Data Analysis Report

by

Team Neural Demons

Tanish

Vaibhav Tripathi

1. **Key questions and Objectives**

The key objective of this report is to analyze the given dataset and generate insights to figure out correlations between variables in the dataset. The analysis will be largely focused on order cancellations and factors associated with it. Our analysis is generalized over orders and riders, making it applicable to any new order or rider outside of dataset.

The key questions that this report will answer are:

- What properties of orders and riders indicate cancellation and upto what extent?
- How are variables that affect cancellation related with each other?

1. **Feature Engineering**

We created some new_features to enhance the analysis:

- “allot_order_delta”: difference between “allot_time” and “order_time”.
- “accept_allot_delta”: difference between “accept_time” and “allot_time”.
- “pickup_accept_delta”: difference between “pickup_time” and “accept_time”.
- “total_distance”: sum of first mile and last mile distances.
- “delivered_allot_ratio”: ratio of “delivered_orders” and “alloted_orders”.

1. **Methodology**

As the analysis is focused on cancellations, it makes sense to study how many cancellations are made subject to certain conditions and compare them with how many deliveries are made under same conditions. We define a variable named “cancellation_rate” defined as the probability that a randomly sampled order from the dataset under some condition S on variable X was cancelled :

$$cancellation\,rate \;=\; P(cancelled \mid X \in \; S\,)$$

For e.g, if we wanted to quantify cancellations when first_mile_distance is between 1 and 2 units, we would calculate cancellation rate as:

$$P(cancelled \;|1 < \; first\,mile\,distance \; < 2) \;=\; \frac{cancelled\,orders\,with\,(1 < first\,mile\,distance \; < 2)}{total\,orders\,with\,(1 < first\,mile\,distance \; < 2)}$$

We use this methodology throughout the report to analyze how the variables in dataset affect cancellation. It is very important to note that cancellation rate is very sensitive to the denominator, say there is a certain condition on a variable that is satisfied by only one sample in the dataset and if that one sample turns out to be a cancelled order then cancellation rate is simply 100% which is absurd. So for an inference based on cancellation rate to be statistically significant, total number of samples satisfying the given condition must be sufficiently large, we chose this minimum limit as 100. The report only contains statistically significant inferences.

3. Order Analysis

In this section we analyze the order specific variables and their relation with cancellation. Below we have a figure for a quick overview of the distributions of order specific features.

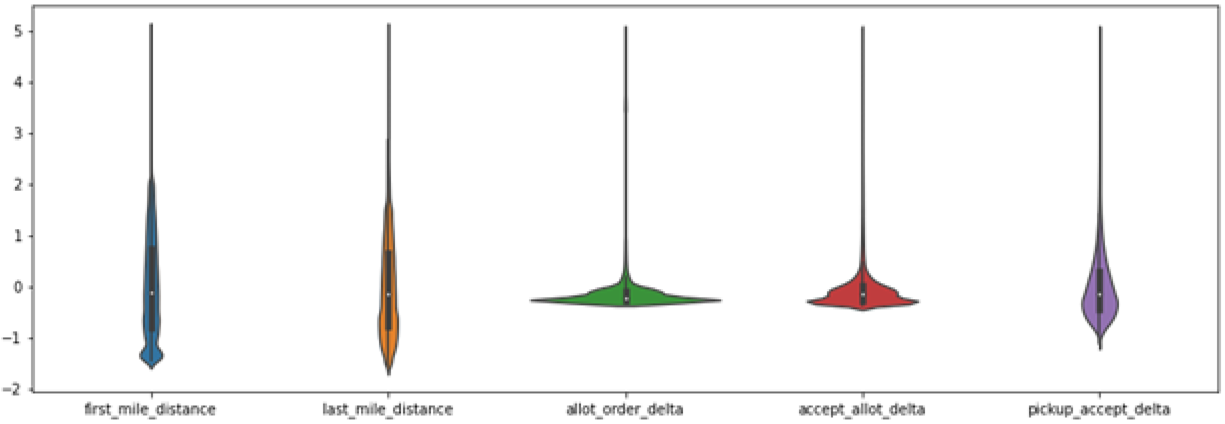
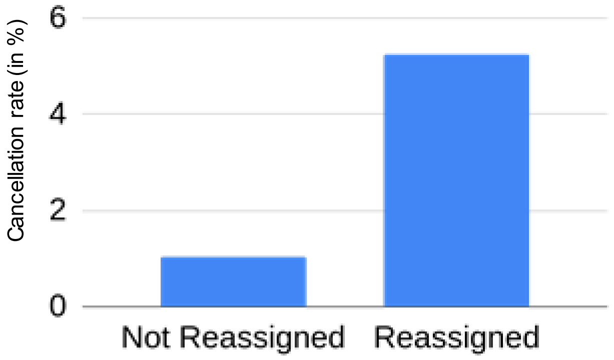
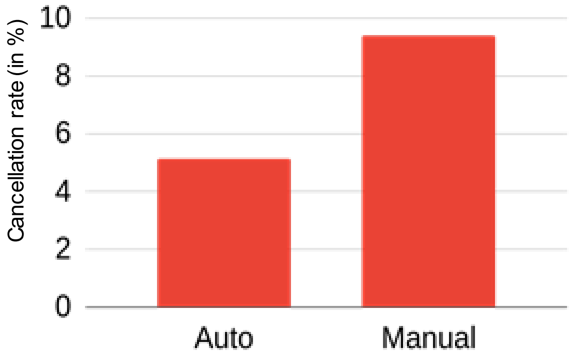


Fig. A violin plot for the numerical features associated with order only, normalized and removed values with z-score more than 5 to neutralize outliers.

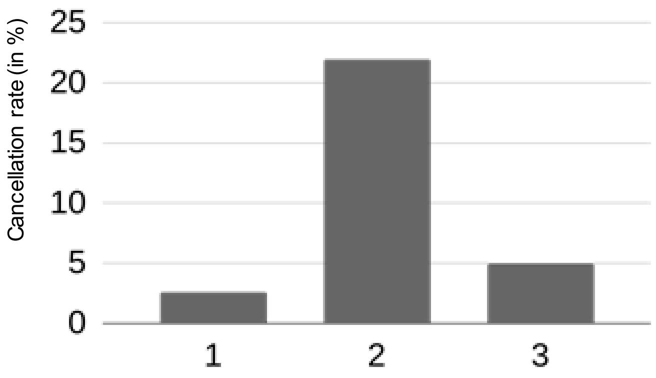
- **Effect of Reassignment:**



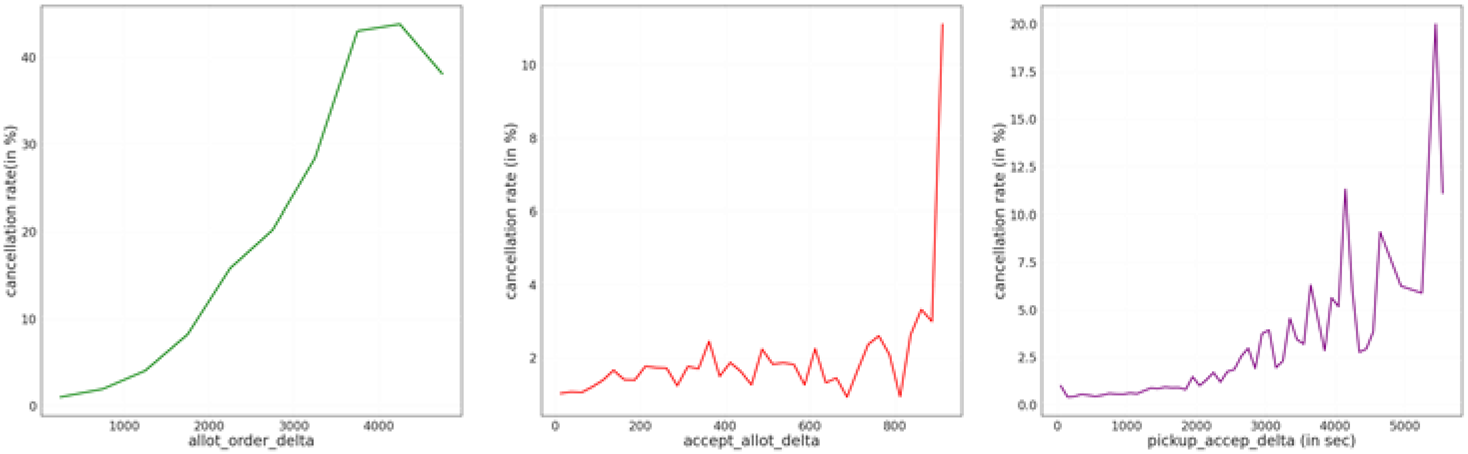


Only 3.05% of samples in the dataset are reassigned orders but the effect of reassignment on cancellation is noticeable. The cancellation rate for reassigned orders (5.24%) is higher than that of orders that are not reassigned (1%). Further, we observe that the cancellation rate for manual reassigned orders (9.41%) are higher than that of orders that are reassigned automatically (5.11%). So reassignment and the method of it can serve as strong indicators of cancellations.

In the bar graph on right, reason 1 is “auto reassigned basis inaction” , reason 2 is “Reassign” and reason 3 is “Reassignment Request from SE portal”. It is clear that cancellation rate for reason 2 goes as high as 22% which is much higher than same for reason 2 than 1 and 3.



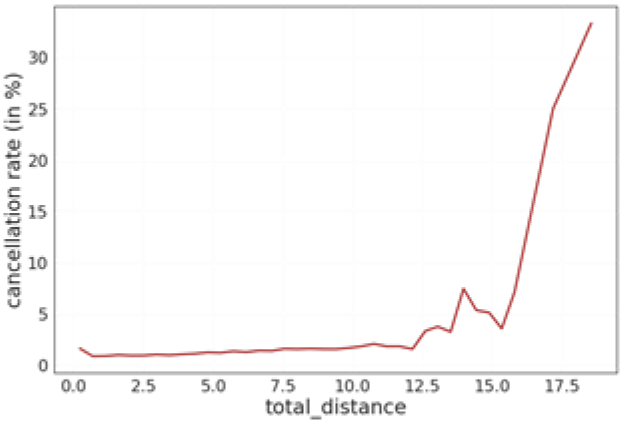
- **Effect of disruptions in order flow:**



In the figures above we observe that if any component of order flow takes unusually long time to complete, then cancellation rates jump suddenly. In the first from left figure we see cancellation rate increases almost linearly with allot_order_delta, going as high as 43.8%. In the second figure we observe a sudden jump after accept_allot_delta crosses 800 seconds. In the third figure we see cancellation rates increasing gradually as it takes longer time for item to get picked, infact 46.39% of cancelled orders are cancelled before getting picked up. So unusually long time difference in any step of order flow indicate that the order is likely to get cancelled.

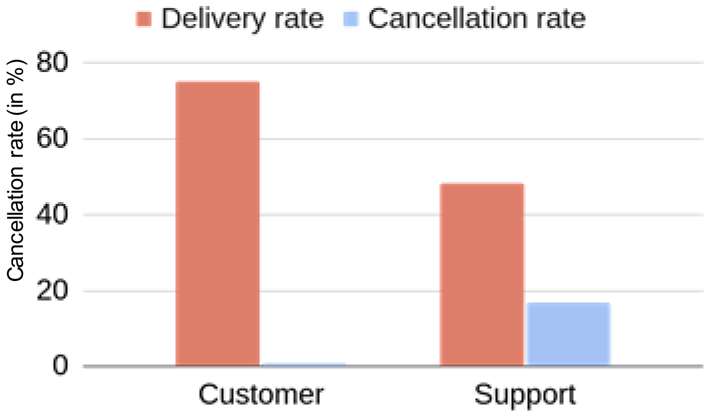
- **Effect of travel distance:**

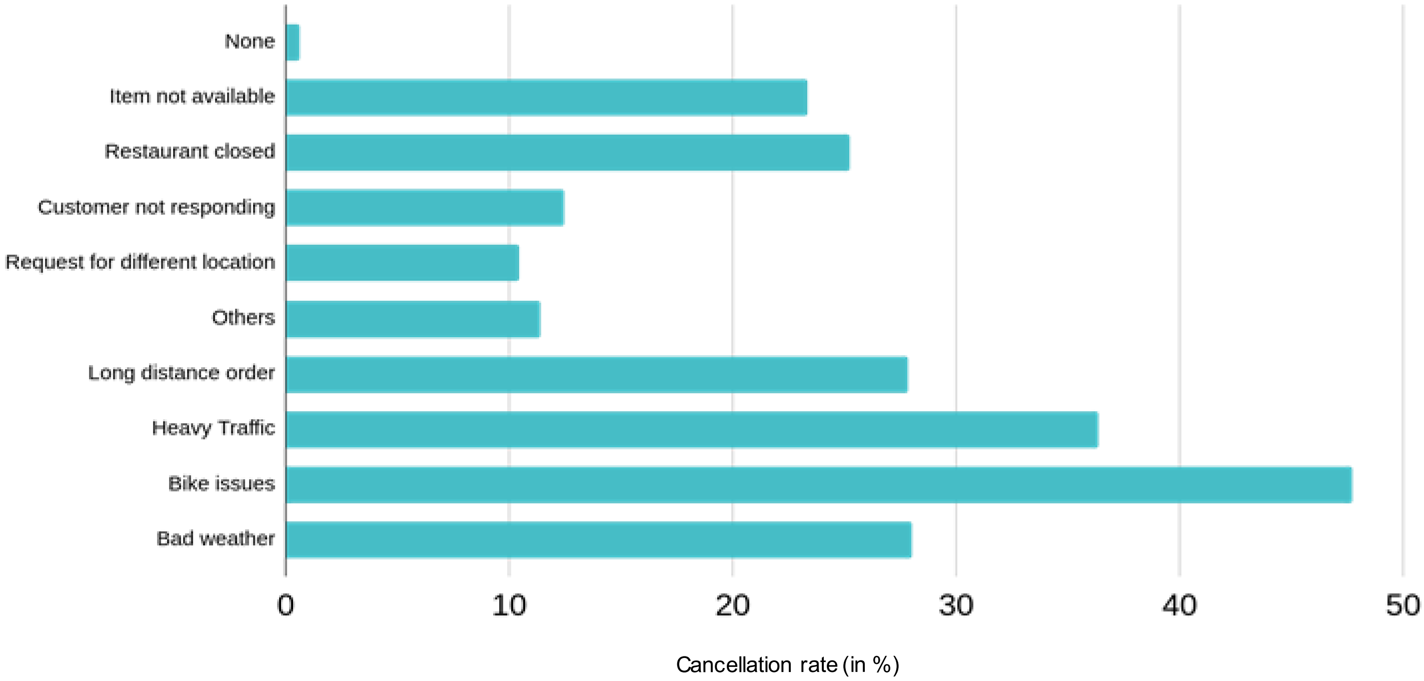
As total distance of the order increases, cancellation rate also increases and we observe a sudden jump after total distance increases 15 units, rate going as high as 30%. First mile and last mile distances have the same relationship with cancellation rate.



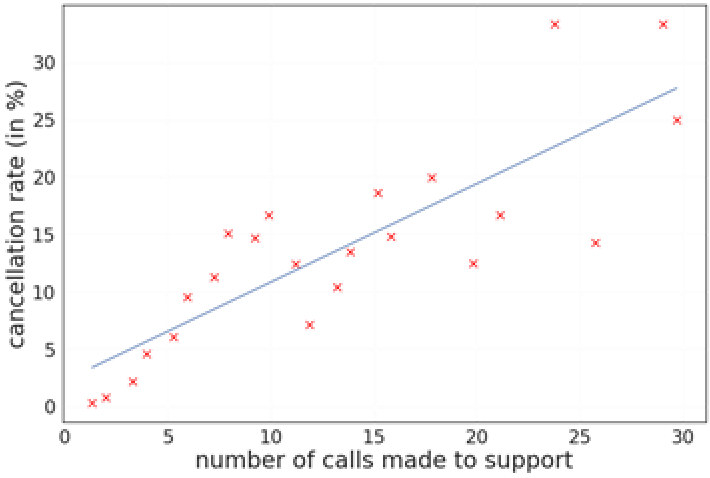
- **Effect of calls:**

Calls made to the support centre or customers have considerable insights about cancellation. Cancellation rate for orders where calls were made to customer is a feeble 0.6% (hardly noticeable in the figure) but the same for orders where calls were made to support centre is 16%.





In the bar graph above we have cancellation rates (in %) for different reasons given for calls to support centre ("Others" is for calls made to customer). It is observable that "Bike issues" and "Heavy Traffic" have highest cancellation rates at around 47% and 36%. So we can infer how likely an order is to be cancelled given the reason for call to support.



In the figure on the right we fit a linear regression line on cancellation rate vs. number of calls made to support centre for an order. The coefficient of regression line is 0.85 and intercept is 2.62, this implies that for every extra call made to support centre, chances of cancellation for that order increase by around 0.85%.

Besides all the insights mentioned above, there are some relationships that we expected due to rationale, but observed otherwise. Some of these are as follows:

1. The weekdays have no effect on the cancellation rate, the value is around 1% throughout the week.
2. The hour at which order is placed also has surprisingly no effect on the cancellation, the rates might seem high for 2AM, 7PM and 8PM but it is only because total samples for these hours are very few.

4. Rider Analysis:

In this section we analyze the rider specific variables and their relation with cancellation. We have a figure below for a quick overview of rider related variables and their distributions.

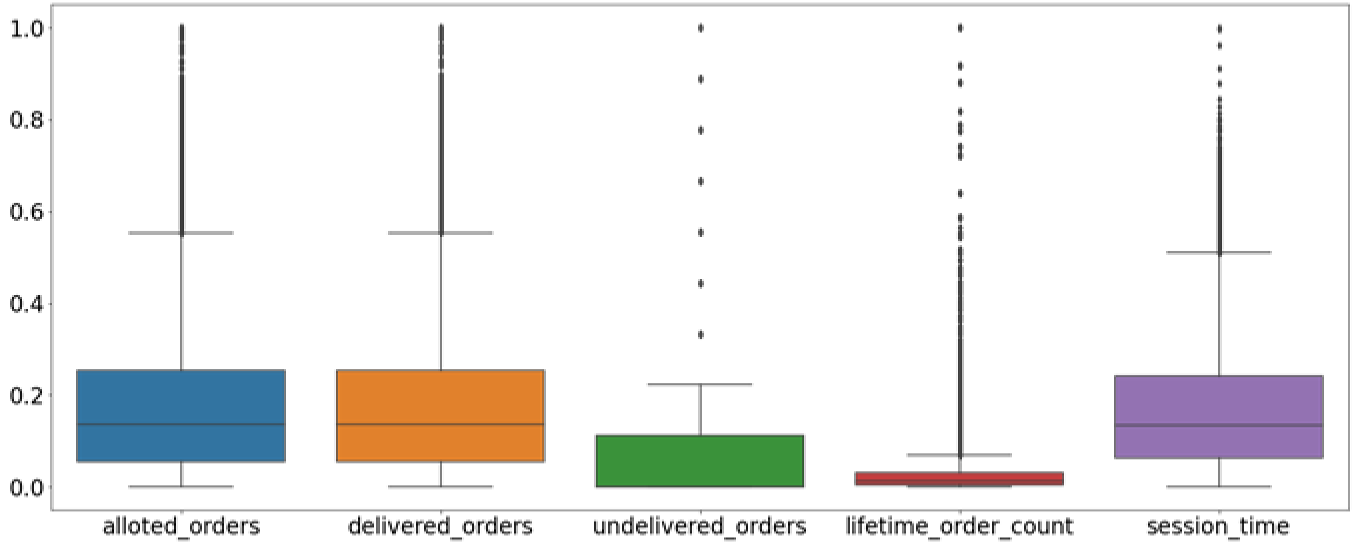
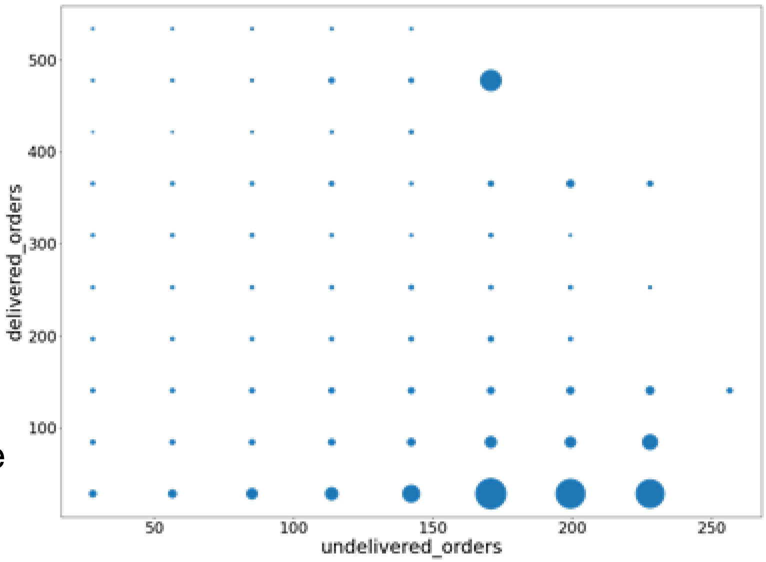
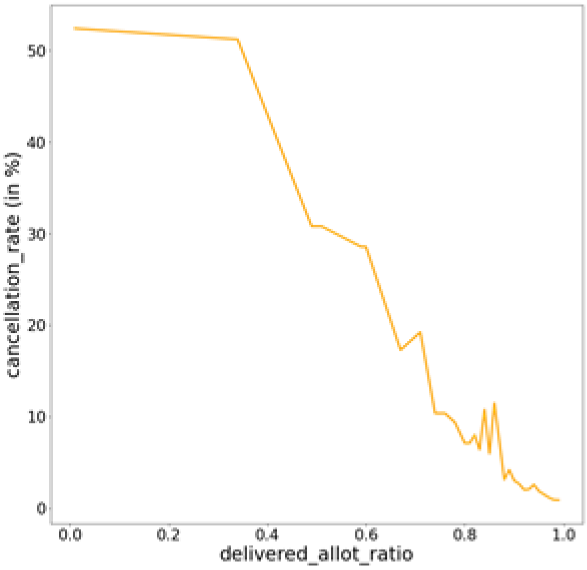


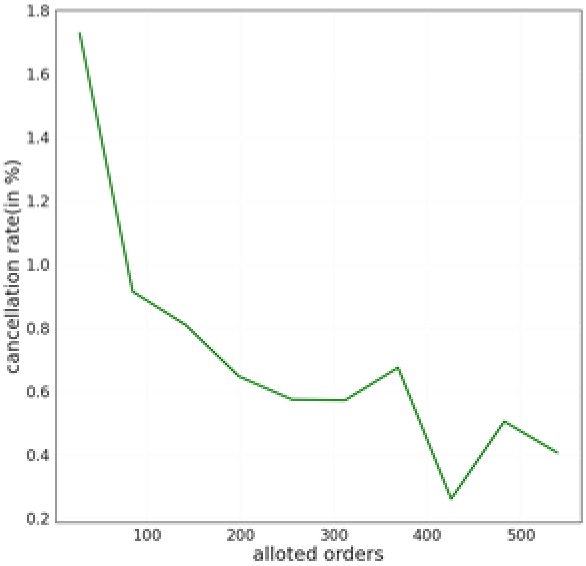
Fig. A box plot for the numerical features associated with rider only, min-max scaled for visual convenience.

- **Effect of past delivered orders:**

On the right we have a plot between delivered and undelivered orders, the size of bubbles are proportional to the cancellation rate. We observe as we move towards bottom and right, the bubbles become bigger which implies that riders with more undelivered orders and less delivered orders have higher cancellation rates, as high as 20% in the bottom right corner.

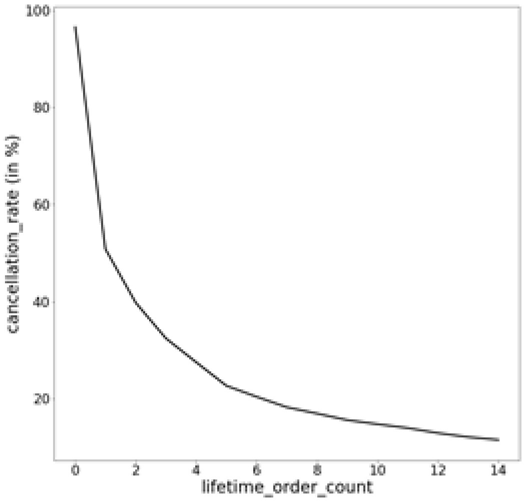






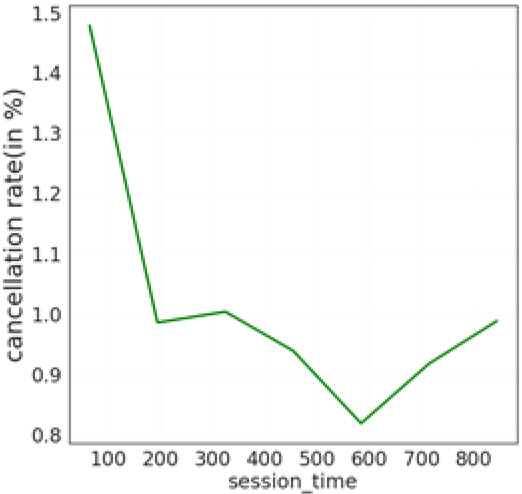
It is observable in the figure on the left that cancellation rates drop drastically as delivered_allot_ratio increases, infact for every 0.2 unit increment in delivered_allot_ratio the cancellation rate dips by 10 % Also, figure on the right shows that if the rider has been allotted more orders their cancellation rate goes down exponentially. This is expected as new riders are in learning phase and are not acclimatized to the job, but experienced riders tend to have low cancellation rates.

Expectedly ,cancellations are very common for new riders, chart on the right shows cancellation rates for riders with low lifetime_order_count. It is worth mentioning that for riders with zero lifetime orders, 56 out of 58 orders were cancelled. Also, for every order delivered by a new rider (lifetime_order_count < 15) the cancellation rate dips by 6% on average.



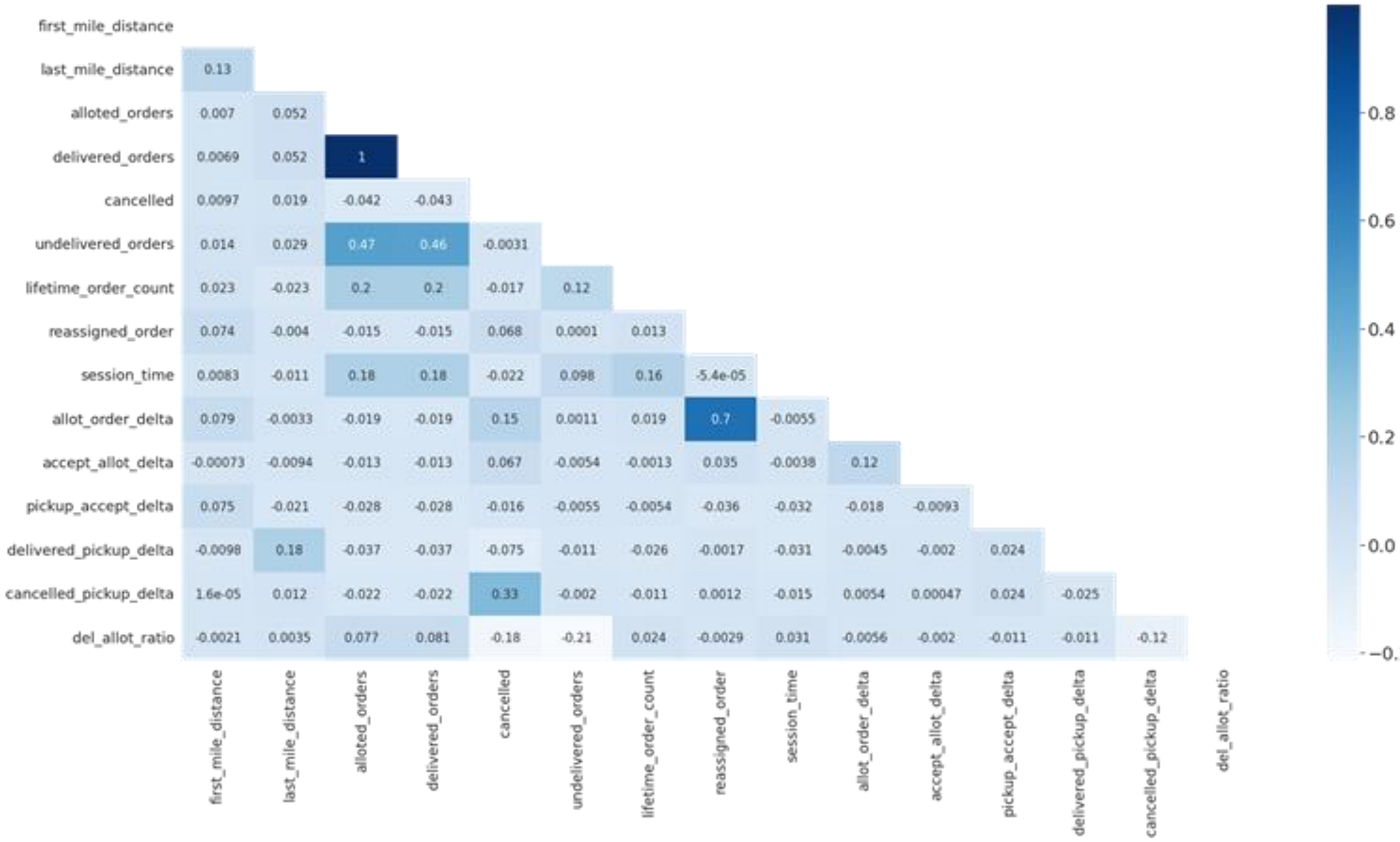
- **Effect of session time:**

Session time graph displays how if a rider is online for a long period of time on that particular day, their cancellation rate decreases. This shows that if a rider is not online infers he is not able to react or confirm the order and can show their disinterest in their job. But remaining online for a longer period of time like more than 600 minutes can have a negative effect on cancellation rate



3. Relationship between variables

In this section we analyze the relationships between variables affecting cancellation.

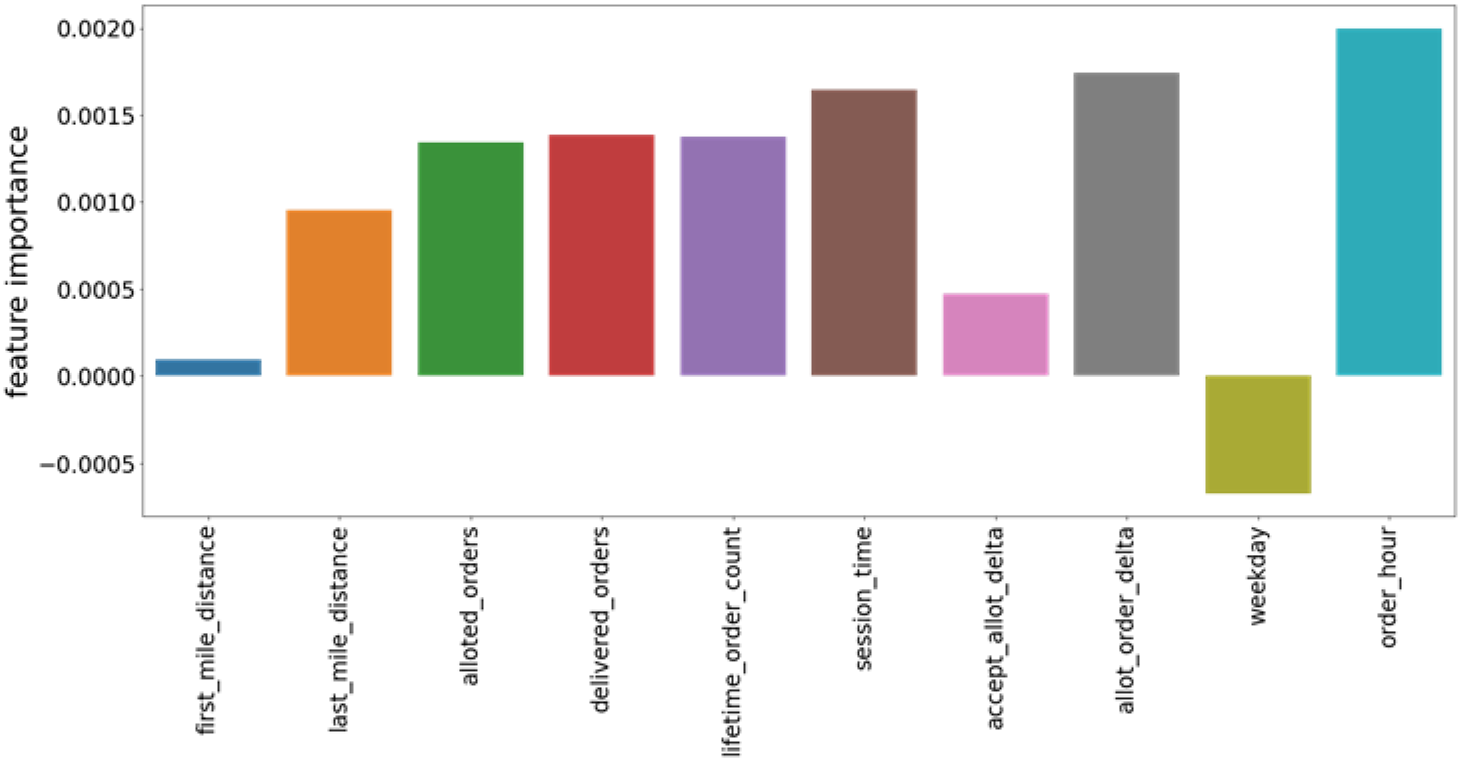


Above, we have a heatmap of pearson’s correlation coefficient between all pairs of variables.

- Most of the map is light blue indicating low linear correlation between most variables.
- There is an unusually strong correlation between “reassign_order” and “allot_order_delta”.
- Correlation between delivered, undelivered and allotted orders is high due to obvious reasons.
- We can observe moderate amount of correlation between “lifetime_order_count” and “alloted_orders”, “delivered_pickup_delta” and “last_mile_distance”, “cancelled_pickup_delta” and “cancelled”.
- It is important to note that person’s correlation only captures linear relationships and some correlations are low in the heatmap despite a strong relationship.

4. Conclusion:

The probability of cancellation of an order beforehand can be estimated to a reasonably good accuracy considering the relationship between given variables. We trained a stacked classifier containing XGBClassifier, RandomForestClassifier, DecisionTreeClassifier, AdaboostClassifier, GaussianNB and finally a Logistic Regression as final estimator. Our model was able to achieve an accuracy of 83% on test data. In the bar graph below we have plotted the feature importances of our model computed using permutation importance method.



These feature importances agree very well with our analysis and the insights mentioned in the sections above. These insights and figures could be used to predict cancellations beforehand and optimize the overall delivery process.

THANK YOU

by

Team Neural Demons

Tanish

Vaibhav Tripathi