Analizator wyników - Lab3

Eksploracja i wstępna analiza danych

Wczytywanie danych i analiza wstępna były wykonane przy użyciu pakietu pandas, umożliwiając identyfikację brakujących wartości oraz usunięcie zbędnej kolumny 'rownames'. Wstępna analiza wykazała obecność zmiennych kategorycznych, które zostały zakodowane.

Kod:

```
data = pd.read_csv('data.csv')
data = data.drop(columns=['rownames'])
```

Przygotowanie danych i inżynieria cech

Przeprowadzono skalowanie cech numerycznych i kodowanie zmiennych kategorycznych. Dodano również cechy wielomianowe, aby umożliwić modelowi lepsze dopasowanie nieliniowe.

```
# Definicja kolumn numerycznych i kategorycznych
numeric_features = ['unemp', 'wage', 'distance', 'tuition', 'education']
categorical_features = ['gender', 'ethnicity', 'fcollege', 'mcollege', 'home', 'urban', 'income', 'region']

# Przetwarzanie wstępne
preprocessor = ColumnTransformer(
   transformers=[
      ('num', StandardScaler(), numeric_features),
      ('cat', OneHotEncoder(drop='first'), categorical_features)
   ]
)
poly_features = PolynomialFeatures(degree=2, include_bias=False)
```

Wybór i trenowanie modelu

Wybrano modele Ridge Regression oraz Gradient Boosting Regressor. Ridge wykorzystano z cechami wielomianowymi, a Gradient Boosting do bardziej kompleksowych predykcji.

```
# Rozdział danych na zbiory treningowe i testowe
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Trenowanie modelu Ridge
ridge_model = Pipeline(steps=[
   ('preprocessor', preprocessor),
   ('poly', poly_features),
   ('ridge', Ridge(alpha=1.0))
])
ridge_model.fit(X_train, y_train)

# Trenowanie modelu Gradient Boosting
gb_model = Pipeline(steps=[
   ('preprocessor', preprocessor),
   ('gb', GradientBoostingRegressor(n_estimators=100, learning_rate=0.1,
random_state=42))
])
gb_model.fit(X_train, y_train)
```

Ocena i optymalizacja modelu

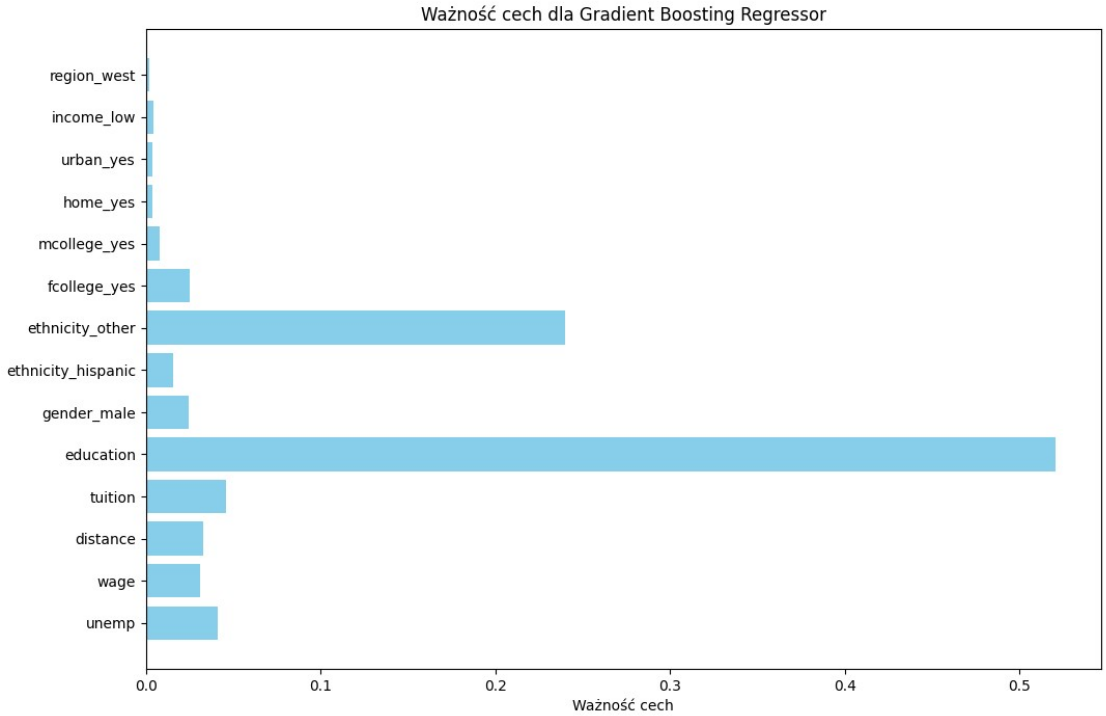
Ocena jakości modeli na zbiorze testowym była przeprowadzona przy użyciu metryk R^2, MAE oraz MSE.

```
Regresja Ridge z cechami wielomianowymi:
Średni R² (kroswalidacja): 0.3256096544889415

Gradient Boosting Regressor:
R²: 0.3670534571944388
MAE: 5.686335982877614
MSE: 47.99783607869402
```

Wizualizacja ważności cech

Ważność cech była zidentyfikowana dla Gradient Boosting Regressor, umożliwiając analizę kluczowych zmiennych.


Ważność cech dla Gradient Boosting Regressor
region_west
income_low
urban_yes
home_yes
mcollege_yes
fcollege_yes
ethnicity_other
ethnicity_hispanic
gender_male
education
tuition
distance
wage
unemp
0.0
0.1
0.2
0.3
0.4
0.5
Ważność cech