Ride Hailing in Chicago:

A Classification Model for Tipped Rides

John Villanueva | Thinkful - Data Science | September 2019
Supervised Learning Capstone Presentation

For iPython Notebook and Presentation Slides:
https://tinyurl.com/SMLCapstoneJV

Motivations and Usage

Business Use

- Understanding tipping behavior so that ride hailing services find ways to incentivize and encourage their customers to tip their drivers. Implementing Tipping campaigns. Understanding tipping behavior to evaluate regulatory measures regarding appropriate driver compensation

Driver Use

- Understanding the tipping behaviors of passengers can help drivers to refine their driving strategy.

The Data: Raw

Ride Trips

- Source: The City of Chicago Data Website
- Includes Rides Starting November 1, 2018 to March 31, 2019
- Rides from Lyft and Uber
- All rides originate or end within the Chicago City Limits
- More detailed data documentation linked here

Weather

- Source: National Oceanic and Atmospheric Administration (NOAA)
- Weather Data
- Precipitation, Temperature, and Snow
- Across All Local Weather Stations for Each Day

The Data: Cleaning for Exploration

Ride Trips

- All Rides Originating AND Ending within Chicago City Limits
- Parsing date-time data into constituent date and time units

Weather

- Daily Mean Aggregation: Precipitation, Snowfall, and Average Temperature
- Aggregated across all weather stations to return averages for each date

→ Data frames joined by date to create a single data frame

The Data: Engineered Features

Average Trip Speed (MPH)

- Calculated using trip distance by trip time

Cost Per Mile ($/mi)

- Divided Sum of fare and additional charges by trip distance
- Tip not included in cost per mile

Tipped (Binarized))

- 0: Driver not tipped; 1: Driver Tipped

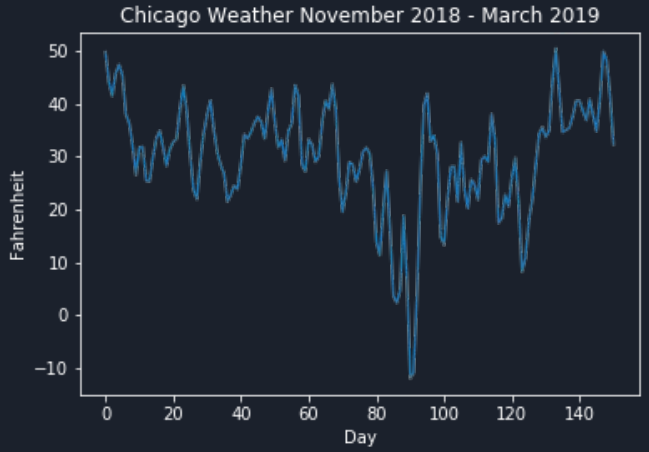
Warm Weather (Binarized)

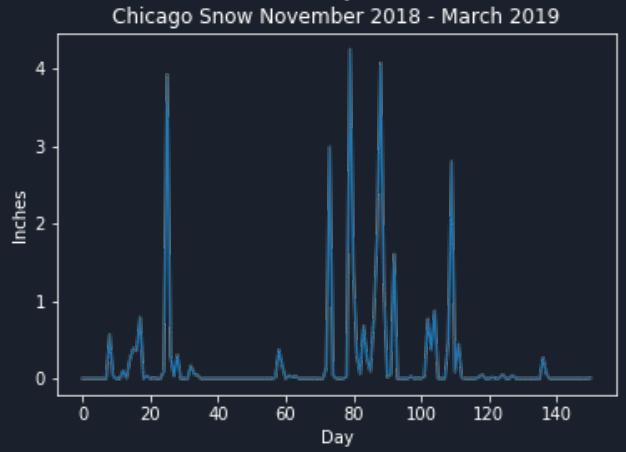
- If the weather was greater than 60 degrees Fahrenheit, then assigned value of 1

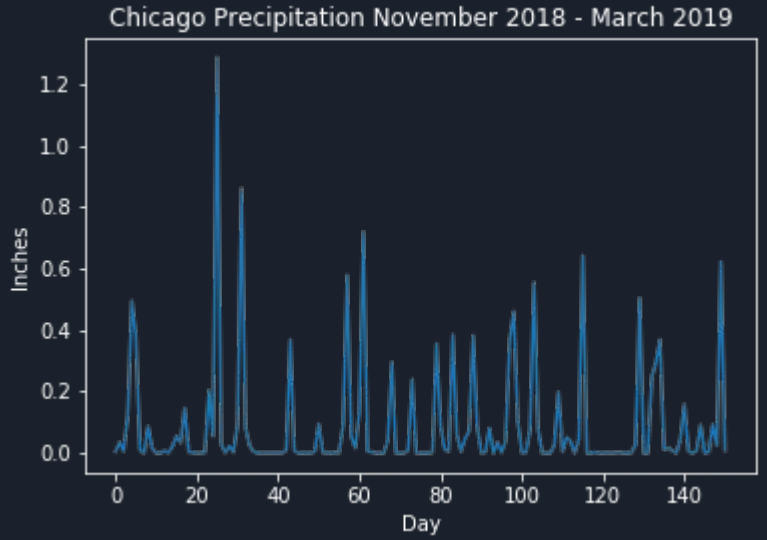
The Data: Ready for Visualization

Fare ($)	Date (MM-DD-YYYY)
Additional Charges ($)	Trip Start Month, Trip Start Day (Int)
Shared Trip Authorized ($)	Trip Start Hour, Trip Start Minute (Int)
Trip Total ($)	Trip End Month, Trip End Day (Int)
Trip Miles	Trip End Hour, Trip End Minute (Int)
Trip Seconds	Average Ride Speed (MPH)
Trips Pooled	Cost per mile ($/mi)
Shared Trip Authorized (Binary)	Warm Weather (Binary)
Pickup Community Area	Precipitation (In.)
Drop Off Community Area	Snowfall (In.)
Tip ($), Tipped (Binary)	Average Daily Temperature (Fahrenheit)

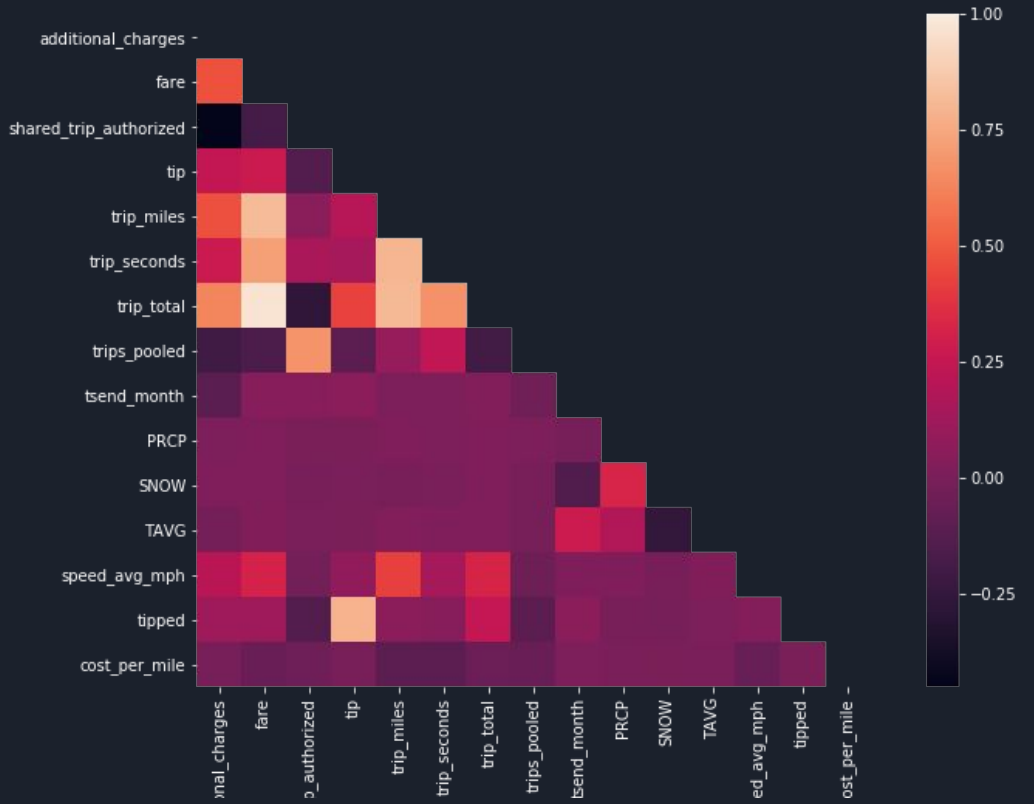
Visual Exploration - How's the weather?



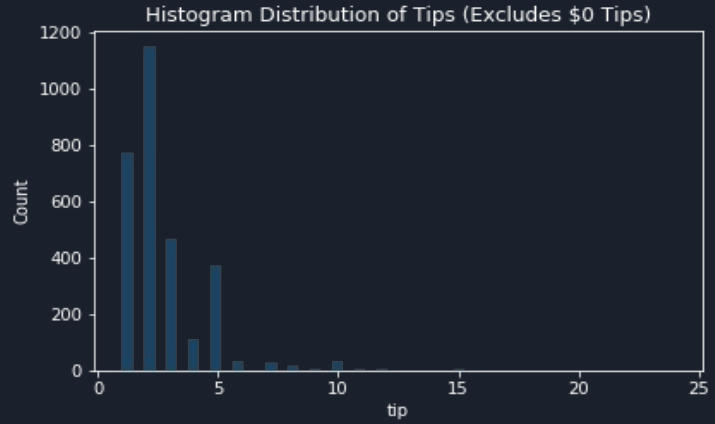


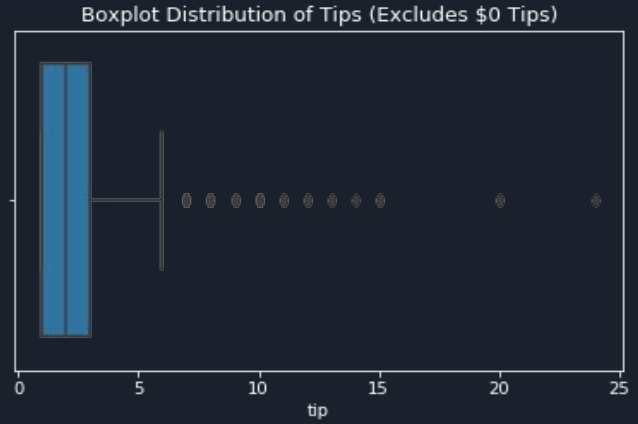


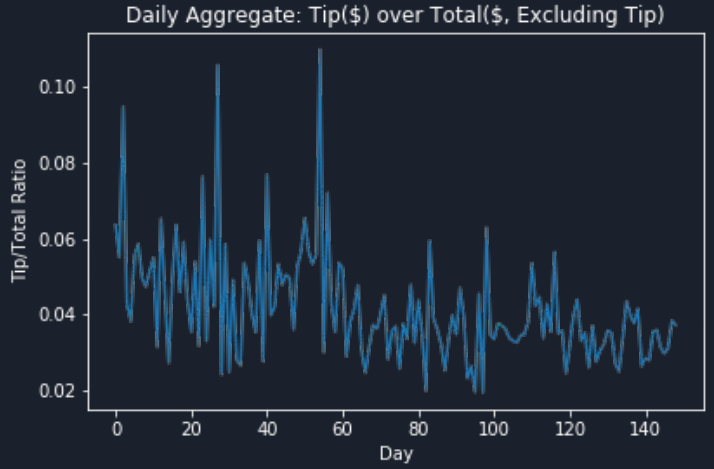
Visual Exploration - Quick Correlations

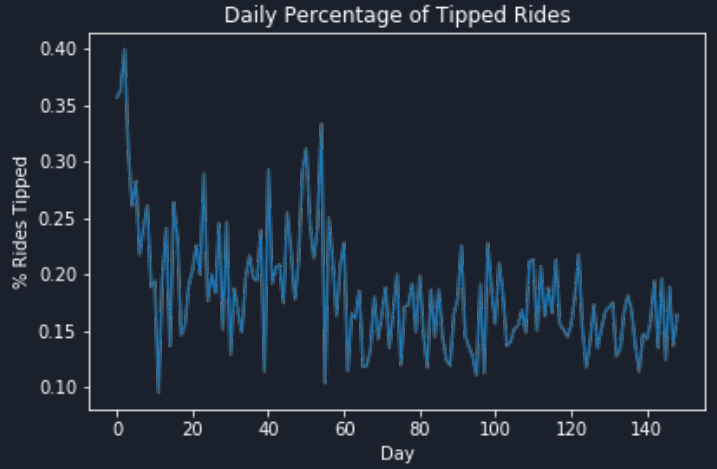


Visual Explorations - Tipping Behavior

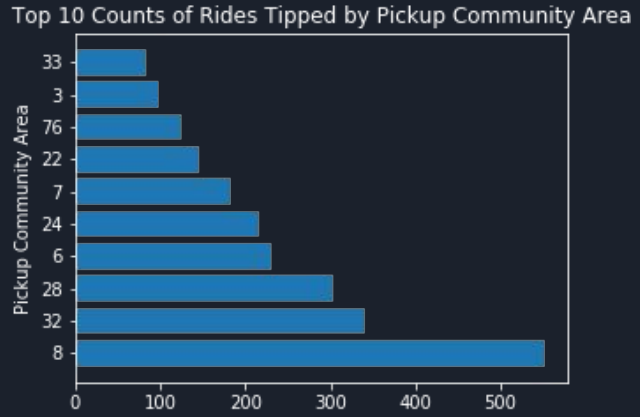


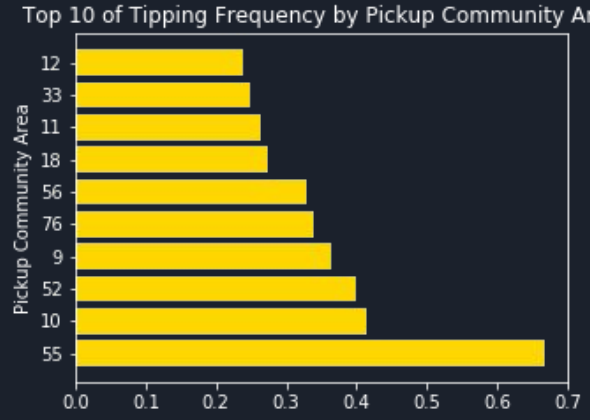


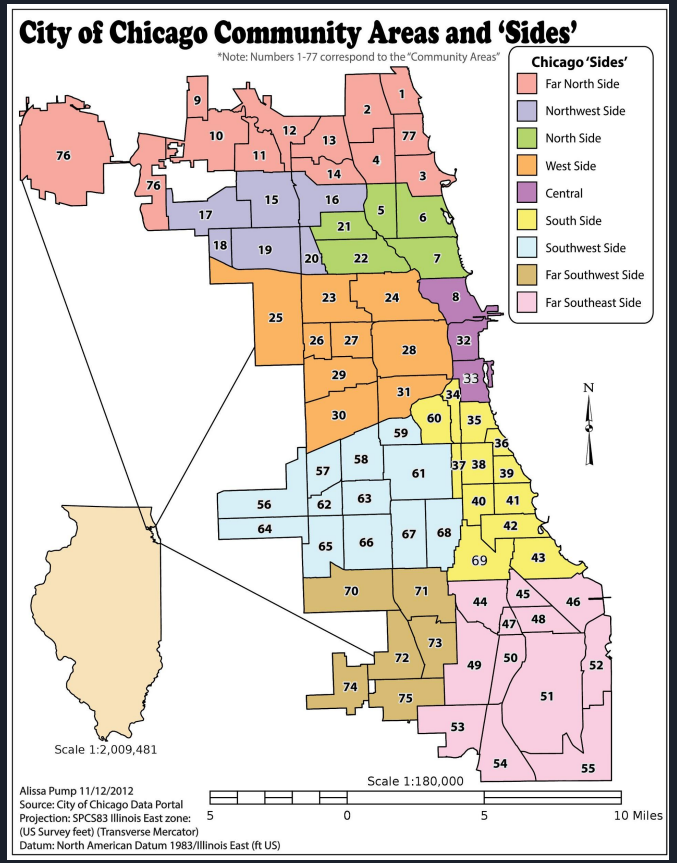




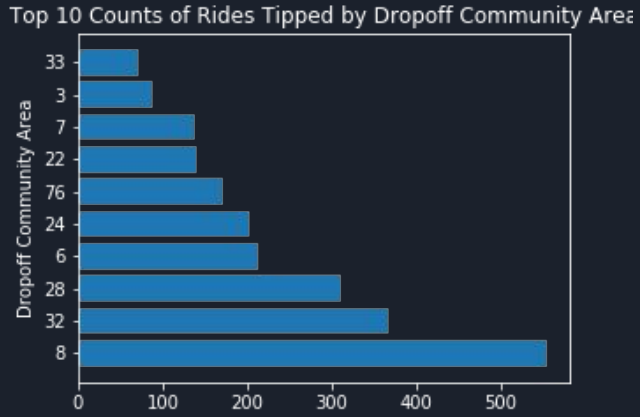
Visual Explorations - Pickups

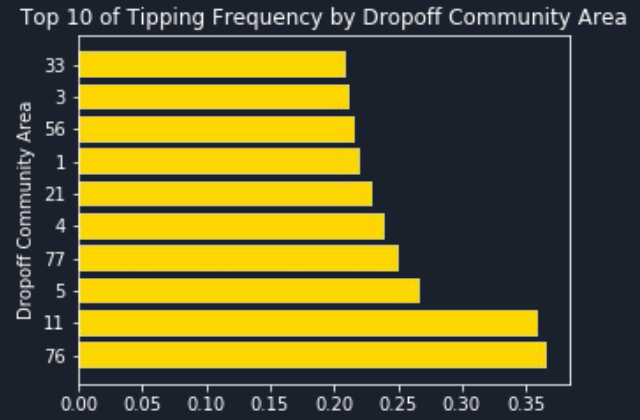


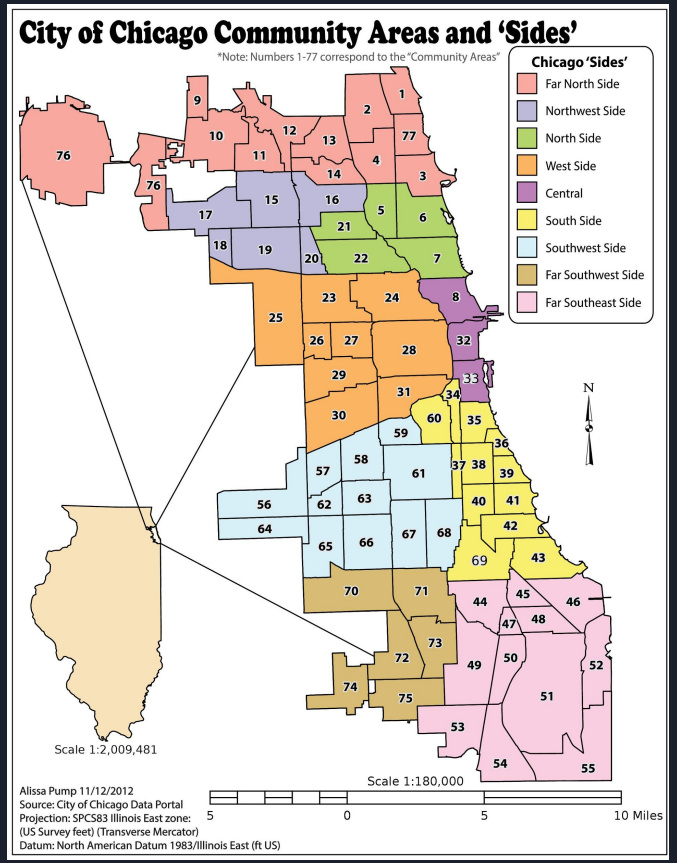




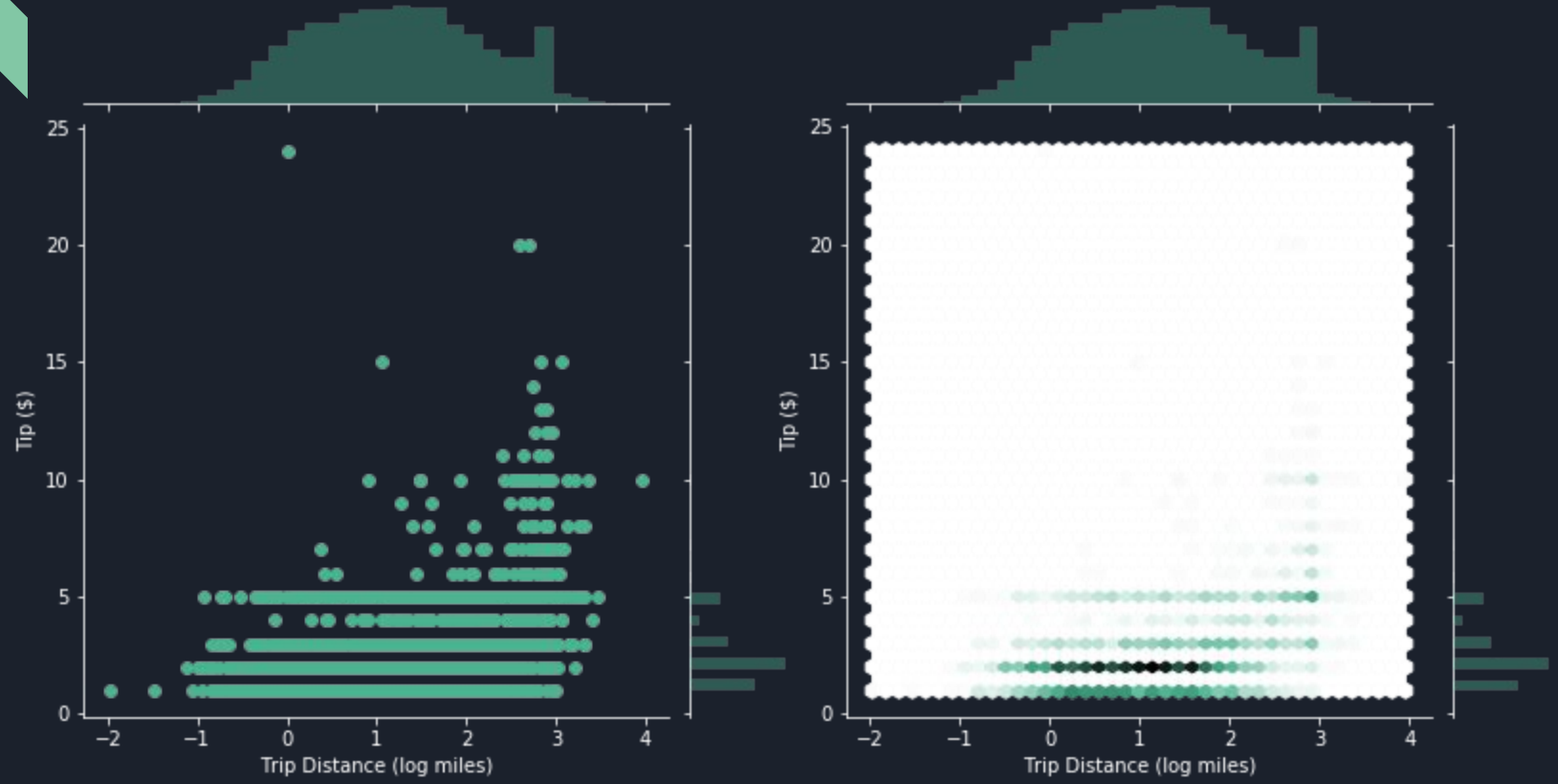
Visual Explorations - Drop Offs







Visual Explorations - Tips vs. Distance



The Model: Preprocessing

1. Winsorized for Outliers
    - Average Ride Speed
    - Average Cost Per Mile

2. Normalized Numeric Columns to reduce skew as necessary
    - Log Normalization
    - Cube Root

3. Binarizing Community Areas
    - Pickup Areas
    - Drop Off Areas

4. Dimensionality Reduction
    - Principal Components Analysis; Reduced 169 Features to 15
    - Captured 99.26% of original variability

The Model: Optimization

SMOTE

The data was imbalanced, necessary for understanding overall accuracy better and creating novel data points in a resampling set

HyperOpt

Ran 1000 iterations over hyperparameter space; smaller resampled train and tests sets for faster optimization.

The Model: Results

 99%	• Random Forest Classifier	• Trained and Tested on Same Data • Model Memorized Data Set
 85%	• Random Forest Classifier • SMOTE	• Subsequent Models: Train/Test Split • Favorable Accuracy, Realistic
 70%	• Support Vector Classifier • SMOTE	• Performed Worse than RFC
 87%	• Gradient Boosting Classifier • SMOTE	• Best Accuracy of Attempted Models
 88%	• Gradient Boosting Classifier • SMOTE • HyperOpt	• Improved Accuracy by 1% • Precision: 87.5% • Recall: 88.2%

Conclusions and Reflections

- External Ride Attributes Provide Great Tip Classification Accuracy
  - Business Initiatives; Incentivizing passengers to tip
- Precision and Recall Scores Validate Significance of Overall Accuracy based on class imbalance of data.
- Tips best predicted using pickup and dropoff locations, Ride Distance

With More Time:

- Better Understanding of API for City of Chicago
- Find most salient indicators for regression of total tips
- More creative feature engineering for exploration