Praca Domowa 1

Weronika Orzechowska

1 Wstępna analiza danych

Pracę z danymi rozpoczęłam od wstępnej analizy, aby wykryć potencjalne problemy, takie jak braki danych, nierównomierny rozkład klas czy skośność rozkładów zmiennych.

W zbiorze X nie występowały braki danych, natomiast w zbiorze Y były obecne. Spośród 949 obserwacji, tylko trzy nie miały przypisanej klasy, co nie stanowiło dużej straty, więc zostały one usunięte.
Dodatkowo, rozkłady zmiennych w zbiorze X przypominają rozkład normalny i nie wykazują skośności, dlatego nie było potrzeby przeprowadzania transformacji logarytmicznych. Rozkład klas również jest zrównoważony, co pozwala na stosowanie kroswalidacji bez ryzyka wystąpienia problemów z klasyfikacją.

2 Eksperyment

Eksperyment podzieliłam na dwie części. Pierwsza część polegała na ustaleniu odgórnych ograniczeń podczas budowy drzewa decyzyjnego. Druga część skupiła się na przycinaniu drzewa.

2.1 Kryterium podziału

Zbadałam dwa kryteria podziału: gini oraz entropy.

Kryterium entropy uzyskało wyższą średnią AUC (0.8881) niż kryterium gini (0.8685), co wskazuje, że na zbiorze treningowym model z kryterium "entropy" lepiej klasyfikuje dane. Odchylenie standardowe było w obu przypadkach zbliżone, co sugeruje podobną stabilność modeli. Kryterium entropy osiągnęło również wyższą wartość AUC na zbiorze testowym (0.8592) w porównaniu do gini (0.8521), co sugeruje, że model z tym kryterium lepiej przewiduje na nowych danych.

Wyniki wskazują, że kryterium entropy jest lepsze, ponieważ uzyskało wyższe AUC zarówno na zbiorze treningowym, jak i testowym.

2.2 Głębokość drzewa

Badanie głębokości drzewa rozpoczęłam od wartości 5. Ponieważ w zbiorze X jest 8 zmiennych, odrzuciłam głębokości równe 3 i 4 jako zbyt niskie do analizy.

Przy małych głębokościach drzewa (do 10) model osiąga wyższe wyniki AUC na zbiorach treningowym i testowym, ale na zbiorze treningowym widoczne są większe wahania (odchylenie standardowe). Spadek wartości AUC na obu zbiorach przy niższych głębokościach wskazuje na możliwość niedouczenia modelu.
Przy większych głębokościach drzewa (powyżej 12) AUC na zbiorze treningowym pozostaje stabilne, co sugeruje, że drzewo dobrze dopasowuje się do danych treningowych. Jednakże stabilizacja wartości AUC na poziomie około 0.85 na zbiorze testowym może oznaczać, że model zaczyna być podatny na przeuczenie. Mimo to, odchylenie standardowe maleje przy większych głębokościach, co oznacza większą stabilność modelu, choć nie przynosi to dalszej poprawy na zbiorze testowym.

Wartości AUC wskazują, że optymalna głębokość drzewa znajduje się w przedziale 8-10, gdzie wyniki na zbiorze testowym są najwyższe. Dalsze zwiększanie głębokości nie poprawia wyników, co oznacza, że optymalna głębokość dla tego modelu wynosi około 8-10. Przeuczenie modelu pojawia się po przekroczeniu głębokości 10, co jest widoczne w stabilizacji i wahaniach AUC na zbiorze testowym. Zwiększanie głębokości drzewa po tym punkcie nie przynosi dodatkowej poprawy.

2.3 Minimalna liczba obserwacji w liściu

Odchylenie standardowe AUC na zbiorze treningowym maleje wraz ze wzrostem minimalnej liczby obserwacji w liściu od około 5 do 20. Największe odchylenie standardowe występuje przy niskich wartościach (5-10), co może oznaczać większą niestabilność modelu. Dla wartości powyżej 20 odchylenie standardowe stabilizuje się, sugerując bardziej stabilny model.
Wartość AUC na zbiorze treningowym początkowo rośnie i osiąga maksymalne wartości dla przedziału 15-20, po czym zaczyna maleć. Przy tych wartościach model uzyskuje najlepsze wartości na zbiorze treningowym. AUC na zbiorze testowym wzrasta dla minimalnej liczby obserwacji w liściu do około 10, a potem oscyluje bez wyraźnej poprawy. Dla wartości 50 obserwacji w liściu, AUC na zbiorze testowym gwałtownie spada, co sugeruje, że zbyt duża liczba obserwacji w liściu prowadzi do niedouczenia.

Optymalna minimalna liczba obserwacji w liściu wynosi około 10-15, gdzie zarówno na zbiorze treningowym, jak i testowym AUC osiąga najwyższe wartości. Wzrost wartości tego parametru powyżej 50 prowadzi do znacznego pogorszenia wyników na zbiorze testowym, co oznacza uproszczenie modelu.

2.4 Dodatkowe parametry

Do przeprowadzonej analizy włączyłam jeszcze dwa parametry: strategia dzielenia węzła (splitter) oraz minimalny spadek niepewności (min impurity decrease).
Parametr splitter w opcji random, mimo nieco wyższego AUC na zbiorze testowym, pokazuje większą niestabilność na zbiorze treningowym, co sugeruje, że model nie jest wystarczająco przewidywalny. Z kolei minimalny spadek niepewności nie wnosi znaczącej poprawy w wynikach AUC, co czyni go mało wartościowym w dalszej optymalizacji.

2.5 Połączenie części 1-3

Na podstawie analizy w podrozdziałach 2.1, 2.2 i 2.3 wybrałam zakresy współczynników maksymalnej głębokości i minimalnej liczby obserwacji w liściu, które wykazały najlepsze wyniki.

Najwyższe AUC na zbiorze testowym występuje w wierszach 35 i 50, gdzie:

- Kryterium: entropy
- Głębokość: 8 / 9
- Liczba liści: 15

Te parametry oferują najlepsze wyniki, ponieważ AUC na zbiorze testowym jest najwyższe, a odchylenie standardowe na zbiorze treningowym niskie, co wskazuje na stabilność modelu. Różnica między AUC na zbiorze treningowym a testowym jest niewielka, co sugeruje dobrą generalizację modelu.
Zatem optymalny zestaw parametrów to kryterium: entropy, głębokość: 8 oraz liczba liści: 15.

2.6 Wczesne przycięcie drzewa

W dodatkowym eksperymencie skupiłam się na przycinaniu drzewa za pomocą współczynnika ccp_alpha. Przeanalizowałam wartości tego współczynnika dla drzew budowanych przy użyciu kryteriów gini i entropy.

Pierwszym badanym modelem był ten z kryterium gini. Odchylenie standardowe pozostaje stabilne i niskie dla większości wartości ccp_alpha, ale gwałtownie rośnie dla końcowych wartości. Najniższe odchylenie standardowe obserwowane jest dla parametru o indeksie 29, co sugeruje, że model jest najbardziej stabilny w tym punkcie.
AUC na zbiorze testowym osiąga swoje maksimum dla indeksów 25-30. Po przekroczeniu indeksu 30 następuje gwałtowny spadek AUC zarówno na zbiorze treningowym, jak i testowym, co wskazuje, że model staje się zbyt uproszczony.
Optymalny punkt to ccp_alpha o indeksie 29, gdzie AUC na zbiorze testowym jest najwyższe, a model najlepiej przewiduje. Niskie odchylenie standardowe w tym punkcie wskazuje na stabilność modelu.

Przeprowadzając analogiczną analizę z kryterium entropy, otrzymujemy, że współczynnik ccp alphas o indeksie 17 jest najbardziej optymalny.

3 Wybór modelu i analiza jakości predykcyjnej

Z eksperymentu otrzymałam trzy optyamlne modele:

- MODEL 1: kryterium: entropy, głębokość: 8 / 9, liczba liści: 15
- MODEL 2: kryterium: gini, ccp_alphas = 0.01153766
- MODEL 3: kryterium entropy, ccp_alphas = 0.01164323

Dla takich współczyyników otrzymałam następujące wartości miaru AUC:

- MODEL 1: 0.936074
- MODEL 2: 0.9165096
- MODEL 3: 0.9165096

Na tej podstawie, wybrałam model 1, którego wartość parametru AUC jest największa.

3.1 Macierz pomyłek

Otrzymałam następująca macierz pomyłek dla zbioru testowego: $\begin{bmatrix} 131 & 11 \\ 31 & 111 \end{bmatrix}$ oraz $\begin{bmatrix} 317 & 16 \\ 32 & 297 \end{bmatrix}$ dla zbioru treningowego.

Na jej podstawie można stwierdzić, że model jest bardziej dokładny w klasyfikowaniu przypadków klasy 1 (stosunkowo mniej fałszywych pozytywnych i fałszywych negatywnych przypadków), natomiast ma nieco większe problemy z poprawnym rozpoznawaniem wszystkich przypadków klasy 0 (31 fałszywych negatywów).

3.2 Dokładność, czułość, precyzja

Model otrzymał następujące wartości:

- ACC_test = 0.8521, ACC_train = 0.9275
- recall_test = 0.78169, recall_train = 0.9027
- precision_test = 0.90984, precision_train = 0.9489

Dokładność (accuracy) to odsetek poprawnych przewidywań (zarówno dla klasy 0, jak i 1) względem wszystkich przewidywań. Oba rodzaje pomyłek traktujemy równo. Wartość 0.8521 oznacza, że model poprawnie sklasyfikował 85.21% przypadków.
Recall mówi o tym, jak czuły jest model w wykrywaniu pozytywów. 0.78169 oznacza, że model poprawnie wykrył 78.17% wszystkich rzeczywistych przypadków klasy 1.
Precision (precyzja) to miara dokładności modelu w przewidywaniu klasy 1. Precyzja mówi nam, jaka część przypadków, które model zaklasyfikował jako 1, faktycznie należy do klasy 1. 0.90984 oznacza, że model przewidział klasę 1 z dokładnością 90.98%.

Na podstawie powyższych można stwierdzić, że model cechuje się wysoką precyzją w klasyfikacji klasy 1 (nie myli jej z klasą 0), ale ma niższy recall, co oznacza, że model nie wykrywa wszystkich przypadków klasy 1. Może to sugerować, że model bardziej stara się unikać fałszywych pozytywów (błędne przypisanie klasy 0 do klasy 1) niż fałszywych negatywów. W przypadku problemu, w którym ważniejsze jest minimalizowanie liczby pominiętych przypadków klasy 1, należałoby zwiększyć recall, ale kosztem precyzji.

3.3 Krzywa ROC, wartość AUC

Krzywa ROC jest blisko lewego górnego rogu wykresu, co wskazuje na dobry balans między TPR a FPR. Im bardziej krzywa jest "wyginana" w stronę górnego rogu, tym lepiej model rozpoznaje klasy.
AUC na poziomie 0.9361 oznacza, że model jest bardzo skuteczny w klasyfikacji, co jest potwierdzeniem dobrej jakości przewidywań. Przy wartości bliskiej 1, model ma wysoką zdolność do rozróżniania pozytywnych i negatywnych przypadków.
Warto zauważyć, że model ma bardzo niski FPR na początku krzywej, co oznacza, że popełnia niewiele błędów przy klasyfikacji przypadków negatywnych jako pozytywnych.
Krzywa ROC potwierdza wysoką jakość modelu, ponieważ jest blisko ideału, który znajduje się w lewym górnym rogu wykresu

4 Wpływ próbki danych na jakość predykcyjną modelu

Eksperyment przeprowadziłam na dwa sposoby - z wykorzystaniem kroswalidacji i bez. Dzięki temu mogłam dodatkowo zaobserwować wpływ losowego podziału danych na wyniki modelu.

4.1 Bez użycia kroswalidacji

Z przedstawionego wykresu wynika, że wpływ rozmiaru próbki danych na jakość predykcyjną modelu jest zauważalny zarówno na zbiorze treningowym, jak i testowym, ale brak zastosowania kroswalidacji powoduje, że te wyniki mogą być mniej stabilne, szczególnie przy mniejszych rozmiarach próbek.
Zacznę od przeanalizowania sytuacji na zbiorze treningowym.
Zwiększanie rozmiaru próbki prowadzi do systematycznego wzrostu wartości AUC. Model lepiej dopasowuje się do danych wraz ze wzrostem ich ilośc. Wartości AUC stabilizują się przy większych próbkach (powyżej 50%), co sugeruje, że model osiąga dobre dopasowanie już przy tej wielkości danych. Jednakże brak kroswalidacji powoduje, że te wyniki mogą być nadmiernie optymistyczne, ponieważ model jest oceniany na tych samych danych, na których został wytrenowany. Kroswalidacja mogłaby dać bardziej wiarygodny obraz zdolności predykcyjnych modelu.
Wartości AUC na zbiorze testowym (niebieska linia) mają zmienny charakter, szczególnie przy mniejszych rozmiarach próbek. Zwiększanie próbki danych prowadzi do wzrostu AUC, ale widać większe wahania przy mniejszych próbkach, takich jak 5% czy 10%. AUC stabilizuje się przy większych próbkach, co oznacza, że model lepiej generalizuje na danych testowych przy wykorzystaniu większej ilości danych.
Kroswalidacja (np. 10-krotna) pozwoliłaby na bardziej stabilne wyniki, gdyż zapewnia wielokrotny podział danych na zbiory treningowe i testowe. Dzięki temu zmniejszylibyśmy wpływ losowości na wyniki.

Rozmiar próbki ma wyraźny wpływ na jakość predykcyjną modelu. Model potrzebuje co najmniej 25-50% dostępnych danych, aby stabilnie i dobrze generalizować, a większe próbki prowadzą do bardziej stabilnych wyników.

4.2 Z użyciem kroswalidacji

Wpływ rozmiaru próbki danych na jakość predykcyjną modelu jest zauważalny zarówno na zbiorze treningowym, jak i testowym. Dzięki zastosowaniu kroswalidacji wyniki te są bardziej stabilne, szczególnie przy mniejszych rozmiarach próbek.

Zwiększanie rozmiaru próbki prowadzi do systematycznego wzrostu wartości AUC na zbiorze treningowym. Model lepiej dopasowuje się do danych wraz ze wzrostem ilości danych treningowych. Wartości AUC stabilizują się przy większych próbkach (powyżej 50%), co sugeruje, że model osiąga dobre dopasowanie już przy tej wielkości danych.
Wartości AUC na zbiorze testowym rosną wraz z wielkością próbek i stają się bardziej stabilne przy większych próbkach (powyżej 50%). Dzięki kroswalidacji widać, że model osiąga przewidywalne wyniki, niezależnie od wielkości próbki. Wahania AUC na zbiorze testowym, które widzieliśmy przy mniejszych próbkach (np. 5% i 10%), są teraz znacznie zredukowane dzięki kroswalidacji, co sugeruje, że model nie jest zbyt wrażliwy na przypadkowe wybory danych. Stabilizacja wyników AUC na poziomie około 0.90 sugeruje, że model dobrze generalizuje, a kroswalidacja pozwala na bardziej ujednoliconą ocenę jego wydajności.