**Prediction of Number of Trains Departing a Station**

**PROBLEM STATEMENT:**

We need to predict the number of trains departing a station (say, Needles) at each 4-hour interval for all days in the next week. The trains could be bound in either direction from the station of interest. In addition, we need to predict how much time in terms of 4-hour buckets needed to reach a destination station (say, Winslow). We are doing these predictions in two phases.

1. Predict number of trains departing from station 1 in each time buckets of day in next week
2. Split the above number of trains into transit time buckets by predicting the transit time it takes to reach station 2.

PHASE 1 PHASE 2



**PHASE 1 – Departure Prediction**

For estimating the departure, we plan to use the active schedule as the baseline guide, and try to estimate the deviations from it. Three types of deviations are possible –

1. Trains, which were supposed to come as per schedule but didn’t come.
2. Trains which came, but not on schedule (either earlier or later than schedule).
3. Trains, which were not scheduled but arrived in the station.

**Note**: The document and Methodology proposed here is prepared after studying the broad patterns in the March 2017 +1 day Active and Actual Schedule. It might happen that the proposed methodology might have to be changed in view of the insights presented by a larger dataset.

**As Deviation from Active Schedules**

Where,

We would need to estimate each of the factors separately to arrive at the number of trains departing the station.

1. **Trains which were supposed to depart as per schedule but didn’t arrive ():**

A few trains were supposed to come, as per the Active Schedule, but the Actual event database did not have any mention of them. There are a few patterns in the March 2017 data, with the +1 schedule, which could be used for prediction:

* + Trains Priority: Trains of Priority A and L
  + Train Section: TRN\_SECT=1 has a higher incidence of trains not reaching
  + Train Schedule Departure date:
    1. About 77% of the trains that never reached needles, were supposed to have left the origin on or after the day, the estimates (Active schedules) were generated.
    2. About 14% of the trains that didn’t reach needles, were supposed to have left the origin one day prior the estimates (Active schedules) were generated.
  + Weather related information (Snow caused a few trains to be cancelled, for example)

Some of these factors are conditional on trains never coming. We need to check if these patterns actually predict trains not coming.

**Methodology**: On the baseline model, use **statistical mean** to estimate the number. To improve it, we can run a classification algorithm, like **logistic regression** or **Random Forest**, on the data with dependent variable as scheduled train came/didn’t come.

1. **Trains which were in Actuals but not actives**

Some trains made unscheduled stops at Needles. The following variables could be used to predict unscheduled trains arriving:

* Train priority - Trains of Priority A and L were more likely to arrive unscheduled.
* Length and Tonnage: Unscheduled trains have a high incidence of trains with 0 length and tonnage – probably locomotives being moved around.
* Weather information: bad weather causes train to be re-routed at the last minute.
* Increased demand of locomotives or empty cars due to economic activity picking up – this can be measured using quarterly US/state-level GDP data.

**Methodology**: This one is difficult because we don’t have a base schedule to compare with. On the baseline model, we will use **statistical mean** to estimate the number. We could try to see if there is any predictable pattern in the number of such trains, regressed on the factors above.

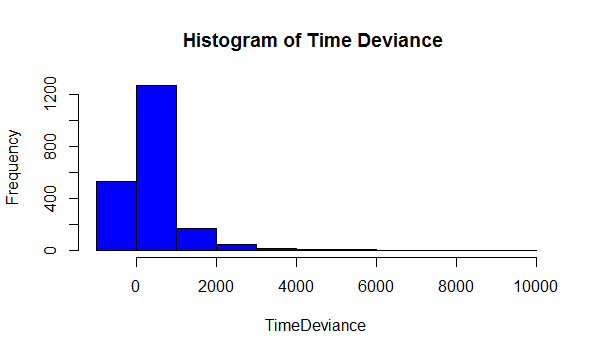
1. **Trains which departed on schedule (**

This includes the trains that were scheduled to depart in bucket i and ended up departing as scheduled. This is likely to be correlated with the scheduled trains based on Active Schedule. We can look at patterns based on day of week, bucket time etc to improve the correlation

**Methodology: Linear regression**, with appropriate transformations on the variables.

1. **Trains which departed from station, but not on scheduled bucket (either earlier or later than schedule) (**

Here trains scheduled actually turn up, but not on time. We could measure this deviance from Schedule by taking the difference from the estimated time of arrival with the actual time of arrival. This quantity –lets term it ‘**TimeDeviance’** - is a heavily right-skewed variable, because trains generally tend to come late. But the negative numbers also indicates that some trains come before time! We need to take this into account when deciding upon the modeling technique.



A few variables, which might be influential in predicting time variance:

* Train load
* Train characteristics, like:
  + Train Type
  + Train Priority
  + Train Symbol
* Weather – inclement weather will cause delay in thee trains (from the Estimated time)

There could be more variables which could be examined for its predictive ability of Time Deviance.

**Methodology:** The method to use would be **linear regression** or a classifying technique like **Random Forest**, with appropriate transformations on the variables.

1. **Trains which departed from station at bucket i, but scheduled on an earlier or later bucket (**

This refers to trains delayed from an earlier bucket which departs from the station. This can be derived using the same analysis from point 4, applied on earlier buckets.

**Methodology:** Similar to point 4 and derived from the analysis

**PHASE 2 – Transit Prediction**

From Phase 1, we have predicted number of trains departing from station 1 in each 4-hour buckets for the next one week. In Phase 2, for the trains departed from station 1, we need to predict the number of 4-hour buckets needed to reach station 2. If a train starts at bucket and it takes buckets to reach station 2, then the arrival time at station 2 is.

Consider the transit time histogram for March-April 2017 period for Needles -Winslow in figure 1. You can see that the histogram is right skewed compared to the planned schedule (Actives) shown in figure 2. This is because of most of the trains are delayed (some can be early as well).

We need to predict the number (or probability) of trains reaching in each bucket. For modelling, we will select those buckets in which atleast 80% of the trains comes in. As shown in figure 1, we can consider the 7 buckets as the dependent variables. We will model them separately.

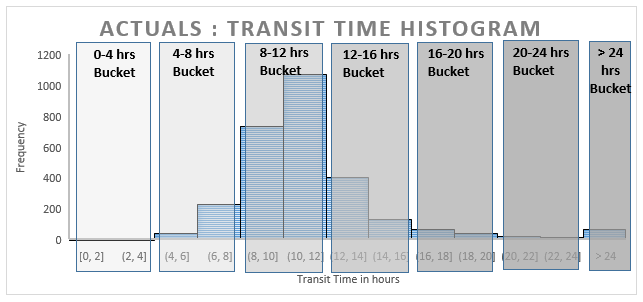
****

Figure : Actual Transit time Histogram

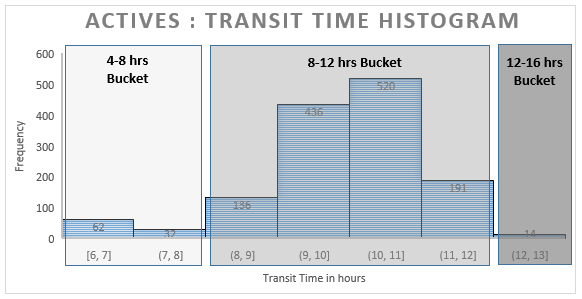
****

Figure : Planned Transit time Histogram

Planned Transit time given by “Actives” will be a good base for these predictions. Ultimately, we need to model the deviance from what has planned and what has actually happening. The other predictors, which we are considering for the prediction, are

* Weekday: The transit time is expected to the correlated to the corresponding weekday. The data is assumed to be weekly seasonal. An average of weekday probability for each bucket multiplied by total trains itself will give a good estimate.
* Departing Bucket of station 1
* Added, Deleted and Shifted trains
* Train characteristics, like:
  + Train Type
  + Train Priority
  + Train Symbol
* Weather – inclement weather will cause delay.

**Methodology:** The method to use would be **linear regression** or a classifying technique like **Random Forest**, with appropriate transformations on the variables.

**Data sources:**

1. We would be using the Actual train arrivals and departures from ‘cwviews. vtrn\_evnt\_bnsf’ table.
2. For the Estimated Schedule, We would use Active Schedule data. Work with +7 Day schedules, and +3 day schedules since our exploratory analysis indicates that +3 closely mimics profile of the Actuals.
3. Train characteristics can be obtained from cwviews.vtrn\_evnt\_bnsf
4. For the weather data we propose to scrap data from a website like Weather Underground which has an API for pulling the data.

**Caveats**

Predicting the breakup of the data by daily buckets would be a challenge as the noise in the data is high the more granular we go. Conversely, the higher the level of aggregation (say weekly vs daily) the accuracy of prediction would tend to be higher as much of the noise is cancelled out.