Programmentwurf Data Science Prototyp v1.0 vom 3.11.2022

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, in der neben den gegebenen Daten auch Ihre Vorhersagen ergänzt sind. Prüfe Sie, ob sich diese Datei 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 Immobilienspekulant*innen sinnvoll sind. Bei Spekulationen werden typischerweise Immobilien erstanden, die wieder mit Gewinn abgestoßen werden. Beginnen Sie 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 (9 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.“).

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). Untersuchen Sie Varianz und Verzerrung in der Vorhersage.

Schreiben Sie in die `data_for_test.csv` die auf Basis Ihres besten Modells vorhergesagt Werte in eine neue Spalte 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. Classification (3 Punkte): Versuchen Sie Immobilien dem richtigen Bezirk zuzuordnen, dabei können Sie den Preis als Eingabewert nehmen. Bewerten Sie die Qualität Ihrer Lösung und kommentieren Sie Ihre Erkenntnisse aus diesem kleinen Test. Erwarteter Umfang entspricht 3 Punkten von 30.

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 22.12.2022 um 12:00 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

Anhang: Beschreibung der Datenfelder

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

Anzahl Zimmer: Gesamtanzahl der Zimmer (keine Küchen und Bäder eingerechnet)

Ausbaustufe: Anzahl der Ebenen oberhalb des Kellers

1 Ebene

2 Ebenen

Baeder: Anzahl der Badezimmer die nicht im Kellergeschoss (KG) liegen, Toiletten eingerechnet

BaederKG: Analog Baeder, aber im KG

Baujahr: Jahr in dem das Gebäude gebaut wurde

EG_qm: Größe der Wohnfläche in qm im Erdgeschoss

Garage_qm: Größe der Garage in qm

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

Gesamteindruck: Eindruck des Gesamtzustandes des Gebäudes insgesamt

5 Sehr gut

4 Gut

3 Durchschnitt

2 Schlecht

1 Sehr schlecht

Keller_Typ_qm: Anzahl der qm im Typ des Kellers (siehe „Kellertyp“ unten)

Keller_qm: Anzahl der qm des gesamten Kellers

Kellerhoehe: Höhe des Kellers

Sehr gut: ca. 250 cm

Gut: ca. 225 cm

Durchschnitt: ca. 200 cm

Schlecht: ca. 175 cm

Sehr schlecht: niedriger als 175 cm

Keine Angabe: kein Keller

Kellertyp: Typ des Kellers

Guter Wohnraum

Mittlerer Wohnraum

Kein Wohnraum

Freizeitraum

Niedrige Qualität

Rohbau

Lage: Bezirk, in dem die Immobilie steht

OG_qm: Quadratmeter des Geschosses oberhalb des OG

Umgebaut: Jahr, in dem größere Umbauten / Anbauten / Renovierungen stattfanden, wenn keine durchgeführt wurden entspricht dies dem Baujahr

Verkaufsjahr: Jahr des Verkaufs

Verkaufsmonat: Monat des Verkaufs

Wohlflaeche_qm: Wohnfläche in qm

Z_Verkaufspreis: Verkaufspreis in Euro