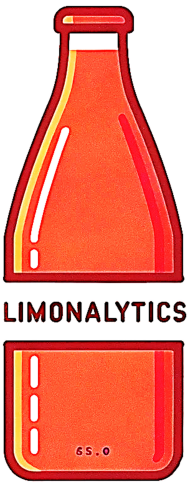
Supply Chain Analytics - Sommersemester 2024 - Hausaufgabe 3

Contents



1 Einführung

Es gelten weiterhin die Informationen, die bereits in Hausaufgabe 1 vorgestellt wurden. Stellen Sie sich vor, dass Sie nach den letzten Analyseergebnissen (HA 2) nun weitere Analysen durchführen möchten, die Ihre Entscheidungsfindungen weiter verbessern sollen.

Die Leistung von Logistikdienstleistern hängt stark von externen Faktoren ab, was verschiedene Ursachen hat. Zum Beispiel verfügen manche Logistikunternehmen über veraltete Fahrzeugflotten, die bei extremen Wetterbedingungen ausfallen oder nicht funktionieren, und Mitarbeiter, die bei extremen Wetterbedingungen nicht zur Arbeit kommen. Unterschiede in den operativen Abläufen oder der technologischen Unterstützung führen dazu, dass Faktoren wie Verkehrsstaus, Internetverfügbarkeit, Stromnetzstabilität, Parkplatzverfügbarkeit oder der Zustand der Infrastruktur unterschiedlich stark die Abläufe und somit die Verfügbarkeit der Logistikdienstleister beeinflussen.

Die externen Effekte, zu denen Ihr Unternehmen Daten sammelt, sind wie folgt:

Wetter

- `Temperature_C`: Durchschnittliche Temperatur im Monat in Grad Celsius
- `Rain_mm`: Summe des Niederschlags pro m^2 in mm im Monat
- `Sunshine_h`: Summe der Sonnenscheinstunden im Monat
- `Humidity`: Durchschnittliche Luftfeuchtigkeit im Monat

Infrastruktur

- `Congestion`: Durchschnittlicher Anstieg der Reisezeit in Spitzenzeiten in Prozent (abhängig vom Pendlerverkehr, Tourismus u.a.)
- `InternetStability`: Internet-Server pro 1 Mio. Einwohner (abhängig von Wartung, Reparatur und anderen)
- `PowerGridStability`: Stromausfalltage pro Monat (abhängig vom Ausfall der Stromversorgung)
- `ParkingSpaceAvailability`: Verfügbare Parkfläche in Prozent der Stadtfläche (abhängig von Veranstaltungen, Tourismus, Strassenerhaltung u.a.)
- `RoadCondition`: Gemessen an der 7-Punkt-Likert-Skala als Befragung von gewerblichen Fahrern (beeinflusst durch Wartung und andere)

Gesellschaft

- `PoliticalStability`: Gemessen an der 7-Punkte-Likert-Skala als Befragung von politischen Experten (beeinflusst durch aktuelle Entscheidungen, Gesetzesänderungen, politische Ereignisse u.a.)
- `AvgHealth`: Index in Prozent auf Grundlage einer Befragung der Bevölkerung zu Gesundheitsthemen (beeinflusst durch Krankheitsbelastung, Verletzungen, Risikofaktoren und andere)
- `Criminality`: Index in Prozent auf Grundlage von Bevölkerungsbefragungen und Statistiken zu kriminalitätsbezogenen Themen (beeinflusst durch wahrgenommene Kriminalität, Besorgnis über Raubüberfälle, Anzahl der Raubüberfälle pro Kopf, Anzahl der Gewaltverbrechen pro Kopf und andere)
- `AirPollution`: kt CO2-Emission pro Quadratkilometer pro Monat
- `WaterQuality`: Index in Prozent basierend auf der Erhebung der Bevölkerung über die Wasserqualität (beeinflusst durch Farbe, Geschmack, Behandlung vor dem Verzehr und andere)
- `leisureAndSocialInteractions`: Index von 1 bis 10 auf der Grundlage einer Bevölkerungsbefragung zum Thema Freizeit (beeinflusst durch soziale Aktivitäten, Armutsrisiko, Teilnahme an Freiwilligentätigkeiten und andere)

Arbeit

- `SkilledLaborAvailability`: Index in Prozent auf Basis einer Unternehmensbefragung zur Wahrscheinlichkeit, Fachkräfte für offene Stellen mit dem ersten Stellenangebot zu finden
- `UnskilledLaborAvailability`: Index in Prozent auf Basis einer Unternehmensbefragung zur Wahrscheinlichkeit, ungelernte Arbeitskräfte für offene Stellen mit dem ersten Stellenangebot zu finden
- `WorkerMotivation`: Gemessen an der 7-Punkte-Likert-Skala als Befragung von Mitarbeitern in der Logistik (Fach- und ungelernte Arbeitskräfte)
- `Overtime`: Durchschnittliche überstunden pro Tag auf Basis der Befragung von Mitarbeitern in der Logistik (Fach- und ungelernte Arbeitskräfte)

Wirtschaft

- `Inflation`: Monatliche Inflationsrate in Prozent
- `BusinessConfidence`: Geschäftsklimaindex auf der Grundlage der Erhebung der Unternehmen über Produktion, Aufträge und Bestände sowie der aktuellen Lage und der Erwartungen für die unmittelbare Zukunft. (normiert auf den langfristigen Durchschnitt = 100)
- `FuelPrice`: US$ Preis von Benzin pro Gallone

Wie die Logistikdienstleiter auf diese externen Effekte reagieren, ist Ihnen - bisher - nicht bekannt.

Für diese Hausaufgabe wurde der Einfluss der externen Effekte auf die Logistikdienstleister zufällig festgelegt. Dabei haben (zufällig) 1 bis 5 externe Effekte einen deutlichen Einfluss auf die Logistikleistung, während der Rest einen verschwindend geringen Einfluss hat.

2 Aufgaben

Hinweis: Die Hinweise aus HA 1 zur allgemeinen Bearbeitung der Aufgaben gelten weiterhin.

2.1 Datenanalyse (10 Punkte)

Sie haben bereits einige Erfahrung im Bereich der Verteilung von "Limonalytics" gesammelt. Obwohl Sie stets sorgfältig mehrere Monate im Voraus planen, variiert die Leistung der Logistikdienstleister erheblich. Deshalb sollen Sie ein Modell entwickeln, um deren Leistung unter Berücksichtigung externer Faktoren abzuschätzen. Dafür erhalten Sie Daten zu externen Faktoren für die Jahre 2019 bis 2024, wobei die zukünftigen Monate auf Prognosen basieren. Diese Vorhersagen sind selbstverständlich etwas ungenau und wurden mittels statistischer Verfahren erstellt. Dies ist jedoch für Ihre Analyse nicht entscheidend, sollte aber in Ihrer Entscheidung kurz *reflektiert* werden.

Daten für die Modellierung vorbereiten

1) Laden Sie die Datensätze `externals` und `services`. Erstellen Sie aus gesamten Datensatz `services` jeweils ein Dataframe für Shipping- und Warehousing-Dienstleistungen. Berechnen Sie anschließend für jede durchgeführte Dienstleistung die On-Time-Delivery Status (d.h. 0 oder FALSE, wenn unpünktlich; 1 oder TRUE wenn pünktlich) beziehungsweise die Item Fill Rate (IFR). Stellen Sie anschliessend jeweils die Kennzahlen der durchschnittlichen OTD-Rate und der durchschnittlichen IFR als Kennzahl je Logistikdienstleister aggregiert dar. Geben Sie diese Werte in zwei Tabellen aus. Die Tabellen sollen einen einfachen Vergleich der LDL ermöglichen. Bewertungsrelevant: Output, Code.

   Hinweis: Erneut bietet es sich an, eine Variable `Periode` dem Datensatz hinzu zu fügen, welche aus Jahr und Monat besteht (im Format YYYYMM, z.B. Februar 2019 –> 201902)

2) Erzeugen Sie ein neues Dataframe, welches die aggregierte IFR je **Warehousing-Logistikdienstleister** enthält. Die IFR soll je Warehousing-LDL, Region und Periode (eine Periode = ein Monat eines einzelnen Jahres) aggregiert werden. Nehmen Sie *kurz* Stellung, wie Sie die Qualität dieser Dienstleistungen allgemein einschätzen. Identifizieren Sie danach den insgesamt **schlechtesten** Warehousing-DL. Geben Sie anschliessend den **besten IFR-Wert** und die entsprechende Periode aus, den **dieser** in der Region Japan jemals erreicht hat. Bewertungsrelevant: Kommentar, Output, Code.

3) Erzeugen Sie ein neues Dataframe, welches die aggregierte OTD je **Shipping-Logistikdienstleister** enthält. Die OTD soll je Shipping-LDL, Region und Periode (eine Periode = ein Monat eines einzelnen Jahres) aggregiert werden. Nehmen Sie *kurz* Stellung, wie Sie die Qualität dieser Dienstleistungen allgemein einschätzen. Geben Sie anschliessend den OTD-Wert (und die entsprechende Periode) aus, den der **beste** Shipping-DL im April 2022 in der Region Shanghai erreicht hat. Bewertungsrelevant: Output, Code.

Modellierung: Warehousing

4) Wählen Sie den Warehousing-DL “Gifter Warehousing” aus. Vereinigen Sie das eben erzeugte Dataframe (genauer: ein Subset dieses Dataframes bezüglich des gewählten Warehousing-DL) mit den externen Faktoren der jeweiligen Periode und Region in einem neuen Dataframe. Zeigen Sie davon den Tabellenkopf. Bewertungsrelevant: Output.

   Hinweis: In der Funktion merge() können mehrere überschneidende Spalten genutzt werden, indem dem “by =”-Parameter ein Vektor der Spalten übergeben wird. Ihnen steht frei, andere Funktionen zu verwenden.

5) Sie möchten sich eine Übersicht zu der Korrelation zwischen den externen Faktoren und der IFR des Warehousing-Dienstleister schaffen. Führen Sie dazu die folgenden Schritte aus:

   (a) Geben Sie eine unsortierte Tabelle aus, in der die externen Effekte und deren Korrelation zur IFR abgebildet sind.
   (b) Geben Sie eine Tabelle aus, in der die 5 am stärksten zur IFR korrelierenden externen Effekten und deren Korrelation zur IFR abgebildet sind. Wie bewerten Sie die Korrelation zwischen diesen 5 Faktoren und der IFR?
   (c) Erstellen Sie ein Korrelations-Plot für diese 5 externen Faktoren. Bewertungsrelevant: Kommentar, Output.

6) Sie möchten nun eine Lineare Regression durchführen, um die IFR mit Hilfe der externen Effekte vorherzusagen. Um die Güte Ihrer Modelle vergleichen zu können, benötigen Sie eine geeignete Baseline. Erzeugen Sie eine sinnvolle Baseline in dem Dataframe zu Ihrem gewählten Warehousing-DL in einer Variable `Baseline`. Begründen Sie Ihre Wahl. Geben Sie von dem DataFrame den Tabellenkopf aus. Geben Sie Sie nur die Spalten ‘Periode’, ‘Region’, ‘IFR’ und ‘Baseline’ aus. Bewertungsrelevant: Output, Begründung.

7) Visualisieren Sie die Baseline Ihres gewählten LDL für den Zeitraum von 2019 bis 2023 sowie die IFR in der Region Shanghai und die IFR in der Region Peking Bewertungsrelevant: Output.

8) Bewerten Sie die Baseline für Ihren gewählten Warehousing-Logistikdienstleister nach MAE und MAPE und speichern Sie diese in einem Dataframe (z.B. “evaluation”) ab. Dieses Dataframe soll später auch für die Bewertung der Regressionsmodelle genutzt werden. Fügen Sie zudem auch eine Spalte für das Bestimmtheitsmass (R^2) sowie das adjustierte Bestimmtheitsmass (adj. R^2) hinzu, welche im Falle der Baseline 0 sein können. Bewertungsrelevant: Output.

9) Teilen Sie das Dataframe Ihres gewählten Warehousing-Logistikdienstleisters in ein Trainings- (80%) und ein Test-Set (20%) auf. Geben Sie von beiden den Tabellenkopf aus. Setzen Sie vorher den Seed 4141. Bewertungsrelevant: Code, Output.

10) Wenden Sie die Forward Selection Variante der Wrapper Methode an (siehe Vorlesung). D.h. erstellen Sie zunächst alle uni-variaten Modelle, bewerten Sie diese Modelle und wählen Sie das Modell mit der besten Bewertung aus. Erstellen Sie - basierend auf dem besten Modell der ersten Iteration - alle bi-variaten Modelle (das Modell der vorherigen Wrapper-Iteration wird jeweils um eine Variable erweitert), bewerten Sie diese Modelle und wählen Sie das Modell mit der besten Bewertung aus. Führen Sie dies so lange fort, bis keine Verbesserung mehr erreicht wird. Nutzen Sie zur Modellierung die lineare Regression. Bewerten Sie die Modelle entsprechend nach MAE und MAPE sowie nach regressionsspezifischen Kennzahlen. Nutzen Sie nur die 5 externen Faktoren als Features, die Sie oben als am stärksten korrelierende externe Faktoren identifiziert haben. Kommentieren Sie Ihr Vorgehen zwischen den Iterationen. Bewertungsrelevant: Output, Vorgehen (einschliesslich Kommentare).

    Hinweis: Tritt eine starke Multikollinearität (“strong multicollinearity”) auf, so können Sie alle Modellierungen mit der entsprechenden Variablen-Kombination unter Bezug auf diesen Hinweis auslassen (siehe Vorlesungsinhalte zu Korrelation).

    Hinweis 2: Für das Erstellen der Modelle reicht es aus, zunaechst die Trainings-Daten zu nutzen. Überprüfen Sie ihr endgültiges Modell jedoch am Ende auf Overfitting, indem Sie die Test-Daten nutzen!

    Hinweis 3: Sie müssen kein Feature Engineering betreiben. Sie müssen auch nicht die Residuenplots überprüfen.

11) Bewerten Sie ihr Modell quantitativ im Vergleich mit der Baseline. Bewertungsrelevant: Output, Kommentar.

12) Ihre Chefin kommt auf der Firmenfeier zu Ihnen und schlägt Ihnen eine Wette vor. Sie sagt: “Ich wette mit Ihnen, dass die durchschnittliche IFR des oben betrachteten WH-DL im April 2024 in Japan höher sein wird, als in Shanghai. Sollte dies nicht der Fall sein, gebe ich Ihnen 400 Euro. Habe ich jedoch Recht, müssen Sie mir die 400 Euro geben.” Sollten Sie die Wette eingehen? Bewertungsrelevant: Output, Kommentar.

Entscheidung

13) Ihr Regressionsmodell soll im kommenden Jahr implementiert und langfristig in die Unternehmensprozesse integriert werden. Beschreiben Sie, welche Nutzer und Prozesse davon profitieren könnten und in welcher Form die Lösung bereitgestellt werden könnte. Nehmen Sie ausserdem ausführlich zur Phase der Datenbeschaffung Stellung. Bewertungsrelevant: Kommentar.

3 Formale Anforderungen

Abzugeben ist ein R-Markdown Notebook als Rmd- sowie als PDF-Datei, welches Ihre Antworten auf sämtliche Aufgabenstellungen enthält und vollständig ausführbar ist.

Für den Umgang mit den R-Markdown Notebooks beachten Sie bitte das Reference Handbook unter diesem Link.

Benennen Sie die das Markdown-Notebook und die PDF-Datei wie folgt:

`SCA_SS24_Gruppe###_HA3`

Spätester Abgabezeitpunkt: 19.06.2024 / 10:00 Uhr

Viel Spaß und Erfolg!