Data Science Initiative
BROWN

Uber/Lyft Price Prediction

Jinjia Zhang
Data Science Institute
Oct 21, 2024

https://www.reuters.com/business/autos-transportation/group-backed-by-uber-lyft-pushes-massachusetts-gig-worker-ballot-measure-2021-08-04/

Data Science Initiative
BROWN

Introduction

Problems Trying to Solve and the Importance

- How do the riding apps determine the price of a ride?
- How do the prices change in different apps/locations/weather conditions…?
- For passengers: to minimize the cost of a ride
- For drivers: to optimize the revenue of a ride

Type of the Problem

- Predict and compare the price of a ride of Uber/Lyft —— regression problem
- Dataset: 693071 ride instances with 57 features and missing values

Data Source and collection

- The dataset comes from the ***Kaggle*** website.
- The data was gathered from various entities including Uber and Lyft from 11-26-2018 to 12-18-2018 in Boston, MA.

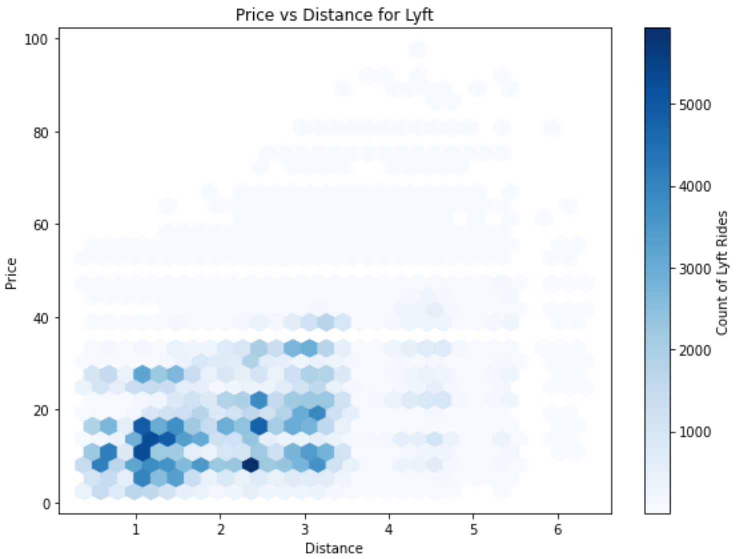
Data Science Initiative
BROWN

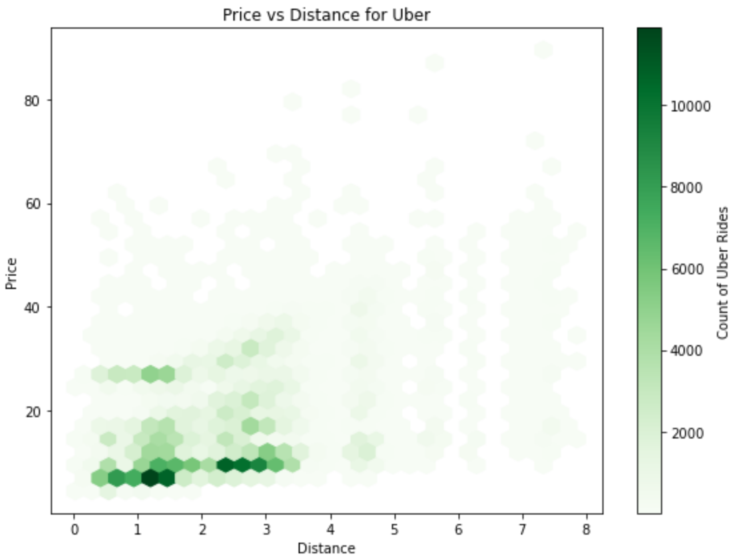
EDA: Part I

- 693071 ride instances with 57 features
- Target feature: **price**
- 55095 missing values, only in the **price** column
- Some important features: source, destination, cab_type, distance, temperature, precipitation, and other weather features
- Exists irrelevant features (timezone, latitude/longitude) , and redundant features about weather information

EDA: Part II

Visualization I: Price vs Distance for Lyft and Uber

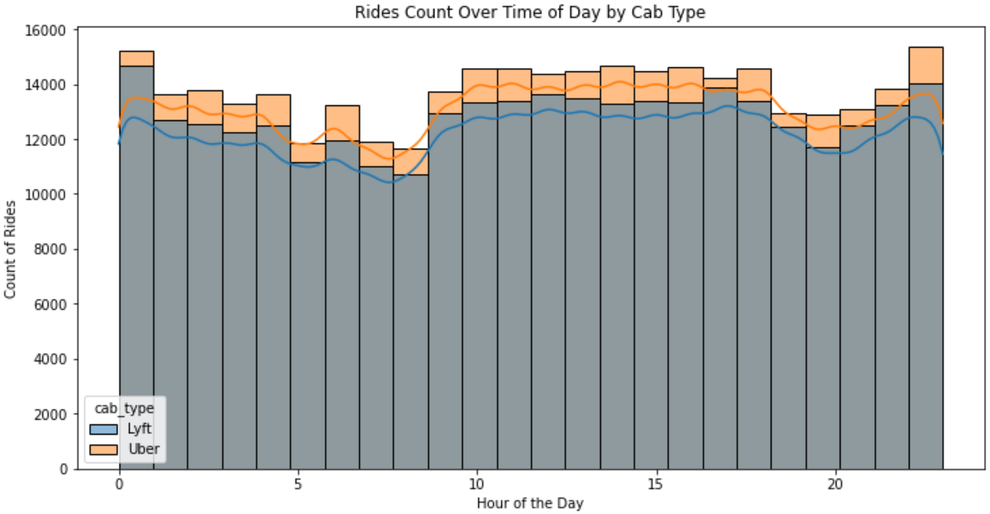




- Lyft has a wider range of prices compared to Uber
- For both Lyft and Uber, Price is correlated with distance, but not so strong

EDA: Part II

Visualization II: Distribution of rides over one day

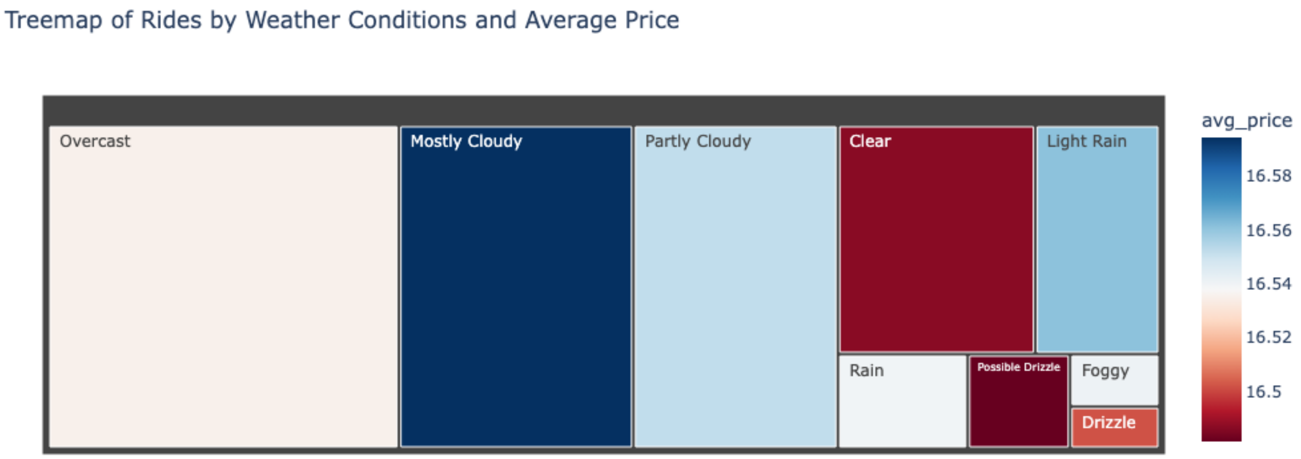


- At any time, Uber has more rides than Lyft
- 10:00 am to 7 pm and 11:00 pm to 1 am are peak periods, while 5 am to 9 am is off-peak hours.

Data Science Initiative
BROWN

EDA: Part II

Visualization III: Treemap of price vs weather condition



- The size of blocks represents the number of rides
- People tend to call a cab on cloudy days
- In cloudy days, the price is a little bit higher than usual

Data Science Initiative
BROWN

Splitting and Preprocessing

Splitting

- The dataset is **iid** since each instance in the dataset is an independent ride
- Apply **basic split** to the dataset (train 70%, validation 15%, test 15%, random_state=42)

Preprocessing

- Data shape **before** preprocessing: (693071, 57)
- Remove rides with missing prices
- Drop redundant and irrelevant columns

Data Science Initiative
BROWN

Splitting and Preprocessing

Preprocessing

- **OneHotEncoder** for categorical features, and **StandardScaler** for continuous features
- Categorical features: hour, day, month, cab_type, source, destination, short_summary, name
- Continuous features: surge_multiplier, distance, temperature, precipIntensity, precipProbability, humidity, windSpeed, windGust, visibility, dewPoint, pressure, cloudCover, uvIndex, ozone, moonPhase
- Data shape **after** preprocessing: X_train: (446583,105), X_val:(95696,105), X_test:(95697,105)