NYCe TAXI !

Anand Anubhav Sindhuja

PROJECT GOALS

- Maximize tip benefits for taxi drivers
- A generic algorithm that can be extended to any city

1

DATASET

Source, Size, Relevant Fields

SOURCE

- NYC Taxi and Limousine Commission
- The data comes from several vendors who manage the meter/gps systems in the cabs

Source: http://publish.illinois.edu/dbwork/open-data/

SIZE

- Merged from 2 datasets each spanning over 4 years - 2010 to 2013
  - Trips - 116 GB (uncompressed)
  - Fares - 75 GB (uncompressed)

- Subset Chosen
  - Year 2013
  - Randomly sampled 10% of each month
  - 200,000 records from above

Source: http://publish.illinois.edu/dbwork/open-data/

RELEVANT FIELDS

Unique Fields

- medallion
- hack_license
- vendor_id

Trip Fields

- pickup_datetime
- dropoff_datetime
- trip_time
- trip_distance
- pickup_latitude
- pickup_longitude
- dropoff_latitude
- dropoff_longitude
- passenger_count

Fare Fields

- fare_amount
- **tip_amount**
- tolls_amount
- total_amount
- mta_tax
- payment_type

ADDITIONAL SOURCE

- US Census Data
  - Cost of living index
  - Median household income
  - Population Density
- Available per Zip code

2

TOOLS USED

Libraries, Infrastructure

LIBRARIES & INFRASTRUCTURE

- Language & Libraries
  - Python, BASH Scripting
  - Scikit, matplotlib
- Infrastructure
  - IBM Softlayer Cloud
  - 16 cores, 16GB RAM, 100 GB Hard Disk

3

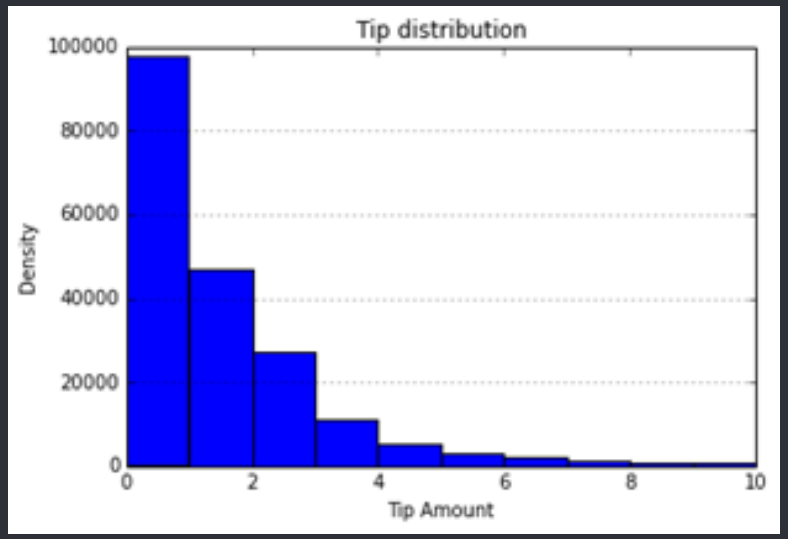
DATA EXPLORATION & FEATURES

Data Cleaning, Data Visualization, Feature Engineering

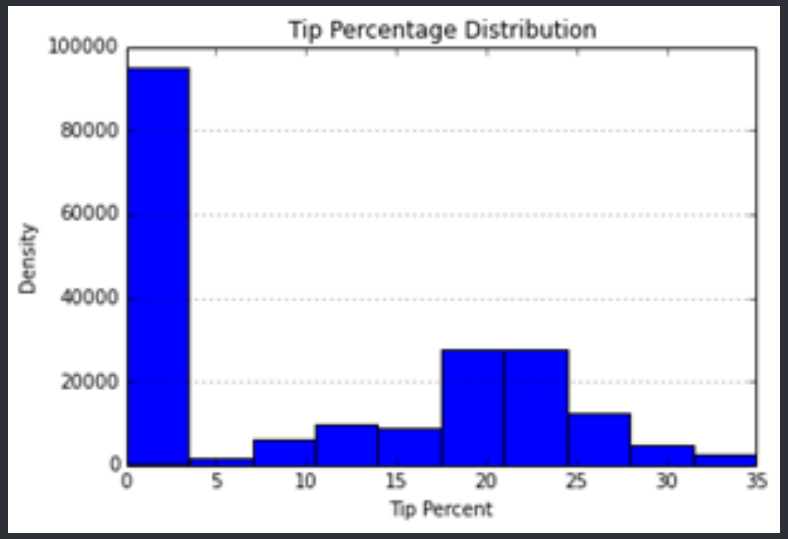
DATA EXPLORATION

- Filtering NA and missing values
- Identifying outliers and incorrect values
  - Plots to detect outliers
  - Removed incorrect values Eg. Passenger count of 250
- Visualizations to identify correlations with tip amount.

VISUALIZATIONS

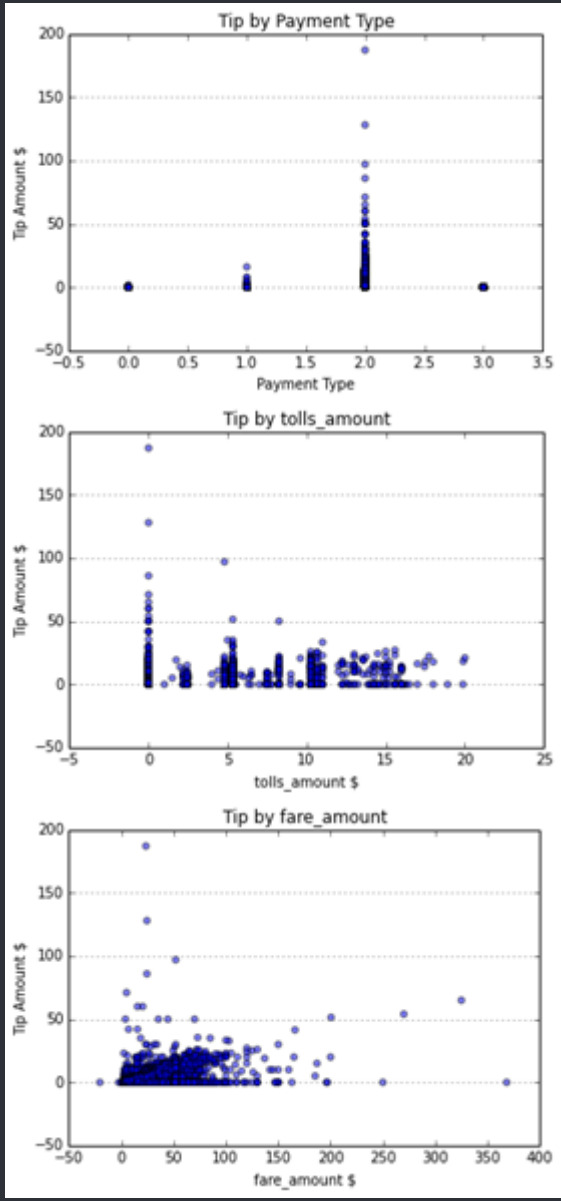


Tip Amount Density Function



Tip Percentage Density Function

CORRELATIONS



Tip Amount vs Payment Type

Tip Amount vs Tolls Amount

Tip Amount vs Fare Amount

FEATURE ENGINEERING

- Date-Time Based
  - Weekend/ Weekday
  - Time of the day
- Location Based
  - Latitude-Longitude to Zip code
  - Zip code to boroughs
  - Boroughs to demographic data

4

MODELS & METRICS

Models, Algorithms, Analysis

MODELS

- Tip - No Tip
- Tip Class w/ Zero Tip Data
- Tip Class w/o Zero Tip Data
- Tip Percent Class w/ Zero Tip Data
- Tip Percent Class w/o Zero Tip Data
- Tip Amount

ALGORITHMS

- Classification Models
  - SVM
  - Decision Tree
  - Random Forest
  - Adaboost
- Regression Models
  - Linear Regression
  - SVM Regression
  - Lasso Regression

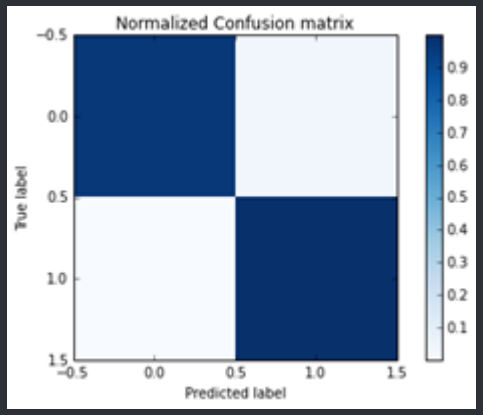
METRICS - CLASSIFICATION & REGRESSION

	Baseline	SVM	Decision Tree	Random Forest	Adaboost
Tip - No Tip	52.33	98.516	98.249	98.259	98.299
Tip Class w/	47.66	81.253	81.565	81.593	81.208
Tip Class w/o	45.07	67.98	68.002	68.002	66.72
Tip % w/	47.66	68.10	68.08	68.097	68.013
Tip % w/o	41.47	41.58	42.12	42.14	41.97

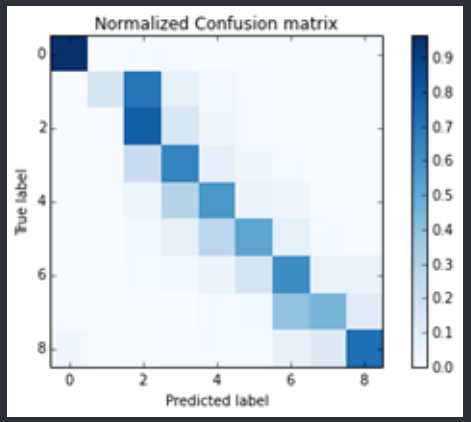
	Baseline(Mean Abs Error)	Linear	SVM	Lasso
Tip	1.38	0.75	0.79	1.254

CONFUSION MATRIX

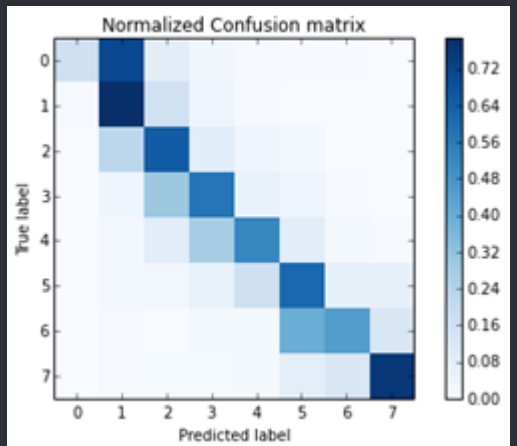
Tip - No Tip



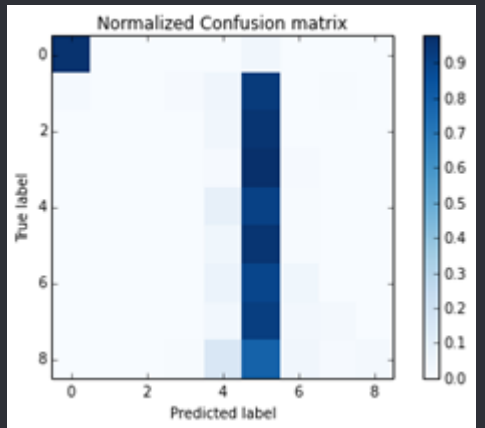
Tip Class w/o Zero Tip



Tip Class w/ Zero Tip



Tip % w/ Zero Tip



5

INFERENCES & ROADMAP

Insights Gained, Future Work

INFERENCES

- Card Payments have higher tip
- Tip varies directly with fare
- Routes with tolls yield less tip
- Tip directly varies with the cost of living index

FUTURE WORK

- Larger Data
- More feature engineering
- Apply the model to another city

QUESTIONS?

Thanks!

CREDITS

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash