Predicting Near-Term Train Schedule Performance and Delay Using Bi-Level Random Forests

Mohammad Amin Nabian, Negin Alemazkoor, Hadi Meidani

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

INTRODUCTION

- In Netherlands Passenger train transportation is an important mode of transport.
- Over a million passengers traveling by train daily [1] .
- Train delays cause significant monetary and non-monetary losses.
- About 22% of trains in British national railways system are delayed in 2006-07, which resulted in a total delay of 14 million minutes [2].
- It is important to have an accurate estimation of delays to alleviate such high cost.

OBJECTIVE

Objective: Given the current delay of a train, forecast the delay for that train 20 minutes later.

State-of-the-practice: It is assumed that The current delay remains unchanged [1].

Tasks:
- Prediction of direction of delay change.
- Prediction of delay jumps.
- Prediction of actual delay (in minutes).

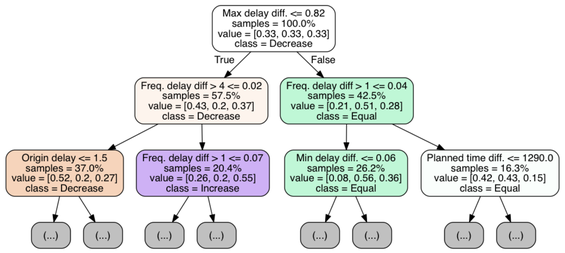
DATA DESCRIPTION

4 September 2017

9 December 2017

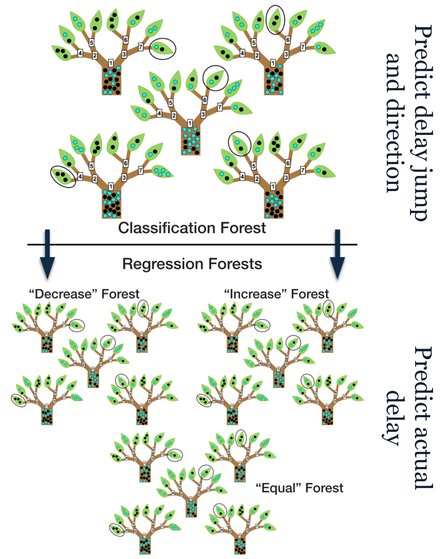
Includes 10,000,000 data points for a thirteen-week period.

DECISION TREES



- At each node, select the feature and threshold that **minimizes uncertainty**.
- Use **cross-validation** to find the optimal depth.
- **Extremely effective strategy**: Build many trees, each for a subsample of data, and classify by merging the results (**random forest**).
- Using classification and regression forests independently may result in **inconsistent predictions**.

BI-LEVEL RANDOM FORESTS



FEATURE CONSTRUCTION

We randomly select an entry from the raw data. Then we search raw data for train location with a planned time 20 minutes. For each sample, given the raw data, we construct a number of features:

	Feature	Notation	Description
Number of arrival-departures	Origin delay	d_s	Delay at the origin station.
	Distance	ℓ	Distance between origin and destination stations.
	Planned time diff.	δP	Difference between the planned time at origin and destination stations.
	Num. A-V	N_{A-V}	Number of train arrivals (A) and departures (V) between the origin-destination stations with a long (i.e. 5 minute) stop.
	Num. KA-KV	N_{KA-KV}	Number of train arrivals (KA) and departures (KV) between the origin-destination stations with a short (i.e. 1 minute) stop.
	Composite change	I_c	Whether the composite has changed during the trip between origin and destination stations. A binary variable.
	Driver change	I_d	Whether the driver has changed during the trip between origin and destination stations. A binary variable.
	Rush hour	I_r	Whether the trip is during the rush hour (i.e. 7-9 AM or 4-6 PM). A binary variable.
Historical delay stats	Delay diff. mean	δd_{avg}	Mean value of the difference between historical delays at origin and destination stations.
	Delay diff. mode	δd_{mode}	Mode of the difference between historical delays at origin and destination stations.
	Max delay diff.	δd_{max}	Maximum of the difference between historical delays at origin and destination stations.
	Min delay diff.	δd_{min}	Minimum of the difference between historical delays at origin and destination stations.
Historical delay stats	Freq. delay diff. > 1	δd_{f_1}	Frequency of the historical events with difference between delays at origin and destination stations greater than 1.
	Freq. delay diff. > 4	δd_{f_4}	Frequency of the historical events with difference between delays at origin and destination stations greater than 4.
	Freq. delay diff. < -1	$\delta d_{f_{-1}}$	Frequency of the historical events with difference between delays at origin and destination stations less than -1.
	Freq. delay diff. < -4	$\delta d_{f_{-4}}$	Frequency of the historical events with difference between delays at origin and destination stations less than -4.
Historical front train delay stats	Front train mean delay	δf_{avg}	Mean of the historical front train delays.
	Front train delay mode	δf_{mode}	Mode of the historical front train delays.
	Freq. front train delay > 1	δf_{f_1}	Frequency of the historical events with front train delays of greater than 1.
	Freq. front train delay > 4	δf_{f_4}	Frequency of the historical events with front train delays of greater than 4.
	Freq. front train delay < -1	$\delta f_{f_{-1}}$	Frequency of the historical events with front train delays of less than -1.
	Freq. front train delay < -4	$\delta f_{f_{-4}}$	Frequency of the historical events with front train delays of less than -4.
Weather conditions	Avg wind speed	W_{avg}	Wind speed daily average.
	Max wind speed	W_{max}	Maximum daily wind speed.
	Avg temperature	T_{avg}	Temperature daily average.
	Min temperature	T_{min}	Minimum daily temperature.
	Max temperature	T_{max}	Maximum daily temperature.
	Rain depth	R	Average daily rain depth (mm).

RESULTS

P_D	P_E	P_I	F_j	F_d	α_{RWMS}	α
0.93	0.67	0.62	0.84	0.77	2.37	9.88

Performance of the bi-level random forest model.

Rank	Classifier	P_D	P_E	P_I	$10F_J + 5F_d$
1	Random Forest	0.93	0.67	0.62	12.25
2	Gradient Boosting	0.92	0.67	0.55	12.22
3	Adaboost	0.91	0.65	0.55	12.16
4	SVM	0.91	0.70	0.55	11.95
5	Extra Tree	0.90	0.68	0.57	11.93
6	Logistic Regression	0.85	0.67	0.55	11.66
7	Decision Tree	0.82	0.59	0.57	11.55
8	KNN	0.78	0.55	0.53	10.91
9	Naive Bayes	0.48	0.84	0.40	8.83

Performance of a variety of classification models.

Model	RF	SVR	Polynomial (d = 2)	Linear	Polynomial (d = 3)
α_{RWMS}	2.12	2.46	2.32	2.48	3.33

Performance of a variety of regression models.

Notation:

P_D: Percentage of correct predictions for 'Decrease' class.
P_I: Percentage of correct predictions for 'Increase' class.
P_E: Percentage of correct predictions for 'Equal' class.
F_j: F-score for delay jump predictions.
F_D: F-score for delay direction predictions.
α_{RWMS}: Root weighted mean square error.
α: Overall prediction accuracy score, defined as:

$$\alpha := 10F_j + 5F_d - \alpha_{RWMS}$$

CONCLUSION

Summary

- We presented a bi-level random forest model to predict near-term passenger train delays in Netherlands.
- At primary level, the model predicts whether the current delay will decrease, increase, or remain unchanged in the next 20 mins.
- At secondary level, the model quantifies the amount of delay (in minutes).
- The proposed model was compared with alternative approaches in the literature.

Findings

- Using the proposed method, the overall prediction accuracy in 9.88.
- The proposed method provided the most accurate predictions on the given dataset, compared to the other machine-learning-based models considered here.
- The proposed bi-level random forest model effectively avoids the inconsistency between classification and regression results.
- The significant features are found to be origin delay, distance, planned time difference, number of arrival-departures, and historical delay stat.

Computation time

- With 150,000 training data, it took 1.84s to train the model.

REFERENCES

[1] 2018 RAS problem solving competition: Train delay forecasting, http://connect.informs.org/railway-applications/awards/problem-solving-competition/new-item2, Accessed: 2018-07-23.
[2] T. Burr, S. Merrifield, D. Duffy, J. Griffiths, S. Wright, G. Barker, Reducing passenger rail delays by better management of incidents, National Audit Office for the Office of Rail Regulation (2008).
[3] D. Forsyth, Probability and Statistics for Computer Science ,Springer ,2018.

ILLINOIS