Optimizing Bike-Sharing Systems: A Machine Learning Approach to Predict Station Imbalances

Ruixuan Tu, Larissa Xia, Steven Haworth, Jackson Wegner, Yuzhe Zhang

GitHub: https://github.com/TURX/451_divvy_bike

INTRODUCTION

This study analyzes Divvy Bike Station, Trip Data, and American Community Survey Data to predict bike station flow imbalances (overflow/underflow). The key questions are: How can demographic data and machine learning predict bike availability? Is the status of existing stations a reliable indicator for nearby stations? Using Logistic Regression, Decision Tree, SVM for demographic data, and kNN for geographic data, with Recursive Feature Elimination and Grid Search with Cross-Validation, SVM was the most effective. The status of existing Divvy stations reliably predicts the status of nearby stations.

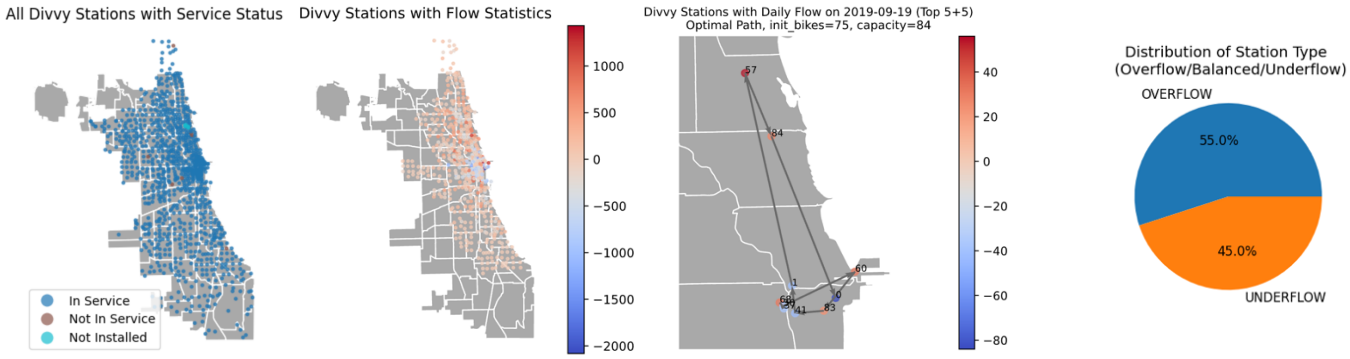
DATA

Divvy Bike Station Data (2022)

Sourced from the *Chicago Data Portal*, this dataset contains 1,419 observations and 8 variables on Divvy Bike stations in Chicago with no missing data or feature engineering needed.

Divvy Trips Data (2022)

This dataset, also from the *Chicago Data Portal*, includes 21.8 million records of Divvy bike-sharing trips with 18 variables (these variables not used in this project). Used to define station categories (overflow/balanced/underflow) based on one-day trip data, it has no missing data or need for feature engineering.



American Community Survey Data (2022)

The data for this project comes from the 2022 American Community Survey (ACS 1-Year Estimates) by the *US Census Bureau*, focusing on datasets: Age and Sex (S0101), Geographic Mobility (S0701), Means of Transportation to Work (S0802), and Commuting Characteristics by Sex (S0801). Filtered by Divvy bike station zip codes in Chicago and merged with Divvy Bike Station Data, it includes 68 observations and 1,054 variables on age, sex, population, education, income, transport modes, work locations, travel patterns, and housing.

Missing values were imputed using the **median** for numerical columns and the **mode** for categorical columns, and numerical values were **standardized** ($\mu = 0, \sigma = 1$).

Numeric Summary of Demographic Data

	mean	std	min	25%	50%	75%	max
total_population (count)	52016.18	22175.31	737.0	34459.0	51961.0	71192.0	92235.0
percent_under_18 (portion)	16.03	6.36	2.0	11.8	15.7	21.8	27.4
percent_21_and_over (portion)	80.15	7.01	68.6	74.8	79.6	84.8	96.4
percent_60_and_over (portion)	16.7	5.61	3.4	13.5	16.0	20.7	28.8
median_age (portion)	34.42	2.59	30.9	32.2	34.4	36.1	41.3
sex_ratio (count)	96.92	10.12	72.4	92.5	97.8	100.8	126.9
bachelors_degree (count)	11536.42	7259.53	142.0	5260.0	10049.0	18170.0	25966.0
graduate_degree (count)	9056.63	5505.2	278.0	4313.0	7542.0	12381.0	20163.0
income_75000_or_more (count)	12904.85	8315.2	279.0	6058.0	10714.0	18315.0	30306.0
below_poverty_level (count)	8130.58	4793.75	18.0	4667.0	7675.0	12050.0	20274.0
public_transportation (portion)	23.83	6.62	8.8	18.8	23.0	27.4	37.1
walked (portion)	9.36	10.23	0.8	2.8	4.8	12.0	49.8
bicycle (portion)	1.94	1.22	0.0	0.9	1.8	3.0	4.0
taxicab_motorcycle_other (portion)	2.23	1.08	0.0	1.6	2.1	2.4	6.2
worked_from_home (portion)	19.07	7.25	3.9	12.4	20.9	24.4	38.0
mean_travel_time (portion)	31.79	4.62	19.9	27.9	31.4	35.0	41.8
owner_occupied_housing (count)	22005.31	11726.05	283.0	12031.0	21784.0	31792.0	49814.0
renter_occupied_housing (count)	27701.05	11089.71	323.0	19175.0	29124.0	35333.0	46362.0
moved_within_same_county (portion)	13.78	3.22	6.0	11.2	14.6	16.6	19.5
moved_different_county_same_state (portion)	1.38	1.06	0.0	0.6	1.0	1.9	4.6
moved_different_state (portion)	4.15	3.13	0.2	1.6	3.3	6.6	12.4
moved_from_abroad (portion)	1.29	2.48	0.0	0.4	0.9	1.4	31.2

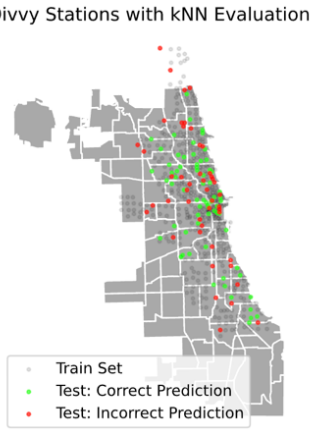
METHODOLOGY

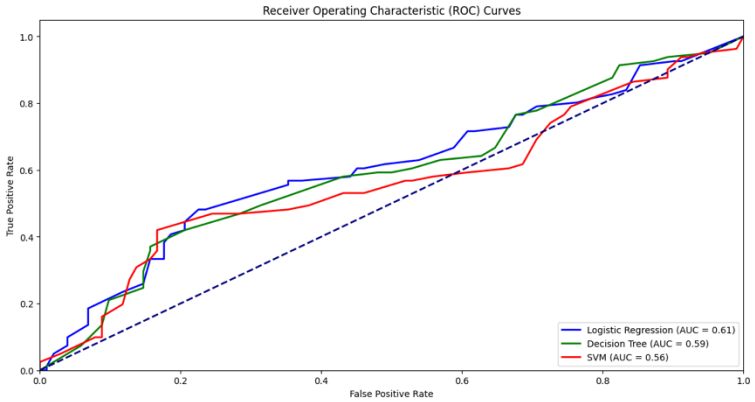
For k-Nearest Neighbors, latitude and longitude predict nearby station status, with hyperparameters optimized via **Grid Search** with **Cross-Validation**. Logistic Regression, Decision Tree, and SVM models use **Recursive Feature Elimination** and **Grid Search** with **Cross-Validation** for demographic data. **Recall** is prioritized to minimize false negatives, ensuring true overflow stations are identified, reducing maintenance costs.

RESULTS AND DISCUSSION

Logistic Regression selected features like population, age, education, income, and transportation types. With parameters `C=10` and `solver='saga'`, it achieved moderate performance, indicating potential underfitting (recall=0.55) due to its linear nature and inability to capture complex relationships.

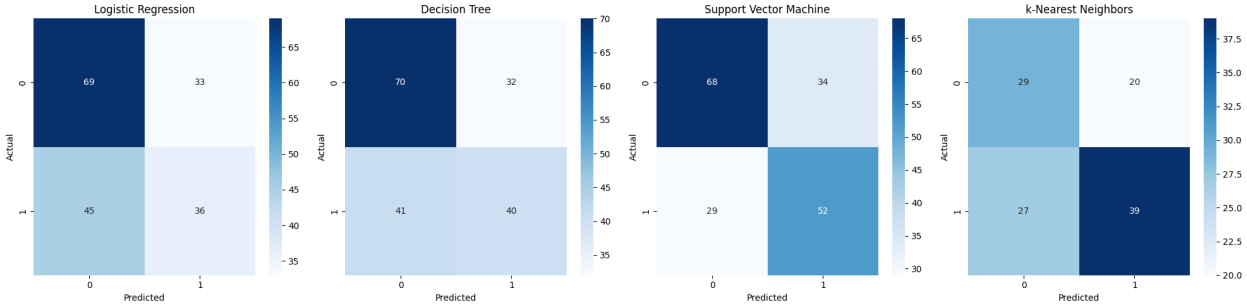
The Decision Tree model selected age, income, transportation types, work location, and housing information. Using `criterion='entropy'`, `max_depth=10`, and `min_samples_split=2`, it captured non-linear patterns better but showed moderate performance (recall=0.59), suggesting some overfitting control.





SVM selected population, education, income, transportation types, travel time, and housing information. With `C=0.1`, `gamma='scale'`, and `kernel='sigmoid'`, it achieved the best performance (recall=0.65), effectively handling non-linear decision boundaries and balancing complexity and generalization, making it the most robust model.

kNN has high precision but moderate recall, meaning it correctly identifies positive cases more often but misses some true positives (recall=0.59). Considering only latitudes and longitudes are used for modeling, we may conclude the **status of existing Divvy stations is a reliable indicator** for predicting the status of nearby stations.



Limitations include reliance on static demographic data, potential overfitting due to model complexity, and the exclusion of external factors like weather and special events, which could significantly impact bike usage.

CONCLUSION

Machine learning can use demographic data to predict Divvy Bike Station imbalances, with SVM being the most effective model. The status of existing stations reliably predicts nearby stations through kNN. Future work should incorporate dynamic factors like weather and address overfitting for improved accuracy and robustness.

Contributions

Member	Proposal	Coding	Presentation	Report
Ruixuan Tu	1	1	1	1
Larissa Xia	1	1	1	1
Steven Haworth	1	1	1	1
Jackson Wegner	1	1	1	1
Yuzhe Zhang	0	0	0.1	0

Notes:

- In the chart above, 1 = full contribution, 0.1-0.9 = partial, 0 = no contribution.
- Yuzhe Zhang has never attended group meetings and contributed to any part of the project.