Programmentwurf Data Science Prototyp v1.0

Es ist ein Immobiliendatensatz gegeben in der Datei `data_for_training.csv`, in dem verschiedene Merkmale von Häusern gegeben sind. Die Beschreibung der Merkmale folgt in diesem pdf unten. Die Daten sind **fiktiv**.

Testdaten sind in der gelöschten Spalte von `data_for_test.csv` zurückgehalten – geben Sie hierfür eine csv-Datei mit ab (siehe Aufgabe 5), in der neben den gegebenen Daten auch Ihre Vorhersagen aus Aufgabe 4 und 5 ergänzt sind. Prüfe Sie, ob sich diese Datei korrekt formatiert und auch fehlerfrei auf einem Windows-Recher öffnen lässt und die Vorhersagen auch korrekt enthält.

1. Business Understanding (3 Punkte): Formulieren Sie ein Ziel oder mehrere Ziele nach dem CRISP-DM Prozess, die für Haustürhandwerksleistungs-vermittlungen sinnvoll sind. Das Geschäftsmodell lautet: Gehe von Tür zu Tür und biete Handwerksdienstleistungen (Streichen, Tür reparieren, neues Vordach, neuer Garten, ..) an. Die Liste der Dienstleistungen ist hier nicht abschließend, gerne dürfen Sie diese auf Basis der Daten sinnvoll ergänzen. Für jede vermittelte Dienstleistung gibt man diese als Tipp an einen Handwerksbetrieb weiter und erhält anteilig Einnahmen für die Vermittlung.

Man kann annehmen, dass Personen in teureren Häusern mehr Geld für Dienstleistungen ausgeben können. Beginnen Sie daher mit der Idee „Wir brauchen mehr Verständnis des Verkaufspreises (`Z_Verkaufspreis`)!“. Geben Sie Ihre Ziele in Ihrem Jupyter-Notebook als Markup an (max. ½ Seite). Wichtig ist hier, eigene zu untersuchende Hypothesen aufzustellen, die dann in Aufgabenteil 2 untersucht werden. Nutzen Sie auch die vorhandenen Daten, um die Hypothesen zu ergänzen oder anzupassen, wenn notwendig.

2. Data Exploration und Analyse (6 Punkte): Laden und untersuchen Sie den Datensatz in `data_for_training.csv` nach den Regeln wie in der Vorlesung gelehrt. Nutzen Sie Mark-Up, um wichtige Erkenntnisse 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 kenntlich machen und zusammenfassend aufführen.

4. Modeling – Regression mit Inferenz (3 Punkte): Führen Sie mit einem geeigneten Verfahren der linearen Regression eine Vorhersage des Preises (`Z_Verkaufspreis`) durch. Ggfs. brauchen Sie dafür mehrere Versionen der „einfachen“ Regressionslösungen, um eine akzeptable Performance zu erreichen. Erklären Sie wichtige identifizierten Zusammenhänge menschenverständlich als Text (z. B. „Eine Haustür erhöht den Preis um 2,75 EUR.“). Geben Sie Qualitätsmetriken für Ihre Lösung (Training- und Validierungsset) an.

5. Modeling und Evaluation (6 Punkte): Vergleichen und optimieren Sie ein oder mehrere weitere Verfahren zur Vorhersage des Verkaufspreises. Gehen Sie vor wie in der Vorlesung gelehrt mit Trainings- und Validierungsdaten (80-20). Optimieren Sie Ihre Vorhersage wenn sinnvoll.

Geben Sie für den **Trainings- und Validierungsdatensatz** die Zielwerte R^2, MSE, RMSE, MAPE, MAX aus. Dokumentieren Sie dies auch.

Interpretieren Sie das Ergebnis und den Einfluss der Features (falls möglich). Kommentieren Sie Varianz und Verzerrung in der Vorhersage.

Schreiben Sie in die `data_for_test.csv` die auf Basis Ihres besten Modells aus Aufgabe 5 sowie Ihres besten Modells aus Aufgabe 4 vorhergesagt Werte in zwei neuen Spalten und geben Sie diese Datei mit ab. (Hinweis: Sortieren Sie nicht um).

6. Deployment (3 Punkte): Erstellen Sie eine Anleitung oder Handreichung für die in Aufgabe 1 genannte Zielgruppe. Dies soll aus Zielgruppensicht wichtige Erkenntnisse der Aufgaben 2 bis 5 zusammenfassen und maximal 2 Seiten im pdf-Ausdruck umfassen, welche auf Basis der Texte aus Aufgabe 1 dann komplett eigenständig lesbar sein sollen.

7. Unsupervised Learning (6 Punkte): Clustern Sie die Immobilien einmal mit und einmal ohne die Features der Bezirke mit aufzuehmen und kommentieren Sie eventuelle Unterschiede. Sie können auch weitere Features hinzunehmen oder weglassen, wenn Sie das für sinnvoll erachten.

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 der Endlösung, Nutzung der erworbenen Kenntnisse aus der Vorlesung, Hinweis: es gibt keine Abzüge für redundanten Code, es ist von Vorteil, wenn die Aufgabe von oben nach unten komplett einfach lesbar ist
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

Abgabe bis zum 8.5 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)
    b. Quellcode in genau einer Jupyter-IPython-Notebook-Datei (.ipynb)
    c. csv-Dateien mit abgeben mit den gegebenen Daten im gleichen Ordner liegend (keine Unterordnerstrukturen), besonders Ihre Vorhersagen in der `data_for_test_filled.csv`
    d. Lauffähig
    e. Einschränkung auf die in der Vorlesung genutzten Bibliotheken (kein Catboost oder neuronale 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. 2 Minuten, .mp4) als Alternativlösung zur Sicherstellung der Korrekturmöglichkeit in jedem technischen Problemfall

! Datenbeschreibungen folgen – siehe folgende Seite !

Allgemein alle N/A entfernen, außer bei Kaminqualität und Besonderheiten => N/A in KQ und B zu -1 umwandeln und 0 in Renovierungsjahr zu -1 umwandeln	Unterschiedlichen Skalen der Attribute standardisieren	Umwandlung Nominal- in Kardinalskala, wenn sinnvoll

A_Index: Eindeutige Identifikationsnummer, nicht fortlaufend (durch Sampling in die ausgegebenen und zurückgehaltenen Daten)

Aussenqualität: Angabe der Qualität der Außenansicht (z. B. wie hochwertig das Material ist)

Aussenzustand: Angabe über den Zustand der Außenansicht (z. B. ob der hochwertige Mamor bereits Schäden hat)

Baujahr: Jahr in dem das Gebäude gebaut wurde

Besonderheiten: Besondere und seltene Eigenschaften des Gebäudes/Verkaufs

Garagen: Anzahl der Fahrzeuge, die in der Garage abgestellt werden können

Garagenbaujahr: Jahr in dem die Garage gebaut wurde

Grundstueck_qm: Grundstücksfläche in qm

Heizungszustand: Zustand der eingebauten Heizung

Kamine: Anzahl der Kamine

Kaminqualitaet: Qualität der Kamine (wenn vorhanden)

Keller: Typ des Kellers

- Guter Wohnraum
- Mittlerer Wohnraum
- Kein Wohnraum
- Freizeitraum
- Niedrige Qualität
- Rohbau

Kuechenzsutand: Zustand der Küche

Lage: Bezirk, in dem die Immobilie steht

Qualitaet: Gesamtqualität (Innen und Außen), je höher desto besser

Renovierungsjahr: Jahr, in dem größere Umbauten / Anbauten / Renovierungen stattfanden (sofern welche durchgeführt wurden, sonst „0“)

Steuerzuordnung: Zuordnung zum Grundsteueramt

Verkaufsjahr: Jahr des Verkaufs

Wohlflaeche_qm: Wohnfläche in qm

Zustand: Gesamtzustand (Innen und Außen), je höher desto besser

Zz_Verkaufspreis: Verkaufspreis in Euro

HINWEIS: Es gilt immer: Sehr gut > Gut > Durchschnitt > Schlecht > Sehr Schlecht