Rideshare & Taxi Service Tipping Behavior

Ezra Kim, Josefina Bollini, Whitney Schreiber, Ling Yang

Big Data Platforms (MSCA 31013) Final Project

Autumn 2022

Agenda

1. Business Problem & Data Introduction
2. Solution Architecture & Data Engineering
3. Data Exploration
4. Machine Learning Models

Executive Summary

Our team investigated the tipping behavior of taxi/rideshare consumers in Chicago.

During exploratory data analysis, we discovered only ~28% of rides have tips and there were a few features that made the difference – like whether the ride was in a taxi vs rideshare.

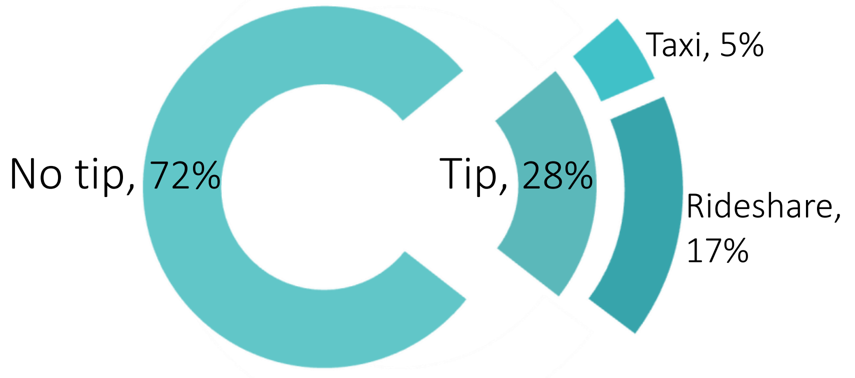
The team used classification models to determine whether a ride will receive a tip or not. The best performing model was the Gradient Boosted Tree Classifier.

The team then chose a Random Forest Regression model to predict how much (*percent of ride cost*) a rider would tip.

Business Problem & Data Introduction

Business Problem & Goals of Analysis

Every day, over 200,000 customers take a rideshare or taxi in Chicago but only 28% of those trips end with a customer happy enough to tip their driver.



Business Problem

Predict **tipping behavior** for **transportation services** in Chicago [taxi and rideshare]

- ➢What is the probability that a rider tips?
- ➢What is the predicted tip percent on a given ride?

Scope

- Transportation service: Rideshare (Uber and Lyft), Taxi
- Location: Chicago
- Time period: Jan. 2019 – Dec. 2021

Data Profile

Rideshare

Data includes all rideshare trips

Source: Chicago Data Portal (Transportation Network Providers)

Size: 66 GB

Key Features:

- `` `Trip ID` `` – unique trip identifier
- `` `Fare` `` – rounded to nearest $2.50
- `` `Tip` `` – rounded to nearest $1.00, *cash tips <u>not</u> recorded
- `` `Additional Charges` `` – taxes, fees, and any other charges
- `` `Trip Total` `` – calculated as `` `Fare` `` + `` `Tip` `` + `` `Additional Charges` ``
- `` `Trip Start Timestamp` ``, `` `Trip End Timestamp` ``
- `` `Trip Seconds` `` – amount of time, in seconds
- `` `Trip Miles` `` – distance in miles
- `` `Pickup Community Area` ``, `` `Dropoff Community Area` ``
- `` `Shared Trip Authorized` `` – customer agreed to a shared trip with another customer
- `` `Trips Pooled` `` – number of trips pooled [includes this trip]

Data Profile

Taxi Trips

Data includes all taxi trips

Source: Chicago Data Portal

Size: 81 GB

Key Features:

- `Trip ID` – unique trip identifier
- `Fare` – exact fare
- `Tips` – *cash tips are generally <u>not</u> recorded
- `Tolls` – tolls for the trip
- `Extras` – extra charges
- `Trip Total` – calculated as `Fare` + `Tip` + `Tolls` + `Extras`
- `Trip Start Timestamp`, `Trip End Timestamp`
- `Trip Seconds` – amount of time, in seconds
- `Trip Miles` – distance in miles
- `Pickup Community Area`, `Dropoff Community Area` – *locations outside Chicago are left blank
- `Payment Type` – type of payment for the trip
- `Company` – taxi company

Data Profile

Covid-19

Data includes new Covid-19 cases, hospitalizations and deaths in Chicago

Source: Chicago Data Portal (Department of Public Health)

Size: 152 KB

Key Features:

- `Date` – dd/mm/yy format
- `Cases-Total` – number of new cases in Chicago
- `Deaths-Total` – number of new deaths in Chicago
- `Hospitalizations-Total` – number of new hospitalizations due to Covid-19 in Chicago

*hospitalizations are based on the date of first hospitalization

Data Profile

Weather

Data includes precipitation and snowfall measurements at various locations in the Chicagoland area

Source: Global Historical Climatology Network

Size: 10.6 MB

Key Features:

- `Date` – unique trip identifier
- `Station_Name` – station where measurements occurred
- `PRCP` – precipitation in inches
- `SNOW` – snowfall in inches
- `SNWD` – snow depth in inches

Data Profile

Chicago Events

Data includes all sports events in Chicago

Source: MLB, NFL, NBA, NHL schedules

Size: 33 KB

Key Features:

- `Location` – arena/field name
- `Day` – day of the event
- `Month` – month of the event
- `Year` – year of the event
- `Neighborhood` – neighborhood of the arena/field
- `Team` – team played in the event
- `nb_code` – unique identified of the neighborhood

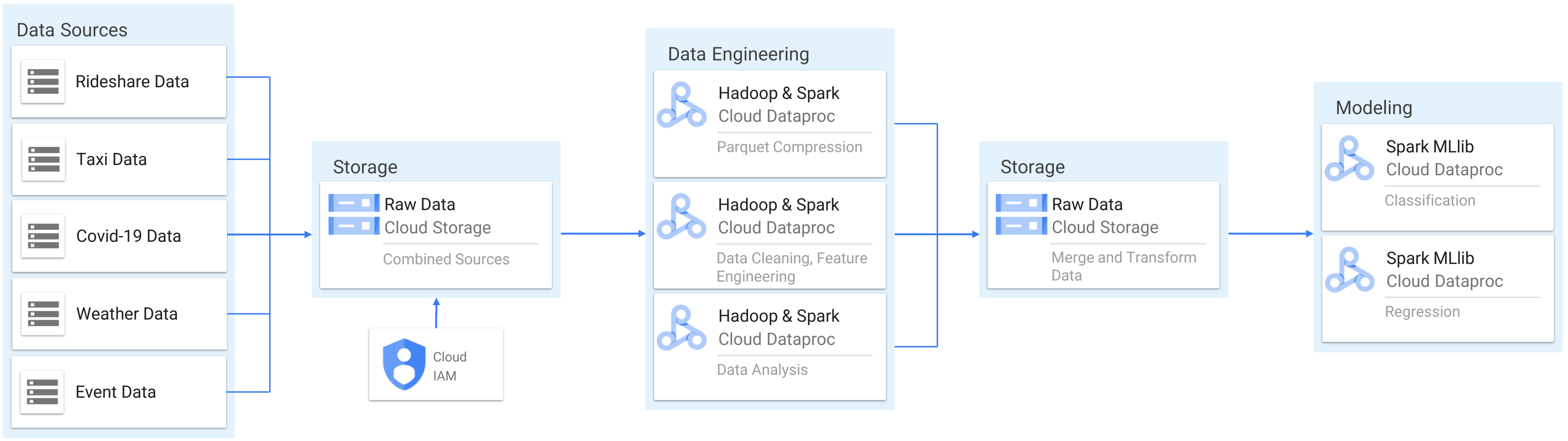
Data Challenges

- There are many macro trends to consider that affect intra city travel. Most of these events are not centralized anywhere so data is hard to collect.
- Non-City of Chicago datasets are often not available, so data had to be extracted manually i.e., the sport events data.
- Cash tips may not be recorded

Solution Architecture & Data Engineering

Google Cloud

Solution Architecture



Data Compression

Compression method: Parquet

140GB → **14**GB

Raw File Size

Compressed Parquet File Size

Data Engineering Considerations:

Data refresh rate for training data would be monthly

Delete source csv after compression

Use Google Functions to orchestrate data extraction and data compression

Joining Datasets

1. Stack datasets

Taxi Trips

+

Rideshare Trips

(filter for 2019 – 2021)

2. Left Join on Date

- Covid-19
- Events in Chicago
- Weather
- Chicago Communities

Data Engineering

Feature Generation

Tip features

`tip_label`

{Y/N}, ride resulted in a tip

`tip_ percent `

percent of ride cost the user tipped

*calculated as: tip value / (fare + additional charges)

Time features

`dow`

day of week the trip started

`weekend`

{Y/N}, trip start occurred between Friday – Sunday

`year`

year the trip occurred

`hour`

hour of the day the trip occurred

`season`

{winter, spring, summer, autumn}

Ride condition features

`outside_chicago`

{Y/N}, trip pickup or drop-off includes a location outside Chicago proper

`ride_type`

{rideshare (1), taxi (0)}, whether rider used a rideshare service or a taxi service

`rain_snow`

{Y/N}, weather conditions include snow and/or rain

Covid-19 features

`covid_cases_7dayAvg`

rolling 7-day average of new Covid-19 cases in Chicago

`covid_deaths_7dayAvg`

rolling 7-day average of new Covid-19 deaths in Chicago

`covid_hosp_7dayAvg`

rolling 7-day average of new Covid-19 hospitalizations in Chicago

Data Exploration

Exploratory Data Analysis

220,167,785 rides
over a 3-year period

riders provided a tip in
28% of the rides

riders provided no tip in
72% of the rides

- *Imbalanced dataset! Must be accounted for during modeling.*

Full Dataset
[Rides with Tips & No Tips]

1. **Hour**: little impact on tips
2. **Rain_Snow**: On average, tips are larger on rainy/snowy days
3. **Season**: Winter has the largest average tip. All other seasons had similar amounts of tips
4. **DOW**: Friday is the only day to have a difference in average tip
5. **Year**: average tip is noticeable higher in 2021 ($1.03), with 2020 being lower than expected ($0.61). 2021 ($0.798)

Filtered Dataset
[Rides with Tips]

1. **Taxis vs Rideshare**: Riders tend to tip more when in taxis vs rideshare ($4.11 vs $3.72). Taxis have higher fares but go less miles and seconds.
2. **Payment Type**: No Charge had the most tip, followed by Cash. Credit Card had the highest tip percentage.
3. **Weekend**: No variation in average tips between weekend and non-weekend rides.
4. **Rain_Snow**: little impact on average tip.
5. **Season**: Winter and Summer had slight differences in average tip.
6. **DOW**: Largest variations on Sunday, Monday, Thursday, and Saturday.
7. **Month**: August, September, and October are the months with the highest average tip. January, March, and February have the lowest average tip.
8. **Year**: Similar numbers to Full Dataset.

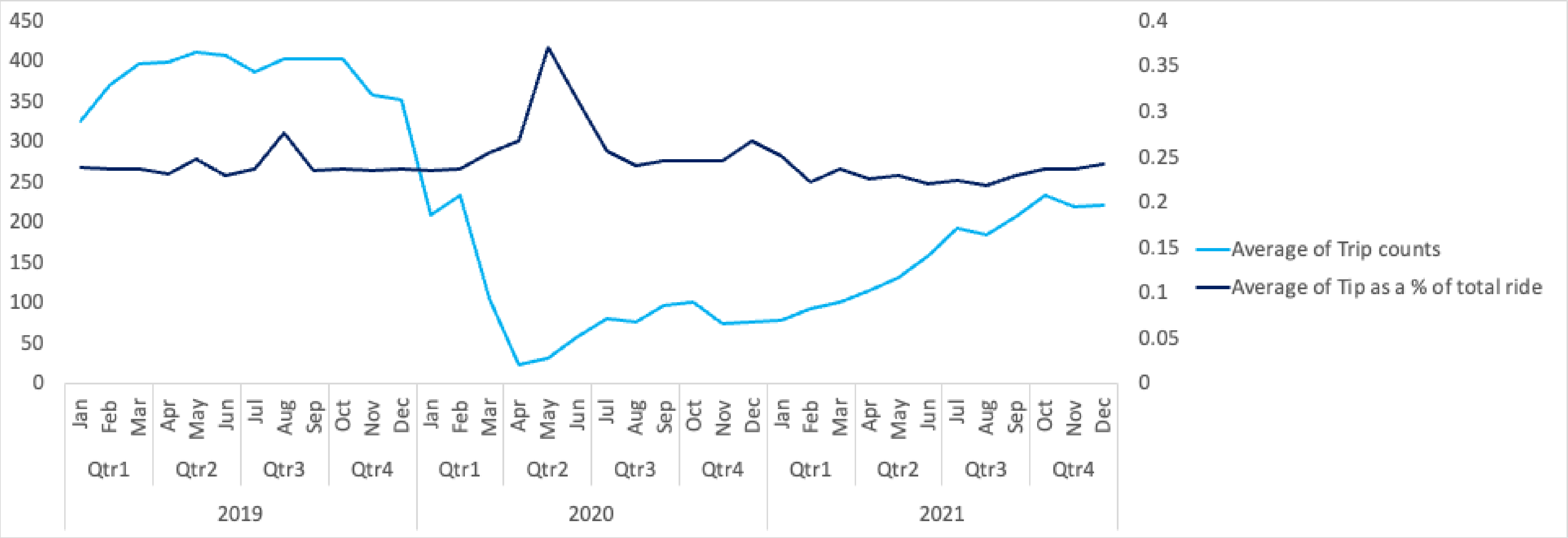
Behavior between Tipped and Non-Tipped Rides

	Average Tip Amount	Average Fare	Average Tip Percentage	Average Trip Seconds	Average Trip Miles
All Rides	**$0.82**	**$13.97**	**5.87%**	**1,051.34**	**6.2**
Tipped Rides	$3.81	$16.22	23.49%	1,138.90	6.9
Non-Tipped Rides	$0.00	$13.35	0.00%	1,027.27	6.0

On average, tips are typically given on rides with longer durations and distances, and pay more in fare

*Tip Percent is calculated as: Tip / Ride Cost

Trips & Tip Percentage 2019-2021

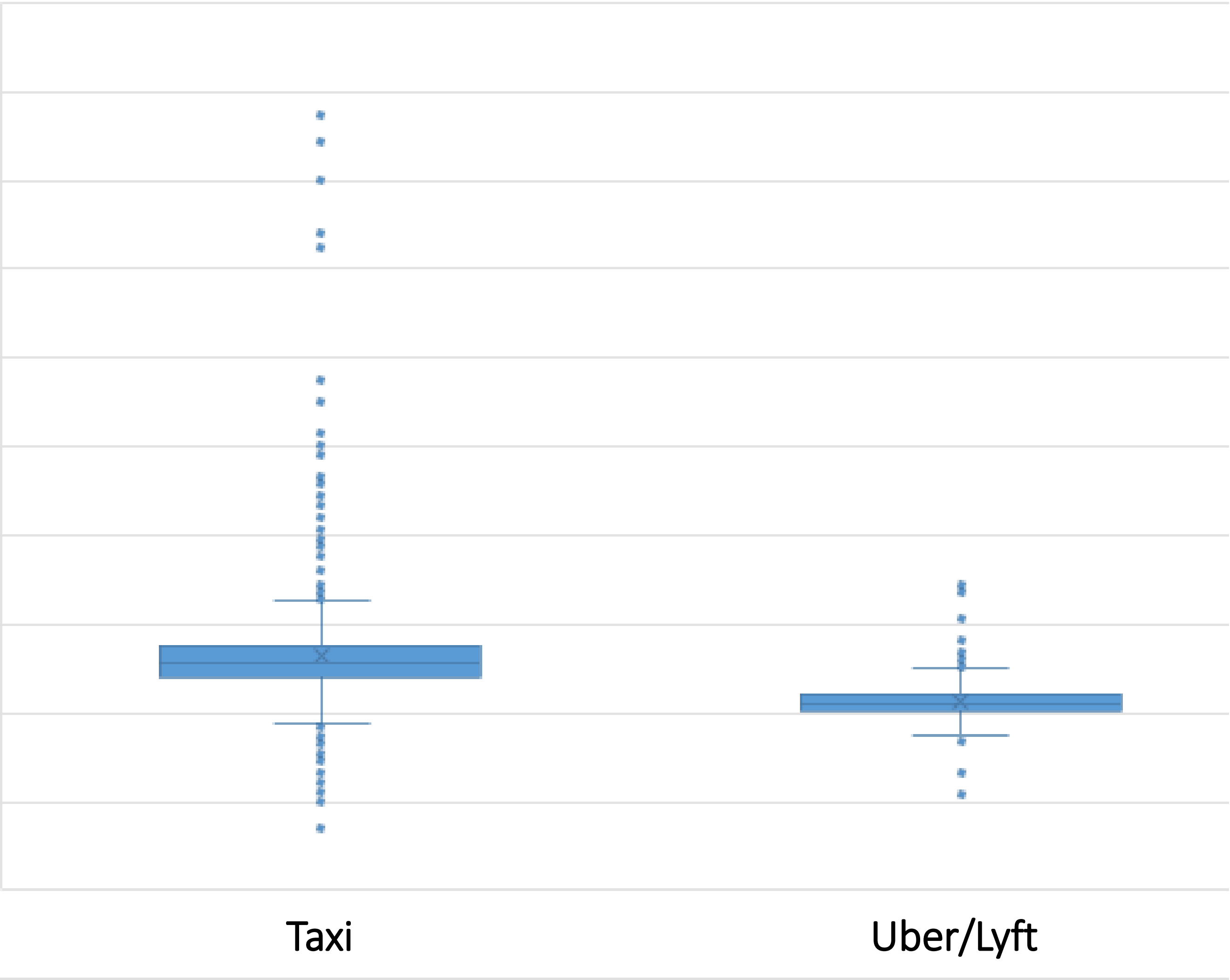


In Q2 2020, the average number of trips by day decreased drastically while tip percentage saw an increase during April and May 2020. The tip percentage returned to normal levels by Q3 2020.

*Tip Percent is calculated as: Tip / Ride Cost

Tip Percent* Behavior

Taxi v. Rideshare



*Tip Percent is calculated as: Tip / Ride Cost

Taxi rides tend to have higher tips than rideshare rides.

The dispersion in these distributions is noticeable: the presence of outliers show that several taxi riders tipped up to **90%** of the cost of the trip. This phenomenon is not observed in the Uber/Lyft rides.

From this analysis, the difference between taxi and rideshare makes this a feature worth watching when modeling

Ride Type Deep Dive

Tipped Rides

Ride Type	Average Tip Amount	Average Fare	Average Tip Percentage	Average Trip Seconds	Average Trip Miles
Taxi	$4.12	$17.38	23.71%	1,007.89	4.84
Rideshare	$3.72	$16.22	22.93%	1,138.9	6.94

Non-Tipped Rides

Ride Type	Average Tip Amount	Average Fare	Average Tip Percentage	Average Trip Seconds	Average Trip Miles
Taxi	$0	$14.98	0%	871.16	3.37
Rideshare	$0	$13.21	0%	1,040.6	6.23

All Rides

Ride Type	Average Tip Amount	Average Fare	Average Tip Percentage	Average Trip Seconds	Average Trip Miles
Taxi	$1.81	$16.04	11.28%	931.42	4.02
Rideshare	$0.70	$13.71	5.11%	1,066.22	6.47

Tipping Behavior across Months

Month	Average Tip Amount	Average Fare	Average Tip Percentage	Average Trip Seconds	Average Trip Miles
September	$4.08	$17.23	23.68%	1,192.17	7.27
August	$4.04	$17.29	23.37%	1,172.37	7.19
October	$4.01	$16.81	23.85%	1,183.49	7.22
July	$4.01	$17.08	23.48%	1,171.37	7.06
June	$3.99	$17.61	22.66%	1,200.79	6.97
November	$3.93	$16.16	24.32%	1,143.52	7.20
December	$3.90	$15.69	24.86%	1,093.17	6.94
May	$3.81	$17.06	22.33%	1,191.27	6.98
April	$3.59	$16.16	22.22%	1,136.8	6.90
February	$3.44	$14.61	23.55%	1,065.15	6.39
March	$3.44	$14.86	23.15%	1,076.79	6.59
January	$3.34	$13.90	24.03%	1,027.74	6.41

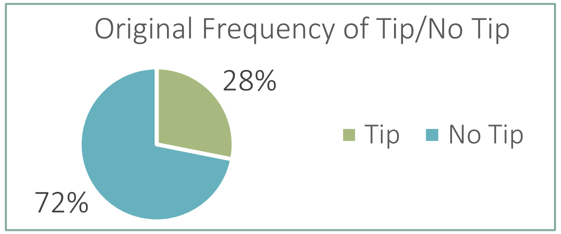
- Tips tend to be highest during late summer/early fall
- First 3 months of the year tend to have the lowest tips
- There is little variation in tip percentage

Machine Learning Models

Tip Prediction {Y/N}

Model Preparation

1. Correct for unbalanced data *[28% of rides resulted in a tip]*
   Under-sampled observations with no tip so the labels 1 and 0 occur with equal frequency
2. StringIndexer
   Map categorical variables to label indices
3. OneHotEncoder
   Map category label indices to a column of binary vectors
4. VectorAssembler
   Place features into a single feature column
5. StandardScaler
   Standardize numeric features
6. Pipeline
   Build pipeline to perform StringIndexer, OneHotEncoder, VectorAssembler, and StandardScaler operations
   Fit and transform the data



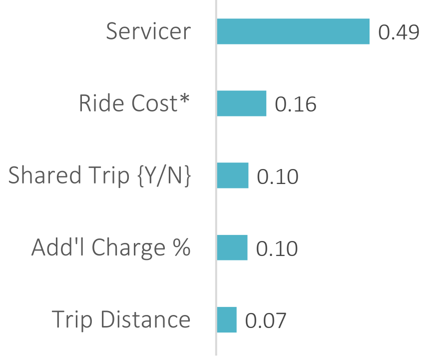
Tip Prediction {Y/N}

Classification Models

Random Forest

	Training	Test
AUC	0.6507	0.6508
Accuracy	0.6436	0.6437
f1	0.5977	0.5978
Weighted Precision	0.6523	0.6524
Weighted Recall	0.6436	0.6437

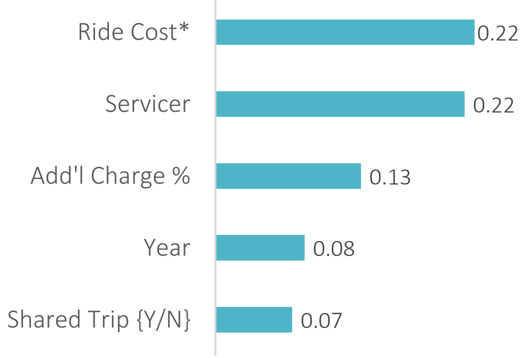
Feature Importance



Gradient-Boosted Tree Classifier

	Training	Test
AUC	0.6671	0.6670
Accuracy	0.6530	0.6530
f1	0.6217	0.6218
Weighted Precision	0.6543	0.6543
Weighted Recall	0.6530	0.6530

Feature Importance



Logistic Regression

	Training	Test
AUC	0.5000	0.5000
Accuracy	0.5870	0.5871
f1	0.4343	0.4344
Weighted Precision	0.3446	0.3447
Weighted Recall	0.5870	0.5871

**Ride Cost = Fare + Additional Charges*

Tip Prediction {Y/N}

Classification Models

Random Forest

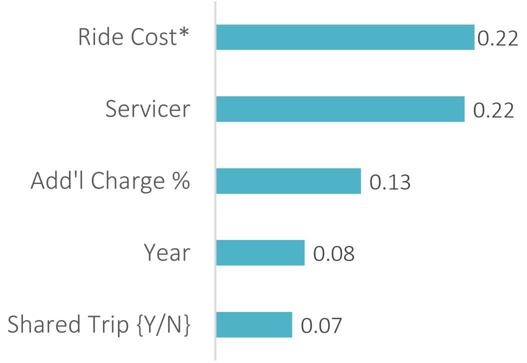
	Training	Test
AUC	0.6507	0.6508
Accuracy	0.6436	0.6437
f1	0.5977	0.5978
Weighted Precision	0.6523	0.6524
Weighted Recall	0.6436	0.6437

Feature Importance

Gradient-Boosted Tree Classifier

	Training	Test
AUC	0.6671	0.6670
Accuracy	0.6530	0.6530
f1	0.6217	0.6218
Weighted Precision	0.6543	0.6543
Weighted Recall	0.6530	0.6530

Feature Importance



Logistic Regression

	Training	Test
AUC	0.5000	0.5000
Accuracy	0.5870	0.5871
f1	0.4343	0.4344
Weighted Precision	0.3446	0.3447
Weighted Recall	0.5870	0.5871

Gradient-Boosted Tree Classifier is the top performing classifier

**Ride Cost = Fare + Additional Charges*

Tip Percent Prediction

Model Preparation

1. StringIndexer

   Map category label indices to a column of binary vectors

2. OneHotEncoder

   Convert relevant categorical features into one hot encoded vectors

3. VectorAssembler

   Place features into a single vector

4. Pipeline

   Build pipeline to perform StringIndexer, OneHotEncoder, and VectorAssembler

5. Fit and transform

   Get data ready for ML model

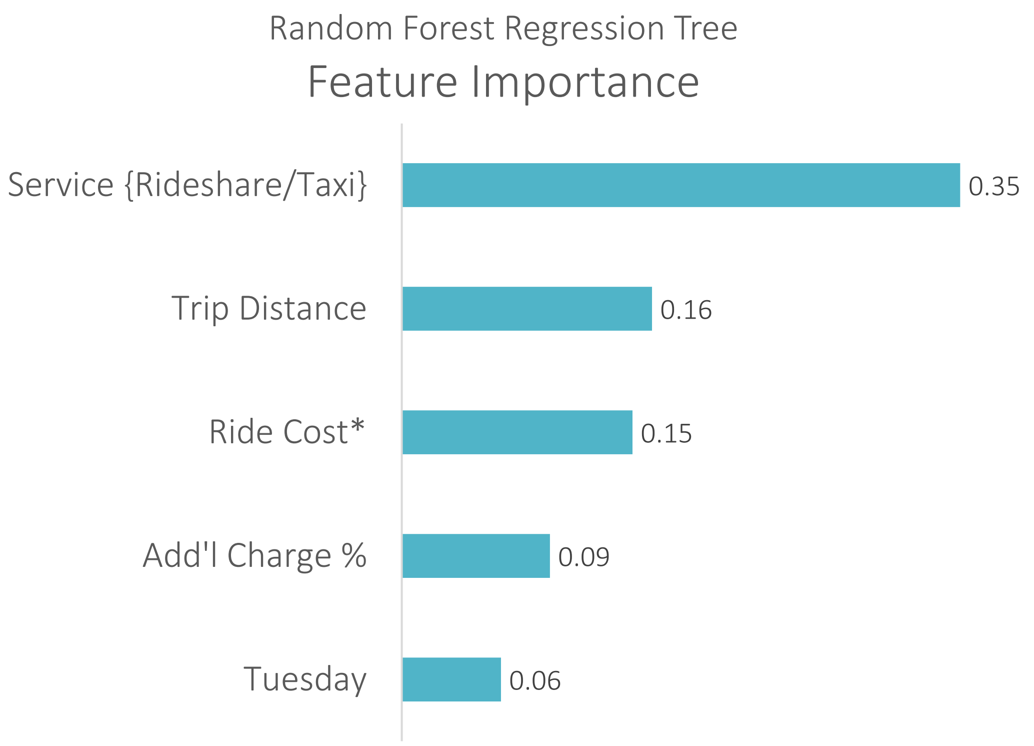
Tip Percent Prediction

Random Forest Regression

Performance Evaluation

	Training	Test
RMSE	0.7730	0.8729
R-squared	0.0035	0.0027

Tip Percent	
predicted	actual
0.0010	0.0000
0.0010	0.0000
0.3338	0.3601
0.0065	0.0000
0.0010	0.0000



Ride Cost = Fare + Additional Charges

Future Work & Considerations

Automate of Data Engineering tasks:

- Compression
- Data ingestion

Modeling

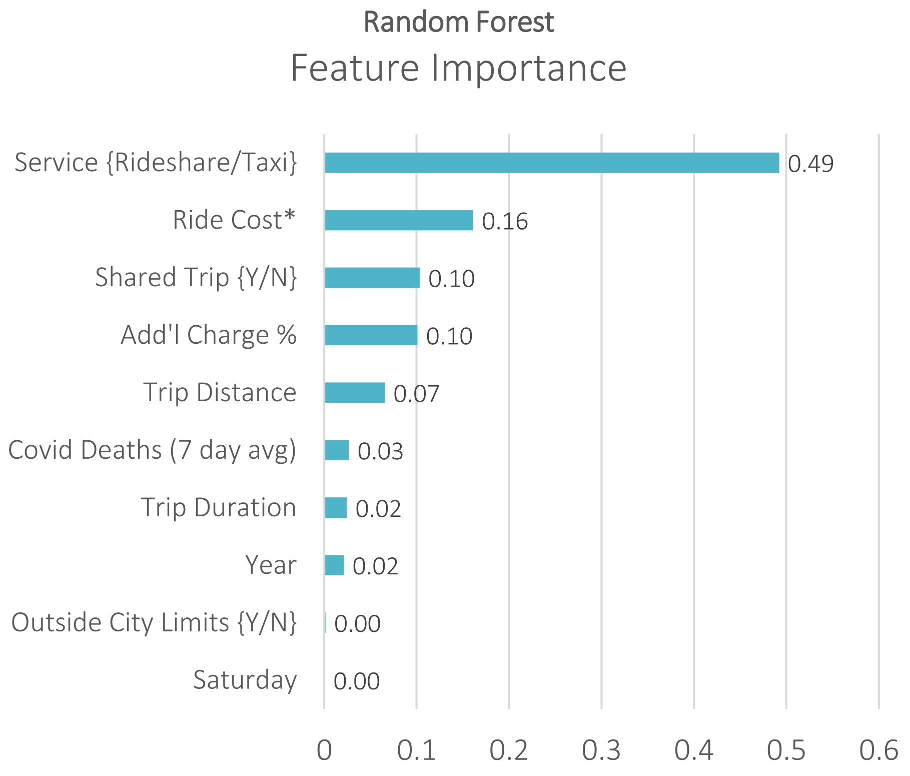
- Test methods to better model a zero-inflated continuous response variable, such as:
  - Stacking models
  - Neural Network
- Model rideshare and taxi service rides separately
- Graph frame of trips by neighborhood
- Outlier detection
- Apply cross-validation to find out the optimal parameters in models
- Incorporate additional features, such as:
  - CTA outages
  - Incorporate “holiday” flags for Thanksgiving and Christmas

Appendix

Tip Prediction {Y/N}

Random Forest Classifier

Model Evaluation Metrics	Training	Test
AUC	0.6507	0.6508
Accuracy	0.6436	0.6437
f1	0.5977	0.5978
Weighted Precision	0.6523	0.6524
Weighted Recall	0.6436	0.6437

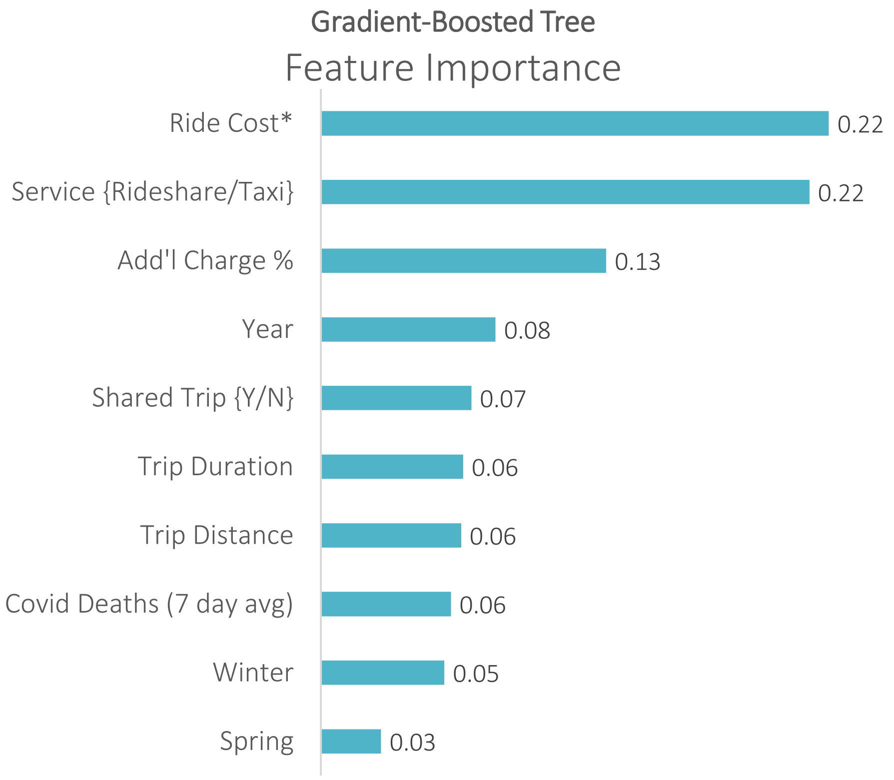


**Ride Cost = Fare + Additional Charges*

Tip Prediction {Y/N}

Gradient-Boosted Tree Classifier

Model Evaluation Metrics	Training	Test
AUC	0.6671	0.6670
Accuracy	0.6530	0.6530
f1	0.6217	0.6218
Weighted Precision	0.6543	0.6543
Weighted Recall	0.6530	0.6530



*Ride Cost = Fare + Additional Charges