Wstęp do Uczenia Maszynowego

Projekt

Termin oddania projektu: 23.12.2024

Termin prezentacji: 15.01.2025

1 Wstęp

Celem jest zaproponowanie metody klasyfikacji, która pozwoli zbudować model o jak największej mocy predykcyjnej. Dysponujemy zbiorem danych dotyczącym przewidywania szans na niewypłacalność (scoring ryzyka kredytowego). Należy dokonać klasyfikacji do dwóch klas. Dokładność modelu będzie mierzona za pomocą miary zrównoważonej dokładności (*balanced accuracy*).

Projekt jest wykonywany samodzielnie!

2 Zbiór danych

Dane do projektu to zbiór, który zawiera 23 zmienne objaśniające. Zbiór treningowy zawiera 3660 obserwacji, natomiast zbiór testowy 1360.
Dostępne są następujące pliki:

- zbiór treningowy: `X_train.csv`,
- etykiety zbioru treningowego: `y_train.csv`,
- zbiór testowy: `X_test.csv`.

Opis kolumn wraz z ich znaczeniem jest zawarty w tabeli na końcu treści pracy domowej. Wartości takie jak $-9, -8, -7$ należy traktować jako brak danych.
Aby wczytać zbiór danych w języku Python wystarczy użyć funkcji `read_csv` z pakietu `pandas`.

3 Oczekiwany wynik

Na przygotowanie rozwiązania projektu będą składały się następujące elementy:

- jakość predykcji na zbiorze testowym mierzona przez *balanced accuracy*, tj.

$$BA = \frac{1}{2}\Big(\frac{TP}{P} + \frac{TN}{N}\Big),$$

- raport opisujący wykorzystane metody, decyzje podjęte prowadzące do wyboru ostatecznego modelu i wyniki eksperymentów (maksymalnie 4 stron A4),
- krótka prezentacja podsumowująca rozwiązanie (maksymalnie 4 minuty).

4 Szczegóły rozwiązania

Zbiór treningowy oraz etykiety do zbioru treningowego należy wykorzystać do przygotowania modelu. Oczekiwany wynik to wektor prawdopodobieństw przynależności do klasy 1 dla obserwacji ze zbioru testowego.

Rozwiązanie powinno zawierać pliki:

- `NUMERINDEKSU_prediction.txt` - prawdopodobieństwo przynależności do klasy 1 dla danych testowych, gdzie `1 = "Bad"` (przykładowy plik `example_prediction.txt`, trzeba pamiętać, aby kolejność obserwacji pozostała niezmieniona),
- folder `Kody` zawierający wszystkie potrzebne kody do przygotowania rozwiązania projektu,
- plik `NUMERINDEKSU_raport.pdf` opisujący wykorzystane metody, decyzje podjęte prowadzące do wyboru ostatecznego modelu i wyniki eksperymentów (maksymalnie 4 strony),
- plik `NUMERINDEKSU_prezentacja.pdf` krótka prezentacja podsumowująca rozwiązanie (maksymalnie 4 minuty)

5 Ocena

Łączna liczba punktów do zdobycia jest równa 40, w tym:

- jakość kodu (porządek, czytelność, obszerność eksperymentów) - 12 punktów,
- jakość predykcji rozwiązania* - 5 punktów,
- raport - 18 punktów,
- prezentacja - 5 punktów.

* - jakość predykcji jest oceniana miarą *balanced accuracy* na zbiorze testowym. Wyniki zostaną ustawione w ranking (od najlepszego do najgorszego). Osoba z najlepszym wynikiem (najbliższym wartości 1) zyskuje 5 punktów. Osoba z najgorszym wynikiem (najbliższym wartości 0) zyskuje 2.5 punktów. Pozostałe wyniki zostaną przeskalowane i zaokrąglone do wartości 0.1.

6 Oddanie projektu

Wszystkie punkty (oprócz prezentacji) z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NUMERINDEKSU_GR_projekt`, gdzie

$$\texttt{GR} = \begin{cases} 1 & \text{dla środa, 12:15,} \\ 2 & \text{dla środa, 14:15.} \\ 3 & \text{dla środa, 16:15.} \end{cases}$$

Tak przygotowany katalog należy przesłać na adres *anna.kozak@pw.edu.pl* do dnia 23.12 do godziny 23:59. Tytuł wiadomości: *[WUM][Projekt] Nazwisko Imię, Numer grupy: GR.*

Prezentację należy wgrać do folderu `Prezentacje projektu/GR_` dostępnego na MS Teams do dnia 14.01.2025.

7 Terminy

1. Oddanie projektu - 23.12.2024,
2. Wyniki projektu - 01.01.2025,
3. Prezentacje na zajęciach laboratoryjnych - 13 tydzień zajęć, 15.01.2025.

Tabele

Tabela 1: Opis zmiennych.

Nazwa zmiennej	Opis
X1	Zbiorcza ocena wskaźników ryzyka.
X2	Ile miesięcy minęło od otwarcia najstarszego konta.
X3	Ile miesięcy minęło od otwarcia najnowszego konta.
X4	Średnia liczba miesięcy, od kiedy istnieją wszystkie konta.
X5	Liczba kont z pozytywną historią spłat.
X6	Liczba kont, na których zalegano z płatności co najmniej 60 dni.
X7	Liczba kont, na których zalegano z płatności co najmniej 90 dni.
X8	Procent kont, na których nigdy nie było opóźnień w płatnościach.
X9	Ile miesięcy minęło od ostatniego opóźnienia w płatnościach.
X10	Najpoważniejsze opóźnienie lub wpis w rejestrze w ostatnich 12 miesiącach.
X11	Najpoważniejsze opóźnienie w historii konta.
X12	Liczba wszystkich kont kredytowych.
X13	Liczba kont otwartych w ciągu ostatnich 12 miesięcy.
X14	Procent kont ratalnych (np. kredytów).
X15	Ile miesięcy minęło od ostatniego zapytania o kredyt (bez zapytań z ostatnich 7 dni).
X16	Liczba zapytań o kredyt w ostatnich 6 miesiącach.
X17	Liczba zapytań o kredyt w ostatnich 6 miesiącach (bez zapytań z ostatnich 7 dni, które mogą wynikać z porównywania ofert).
X18	Procent wykorzystania dostępnego limitu na kontach obrotowych (saldo podzielone przez limit kredytowy).
X19	Procent spłaty kredytów ratalnych (saldo podzielone przez początkową kwotę pożyczki).
X20	Liczba kont obrotowych z niespłaconym saldem.
X21	Liczba kont ratalnych z niespłaconym saldem.
X22	Liczba kont bankowych/krajowych z wysokim wskaźnikiem wykorzystania limitu.
X23	Procent kont z niespłaconym saldem.

Tabela 2: Opis zmiennej `X10`.

Wartość	Znaczenie
0	poza skalą
1	120+ dni zaległości
2	90 dni zaległości
3	60 dni zaległości
4	30 dni zaległości
5, 6	nieznane zaległości
7	bieżące i nigdy nie zaległe
8, 9	wszystkie inne

Tabela 3: Opis zmiennej `X11`.

Wartość	Znaczenie
1	nie ma takiej wartości
2	poza skalą
3	120+ dni zaległości
4	90 dni zaległości
5	60 dni zaległości
6	30 dni zaległości
7	nieznane zaległości
8	bieżące i nigdy nie zaległe
9	wszystkie inne