Predicting NYC Green Taxis Tip Percentage

AN APPLICATION OF LINEAR REGRESSION MODEL

WEINING HU

MARCH 8, 2020

Agenda

- Problem Statement
- Methodology
  - Dataset Description
  - Feature Engineering
  - Data Preprocessing
  - Visualization
- Future Work
  - Exploratory Data Analysis
  - Model fitting and evaluation
- Reference

Problem Statement

General Problem Statement

In 2013, the New York City government launched the Green Taxis program as a mechanism to serve the outer boroughs (Bronx, Queens, Brooklyn, Staten Island and Manhattan above the 110th Street) and to tackle the illegal taxi problem. Analyzing the trip data fetched by the green taxi meters can help policy makers to understand how the green taxis are performing to serve the program's initial purposes.

Task 1:

Are green taxis still predominantly serving the outer boroughs of Manhattan?

Task 2:

Predicting tip percentage

Dataset Description

- New York City Taxi and Limousine Commission (TLC)
- Green taxi trips in June 2019
- 471,052 observations
- 20 columns
  - **Admin related (2)**: VendorID, Store_and_fwd_flag
  - **Trip related (7)**: lpep_pickup_datetime, lpep_dropoff_datetime, passenger_count, Trip_distance, PULocationID, DOLocationID, Trip_type
  - **Trip fare related (11)**: RateCodeID, Payment_type, Fare_amount, Extra, MTA_tax, Improvement_surcharge, Tip_amount, Tolls_amount, Total_amount, congestion surcharge, ehail_fee
- Column datatypes: categorical, numeric (int64 and float64)
- All columns are non-null (except for ehail_fee)

- Taxi Zone Shapefile
- Taxi Zone Lookup Table

locationid	borough	zone	Service zone
1	EWR	Newark Airport	EWR
2	Queens	Jamaica Bay	Boro Zone

Feature Engineering

PULocationID	DOLocationID
74	263
75	74

df.merge

pu boro	pu zone	pu sz	do boro	do zone	do sz
Manhattan	East Harlem North	Boro Zone	Manhattan	Yorkville West	Yellow Zone
Manhattan	East Harlem South	Boro Zone	Manhattan	East Harlem North	Boro Zone

PU Datetime	DO datetime
2019-06-01 00:25:27	2019-06-01 00:33:52
2019-06-01 00:39:13	2019-06-01 00:46:38

pd.to_datetime
df.dt.hour
df.dt.weekday_name

pu hour	pu day	do hour	do day	trip time
0	0	Saturday	Saturday	8.0 (minutes)
0	0	Saturday	Saturday	7.0

DOLocationID
1
2

gpd.read_file
df.merge

DOLocationID	geometry	borough	zone
1	POLYGON ((933100.9183527103 192536.0856972019,...	EWR	Newark Airport
2	(POLYGON ((1033269.243591294 172126.0078125, 1...	Queens	Jamaica Bay

Data Preprocessing

Raw dataset (471,052 Observations)

Remove Criteria	# of R	Pct of S
Distance of 0 mile or > 40 miles (99.99% pctl was 39.27 miles)	6,671	1.42%
Trip time of 0 min or > 82 minutes (99% pctl was 82 mins)	8,164	1.73%
Fare amount < USD 2.50	1,628	0.35%
Tip amount > twice the fare amount	81	0.01%
Trip speed >=240 MPH	36	~0.01%

Analytical dataset (454,472 Observations)

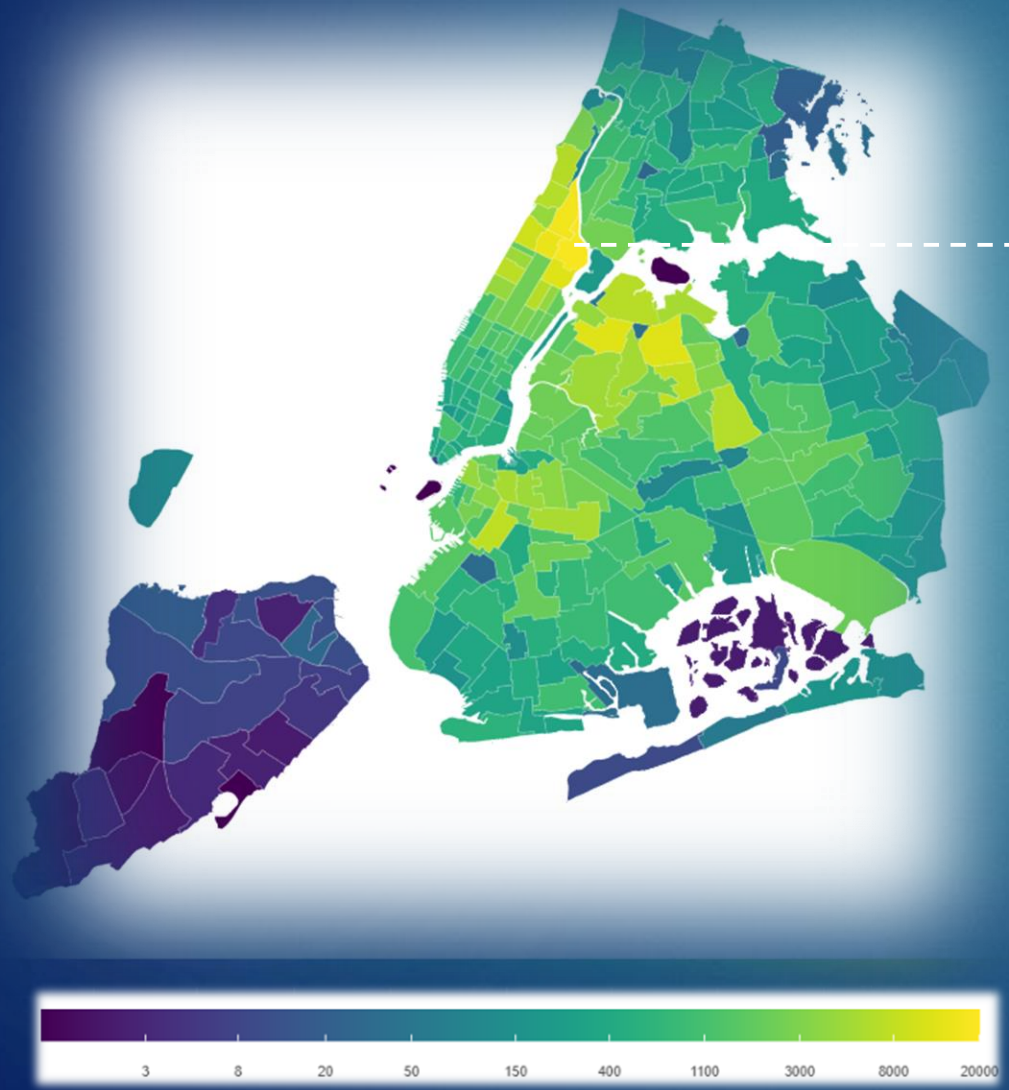
Impute Data	# of R	Pct of DS
Impute valid observations with distance of 0 mile and/or 0 min	4,484	0.99%
Replace ['passenger_count'] == 0 by 1 (mode)	672	0.15%

Visualization – Popular PU Locations

Name: East Harlem North
Borough: Manhattan
Zone ID: 74
# of Trips: 38,459

Rank	Popular PULoc	# Trips
1	East Harlem North	38,459
2	East Harlem South	29,943
3	Central Harlem	27,901
4	Elmhurst	22,084
5	Astoria	20,849

Visualization – Popular DO Locations



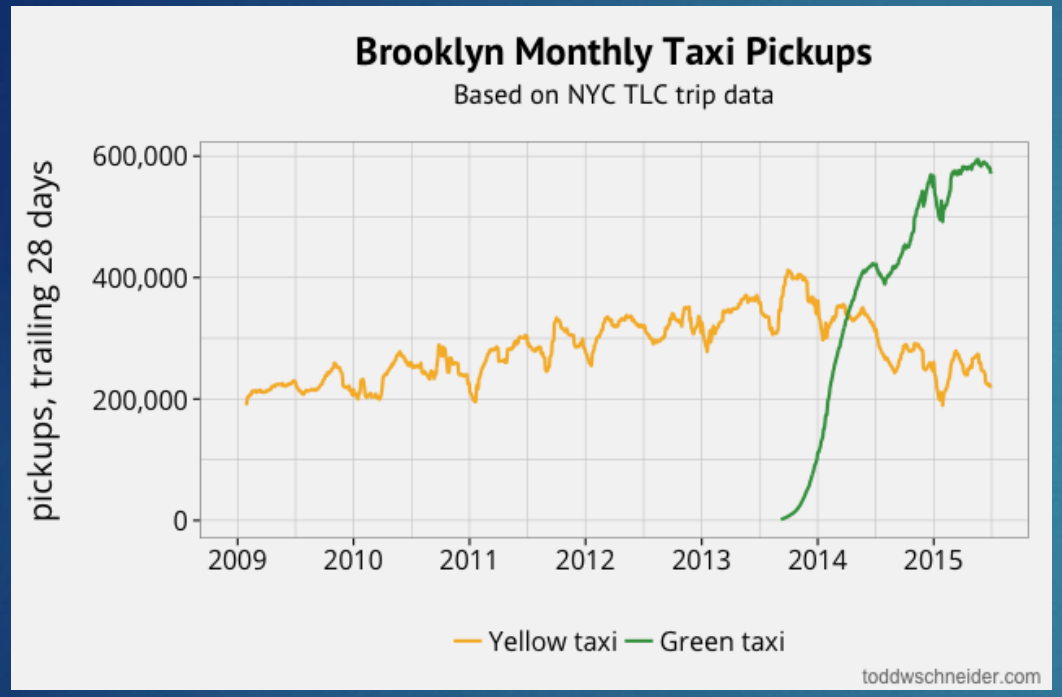
Name: East Harlem North
Borough: Manhattan
Zone ID: 74
# of Trips: 18,586

Rank	Popular DOLoc	# Trips
1	East Harlem North	18,586
2	Central Harlem North	16,854
3	Central Harlem	15,057
4	East Harlem South	12,888
5	Astoria	12,662

Future Work

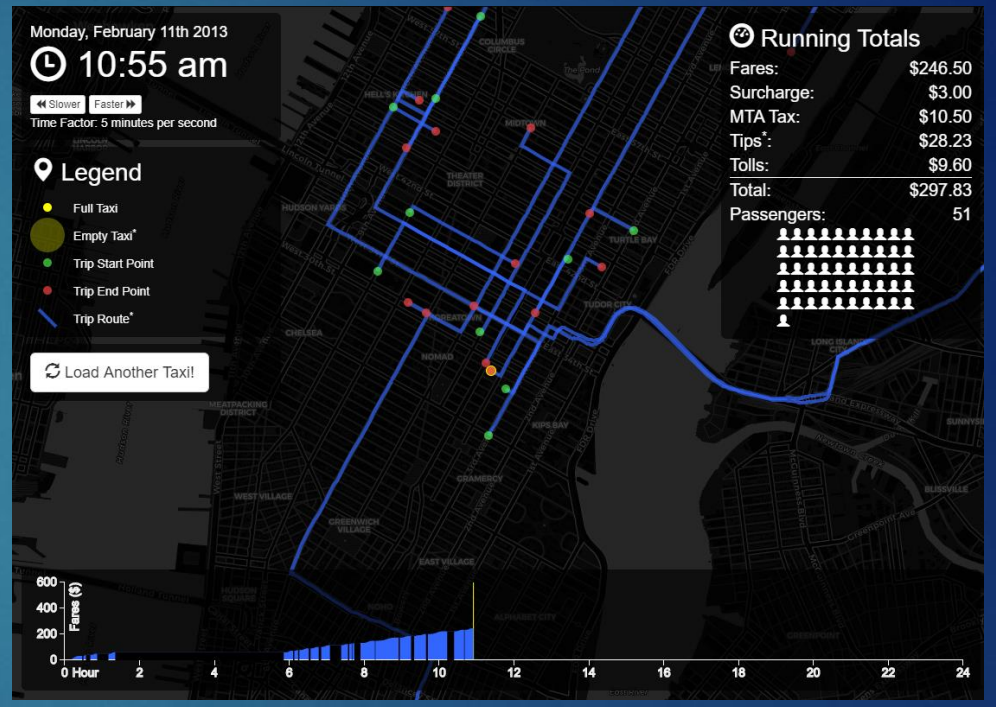
- Exploratory Data Analysis
  - Tips distribution by hour and by day of week
- Modeling
  - Prepare dataset for predicting tip percentage
  - Feature selection
  - Model fitting
    - Linear regression with ordinary least square
    - Logistic regression + linear regression
    - Polynomial model
    - Regularization and gradient boosting
  - Model evaluation
    - Impact of extreme data
    - Design evaluation matrix (R^2)

Reference



Borough Trends, and the Rise of Uber

Todd Schneider



NYC Taxis: A Day in the Life

Chris Whong

Reference

- NYC Green Taxi Data Visualization – Gshahane link
- NYC Green Taxi of September 2015 Data Analysis – kthouz link
- Looking Through the Taxi Meter – Analysis of NYC Green Taxi data of September 2015 – Jiamin Han link
- Predict New York City Taxi Demand – Yunrou Gong, Bin Fang, Shuo Zhang, and Jingyu Zhang link
- Visualizing NYC Taxi Cab Data – Network Graph link
- Geospatial joining and Bokeh heatmap visualization – Shekhar link