Feature Engineering und Zeitreihen

Im Rahmen des Feature Engineering werden vier neue Features erstellt. Zunächst wird jeder Person ein BMI zugeordnet. Dieser entspricht jeweils dem Median der einzelnen Klassifikationen. Untergewichtigen Personen wird ein BMI von 17, normalgewichtigen Personen ein BMI von 21,7, übergewichtigen Personen ein BMI von 27,5 und adipösen Personen ein BMI von 35 zugewiesen. Aus der Körpergröße und dem BMI lässt sich daraufhin das Körpergewicht berechnen. Dieses setzt sich aus der quadrierten Körpergröße mal BMI geteilt durch 10.000 zusammen. Zusätzlich werden die Features basierend auf dem Ernährungsverhalten und die Features basierend auf der körperlichen Aktivität zusammengefasst. Das Feature „FoodConsumption“ addiert den Fast-Food-Konsum mit dem Faktor für die Menge der täglichen Mahlzeiten und der Prävalenz von Zwischenmahlzeiten und subtrahiert davon die tägliche Wasseraufnahme und die Häufigkeit des Gemüseverzehrs. Das Merkmal „Aktivität“ addiert die Faktoren für die Häufigkeit der körperlichen Aktivität und das bevorzugte Verkehrsmittel und subtrahiert davon die Bildschirmzeit. Ein niedriger Wert für „FoodConsumption“ und ein hoher Wert für „Activity“ kann somit als positiv gewertet werden.

Zwischen dem Verzehr von Fast Food und körperlicher Aktivität besteht eine mäßige Korrelation: Personen, die häufig Fast Food konsumieren, sind im Durchschnitt an fünf bis sechs Tagen pro Woche körperlich aktiv, während Befragte, die angaben, kein Fast Food zu konsumieren, im Durchschnitt an drei bis vier Tagen pro Woche körperlich aktiv sind. Zwischen Fast-Food-Konsum und Gewicht besteht ein leichter Zusammenhang. Während Personen, die Fastfood konsumieren, im Durchschnitt etwa 80 kg wiegen, liegt das Durchschnittsgewicht bei Personen, die kein Fastfood konsumieren, bei etwa 75 kg. Zwischen der Summe der Ernährungsfaktoren und dem Gewicht besteht ein mäßiger Zusammenhang. Personen mit höherem Gewicht essen mehr Fastfood, haben mehr Hauptmahlzeiten oder essen weniger Gemüse. Im Gegensatz dazu gibt es keine klare Korrelation zwischen den summierten Faktoren körperliche Aktivität und Gewicht. Schwere Personen scheinen insgesamt ähnlich aktiv zu sein wie leichte Personen.

Da die Datensätze keine Informationen zu Kalenderdaten enthalten, werden die Aufgaben zu den Time Series mit anderen Daten bearbeitet (Shandeep Raula 2024). Diese zeigen die monatlichen Verkäufe eines Amerikanischen Unternehmens in vier Regionen (Central, East, South, West). Die Verkaufszahlen schwanken stark, wobei die Regionen "East" und "West" insgesamt höhere Trends aufweisen. "Central" zeigt moderate Steigerungen, während "South" relativ flach bleibt. Insgesamt steigt der Verkaufstrend in allen Regionen leicht an.

Quelle der Daten für die Time Series:

Shandeep Raula (2024): Retail Supply Chain Sales Dataset. kaggle. Online verfügbar unter https://www.kaggle.com/datasets/shandeep777/retail-supply-chain-sales-dataset.