

Hausaufgabe 3

Contents

1	Simulationsspiel	2
1.1	Einfuehrung in tiefergehende Simulationsmechanismen	2
1.1.1	Externe Effekte	2
1.1.2	Externe Effekte in der Simulation	3
2	Aufgaben	3
2.1	Aufgabe 1 - Datenanalyse (10 Punkte)	3
2.2	Aufgabe 2 (5 Punkte, gehoerig zur Case Study)	5
3	Formale Anforderungen	6

1 Simulationsspiel

Die nachfolgenden Aufgaben beziehen sich auf das Simulationsspiel. Es gelten weiterhin die Informationen, die bereits in Hausaufgabe 1 vorgestellt wurden. Es fand jedoch keine Simulation von neuen Monaten statt. Stellen Sie sich stattdessen vor, dass Sie nach den letzten Analyseergebnissen (HA 2) nun weitere Analysen durchführen möchten, die Ihre Entscheidungsfindungen weiter verbessern sollen.

1.1 Einfuehrung in tiefergehende Simulationsmechanismen

Die Logistikdienstleister sind in ihrer Leistung von externen Faktoren abhaengig. Dies hat verschiedene Ursachen und Gruende. Zum Beispiel haben Logistikdienstleister zu alte Flotten, die bei extremen Wetterbedingungen nicht funktionieren oder verhaeuft ausfallen, sowie Mitarbeiter, die bei zu extremen Wetterbedingungen nicht zur Arbeit erscheinen. Ebenso haben Unterschiede in operativen Ablaeufen oder Technologieunterstuetzung der Prozesse zur Folge, dass Stau, Internetverfuegbarkeit, Stabilitaet des Stromnetzes, Parkplatzverfuegbarkeit oder Infrastrukturzustand eine unterschiedliche Wirkung auf die Ablaeufe der Logistikdienstleister und somit deren Verfuegbarkeit haben.

1.1.1 Externe Effekte

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 (abhaengig vom Pendlerverkehr, Tourismus u.a.)
- `InternetStability`: Internet-Server pro 1 Mio. Einwohner (abhaengig von Wartung, Reparatur und anderen)
- `PowerGridStability`: Stromausfalltage pro Monat (abhaengig vom Ausfall der Stromversorgung)
- `ParkingSpaceAvailability`: Verfuegbare Parkflaeche in Prozent der Stadtflaeche (abhaengig 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, Gesetzesanderungen, politische Ereignisse u.a.)
- `AvgHealth`: Index in Prozent auf Grundlage einer Befragung der Bevoelkerung zu Gesundheitsthemen (beeinflusst durch Krankheitsbelastung, Verletzungen, Risikofaktoren und andere)
- `Criminality`: Index in Prozent auf Grundlage von Bevoelkerungsbefragungen und Statistiken zu kriminalitaetsbezogenen Themen (beeinflusst durch wahrgenommene Kriminalitaet, Besorgnis ueber Raubueberfaelle, Anzahl der Raubueberfaelle 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 Bevoelkerung ueber die Wasserqualitaet (beeinflusst durch Farbe, Geschmack, Behandlung vor dem Verzehr und andere)
- `leisureAndSocialInteractions`: Index von 1 bis 10 auf der Grundlage einer Bevoelkerungsbefragung zum Thema Freizeit (beeinflusst durch soziale Aktivitaeten, Armutsrisiko, Teilnahme an Freiwilligentaetigkeiten und andere)

Arbeit

- **SkilledLaborAvailability:** Index in Prozent auf Basis einer Unternehmensbefragung zur Wahrscheinlichkeit, Fachkraefte fuer offene Stellen mit dem ersten Stellenangebot zu finden
- **UnskilledLaborAvailability:** Index in Prozent auf Basis einer Unternehmensbefragung zur Wahrscheinlichkeit, ungelernte Arbeitskraefte fuer 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 Arbeitskraefte)
- **Overtime:** Durchschnittliche ueberstunden pro Tag auf Basis der Befragung von Mitarbeitern in der Logistik (Fach- und ungelernte Arbeitskraefte)

Wirtschaft

- **Inflation:** Monatliche Inflationsrate in Prozent
- **BusinessConfidence:** Geschaeftsklimaindex auf der Grundlage der Erhebung der Unternehmen ueber Produktion, Auftraege und Bestaende sowie der aktuellen Lage und der Erwartungen fuer die unmittelbare Zukunft. (normiert auf den langfristigen Durchschnitt = 100)
- **FuelPrice:** US\$ Preis von Benzin pro Gallone

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

1.1.2 Externe Effekte in der Simulation

Fuer die Simulation wurde der Einfluss der Externen Effekte auf die Logistikdienstleister zufaellig festgelegt. Dabei haben (zufaellig) 1 bis 5 externe Effekte einen deutlichen Einfluss auf die Logistikleistung, waehrend der Rest einen verschwindend geringen Einfluss hat.

2 Aufgaben

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

2.1 Aufgabe 1 - Datenanalyse (10 Punkte)

Sie haben nun etwas Erfahrung auf dem Gebiet der Distribution von "Limonalytics". Obwohl Sie stets gewissenhaft fuer die naechsten Monate im Voraus planen, schwankt die Leistung der Logistikdienstleister stark. Daher sollen Sie ein Modell aufstellen, um deren Leistung auf Basis von externen Effekten abzuschuetzen. Sie erhalten daher die externen Effekte der letzten 60 Monate (2016-2020) sowie die Vorhersage fuer die kommenden 12 Monate (2021). Die Vorhersage der externen Effekte fuer die naechsten Monate ist - selbstverstaendlich - ungenau und ebenso mit statistischen Verfahren erstellt. Dies ist fuer Sie jedoch innerhalb der Analyse nicht relevant. Es sollte jedoch in Ihrer Entscheidung kurz *reflektiert* werden.

Daten fuer die Modellierung vorbereiten

- 1) Laden Sie die Datensaeetze `externals` und `services`. Berechnen Sie ueber den gesamten Datensatz `services`, d.h. fuer jede durchgefuehrte Dienstleistung, den On-Time-Delivery (OTD) Status (d.h. 0 oder FALSE, wenn unpuenktlich; 1 oder TRUE wenn puenktlich) sowie die Item Fill Rate (IFR). Stellen Sie anschliessend jeweils die Kennzahlen der durchschnittlichen OTD-Rate und der durchschnittlichen Item Fill Rate als Kennzahl je Logistikdienstleister aggregiert dar. Geben Sie diese Werte in zwei Tabellen aus. Die Tabellen sollen einen einfachen Vergleich der LDL ermoeglichen. Bewertungsrelevant: Output, Code.

Hinweis: Erneut bietet es sich an, eine Variable Periode dem Datensatz hinzu zu fuegen, welche aus Jahr und Monat besteht (im Format YYYYMM, z.B. Februar 2014 → 201402)

- 2) Erzeugen Sie ein neues Dataframe, welches die aggregierte IFR je **Warehousing-Logistikdienstleister** enthaelt. Die IFR soll je LDL (nur Warehousing), Region und Periode (eine Periode = ein Monat eines einzelnen Jahres) aggregiert werden. Geben Sie anschliessend eine Tabelle aus, die den einfachen Vergleich der WH-DL fuer die Periode 2020/12 in Japan ermoeglicht. Bewertungsrelevant: Output, Code.

- 3) Erzeugen Sie ein neues Dataframe, welches die aggregierte OTD je **Shipping-Logistikdienstleister** enthaelt. Die OTD soll je LDL (nur Shipping), Region und Periode (eine Periode = ein Monat eines einzelnen Jahres) aggregiert werden. Geben Sie anschliessend eine Tabelle aus, die den einfachen Vergleich der Shipping-DL fuer die Periode 2020/12 in Japan ermoeoglicht. Bewertungsrelevant: Output, Code.
-

Modellierung: Warehousing

- 4) Waehlen Sie den Warehousing-DL "CPS Warehousing" aus. Vereinigen Sie das eben erzeugte DataFrame (genauer: Ein Subset dieses Dataframes bezueglich des gewaehlten 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() koennen mehrere ueberschneidende Spalten genutzt werden, indem dem "by ="-Parameter ein Vektor der Spalten uebergeben wird. Ihnen steht frei, andere Funktionen zu verwenden.

- 5) Sie moechten sich eine Uebersicht zu der Korrelation zwischen den externen Faktoren und der IFR des Warehousing-Dienstleister schaffen. Fuehren 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 starksten zur IFR korrelierenden externen Effekten und deren Korrelation zur IFR abgebildet sind.
- (c) Erstellen Sie ein Korrelations-Plot fuer diese 5 externen Faktoren. Bewertungsrelevant: Output.

- 6) Sie moechten nun eine Lineare Regression durchfuehren, um die IFR mit Hilfe der externen Effekte vorherzusagen. Um die Guete Ihrer Modelle vergleichen zu koennen, benoetigen Sie eine geeignete Baseline. Erzeugen Sie eine sinnvolle Baseline in dem DataFrame zu Ihrem gewaehlten Warehousing-DL in einer Variable `Baseline`. Begrunden 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, Begrueundung.

- 7) Visualisieren Sie die Baseline Ihres gewaehlten LDL fuer den Zeitraum von 2016 bis 2020 sowie die IFR in der Region Shanghai. Bewertungsrelevant: Output.

- 8) Bewerten Sie die Baseline fuer Ihren gewaehlten Warehousing-Logistikdienstleister nach MAE und MAPE. Bewertungsrelevant: Output.

Hinweis: Es bietet sich an, die Werte der Bewertungen in einem Dataframe 'evaluation' zu speichern.

- 9) Teilen Sie das Dataframe Ihres gewaehlten 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 zunaechst alle uni-variaten Modelle, bewerten Sie diese Modelle und waehlen 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 waehlen Sie das Modell mit der besten Bewertung aus. Fuehren 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 regressionspezifischen Kennzahlen. Nutzen Sie nur die 5 externen Faktoren als Features, die Sie oben als am staerksten korrelierende externe Faktoren identifiziert haben. Kommentieren Sie Ihr Vorgehen zwischen den Iterationen. Bewertungsrelevant: Output, Vorgehen (einschliesslich Kommentare).

Hinweis: Tritt eine starke Multikollinearitaet ("strong multicollinearity") auf, so koennen Sie alle Modellierungen mit der entsprechenden Variablen-Kombination unter Bezug auf diesen Hinweis auslassen

(siehe Vorlesungsinhalte zu Korrelation).

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

- 11) Bewerten Sie ihr Modell quantitativ im Vergleich mit der Baseline. Bewertungsrelevant: Output, Kommentar.
- 12) Ihre Chefin kommt auf der Firmenfeier zu Ihnen und schlaegt Ihnen eine Wette vor. Sie sagt: "Ich wette mit Ihnen um 100 Euro, dass die IFR des oben betrachteten WH-DL im naechsten Monat (Januar 2021) in allen Regionen ueber 0.85 sein wird." Sollten Sie die Wette eingehen? Bewertungsrelevant: Output, Kommentar.

Shipping

Aufgaben entfallen

Entscheidung

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

2.2 Aufgabe 2 (5 Punkte, gehoerig zur Case Study)

Erheben Sie ueber einen Zeitraum von 14 Tagen Daten aus Ihrem Alltag und erstellen Sie anschliessend eine Visualisierung, um die gewonnen Erkenntnisse visuell darzustellen. Beachten Sie dabei die folgenden Anweisungen und Hinweise:

Anweisungen:

- Beachten Sie insbesondere den Inhalt der Vorlesung "Ergebniskommunikation von Analytics".
- Die Beobachtungen sollten im weitesten Sinne einen Bezug zu Logistik oder logistischen Aktivitaeten besitzen (z.B. Lagerung, Umschlag, Transport/Bewegungen, Beschaffung, Produktion, ...). Dazu koennen beispielsweise Transporte zaehlen, die Sie vornehmen sowie deren Entfernung, Dauer oder Tageszeit; Dinge, die Sie besitzen sowie deren Kategorie, Lagerort oder Mengen ueber die Zeit; Dinge, die Sie beschaffen sowie deren Mengen ueber die Zeit, deren Kategorien ueber die Zeit, deren Volumina oder deren Wert; Dinge, die Sie produzieren (z.B. Textdokumente), deren Produktionsfortschritt ueber die Zeit, Bearbeitungszeitpunkte, Dauer von Teilschritten; ...
- Die aufgenommenen Datenpunkte/Beobachtungen sollen mindestens 2 Variablen haben.
- Die Visualisierung soll die Aufmerksamkeit des Betrachters auf einen von Ihnen gewuenschten Sachverhalt oder eine Besonderheit in den Daten lenken.
- Neben der digitalen Abgabe soll eine Abgabe als Postkarte bzw. in Postkartengroesse (**DIN A6!**) als Hardcopy erfolgen, mit der Visualisierung auf der Vorderseite und einer Erlaeuterung auf der Rueckseite. Die Intention dahinter ist wie folgt: Sie haben auf der Postkarte nur einen sehr begrenzten Platz mit wenig Raum fuer Text. Es besteht fuer den Betrachter ausserdem nicht die Moeglichkeit, eventuellen Text oder die Visualisierung digital zu vergroessern. Um eine sinnvolle Gestaltung vorzunehmen, muessen Sie sich von daher genau ueberlegen, welche Aussage Sie mit der Visualisierung machen moechten, und wie Sie diese erreichen.

Hinweise:

- Die Visualisierung muss kein Graph sein, sie kann auch eine Infografik sein. Sie muss ebenso nicht mit R erstellt werden - PowerPoint, Excel oder ein Programm Ihrer Wahl sind ebenso zugelassen.

- Bitte anonymisieren Sie die Visualisierung (keine Namen, keine Matrikelnummern, keine Adressen von Wohnorten)
- Als Inspiration sind empfohlen: Pinterest.com (u.a. "visualization", "infographic"), David McCandless, das Dear Data Project, Subreddit "Data is Beautiful".

Bewertungsrelevant: Vollstaendigkeit der Hardcopy: Visualisierung (mit erkennbarer Datenerhebung) und Er-laeuterung, Vollstaengigkeit bei Beschreibung und Vorgehen, Begrueendung der Visualisierung, Begrueendung des Bezugs zur Logistik, Einhaltung der Formalen Vorgaben (DIN A6).

3 Formale Anforderungen

Abzugeben ist ein R-Markdown Notebook als Rmd- sowie als PDF-Datei, welches Ihre Antworten auf saemtliche Auf-gabenstellungen enthaelt und vollstaendig ausfuehrbar ist.

Fuer den Umgang mit den R-Markdown Notebooks beachten Sie bitte das Cheat Sheet unter diesem Link und das Ref-erence Handbook unter diesem Link. Wie in der Vorlesung angesprochen, koennen Sie erfolglose Code-Fragmente im R-Notebook lassen, beachten Sie jedoch, dass diese auf `eval = FALSE` gesetzt sind.

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

SCA_WS2021_Gruppe###_HA3

Bitte versuchen Sie Ihre Texte sinnvoll zu strukturieren!

Spaetester Abgabezeitpunkt (Aufgabe 1): 13.01.2020 / 14:00 Uhr

Spaetester Abgabezeitpunkt (Aufgabe 2): 27.01.2020 / 23:55 Uhr (aufgrund der besonderen Bedingungen in diesem Semester kann die Abgabe auch digital erfolgen)

Viel Erfolg!