

Applied Analytical Data Science

- Übung 2 -

Dr. Jörg-Uwe Kietz,
Übung zur Vorlesung,
Mittwoch, 14-16 Uhr Vorlesung,
16-18 Uhr Übung (alle 2 Wochen)

<http://www.kietz.ch/AADS/>

Vorhanden von der letzten Übung

- 1) RapidMiner installiert.
- 2) Daten importiert
- 3) Business Ziele unserer Aufgabe
- 4) Business Erfolgskriterium
- 5) Data Mining Aufgabe
- 6) Data Mining Erfolgskriterium
- 7) Vorhandene Daten
- 8) Relevanz der Daten für die Aufgabe
- 9) Qualität der Daten

2. Aufgabe: Daten Qualität verbessern

Role	Name	Type	Missings
regular	WEALTH1	integer	44672
regular	WEALTH2	integer	43764
regular	AGEINYEARS	integer	23634
regular	INCOME	integer	21255
regular	HOMEOWNR	binominal	20803
regular	TIMELAG	integer	9954
regular	NEXTDATE_Y	real	9954
regular	GENDER	binominal	5035
regular	CLUSTER	integer	2315
regular	DOMAIN1	polynomial	2315
regular	DOMAIN2	integer	2315
regular	LASTGIFT_Y	real	837
regular	GEOCODE2	polynomial	186
regular	MDMAUD_R	binominal	185
regular	MSA	integer	131
regular	ADI	integer	131
regular	DMA	integer	131
regular	CLUSTER2	integer	131
regular	MDMAUD_F	binominal	107
regular	MDMAUD_A	binominal	45
regular	NOEXCH	integer	42
regular	FISTDATE_Y	real	2

- Die Daten haben viele fehlende Werte.
 - Viele DM-Tools haben Probleme mit fehlenden Werten
- ⇒ Schätzt die fehlenden Werte
- ⇒ Benutzt die in der Vorlesung besprochenen Methoden (Teil 3, Seite 14-20)
- ⇒ Evaluiert eure Methoden auf den Eval-daten

Die einfache Lösung in RM

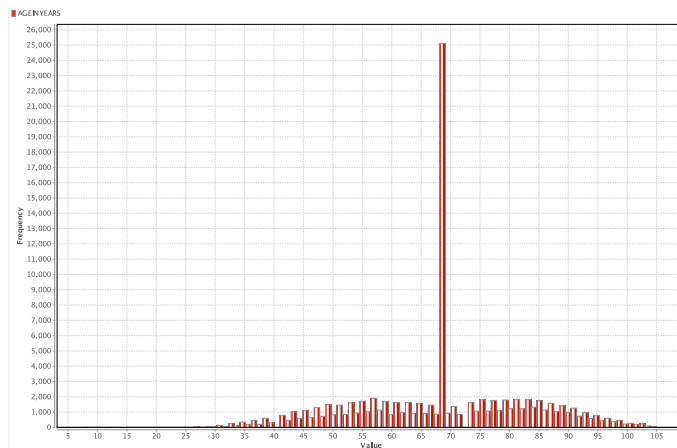
The screenshot displays the Rapid Miner software interface. On the left, a tree view shows the 'Operators' pane with 'Data Cleansing' expanded to 'Replace Missing Values'. The main workspace shows a 'Main Process' diagram with a 'Retrieve' operator connected to a 'Replace Missing Values' operator. The configuration panel for 'Replace Missing Values' is open on the right, showing the following settings:

- attribute filter type: all
- invert selection
- include special attributes
- default: average
- 2 hidden expert parameters
- Compatibility level: 5.2.008

Below the configuration panel, there is a 'Synopsis' section stating 'Replaces missing values in examples.' and a 'Description' section explaining that missing values are replaced by functions like 'minimum', 'maximum', 'average', and 'none'.

Die einfache Lösung hat ein Problem

AGEINYEARS nach Fill Mean Value:



Vortrag jeder Gruppe in der nächsten Übung

- Welche Attribute habt ihr warum gewählt?
- Was habt ihr mit ihnen probiert?
- Wie gut bzw. wieviel besser war es als die “Einfache Methode”

- Mindestens einfache Lösung, so dass alle Daten gefüllt sind
- Auserdem:
 - 1 mal numerisch per regression gefüllt und evaluiert
 - 1 mal categorial per classification gefüllt und evaluiert

Abgabe der 2. Übung

- Vortrag am **28.03.18** (Mittwoch in der Übung)
- RM Process, Modelle und Resultate (Report, Graphs, Validation) via Email bis **26.03.18**(Montag abend)
Bitte einen gezippten Folder namens GrXUeb2 mailen.