EDWARD JACKSON

INTELLIGENT RAIL

Introduction

With delays and cancellations on the French high-speed rail network soaring, I set out to address the problem with two clear objectives:

1. Use machine learning to predict routes with delays affecting more than 10% of services;
2. Evaluate & analyse data to identify attributes of journeys contributing most to these delays.

The Netherlands has taken third spot in global punctuality rankings, embracing the power of data and digital innovation in recent years. Dutch operators are now more proactive, developing predictive models to determine maintenance schedules and flexible timetabling. However, many countries are falling behind the curve.

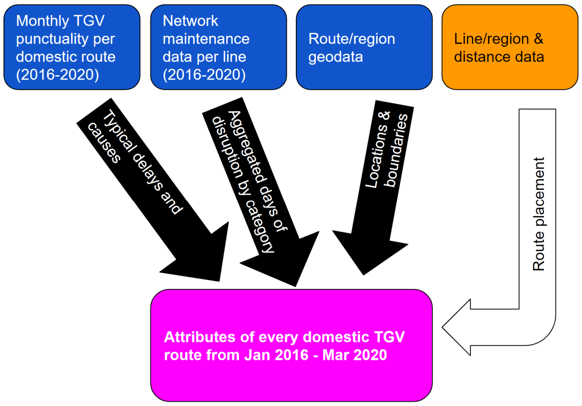
The data

The French national operator - **SNCF** - have embarked on an open data project in recent years, releasing and updating hundreds of datasets about their network. This data is free to share under the *Open Database License*.

Using **git bash**, I was able to download numerous datasets speedily, most notably the following:

- TGV monthly punctuality by connections
- Network maintenance records
- Rated maximum line speed

In addition, I created a separate table using an SNCF route map (both provided in the pack) and a geojson file from data.gouv.fr to describe each route in terms of distance, lines used and regions passed through, very much like a flexible grid system.


Monthly TGV punctuality per domestic route (2016-2020)
Network maintenance data per line (2016-2020)
Route/region geodata
Line/region & distance data
Typical delays and causes
Aggregated days of disruption by category
Locations & boundaries
Route placement
Attributes of every domestic TGV route from Jan 2016 - Mar 2020

Others' approaches to this problem space

In the background, I managed to synthesise over 1.8 million individual journeys representing five years using (derived from SNCF's monthly aggregations). However, it became very difficult to infer the degree of delay for individual journeys while being faithful to the aggregations but not having the distribution of delays.

Most other work on predicted train delays is - of course - based on individual journeys. One example by *Masoud Yaghini* (entitled **Railway passenger train delay prediction via neural network model**) who used neural networks to achieve accuracy of up to 90%, predicting train delays on the Iranian passenger rail network. Another example is by *Chao Wen et al* (entitled **Predictive Model of Train Delays in a Railway System**) who used a Long Short Term Memory network to predict delays on the Dutch rail network, achieving 87.6% accuracy on test data.

Methodology

Data was joined using regions and line codes as keys to build the final dataset. The most complex part of the extensive feature engineering in this project was extracting and collating the maintenance disruptions and categories attributed to each journey.

Extensive data cleaning was necessary on much of the SNCF-sourced data, with examples including:

- imputing missing values where possible
- inconsistencies in data entry (e.g. negative values instead of positives)
- entries for region names varied from table to table
- French language data had to be translated and categorised
- routes had to be one-hot encoded in preparation for modelling

All data was scaled using RobustScaler which proved to reduce the influence of some outliers in the maintenance data and it also returned better results in principal component analysis (PCA): far fewer principal components captured far higher cumulative variance in the data. Test data was always transformed using transformers fitted to training data.

Exploratory data analysis identified major disruption to the rail network from April 2020 onwards due to the pandemic. This period was dropped as it could not be considered ‘predictable’ by route attributes. Another example of effective analysis was the identification of outliers in some of the punctuality data e.g. routes which were ‘early’ by over 100 minutes on average. This allowed me to explore data entry errors and make logical adjustments.

Modelling and evaluation

In an effort to predict route-months with more than 10% of journeys delayed, I employed **five** model types, each highly appropriate for a binary classification problem. Logit was the only ‘white-box’ model type, allowing me to extract log-odds (coefficients) for each data feature.

Each model type was fitted twice (separately) to scaled training data and scaled principal components of the training data (arising from PCA). Validation subsets and k-fold cross-validation was employed to improve models without exposing them to the test data subset.

Recall score was important in this project because I felt that industry experts (and even customers) would be more trusting of a model capable of reliably predicting ACTUAL delays correctly (i.e. limiting false negatives as much as possible). Model results are as follows:

MODEL	DATA	Validation	Test	Validation	Test	Validation	Test
		ACCURACY		**RECALL SCORE**		**PRECISION SCORE**	
Logit	Scaled	86.5	86.5	85	85	88	88
	PCA	83.2	82.3	80	80	86	84
SVM	Scaled	86.0	85.3	82	83	89	87
	PCA	86.9	85.8	82	82	91	88
Random Tree	Scaled	81.5	80.2	80	80	82	80
	PCA*	80.4	78.5	77	75	83	81
XG Boost	Scaled	86.3	85.3	85	84	87	87
	PCA*	79.0	79.5	78	78	80	81
ConvNet	Scaled	90.0	88.9	94	85	97	93
	PCA	87.8	87.3	90	84	93	90

** model uses same hyperparameter settings as model on scaled data*

We can see that the neural network has returned superior performance in all areas, albeit further work is required to fully mitigate any overfitting (note test scores lower than validation scores). However, we can see that logistic regression returned the same recall score of 85% on the test data.

Insights

Using the log-odds (coefficients) from the first logit model, we can take their exponent to find route attributes increasing and decreasing odds of routes breaching the 10% delay threshold.

It would be very difficult to draw clear conclusions from these coefficients with stations on and off the main LGVs increasing and decreasing the risk of route delay. However, I do feel that routes in the South-East and served by Lyon are vulnerable to increased delays due to the frailties of the regional line 930 running from Marseille to Monaco. There are no escape routes on this line and there is also no through-route beyond Monaco. Line 930 is indicated in our analysis (and repeatedly in many previous iterations) as increasing the odds of breaching the route delay threshold.

Looking ahead

We have been able to use machine learning to predict routes with delays affecting more than 10% of services but with many areas for improvement:

- further hyperparameter optimisation, including restructuring the CNN
- Experimenting with different definitions of delay (e.g. over 15 minutes)
- Experimenting with the arbitrary 10% threshold for delays on a route

The log-odds coefficients also highlight that a few features have very high association with routes breaching the 10% delay threshold. Further features might be introduced (e.g. weather conditions) while others could be removed. The list of features does need to be manageable if any model is to be sufficiently flexible and portable for use by industry experts remotely and on different networks.

I would be especially interested in developing models against datasets comprising individual journeys, the timings at different station stops, leg split times and more granular track sectional data. The possibilities are very exciting… but the industry as a whole must adapt to embrace the potential of data and digital innovation. The successes in the Netherlands and Japan (to name but a few) are testament to that.