Data Wrangling (daw)
HS 2021
Cédric Huwyler

Fachhochschule Nordwestschweiz

Prüfung

Nachname	Vorname

Hinweise

Datum, Zeit, Ort: 10. Februar 2022, 08:15 Uhr, Zimmer 5.2B51

Prüfungsdauer: 90 Minuten

Rahmenbedingungen

- Sie bestreiten die Prüfung auf Ihrem eigenen Computer und geben am Schluss diese Prüfungsblätter und ein begleitendes Jupyter-Notebook (an `cedric.huwyler@fhnw.ch`) ab.
- Benötigte Datensets werden vom Dozierenden zum Start der Prüfung zur Verfügung gestellt.
- Für die maximale Punktzahl wird ein sauber dokumentierter Lösungsweg im Notebook erwartet. Zusätzliche Erklärungen im Code sollen in Markdown erstellt werden.
- Die Prüfung ist Open-Book und Sie dürfen das Internet zur Informationssuche benutzen.
- Kommunikation mit anderen Studierenden oder aussenstehenden Personen während der Prüfung ist nicht erlaubt. Zuwiderhandlung zieht die Note 1 für alle Beteiligten mit sich.

Benotung

Aufgabe	1	2	Total	Note
Maximale Punktzahl	49	24	73	
Erreichte Punktzahl				

Die Note 6 ist bei 57 Punkten angesetzt.

Aufgabe 1. (49 Punkte)

Eine Firma die mit Oldtimern handelt möchte den An- und Verkauf von Autos optimieren und beschliesst sich, eine Data Science - Position auszuschreiben, für die Sie schliesslich eingestellt werden. Die Firma möchte mittelfristig ein System aufbauen, das den Wert von Oldtimern aufgrund verschiedener Parameter schätzen kann und so die durch den Ankauf und Verkauf von Fahrzeugen erzielten Gewinne optimiert. An Ihrem ersten Arbeitstag übergibt Ihnen Ihre Vorgesetzte ein Beispiel-Datenset mit Verkaufspreisen und weiteren Features, das der Informatiker für Sie zusammengestellt hat. Der Informatiker ist kein Data Scientist und schon länger im Dienst und hat entsprechend ein etwas 'altmodischeres' Format gewählt. Das Datenset kommt als Textdatei `cars.txt` daher und ist im Prüfungsmaterial zu finden.

a) **(6 Punkte)** Sichten Sie die Datei mit einem Texteditor und lesen Sie sie dann in ein Data Frame ein. Machen Sie das ganze ohne manuelle Arbeit mit dem Texteditor, sondern versuchen Sie die Spaltennamen ebenfalls mit Pandas einzulesen.

   Hinweis: Falls Sie es nicht schaffen, die Spaltennamen mit Pandas einzulesen, können Sie die Spaltennamen auch per Copy&Paste setzen, mit entsprechendem Punkteverlust natürlich.

b) **(13 Punkte)** Bereinigen Sie nun das Datenset. Stellen Sie dabei sicher, dass

   - fehlende Werte mit `NaN` als solche markiert sind,
   - jede Spalte den passenden Datentyp besitzt,
   - der Schreibfehler 'alfa romero' zu 'alfa romeo' korrigiert ist,
   - die Anzahl der Zylinder als Integer statt als String vorliegt,
   - alle diskreten, nicht-numerischen Spalten als Kategorien definiert sind (je nach Skalierung mit oder ohne Ordnung).

c) **(2 Punkte)** Quantifizieren Sie die prozentuale Anzahl der fehlenden Werte mit einem Barplot. Welche Spalte enthält am meisten fehlende Werte?

d) **(6 Punkte)** Sie verschaffen sich einen ersten Überblick über die Daten und haben sich dazu folgende drei Fragen überlegt, die Sie mit Box- oder Scatterplots grob (also ohne statistische Tests) beantworten möchten:

   1. Hat der Wertverlust prozentual vom Kaufpreis einen Zusammenhang mit der Automarke?
   2. Hat der Risikofaktor des Autos einen Zusammenhang mit dem Kaufpreis?
   3. Hat die Anzahl der Zylinder einen Zusammenhang mit den Pferdestärken?

e) **(6 Punkte)** Beantworten Sie ausserdem die folgenden Fragen mit entsprechenden Aggregationen:

   1. Welches ist das teuerste Auto mit Risikofaktor `+3`?
   2. Welche Automarke ist durchschnittlich am günstigsten?
   3. Welche Automarke bietet das Cabriolet (`convertible`) mit am meisten PS?

f) **(9 Punkte)** Der Preis (`price`) und der Wertverlust pro Jahr (`normalized-losses`) sind unentbehrliche Grössen zum Treffen einer Kaufentscheidung aufgrund des Typs und des Alters eines Autos. Leider fehlt ein beachtlicher Teil der Wertverluste. Für die Proof-of-Concept Studie verfolgen Sie zuerst einmal eine relativ einfache Imputationsstrategie, bevor Sie sich komplizierteren Strategien widmen:

Sie imputieren den Wertverlust mit dem Medianwert **pro Automarke** falls möglich, **sonst** mit dem insgesamten Medianwert vor der ersten Imputation, falls keine Wertverluste für eine Automarke verfügbar sind.

Implementieren Sie dafür eine Funktion `impute_medianloss( df )` die als Argument das Data Frame erhalten soll und die Spalte `normalized-losses` wo nötig mit imputierten Werten zurückgeben soll.

Die Funktion soll ausserdem die Anzahl der fehlenden Werte vor der ersten Imputation, nach der ersten Imputation und nach der zweiten Imputation per `print()` ausgeben. Der letzte Wert soll natürlich auf Null kommen.

g) **(7 Punkte)**

Um den Erfolg der gewählten Strategie zu evaluieren, haben Sie sich folgendes Setting überlegt: Sie erstellen zuerst eine geshuffelte Version des Data Frames (Zeilen zufällig neu sortiert), speichern dann die ersten 50 Werte von `normalized-losses` separat ab und setzen diese anschliessend auf `NaN`. Im Anschluss imputieren Sie mit der obigen Strategie und berechnen anschliessend die Wurzel des quadratisch aufsummierten Fehlers (RMSE) auf die tatsächlichen Werte:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_i (\hat{x}_i - x_i)^2},$$

wobei x_i den tatsächlichen Wert (den wir gespeichert haben) und $\hat{x}_i$ den imputierten Wert und n die Anzahl der Werte bezeichne. Nehmen Sie Werte, die schon vorher gefehlt haben aus der Rechnung, da dort keine tatsächlichen Werte verfügbar sind. Beurteilen Sie die Qualität dieser Strategie, in welchem Prozentbereich der tatsächlichen Werte liegt das RMSE ungefähr?

Hinweis:

- Vergessen Sie nicht, dass Slices aus Data Frames je nachdem Views bleiben, d.h. Veränderungen auf dem ursprünglichen Data Frame auch den Slice ändern. Machen Sie wo nötig eine Kopie des Objekts.
- Sie können die Evaluation auch mehrmals ausführen für verschiedene Shufflings, so erhalten Sie ein besseres Bild der Qualität und sind weniger vom Zufall abhängig.

Aufgabe 2. (24 Punkte)

Zu Beginn von Covid-19 wurde diskutiert, dass vielleicht einfach die 'Alten und Schwachen' mit dem Virus etwas früher 'wegsterben', die Sterblichkeit dadurch zwischenzeitlich etwas höher werde aber dafür im Nachhinein unter die Erwartung sinken könnte. Als Data Scientist möchten Sie diese Diskussion kurz mit Zahlen untermauern und zeigen, dass dies mindestens auf den ersten Blick nicht so erscheint.

a) **(2 Punkte)**

In der beiliegenden Excel-Datei `Sterblichkeit.xlsx` finden Sie die entsprechenden Zahlen zur Sterblichkeit des Bundesamts für Statistik. Lesen Sie die Datei in ein Data Frame ein.

b) **(9 Punkte)**

Bringen Sie alle Spalten in ein passendes Format. Insbesondere:

- Ganze Zahlen sollen wenn möglich als Integer vorliegen (sonst als Float).
- Das Datum 'endend' soll im Datumsformat vorliegen.
- Alle Zeichenketten sollen am Anfang und Ende keine Leerzeichen vorweisen.
- Fehlende Werte sollen entsprechend mit `NaN` gekennzeichnet werden.

c) **(4 Punkte)**

Stellen Sie die tatsächliche Anzahl der Todesfälle pro Woche, die erwartete Anzahl und deren untere und obere statistische Grenze in jeweils einem separaten Plot pro Altersklasse dar.

d) **(4 Punkte)**

Berechnen Sie nun für jede Woche die Differenz der registrierten Todesfälle zum erwarteten Wert und visualisieren Sie diese Differenz für die über-65-Jährigen pro Woche. Rechnen Sie die überdurchschnittlichen Todesfälle (Übersterblichkeit, alle Werte über der Nulllinie) und die 'eingesparten' Todesfälle (Untersterblichkeit, alle Werte unter der Nulllinie) über die ganze erfasste Zeit zusammen und vergleichen Sie.

e) **(5 Punkte)**

Berechnen Sie nun, wieviele Personen absolut und in Prozent in den Jahren 2020 und 2021 mehr gestorben sind als sonst. Hat sich der 'Covid19-Effekt' wieder aufgehoben und sind insgesamt etwa gleich viele Personen gestorben?

Erstellen Sie eine Tabelle der folgenden Struktur:

		Erwartung	Diff Absolut	Diff Prozent
Jahr	**Alter**			
2020	**0-64**			
	65+			
2021	**0-64**			
	65+			