Allgemeine Anforderungen

Im Folgenden sind allgemeine Anforderungen an und Hinweise für die Studienarbeit zusammengefasst:

Bestandteile

Die Studienarbeit muss mindestens folgende Bestandteile enthalten:

- Analyse der Daten inkl. Bewertung der vorliegenden Datenqualität
- Erstellung und Evaluation eines Modells zur Bearbeitung der Fragestellung inkl. Begründung der Modellwahl und Beschreibung möglicher Defizite bzw. Verbesserungspotentiale
- Diskussion der Resultate in Worten, Zahlen und Grafiken

Quellen

Wie über das gesamte Semester möchte ich Sie auch hier explizit ermuntern, Quellen zu verwenden. Sie brauchen das Rad (oder die Versicherung) nicht neu zu erfinden, sondern sollen vielmehr sinnvolle Fragen stellen und diese dann (ggf. unter Zuhilfenahme von Quellen) beantworten. Ergiebige Quellen in diesem Umfeld sind z.B.:

- Im Kurs genannte Literatur
- Dokumentation von Bibliotheken wie ScikitLearn
- Data Science Plattformen wie z.B. kaggle.com
- Blogs wie z.B. Towards Data Science

Sollten einzelne (Teile von) Analysen aus externen Quellen verwendet werden, so sind diese Quellen zwingend zu zitieren. Bei konkreten Fragen, wie man etwas programmiertechnisch umsetzt (z.B. "Wie füge ich einem DataFrame weitere Zeilen hinzu?") können Suchmaschinen oder einschlägige Foren wie z.B. Stackoverflow nützlich sein. Diese brauchen Sie natürlich nicht zu zitieren.

Abgabe

Das Abgabeformat ist ein Jupyter-Notebook, welches in einer Python 3 Umgebung unverändert fehlerfrei laufen muss. Sollten Sie die Aufgabe im Team bearbeitet haben, so ist nur eine Abgabe unter Nennung aller Teammitglieder notwendig. Die Abgabe des Jupyter-Notebooks erfolgt per Upload im Moodle-Kurs. Der Dateiname sollte `X_nachname.ipynb` sein, wobei $X \in \{A, B, C\}$ die gewählte Aufgabe angibt und `nachname` Ihren Nachnamen (im Fall von Abgaben im Team entsprechend all Ihre Nachnamen). Offizieller Abgabetermin ist der letzte Vorlesungstag des Semesters, also der 24. Januar 2023. Nachträgliche Abgaben bis 28. Februar 2023 werden akzeptiert.

Bewertung

Die Bewertung erfolgt anhand der in den folgenden Kategorien erzielten Punkte:

- Form der Arbeit – berücksichtigt insbesondere Sprache, Lesbarkeit des Codes (Kommentierung), Qualität der Grafiken
- Logischer Aufbau und Argumentation – berücksichtigt insbesondere die Struktur der Arbeit ("roter Faden") sowie die Motivation und Interpretation von Analysen
- Einsatz von Methoden – berücksichtigt methodische Aspekte insbesondere bei der Analyse und Aufbereitung der Daten und bei Auswahl und Training von Modellen
- Ergebnisse des Modells – berücksichtigt die Güte der vom Modell getroffenen Vorhersage.

Dabei wird jede der genannten Kategorien mit 0-5 Punkten bewertet. Zusätzlich gibt es einen Überhang: Jede der drei Aufgabenstellungen enthält [*eine Zusatzfrage, die kursiv in eckigen Klammern*] gestellt ist. Die Bearbeitung dieser Frage ist optional und wird ggf. ebenfalls mit bis zu 5 Punkten bewertet. Die Kategorie "Einsatz von Methoden"wird doppelt gewichtet, alle anderen einfach. Insgesamt sind somit maximal 30 Punkte erreichbar, wobei 25 Punkte als 100% angesehen werden.

Viel Freude und Erfolg bei der Erstellung der Studienarbeit!

Aufgabenstellung

Im Rahmen dieser Studienarbeit haben Sie die Auswahl zwischen einem Klassifikations- und zwei Regressionsproblemen. Sie bearbeiten *eine* der drei folgenden Aufgaben.

Aufgabe A – Regression

Allgemeines Setting

Wir betrachten die Daten der Münchener Raddauerzählstellen, die an sechs Standorten im Stadtgebiet installiert sind. Ein unter der Straßenoberfläche verlegter Sensor erfasst die Anzahl der Radlerinnen und Radler. Die Zählergebnisse werden in einer Auflösung von 15 Minuten gespeichert. Basierend auf den verfügbaren historischen Daten soll ein Modell entwickelt werden, dass die Hochrechnung von Kurzzeitzählungen (zwei mal vier Stunden) auf volle Tageswerte zuverlässig ermöglicht.

Beschreibung der Daten

Die Daten der Raddauerzählstellen sind über opendata.muenchen.de verfügbar. Um Ihnen die Arbeit zu erleichtern, erhalten Sie via GitLab sämtliche dort verfügbare Daten zusammengefasst in der Datei `bike_15min.csv`. Diese Datei enthält neben Datum und Zeitraum der Zählung die Zählstelle sowie die gezählten Fahrräder pro Fahrtrichtung und die Summe beider Richtungen. Im Rahmen dieser Aufgabenstellung ist nur die Summe beider Richtungen relevant.

Beschreibung der Aufgabe

Untersuchen Sie die vorliegenden Daten und gehen Sie dabei insbesondere auf fehlende Daten ein. Verdeutlichen Sie, dass die Baustelle an der Arnulfstraße seit Anfang 2021 die dortige Radzählung stark beeinflusst. Aus diesem Grund sollen die Daten der Messstelle Arnulfstraße nicht weiter berücksichtigt werden. Außerdem soll die zeitliche Auflösung der Daten auf Stundenintervalle (statt 15min) angepasst werden.
Für eine Kurzzeitzählung wird der Radverkehr von 08:00 – 12:00 Uhr sowie 14:00 – 18:00 Uhr erfasst und aus diesen Daten eine Vorhersage für den gesamten Radverkehr am jeweiligen Tag bestimmt. Erstellen Sie ein lineares Modell das diese Vorhersage durchführt. Führen Sie dazu mindestens ein neues, in diesem Zusammenhang sinnvoll erscheinendes Feature ein. Teilen Sie die vorliegenden Daten sinnvoll in Trainings- und Testdaten auf und trainieren Sie neben dem linearen Modell noch ein weiteres frei zu wählendes Modell.
Die Modelle sollten für jeden der fünf verbleibenden Standorte separat trainiert und bewertet werden. Verwenden Sie für die Bewertung die RMSE-Score. Interpretieren und beschreiben Sie insbesondere das lineare Modell [*und vergleichen Sie die resultierende "lineare Hochrechnungsmethode" zwischen den Standorten. Ergeben sich ähnliche Hochrechnungsmethoden? Warum (nicht)?*]

Aufgabe B – Regression

Allgemeines Setting

Wir betrachten wie in Aufgabe A die Daten der Münchener Raddauerzählstellen, die an sechs Standorten im Stadtgebiet installiert sind. Ein unter der Straßenoberfläche verlegter Sensor erfasst die Anzahl der Radlerinnen und Radler. Die Zählergebnisse werden auf Tageswerte aggregiert und zusammen mit Informationen über das lokale Wetter gespeichert. Basierend auf den verfügbaren historischen Daten soll ein Modell entwickelt werden, dass den Radverkehr für einen gegebenen Tag vorhersagt.

Beschreibung der Daten

Die Daten der Raddauerzählstellen sind über opendata.muenchen.de verfügbar. Um Ihnen die Arbeit zu erleichtern, erhalten Sie via GitLab sämtliche dort verfügbare Daten zusammengefasst in der Datei `bike_daily.csv`. Diese Datei enthält neben Datum und Zeitraum der Zählung die Zählstelle sowie die gezählten Fahrräder pro Fahrtrichtung und die Summe beider Richtungen. Im Rahmen dieser Aufgabenstellung ist nur die Summe beider Richtungen relevant. Außerdem sind die Tagestiefst- und Tageshöchsttemperatur, die Niederschlagsmenge, die prozentuale Bewölkung sowie die Sonnenstunden verfügbar.

Beschreibung der Aufgabe

Untersuchen Sie die vorliegenden Daten und gehen Sie dabei insbesondere auf fehlende Daten ein. Verdeutlichen Sie, dass die Baustelle an der Arnulfstraße seit Anfang 2021 die dortige Radzählung stark beeinflusst. Aus diesem Grund sollen die Daten der Messstelle Arnulfstraße nicht weiter berücksichtigt werden. Außerdem sind die verfügbaren Daten nicht vollständig. Bestimmen Sie die fehlenden Tage.
Ziel dieser Aufgabe ist es, für einen gegebenen Tag die Anzahl der Fahrradfahrer an einer gegebenen Messstelle zu prognostizieren. Erstellen Sie ein lineares Modell das diese Vorhersage durchführt. Führen Sie dazu mindestens ein neues, in diesem Zusammenhang sinnvoll erscheinendes Feature ein. [*Untersuchen Sie in diesem Zusammenhang, welchen Einfluss das Oktoberfest auf der Radverkehr hat.*] Teilen Sie die vorliegenden Daten sinnvoll in Trainings- und Testdaten auf und trainieren Sie neben dem linearen Modell noch ein weiteres frei zu wählendes Modell.
Die Modelle sollten für jeden der fünf verbleibenden Standorte separat trainiert und bewertet werden. Verwenden Sie für die Bewertung die RMSE-Score. Interpretieren und beschreiben Sie insbesondere das lineare Modell. Nutzen Sie die beiden erstellten Modelle, um die fehlenden Daten zu füllen. Führen Sie außerdem eine Prognose für den Zeitraum 01.11.–31.12.2022 am Standort “Olympia” durch und stellen Sie für diesen Standort das gesamte Jahr 2022 grafisch dar.

Aufgabe C – Klassifizierung

Allgemeines Setting

Wir betrachten die Daten von knapp 60.000 Kfz-Versicherungspolicen. Diese Daten beinhalten u.a. die Alter der jeweiligen Police, das Alter des Versicherungsnehmers, Daten zur Region, in der das Fahrzeug angemeldet ist, sowie Daten zum Fahrzeug selbst. Ziel ist es, ein Modell zu entwickeln, welches basierend auf diesen Daten vorhersagt, ob für eine Police innerhalb der nächsten sechs Monate ein Schaden gemeldet wird oder nicht.

Beschreibung der Daten

Es werden Ihnen drei Dateien via GitLab zur Verfügung gestellt:

- `ins_train.csv` enthält sämtliche Informationen zu knapp 60.000 Policen sowie die Zielvariable `is_claim`.
- `ins_live.csv` enthält die analogen Informationen zu weiteren knapp 40.000 Policen, allerdings ist hier die Zielvariable `is_claim` nicht verfügbar. Diese soll im Rahmen der Aufgabe bestimmt werden.
- `ins_pred_sample.csv` ist ein Beispiel für die Ergebnisdatei, die von Ihnen abgegeben werden soll.

Die Daten stammen aus dem Dataverse Hack 2022. Eine genauere Beschreibung der verfügbaren Features finden Sie hier.

Beschreibung der Aufgabe

Untersuchen Sie die vorliegenden Daten und gehen Sie dabei insbesondere auf die Relevanz der vielen verfügbaren Features ein. Beachten Sie auch, dass die vorliegenden Daten stark unausgeglichen sind, da es wesentlich mehr Policen gibt, die keinen Schaden melden werden. Diskutieren Sie mögliche damit in Verbindung stehenden Probleme. [*Wie könnte man mit dieser Problematik umgehen? Probieren Sie Ihre Idee aus.*]
Die Zielvariable `is_claim` ist binär, der Wert 0 steht für "keine Schadensmeldung", der Wert 1 steht für "Schadensmeldung". Es sollen zwei Modelle zur Vorhersage von `is_claim` entwickelt werden: ein (möglicherweise regularisiertes) lineares Modell sowie ein frei zu wählendes weiteres Modell. Interpretieren und beschreiben Sie insbesondere das trainierte lineare Modell und vergleichen Sie dessen Performance mit der des frei gewählten Modells. Als Metrik zur Bewertung soll das sog. F_1-Maß verwendet werden, welches definiert ist als harmonisches Mittel von Precision P und Recall R:

$$F_1 = 2\frac{PR}{P+R}$$

Verwenden Sie das bessere der beiden entwickelten Modelle, um für die in `live.csv` gegebenen Daten die Vorhersage für `is_claim` zu bestimmen. Das Ergebnis soll in die Datei `pred_live.csv` ausgegeben werden, die wie die Beispieldatei `pred_sample.csv` aufgebaut ist. Diese Datei ist neben dem Jupyter-Notebook, das die Studienarbeit enthält, abzugeben.