NYC Ride Hailing Optimization – Fare Amount, Pickup Points and Collision Predictive Analytics

Amol Wadwalkar, Jagannadh Commuri, Sriram Ayyagari, Suresh Kadavath, Promodh N Ravichandran

Motivation

- Traditional Taxi drivers lack the data-driven insights available to Uber/Lyft
- Unavailability of such insights put Taxi drivers in a significant disadvantage to improve their profit margin and compete with the Tech Giants.
- Our tool will help taxi drivers to maximize profit and reduce risk of collision.

Data

- Datasets were downloaded from
  - NYC TLC – Taxi dataset
  - NYC Open Data – Collision dataset
  - Five-Thirty-Eight – Uber dataset
- Dataset Characteristics
  - Size after data cleaning and feature selection – 15 GB
  - 180 mil records – not temporal

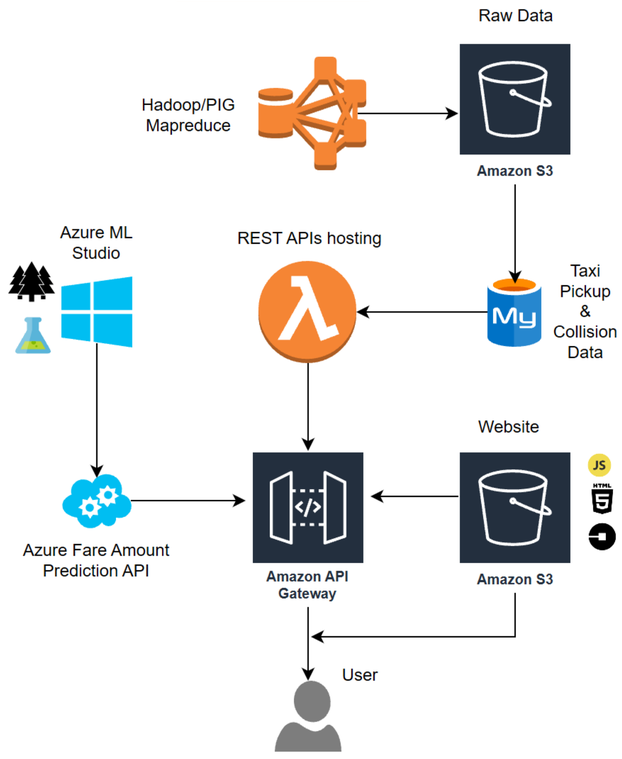
Innovation

- Provide a website for traditional taxi drivers with diverse datapoints impacting profitability and safety.
- Incorporate collision prediction while displaying high pickup zones and price per mile to avoid pickup zones with a high probability of collision
- Ability to check prediction for any location within NYC and for any day and hour for better day planning

Machine Learning Algorithms

- **Fast Forest Quantile Regression** to predict fare amount distribution over quantiles – uses Random Forest of decision trees with bagging.
- **Multi Class Decision Jungle** to predict pickup counts and collision probability – uses an ensemble of decision directed acyclic graphs. Provides lower memory footprint and better generalization at the cost of higher training time and is also resilient to noisy features

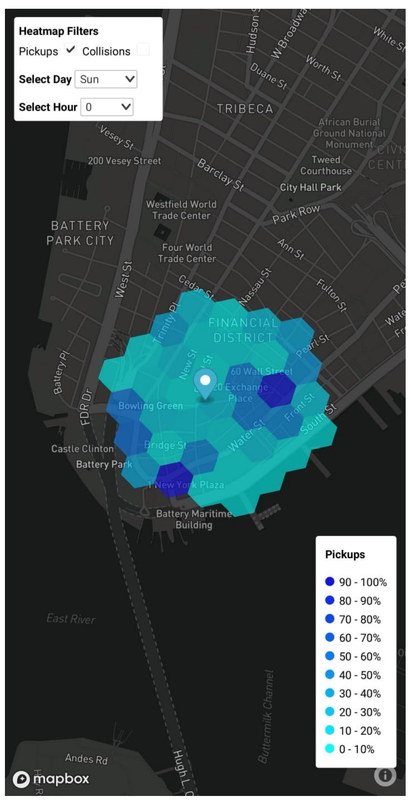
Architecture



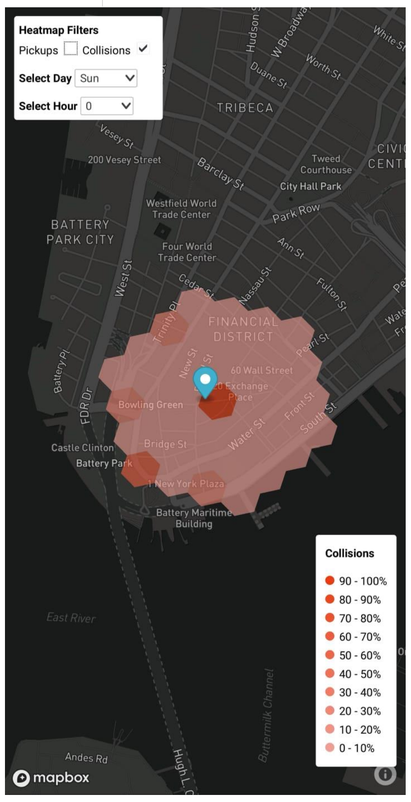
Data Visualization Algorithms

- **H3** – Hexagonal hierarchical geospatial Indexing system – using hexagons provides the benefit of equidistant neighbours as compared to squares or triangles and have a property of expanding ring of neighbours.

Pickup Points Visualization



Collision Visualization



Combined Weighted Visualization

Experiments

- Linear regression with cross-validation model had a very low coefficient of determination (0.2 - 0.4). So we used the Multiclass Decision Jungle model which provided accuracy of prediction as 80%.
- Tableau was evaluated for visualization, but due to the licensing cost involved and absence of H3 support, JS was finalized.

Results

Prediction	Mean	Median	Min	Max	StdDev
Pickup	25.21	26.82	3.25	33.71	7.51
Fare	2.49	2.70	0.66	2.78	0.76
Collision(R)	0.001	0.0008	0.00	0.12	0.001
Collision(O)	0.19	0.187	0.04	0.40	0.04
Collision(Y)	0.8	0.81	0.60	0.95	0.04

Technology Stack