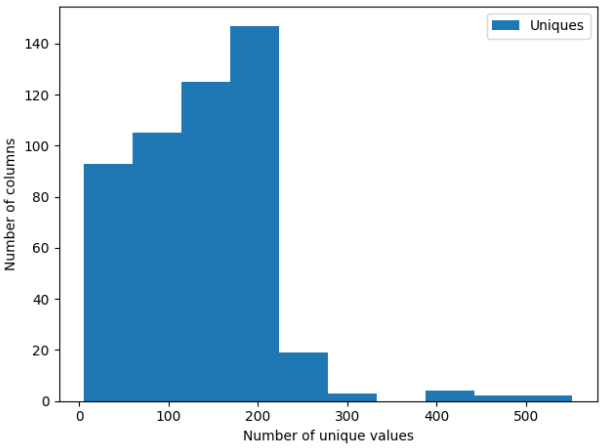
Zaawansowane metody uczenia maszynowego
Projekt 2

Mikołaj Małkiński
malkinskim@student.mini.pw.edu.pl

19 maja 2019

1 Wstępna analiza danych

Oryginalny zbiór danych został podzielony na treningowy (2000 obserwacji) oraz testowy (600 obserwacji). Oba zbiory zawierają 500 kolumn numerycznych. Zbiory nie posiadają żadnych brakujących wartości. Rysunek 1 przedstawia stosunek liczby unikalnych wartości w kolumnie, dla liczby kolumn.



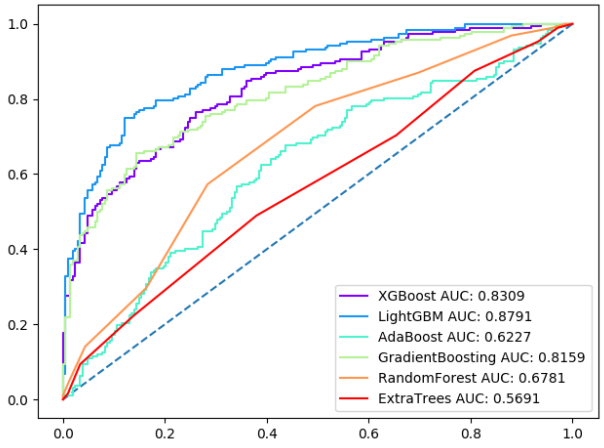
Rysunek 1: Unikalne wartości

Na początku, porównano rozkłady zmiennych ze względu na zbiór danych (Dodatek A) oraz klasę (Dodatek B). Wyraźnie można zauważyć że zmienne zostały wygenerowane z rozkładów normalnych z różną średnią i odchyleniem standardowym. Podział ze względu na zbiór dla niektórych zmiennych pokazuje delikatne różnice, jednak wynika to prawdopodobnie z małych rozmiarów zbiorów. Natomiast, podział ze względu na klasę nie ujawnił żadnych widocznych różnic.

	AUC	Dokładność	Zbalansowana dok.	F1	Precyzja	Czułość
XGBoost	0.8309	0.7400	0.7390	0.7249	0.7366	0.7135
LightGBM	**0.8791**	**0.8025**	**0.8023**	**0.7948**	**0.7927**	**0.7969**
AdaBoost	0.6227	0.6125	0.6138	0.6154	0.5877	0.6458
GradientBoosting	0.8159	0.7275	0.7272	0.7169	0.7150	0.7188
RandomForest	0.6781	0.6475	0.6446	0.6094	0.6509	0.5729
ExtraTrees	0.5691	0.5575	0.5549	0.5151	0.5434	0.4896

Tablica 1: Metryki podstawowych modeli klasyfikacyjnych

2 Podstawowe modele w domyślnych konfiguracjach



Rysunek 2: Porównanie krzywych ROC podstawowych modeli klasyfikacyjnych

Po odpowiednim przygotowaniu danych, w prosty sposób można sprawdzić jakość różnych modeli z domyślnymi parametrami. Zbiór danych został podzielony na treningowy i walidacyjny w stosunku 4:1. Zbadano jakość następujących klasyfikatorów: *XGBoost* [2], *LightGBM* [6], *AdaBoost* [3], *GradientBoosting* [4], *RandomForest* [1], *ExtraTrees* [5].

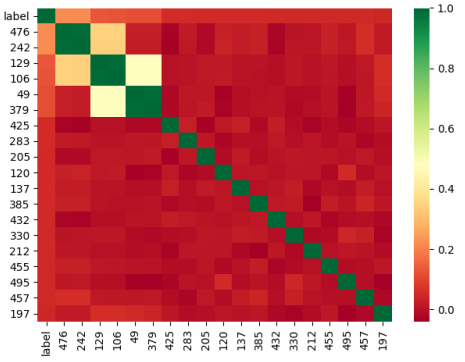
Krzywe ROC tych modeli zaprezentowane są na Rysunku 2. Łatwo zauważyć rozbieżność w otrzymanych krzywych ROC. Przypuszcza się, że jest to spowodowane dużą liczbą nieistotnych kolumn w zbiorze danych - niektóre z użytych modeli nie są w stanie dobrze wybrać wartościowe kolumny. Najlepsze wyniki osiągnęły modele LightGBM oraz XGBoost, więc to one zostaną użyte do przeprowadzenia kolejnych eksperymentów. Dokładniejsze statystki przedstawione są w Tablicy 1. Porównano w niej takie metryki jak: AUC, dokładność, zbalansowaną dokładność, F1, precyzję oraz czułość.

3 Wybór cech

Oryginalny zbiór danych posiadał 500 kolumn numerycznych. Aby ograniczyć ich liczbę, wybrano te, które są najważniejsze do dokonania poprawnej klasyfikacji. Zmniejszenie liczby zmiennych przede wszystkim pomaga zmniejszeniu zjawiska przetrenowania, co pozytywnie wpływa na dokładność klasyfikacji, oraz znacząco redukuje czas trenowania. W tej sekcji zostały opisane metody które zostały wykorzystane do ostatecznego wyboru zmiennych.

3.1 Macierz korelacji

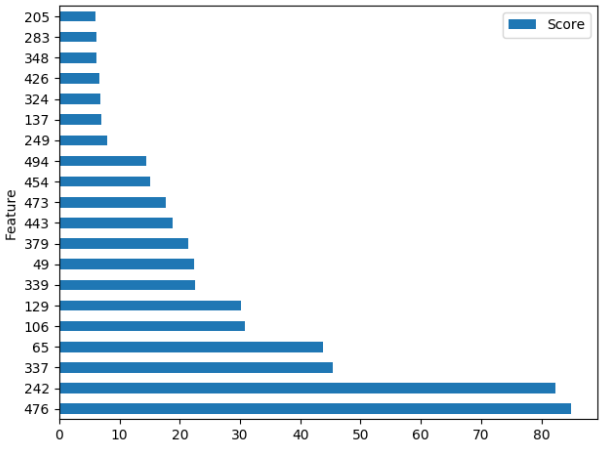
Pierwszym sposobem było zbadanie korelacji między zmiennymi. Wynik dla 20 zmiennych zawierających najwyższe wartości zaprezentowany jest na Rysunku 3. Wizualizacja pokazuje że korelacja jest mała, jednak warto zwrócić uwagę na zmienne: 476, 242, 129, 106, 49 oraz 379. Pomimo tego że wybrane pary są ze sobą wyraźnie bardziej skorelowane niż inne, to są one także zależne od zmiennej *label*, która określa klasę do której przynależy dana obserwacja. Eksperymenty omówione w Sekcji 3.3 wykazały że zmienne z wyjątkiem 129 zostały wybrane w jednym bądź obu przypadkach.



Rysunek 3: Macierz korelacji ograniczona do 20 największych wartości

3.2 Analiza jednowymiarowa

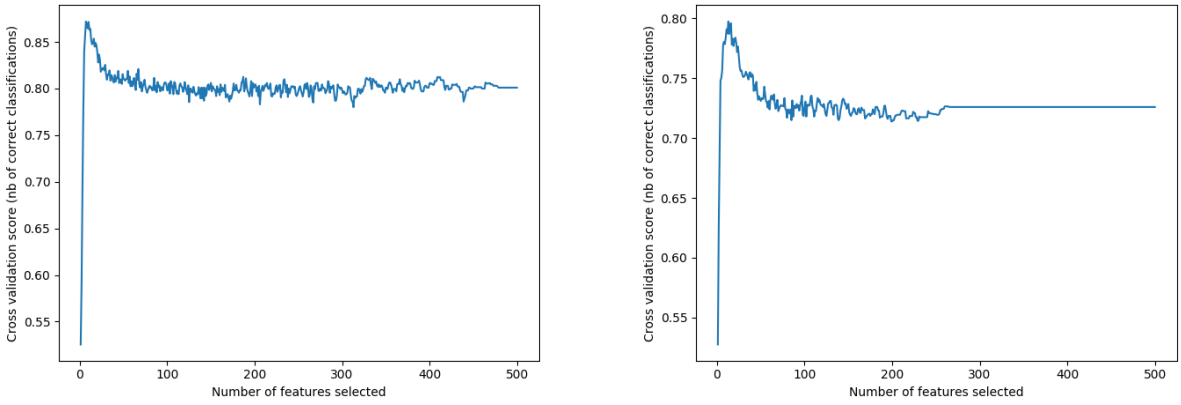
Kolejną metodą było przeprowadzonenie testu statystycznego, który osobno sprawdzał zależność między każdą zmienną a zmienną określającą przynależność do danej klasy. Zmienne zostały wybrane na podstawie F-wartości osiągniętej z testu ANOVA (analiza wariancji). Rysunek 4 przedstawia otrzymane wyniki. Warto zwrócić uwagę, że ponownie wyróżniły się zmienne 476, 242, 106, 129, 49 oraz 379, które zostały wyszczególnione w Sekcji 3.1. Dodatkowo, dobre wyniki osiągnęły zmienne 337, 339, 443, 473 oraz 454, które także pojawią się w Sekcji 3.3.



Rysunek 4: 20 największych wartości otrzymanych za pomocą testu ANOVA

3.3 Rekurencyjna eliminacja cech

Oba modele, LightGBM oraz XGBoost, po zakończonym treningu są w stanie określić które cechy są najbardziej istotne dla wytrenowanego modelu. Wykresy pokazujące przykładowy ranking przedstawione są w Dodatkach C oraz D. Na podstawie takich metryk, przeprowadzona została eliminacja cech, używając 5-krotnej kroswalidacji. Polega ona na wielokrotnym trenowaniu modelu, usuwając najmniej istotne cechy po każdej iteracji. Rysunek 5 przedstawia zbalansowaną dokładność modeli LightGBM oraz XGBoost w zależności od ilości wykorzystanych cech. W pierwszym przypadku, najlepszy wynik został osiągnięty korzystając z 7 cech (106, 154, 319, 337, 379, 443, 454), a w drugim korzystając z 13 cech (29, 49, 106, 154, 242, 282, 319, 339, 379, 443, 452, 473, 476). Wyraźnie widać że wykorzystane cechy zostały już zauważone w Sekcjach 3.1 oraz 3.2.



(a) LightGBM

(b) XGBoost

Rysunek 5: Zbalansowana dokładność w zależności od liczby zmiennych

4 Dobór hiperparametrów

Następnie, dla obu modeli LightGBM oraz XGBoost dokonano doboru hiperparametrów korzystając z 5-krotnej kroswalidacji, tak aby osiągnąć jak najwyższą zbalansowaną dokładność. Dla każdego z modeli użyto innego zbioru danych, złożonego tylko z cech które zostały wybrane dla danego modelu w Sekcji 3.3. LightGBM, korzystając z 7 cech, osiągnął średnią zbalansowaną dokładność równą 87.2%, a XGBoost, korzystając z 13, równą 85.9%. Na tej podstawie, do ostatecznej predykcji klas dla zbioru testowego użyto modelu LightGBM.

Literatura

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[2] Tianqi Chen and Carlos Guestrin. XGBoost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[3] Yoav Freund and Robert E Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14:771–780, 10 1999.

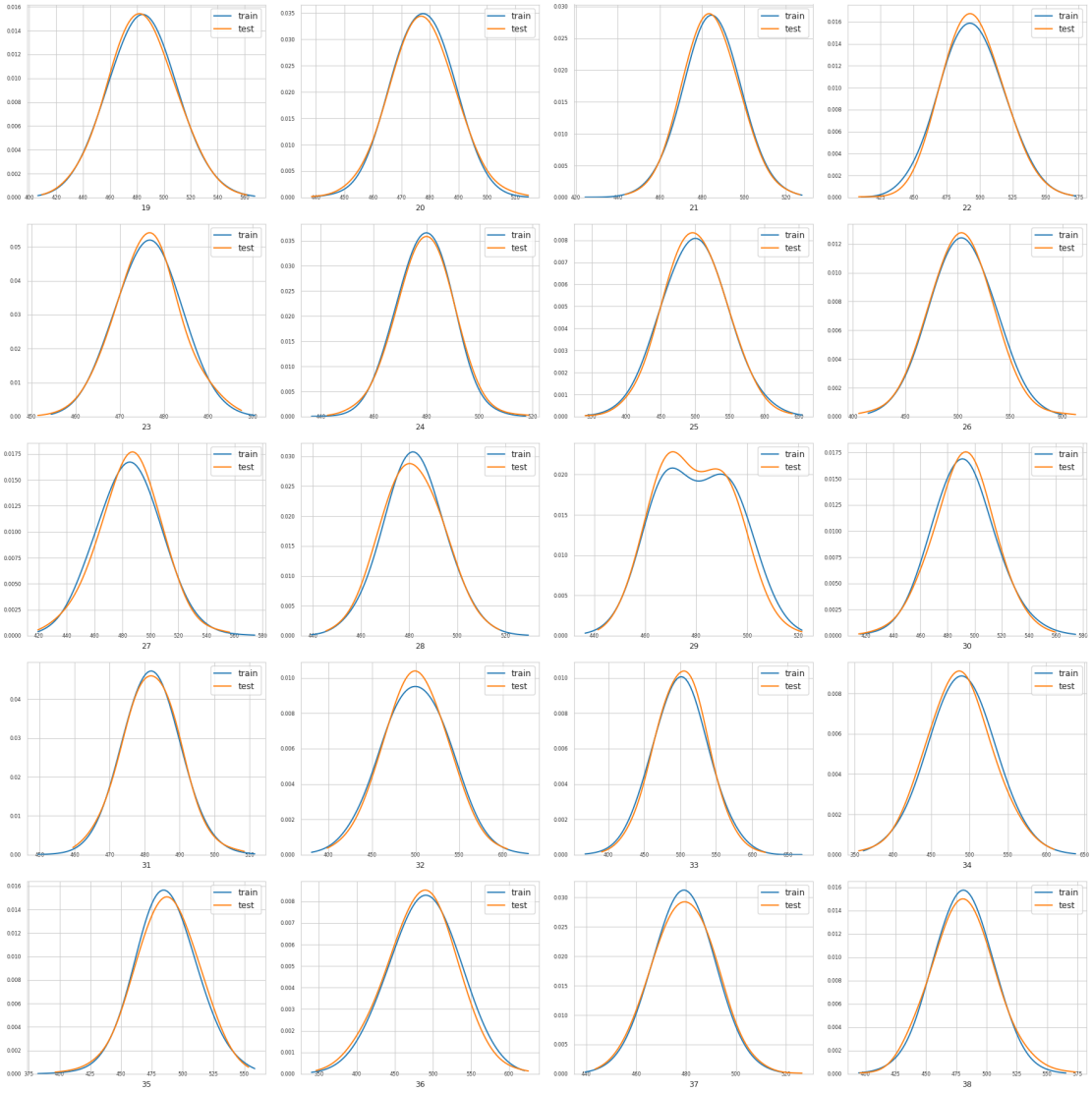
[4] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 10 2001.

[5] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, Apr 2006.

[6] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017.

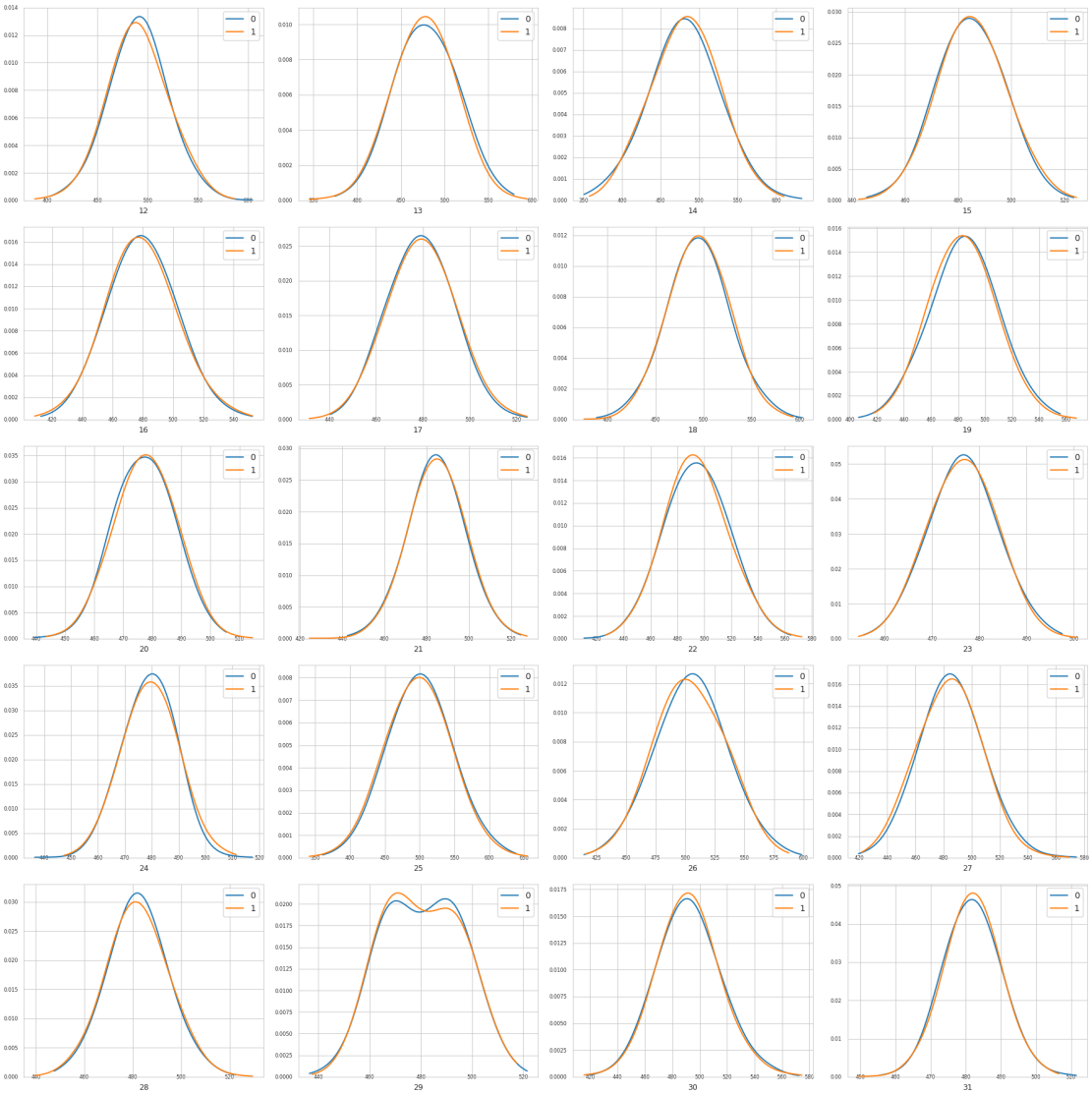
Dodatki

A Rozkład cech z podziałem na zbiór



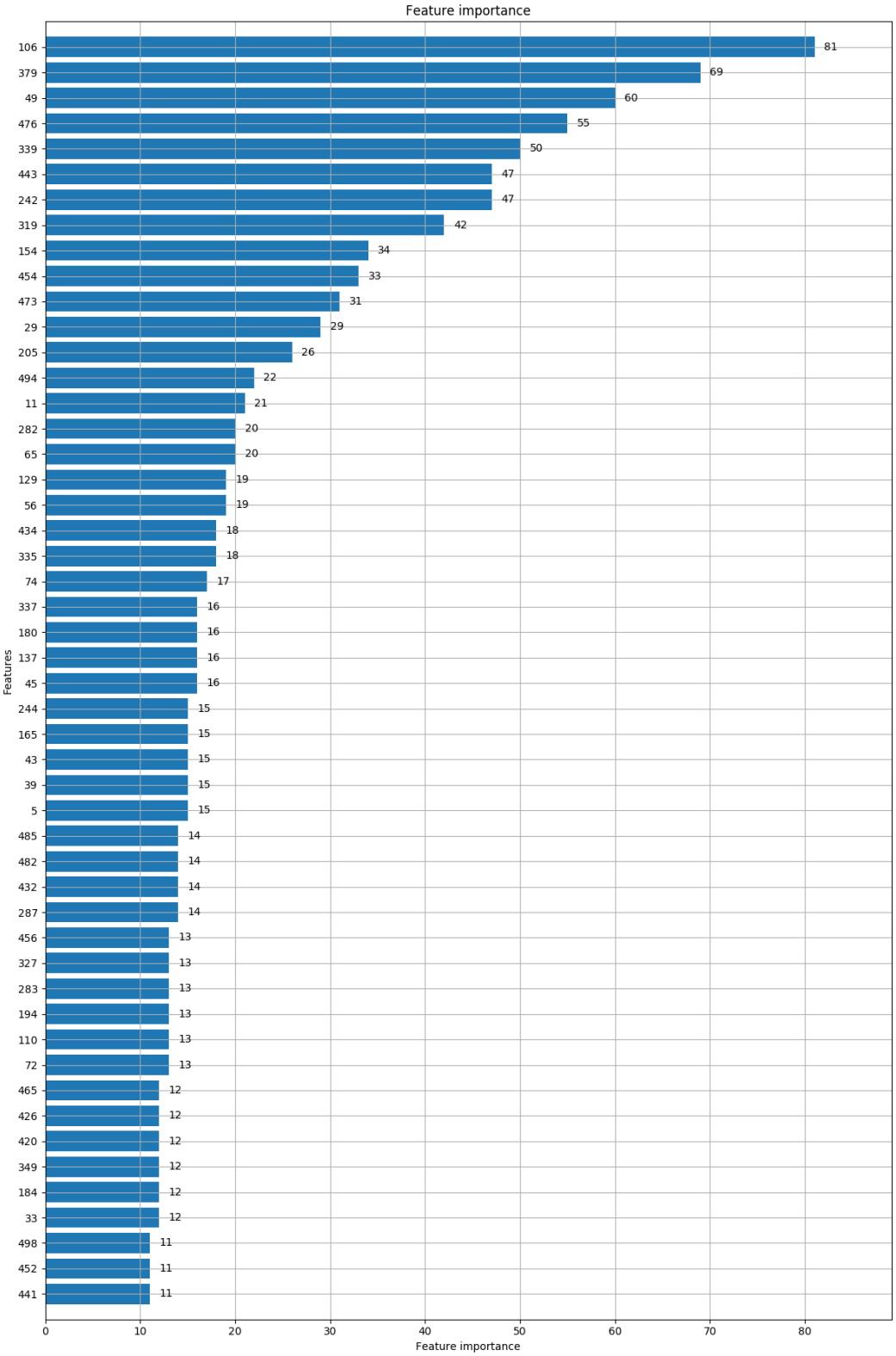
Rysunek 6: Rozkład cech z podziałem na zbiór

B Rozkład cech z podziałem na klasę



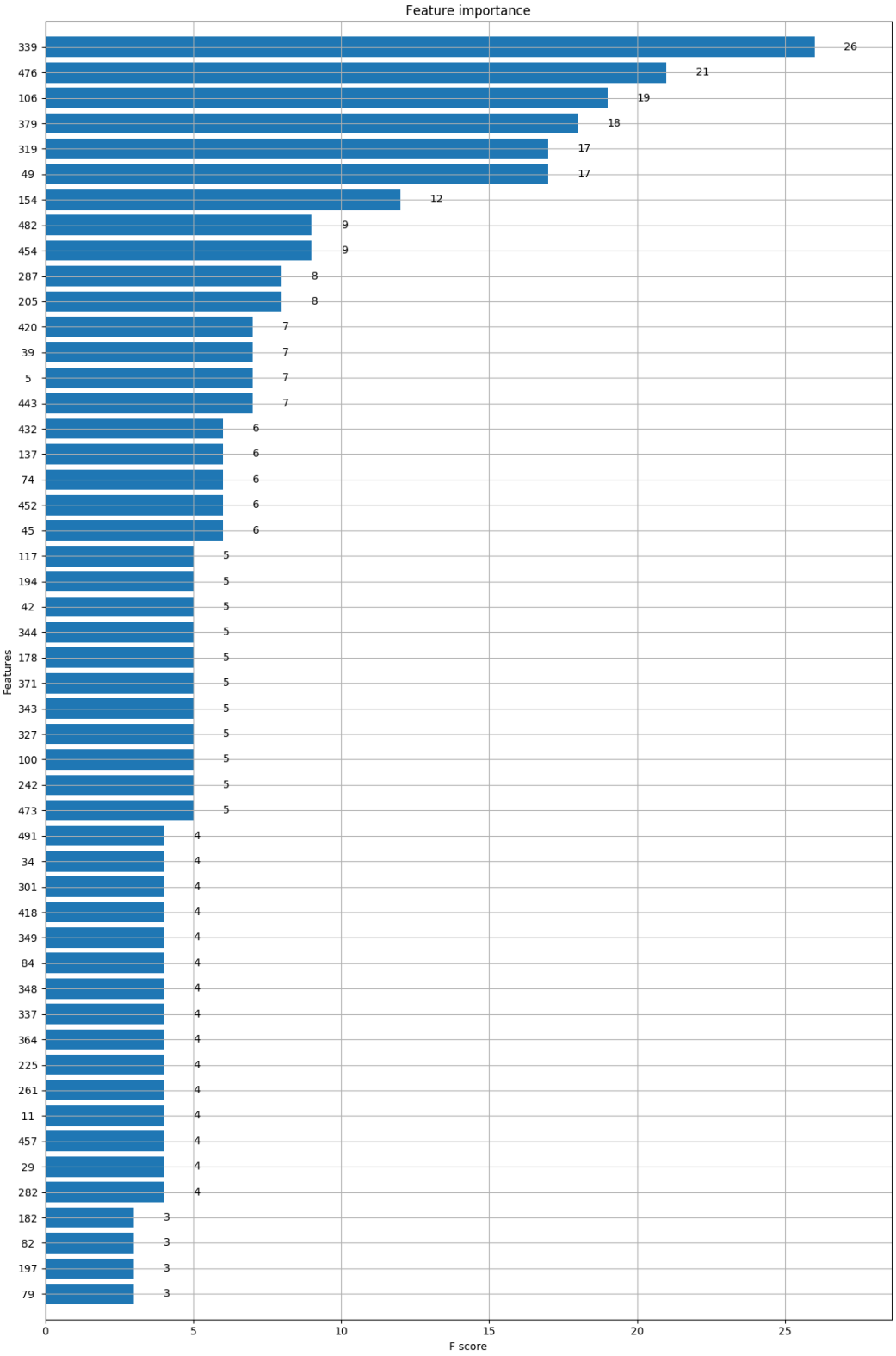
Rysunek 7: Rozkład cech z podziałem na klasę

C Istotność cech dla modelu LightGBM



Rysunek 8: Istotność najważniejszych 30 cech dla modelu LightGBM

D Istotność cech dla modelu LightGBM



Rysunek 9: Istotność najważniejszych 30 cech dla modelu XGBoost