Protokoll 14.12.15

Vorgang:

KDD Prozess, wir gehen mit diesem vorran! FELIX, notizen vom ersten mal

Frederik: explorative Datenanalyse

Verteilung der Klassen

Alle Numerische Histogramme gezeigt, Anzeige von Outliern ganz am ende sowie Boxplots

Die codierung von Default werten, wie wurden diese Behandelt (verschiedene daten sets erstellt z.B nur numerische Var.)

Correlation angeschaut (corplot)

2. Preprocessing

Missing Value handling

N/A Methode is N/A, das auch alle diese N/A haben Funktion geschrieben im Utils die Methode geschrieben, setNA.(z.B bei alter anstatt 0 gleich NA), Categroischen Haben 00 als default. Um diese besser (bewusster) zu behandeln. Dadruch ALLE default werte zu NA gesetzt. Darauf auf bauend missing value Handling

Kommentar( versagt bei zu viel default values), deshalb methode geschrieben, EXCLUDESPrseData (dünnbesetzte Matrix). Wir schmeißen diese Variablen raus. Nur Variavlen mit unter x (z.z 10% werden berücksichtigt).

* Fertig vorbereitet für das IMputing (Missing value handling)

Dann hässlich programmieren, immer noch 130 spalten; Wir konnten den kompletten dataframe übergeben. Wir konnten bis 113 übergegen WHY? Kein Ergebnis (warum) deshalb alle einzeln

Als schnittstelle funktion getImputeddata, für numerische und kategorische

Outlier detection

1D , wie kann man rausfinden ob var. Outlier; wenn Mean und Median stark auseinander liegen; complete cases. Wo liegt mean und Median auseinander, wo der Beträg größer als der halbe median ist.

Outlier detection with Alex, take care of each variable: Summary the observations!.

Dann die Funktion, die 2 quantile bestimmt (025%, Antenne Wikness, 1,5 abstand der Quantile: PROBLEM! Keine correlation oder clustering .. besprochen 50% etc.

* + 1. Iteration Missing value and Boxplot detection outlier betrachtet keine Correlation aber wir akzeptieren.!!!

Wie setzten die Extremwerte auf die oberen antennen und nicht die auf den Median, WARUM? Aufgrund das sich der Median dadurch nicht so stark verändert und es keine ausreißer mehr gibt.

Wir versuchen zu untersuchen alle Variablen auf ihre Extremwerte, um diese auf eventuele atificial (messfehler) zu untersuchen und evtl auf den Median zu setzten.

Daten outlier alle Variablen angeschaut; rev\_mean ttotmrc\_mean, beide min unter 0; nicht sinnvoll da charge, entdeckt aber als unter quantil outlier genommen. Trotzdem bemerkt!!!