Die Dokumentation vom Notebook „previous application”

**Diese Tabelle enthält die statischen Daten in Bezug auf Kunden und ihre früheren Kredite bei der Home Credit Group.**

Liste der Merkmale:

Der Datensatz enthält 1670214 Reihen und 37 Spalten. Insgesamt wurden 37 Spalten aus der Tabelle „previous application“ abgelesen. Nach der Prüfung von mehr als 40 % fehlenden Werten wurden 11 Spalten rausgeworfen, da es zu viel eingesetzt werden musste.

Die folgenden Spalten: (RATE\_INTEREST\_PRIVILEGED, RATE\_INTEREST\_PRIMARY, AMT\_DOWN\_PAYMENT, RATE\_DOWN\_PAYMENT, NAME\_TYPE\_SUITE, NFLAG\_INSURED\_ON\_APPROVAL, DAYS\_TERMINATION, DAYS\_LAST\_DUE, DAYS\_LAST\_DUE\_1ST\_VERSION, DAYS\_FIRST\_DUE, DAYS\_FIRST\_DRAWING).

Dann wurden zwei weitere Spalten entworfen, die als irrelevant betrachtet wurden, und zwar der Wochentag der Anmeldung vom Antrag und die Uhrzeit der Anmeldung vom Antrag. Danach wurde die Reihen für die Duplikate geprüft, wobei mit 1331357 Duplikaten von vorheriger Antragstellung gefunden wurden. Weiterhin wurden die Duplikate rausgeworfen und mit 338857 Reihen und 24 Spalten mit der Tabelle weitergegangen.

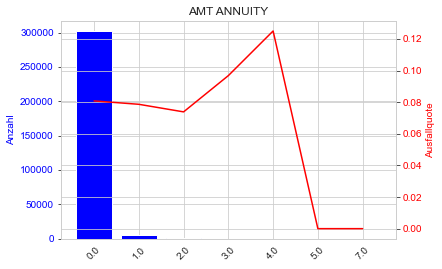
Unter der Annahme, dass monatliche Rückgabe und anfängliche Anzahl der Raten zur Tilgung von Schulden in hohem Maße miteinander verbunden sind, da CNT PAYMENT möglicherweise vollständig durch AMT ANNUITY beschrieben werden kann, werde ich die CNT PAYMENT-Variable entfernen. Anschließend wurde Spalte für Spalte durchgegangen.

Anstatt von NaN-Werten wurden oft die Werten mit XNA und XPA gesehen, die dann den Inhalt falsch repräsentieren, ausgehend davon, wie viele XPA und XPA vorhanden sind, wurden die Spalten entweder entworfen oder mit dem oft vorkommenden Wert eingesetzt. Spalten, die größten Anteil von XPA - und XNA – Werten hatten, waren - gruppierter Zinssatz (Name yield group), die Industrie der Verkäufer, der Name vom Portfolio, Waren von vorherigem Antrag, die Ursache von der vorherigen Absage, die Bezahlungsart, Zweck des Bardarlehens und die Art vom Produkt, insgesamt 8 Spalten. Als nächste sind die Werten mit niedrigem Anteil von XNA und die wurden mit häufigstem vorkommendem Wert ersetzt. Ein anderes Merkmal war, dass das Kennzeichnen, wenn es sich um die letzte Bewerbung für den vorherigen Vertrag handelte und wenn die Anwendung die letzte Anwendung pro Tag des Kunden ging, hat 99,7% als „ja“ Zustimmung, daher hätten diese Variablen keinen großen Einfluss gezeigt. Daher wurden die zwei „FLAG“ Spalten entfernt. Die Tage von der Entscheidung sind als negative Werten angegeben und um die Rechnung weiter zu erleichtern, wurden die negative Werten zu positiven Werten umgewandelt. Am Ende der Reinigung sind insgesamt 338857 Reihen und 16 Spalten.

Die erste Reinigung wurde erfolgreich ohne NaN - oder unlogische Werten und alle Null-Werten durchgeführt. Im Anschluss wurden die Spalten in numerischen und kategorischen Variablen gruppiert. Die kategorischen Variablen wurden mittels Dummies in viele neue Spalten geteilt, die folgenden Spalten: der Type vom Vertrag, der Status vom Vertrag, der Type vom Kunde, der Type vom Kanal und die Kombination vom Produkt. Da zu viele Spalten erzeugt wurden, wurden die kategorischen Variablen in den einzelnen Datensatz als df.csv gespeichert. \*Zu EDA

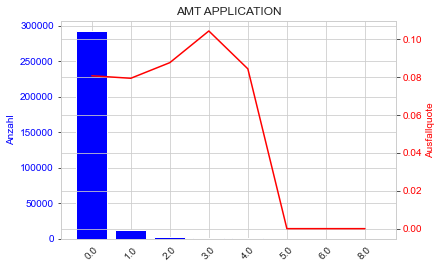
Die numerischen Variablen sind insgesamt 6 und der Bereich vom Verkaufsort hat zu wenig Information und daher wurde entworfen. Folgend wurden einige Statistiken für jede Gruppe extrahiert. Die monatliche Bezahlung wurde zu dem Durchschnitt gruppiert und schließlich mit den anderen Gruppierungen von numerischen Variablen, die summiert wurden, aggregiert. Der vorherige/frühere ID ist schließlich raus und die Variablen basierend auf dem aktuellen ID wurden erzeugt. Die Variablen wurden unter dem agg\_df.csv abgespeichert.

Durch die deskriptiven Statistiken war es auffällig, dass die Minimal- und Maximalwerte in unseren numerischen Spalten weit auseinander lagen. Diese "extreme" Werte wurden anhand der Boxplots und Histogrammen angesehen. Da es um den Ausfall vom Kredit geht, können diese "extreme" Werte eine wichtige Rolle spielen. Der genaue Anteil von den Outliers wurde mittels einer Funktion berechnet.

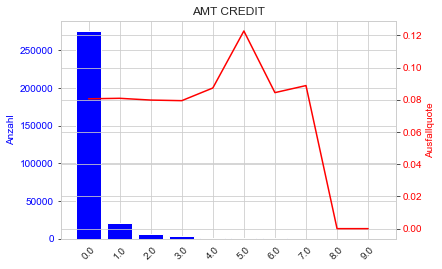
Visualisierung:

*Abbildung 1: Die durchschnittliche Rückzahlung im Zusammenhang mit der Ausfallquote*

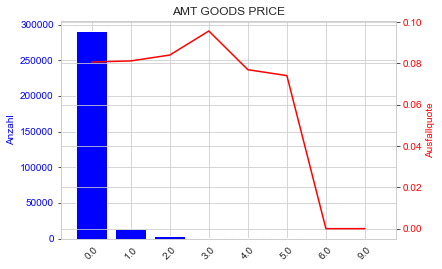
In der Abbildung 1 ist die durchschnittliche Bezahlung im Zusammenhang mit dem Ausfall gezeigt.



*Abbildung 2: Die Summe vom gefragten Anteil des Kredits im Zusammenhang mit der Ausfallquote*



*Abbildung 3: Die Summe vom gegebenen Anteil des Kredits im Zusammenhang mit der Ausfallquote*



*Abbildung 4: Die Summe vom Anteil des Waren Prices im Zusammenhang mit der Ausfallquote*

In der Abbildung 3

In der Abbildung 4

AMT\_ANNUITY

AMT\_APPLICATION

AMT\_CREDIT

AMT\_GOODS\_PRICE

DAYS\_DECISION

SELLERPLACE\_AREA