Programmentwurf Data Science Prototyp v1.0

Es ist ein Hausdatensatz gegeben in der Datei `data_for_training.csv`, in dem verschiedene Merkmale von Häusern gegeben sind sowie eine Beschreibung der Merkmale. Die Daten sind fiktiv, d.h. keiner realen Stadt zuzuordnen, jedoch orientieren die Daten sich am nordamerikanischen Markt. Diesen Datensatz nutzen Sie für Training und Validierung. Echte Testdaten werden im Format wie `data_for_testing_test.csv` zurückgehalten.

1. Business Understanding (3 Punkte): Formulieren Sie ein Ziel oder mehrere Ziele nach dem CRISP-DM Prozess, die für einen Immobilienverein von Privatpersonen sinnvoll sind. Diese Personen kaufen und verkaufen Häuser zur Eigennutzung oder für die Familie, manchmal mehrfach. Beginnen Sie mit der Idee „Wir brauchen mehr Verständnis und 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 Text).

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. Ändern Sie hierbei nicht die einzulesende csv-Datei! Schreiben Sie die wichtigsten Erkenntnisse für die in Aufgabe 1 definierten Ziele als Summary auf (max. ½ Seite Text). Passen Sie ggfs. Ihre Ziele im Business Understanding an. Kommentieren Sie Rückfragen für Fachexperten sowie mögliche zusätzliche Datenquellen und Auswertungen, die Sie damit ausführen würden.

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 dann nochmals zusammenfassend auf).

4. Modeling – Regression mit Inferenz (3 Punkte): Führen Sie mit geeigneten Verfahren der Regression (Linear / Lasso / Ridge) eine Vorhersage des Preises (`Z_Verkaufspreis`) durch, der akzeptabel in der Evaluation abschneidet und Verständnis ermöglicht. Erklären Sie die identifizierten Zusammenhänge menschenverständlich als Text (mit nachvollziehbarer Anzahl der Merkmale).

5. Modeling – Best of Class (3 Punkte): Vergleichen und optimieren Sie ggfs. ein oder mehrere andere Verfahren zur Vorhersage des Verkaufspreises (`Z_Verkaufspreis`). Gehen Sie vor wie in der Vorlesung gelehrt mit Trainings- und Validierungsdaten. Vergleichen und optimieren Sie. Interpretieren Sie das Ergebnis und den Einfluss der Dimensionen (falls möglich).

6. Evaluation und Test (3 Punkte): Schließen Sie die Aufgabe mit einer finalen Evaluation der Vorhersagequalität ab. Stellen Sie sicher, dass die Testdatei (`data_for_testing_test.csv`) ladbar ist und mit zurückgehaltenen Testdaten `data_for_test.csv` ersetzt werden kann, um einen Benchmark zu erstellen. Kommentieren Sie die Stelle mit dem Kommentar „#HIER DATEINAMEN ERSETZEN“. Geben Sie für die Testdaten aus: R^2, MSE, RMSE, MAPE, MAX.

7. Deployment (3 Punkte): Erstellen Sie eine Anleitung oder Handreichung für den Immobilienverein aus Aufgabe 1 basierend auf allen Erkenntnissen. Sie können Markup und erklärende Visualisierungen innerhalb des .ipynb nutzen. Dies soll alle

für den Verein wichtigen Erkenntnisse zusammenfassen (auch wenn dadurch Redundanz in der Abgabe entsteht) und maximal 2 Seiten im pdf-Ausdruck umfassen, welche komplett eigenständig lesbar sein sollen.

Add-On: Modeling – Klassifikation (3 Punkte): Versuchen Sie eine Vorhersage der Kamine (ja / nein, alternativ: Anzahl der Kamine, Feld: `Kamine`). Evaluieren Sie die Vorhersage. Dieser Punkt fließt nicht in alle anderen Aufgabenteile ein.

Bewertungskriterien

1. **Fachliche Bewertung (50%):** 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, Vollständigkeit der Lösung in Bezug auf die Aufgabenstellung
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 1.12.2021 16:00 Uhr

1. **Programm:**
    a. Matrikelnummer statt Name nutzen (Anonymisierung)
    b. Quellcode in genau einer Jupyter-IPython-Notebook-Datei (.ipynb)
    c. Original-csv-Dateien mit abgeben mit den gegebenen Daten im gleichen Ordner liegend (keine Unterordnerstrukturen)
    d. Lauffähig
    e. Einschränkung auf die gegebenen Bibliotheken wie in der Vorlesung angegeben
    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)
2. **pdf-Ausdruck des kompletten Notebooks**
    a. Primärquelle für Korrektur ist das pdf!
    b. Text / Grafiken / Code einzeln prüfen auf Lesbarkeit / Optik
    c. Genau eine große pdf-Datei pro Team
3. **Video** des Ablaufens Ihres Notebooks ohne Ton (max. 2 Minuten, .mp4)

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

Anhang: Beschreibung der Datenfelder

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

AnzahlZimmer: Anzahl der Zimmer, Keller nicht eingerechnet

Ausbaustufe: Art des Gebäudes

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

Baeder: Anzahl der Bäder, Keller nicht eingerechnet

BaederUG: Anzahl der Bäder, nur Keller

Baujahr: Jahr in dem das Gebäude gebaut wurde

Bebauungsdichte: Dichte des bebauten Wohngebietes

- Niedrig
- Mittel
- Hoch
- Seelage: Spezielle Lage an einem See
- Unbekannt: Lage außerhalb der ausgewiesenen Wohngebiete, z. B. Gewerbegebiete oder auf Feldern

Besonderheiten: Besonderheiten die beim Kauf relevant waren

Dachtyp: Dach der Immobilie

- Flachdach
- Satteldach
- Mansarddach
- Pultdach
- Walmdach

EG_qm: Größe der Wohnfläche in qm im Erdgeschoss

Fassadenqual: Qualität der Verkleidung der Fassade

- Sehr gut
- Gut
- Durchschnitt

Schlecht

Sehr schlecht

Fassadenzustand: Zustand der Verkleidung der Fassade

Gleiche Abstufung wie Fassadenqualität

Garagen: Anzahl der Fahrzeuge, die in die Garage passen, dabei ist 0 keine Garage

Gesamteindruck: Eindruck des Zustands der gesamten Immobilie

1 Sehr schlecht

2 Schlecht

3 Durchschnitt

4 Gut

5 Sehr Gut

Grundstueck_qm: Größe des Grundstücks in qm

Kamine: Anzahl der Kamine

KellerQual: Eindruck vom Zustand des Kellers

Keller_qm: Größe des Keller in qm

Kellerhoehe: Höhe des Kellers

Sehr gut: über 250 cm

Gut: über 225 cm

Durchschnitt: über 200 cm

Schlecht: über 175 cm

Sehr schlecht: darunter

Lage: Stadtteillage in der fiktiven nordamerikanischen Stadt Neu-Stuttgart

ToilettenEG: Anzahl der Halbbäder und Toiletten, Keller nicht eingerechnet

ToilettenUG: Anzahl der Halbbäder und Toiletten, nur Keller

Typ: Typ des Hauses

Freistehend

Doppelhaus

Reihenhaus

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