Programmentwurf Data Science Prototyp v1.0

Es ist ein Hausdatensatz gegeben, in dem verschiedene Merkmale von Häusern gegeben sind mit 2000 Datenpunkten sowie einer Beschreibung der Merkmale. Die Daten sind fiktiv, d.h. keiner realen Stadt und keinem realen Zeitpunkt zuzuordnen.

1. Business Understanding (3P): Formulieren Sie ein Ziel oder mehrere Ziele nach dem CRISP-DM Prozess, die für **Investor*innen** sinnvoll sind. Diese Personen investieren Geld mit dem Ziel, dieses zu vermehren, z. B. durch Kauf/Verkauf, Renovierung, Umbau. Beginnen Sie mit der Idee „Wir brauchen mehr Verständnis und eine Vorhersage des Verkaufspreises (`Preis`)!“, welche auf jeden Fall zu bearbeiten ist.

Für Aufgabe 1 ist eine Zwischenabgabe Pflicht bis zum **16.11.2020 16:00 Uhr.** Geben Sie Aufgabe 1 nochmals als Teil der gesamten Abgabe ab (Sie dürfen die Ergebnisse nochmals ändern, wenn es Feedback / neue Informationen gibt).

2. Data Exploration und Analyse (10P): Untersuchen Sie den Datensatz. Analysieren Sie die gegebenen Informationen und suchen Sie Anomalien. Behalten Sie dabei die Zielstellung aus Teil 1 im Auge. Bewerten Sie Schlussfolgerungen daraufhin, wie viel Information Sie haben, um diese zu stützen. Beschreiben Sie die Ergebnisse in ganzen Sätzen auf Deutsch. Nutzen Sie Diagramme wo sinnvoll – achten Sie darauf, diese gut zu formatieren, zu beschriften und korrekt einzusetzen. Schreiben Sie die allerwichtigsten Erkenntnisse für die in Teil 1 definierten Ziele als Summary auf (3 bis 5 Erkenntnisse).

3. Data Preparation und Modeling (6 Punkte): Bereinigen Sie die Daten, falls notwendig. Führen Sie Feature Engineering durch, wenn notwendig. Führen Sie mit geeigneten Verfahren eine Vorhersage des Preises (`Preis`) durch. Eine davon soll eine erklärbare, verständliche und interpretierbare lineare Regression sein. Erklären Sie diese im Detail in Bezug auf die Ziele aus Aufgabe 1. Wählen Sie mehrere geeignete Regressionsverfahren. Vergleichen Sie die angewendete Verfahren grafisch. Begründen Sie Ihre Wahl in Bezug auf die Ziele.

4. Evaluation (3 Punkte): Evaluieren Sie ihr finales Modell. Quantifizieren Sie bei der Evaluation die Konfidenz ab. Achten Sie auf die Sinnhaftigkeit der Bewertung im Bezug zu den Geschäftszielen aus Aufgabe 1. Stellen Sie 3 bis 5 wichtige Erkenntnisse in einem Summary heraus. Beschreiben Sie diese in ganzen Sätzen auf Deutsch.

5. Deployment (3+3 Punkte): Erstellen Sie eine Anleitung / Handreichung / Vorgehen für Investor*innen basierend auf den Erkenntnissen von Aufgabe 1 bis 4. Sie können dies per Hand, mit Software, als Formel, als Folie, oder anders umsetzen. Wählen Sie eine geeignete Präsentationsform für Ihre Zielgruppe. Stellen Sie Ihre gesamten in Punkt 1 – 4 erarbeiteten Ergebnisse in 5 bis 7 Folien Ihren Auftraggeber*innen als Videopräsentation vor.

Technische Anforderung (2 Punkte): Stellen Sie die Ablauffähigkeit sicher. Ändern Sie nichts an der csv-Datei, die vorgegeben ist. Ermöglichen Sie es, eine zusätzliche csv-Datei in dem gegebenen Format einzulesen, mit der dann das finale Modell durchlaufen und für die Hauspreisvorhersage folgende fünf Werte berechnet werden:

R^2, MSE, RMSE, MAPE, MAX.
Sie können dies testen, indem Sie von der vorgegebenen csv-Datei z. B. die ersten 3 oder die letzten 3 Zeilen einlesen und prüfen, ob das Programm sich erwartungskonform verhält. Dokumentieren Sie die Code-Stelle und Verwendung.

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, Videopräsentation

Abgabe

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

Für **Aufgabe 1** ist eine Zwischenabgabe Pflicht bis zum **16.11.2020 16:00 Uhr.** Hier ist die Abgabeform ein pdf-Dokument, 1 Seite DIN A4.

Komplettabgabe ist fällig bis zum **30.11.2020 16:00 Uhr,** abzugeben sind:

1. **Programm:** Quellcode in genau einer Jupyter-IPython-Notebook-Datei (.ipynb), original-csv-Datei mit Daten im gleichen Ordner liegend, lauffähig, klare Markierung der Aufgabenteile, Dokumentation (direkt als Markup enthalten im .ipynb, Beschriftungen direkt an Diagrammen, Codekommentare in Codezellen wo notwendig, Matrikelnummer statt Name nutzen, achten Sie auf Eleganz und Lesbarkeit
2. **pdf-Ausdruck des kompletten Notebooks** mit Grafiken (achten Sie darauf, dass alles vollständig enthalten ist, z. B. dass alle Diagramme lesbar sind, dass der Zeilenumbruch klappt)
3. **Video** der Präsentation der Ergebnisse mit Hilfe der in Punkt 5 erstellten Handreichung. Präsentation selbst durch beide Teammitglieder in 3 – 4 Minuten (Abzug für das Über- oder Unterschreiten der Zeit). Anschließend Video-Demo des Programmablaufs (Echtzeit, maximal 2 Minuten).