Trainingsdaten

Bei den Trainingsdaten habe ich mir überlegt, was für Werte relevant sein könnten, um ein maschinell lernendes Modell zu trainieren, welches die Anzahl an Waldbränden pro Monat vorhersagen kann.

In einem maschinellen Lernmodell sind die Trainingsdaten von zentraler Bedeutung. Sie setzen sich zusammen aus einem bestimmten Wert, den das Modell vorhersagen soll, bezeichnet als „Zielwert“ oder „abhängige Variable“. Neben diesen Werten benötigt das Modell diverse „Features“ oder „unabhängige Variablen“, die entscheidenden Informationen liefern, welche das Modell zur Prognose des Zielwertes heranzieht. Die Auswahl der Features ist von kritischer Bedeutung, denn sie hat unmittelbaren Einfluss auf die Genauigkeit der vom Modell getroffenen Vorhersagen. Es ist daher unerlässlich, eine angemessene Menge an relevanten und aussagekräftigen Features sorgfältig auszuwählen und in das Modell zu integrieren, um dessen maximale Effektivität sicherzustellen.[1][2]

Bei der Auswahl der notwendigen Daten habe ich festgestellt, dass ich die monatliche Anzahl der Waldbrände in jeder Region sowie die dazugehörigen Wetterdaten benötige, da diese in direktem Zusammenhang mit der Entstehung von Waldbränden stehen. Es ist sinnvoll, kleinere Bereiche wie Bundesländer statt ganzer Länder wie Deutschland zu betrachten, da sich das Wetter im Norden Deutschlands stark von dem Wetter im Süden unterscheiden kann. Die Wetterdaten für ein zu großes Gebiet wären nicht aussagekräftig genug.

Des Weiteren ist ein kürzerer Zeitraum für die Datenbasis empfehlenswert. Statt ein ganzes Jahr sollte man besser einen Monat betrachten, um die Wetterdaten effektiver in das Modell integrieren zu können. Würde man beispielsweise die Anzahl der Waldbrände pro Jahr und die dazugehörigen durchschnittlichen Wetterdaten verwenden, könnte dies zu ungenauen Vorhersagen führen. Die Durchschnittstemperatur eines ganzen Jahres würde auch die Wintermonate umfassen, in denen Waldbrände weniger wahrscheinlich sind als im Sommer.

[1] (cloudfactory, 2023)
[2] (learn.g2, 2023)

Ich habe dann nach aussagefähigen Statistiken gesucht zum Thema Anzahl an Waldbränden pro Monat. Dort bin ich dann auf die Statistiken von der Bundesanstalt für Landwirtschaft und Ernährung gestoßen welche Waldbrandstatistiken von 2010 bis 2022 zur Verfügung stellt. Diese Statistiken enthielten zahlreiche Informationen zu Waldbränden in Deutschland. Unteranderem auch die Anzahl an Waldbränden pro Monat pro Bundesland.

Nachdem ich dann die Werte für die Zielparameter hatte, habe ich mich dann auf die Suche nach einer Wetter-API gemacht.

Eine Wetter-API (Application Programming Interface) ist eine Schnittstelle, die es Entwicklern ermöglicht, auf Wetterdaten von einem externen Dienst zuzugreifen. Diese Daten können Temperatur, Luftfeuchtigkeit, Niederschlagsmenge, Windgeschwindigkeit, Windrichtung, Luftdruck und andere Wetterbedingungen umfassen. Es kann sich auch um andere Daten handeln, wenn es eine andere API ist.

Dabei war es mir persönlich sehr wichtig, eine Wetter-API zu finden, die Wetterdaten seit 2010 anbietet, da ich die Waldbrandstatistiken ab diesem Jahr habe und es wünschenswert wäre, so viele Datensätze wie möglich zu erstellen. Außerdem sollte die Wetter-API möglichst viele relevante Wetterparameter anbieten.

Nach einiger Zeit fand ich die API weatherapi.com[3], welche seit Januar 2010 Wetterdaten anbot. Bei dieser API musste ich mir aber, um auf die historischen Wetterdaten zugreifen zu können, ein Business Account für 65 Dollar kaufen, worauf ich 10% Rabatt bekam, da ich ein Schüler bin.

Nachdem ich Zugang zur Wetter-API erhalten hatte, übertrug ich die Daten aus den Waldbrandstatistiken in eine Excel-Tabelle im folgenden Format: Startdatum, Enddatum, Ort, Anzahl der Waldbrände. Dadurch ergaben sich insgesamt 2496 Datensätze. Zur Tabelle fügte ich noch die Spalten 'Jahreszeit' und 'Dauer' hinzu. Für die metrologischen Jahreszeiten verwendete ich das folgende Schema: Frühling (März bis Mai), Sommer (Juni bis August), Herbst (September bis November), Winter (Dezember bis Februar). Im nächsten Schritt verwendete ich einen agilen Ansatz, indem ich mehrere Testprogramme schrieb. Aus der Excel-Tabelle erstellte ich mehrere Testdatensätze, welche ich als CSV-Datei exportierte, um

[3] (weatherapi.com, 2023)

sie in meinem Programm einlesen zu können. Nach zahlreichen erfolgreichen Tests entwickelte ich das Hauptprogramm, welches die CSV-Datei einliest, daraufhin die Wetterdaten abruft und schließlich die Ergebnisse in der CSV-Datei abspeichert. Dies erfolgte, indem das Programm Startdatum, Enddatum und Ort aus der CSV-Datei extrahierte, welche anschließend in einer Liste abgelegt wurden. Für jeden Eintrag in der Liste bezog das Programm die historischen Wetterdaten des jeweiligen Ortes für jeden Tag im Zeitraum zwischen Start- und Enddatum.

Die Datenpunkte durchschnittliche Temperatur (in Celsius), maximale Windgeschwindigkeit (in km/h), durchschnittlicher Niederschlag (in mm) und durchschnittliche Luftfeuchtigkeit (in %) wurden jeweils in eigenen Listen gespeichert. Auf Basis dieser Daten wurden anschließend die Durchschnittswerte für den Monat berechnet und der CSV-Datei hinzugefügt, welche am Ende des Prozesses gespeichert wurde. Dann startete der gesamte Vorgang von vorn für den nächsten Eintrag in der Liste. Während der zahlreichen API-Anfragen kam es gelegentlich vor, dass für bestimmte Monate keine Wetterdaten verfügbar waren oder dass einzelne Tage aufgrund von Timeouts oder anderen Fehlern nicht abgefragt werden konnten. Um diese Probleme zu beheben, schrieb ich ein zusätzliches Programm, welches die fehlenden Werte für die einzelnen Monate erneut abfragte und berechnete.

Die berechneten Durchschnittswerte wiesen häufig viele Nachkommastellen auf, daher entschied ich mich bewusst gegen eine Rundung, um die Genauigkeit der Modelle zu erhöhen und somit einen effektiveren und verantwortungsbewussteren Beitrag zu leisten. Zudem entwickelte ich ein Programm zur Kodierung der Jahreszeiten, da maschinelle Lernmodelle ausschließlich mit numerischen Werten arbeiten. Die Kodierung erfolgte wie folgt: Frühling = 1, Sommer = 2, Herbst = 3 und Winter = 4.

Abschließend hielt ich es für notwendig, ein Programm zu schreiben, das die Waldfläche der einzelnen Orte zu der CSV-Datei hinzufügt. Dabei stieß ich allerdings auf das Problem, dass ich die Waldfläche in Hektar nur für die Jahre 2016 und 2021 für die einzelnen Bundesländer finden konnte. Für die Jahre 2016 bis 2020 verwendete ich die Waldflächendaten von 2016, für 2021 und 2022 die Daten von 2021. Trotz dieser Einschränkung entschied ich mich für die Verwendung dieser Werte, da ich überzeugt war, dass sie zur Genauigkeit der Modelle beitragen würden, weil die Größenordnung der Waldfläche stimmig war.

Nun waren die 2480 Trainingsdatensätze bereit zur Verwendung.

Literaturverzeichnis

cloudfactory. (19. 11 2023). Von https://www.cloudfactory.com/training-data-guide abgerufen

learn.g2. (19. 11 2023). Von https://learn.g2.com/training-data abgerufen

weatherapi.com. (19. 11 2023). Von https://www.weatherapi.com/ abgerufen