„Analiza porównawcza wybranych algorytmów data mining”

Autor pracy: mgr inż. Łukasz Smyła
Opiekun pracy: dr inż. Paweł Dymora

Cel i zakres pracy

Głównym celem pracy było zbadanie skuteczności dziesięciu wybranych algorytmów data mining w kontekście klasyfikacji i regresji oraz porównanie ich efektywności między dwoma popularnymi językami programowania używanymi w data science: Pythonem i R

Badanie miało na celu ocenę, jak różne algorytmy radzą sobie w zadaniach klasyfikacji i regresji w kontekście konkretnego problemu w tym wypadku wykrywaniu oszustw. Skuteczność algorytmów zostanie oceniona na podstawie kluczowych metryk, takich jak dokładność, czas wykonania, różnica pomiędzy najlepszym i najgorszym wynikiem oraz wybrane algorytmy Regresji Liniowej zostały przebadane pod kątem błędu średniokwadratowego (MSE).

Porównanie będzie przeprowadzone zarówno w ramach porównania poszczególnych algorytmów, jak i między językami Python i R, co pozwoli na ocenę ich wydajności, łatwości implementacji oraz jakości wyników.

Co to data mining?

Data mining, czyli eksploracja danych, to proces wydobywania użytecznych informacji i wzorców z dużych zbiorów danych. Polega na zastosowaniu metod statystycznych, uczenia maszynowego oraz systemów baz danych w celu odkrywania wzorców, korelacji, trendów i anomalii w danych.
Celem jest przekształcenie surowych danych w wartościowe i praktyczne informacje, które mogą wspierać procesy decyzyjne i strategię biznesową. Proces ten obejmuje zbieranie i przygotowanie danych, ich eksplorację, modelowanie, ocenę wyników oraz wdrożenie odkrytej wiedzy do praktyki.

Data mining znajduje zastosowanie w wielu dziedzinach, takich jak marketing, gdzie pomaga w segmentacji klientów i personalizacji ofert, w finansach do wykrywania oszustw i analizy ryzyka, w opiece zdrowotnej do diagnozowania chorób i optymalizacji leczenia, oraz w handlu detalicznym do analizy zachowań zakupowych i zarządzania zapasami.

Przyszłość data mining zapowiada się obiecująco dzięki postępom w technologii, w tym sztucznej inteligencji i uczeniu maszynowemu. Te innowacje pozwolą na jeszcze bardziej precyzyjne analizy, automatyzację procesów oraz odkrywanie bardziej złożonych wzorców, co zwiększy wartość biznesową i operacyjną danych.

Wybrane do badań algorytmy data mining

ALGORYTMY KLASYFIKACJI:

- Naiwny klasyfikator Bayesa
- K-Najbliższych Sąsiadów
- SVM
- Drzewo decyzyjne
- Las Losowy
- GBM

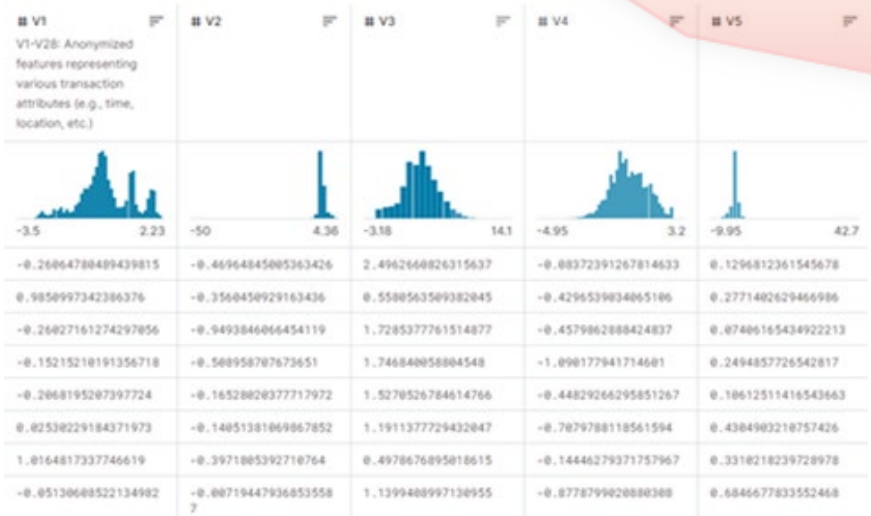
ALGORYTMY REGRESJI:

- Regresja Liniowa
- Regresja Logistyczna
- Regresja Grzbietowa
- Regresja LASSO

Opis użytych danych

Baza danych „Credit Card Fraud Detection Dataset 2023” użyta w tej pracy, zawiera informacje dotyczące transakcji kart kredytowych. Dane te zostały zebrane w celu analizy i eksploracji w dziedzinie uczenia maszynowego, a także do przeprowadzenia różnorodnych analiz dotyczących transakcji finansowych.

- Dane po anonimizacji i normalizacji
- Zawiera 568 630 rekordów
- V1-V28 – Atrybuty na podstawie, których jest dokonywana klasyfikacja
- Class - jest etykietą binarną wskazującą, czy transakcja jest fałszywa (1), czy niefałszywa (0)



Analiza porównawcza algorytmów data mining – porównanie algorytmów

ALGORYTM	SKUTECZNOŚĆ [%]
NAIWNY KLASYFIKATOR BAYESA	91,902
K-NAJBLIŻSZYCH SĄSIADÓW	99,754
SVM	99,180
DRZEWO DECYZYJNE	95,739
LAS LOSOWY	99,754
GBM	92,604
REGRESJA LOGISTYCZNA	96,472

Skuteczność liczona na podstawie poprawnie sklasyfikowanych przypadków. W porównaniu między algorytmami nie rozróżniany został język programowania dlatego każdy algorytm wykonany został 200 razy.

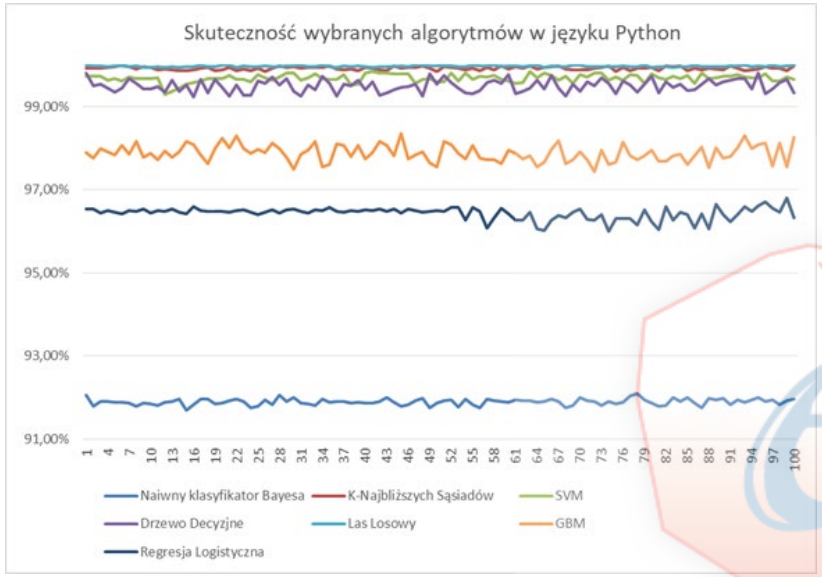
Podobnie jak w przypadku skuteczności, błąd średniokwadratowy został policzony na podstawie 200 prób wybranych algorytmów. Końcowe wyniki zostały wyliczone przy pomocy średniej arytmetycznej.

ALGORYTM	ŚREDNI MSE
REGRESJA LINIOWA	0,05907873
REGRESJA GRZBIETOWA	0,05941092
REGRESJA LASSO	0,154563266

ALGORYTM	SKUTECZNOŚĆ PYTHON [%]	SKUTECZNOŚĆ R [%]
NAIWNY KLASYFIKATOR BAYESA	91,893	91,912
K-NAJBLIŻSZYCH SĄSIADÓW	99,918	99,589
SVM	99,684	98,677
DRZEWO DECYZYJNE	99,498	91,979
LAS LOSOWY	99,969	99,539
GBM	97,875	87,332
REGRESJA LOGISTYCZNA	96,428	96,515

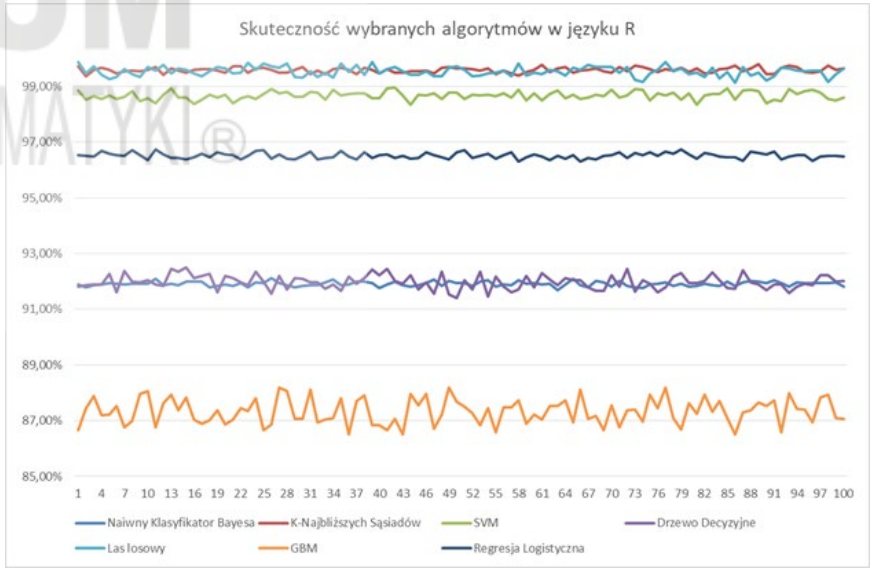
W porównaniu pomiędzy językami programowania, algorytmy zostały podzielone po 100 wykonań dla każdego języka, a następnie wyliczona została średnia arytmetyczna wyników dla Python oraz R.

ALGORYTM	MSE PYTHON	MSE R
REGRESJA LINIOWA	0,059040064	0,059117396
REGRESJA GRZBIETOWA	0,059047293	0,059774547
REGRESJA LASSO	0,250000999	0,059125534



vs

Oba wykresy przedstawiają jak wyglądała skuteczność algorytmów dla każdej ze 100 wykonanych prób. Taki wykres pokazuje jak bardzo wyniki algorytmów były zróżnicowane.



Podsumowanie

Samo pojęcie data mining, może mówić całkiem niewiele dla człowiek nie interesującym się tym tematem. Jednak w miarę zgłębiania się w ten temat okazuje się, że eksploracja danych towarzyszy nam wszędzie. Na podstawie tej pracy zostało pokazane, że eksploracja danych towarzyszy nam na co dzień mimo, że większość z nas nawet sobie nie zdaje z tego sprawy.

Po dokonaniu analizy okazało się, że najskuteczniejszymi algorytmami do wykrywania oszustw w mojej pracy były Las Losowy oraz K-Najbliższych Sąsiadów. Jeśli chodzi o dobór języka programowania, lepiej z zagadnieniem poradził sobie język Python.