# Bold Data Analyst Assignment

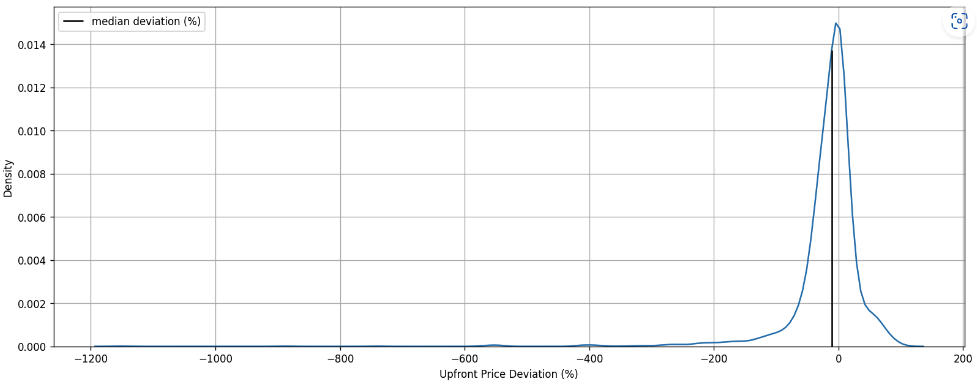
## Introduction

With an effort to provide customers with better transparency of prices. Riding-hailing apps such as Bolt have resorted to providing upfront pricing to customers prior to them booking rides. However, over a few months in early 2020, Bolt’s pricing algorithm tended towards predicting lower upfront prices, which leads to customers paying a higher price at the end of the ride – this ruins the customer experience. This assignment helps identify those cases and attempts to uncover their root causes.

## Impact

**What does the deviation look like?**

After plotting the deviations between the *upfront price* and the *metered price*, i.e. (upfront\_price – metered\_price)/upfront\_price, we get a distribution as such:



From the above, it’s clear that most cases experience higher metered prices as the distribution is left-skewed and long-tailed. Furthermore, this isn’t an experience you want as a customer – planning your journey only having to pay over 20% more.

**How many customers does this deviation impact?**

Assuming the dataset is independent and represents the population of the 4270, **70% see upfront pricing**. The others could result from pre-booking or using other means of transportation that don’t necessarily display prices at the start of the journey.

Of the 70%, around **60%** of rides fall into a higher metered price than the upfront price, and **35%** of rides get charged more at the end of the journey[[1]](#footnote-1). Of the 35%, 4% went to the extent of complaining that they overpaid at the end of the journey.

Therefore, this problem is prevalent in the Bolt ecosystem and needs to be addressed.

## Identifying the source of the deviation

During upfront pricing, a change in the following factors could cause our services to update the metered pricing:

1. Geographical factors:
   1. Updating the destination – this causes the distance to update and requires an update in the predicted price
   2. New routes – if drivers take longer routes (deviating from the predicted one), customers are charged more.
   3. Tolls – unaccounted toll gates
   4. Inaccurate GPS – the final/source destination could be captured incorrectly
2. Traffic:
   1. Wait time due to incoming traffic
   2. Slower speeds and hence higher ride durations
3. Surge:
   1. Poor weather (such as rain) causes vehicles to slow down
   2. Increase demand or change in time of day – which causes pricing to start using a surge multiplier

Where do we observe the maximum deviation in values with our dataset, and by how much?

**Note:** Mean/Median Abs deviation (%) = Mean/Median of abs((upfront\_price – metered\_price) \* 100/upfront\_price)

### GPS confidence

A picture containing table

Description automatically generated

There is an increase in the mean and median absolute deviation (%), as the GPS confidence is poor. **Could this be due to a particular manufacturer?**

Table

Description automatically generated

Manufacturers such as **Tecno, Infinix, and Itel constantly provide poor GPS confidence**. We could:

* Investigate if the app is running properly on these devices, e.g., using the GPS API correctly on Android.
* Warn customers about inaccurate GPS on these devices.
* Predict a certain percentage higher (explored in the next section)

### Device Manufacturer

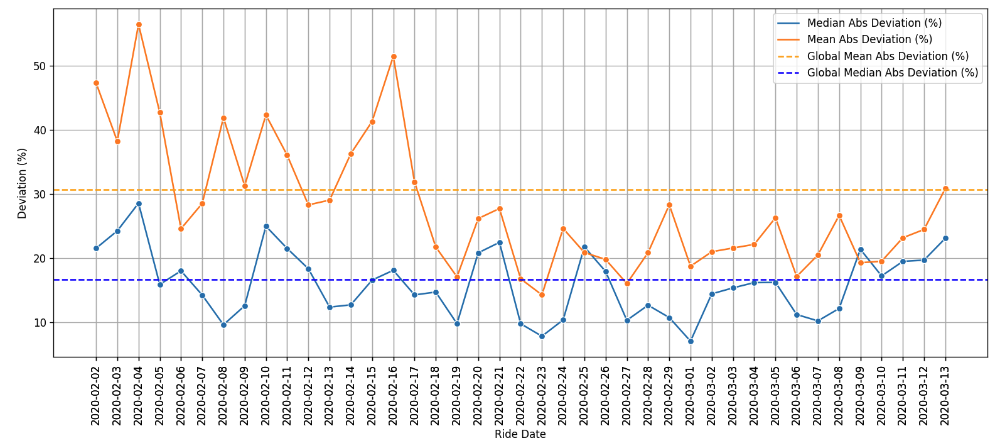
Table

Description automatically generated

**Tecno, Infinix, and Itel are repeat offenders** with up to 20 pp increase in the median deviation for Itel. We can circumvent this by predicting a 5-10 pp higher upfront prices for these devices in our upfront pricing model.

### Date

**Is there any particular day (such as holidays) when prices deviate more?**



The mean deviation peaks more than usual on the 4th and 16th of February. However, the median deviation on the 16th remains around the same upon closer inspection of the coefficient of variance (given below). The 16th was disproportionately affected by outliers. The date alone doesn’t look like a factor that’s impacting prices. Furthermore, there were no global holidays on these days–the data might need to be bifurcated further into countries to understand.

Chart, line chart

Description automatically generated

### Destination Changes

**Are more destination changes leading to higher prices?**

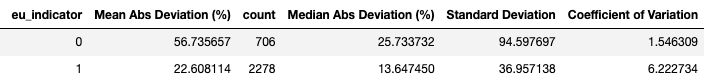
Table

Description automatically generated

* There is an increase of 2 pp in median deviation with one destination change and an additional 18 pp with three destination changes.
* However, it's difficult to confidently say that this trend will continue as the number of ride changes increases due to the unavailability of sufficient data.
* The increase should lead us to investigate how much the driver deviated from the route after being assigned the new destination, which led to higher pricing.

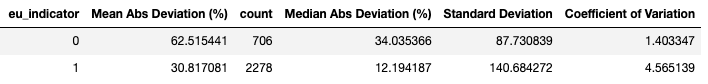
### EU Indicator

**How non-EU countries comparing to EU countries in terms of deviation?**



* There is a higher price deviation in both mean and median in non-EU member countries.
* This could be because the road infrastructure is better developed in the EU. Frequent road closures and deviations in distances cause prices to increase.
* It could also be due to regulations that do not allow price deviations in the EU.

Looking at the deviation in distances below:



It's evident that distances deviate more likely in non-EU countries than in EU ones.

## Additional Insights

### App Version

**Is there a particular rider app version that’s buggy?**

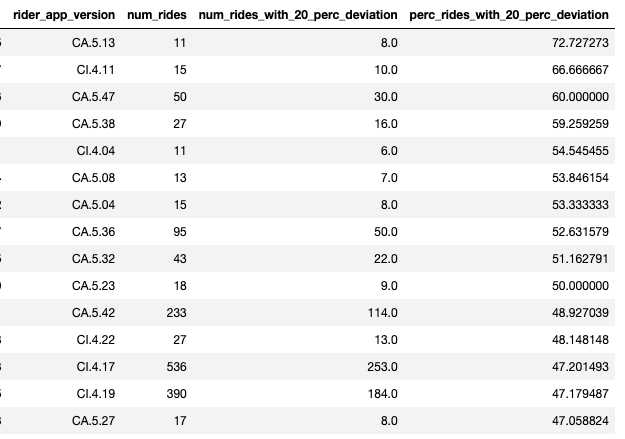
Although pricing is calculated on the backend, is something inherently wrong with how a particular app version extracts location or duration information?

Table

Description automatically generated

Some app versions, such as CA.5.47, typically perform poorly compared to others. However, it’s also important to note that the app version is a function of adoption–this means the problem is only apparent as most customers are on this version.

We can normalize this data and see the percentage of rides on app versions with over 20% deviation.



We notice a similar trend here, with versions such as CA.5.47 being repeat offenders that might require additional investigation.

### Why are there 0 distances and 0 duration?

There are roughly 35 rides, or 1% of the ride population, with 0 distances throughout the ride. These could be a result of the following:

* GPS was malfunctioning, and the actual distance didn't get recorded.
* The driver took the ride off of the app.

In the second point, we also notice durations that are 0. Roughly 50% of the 0 distance cases. These drivers could have reached the pick-up point and immediately canceled the ride. In India, Uber and Ola drivers cancel the ride and take the commission[[2]](#footnote-2); customers are indifferent as they’re usually charged the same amount.

1. Assuming at least a 20% increase in deviation. [↑](#footnote-ref-1)
2. <https://entrackr.com/2021/12/why-do-uber-and-ola-drivers-ask-users-to-cancel-rides/> [↑](#footnote-ref-2)