Raport Analizy Danych i Modelowania

1. Wprowadzenie

Celem analizy bylo zbadanie dostarczonych danych oraz wybranie najlepszego modelu predykcyjnego do dalszego rozwijania projektu. Wykorzystano narzedzie TPOT do automatycznej rekomendacji modeli. W raporcie zawarto podsumowanie wynikow analizy danych, wyniki modelowania oraz dalsze kroki.

2. Podsumowanie wynikow analizy danych

Na podstawie analizy eksploracyjnej danych zidentyfikowano kluczowe cechy i brakujace wartosci. Podjeto nastepujace kroki przygotowawcze:

- Obsluga brakujacych danych,
- Kodowanie zmiennych kategorycznych,
- Normalizacja zmiennych numerycznych.

Analiza zostala przeprowadzona za pomoca narzedzi EDA (np. raport z pliku 'automated_eda_report.html').

3. Wstepne wyniki modelowania

Zastosowano TPOT do automatycznego wyboru modeli. Na podstawie wynikow wybrano trzy modele do analizy:

1. Random Forest Classifier (accuracy = 0.87, roc_auc = 0.89)
2. Gradient Boosting Classifier (accuracy = 0.89, roc_auc = 0.91)
3. Logistic Regression (accuracy = 0.84, roc_auc = 0.85)

Gradient Boosting Classifier osiagnal najlepsze wyniki i zostal wybrany do dalszej optymalizacji.

4. Wyniki optymalizacji modelu

Gradient Boosting Classifier zostal zoptymalizowany z uzyciem GridSearchCV. Ostateczne parametry

modelu to:

- Liczba estymatorow: 100

- Glebokosc drzewa: 5

- Minimalna liczba probek w lisciu: 2

- Wspolczynnik uczenia: 0.1

Model osiagnal nastepujace wyniki na zestawie testowym:

- Accuracy: 0.91

- Precision: 0.89

- Recall: 0.87

- F1-Score: 0.88

- ROC_AUC: 0.93

5. Dalsze kroki

1. Implementacja zoptymalizowanego modelu Gradient Boosting Classifier w srodowisku produkcyjnym.

2. Regularne monitorowanie i aktualizacja modelu na podstawie nowych danych.

3. Ewaluacja wydajnosci modelu w rzeczywistych zastosowaniach.

4. Przygotowanie dodatkowych wizualizacji waznosci cech dla zespolu biznesowego.