Pricing of New York City Taxi Rides

Christophoros Antoniades | Delara Fadavi | Antoine Foba-Amon Jr.
chranto@stanford.edu | dfadavi@stanford.edu| amonjr@stanford.edu

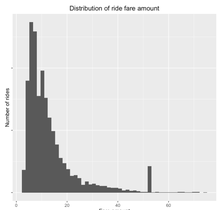
ENGINEERING

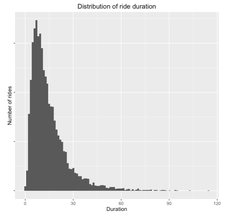
Taxi & Limousine Commission

Objective

This project uses publically available taxi data from New York City Taxi & Limousine Commission to extract insights about ride fare and duration. This information can be useful in helping drivers decide between rides to accept to increase profit or to help passengers choose times of day to minimize fare or ride time.

Taxi Pickups (blue) and Dropoffs (Yellow)[1]





Dataset

Each observation represents a single taxi ride and includes feature information such as pickup/dropoff location, time of ride, fare, tip, payment type, and more.

The dataset was cleaned to have clear covariates delineating exact times and dates of each ride. Data from May 2016 was used, which contained approximately 12 million observations of taxi rides.

8,000 observations were used as training data and 2,000 observations were used as a validation set

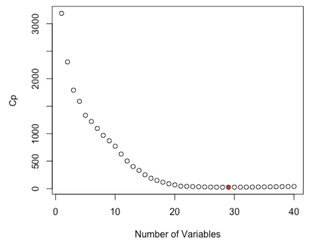
Covariates

trip_distance	pickup_longitude	pickup_latitude	dropoff_longitude
dropoff_latitude	fare_amount	extra	mta_tax
tip_amount	tolls_amount	improvement_surcharge	total_amount
manhattan_dist	shortest_dist	pickup_month	dropoff_month
pickup_year	dropoff_year	pickup_day	dropoff_day
pickup_weekday	dropoff_weekday	pickup_hour	dropoff_hour
pickup_minute	dropoff_minute	passenger_count	RatecodeID
payment_type			

Methods

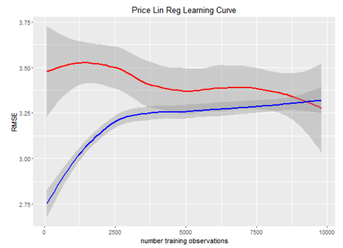
Forward Search and Lasso

- Forward search suggest keeping nearly all variables
  - Trip distance and rides in hour most important variables
- Lasso resulted in small lambda parameter and hence no significant increase in prediction accuracy



Linear regression

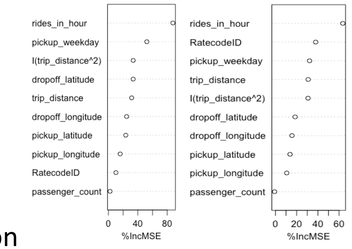
- Predicting duration and fare
  - Linear regression gives reasonable results, but has a limit to its accuracy
  - Coordinate system variables are not linear and do therefore not give significant results

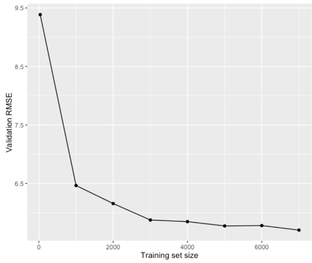


Validation RMSE (red), Training RMSE (blue)

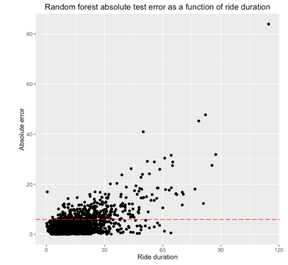
Random Forest

- 500 trees and m = n / 3 predictors per split
- Random Forest outperforms all linear regressions and Lasso
- Manages to model nonlinearity in location coordinates
- Error likely to depend on traffic and individual driving characteristics





Random Forest Validation RMSE



Absolute prediction error is proportional to ride duration

Additional Model Modifications

- Transformation of latitude/longitude coordinates
- Traffic modeling by considering rides per hour (yields small prediction improvement)

Results

Model	RMSE Validation	RMSE Train
Fare, Baseline Mean	$10.45	$10.36
Fare, Linear Regression	$3.52	$3.04
Fare, Random Forest	$2.28	$2.16
Duration, Baseline Mean	11.95 min	11.43 min
Duration, Linear Regression	6.51 min	6.17 min
Duration, Random Forest	5.24 min	5.09 min

Conclusions and Future direction

- The Random Forest model performs the best, because of the nonlinear influence of location patterns on trip duration and fare
- Prediction accuracy flattens with more variables from this data set, implying need for additional predictive variables
- Analyze more data to infer traffic conditions or other variabilities that can affect duration and fare
- Consider modelling traffic between pickup and dropoff locations

[1]N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips," *IEEE Transactions on Visualization and Computer Graphics*, 2013. [Online]. Available: https://vgc.poly.edu/~juliana/pub/taxivis-tvcg2013.pdf. Accessed: Dec. 11, 2016.