Team: Phine und Marius

Offene Fragen

1. Erst horizontales oder vertikales löschen von fehlenden Werten?
2. Na hand\_webcap ist kein Problem. Wird als leeres Feld eingelesen.
3. RETDAYS -> NA zu 0? Nein, weil falsch.
4. AGE1 und AGE2 verstehen
5. Manche Categorielle sind ordinal manche nominal. Wie geht R damit um?
6. Unkown zu Missing ->Check
7. Hndset-price ist numerisch
8. Kid6\_10 mit anderen kid variablen mergen zu Kind Y/N?

Aufgaben

1. Age1 und Age2 sind metrisch und enthalten Wert 0 -> sollte zu Missing werden? Oder doch selbständige Information? CHECK! ->ABER ERSTELLE NEUE VARIABLE DIE KATEGORIELL IST: MISSING Y/N
2. Check ob Daten bei educ1 in ordinaler Reihenfolge sind
3. Hnd\_price: Odinal oder Metrisch?
4. Last swap ist Datum ->Muss konvertiert werden, zudem Missing ist hier kein Swap: Wie damit umgehen? Check- zwei Variablen gemacht, eine numerisch eine factor (binning

Aber - > Prolbem: bei Binning 3 Missings mehr als ohne Binning

1. Ref\_qty: Was ist referrel? Konvertiere Missing zu 0. CHECK
2. Tot\_acpt: Konvertieren zu Binary?
3. Uniqsubs: Konvertieren zu Binary?
4. Retdays (cont.): Konvertieren zu Binary?

Alter wird jetzt als continuous eingelesen

\_------------------------------

Marco & Nikoleta

Chi-Quadrat Test für Ausreißer

Was ist mit Ausreißern

Range ist hoch korelliert mit Mean: Guck welche am besten ist für Churn-Prediction und hau die anderen raus

CHECK FÜR TEST DATEN AUF AUSREIßER; FEHLER, MISSINGS

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

Gabriel: Stepwise und Korrelation um Variablen zu Minimieren

Marius: Parallelisierung

Marius & Phine: Imputation ->

Marius & Phine: Entfernte Variablen checken und hinzufügen, wenn Gabriel Signal, dass es wir sie brauchen

Lift Score, AUROC, …

Sonntag 12:30

Faustregeln für die Dauer der Berechnung in Abhängigkeit von Prozessor, Model, Variablen und Beobachtungen:

1. Outlier Detection: Man muss schauen, ob Model besser wird oder nicht. Nicht so viel Zeit in Ausreißer investieren. Random Forest kann hierfür auch verwendet werden. Rechne dann einmal mit und ohne Ausreißer und Vergleiche die Performance. Entscheidungsbaumverfahren sind robust gegen Ausreißer. Grundsätzlich auch ANN und SVM. Wirklich schlecht ist es nur bei linearen Verfahren.
2. Grenzen für NAs: Es ist immer heuristisch. Probiere verschiedene Grenzen und versuche, welcher am besten funktioniert. Grundsätzlich ist es aber vollkommen in Ordnung einfach Variablen zu entfernen. Allerdings kann die Information als Dummy-Variable im Model enthalten bleiben. Um bei den Tests nicht zu viel Rechenzeit zu verlieren, kann der Datensatz verringert werden.
3. Information basiert auf erster Kategorisierung, Korrelation aber nicht.
4. Variablenanzahl im Modell: Dies wird immer empirisch gemacht und es gibt keine Faustregel. Dies steht im Gegensatz zu der üblichen linearen Regression.
5. Imputation: Ist Mean oder Median ausreichend?: Immer sollte damit angefangen werden. Wenn eine Variablen großen Erklärungsgehalt hat, sollten bessere Imputationsmethoden verwendet werden. Empirische Evidenz zeigt, dass es nicht viel bringt, für alle komplizierte Imputationsverfahren zu nehmen.
6. WOE: Alles Kategoriell Information Value, für numerische Korrelation. Diese werden dann miteinander verrechnet. WOE kann berechnet werden indem numerische Variablen mit Equal-Size gebint werden. Niemals Equal-Width! Oder erst gute numerische, dann gute kategorielle und dann diese bei der Modelperformance zu werden.
7. Forward/Backward Selection: Backward funktioniert nicht, weil alle korrelierte Daten drin sind. Lineare SVM, R-Pkg von liplinear. Das ist sehr schnell und super für eine Forwardselection. Alternative ist logistische Regression, weil es auch schnell ist. Ein Random Forest ist zu langsam.
8. Korrelationsmatrix mit abhängiger Variable und lösche niedrige Korrelation. Faustregel Wikipedia?
9. Korrelationsmatrix unabhängiger Variablen und lösche Variablen die untereinander eine Korrelation nahe |1| haben.
10. Wrapper Approach mit Forward Selection oder linear SVM. Siehe Punkt 7 oben.
11. Maß für Korrelation Kategorielle Variablen. Bei Binärer funktioniert das mit Pearson. Bei Ordinalen mit Rangkorrelation.
12. Einmal alles mit kategoriellen und einmal mit alles numerisch.
13. **Forward Selectin mit LipLinear für alle Variablen.**

**Marco: Random Forest Selection**

**Gabriela: Korrelation und WOE**

**Forward Selection: Marius und Phine**