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**, d.h. keiner realen Stadt zuzuordnen, jedoch orientieren sie sich am nordamerikanischen Markt. Echte 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 und genauso formatiert ist wie die gegebenen Daten (bitte manuell überprüfen).

1. Business Understanding (3 Punkte): Formulieren Sie ein Ziel oder mehrere Ziele nach dem CRISP-DM Prozess, die für mittelfristige Investoren (Investitionen für maximal 1-2 Jahre) auf Basis der gegebenen Daten sinnvoll zu beantworten sind. Beginnen Sie mit der Idee „Wir brauchen eine Vorhersage des Verkaufspreises (`Z_Verkaufspreis`)!“, welche auf jeden Fall zu bearbeiten ist. Geben Sie dies in Ihrem Jupyter-Notebook als Markup an (max. ½ Seite, Fließtext).

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. Hierzu erstellen Sie sinnvolle Charts in verschiedenen Detailstufen und kommentieren Ihre Ergebnisse als Markup.

3. Data Preparation (3 Punkte): Bereinigen Sie die Daten und führen Sie ein sinnvolles Feature Engineering durch. Hinweis: Das kann auch für Punkt 2 bereits relevant sein (führen Sie das dann hier zusammenfassend auf, z. B. „X schon erledigt oben“.).

4. Modeling – Regression mit Inferenz (6 Punkte): Führen Sie eine Regression durch für die Zielvariable mit mindestens 3 bekannten Verfahren – davon soll eins eine lineare Annährung sein. Kommentieren Sie Varianz und Verzerrung der Vorhersage. Untersuchen Sie den Einfluss der Features für die Vorhersage. Vergleichen Sie die untersuchten Verfahren in Bezug auf gängige Zielwerte. 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. Das Format dieser Datei soll identisch sein mit den erhaltenen Dateien. Sortieren Sie nicht um.

5. Klassifikation (3 Punkte): Führen Sie eine Vorhersage der Lage einer Immobilie ausschließlich auf Basis des Preises und der Anzahl der Stockwerke durch. Wählen Sie dafür mindestens 2 Verfahren. Bewerten Sie ihr Modell. Beurteilen Sie ihr Modell natürlichsprachlich selbst (sind Sie zufrieden? Warum? Warum nicht?).

6. Clustering (3 Punkte): Führen Sie ein Clustering der Immobilien durch mit allen Features außer dem Bezirk. Versuchen Sie, die Cluster mit Charts und natürlichsprachlich zu beschreiben. Korrelieren diese mit den Bezirken?

7. Deployment (3 Punkte): Erstellen Sie eine Anleitung. Dies soll aus Kundensicht wichtige Erkenntnisse zusammenfassen (auch wenn dadurch Redundanz in der Abgabe entsteht). Achten Sie auf präzise und kurze, bündige Aussagen die sich ausschließlich auf die Fragestellung beziehen und auf Ihren Daten basieren!

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, Präzision der Aussagen, Belegung mit Daten und Diagrammen (kein „Erraten“)

Abgabe bis zum 19.12.2025 um 19:00 Uhr im

Modus Zweiergruppen oder freiwillige Einzelarbeit

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
   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 hier keine Abzüge für redundanten Code.
2. **pdf-Ausdruck des kompletten Notebooks**
   a. Genau eine große pdf-Datei
   b. Hochformat
   c. A4
   d. Einzelseiten
   e. Primärquelle für Korrektur ist das pdf!
3. **Video** des Ablaufens Ihres Notebooks ohne Ton (max. 2 Minuten, .mp4)

!Datenbeschreibungen folgen – siehe folgende Seiten!

Anhang: Beschreibung der Datenfelder

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

Ausbaustufe: Art des Gebäudes

- 1 Ebene
- 1,5 Ebenen
- 2 Ebenen
- 2,5 Ebenen

Baujahr: Jahr in dem das Gebäude gebaut wurde

Besonderheiten: Besonderheiten die beim Kauf relevant waren

EG_qm: Größe der Wohnfläche in qm im Erdgeschoss

Gesamtqual: Eindruck des ersten Eindrucks an Qualität des Gebäudes insgesamt

- Sehr gut
- Gut
- Durchschnitt
- Schlecht
- Sehr schlecht

Gesamtzustand: Eindruck des Gesamtzustandes des Gebäudes insgesamt

- Sehr gut
- Gut
- Durchschnitt
- Schlecht
- Sehr schlecht

Grundstueck_qm: Größe des Grundstücks in qm

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

Lage: Stadtteillage in der fiktiven nordamerikanischen Stadt Neu-Berlin

QualInnenfarbe: Qualität des Innenanstrichs

1: sehr niedrig, bis 5: sehr hoch

Steigung: Steigung des Grundstücks

Keine/Kaum: Fast keine merkliche Steigung

Mittel: Moderate Steigung

Stark: Auffällige Steigung

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

Verkaufsjahr: Jahr des Verkaufs

Wohlflaeche_qm: Wohnfläche in qm

Z_Verkaufspreis: Verkaufspreis in Euro