Predicting Trip Destinations with BIXI Data

Concordia University
SOEN 499

April 8th, 2020

Krishna Patel 40031019
Manpreet Singh 27517378
Derek Yu 40022110

Introduction

Presentation of the problem

- BIXI trip history dataset (2014 to 2019)
- Predict the destination of a trip based on:
  - Starting location
  - Time
  - *Extra: weather data*
- Classification vs. Regression

Materials and Methods

- ➢ **Dataset**
- ➢ Technologies
- ➢ Data Preprocessing
- ➢ Algorithms

Two types of CSVs per year:

- Monthly trip histories
  - *start_date*
  - *start_station_code*
  - *end_date*
  - *end_station_code*
  - *duration_sec*
  - *is_member*
- Bike Stations
  - *code*
  - *name*
  - *latitude*
  - *longitude*

Materials and Methods

- ➢ Dataset
- ➢ **Technologies**
- ➢ Data Preprocessing
- ➢ Algorithms

- Apache Spark
  - Using the Python library *pyspark*
  - Data preprocessing
  - Machine learning with MLlib
- Folium
  - Python data visualization library for creating Leaflet.js maps
  - Helpful for viewing bike stations when evaluating predictions.

Materials and Methods

- ➢ Dataset
- ➢ Technologies
- ➢ **Data Preprocessing**
- ➢ Algorithms

- Replace start/end station codes with the coordinates of their corresponding station.
- Skip samples with missing features.
- Break down *start_date* into hour, day of the week, and month.
- Encode the hour to account for its cyclical nature.

—

Dealing with Imbalanced Data

Classifying into clusters:

- Use random undersampling to balance the number of trips ending in each cluster.

Unsampled data

End Cluster	Count
1	947,379
2	25,010,384
3	131,229
4	15,998
5	3,178,611
6	5,532,257
7	12,283,466
8	8,444,028
9	13,358
10	1,173,774

Resampled with 1:2 Ratio

End Cluster	Count
1	26,556
2	26,630
3	26,467
4	15,998
5	26,878
6	26,739
7	26,711
8	26,626
9	13,358
10	26,882

Materials and Methods

- ➢ Dataset
- ➢ Technologies
- ➢ Data Preprocessing
- ➢ Algorithms

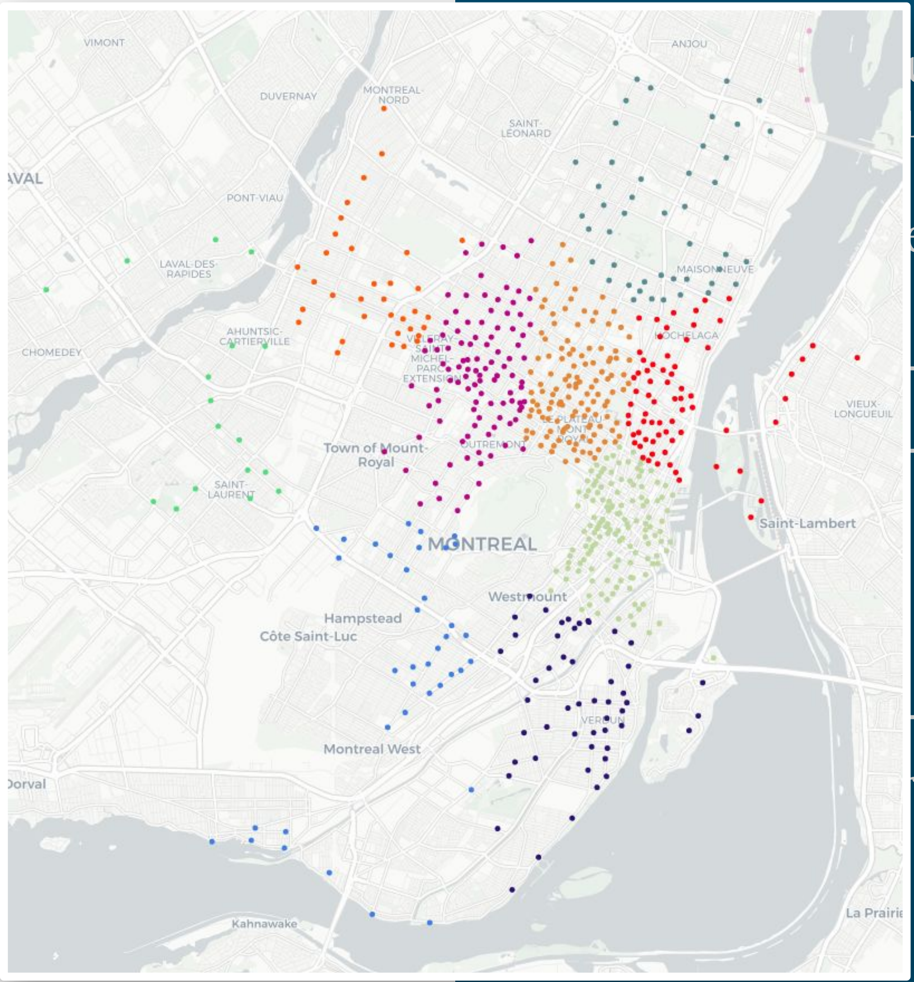
- K-Means Clustering
  - With K = 10
- Decision Tree
  - Features:

Classification	Regression
❑ *start_name/ start_cluster* ❑ *month* ❑ *day_of_week* ❑ *hour_sin* ❑ *hour_cos*	❑ *start_longitude* ❑ *start_latitude* ❑ *month* ❑ *day_of_week* ❑ *hour_sin* ❑ *hour_cos*

- Random Forest
  - Ensemble of 20 trees

Materia
Meth

- ➢ Dataset
- ➢ Technolog
- ➢ Data Prep
- ➢ **Algorithm**



ustering

10

ee

Regression

- ❏ *start_longitude*
- ❏ *start_latitude*
- ❏ *month*
- ❏ *day_of_week*
- ❏ *hour_sin*
- ❏ *hour_cos*

rest

of 20 trees

Results

- ➢ **Classification**
- ➢ Regression

Ran 4 versions of the classifiers to compare performance:

1. Using individual start station to predict end station
2. Using start cluster to predict an end cluster
3. Using start cluster and temperature to predict end cluster
4. Using resampled clusters to predict end cluster

Classification Metrics

Version 1	Decision Tree	Random Forest
Accuracy	1.88%	2.19%
Precision	0.82%	1.43%
Recall	1.89%	2.20%
F1-Score	0.53%	0.83%

Version 2	Decision Tree	Random Forest
Accuracy	52.90%	51.25%
Precision	52.65%	50.52%
Recall	52.87%	51.20%
F1-Score	52.48%	48.22%

Version 3	Decision Tree	Random Forest
Accuracy	52.87%	49.83%
Precision	52.64%	45.05%
Recall	52.92%	49.88%
F1-Score	52.50%	45.66%

Version 4	Decision Tree	Random Forest
Accuracy	52.77%	51.54%
Precision	61.72%	61.86%
Recall	52.66%	51.08%
F1-Score	55.30%	53.09%

Results

- ➢ Classification
- ➢ **Regression**

Ran 2 versions of the regressors to compare performance:

1. Using provided features from data to predict latitude and longitude
2. Using temperature to predict latitude and longitude

Regression Metrics

Version 1	Decision Tree	Random Forest
Lat. RMSE	0.01496307668	0.01597791819
Long. RMSE	0.01744732295	0.01804911382

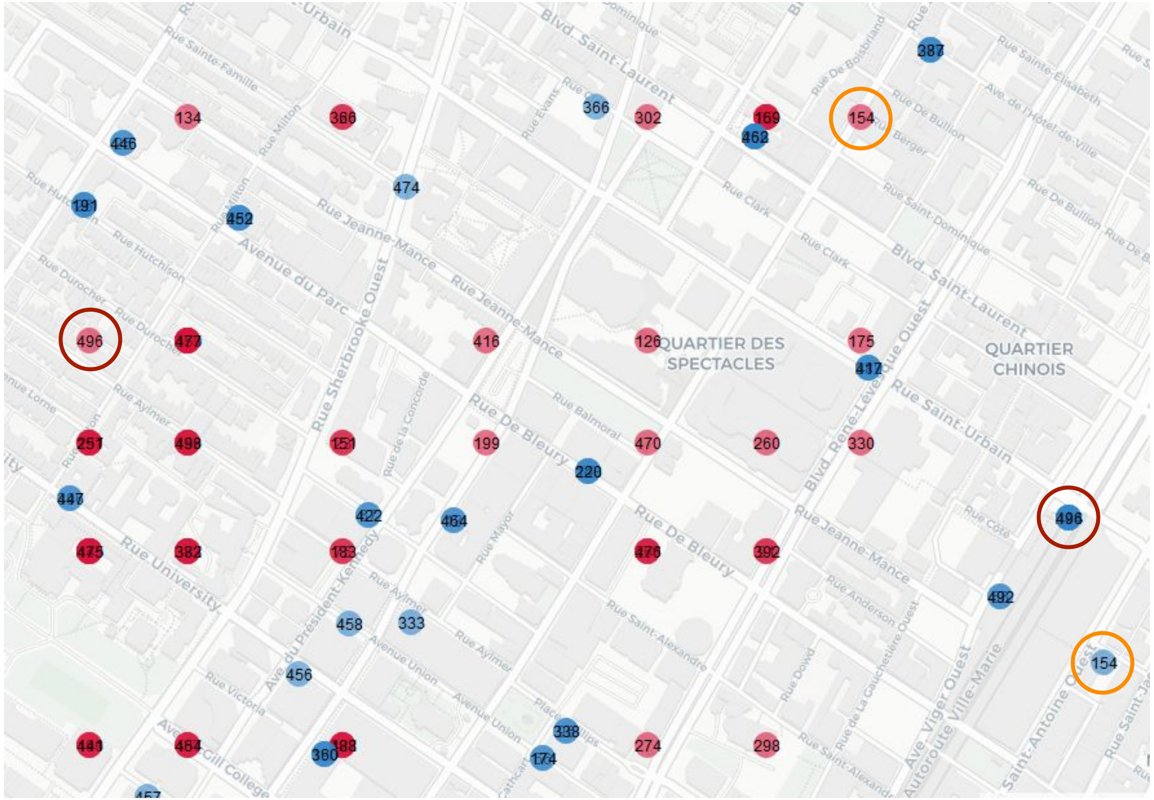
Version 2	Decision Tree	Random Forest
Lat. RMSE	0.0149691975	0.01576569645
Long. RMSE	0.01743998037	0.01791527498

Regression Metrics

- Decision Tree V2
  Visual Result

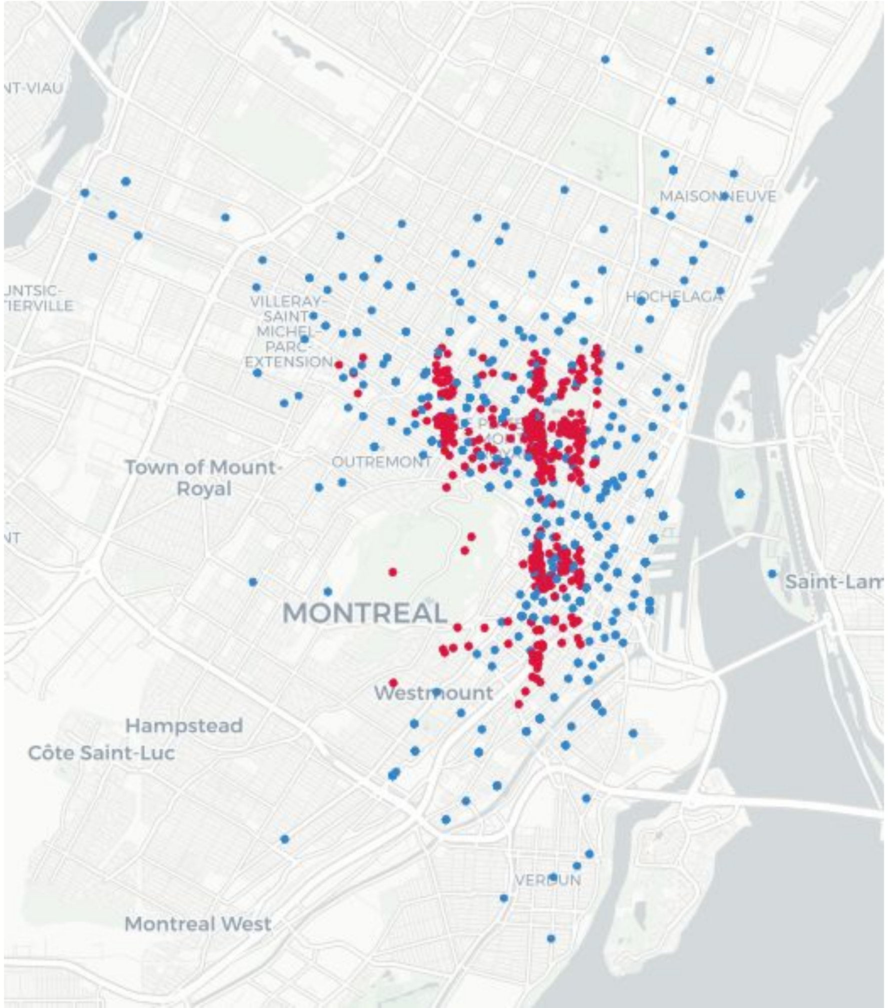
actual location

predicted location



actual location

predicted location



Conclusions

- Predicting end stations is a hard problem
- Used clustering to improve classification
- Used separate regressors for lat/long
- Accuracy/RMSE is better with Decision Tree

Thank You!