Avi Kumar Talaviya Rail Riddle Report

Major Predictions and Approach

What trends or patterns influence rail delays?
A: There are multiple features and factors that influence rail delays. It is hard to say that if there's only one feature that directly impacts the rail delay. In both the datasets, commonly I found that multiple or one major feature can influence the train delay. Most commonly Business unit, customer, route of the train, load, and wagon deviations, and possible weather-related factors were involved.
I have written feature names from both datasets that impacted the train delay here.

1) Qmaster dataset: ['NWB trains.NWBPreliminaryDate','NWB trains.NWBConfirmedDate', 'CargoMovementEndDate',' Destination.DestinationTransactionDate','BU','Customer','Route',' Destination.Split','Consignment','Destination.DestinationQuantSum',' Wagons.Planned','Wagons.Tonnes','Wagons.Actual',' Wagons.Tipped',' Wagons.Bypassed'] and ['Cancelled.ReasonGroup',' Cancelled.ReasonSubgroup']

Apart from this, weather and load shedding are also important factors in train delays.

2) Old system dataset: ['BU_Name','SalesGrouping,'Flow_Name','Flow_Source','Flow_Destination',' SP_Name',' BU Plan Date','Actual Volume Date','Train Status','Wagon_Type','CNWB_Volume','Act_Volume',' CNWB_Wagons','Act_Wagons', 'Source_TAT_SD','THT','CX_Group','CX_Reason']

What is the reasoning behind your answer?
A: After analyzing both datasets, a common observation is that there can be multiple reasons behind the train delay on any particular day.

- In the Qmaster dataset, I found that total of 7008 trains were delayed after subtracting the train's planned run time from the actual run time. Among them, feature 'BU' indicated that around 9 BU locations showed train delays, and one particular location called 'Location54 Location194 Coal' had the most number of delays.
- In the same way, using the crosstab method Few customers and route locations showed train delays.
- In a number of product ('Destination.Split') features, most trains carrying 1 item were delayed.
- 'Consignment' feature also showed some train delay but it had many dummy or incorrect entries
- Around 3812 trains were delayed when the actual load was higher than planned.

- Around 137 trains were delayed when actual wagons were greater than planned.
- Wagons.Bypaased also showed if bypassed wagons are low then trains can be delayed.
- In the Weather dataset, Cloudy weather and low temperature showed several train delays and according to available data, it was statistically significant.

- In the Old system dataset, it was very similar patterns that affected the train delays. Like BU name, customer, route, origin and destination of the train, train status, and sales grouping all had a relationship with train delays. However, Wagon and load tonne deviation showed a positive correlation with train delays. If the actual load or wagon on the train is higher than planned than train is likely to be delayed.

How robust are these patterns, and how might they be improved?
A: Most of these patterns are robust across both datasets, but there was some incorrect information about the Customers name it showed the train delay in Qmaster dataset.

- Wagon and load tonnes deviation was showing a positive correlation with the train being delayed in Old system dataset but it was not the case on the Qmaster dataset.
- In Qmaster dataset only wagon deviation showed positive correlation with train delays. It seemed that Qmaster dataset has much more incorrect information than Old System dataset.
- Adding to that, Load shedding data did not show any significant relationship with train run times. Only two trains were about to run during the load-shedding time. But in both cases, load shedding was at a 0th level
- Weather data that I used showed a significant pattern to cause a train delay.
- Customer named ‘Location35 EITAG’ showed the most train delays in Qmaster dataset.
- Other features like BU names, Route, Customers, Train Status, and Items loaded on trains showed a correlation with train delays in both the dataset equally.

Some types of delay may be more predictable than others. Which types of delay are most and least predictable?
A: Train delays which didn’t have any cancellations were most predictable than otherwise.

- Those trains which were cancelled and delayed both at the same is the point of concern. We cannot say if they were cancelled only because of train delay only. But one thing is that all canceled trains were under train delay category as well. So, it could be possible that if trains are delayed due to some reason then it might get canceled. Here our target variable itself changes so such cases cannot be analysed as a train delayed as it changes the objective itself.

Key Features or Metrics

What are the key features or signatures in the data that help us predict rail delays?
A: I wrote key features as a yellow highlighted text above. Those features helped in data analysis to identify potential causes behind the train delays.

- Apart from that, I created a few new features from the dataset like train run deviation from planned run time, wagons deviation, and load deviation which helped in data analysis
- I also used load shedding and external weather data sources with features like temperature, cloud density, Wind gusts, etc.

Did you use any additional external data?
A: Yes, I used Weather data sourced from below website:
https://www.meteoblue.com/en/weather/historyclimate/weatherarchive/south-africa_south-africa_8335359

Did this data improve your ability to predict rail delays?
A: Yes, certainly this dataset had temperature and cloud cover features. Which proved to be potential reasons behind the train delays but this is an average weather measurement across the regions. We cannot say about the temperature at any particular location as they are anonymized.

Methods and Approach

What is the best analytic approach to take?
A: I would say, analyze both the datasets separately as both of them are having different trains recorded and it also have some different features. So there is no need to merge the datasets.

- Afterwards, I recommend finding the target variable with train delay in our case. But we may need to create this feature from existing features.
- Once the target feature is identified, find the best possible predictors based on domain knowledge and do the univariate analysis to check the distributions of the features
- Then, we can do bi-variate and multivariate analysis keeping the target variable in analysis to identify potential causal relationships that help to understand the train delays.
- We can use the crosstab method of pandas and pointplot/stripplot for such analysis. For multivariate also we can use pandas' crosstab method.
- It is recommended that perform a statistical test for critical variables is to see if the results are significant or not.
- In this case, date time data is at the core of the analysis, so make sure you identify important time series columns and utilize appropriate charts for the analysis.

If you think machine learning is viable, what type of algorithm would you use?
A: I do not think that machine learning is viable based on a given dataset, knowing that we want to predict potential delays at least 3 weeks prior to the train run. Some variables may change in last few days of train run while load shedding, the weather-related variable can change on the day of train run.

- It is better to identify the root cause of the train delays, like certain location, BU or route causing the delay then identify that cause and do the maintenance or correction to avoid the train delay.

- Having said that, there’s still the possibility of developing a model that can help in decision-making rather than predicting train delays explicitly. We can build some decision tree-like model to know what factors are likely to cause delay but not necessarily at least 3 weeks before the train run.

What did you try?
A: No, I haven’t tried any ML modeling in this case.

Did anything not work?
A: There are many incorrect or mispresented data recorded especially in Qmaster dataset causing difficulty in analysis.

- In Qmaster dataset train start time and end time were the same which was obviously false as both times cannot be the same. This wasn't supposed to be there

Data Preparation

Were there any variables or variable sets that required significant cleaning, or contained significant missing values? How would you recommend dealing with these?
A: There are many features with missing values as well as false information input which was recorded in the dataset.

- Considering the project’s objective there was no such need to clean any variable except train run time variables needed to change data time column data type from “Object” to “Datetime” using pandas ‘to_datetime’ function. The same issue was there in load shedding dataset where it also needed a datetime column to change its data type.
- I removed columns that were filled with many missing values like column “Wagons.DTK” in the Qmaster dataset.
- Other variables which have missing values but they were not useful in analysis for there wasn’t any need for cleaning.

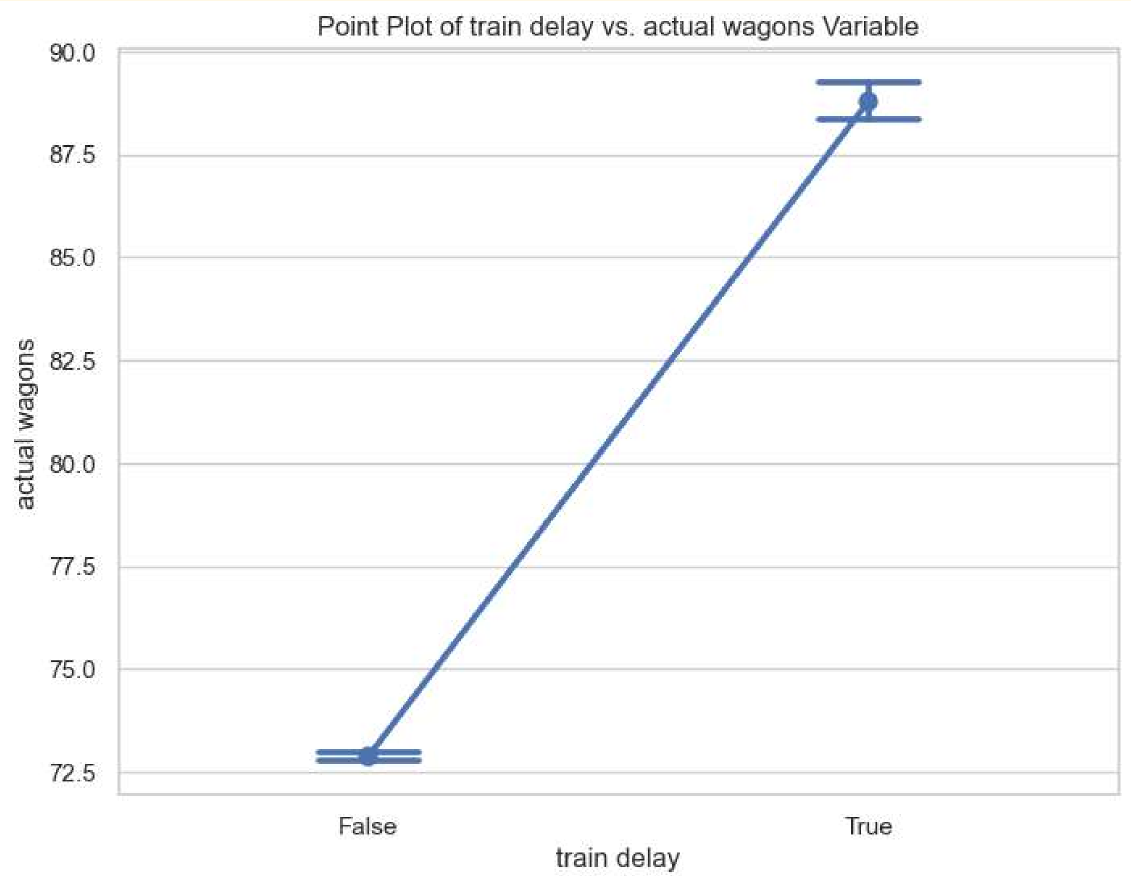
What other sources of data would be most helpful in predicting rail delays?
A: Weather measurements like temperature, Wind, cloud cover, rain, and any natural calamities dataset would be helpful in predicting rail delays.

Data Understanding

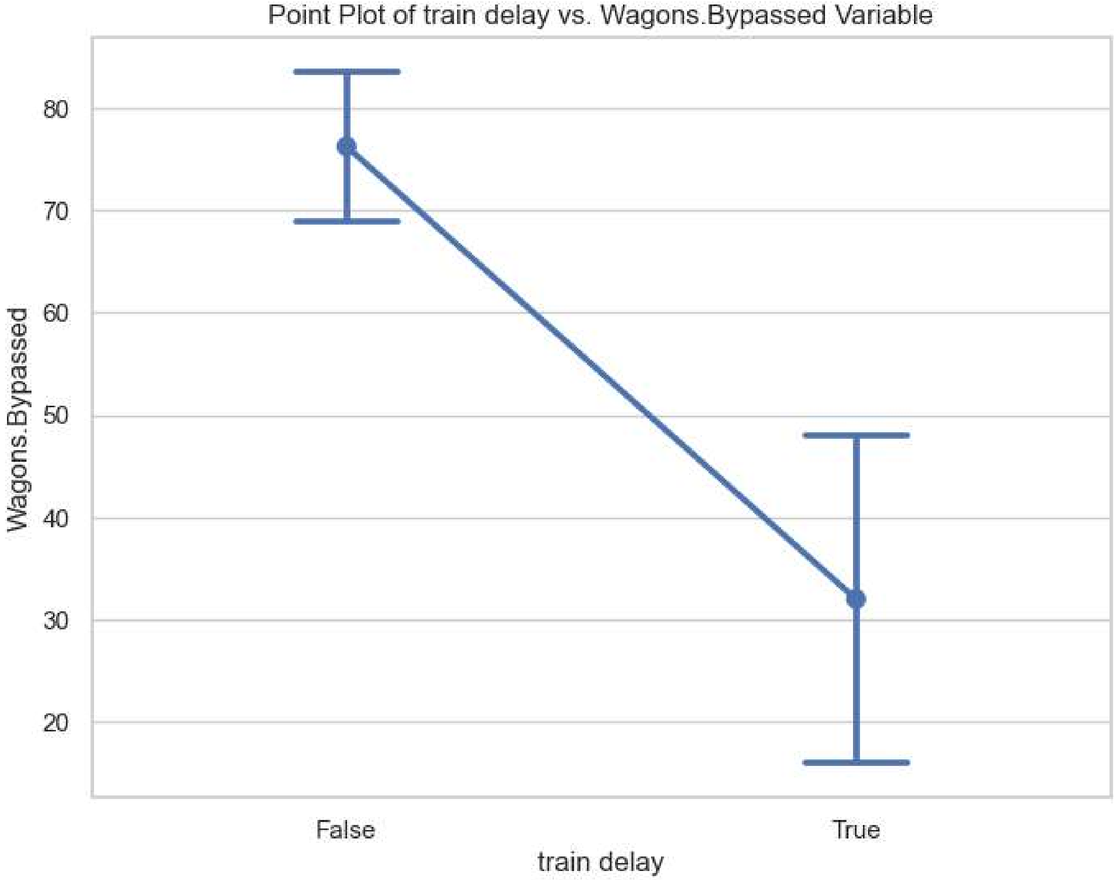
Include 1 or 2 visuals that demonstrate the relationships between variables and the target variable.

A:

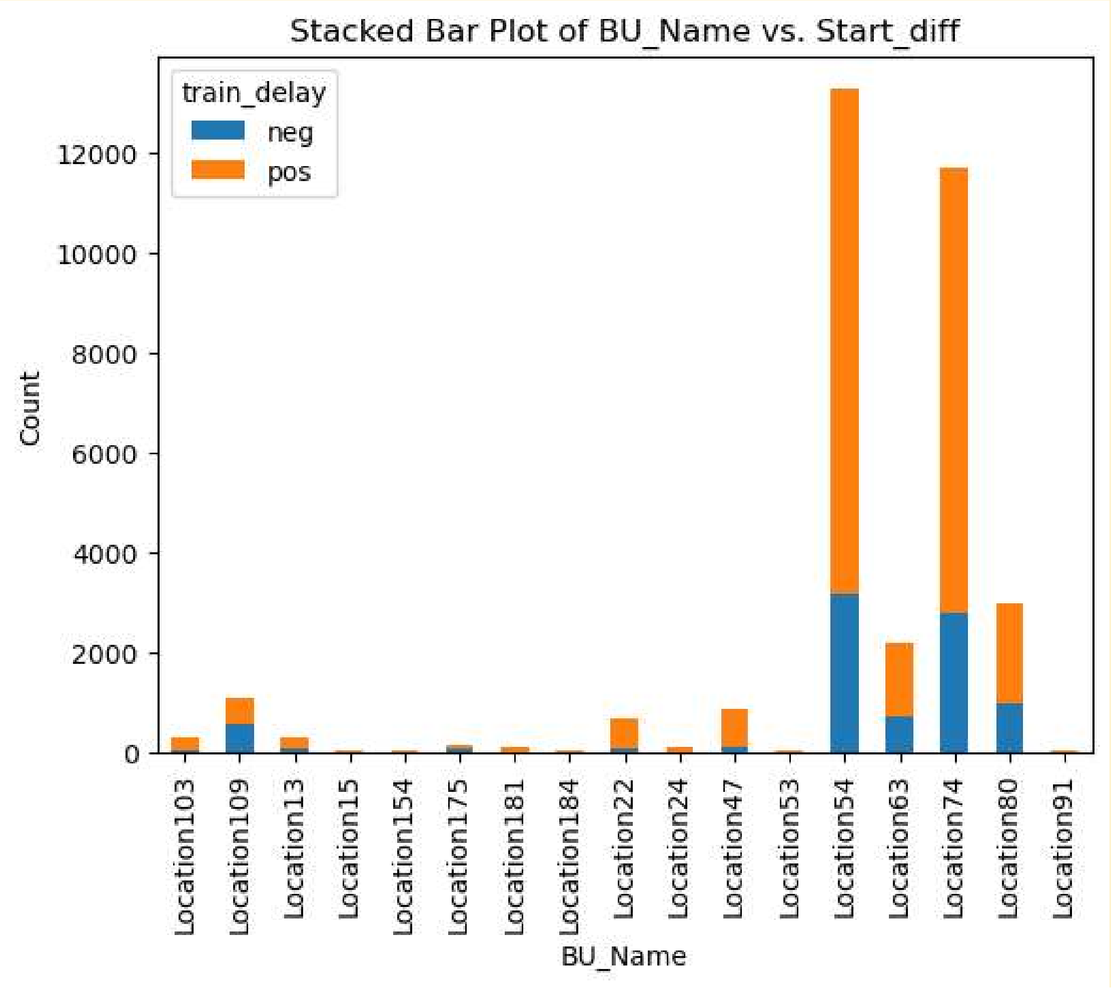
1) Plot of train delay w.r.t actual wagons loaded on the train



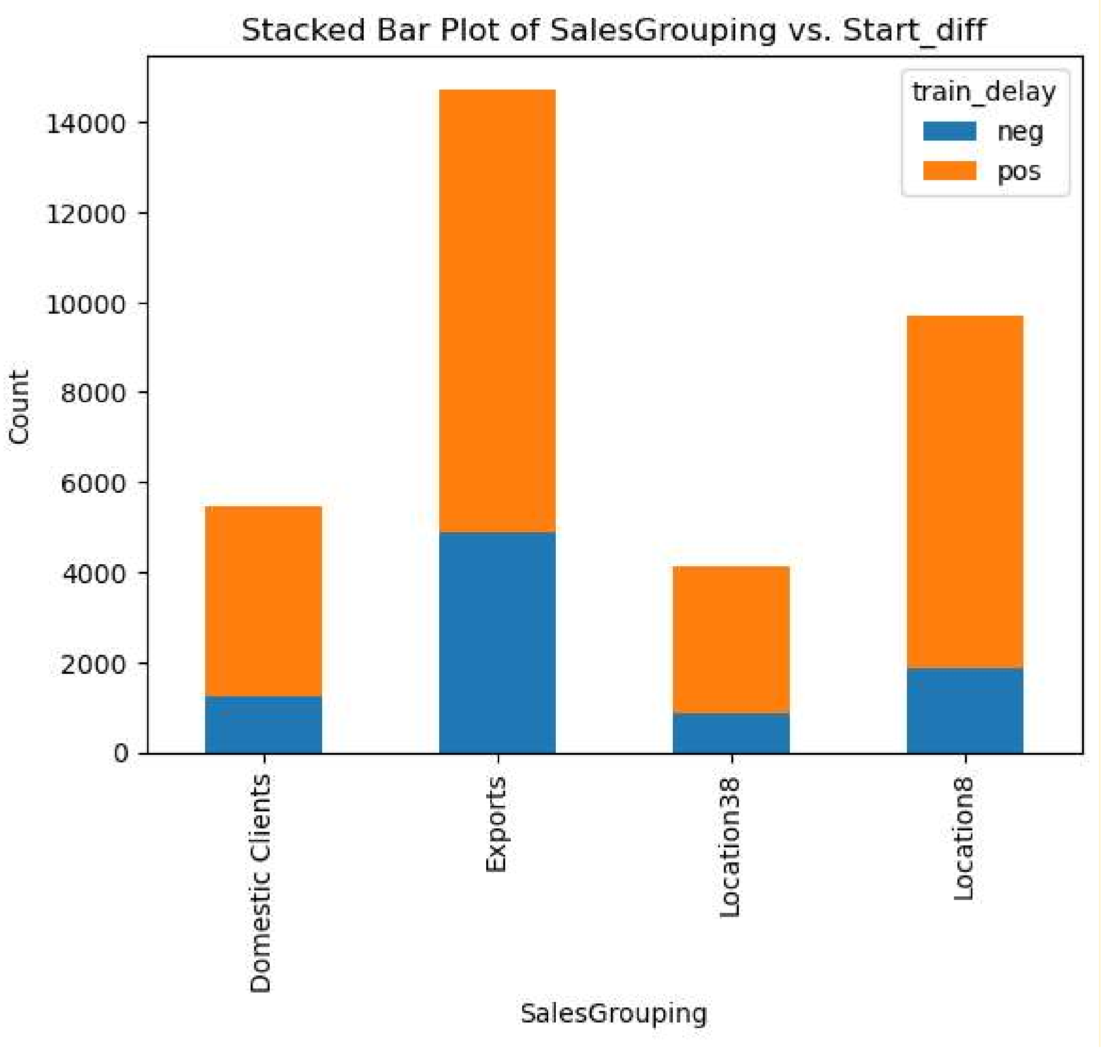
2) Plot of train delay w.r.t wagons bypassed



3) Stacked bar chart of BU_name w.r.t train delay (here “post” means train delay and “neg” mean train ran earlier than planned)



4) Train delay w.r.t SalesGrouping



Did you create any other features from the dataset?
A: Yes, I created multiple features to measure train delay, load (in tonnes) deviation and wagons deviation.

- We have planned and actual time of the train run so this gives a train delay feature
- We have a planned and actual wagon as well as a load (in tonnes) so these can provide any deviation in these two features that can cause potential delay

Other Findings

Have you found any other valuable insights from the data?
A: QMASTER DATASET INSIGHTS

1) In Qmaster, location 54-194 Coal has the greatest number of train delays
2) In Qmaster, location 13-125 and location 80-125 (84t) have a 15 and 54 train delays respectively.
3) There are outliers in the number Wagaons.planned column in Qmaster dataset, some values are above 1000 which is not practical at all.
4) Around 91 delayed trains are from Location 172
5) When there are changes in wagons' load and number of wagons. And it it's higher then trains are likely to delay.
6) Most numbers of delays happened with customers named `Location35 EITAG` and it has different routes, so major cause of delay could be from customer side
7) Cancel reason `Location172` AND cancel subgroup reason `Marketing` are the consequence of delayed trains. well, it sounds contradictory that if trains are delayed then why are they cancelled as well? but here the assumption is that delay happens first then cancellation.
8) If wagons are bypassed are low then train is highly likely while if bypassed wagons are high then trains are not likely to be delayed

OLD SYSTEM DATASET INSIGHTS

1) 25197 rows are having train delays among all the records
2) On average actual delay is much higher than allowed delay from above two histograms
3) Delayed trains are not showing under train cancellation which suggests that all the canceled trains must have a different reason of cancellation than train delay. But there are some discrepancies as well when analyse with other columns of train cancellation.
4) Salesgroup Exports and location8 have the highest number of delayed trains which is greater than on-time trains.
5) Train origin location 54 and location 74 have the majority of train delays in this dataset.
6) Train destination locations 125, 179, and 82 have the most delays among all the destinations.
7) if load tonnes deviation is greater than zero then trains are more likely to get delayed.
8) if wagon deviation is positive from planned then it can have high chance of delay, but it is not alone the reason behind the train delays

Recommendations

What do you recommend as the next steps?
A: Overall dataset is of good quality but it still needs some clarity with missing values of train timing and correction of incorrect values. I would recommend this to be checked so that the quality of the analysis can be improved.

What information or data is missing that would be needed to develop a machine-learning solution?
A: ML modeling is complex to develop in this case to predict train delay in the required time frame as I said earlier. For precise ML modeling data needs to be recorded in real-time by incorporating all the factors like weather, load shedding, BU or customer side factors etc.