Regression Challenge

Aufgabe

Sie arbeiten in ein Unternehmen, das Schmuck direkt an Endkonsumenten über einen Online-Shop verkauft. Innerhalb der letzten 5 Jahre ist das Unternehmen rasant gewachsen. Die Geschäftsführung stellt nun fest, dass es notwendig ist, zentrale Prozesse zu optimieren und mit geeigneten Tools zu unterstützen. Ein solcher Prozess ist das Demand Planning. Hierbei geht es darum, die zukünftige Nachfrage nach Produkten zu prognostizieren, um Warenbestände, Lieferketten und Ressourcen optimal zu steuern. Ziel ist es, das richtige Produkt in der richtigen Menge zur richtigen Zeit am richtigen Ort verfügbar zu haben. Bisher wurde dies eher nach Bauchgefühl gemacht. Sie haben nun die Aufgabe ein ML-Modell zu entwickeln, welches die monatlichen Absätze von Produkten für die nächsten drei Monate vorhersagt.

Als Grundlage für die Prognose erhalten Sie zwei Datensätze. In der Datei „*MasterData.csv*“ finden Sie die Stammdaten zu den Produkten, für die die zukünftigen Absätze zu prognostizieren sind. Die Spalten dieser Tabelle sind die folgenden:

- *TimeSeriesId*: Eindeutiger Kenner für die Zeitreihe eines Produktes
- *Region*: Region der Absätze (Märkte)
- *Product_Hierachy1*: 1-te Produkthierarchie (z.B. Kette)
- *Product_Hierachy2*: 2-te Produkthierarchie (z.B. Halskette)
- *Product_Hierachy3*: 3-te Produkthierarchie (z.B. Klassik)
- *ABC_Classification*: ABC-Klassifikation nach Produktabsätzen, d.h. A-Produkte = High-Runner und C-Produkte = Low-Runner
- *XYZ_Predictability*: XYZ-Vorhersagbarkeit, d.h. X-Produkt = gleichmäßiger Verlauf und Z-Produkt = sporadischer Verlauf
- *Price_Normalized*: Normalisierter Produktpreis zw. 0 und 1
- *First_Goods_Issue_Date*: Datum des ersten Verkaufs
- *Effective_Out_Date*: Datum des letzten Verkaufs

Die eigentlichen (historischen) Verkaufszeitreihen finden Sie in der Datei „*Sales.pkl*“. Sie finden dort Tagesabsätze im Zeitraum 2018-04-01 bis 2022-08-31 für insgesamt 6638 Produkte. Eine Zeitreihe ist eine geordnete Folge von Datenpunkten, hier Absätzen, die in regelmäßigen Zeitintervallen erfasst wurden, um Entwicklungen oder Muster über die Zeit hinweg zu analysieren. Die Spalten in der Tabelle sind die folgenden:

- *TimeSeriesId*: Eindeutiger Kenner für die Zeitreihe eines Produktes
- *Date*: Verkaufsdatum
- *Quantity*: Verkaufsmenge

Evaluation

- Die Prognosegüte wird anhand des RMSE gemessen.
- Entwickeln Sie ein Modell zur Vorhersage zukünftiger Absätze, welches den RMSE optimiert.
- Machen Sie anschließend eine Prognose für den Zeitraum 2022-09 bis 2022-11 und exportieren Sie diese als csv-Datei. Es muss ersichtlich sein, welcher Zeitreihe (TimeSeriesId) und Monat sich ein Prognosewert zuordnet.

Abgabe

- Bearbeiten Sie die Aufgabenstellung in einem Colab-Notebook.
- Verschaffen Sie sich einen ausreichenden Überblick zu den Daten, z.B. durch visuelle Darstellungen oder Kennzahlen.
- Testen Sie verschiedene Modellvarianten – z.B. Modelle nach Region oder Produkthierarchie neben einem globalen Modell.
- Achten Sie darauf, dass Ihr Code gut lesbar und verständlich ist.
- Ein **Link auf Ihr Colab-Notebook** ist bis zum **23.03.25, 20 Uhr** an alexander.kressner@dhbw-stuttgart.de zu senden. Achten Sie bitte darauf, dass das Notebook nach laden der Daten fehlerfrei durchläuft und eine Prognose erstellt.
- Weiterhin erstellen Sie bitte eine **15-minütige Präsentation**, die Ihre Arbeitsergebnisse dokumentiert (Problemstellung, Lösungsansatz, Ergebnisse). Diese schicken Sie bitte als pdf-Datei ebenfalls bis zum **23.03.25, 20 Uhr** an alexander.kressner@dhbw-stuttgart.de
- Bei der **Abschlusspräsentation am 26.03.25** übernimmt bitte jedes Gruppenmitglied einen Präsentationsteil. Anschließend findet eine ca. 35–40-minütige Diskussion zur Fallstudie statt.