CODING CHALLENGE – DATA SCIENTIST

Aufgabe

Unter https://www.openml.org/d/41214 und https://www.openml.org/d/41215 finden Sie zwei Datensätze eines französischen Automobilversicherers. Diese beinhalten Risikomerkmale und Schadeninformationen zu Kraftfahrt-Haftpflicht-Versicherungsverträgen (eine Datensatzbeschreibung finden Sie am Ende dieses Textes). Ihre Aufgabe besteht in der Modellierung der zu erwartenden Schadenhöhe pro Versicherungsnehmer und Jahr anhand der Risikomerkmale der Kunden. Dieser Wert ist Basis für die Berechnung eines fairen Versicherungsbeitrags.

Die Ergebnisse ihrer Auswertungen stellen Sie im Rahmen des 60-minütigen technischen Interviews vor. Dabei haben Sie zunächst 15 Minuten Zeit, um die aus ihrer Sicht wesentlichen Ergebnisse vorzustellen. Die Form der Präsentation (frei, Folien, Jupyter-Notebook, R-Markdown…) wählen Sie dabei selbst. Anschließend findet eine Diskussion statt, in deren Rahmen Ihnen Fragen zu ihrer Vorgehensweise sowie zu Ihrem Programmcode gestellt werden. Es sollte anhand ihrer Arbeit gut nachvollziehbar sein, wie sie vorgegangen sind (beispielsweise anhand eines Jupyter-Notebooks oder R-Markdowns).

Gehen Sie dabei in folgenden Teilschritten vor:

- **Explorative Datenanalyse:** Machen Sie sich mit dem Datensatz vertraut. Identifizieren Sie dabei mögliche Probleme sowie grundlegende statistische Zusammenhänge, welche für die anschließende Modellierung wichtig sein könnten.
- **Feature Engineering:** Bereiten Sie, soweit für ihre Modellierung nötig, die Variablen geeignet auf.
- **Modellvergleich:** Entscheiden Sie sich für ein geeignetes Modell anhand einer dafür geeigneten Metrik. Erläutern Sie wie Sie dabei vorgehen und begründen Sie ihre Entscheidung.
- **Modellbuilding:** Trainieren Sie unter Berücksichtigung der vorangegangenen Schritte das von Ihnen gewählte Modell zur Vorhersage der erwarteten Schadenhöhe pro Kunde und Jahr. Ihr Ziel ist es, einen möglichst fairen Versicherungsbeitrag pro Jahr für einzelne Kunden anhand der Ihnen zu Verfügung stehenden Merkmale zu bestimmen. Wählen Sie mindestens eine geeignete Metrik, um die Güte des finalen Modells zu beurteilen. Zeigen Sie, welche Variablen und Zusammenhänge für Ihr finales Modell relevant sind. Überlegen Sie sich (ohne dies umzusetzen) wie Sie Ihr Modell weiter optimieren könnten.

Diese Coding Challenge ist in etwa 5h gut zu erledigen. Wir erwarten nicht, dass Sie mehr Zeit investieren. Bitte teilen Sie sich ihre Zeit entsprechend ein und konzentrieren Sie sich auf das – aus ihrer Sicht - Wesentliche. Eine erschöpfende Tiefenanalyse der Daten und oder eine aufwändige Optimierung vieler verschiedener Modelle werden dementsprechend nicht von Ihnen erwartet.

Datensatzbeschreibung

freMTPL2freq:

- IDpol: ID des Vertrags
- ClaimNb: Anzahl Schäden im Versicherungszeitraum
- Exposure: Länge des Versicherungszeitraums (in Jahren) [Komponente der Zielvariable]

- Area: Area-Code des Versicherungsnehmers [unabhängige Variable]
- VehPower: Leistung des versicherten Kfz [unabhängige Variable]
- VehAge: Alter des versicherten Kfz [unabhängige Variable]
- DrivAge: Alter des Versicherungsnehmers [unabhängige Variable]
- BonusMalus: Schadenfreiheitsrabatt (französische Entsprechung der Schadenfreiheitsklasse) [unabhängige Variable]
- VehBrand: Marke des versicherten Kfz [unabhängige Variable]
- VehGas: Antrieb des versicherten Kfz [unabhängige Variable]
- Density: Anzahl der Einwohner pro km^2 im Wohnort des Versicherungsnehmers [unabhängige Variable]
- Region: Region des Versicherungsnehmers [unabhängige Variable]

freMTPL2sev:

- IDpol: ID des Vertrags
- ClaimAmount: Höhe der einzelnen Schadenaufwände (mehrere Einträge pro Vertrag, falls im Zeitraum mehrere Schäden vorhanden waren.) [Komponente der abhängigen Variable]

Die Zielvariable ist definiert als ClaimAmount je Versicherungsnehmer geteilt durch Exposure des Versicherungsnehmers

Hinweis:

Der Datensatz steht öffentlich nur im .arff Format zur Verfügung. Sie können z.B. mit Hilfe folgender Code-Zeile die Daten in Python einlesen und zu einem Pandas Dataframe konvertieren:

```
!pip install arff

import pandas as pd
import arff
data_freq = arff.load('freMTPL2freq.arff')
df_freq = pd.DataFrame(data_freq, columns=["IDpol", "ClaimNb", "Exposure", "Area", "VehPower", "VehAge","DrivAge", "BonusMalus", "VehBrand", "VehGas", "Density", "Region"])

data_sev = arff.load('freMTPL2sev.arff')
df_sev = pd.DataFrame(data_sev, columns=["IDpol", "ClaimAmount"])
```