Badanie skuteczności modelu w zależności od jego parametrów

Jan Pogłód

1. Wstęp

Celem pracy domowej było zbadanie jak poszczególne parametry drzewa decyzyjnego wpływają na jego dokładność oraz zdolność do predykcji. W tym celu przeprowadziłem poniższe kroki, analizując działanie każdego z badanych modeli, aby ostatecznie wybrać najlepszy i podsumować moje odkrycia. Głównymi parametrami drzewa decyzyjnego, jakie brałem pod uwagę były:

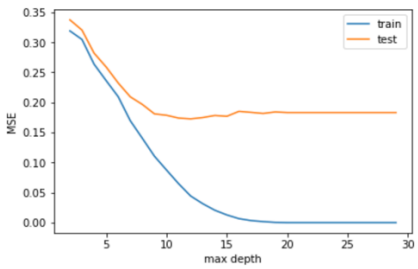
- *Criterion* – kryterium używane do podziałów węzłów. Analizowałem jaka jest różnica pomiędzy kryterium „Gini" i „Entropy".
- *Max_depth* – maksymalna głębokość, do której drzewo powinno dojść aby model osiągnął najlepszy wynik.
- *Min_samples_leaf* – minimalna liczba obserwacji w liściu.
- *Min_samples_split* – minimalna liczba obserwacji wymagana do podziału węzłów.
- *Max_features* – Maksymalna liczba cech, która jest brana pod uwagę przy podziale węzłów.
- *Ccp-alpha* – Parametr kontrolujący przycinanie drzewa
- *Random-state* ustawione na 320575 – Początkujący stan generatora

2. Kroki do osiągnięcia najlepszego modelu

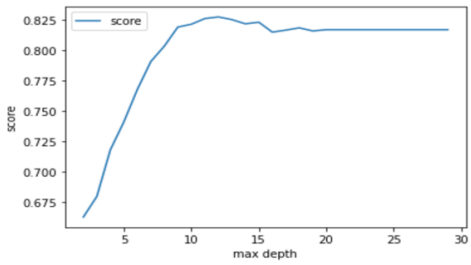
2.1 Osiągnięcie najlepszych parametrów

Po testach przeprowadzonych na danych zauważyłem, że kryterium podziału „gini" w przypadku badanych obserwacji ma lepszą jakość predykcyjną od entropii. Następnie zacząłem szukać najlepszej wartości dla parametru *max_depth*. W tym celu narysowałem wykresy zmiany błędu średniokwadratowego na zbiorach treningowym i testowym oraz zmiany dokładności modelu na zbiorze testowym.

Wykres 1. Zależność błędu średniokwadratowego od głębokości drzewa



Wykres 2. Zależność wartości dokładności głębokości drzewa

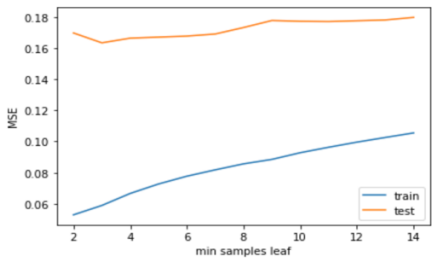


W ten sposób dokładny parametr optymalnej głębokości drzewa wyniósł *max_depth = 12*. Na tak utworzonym drzewie decyzyjnym przeprowadziłem 6-krotną walidację krzyżową (z 6 podziałami) otrzymując następujące wartości średniej oraz odchylenia standardowego z wyników dokładności modelu:

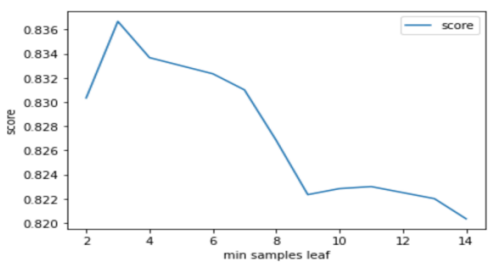
- Wartość średniej = 0.832
- Odchylenie standardowe = 0.0048

Następnie dla tak dobranych parametrów postępowałem analogicznie wyznaczając nowy parametr *min_samples_leaf*. Wyniki z przeprowadzonego eksperymentu prezentują się następująco.

Wykres 3. Zależność błędu średniokwadratowego od minimalnej liczby obserwacji w liściu



Wykres 4. Zależność wartości dokładności od minimalnej liczby obserwacji w liściu



W szczególności na wykresie 4 odznacza się jak przydatny jest parametr *min_samples_leaf* w doborze idealnego drzewa decyzyjnego do jakości predykcji danych. Można dostrzec na tym wykresie punkt górowania w punkcie 3, co oznacza, że parametr *min_samples_leaf = 3* przysłużył mi do uzyskania najlepszego drzewa.

2.2 Wyniki najlepszego drzewa decyzyjnego

Skuteczność drzewa decyzyjnego dla parametrów *max_depth=12, min_samples_leaf=3, criterion="gini"* oceniłem na podstawie jego dokładności, czułości, precyzji, krzywej ROC, wartości AUC oraz macierzy pomyłek. Oto wyniki eksperymentu:

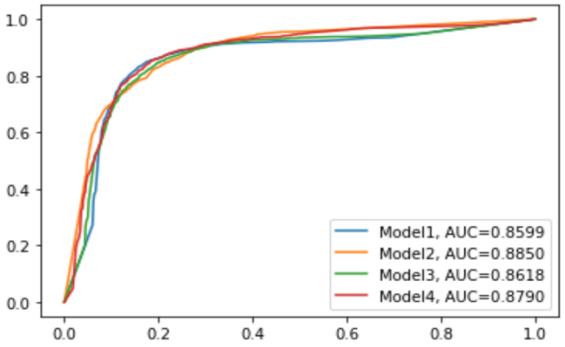
- Dokładność = 83,78%
- Czułość = 84,2%
- Precyzja = 83,42%
- AUC = 85,99% (jako stosunek pola pod wykresem)
- Macierz pomyłek: [[2484 504] [476 2536]]

2.3 Porównanie z innymi modelami, branymi pod uwagę w eksperymencie

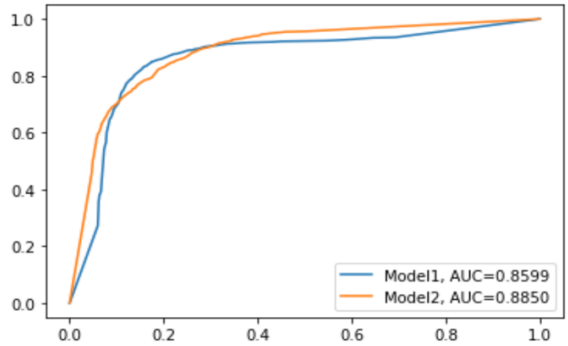
Dalej badałem również kryterium entropię, nie dodając innych parametrów prócz tych z powyższego modelu, ale zmieniając ich wartości na bardziej optymalne. Dostałem niższe wartości powyższych miar, ale otrzymałem wartość AUC równą 88,5%, co okazało się być najwyższą ze wszystkich modeli.

W dalszych krokach za pomocą 6-krotnej walidacji krzyżowej starałem się zwiększyć dokładność mojego najlepszego drzewa decyzyjnego dodając takie parametry jak *min_samples_split*, *max_features* lub *cppalpha* porównując wszystkie wartości dokładności w tych modelach. Ku mojemu rozczarowaniu parametry te jednak obniżały poziom dokładności, czułości i precyzji, natomiast zwiększała się wartość pod krzywą ROC. Niemniej najlepsze wyniki dokładności, precyzji i czułości jakie otrzymałem przy innych modelach to odpowiednio 82,13%, 82,44% i 82,6%. Te wartości wskazują również na fakt, że w najlepszej jakości macierz pomyłek wystąpiła w moim pierwszym modelu. Tam zaobserwałem także najwięcej wartości *True Positive i False Negative*. Również dodanie parametru *ccp_alpha* nie wpłynęło na poprawę modelu, ale ponownie powiększyła się wartość pola pod wykresem krzywej ROC. Najciekawsze i najlepszej jakości modele które stworzyłem zobrazowałem na poniższych krzywych ROC.

Wykres 5. Wszystkie modele



Wykres 6. Model1 i Model2



Powyższe modele zostały stworzone przez dodawanie i optymalizowanie parametrów:
Model 1: *max_depth, min_samples_leaf, criterion = „gini"*
Model 2: *max_depth, min_samples_leaf, criterion = „entropy"*
Model 3: *max_depth, min_samples_leaf, criterion = „gini", min_samples_split, max_features*
Model 4: *max_depth, min_samples_leaf, criterion = „gini", ccpalpha*

3. Podsumowanie

Drzewo decyzyjne, które otrzymałem i przedstawiłem w sekcji *2.2 wyniki modelu* uważam za najlepsze ze wszystkich testowanych podczas eksperymentu, ponieważ przeważało ono w każdej z miar: dokładność, precyzja, czułość oraz w wyniku dało najlepszą macierz pomyłek. Zostało ono stworzone na bazie parametrów criterion, max_depth i min_samples_leaf, których wpływ na jakość drzewa jest znaczący, co widać na wykresach 1-4. Eksperymentowanie z innymi parametrami dało mi osiągnąć wyższą jakość wartości AUC, ale dokładność modelu była niższa. Moje wyniki potwierdzają, że dobór optymalnych parametrów jest bardzo istotny w doborze drzewa decyzyjnego o najlepszej jakości predykcji.