Lab3
Analizator wyników

Jakub Wójcik

1.Eksploracja i wstępna analiza danych

-Plik zawiera 4739 wierszy i 15 kolumn. Oto główne cechy danych:

a)Kolumny typu liczbowego:

- rownames (numery wierszy),
- score (najprawdopodobniej wynik testowy) - atrybut decyzyjny,
- unemp (najprawdopodobniej wskaźnik bezrobocia),
- wage (wynagrodzenie),
- distance (odległość najprawdopodobniej od szkoły),
- tuition (czesne),
- education (lata edukacji).

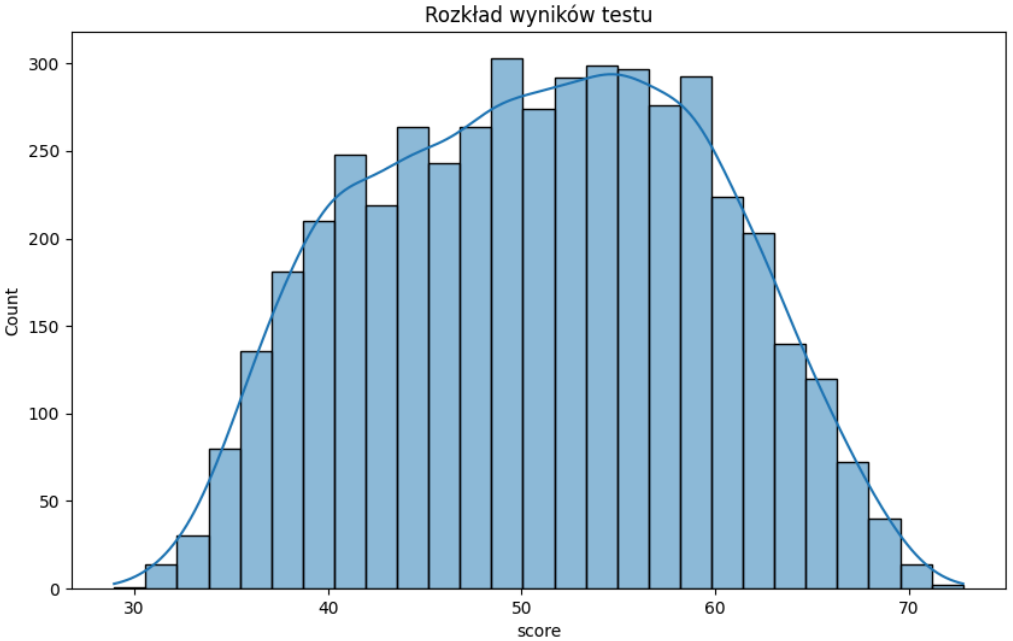
b)Kolumny typu kategorycznego:

- gender (płeć),
- ethnicity (grupa etniczna),
- fcollege, mcollege (czy ojciec/matka ukończyli studia),
- home, urban (miejsce zamieszkania: czy otoczenie miejskie),
- income (dochód),
- region (region zamieszkania).

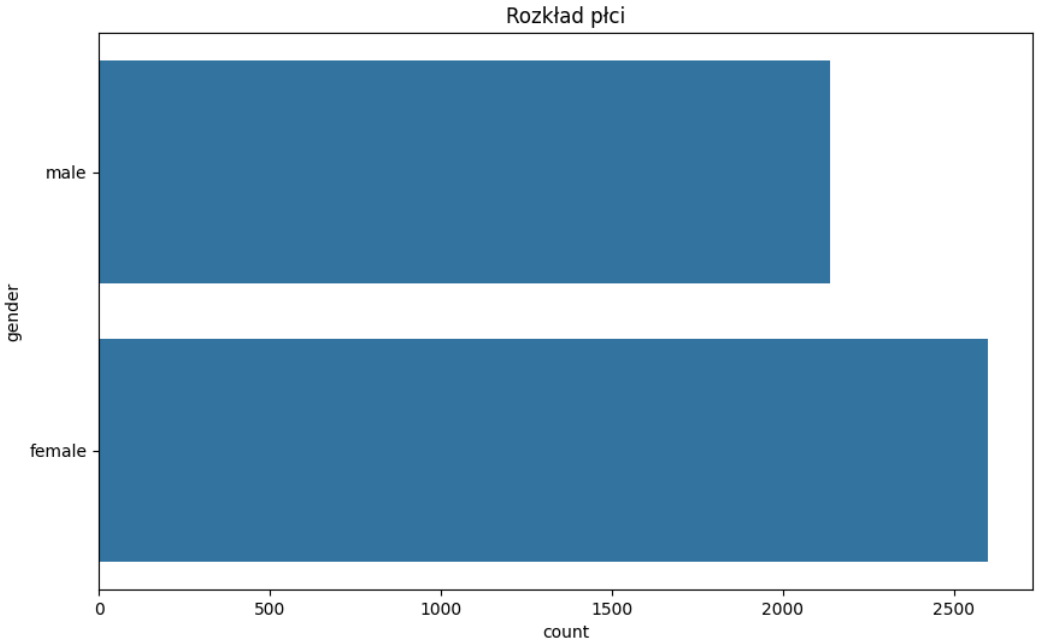
Dane nie zawierały brakujących wartości oraz były już dość dobrze przygotowane. Nie pozwoliło mi to jednak pozostawić ich bez obróbki co zrobię już w kolejnym etapie.

-Poniżej przygotowałem trzy najciekawsze moim zdaniem wykresy obrazujące następująco:

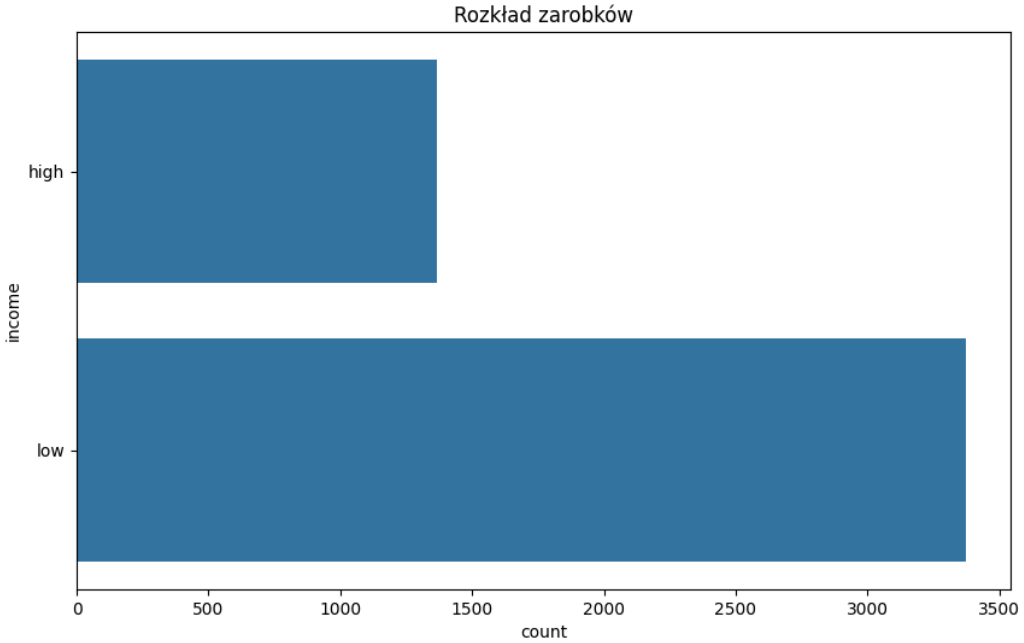
a)Rozkład wyników testu



b)Rozkład płci (z dość wyrównanymi wynikami)



c)Rozkład zarobków (ze znaczną dominacją niskich zarobków)



2.Inżynieria cech i przygotowanie danych

Jako że dane nie zawierały brakujących wartości, nie musiałem ich uzupełniać czy też usuwać wierszy z brakującymi wartościami. Przeprowadziłem za to normalizację danych numerycznych od 0 do 1 z wyłączeniem kolumn: *score*, ponieważ to właśnie ją będziemy przewidywać oraz *rownames*, ponieważ mimo wartości mocno odznaczających się na tle innych, nie będziemy brali jej pod uwagę przy przewidywaniu. Zmienne kategoryczne zaś zakodowałem. Po przeprowadzeniu odpowiedniej inżynierii cech przystąpiłem do podziału zbioru na zbiór treningowy i testowy gdzie postawiłem na sprawdzony podział 80/20 zastanawiając się również nad podziałem 75/25, lecz dawał on niższe wyniki podczas przewidywania.

3.Wybór i trenowanie modelu

Ze względu na ciągły charakter zmiennej *score*, model regresji liniowej był naturalnym wyborem. Zdecydowałem się na ten model, ponieważ oferuje kilka korzyści:
a)Prostota i interpretowalność: Model regresji liniowej jest łatwy do interpretacji, co pozwala lepiej zrozumieć wpływ poszczególnych zmiennych na score.
b)Efektywność obliczeniowa: Jest szybki w treningu, co jest korzystne szczególnie dla mniejszych i średnich zbiorów danych.
c)Przystosowanie do zmiennych ciągłych: Regresja liniowa dobrze sobie radzi z przewidywaniem wartości ciągłych, co sprawia, że jest odpowiednia do przewidywania *score*.

4.Ocena i optymalizacja modelu

a)Wyniki modelu regresji liniowej są następujące:

```
Szczegółowy raport oceny modelu:
Mean Squared Error (MSE): 49.15
Mean Absolute Error (MAE): 5.76
Mean Absolute Percentage Error (MAPE): 11.94%
R-squared (R2): 0.35
```

- **Mean Squared Error (MSE):** 49.15 – wskazuje na średni kwadratowy błąd przewidywań, gdzie mniejsza wartość oznacza lepszą wydajność.
- **Mean Absolute Error (MAE):** 5.76 – przeciętna bezwzględna różnica między przewidywanymi i rzeczywistymi wynikami score, co wskazuje na typowy błąd modelu w przewidywaniach.
- **Mean Absolute Percentage Error (MAPE):** 11.94% – średni błąd procentowy w przewidywaniu *score*, pozwala ocenić odchylenie przewidywań względem rzeczywistych wyników.
- **R-squared (R2):** 0.35 – współczynnik determinacji,czyli statystyczna miara jakości dopasowania modelu do danych.

b)Podsumowanie:
Model osiągnął przyzwoity wynik, lecz postanowiłem spróbować jeszcze go podnieść sprawdzając inne możliwości czy też modele. Pomimo próby zastosowania GradientBoostingRegressor, XGBoost oraz przeszukiwania Grid Search zamiast zwykłej regresji liniowej nie udało mi się poprawić znacząco wyników przewidywania. Wyniki uzyskiwane poprzez zastosowanie tych o wiele bardziej skomplikowanych modeli były lepsze o zaledwie około 1/1.5%.