Efekt 8, Dodatkowa analiza

Mateusz Nowak, Damian Okoń, Robert Zamiar

# Pytanie badawcze

Jak modele klasyfikacyjne radzą sobie z rozpoznawaniem krajów pod względem posiadanej waluty na podstawie danych makroekonomicznych?

# Wstęp do analizy

Do budowy modeli jako zmienną celu wykorzystano predyktor binarny – is\_euro\_currency. Zmienna przyjmuje wartości 0 – dla krajów nie będących w systemie walutowym euro oraz 1 – dla krajów posiadających walutę euro jako walutę główną.

# Podstawowe modele klasyfikacyjne

Wstępną klasyfikację wykonano na podstawie modelu k-najbliższych sąsiadów. Model ten wykorzystuje bliskość sklasyfikowanych już rekordów na podstawie zadanych predyktorów – odległości w przestrzeni przez nie wyznaczonej.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 1. Model k-najbliższych sąsiadów

Model wykazuje dokładność na poziomie 88%, najtrafniej przewiduje przypadki pozytywny (sensitivity), natomiast w przypadku przypadków negatywnych model pomylił się trzykrotnie (specitifity).

Jednym z podstawowych modeli klasyfikacji jest również model “Naiwnego Bayesa”. Do predykcji zmiennej celu wykorzystuje on prawdopodobieństwo warunkowe.

**Obraz zawierający tekst, paragon, zrzut ekranu

Opis wygenerowany automatycznie**

Rysunek 2. Model Naive-Bayes

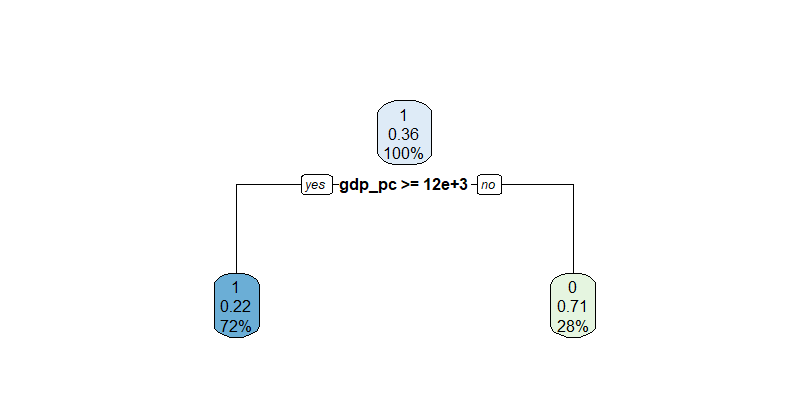
Model ten wypadł gorzej niż poprzedni. Jednym z powodów gorszej predykcji może być fakt, że wartości predyktorów użytych do klasyfikacji nie są jednorodne. *Naive Bayes* jest używany zazwyczaj do predykcji danych tekstowych i kategorycznych, gorzej sprawdza się przy danych niejednorodnych jak na przykład dane makroekonomiczne. Z kolei drzewa klasyfikacyjne i regresyjne są formą drzew decyzyjnych, wykorzystujące predyktory do jak najlepszego podziału obserwacji ze względu na zmienną celu.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 3. Model - drzewo klasyfikacyjne

Predykcje modelu są gorsze niż w poprzednich przypadkach. Ma to związek z podziałem drzewa według zmiennej GDP per capita, która różni zdecydowanie w poszczególnych rekordach. Na łącznie sześć nietrafnych predykcji mogą więc mieć wpływ obserwacje odstające. Model przewidział nietrafnie 4 przypadki negatywne (zakwalifikował je jako pozytywne), a dwa pozytywne zakwalifikował jako negatywne.



Rysunek 4. Drzewo klasyfikacyjne - graficznie

Drzewo decyzyjne należy rozumieć następująco. Jeżeli PKB per capita jest większe niż 12 tys. Euro, to z prawdopodobieństwem 72% rekord (kraj) posiada euro jako walutę, jeżeli wartość PKB per capita jest mniejsza niż 12 tys. Euro, to prawdopodobieństwo to ma wartość 28%.

# Modele usprawniające klasyfikację

Modelami usprawniającymi klasyfikację są modele lasu losowego oraz boosting. Las losowy bazuje na baggingu – losuje dane ze zwracaniem tworząc drzewa decyzyjne. Oprócz próbki danych las losowy za każdym razem losuje inny dobór predyktorów. Jego praca się kończy, gdy utworzy zadaną liczbę drzew, a stworzone przez algorytm modele wskażą większościa głosów najlepsze rozwiązanie. Boosting z kolei to uczenie modelu na próbce danych. Działa on iteracyjnie, za każdym razem “poprawia” wcześniejsze błędy – przywiązuje do nich większe wagi.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 5. Wyniku boostingu

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 6.Wyniki lasu losowego

Z powodu małej próbki danych oba algorytmy przedstawiają te same wyniki – wyszły jednak gorzej niż k-najbliższych sąsiadów. Może być to spowodowane tzw. “przetrenowaniem” modelu.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 7. Porównanie modeli k-ns i boostingu

Modele te mylą się w przypadku Czech, czego powodem może być wyższe od średniej PKB per capita. K-najbliższych sąsiadów myli się w przypadku Węgier i Rumunii, co także jest spowodowane bliskością niektórych predyktorów w przestrzeni, kraje te zostały niepoprawnie sklasyfikowane jako należące do strefy walutowej euro. Boosting z kolei pomylił się w przypadku krajów posiadających euro – Litwy i Królestwa Niderlandów, które zostały sklasyfikowane niepoprawnie jako kraje nieposiadające europejskiej waluty jako środka płatniczego.

# Odpowiedź na pytanie badawcze

Wykonane analizy w większości przypadków odzwierciedlają rzeczywistość, wykorzystane modele w większości przypadków trafnie klasyfikują kraje ze względu na przynależność do strefy euro, jednakże w niektórych przypadkach należałoby użyć innego doboru predyktorów.