University of Illinois at Urbana-Champaign
Department of Civil and Environmental Engineering
Uncertainty Quantification Group

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

INFORMS RAS Problem Solving Competition
Final Round
INFORMS Annual Meeting 2018, Phoenix, Arizona
November 4

Mohammad Amin Nabian
PhD Candidate

Negin Alemazkoor
PhD Candidate

Hadi Meidani
Assistant Professor

Contents

- Problem description
- Feature construction
- Methodology: Bi-Level Random Forrests
- Results and comparison with other methods
- Conclusion

Problem Description

- 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.

Goal of this competition:

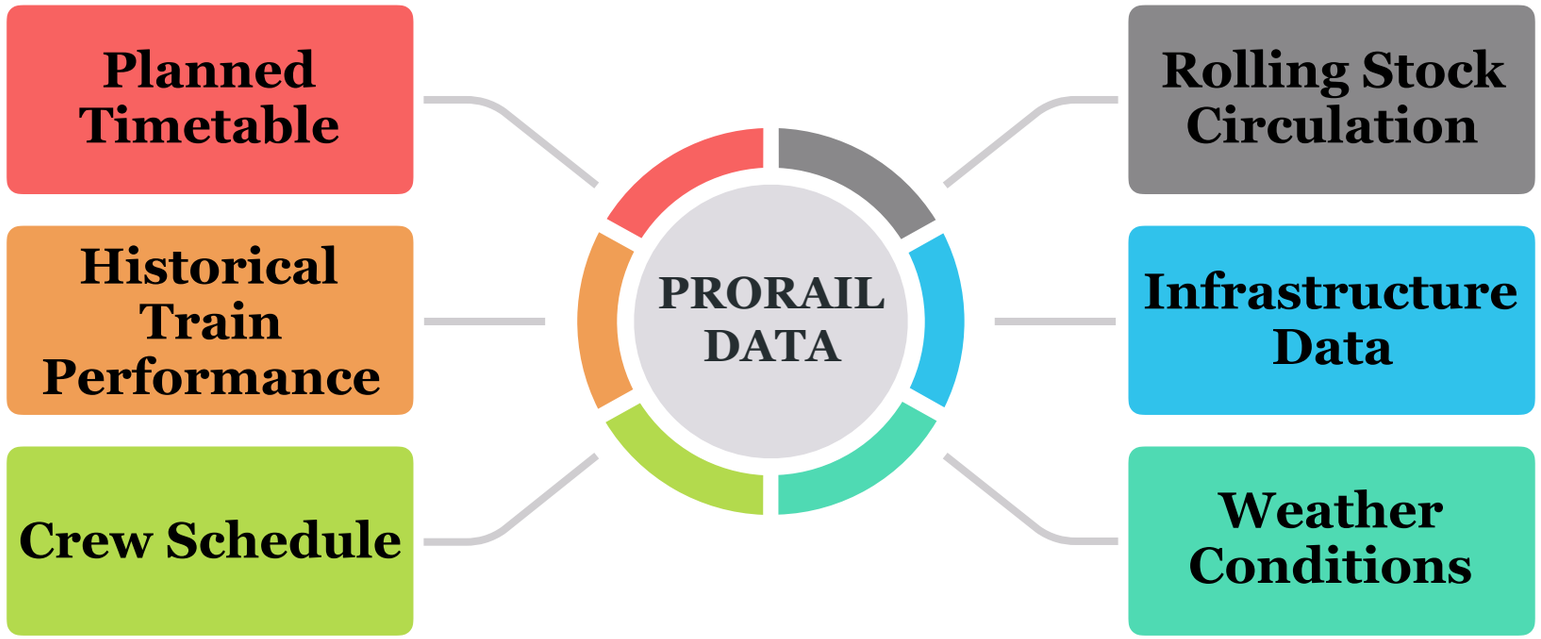
Given the current delay of a train, forecast the delay for that train 20 minutes later.

State of Practice (In Netherlands):

Assumption: The current delay remains unchanged (delay=3 mins at B)

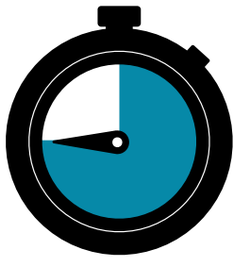
Tasks:

1. Prediction of direction of delay change: Classification
   - Decrease, increase, or equal
2. Prediction of delay jumps: Yes, or No. Classification
3. Prediction of actual delay (in minutes) Regression



4 September 2017

9 December 2017



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

Feature Construction

We exclude:

- Wednesdays data
- Trips with delays of more than 15 mins

To construct our training & test data:

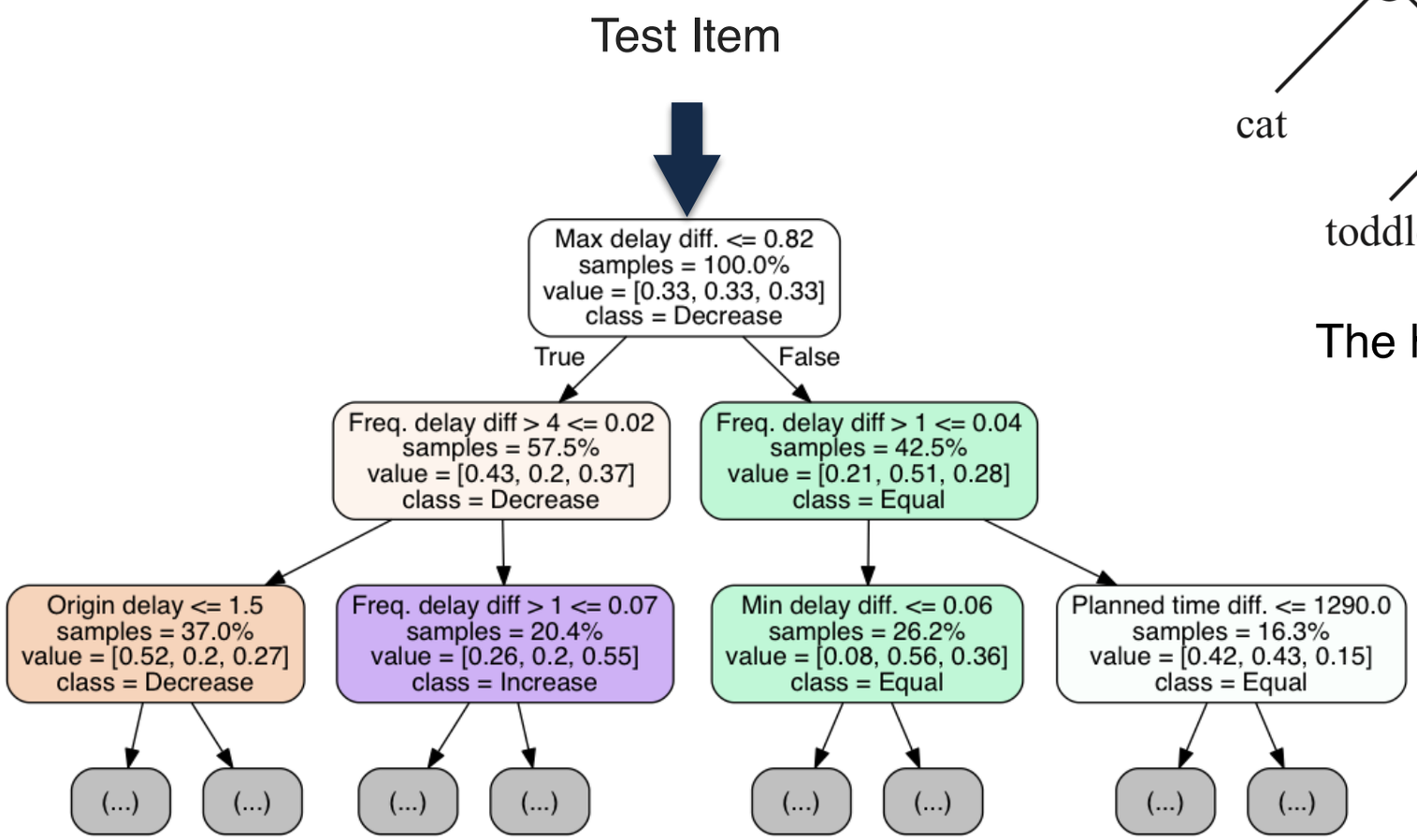
- Randomly select an entry from the raw data (source station).
- Search raw data for train location with a planned time 20 minutes late (terminal station).
- For each sample, given the raw data, we construct a number of features.

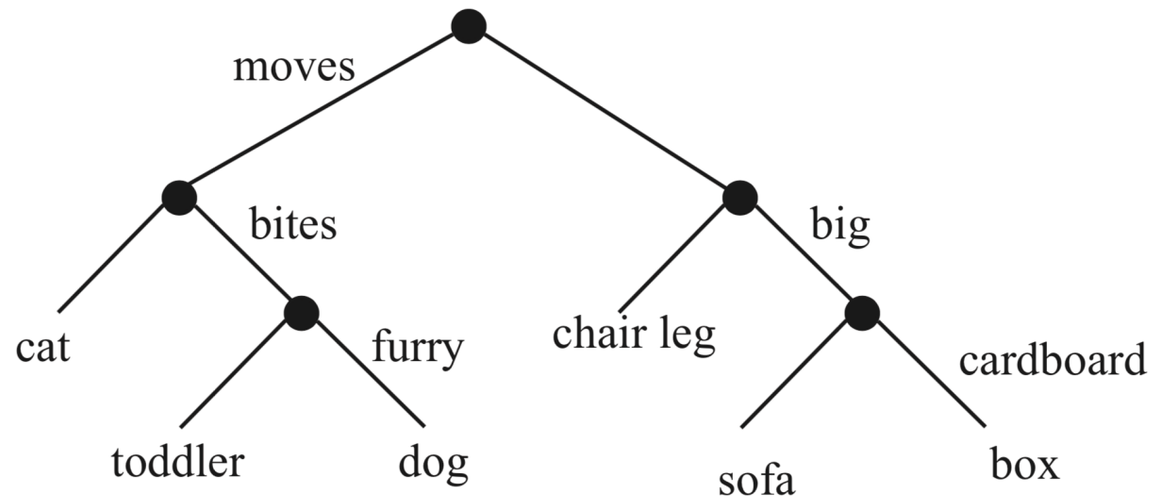
- Size of training data: 150,000
- Size of test data: 30,000

	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.

	Feature	Notation	Description
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).

Bi-Level Random Forests





The household robot's guide to obstacles

Training questions:

1. Depth of the tree?
2. Feature and threshold at each node?

Image is taken from [3]

Depth of tree

- Constructing a too deep tree results in over-fitting
- We use cross-validation to select the best tree depth.

Feature and threshold at each node

- At each node, select the feature and threshold that minimizes the uncertainty:

Classification

$$p_{jm} = \frac{1}{n_j} \sum_{\mathbf{x}_i \in D_j} I(\mathbf{y}_i = m)$$

$$H_E(D_j) := -\sum_m p_{jm} log(p_{jm})$$

Regression

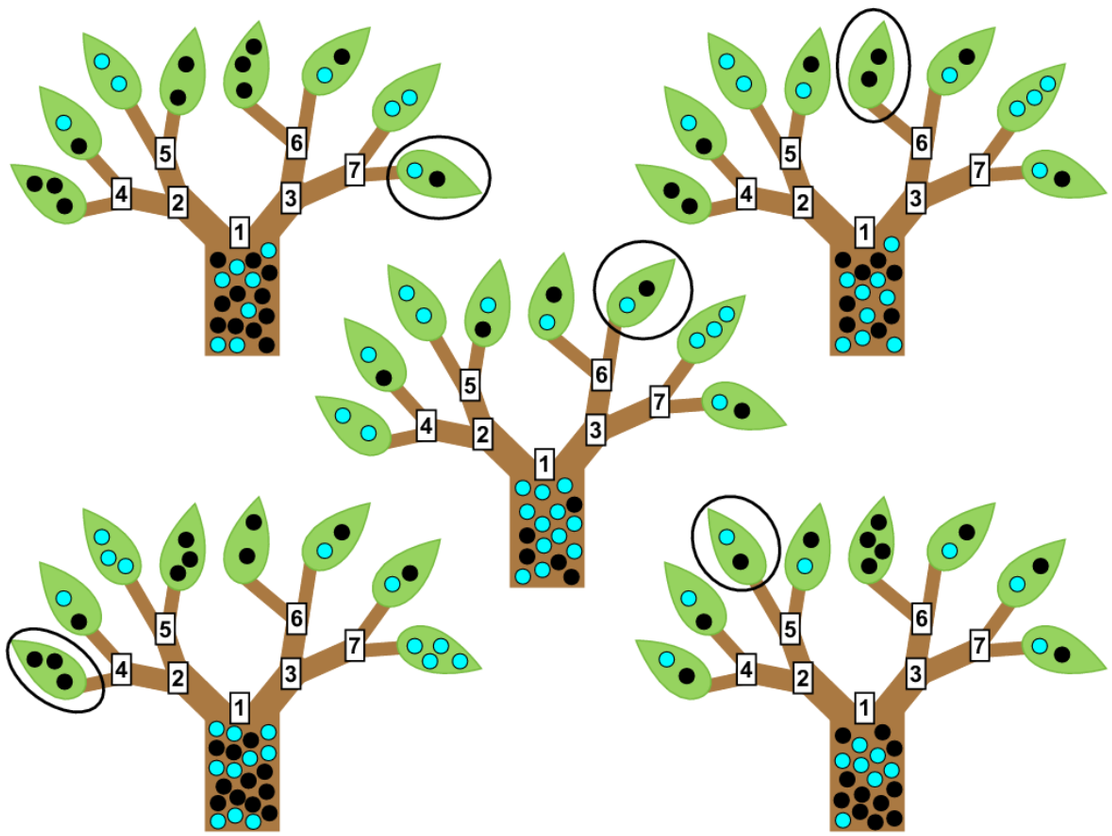
$$\bar{y}_{jm} = \frac{1}{n_j} \sum_{\mathbf{x}_i \in D_j} (y_i)$$

$$H_{MSE}(D_j) := \frac{1}{n_j} \sum_{\mathbf{x}_i \in D_j} (y_i - \bar{y}_{jm})^2$$

- The optimal split is obtained by

$$\theta_j^* = \arg\min_{\theta_j} G(D_j, \theta_j)$$

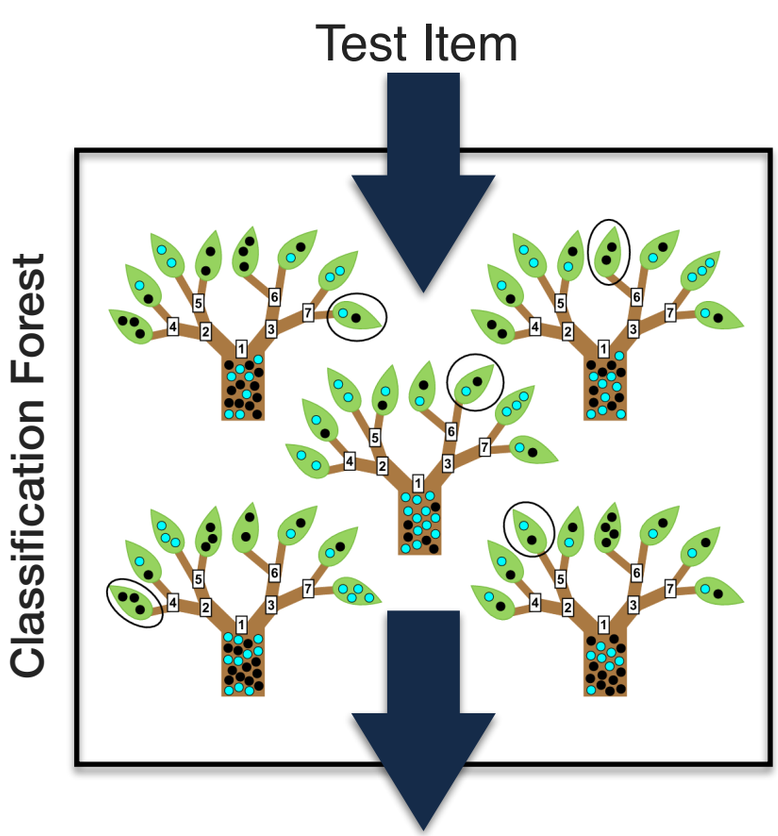
$$G(D_j, \theta_j) := \frac{|D_j^{(l)}(\theta_j)|}{|D_j(\theta_j)|} H(D_j^{(l)}(\theta_j)) + \frac{|D_j^{(r)}(\theta_j)|}{|D_j(\theta_j)|} H(D_j^{(r)}(\theta_j))$$

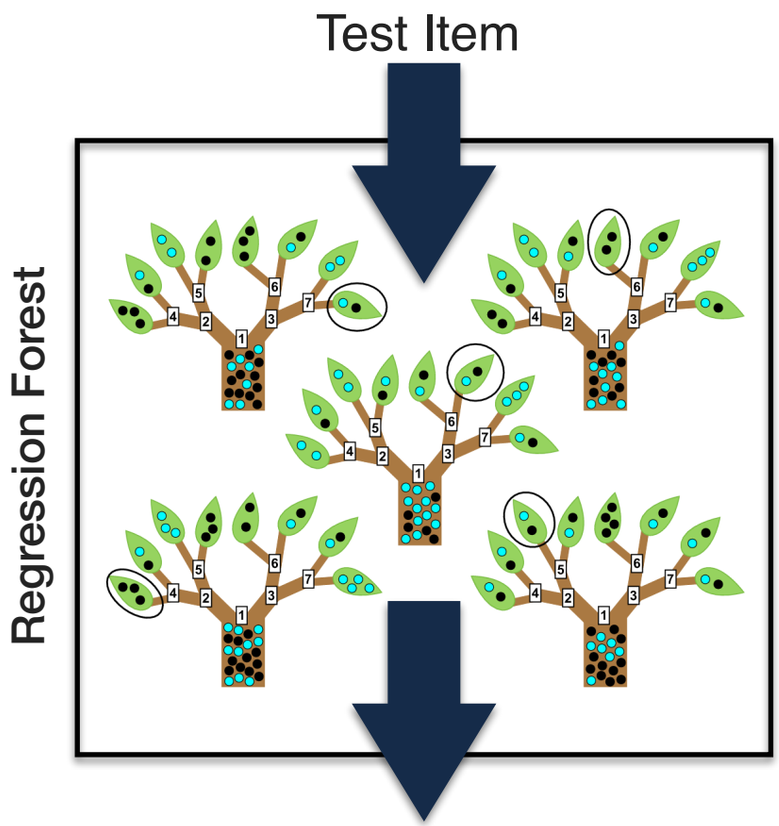


- The tree that gives the best possible results on the training data can perform rather poorly on test data due to overfitting.
- **Extremely effective strategy:** build many trees, each for a subsample of training data, and classify by merging their results.
- **Vote for classification**
- **Take average for regression**

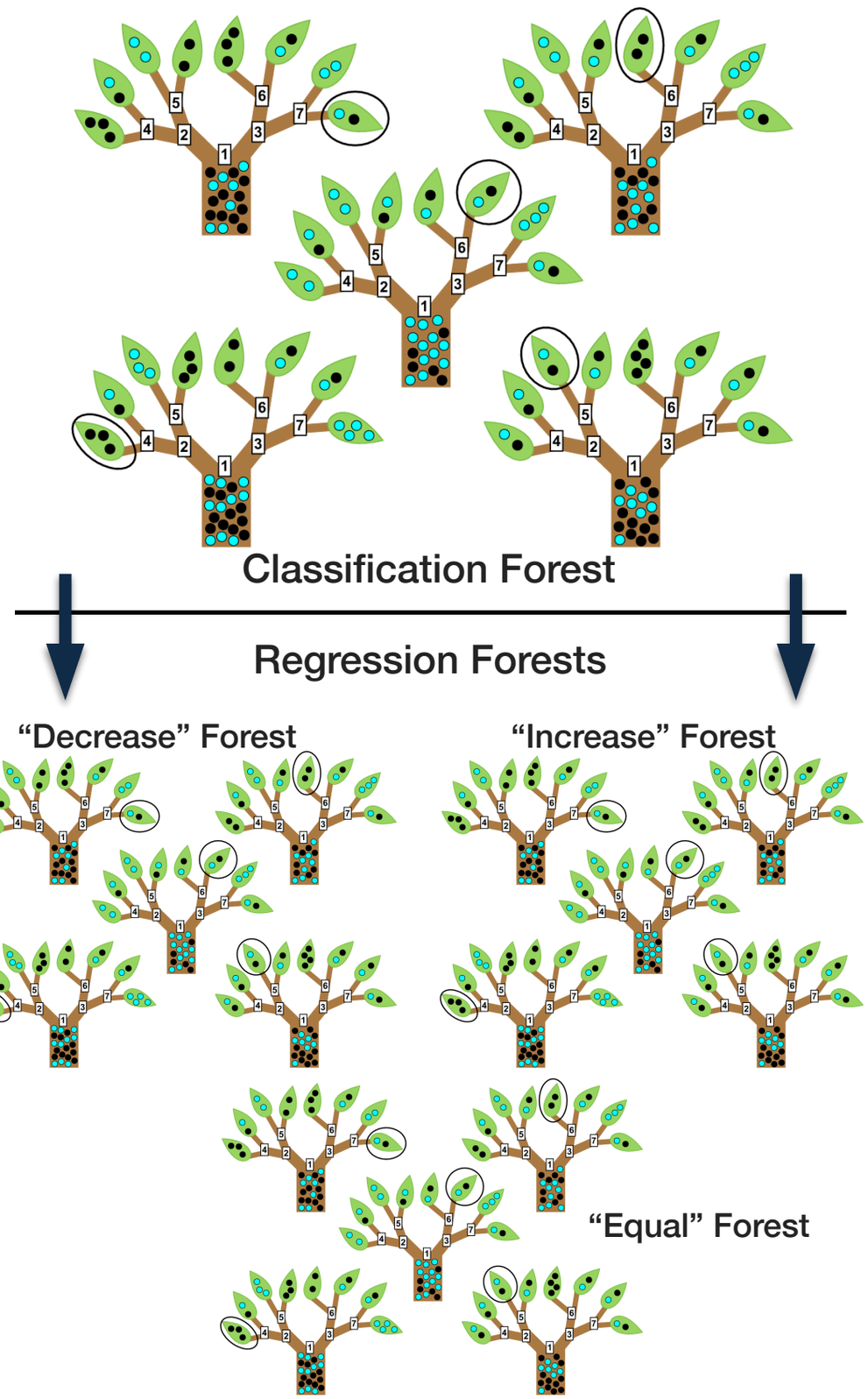
Image is taken from: http://inspirehep.net/record/1335130/plots

- Using classification and regression forests independently may result in contradictory predictions.





- The proposed model inherently avoids this issue by training a unique regression model for each class.
- Therefore, **the proposed algorithm is suitable for coupled classification-regression tasks.**



Results

Performance Measures Selected by Committee:

- Jump and its direction $\alpha_P := \frac{TP}{TP+FP}$ $\alpha_R := \frac{TP}{TP+FN}$ $F = \frac{2\alpha_P\alpha_R}{\alpha_P+\alpha_R}$
- Delay forcecasting $\alpha_{RWMS} = \sqrt{\frac{\sum_{i=1}^{n} w_i(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} w_i}}$
- Overall prediction accuracy $\boxed{\alpha := 10F_j + 5F_d - \alpha_{RWMS}}$

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 predictive model on the test data.

Results are reported on our own test data.

Significant features

Feature	Notation	Description
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.
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.

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.
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.
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).

Performance comparison with other methods

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

A summary of the performance of a variety of classification models.

Regression model	RF	SVR	2nd order polynomial	Linear	3rd order polynomial ($d = 3$)
α_{RWMS}	2.12	2.46	2.32	2.48	3.33

A summary of the performance of a variety of regression models.

Conclusion

- 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.

- The proposed bi-level model provided the most accurate predictions on the given dataset.

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.

Thank you!