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.

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

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

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

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE Validation</th>
<th>RMSE Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fare, Baseline Mean</td>
<td>$10.45$</td>
<td>$10.36$</td>
</tr>
<tr>
<td>Fare, Linear Regression</td>
<td>$3.52$</td>
<td>$3.04$</td>
</tr>
<tr>
<td>Fare, Random Forest</td>
<td>$2.28$</td>
<td>$2.16$</td>
</tr>
<tr>
<td>Duration, Baseline Mean</td>
<td>$11.95$ min</td>
<td>$11.43$ min</td>
</tr>
<tr>
<td>Duration, Linear Regression</td>
<td>$6.51$ min</td>
<td>$6.17$ min</td>
</tr>
<tr>
<td>Duration, Random Forest</td>
<td>$5.24$ min</td>
<td>$5.09$ min</td>
</tr>
</tbody>
</table>

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`
- `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`

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

---