CUSTOMER ANALYTICS – **CAPSTONE PROJECT**

CAPSTONE PROJECT

CUSTOMER SEGMENTATION WITH ML

By Mansi More
#989469774:
Customer Analytics /
Capstone Project

DATASET

The Uber and Lyft Dataset includes all data regarding the ride-hailing services in Boston, Massachusetts provided to study the behaviors for similar rental services. The available data enable understanding of factors that affect ride selection, leading to its categorization and corresponding business models.

1.BUSINESS LOGIC:

- The dataset further helps capture the most important features that define the ride-sharing business in Boston. When did, it will enable the ride-sharing corporations to categorize customers, design the most suitable promotions, and design alterations that could augment consumers' satisfaction and boost growth.
- Without this data it is impossible to develop the right business logic, and, at the same time, this data provides the basis for data-driven decision-making that can result in gaining a competitive advantage in the ride-sharing segment.

- **Customer Segmentation Stratgies:**

- We will study the dataset for all kinds of cab_types which will help us understand what possible strategies we can use to achieve the target.
- Customer Segmentation means dividing customers with similar features into distinct groups so that we can provide services according to their requirements.
- In our Dataset, customer segments will help us to find the pattern of the dataset, for example, different kinds of rides, location, time, and many other features.
- Let's find what all is there in the dataset. Companies can offer a membership or a discount services to the customers who have a long-distance journey to hold the customers.
- In our dataset, we're looking to segment ride-sharing customers based on various segmentation ideas like hour, day of the week, common pick-up point, common destination point, cab types, and distances traveled. Let's see them one by one:

- **Time Segment**
- **Day segment**
- **Source Segment**
- **Destination Segement**
- **Cab Segment**
- **Distance Segment**

- This segmentation groups the distance from the source to the destination for each customer, it will help to decide short, mid, and long-distance riders.

- Short-distance riders can complete their rides witin few minutes while long distance riders will take a few hours to complete the journey.
- Companies can offer membership or a discount service to the customers who have a long-distance journey to hold the customers.

2. DATA UNDERSTANDING

The dataset contains 0 duplicate values. 46 numerical columns and 11 categorical columns. The complete dataset contains 693071 rows, in which 385663 contains information about Uber's personal data while 307408 provides information about lyft's data. 57 features are available in the dataset which give detailed information about the ride including the weather, time, start and end location. Let's see the difference through visualization.

We don't have appropriate product information in the dataset,it would have been the most essential feature for the segmentation but now because of inappropriate data we can drop this variable.

- **Dropping Unnecessary columns**

Since we have 57 columns in the dataset, Some columns are not necessary for our simple use case. The necessary features which are actually making an impact on our predictions we will keep these columns and rest we will drop.

3. EXPLORATORY DATA ANALYSIS:

After dropping a few columns, we have 11 columns in our dataset.

hour:
Represents the exact hour of the day when a ride was ordered. It is useful in distinguishing the periods where many/stable/ few people may require a ride and these include morning, evening, night and any other time of the day.

day:
The date of the month when the ride was arranged. It can also assist in finding out the actual day patterns which might be special date and thus form part of the demand.
month:

The month of the year the ride was requested Putting more emphasis on the best months of the year to request for a ride. Daily and weekly data can show oscillation in the ride demand throughout the week and even throughout the year due to the changing weather.

datetime:
A single timestamp for the ride request by combining the date and time for the ride request. The timing of this column can be stated with much accuracy though it can be split down to the hour, day and month dimensions for further analysis.

source:
It Indicates the start point of the ride usually in form of an integer. Relative drop off zones might show some area attractiveness and potentially group passengers by the need-based location.

destination:
The place where a ride is being discontinued and also represented by an integer. Popular destinations could be useful to help segment destinations of customer or to analyze patterns of riding.

cab_type:
Specifies whether the person called for an Uber or Lyft, for instance. This can help in understanding the customers own between service providers.

name:
The subclass of services within a cab service, usually denoted by an app or company name (Uber, Lyft). This extends specific customers' preference and ride categorization according to service classes or automobiles' categories.

price:
The time that the ride takes, usually expressed in hours and in exceptional circumstances accompanied by the distance of the ride, often in miles or kilometers. This can have a direct relationship with the time spent offering a ride or the amount of money that's to be charged and helpful in anticipating the demand per short distance riding as opposed to longer distance.

distance:
The distance of the ride, typically in miles or kilometers. This can directly impact ride duration and price and is useful in predicting demand for short vs. long-distance travel.

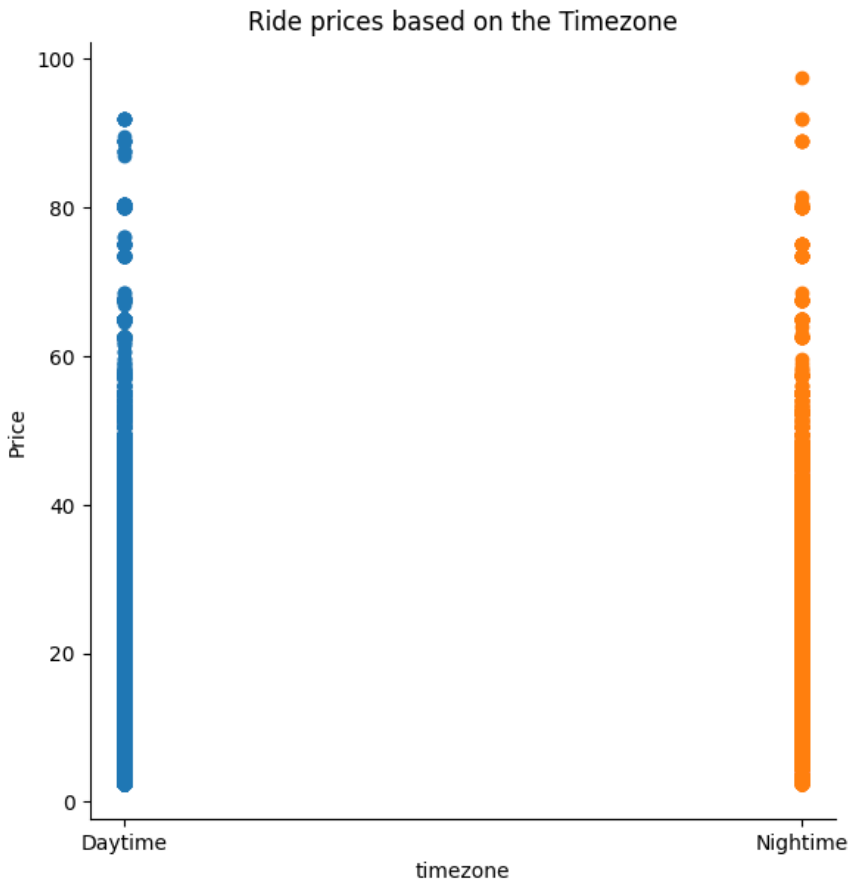
surge_multiplier:
An extra charge that is added when there are a lot of people using a given Uber car to travel to a particular location, which will help to push up the price of a ride. This earns the case surge pricing as a demand pressure worth using in identifying and forecasting changes in prices based on varying demand levels.

- **Checking and Handling Missing values**

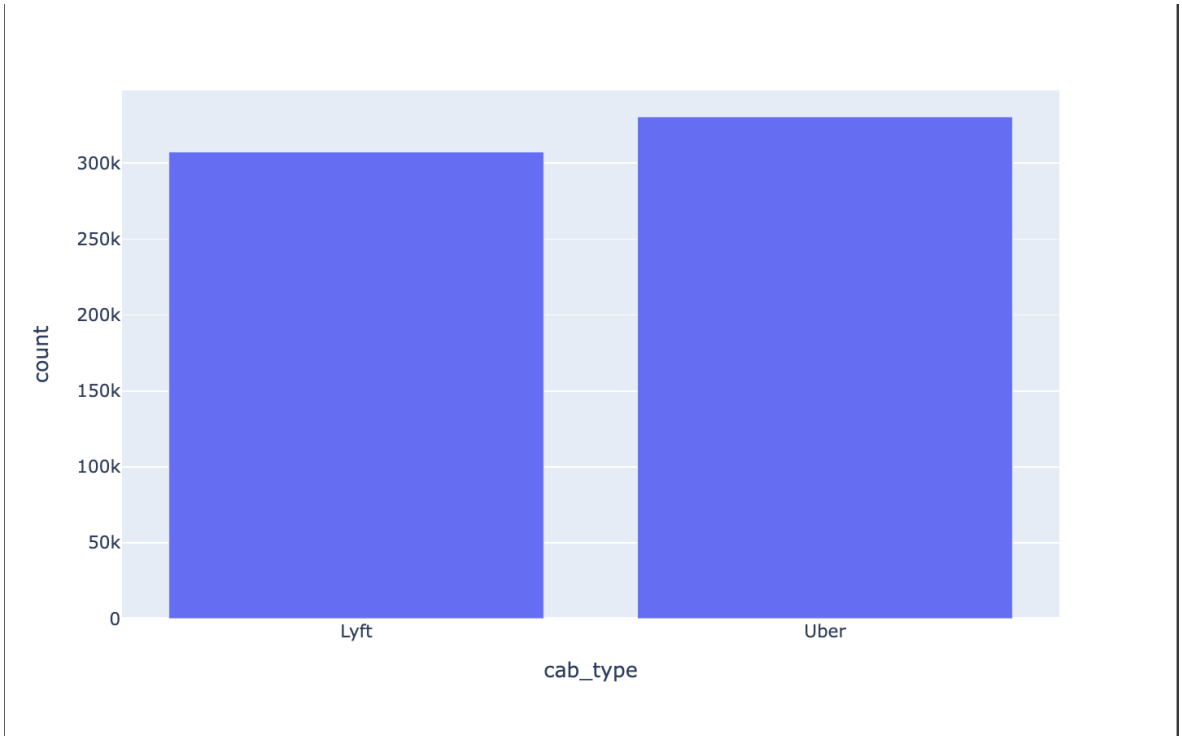
- Since we have 6,93,071 total rows and in that 55,095 rows contain missing values (around 9% of the dataset), Let's discuss how dropping missing value-rows won't impact our dataset:
- Even after dropping 55,095 rows we still will have suffiecient data for training Machine Learning models.
- Our dataset is based on real-time scenario which means rows are not connected with each other and dropping rows will not affect the segmentation.

DATA VISUALIZATION:

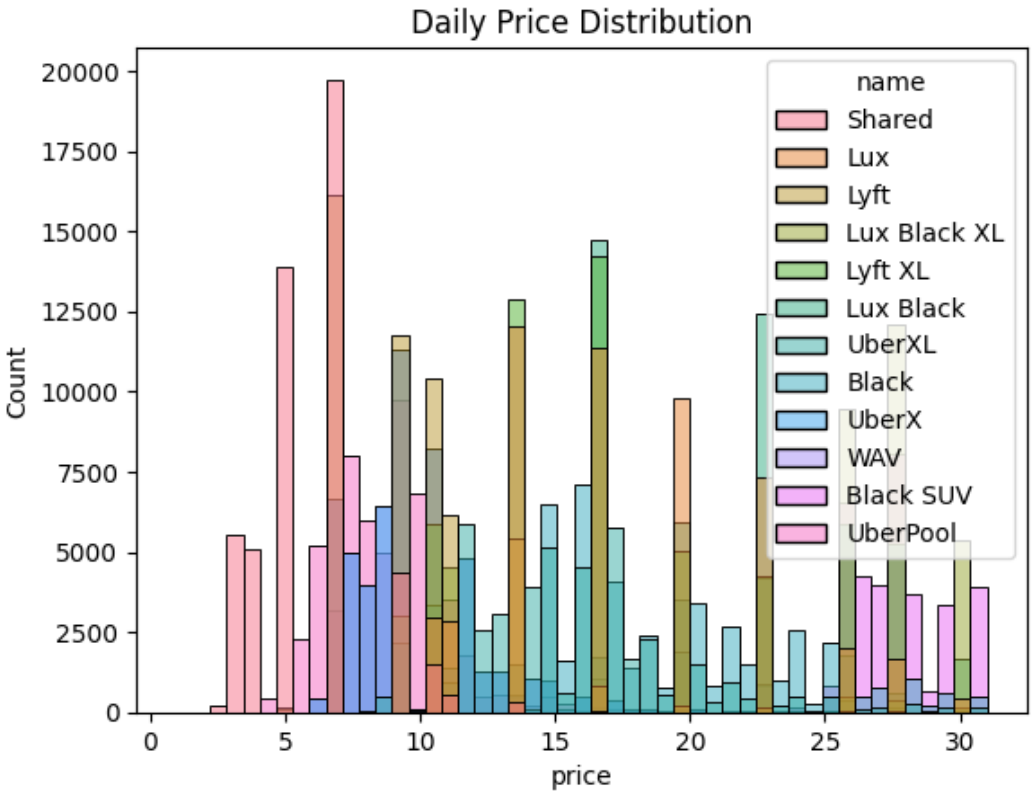
- **Ride prices based on the Timezone**



- **Counting the frequency of each cab type**



- **Price vs cab_type**



- **Average Ride Price Over Months**

CUSTOMER SEGMENTATION IDEAS:

Customer Segmentation means dividing customers with similar features into distinct groups so that we can provide services according to their requirements. In our Dataset, customer segments will help us to find the pattern of the dataset, for example, different kinds of rides, location, time, and many other features. Let's find what all are there in the dataset. Companies can offer a membership or a discount services to the customers who have a long distance journey to hold the customers. In our dataset, we're looking to segment ride-sharing customers based on various segmentation ideas like hour, day of the week, common pick-up point, common destination point, cab types, and distances traveled. Let's see them one by one:

1. **Time Segment:**

- It will fetch information about hour of a day, whether the customer is a Morning commuter or an Evening commuter
- It will help to merge the same group of people and increase the number of customers and we can supply offers on peak demand times.

2. **Day segment:**

- It will fetch information about a day of the week, whether the customer commuting on weekdays or weekdays for school or a work
- The goal is to group all the same group of people and increase the number of customers and we can supply offers on peak demand times.

3. **Source Segment:**

- It will identify common pick-up locations so that in the future we can do uberShare for people who live nearby locations while going to work and school.

4. **Destination Segement:**

- It will identify common drop-off locations so that in the future
- We can do uberShare for people who live nearby locations while coming back from the workplace or from the university.

5. **Cab Segment:**

- This segmentation is based on the cab type, just to see how many people likes to commute by lyft and uber.
- when we understand this preferences it will be easy for a particular compnies with the offers and discounts and it will help to add on more marketing strategies to gain more customers.

6. **Distance Segment:**

- This segmentation groups the distance from the source to the destination for each customer, it will help to decide short, mid, and long-distance riders.
- Short-distance riders can complete their rides witin few minutes while long distance riders will take a few hours to complete the journey.
- Companies can offer a membership or a discount services to the customers who have a long distance journey to hold the customers.

DATA MODELING:

- I wanted to check the results for different Machine Learning models, I tried Random Forest Regressor, Linear Regression, and XG Boost algorithms. We got good results with the Random Forest Regressor and XB Boost model.
- This is a clustering analysis. It is an unsupervised machine learning method that groups customers based on their similarities, ensuring that customers within the same group (or cluster) are more alike than those in other groups.
- Splitting data into train test split

MACHINE LEARNING MODELS:

- Random Forest Regressor:
- XGBoost:
- Linear Regression:

5. MODEL EVALUATION:

	MAE	MSE	R-squared:
Random Forest REgressor	7.09	76.66	0.12
XG Boost	7.09	76.66	0.12
Linear Regression	7.17	78.64	0.09

1. Mean Absolute Error (MAE):

- **Random Forest and XGBoost:** It is also to be noted that both the models have same MAE that is 7.09; which means that the average prophecy error is similar between two models. This implies that on average their predictions are approximately equal to 7.09 units away from the actual values.
- **Linear Regression:** The other evaluation metric was mean absolute error (MAE) which is slightly higher at 7.17, implying that it has relatively lower ability in predicting the performance of the models than LM and MBD.

2. Mean Squared Error (MSE):

- **Random Forest and XGBoost:** They also both have the same MSE of 76.66, which shows the amount of the squared error in the average of the prediction in both models were similar.
- **Linear Regression:** The MSE of the model is 78.64 Ky for the test data, and is again slightly larger compared with the other two models. This means that, the MSE of Linear Regression model is high due to large errors as compared to Random Forest and XGBoost models as MSE consider large errors par to small ones.

3. R-squared (R^2):

- **Random Forest and XGBoost:** As for the models, the models are compared by R-squared equal to 0.12 for both, which means that only 12% of the changes in the target variable are caused by the used features. This implies that the models can be said to have low reliability on the variance in the data that is being analyzed.
- **Linear Regression:** Linear Regression has the least amount of variance explained in the data the R^2 of 0.09 which is equivalent to 9% hence proving to be the weakest as an ensemble method.

6. SUMMARY AND RECOMMENDATIONS:

The patterns of ride behaviors detected in the Uber and Lyft dataset include the timing of commutation, the choice of the pickup/drop location, and the Uber/Lyft split. More specific market segments are morning and evening rush hour travelers, weekend travelers, short distance travelers and long distance travelers.

- **Targeted Marketing:** Create appealing campaigns depending on the dividing users into some groups; for instance, offering special offers for the weekend trippers or bonuses for early morning rush-hour travelers.

- **Service Optimization:** To reduce delay during peak hours and on crowded stations, there should be an optimization of the fleet circulation.

- **Dynamic Pricing:** Make sure pricing and revenue strategies use demand patterns to continue to maintain and maximize the income from customers.

- **Customer Loyalty Programs:** Implement a system to give incentives to regular commuters and customers in order to encourage them to stick around. They can help to enhance customer acquisition and business effectiveness with reference to the intensely competitive environment of the ride-sharing industry.