Data Mining & Grundlagen Maschinelles Lernen 1

Wintersemester 2022/23

Semesterprojekt: Retourenvorhersage

1 Organisatorisches

1.1 Zweck und Scope

Im Modul Data Mining & Grundlagen Maschinelles Lernen 1 sollen Sie unter anderem lernen, mathematische Vorhersagemodelle für Klassifizierungs- oder Regressionsprobleme zu entwickeln und zu bewerten. Sie sollen dabei die Kenntnisse und Techniken, die im Laufe der Vorlesung vermittelt wurden, auf ein konkretes Problem anwenden, um ein Modell zu entwickeln, das auf unbekannten Daten möglichst gute Vorhersagen liefert.

Die einzelnen benötigten Techniken wurden im Laufe des Semesters bereits in kleineren Übungsprojekten angewendet. Die Lernziele der vorliegenden Aufgabe sind:

- Das Erlernte mit relativ wenigen Vorgaben für ein neues Problem einzusetzen und dabei erlerntes Wissen aus verschiedenen Einheiten zu verknüpfen.
- Sauber strukturierten und kommentierten Code zu erzeugen, der von anderen Personen übernommen und ggf. weiterentwickelt werden könnte.

1.2 Organisation und Termine

Die Übung ist durch Projektgruppen bestehend aus **bis zu vier** Studierenden umzusetzen. Die Gruppen müssen in ILIAS registriert sein.

- **Start**: 30.11.2022
- **Abgabe Jupyter Notebook (Detailhinweise s.u.): 13.12.2022, 18 Uhr.**
  Über ILIAS muss ein lauffähiges Jupyter Notebook bis zur Deadline eingereicht werden (ggf. mit Hinweisen auf benötigte Pakete/Versionen).
- **Abgabegespräch (Detailhinweise s.u.): 14.12.2022**

Das Projekt wird basierend auf dem Notebook und dem Abgabegespräch benotet. Die Projektnote macht **30% der Modulnote** aus.

1.3 Anforderungen

1.3.1 Jupyter Notebook

Das Notebook sollte sauber strukturiert und lauffähig sein (ggf. Hinweise auf verwendete Pakete und Versionsnummern). Benutzen Sie sowohl Kommentare im Code als auch Markdown Zellen um ihr Vorgehen zu erläutern. Achten Sie auch auf sinnvolle Bezeichner für die Variablen. Grundsätzlich gehe ich davon aus, dass sie die in der Vorlesung behandelten Pakete wie `scikit-learn` , `Pandas` und `NumPy` für die Analysen benutzen. Falls Sie gänzlich andere Pakete benutzen möchten, ist das grundsätzlich auch möglich. Sprechen Sie dies dann aber bitte vorher mit mir ab. Grundsätzlich gilt: benutzen Sie nur Methoden, die sie auch erklären können.

1.3.2 Vorstellung der Analysen

Jede Gruppe soll ihre Lösung in einem Abgabengespräch vorstellen. In dem Gespräch sollen sie ihr Notebook vorstellen und ihre Analysen sowie ihr Vorgehen erläutern. Ich möchte dabei verstehen

- welche Analysen und Verfahren Sie eingesetzt haben,
- warum sie sie eingesetzt haben und
- wie sie diese im Code umgesetzt haben.

Ich gehe davon aus, dass jedes Teammitglied gleichermaßen mit dem abgegeben Code vertraut ist und werde ggf. Verständnis durch Rückfragen überprüfen. Als Richtwert für das Gespräch plane ich 20-30 Minuten pro Gruppe ein.

2 Aufgabe

Im Onlinehandel wird ein signifikanter Teil der gekauften Produkte zurückgesendet. Retourenversand ist für Online-Plattformen sehr teuer, gleichzeitig erwarten Kunden einen einfachen, kostenloser Retourenprozess.

Es gibt allerdings Möglichkeiten, das Kaufverhalten von Kunden zu beeinflussen. Zum Beispiel könnte ein Onlineshop in Fällen, in denen eine Retoure wahrscheinlich ist, Zahlungs- und Versandoptionen einschränken, zusätzliche Werbung oder Pop-ups anzeigen oder ggf. den Preis erhöhen. Dafür muss für einen gegebenen Artikel und Kunden vorhergesagt werden, ob der Artikel zurückgeschickt werden wird oder nicht.

2.1 Datensatz

Der vorliegende Datensatz beschreibt 75,007 bestellte Artikel, von denen bekannt ist, ob sie zurückgesendet wurden oder nicht. Für einen bestellten Artikel (1 Zeile) existieren folgende Felder in den Daten:

- `order_item_id`: ID des bestellten Artikels
- `order_date`: Datum der Bestellung
- `delivery_date`: Lieferdatum des Artikels
- `item_id`: ID des Artikels
- `item_size`: Größe des Artikels
- `item_color`: Farbe des Artikels
- `brand_id`: Hersteller ID
- `item_price`: Preis des Artikels
- `user_id`: ID des Nutzers
- `user_title`: Titel des Nutzers
- `user_dob`: Geburtsdatum des Nutzers
- `user_state`: Bundesland, in dem der Nutzer wohnt
- `user_reg_date`: Datum, an dem sich der Nutzer registriert hat
- `return`: Wurde der Artikel zurückgeschickt (1) oder nicht (0)

	order_item_id	order_date	delivery_date	item_id	item_size	item_color	brand_id	item_price	user_id	user_title	user_dob	user_state	user_reg_date	return
0	1	2016-06-22	2016-06-27	643	38	navy	30	49.90	30822	Mrs	1969-04-17	Saxony	2016-06-23	0
1	2	2016-06-22	NaN	337	152	grey	30	19.95	30822	Mrs	1969-04-17	Saxony	2016-06-23	0
2	3	2016-06-22	2016-06-27	270	xxl	grey	49	79.90	30823	Mrs	1970-04-22	Baden-Wuerttemberg	2015-03-15	1
3	4	2016-06-22	2016-06-27	142	xxl	grey	49	99.90	30823	Mrs	1970-04-22	Baden-Wuerttemberg	2015-03-15	0
4	5	2016-06-22	2016-06-27	561	xxl	grey	3	14.90	30823	Mrs	1970-04-22	Baden-Wuerttemberg	2015-03-15	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

2.2 Aufgaben

Das Ziel der Übung ist es, ein Klassifikationsmodell auf den Daten zu trainieren, das für einen Artikel *am Ende eines Bestellvorgangs* vorhersagen kann, ob der Artikel zurückgesendet werden wird oder nicht. Hierzu wird auf unbekannten Daten eine möglichst gute Performance angestrebt.

Es gibt einen zweiten Datensatz im gleichen Format wie die Trainingsdaten. Dieser steht ihnen allerdings nicht zur Verfügung, sondern ich werde damit ihr Modell nach der Abgabe testen. Achten Sie also darauf, alle Datenverarbeitungsschritte so zu spezifizieren, dass sie automatisch auf neue, unbekannte Daten angewendet werden können.

Für die Entwicklung des Modells muss der gesamte Machine Learning Workflow durchlaufen werden, von der Datenvorverarbeitung und -analyse über das Modelltraining verschiedener Modelle bis zur Auswertung der Modellgüte.

Bearbeiten Sie dazu folgende Aufgaben:

1. Laden Sie die Daten und verschaffen Sie sich einen Überblick (Wertebereich einzelner Features, mögliche Korrelationen, Ausreißer, fehlende Werte).

2. Führen Sie geeignete Vorverarbeitungsschritte durch, z.B. Behandlung von Ausreißern und fehlenden Werten, Skalierung der Features sowie Generierung neuer Features.

3. Trainieren Sie drei verschiedene Klassifikationsmodelle, die in der Vorlesung behandelt wurden.

4. Optimieren sie die Hyperparameter der Modelle mittels Suche und Kreuzvalidierung. Überlegen Sie dazu zunächst (mit Hilfe der Vorlesungsunterlagen und der Dokumentation der Methoden in `scikit-learn`), was für die jeweiligen Modelle Hyperparameter sind und für welche sich eine Optimierung ggf. lohnen könnte. Wählen Sie die Accuracy als Zielgröße für die Hyperparameteroptimierung.

5. Erstellen Sie einen ROC Plot, in dem Sie die verschiedenen Modelle vergleichen. Welche Aussagen können Sie aus dem Plot ableiten?

6. Nehmen Sie an, dass eine Retoure Kosten von durchschnittlich 7€ verursacht (Transportkosten, evtl. Minderung des Wiederverkaufswertes). Nehmen Sie außerdem vereinfachend an, dass die weiter oben beschriebenen Maßnahmen (Einschränkung der Zahlungsoptionen etc.) wirksam sind und so eine *vorhergesagt* Retoure dazu führt, dass der Kunde den Artikel in 50% der Fälle nicht bestellt. In diesem Fall entsteht dem Händler ein Verlust von 10% des Artikelpreises, falls der Artikel fälschlicherweise als Retoure vorhergesagt wird.

   Welches ihrer Modelle wird voraussichtlich die geringsten Kosten verursachen?