Taxi Tip Prediction

A Case Study of Prediction of Taxi Tips Using NYC Taxi Data

Mike Amodeo

March 14, 2019

The problem

Company

The company wants to start a ride-hailing app and develop a tipping recommendation for riders

Context

The New York City Taxi and Limousine Commission makes all taxi ride data available

Problem statement

Using three months of taxi ride data, create a model that will recommend a tip amount to riders at the end of their trip

Data Source

NYC TLC

New York City Taxi & Limousine Commission Trip Record Data

29 Million Records in March, June, and November of 2017

Data Dictionary

Contents

- Pickup Neighborhood/Time
- Dropoff Neighborhood/Time
- Fare
- Tip
- Fare Type
- Payment Method
- Tolls, Taxes, Fees

Issues

Erroneous Data

- Negative Values
- Unfeasibly Large
- Invalid Values

Missing Data

- No Cash Tips

Confounded Data

Data Challenges: Erroneous Data

Negative Values

> 14,000 Negative Fares

~200 Negative Tips

Unfeasibly Large

192 Passengers in one ride

Trips of nearly 10,000 miles

Fares over $600,000

Invalid Values

Invalid Rate Codes (only 6 standard types)

Invalid Extras and Taxes (standard rates)

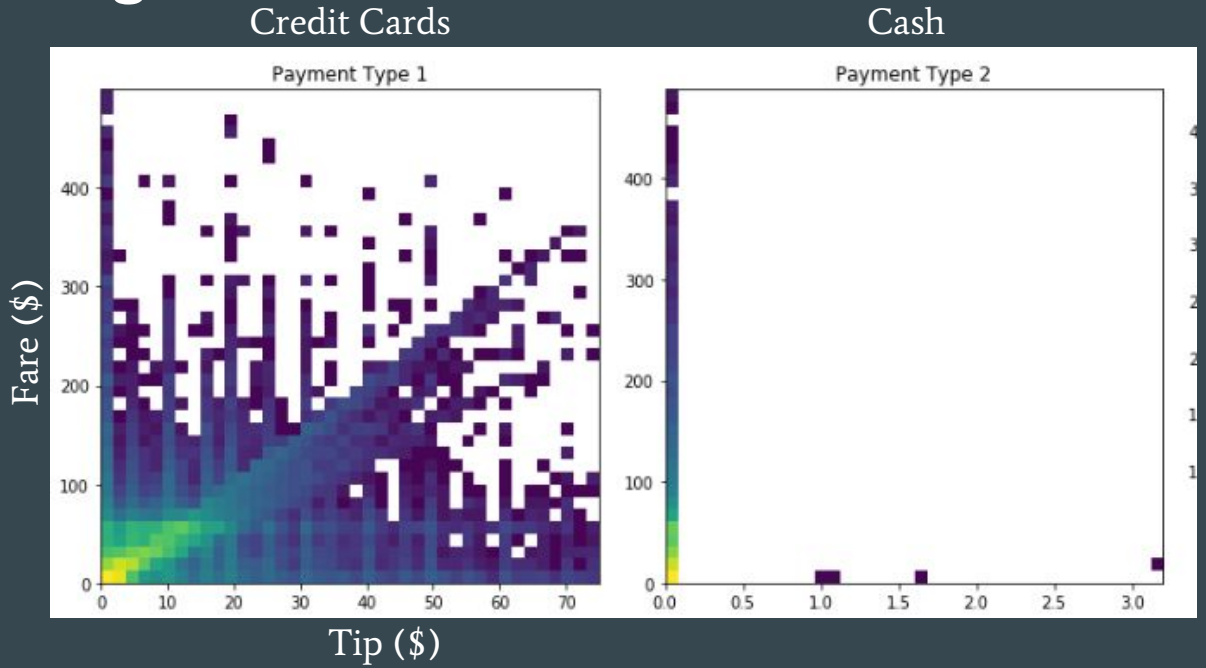
Solution: Basic filtering applied to remove obviously false records

For more detail, see the accompanying Jupyter notebook, Taxi Fares and Tips 1 EDA

Data Challenges: Missing Data

No Cash Tips

Trips paid in cash do not include any tip amount. This is most likely because drivers do not record cash tips. These trips cannot be used for model building



Solution: Cash payments not used in model building or evaluation

For more detail, see the accompanying Jupyter notebook, Taxi Fares and Tips 1 EDA

Data Checks

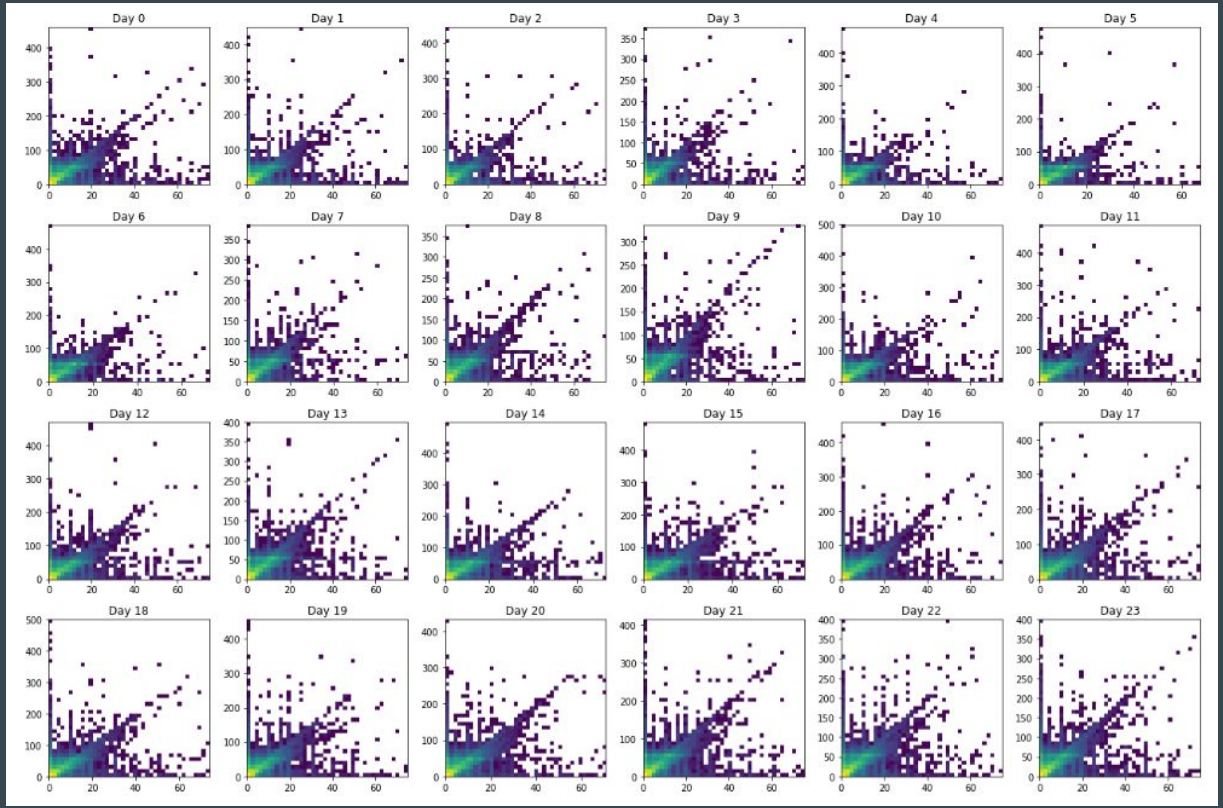
Temporal Checks

No Effects Found:

- Monthly
- Day of the Week
- Hour by Hour

Similar Tipping Patterns found.

Date and time not used for modeling.

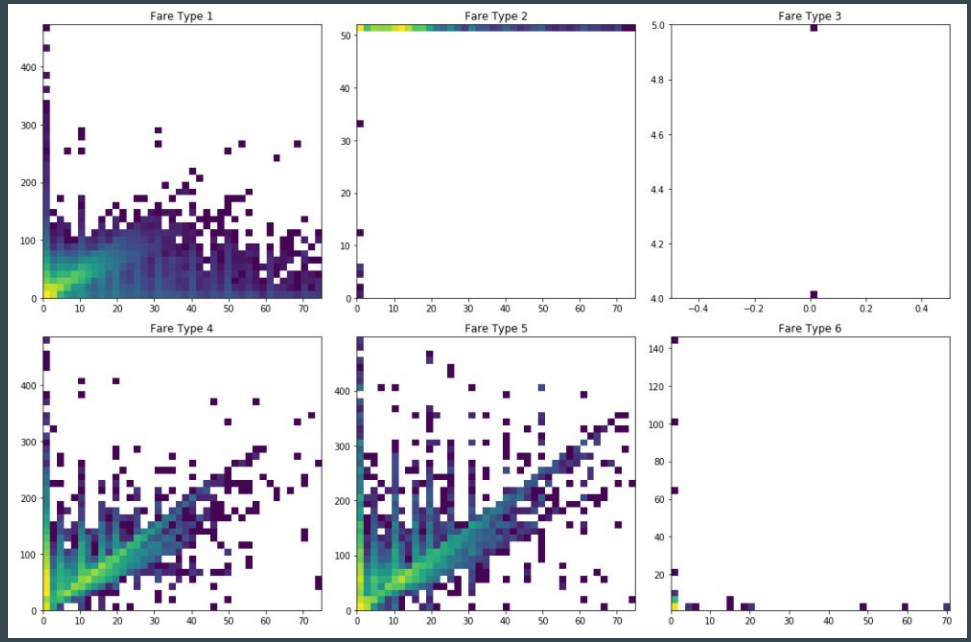


Data Checks

Fare Types

JFK Airport (2) has a fixed fare.
Will have different tipping pattern.

Long distance trips (4) and negotiated fares (5) have their own patterns, similar to standard fare (1) but slightly different.



Solution: Create indicator variables for regression modeling

For more detail, see the accompanying Jupyter notebook, Taxi Fares and Tips 1 EDA

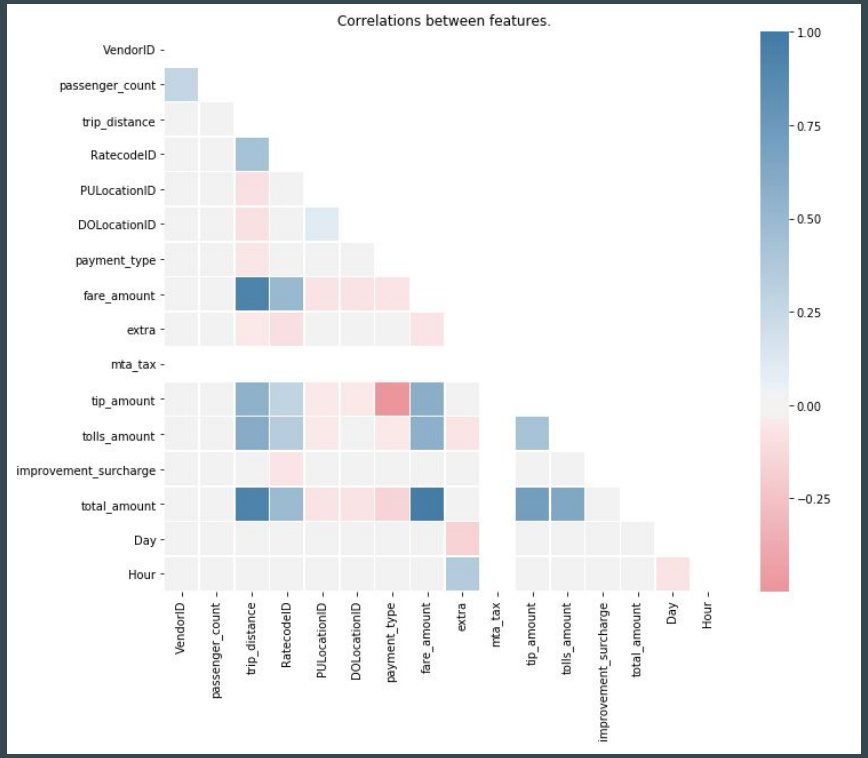
Modeling Considerations

Correlation

Check for confounded variables

Strong correlations between trip distance, fare amount, total amount, and tip amount

Total amount is also a function of tip



Solution: Only use one of the correlated cost and distance variables

For more detail, see the accompanying Jupyter notebook, Taxi Fares and Tips 1 EDA

Solution

Linear Regression Model

Four models were created with the intention of predicting a continuous variable of tip based on pre-tip fare and fare type.

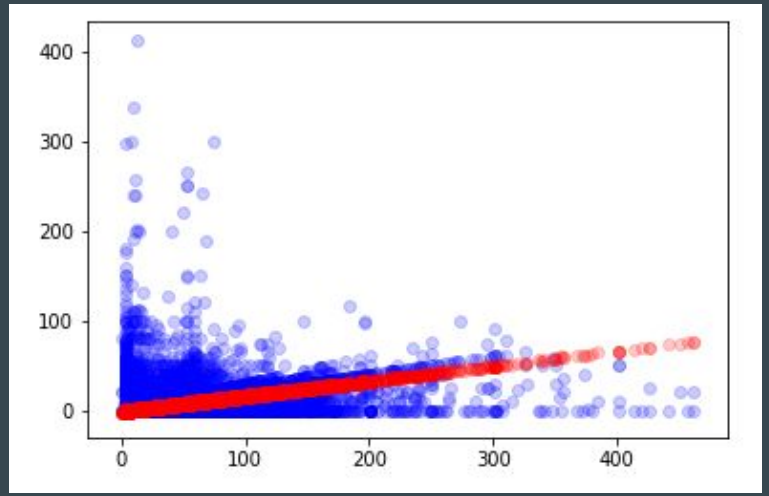
R-squared values:

- **Linear Regression - 0.59**
- SGD Regressor - 0.60
- SGD Regressor w/ Bonus Features - 0.59
- Tree Regressor - 0.20

For more detail, see the accompanying Jupyter notebook “Taxi Fares and Tips 2 Model

Evaluation

Linear Regression Model



The linear regression is the simplest model and most intuitive, so is a favored choice.

Predicted values (red) follow a straight line, but actual values show a large amount of variance, especially at lower fares.

Tips start at 17% of the pre-tip total, with some adjustments for fare type.

For more detail, see the accompanying Jupyter notebook “Taxi Fares and Tips 2 Model

Next Steps: Room for Improvement

More Data

More data sources can inform tipping:

- User profiles
- User ratings
- Driver ratings
- Location data used as categorical data

Refined Models

Rounded predictions at $0.25 intervals would allow for data to be binned and different kinds of models to be used

Decision rules like minimum tip amount

Interaction of fare and indicator variables

Polynomial regressions and higher order regression