Zaawansowane metody uczenia maszynowego

Projekt 1

Mikołaj Małkiński

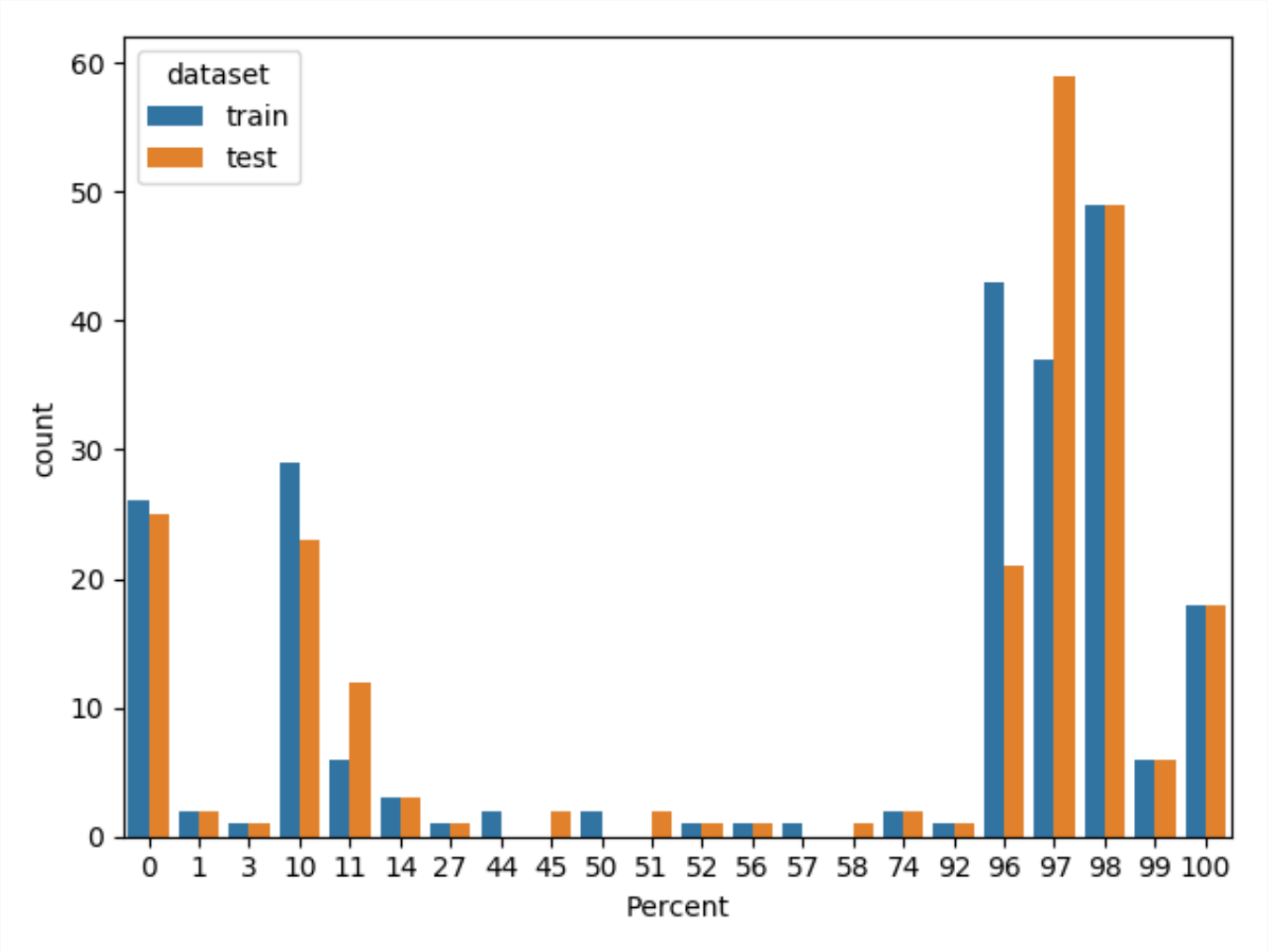
29 kwietnia 2019

Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych

Spis treści

Przygotowanie danych

Brakujące dane



Rysunek 1: Liczba kolumn posiadających dany procent brakujących danych z podziałem na zbiór treningowy i testowy

Brakujące dane - rozwiązanie

- Usunięcie kolumn które mają więcej niż 90% braków
- Kolumny numeryczne - wypełnienie medianą
- Kolumny kategoryczne - dodanie nowej kategorii: *unknown*

Unikalność danych - rozwiązanie

- Usunięcie kolumn kategorycznych z 1 unikalną wartością
- Usunięcie kolumn kategorycznych z ponad 100 unikalnymi wartościami
- (Opcjonalnie) - potraktowanie kolumn numerycznych mających mało unikalnych danych jako kategoryczne

Przekształcenie kolumn kategorycznych - One-hot encoding

Var42
foo
bar
bar
foo
foo

(a) Przed

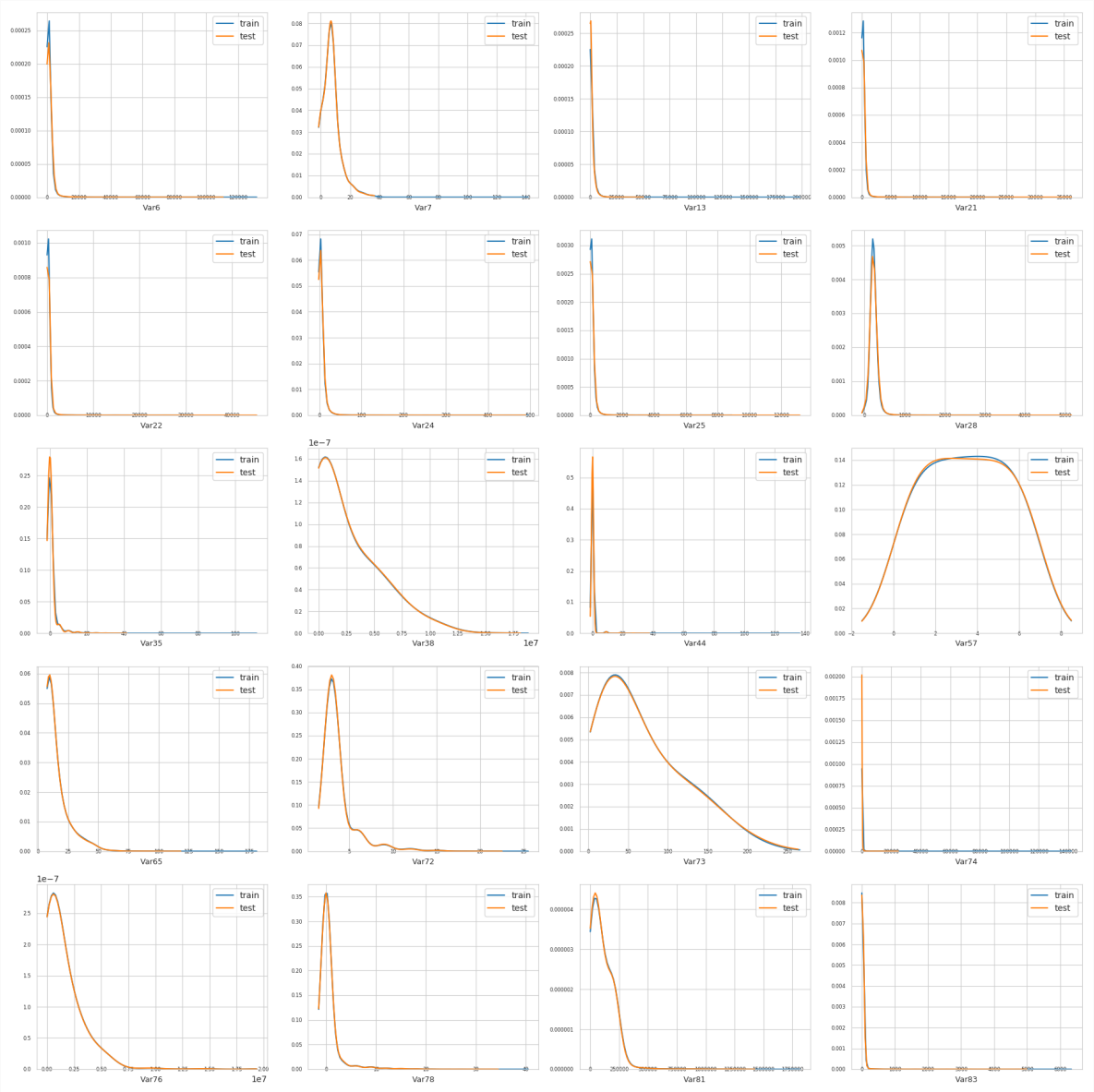
Var42_foo	Var42_bar
1	0
0	1
0	1
1	0
1	0

(b) Po

Rysunek 2: One-hot encoding na kolumnie posiadającej 2 różne wartości

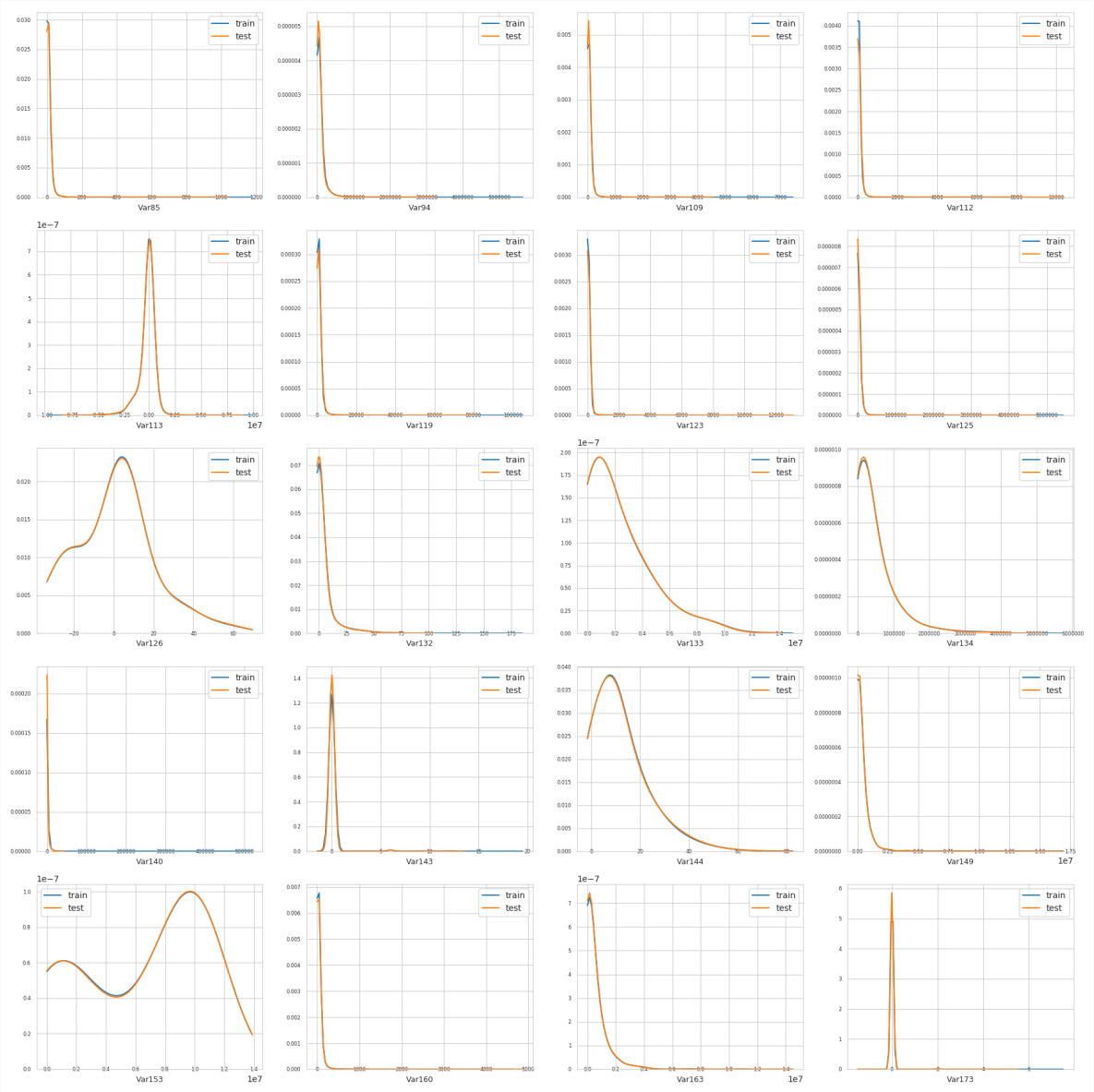
Analiza danych

Porównawcza analiza danych



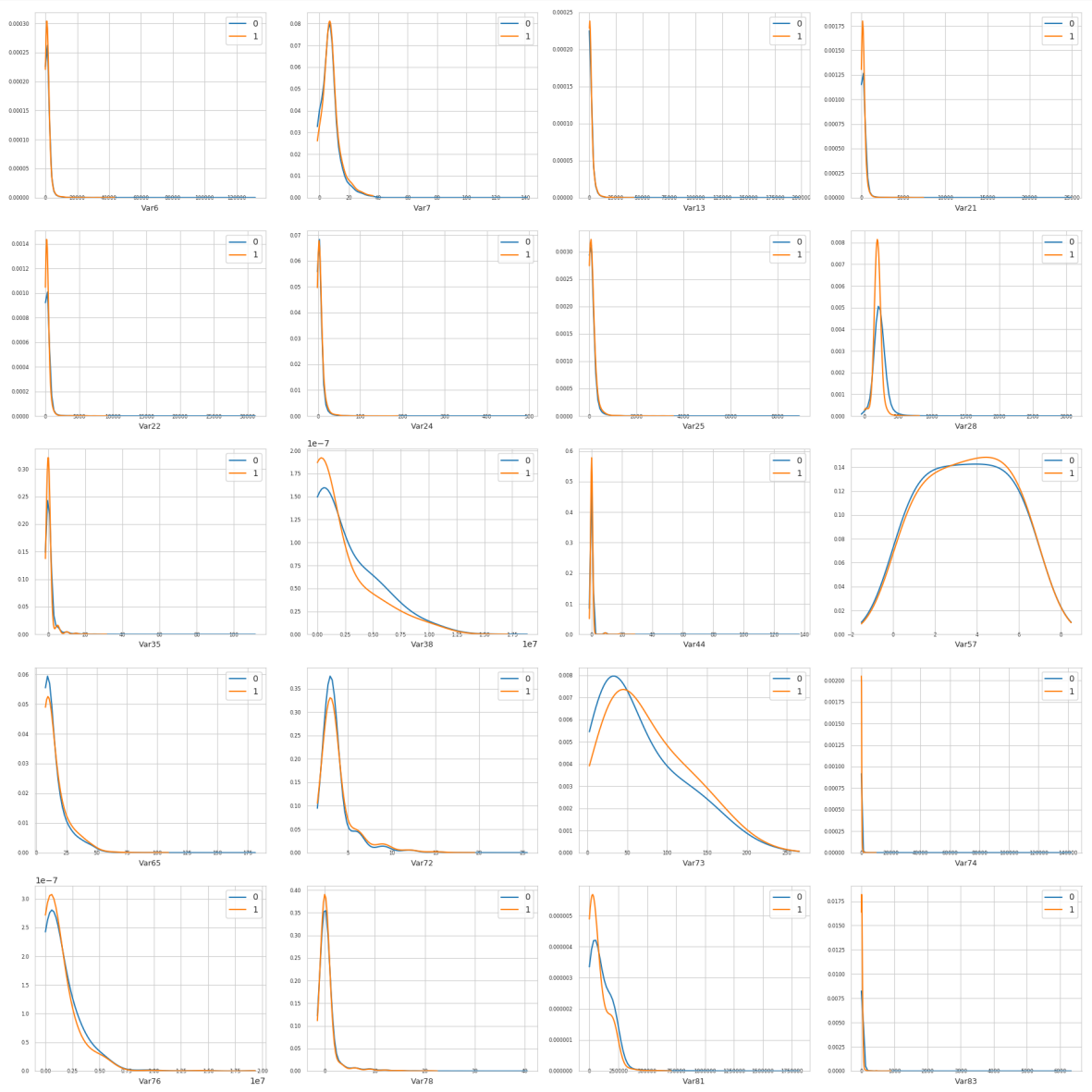
Rysunek 3: Porównanie rozkładów cech ze względu na zbiór

Porównawcza analiza danych



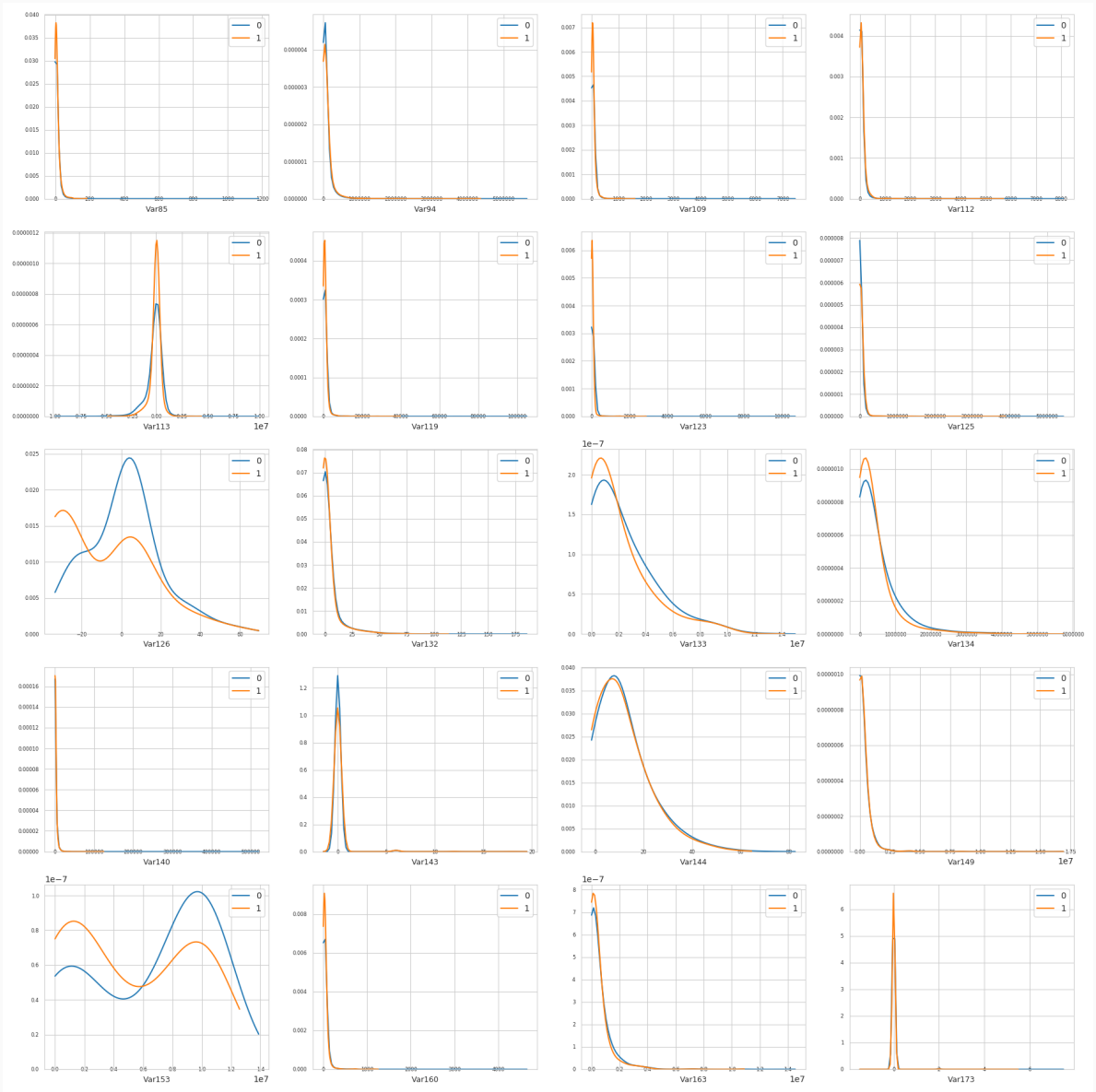
Rysunek 4: Porównanie rozkładów cech ze względu na zbiór

Porównawcza analiza danych



Rysunek 5: Porównanie rozkładów cech ze względu na klasę

Porównawcza analiza danych



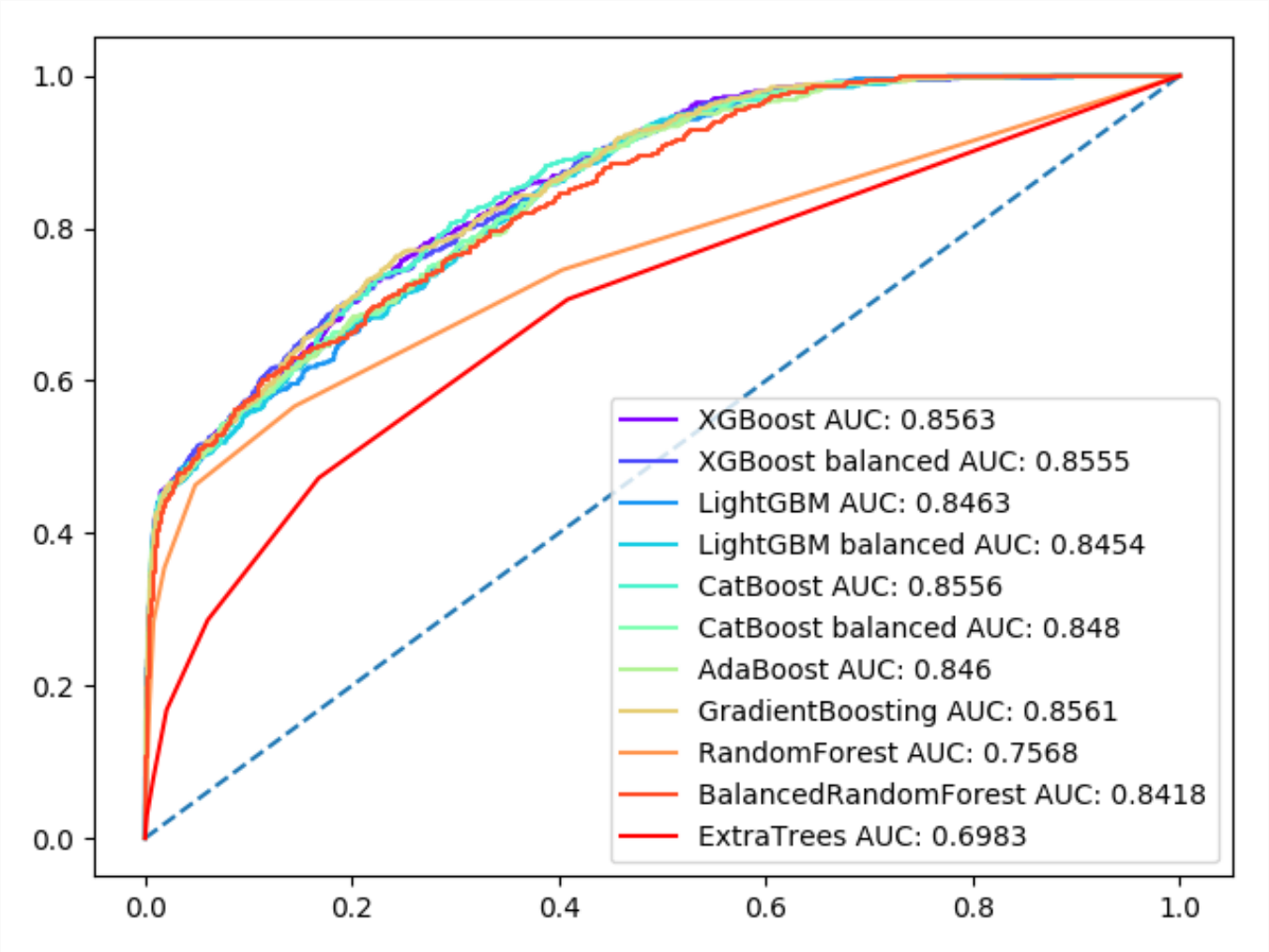
Rysunek 6: Porównanie rozkładów cech ze względu na klasę

Podstawowe klasyfikatory

Podstawowe klasyfikatory

- XGBoost
- XGBoost balanced
- LightGBM
- LightGBM balanced
- CatBoost
- CatBoost balanced
- AdaBoost
- GradientBoosting
- RandomForest
- BalancedRandomForest
- ExtraTrees

Podstawowe klasyfikatory - krzywe ROC



Rysunek 7: Porównanie krzywych ROC podstawowych modeli klasyfikacyjnych

Podstawowe klasyfikatory - rezultaty

	AUC	Dokładność	Precyzja@10
XGBoost	**0.8563**	**0.9516**	**0.3600**
XGBoost balanced	0.8555	0.7742	0.3588
LightGBM	0.8463	0.9508	0.3575
LightGBM balanced	0.8454	0.8654	0.3488
CatBoost	0.8556	0.9514	0.3500
CatBoost balanced	0.8480	0.8450	0.3563
AdaBoost	0.8460	0.9505	0.3550
GradientBoosting	0.8561	0.9510	0.3588
RandomForest	0.7568	0.9404	0.3250
BalancedRandomForest	0.8418	0.7332	0.3575
ExtraTrees	0.6983	0.9328	0.2162

Tablica 1: Metryki podstawowych modeli klasyfikacyjnych

Podstawowe klasyfikatory - rezultaty

	F1	Precyzja	Czułość
XGBoost	**0.5442**	**0.7524**	0.4262
XGBoost balanced	0.3043	0.1923	0.7288
LightGBM	0.5310	0.7483	0.4114
LightGBM balanced	0.3631	0.2672	0.5664
CatBoost	0.5418	0.7492	0.4244
CatBoost balanced	0.3467	0.2426	0.6070
AdaBoost	0.5308	0.7417	0.4133
GradientBoosting	0.5410	0.7404	0.4262
RandomForest	0.2891	0.7519	0.1790
BalancedRandomForest	0.2707	0.1661	**0.7306**
ExtraTrees	0.0561	0.5714	0.0295

Tablica 2: Metryki podstawowych modeli klasyfikacyjnych

Próby poprawienia wyników

Metody balansowania klas

- Zmniejszenie ilości obserwacji z klasy 0
- Zwiększenie ilości obserwacji z klasy 1
- Dodanie sztucznych obserwacji z klasy 1 używając techinki SMOTE

Wybór cech

- Przygotowany zbiór danych posiadał 470 różnych kolumn
- Wytrenowanie modeli XGBoost, LightGBM oraz CatBoost
- Wybranie 30 najbardziej istotnych cech dla każdego modelu
- Połączenie zbiorów dało 57 najważniejszych cech

Optymalizacja najlepszego modelu

Dobór hiperparametrów

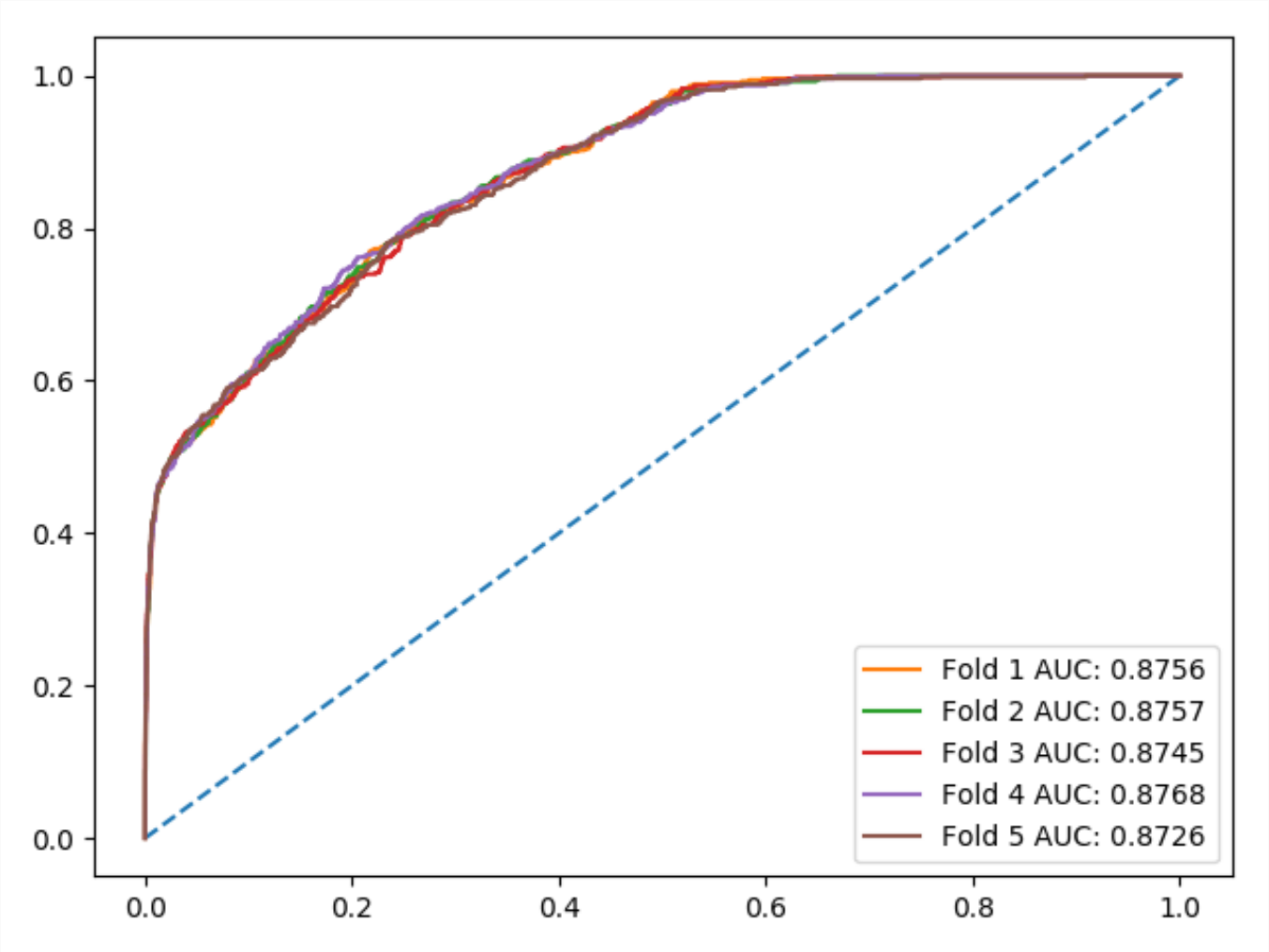
Aby dobrać hiperparametry, użyto 3-krotnej kroswalidacji oraz kierowano się jak największą wartością metryki precyzja@10.
Metody:

- Random search
- Grid search

Hiperparametry:

- stosunek klas
- maksymalna głębokość drzew wchodzących w skład komitetu (base learners)
- współczynnik uczenia
- liczba członków komitetu
- współczynniki próbkowania

Ostateczna ewaluacja - krzywe ROC



Rysunek 8: Porównanie krzywych ROC zoptymalizowanego modelu XGBoost używając kroswalidacji

Ostateczna ewaluacja - metryki

	AUC	Dokładność	Precyzja@10
Próba 1	0.8605	0.7779	0.4000
Próba 2	0.8757	0.7960	0.4263
Próba 3	0.8702	0.7770	0.4075
Próba 4	0.8621	0.7758	0.3987
Próba 5	0.8569	0.7755	0.3812
Średnia	0.8651	0.7804	**0.4027**

Tablica 3: Metryki zoptymalizowanego modelu XGBoost używając kroswalidacji

Ostateczna ewaluacja - metryki

	F1	Precyzja	Czułość
Próba 1	0.3210	0.2071	0.7131
Próba 2	0.3625	0.2395	0.7448
Próba 3	0.3283	0.2099	0.7530
Próba 4	0.3215	0.2055	0.7378
Próba 5	0.3087	0.1971	0.7110
Średnia	0.3284	0.2118	0.7319

Tablica 4: Metryki zoptymalizowanego modelu XGBoost używając kroswalidacji

Podsumowanie

Podsumowanie

- Osiągnięty wynik (40.27%) znacznie odbiega od idealnego (70%)
- Wyniki zastosowanych modeli klasyfikacyjnych były do siebie zbliżone
- Optymalizacja hiperparametrów dla wybranego modelu (XGBoost) przyniosła poprawę jego jakości zbliżoną do 4%
- Główną przeszkodą było silne zanonimizowanie zbioru danych

Dziękuję za uwagę