Dokumentacja do LAB-3

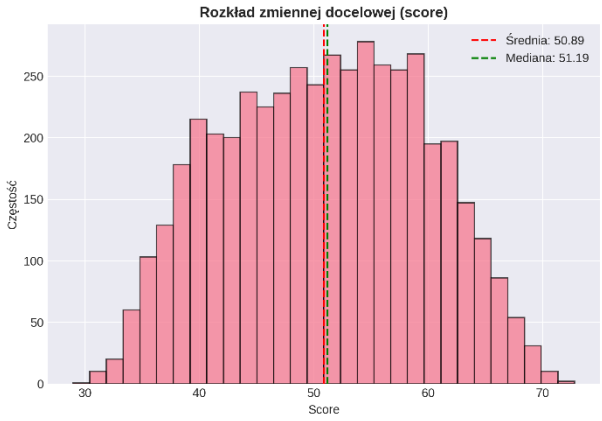
1. Eksploracja i wstępna analiza danych

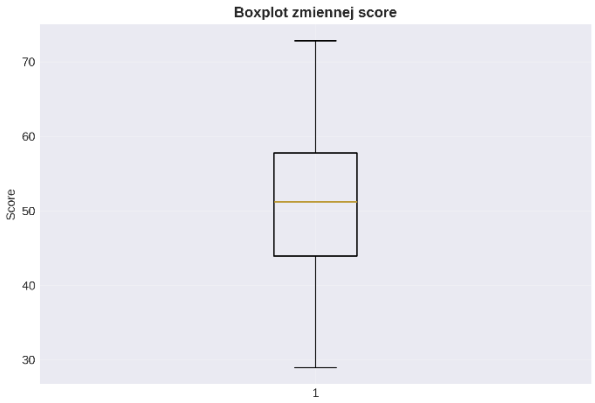
1.1 Czyszczenie danych

- Kolumna „rownames” została usunięta, ponieważ nie zawiera informacji predykcyjnej.
- Dane nie posiadały brakujących wartości
- Kolumny logiczne: (fcollege, mcollege, home, urban) przekształcono w wartości binarne (1 – „yes”, 0 – „no”)

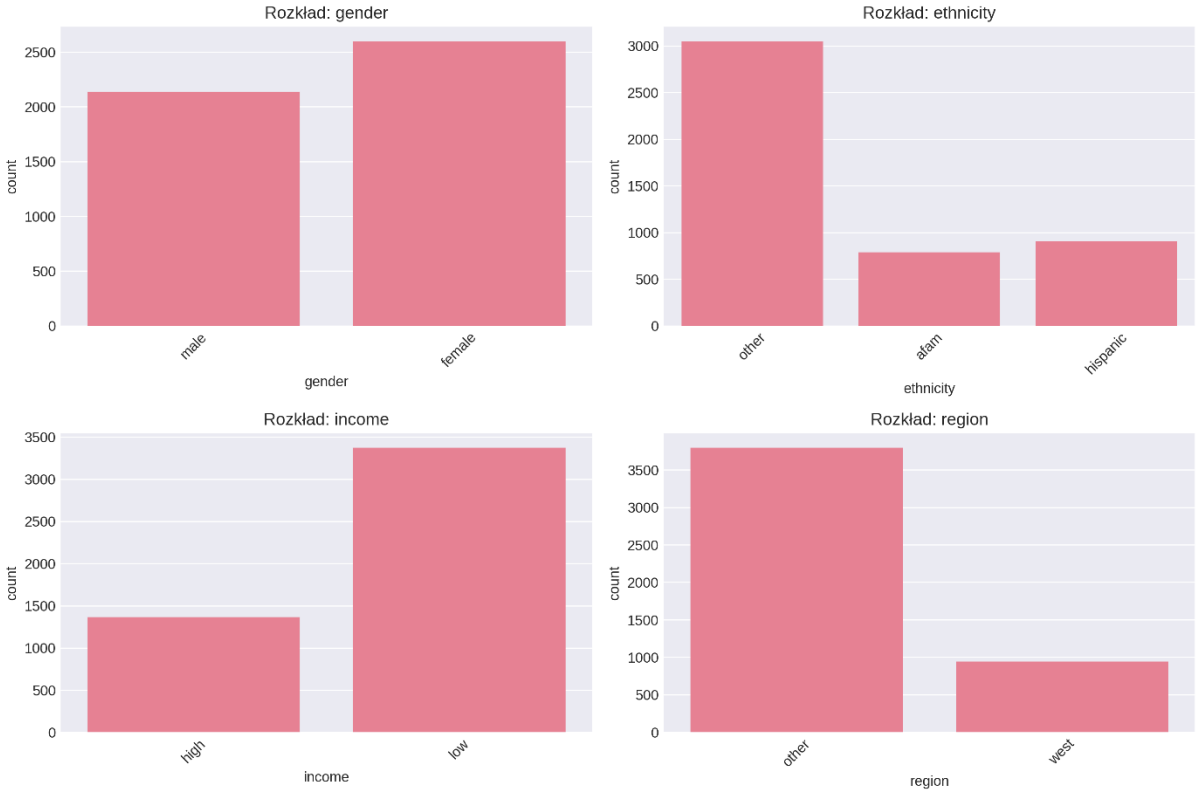
1.2 Analiza statystyczna i wizualizacja

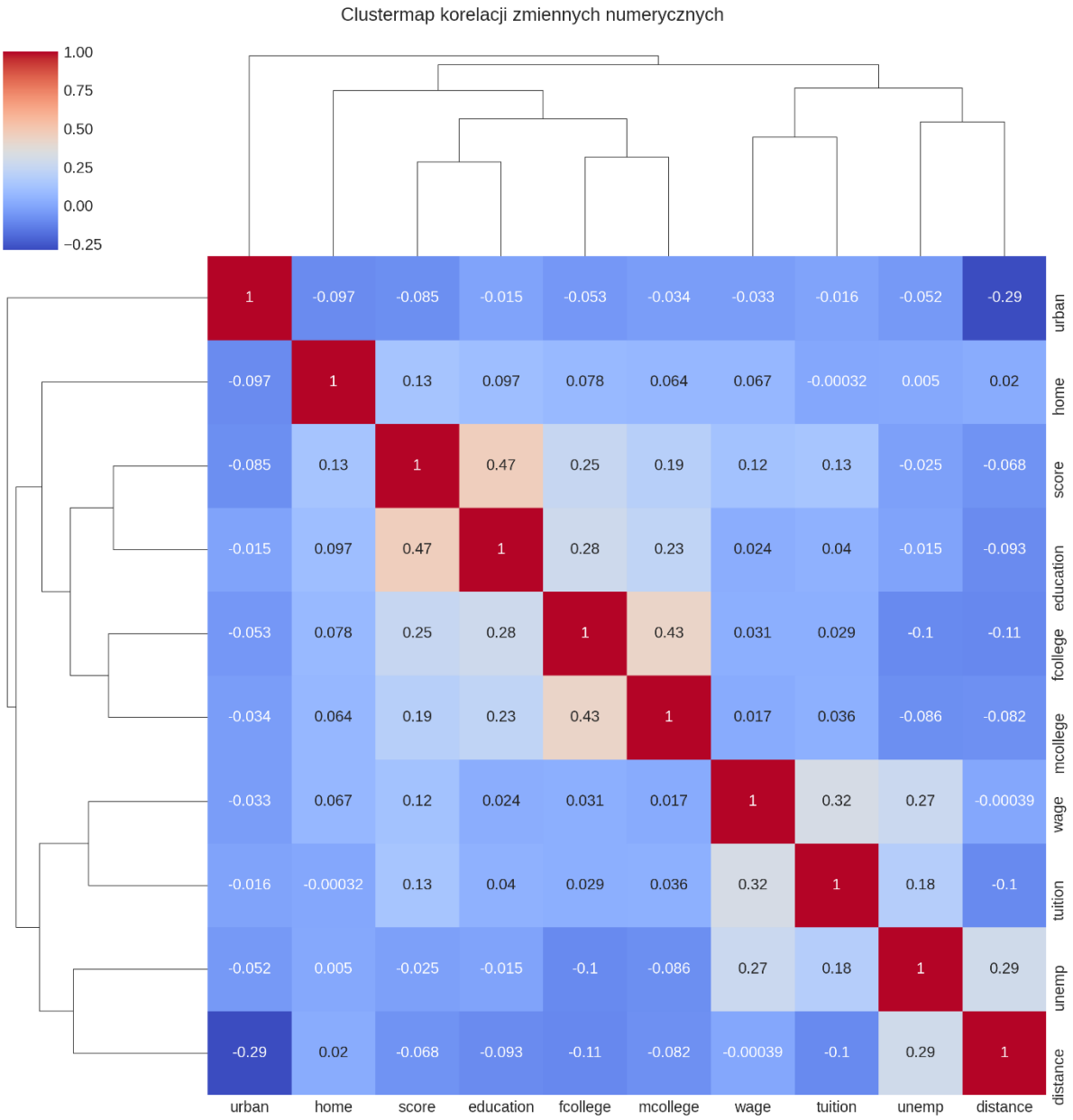
- Wyświetlono rozkład zmiennej docelowej score (histogram + boxplot).
- Wyznaczono macierz korelacji zmiennych numerycznych i wybrano top 5 cech najbardziej skorelowanych ze score.
- Dla tych cech wygenerowano pairplot z regresją liniową.
- Wykresy rozkładów zmiennych kategorycznych (countplot) i clustermap korelacji.

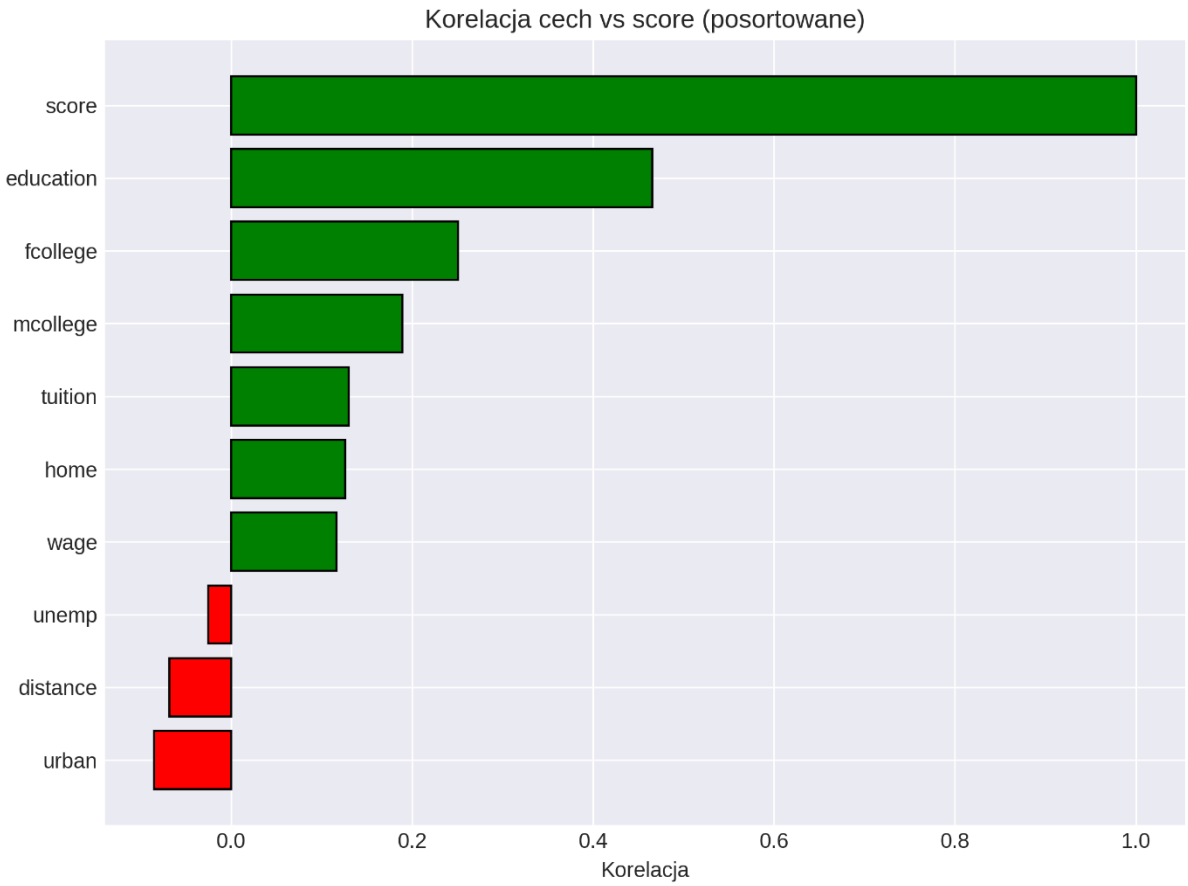


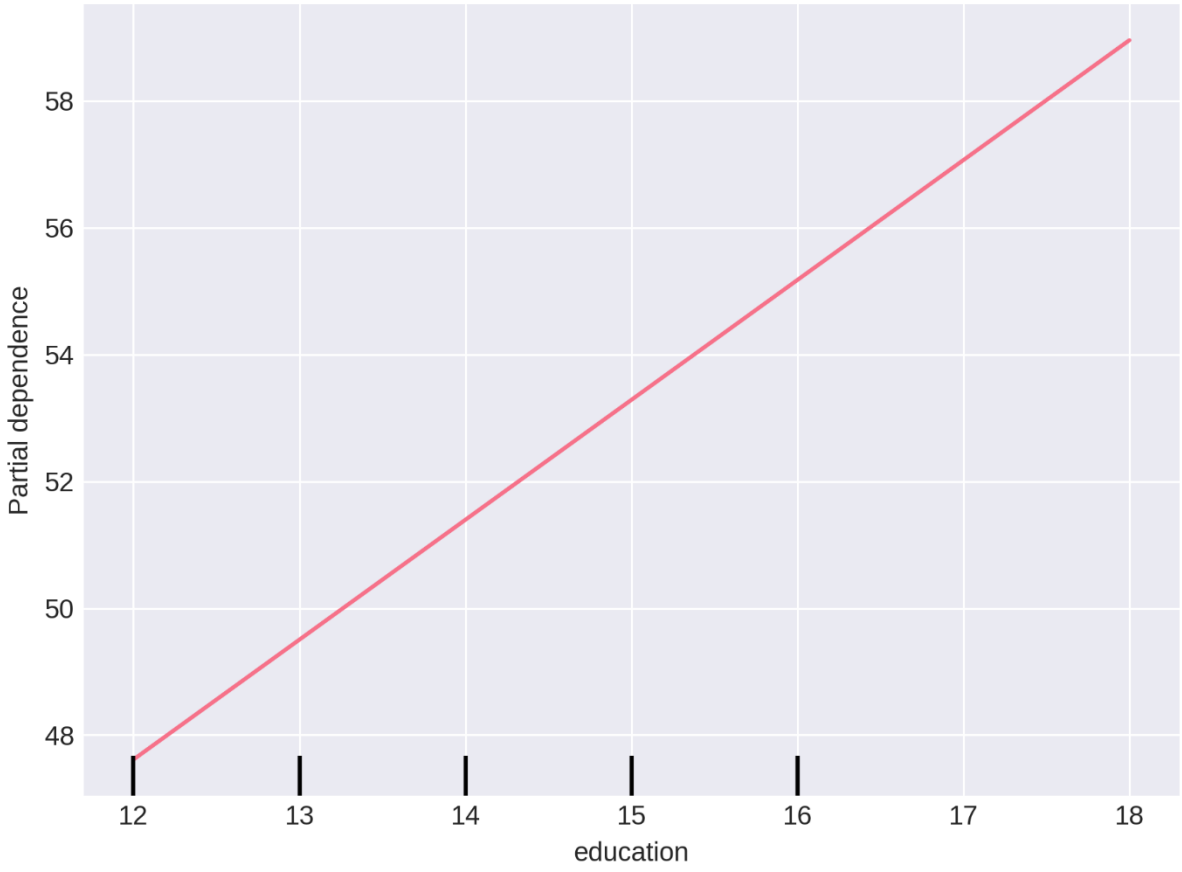


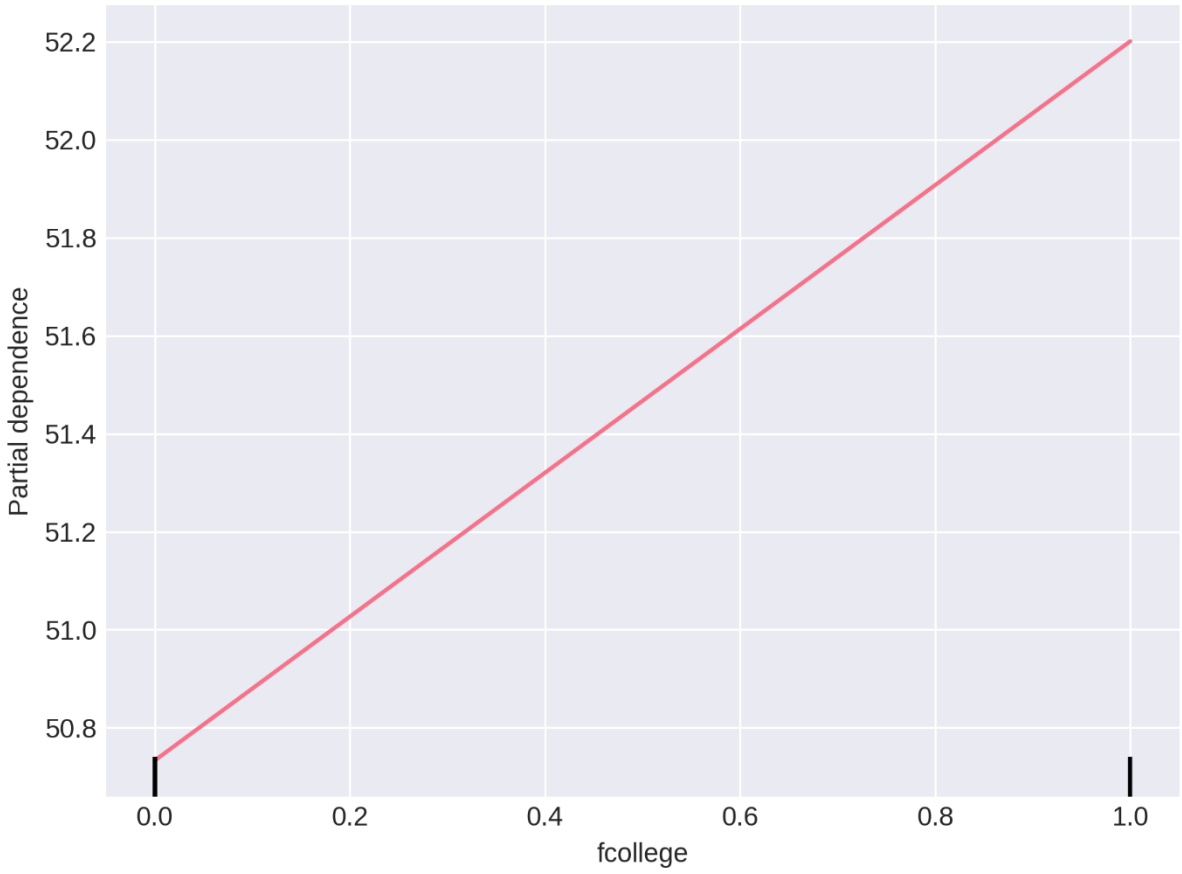

Pairplot: top cech vs score
education
fcollege
mcollege
tuition
home
score


Rozkład: gender
count
gender
male
female
Rozkład: ethnicity
count
ethnicity
other
afam
hispanic
Rozkład: income
count
income
high
low
Rozkład: region
count
region
other
west


Clustermap korelacji zmiennych numerycznych
1.00
0.75
0.50
0.25
0.00
−0.25
urban
home
score
education
fcollege
mcollege
wage
tuition
unemp
distance


Korelacja cech vs score (posortowane)
score
education
fcollege
mcollege
tuition
home
wage
unemp
distance
urban
0.0
0.2
0.4
0.6
0.8
1.0
Korelacja





2. Inżynieria cech i przygotowanie danych

2.1 Podział danych

- Dane podzielono na:
  - **Treningowy:** 80% próbek,
  - **Testowy:** 20% próbek.

2.2 Standaryzacja zmiennych numerycznych

- Zmiennym numerycznym nadano średnią 0 i odchylenie standardowe 1.

2.3 Kodowanie zmiennych kategorycznych

- Zastosowano One-Hot Encoding z opcją handle_unknown='ignore', aby poprawnie obsłużyć kategorie występujące wyłącznie w zbiorze testowym.

2.4 Pipeline przetwarzania danych

- ColumnTransformer scala transformacje numeryczne i kategoryczne w jeden preprocessor, zachowując kolejność kolumn i zapewniając spójne przygotowanie danych dla wszystkich modeli.

3. Wybór i trenowanie modeli

W celu przewidzenia wartości zmiennej score, opracowano i porównano trzy różne modele regresyjne, reprezentujące różne podejścia do analizy danych:

1. **Linear Regression**
   - Charakterystyka: Model parametryczny zakładający liniową zależność między zmiennymi.
   - Zalety: Prosty, szybki w trenowaniu i bardzo dobrze interpretowalny — umożliwia bezpośrednie zrozumienie wpływu poszczególnych cech na wynik.
   - Wynik:
     - R^2 (train): 0.3315
     - R^2 (test): 0.3523
     - MAE: 5.75
     - RMSE: 7.01
       Model ten okazał się stabilny, ale jego dopasowanie (R^2 ~0.35)

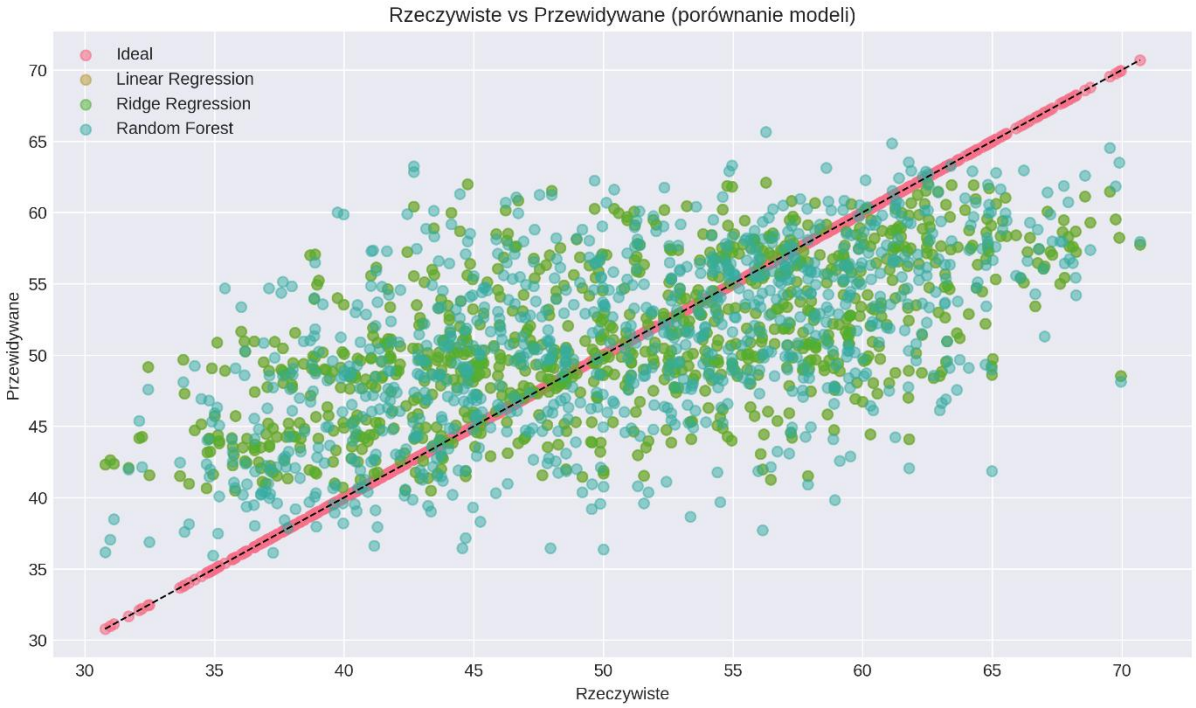
wskazuje, że jedynie część zmienności zmiennej score da się wyjaśnić liniową relacją cech wejściowych.

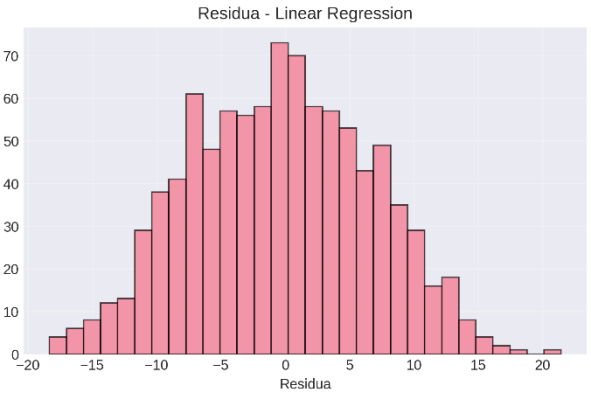
2. **Ridge Regression**
    - Charakterystyka: Rozszerzenie regresji liniowej z regularyzacją L2, która ogranicza wpływ współliniowości między cechami i redukuje przeuczenie.
    - Zalety: Stabilniejszy od klasycznej regresji, zwłaszcza przy dużej liczbie powiązanych predyktorów.
    - Wynik:
        - R^2 (train): 0.3315
        - R^2 (test): 0.3523
        - MAE: 5.75
        - RMSE: 7.01
          Ridge Regression osiągnęła niemal identyczne wyniki jak model liniowy, co sugeruje, że wpływ regularyzacji był minimalny — dane nie wykazywały silnej współliniowości.
3. **Random Forest**
    - Charakterystyka: Model typu ensemble, łączący wiele drzew decyzyjnych dla uzyskania większej dokładności i odporności na szumy.
    - Zalety: Umożliwia modelowanie nieliniowych zależności, jest odporny na wartości odstające i pozwala analizować znaczenie poszczególnych cech.
    - Wynik:
        - R^2 (train): 0.8615
        - R^2 (test): 0.2890
        - MAE: 5.87
        - RMSE: 7.34
          Model ten znacząco przeuczył się (overfitting) — bardzo wysokie R^2 na danych treningowych i spadek na testowych wskazują, że zbyt mocno dopasował się do treningu.

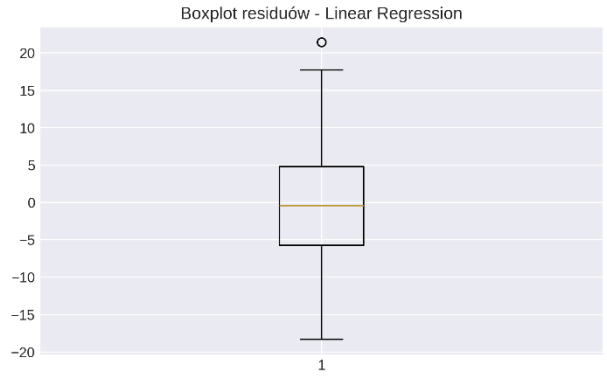
Wybór najlepszego modelu:

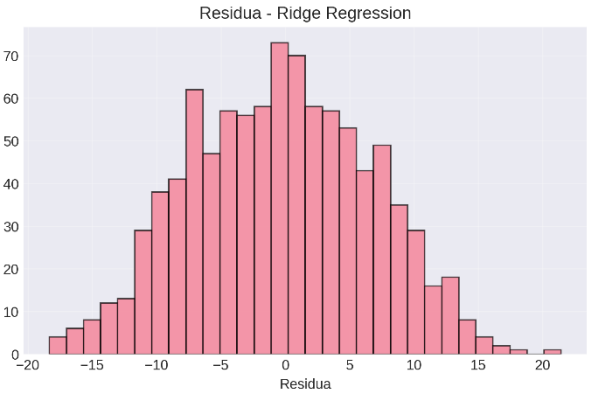
Pomimo wyższej złożoności, Random Forest nie okazał się najlepszym rozwiązaniem ze względu na przeuczenie i spadek dokładności na danych testowych.
Najbardziej zrównoważone i stabilne wyniki uzyskały modele Linear Regression oraz Ridge Regression, osiągając $R^2 \approx 0.35$ zarówno na zbiorze treningowym, jak i testowym.

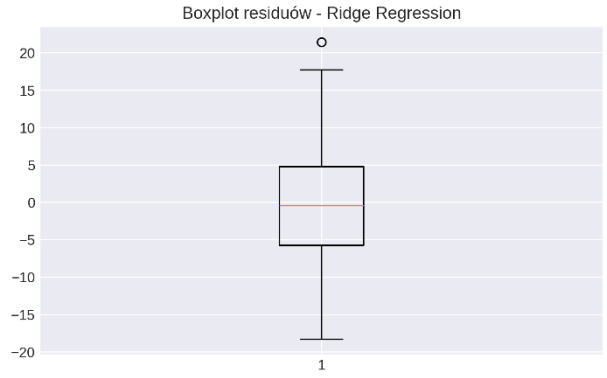
Z tego względu Ridge Regression została uznana za model najlepiej generalizujący dane - zapewnia prostotę interpretacji, stabilność i odporność na współliniowość, przy zachowaniu podobnej jakości predykcji jak model liniowy.

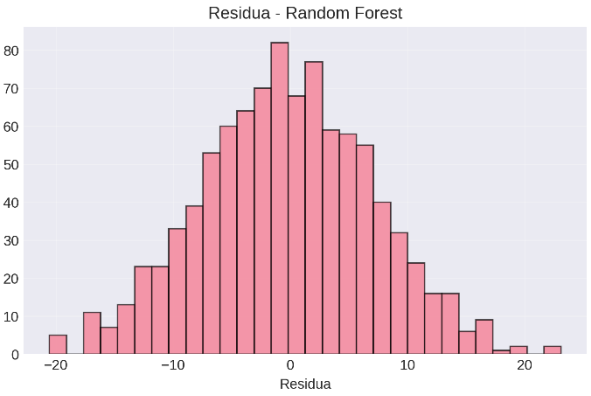

Rzeczywiste vs Przewidywane (porównanie modeli)
Ideal
Linear Regression
Ridge Regression
Random Forest
Przewidywane
Rzeczywiste
30
35
40
45
50
55
60
65
70

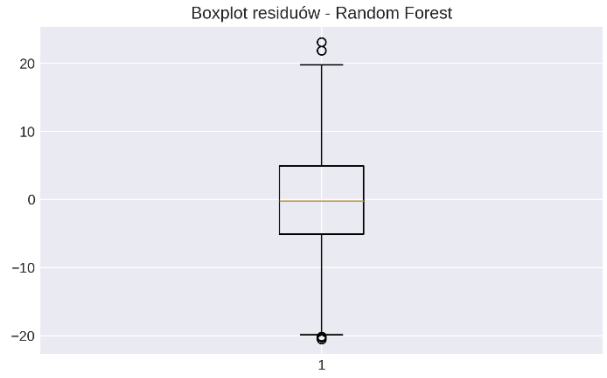

Residua - Linear Regression
Residua

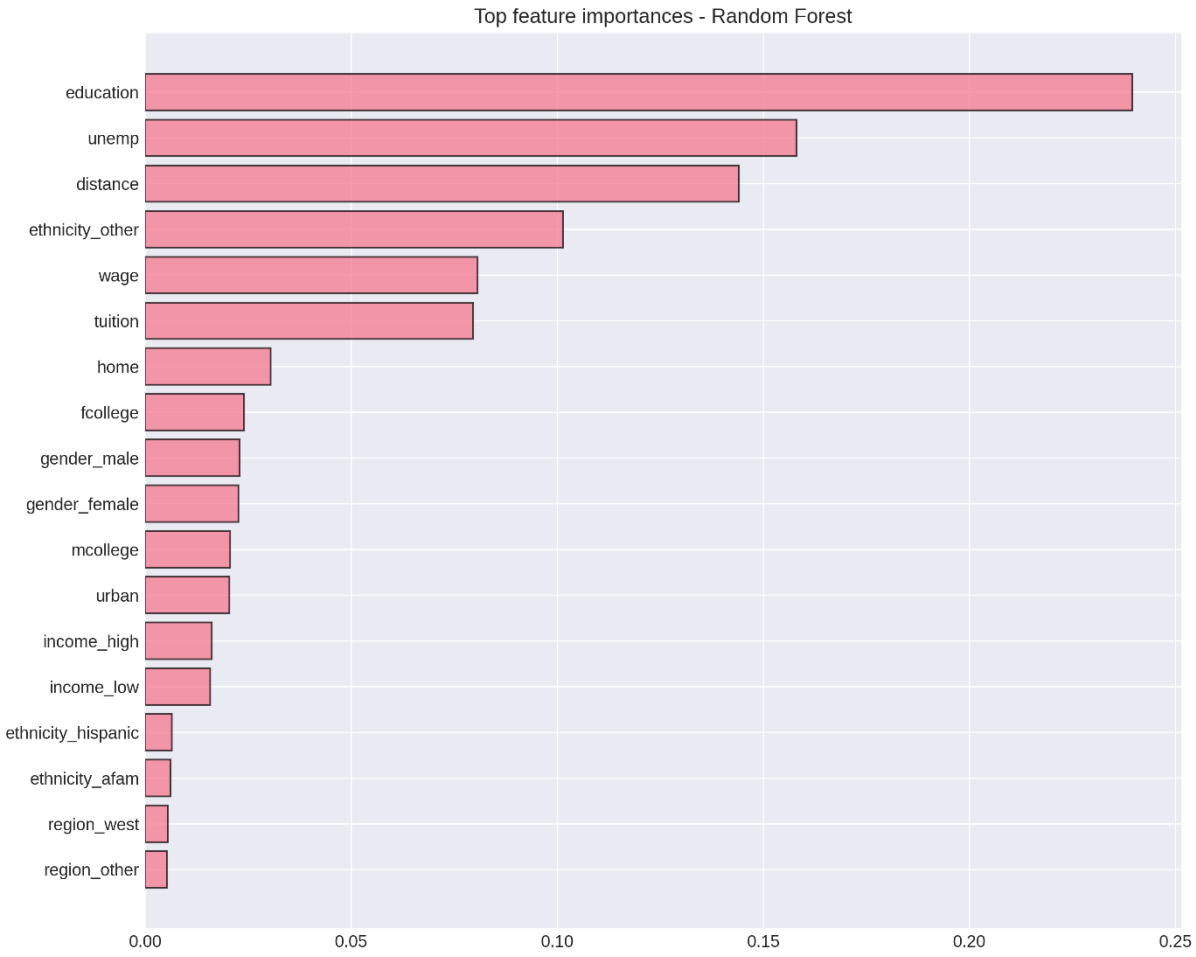

Boxplot residuów - Linear Regression

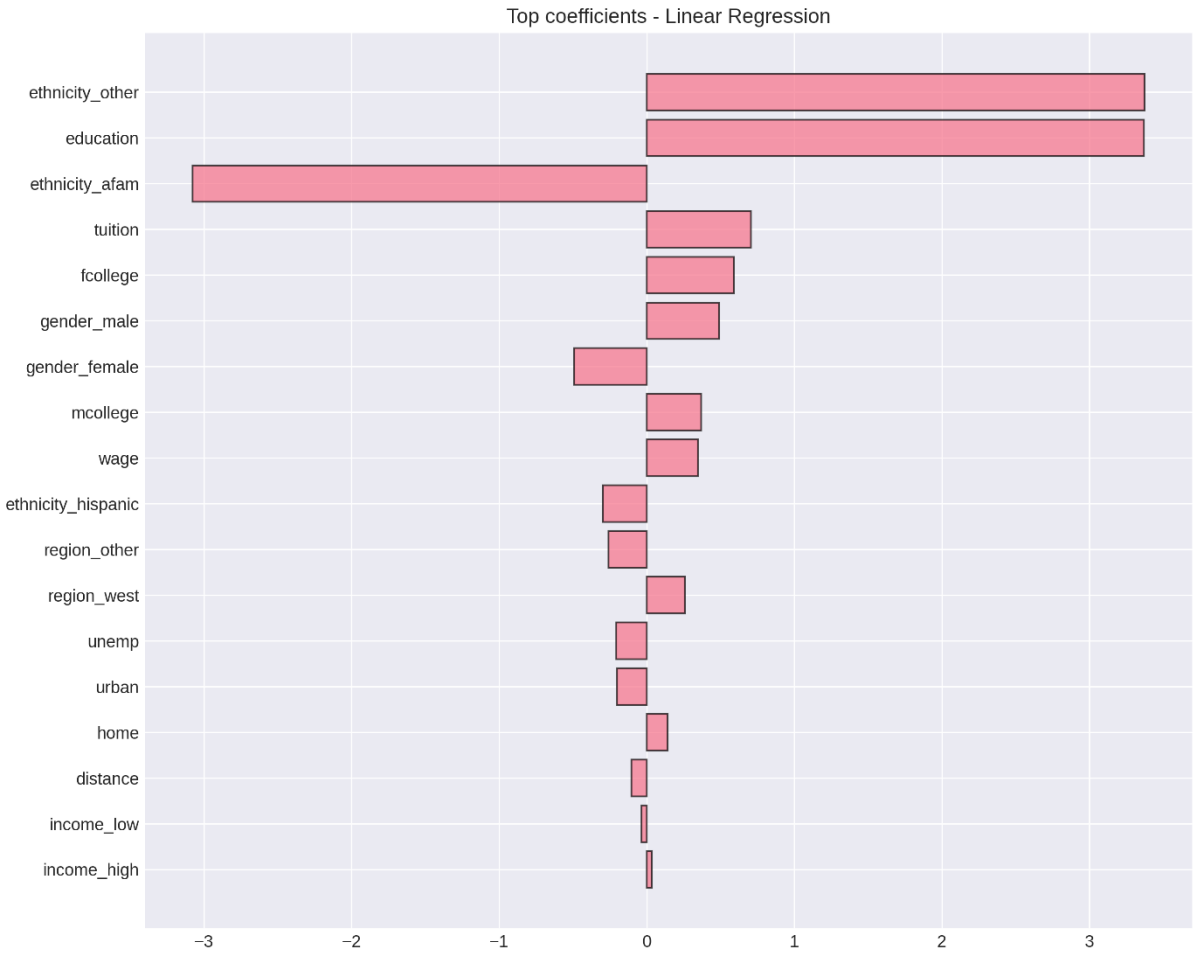

Residua - Ridge Regression
Residua

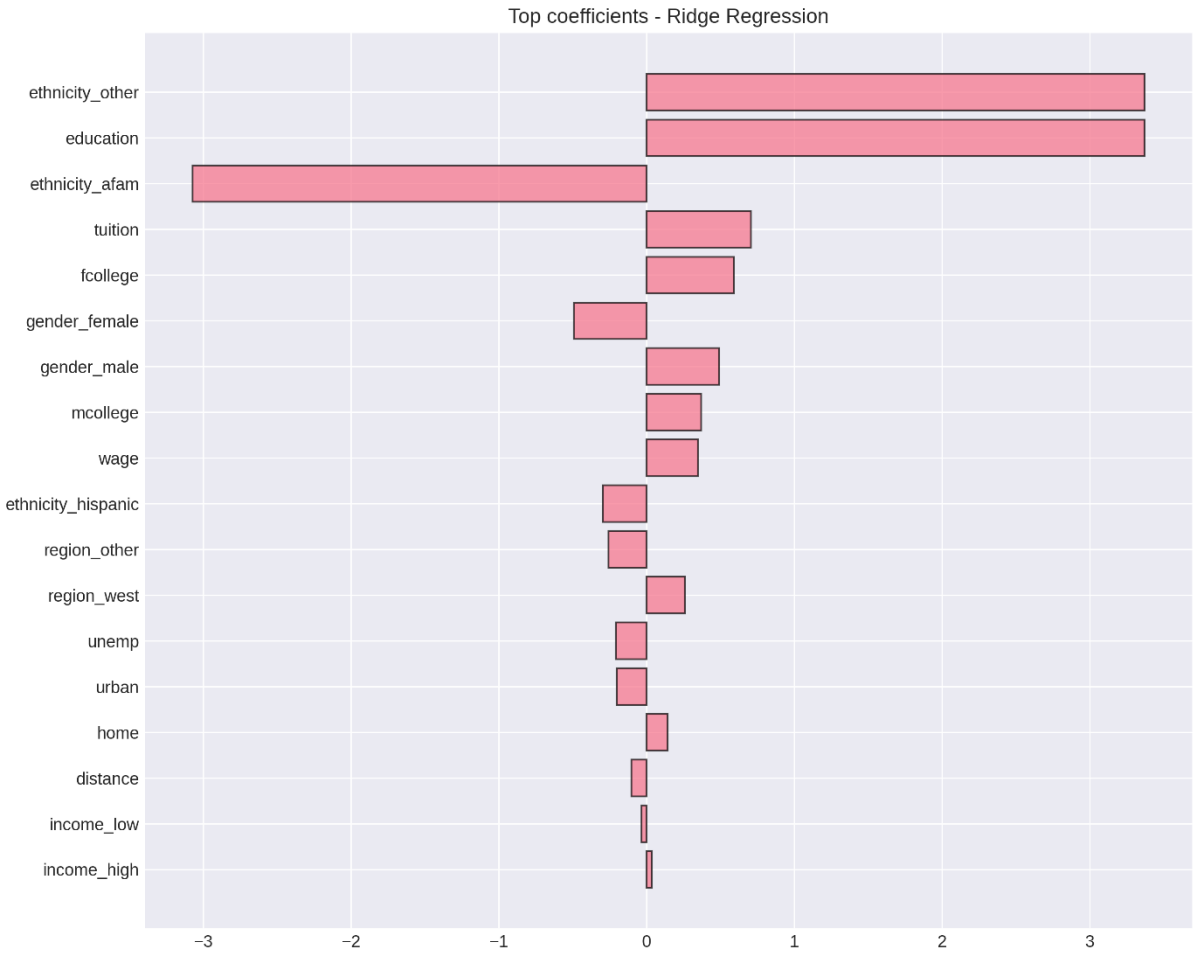

Boxplot residuów - Ridge Regression


Residua - Random Forest
Residua


Boxplot residuów - Random Forest


Top feature importances - Random Forest
education
unemp
distance
ethnicity_other
wage
tuition
home
fcollege
gender_male
gender_female
mcollege
urban
income_high
income_low
ethnicity_hispanic
ethnicity_afam
region_west
region_other
0.00
0.05
0.10
0.15
0.20
0.25


Top coefficients - Linear Regression
ethnicity_other
education
ethnicity_afam
tuition
fcollege
gender_male
gender_female
mcollege
wage
ethnicity_hispanic
region_other
region_west
unemp
urban
home
distance
income_low
income_high
−3
−2
−1
0
1
2
3



4. Ocena i optymalizacja modelu

Metryki oceny

Dla każdego modelu obliczono cztery miary jakości:

- **R^2** – współczynnik determinacji, określający, jaka część zmienności zmiennej score jest wyjaśniana przez model.
- **MAE (Mean Absolute Error)** – średni bezwzględny błąd predykcji.
- **MSE (Mean Squared Error)** – średni błąd kwadratowy.
- **RMSE (Root Mean Squared Error)** – pierwiastek z MSE, interpretowany w jednostkach zmiennej wynikowej.

Wyniki modeli

Model	R^2 (train)	R^2 (test)	R^2 (CV)	MAE	MSE	RMSE
Linear Regression	0.3315	0.3523	0.3252	5.75	49.11	7.01
Ridge Regression	0.3315	0.3523	0.3252	5.75	49.11	7.01
Random Forest	0.8615	0.2890	0.2358	5.87	53.92	7.34

Interpretacja wyników

Modele **Linear Regression** i **Ridge Regression** uzyskały bardzo zbliżone wyniki. Obydwa wyjaśniają około 35% zmienności zmiennej score.
Oznacza to umiarkowaną skuteczność predykcji przy jednoczesnym zachowaniu prostoty i interpretowalności modelu.

Model **Random Forest** osiągnął znacznie wyższy wynik R^2 na zbiorze treningowym (0.86), lecz dużo niższy na zbiorze testowym (0.29).
Taka różnica wskazuje na przeuczenie modelu - bardzo dobre dopasowanie do danych treningowych, ale słabe uogólnienie na nowe obserwacje.

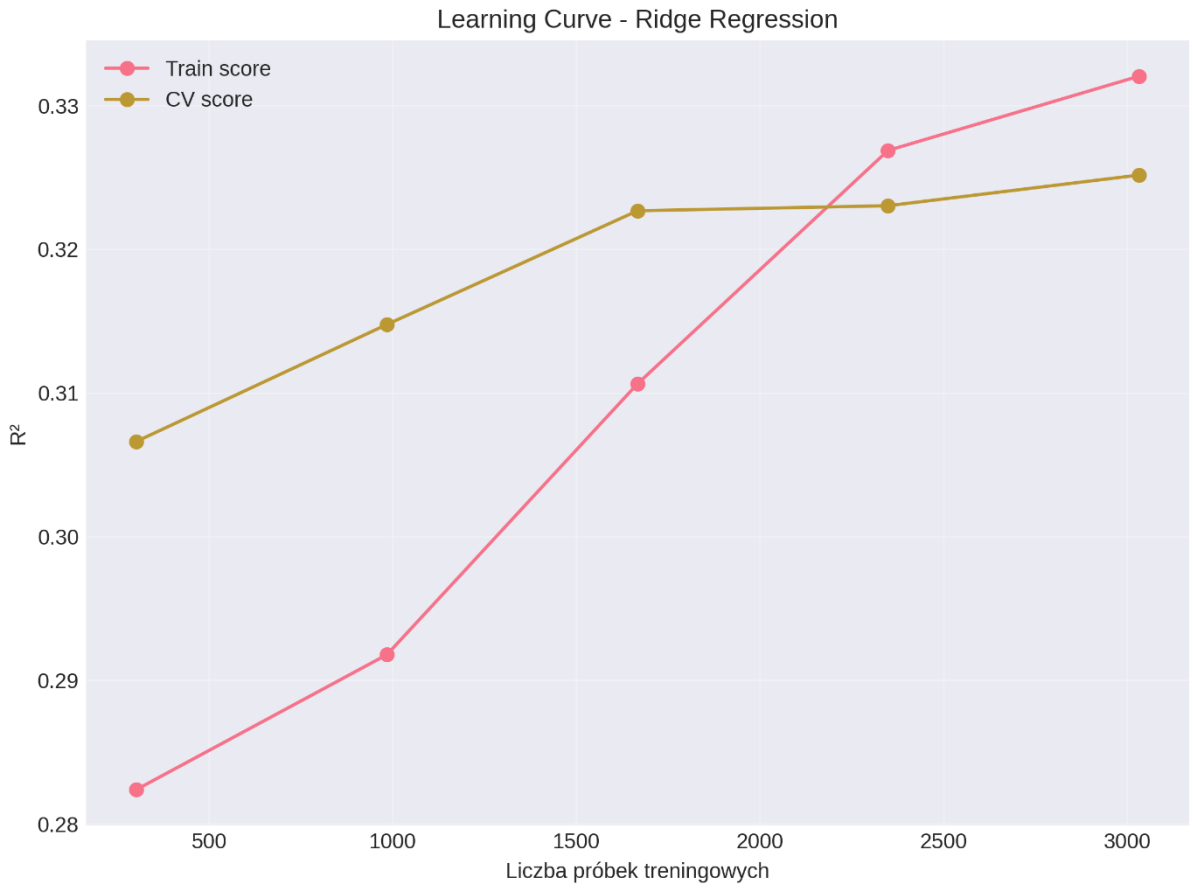
Wybór najlepszego modelu

Za **model końcowy** wybrano **Ridge Regression**, ponieważ:

- Oferuje najlepszy kompromis między dokładnością a stabilnością.
- Jest odporny na współliniowość zmiennych.
- Daje wyniki porównywalne z klasyczną regresją liniową, ale z mniejszym ryzykiem przeuczenia.
- Charakteryzuje się prostą interpretacją współczynników, co ułatwia analizę wpływu poszczególnych cech.

Podsumowanie

Przeprowadzona ewaluacja potwierdziła, że modele liniowe sprawdzają się lepiej w tym zadaniu niż modele złożone, takie jak Random Forest.
Ridge Regression stanowi optymalne rozwiązanie – jest stabilna, dobrze dopasowana i zachowuje równowagę między prostotą a skutecznością.


Learning Curve - Ridge Regression
Train score
CV score
0.33
0.32
0.31
0.30
0.29
0.28
R²
500
1000
1500
2000
2500
3000
Liczba próbek treningowych