S25278 Marcin Milewski

Projekt Przewidywania Wyniku score w Danych o Dystansie do Uczelni

Opis Projektu

Celem tego projektu jest zbudowanie modelu predykcyjnego, który na podstawie zestawu danych CollegeDistance.csv będzie przewidywał zmienną score. Zmienna score odnosi się do określonego wyniku powiązanego z dystansem do uczelni dla każdego ucznia. Dane obejmują informacje demograficzne, społeczno-ekonomiczne i geograficzne, które służą jako cechy wejściowe dla modelu.

Projekt jest podzielony na cztery etapy, obejmujące eksplorację danych, inżynierię cech, wybór i trening modelu, ocenę oraz optymalizację modelu.

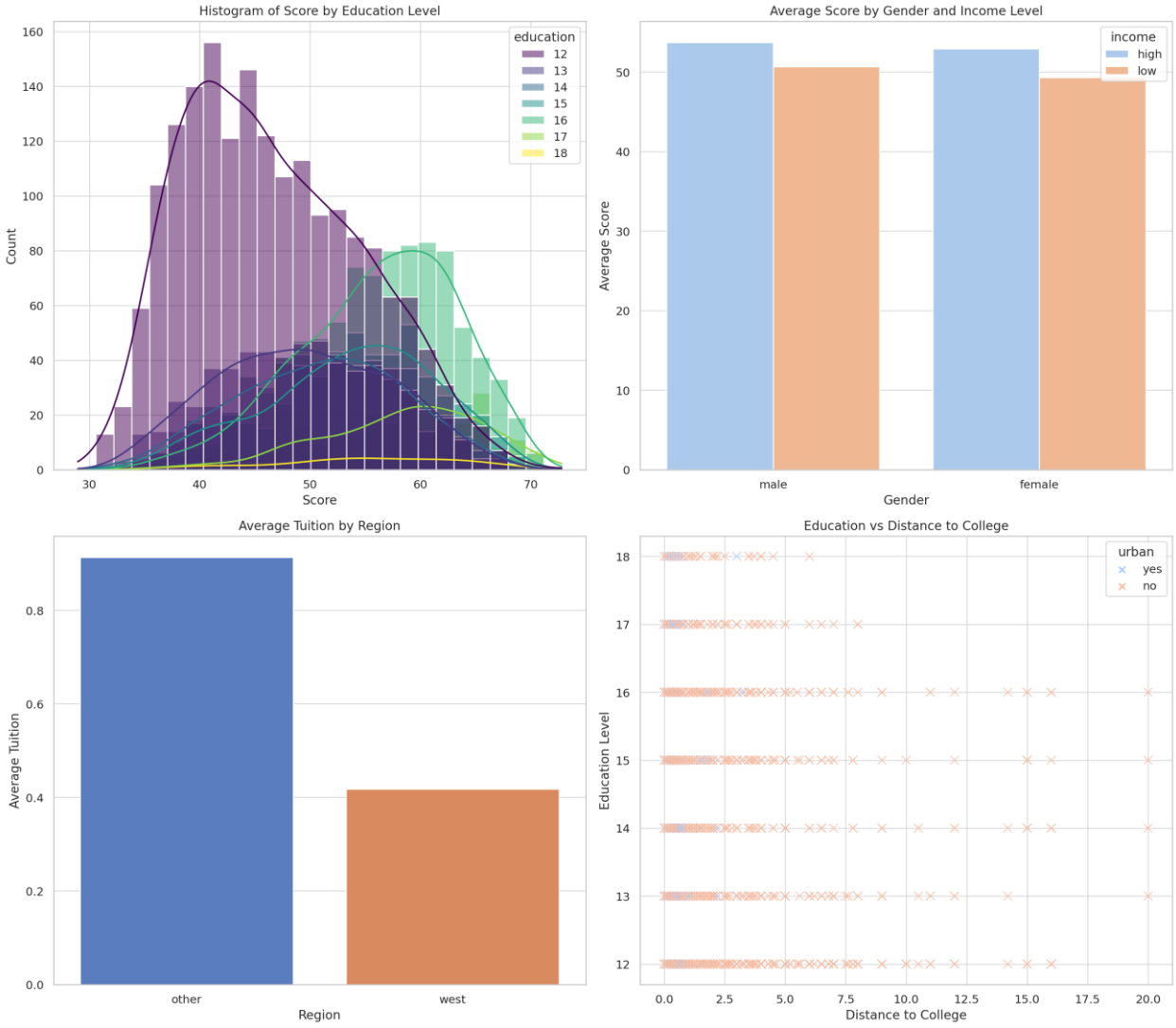
Eksploracja i wstępna analiza danych

Plik zawiera 4,739 wierszy i 15 kolumn, a dane opisują różne cechy związane z edukacją i środowiskiem życia osób. Oto kluczowe kolumny:

- **Demografia i środowisko**: gender, ethnicity, home, urban, region.
- **Edukacja**: score (wynik), education (liczba lat edukacji).
- **Sytuacja rodzinna i ekonomiczna**: fcollege (czy ojciec ukończył college), mcollege (czy matka ukończyła college), income (poziom dochodu), unemp (stopa bezrobocia), wage (wynagrodzenie).
- **Czynniki kosztowe**: distance (odległość do college'u), tuition (czesne).

Wykresy:

- Histogram wyników (score) w zależności od poziomu wykształcenia (education).
- Średnie wyniki (score) w podziale na płeć (gender) i dochód (income).
- Średnie czesne (tuition) w różnych regionach (region).
- Wykres zależności między odległością do college'u (distance) a latami edukacji (education).



Inżynieria Cech i Przygotowanie Danych

- **Kodowanie zmiennych kategorycznych**: Wykorzystano OneHotEncoder do przekształcenia kolumn kategorycznych (gender, ethnicity, region, fcollege, mcollege, home, urban, income) na reprezentacje numeryczne.
- **Standaryzacja cech numerycznych**: Skorzystano z StandardScaler, aby znormalizować wszystkie zmienne numeryczne. Dzięki temu każda z nich ma średnią równą 0 i odchylenie standardowe równe 1, co jest szczególnie ważne dla poprawnego działania modeli takich jak sieci neuronowe.
- **Podział danych**: Dane zostały podzielone na zbiór treningowy (80%) i testowy (20%) przy użyciu train_test_split, co umożliwia niezależną walidację modelu na nieznanych danych.

Wybór i Trening Modelu

Modele zastosowane:

- Regresja liniowa: Prosty model predykcyjny bazujący na założeniu liniowej zależności między zmiennymi. Może być interpretowany i łatwy do obliczenia, jednak może nie najlepiej pasować do bardziej złożonych wzorców.
- Lasy losowe: Model RandomForestRegressor, który tworzy zbiór drzew decyzyjnych, aby osiągnąć lepsze dopasowanie i radzić sobie z nieregularnymi danymi. Jest bardziej odporny na przetrenowanie i może uchwycić nieliniowe zależności.
- Sieć neuronowa (MLP): Model MLPRegressor, który pozwala uchwycić skomplikowane zależności w danych. Model został przetrenowany z maksymalnie 500 iteracjami, jednak nie osiągnął pełnej konwergencji, co sygnalizuje, że dalsze zwiększenie iteracji lub dostrojenie parametrów mogłoby poprawić wyniki.

Wyniki modeli:

- Regresja liniowa: MSE = 49.13, R^2 = 0.35
- Lasy losowe: MSE = 52.02, R^2 = 0.31
- Sieć neuronowa: MSE = 50.62, R^2 = 0.33

Ocena i Optymalizacja Modelu

Ocena:

- Wyniki zostały ocenione przy użyciu dwóch metryk: MSE (średni błąd kwadratowy) i R^2 (współczynnik determinacji). MSE pokazuje, jak bardzo przewidywania różnią się od rzeczywistych wartości, natomiast R^2 określa, jaka część wariancji zmiennej score jest wyjaśniona przez model.
- Najlepiej wypadł model regresji liniowej, uzyskując MSE = 49.13 i R^2 = 0.35, co sugeruje, że ten model wyjaśnia 35% wariancji zmiennej score.

Optymalizacja:

- Sieć neuronowa osiągnęła maksymalną liczbę iteracji (500) bez pełnej konwergencji, co sugeruje możliwość poprawy wyników poprzez zwiększenie liczby iteracji lub dostrojenie hiperparametrów.
- Modele takie jak lasy losowe mogłyby również skorzystać z dostrajania parametrów, np. liczby drzew lub głębokości drzew, aby poprawić dopasowanie.