INTERNATIONALE HOCHSCHULE

FALLSTUDIE

Aufgabenstellung zum Kurs:

DLMDWME01 – Model Engineering

INHALTSVERZEICHNIS

1. Aufgabenstellung 2

1.1. Aufgabenstellung 1: Erstellen eines Prognosemodells des Kreditkartenzahlungsverkehr für Online-Einkäufe 2

1.2. Aufgabenstellung 2: Automatisierung des Bereitschaftsdienstplans für Einsatzfahrer mithilfe eines Vorhersagemodells 3

1.3. Aufgabenstellung 3: Effizienter Flugbetrieb bei East Carmen Airlines mittels eines Machine-Learning-Modells zur Ankunftszeit-Vorhersage 4

2. Zusatzinformationen zur Bewertung der Fallstudie 6

3. Betreuungsprozess 7

1. AUFGABENSTELLUNG

Wähle für Deine Bearbeitung eine der folgenden Fallstudien aus.

Bitte berücksichtige bei Deiner Bearbeitung die in der jeweiligen Fallstudie selbst beschriebene Aufgabenstellung.

Zusätzlich besteht jede Fallstudie aus einem zip-Ordner, der einen csv-Datensatz und ein txt-Dokument mit weiterführenden wichtigen Informationen, wie zum Beispiel der Beschreibung des Datensatzes, beinhaltet. Der zip-Ordner wird zum Kursmaterial angehängt.

Als Grundlage dienen hierbei das Studienskript sowie vertiefende Information aus dem Internet oder geeigneten Fachbüchern. Beispielsweise bieten Data Science Community-Webseiten, wie **kaggle** (https://www.kaggle.com/), eine reichhaltige Sammlung an frei verfügbaren Datensätzen und öffentlichen Notebooks. Es wird erwartet, dass für das gewählte Thema selbst weiterführende, relevante Literatur und aktuelle wissenschaftliche Quellen recherchiert und für die Erarbeitung zu Grunde gelegt werden. Diese sind im Literaturverzeichnis korrekt zu referenzieren.

Insbesondere die Fallstudien im Modul „Model Engineering“ bestehen aus konzeptionellen Aufgaben sowie Programmieraufgaben. Deine Erkenntnisse und Visualisierungen sollen in ein finales Dokument eingearbeitet werden. Das Format dieses finalen Dokuments wird im Prüfungsleitfaden (s. 2.3, „Formalia“) ausführlich beschrieben. Für die Programmierung werden die Sprachen **python** und **R** empfohlen, weil sie mittlerweile die gängigsten Programmiersprachen der Data Science Community sind. Außerdem ist die frei zugängliche Web-Oberfläche **Jupyter Notebook** (https://jupyter.org/) vor allem für eine erste, explorative Untersuchung des Datensatzes sehr hilfreich. In dieser Web-Oberfläche lassen sich Programmteile, Visualisierungen und Text in ein einzelnes Notebook integrieren.

Vergiss nicht, Deinen gut dokumentierten Code als Anhang in die finale Einreichung einzufügen, da es über Turnitin nur möglich ist, ein Dokument einzureichen.

1.1. Aufgabenstellung 1: Erstellen eines Prognosemodells des Kreditkartenzahlungsverkehr für Online-Einkäufe

Beschreibung der Fallstudie:

Bereits an Deinem allerersten Tag als Data Scientist bei einem der weltweit größten Einzelhandelsunternehmen wirst Du zu einem Treffen mit Experten für Onlinezahlungen eingeladen, die Dich um Unterstützung bitten. Im letzten Jahr war die Ausfallsrate an Online-Kreditkartenzahlungen besonders hoch. Wegen dieser vielen fehlgeschlagenen Online-Überweisungen verliert das Unternehmen einerseits sehr viel Geld und andererseits werden die Kunden zunehmend unzufrieden mit dem Online-Shop des Unternehmens.

Online-Kreditkartenzahlungen werden mithilfe sogenannter Zahlungsdienstleister, abgekürzt als „PSPs“ (=payments service providers), durchgeführt. Dein neuer Arbeitgeber hat mit vier verschiedenen Zahlungsdienstleistern Verträge abgeschlossen und muss für jede einzelne Überweisung Servicegebühren an diese Unternehmen zahlen.

Die Logik, welcher PSP für eine bestimmte Überweisung am geeignetsten ist, basiert aktuell auf einem fixen Regelwerk und wird manuell durchgeführt. Die Entscheidungsträger innerhalb des Fachbereichs für Online-Zahlungen sind aber der Überzeugung, dass ein Prognosemodell zu besseren Entscheidungen, als ein fixes, manuelles Regelwerk, führen kann.

Ziel des Projekts:

Unterstütze den Fachbereich für Online-Zahlungen durch ein Prognosemodell, um die Zuweisung einer Kreditkartenzahlung zu einem PSP zu automatisieren. Das Modell soll einerseits **die Erfolgsrate der Transaktionen erhöhen** und andererseits die **Transaktionskosten geringhalten**.

Datensatz:

Der Datensatz und weiterführende Informationen aus dem Fachbereich (Name der Zahlungsdienstleister („PSPs"), Transaktionsgebühren) sind in einem separaten zip-Dokument hinterlegt, das zum Kursmaterial angehängt wird.

Detaillierte Beschreibung der Aufgabe:

Die Aufgabe besteht sowohl aus einem Programmierteil als auch aus konzeptionellen Teilaufgaben. Hier folgt eine detaillierte Beschreibung an offenen Fragen, die im finalen Dokument beantwortet werden sollen:

- **Organisiere das Projekt** mithilfe der CRISP-DM oder der MS Team Data Science Methode. Mache einen Vorschlag, wie die Ordnerstruktur eines Git-Repositories für das Projekt aufgebaut werden soll. Beachte, dass Du den finalen Code des Projekts nicht nach dieser Ordnerstruktur aufbauen musst.
- **Beurteile die Qualität** des zur Verfügung gestellten Datensatzes. **Bereite Deine Erkenntnisse so auf und visualisiere sie so,** dass Businesspartner in einer klaren und einfachen Weise die wichtigen Zusammenhänge verstehen können.
- Stelle ein erstes Basismodell (ein sogenanntes **Baseline-Modell**) auf, sowie ein **präzises Vorhersagemodell, das den Businessanforderungen genügt,** nämlich die Erfolgsrate der Kreditkartenzahlungen zu erhöhen und gleichzeitig die Transaktionskosten gering zu halten.
- Damit die Businesspartner Vertrauen in Dein neues Modell entwickeln, solltest du die Wichtigkeit der einzelnen erklärenden Variablen diskutieren und die Modellresultate so interpretierbar wie möglich gestalten. Außerdem ist eine detaillierte Fehleranalyse sehr wichtig, damit die Businesspartner auch die Schwachstellen Deiner Herangehensweise verstehen.
- Im letzten Schritt des Projekts soll ein Vorschlag unterbreitet werden, wie Dein Modell in die tägliche Arbeit des Fachbereichs eingebunden werden kann, beispielsweise wie eine graphische Benutzeroberfläche (GUI) aussehen könnte.

1.2. Aufgabenstellung 2: Automatisierung des Bereitschaftsdienstplans für Einsatzfahrende mithilfe eines Vorhersagemodells

Beschreibung der Fallstudie:

Als Unternehmensberater im Bereich Data Science wirst Du eingeladen, bei einem Projekt des Berliner Rotkreuz-Rettungsdienstes mitzuwirken. Die Personalplanung hat Probleme, den Bereitschaftsdienst der Einsatzfahrenden effizient zu planen. Mithilfe Deines Fachwissens soll die aktuelle Planungslogik verbessert werden.

Tagtäglich sind eine bestimmte Anzahl an Einsatzfahrerenden im Dienst. Jedoch kann es wegen kurzfristiger Krankenstände der Einsatzfahrenden oder einer unerwartet hohen Anzahl an Notrufen zu Kapazitätsengpässen kommen und es werden mehr Einsatzfahrende benötigt als ursprünglich angenommen. Folglich muss eine bestimmte Anzahl an Einsatzfahrenden in Bereitschaft gehalten und bei Bedarf aktiviert werden. In der aktuellen Planungslogik werden pro Tag 90 Einsatzfahrende in Bereitschaft gehalten.

Kollegen aus dem Planungsbereich sind davon überzeugt, dass es bestimmte saisonale Muster gibt, wie zum Beispiel, dass in den Wintermonaten mehr Einsatzfahrende aufgrund von Kurzfristkrankenständen ausfallen.

Diese Muster sind im bisherigen Planungsprozess nicht abgebildet. Außerdem fehlen an manchen Tagen sogar Einsatzfahrende, wenn die Höchstzahl der 90 Ersatzfahrenden erreicht ist, aber mehr Fahrende benötigt werden. Dann müssen Einsatzfahrende, die keinen Dienst haben, in den Dienst gerufen werden.

Organisatorisch ist es ist sehr wichtig zu wissen, dass der Dienstplan am 15. des laufenden Monats für den Folgemonat fertiggestellt sein muss. Der Dienstplan für November muss also am 15. Oktober fertiggestellt sein.

Ziel des Projekts:

Unterstütze die Kolleg:innen der Personalplanung mithilfe eines **Vorhersagemodells, das auf Tagesbasis die Anzahl an Einsatzfahrenden im Bereitschaftsdienst effizient abschätzt**. In diesem Zusammenhang bedeutet *effizient*, dass der Prozentsatz an aktivierten Einsatzfahrenden höher als im bisherigen Planungsprozess mit konstant 90 Bereitschaftsfahrenden pro Tag ist. Außerdem sollen Tage mit zu wenig Einsatzfahrenden seltener als im aktuellen Planungsprozess vorkommen. Beachte, dass der Bereitschaftsplan am 15. des laufenden Monats für den Folgemonat fertiggestellt sein muss.

Datensatz:

Der Datensatz und weiterführende Informationen aus dem Fachbereich (Abwesenheitsdaten, Anzahl an Notrufen auf Tagesbasis) sind in einem separaten zip-Dokument hinterlegt, das zum Kursmaterial angehängt wird.

Detaillierte Beschreibung der Aufgabe:

Die Aufgabe besteht sowohl aus einem Programmierteil als auch aus konzeptionellen Teilaufgaben. Hier folgt eine detaillierte Beschreibung an offenen Fragen, die im finalen Dokument beantwortet werden sollen:

- **Organisiere das Projekt** mithilfe der CRISP-DM oder der MS Team Data Science Methode. Mache einen Vorschlag, wie die Ordnerstruktur eines Git-Repositories für das Projekt aufgebaut werden soll. Beachte, dass Du den finalen Code des Projekts nicht nach dieser Ordnerstruktur aufbauen musst.
- **Beurteile die Qualität** des zur Verfügung gestellten Datensatzes. **Bereite Deine Erkenntnisse so auf und visualisiere sie so,** dass Businesspartner in einer klaren und einfachen Weise die wichtigen Zusammenhänge verstehen können.
- Stelle ein erstes Basismodell (ein sogenanntes **Baseline-Modell**) auf, sowie ein **präzises Vorhersagemodell, das den Businessanforderungen genügt**, nämlich die Anzahl an (aus dem Bereitschaftsdienst) aktivierten Einsatzfahrenden zu maximieren und gleichzeitig die Tage mit zu wenigen Einsatzfahrenden zu minimieren.
- Damit die Businesspartner Vertrauen in Dein neues Modell entwickeln, solltest du die Wichtigkeit der einzelnen erklärenden Variablen diskutieren und die Modellresultate so interpretierbar wie möglich gestalten. Außerdem ist eine detaillierte Fehleranalyse sehr wichtig, damit die Businesspartner auch die Situationen, in denen Dein Ansatz möglicherweise ungeeignet ist, nachvollziehen können.

Im letzten Schritt des Projekts soll ein Vorschlag unterbreitet werden, wie Dein Modell in die tägliche Arbeit des Fachbereichs eingebunden werden kann, beispielsweise wie eine graphische Benutzeroberfläche (GUI) aussehen könnte.

1.3. Aufgabenstellung 3: Effizienter Flugbetrieb bei East Carmen Airlines mittels eines Machine-Learning-Modells zur Ankunftszeit-Vorhersage

Beschreibung der Fallstudie:

Willkommen bei East Carmen Airlines! Als Data Scientist innerhalb der Strategieabteilung des Unternehmens kommst Du mit vielen Projekten in Berührung, die das Airline-Geschäft nachhaltig transformieren und automatisieren. Von erheblicher Bedeutung für Fluglinien ist ein stabiler Flugbetrieb und demzufolge präzise

Vorhersagen der Ankunftszeiten während eines Betriebstages. Passagiere würden durch präzise Vorhersagen der Ankunftszeiten im Verspätungsfall eher Anschlussflüge erreichen und damit Kosten für die Fluglinie reduzieren.

Dein Vorhersagemodell soll sich auf jedes Flugzeug innerhalb der Flotte anwenden lassen. Nehmen wir als Beispiel das Flugzeug mit dem Kennzeichen EC-LPD. Beim geplanten Abflug des allerersten Fluges eines Betriebstages sollte das Modell alle Ankünfte der EC-LPD an diesem Tag vorhersagen.

Viele Kolleg:innen innerhalb der Fachbereiche stehen einem Ankunftszeit-Vorhersagemodells skeptisch gegenüber. Sie behaupten, dass es zwar möglich sei, die Ankunft des allerersten Fluges innerhalb eines Betriebstages mit annehmbarem Fehler vorherzusagen, aber die Vorhersagen des zweiten und dritten Fluges werden schon zu ungenau sein. Es ist nun Deine Aufgabe, die Kollegen der einzelnen Fachbereiche davon zu überzeugen, dass mithilfe eines geeigneten Datensatzes und sinnvoller erklärender Variablen ein präzises Vorhersagemodell entwickelt werden kann.

Ziel des Projekts:

BI-Analysten haben aus unterschiedlichen Fachbereichen Informationen über mögliche erklärende Variablen für Flugvorhersagen gesammelt. Nun ist es Deine Aufgabe, mithilfe dieses Datensatzes **ein Prototyp-Vorhersagemodell zu erstellen, das alle Ankunftszeiten eines Flugzeuges der Flotte innerhalb eines Betriebstages vorhersagt. Die Vorhersagen sollen zum Zeitpunkt des geplanten Abflugs des allerersten Fluges am Betriebstag stattfinden.** Als Beispiel ist hier das Flugzeug mit der Kennzahl EC-LPD angefügt, das den ersten Abflug des Tages um 5:30 Uhr von MAD nach VIE hat und die Kette MAD-VIE-MAD-CDG-MAD-LHR-MAD fliegt. Für diesen Betriebstag und dieses Kennzeichen soll Dein Modell die aufeinanderfolgenden Ankunftszeiten in VIE, MAD, CDG, MAD, LHR, MAD um exakt 5:30 Uhr vorhersagen. Die Kolleg:innen des Flugbetriebs sind vor allem daran interessiert, wie sich das **Konfidenzintervall eines solchen Modells innerhalb eines Betriebstages entwickelt.**

Datensatz:

Der Datensatz und weiterführende Informationen aus dem Fachbereich (Flugbetriebsdaten, Wetterdaten pro Flug, Flottenliste) sind in einem separaten zip-Dokument hinterlegt, das zum Kursmaterial angehängt wird.

Detaillierte Beschreibung der Aufgabe:

Die Aufgabe besteht sowohl aus einem Programmierteil als auch aus konzeptionellen Teilaufgaben. Hier folgt eine detaillierte Beschreibung an offenen Fragen, die im finalen Dokument beantwortet werden sollen:

- **Organisiere das Projekt** mithilfe der CRISP-DM oder der MS Team Data Science Methode. Mache einen Vorschlag, wie die Ordnerstruktur eines Git-Repositories für das Projekt aufgebaut werden soll. Beachte, dass Du den finalen Code des Projekts nicht nach dieser Ordnerstruktur aufbauen musst.
- **Beurteile die Qualität** des zur Verfügung gestellten Datensatzes. **Bereite Deine Erkenntnisse so auf und visualisiere sie so,** dass Businesspartner in einer klaren und einfachen Weise die wichtigen Zusammenhänge verstehen können.
- Stelle ein erstes Basismodell (ein sogenanntes **Baseline-Modell**) auf, sowie ein **präzises Vorhersagemodell, das den Businessanforderungen genügt,** wie sie im Unterabschnitt „Ziel des Projekts“ beschrieben sind.
- Damit die Businesspartner Vertrauen in Dein neues Modell entwickeln, solltest Du die Wichtigkeit der einzelnen erklärenden Variablen diskutieren und die Modellresultate so interpretierbar wie möglich gestalten. Für die Fachbereichskolleg:innen ist es vor allem essenziell zu verstehen, wie sich der Vorhersagefehler für aufeinanderfolgende Flüge fortpflanzt.

Im letzten Schritt des Projekts soll ein Vorschlag unterbreitet werden, wie Dein Modell in die tägliche Arbeit des Fachbereichs eingebunden werden kann, beispielsweise wie eine graphische Benutzeroberfläche (GUI) aussehen

könnte. Diskutiere mögliche wichtige Information, die im aktuellen Datensatz noch nicht enthalten sind, jedoch dabei helfen würden, die Vorhersagequalität weiter zu erhöhen.

2. ZUSATZINFORMATIONEN ZUR BEWERTUNG DER FALLSTUDIE

Bei der Analyse der Fallstudie und ihrer Bearbeitung sollten die aufgeführten Bewertungskriterien und Erläuterungen berücksichtigt werden (vgl. Prüfungsleitfaden, 2.4 „Bewertung der Fallstudienbearbeitung“).

Erfassung: Wie wird die Problemstellung aus einem Geschäftsfeld in ein Data Science Projekt übersetzt, d.h. welcher Data Science Use Case lässt sich aus der Problemstellung ableiten?

- Beispiel: „Für die vorliegende Aufgabenstellung wenden wir zuerst einen Klassifikationsalgorithmus des überwachten Lernens an. Von den prognostizierten Wahrscheinlichkeiten leiten wir dann ein Regelwerk ab.“

Konzepte: Anwenden und Einbeziehen der relevanten Konzepte des Kurses in die Studie. Beispiele hierfür sind:

- Durchführen einer explorativen Datenanalyse, wie zum Beispiel das Berechnen von Korrelationen.
- Welche Modelle werden für die Vorhersage verwendet (z.B. Lineare Regression, Entscheidungsbaumverfahren…)?
- Anhand welcher Techniken werden die erklärenden Variablen („Features“) ausgewählt?
- Wie verhindert man die Überanpassung („Overfitting“) des Systems (z.B. durch das Trainieren mit mehreren Hyperparametern und das Anwenden von Regularisierungsmethoden, …)?
- …

Analyse: Erstelle eine genaue und detaillierte Analyse der Daten und des Modellierungsansatzes.

- Beispiel: „In unserer ersten, explorativen Datenanalyse stellen wir fest, dass für Argentinien die Zahl der Kund:innen im letzten Jahr um 50% zurückgegangen ist, während in allen anderen Ländern der Rückgang im Durchschnitt 10% beträgt. Das bedeutet also, dass wir gerade in Argentinien große Probleme mit unserem Geschäftsmodell haben.“
- Beispiel: „Wir beobachten, dass ein Entscheidungsbaumverfahren die Gesamtgenauigkeit des Modells um mehr als 5% erhöht. Außerdem erhöht sich die Genauigkeit um weitere 5%, wenn wir trigonometrische Funktionen auf die Zeitvariablen anwenden.“

Abschluss: Beschreibe die wichtigsten Ergebnisse Deiner Studie in einer klaren und quantitativen Weise. Gehe dabei genauso strukturiert wie bei der Analyse der Daten und des Modellierungsansatzes vor. Beantworte Fragen, die über das reine Modellieren hinausgehen, wie zum Beispiel: „Wie kann das Modell in der täglichen Arbeit genutzt werden?“, „Sind Weiterentwicklungen notwendig, um das Modell in einen operativen Modus zu überführen?“

- Beispiel: „Nach eingehender Analyse des Modells kommen wir zum Schluss, dass die Genauigkeit des neuen, datengetriebenen Ansatzes den bisherigen, regelbasierten Ansatz auf einem Testdatensatz um 20% übertrifft. Wir konnten keinen Fall entdecken, in dem der regelbasierte Ansatz zu besseren Ergebnissen führt. Deshalb sind wir zuversichtlich, dass das Prognosemodell zu einer erheblichen Kostensenkung führen wird. Unsere vorsichtige Schätzung für die nächsten zwei Monate geht von einer Kostenreduktion von mehr als 6% durch das Einbinden des Vorhersagemodells in den laufenden Prozess aus.“

3. BETREUUNGSPROZESS

Für die Betreuung der Fallstudie stehen grundsätzlich mehrere Kanäle offen; die jeweilige Inanspruchnahme liegt dabei im eigenen Verantwortungsbereich. Der/die Tutor:in steht per E-Mail für fachliche Rücksprachen zur Themenwahl einerseits sowie für formale und allgemeine Fragen zum wissenschaftlichen Arbeiten andererseits zur Verfügung. Eine Abnahme von Gliederungen, Textteilen oder –entwürfen durch den/die Tutor:in ist hierbei jedoch nicht vorgesehen, da die eigenständige Erstellung Teil der zu erbringenden Prüfungsleistung ist und in die Gesamtbewertung einfließt. Es werden jedoch Hinweise zu Gliederungsentwürfen gegeben, um den Einstieg in die Strukturierung einer wissenschaftlichen Arbeit zu erleichtern.