Random forest, **source:** http://blog.yhat.com

Handling Missing Values in Decision Forests in the Encrypted Network Traffic

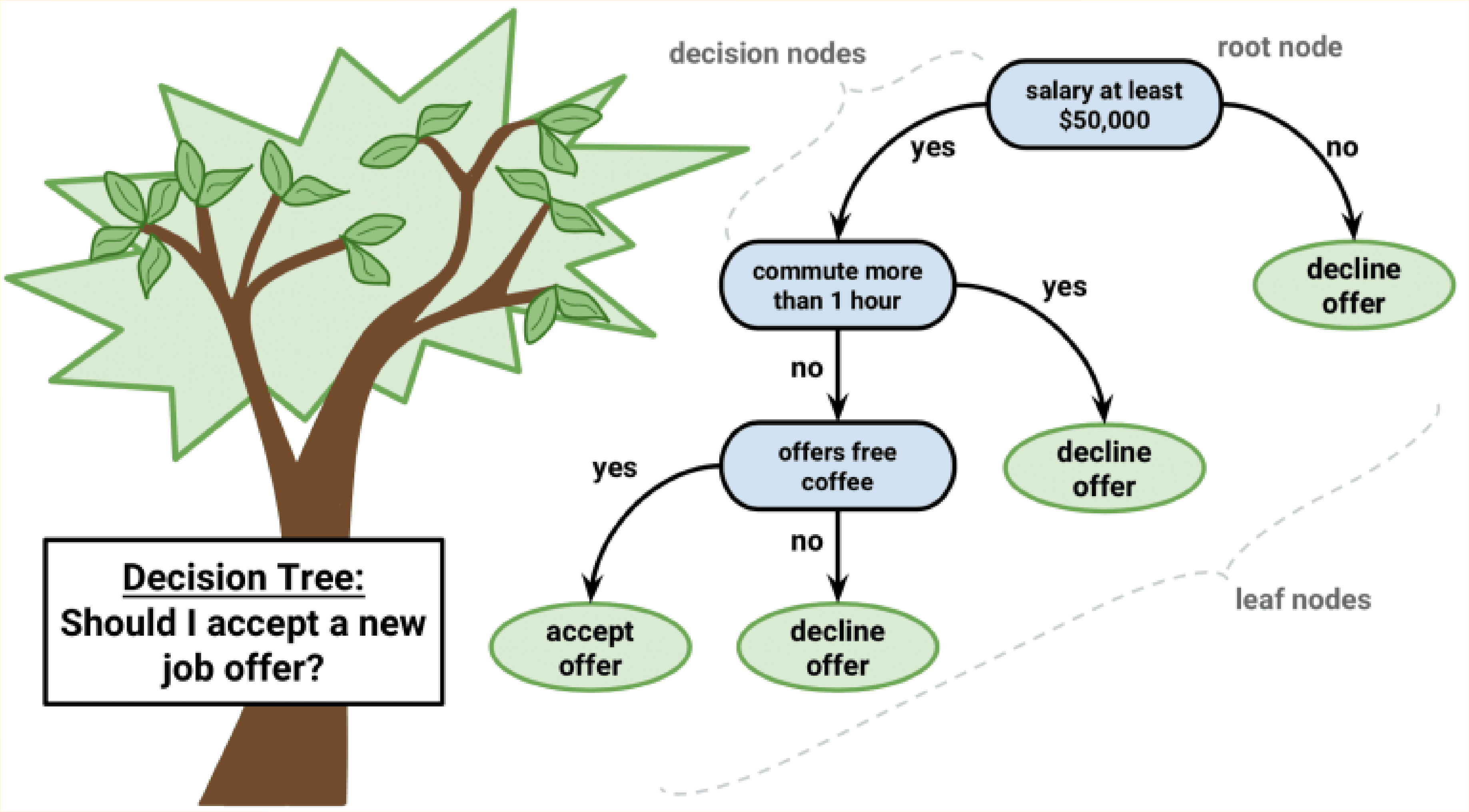
Author: Lukáš Sahula
Supervisor: Ing. Jan Brabec
Bachelor thesis

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Science

Handling missing values...

Name	Animal	Age	Gender	Description
Rex	Dog	3	Male	A good boy
Lady	Dog	X	Female	X
Cat	Cat	4	Male	X
Kitty	Cat	X	Female	Likes to cuddle
Gizmo	X	X	Male	X

… in Decision Forests …

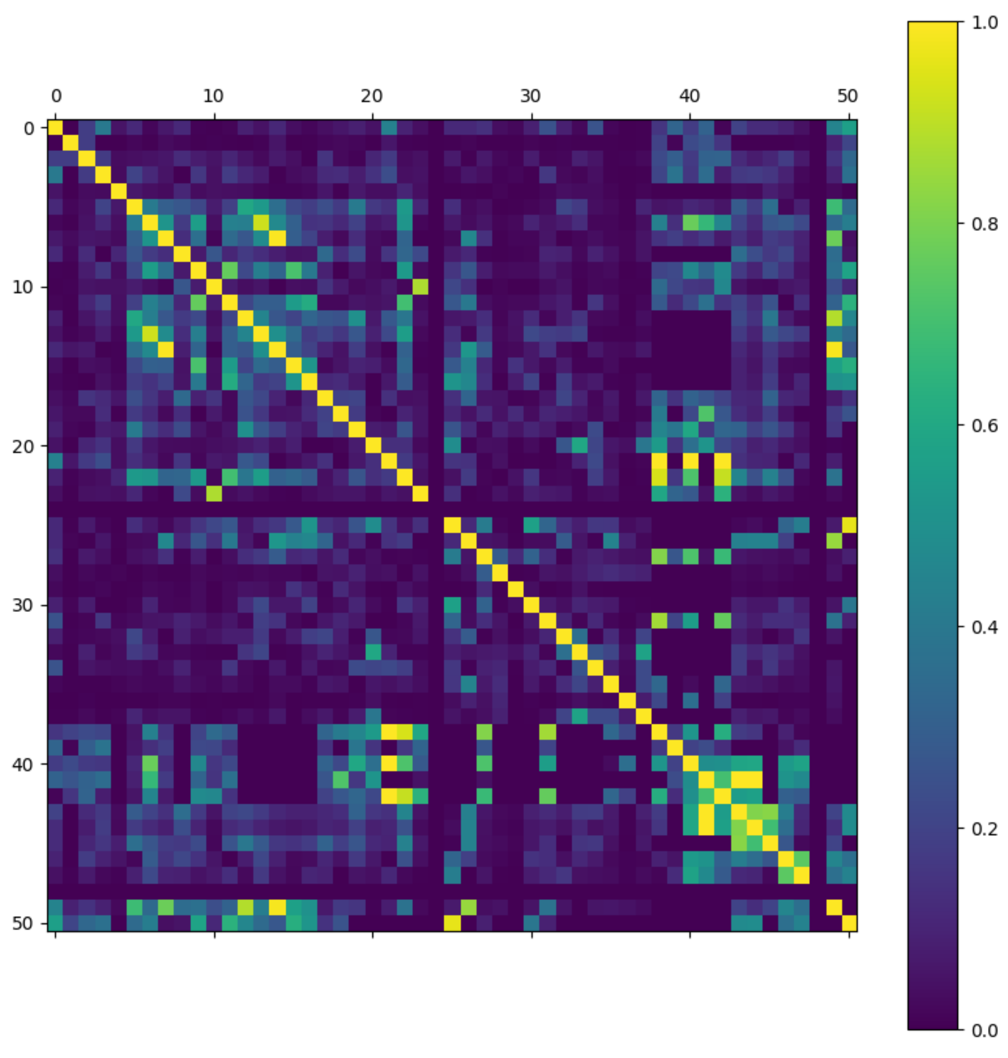


Decision Tree classifier, **source:** http://packtpub.com

… in the Encrypted Network Dataset

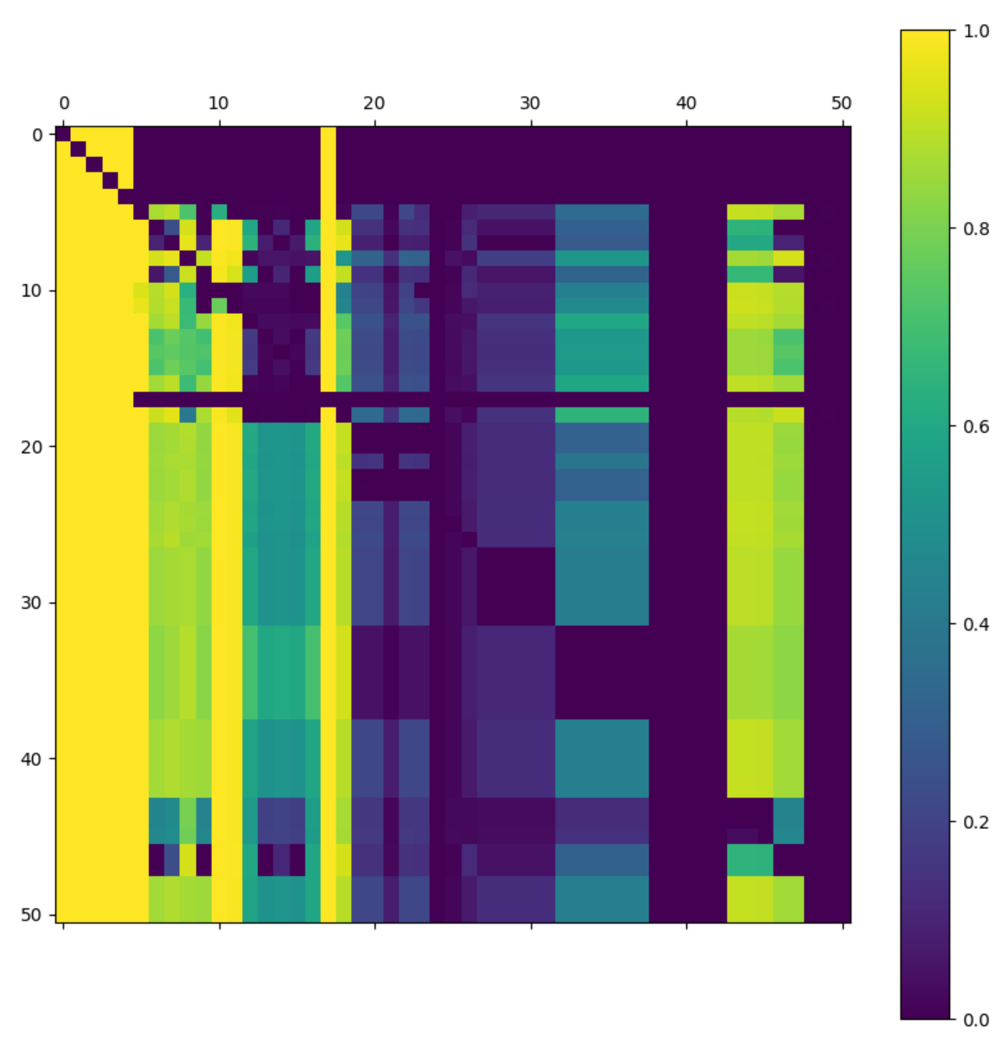
- Data from network proxy logs
- Classification of malware
- Over 100 classes of malware
- 50 features
- Data missingness over 50%

Dataset correlation analysis



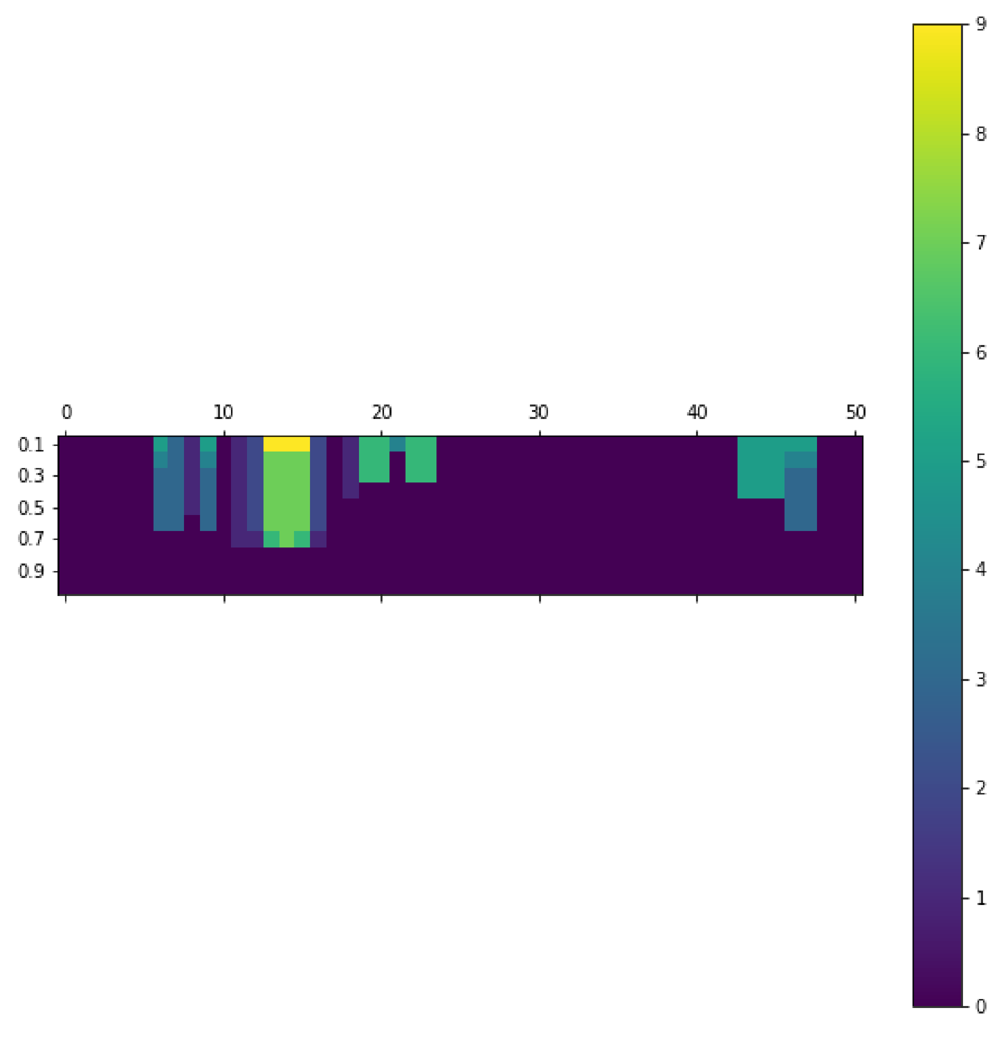
Heatmap of feature pairs correlations (Pearson)

Conditional probabilities of missingness



P(i_missing | j_not_missing)

Feature substitution

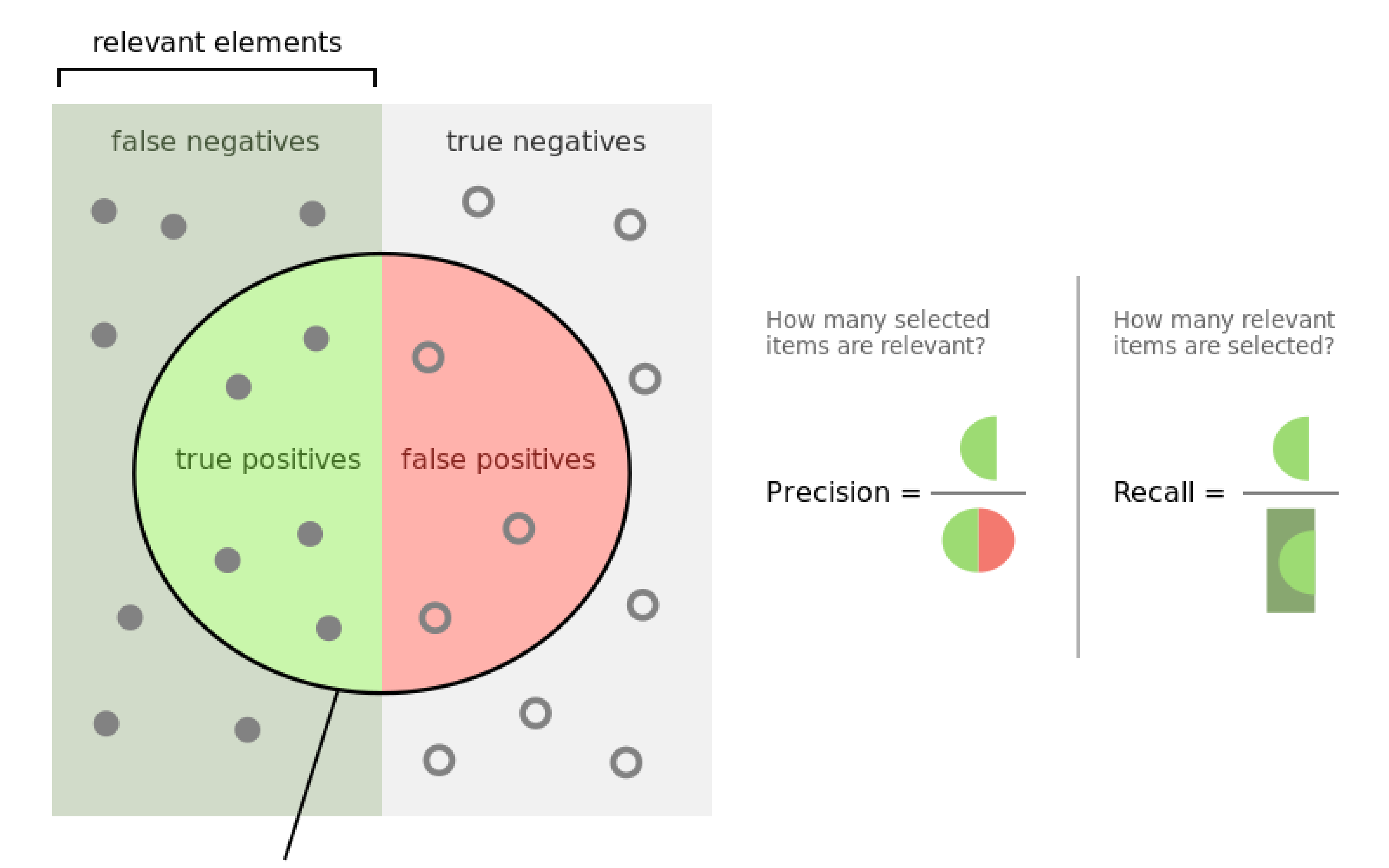


Feature pairs with PCC above 0.3

Existing methods for missing data imputation

- **Baseline**
- **Strawman imputation (mean or median)**
- **On-the-fly-imputation method**
- **Missingness incorporated in attributes**
- MissForest
- Surrogate splits
- ...

Evaluation metrics



Precision and recall, **source:** http://wikipedia.org

Experiments with random forests

- Number of trees: 100
- Minimal number of samples for a split: 2
- Maximal number of features for a split: sqrt
- Maximal depth of trees: unlimited
- Trained on data from three days in January 2017
- Tested on data from one day in March 2017
- Randomness factor: 1% of variance in recall and precision

Results

Method	Precision	Recall	Prec = 1.0	Prec > 0.8	Prec > 0.5
Baseline	0.61	0.57	22	54	70
Mean	0.59	0.54	21	54	70
Median	0.56	0.49	19	45	65
OTFI	0.23	0.06	18	25	25
MIA	0.65	0.58	28	60	74

Average precision, recall, and number of classes with precision above a certain threshold

Contributions

- Correlation of datasets features analysed
- Algorithms compared on real data
- On-the-fly-imputation found not suited for data with heavy missingness
- Missingness incorporated in attributes slightly improves the baseline method
- Python framework for further experiments implemented

Answers

Method speed comparison

- Baseline: ~18 hours
- Strawman: ~18 hours
- MIA: ~45 hours
- OTFI: ~100 hours