Porównanie klasycznych modeli ML i pretrenowanych modeli tabelarycznych w analizie nierówności płac w zbiorach danych HR

1. Wstęp

W ostatnich latach rośnie znaczenie analityki HR w wykrywaniu nierówności płac oraz w podejmowaniu decyzji menedżerskich opartych na danych. Tradycyjne metody statystyczne i regresyjne często zawodzą w przypadku danych tabelarycznych o relatywnie dużej liczbie zmiennych, wymagających jednoczesnego uwzględnienia cech demograficznych, stanowiskowych oraz informacji dotyczących ścieżki zawodowej i sytuacji pracowników. Dodatkowym wyzwaniem jest złożona struktura zależności pomiędzy tymi cechami, obejmująca interakcje nieliniowe oraz efekty pośrednie, które trudno uchwycić przy użyciu klasycznych modeli liniowych.

Jednocześnie kluczową, często pomijaną cechą zbiorów danych HR jest ich ograniczona liczebność. W organizacji zatrudniającej przykładowo 2 000 pracowników liczba dostępnych obserwacji dotyczących wynagrodzeń jest z reguły zbliżona do liczby zatrudnionych osób, o ile nie uwzględnia się danych historycznych lub panelowych. W praktyce oznacza to, że analityka HR operuje na zbiorach liczących od kilkuset do kilku tysięcy rekordów, co istotnie ogranicza możliwość stosowania złożonych modeli uczonych od zera i sprzyja nadmiernemu dopasowaniu (overfitting).

Dodatkowym, niezwykle istotnym aspektem jest wrażliwy charakter danych płacowych i personalnych. Dane HR należą do najbardziej wrażliwych kategorii informacji w organizacjach i podlegają rygorystycznym regulacjom prawnym oraz wewnętrznym politykom bezpieczeństwa. W konsekwencji wiele organizacji — zwłaszcza małych i średnich przedsiębiorstw — nie może lub nie chce korzystać z rozwiązań opartych na zewnętrznych interfejsach API, przetwarzaniu danych w chmurze czy modelach wymagających przesyłania danych poza infrastrukturę firmy. Powoduje to rosnące zapotrzebowanie na rozwiązania analityczne możliwe do uruchomienia w całości in-house, bez ryzyka wycieku danych i bez konieczności udostępniania wrażliwych informacji podmiotom trzecim.

W tym kontekście pojawia się potrzeba metod, które są jednocześnie odporne na małą liczebność danych, zdolne do modelowania złożonych, nieliniowych zależności, dobrze radzące sobie ze zmiennymi kategorycznymi i brakami danych, a przy tym możliwe do efektywnego wykorzystania w środowisku lokalnym, bez zaawansowanej infrastruktury obliczeniowej. Szczególnie obiecującym kierunkiem są pretrenowane modele tabelaryczne, które przenoszą wiedzę zdobytą na dużej liczbie zróżnicowanych zbiorów

danych na nowe, niewielkie problemy predykcyjne, umożliwiając skuteczne uczenie nawet przy bardzo ograniczonej liczbie obserwacji.

W niniejszym artykule porównujemy klasyczne modele uczenia maszynowego (Random Forest, XGBoost) z pretrenowanymi modelami tabelarycznymi - TabPFN, LimiX w kontekście predykcji nierówności płac. Oceniamy ich dokładność predykcji, interpretowalność oraz stabilność wyników w warunkach ograniczonej liczby obserwacji. Analogicznie do modeli BERT w przetwarzaniu języka naturalnego czy Vision Transformer (ViT) w analizie obrazu TabPFN i LimiX stanowią przykłady modeli pretrenowanych dedykowanych danym tabelarycznym, które umożliwiają demokratyzację zaawansowanej analityki ML - pozwalając organizacjom samodzielnie, lokalnie i w sposób bezpieczny analizować wrażliwe dane HR, bez konieczności korzystania z zewnętrznych usług czy rozwiązań typu „black-box".

2. Ramy pojęciowe

- Klasyczne ML vs pretrenowane modele tabelaryczne:
    - Klasyczne: Linear, Bayesian Regression, Random Forest, XGBoost, BART – trenowane od zera na konkretnym zbiorze danych.
    - Pretrenowane: TabPFN, LimiX – pretrenowane na wielu datasetach tabelarycznych, umożliwiające szybsze uczenie i lepszą generalizację, zwłaszcza przy ograniczonej liczbie obserwacji.
- Interpretowalność i znaczenie dla HR analytics: SHAP, feature importance.
- Zalety pretrenowanych modeli: odporność na małe dane, szybszy trening, lepsze wykrywanie złożonych zależności między cechami.

3. Research Questions

1. Jak różne klasy modeli tabelarycznych (klasyczne modele ML vs pretrenowane modele tabelaryczne) różnią się pod względem jakości predykcji nierówności płac w zbiorach HR o małej liczbie obserwacji (≈2 000 rekordów)?
2. W jakim stopniu pretrenowane modele tabelaryczne (np. TabPFN, LimiX) wykazują większą stabilność predykcji niż klasyczne modele ML w obecności zmiennych kategorycznych oraz brakujących danych? (współczynnik MAE, CV)
3. Jak zmniejszająca się liczba obserwacji wpływa na względną przewagę modeli pretrenowanych nad klasycznymi metodami ML w zadaniu regresji płac? (np. 500 / 1 000 / 2 000)

4. Hipotezy

- H1: Pretrenowane modele tabelaryczne (TabPFN, LimiX) osiągają istotnie lepszą jakość predykcji oraz lepszą kalibrację predykcji niż klasyczne modele ML w zadaniu regresji płac dla małych zbiorów danych.
- H2: Pretrenowane modele tabelaryczne wykazują istotnie mniejszą wariancję wyników predykcyjnych w walidacji krzyżowej niż klasyczne modele ML w warunkach bardzo małej liczby obserwacji (N < 500).
- H3: W typowych zbiorach danych HR o wielkości około 2 000 obserwacji, modele pretrenowane na zadaniach small-data (np. TabPFN) nie tracą przewagi predykcyjnej względem innych nowoczesnych modeli tabelarycznych.

5. Dane (Dataset)

Zbiór danych obejmuje publiczne dane HR dostępne na platformach typu Kaggle oraz syntetyczne dane generowane przy użyciu SDV w celu uzupełnienia braków i zwiększenia różnorodności. Łączna liczba rekordów w eksperymencie nie przekracza 2 000 wierszy, co odpowiada typowym datasetom średniej wielkości organizacji.

Minimalny zestaw cech dla skorygowanej luki płacowej:

Kolumna	Typ | Przykład
gender	Male / Female
age	35 (liczba lat)
education_level	1–4 (Bachelor’s, Master’s, PHD, inny)
experience_years	10 (liczba lat)
job_level	1– 4 (Junior / Mid / Senior / Manager)
child	2 (liczba dzieci)
distance_from_home	0 / 1 (poniżej/powyżej 15 km)
income	10 000 (miesięczny w PLN)

Dane przygotowano tak, aby umożliwić porównanie modeli pod kątem predykcji nieskorygowanej i skorygowanej luki płacowej. Braki danych uzupełniono imputacją, zmienne kategoryczne zakodowano odpowiednio do wymagań modeli.

6. Metody

6.1. Do benchmarku wybrano siedem modeli, pogrupowanych według typu:

- Klasyczne ML:
    - Linear Regression
    - Random Forest
    - XGBoost
- Modele bayesowskie / probabilistyczne [posterior P(y|X)]:
    - Bayesian Regression
    - BART (Bayesian Additive Regression Trees)
- Pretrenowane modele tabelaryczne (Transformery):
    - TabPFN
    - LimiX

Ocena jakości predykcji obejmowała metryki: CV, MAE, a interpretowalność analizowano przy użyciu SHAP i feature importance. Eksperymenty przeprowadzono przy różnych konfiguracjach zmiennych i brakujących danych, a stabilność wyników badano poprzez kroswalidację (k=5).

6.2. TabPFN vs LimiX

TabPFN i LimiX reprezentują odmienne paradygmaty pretrenowanego modelowania danych tabelarycznych, mimo że oba modele są projektowane z myślą o zadaniach regresji i klasyfikacji w reżimie ograniczonej liczby obserwacji. TabPFN można interpretować jako model realizujący implicit Bayesian inference, który uczy się rozkładu funkcji predykcyjnych na podstawie dużej liczby syntetycznych zadań tabularnych. W konsekwencji, w fazie inferencji TabPFN nie wymaga dodatkowego trenowania, lecz bezpośrednio wykorzystuje wyuczony priorytet do estymacji predykcji oraz niepewności, co czyni go szczególnie skutecznym i stabilnym w scenariuszach small-data. Z kolei LimiX opiera się na generatywnym podejściu wykorzystującym latent diffusion models, których celem jest modelowanie struktury i rozkładu danych tabelarycznych w przestrzeni latentnej. Takie podejście umożliwia LimiX skuteczniejsze radzenie sobie z brakującymi wartościami oraz złożonymi zmiennymi kategorycznymi, kosztem większej

złożoności obliczeniowej i większej wrażliwości na konfigurację inferencji. W praktyce różnice te przekładają się na przewagę TabPFN w zadaniach predykcyjnych przy bardzo małych, relatywnie czystych zbiorach danych, natomiast LimiX wykazuje większą odporność w warunkach heterogenicznych danych HR, gdzie występują liczne zmienne kategoryczne oraz braki danych.

7. Eksperymenty

Eksperymenty obejmowały benchmark wszystkich siedmiu modeli na datasetach ~2 000 (i mniej) wierszy. Analizowano wpływ liczby zmiennych i brakujących danych na dokładność predykcji, porównywano interpretowalność modeli poprzez wizualizację SHAP oraz feature importance. Sprawdzono stabilność wyników przy ograniczonym rozmiarze danych i dla minimalnego zestawu cech niezbędnych do obliczenia skorygowanej luki płacowej.

8. Wyniki

8.1. Skuteczność predykcyjna modeli

W przeprowadzonych eksperymentach pretrenowane modele tabelaryczne wykazały przewagę nad klasycznymi metodami uczenia maszynowego w warunkach ograniczonej liczby obserwacji.

W szczególności TabPFN osiągał najlepszą jakość predykcji w zbiorach danych o wielkości około 2 000 obserwacji, uzyskując najniższe wartości błędu MAE oraz RMSE w porównaniu z pozostałymi modelami.

Model LimiX osiągał wyniki porównywalne do TabPFN, jednak jego przewaga była bardziej widoczna w scenariuszach z brakującymi danymi oraz przy większej liczbie zmiennych kategorycznych. W takich warunkach LimiX charakteryzował się mniejszą degradacją jakości predykcji.

Klasyczne modele uczenia maszynowego (np. regresja liniowa, XGBoost) wykazywały większą wrażliwość na zmienne kategoryczne oraz wyraźnie większą wariancję wyników w walidacji krzyżowej, co wskazuje na tendencję do overfittingu w warunkach ograniczonej liczby obserwacji.

8.2. Stabilność predykcji

Analiza wariancji wyników w powtarzanej walidacji krzyżowej wykazała, że pretrenowane modele tabelaryczne charakteryzują się istotnie mniejszą zmiennością wyników w porównaniu do klasycznych metod ML.

Efekt ten był szczególnie widoczny w eksperymentach z losowym podpróbkowaniem danych treningowych oraz przy wprowadzaniu brakujących wartości.

8.3. Analiza istotności zmiennych

Analiza istotności cech (feature importance) wykazała, że do najważniejszych zmiennych wpływających na predykcję nierówności płac należą:

- Job level
- Child
- Gender
- Education level

Zmienne te pojawiały się konsekwentnie jako istotne w większości analizowanych modeli, co wskazuje na ich kluczową rolę w modelowaniu różnic płacowych w badanym zbiorze danych.

9. Dyskusja

Pretrenowane modele tabelaryczne mogą znacząco poprawić predykcję i interpretowalność w analizie nierówności płac, nawet przy bardzo małych datasetach (~2 000 rekordów). TabPFN jest szczególnie efektywny przy ograniczonych danych, natomiast LumiX lepiej radzi sobie przy większej liczbie braków w danych. Połączenie klasycznych metod ML z pretrenowanymi modelami tabelarycznymi może być korzystne dla praktyków HR, którzy potrzebują zarówno dokładnych, jak i interpretowalnych predykcji.

10. Wnioski

Artykuł prezentuje porównanie klasycznych modeli ML i pretrenowanych modeli tabelarycznych w kontekście analizy skorygowanej i nieskorygowanej luki płacowej przy małych datasetach. Wyniki wskazują na potencjał TabPFN i LimiX w predykcji pay gap oraz interpretowalności modeli dla HR. Wyniki te stanowią fundament do dalszych badań, w tym integracji z systemami wspomagania decyzji w HR.