Herta-Projektbeleg SoSe 2015 Richard Remus, Claus Holland

Anforderung: 15-20 Seiten, inkl Abbildungen, Schriftgröße 12-13, Zeilenabstand wissenschaftlich, also 1.5

Aufgaben:

1. Einen geeigneten Datensatz wählen
2. Explorative Datenanalyse durchführen
3. Data Mining/Data Wrangling (Clean, Transform, Merge, Reshape)
4. Feature Engineering
5. Klassifikatoren testen, mehrere Klassifikatoren ausprobieren und die besten bezüglich einer bestimmten Fragestellung bewerten
6. Evaluation der Ergebnisse (positiv, negativ, Verbesserungsmöglichkeiten)

zu 1.: Vorschlag Datensatz Kreditwürdigkeit

https://www.kaggle.com/c/GiveMeSomeCredit

- Ziel: Wahrscheinlichkeit von Zahlungsproblemen bei der Bedienung eines Kredits feststellen ("financial distress")
- Analyse anhand von Kreditnehmer-Datensätzen
- Vorhersage, welche "typischen" Personen sich mit hoher Wahrscheinlichkeit zu Problemfällen für Banken (Kreditgeber) entwickeln
- man kann also viele Mekrmale betarchten, unsinnge aussortieren
- hoher Datensatz von 250,000 Menschen, also Potenzial fehlerhafte Datensätze zu entfernen, ohne zu viel "Rauschen" in die Analyse zu bringen
- Daten sind gut aufbereitet

zu 2.:

- Lerne den Datensatz kennen
- Ermittlung wichtiger Werte
- Ermittlung irrelevanter Werte (kaum, da die gegebenen Werte sich bereits auf finanzielle Abhängigkeiten beziehen)
- verschiedene Visualisierungen (Histogramme, Boxplot... PANDAS-Plots)

Ideen:

ThinkStats2 Nist Handbook

The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. conclusions as to whether individual factors are statistically significant
8. optimal settings

Mögliche ToDos:

- Wie ist die Altersverteilung beim financial distress (dictionary: Zeile 6,10,12)
- Welches Einkommen haben die schlimmen Schuldner (Ist es ein arme Leute Problem, oder ein Mittelstandsproblem?, Auch reiche Leute könnten versuchen, sich aus Krediten mit Hilfe ihrer Anwälte herauszuklagen....)
- Wo gibt es viele fehlende Angaben?
- Welche von den 3 genantne financial distresses (Zeile 6,10,12) komm am häufigsten vor, bereitet den Banken am meisten Probleme?
- Sind die Leute häufig mit ihren Raten im Rückstand? Oder einmal und nie wieder? wieder?
- Was passiert, wenn wir einzelne Werte modifizieren? (Geht dann schon in richtung Data Cleaning)
- Gibt es (schienbar) sehr aussagekräftige Merkmale?
- Gibt es Aussreißer?
- Gibt es Muster?

Geeignete Darstellungsformen:

http://www.itl.nist.gov/div898/handbook/eda/section1/eda17.htm

Als nicht finanzielle Variablen haben wir vor allem Alter und Anzahl von Haushaltsmitgliedern. Was steht nicht in dne Daten, was wir aber eigentlich gebrauchen könnten und somit die Zahl möglicher Hypothesen einschränkt? Müssen wir bereits für die explorative Analyse Metriken anpassen und umrechnen, damit die Plotter damit etwas anfangen können?

zu 3.: Verwenden von Pandas

- abhängig von den möglichen Klassifizierungsalgorithmen
- z.B. NAN-Werte umwandlen in mean ode rmedian (Vorsicht: Datenverfälschung)
- fhelerhafte Datensätze kann man auch entfernen, wenn man genügend besitzt

zu5.:

- bei der Verwendung eines Decision-Trees mehrere Merkmale untersuchen

zu6.:

https://moodle.htw-berlin.de/pluginfile.php/216933/mod_resource/content/1/kontingenz-tabellen.pdf