Raport
Podstawy Uczenia Maszynowego
Projekt
Bankructwo

Paulina Reus
Piotr Wieczorek

Spis treści

1 Wstęp

Niniejsze badanie polegało na klasyfikacji bankructwa przedsiębiorstwa na podstawie jego danych finansowych. Zagadnienie ma charakter dychotomiczny, ponieważ spółka może przyjąć tylko dwie wartości (0 nie zbankrutowała, 1 zbankrutowała). Głównym celem badania jest opracowanie modelu, który skutecznie przewidzi upadłość przedsiębiorstwa.

Tego typu badania są niezwykle istotne dla inwestorów, którzy decydują się na wykup udziałów w danej firmie.

2 Eksploracja Danych

2.1 Zbiór danych

W modelu został wykorzystany zbiór danych zawierających takie informacje na temat spółek jak: wskaźnik rentowności aktywów przed odsetkami i amortyzacją, stopa zysku operacyjnego, dochody i wydatki/przychody pozaprzemysłowe, stosunek dochodu netto do aktywów ogółem, stosunek zysk brutto do sprzedaży i wiele innych oraz informacje na temat tego czy dany podmiot zbankrutował czy nie. Na podstawie tym informacji modele miały być trenowane w celu predykcji bankructwa danej firmy.

2.2 Charakterystyka zbioru danych

W zbiorze znajdowało się 20 brakujących wartości (najwięcej braków zawierała kolumna “Quick Asset Turnover Rate” było ich 8). Tak niewielka ilość braków pozwoliła na bezproblemowe usunięcie wierszy je zawierające.

Kolejnym ważnym aspektem na jaki warto było zwórcic uwagę w trakcie eksploracji danych byłu wartości odstające. Na istnienie takowych wartości wskazuje spora różnica pomiędzy średnią a maksymalną wartością, w danej kolumnie, tak jak w przypadku kolumny “ Research and development expense rate” 1.

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Non-industry income and expenditure/revenue	Operating Expense Rate	Research and development expense rate	Cash flow rate
count	498.000000	498.000000	498.000000	498.000000	498.000000	498.000000	498.000000	4.980000e+02	4.980000e+02	498.000000
mean	0.437751	0.459373	0.506949	0.601942	0.601963	0.998885	0.303116	2.408321e+09	1.158767e+09	0.464614
std	0.496609	0.068715	0.076315	0.012022	0.012007	0.000491	0.003419	3.491129e+09	1.965044e+09	0.010459
min	0.000000	0.024277	0.033514	0.532906	0.532906	0.991888	0.235090	1.011819e-04	0.000000e+00	0.343818
25%	0.000000	0.441878	0.492411	0.596704	0.596726	0.998891	0.303295	1.571500e-04	0.000000e+00	0.460881
50%	0.000000	0.476040	0.528535	0.601407	0.601364	0.998973	0.303466	3.591091e-04	2.175000e+08	0.463302
75%	1.000000	0.496539	0.548745	0.605985	0.605985	0.999029	0.303541	5.625000e+09	1.360000e+09	0.467335
max	1.000000	0.589139	0.632743	0.665151	0.665151	0.999254	0.305396	9.960000e+09	9.920000e+09	0.545963

Rysunek 1: Opis zawartości poszczególnych kolumn

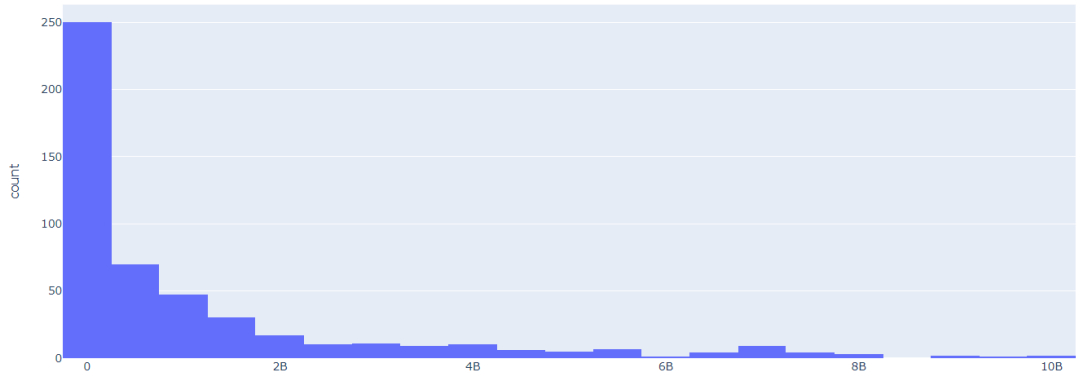
Występowanie wartości odstających widoczne jest również na wykresie prezentującym rozłożenie wartości zmiennej. Nagłe skoki/spadki wśód danych wskazują na wartości odstające 2.

Takie wartości zazwyczaj wynikają z nieschematycznych zdarzeń, których uwzględnienie w modelu mogłoby spowodować zmniejszenie jego efektywności. Aby zmniejszyć wartości odstające zostały one zastąpione przez średnia wartość. Wykres tej samej zmiennej po zastąpieniu wartości odstających został zaprezentowany poniżej 3.

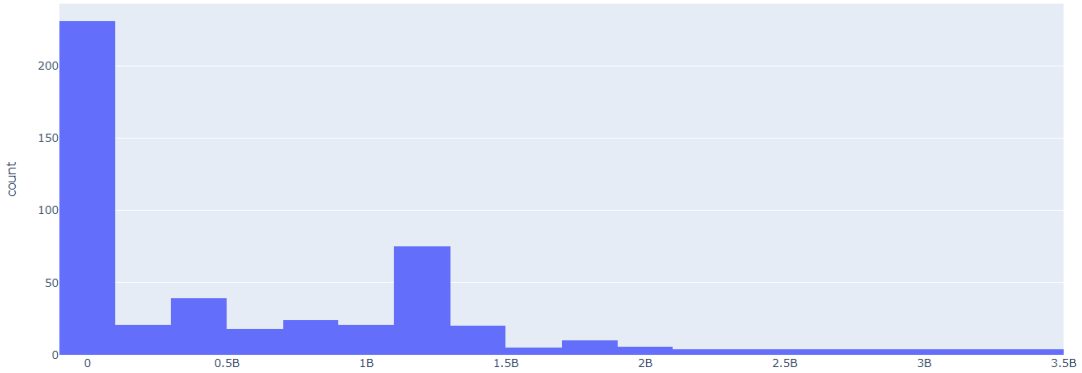
Mimo zastąpienia wartości odstających w zbiorze, wciaż wśród danych występują pewne skoki (wynikają z osiągnięcia maksimun przez pojedyńcze rekordy).

2.3 Rozkład normalny, kurtoza i skośność

Na podstawie przerpwadzonego testu Shapiro-Wilka udało się zbadać czy zmienne mają rozkład normalny. Niestety jedynie zmienna ’Working capitcal Turnover Rate’ okazała się mieć rozkład normalny. Również istotne było zbadanie skośności i kurtozy zmiennych. W przypadku większości zmiennych kutroza byłą dodatnia co wskazuję na, iż w danych istnieje więcej dodatnich wartości



Rysunek 2: Research and development expense rate”



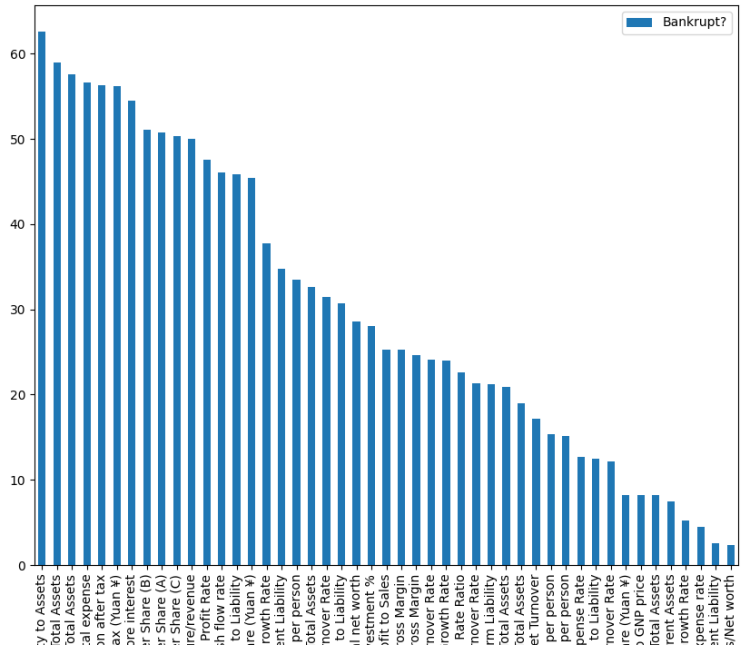
Rysunek 3: Research and development expense rate” po zastąpieniu wartości odstających

odstających niż w przypadku rozkładu normalnego. Również w przypadku skośności w większości przypadków jest ona dodatnia co sugeruje asymetrię prawostronną.

2.4 Korelacja między zmiennymi a zmienną Bankrupt?

Po wstępnym zapoznaniu się ze zmiennymi i tym jak one wyglądają “ wewnątrz”, należało również sprawdzić jak korelują ze zmienną zależną ”Bankrupt?”. Aby to sprawdzić została zastosowana funkcja corr. Poniższy wykres numer 4. prezentuje wielokość korelacji pomiędzy zmienną zależną i pozostałymi zmiennymi.

Okazało się, iż najsilniej skorelowane z zmienną “Bamkrupt?”, są “Current Liability to Assets”, “Net Income to Total Assets”, “Retained Earnings to Total Assets “, “Total income/Total expense” (powyżej 0,5). Zdecydowaliśmy, iż w modelach nie będziemy uwzględniać zmiennych których korelacja z zmienną objaśnianą jest poniżej 0.1.



Rysunek 4: Korelacja pomiedzy zmienną Bankrupt? i pozostałymi zmiennymi

3 Przygotowanie danych

3.1 Przygotowanie danych do modelu

Aby przygotować dane do modelu brakujące wartości zostały usunięte, zaś wartości odstające zostały zamienione na uśrednione wartości. Ze względu na to, iż większość zmiennych nie charakteryzowała się rozkładem normalnym, została przeprowadzona normalizacja.Następnie dane zostały poddane standaryzacji

3.2 Dobór zmiennych do modelu

Aby wyznaczyć zmienne uwzględniane w modelach została zastosowana analiza głównych składowych. Dla zmiennych których korelacja była większa niż 0.1 została zastosowana funkcja PCA z biblioteki sklearn. Wykres numer 5 prezentuje jaka cześć wariancji została wyjaśniona przy uwzględnieniu daje liczby zmiennych.Na wykresie jest to bardzo widoczne jak z początku wrasta poziom wyjaśnionej wariancji, aby przy 5 zmiennych osiągnoć 100%. Aby dokładniej zidentyfikować adekwatną ilość zmiennych, które należy uwzględnić w modelu, do funkcji PCA zostały dodane dwa argumenty svdsolver oraz ncomponets. Svdsolver jest to typ dekompozycji macierzy. W naszym przypadku zastała mu przypisana wielkość “full”, dzięki czemu funkcja PCA wybiera liczbę komponentów tak, aby wielkość wariancji, którą należy wyjaśnić, była większa niż procent określony przez ncomponents, w naszym przypadku był on ustalony na poziomie 0.95. Ostateczna ilość zmiennych uwzględnionych w modelu była równa 5.

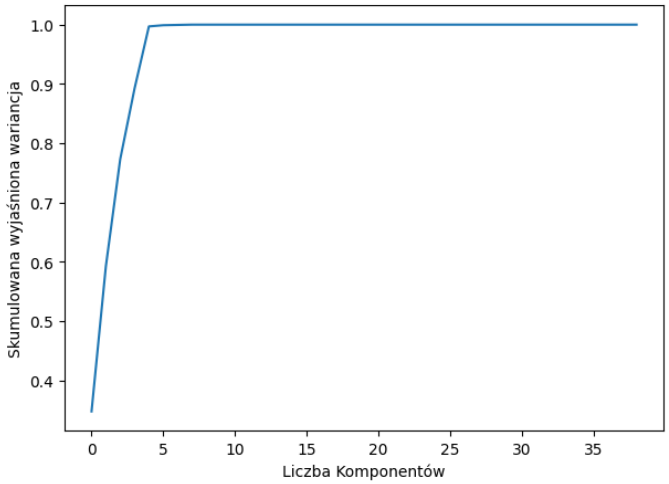
3.3 Podział na próbę ucząca i testową

Do próby testowej zostało zaliczone 20%(100) przypadków z całego zbioru danych z kolei pozostałe 80%(398) przypadków znalazło się w próbie uczącej/treningowej.

4 Wybrane modele

4.1 Regresja logistyczna

Regresja logistyczna (ang. Logistic Regression) – algorytm klasyfikacji binarnej oszacowujący prawdopodobieństwo przynależności do jednej z dwóch klas. Działanie modelu polega na stworzeniu funkcji logistycznej wynikającej z wyliczenia ważonej sumy cech wejściowych. Uczenie modelu



Rysunek 5: Zależność wyjaśnionej wariancji od liczby uwzględnianych zmiennych

polega na dostosowaniu takiego wektora parametrów, aby model szacował wysokie prawdopodobieństwo dla pozytywnych próbek a niskie dla próbek negatywnych. Jest to jeden z podstawowych modeli wykorzystywanych do klasyfikacji binarnej.

4.2 Maszyna wektorów nośnych

Maszyna wektorów nośnych (ang. Support Vector Machine) – algorytm klasyfikacji, którego działanie polega na wyznaczeniu linii decyzyjnej pomiędzy dwoma klasami, która jednocześnie jak najdokładniej rozdziela obie klasy oraz utrzymuje jak największy dystans pomiędzy nimi. W badaniu wykorzystano dwie metody nieliniowe algorytmu, wykorzystujące funkcje jądra wielomianowego (parametr kernel = 'poly') oraz jądra gaussowskiego RBF (parametr kernel = 'rbf'). Model ten bardzo dobrze sprawdza się w przypadku małych lub średnich zbiorów danych.

4.3 Metoda ekstremalnego wzmacniania gradientu

Metoda ekstremalnego wzmacniania gradientu (ang. Extreme Gradient Boosting - XGBoost) – algorytm zespołowy klasyfikacji wykorzystujący domyślnie drzewa decyzyjne. Działanie metody polega na iteracyjnej poprawie mniejszych modeli na podstawie błędu resztowego popełnionego przez poprzednika, minimalizując w ten sposób błąd resztowy całego modelu zespołowego. Jest to niezwykle szybki i skalowalny model, lecz jego trenowanie wymaga wielu zasobów obliczeniowych.

5 Uczenie i hiperparametryzacja modeli

Każda z metod została wykorzystania do wytrenowania modeli z użyciem dwóch zestawów: danych po przeprowadzonej redukcji wymiarowości oraz ustandaryzowanych danych bez redukcji wymiarowości.

Hiperparametry modeli zostały wybrane z wykorzystaniem metody przeszukiwania siatki (ang. Grid Search). Dobór przeszukiwanych zakresów hiperparametrów został wybrany metodą prób i błędów.

- Parametry regresji logistycznej:
  PCA: 'C': 0.2, 'solver': 'lbfgs', 'tol': 5, NONPCA: 'C': 0.5, 'solver': 'lbfgs', 'tol': 5
  - solver – algorytm używany do optymalizacji modelu,
  - tol – tolerancja zatrzymania im większy tym algorytm szybciej zatrzyma optymalizowanie modelu,
  - C – odwrotność siły regularyzacji, im większa wartość tym mniejsza regularyzacja,
- Parametry maszyny wektorów nośnych:
  PCA: 'C': 6, 'gamma': 'scale', 'kernel': 'rbf', NONPCA: 'C': 2, 'gamma': 0.01, 'kernel': 'rbf'
  - gamma – współczynnik jądra, im większy tym mniejszy promień wektora nośnego,

  - C – parametr regularyzacji, im większy tym mniejszy margines błędu,
  - kernel – funkcja jądra, w tym przypadku rbf lub poly.

- Parametry XGBoost:
  PCA: 'eta': 0.01, 'gamma': 0, 'max_depth': 4, 'n_estimators': 100,
  NONPCA: 'eta': 0.3, 'gamma': 1, 'max_depth': 5, 'n_estimators': 9
  - eta – określa wagę pojedynczego drzewa w zespole, im większe tym model dokładniej dopasowuje się do danych i słabiej generalizuje,
  - gamma – parametr regularyzacji, wskazuje o ile musi zmniejszyć się strata aby można było dodać kolejny węzeł w drzewie,
  - max depth – maksymalna głębokość drzewa w zespole,
  - n estimators – liczba drzew w zespole.

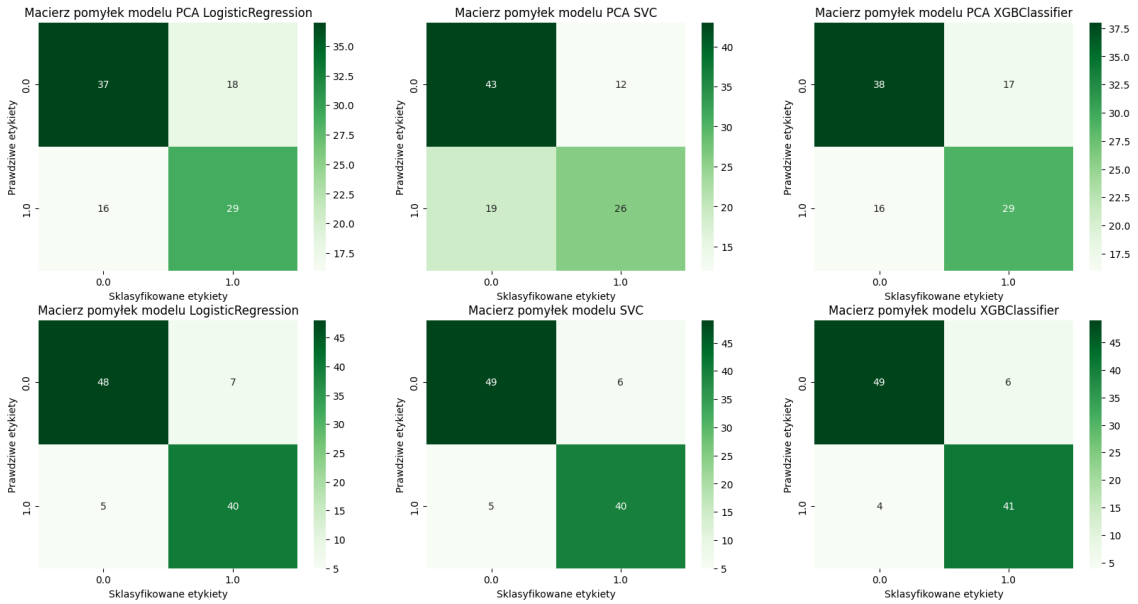
6 Ocena

6.1 Ocena modeli

Ze względu na niezbalansowaną liczebność klas w zestawie danych (klasa 0: 56%, klasa 1: 44%), modele zostały ocenione z wykorzystaniem miary F1, będącej średnią harmoniczną precyzji oraz czułości. Ponadto skuteczność modeli porównano z wykorzystaniem miary dokładności, macierzy pomyłek oraz miar czułości i precyzji.

Skuteczniejszymi modelami okazały się modele wyuczone na danych bez redukcji wymiarowości. Każdy z modeli wyuczonych na danych bez użycia PCA osiągnął bardzo podobne miary F1, które wyniosły 86-89%. Różnice pomiędzy modelami wyuczonymi na różnych zbiorach wyniosły około 20% dla każdego modelu.

Najskuteczniejszym modelem pod względem miary F1 oraz dokładności był model XGBoost, dla którego wartość miary F1 wyniosła 0,89% a dokładność 90%. Był to również najskuteczniejszy model jeśli chodzi o poprawne przewidywanie mniej liczebnej klasy 1, co wskazuje wartość miary czułości równa 0,91%



Rysunek 6: Stosunek liczebności klas

6.2 Ważność atrybutów

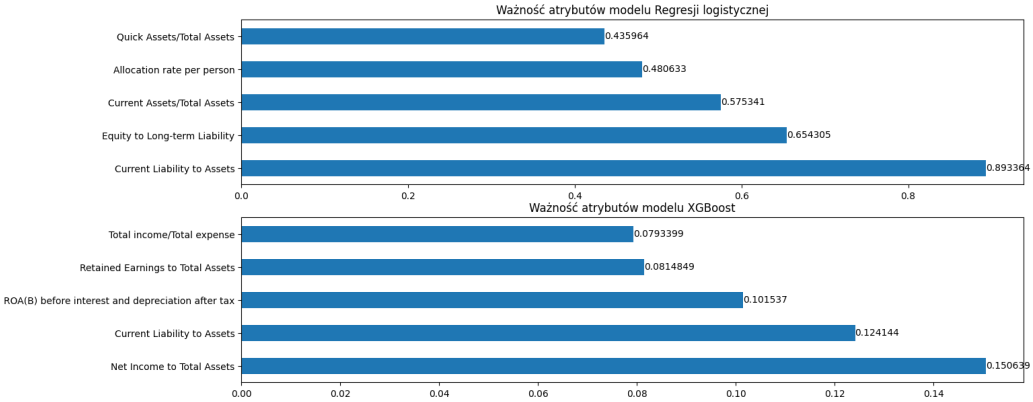
Ważność atrybutów została wyliczona dla modelu regresji logistycznej oraz XGBoost dla danych bez PCA. Transforamcja danych przez wybraną metodę kernel 'rbf' wykorzystywaną w algorytmie wektorów nośnych uniemożliwia znalezienie najważniejszych zmiennych wpływających na skuteczność modelu. Ważności atrybutów przedstawione zostały na rysunku 7.

Model	Dokładność	F1-Score	Czułość	Precyzja
PCA Regresja logistyczna	0.66	0.6304	0.6444	0.6981
PCA SVM	0.69	0.6265	0.5778	0.6935
PCA XGBoost	0.67	0.6373	0.6444	0.7037
Regresja logistyczna	0.88	0.8696	0.8889	0.9057
SVM	0.89	0.8791	0.8889	0.9074
XGBoost	0.90	0.8913	0.9111	0.9245

Tabela 1: Tabela porównawcza miar skuteczności modeli

Dla modelu regresji logistycznej najważniejszym atrybutem był stosunek bieżącego zobowiązania do wartości aktywów przedsiębiorstwa (ang. Current Liability to Assets). Kolejnymi ważnymi atrybutami w przypadku tego modelu były m.in. stosunek kapitału własnego do zobowiązań długoterminowych (ang. Equity to Long-term Liability) oraz stosunek aktywów obrotowych i aktywów razem.

Dla modelu XGBoost najważniejszym atrybutem był stosunek dochodu netto i wartości aktywów (ang. Net Income to Total Assets). Drugim najważniejszym atrybutem okazał się stosunek bieżącego zobowiązania do wartości aktywów przedsiębiorstwa (ang. Current Liability to Assets)



Rysunek 7: Ważność atrybutów dla modelu XGBoost i regresji logisycznej

Rożnice pomiędzy listami najważniejszych atrybutów obu modeli wynikają prawdopodobnie z różnego podejścia algorytmów do znajdywania związków pomiędzy wartościami cech a przynależnością do klasy. Model XGBoost o wiele lepiej radzi sobie w znajdywaniu nieliniowych powiązań, z kolei model regresji logistycznej lepiej znajduje zależności liniowe. Ponadto w przypadku algorytmu XGBoost, ze względu na występującą współliniowość wśród atrybutów, ich ważność może być zniekształcona.

7 Wnioski

Mimo, że redukcja wymiarowości pozwala na o wiele szybsze uczenie modeli to uniemożliwia ona zbadanie, które atrybuty mają największy wpływ na ich skuteczność, tworząc w ten sposób z modeli „czarne skrzynki". Ponadto każda redukcja wymiarowości zmniejsza liczbę dostarczanych przez dane informacji, co w tym przypadku znacząco wpłynęło na skuteczność modeli.

Wysokie skuteczności klasyfikacji modeli wskazują, że wskaźniki finansowe firmy pozwalają na efektywną predykcję jej upadłości. Każdy z trzech modeli osiągnął podobne i zadowalające wyniki, które prawdopodobnie mogłyby być o wiele wyższe wraz z powiększeniem zbiorów danych uczących.