Rail Riddle By Amit jakhar

My work is focused on studying rail cancellation patterns and findings.

Major Predictions and Approach

Approach will be explained in later sections. Following are some insights that I gained:

- Some values in **Flow_Name** can be clustered efficiently to help studying cancellations that happened because of **low stocks, no locos and slow offloading.**
- Some values can be clustered from possible **Flow_Destination**, to gain insights on cancellations due to **no locos.**
- Some **NWB export accounts** are linked to cancellation due to **overhead cable damage.**

Key Features or Metrics

Predictions I made are done by using a heatmap generated for the normalized standard deviation of the percentage of contribution of each cancellation reason for every feature. **Key idea** is that ideally, every cancellation reason should be temporary, or should be seen rarely in only a few instances. Calculating the standard deviation of the normalized count of use of different cancellation reasons detects the reality of the aforementioned points. If the deviation is high, that means there are many instances deviating away from the mean value, i.e., the problem is more prominent or the problem is long enough to be persistent and affect in standard deviation.

Methods and Approach

Per feature, I grouped the date with respect to **CX_Reason** and elements that exist in the feature. I counted the total occurrence of a reason corresponding to that value. Corresponding to every value, I calculated the percentage use of each reason in the cancellations. Different elements for a feature will use a reason for different percentage amounts of time. I calculated the standard deviation of the contribution of a reason for a feature. As the percentage had a different order of magnitude for different elements, it needed to be normalized, for that, I tried dividing the standard deviation by the mean of the contributions for different elements of the feature, to normalize the order of magnitude. This process was done over almost every feature. Finally, I made a data frame containing standard deviation in the contribution of the use of a reason for every feature and plotted a heatmap for it. Heatmap was preferred because it is better to convenient for 2D plots and is easy to read due to the existence of sharp contrasts.

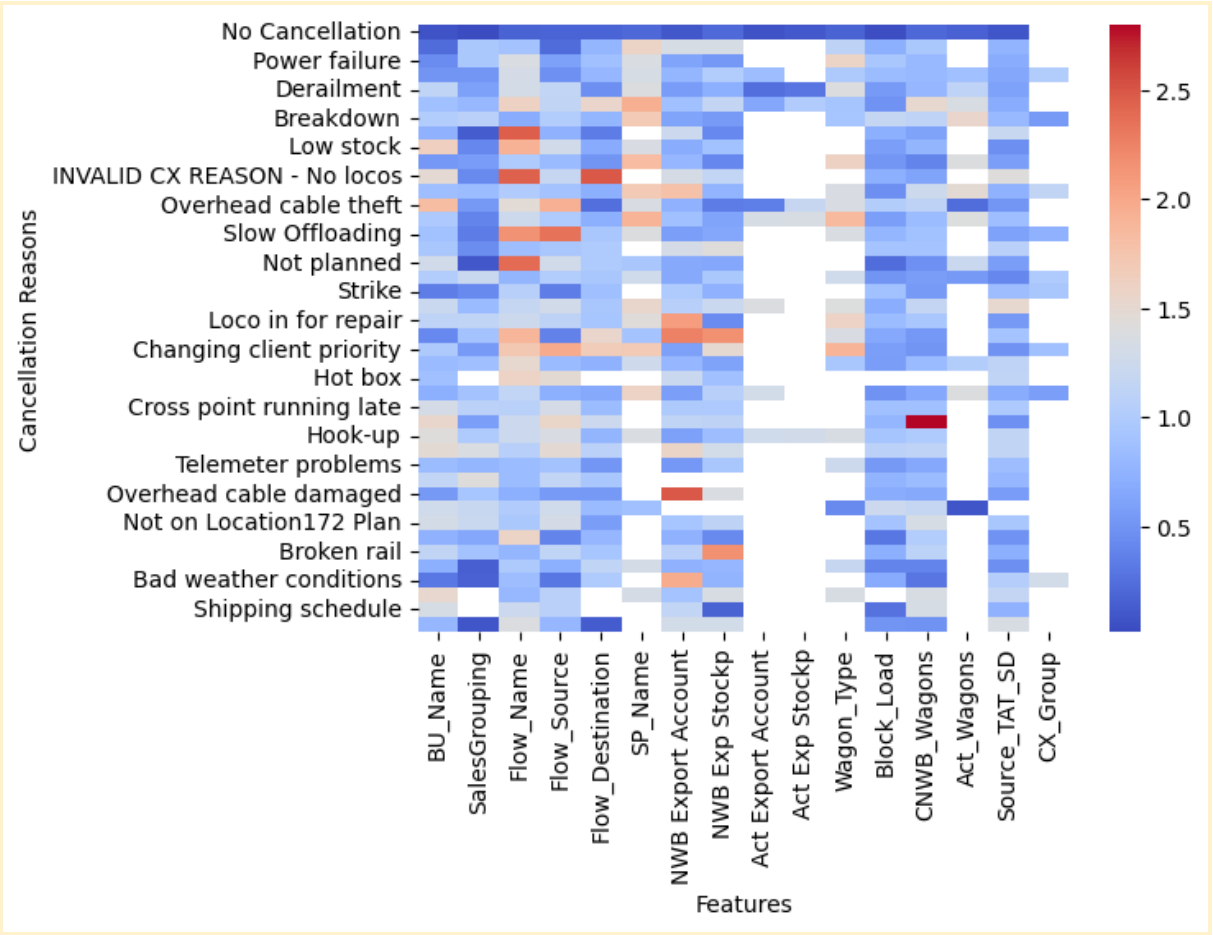
Data Preparation

- I converted features containing details related to time to datetime datatype for ease of their usage.

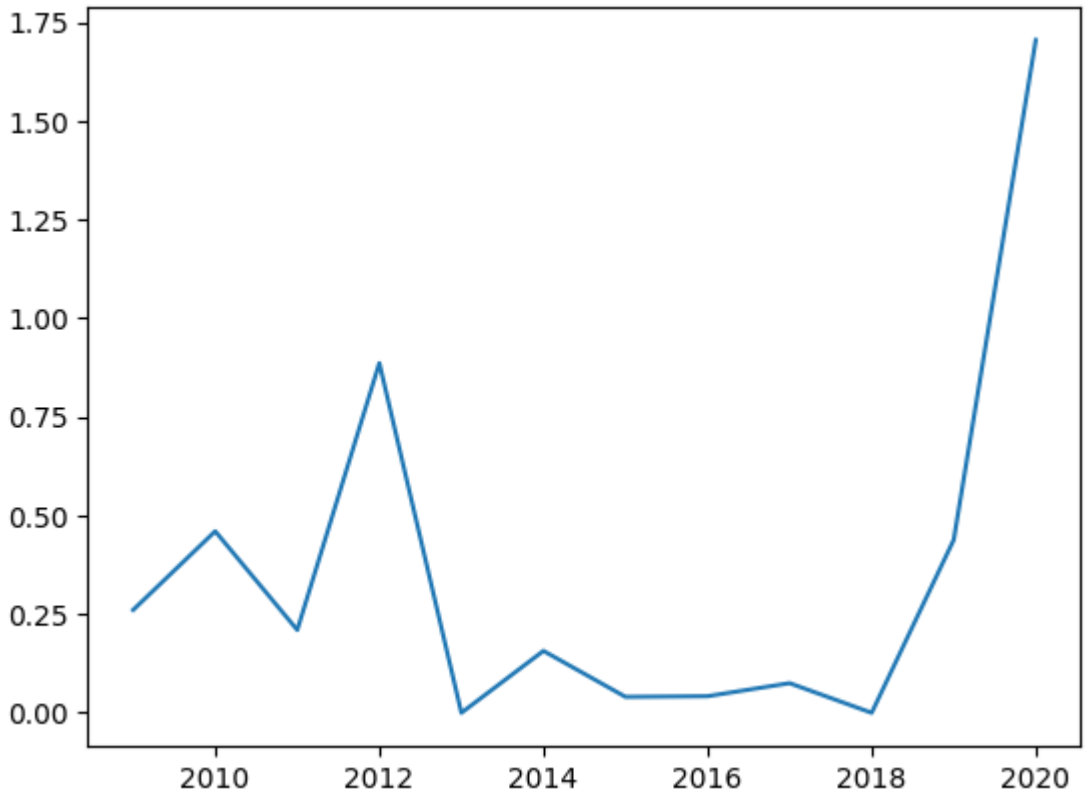
- I removed the following features because of their irrelevance in the approach that I used: Act Acc No, RBCT IngoMass, TFR Delay subtracted, Plan Note External, Act Note External, Cancellation Notes.
- I made some features related to the count and standard deviations and normalizations, which I have discussed earlier as well.

Data Understanding

Following is the heatmap for visualizing the normalized standard deviation of the percentage contribution of the reasons for every feature.



Following is the trend of “low stock” being used as a reason for cancellation



Other Findings

Low stock being used as a reason for cancellation had a small contribution to the overall reasons for cancellations. It was almost null from the year 2013 to the year 2018. But it had a sharp increase in 2018 and has been showing the same trend since then.

Recommendations

- I yet have to explore the time dependence of the reason per feature. High deviation might be explained by the time dependencies of the count of the use of the reasons.
- I can try clustering the data that contributed to high deviations and find some link between them via other features.