Programmentwurf Data Science Prototyp

Gegeben ist ein Datensatz zum Thema „Cost of Living“ (Datei im Moodle). Die Beschreibung der Features ist unter Kaggle zu finden:
https://www.kaggle.com/datasets/mvieira101/global-cost-of-living/data

Szenario: Sie sind Remote Worker in der Kreativbranche und möchten für 6 Monate mit einem Freund / einer Freundin ins Ausland. Dabei möchten Sie fünf Städte identifizieren, die in Frage kommen, und dabei möglichst verschiedene Optionen betrachten. Gruppenzusatz:

G1: am Meer / Festland

G2: tropisch warm

G3: möglichst exotisch

G4: an einem Fluss

G5: an einem See

G6: in den Bergen

G7: möglichst abgelegen

G8: möglichst kühl / nicht zu heiß

G9: nur kleine Inseln

G10: nur Großstädte

1. Business Understanding (3 Punkte):

Formulieren Sie ein Ziel oder mehrere Ziele nach dem CRISP-DM Prozess, die sinnvoll mit dem gegebenen Datensatz bearbeitbar sind. Geben Sie dies in Ihrem Jupyter-Notebook als Markup an (½ Seite). Wichtig ist hier, eigene zu untersuchende Fragen/Hypothesen/Annahmen aufzustellen, die dann untersucht werden. Begleitende Recherche ist erlaubt aber nicht im Fokus der Arbeit.

2. Data Preparation & Feature Engineering (3 Punkte): Lesen Sie den Datensatz der in der zip enthaltenen csv ein. Fügen Sie möglichst elegant die jeweiligen Kontinente hinzu (Asien, Afrika, Nordamerika, Südamerika, Europa und Australien).

Löschen Sie alle Daten mit schlechter Qualität (in der entsprechenden Spalte, einfach vorgegeben).

Bewerten Sie nun die Qualität aller verbleibenden Spalten und bewerten Sie, wie nützlich diese für Ihre Hypothesen sind. Bereinigen Sie die Daten wie in der Vorlesung gelehrt. Prüfen Sie die danach vorliegenden Daten auf mögliche Probleme / Fehler. Sie sollen selbst entscheiden, was hier das richtige Vorgehen ist.

3. Data Exploration und Analyse (9 Punkte): Untersuchen Sie den Datensatz in Bezug auf das Ziel nach den Regeln wie in der Vorlesung gelehrt.

Nutzen Sie Markup, um wichtige Erkenntnisse zu dokumentieren.

Werten Sie sowohl Kontinente, Länder, Städte sinnvoll nach Ihren Zielen generell aus. Berücksichtigen Sie ihr Gruppenziel dabei spezifisch.

4. Modeling und Evaluation Regression (6 Punkte): Sagen Sie vorher, wieviel ein Apfel kostet. Arbeiten Sie wie gelehrt mit Trainings-, Validierungsdaten und Testdaten. Optimieren Sie Ihre Vorhersage, wenn sinnvoll. Eines der Verfahren soll eine möglichst einfache lineare Regression sein, die nur wenige Eingangsspalten verwenden und leicht verständlich ist, sonst bitte aktuelle Verfahren verwenden (keine NNs!).

Geben Sie für die oben genannten Datensätze die Bewertungsmetriken R^2 und RMSE aus. Dokumentieren Sie dies. Interpretieren Sie das Ergebnis und untersuchen Sie den Einfluss der einzelnen Features. Kommentieren Sie jeweils Varianz und Verzerrung der Vorhersage.

5. Unüberwachtes Lernen - Clustering (6 Punkte): Clustern Sie die gegebenen Daten mit einem oder mehreren unüberwachten Lernverfahren (keine NNs) in eine überschaubare, sinnvolle Anzahl und mathematisch guter Cluster. Geben Sie sinnvolle Visualisierungen und Bewertungsmetriken Ihres Ergebnisses an. Zeigen Sie Ihre gewählten Städte innerhalb der Cluster. Interpretieren Sie die Cluster sinnvoll im Sinne von Aufgabe 1.

5. Deployment (3 Punkte): Erstellen Sie eine „Anleitung“ für die im Szenario genannte Zielgruppe. Dies soll die für die Zielgruppe wichtigsten Erkenntnisse zusammenfassen und maximal 2 Seiten im pdf-Ausdruck umfassen.

Bewertungskriterien

1. **Fachliche Bewertung (50%):** Vollständigkeit, Korrektheit, Lösungsqualität und Eleganz sowie Klarheit und Umfang der Betrachtung, Umsetzung von Data Science wie in der Vorlesung gelehrt in einem Code-Prototyp, korrekte Verwendung von wichtigen Funktionen / Bibliotheken, Güte,
2. **Dokumentation (50%):** Dokumentation des Vorgehens der Datenauswertung im Sinne von Data Science, Codekommentare wie in der Informatik üblich wo notwendig, Qualität der Diagramme, Markup, Texte, pdf.

Beachten Sie unsere **Regeln** zur Nutzung von ChatGPT. Analog zu normalen Internetquellen (dort Datum und Link angeben) dürfen Sie ChatGPT nutzen, sofern Sie sich an unsere Regeln halten. Achten Sie auf präzise Aussagen (be concise!).

Abgabe bis 11.04.2025 um 18 Uhr

Bearbeitung findet in Gruppen mit jeweils **genau 2 Personen** statt oder als freiwillige **Einzelarbeit**. Alle Ergebnisse sind einzureichen über **Moodle**.

1. **Programm:**
    a. Matrikelnummer statt Name nutzen (Anonymisierung), Achtung es gibt sonst Abzug!
    b. Quellcode in genau einer Jupyter-IPython-Notebook-Datei (.ipynb)
    c. Lauffähig, Einschränkung auf die in der Vorlesung genutzten Bibliotheken (kein Catboost, keine neuronalen Netze)
    d. Klare Markierung der Aufgabenteile

    e. Dokumentation direkt als Markup enthalten im .ipynb-Notebook
    f. Beschriftungen direkt an Diagrammen
    g. Codekommentare in Codezellen (nur wenn und wo notwendig)
    h. Primäres Ziel des Codes ist die **Lesbarkeit** (nicht Wiederverwendbarkeit), es gibt daher keine Abzüge für redundanten Code.
2. **pdf-Ausdruck des kompletten Notebooks**
    a. Genau eine pdf-Datei pro Team
    b. Hochformat
    c. A4
    d. Einzelseiten (wenn möglich), nur als Notlösung verbunden
    e. Primärquelle für Korrektur ist das pdf!
    f. Erwartete Länge bei sinnvoller Informationsdichte, Schriftgröße etc: 10-20 Seiten (Soll, kein Muss), fokussieren Sie sich, achten Sie darauf, dass es weder „Füllseiten“ noch unlesbare Informationen gibt.
3. **Video** des Ablaufens Ihres Notebooks ohne Ton (max. 3 Minuten, .mp4) als Alternativlösung zur Sicherstellung der Korrekturmöglichkeit in jedem technischen Problemfall (leider bewährt).