Holly Cohan,Yesh Onipede, Yu Xia

Week 2 Report: Identify the Problem Statement and Dataset

Dataset Overview

The dataset we're using from KaggleLinks to an external site. It contains 57 columns, which provide a rich variety of information to predict rideshare prices in Boston during November and December 2018. Alongside the unique identifier and target variable (price), the dataset includes 55 predictive features that capture details about the time, location, weather conditions, and rideshare-specific information.

Target Variables

The target variable in our dataset is **'price'**, which represents the rideshare fare in US dollars. This variable is chosen as the target because our primary objective is to predict the cost of a rideshare based on various factors. Accurately predicting ride prices is crucial for understanding how different elements, such as time, location, and weather, affect fares. By modeling the price, we can better forecast pricing trends, optimize ride scheduling, and potentially provide cost estimates to users in real-time.

Predictor Variables

The **predictors** are the variables that help explain or influence the target variable, 'price.' In our dataset, the predictors include:

- **Temporal variables** (e.g., *'hour'*, *'day'*, *'month'*, *'timestamp'*, *'datetime'*): These help capture the time-based factors that influence rideshare pricing, such as peak hours, days of the week, or seasonal effects during November and December.
- **Location variables** (e.g., *'source'*, *'destination'*, *'latitude'*, *'longitude'*): These provide insight into the geographical factors affecting pricing. Certain areas may have higher demand, leading to increased prices.
- **Ride-specific variables** (e.g., *'cab_type'*, *'product_id'*, *'name'*): These describe the type of rideshare service being used (e.g., UberX, Lyft Lux), which impacts pricing. Different service levels have distinct fare structures.
- **Distance and surge pricing** (e.g., *'distance'*, *'surge_multiplier'*): Distance traveled and surge pricing directly influence the final fare. Surge pricing reflects periods of high demand, and longer distances naturally lead to higher costs.

- **Weather conditions** (e.g., *'temperature'*, *'precipProbability'*, *'humidity'*, *'windSpeed'*): Weather factors, such as temperature or rain, can impact the ease and safety of travel, as well as demand, potentially leading to price fluctuations.

These predictors were chosen because they offer a comprehensive view of the elements influencing rideshare pricing. Incorporating a variety of features, from temporal and geographical data to weather conditions and ride-specific details, enables us to build a robust model capable of capturing the many factors that affect ride costs.

Data Types

The original data types imported csv are shown in the data dictionary (Appendix 1.A), but `id` and `datetime` shouldn't be objects rather than integers and datetime. While some of the data types needed to be changed it is also helpful to refer back to the data dictionary when interpreting models. In addition, we are not going to include`id` and `datetime` as our features, because `id` is just a unique identifier, and datetime is already split into variables, like 'timestamp', 'hour', 'day', and 'month'.

Here are the numerical variables and the categorical variables for future reference.

- **Numerical Variables:** ['timestamp', 'hour', 'day', 'month', 'price', 'distance', 'surge_multiplier', 'latitude', 'longitude', 'temperature', 'apparentTemperature', 'precipIntensity', 'precipProbability', 'humidity', 'windSpeed', 'windGust', 'windGustTime', 'visibility', 'temperatureHigh', 'temperatureHighTime', 'temperatureLow', 'temperatureLowTime', 'apparentTemperatureHigh', 'apparentTemperatureHighTime', 'apparentTemperatureLow', 'apparentTemperatureLowTime', 'dewPoint', 'pressure', 'windBearing', 'cloudCover', 'uvIndex', 'visibility.1', 'ozone', 'sunriseTime', 'sunsetTime', 'moonPhase', 'precipIntensityMax', 'uvIndexTime', 'temperatureMin', 'temperatureMinTime', 'temperatureMax', 'temperatureMaxTime', 'apparentTemperatureMin', 'apparentTemperatureMinTime', 'apparentTemperatureMax', 'apparentTemperatureMaxTime']
- **Categorical Variables**: ['timezone', 'source', 'destination', 'cab_type', 'product_id', 'name', 'short_summary', 'long_summary', 'icon']

Next Step

The data ingestion and exploration process also provided us some insights about what we need to do for data preparation. Some takeaways includes:

- There are 55,095 rows with missing price, which we might want to drop or impute with median/mean.
- Many of the numerical variables in the dataset show skewness and contain a significant number of outliers, suggesting that their distributions differ from normality. This could affect the performance of various statistical models and machine learning algorithms. To mitigate these issues, it's recommended to standardize the numerical variables. Thus, standardization is likely necessary.
- A lot of variables are highly correlated with each other, which can lead to multicollinearity for regression models. It also indicates needs for feature engineering, and there are several methods we are considering to approach this problem, including removing variables, combining variables, applying regularization methods, and etc.
- There are 10 categorical variable that needed to be encoded, some of which are not very intuitive. For example, `short_summary`, `long_summary`, `icon` are all summary for weather, which seems a bit redundant considering we have a lot of numeric weather variables, so it indicates additional exploration and correlation analysis needed with other numeric weather variable.

Appendix 1.A Data Dictionary

Column Name	Data Type	Null Values	Description
id	object	0	Unique Identifier for each column
timestamp	float64	0	Unix Timestamp
hour	int64	0	Hour of the day
day	int64	0	Day of the week
month	int64	0	Month in a year
datetime	object	0	Date value
timezone	object	0	Timezone
source	object	0	Initial source of the ride
destination	object	0	Destination of the ride
cab_type	object	0	The type of cab (either an Uber or Lyft cab)
product_id	object	0	Uber/Lyft identifier for cab-type
name	object	0	Type of ride (i.e. UberX, UberXL, Lyft Lux, Lyft Shared)
price	float64	55095	Price of the ride
distance	float64	0	Total distance of the requested ride
surge_multiplier	float64	0	The multiplier by which price was increased (the default is 1)
latitude	float64	0	The latitude of the ride's pick up/drop off
longitude	float64	0	The longitude of the ride's pickup/drop off
temperature	float64	0	Temperature in Fahrenheit
apparentTemperature	float64	0	The temperature that it felt like at the time of the ride
short_summary	object	0	Short written summary of the weather conditions in the hour of the ride
long_summary	object	0	Longer written summary of the

			weather conditions in the hour of the ride
precipIntensity	float64	0	The rate at which precipitation fell throughout the hour in which the ride took place
precipProbability	float64	0	Whether or not there is a chance of precipitation (0 or 1)
humidity	float64	0	Humidity (%)
windSpeed	float64	0	The sustained, constant wind speed (mph)
windGust	float64	0	The speed of the peak wind gust(mph)
windGustTime	int64	0	The time that the peak wind gust was recorded (in Unix timestamp)
visibility	float64	0	The maximum distance at which objects could be clearly seen
temperatureHigh	float64	0	The high temperature in the hour of the ride
temperatureHighTime	int64	0	The Unix timestamp of when the high temperature was recorded
temperatureLow	float64	0	The low temperature during the hour of the ride
temperatureLowTime	int64	0	The Unix timestamp of when the low temperature was recorded
apparentTemperatureHigh	float64	0	The “feels-like” high temperature when taking into account variables like wind and humidity
apparentTemperatureHighTime	int64	0	The Unix timestamp recorded when the high temperature was taken
apparentTemperatureLow	float64	0	The “feels-like” low temperature when taking into account t variables like wind and humidity
apparentTemperatureLowTime	int64	0	The Unix timestamp recorded from the low temperature
icon	object	0	The icon on the weather app during the time of the ride
dewPoint	float64	0	Relates to the absolute moisture

			content in the air
pressure	float64	0	Pressure (mb)
windBearing	int64	0	he direction from which the wind is blowing, measured in degrees from true north
cloudCover	float64	0	The fraction of the sky covered by clouds
uvIndex	int64	0	The strength of UV radiation from the sun
visibility.1	float64	0	The maximum distance at which objects could be clearly seen
ozone	float64	0	The concentration of ozone in the atmosphere
sunriseTime	int64	0	The time that the sun rose on the day of the ride
sunsetTime	int64	0	The time that the sun set on the day of the ride
moonPhase	float64	0	The current phase of the moon
precipIntensityMax	float64	0	The maximum rate of precipitation observed
uvIndexTime	int64	0	The time at which the UV index measurement was reported
temperatureMin	float64	0	The minimum temperature recorded
temperatureMinTime	int64	0	The time at which the minimum temperature was recorded
temperatureMax	float64	0	The maximum temperature recorded
temperatureMaxTime	int64	0	The time at which the maximum temperature was recorded
apparentTemperatureMin	float64	0	The minimum temperature as it feels to people considering other weather factors
apparentTemperatureMinTime	int64	0	The time at which the apparent minimum temperature was recorded
apparentTemperatureMax	float64	0	The maximum temperature as it feels to people considering other weather factors

apparentTemperature MaxTime	int64	0	The time at which the apparent maximum temperature was recorded

Notes about the variable descriptions:

- Weather data was queried every one hour.
- The dataset only provides one latitude variable and one longitude variable per ride. Therefore we are unsure if these coordinates refer to the pick up location or the drop off location of the ride.
- There are two columns related to visibility. One is called visibility and the other is called visibility.1. We are unsure about the difference between these two