Allgemeines

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

Quellen

Wie über das gesamte Semester möchte ich Sie auch hier explizit ermuntern, Quellen zu verwenden. Sie brauchen das Rad 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. Auch Tools wie ChatGPT oder GitHub Copilot sind hierfür sehr nützlich. Diese brauchen Sie natürlich nicht zu zitieren.

Abgabe

Abzugeben sind die vier Jupyter-Notebooks

- `daten_pv.ipynb`
- `daten_wetter.ipynb`
- `model_energiebedarf.ipynb`
- `model_PV.ipynb`,

welche in einer Python 3 Umgebung unverändert fehlerfrei laufen müssen. Rohlinge dieser vier Jupyter Notebooks finden Sie in GitLab unter `Studienarbeit/abgabe`. Bitte verwenden Sie diese Dateien und tragen Sie in der ersten Zelle an der **vorgesehenen Stelle Ihren Namen** ein.
Außerdem sind die vier .csv Dateien

- `PV.csv`
- `wetter.csv`
- `model_energiebedarf.csv`

- `model_PV.csv`

abzugeben. Beispiele für diese Dateien finden Sie ebenso bei GitLab unter `Studienarbeit/abgabe`.
Die Abgabe aller acht Dateien erfolgt per Upload der einzelnen Dateien (nicht gezippt etc.) im Moodle-Kurs.
In die Bewertung gehen insbesondere die Antworten auf die unten explizit gestellten Fragen ein. Ebenso wird die Qualität der .csv Dateien bewertet. Auch die Einhaltung der beschriebenen Anforderungen an Datenformate, Upload etc. wird bewertet.
Abgabetermin ist der letzte Vorlesungstag des Semesters, also der **23. Januar 2024**.

Viel Freude und Erfolg bei der Erstellung der Studienarbeit!

Aufgabenstellung

Allgemeines Setting

Im Rahmen dieser Studienarbeit betrachten wir ein Musterhaus, das zwischen München, Augsburg und Landsberg auf dem Lechfeld gelegen ist (geografische Breite 48.2° Nord, bzw. 48°12′N) und von einer vierköpfigen Familie bewohnt wird. Das Haus verfügt über eine Photovoltaik-Anlage mit einer Peak-Leistung von 9.9kWp und wird mittels einer Wärmepumpe mit Wärme versorgt. Es stehen Daten zum Stromverbrauch und zur Stromproduktion durch die PV-Analge zur Verfügung. Ebenso stehen Wetterdaten der DWD-Stationen Augsburg und Lechfeld bereit. Basierend darauf sollen zwei Modelle erstellt werden:

- Ein Modell, das den Tages-Energiebedarf prognostiziert.
- Ein Modell, das den Verlauf der produzierten Photovoltaik-Energie vorhersagt.

Modelle dieser Art können für das Energie-Management im Rahmen einer Smart-Home-Lösung verwendet werden. Strategien zur Nutzung eines Akkus bzw. zur Steuerung von elektrischen (Groß-)verbrauchern können daraus abgeleitet werden.
Natürlich geht es im Rahmen dieser Studienarbeit nur um sehr einfache Modelle. Insbesondere sollen diese der Einfachheit halber *nicht* die Daten des aktuellen Tages explizit nutzen, um eine Vorhersage für den Folgetag zu errechnen.

Beschreibung der Daten

Für die gesamte Studienarbeit sollen ausschließlich folgende externe Daten verwendet werden.

PV-Daten Die Datei `PV.csv` enthält die vom Wechselrichter des Musterhauses bereitgestellten Daten ab dem 01.04.2022 in 15 Minuten Intervallen. Die Daten sind selbsterklärend beschriftet und in der Einheit Wh (Wattstunden) angegeben; die Daten in der Spalte „Ladezustand“ beziehen sich auf einen installierten Akku und sind als Prozentwerte angegeben. Sie sind für diese Arbeit nicht relevant.

Wetterdaten Die Dateien `wetter_A.csv` und `wetter_L.csv` enthalten aggregierte Daten des Deutschen Wetterdienstes für die Messstationen **A**ugsburg (ID 0232) bzw. **L**echfeld (ID 2905). Das sind alle Wetterdaten, die Sie benötigen.
Der Vollständigkeit halber finden Sie jedoch die Rohdaten, aus denen die beiden genannten `.csv` Files generiert wurden in den Ordnern `akt` (aktuelle Daten) und `hist` (historische Daten). Dort sind auch die Beschreibungen der Daten, der Messgeräte etc. enthalten.
Die Spalten der `.csv` Files haben folgende Bedeutung, siehe auch die entsprechenden Dateien `Metadaten_Parameter_*`:

F	Windgeschwindigkeit in $\frac{\text{m}}{\text{s}}$
D	Windrichtung in Grad
V_N	Bedeckungsgrad in Achtel
R1	Niederschlagshöhe in mm
RS_IND	Indikator Niederschlag ja/nein
TT_TU	Lufttemperatur in °C
RF_TU	Relative Luftfeuchte in %

Für die Messstation Lechfeld stehen keine Niederschlagsdaten zur Verfügung.

Prüfen und Aufbereiten der Daten

PV-Daten *Führen Sie diese Aufgabe im Notebook `daten_pv.ipynb` durch.*
Untersuchen Sie die Datenqualität der vorliegenden Photovoltaik-Daten. Dokumentieren Sie Ihre Ergebnisse (rein in Textform) knapp in der Zelle `A1 - Datenqualität PV`. Exportieren Sie die (möglicherweise aufbereiteten) Daten in eine Datei `PV.csv`, die analog zur Originaldatei aufgebaut ist.

Wetter-Daten *Führen Sie diese Aufgabe im Notebook `daten_wetter.ipynb` durch.*
Bereiten Sie die Wetter-Daten für die weitere Verwendung auf. Da das Musterhaus zwischen den Messstationen Lechfeld und Augsburg liegt, soll für alle Wetterdaten, die von beiden Stationen vorliegen, der Mittelwert aus den beiden Messungen verwendet werden. Für Daten, die nur von einer Station zur Verfügung stehenen, soll entsprechend dieser Wert direkt genutzt werden.
Prüfen Sie insbesondere auch, ob Daten zu einzelnen Zeitpunkten fehlen und füllen Sie diese etwaigen Lücken in einer inhaltlich sinnvollen Art. Nutzen Sie an geeigneter Stelle die Methode `interpolate`. Behandeln Sie auch möglicherweise enthaltene fehlerhafte Daten (ein Wert von z.B. -120 für die relative Luftfeuchte muss offensichtlich fehlerhaft sein).
Dokumentieren Sie alle durchgeführten Anpassungen strukturiert und knapp in der Zelle `A2 - Datenqualität Wetter`. Exportieren Sie die aufbereiteten Wetterdaten in eine Datei `wetter.csv`, analog zur Beispieldatei `wetter_dummy.csv`.

Modell 1 - Ermittlung des Energiebedarfs

Führen Sie diese Aufgabe im Notebook `model_energiebedarf.ipynb` durch.
Erstellen Sie ein lineares Regressionsmodell zur Vorhersage des täglichen Energiebedarfs. Als Features sollen die zur Verfügung stehenden Wetterdaten des jeweiligen Tages zu den Zeitpunkten 0:00 Uhr, 4:00 Uhr, 8:00 Uhr, ..., 20:00 Uhr genutzt werden. Die Daten zum Niederschlag sollen nicht berücksichtigt werden. Zur Erzeugung dieser Features ist die Methode `pivot_table` nützlich. Als weitere Features sollen Indikatoren für den Wochentag sowie die Tageslänge verwendet werden.
Reichern Sie die Features polynomiell bis zum Grad 2 an. Da die einzelnen Features bereits ohne diese quadratische Anreicherung unterschiedliche Skalen hatten (z.B. ist der Bedeckungsgrad zwischen 0 und 8 [Achtel], die Windrichtung aber zwischen 0 und 360 [Grad]), gehen die Skalen der quadratisch angereicherten Daten sehr weit auseinander. Daher müssen die einzelnen Features nach dem Anreichern skaliert werden. Dafür können Sie z.B. den `StandardScaler` verwenden.
Nun soll als regularisiertes lineares Modell Lasso verwendet werden. Bestimmen Sie per Cross Validation den besten Wert für den Regularisierungsparameter $\alpha \in \{1000, 500, 100, 50, 10, 1\}$ in Bezug auf den RMSE (Root Mean Squared Error). Trainieren Sie das entsprechende Modell auf allen vorliegenden Daten und geben Sie den RMSE Score an. Nutzen Sie das Modell, um den Hausverbrauch an den in den Daten fehlenden Tagen zu schätzen. Erstellen Sie einen Plot des zeitlichen Verlaufs des geschätzten und tatsächlichen Energiebedarfs. Exportieren Sie den vom Modell geschätzten täglichen Energiebedarf in die Datei `model_energiebedarf.csv`, analog zur vorliegenden Beispieldatei.

Modell 2 - Ermittlung des PV-Ertrags

Führen Sie diese Aufgabe im Notebook `model_PV.ipynb` durch.
Erstellen Sie ein Random Forest Modell, das basierend auf den Wetterdaten die Produktion an Photovoltaik-Energie pro Stunde prognostiziert. Ein neben dem Wetter wesentliches Feature ist dabei die extraterrestrische solare Einstrahlung, also die Strahlung, die direkt

von der Sonne auf die Erde trifft. Ein stark (!) vereinfachtes Modell für den Verlauf dieser Einstrahlung zur Zeit t ist gegeben durch

$$E_{\text{solar}}(t) = S\left(\sin(\phi)\sin(\delta) + \cos(\phi)\cos(\delta)\sin\left(\frac{2\pi(t - t_{\text{Mittag}})}{24}\right)\right).$$

Dabei ist S die durchschnittliche Strahlungsleistung der Sonne (ca. 1360 $\frac{\text{W}}{\text{m}^2}$). Diese setzen wir im Modell jedoch auf 1, da wir nur den relativen Verlauf als Feature benötigen. ϕ bezeichnet den Breitengrad. δ gibt die Sonnendeklination an, also den Winkel unter dem die Sonnenstrahlen auf die Erdoberfläche treffen. Sie lässt sich (in Grad) grob annähern durch

$$23.45° \sin\left(\frac{284 + d}{365} \cdot 360°\right),$$

wobei d die Zahl des Tages im Jahr ist (vgl. `tm_yday`). t ist die Tageszeit.
Dass dieses Modell stark vereinfacht ist, lässt sich schon alleine daran sehen, dass auch für Zeiten, die in der Nacht liegen, eine positive Einstrahlung berechnet wird. Eine exaktere Bestimmung des Verlaufs der solaren Einstrahlung z.B. unter Berücksichtigung von Sonnenauf- und -untergangszeiten soll jedoch nicht durchgeführt werden.
Verwenden Sie die solare Einstrahlung E_{solar} als weiteres Feature und bestimmen Sie mittels Cross Validation eine geeignete Anzahl `n_trees` $\in \{10, 50, 75, 100, 200, 500\}$ an Bäumen für einen Random Forest Regressor. Verwenden Sie als Fehlermetrik RMSE. Trainieren Sie anschließend das Modell auf allen verfügbaren Daten und erzeugen Sie geeignete Plots der berechneten und gemessenen Photovoltaik-Produktion. Exportieren Sie den stündlichen Photovoltaik-Ertrag in die Datei `model_PV.csv`, analog zur vorliegenden Beispieldatei.