Temat: Analiza danych i modelowanie regresyjne z obsługą wartości odstających.

1. Wstęp i opis zbioru danych

Celem projektu było zbudowanie modelu regresyjnego przewidującego dochód (Income) na podstawie cech demograficznych i finansowych (np. wiek, wykształcenie, stan cywilny). Analiza miała na celu zbadanie wpływu wartości odstających (outliers) na jakość predykcji. Do projektu wykorzystano zbiór danych data.csv. Zmienna celowa Income jest zmienną ciągłą, co determinuje użycie algorytmów regresji.

2. Eksploracja danych i analiza Outlierów

Wstępna analiza (histogramy, wykresy pudełkowe) wykazała występowanie w zmiennej Income wartości odstających – relatywnie niewielkiej liczby osób o zarobkach znacznie przekraczających średnią.
Do identyfikacji tych anomalii zastosowano metodę IQR (Rozstęp ćwiartkowy). Wyznaczono granice odcięcia, a obserwacje leżące poza nimi oznaczono jako outliery.

- Wpływ outlierów na dane: Obecność skrajnie wysokich dochodów powoduje silną asymetrię rozkładu (prawostronną) i sztucznie zawyża średnią, czyniąc ją mało reprezentatywną dla ogółu populacji.

3. Przygotowanie danych

Przygotowano dwa warianty zbioru danych do eksperymentu:

1. Zbiór A (Surowy): Zawierający wszystkie obserwacje, włącznie z wartościami odstającymi.
2. Zbiór B (Oczyszczony): Zbiór, z którego usunięto rekordy zidentyfikowane jako outliery.

Dla obu zbiorów wykonano:

- Uzupełnianie braków: Mediana dla zmiennych liczbowych, moda dla kategorycznych.
- Kodowanie: Zastosowano One-Hot Encoding dla zmiennych kategorycznych (np. Education, Marital_Status).
- Standaryzację: Przeskalowano cechy numeryczne za pomocą StandardScaler.
- Podział: Podział na zbiór treningowy (80%) i testowy (20%).

4. Budowa modeli regresyjnych

Zastosowano algorytm Regresji Liniowej dla obu wariantów danych. Wybór tego modelu podyktowany był chęcią wyraźnego zaobserwowania wpływu outlierów, na które metoda najmniejszych kwadratów (stosowana w regresji liniowej) jest szczególnie wrażliwa.

5. Wyniki i Porównanie Modeli

Modele oceniono na zbiorze testowym przy użyciu metryk RMSE (Błąd średniokwadratowy) oraz R2 (Współczynnik determinacji).

6. Wnioski

1. Analiza błędów (RMSE): Usunięcie wartości odstających przyniosło znaczącą poprawę precyzji modelu. Błąd RMSE spadł z ~910 na ~729 jednostek walutowych. Oznacza to, że model uczony na "czystych" danych myli się średnio o znacznie mniejszą kwotę przy przewidywaniu pensji typowej osoby.
2. Interpretacja R2: Oba modele osiągnęły ekstremalnie wysoki wynik R2 (~0.99), co świadczy o tym, że w tym syntetycznym zbiorze danych cechy (np. Wykształcenie, Wiek) prawie idealnie tłumaczą zarobki. Minimalnie niższy R2 w modelu drugim wynika z faktu, że usunęliśmy część wariancji (skrajne przypadki), więc model opisuje węższy zakres rzeczywistości, ale robi to precyzyjniej.
3. Wpływ Outlierów: Wartości odstające działały jak "magnes", odciągając linię regresji od głównego trendu danych. Próba dopasowania się modelu do milionerów powodowała większe błędy dla osób o przeciętnych dochodach. Oczyszczenie danych metodą IQR pozwoliło stworzyć model bardziej stabilny i wiarygodny dla większości populacji.