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

Handling Missing Values in Decision Forests in the Encrypted Network Traffic

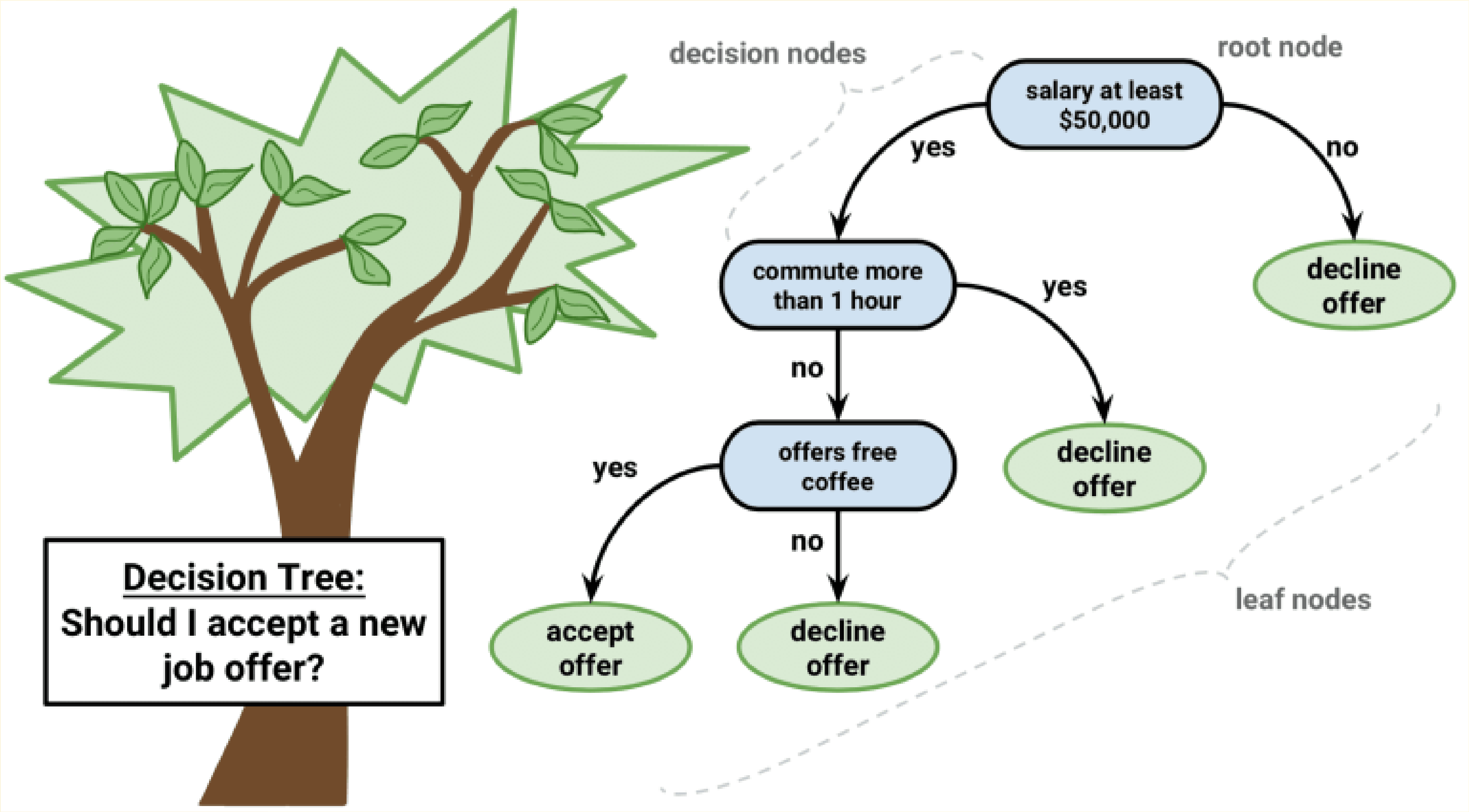
Author: **Lukáš Sahula**
Supervisor: **Ing. Jan Brabec**
Bachelor thesis

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

Handling missing values...

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

… in Decision Forests …

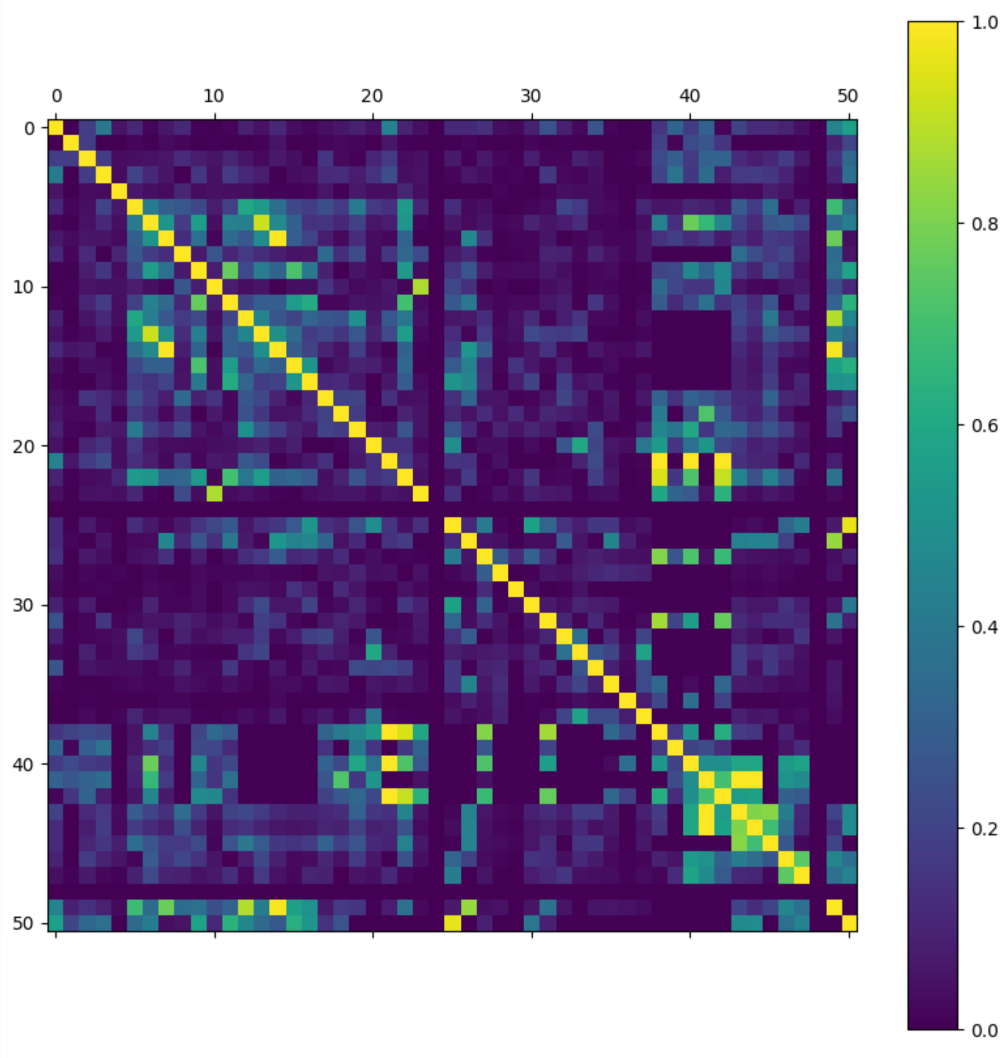


Decision Tree classifier, **source:** http://packtpub.com

… in the Encrypted Network Dataset

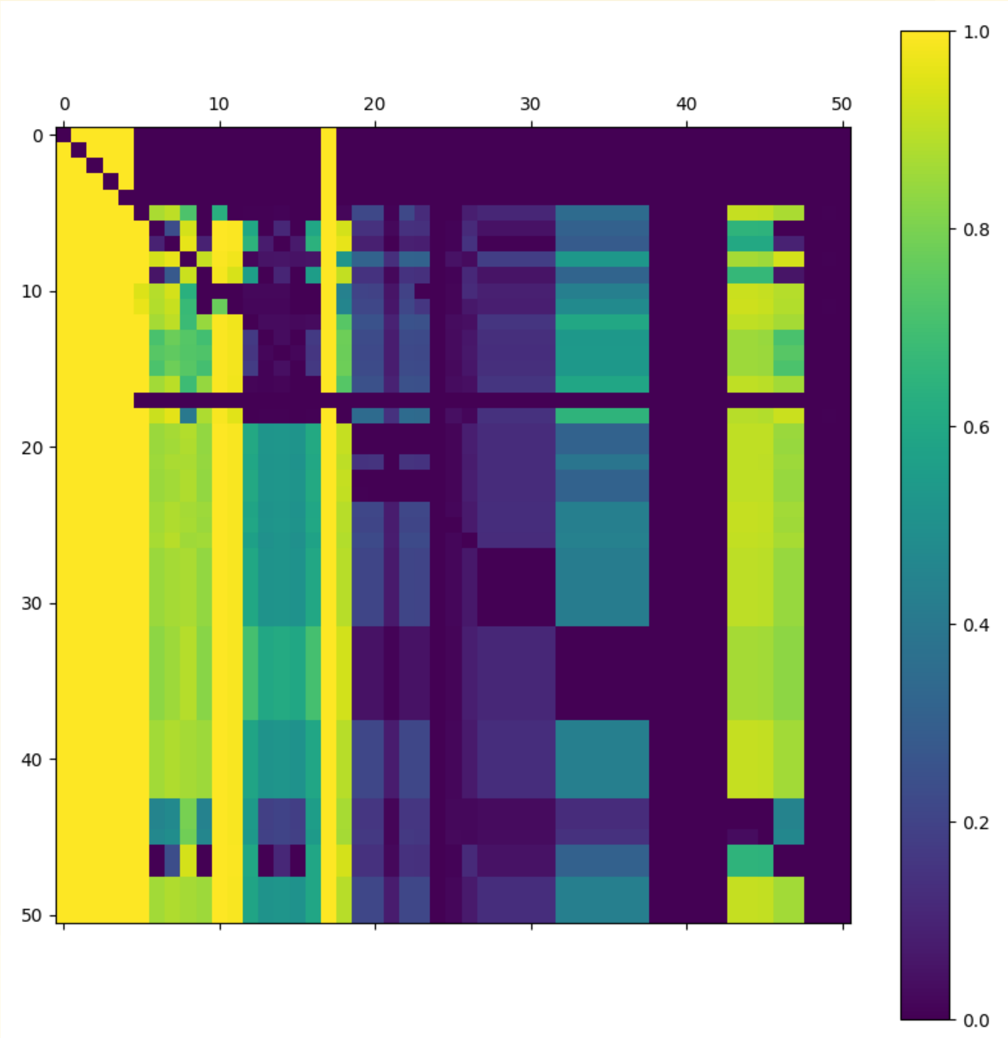
- Data from network proxy logs
- Detection and classification of malware
- Over 100 classes of malware
- 50 features
- Data missingness over 50%
- 600 million of records

Dataset correlation analysis



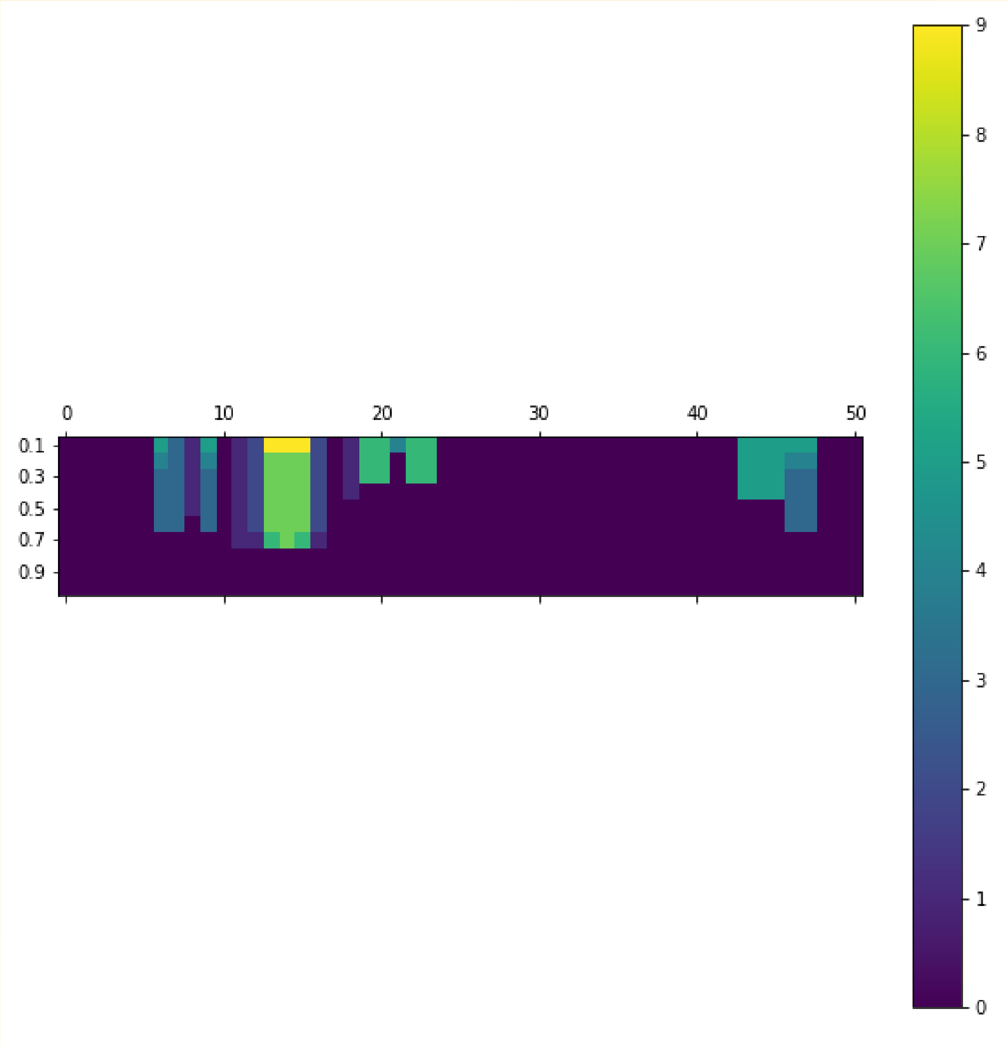
Heatmap of feature pairs correlations (Pearson)

Conditional probabilities of missingness



P(j_not_missing | i_missing)

Feature substitution



Number of features with PCC > 0.3 that can replace each feature

Existing methods for missing data imputation

- **Baseline**
- **Strawman imputation**
- **On-the-fly-imputation method**
- **Missingness incorporated in attributes**
- MissForest
- mForest
- Surrogate splits
- …

Baseline method

- Compute the best split with all the missing values replaced by a constant value smaller than all other values

Strawman imputation

- Impute the missing values using the mean or median value

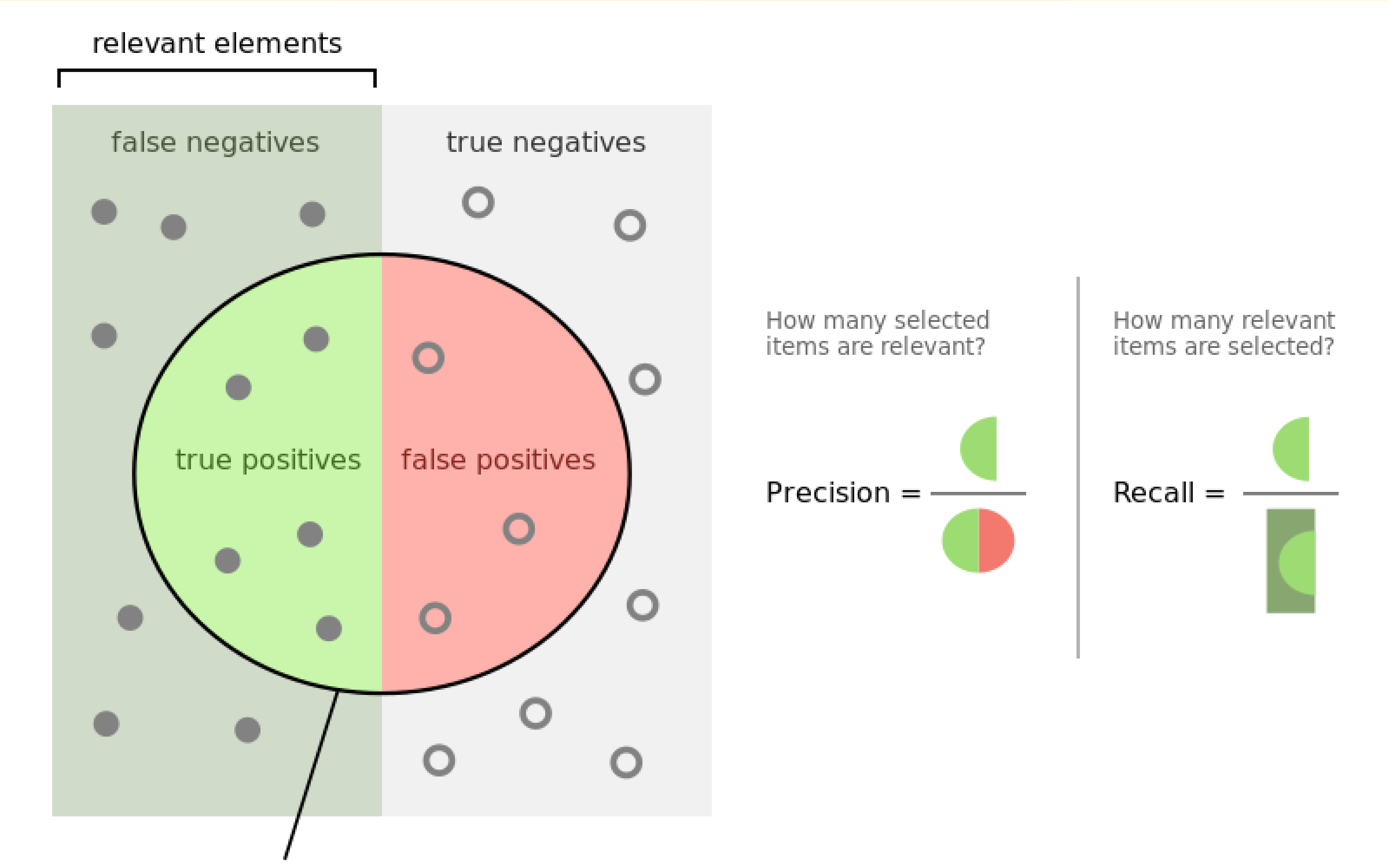
On-the-fly-imputation

- Impute the missing value using other values of the inbag data (at the current node) and their frequency

Missingness incorporated in attributes

- Similar to the baseline methods
- Compute the best split with all the missing values replaced by a constant value smaller than all other values first
- Compute the best split again with missing values replaced by a constant value **bigger** than all other values
- Compute the best split treating all missing values as -1 and all non-missing values as 1
- Of these 3 splits, choose the one with the biggest information gain

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

Answers

Method speed comparison

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

Answers

Scaling

- **Most of the algorithms do not scale well with bigger amounts of data missing or they run very slow on big datasets, so only the relevant were implemented.**
- Most of the algorithms perform worse as the amount of missing data increases or they run very slow on big datasets, so only the relevant were implemented.

Answers

Wrong entropy equation

$$\mathrm{H}(\mathrm{X}) = -\sum_{c \in \mathbf{C}}^{n} p(c) \log p(c)$$

Wrong

$$\mathrm{H}(\mathrm{X}) = -\sum_{c \in \mathbf{C}} p(c) \log p(c)$$

Correct