**DATA PREPARATION**

**i) Missing Values** (different datasets with different approaches to imputation of missing values):

Step i.1

– Everything over 50% . Then calculate correlation of the rest / metric from Übung

– All observations with missing values are rejected

–

Step i.2 Imputation

– Mean / Median

– Advanced Methods (Random forest?, nearest neighbor?,

Team: Phine und Marius

**ii) Noise (weird values - mistakes) & iii) Outliers and Noise**

– Observations having z-score bigger than 3 (or  2.5) - function to apply it to all of the columns.

– Histogram / boxplot

– Random forest outlier measure?

How to handle outliers? Slide 19 —> Replace the outlier with the maximum value

Variables?

 – Age - zwischen 17 - 75?

Team: Nikoleta und Marco

**iv) Variable reduction**

Step 1) Filter approach

– Correlation / Information Value (threshold for IV predictiveness - slide 45)

– WOE ( Übung)

Team: Gabriele

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Step 2) Wrapping approach

– Forward selection (logistische regression?)

– LASSO regression?

– Random Forest Variable Importance (Übung) - Partial Dependence Plot

**v) Variable transformation for modeling?**

–

**MODELS**

**i) Classification Trees**

**ii)  Logistic Regression**

**iii) SVM**

**iv) Artificial Neural Networks**

**v) Random Forests**

**vi) Naive Bayes**

**vii) Ensemble**

**Random**

– Horizontal reduction - sampling cases

– Parallelisierung R  (Ramona)

– Korbi  Lösungen Fragen (Marco)

– 5-Crossvalidation vs 2-Crossvalidation & Crossvalidation Variable Selection - Modelling

– Principal Component Analysis