Duale Hochschule Baden-Württemberg

Data Exploration Project

Wettervorhersage

Kurs Wirtschaftsinformatik

Studienrichtung Data-Science

Verfasser(in):	Rico Joel Siegelin, Nils-Jannik Klink, Marvin Spurk
Matrikelnummer:	6577951, 2538158, 1573694
Kurs:	WWI-20-DSB
Studiengang:	Wirtschaftsinformatik Data Science
Abgabedatum:	13.07.2022

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Motivation

Unser Projekt setzt sich mit dem Thema der Wettervorhersage auseinander. Jeden Tag schauen wir uns die Wettervorhersage für heute oder die nächsten Tage an und verlassen uns auf diese Vorhersage. Auch laut einer Studie aus den USA vom Jahr 2009, schauen sich Erwachsene in den USA jährlich rund 300 Mrd. Mal die Wettervorhersagen. Da dieses Thema also auch aus unserem Alltag kommt, fanden wir es umso interessanter uns selbst mit diesem Thema zu beschäftigen und eine eigene Vorhersage zu machen. Außerdem könnten solche Vorhersagen wichtig werden, sodass man im Notfall auch gefährliche Unwetter/Katastrophen im Voraus erkennen kann. Diese Arbeit beschränkt sich jedoch auf die Vorhersage der Temperaturentwicklung. Der Klimawandel schreitet weiter voran und wir haben es selbst mit Temperaturen von bis zu 40°C in Deutschland zu tun. Demnach geben wir auch ein Ausblick auf die Temperaturentwicklung für ein Jahr. Zudem dachten wir, dass es sinnvoll wäre, auch die Klimaentwicklung das nächste Jahr mit aufzuzeigen.

2. Related Work

Der Datensatz sowie die Datenaufzeichnung stammen alle vom Deutschen Wetter Dienst. Der Deutsche Wetter Dienst setzt auf Modelle mit numerischer Modellierung. Beim Einsatz von NWV-Modellen unterscheidet man dann nochmals in deterministische und probabilistischer Vorhersagerechnung.
Die **deterministische Vorhersage,** berechnet das Modell ausgehend vom Anfangszustand eine zukünftige Weiterentwicklung.

Für die **probabilistische Vorhersage** (Ensemble-Vorhersage) werden gleichzeitig, ausgehend von mehreren leicht unterschiedlichen Anfangszustände, einfach mehrere Modelle aufgestellt. Hiermit kann man verschiedene Entwicklungen erfassen, um bei Unsicherheiten die Entwicklung besser abschätzen zu können.

Des Weiteren gibt es eine Studie von Weyn (Vgl. „The AI forecaster: Machine learning takes on weather prediction”), die auf einem neuen Ansatz mit neuronalen Faltungsnetzwerkes basierend. Diese maschinelle Wettervorhersagesystem nennt sich **Deep Learning Weather Prediction (DWLP)**. Dieses Modell wird anhand von historischen Wetterdaten trainiert. Diese basiert nämlich auf der mathematischen Darstellung von physikalischen Gesetzen. Jedoch weist diese Methode zum aktuellen

Zeitpunkt, nur eine hohe Genauigkeit der Vorhersage in einem Zeitraum bis zu 4,5 Tage auf. Sobald eine Vorhersage für 1-2 Wochen erstellt werden soll, schwindet die Genauigkeit drastisch. Aus diesem Grund, wurde dieser Ansatz nicht weiterverfolgt und sich für eine andere Methode entschieden.

3. Verwendete Technologien

Die verwendeten Bibliotheken sind Pandas, Numpy, Mathplotlib, Sklearn & Datetime. Die Support Vektor Machine (SVM) ist unsere gewählte Methode. Diese mathematische Methode des maschinellen Lernens, dient der Klassifikation von Objekten. Anhand eines gesplitteten Trainings-/Testdatensatz, wird entweder das Modell trainiert oder getestet. Anhand des Scores, wird die Güte des Modells bestimmt. Das Modell, versucht jetzt in den bereits bekannten Zuordnungen in den Trainingsdaten sog. Trennlinien bzw. Trennflächen zu finden. Alles was im Testfall beispielsweise über dieser Linie liegt gehört zur Klasse 1 und alles darunter, zur Klasse 2. Das Ganze ist dann noch im mehrdimensionalen Fall anwendbar mit mehr als 2 Klassen.

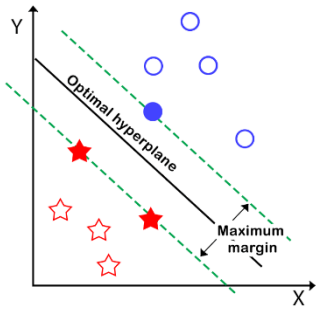

Abbildung 1: SVM Beispiel Modell

4. Ergebnisse

Da die Wetterdaten ab dem 01.01.1948 jede Stunde erfasst wurden, sind insgesamt 648.672 Dateneinträge für unser Projekt vorhanden. Abzüglich der Schaltjahre, welche entfernt werden um ein besseres Modell zu trainieren, verbleiben noch 648.240 Dateneinträge. Sobald mit dieser großen Datenmenge das Modell trainiert werden soll, dauert dies mindestens länger als 17 Stunden. Nach dieser Zeit, wurde das Training abgebrochen, da noch kein fertiges Modell zustande kam. Daher ist die kritische Frage zu stellen, wie groß der Datensatz sein muss. Des Weiteren muss Gamma, welche den x-Wert beschreibt, gewählt werden. Um das richtige Gamma für dieses Projekt zu erhalten, wurden dafür einmal 0,0001 und einmal 0,1 gewählt. Um Zufälle bezüglich

der Datengröße ausschließen zu können, wurde zusätzlich als Datenmenge 4 Jahre, 7 Jahre und 20 Jahre gewählt. Der Vorhersagezeitraum beträgt 365 Tage, also 1 Jahr.

Score in %	4 Jahre	7 Jahre	20 Jahre
Gamma: 0,0001	47,72%	51,58%	55,29%
Gamma: 0,1	48,84%	52,11%	55,76%

Die Ergebnisse des Scores, welche die Genauigkeit zwischen Realdaten und vorgesagten Daten darstellt, zeigt, dass je größer der Datensatz und das Gamma sind, desto besser ist der Score. Daher wurde fortfolgend als Datensatzgröße die letzten 20 Jahre und für Gamma 0,1 gewählt.

Bei Betrachtung des Datensatzes sind Temperaturwerte von –999 °C aufgefallen. Diese Dateneinträge wurden alle gelöscht. Jedoch gibt es noch eine weitere Lösung, um mit diesem Problem umzugehen. Es existiert der sogenannte „Forward Fill“, welcher NaN-Values, durch den vorherigen Wert ersetzt. Wir nehmen an, dass wir einen Datensatz haben der wie folgt aussieht: [1,4,3,7,8, NaN,6,9, NaN,0]. Wird dieser nun mithilfe des Forward Fill bearbeitet, sieht der Datensatz wie folgt aus: [1,4,3,7,8,8,6,9,9,0]. Somit kann das Modell, welches trainiert werden soll, besser mit diesen Dateneinträgen umgehen und eine genauere Vorhersage treffen. Dies wird durch unser Modell bewiesen. Der Vorhersagezeitraum beträgt 365 Tage, also 1 Jahr. Für die Datenbasis wurden die letzten 20 Jahre gewählt.

Score in %	-999 Werte -> löschen	-999 Werte -> Forward Fill
Gamma: 0,0001	55,29%	58,43%
Gamma: 0,1	55,76%	58,83%

Die Datenbasis, das Gamma und falsche Werte wurden bis jetzt als Problem erkannt und optimiert. Fortfolgend wird nun das Modell auf verschiedene vorherzusagende Zeiträume trainiert und getestet. Bis jetzt wurde als Zeitraum für die Vorhersage immer 1 Jahr gewählt. Um den Score zu vergleichen, werden im Folgenden die Zeiträume 1 Stunde und 1 Monat vorhergesagt. Für die Datenbasis, wurden die letzten 20 Jahre und für Gamma wurde 0,1 gewählt. Die falschen –999 Werte, wurden durch den Forward Fill ausgetauscht.

Vorhersagezeitraum	1 Jahr	1 Monat	1 Stunde
Score:	58,83%	47,10%	97,97%

Bei diesem Vergleich fällt auf, dass ein sehr kurzer Vorhersagezeitraum von einer Stunde eine hohe Genauigkeit aufweist. Des Weiteren kann ein Zusammenhang zwischen der Genauigkeit und dem Zeitraum der Vorhersage festgestellt werden. In diesem Fall, fällt die Vorhersage für den kompletten Dezember 2021 aus dem Raster heraus. Um das ganze weiter zu verfolgen, stellen wir das Modell um. Aktuell benötigt das Modell ca. 70 Minuten bis es vollständig trainiert ist. Der Datensatz umfasst jede Stunde der letzten 20 Jahre und besteht somit aus 175.200 Dateneinträgen. Um diesen auf 7.300 Dateneinträge zu verkleinern, wird für jeden Tag die minimale und die maximale Temperatur berechnet und abgespeichert. Fortfolgend konzentrieren wir uns nur auf den Datensatz mit den minimalen vorverarbeiteten Temperaturwerten. Um zu beweisen, dass der Datensatz mit der Mindesttemperatur eine vergleichbare Aussagekraft wie der Datensatz mit jeder Stunde der letzten 20 Jahre hat, werden diese nun beide verglichen. Damit das Modell schneller trainiert, wird eine Datenbasis von 6 Jahren für den Vergleich und für Gamma 0,1 gewählt. Die falschen –999 Werte , werden durch den Forward Fill ausgetauscht. Der Vorhersagezeitraum beruft sich auf 7 Tage.

Datenbasis:	6 Jahre, jede Stunde	6 Jahre, minimal Wert jeden Tages
Score:	62,88%	62,19%

Daraus ist zu schließen, dass der Datensatz mit der Mindesttemperatur eine vergleichbare Aussagekraft, wie der Datensatz mit jeder Stunde der letzten 20 Jahre hat. Somit können nun die unterschiedlichen Vorhersageräume miteinander verglichen werden, ohne dass dies mehrere Stunden dauert, da nur noch wenige Minuten benötigt werden um das Modell zu trainieren.

Für die Datenbasis, wurden die minimalen Temperaturen für jeden Tag der letzten 20 Jahre und für Gamma wird 0,1 gewählt. Die falschen –999 Werte werden durch den Forward Fill ausgetauscht. Unsere Scores im Überblick auf die verschiedenen Vorhersagezeiträume auf der Datenbasis der Mindesttemperatur eines Tages:

Zeiträume:	1 Tag	7 Tage	14 Tage	1 Monat	1 Jahr	4 Jahre
Score:	**97,9%**	**55,8%**	**51,9%**	**40,3%**	**58,4%**	**53,1%**

Auch hier fällt die Vorhersage für den gesamten Dezember 2021 aus dem Raster. Jedoch ist der Score visuell anschaulicher, um die Performance eines Modells zu erkennen und diese mit anderen Modelle zu vergleichen. Ein besseres Verständnis für die Daten, kann hier nicht erlangt werden. Hierfür eignet sich die visuelle Ausgabe der erzielten Temperaturdaten, im Vergleich zu den wirklichen Temperaturdaten besser.

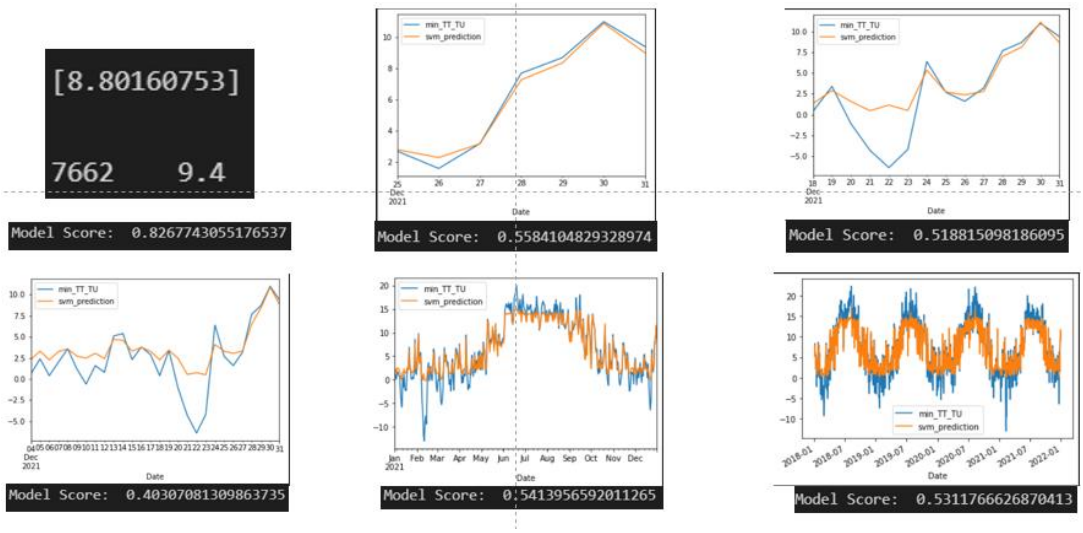

Abbildung 2: Vergleich der Scores

Nun ist zu erkennen, dass die Modelle recht nah an den tatsächlichen Werten liegen und durch Ausreißer der Modellscore verfälscht wird. Ebenfalls ist zu erkennen, dass eine Vorhersage für mehr als 4 Jahre keinen Sinn ergibt, da eine Sinusförmige Formel durch das Modell abgeleitet wird und diese sich durch Erhöhung des Vorhersagezeitraums nur verfestigen würde.

5. Kritische Bewertung (Lessons Learned)

Über die Dauer des Projektes ist uns vieles klarer geworden, was das arbeiten mit großen Datenmengen und vor allem das Arbeiten mit machine learning Algorithmen angeht. Da wir unsere Daten nicht von Kaggle o.Ä., sondern vom Deutschen Wetterdienst bezogen haben, mussten wir diese noch vorverarbeiten, bereinigen und labeln. Wir mussten entscheiden, was für unseren konkreten Fall von Bedeutung ist. So war es zum Beispiel nach längerer Überlegung nicht notwendig, alle Daten von 1948 bis 2021 zu verwenden. So haben wir uns im Folgenden zunächst auf 20, 7 oder 4 Jahre beschränkt um diese Theorie zu verfolgen. Nachdem wir alle Ergebnisse der Modelle miteinander verglichen haben, konnten wir feststellen, dass eine Datenbasis

von den letzten 20 Jahren die besten Ergebnisse erzielt. Anzumerken ist jedoch, dass hier kein Optimum gefunden wurde, sondern nur eine Verbesserung erzielt und eine Kenntnis gewonnen. Trotzdem dauerte es noch zu lange, bis der Algorithmus durchgelaufen ist. Deshalb haben wir uns des Weiteren auf die Minimalwerte beschränkt. Der Algorithmus brauchte so nur noch wenige Sekunden bis Minuten, je nach Vorhersagedauer. Daraus nehmen wir also mit, dass in unserem Fall die Datenbasis stark ausschlaggebend dafür ist, wie gut die Vorhersage ist. Somit haben wir einen minimal schlechteren Score in Kauf genommen, jedoch konnten wir so mehr Vorhersagen treffen und die Trainingsdauer reduzieren.

Des Weiteren haben wir im Fehlerhafte Werte bzw. Fehlaufzeichnungen gefunden, die den Algorithmus zu stark beeinflusst haben. Zunächst wollten wir die Daten rauswerfen, haben jedoch dann den Forward Fill entdeckt, der die fehlerhaften Werte mit dem vorherigen Wert auffüllt. Hiermit konnten wir Ausreißer entfernen, ohne wirklich den Datensatz zu löschen.So haben wir eine Verbesserung von 52% auf 58% erzielt. Damit ist festzuhalten, dass nicht einfach willkürlich Dateneinträge gelöscht werden dürfen, auch wenn diese fehlerhaft sind.

Außerdem haben wir die Sinnhaftigkeit von längeren Vorhersagen angezweifelt. Vorhersagen über mehr als eine Woche sind schon weniger sinnvoll. Jedoch sind Vorhersagen von einem Zeitrahmen, der noch größer als eine Woche ist, noch viel ungenauer. Wie oben bereits erwähnt, sind die Vorhersagen dafür sehr ungenau und man kann sich schwer darauf verlassen. Außerdem können auch unvorhergesehene Dinge passieren, wie Temperatureinstürze von 12°C von einem auf den nächsten Tag. Hier würde jedes Modell an seine Grenzen stoßen, dies langfristig vorherzusagen. Außerdem sind alle Wettervorhersagen, die über den Zeitraum von 1 Monat hinaus gehen, nicht relevant, da sie weder im Wetterbericht, noch in verschiedenen Wetterapps aufgeführt werden.

Des Weiteren haben uns die Datenvisualisierungen gezeigt, wie unsere Daten aussehen und uns gelehrt, dass nur der Score eines Modelles nicht aussagekräftig genug ist. Durch die visuelle Darstellung der Vorhersagedaten und die Gegenüberstellung der realen Daten wird erkennbar, dass unser Modell recht genaue Vorhersagen trifft, jedoch die Temperatur von Tagen, welche kurzfristig einen hohen oder niedrigen Ausschlag haben, nicht vorhersagen kann. Somit kann unser Modell einen Trend der Temperatur vorhersagen, jedoch nicht jeden Tag zu 100%. Deshalb

sollte sich nicht ausschließlich auf unser Modell verlassen werden, wenn man das Haus verlässt oder die nächste Geburtstagsfeier plant.

Abschließend ist zu erwähnen, dass wir für langfristige Vorhersagen besser tägliche Daten, statt stündliche Aufzeichnung ausgewählt hätten. Diese sind zwar nicht so feingranular, jedoch sinnvoller für langfristige Vorhersagen. Hiermit hätte man die Datenbasis durch 24 teilen können, ohne einen großen Aufwand zu betreiben.

6. Ausführung des Codes

Für die Ausführung des Codes sind die unter Punkt 2 aufgeführten Bibliotheken zu installieren, falls diese nicht vorhanden sind. Des Weiteren muss das gesamte Projekt von GitHub heruntergeladen werden. Es existieren zwei ausführbare jupyter Notebooks, welche einmal auf das Modell mit Temperaturdateneinträgen von jeder Stunde und einmal mit der minimalen Temperatur jeden Tages abgestimmt sind. Durch das ändern der Jahreszahl für den Datensatz, das ändern des Gammas für die SVM und durch ändern des Vorhersagezeitraums können somit alle Ergebnisse reproduziert werden.