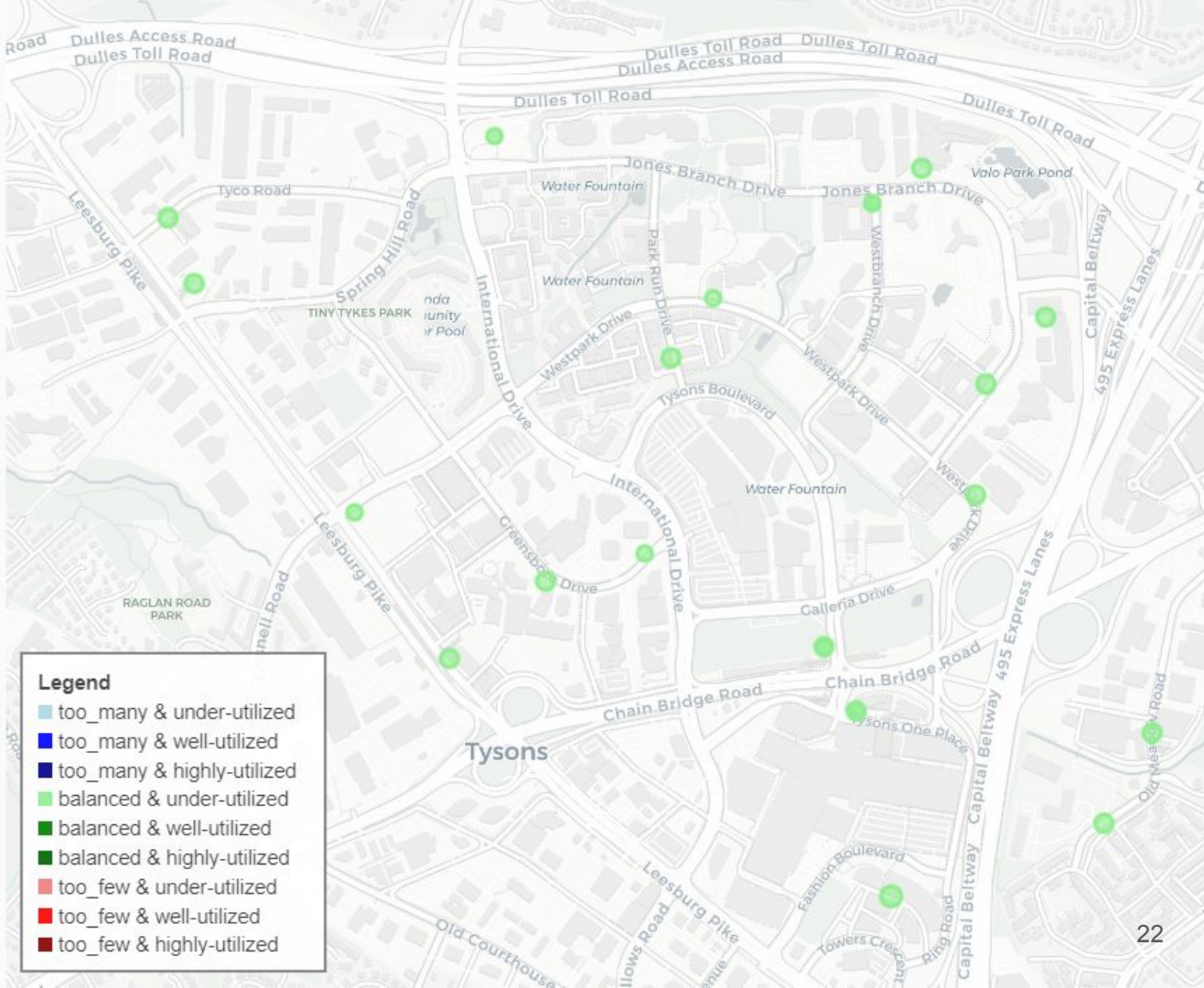
KPI | ZOOM 1

- Dupont Circle
- Imbalanced (!)
- Well-/Highly-utilized



KPI | ZOOM 2

- Tysons
- Balanced
- Under-utilized (!)



KPI | USER PATTERNS

Understanding user patterns leads to Feature Creation

Location proximity to city centre in general [in km]

Bike station near:

- Residential areas
- Workplaces
- Attractions
- Metrorail Stations
- Popular Venues

MACHINE LEARNING | INTRO

Plotted feature relationships, correlation matrix

Experimented with various Supervised Regression Models

Used Univariate, Polynomial, Multi-variate

ML | FEATURE EXPLORATION *

Left:

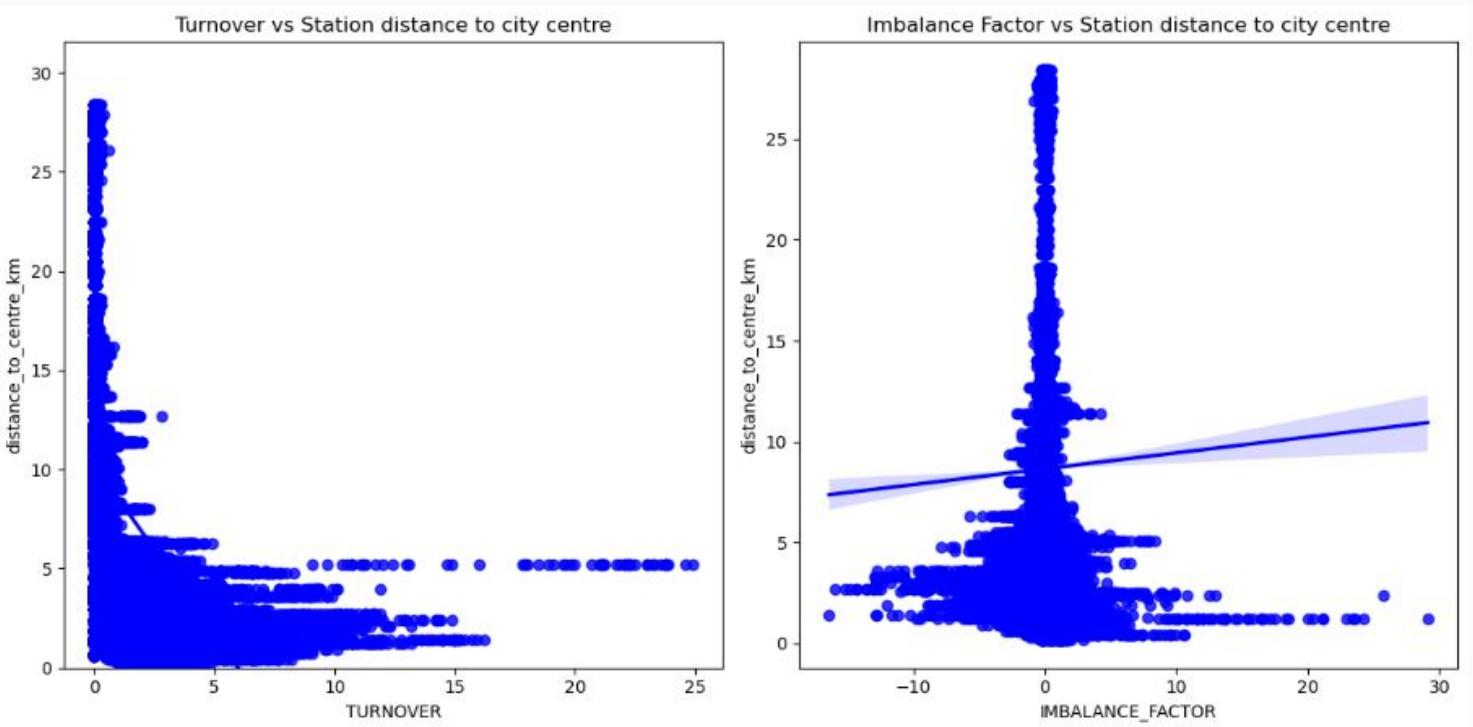
Distance ↑

Turnover ↓

Right:

Distance ↑

Imbalance ↓



(*) Based on weekly station data

ML | FEATURE EXPLORATION *

Left:

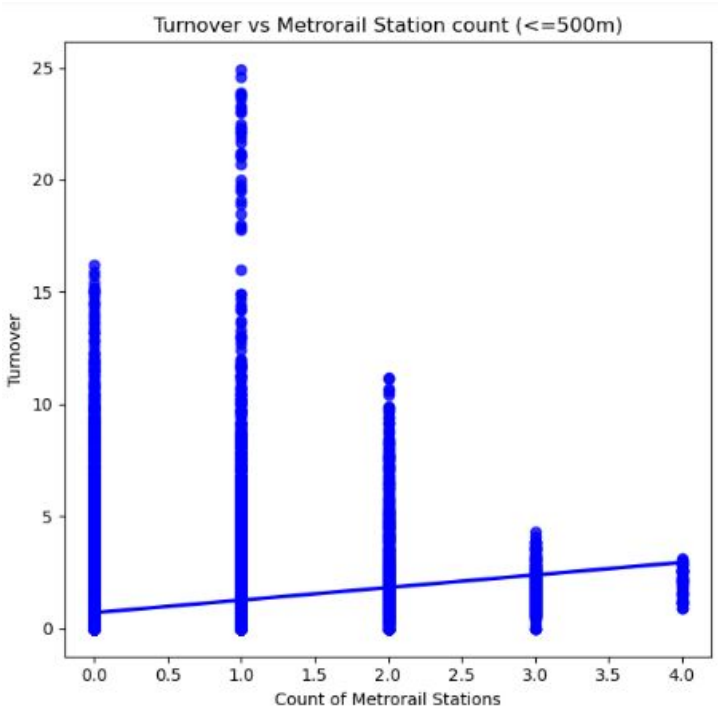
 # Metrorail ↑

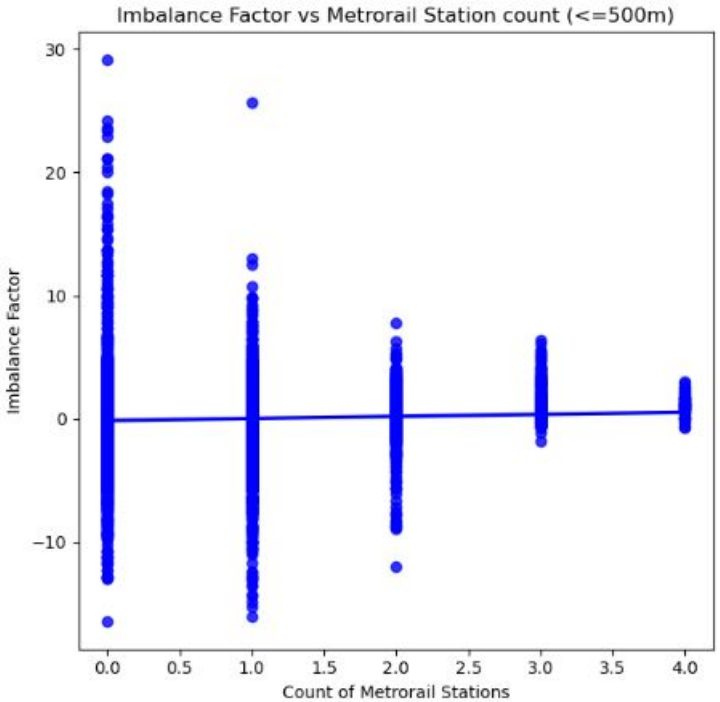
Turnover ↗

Right:

 # Metrorail ↑

Imbalance →





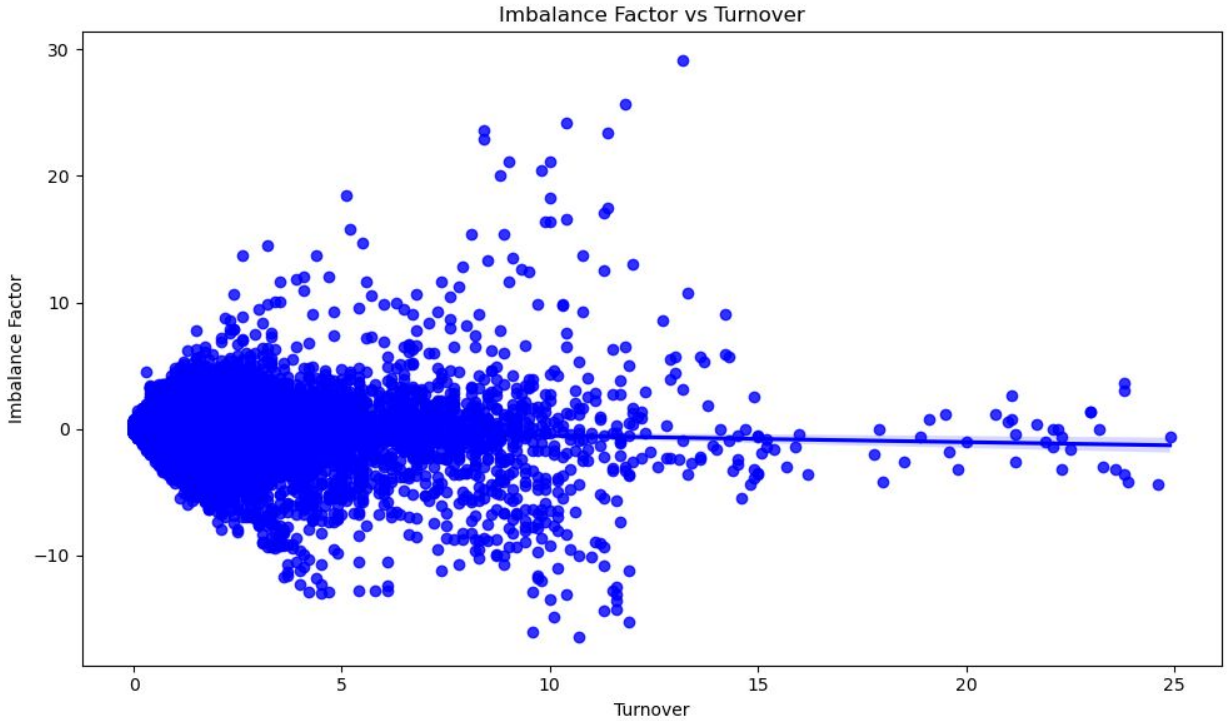
(*) Based on weekly station data

ML | FEATURE EXPLORATION *

Imbalance vs. Turnover

Turnover ↑
Imbalance (-)

Turnover ↓
Imbalance (+/-)



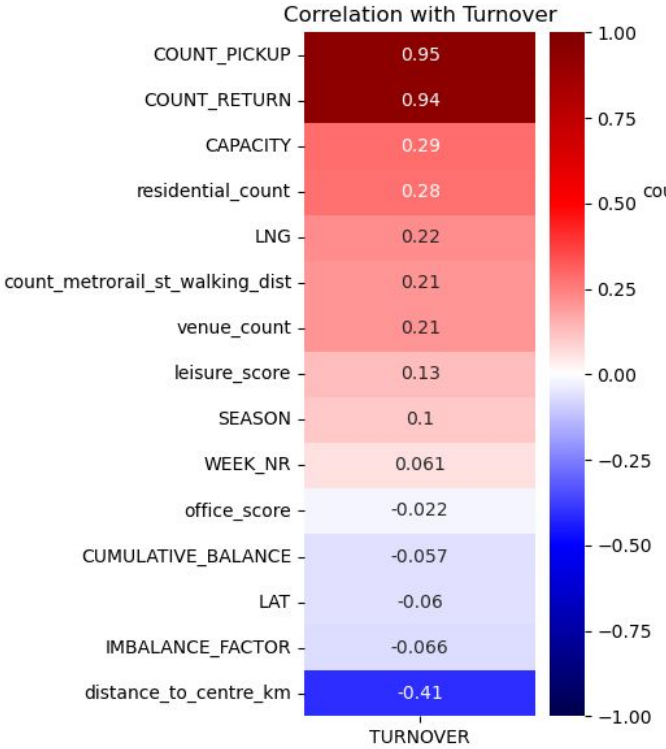
(*) Based on weekly station data

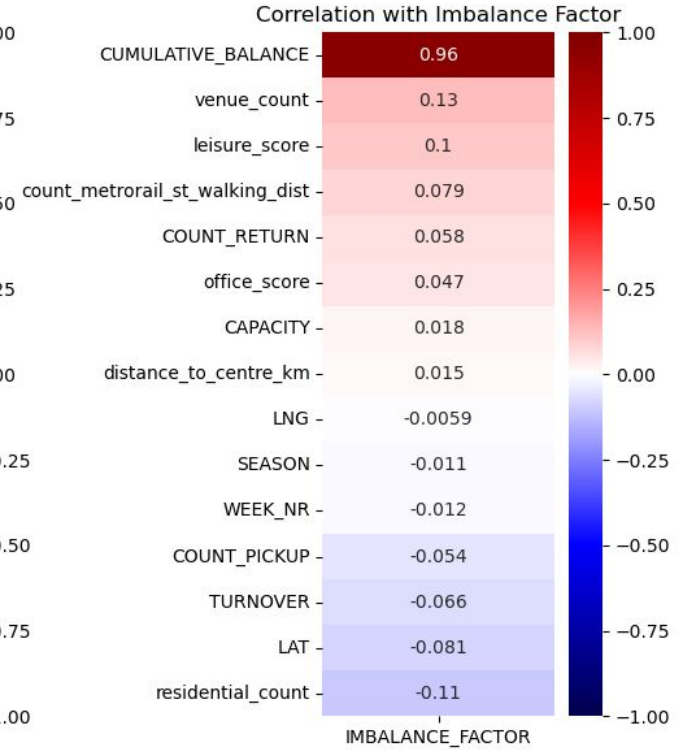
ML | FEATURE EXPLORATION *

Correlation Matrix
Focus: Turnover

(+) Capacity
(+) # Residential
(+) # Metrorail
(+) # Venue

(-) Distance to centre





(*) Based on weekly station data

ML | MODEL EVALUATION | METRICS

R^2 = R-Squared, Coefficient of Determination

MAE = Mean Absolute Error

MSE = Mean Squared Error

RMSE = Root Mean Squared Error

ML | MODEL EVALUATION * | TURNOVER

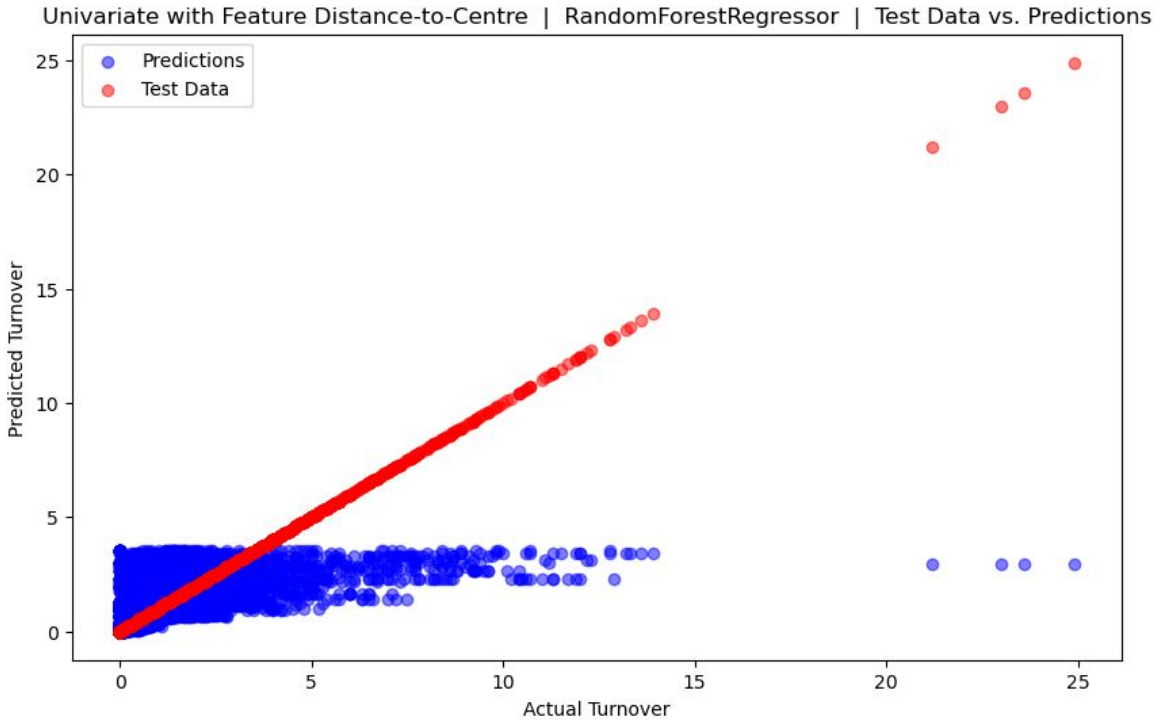
Features: 1

R^2: 36.0%

MAE: 0.63

MSE: 1.68

RMSE: 1.29



(*) Based on weekly station data

ML | MODEL EVALUATION * | TURNOVER

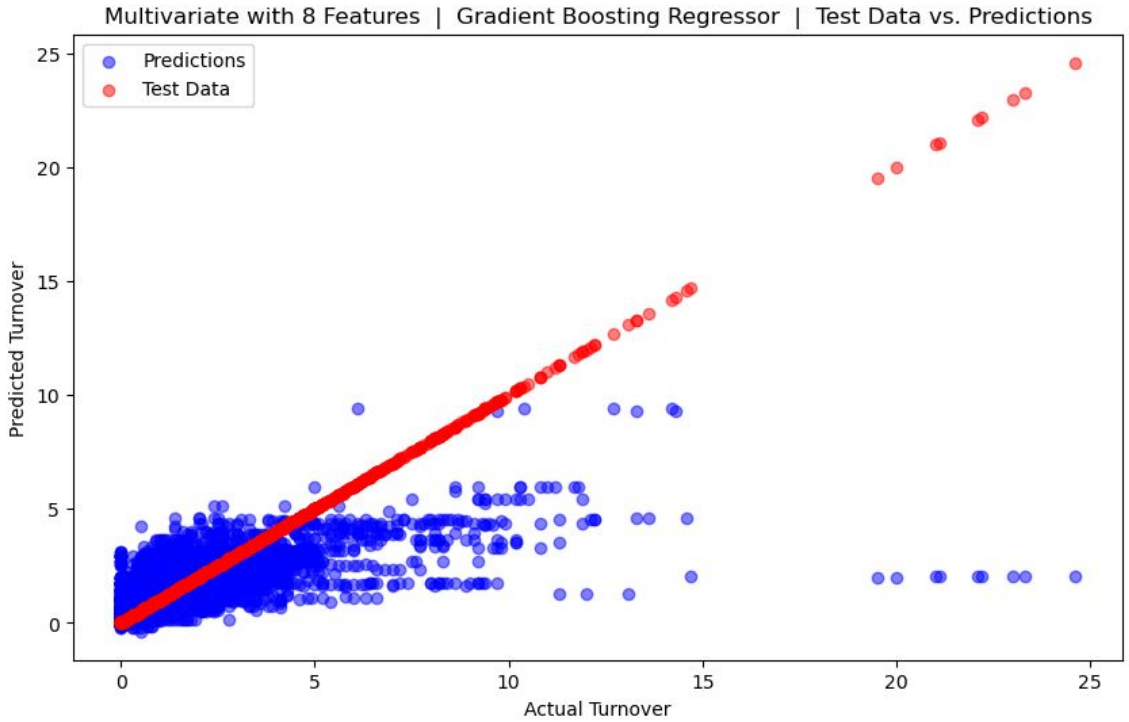
Features: 8

R^2 : 50.6% ↗

MAE: 0.55 ↘

MSE: 1.48 ↘

RMSE: 1.22 ↘



(*) Based on weekly station data

ML | MODEL EVALUATION * | TURNOVER

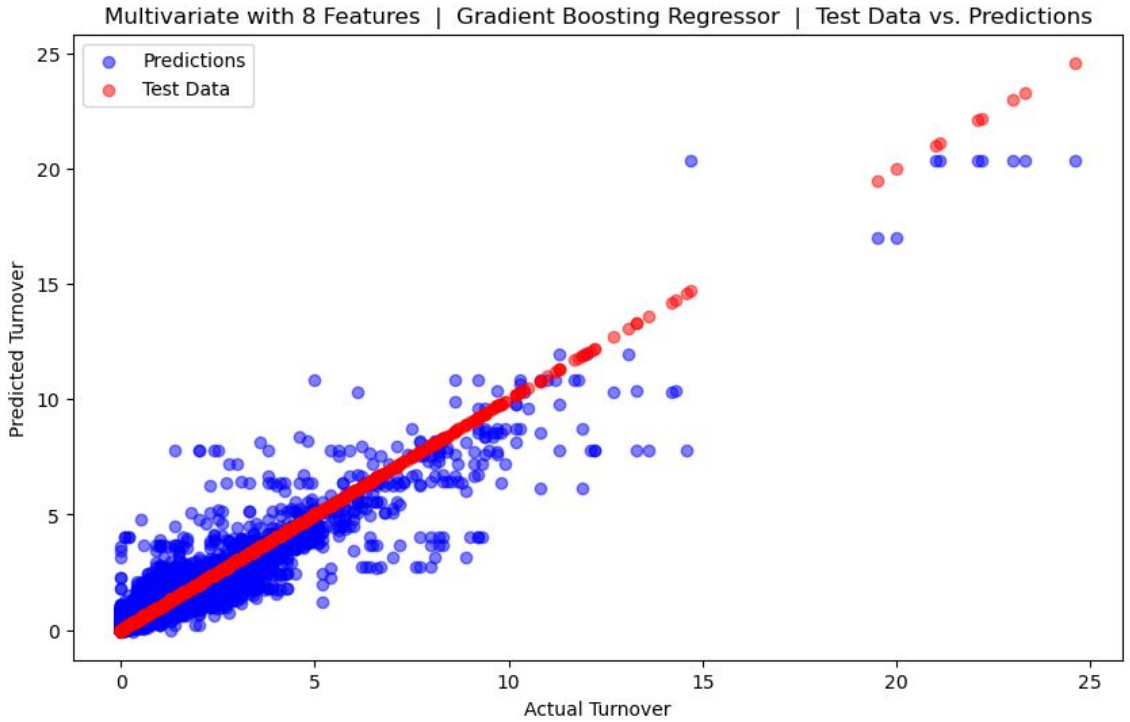
Features: 8
Hyperparameter-Tuning

R^2 : 88.2% ↑

MAE: 0.26 ↓

MSE: 0.35 ↓

RMSE: 0.59 ↓



(*) Based on weekly station data

ML | MODEL EVALUATION * | IMBALANCE

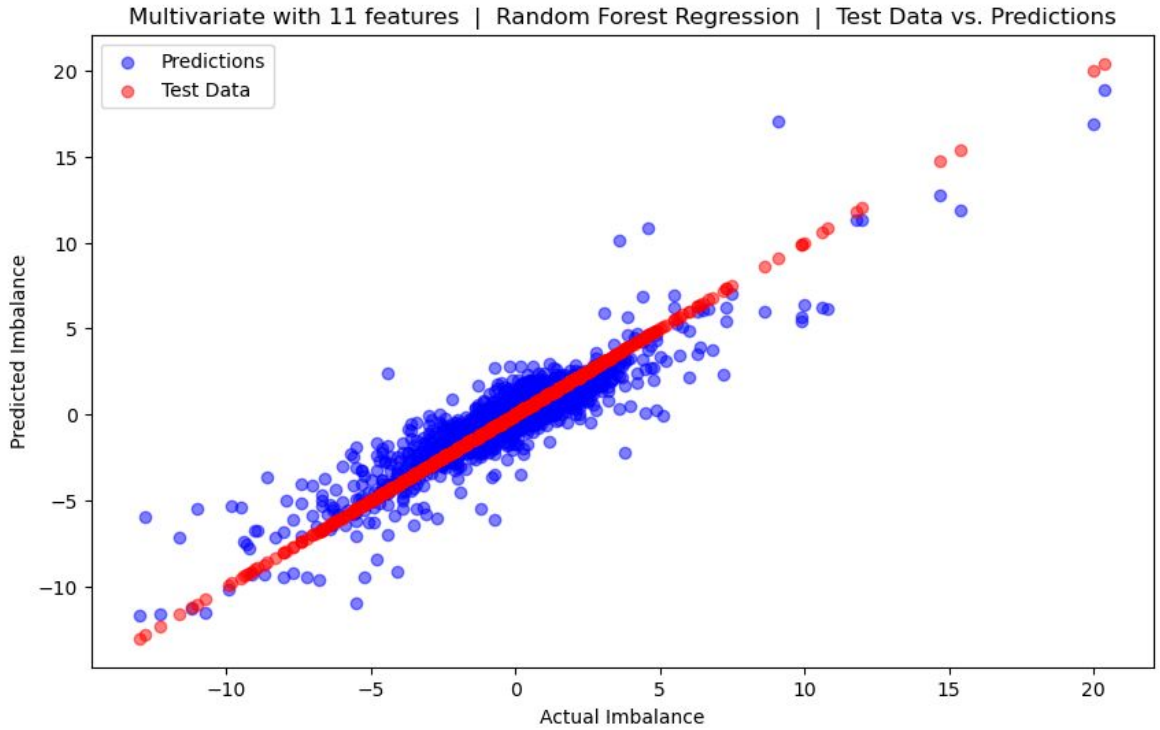
Features: 11
Hyperparameter-Tuning

R^2 : 81.3%

MAE: 0.27

MSE: 0.30

RMSE: 0.55



(*) Based on weekly station data

ACTION PLAN | PART 1

Department	Action	Deadline
Marketing	1. Develop incentive scheme for users returning to under-utilized station (discount, more time, other means of credit) 2. Promote physical rebalancing pilot to user - Motto: “You give us feedback, we do something about it!” 3. Develop image campaign “change your routine”, “ Re-discover your neighborhood”, use testimonials: “through CBS I discovered this shop in my neighborhood” → Appeal to consumers who want to shop locally & support local tradespeople.	15.04.2025 30.04.2025 30.05.2025
Sales	1. Identify local businesses around under-utilized stations in the city center 2. Negotiate product discounts with businesses to get CBS customers to park bikes there 3. Investigate current special event offers by museums, libraries, etc. and propose to CBS users via App to steer the return pattern of CBS customers	15.04.2025 30.04.2025 15.05.2025
IT / App Development	1. Develop app feature to direct customers towards under-utilized stations (Discount scheme to be provided by Marketing) 2. Develop app feature (‘Re-Discover your neighborhood’) 3. Develop dashboard for weekly KPI tracking	15.05.2025 30.05.2025 15.06.2025

ACTION PLAN | PART 2

Department	Action	Deadline
Legal	1. Review T&C with contractor for manual re-balancing of stations → Price structure might have to be amended, Required services and resources will likely be lower in the future	15.05.2025
Operations	1. Select focus area for pilot of physical capacity re-balancing. Develop project plan. 2. Execute physical re-balancing pilot. 3. Calculate required additional capacity required at key stations in city center. 4. Start to integrate progress-rebalancing via dashboard in weekly KPI review meeting	30.04.2025 30.05.2025 15.04.2025 25.06.2025

KEY LEARNINGS

Feature development : From idea to data w/ aid of Chat GPT

geo.py & folium powerful for allocation problem

Importance of hyperparameter calibration

WITH MORE TIME

Inclusion of feature proximity to hotels (casual users)

Min / Max / Avg Turnover or Imbalance per Station

More cross-checking of data in Chat GPT-generated lists

Analysis of cost structure of rebalancing

Feasibility of non-permanent docking stations for special events

THANK YOU

RESOURCES

Chat GPT → Prompting for Feature development

2014, Virginia Tech Study → Balancing in Bike-share schemes

2016, CBS User Survey → User patterns

CAB LMS Ressources → ML

iStock & Alamy → CBS-Photos

Google Maps

opendata.dc.gov → CBS Station Capacity, Congestion

BACK-UP

Movie time?

Siri! Tell me, if a station is 2km away from the city center, has 2 metro rail stations within 500m distance, … → What is the expected turnover?

INTRODUCTION OF DATASET

- ➔ Capital Bikeshare program in Washington DC
- ➔ Time period 2021 - 2023
- ➔ Three datasets:
- ➔ 1. Hourly, weather, casual/member (1.095 rows)
- ➔ 2. Daily, weather, casual/member (26.280 rows)
- ➔ 3. Daily, Start & End stations, casual/member (10.693.996
- ➔ rows)

KPI | USER PATTERNS → FEATURE CREATION

Understanding user patterns → FEATURE CREATION

Location proximity to city centre in general [in km]

Bike station near:

- Residential areas [top-20, within 500m of bike station]
- Workplaces [top-20 based on employee count, within …]
- Attractions [top-20 based on annual visitors, within …]
- Metrorail Stations [98 stations, within …]
- Popular Venues [top-20 based on annual guest count, within …]

KPI | FIRST RESULTS

Too many bikes

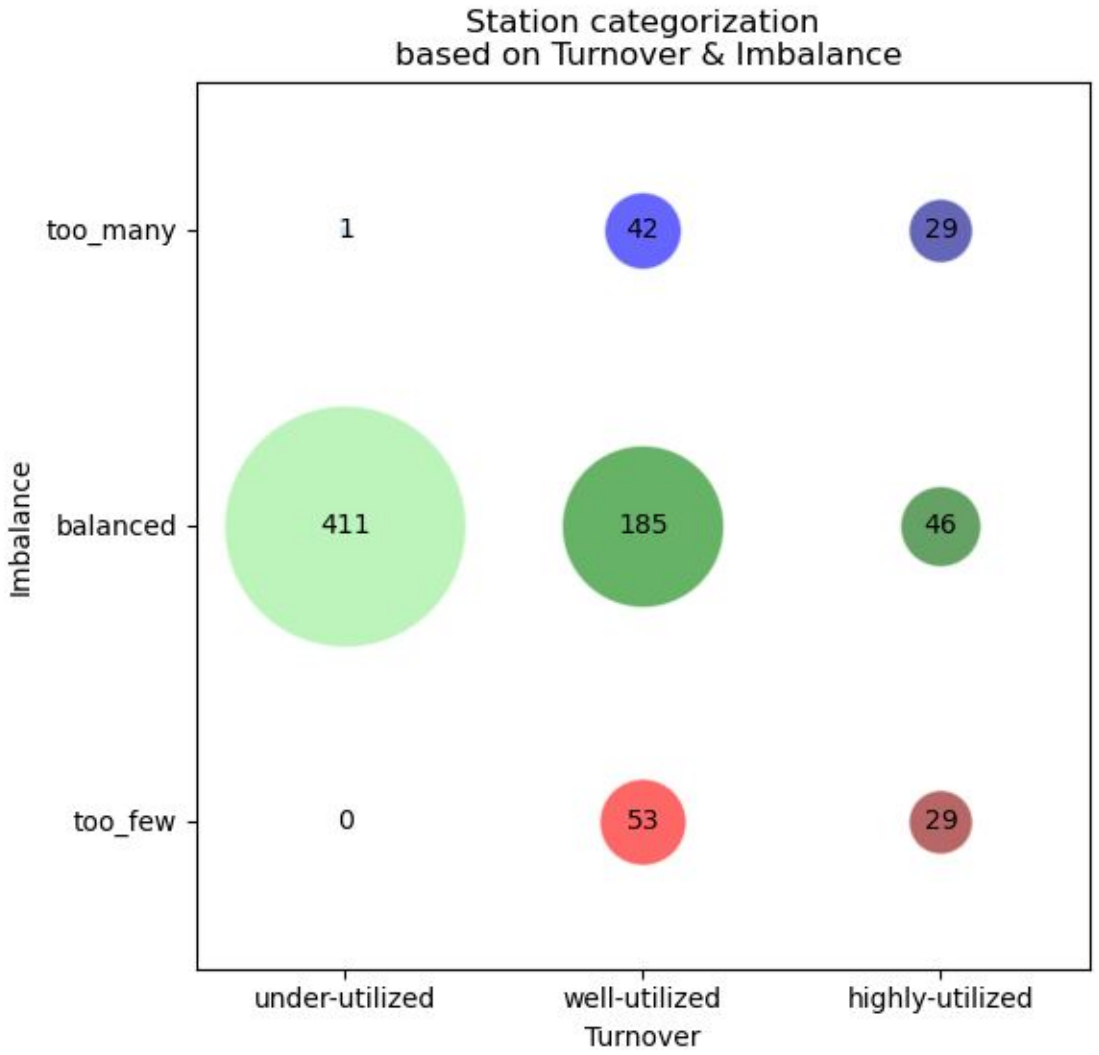
Balanced

Too few bikes

Priorities:

Too_few & Highly-utilized

Balanced & Under-utilized

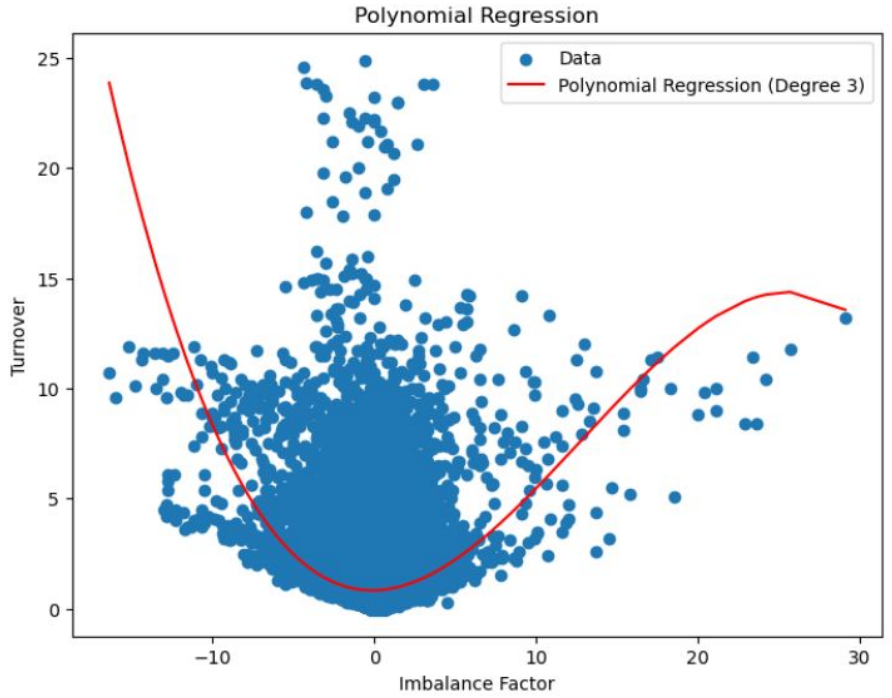


ML | FEATURE EXPLORATION

One

Two ↗

Three ↑



ML | SUMMARY

Univariate

→ Start w/ **annual** station data, **low** R^2 score = 19%

→ After switching to **weekly** data, increase to 36%

Polynomial, Univariate

→ seemed visually better fitted, however R^2 = 22%

Multivariate

→ with **8 location features,** R^2 jumped to 80%

→ highest score after **hyperparameter-tuning** = 88%