Projekt „Data Science“ (Mission2Move)

Sie erhalten einen Datensatz zur Analyse im Rahmen des Projekts. Der Datensatz ist relativ neu und wurde auf der Konferenz NeurIPS 2022 vorgestellt. Es handelt sich um einen synthetischen Datensatz, der mithilfe eines CTGAN erzeugt wurde. Das Anwendungsgebiet ist der Betrug mit Bankkonten. Die daraus abgeleitete Aufgabe ist eine Klassifikationsaufgabe: Die Zielvariable gibt an, ob bei dem jeweiligen Bankkonto ein Betrug stattgefunden hat oder nicht.

Der Datensatz ist realistisch hinsichtlich seiner Größe, der verwendeten Merkmale und seiner Zusammensetzung. Er wurde vor allem in Hinblick auf Fairness von Modellen entwickelt und enthält drei sog. geschützte Attribute (Merkmale): *Altersgruppe*, *Beschäftigungsstatus* und *Einkommen*. In diesem Projekt soll der Einfachheit halber aber nur die Leistungsfähigkeit des finalen Modells bewertet werden.

Der Datensatz existiert in verschiedenen Varianten bzgl. seiner Zusammensetzung hinsichtlich der geschützten Attribute. Sie erhalten eine Variante, bei der die Gruppenzusammensetzung gleich ist und die Prävalenz, also das relative Vorkommen von Betrug in beiden Gruppen etwa gleich hoch. Wie oben bemerkt, ist auch lediglich die Klassifikationsleistung des finalen Modells wichtig. Da es sich um einen unbalancierten Datensatz bzgl. der Zielgröße handelt, ist hier die Frage der Metrik zur Beurteilung der Klassifikationsleistung besonders wichtig.

Der Datensatz steht unter dem Namen **`bankaccounts.csv`** auf dem Jupyterhub unter dem Datenverzeichnis zur Verfügung.

Sie finden weitere Informationen zur Erzeugung des Datensatzes und vor allem zur Bedeutung der Merkmale unter

`https://github.com/feedzai/bank-account-fraud/blob/main/documents/datasheet.pdf`

Die einzelnen Merkmale bedeuten (Auszug aus dem verlinkten Dokument, die geschützten Attribute sind rot gedruckt):

- **income** (numeric): Annual income of the applicant (in decile form). Ranges between [0.1, 0.9].
- **name_email_similarity** (numeric): Metric of similarity between email and applicant’s name. Higher values represent higher similarity. Ranges between [0, 1].
- **prev_address_months_count** (numeric): Number of months in previous registered address of the applicant, *i.e.* the applicant’s previous residence, if applicable. Ranges between [−1, 380] months (-1 is a missing value).
- **current_address_months_count** (numeric): Months in currently registered address of the applicant. Ranges between [−1, 429] months (-1 is a missing value).
- **customer_age** (numeric): Applicant’s age in years, rounded to the decade. Ranges between [10, 90] years.
- **days_since_request** (numeric): Number of days passed since application was done. Ranges between [0, 79] days.
- **intended_balcon_amount** (numeric): Initial transferred amount for application. Ranges between [−16, 114] (negatives are missing values).
- **payment_type** (categorical): Credit payment plan type. 5 possible (annonymized) values.

- **zip_count_4w** (numeric): Number of applications within same zip code in last 4 weeks. Ranges between [1, 6830].
- **velocity_6h** (numeric): Velocity of total applications made in last 6 hours *i.e.*, average number of applications per hour in the last 6 hours. Ranges between [−175, 16818].
- **velocity_24h** (numeric): Velocity of total applications made in last 24 hours *i.e.*, average number of applications per hour in the last 24 hours. Ranges between [1297, 9586]
- **velocity_4w** (numeric): Velocity of total applications made in last 4 weeks, *i.e.*, average number of applications per hour in the last 4 weeks. Ranges between [2825, 7020].
- **bank_branch_count_8w** (numeric): Number of total applications in the selected bank branch in last 8 weeks. Ranges between [0, 2404].
- **date_of_birth_distinct_emails_4w** (numeric): Number of emails for applicants with same date of birth in last 4 weeks. Ranges between [0, 39].
- **employment_status** (categorical): Employment status of the applicant. 7 possible (annonymized) values.
- **credit_risk_score** (numeric): Internal score of application risk. Ranges between [−191, 389].
- **email_is_free** (binary): Domain of application email (either free or paid).
- **housing_status** (categorical): Current residential status for applicant. 7 possible (annonymized) values.
- **phone_home_valid** (binary): Validity of provided home phone.
- **phone_mobile_valid** (binary): Validity of provided mobile phone.
- **bank_months_count** (numeric): How old is previous account (if held) in months. Ranges between [−1, 32] months (-1 is a missing value).
- **has_other_cards** (binary): If applicant has other cards from the same banking company.
- **proposed_credit_limit** (numeric): Applicant's proposed credit limit. Ranges between [200,2000].
- **foreign_request** (binary): If origin country of request is different from bank's country.
- **source** (categorical): Online source of application. Either browser (INTERNET) or app (TELEAPP).
- **session_length_in_minutes** (numeric): Length of user session in banking website in minutes. Ranges between [−1, 107] minutes (-1 is a missing value).
- **device_os** (categorical): Operative system of device that made request. Possible values are: Windows, macOS, Linux, X11, or other.
- **keep_alive_session** (binary): User option on session logout.
- **device_distinct_emails** (numeric): Number of distinct emails in banking website from the used device in last 8 weeks. Ranges between [−1, 2] emails (-1 is a missing value).
- **device_fraud_count** (numeric): Number of fraudulent applications with used device. Ranges between [0, 1].
- **month** (numeric): Month where the application was made. Ranges between [0, 7].
- **fraud_bool** (binary): If the application is fraudulent or not.

Hinweis: Der Ihnen zur Verfügung gestellte Datensatz enthält noch zwei weitere Merkmale x1 und x2. Bitte löschen Sie die entsprechenden beiden Spalten nach dem Lesen der csv-Datei aus dem Datensatz!

Bevor Sie ein Modell mit dem Datensatz trainieren, beachten Sie mindestens die folgenden Empfehlungen:

- Schauen Sie sich an, wie die Verteilung hinsichtlich der geschützten Attribute aussieht.
- Schauen Sie sich an, wie die Häufigkeiten der Werte der Zielgröße insgesamt und bzgl. den Gruppen bei den geschützten Attributen sind.
- Betrachten Sie, welche Merkmale fehlende Werte aufweisen.
- Überlegen Sie sich, ob Sie bestimmte Merkmale (zusätzlich zu x1 und x2) vor der weiteren Verwendung des Datensatzes entfernen.
- Betrachten Sie, welche Merkmale kategoriale Merkmale sind und überlegen Sie sich, ob Sie diese sinnvoll durch Dummyvariablen ersetzen können.
- Überlegen Sie sich abhängig vom zu trainierenden Modell, ob Skalierungen bei den einzelnen Merkmalen sinnvoll sind.
- Überlegen Sie sich, wie Sie eine Trennung in Trainings- und Testdatensatz gestalten.
- Überlegen Sie sich auch, ob Sie mithilfe der im Tutorium besprochenen Methode eine Hyperparameteroptimierung durchführen.
- Überlegen Sie sich, welche Metrik für die Leistung Ihres Modells in einem solchen unbalancierten Datensatz sinnvoll ist.
- Überlegen Sie sich, welche Klassifikationsmodelle infrage kommen und testen sie die Entwicklung mit zwei oder drei Modellen.

Bereiten Sie zusätzlich zu Ihrem Notebook für den abschließenden Vortrag **max. 8 Folien** vor, in denen Sie Ihre Überlegungen und Analysen zusammenfassen. Geben Sie bitte spätestens 2 Tage vor dem Vortragstermin Ihr Notebook per Email (ulrich.klauck@hs-aalen.de) bei mir ab. Beachten Sie bitte, dass Sie alle Ausgabezellen vor dem Versand löschen, damit das Notebook nicht zu groß wird. Die Folien können Sie mir gerne nach dem Vortrag schicken.

Nun wünsche ich Ihnen viel Freude bei Ihrem Projekt.