Divvying Divvy Bikes

A Report from:

Larissa Xia
Ruixuan Tu
Steven Haworth
Yuzhe Zhang
Jackson Wegner

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

Background & Problem

- Chicago
- City Commuting
- Divvy Bike Rentals
- Crowded and Empty Stations

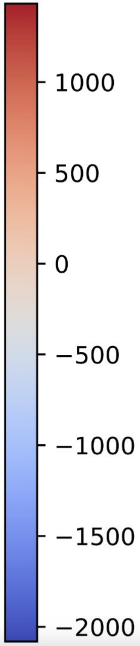
Addressing the Problem

- Balancing out inflows and outflows of rentals
- Helping the city plan for rental allocation
- Mapping out the optimal way of refilling low-inventory stations

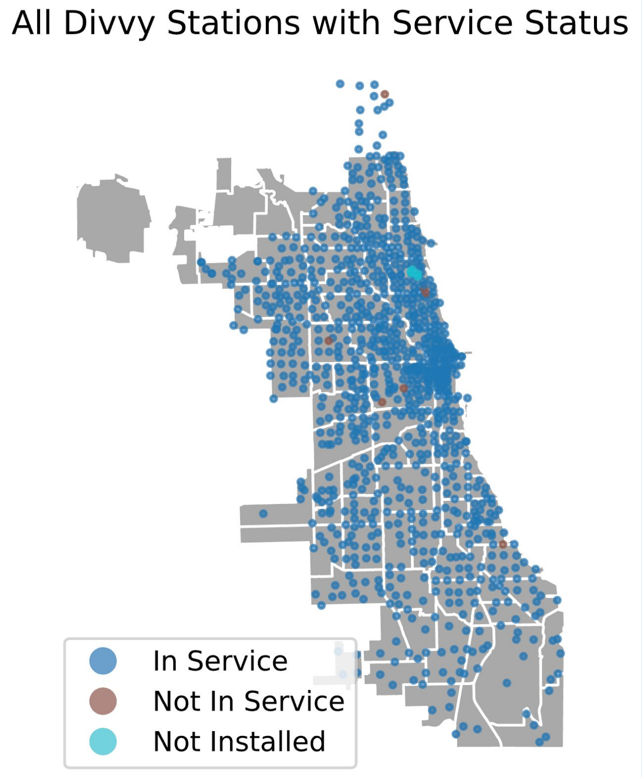
Overflow and Underflow Measurements

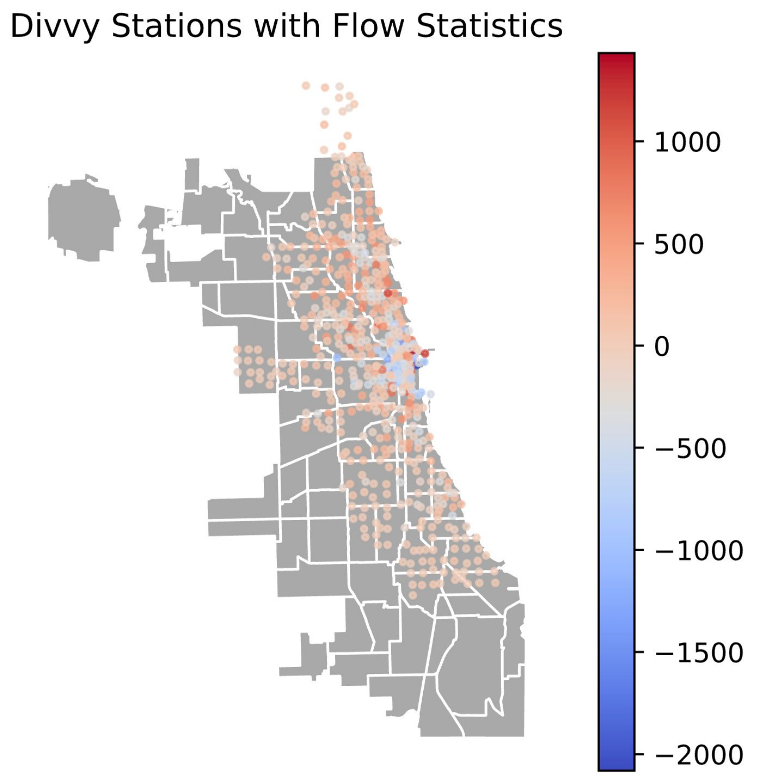
- Tracked flows from Chicago’s public dataset for Divvy Bikes
- Measured out stations net flow of bikes
- Marked category and extremity

	STATION ID	STATION NAME	DAILY FLOW
0	114	Sheffield Ave & Waveland Ave	56
1	91	Clinton St & Washington Blvd	31
2	35	Streeter Dr & Grand Ave	30
3	220	Clark St & Drummond Pl	27
4	90	Millennium Park	23
5	195	Columbus Dr & Randolph St	-84
6	287	Franklin St & Monroe St	-42
7	100	Orleans St & Merchandise Mart Plaza	-40
8	174	Canal St & Madison St	-38
9	191	Canal St & Monroe St	-38



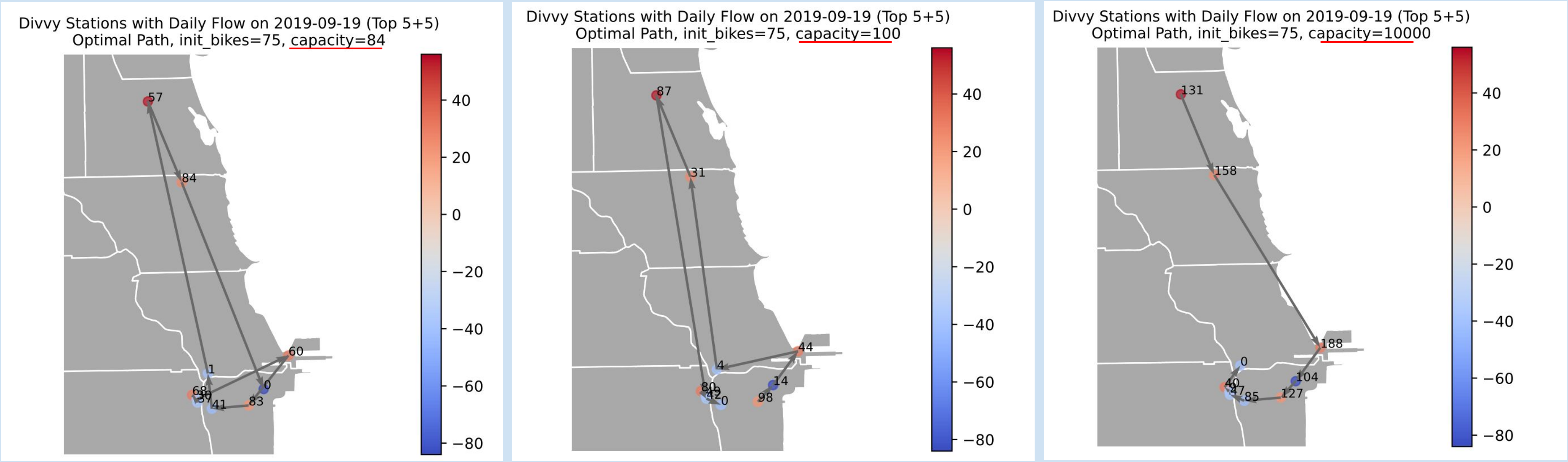
Mapped Out Groupings





Ideal Routing

Method: Breadth-First Search for Constrained Shortest Hamiltonian Path (the shortest path that visits every node once)
Init bikes: -sum(all bikes on graph) or 0 whichever smaller to avoid global deficiency
number of outside bikes to carry before visiting the first station
Capacity: maximum number of bikes the relocation truck can carry

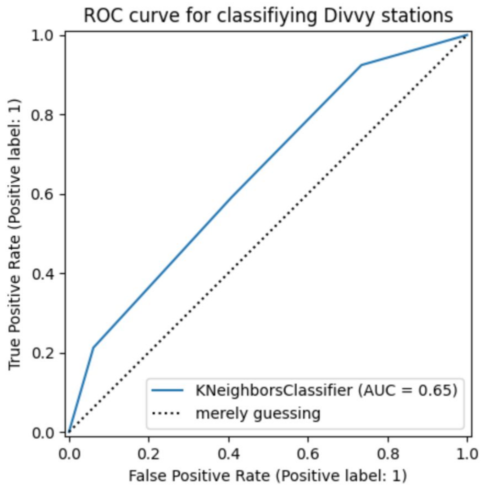


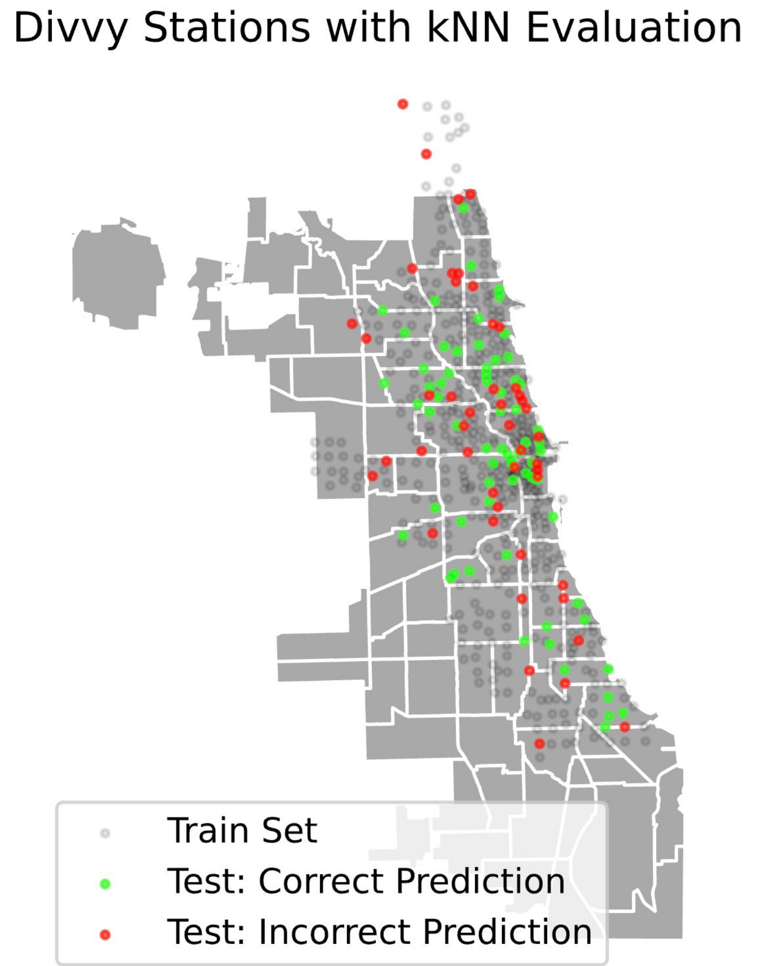
3-NN

X: Station ID, Longitude, Latitude
Y: Overflow (1) / Underflow (-1)

Metrics:

- **Accuracy:** 59%
- **Precision:** 66%
- **Recall:** 59%
- **F1:** 62%





Training Models Used & Effectiveness

Logistic Regression

Accuracy: 57%

Precision: 56%

Recall: 56%

F1: 56%

Support Vector Machine

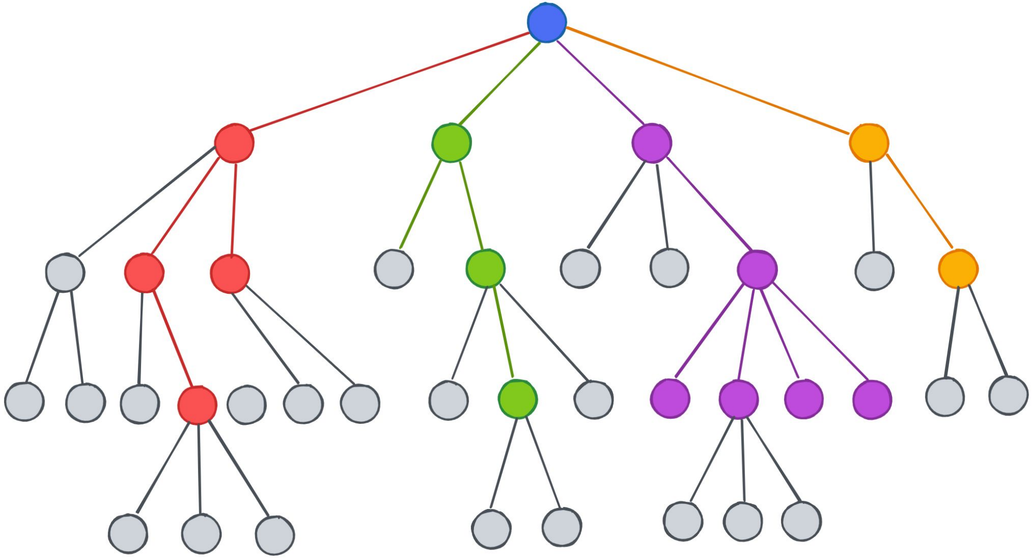
Accuracy: 66%

Precision: 65%

Recall: 65%

F1: 65%

Decision Tree



Underflow

Accuracy: 79%

Precision: 61%

Recall: 35%

F1: 87%

Overflow

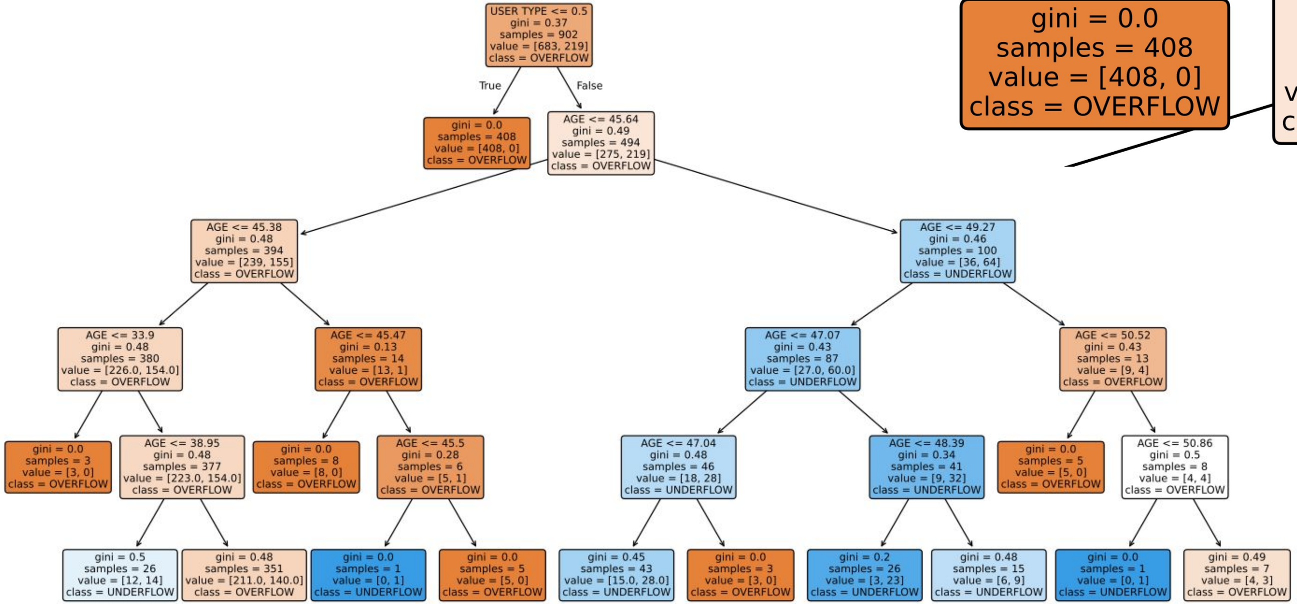
Accuracy: 79%

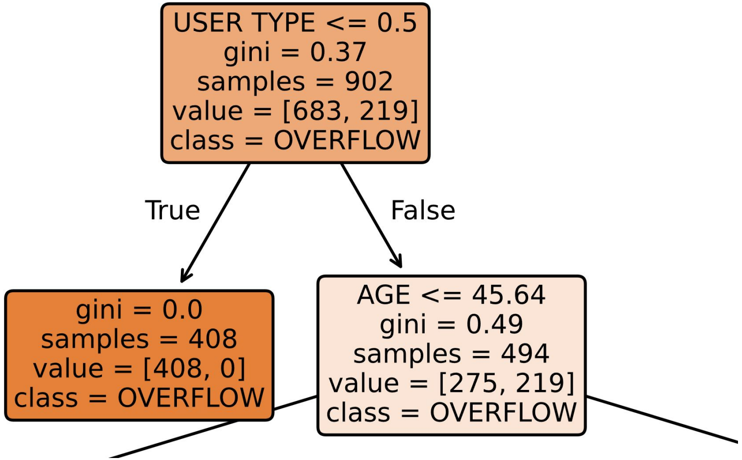
Precision: 82%

Recall: 93%

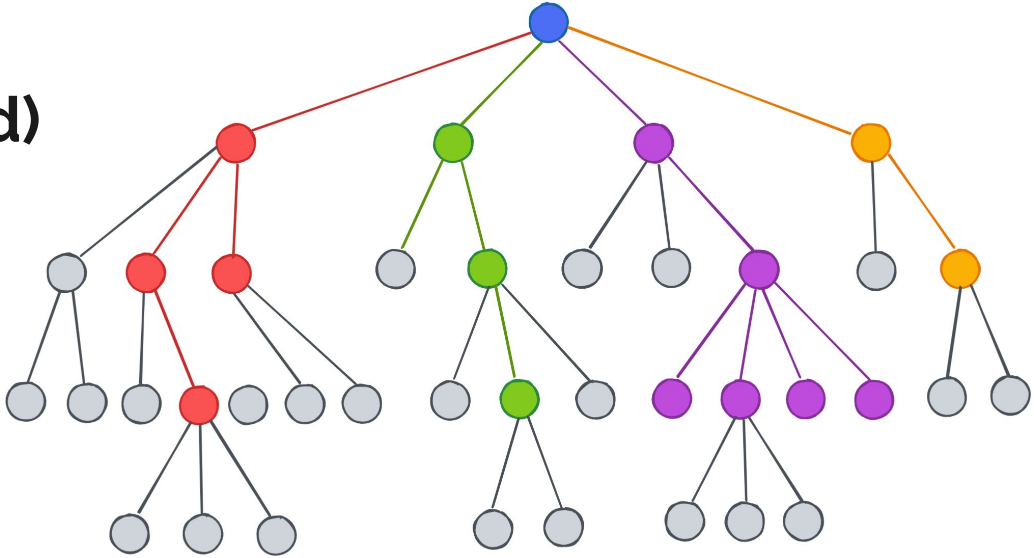
F1: 87%

Decision Tree Visualization





Decision Tree (oversampled)



Underflow

Accuracy: 64%

Precision: 66%

Recall: 35%

F1: 46%

Overflow

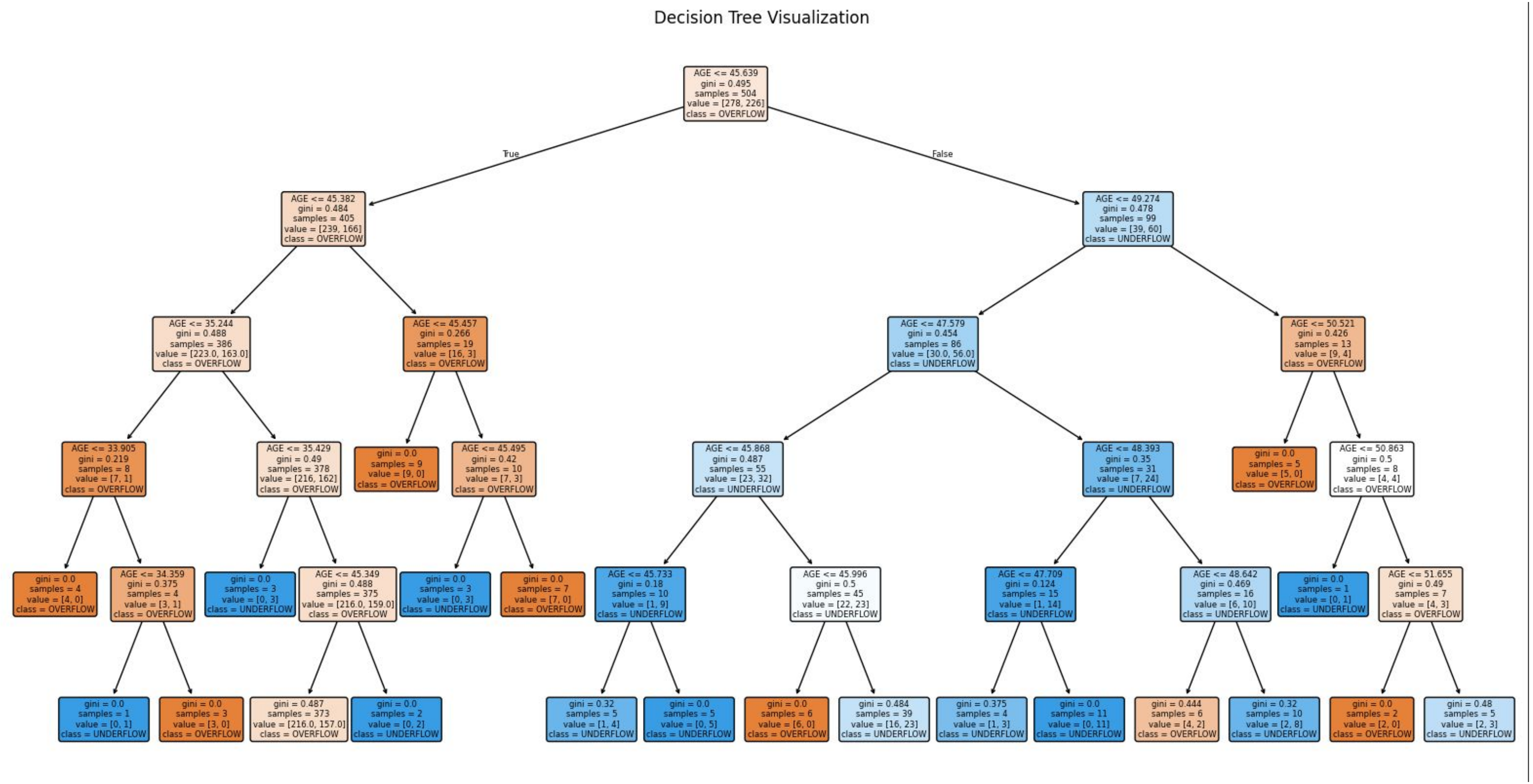
Accuracy: 64%

Precision: 63%

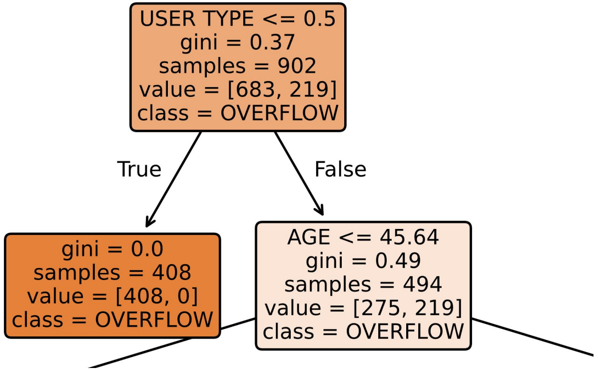
Recall: 87%

F1: 73%

Decision Tree Visualization (Oversampled)



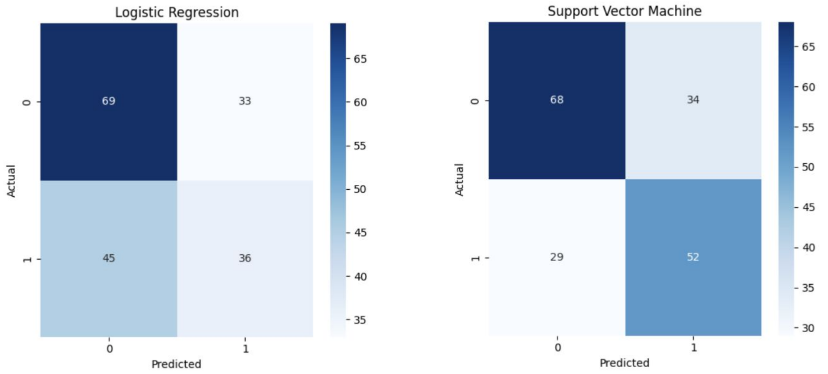
Accuracy and Metrics Explained



- Why is Recall on UNDERFLOW low?
  - What insights does it provide about Divvy-stations?
  - What did we do to provide balance ?
- Oversampling
  - Our dataset may lack pertinent information to classify UNDERFLOW
  - Increasing model decision boundary complexity yielded no real change
  - Other hyper-parameters

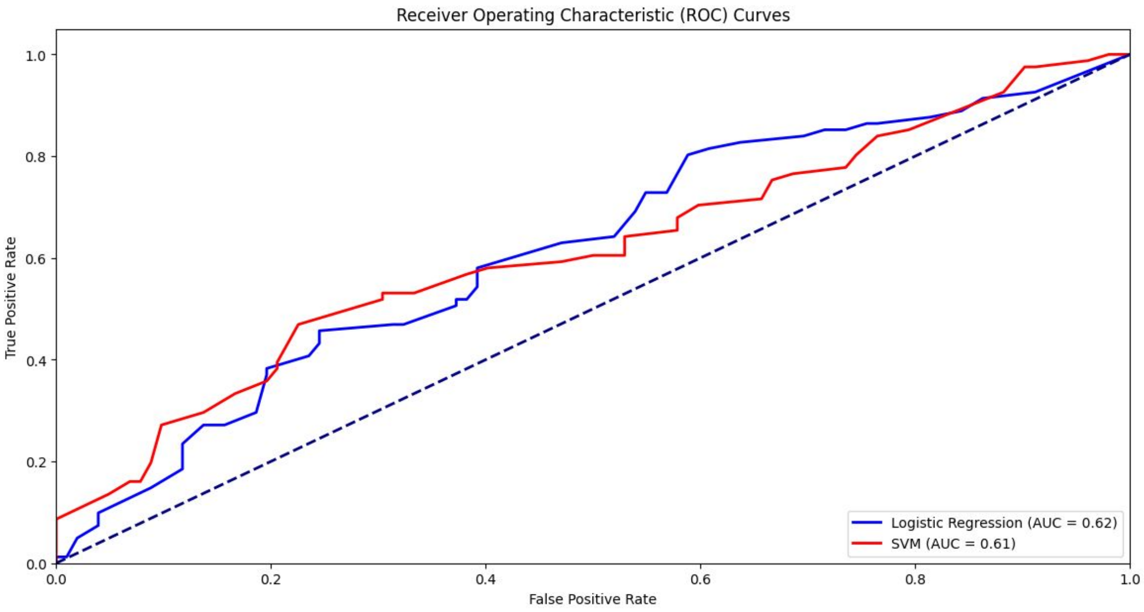
Logistic Regression and Support Vector Machine

Feature	Logistic Regression	SVM
Total Population	✓	✓
Percent Under 18	✓	
Percent 21 and Over	✓	
Percent 60 and Over	✓	
Median Age	✓	
Graduate Degree	✓	✓
Income $75,000 or More	✓	✓
Walked	✓	✓
Taxicab, Motorcycle, Other	✓	
Moved from Abroad	✓	✓
Bachelors Degree		✓
Mean Travel Time		✓
Owner-Occupied Housing		✓
Renter-Occupied Housing		✓
Moved Different State		✓



Metric	Logistic Regression	SVM
Best Parameters	{'C': 10, 'solver': 'saga'}	{'C': 0.1, 'gamma': 'scale', 'kernel': 'sigmoid'}
Accuracy	57.38%	65.57%
Precision	56.35%	65.28%
Recall	56.05%	65.43%
F1 Score	55.94%	65.31%
AUC	0.62	0.61

ROC/AUC



Receiver Operating Characteristic (ROC) Curves
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Logistic Regression (AUC = 0.62)
SVM (AUC = 0.61)

Conclusion

- **How can we use demographic data and machine learning algorithms to predict bike availability at Divvy stations in Chicago over a defined period?**

  Demographic Data around the Bike Station could be used to predict Bike Station Overflow/Underflow. However, due to the static nature of demographic data, they may not be very effective.

- **Is the status (overflow, underflow, balanced) of existing Divvy stations a reliable indicator for predicting the status of nearby stations?**

  The status of existing stations a reliable indicator, but external factors like weather and special events could be included to improve performance.