CASCADE CUP 2022 ROUND 3

DATA ANALYSIS REPORT

RIDER-DRIVEN CANCELLATION

Introduction

Shadowfax's business includes delivering food orders from clients to customers. Through this competition, we had to analyse the order data and predict when a rider is likely to cancel the order so that another rider can be assigned beforehand. In this data analysis report, we have tried to search for and establish patterns and trends that may be of interest to Shadowfax in this Rider-Driven Cancellation Problem.

Team CP_Karlo

Gunjan Dhanuka | Harsh Agrawal

Indian Institute of Technology Guwahati (IITG)

Data Preprocessing

Before we can start working with the data, it is essential to pre-process it and make it suitable for analysis.
We followed these steps for preprocessing:

1. **Parsing the Dates:** We needed to convert the dates and times to the DateTime object in Python suited from the object type so that we can use the functions to extract day and hour from these timestamps.

```
In [16]: def parse_dates(train):
             train['order_time'] = pd.to_datetime(train['order_time'], format = "%Y-%m-%d %X")
             train['order_date'] = pd.to_datetime(train['order_date'], format = "%Y-%m-%d %X")
             train['allot_time'] = pd.to_datetime(train['allot_time'], format = "%Y-%m-%d %X")
             train['accept_time'] = pd.to_datetime(train['accept_time'], format = "%Y-%m-%d %X")
             train['pickup_time'] = pd.to_datetime(train['pickup_time'], format = "%Y-%m-%d %X")
             train['delivered_time'] = pd.to_datetime(train['delivered_time'], format = "%Y-%m-%d %X")
             train['cancelled_time'] = pd.to_datetime(train['cancelled_time'], format = "%Y-%m-%d %X")

             return train
```

2. **Fixing the Null Values:** There were a lot of null values present in the DataFrame. The columns *accept_time, pickup_time, delivered_time, alloted_orders, delivered_orders, undelivered_orders, lifetime_order_count, reassignment columns, session_time* and *cancelled_time* have null values.
    a. **Filling with zeros:** The null values in columns *reassigned_order, delivered_orders, undelivered_orders, alloted_orders* were filled by 0 since there wasn't any other way to estimate the missing data and zero is logical in this scenario.
    b. **Filling string fields with 'None':** The columns *reassignment_method and reassignment_reason* had their NA values replaced by 'none' to store the information that their values were missing initially.
    c. **Filling with Median:** Some fields like *session time* had a wide range of values, possibly including outliers, so we resorted to using Median to fill the null values.
3. **Outlier Detection:** Columns like *session_time, total_distance* have a few exceptionally large values that are potential outliers but we could find no way to fix them in a meaningful way, or else it would spoil the real-world nature of the data. Hence, we let them be a part of data as it is.

Feature Engineering

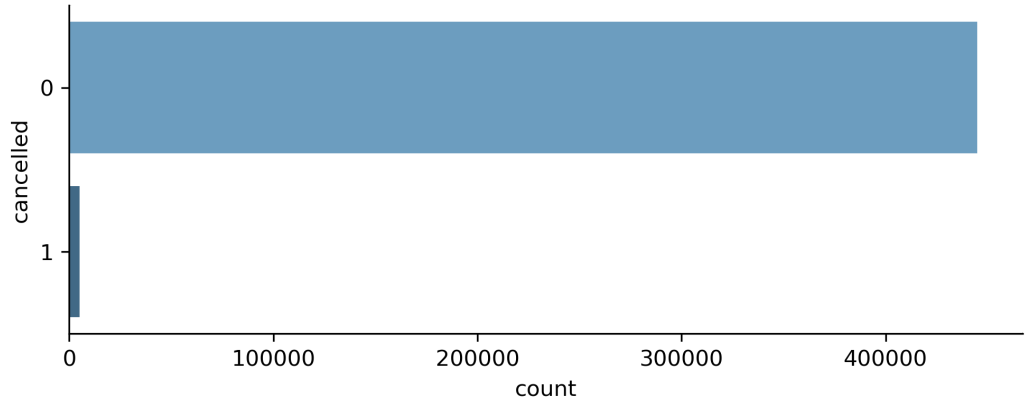
It is important to create new features which might come in handy for further analysis. Here are the features we created for better insights into the given data:

- **Total Distance:** Stores the sum of *first_mile_distance* and *last_mile_distance,* which can negatively impact the rider's will to deliver the order if the total distance is too large.
- **Weekday:** Stores the Day of the Week when the order was created. Can quantify trends based on the day of the week.
- **Order Time Hour:** This contains the hour when the order was placed by the client, helpful for seeing hourly trends in cancellation behaviour.
- **First Order:** Indicates if this order is the first order for that rider (if *lifetime_order_count* = 0), which can increase the chances of cancellation due to technical difficulties.
- **Cancel Before Accept:** If the *accept_time* was null for a particular entry, we have assumed that the particular order was cancelled before the rider accepted it. And hence this feature stores this trait.
- **Delivered Fraction:** It stores the ratio of *delivered_orders* to *allotted_orders* in the last 30 days. Hence somewhat of a measure for the reliability of the rider.
- **Accept Order Time Difference**: In cases where there is a difference in time between order creation and acceptance time of the rider, there can be situations like network issues, confusion regarding address, etc. To capture this difference and use it for our analysis, this feature is used!
- **Cancelled Hour:** To store the hour of the day when the order was cancelled (for cancelled orders only), to check if the cancellation behaviour is associated with the time of the day.
- **Total Calls:** For each order, using the *call_data.csv* file, we have stored the total number of calls that were made against that order. An exceptionally high number of calls can indicate that there were technical troubles that the rider was facing, or too few calls can indicate that the rider delivered the order smoothly.

Data Analysis

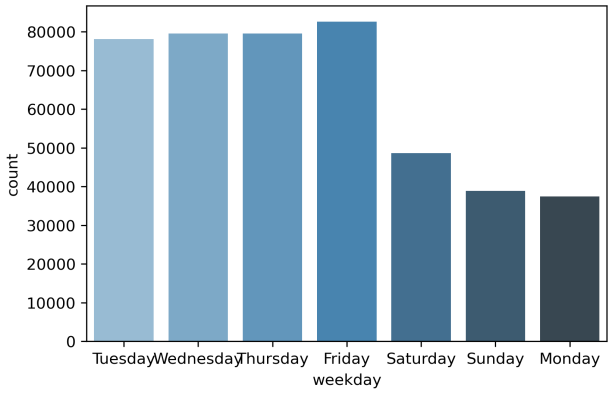
Basic Plots

- First of all, we observe that the data is **highly imbalanced**, containing very few cancelled entries (*~5000*) as compared to delivered entries (*~4,44,782*). Thus we must be very careful while making comparisons in any kind of trends.

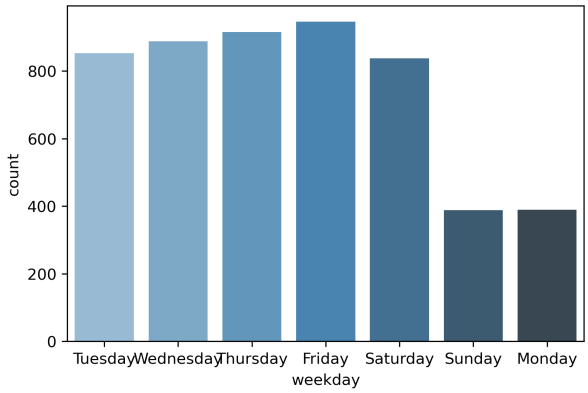


- Next, we try to compare the **trends of cancellations on weekdays**.

Delivered Orders

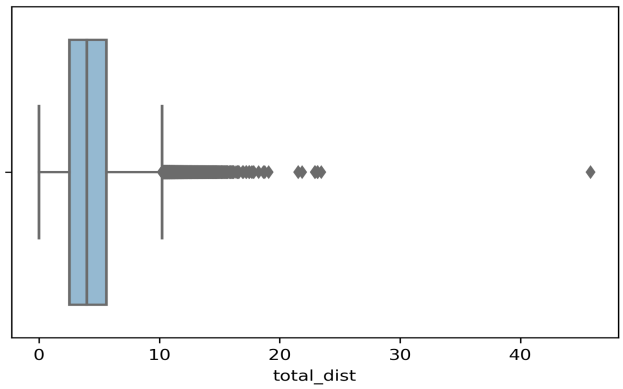


Cancelled Orders



However, we see that the distribution is almost similar for cancelled and delivered orders. Hence **the day of the week does not play a major role in the cancellation trait.**

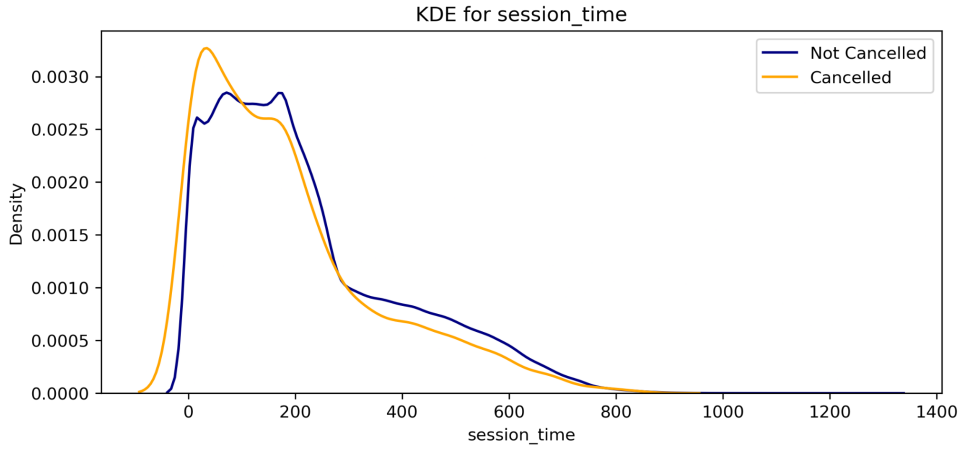
- In the Box Plot for the total distance to be covered by the rider, we see that the



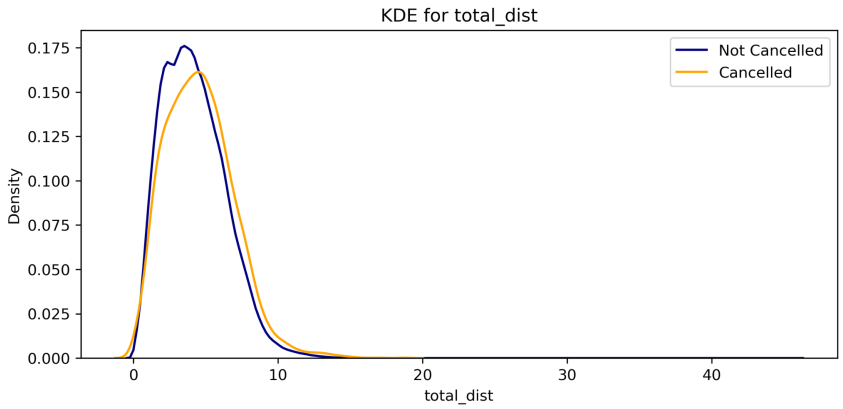
median lies at 4 miles. It shows that most ***orders demand ~4 miles of travel for the rider*.** However, there are also outliers, for eg. one point at 40+ miles which is very unlikely! This is most likely due to a discrepancy in data entry.

KDE Plot

It stands for Kernel Density Estimate and is a method to **visualize the distribution of observations in a dataset**. It indicates the *probability of seeing a point at a particular location*. We will be going through some KDE Plots now and try to explain the trends we see.

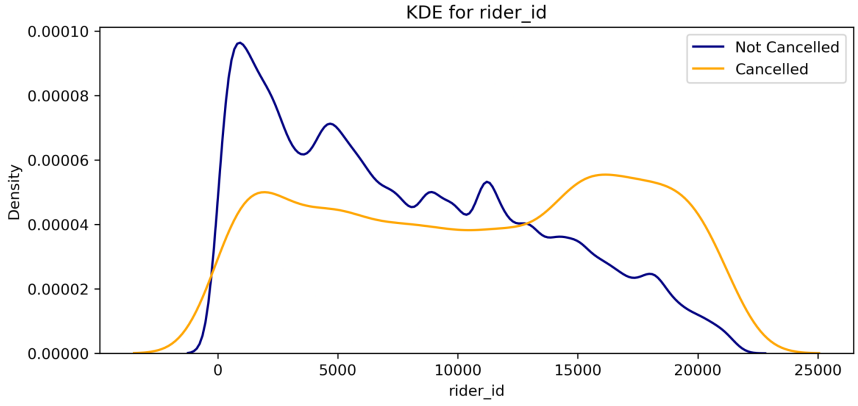


In this figure, we can see that there is a **high possibility of cancellation where the session time is close to zero**. Also, the possibility of cancellation as well as delivery drop sharply for a certain time (around 200), since the riders would be looking for a delivery that is more optimal to them, considering the distance and location. However, after some time of idling, they accept the allotted orders and hence we see the slow gradual decay in the probability. ***This shows that some riders can be picky especially at the start of the day.***

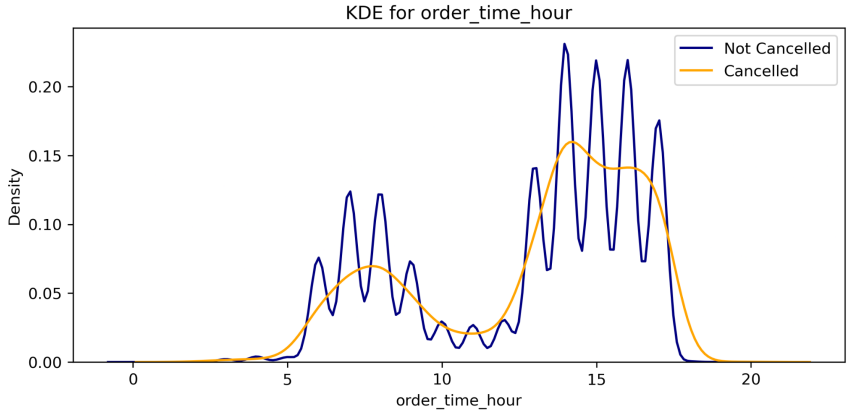


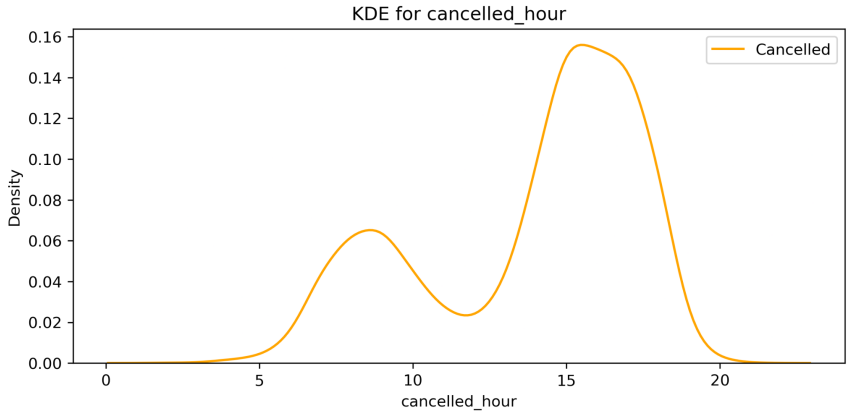
We see the distribution is similar for cancelled and not cancelled in total distance to be travelled by rider. Hence it is probable that **the distance is not always a major factor** in driving cancellation from the side of the rider.

This plot indicates a certain kind of reliability in the rider. **If the rider has delivered more orders in the month, he is more probable to deliver the order.** If a rider hasn't delivered any order in the previous 30 days, he is more likely to be inactive or busy, and thus unable to complete the allotted order deliveries.



An interesting observation is the rider id, *which may be allotted on the basis of when the rider registered with the company*. **We see that the newer riders (i.e. having larger rider id's) tend to cancel more, while the "experienced" riders, who might have registered early, tend to complete the deliveries.**

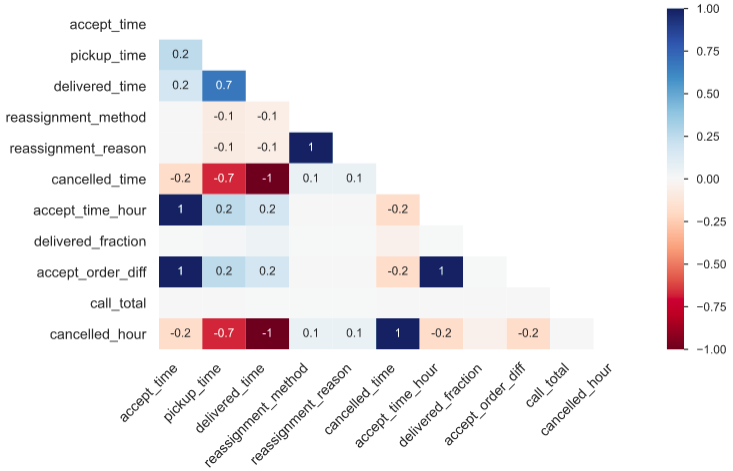




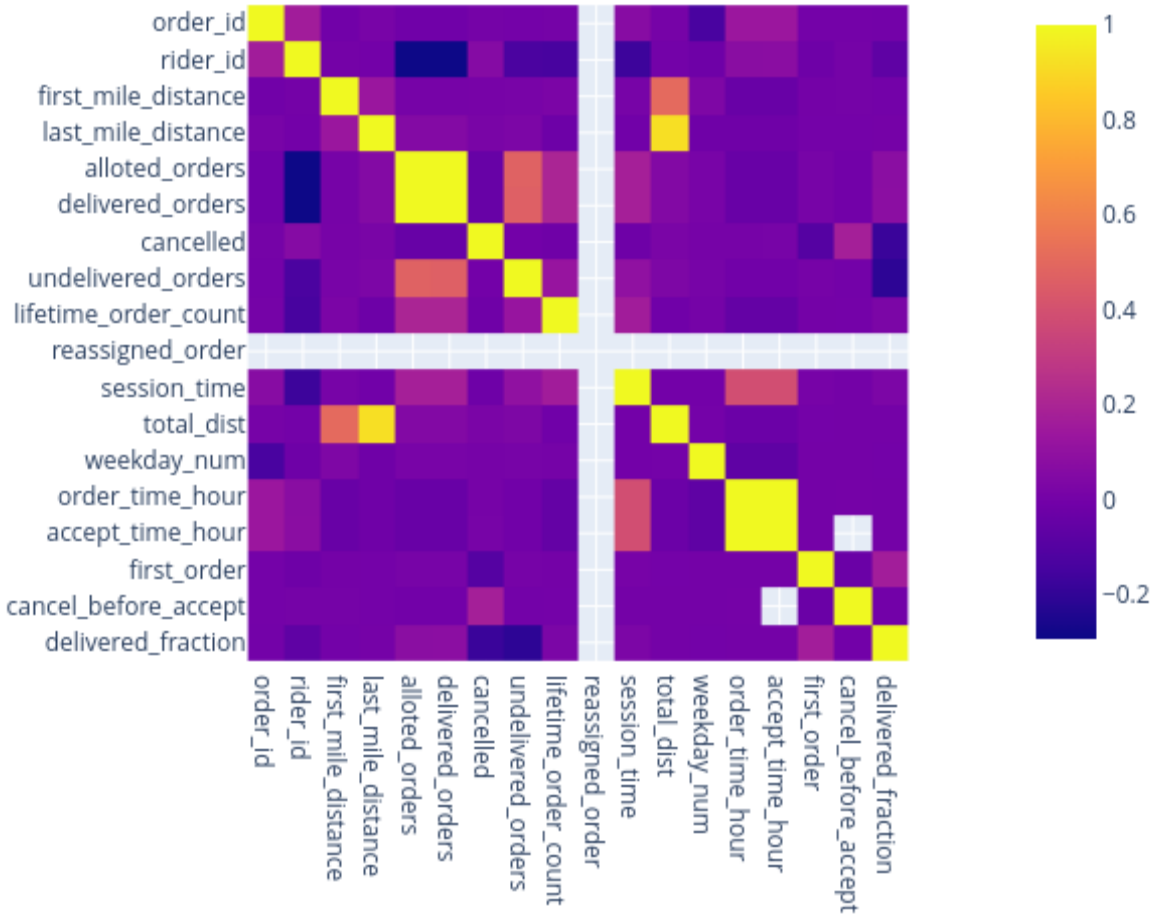
In the adjoining plots, we have tried to see the trends in the hour of the day and cancellation tendency. We hypothesised that late-night orders may be more likely to be cancelled, but **there was a lack of data of order data beyond 8 PM**. *Since the data is from around early 2021, it may be due to a night curfew and restrictions in many parts of India.*

Also, we observe that the peak probability of cancellation is during afternoon times, which is trivial since that is the time most of the orders were placed.

Correlation Matrix



The correlation heatmap measures nullity correlation: how strongly the presence or absence of one variable affects the presence of another. This can help us in judging which null values hurt our analysis the most.



From this correlation matrix, we can see the total linear correlation between two variables. *-1 indicating total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation.* **From the matrix, we can see that there are not many multivariate features present, and the correlation is quite absent** except the obvious ones like *first_mile_distance* and *total_dist.* This observation confirmed that the data will not show interesting results on scatter plots, or on dimensionality reduction. The bright yellow spots indicate a high positive correlation, while the dark spots show a high negative correlation.

Inferences

From the aforementioned analysis of the entire data, we have arrived at the following conclusions and comments:

- The data provided was highly imbalanced and also consisted of many missing values. Therefore, it is essential to preprocess this data before any further use.
- The data we received was recorded in early 2021 when there was a decrease in Covid cases in India, and hence the delivery services had just resumed. However there were restrictions in many parts of India, and lack of late-night data adds to it.
- There was little to no multivariate relationship between features, and hence it was difficult to come up with deeper trends in the underlying data.
- There are a number of outliers in features like distance, session time, calls, etc. This made it difficult to scale the data and hence, almost all of the plots look skewed and crowded near some common points and scarce at the extremities.
- The rider id was an interesting feature that we found could suggest cancellation. Since rider id is allotted according to the region or registration time, this can very well serve as an indicator for further analysis.
- During the early active hours, riders tend to be picky in accepting delivery orders, and as the day passes on, the reluctance decreases and the rate of cancellation starts falling.
- If the rider is newly registered or inexperienced in the past month, he is more probable to cancel the delivery due to technical difficulties or navigation problems since he may be new to the city.
- We might have gotten more insights if the data was from around the year, so we can capture seasonal trends like strikes, diseases, monsoons and movement of traffic.

From extensive data analysis, we have come to the conclusion that rider-driven cancellation doesn't depend largely on one or two selected factors, but on the overall set of features by a small per cent.
Also, cancellation is an act of human decision-making that cannot be captured solely with data, since the situation and condition can be different for the same rider at different times.

THANK YOU.