Data Mining & Grundlagen Maschinelles Lernen 1

Wintersemester 2023/24

Semesterprojekt: Umsatzvorhersage

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. Die einzelnen Schritte sollen dabei nachvollziehbar dargestellt und begründet werden.
- 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: 06.12.2023**
- **Abgabe Jupyter Notebook und Vorhersagen (Detailhinweise s.u.): 12.12.2023, 19 Uhr**
- **Präsentation (Detailhinweise s.u.): 13.12.2023**

Das Projekt wird basierend auf dem Notebook, der Vorhersage auf den Testdaten und der Präsentation 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.

Versuchen Sie, alle Vorverarbeitungsschritte sowie das Modelltraining mit den von `scikit-learn` bereitgestellten Klassen wie ColumnTransformer und Pipeline abzubilden.

1.3.2 Vorstellung der Analysen

Jede Gruppe soll ihre Lösung der Aufgaben in einer Kurzpräsentation vorstellen. *Wichtig:* fassen Sie sich kurz, auf den Folien sollen komprimiert die wichtigsten *Ergebnisse* stehen (Details kann ich im Notebook nachschauen).

Halten Sie sich an folgende Vorgaben:

- Aufgabe 1: 2-3 Slides
- Aufgabe 2: 1 Slide
- Aufgabe 3: 1 Slide
- Aufgabe 4: 3 mal 1 Slide
- Aufgabe 5: 1 Slide

Die Präsentation der einzelnen Aufgaben fließt in die entsprechenden Punktzahlen (s.u.) mit ein.

Die Präsentation soll **nicht mehr als 10 Minuten** dauern. Sie kann von einem oder mehreren Teammitgliedern gehalten werden. Für Rückfragen und Diskussion (auch zum Code) plane ich weitere 10-20 Minuten ein. Ich gehe davon aus, dass jedes Teammitglied gleichermaßen mit dem abgegeben Code vertraut ist und werde ggf. Verständnis durch Rückfragen überprüfen.

1.3.3 Vorhersage auf den Testdaten

Die Vorhersage des besten Modells soll als `.csv` Datei mit abgegeben werden. Fügen Sie dazu einfach eine Spalte `Sales` zur Datei `dmml1_test.csv` hinzu, in der Sie Ihre Vorhersagen speichern.

2 Das Projekt

Eine große Supermarktkette betreibt 300 Filialen in Deutschland. Als Grundlage für die Personalplanung muss für jede Filiale entschieden werden, wie viele MitarbeiterInnen an einem bestimmten Tag in der Filiale sein müssen. Der Bedarf richtet sich dabei nach der erwarteten Anzahl von Kunden bzw. dem erwarteten Umsatz. Bisher wird der Bedarf individuell von den FilialleiterInnen ermittelt.

In dieser Aufgabe soll anhand von historischen Umsatzzahlen, Filial-, Angebots- und Wettbewerberdaten ein Modell entwickelt werden, das für jede Filiale den erwarteten Umsatz möglichst gut vorhersagt und damit allen FilialleiterInnen eine passgenaue Personalplanung ermöglicht.

2.1 Aufgaben

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:

Aufgabe 1 (10 Punkte)

Untersuchen Sie die Daten. Beantworten Sie dabei folgende Fragen bzw. Aufgaben:

1. Wie viele Datenpunkte gibt es pro Store?
2. Visualisieren Sie die Verteilung der Verkaufszahlen für einige zufällig ausgewählte Stores und beschreiben Sie typische Muster, die Sie erkennen.
3. Beschreiben Sie einige prägnante Zusammenhänge, die Ihnen zwischen verschiedenen Features auffallen.

Aufgabe 2 (10 Punkte)

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

Aufgabe 3 (15 Punkte)

Generieren Sie neue Features, die für die Vorhersage des Umsatzes aussagekräftig sein könnten.

Aufgabe 4 (30 Punkte)

Trainieren Sie drei verschiedene Modelle, die in der Vorlesung behandelt wurden: ein lineares Modell (einfache lineare Regression, Ridge, Lasso), einen Entscheidungsbaum und ein Ensemble-Modell (Gradient Boosting oder Random Forest)

1. Optimieren sie 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.
2. Welches sind die wichtigsten Features für die jeweiligen Modelle?

Aufgabe 5 (15 Punkte)

Das Ziel der Modelle ist es, die FilialleiterInnen bei der Erstellung der Schichtpläne zu unterstützen. Dazu leiten diese aus der Umsatzvorhersage ihres Modells einen Bedarf an MitarbeiterInnen ab, die an diesem Tag zur Schicht eingeteilt werden. Wenn die Umsatzvorhersage deutlich höher ist als der tatsächliche Umsatz, so werden zu viele Mitarbeiter in der Filiale sein und es entstehen unnötige Lohnkosten. Nehmen Sie an, dass pro Tag, an dem die Vorhersage den tatsächlichen Umsatz um mehr als 3000 EUR überschätzt 150 EUR Kosten (zusätzliche Lohnkosten) anfallen. Wenn die Vorhersage um mehr als 6000 EUR höher liegt, betragen die Kosten sogar 250 EUR. Falls die Vorhersage den tatsächlichen Umsatz unterschätzt, so sind die Kosten schwieriger zu quantifizieren. Darunter fallen Kosten für etwaige Überstunden aber auch Kosten durch Unzufriedenheit der Kunden und der Mitarbeiter (da z.B. lange Wartezeiten in Kauf genommen werden müssen), die sich langfristig negativ auswirken. Nehmen Sie vereinfachend an, dass 100 EUR Kosten entstehen, wenn das Modell den Umsatz um mehr als 4000 EUR unterschätzt. Welches ihrer Modelle wird nach diesen Annahmen voraussichtlich die geringsten Gesamtkosten verursachen?

Aufgabe 6 (20 Punkte)

Erstellen Sie eine Vorhersage für die Testdaten mit dem besten Modell (bezüglich ihrer Metrik aus der vorherigen Aufgabe). Geben Sie diese Vorhersage als `.csv` Datei mit ab. Fügen Sie dazu einfach eine Spalte `Sales` zur Datei `dmml1_test.csv` hinzu.

2.2 Beschreibung der Daten

Die Trainingsdaten bestehen aus zwei Dateien: `dmml1_stores.csv`, die Stammdaten zu den Filialen enthält und `dmml1_sales.csv`, die Umsatzdaten (pro Filiale und Tag) enthält. Darüber hinaus gibt es eine Datei `dmml1_test.csv`, die genauso strukturiert ist wie die Datei `dmml1_sales.csv`, aber in der die Spalten `Sales` und `Customers` fehlen.

Die Spalten in diesen Dateien haben folgende Bedeutung:

- Store ID - eine eindeutige Kennung für jedes Geschäft
- Sales - der Umsatz für einen bestimmten Tag (das ist das, was Sie vorhersagen sollen)
- Customers - die Anzahl der Kunden an einem bestimmten Tag (ist im Voraus natürlich nicht bekannt)

- Open - ein Indikator dafür, ob der Laden geöffnet war: 0 = geschlossen, 1 = geöffnet
- StateHoliday - gibt einen gesetzlichen Feiertag an. Normalerweise sind alle Geschäfte, mit wenigen Ausnahmen, an gesetzlichen Feiertagen geschlossen. a = Feiertag, b = Ostern, c = Weihnachten, 0 = kein Feiertag
- SchoolHoliday - gibt an, ob im Bereich des Geschäfts an dem betreffenden Datum Schulferien waren
- StoreType - unterscheidet zwischen 4 verschiedenen Ladenmodellen: a, b, c, d
- Assortment - beschreibt eine Sortimentsstufe: a = basic, b = extra, c = extended
- CompetitionDistance - Entfernung in Metern zum nächstgelegenen Konkurrenzmarkt
- CompetitionOpenSince[Month/Year] - gibt das ungefähre Jahr und den Monat an, in dem der nächstgelegene Konkurrenzmarkt eröffnet wurde
- Promo - gibt an, ob ein Geschäft an diesem Tag eine Werbeaktion durchführt
- Promo2 - Promo2 ist eine fortlaufende und aufeinanderfolgende Werbeaktion für einige Geschäfte: 0 = Geschäft nimmt nicht teil, 1 = Geschäft nimmt teil
- Promo2Seit[Jahr/Woche] - beschreibt das Jahr und die Kalenderwoche, in der die Filiale mit der Teilnahme an Promo2 begonnen hat
- PromoInterval - beschreibt die aufeinanderfolgenden Intervalle, in denen Promo2 gestartet wird, wobei die Monate genannt werden, in denen die Aktion neu gestartet wird. Z.B. "Feb,May,Aug,Nov"bedeutet, dass jede Runde im Februar, Mai, August, November eines jeden Jahres beginnt.