PFR X CHALLENGE IT

ZADANIE DLA ŚCIEŻKI ANALITYCZNO-PROGRAMISTYCZNEJ

CZEŚĆ!

Jeśli czytasz ten tekst to prawdopodobnie bierzesz udział w **PFR x Challenge IT**, czyli hackatonie organizowanym przez PFR oraz SKN-K. Poniżej znajdziesz krótkie wprowadzenie do zadania, jego treść oraz krótki słownik dla danych, które są niezbędne do pracy.

Zadanie dla ścieżki analityczno-programistycznej opiera się o dane pochodzące z programu PPK, o którym kilka słów znajdziecie poniżej.

Pracownicze Plany Kapitałowe to prywatny i dobrowolny dla pracownika system długoterminowego oszczędzania. Oszczędności budowane są wspólnie przez pracowników, pracodawców oraz państwo.

Rachunki są zasilane wpłatami pracownika i pracodawcy oraz wpłatą powitalną i dopłatami rocznymi od państwa. Wpłaty pracownika oraz pracodawcy są naliczane procentowo od wysokości wynagrodzenia. Państwo z kolei przekazuje ustalone kwoty (po spełnieniu odpowiednich warunków) – niezależne od dochodów pracownika.

Uczestnictwo w PPK jest dobrowolne. Oznacza to, że pracownik w każdej chwili może zarówno zrezygnować z oszczędzania w PPK, jak i do niego wrócić. Osoba, która nie chce zostać zapisana do PPK, może złożyć deklarację o rezygnacji z dokonywania wpłat do PPK jeszcze przed zawarciem w jej imieniu i na jej rzecz umowy o prowadzenie PPK (wówczas nie stanie się uczestnikiem PPK).

Wszelkie dodatkowe informacje możecie znaleźć pod linkiem.

ZADANIE

Jak wynika ze wstępu, pracownik zapisany do PPK może w dowolnym momencie zrezygnować z uczestnictwa w programie i właśnie wokół tego problemu zorientowane jest Wasze zadanie.

Waszym głównym celem jest analiza danych oraz budowa klasyfikatora służącego do określenia, czy dany uczestnik zrezygnuje z programu w pewnym, nieznanym horyzoncie czasowym.[1] Predykcja powinna być połączona z szeroką analizą problemu, a finalnym produktem powinna być prezentacja, którą podzielicie pomiędzy część techniczną i część biznesową.

Ponadto, osobno punktowanym zadaniem jest stworzenie jak najciekawszej wizualizacji opartej o dane załączone do zadania. Dane są spseudonymizowane, co niestety trochę ogranicza możliwości, ale liczymy na Waszą kreatywność! Staraliśmy się zachować ich strukturę, lecz ich wartości nie mają one odzwierciedlenia w rzeczywistości.

Dane treningowe zamieściliśmy w pliku **PPK_Uczestnicy.csv**, dane do przeprowadzenia predykcji zamieściliśmy w pliku **PPK_Uczestnicy_TEST.csv**, który ma analogiczną strukturę, oczywiście po za zmienną objaśnianą. Plik **PPK_Pracodawcy.csv** jest uniwersalny dla obu plików, łączy się go po odpowiedniej zmiennej.

[1] Tego typu analizy znane są w literaturze pod nazwą *churn analysis*.

ROZWIĄZANIE

W ramach rozwiązania oczekujemy od Was:

- Pliku z predykcją – pliku CSV o strukturze analogicznej do **Submission_template.csv**, ważne by nazywał się **Nazwa_drużyny_submission.csv**.
- Kodu źródłowego – plików, które będą stanowiły rodzaj dokumentacji wykonanej przez Was pracy. Może to być czysty kod źródłowy, może to być raport w stylu Jupyter/Markdown Notebook, chcielibyśmy poznać wasz tok rozumowania oraz podejście do problemu. Pliki nazwijcie **Nazwa_drużyny_code_XX.***, gdzie XX to numer pliku.
- Wizualizacji – pojedynczej prezentacji danych, w dowolnym formacie. Może być to plik **.png*, może być to np. dokument tekstowy z linkiem do serwisu *plotly*. Ważne, aby nazywał się **Nazwa_drużyny_visualization.*** i był możliwy do otworzenia bez specjalistycznego oprogramowania.
- Prezentacji – krótkiej, około 15-minutowej prezentacji w dowolnej formie (zachęcamy do jej nagrania i udostępniania nam linku), podczas której wyjaśnicie podjęte przez siebie decyzje na etapie analizy i budowy modelu, a także przedstawicie rekomendację biznesową. Plik nazwijcie **Nazwa_drużyny_presentation.***
  (jeśli nagracie prezentację to umieśćcie link do np. YT w pliku Nazwa_drużyny_presentation.txt).

OCENA

Za rozwiązanie możecie otrzymać maksymalnie 100 punktów, które podzielone są wg schematu

Predykcja	max. 50 pkt
Kod źródłowy	max. 20 pkt
Prezentacja	max. 15 pkt
Wizualizacja	max. 15 pkt

Nie ma obowiązku przesyłania każdej części rozwiązania.

NARZĘDZIA

Dopuszczamy korzystanie z dowolnego oprogramowania, które jest dostępne za darmo, lub posiada co najmniej 30 dniową darmową wersję próbną. Preferujemy jednak do korzystanie z narzędzi open-source oraz języków programowania R/Python.

Wyjątkiem jest pakiet MS Office, z którego także możecie korzystać.

TERMIN I SPOSÓB PRZESYŁANIA ROZWIĄZAŃ

Wszystkie pliki niezbędne do oceny rozwiązania spakujcie w archiwum, nazwijcie plik **ANALITICS_Nazwa_drużyny.*** i prześlijcie na adres sknkonsultingu.cc@gmail.com do godz. 15:00 dnia 23.05.2021r.

DANE

Dane zawierają się w dwóch plikach płaskich, poniżej znajdziecie ich słownik.

Dane zostały częściowo oczyszczone oraz poddane pseudonimizacji, ale staraliśmy się odzwierciedlić i faktyczny kształt. Otrzymujecie dane, na których faktycznie pracujemy w PFR.

Zmienne z pliku **PPK_Pracodawcy.csv**:

Nazwa kolumny	Opis	Dane spseudonimozowane?	Typ
ID	Unikalny identyfikator pracodawcy	Tak	CHAR[8]
PPK_STAGE	Etap PPK w którym pracodawca dołączył do programu	Nie	INTEGER
LOGICAL_FACTOR_1	Zmienna binarna stosowana wewnątrz systemów PFR	Nie	BOOLEAN (0/1)
LOGICAL_FACTOR_2	Zmienna binarna stosowana wewnątrz systemów PFR	Nie	BOOLEAN (0/1)
REGION_CODE	Spseudonimizowany kod regionalny wg schematu: **AABBB**, gdzie: • **A** oznacza powiat • **B** oznacza gminę To nie jest kod pocztowy!	Tak	CHAR[5]
VOIVODESHIP	Identyfikator województwa	Tak	CHAR
PKD_CODE	Identyfikator PKD – struktura jest taka sama *faktycznego* identyfikatora, ale jego znaczenie nie pokrywa się z oficjalnymi kodami PKD.	Tak	CHAR[5]
COMPANY_SIZE	Wielkość firmy zdefiniowana jako: • A – do 10 pracowników • B – od 11 do 25 pracowników • C – od 26 do 50 pracowników • D – od 51 do 100 pracowników • E – od 101 do 250 pracowników • F – od 251 do 500 pracowników • G – od 501 do 1000 pracowników • H – powyżej 1000 pracowników	Tak	CHAR[1]
COMPANY_TYPE	Rodzaj spółki	Tak	CHAR[2]

Zmienne z pliku **PPK_Uczestnicy.csv**:

Nazwa kolumny	Opis	Dane spseudonimozowane?	Typ
IS_SUSPENDED	**Zmienna objaśniana** Wartość równa 1 oznacza, że rachunek został zawieszony	Nie	LOGICAL
MEMBER_ID	Identyfikator pracownika	Nie	CHAR[8]
EMPL_ID	Identyfikator pracodawcy	Tak	CHAR[8]
SEX	Płeć pracownika	Nie	CHAR[1]
NATIONALITY	Narodowość pracownika	Tak	CHAR[2]
WORK_START	Data rozpoczęcia zatrudnienia	Tak	DATE
WORK_STOP	Data zakończenia zatrudnienia	Tak	DATE
PPK_BANK	Identyfikator banku pośredniczącego	Tak	INTEGER
AGE	Wiek pracownika	Tak	DOUBLE
CREATED_AT	Data utworzenia rachunku	Tak	DATE
NUMERICAL_VALUE	Zmienna liczbowa stosowana wewnątrz systemów PFR	Tak	DOUBLE
UOZ_SIGN_DATE	Data podpisania umowy o zarządzanie PPK przez pracodawcę w imieniu pracownika	Tak	DATE
UOP_SIGN_DATE	Data podpisania umowy o prowadzenie PPK przez pracodawcę w imieniu pracownika	Tak	DATE
SIGN_DATE	Data podpisania umowy przez pracownika	Tak	DATE
RESIGN_DATE	Data zawieszenia umowy przez pracownika	Tak	DATE
HAS_AE	Czy pracownik otrzymuje wpłatę dodatkową od pracodawcy?	Nie	LOGICAL
HAS_AW	Czy pracownik wpłaca własną dodatkową wpłatę?	Nie	LOGICAL
HAS_IP	Czy pracownik otrzymał wpłatę powitalną?	Nie	LOGICAL

Zależności pomiędzy datami są dość skomplikowane, lecz należy mieć na uwadze, że faktyczny moment przystąpienia przez pracownika do PPK jest określany przez zmienną SIGN_DATE, a moment rezygnacji/zawieszenia rachunku PPK jest zdefiniowany przez zmienną RESIGN_DATE. Pozostałe daty są zależne od pracodawcy (UOZ_SIGN_DATE, UOP_SIGN_DATE). Dane dotyczące zatrudnienia pochodzą spoza rejestrów PFR i ich jakość nie zawsze jest użyteczna. Zmienna CREATED_AT oznacza moment utworzenia rachunku – warto mieć na uwadze, że pracownik może ponowić wpłacanie składek na rachunek, który został wcześniej zawieszony.