Predicting Tips: New York City Taxis

Vincent Morin - vem14

12/4/2019

Statement of Purpose

Goal:

Conduct analysis into the relationships between average tip-size per day and other external variables such as weather, time/day, number of passengers, etc..

What is our best predictor variable?

Our Data

“2016 Green Taxi Trip Data”

- Includes records from all trips completed by New York’s Green Taxis (Boro Taxis) in 2016.
- The Green Taxi was created in 2011 after analysis into the transit system which showed a lack of available cabs in Upper Manhattan and the outer-boroughs.

Our Data

- Available via **NYC OpenData**
- Provided by the Taxi and Limousine Commission
- **16.4 million rows**
- **23 columns** (how many are relevant?)

Wrangling

Due to size of the data, need to scale down:

- **Criteria for removing columns:** uncertainty/unknown descriptor, constant value, and location (Note: holding location constant - NYC). - Only focus on columns which include tips.
- **Scaling:** Aggregate data around shared date and remove non-tip entries.

Our Data

```
PresData %>% glimpse()
```

```
Observations: 366
Variables: 12
$ date           <date> 2016-01-01, 2016-01-02, 2016-01-03,
$ Avg_Passengers <dbl> 1.414673, 1.385850, 1.374284, 1.3584
$ Avg_Distance   <dbl> 3.710635, 3.312719, 3.337866, 3.0135
$ Avg_Fare       <dbl> 14.21573, 13.12118, 13.16626, 12.896
$ Avg_Extra      <dbl> 0.3363794, 0.2281322, 0.2351443, 0.4
$ Avg_MTA_tax    <dbl> 0.4941176, 0.4943051, 0.4939318, 0.4
$ Avg_Tip        <dbl> 3.230683, 2.889349, 2.943467, 2.9025
$ Avg_Tolls      <dbl> 0.13535186, 0.13203838, 0.13855564,
$ Avg_Surcharge  <dbl> 0.2964706, 0.2964652, 0.2964278, 0.2
$ Avg_Total      <dbl> 18.70873, 17.16147, 17.27379, 17.173
$ Avg_Duration   <dbl> 0.3691987, 0.3268979, 0.3345052, 0.2
$ Total_Trips    <dbl> 23035, 17823, 17468, 16508, 17177, 1
```

More Wrangling

One issue: lack of variables/observations.

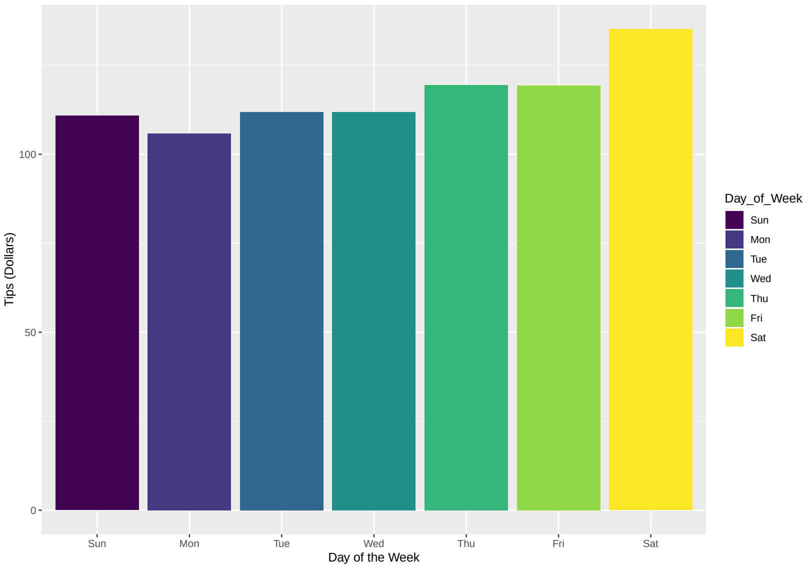
- We'll pull in 2016 New York Weather Data, and combine with inner_join.
    - Create more variables

```
PresWeather %>% glimpse()
```

```
Observations: 366
Variables: 7
$ date                  <chr> "1-1-2016", "2-1-2016", "3-1-
$ `maximum temperature` <dbl> 42, 40, 45, 36, 29, 41, 46, 4
$ `minimum temperature` <dbl> 34, 32, 35, 14, 11, 25, 31, 3
$ `average temperature` <dbl> 38.0, 36.0, 40.0, 25.0, 20.0
$ precipitation         <chr> "0.00", "0.00", "0.00", "0.0
$ `snow fall`           <chr> "0.0", "0.0", "0.0", "0.0", "
$ `snow depth`          <chr> "0", "0", "0", "0", "0", "0"
```

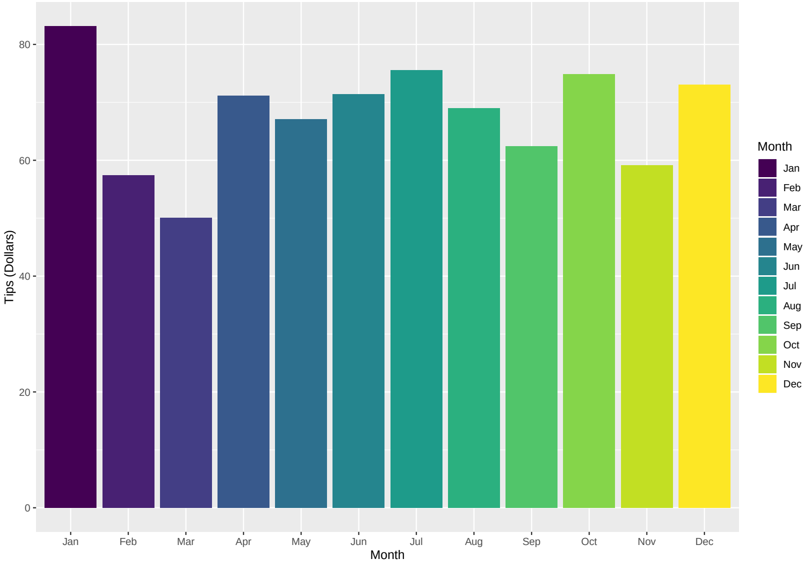
Data Analysis

Total Tips by Day of Week



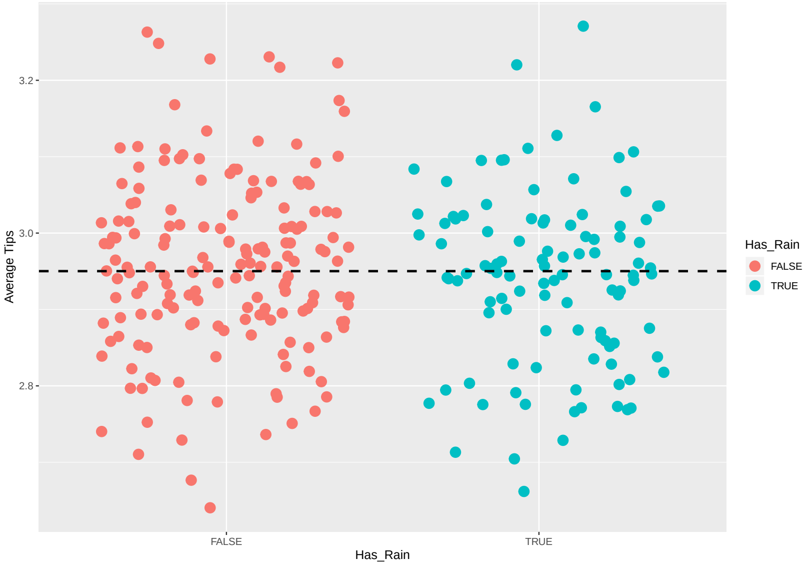
Data Analysis

Total Average Tips per Ride by Month



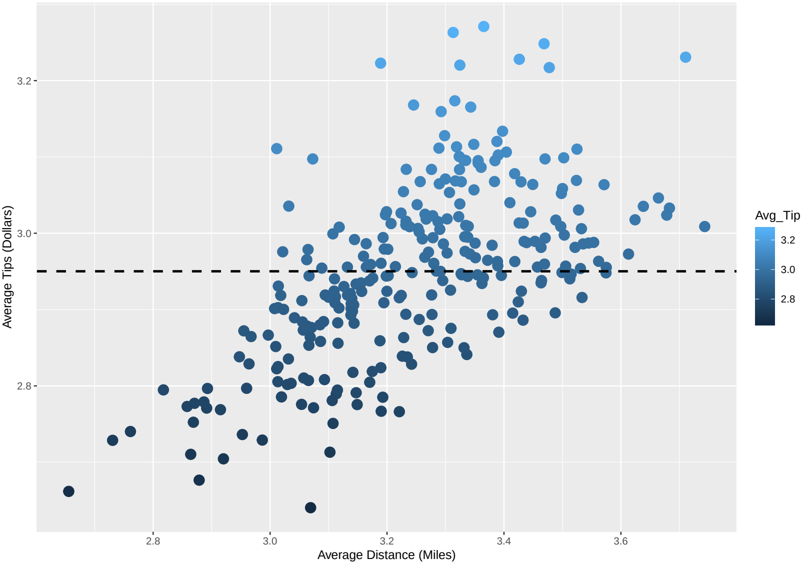
Data Analysis

Tip Variation for Rain



Data Analysis

Tips and Distance of Ride



Methods/Tools

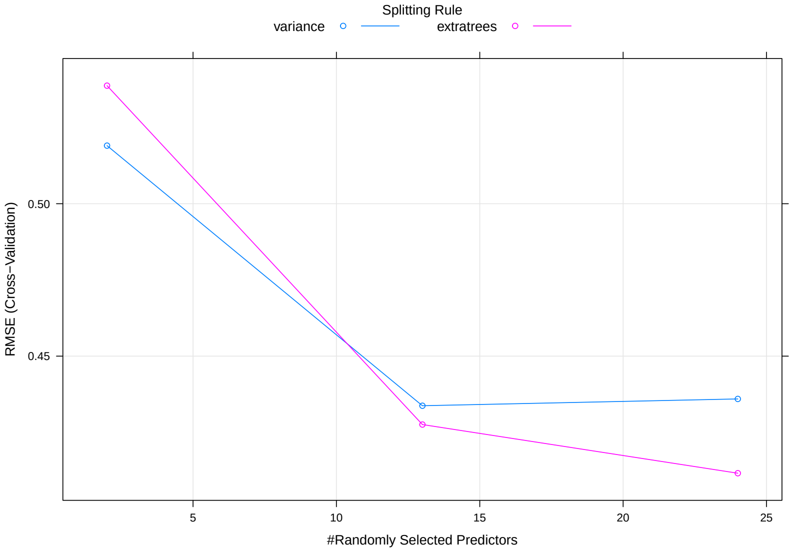
Supervised Machine Learning

Goal: build models which best predict the average tip size per ride.

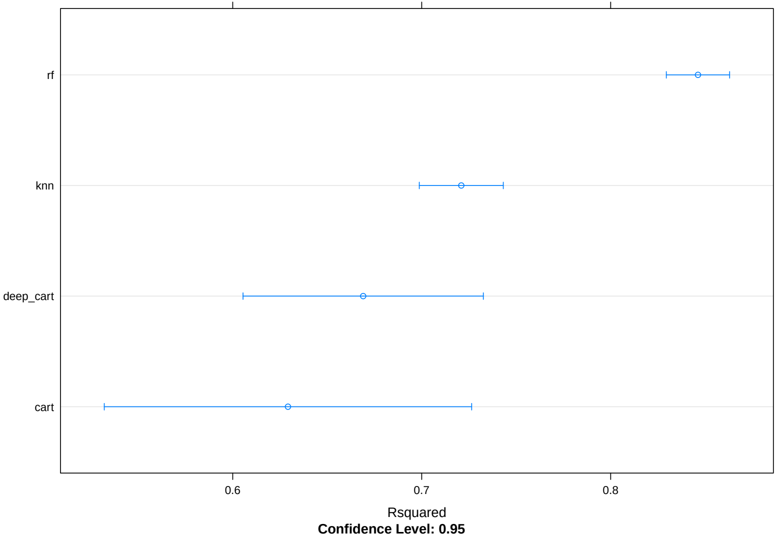
- Regression methods:
  - Linear Regression - predict Y, based on predictor X.
  - K-Nearest Neighbors - predict Y, based on similar observations in proximity.
  - Classification and Regression Trees (CART)
  - Random Forest - Decision trees acting as an ensemble.

Results

Machine Learning!



Results



Conclusions

Preliminary Results

- Results from models are not very conclusive,
- Relationships with predictor models are not very strong.
- Need more data/observations.

Lessons

- Aggregating data loses a lot of information: variability, etc. - keeping each ride versus averaging across each day.
- Large datasets will ruin your day. . .
  - Important to save large data often.
  - Manage environment and memory.
- Save models as images.
- Push to Git often.

Thanks!