**Modul**: DLMDWME01 – Fallstudie: Modell Engineering

**Tutor**: Markus Pak

**Matrikelnummer**: 91501239

**Student**: Sebastian Kinnast

**Semester**: 2. Semester Teilzeit II

**Studiengang**: Master of Data Science DEU 60 ETCS (FS MADW-60 2022 SS)

**Fallstudie**

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

Eingereicht am

Inhaltsverzeichnis

[1. Einleitung 3](#_Toc143950732)

[1.1. Ausgangssituation 3](#_Toc143950733)

[1.2. Ziel des Projekts 3](#_Toc143950734)

[2. Hauptteil (Fallvorstellung, -bearbeitung und -lösung) 3](#_Toc143950735)

[2.1. Projektorganisation 3](#_Toc143950736)

[2.1.1. Geschäftsverständnis 3](#_Toc143950737)

[2.1.2. Datenverständnis 4](#_Toc143950738)

[2.1.3. Datenaufbereitung 5](#_Toc143950739)

[2.1.4. Modellierung 5](#_Toc143950740)

[2.1.5. Evaluierung 5](#_Toc143950741)

[2.1.6. Bereitstellung 5](#_Toc143950742)

[3. Schluss (Diskussion und weitere Übertragungen) 5](#_Toc143950743)

[3.1. Analyse und Interpretation Merkmale und Leistungsmetriken 5](#_Toc143950744)

[3.2. Anwendung in der Praxis 6](#_Toc143950745)

[4. Literaturverzeichnis 6](#_Toc143950746)

[5. Abkürzungsverzeich 6](#_Toc143950747)

[6. Abbildungs- und Tabellenverzeichnis 6](#_Toc143950748)

[7. Verzeichnis der Anhänge 6](#_Toc143950749)

# Einleitung

# Ausgangssituation

Im letzten Jahr war die Ausfallsrate an Online-Kreditkartenzahlungen bei einem der weltweit größten Einzelhandelsunternehmen 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. Es wurden mit vier verschiedenen Zahlungsdienstleistern Verträge abgeschlossen und es muss für jede einzelne Überweisung Servicegebühren an diese Unternehmen bezahlt werden. 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

Der Fachbereich für Online-Zahlungen soll durch ein Prognosemodell unterstützt werden, um die Zuweisung einer Kreditkartenzahlung zu einem PSP zu automatisieren. Das Modell soll einerseits die Erfolgsrate der Transaktionen erhöhen und andererseits die Transaktionskosten gering halten.

# Hauptteil (Fallvorstellung, -bearbeitung und -lösung)

# Projektorganisation

Zur Organisation des Projekts wird das Rahmenwerk der CRISP-DM-Methode (Cross Industry Standard Process for Data Mining) verwendet. Diese Methode beinhaltet die sechs Prozessphasen Geschäftsverständnis, Datenverständnis, Datenaufbereitung, Modellierung, Evaluierung und Bereitstellung, die unterschiedliche Wechselwirkungen untereinander haben und Bestandteile eines kontinuierlichen Verlaufs sind, der einen iterativen Charakter hat.

Geschäftsverständnis

Zur Erarbeitung eines Geschäftsverständnisses muss das Projekt zunächst in die Gesamtbetrachtung des Unternehmens eingeordnet sowie der durch das Projekt erzielbare Nutzen herausgestellt werden. Wie in der vorigen Schilderung der Ausgangsituation unter 1.1 bereits erwähnt, betrifft die zu lösende Problemstellung, die Sparte der Online-Verkäufe des Einzelhandelsunternehmens. Die aktuell sehr hohe Ausfallsrate von Zahlungsabwicklungen betrifft Transaktionen mit der spezifischen Zahlungsart „Kreditkarte“. Daraus ergeben sich mehrere negative Effekte. Zum einen fallen Kosten für das Unternehmen an, ohne dass Umsatz in das Unternehmen zurückfließt und zum anderen sinkt die Kundenzufriedenheit, was langfristig zur Kundenabwanderung und weiteren Umsatzeinbußen führen kann. Im Zahlungsabwicklungsprozess erfolgt die Auswahl des passenden Zahlungsdienstleisters zurzeit manuell auf Basis eines fixen Regelwerks. Zur Ablösung dieses Modells durch einen automatisierten Machine-Learning-Ansatz, wird die Nutzenhypothese aufgestellt, dass dieses Auswahlverfahren durch eine auf Basis von Dan trainiertes und automatisiertes Prognosemodell ersetzt werden kann und damit die Ausfallrate reduziert wird. Weiterhin soll durch die bessere Auswahl des Zahlungsdienstleisters auch eine Minimierung der Kosten je Transaktion erreicht werden. Eine Übersicht der Gebühren je PSP ist nachfolgend aufgeführt.

Figure 1: CRISP-DM-Prozess-Modell (in Anl. an Fuchs, 2020)

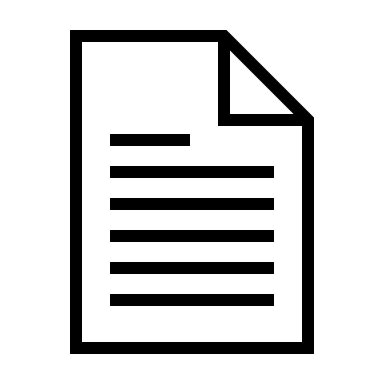
|  |  |  |
| --- | --- | --- |
| PSP | Gebühr für erfolgreiche Transaktionen | Gebühr für fehlgeschlagene Transaktionen |
| Moneycard | 5 Euro | 2 Euro |
| Goldcard | 10 Euro | 5 Euro |
| UK\_Card | 3 Euro | 1 Euro |
| Simplecard | 1 Euro | 0,5 Euro |

Für die weitere Planung des Projekts kann aufgrund des überschaubaren Umfangs der Aufgabenstellung auf die Erstellung eines ausführlichen Projektplans verzichtet werden. Stattdessen werden die wichtigsten Kernpunkte kurz umrissen (Shearer, 2000)

Benötigen Ressourcen:

* Personell: ein Data Scientist
* Technologisch/ Hardware:
* Software: Python: JupyterLab ist die webbasierte interaktive Entwicklungsumgebung für Notebooks, Code und Daten. Für Arbeitsabläufe in den Bereichen Data Science, wissenschaftliches Rechnen, Computerjournalismus und maschinelles Lernen zu konfigurieren und zu gestalten, modulares Design für Erweiterungen ein,*projektspezifische Anforderungen, Annahmen, Einschränkungen und Risiken,*
* *Programmiersprache Python 🡪 Test einzelner Programmteile, Visualisierungen und explorative Datenanalysen sowie erstellung des Benchmarkmodell und des finalen Prognosemodel*

Github wird zur Versionskontrolle des Projekts genutzt. Dazu kann die Ordnerstruktur für das Projekt wird wie folgt aufgebaut werden.



readme.md

Figure 2: Ordenrstruktur des GIT-Repositories

Der Ordner „data“ ist entsprechend so strukturiert, dass die Daten in den verschiedenen Verarbeitungsstadien dort abgelegt werden können. Unter „raw“ werden die ursprünglichen unverarbeiteten Rohdaten abgelegt. Der Ordner „processed“ enthält die für das Modell fertig aufbereiteten Daten. Im Ordner „notebooks“ können Jupyter-Notebooks oder andere interaktive Codedateien aufbewahrt werden. Der Ordner „scripts“ ist für eigenständig ausführbare Skripte gedacht. Im Ordner „models“ werden die entwickelten und evaluierten Machine-Learning-Modelle gespeichert. Dabei wird zwischen der Ablage für die ersten einfachen „Baseline“-Modellen unterschieden, die als Prototyp für eine Exploration der Möglichkeiten und als Referenzbeispiel für das finale präzise Modell dienen. Dieses wird schrittweise entsprechend der Unterordnerstruktur in den Phasen „Dev“ für Entwicklung, „test“ für Evaluierung und Tests und „prod“ für produktive Bereitstellung umgesetzt. Unter „report“ werden projektbezogene Berichte, Präsentationen und Dokumentationen abgelegt. Die Textdatei „readme.md“ im Hauptverzeichnis auf der obersten Ebene soll einen Überblick über die Projektinhalte geben sowie aktuelle technische Informationen, wie Installationsanweisungen und Version der verwendeter Python-Bibliotheken enthalten. Da in diesem Projekt nur ein Data Scientist an dem Projekt arbeitet wird auf das Erstellen verschiedener sogenannter „branches“ verzichtet. In der PrtaxisDiese ermöglichen es, dass verschiedene Teammitglieder in eigenen redundanten Zweigen derselben Struktur parallel arbeiten, um die Änderungen dann am Ende wieder abgestimmt in den Hauptzweig, den „main branch“, zu überführen (Vgl. Araujo, 2020).

* Versionierung
  + Daten (DVC)
  + Data-Science-Prozess (MLFlow/ Feature Store)
* Entwicklungsumgebung isolieren / portierbar machen mit Virtualisierung und freeze von Lirbaries
* Robuster Pipeline

# Datenverständnis

Datenzugriff 🡪 einfach per Excel aus Aufgabenstellung

explorativen Datenanalyse (EDA)

* Beurteilung der Datenqualität
* Untersuchung statistischer Signifikanz und Korrelationen einzelner Merkmale zur Zielvariable

*Beurteile die Qualität des zur Verfügung gestellten Datensatzes, bereite Erkenntnisse so auf und visualisiere sie , dass Businesspartner in klarer und einfacher Weise wichtigen Zusammenhänge verstehen*

Vorgaben zur Datenstruktur

|  |  |
| --- | --- |
| Feld | Strukturvorgabe |
| tmsp | Zeitstempel der Überweisung/Transaktion |
| country | Land der Überweisung |
| amount | Überweisungsbetrag |
| success | wenn “1”, dann ist die Überweisung erfolgreich |
| PSP | Name des Zahlungsdienstleisters (PSP = payments service provider) |
| 3D\_secured | wenn “1”, dann ist der Kunde 3D-identifiziert (dadurch eine noch sicherere Online-Kreditkartenzahlung) |
| card | Kreditkartenanbieter (Master, Visa, Diners) |

*Prüfung auf Vollständigkeit*

* *alle Datenfelder befüllt? wie viele Werte fehlen?*
* *Stimmt Art der Daten mit Spezifikation und Beschreibung der Quellsysteme überein? Systematische Abweichungen 🡪 Fehler in den Datenerfassungs-, Verarbeitungs- oder Übertragungsmechanismen 🡪 Hindernisse, oft nur mit einem erheblichen Arbeitsaufwand überwindbar*

# Datenaufbereitung

* Codierung in numerische Werte für maschinelles Lernmodell
* Keine weiteren Korrekturen 🡪 da keine offensichtlichen inhaltlichen oder Datenformats-Fehler erkennbar

# Modellierung

Klassifizierungsproblem 🡪 Wahrscheinlichkeit eines PSP zur Zugehörigkeit zu Klasse A: Erfolg oder B: Fehlgeschlagen

1. *erstes Basismodell (ein sogenanntes Baseline-Modell)* 🡪*Einfaches Regel-basiertes Modell als Referenz 🡪 Entscheidungsregeln mit IF-Schleifen aus EDA abgeleitet 🡪 z.B. immer Major Class*
2. *präzises Vorhersagemodellauswählen*

* *mit Konfidenzband 🡪 Punktvorhersage ohne Info zur Konfidenz nicht viel wert)*

1. logistische Regression 🡪 One-Hot-Coding von kategorischen Merkmalsvariablen mit Regularisierung
2. logistische Regression mit Random Forest Classifier 🡪 Overfit🡪Verwerfen
3. logistische Regression mit Support Vector Machines (SVM) (Königsklasse)
4. XGBoost (Extreme Gradient Boosting)

* nach Vorhersage 🡪 Regelwerk , das anhand Gebühren entscheiden, ab welchem Delta der Erfolgswahrscheinlichkeit höhere Kosten in Kauf genommen werden

# Evaluierung

* *AUC-Score 🡪 wie gut kann Modell Wahrscheinlichkeit schätzen*
* *Vorhersage-Genauigkeit*
* *Merkmalswichtigkeit*
* *Fehler/ Schwachstellen analysieren/ vergleichen 🡪 Distribution d. Fehler/ nicht nur Durchschnitssfehler!*
* *Kross-Validierung*

*Dann Feintuning…*

* *Hyperparameter*
* *Merkmale reduzieren*
* *Regularisierung*
* *Final mit allen Daten trainieren!*

*Keine allzu hohe Genauigkeit erwarten*

# Bereitstellung

# Schluss (Diskussion und weitere Übertragungen)

# Analyse und Interpretation Merkmale und Leistungsmetriken

*, solltest du die*

* *Wichtigkeit der einzelnen erklärenden Variablen diskutieren und die Modellresultate so interpretierbar wie möglich gestalten, damit die Businesspartner Vertrauen in Dein neues Modell entwickeln*
* *detaillierte Fehleranalyse sehr wichtig, damit die Businesspartner auch die Schwachstellen Deiner Herangehensweise verstehen.*

# Anwendung in der Praxis

* graphische Benutzeroberfläche (GUI) skizzieren 🡪 Tools für Echtzeitdarstellung: Streamlet/ Flask
* Modell prognostiziert pro Transaktion

# Literaturverzeichnis

Fuchs, M. (21. 08 2020). *https://michael-fuchs-python.netlify.app*. Von https://michael-fuchs-python.netlify.app: https://michael-fuchs-python.netlify.app/2020/08/21/the-data-science-process-crisp-dm/ abgerufen

Roland Gabriel, P. G. (2009). Datawarehouse & Data Mining. In P. G. Roland Gabriel, *Datawarehouse & Data Mining* (S. 123 ff.). Witten: W3L GmbH.

1. Abkürzungsverzeichnis

CRISP-DM *Cross Industry Standard Process for Data Mining*

PSP *Payment Service Provider*

# Abbildungs- und Tabellenverzeichnis

[Figure 1: CRISP-DM-Prozess-Modell, in Anl. an (Fuchs, 2020) 3](#_Toc144032474)

# Verzeichnis der Anhänge