Przyjrzyjmy się lasom losowym nieco dokładniej, na konkretnym przykładzie.

STUDIUM PRZYPADKU BCG – wyszukiwanie najlepszych lokalizacji pod nowe oddziały banku

Algorytm lasów losowych świetnie nadaje się do rozwiązywania złożonych problemów, z którymi drzewa decyzyjne nie radzą sobie najlepiej. Gdybyśmy na przykład, mieli znaleźć najlepsze lokalizacje pod nowo otwierane oddziały banku na podstawie wzorców ukrytych w wielu zmiennych, użylibyśmy lasu losowego.

Mieszkam w Australii i kiedy chcę zarejestrować się jako nowy klient banku, szukam takiego, który ma wygodnie zlokalizowane oddziały. Chciałbym, żeby oddział banku znajdował się blisko mojego domu, w pobliżu miejsca pracy i sklepów, w których robię zakupy. Jeśli bank ma też oddział blisko plaży, tym lepiej. Nie ma nic gorszego od konieczności jechania przez całe miasto tylko po to, żeby porozmawiać z konsultantem lub zrealizować czek.

Banki wiedzą, że dostępność ich oddziałów jest głównym powodem, dla którego potencjalni klienci je wybierają. Ale muszą też mieć pewność, że nowo otwarty oddział nie przyniesie im strat. Artem Vladimirov, czołowy konsultant ds. analityki w Boston Consulting Group (BCG), dostał od jednego z banków należących do grupy kapitałowej, który chciał rozwinąć swoją działalność na terenie całej Australii, takie zadanie do rozwiązania.

Artem najpierw przeanalizował dane demograficzne klientów banku, żeby poznać ich liczbę w poszczególnych okręgach mieszkalnych Australii. Zauważył, że ponieważ oddziały baku nie były równomiernie rozmieszczone w całym kraju, brakuje mu danych z niektórych okręgów. Żeby móc obliczyć zyskowność oddziałów otwartych w tych okręgach, Artem przeprowadził analizę porównawczą okręgów „znanych" i „nieznanych" bankowi, używając do tego celu ogólnie dostępnych, rządowych danych demograficznych. Na podstawie takich informacji demograficznych jak średni wiek, procent mężczyzn i kobiet, wykształcenie i koszt życia, Artem uzupełnił braki w danych. Pozwoliło mu to oszacować prawdopodobieństwo otworzenia zyskownego oddziału w nowym okręgu, którego mieszkańcy byli podobni do mieszkańców okręgu „znanego" bankowi.

Żeby rozwiązać postawiony przed nim problem Artem użył lasów losowych:

> Użyliśmy całej bazy klientów banku do nauczenia algorytmu lasów losowych związku między danymi demograficznymi klientów a ich rentownością. Predykcje zostały wykonane dla okręgów, w których istniały już oddziały banku, a więc pozostało nam jedynie sprawdzić, czy otwarcie oddziału w nowym okręgu przyniesie zyski przez porównanie danych demograficznych jego mieszkańców (SuperDataScience, 2016).

---

nie jest gorsza niż trafność losowego zgadywania. Algorytm lasów losowych uwzględnia oba te założenia. (przyp. tłum.).

Po zidentyfikowaniu interesujących dla banku okręgów Artem sprawdził dane na temat konkurencji i liczby oddziałów innych banków w poszczególnych okręgach. Raz jeszcze użył lasów losowych do oszacowania procentowego udziału w rynku, jaki bank może uzyskać w tych okręgach, otwierając w nich swoje oddziały.

Las losowy nie zawiera opisów zwracanych przez siebie predykcji. W tym przypadku Artem nie musiał wytłumaczyć, które zmienne demograficzne miały największy wpływ na otrzymane predykcje, co pozwoliło mu ominąć problem przetwarzania danych osobowych i precyzyjnie wskazać te okręgi, w których otwarcie oddziału będzie dla banku najkorzystniejsze.

Kroki budowania lasów losowych

1. **Wybierz liczbę drzew, które chcesz zbudować.** W wielu programach domyślną wartością tego hiperparametru będzie 10. Liczba, którą wybierzesz, zależy od kontekstu konkretnego zadania. Mniej drzew może oznaczać mniej dokładne predykcje. Jednocześnie możesz bezpiecznie zbudować tyle drzew, ile tylko chcesz, nie martwiąc się, że algorytm lasów losowych nadmiernie dopasuje się do danych treningowych[7].
2. **Naucz klasyfikator na podstawie danych treningowych.** Model nauczony na danych treningowych może być użyty do predykcji, czyli do uzupełnienia danych testowych o wartość zmiennej wyjściowej. Porównując wyniki predykcji z rzeczywistymi wartościami zmiennej wyjściowej, będziemy mogli zmierzyć jakoś (np. trafność) klasyfikatora.

Algorytm lasów losowych wylosuje ze zwracaniem z danych treningowych N podzbiorów, gdzie N jest liczbą drzew decyzyjnych wybranych w pierwszym kroku. Ponieważ używana jest metoda losowania ze zwracaniem, te same obserwacje mogą trafić do wielu podzbiorów, ale żaden z tych podzbiorów nie będzie taki sam jak pozostałe.

Po utworzeniu podzbiorów każdy z nich zostanie użyty do zbudowania osobnego drzewa decyzyjnego. W ten sposób poszczególne drzewa decyzyjne będą nauczone na osobnych podzbiorach danych treningowych, i nie będą znały wszystkich danych. To rozwiązanie pozwala zagwarantować różnorodność i niezależność drzew decyzyjnych – cechy, które dają lasom losowym „siłę tłumu".

[7] Zostało sprawdzone, że o ile tylko dysponujesz co najmniej kilkuset obserwacjami opisanymi przez co najmniej kilkanaście zmiennych, jakość modelu lasów losowych osiągnie maksimum dla mniej więcej 50 drzew decyzyjnych. Dalsze ich zwiększanie nie ma znaczącego wpływu na poprawę wyników, a wydłuża czas uczenia modelu (przyp. tłum.).

Wynika z tego, że do poprawienia dokładności predykcji wystarczy nauczyć algorytm lasów losowych na większym zbiorze danych – im więcej będzie zawierał przykładów, tym dokładniejsze będą predykcje zwracane przez algorytm.

Drzewa decyzyjne czy lasy losowe?

Chociaż algorytm lasów losowych można uznać za „nowszą wersję" drzew decyzyjnych, oba mają swoje zalety i wybór jednego z tych algorytmów zależy od postawionego zadania. Jeśli dysponujesz niewielkim zbiorem danych, użycie lasów losowych może dać nieoptymalne wyniki, bo algorytm niepotrzebne podzieli Twoje dane na podzbiory. W takim przypadku użyj drzew decyzyjnych, które nie tylko są szybsze, ale również pozwalają w prosty sposób interpretować zwracane predykcje. Ale jeśli pracujesz z dużym zbiorem danych, lasy losowe dadzą dokładniejsze, choć trudniejsze w interpretacji wyniki[8].

Algorytm k najbliższych sąsiadów

Algorytm k najbliższych sąsiadów (k-NN) używa wykrytych w danych wzorców do przypisania nowych obserwacji do właściwych kategorii. Powiedzmy, że lekarka z San Francisco przeczytała o rosnącej liczbie zachorowań na cukrzycę w Stanach Zjednoczonych i chce zapobiec epidemii wśród swoich pacjentów. Pani doktor wie, że cukrzycy typu 2 łatwiej jest zapobiegać, niż leczyć. Zwróciła się do specjalisty danych z prośbą o zbudowanie modelu, który na podstawie danych jej pacjentów, z których część choruje na cukrzycę, określi prawdopodobieństwo, z jakim ta choroba może w przyszłości dotknąć nowych pacjentów.

Pani doktor liczy, że na podstawie predykcji modelu będzie mogła wcześniej zidentyfikować pacjentów z grupy podwyższonego ryzyka i pomóc im zachować zdrowie dzięki badaniom profilaktycznym i konsultacjom dotyczącym zdrowego trybu życia. Z doświadczenia wie, że dwoma czynnikami mającymi istotny wpływ na zapobieganie chorobie są liczba ćwiczeń tygodniowo i waga. Zadaniem specjalisty danych jest teraz zbudowanie modelu, który wiarygodnie klasyfikowałby pacjentów jako należących do grupy podwyższonego ryzyka.

Czego możesz się spodziewać po algorytmie k najbliższych sąsiadów?

Algorytm k-NN analizuje „podobieństwo". Jego działanie polega na obliczeniu odległości między nową obserwacją a przykładami treningowymi. Ponieważ w tym

[8] Ponieważ końcowy wynik jest obliczany na podstawie predykcji zwróconych przez poszczególne drzewa decyzyjne, prześledzenie reguł, na podstawie których został on zwrócony, będzie bardzo trudne.