Programmentwurf Data Science Prototyp

Gegeben ist ein Immobiliendatensatz gegeben (Datei im Moodle). In diesem sind verschiedene Merkmale von Immobilien in Deutschland gegeben. Die Beschreibung der Merkmale ist unter Kaggle zu finden:
https://www.kaggle.com/datasets/corrieaar/apartment-rental-offers-in-germany

0. Preprocessing (3 Punkte): Lesen Sie den Datensatz der in der zip enthaltenen csv ein. Löschen Sie alle Spalten außer: `['regio1', 'newlyConst', 'balcony', 'totalRent', 'yearConstructed', 'hasKitchen', 'baseRent', 'livingSpace', 'condition', 'interiorQual', 'petsAllowed','noRooms', 'thermalChar', 'regio2', 'regio3', 'lastRefurbish', 'date'.`
Löschen sie alle Zeilen, die nicht in `Chemnitz` sind. Speichern Sie dies als „neue“ csv `outData.csv` ab, die Sie auch mit abgeben. Das ist Ihre Datenbasis – die Empfehlung ist, im Notebook jetzt diese Datei einzulesen (statt der ursprünglichen).

1. Business Understanding (3 Punkte): Formulieren Sie ein Ziel oder mehrere Ziele nach dem CRISP-DM Prozess, die für Anbieter von diversen Renovierungsdienstleistungen (Innen und Außen) sinnvoll mit dem Datensatz bearbeitbar sind. Zusätzlich kann man annehmen, dass bei Personen in teuren Wohnungen mehr finanzieller Spielraum für Renovierungen besteht (gegebene Annahme).

Geben Sie Ihre Ziele in Ihrem Jupyter-Notebook als Markup an (max. ½ Seite). Wichtig ist hier, eigene zu untersuchende Fragen/Hypothesen aufzustellen, die dann in Aufgabenteil 2 untersucht werden. Nutzen Sie auch die vorhandenen Daten, um die Hypothesen zu ergänzen oder anzupassen, wenn notwendig (sie sollten in der Abgabe dann zu den gegebenen Daten passen und damit bearbeitbar sein – z. B. wäre ein Vergleich von dt. und engl. Wohnungen sinnlos, da Sie hierzu keine Daten haben).

Sie dürfen die Daten oder auch Ihre Erkenntnisse ergänzen um Informationen, die Sie im Internet finden, die bei Ihrer Aufgabe helfen, wo notwendig (z. B. wenn Häuser mit Aufzug grundsätzlich nicht renoviert werden können). Dann bitte Quellen angeben (Link und Datum sind ausreichend).

2. Data Exploration und Analyse (6 Punkte): Laden und untersuchen Sie den Datensatz vollständig wie nach den Regeln wie in der Vorlesung gelehrt. Nutzen Sie Markup, um wichtige Erkenntnisse generell (Basisuntersuchungen) sowie in Bezug zu Aufgabe 1 zu dokumentieren.

3. Data Preparation (3 Punkte): Bereinigen Sie die Daten und führen Sie Feature Engineering durch. Hinweis: Kann bereits für Aufgabe 2 teilweise notwendig sein, dann zusammenfassend aufführen (z. B. in Aufgabe 2.5 wurde bereits X gelöscht, der Datensatz enthält bereits n weniger Datenpunkte).

4. Modeling und Evaluation – Regression (4 Punkte): Vergleichen und optimieren Sie drei geeignete Verfahren zur Vorhersage der `totalRent`. Nutzen Sie einmal alle Spalten („full“) und einmal alle außer `baseRent` („part“). ehen Sie vor wie in der

Vorlesung gelehrt mit Trainings-, Validierungsdaten und Testdaten. Optimieren Sie Ihre Vorhersage, wenn sinnvoll.

Geben Sie für die oben genannten Datensätze die Bewertungsmetriken R^2, RMSE, MAPE aus. Dokumentieren Sie dies. Interpretieren Sie das Ergebnis und den Einfluss der einzelnen Features. Kommentieren Sie Varianz und Verzerrung in der Vorhersage basierend auf Ihren Bewertungsmetriken sowie die beiden Vorhersagen („full“ und „part“ gegeneinander.

5. Modeling und Evaluation – Leere Felder (4 Punkte): Versuchen Sie, den `thermalChar` für nicht vorhandene Werte anhand der anderen Einträge zumindest grob vorherzusagen. Sie haben hier im Vorgehen freie Hand und dürfen die Aufgabe und das Vorgehen sinnvoll eingrenzen. Obwohl hier keine Zielwerte vorliegen, sollen Sie sich ein Vorgehen überlegen, welches insgesamt zu guten, für das Business Understanding hilfreichen Ergebnissen führen könnte und in der Evaluation begründen, ob sich ihr Vorgehen bewährt hat oder nicht.

6. Clustering (4 Punkte): Versuchen Sie, für den in Aufgabe 1 erstellen Business Case die Immobilien sinnvoll zu clustern. Sie können dafür alle oder Teile der Features verwenden, nach eigenem Ermessen. Bilden Sie zwischen 3 und 5 sinnvolle Cluster. Beschreiben Sie die Cluster mit Grafiken und mit natürlichsprachlichem Text. Bewerten Sie ihr Ergebnis und ihr Vorgehen nach Sinnhaftigkeit für die gesetzten Ziele aus Aufgabe 1.

7. Deployment (3 Punkte): Erstellen Sie eine Anleitung oder Handreichung für die in Aufgabe 1 genannte Zielgruppe. Dies soll aus nur die aus Zielgruppensicht wichtige Erkenntnisse zusammenfassen (nicht die für die Vorlesung wichtigen Erkenntnisse, wie z. B. die Laufzeiten Ihrer Notebooks oder technische Details, die haben Sie bereits vorher dokumentiert) 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.

Neu: 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.

Abgabe bis zum 22.4.2024, 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. Die in Aufgabe 0 beschriebene outData.csv (nicht die Ursprungsdatei, da zu groß).
    d. Lauffähig
    e. Einschränkung auf die in der Vorlesung genutzten Bibliotheken (kein Catboost, keine neuronalen Netze)
    f. Klare Markierung der Aufgabenteile
    g. Dokumentation direkt als Markup enthalten im .ipynb-Notebook
    h. Beschriftungen direkt an Diagrammen
    i. Codekommentare in Codezellen (nur wenn und wo notwendig)
    j. 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!
3. **Video** des Ablaufens Ihres Notebooks ohne Ton (max. 3 Minuten, .mp4) als Alternativlösung zur Sicherstellung der Korrekturmöglichkeit in jedem technischen Problemfall (leider bewährt).