Applied Machine Learning - Mini-Challenge

Cross-Selling von Kreditkarten

Daniel Perruchoud, Institut für Data Science

1 AML Leistungsnachweis

Im Modul "Applied Machine Learning (AML)" beruht der Leistungsnachweis auf der Demonstration praktischer und theoretischer Kompetenzen.

Die praktische Kompetenz wird mit Hilfe einer Mini-Challenge geprüft, die theoretische Kompetenz mittels mündlicher MSP, welche nach Abgabe der Mini-Challenge absolviert wird.

Die Gesamtbeurteilung setzt sich zusammen aus benoteter Mini-Challenge und mündlicher MSP (Gewicht je 50%).

2 Mini-Challenge Grundlagen

2.1 Software

Die Mini-Challenge kann mit R oder Python bearbeitet werden, wobei die auf die konsistente Verwendung von Frameworks zu achten ist, insbesondere `tidymodels` für R und `sklearn` für Python. Die durch diese Frameworks bereitgestellten Funktionen sind gezielt einzubinden und mit eigenen Funktionen zu ergänzen.

2.2 Daten

Die Daten für die Mini-Challenge sind abgelegt unter "Aufgaben", deren Inhalte sind beschrieben unter Beschreibung (navigiere zu PKDD'99 Challenge > Data > Financial data description).

3 Mini-Challenge Abgabebedingungen

3.1 Abgabeobjekte

3.1.1 Notebooks

Analysen sind in Form von Notebooks einzureichen, wobei neben dem .Rmd-File oder .ipynb-File auch eine als .html oder .pdf gerenderte Version abzugeben ist. Alle Analysen sind vor der Einreichung am Stück auszuführen, Idee und Durchführung der Analysen sind präzise zu beschreiben, Resultate verständlich zu dokumentieren und interpretieren.

3.1.2 Code Repositories

Zudem ist eine gut strukturierte und dokumentierte Ablage der finalen und lauffähigen Codes (R oder Python) mit Anleitung zugänglich zu machen, wobei Zweck und Ausführung einzelner Codes sowie zusätzlich zu installierender Bibliotheken (s. R's `sessionInfo()` oder requirements.txt oder environment.yml File für Python) bereitzustellen sind. Eine Auslagerung wiederkehrend verwendeter Funktionalitäten ist anzustreben (z.B. Skript-File, Library oder Package).

Titel der Mini-Challenge und Autorenschaft sind im Namen zu vermerken. Die Analysen sind termingerecht per e-mail an "daniel.perruchoud@fhnw.ch" zu senden.

3.2 Erarbeitung

Die Mini-Challenge darf allein oder in 2-er Gruppen erarbeitet werden.

Hinweise:

Eine Zusammenarbeit zwischen Gruppen darf sich nur auf konzeptionelle Aspekte beschränken, insbesondere darf kein Code von anderen Gruppen oder aus dem Internet kopiert werden.

3.3 Hilfsmittel

Die Verwendung von ChatGPT oder vergleichbaren AI-Tools ist erlaubt. Ihre Verwendung ist im Lieferobjekt für entsprechende Code-Stücke zu vermerken und übergeordnet am Ende der Analyse in einem separaten Abschnitt kurz zu beurteilen und diskutieren (Länge 250-500 Wörter). Zu beurteilen sind dabei für welche Task das AI-Tool eingesetzt worden ist, und welche Ansprache-Strategie (Prompting Strategie) verwendet wurde. Zudem soll beschrieben werden, welche Prompting Strategie am erfolgreichsten war, d.h. am meisten a) zum Lösen der Aufgabe und b) zum Kompetenzerwerb beitragen konnte.

3.4 Abgabetermin

Der Abgabetermin für die Mini-Challenge ist der 14. Juni 2024.

4 Mini-Challenge Beurteilungskriterien

Die Notenvergabe beruht auf den unten aufgeführten Beurteilungskriterien.

4.1 Vollständigkeit

Die Analysen sind inhaltlich vollständig gemäss Beschreibung der Mini-Challenge zu lösen.

4.2 Korrektheit

Die abgebenen Analysen werden auf inhaltliche Korrektheit geprüft. Lauffähigkeit der Codes und dokumentierte Tests eigener Funktionen sind notwendige Grundvoraussetzung dafür.

4.3 Nachvollziehbarkeit

Die Analysen sind so zu gestalten, dass sowohl die

- zugrundeliegende Überlegungen,
- deren Implementierung und
- die abgeleiteten Resultate

nachvollziehbar sind. Das setzt voraus, dass die

- Notebooks als Ganzes optimal strukturiert sind,
- Codes gemäss Best Practice Standards strukturiert, formatiert und kommentiert sind,
- Analyseresulate mit Tabellen und Grafiken dargestellt und auch mit Text vollständig diskutiert sind.

4.4 Best Practice Standards

Zudem sollen Wiederholung durch kopieren von Code vermieden und in Funktionen ausgelagert werden, welche vor Verwendung sinnvoll getestet werden.

5 Mini-Challenge Inhalte & Lernziele

Inhalt der Mini-Challenge ist die Entwicklung und Evaluierung von Affinitätsmodellen für personalisierte Kreditkarten-Werbekampagnen im Auftrag einer Bank. Methodisch kommt dabei die binäre Klassifikation für tabellarische Daten zum Einsatz, im Zentrum stehen aber vor allem Aspekte der Anwendung, wie sie in der Praxis anzutreffen sind, nämlich:

- Aufbereitung eines Modellierungsdatensatz aus transaktionellen Datenbeständen,
- Modellentwicklung und systematischer Performance-Vergleich,
- Vergleich der Haupteinflussfaktoren und Top-N Listen der Modelle,
- Modell-Selektion sowie systematische Hyperparameter-Optimierung,
- Modellvereinfachung und -beschreibung für Non-Data Scientist.

5.1 Mini-Challenge Lernziele

Lernziele: Kenntnis von Vor- und Nachteilen verschiedener

- Ansätze zur Behandlung unbalancierter Daten,
- Modellperformance-Metriken,
- Methoden zur Hyperparameter-Optimierung,
- analytischer Verfahren zur Erklärung prädiktiver Modelle,
- praxis-relevanter Ansätze zur Beschreibung prädiktiver Modelle,
- baum-basierter Vorhersagemodelle.

6 Product Affinity Modeling - Spezifikation und Anleitung

6.1 Mini-Challenge Vorgehen

Ziel ist es Kundenlisten für eine personalisierte Kreditkarten-Werbekampagne zu erzeugen, wobei keine Junior-Karten angeboten werden sollen.

Die nachfolgenden Schritte sollen helfen die analytischen Aufgaben im Sinn einer Roadmap zu planen.

Datenaufbereitung

1. Laden, transformieren und überprüfen der Datenqualität mittels explorativer Datenanalyse.
2. Kombinieren der Informationen zu Kunden und Bankdienstleistungen.
3. Bereinigung der Grundmenge in Hinblick auf die Modellentwicklung.

Modellkonstruktion

4. Identifizieren bestehender Kreditkartenkäufer inkl. Bestimmung des Kaufdatums und Rollup-Fensters.
5. Bestimmen der Nicht-Käufer zum Vergleich (inkl. Rollup-Fenster).
6. Erzeugen event-bezogener Kundeninformationen vor Kreditkartenkauf auf Basis der Transaktionshistorie (analog für Nicht-Käufer).

Feature Engineering

7. Herleiten Kunden-spezifischer, statistischer Kennzahlen für Vermögen und Umsätze im Rollup-Fenster.
8. Kombinieren event-bezogener Informationen von Kreditkarten-Käufern und Nicht-Käufern.
9. Bereinigen unnötiger Informationen (z.B. IDs) und Überprüfen der Struktur der Modellierungsdaten mittels explorativer Datenanalyse.

Modellentwicklung

10. Partitionieren der Daten in Trainings- und Testdaten.
11. Erstellen eines Baseline Modelles mittels logistischer Regression und den Informationen “Alter”, “Geschlecht”, “Domizilregion”, “Vermögen” und “Umsatz” vor Kreditkartenkauf.
12. Systematisches Explorieren von Verbesserungsmöglichkeiten des Baseline Modelles durch Erweiterung erklärender Variablen und Verwendung anderer Algorithmen.

Modellvergleich, -selektion und -optimierung

13. Vergleichen der Kandidatenmodelle und identifizieren des bzgl. Performance “besten” Modelles mit ROC, AUC und Precision.
14. Quantitatives Untersuchen der Unterschiede von Top-N Kunden-Listen verschiedener Modelle.
15. Optimieren des “besten” Kandidatenmodelles hinsichtlich Hyperparameter-Einstellungen.

Modellerklärung und -reduktion

16. Untersuchen der globalen Wichtigkeit der Einflussfaktoren des “besten” Modelles und Modellreduktion (Trade-off von globaler Wichtigkeit und Modellperformance).
17. Beschreiben des Mehrwerts des “finalen” Modelles in der Praxis.

Die oben skizzierten, in der Praxis verwendeten Ansätze und Methoden sind in Lehrbüchern nicht einfach auffindbar. Im Verlauf des Semesters werden deshalb gezielt Hinweise in Form von JITTs und Spaces-Posts geliefert.

6.2 Mini-Challenge Lieferobjekte

Umfang und Ausarbeitung der Analysen ist nicht näher definiert. Die folgenden, minimalen Lieferobjekte sind jedoch im Notebook in Form von tabellarischen oder visuellen Artefakten zu integrieren:

- Entity-Relationship Diagramm der Grunddaten.
- Übersicht der Data Pipeline zur Kombination der Daten.
- Anzahl Kreditkarten-Käufer und Nicht-Käufer mit kompletter 12 Monate-Rollup Information.
- Verteilung der Kaufzeitpunkte der Kreditkarten-Käufer bzw. Vergleichszeitpunkte der Nicht-Käufer.
- Übersicht der selber konstruierten Predictive Features.
- Übersicht der zeitlichen Entwicklung von Vermögen und Umsatz für Konto Nummer 14 und Nummer 18 pro Monat.
- Übersicht des Baseline-Modelles und der verwendeten Kandidaten-Modelle inkl. deren Parametrisierung und Predictive Features.
- Performance-Vergleich von Baseline und Kandidaten-Modellen via Kreuzvalidierung mit sinnvollen visuellen Metriken und Kennzahlen (nach Modellselektion auch analog für Testdaten).
- Übersicht der Funktionsweise der Wichtigkeit der Predictive Features für Baseline und Kandidaten-Modellen mittels Interpretable ML / Explainable AI Verfahren.
- Quantifizierender Vergleich der Unterschiede von Top-5%, Top-10% Kunden-Listen für Baseline und Kandidaten-Modelle in Hinblick auf Konsistenz oder Differenz.
- Lift Kurve und quantitative Beschreibung zentraler Predictive Features des finalen Modelles für Non-Data Scientists.

WICHTIG: Studierende planen und kommunizieren in Hinblick auf Klärung des Verständnisses und Abgabe rechtzeitig mindestens einen Kontaktstunden-Termin mit dem Fachexperten.