Individual Assignment 1 – NYC Cabs!

New York. Die Stadt die niemals schläft. Unter Tags ist New York eine der geschäftigsten Städte der Welt. Viele Touristen und ausländische Geschäftsleute nutz(t)en Taxis als Transportmittel, um von Termin zu Termin zu fahren. Auch wenn die gelben Taxis fester Bestandteil der New Yorker Stadtkulisse sind, haben sich in den letzten Jahren viele Konkurrenten fest im Markt etablieren können. Anbieter wie z.B. Uber bieten oftmals nicht nur einen günstigeren Preis. Mit ihrer App bieten sie Kunden oftmals auch ein einfacheres digitales Mobilitätserlebnis. Aus diesem Grund haben sich die verschiedenen Taxi-Unternehmen von New York zusammengeschlossen und eine App in Auftrag gegeben, mit der Taxis bestellt werden können und die für potentielle Fahrten eine Vorhersage über den wahrscheinlich zu entrichtenden Preis ausgibt.

Sie wurden beauftrag einen entsprechenden Algorithmus für die Taxen in NY City zu entwickeln. Für das Assignment 1 werden originale Daten der historischen Taxifahrten in NY City verwendet. Aufgrund der Grösse des Datensatzes werden ca. 0.2 % der Daten aus dem Januar 2016 verwendet. Zudem haben Sie sich entschlossen auch Wetterdaten für NY City sowie Informationen über die amerikanischen Feiertage in Ihr Model mit einfliessen zu lassen.

- Total_amount im Datensatz „rides.csv“: Vom Fahrgast zu entrichtender Fahrpreis.
- Alle anderen Variablen und Datensätze sollten weitgehend selbsterklärend sein.

Bitte erstellen Sie ein R Skript und beantworten Sie die folgenden Fragen:

Verschaffen Sie sich einen ersten Überblick über die Daten und bereiten Sie diese für die spätere Analyse vor:

1) Lesen Sie die Daten ein. Die Variablen sollten als Integer oder Numeric erkannt werden. Einige Variablen könnten als „character“ oder bereits als „factor“ eingelesen worden sein. Die Interpretation von Faktoren in Regressionsmodellen werden wir erst in der nächsten Woche lernen – entfernen Sie diese Variablen daher aus dem Datensatz
   Hinweis: Vergessen Sie nicht das Arbeitsverzeichnis in R zu setzen.
2) Wie viele Variablen und Instanzen befinden sich im Datensatz? Handelt es sich bei dem Datensatz um eine Matrix oder ein Dataframe? Was ist überhaupt der Unterschied?
3) Berechnen Sie den Mittelwert, Median und die Standardabweichung der Variable „Total_Amount“. Was ist der Unterschied zwischen den drei Werten?
   Hinweis: Drücken Sie die konzeptionellen Unterschiede zwischen diesen Grössen in ihren eigenen Worten aus.
4) Erstellen Sie ein Histogramm der Variable „Total_Amount“? Wie würde Sie die Verteilung der Daten beschreiben?
5) Mit welchen Variablen korreliert „Total_Amount“ am stärksten? Könnte diese hohe Korrelation problematisch sein für ein späteres Regressionsmodell? Wenn ja, löschen Sie eine der entsprechenden Variablen.
6) Erstellen Sie einen Scatterplot aus den Variablen „dropoff_latitude“ und „dropoff_longitude“. Was können Sie auf dem Plot erkennen? Es scheint als hätten sich ein paar Datenverarbeitungsfehler bei der Aufnahme der GPS-Koordinaten gegeben – Erstellen Sie ein neues data.frame, in dem Sie die fehlerhaften Werte ausschließen. Was können Sie nun auf dem verbesserten Scatterplot erkennen?
7) Wählen Sie die Instanzen 1-3000 des Datensatzes und speichern Sie diese als Trainingsdaten auf der Festplatte ab. Wiederholen Sie dies für die restlichen Instanzen und speichern Sie

diese als Testdaten auf der Festplatte ab. Achten Sie darauf, dass keine Zeilennummern mit abgespeichert werden. Lesen Sie die beiden Datensätze in R ein und geben Sie ihnen aussagekräftige Namen.

8) Trainieren Sie auf den Trainingsdaten mittels einer linearen Regression ein Vorhersagemodell, mit dem Sie den erzielbaren „Total_Amount“ einer jeden Taxi-Fahrt vorhersagen können.
9) Was sind die wichtigsten Variablen in Ihrem Vorhersagemodell? Warum? Berechnen Sie den RMSE in den Trainings- und Testdaten und vergleichen Sie diese. Wie würden Sie die Ergebnisse interpretieren?
10) Wie hoch ist der R2 und was ist dessen unterschied zum RMSE?
11) Erstellen Sie einen Actual-vs-Predicted-Plot in den Test- und Trainingsdaten und vergleichen Sie diese!
12) Kann die prognostische Qualität des Modells verbessert werden, wenn nicht signifikante Variablen aus dem Modell entfernt werden?

Optional (nicht Teil der Bewertung):

Bei der Spezifikation von Modellen ergibt sich oftmals ein Trade-off zwischen der Modellkomplexität (Anzahl Variablen) und Genauigkeit. Okkham´s Rasiermesser besagt, dass daher immer das „Prinzip der Sparsamkeit“ angewendet werden sollte und nur die unbedingt notwendigen Variablen in das Modell aufgenommen werden sollten.

R bietet eine Funktion mit der das Ausprobieren unterschiedlicher Variablenkombinationen automatisiert werden kann: step(). Der angesprochene Trade-off wird hier mittels des Aikaike Information Criterion (AIC) formalisiert. Das AIC beschreibt die Qualität eines Models, welche mit einem „Strafterm“ für die Anzahl der Variablen belegt wird. Je kleiner das AIC, desto besser das Modell.

Mittels der step()-Funktion kann das „sparsamste“ Modell einfach berechnet werden. Der einzige Input ist ein abgespeichertes (lineares) Modell. Der Output ist ein vereinfachtes Modell. Wenn das Ergebnis der Regressionsanalyse im Model *price.reg* abgespeichert wurde, können wir das Modell mittels step(price.reg) vereinfachen. Das vereinfachte Modell kann wie ein Regressionsmodell in R abgespeichert werden (step.model=step(price.reg))