Taxi Fare Prediction

Ashwath & Cathal

Today’s Agenda

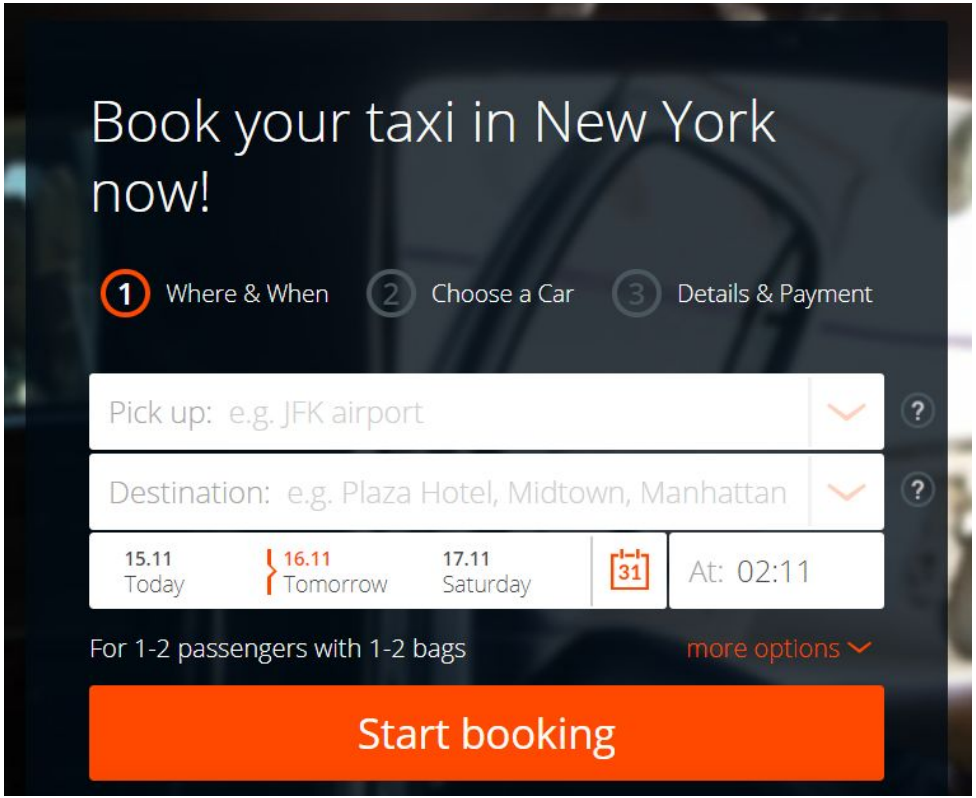
- Introduction
- Datasets
- Preprocessing & Model Selection
- Evaluation
- Business Impact

Problem Statement

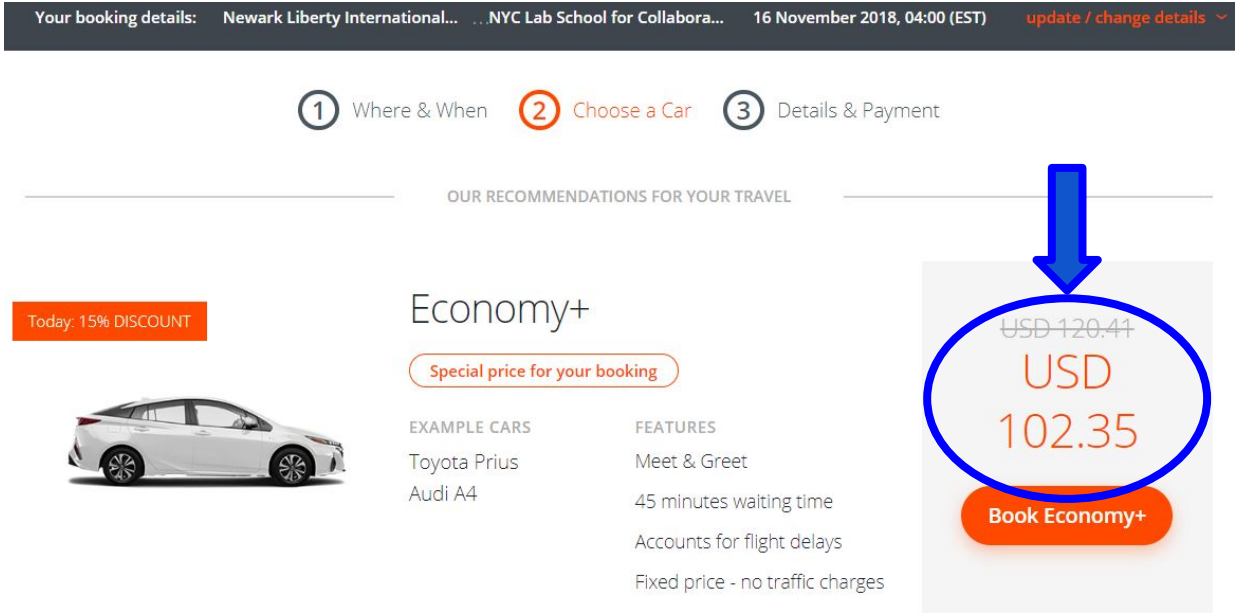
A taxi booking service provider called Talixo wants to increase its market share in New York. The aim is to provide near real-time taxi fare amount to the customers which is near to the Uber's fare prediction. Talixo is unique as it charges fixed taxi fare amount to its customers.

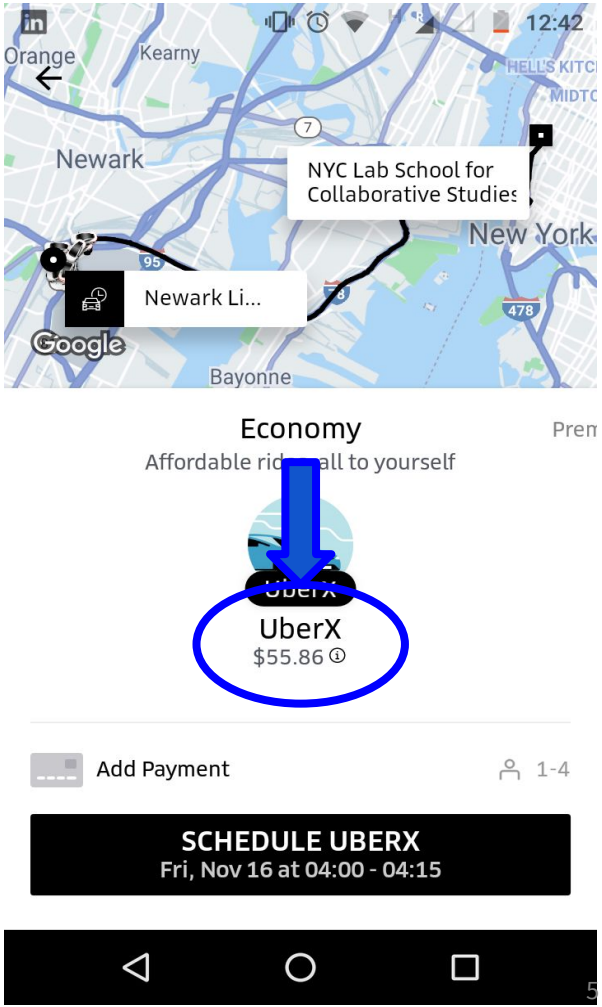
Note: Talixo company logo is used for educational purpose.

Background


Book your taxi in New York now!
1 Where & When
2 Choose a Car
3 Details & Payment
Pick up: e.g. JFK airport
Destination: e.g. Plaza Hotel, Midtown, Manhattan
15.11 Today
16.11 Tomorrow
17.11 Saturday
31
At: 02:11
For 1-2 passengers with 1-2 bags
more options
Start booking

An Example


Your booking details: Newark Liberty International... ...NYC Lab School for Collabora... 16 November 2018, 04:00 (EST) update / change details
1 Where & When
2 Choose a Car
3 Details & Payment
OUR RECOMMENDATIONS FOR YOUR TRAVEL
Today: 15% DISCOUNT
Economy+
Special price for your booking
EXAMPLE CARS
Toyota Prius
Audi A4
FEATURES
Meet & Greet
45 minutes waiting time
Accounts for flight delays
Fixed price - no traffic charges
USD 120.41
USD
102.35
Book Economy+


12:42
Orange
Kearny
Newark
NYC Lab School for Collaborative Studies
New York
Newark Li...
Google
Bayonne
Economy
Affordable ride all to yourself
UberX
$55.86
Add Payment
1-4
SCHEDULE UBERX
Fri, Nov 16 at 04:00 - 04:15

Dataset

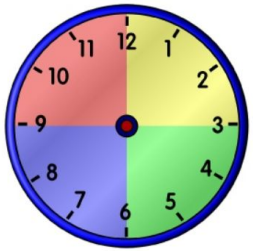
New York Taxi Trip Dataset

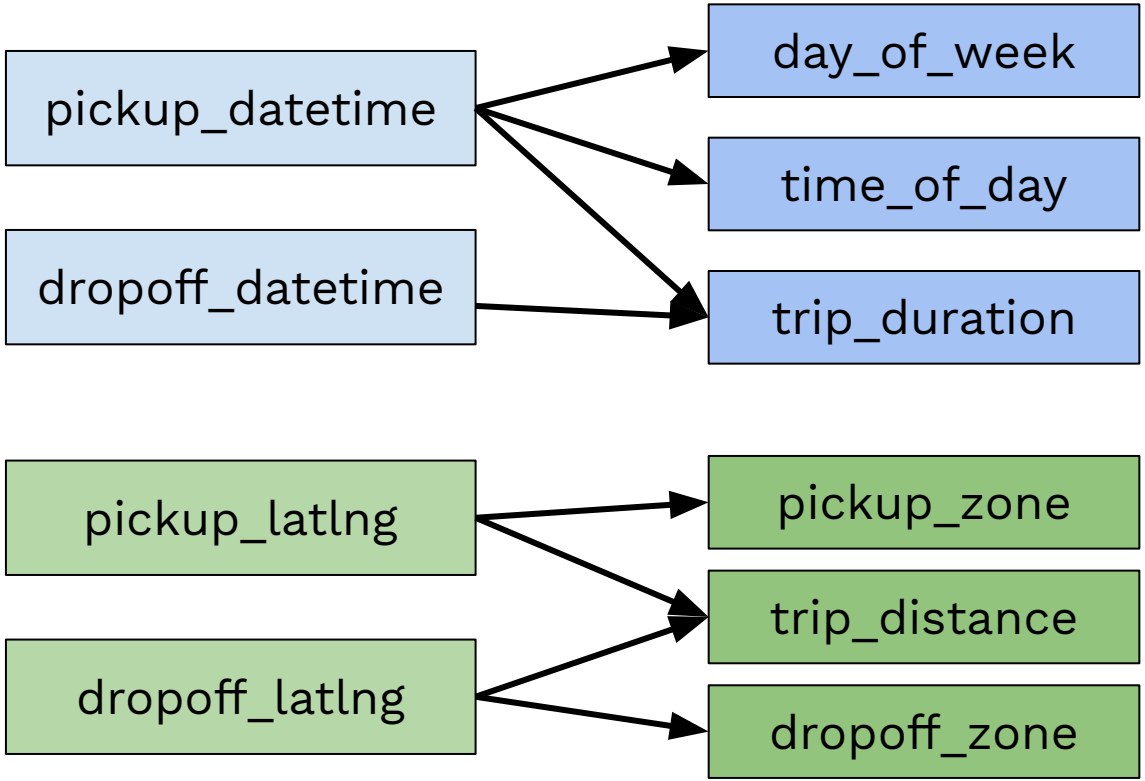
- Provided publicly by the NYC Taxi & Limousine Commission
- 19 Features
- ~10m trips per year
- Released every month (most recent June 2018)

pickup_datetime
dropoff_datetime
passenger_count
trip_distance
pickup_longitude
pickup_latitude
dropoff_longitude
dropoff_latitude
...

Feature Extraction

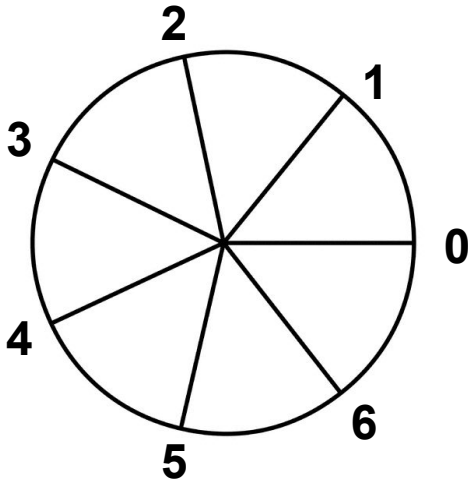
Generated Features

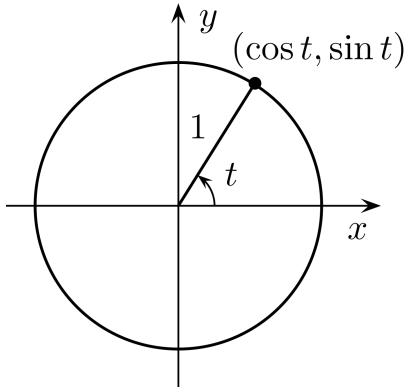




Cyclical Features

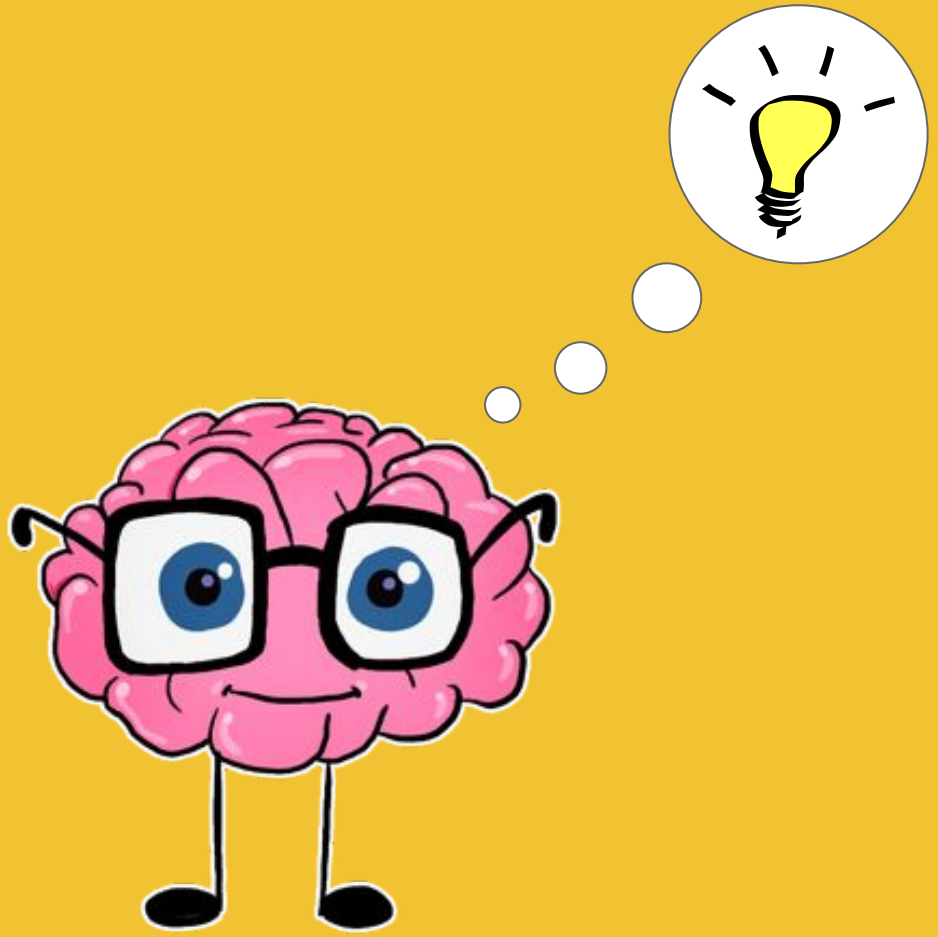
day_of_week	*d*
Monday	0
Tuesday	1
Wednesday	2
Thursday	3
Friday	4
Saturday	5
Sunday	6





$t = 2\pi * d$

Models



Ensemble techniques

- lightGBM and CatBoost
- Both are gradient boosting frameworks which use tree based learning algorithms
- Allow specifying custom loss functions
- Automatic early stopping
- In-built ability to handle categorical variables

Microsoft
LightGBM

Yandex
CatBoost

Evaluation & Results

Metric

Mean Absolute Percentage Error (MAPE)

It usually expresses accuracy as a percentage, and is defined by the formula:

$$M = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

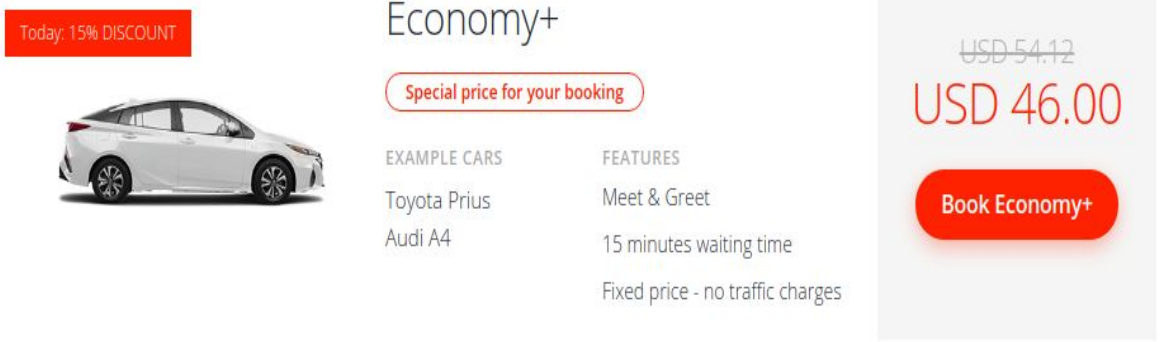
where A_t is the actual value and F_t is the forecast value.

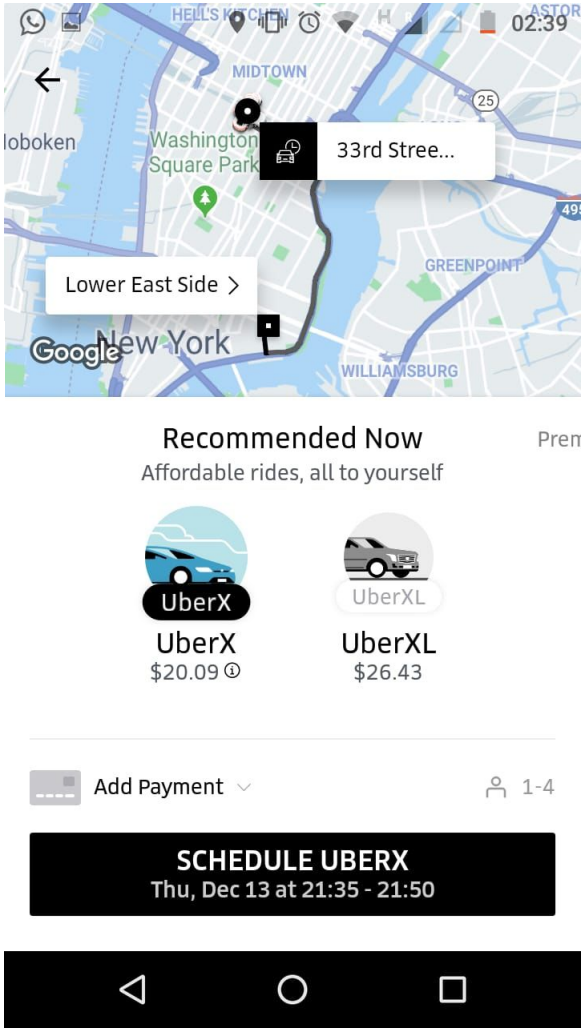
Results

Model	train_RMSE	validation_RMSE	test_RMSE	test_MAPE
lightGBM	1.00717	1.43951	2.45256	4.14%
CatBoost	0.789084	1.45098	2.45798	3.60%

Business Impact

Fare Prediction Comparison





Added Business Value

- Uber predicts around $20 for a 4 mile trip.
- And, Using the newly developed model we predicted $21.574 for the same trip distance, same region and same time window.

New fare = Predicted fare + Threshold

Taxi Fare Prediction

Ashwath & Cathal