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

GitHub: https://github.com/mikolajkuna/pay-gap-ml

1. Wstęp

Analiza nierówności płac w danych HR staje się coraz bardziej istotna w procesach decyzyjnych opartych na danych. Tradycyjne metody statystyczne i regresyjne często zawodzą przy danych tabelarycznych o liczbie zmiennych obejmującej zarówno cechy demograficzne, stanowiskowe, jak i ścieżkę zawodową pracowników. Dodatkowo, ograniczona liczba obserwacji w typowych organizacjach (kilkaset–kilka tysięcy rekordów) utrudnia stosowanie modeli uczonych od zera i zwiększa ryzyko overfittingu.

Pretrenowane modele tabelaryczne, takie jak TabPFN i AutoGluon, stanowią alternatywę umożliwiającą skuteczne uczenie w warunkach small-data. TabPFN, będący Bayesowskim modelem transformera, wykorzystuje wiedzę zdobytą na wielu datasetach tabelarycznych, pozwalając na precyzyjne estymacje predykcji oraz niepewności. AutoGluon Tabular to framework AutoML, który automatycznie dobiera i ensemble'uje wiele modeli, umożliwiając wykrywanie złożonych zależności i obsługę brakujących danych. Oba podejścia działają w pełni lokalnie, co jest kluczowe przy wrażliwych danych HR.

Celem artykułu jest porównanie klasycznych modeli ML (Linear Regression, Random Forest, XGBoost, BART) z pretrenowanymi modelami tabelarycznymi (TabPFN, AutoGluon) pod kątem predykcji nierówności płac, stabilności wyników oraz interpretowalności przy ograniczonej liczbie obserwacji.

2. Metody

Benchmark obejmował trzy grupy modeli:

1. Klasyczne ML: Linear Regression, Random Forest, XGBoost.
2. Modele bayesowskie/probabilistyczne: Bayesian Regression, BART (Bayesian Additive Regression Trees).
3. Pretrenowane modele tabelaryczne:
    - TabPFN (v2.5): Bayesowski transformer, szybkie uczenie, estymacja niepewności, minimalna konfiguracja hiperparametrów.
    - AutoGluon (v1.5): AutoML Tabular, automatyczne testowanie wielu modeli, tuning hiperparametrów, ensembling/staking.

Ocena jakości predykcji:

- MAE, CV, interpretowalność przez SHAP i feature importance.
- Eksperymenty przeprowadzono na zbiorach ~2 000 rekordów i mniejszych, w tym scenariuszach z brakującymi danymi i zmiennymi kategorycznymi.

Analiza kontrfaktyczna (pay gap):

- Predykcja skorygowanej luki płacowej dla scenariuszy gender.
- Wpływ zmiennych na predykcję badano za pomocą SHAP.

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.

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, AutoGluon) 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, AutoGluon) 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.

Kolumna	**Typ**	**Przykład**
gender	Male / Female	Male
age	liczba lat	35
education_level	1–4 (Bachelor/Master/PhD)	2 (Master)
experience_years	liczba lat	10
job_level	1–4 (Junior/Mid/Senior/Manager)	3 (Senior)
child	liczba dzieci	2
distance_from_home	0/1 (<15 km / ≥15 km)	0
income	miesięczny w PLN	10 000

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. Wyniki

1. Skuteczność predykcyjna:
    - TabPFN osiągał najniższy MAE i CV przy ~2 000 obserwacjach.
    - AutoGluon miał porównywalną dokładność, ale lepiej radził sobie z brakami danych i większą liczbą zmiennych kategorycznych.
    - Klasyczne modele ML wykazywały większą wariancję predykcji, szczególnie przy ograniczonej liczbie danych.
2. Stabilność predykcji:
    - Pretrenowane modele tabelaryczne cechowały się mniejszą zmiennością wyników w walidacji krzyżowej (k=5), co potwierdza odporność na small-data.
3. Interpretowalność i istotność zmiennych:
    - Najważniejsze cechy w predykcji płac: Job level, Child, Gender, Education level.
    - SHAP wskazał, że wpływ zmiennych demograficznych i stanowiskowych jest spójny w modelach pretrenowanych i klasycznych.
4. Analiza kontrfaktyczna (gender pay gap):
    - TabPFN: niższy MAE, mniejszy CV, precyzyjne predykcje jednostkowe.
    - AutoGluon: większa wrażliwość na ukryte wzorce, wyższa kontrfaktyczna luka płacowa, ale przy wyższym błędzie przewidywań.

7. Analiza istotności zmiennych

Generalizując 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.

8. Wnioski

Analiza porównawcza dwóch podejść do predykcji dochodów w danych tabelarycznych ujawnia istotne różnice w charakterze ich działania i wynikach. Model TabPFN, oparty na pojedynczym transformerze, charakteryzuje się wysoką precyzją prognoz jednostkowych oraz prostotą użycia, wymagając minimalnej konfiguracji i tuningu hiperparametrów. Z kolei AutoGluon Tabular, jako framework AutoML dla danych tabelarycznych, automatycznie testuje różne modele, może wykonywać tuning hiperparametrów oraz opcjonalnie łączyć wyniki wielu modeli w ensembling lub stacking w celu poprawy jakości predykcji. W kontekście oceny kontrfaktycznej luki płacowej AutoGluon wykazuje większą wrażliwość na złożone wzorce w danych, natomiast TabPFN, mimo wysokiej dokładności predykcji jednostkowych, jest mniej podatny na wychwytywanie subtelnych efektów strukturalnych. W rezultacie wybór między tymi

podejściami zależy od celu analizy: TabPFN sprawdza się w sytuacjach wymagających szybkiej i precyzyjnej prognozy, natomiast AutoGluon jest korzystniejszy przy badaniu ukrytych zależności i kontrfaktycznych scenariuszy.

- Pretrenowane modele tabelaryczne (TabPFN i AutoGluon) przewyższają klasyczne ML w zadaniach predykcji płac przy ograniczonej liczbie obserwacji.
- TabPFN jest idealny do szybkiej, precyzyjnej predykcji jednostkowej i oceny niepewności.
- AutoGluon sprawdza się lepiej przy analizie złożonych interakcji i braków danych, dzięki automatycznemu ensemblingowi wielu modeli.

Modele te umożliwiają bezpieczną, lokalną analizę danych HR, co jest kluczowe przy wrażliwych informacjach płacowych. Wyniki stanowią fundament do integracji pretrenowanych modeli tabelarycznych w systemach wspomagania decyzji HR i dalszych badań nad interpretowalnością oraz kontrfaktyczną analizą nierówności płac.