

Studium Magisterskie

Kierunek: Analiza Danych – Big Data

Marcin Mandziej

Nr albumu: 68373

**Analiza determinant upadku przedsiębiorstw w Polsce. Stworzenie narzędzia do automatycznej predykcji bankructwa przedsiębiorstw na podstawie dostępnych danych historycznych**

Praca magisterska

Napisana w Instytucie Statystyki i Demografii

pod kierunkiem naukowym

dr Karola Przanowskiego

Warszawa 2021

Spis treści.

[Wstęp. - 5 -](#_Toc62574762)

[Uzasadnienie wyboru tematu. - 5 -](#_Toc62574763)

[Struktura pracy. - 7 -](#_Toc62574764)

[I. Pojęcie ryzyka bankructwa przedsiębiorstwa i korzyści z jego predykcji. - 9 -](#_Toc62574765)

[I.1 Pojęcie ryzyka w kontekście prowadzenia działalności gospodarczej. - 9 -](#_Toc62574766)

[I.2 Pojęcie upadłości i likwidacji przedsiębiorstwa. - 11 -](#_Toc62574767)

[I.3 Wymogi regulacyjne odnośnie weryfikacji kontrahentów w Polsce. - 14 -](#_Toc62574768)

[I.4 Korzyści z możliwości automatycznej predykcji bankructwa przedsiębiorstw. - 17 -](#_Toc62574769)

[II. Omówienie danych wykorzystanych do przeprowadzenia analizy. - 21 -](#_Toc62574770)

[II.1 Źródła danych wykorzystane do przeprowadzenia analizy. - 21 -](#_Toc62574771)

[II.2 Struktura i przekształcenia zebranych danych. - 24 -](#_Toc62574772)

[II.3 Analiza eksploracyjna wykorzystanych danych. - 31 -](#_Toc62574773)

[II.4 Wybór zmiennych najlepiej wyjaśniających zjawisko zakończenia działalności gospodarczej. - 43 -](#_Toc62574774)

[III. Analiza właściwa. Budowa i ocena jakości modeli klasyfikacyjnych. - 49 -](#_Toc62574775)

[III.1 Model klasyfikacyjny zbudowany na podstawie algorytmu Gradient Boosting Machine. - 49 -](#_Toc62574776)

[III.2 Model klasyfikacyjny zbudowany na podstawie algorytmu klasyfikacyjnego lasu losowego. - 54 -](#_Toc62574777)

[III.3 Model klasyfikacyjny zbudowany na podstawie algorytmu sieci neuronowej. - 58 -](#_Toc62574778)

[III.4 Porównanie i ocena jakości zbudowanych modeli klasyfikacyjnych - 62 -](#_Toc62574779)

[III.5 Interpretacja uzyskanych wyników i wyciągnięcie wniosków. - 64 -](#_Toc62574780)

[Zakończenie. - 66 -](#_Toc62574781)

[Bibliografia. - 68 -](#_Toc62574782)

[I. Wydawnictwa zwarte. - 68 -](#_Toc62574783)

[II. Artykuły. - 68 -](#_Toc62574784)

[III. Akty prawne. - 69 -](#_Toc62574785)

[IV. Strony internetowe. - 69 -](#_Toc62574786)

[Spis wykresów. - 71 -](#_Toc62574787)

[Spis tabel. - 71 -](#_Toc62574788)

[Streszczenie. - 73 -](#_Toc62574789)

**Wstęp.**

## Uzasadnienie wyboru tematu.

Temat przyczyn upadku przedsiębiorstw jest bardzo chętnie poruszaną kwestią przez ekonomistów. Znajomość powodów tego zjawiska umożliwia zastosowanie posiadanej wiedzy w praktyce do przeprowadzania predykcji na temat podmiotów prowadzących obecnie działalność gospodarczą. W celu poprawnej oceny prawdopodobieństwa upadłości firm przeprowadzono wiele analiz jakościowych i ilościowych, których najbardziej rozpoznawanym rezultatem jest model scoringowy opracowany przez profesora Edwarda Altmana, którego pierwsza wersja została opublikowana już w 1968 roku. Analiza tego tematu jest w pełni uzasadniona z biznesowego punktu widzenia. Możliwość odpowiednio wczesnego oszacowania prawdopodobieństwa zamknięcia działalności gospodarczej w wyznaczonym horyzoncie czasowym z zadowalającą jakością predykcji jest przydatną umiejętnością w kontekście prowadzenia przedsiębiorstwa. Prawidłowa ocena szansy bankructwa kontrahenta, dostawcy czy podmiotu prowadzącego działalność konkurencyjną może pozwolić firmie przygotować się do ewentualnego zagrożenia lub wykorzystania okazji nadarzających się na rynku. Posiadanie takiej umiejętności jest również pożądane z perspektywy udziałowców, właścicieli firmy czy też spekulantów, którzy mogą odpowiednio wcześniej przygotować się na rozwój sytuacji i dzięki posiadanej wiedzy wzbogacić się lub uniknąć strat.

Znajomość sytuacji innych firm, z którymi została podjęta współpraca lub jest takowa rozważana może pomóc w zarządzaniu ryzykiem operacyjnym i finansowym, zarządzaniu zobowiązaniami poprzez skuteczniejszą ocenę możliwości ich ściągalności. Zastosowanie tej wiedzy może być też przydatne do ograniczenia ryzyka nadużyć. Wysoka ocena szansy upadłości firmy powinna pomóc w bardziej skutecznym stopniu dobierać partnerów biznesowych, dostawców (szczególnie będzie to istotne w przypadku przedsiębiorstw zajmujących się handlem dobrami materialnymi i ich przetwarzaniem, które są w pewnym stopniu uzależnione od firm wyżej położonych w łańcuchu dostaw dostarczających surowce i półprodukty). Dzięki odpowiedniej weryfikacji kontrahentów i partnerów biznesowych przedsiębiorstwo ogranicza możliwość poniesienia strat wizerunkowych i reputacyjnych z powodu ograniczenia kontraktów z podmiotami o niskiej wiarygodności finansowej, które zostałyby nawiązane w przypadku braku takiej weryfikacji.

Regulacje zawarte w polskim prawie narzucają przedsiębiorcom obowiązek przeprowadzania formalnej weryfikacji kontrahenta zarówno przed dokonaniem transakcji jak i w trakcie realizacji umowy. Choć według treści ustaw regulujących te kwestie takowa analiza nie musi obejmować sprawdzenia kondycji finansowej partnera biznesowego, to przy wykorzystaniu technologii i dostępnych źródeł danych stosunkowo niewielkim kosztem może zostać ona poszerzona również o ten element a co za tym idzie – na podstawie pozyskanych informacji na temat podmiotu może zostać przeprowadzone oszacowanie prawdopodobieństwa bankructwa. Taka wiedza uzyskana z wyprzedzeniem może okazać się bardzo cenna dla firmy, zwłaszcza, że może zostać pozyskana niewielkim nakładem pracy, kosztów i w sposób zautomatyzowany, czego dowodem ma być narzędzie przedstawione w pracy. Wydaje się, że dzięki coraz lepszej dostępności informacji potrzebnych do wykonania tego rodzaju analizy oraz stosunkowo niewielkiemu kosztowi ich pozyskania przedsiębiorstwa będą dążyć do zwiększania wiedzy na temat innych uczestników rynku. Potencjalne korzyści są wysokie w stosunku do nakładów pracy i kosztów potrzebnych do jej pozyskania.

Na rynku istnieją firmy dostarczające usługi i gotowe rozwiązania związane z tematem masowej weryfikacji kontrahentów. Podobne projekty realizowane są zwykle na zasadzie zebrania informacji z wielu otwartych i płatnych źródeł, rekoncyliacji tych danych i przeprowadzenia analizy wiarygodności i oceny kondycji firmy. Zazwyczaj analiza jest przeprowadzana przy wykorzystaniu modelu eksperckiego i karty scoringowej, gdzie wartości uwzględnionych zmiennych są odpowiednio punktowane ze względu na ich pozytywny bądź negatywny wpływ na ryzyko kontrahenta. W trakcie swojego doświadczenia zawodowego autor nie spotkał się jednak z w pełni zautomatyzowanym rozwiązaniem, które dostarczałoby także oceny prawdopodobieństwa bankructwa. Autor nie spotkał się także z rozwiązaniem wykorzystującym model zbudowany na tak dużym wolumenie danych historycznym złożonym z kilkudziesięciu tysięcy obserwacji historycznych – przedsiębiorstw prowadzących działalność operacyjną w Polsce w ciągu ostatnich lat. Brak istnienia podobnego rozwiązania, które mogłoby przysłużyć się podmiotom prowadzącym działalność gospodarczą stał się inspiracją do przeprowadzenia samodzielnej analizy.

Celem pracy jest próba odpowiedzi na pytanie czy możliwe jest skonstruowanie narzędzia, które wykorzystując ogólnodostępne dane pochodzące ze źródeł otwartych oraz źródeł dostępnych dla studentów Szkoły Głównej Handlowej w Warszawie będzie w stanie z zadowalającą skutecznością ocenić prawdopodobieństwo bankructwa danej firmy. W opinii autora skonstruowanie podobnego narzędzia jest potrzebne i powinno dostarczyć wartość dodaną osobom odpowiedzialnym za kierowanie podmiotami prowadzącymi działalność gospodarczą, ich właścicielom oraz spekulantom. Kryteria oceny szansy bankructwa firmy są w pewnym sensie arbitralne i w przypadku zastosowania narzędzia do rozwiązania rzeczywistego problemu biznesowego będą zależne od tolerancji na ryzyko osób decyzyjnych podejmujących decyzję o współpracy z daną firmą. To osoby, które kierują przedsiębiorstwem i w szczególności są odpowiedzialne za ryzyko operacyjne powinny decydować o preferowanym punkcie odcięcia oceny ryzyka niewypłacalności nadanej przez model, powyżej którego podjęcie współpracy (lub jej kontynuacja) z podmiotem jest zbyt niebezpieczne.

Możliwość odpowiednio wczesnego przewidzenia upadłości lub likwidacji podmiotów, które znajdują się w otoczeniu biznesowym przedsiębiorstwa dostarcza osobom kierującym działalnością wymiernych korzyści poprzez poprawę jakości i przyspieszenie procesów decyzyjnych, co ma duże szanse przełożyć się pozytywnie na wynik finansowy oraz stabilność firmy. W opinii autora są to wystarczające argumenty do budowy zautomatyzowanego narzędzia wykorzystującego źródła jawnego wywiadu gospodarczego do wykonywania prognoz na ten temat.

## Struktura pracy.

W pierwszym rozdziale pracy omówione zostanie pojęcie bankructwa przedsiębiorstwa, a także ryzyka w kontekście zarządzania przedsiębiorstwem i ryzyka upadłości. Poruszony zostanie także temat wymogów regulacyjnych odnośnie weryfikacji kontrahentów i jak przeprowadzanie predykcji ich upadłości wpasowuje się w obowiązki narzucone przez prawo. Przedstawione zostaną również korzyści płynące z możliwości zautomatyzowanej oceny szansy bankructwa bądź likwidacji firmy dla podmiotów prowadzących działalność gospodarczą.

W kolejnej części pracy zostanie przeprowadzony przegląd dostępnych danych i możliwość ich wykorzystania w kontekście predykcji bankructwa. Omówiony zostanie sposób pobierania danych i ich rekoncyliacji w kontekście korzystania z narzędzia w sposób automatyczny. Przeprowadzona zostanie analiza eksploracyjna zebranych danych, wybrane zostaną zmienne o największej zdolności wyjaśniania zjawiska bankructwa. Omówione zostanie znaczenie wybranych czynników ryzyka użytych do budowy modelu klasyfikacyjnego oraz sposób kalkulacji metryk tworzonych na podstawie pobranych surowych danych wykorzystanych następnie jako zmienne objaśniające.

Następnym elementem pracy będzie budowa modeli w oparciu o algorytmy uczenia maszynowego. Przeprowadzona zostanie walidacja modeli i ocena skuteczności prognozy oraz stabilność wyników. Wybranie zostanie najlepszy model, który będzie wykorzystany do przeprowadzania prognoz na temat przyszłości wybranych podmiotów i który powinien potencjalnie zostać zastosowany w środowisku produkcyjnym po przeprowadzeniu wdrożenia narzędzia.

W ostatniej części pracy zostanie przeprowadzona próba odpowiedzi na pytanie, czy zbudowane narzędzie posiada wystarczającą jakość, żeby mogło zostać wdrożone w praktyce i zastosowane przez firmy prowadzące działalność operacyjną. Zostanie wykonana ocena jakości modelu i możliwości jego wykorzystania w kontekście rozwiązywania rzeczywistych problemów biznesowych. Omówiona zostanie przykładowa implementacja i sposób wykorzystania zbudowanego narzędzia.

# I. Pojęcie ryzyka bankructwa przedsiębiorstwa i korzyści z jego predykcji.

## I.1 Pojęcie ryzyka w kontekście prowadzenia działalności gospodarczej.

Niepowodzenie przedsiębiorstwa i zakończenie jego działalności jest nieodłączną i charakterystyczną cechą wszystkich gospodarek wolnorynkowych. Możliwość upadłości ma wpływ zarówno na działanie konkretnej, zagrożonej nią firmy jak i całe otoczenie, w którym dany podmiot prowadzi działalność. W przypadku zrealizowania się scenariusza bankructwa jego konsekwencje ponoszą wszyscy uczestnicy rynku. Zakończenie działalności przedsiębiorstwa wpływa na klientów, konkurentów, dostawców, regulatorów – dotknięci zostają wszyscy interesariusze powiązani z bankrutem. Dlatego należy zwracać uwagę na ryzyko niepowodzenia biznesowego. Punktem wyjścia do oceny ryzyka niepowodzenia działalności gospodarczej jest jego zdefiniowanie.

Ryzyko jest zjawiskiem obecnym w historii od zawsze. Dotyczy ono każdego aspektu ludzkiego życia, jest związane z podejmowaniem decyzji, których skutki mogą być znane lub nieznane. Skutki podjęcia decyzji mogą być mierzalne lub niemierzalne i w zależności od interesów podmiotu, którego one dotyczą - można intepretować je w kontekście pozytywnym lub negatywnym. Zagadnienie ryzyka jest pojęciem wielopłaszczyznowym, które może być interpretowane indywidualnie w zależności od okoliczności, wagi i skutków podejmowanych decyzji. W zależności od kontekstu ryzyko może być definiowane na różne sposoby.

Według teorii gier mianem ryzyka możemy określić wystąpienie sytuacji, w której postąpienie według jednego z możliwych wariantów decyzyjnych prowadzi do wystąpienie różnych pozytywnych i negatywnych konsekwencji, gdzie prawdopodobieństwo wystąpienia każdej z konsekwencji jest znane[[1]](#footnote-1).

Kluczową kwestią w tym kontekście jest informacja, że znany jest rozkład prawdopodobieństwa dla wszystkich skutków, które wystąpią po dokonaniu wyboru danego wariantu decyzyjnego. Taka sytuacja nie ma jednak miejsca w przypadku prowadzenia działalności biznesowej. Osoby odpowiedzialne za decyzje dotyczące przedsiębiorstwa nie mają pełnej wiedzy na temat tego, jakie skutki będzie miała podjęta przez nie decyzja. W przypadku gospodarki wolnorynkowej jest to niemożliwe, ponieważ przedsiębiorstwa działają w otwartym środowisku. Potencjalne skutki nie zależą jedynie od decyzji kierownictwa i działań podjętych przez pracowników. Na skutki podjętych decyzji mają wpływ także wybory pozostałych uczestników rynku, działania wszystkich podmiotów są ze sobą powiązane, powiem uczestnicy rynku wpływają na siebie nawzajem. Przedsiębiorstwo nie wie jak zachowają się pozostałe podmioty, których dotyczą skutki podjętej decyzji. Z pomocą przychodzą modele mające za zadanie upraszczać rzeczywistość i pomóc w przewidzenia zachowania pozostałych uczestników rynku pomagając w ten sposób osobom decyzyjnym podjąć optymalną decyzję. Mimo wszystko, dalej one dostarczają one tylko prognozę a nie rzeczywiste, potwierdzone informacje.

W przypadku prowadzenia działalności gospodarczej wydaje się, że pojęcie ryzyka może zostać zawężone do możliwości wystąpienia negatywnych konsekwencji podjętych decyzji. W takim kontekście ryzyko można określić miarę prawdopodobieństwa i wagi niepożądanych konsekwencji. Ryzyko może zostać być intepretowane jako wystąpienie sytuacji, w której decydent musi podjąć decyzję pozostając w asymetrii informacji i nie posiadając pełnej wiedzy na temat możliwych skutków. Jest on świadomy różnych możliwych konsekwencji podjętej decyzji i jest w stanie (według swojej wiedzy i doświadczenia) jedynie oszacować prawdopodobieństwo wystąpienia danego skutku[[2]](#footnote-2).

W takiej sytuacji właśnie znajdują się podmioty prowadzące działalność gospodarczą. Są one świadome różnych scenariuszy decyzyjnych i na podstawie wiedzy i doświadczenia osób podejmujących decyzje mogą oszacować wagę i prawdopodobieństwo wystąpienia potencjalnych skutków. Jednym z możliwych skutków podejmowania niewłaściwych decyzji biznesowych w długim horyzoncie czasowym jest niepowodzenie przedsięwzięcia i bankructwo przedsiębiorstwa.

Podmioty prowadzące działalność muszą stawić czoła pewnym zagrożeniom wynikającym z czynników finansowych, regulacyjnych, reputacyjnych czy spowodowanych niedostatkiem informacji. Żaden z podmiotów nie jest w stanie przewidzieć swoich osiąganych wyników finansowych tylko na podstawie podejmowanych decyzji, ponieważ wpływają na niego czynniki zewnętrze, w tym regulacyjne i makroekonomiczne, na które zdecydowana większość firm ma bardzo ograniczony wpływ. Pomimo niewielkiej mocy sprawczej i wpływu na środowisko zewnętrzne przedsiębiorstwa mogą skupić się na pozyskaniu wiedzy i budowaniu narzędzi, które pozwolą im w lepszym stopniu oszacować konsekwencje podejmowanych decyzji. Posiadanie lepszej wiedzy na temat czynników prowadzących do bankructwa zmniejsza tę asymetrię informacji i dostarcza decydentowi wiedzy, przez co jest on w stanie podejmować decyzje w bardziej świadomy sposób i ograniczać prawdopodobieństwo zakończenia prowadzonej działalności porażką. W tym celu możliwe jest wykonanie analizy danych historycznych na temat podmiotów, które znajdowały się w podobnej sytuacji do decydenta[[3]](#footnote-3).

Zbudowanie narzędzia, które w oparciu o zgromadzone dane i wykorzystanie algorytmu klasyfikacyjnego pozwala w zautomatyzowany sposób szacować wpływ potencjalnych podjętych decyzji. Dzięki zastosowaniu takiego rozwiązania asymetria informacji jest zmniejszana, a osoby decyzyjne w przedsiębiorstwie niewielkim nakładem kosztów i czasu są w stanie ograniczyć długoterminowe ryzyko negatywnych konsekwencji dla przedsiębiorstwa – w tym jego upadłości – zmniejszając jednocześnie ilość czasu potrzebnego rozważenia wariantów i do podjęcia decyzji. Przy założeniu, że osoby podejmujące decyzje biznesowe są w stanie oszacować ich konsekwencje znajomość czynników wpływających na bankructwo firm oraz ich wag pozwala na przeprowadzenie swego rodzaju symulacji. Po sprawdzeniu tego jak zmienią się wskaźniki wyjaśniające szansę na upadłość firmy decydent może w pewnym stopniu zweryfikować, czy jego zamiary będą trafne.

## I.2 Pojęcie upadłości i likwidacji przedsiębiorstwa.

Według treści polskiego prawa przedsiębiorstwem można określić podmiot, którego podstawowym celem jest osiągnięcie zysku w wyniku prowadzenia działalności gospodarczej poprzez sprzedaż towarów i usług. Przedsiębiorstwem jest zorganizowany zespół składników materialnych i niematerialnych, które są przeznaczone do prowadzenia działalności gospodarczej i umożliwiają pozyskiwanie przychodów. Z praktycznego punktu widzenia oznacza to, że przedsiębiorstwem można określić każdą organizację, której podstawowym celem jest osiąganie zysku i utrzymywanie się z prowadzonej działalności[[4]](#footnote-4).

W świetle polskiego prawa upadłość (nazywana inaczej bankructwem, pojęcia te są wobec siebie tożsame i można posługiwać się nimi wymiennie) ogłaszana jest przez sąd w stosunku do dłużnika, który stał się niewypłacalny – oznacza to, że podmiot nie jest dłużej w stanie terminowo obsługiwać posiadanych zobowiązań pieniężnych w stosunku do wierzycieli. Należy również dodać, że dopuszczalne jest opóźnienie w spłacie zobowiązań do maksymalnie trzech miesięcy od daty wymagalności zadłużenia. W rozumieniu treści ustawy niewypłacalność jest tożsama z utraceniem płynności finansowej przez podmiot prowadzący działalność gospodarczą – tym samym możliwe jest stosowanie obu pojęć zamiennie.

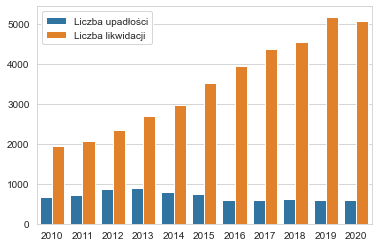
Dłużnikiem można określić osobę fizyczną, osobę prawną lub jednostkę organizacyjną, która posiada osobowość prawną zobowiązaną do wypełnienia świadczenia na rzecz innego podmiotu (wierzyciela) na podstawie ustalonego stosunku zobowiązania. Niewypłacalność może zostać jednak ogłoszona także wobec podmiotu, jeżeli jego zobowiązania przekraczają wartość majątku nawet w sytuacji, gdy obsługuje on posiadane zobowiązania terminowo. Zapis ten nie znajduje jednak zastosowania w przypadku osób fizycznych. W stosunku do podmiotu może zostać wszczęte postępowanie upadłościowe jeżeli posiada on zdolność upadłościową, która oznacza możliwość odzyskania należności przez wierzycieli dzięki spieniężeniu majątku dłużnika. Procedura wszczęcia postępowania upadłościowego regulowana jest przez ustawę Prawo upadłościowe z dnia 28 lutego 2003 r.[[5]](#footnote-5).

W świetle prawa likwidacja to z kolei formalne zakończenie i zamknięcie prowadzonej działalności gospodarczej, czego skutkiem jest zakończenie istnienia firmy oraz spieniężenie jej majątku. Procedura likwidacyjna jest regulowana przez Kodeks spółek handlowych z dnia 15 września 2000 r. Oznacza to, że przepisy dotyczą podmiotów będących spółkami prawa handlowego, nie dotyczą osób fizycznych prowadzących działalność gospodarczą oraz spółek cywilnych. Przyczyny likwidacji nie zawsze są bezpośrednio powiązane z niewypłacalnością. Mogą one wynikać z chęci wspólników spółki do zakończenia działalności bądź też przeniesienia siedziby firmy za granicę. Do rozpoczęcia procesu likwidacji może doprowadzić także utracenie koncesji lub zezwolenia na wykonywanie działalności gospodarczej. Inną z możliwych przyczyn jest ponoszenie strat lub niezadowalających wyników finansowych, co może być powiązane z początkowym stadium niewypłacalności. Możliwe jest także zakończenie działalności z powodu konfliktu między wspólnikami, który uniemożliwia zrealizowanie celu, do którego została powołana spółka. Kolejną z możliwych przyczyn jest osiągnięcie celu firmy, w sytuacji gdy jest ona podmiotem celowym – została utworzona tylko do zrealizowania konkretnych transakcji lub przedsięwzięć. Po wszczęciu postępowania likwidacyjnego przedsiębiorstwo dalej prowadzi działalność gospodarczą w ograniczonym zakresie aż do dnia zakończenia postępowania. Następnie jest ono wykreślane z rejestru przedsiębiorców prowadzonego przez Krajowy Rejestr Sądowy[[6]](#footnote-6).

Istnieje również możliwość zakończenia działalności bez wszczynania postępowania likwidacyjnego. Możliwość taką wprowadzono od stycznia 2016 roku. Jeśli sąd prowadzący sprawę ustali, że spółka spieniężyła swój majątek, nie posiada zbywalnych aktywów oraz zobowiązań do spłaty, nie prowadzi również działalności gospodarczej którą wcześniej deklarowała możliwe jest wykreślenie spółki z rejestru[[7]](#footnote-7).

Jak pokazują statystyki odnośnie liczby podmiotów, wobec których przeprowadzono postępowanie upadłościowe i likwidacyjne w Polsce zagrożenie zakończenia prowadzonej działalności jest znaczące. Raporty na ten temat są publikowane przez Ministerstwo Finansów w cyklicznie wydawanych biuletynach – Monitorach Sądowych i Gospodarczych, które są dostępne bezpłatnie na stronie internetowej Ministerstwa. Dane na ten temat są także gromadzone przez specjalistyczne firmy, agregowane i udostępniane odpłatnie w znacznie bardziej przystępnej formie niż ma to miejsce w przypadku raportów organu rządowego. Informacje na temat upadłości i likwidacji podmiotów, w tym daty uchwalenia i zakończenia postępowań są również przechowywane w bazie danych Głównego Urzędu Statystycznego, do której dostęp jest darmowy. Oznacza to, że w celu budowy narzędzi do predykcji likwidacji firm możliwe jest skorzystanie z tego źródła danych w celu otrzymania danych historycznych na ten temat i przeprowadzenia analizy przyczyn upadku firm, które zostały zlikwidowane. Natomiast, raporty udostępniane przez Ministerstwo Finansów stanowią dobre źródło informacji na temat skali badanego zjawiska i trendu, który panował w ostatnich latach.

Wykres 1. Liczba upadłości i likwidacji firm w latach 2000-2019.



Źródło: Opracowanie własne na podstawie danych udostępnionych przez witrynę www.coig.com.pl na podstawie raportów z lat 2000 - 2017[[8]](#footnote-8).

Z wykresu można odczytać, że w ciągu ostatnich 10 lat liczba firm wobec których ogłoszono likwidację systematycznie rosła. Liczba to w ciągu badanego okresu wzrosła ponad 2,5-krotnie. Oznacza to, że zjawisko zakończenia działalności dotyczy coraz większej liczby firm objętych regulacjami kodeksu handlowego. Liczba podmiotów wobec których ogłoszono upadłość pozostaje natomiast na stosunkowo stabilnym poziomie w badanym okresie. Z przedstawionych danych odczytać można niewielką tendencję wzrostową w latach następujących po kryzysie finansowym z 2008 r., ale na przestrzeni całego analizowanego okresu trend nie jest stały i zmieniał się w badanych latach. Należy zwrócić uwagę jednak przede wszystkim na liczbę podmiotów z zakończonym procesem likwidacyjnym. Zjawisko to skutkuje realnymi konsekwencjami dla uczestników rynku znajdujących się w otoczeniu gospodarczym zlikwidowanych firm oraz podmiotów powiązanych lub prowadzących z nimi współpracę. Gdyby podmioty te mogły przewidzieć wystarczająco wcześnie ich likwidację mogłyby zminimalizować negatywne konsekwencje z tego tytułu, lub wykorzystać nadarzające się okazje w przypadku prowadzenia przez nich działalności konkurencyjnej.

Przewidywanie, czy działalność gospodarcza będzie prowadzona z sukcesem czy z czasem firma stanie się niewypłacalna lub działalność zostanie zakończona z innych przyczyn zanim faktycznie to się stanie doprowadziło do przeprowadzenia na ten temat wielu analiz i powstania różnych teorii. Niepowodzenia wielu przedsiębiorstw i konsekwencje skutków ich upadłości są powodami dla których prowadzone są badania nad przyczynami upadłości. Takie analizy są podejmowane w celu lepszego przygotowania się z wyprzedzeniem na negatywne skutki bankructwa przedsiębiorstw. Dzięki znajomości powodów likwidacji działalności gospodarczych w przeszłości można przypuszczać, że podmioty o podobnych cechach działające obecnie również mogą zakończyć działalność. Jest to cenna wiedza, która umożliwia osobom odpowiedzialnym za kierowanie przedsiębiorstwem ocenę obecnej sytuacji prowadzonej firmy lub ocenę kondycji kontrahentów, klientów czy podmiotów prowadzących działalność konkurencyjną. To z kolei umożliwia podjęcie trafnych decyzji zwiększających zysk firmy i jego wartość. Znajomość tych przyczyn jest również przydatna właścicielom i inwestorom, którzy chcą ocenić perspektywy posiadanego biznesu lub inwestycji.

## I.3 Wymogi regulacyjne odnośnie weryfikacji kontrahentów w Polsce.

Sprawdzenie wiarygodności finansowej potencjalnego nowego kontrahenta jest nie tylko przydatne z perspektywy oceny szansy na ciągłość współpracy oraz predykcji na temat jego wypłacalności. Obecnie sprawdzenie rzetelności partnerów biznesowych jest koniecznością, nie tylko potencjalną dodatkową aktywnością, która może dostarczyć przedsiębiorstwu wartości dodanej w postaci usprawnienia procesów decyzyjnych.

Do weryfikacji wiarygodności finansowej i reputacyjnej kontrahenta zobowiązują regulacje wprowadzone w polskim prawie w 2018 r. Nowe przepisy zostały wprowadzone w celu ograniczenia ryzyka zawierania nieuczciwych transakcji, nadużyć i oszustw finansowych oraz zmniejszenia rozmiaru luki podatkowej i wolumenu pieniędzy wyprowadzanych z legalnego obiegu finansowego z powodu wyłudzeń podatku VAT. Według zaleceń Ministerstwa Finansów podmioty prowadzące działalność gospodarczą mają obowiązek zachowania należytej staranności podczas zawierania transakcji, podczas których wymagane jest odprowadzenie podatku od wartości dodanej od towarów i usług. Jednym z kluczowych elementów stosowania się do tych zaleceń jest formalna weryfikacja kontrahenta zanim transakcja zostanie przeprowadzona. Na mocy obowiązujących zaleceń przedsiębiorstwa zostały zobowiązane do zbadania zarówno potencjalnych jak i aktualnych kontrahentów w kwestii wielu kryteriów. Działania te mają być wykonywane w celu upewnienia się, czy partner biznesowy jest wiarygodny i czy transakcje nie narażą firmę na ryzyka związane z oszustwami podatkowymi oraz nadużyciami finansowymi.

Na przedsiębiorców został nałożony obowiązek weryfikacji tzw. kryteriów formalnych i transakcyjnych. W związku z tym przed podjęciem współpracy z nowym podmiotem przedsiębiorcy zobowiązani są do sprawdzenia wielu czynników, takich jak obecność podmiotu w rejestrze KRS lub CEIDG (w zależności od formy prawnej prowadzonej działalności), informacji na temat kontrahenta w bazie podatników VAT, posiadanych koncesji i zezwoleń koniecznych do zawarcia planowanych transakcji, prawdziwości tożsamości reprezentantów, prokurentów i pełnomocników w bazach rejestrowych.

Dodatkowo przedsiębiorca powinien zatroszczyć się o zaspokojenie kryteriów transakcyjnych, które prowadzą do zmniejszenia ryzyka nadużycia na tle podatkowym poprzez przeprowadzanie płatności bez ryzyka gospodarczego – do najważniejszych zaleceń należą: ograniczenie stosowania płatności podzielonych, wprowadzenie obowiązku transferu pieniędzy jedynie na rachunki bankowe wskazane przez firmy w rejestrze podatników VAT (tzw. białej liście podatników VAT), oraz ograniczenie progu maksymalnej dozwolonej transakcji gotówkowej do 15 000 PLN[[9]](#footnote-9).

Należy także pamiętać, że oprócz omówionych regulacji, które są uniwersalne i niezależne od rodzaju prowadzonej działalności i branży, do której należy przedsiębiorstwo istnieje wiele sektorów, które z powodu wymogów prawnych są zobowiązane do weryfikacji partnerów biznesowych w celu przeciwdziałania praniu pieniędzy i finansowaniu terroryzmu. Przepisy narzucone przez regulatorów na terenie Unii Europejskiej i prawo obowiązujące w Polsce sprawiają, że między innymi podmioty takie jakie banki, instytucje finansowe, fundusze inwestycyjne, fintechy, kancelarie prawne, notariusze, agenci nieruchomości, firmy oferujące usługi hazardowe i bukmacherskie mają prawny obowiązek weryfikacji, czy środki udostępniane klientom nie są następnie wykorzystywane do prowadzenia wspomnianych aktywności. W związku z tym, podmioty te muszą zweryfikować źródło dochodów badanych partnerów biznesowych oraz na co one są przeznaczane. Dlatego, wyżej wymienione podmioty muszą nawet pośrednio sprawdzić kondycję finansową klientów i kontrahentów i mają dostęp do danych finansowych tych podmiotów. Oznacza to, że mogą wykorzystać posiadane informacje także do oszacowania ich wypłacalności i prawdopodobieństwa kontynuacji współpracy w długim okresie[[10]](#footnote-10).

Choć najnowsze regulacje na temat weryfikacji wiarygodności w kontekście podatkowym w pewien sposób przenoszą ryzyko związane z kontaktami handlowymi z nieuczciwymi podmiotami, które występują w obiegu gospodarczym na podatników, którzy należycie wywiązują się ze swoich obowiązków jako płatników to w kontekście weryfikacji finansowej kontrahentów, wypłacalności, bezpieczeństwa i ciągłości współpracy można dostrzec również pozytywne aspekty wprowadzonych zmian. Mimo tego, że na podmioty prowadzące działalność gospodarczą nałożono dodatkowe obowiązki, które w teorii nie powinny sprzyjać rozwojowi firmy i ekspansji działalności to wydaje się, że narzucony obowiązek weryfikacji kontrahentów może jednak dostarczyć wymiernych korzyści biznesowych. Skoro według prawa muszą one i tak w przynajmniej ograniczonym zakresie zbadać sytuację swoich partnerów biznesowych, to zebrane na ich temat informacje mogą wykorzystać do powiększenia wiedzy na ich temat a następnie poprawy procesów decyzyjnych. Zwłaszcza, że zostały im dostarczone przystępne źródła wiedzy na ten temat.

W celu wykonania analizy kontrahentów i zachowania obowiązku należytej staranności przedsiębiorcom zostały udostępnione stosowne bazy i narzędzia takie jak między innymi: rejestr podatników VAT, rejestry GUS, KRS i CEIDG, baza sprawozdań finansowych KRS oraz baza Urzędu Regulacji Energetyki. Oprócz tradycyjnego wykorzystania tych informacji do weryfikacji kontrahentów takie dane po odpowiednim przetworzeniu mogą okazać się przydatne do oceny kondycji finansowej pozostałych uczestników rynku i wykonania predykcji na temat potencjalnego bankructwa lub zakończenia działalności z innych przyczyn. Nie należy zapominać również o możliwości wykorzystania dodatkowych informacji, które mogą zostać zebrane poprzez pobranie ich bezpośrednio ze stron internetowych innych witryn przy wykorzystaniu metod automatyzującego oprogramowania. Ważnym argumentem za wykorzystaniem udostępnionych danych w tym celu jest to, że wszystkie omówione źródła są darmowe i można wykorzystać je w zautomatyzowany sposób przy użyciu dostępnych technologii, z których część już została udostępniona nieodpłatnie w publicznych repozytoriach. Wydaje się, że skoro przedsiębiorcy mają nieograniczony dostęp do tych informacji, to warto jest spróbować wykorzystać je w sposób, który oprócz standardowego zastosowania dostarczy im wartości dodanej i usprawni procesy decyzyjne poprzez możliwość oceny wypłacalności podmiotów, które znajdują się w ich otoczeniu biznesowym.

## I.4 Korzyści z możliwości automatycznej predykcji bankructwa przedsiębiorstw.

Dostatek różnorodnych danych, ich zadowalające pokrycie oraz duży wolumen zgromadzonych obserwacji historycznych sprawiają, że możliwe jest podjęcie próby budowy modelu regresyjnego lub klasyfikacyjnego, którego zadaniem jest rozróżnianie przedsiębiorstw o wysokim prawdopodobieństwie bankructwa od podmiotów, które znajdują się w dobrej sytuacji finansowej i w rozważanym okresie są w niewielkim stopniu narażone na ryzyko niewypłacalności lub zakończenia działalności z innych przyczyn. Przy zastosowaniu rozwiązań technologicznych, które umożliwiają połączenie procesu wprowadzenia danych wejściowych z wykorzystaniem modelu, możliwe jest stworzenie narzędzia, które będzie automatyzować to zadanie i wykonywać je przy minimalnym wkładzie ludzkim. Proces będzie zatem obejmować pobranie potrzebnych informacji, przeprocesowanie ich, wykorzystanie jako dane wsadowe do zbudowanego klasyfikatora i generowanie wyników, które będą dostarczać informacji i wartości dodanej decydentom.

W ten sposób skonstruowane zostanie narzędzie, które przedsiębiorcy mogą zastosować do samoczynnej weryfikacji podmiotów takich jak ich klienci, kontrahenci czy konkurenci. Ograniczenie roli użytkownika do wprowadzenia identyfikatorów analizowanych podmiotów, którymi mogą być numery NIP, REGON czy KRS oraz uruchomienia narzędzia sprawia, że mogą korzystać z niego nawet osoby o ograniczonej wiedzy technicznej. W świetle regulacji, którymi są objęci polscy przedsiębiorcy możliwość zastosowania darmowych źródeł danych do uzyskania cennych informacji na temat ryzyka niewypłacalności podmiotów w ich otoczeniu gospodarczym wydaje się przydatną funkcjonalnością, która dostarcza przedsiębiorstwu dodatkowej wartości dodanej.

Posiadanie podobnego narzędzia pozwala przyspieszyć proces weryfikacji partnerów biznesowych przy niewielkim zaangażowaniu człowieka. Po uwzględnieniu funkcjonalności automatyzacji pobierania i procesowania danych oraz wcześniejszym wytrenowaniu modelu wydajność takiego narzędzia jest wysoka i pozwala na masowe weryfikacje w zasadzie dowolnej liczby firm. Zasadniczym ograniczeniem jest jedynie dostępność informacji na temat analizowanych podmiotów i ich zadowalające pokrycie oraz czas maszynowy i przepustowość łączy, które są potrzebne do pobrania danych niezbędnych do wykonania analizy. Czas oczekiwania na wyniki w większości sprowadza się do konieczności pobrania danych, co przy wykorzystaniu omówionych wcześniej zbiorów danych z powodu zastosowania wydajnych technologii jak relacyjne bazy danych i API nie jest czasochłonne. Samo pobranie, przygotowanie danych do analizy i nadanie oceny ryzyka przez wytrenowany model po zaprojektowaniu i zbudowaniu odpowiedniego procesu odbywa się w sposób zautomatyzowany i nie jest wymagana dodatkowa ludzka praca.

Narzędzie pozwala tym samym zaoszczędzić czas osób, które przeprowadziłyby taką analizę w sposób ręczny. Dzięki automatyzacji znacznej części pracy takie osoby mają możliwość poświęcenia większej uwagi podmiotom o wysokiej ocenie ryzyka nadanej przez algorytm, czy dokładnej weryfikacji czynników ryzyka, które można uznać za potencjalne zagrożenia w kontekście niskiej wiarygodności finansowej i reputacyjnej danego podmiotu. Wydajność takiej analizy zostaje zatem podniesiona dzięki zastosowaniu omawianego narzędzia. Analiza zostaje również pozbawiona ryzyka wystąpienia błędów ludzkich lub przynajmniej zostają one ograniczone. Pomyłki wynikające na przykład ze zmęczenia, braku wiedzy czy uprzedzeń analityka zostają wyeliminowane przez zautomatyzowane rozwiązanie. Ilość czasu potrzebnego na wykonanie takiej analizy może zostać ograniczona lub czas ten może być poświęcony na wykonanie analizy w sposób bardziej staranny niż miałoby to miejsce bez informacji dostarczanych przez narzędzie. Zatem, korzyścią którą przedsiębiorstwo otrzymuje z automatyzacji weryfikacji kontrahentów jest oszczędność czasu i wykorzystanie go w bardziej wydajny sposób.

Podejście do realizacji korzyści z tytułu wprowadzenia omawianego rozwiązania zależy od podejścia i apetytu na ryzyko kierownictwa, które z jednej strony może starać się ograniczyć koszty czasu poświęconego na weryfikację kontrahentów a z drugiej może starać się podnieść jej jakość przy jednoczesnym utrzymaniu stałego nakładu czasu i wydatków na ten cel w porównaniu do okresu przed wprowadzeniem nowego rozwiązania. Skorzystanie z takiego narzędzia pozwala przedsiębiorstwu ograniczyć koszty prowadzonej działalności i stać się bardziej konkurencyjnym podmiotem jednocześnie podnosząc jakość świadczonych usług i zabezpieczając się przed potencjalnymi stratami z powodu współpracy z kontrahentami o niskiej wiarygodności.

Z perspektywy osób odpowiedzialnych za kierowanie przedsiębiorstwem automatyzacja i poszerzenie weryfikacji kontrahentów o możliwość predykcji ich bankructwa daje dwie zasadnicze korzyści. Po pierwsze, zmniejszone zostaje ryzyko podjęcia współpracy z podmiotem o słabej kondycji finansowej, który może w najbliższej przyszłości mieć problemy z dotrzymywaniem terminów płatności i spłacaniem zobowiązań, lub nawet może być postawiony w stan upadłości lub zlikwidowany. Nawiązanie współpracy z takim podmiotem w charakterze klienta lub dostawcy mogłoby wpłynąć negatywnie na funkcjonowanie przedsiębiorstwa. Znajomość jego sytuacji finansowej pozwala jednak uniknąć takiego scenariusza. Podobne rozwiązanie można zastosować także w kontekście aktualnych partnerów biznesowych. Możliwe staje się także przeprowadzanie cyklicznych nisko kosztowych (np. miesięcznych, kwartalnych) weryfikacji firm, z którymi aktualnie prowadzona jest współpraca i zaimplementowanie systemu ostrzegawczego, który w przypadku wykrycia sytuacji wskazujących na zwiększenie się ryzyka niewypłacalności (np. wykrycie straty finansowej w ostatnim okresie sprawozdawczym, spadek wartości aktywów lub innego wcześniej zdefiniowanego zdarzenia, które stanowi sygnał ostrzegawczy dla niekorzystnych zmian) będzie w porę informował osoby kierujące działalnością i pozwalał im na ograniczenie lub zakończenie współpracy z podmiotem zagrożonym bankructwem. Dzięki temu zastosowanie narzędzia pozwala wyeliminować lub co najmniej ograniczyć straty finansowe z powodu zaangażowania się lub kontynuowania niekorzystnej współpracy. Podobne korzyści dotyczą ciągłości łańcucha dostaw, która zostaje zabezpieczona.

Drugą zasadniczą korzyścią z wdrożenia omawianego systemu jest ograniczenie możliwości podjęcia współpracy z podmiotami nieuczciwymi i zmniejszenie ryzyka nadużyć. Niedopełnienie obowiązku zachowania należytej staranności może nieść poważne konsekwencje dla przedsiębiorstwa w postaci utraty reputacji i zaufania, kar finansowych, czy interwencji regulatorów i podmiotów odpowiedzialnych za nadzorowanie rynku co może skutkować cofnięciem bądź zawieszeniem zezwolenia lub koncesji na wykonywanie działalności. Skoro przedsiębiorcy są prawnie zobowiązani do weryfikacji kontrahentów, przydatnym rozwiązaniem powinno być zautomatyzowanie tego procesu w celu uwolnienia części zasobów przedsiębiorstwa do wykonywania innych aktywności, które będą w bardziej wydajny sposób przekładać się na pozytywne wyniki finansowe firmy.

Wdrożenie rozwiązania pozwalającego na automatyczną weryfikację firm i predykcję ich sytuacji finansowej z wyprzedzeniem jest zdaniem autora korzystne dla wszystkich uczestników rynku. Zwłaszcza, jeśli narzędzie będzie mogło w całości opierać się na otwartych źródłach danych, co będzie skutkować brakiem opłat licencyjnych lub subskrypcyjnych. Dzięki otwartemu dostępowi do takich informacji zmniejszona zostaje asymetria informacji wśród uczestników rynku. Przez to staje się on bardziej konkurencyjny, podmioty dzięki lepszej świadomości na temat sytuacji klientów i konkurentów i możliwości przewidzenia ich sytuacji finansowej w niedalekiej przyszłości muszą starać się podnosić jakość świadczonych usług zamiast wykorzystywać przewagę posiadaną dzięki dostępowi do niszowych informacji. Zmniejszony zostaje wpływ tego kontrowersyjnego aspektu przewagi konkurencyjnej podmiotów które stać na dostęp do takich informacji w porównaniu do podmiotów, które z dowolnych przyczyn nie mają do nich dostępu. Taka sytuacja jest korzystna szczególnie dla nowych podmiotów, ponieważ zmniejszona zostaje bariera rywalizacji konkurencji do wejścia na rynek. Dodatkowo, podmioty rozważające dołączenie do rynku będą w stanie efektywniej oszacować wynikające z tego korzyści i zagrożenia. Natomiast, firmy prowadzące już działalność powinny dzięki temu w większym stopniu się starać się osiągać dobre wyniki finansowe poprzez poprawę jakości dostarczanych produktów i usług. Dzięki temu powinien podnosić się poziom konkurencyjności, co będzie pozytywnie wpływać na jakość końcowych produktów oferowanych końcowym konsumentom[[11]](#footnote-11).

Kolejnym argumentem przemawiającym za zastosowaniem zbudowanego narzędzia jest możliwość wykorzystania czasu uczestników rynku w bardziej wydajny sposób. Wprowadzone automatyzacje podczas usprawnienia procesu weryfikacji partnerów biznesowych i oceny ich sytuacji finansowej sprawią, że zainteresowane osoby będą miały więcej czasu na inne zadania i aktywności co przyczyni się do poprawy jakości i konkurencyjności całego sektora biznesowego, a to będzie miało pozytywne przełożenie dla klientów końcowych.

Korzyści z wdrożenia proponowanego narzędzia i jego zastosowania w praktyce przekładają się zatem w realny, pozytywny sposób na sposób prowadzenia działalności biznesowej, co w opinii autora stanowi uzasadnioną przyczynę do podjęcia jego budowy i sprawdzenia czy jego skuteczność będzie zadowalająca z perspektywy podmiotów, które będą chciały wykorzystać je do usprawnienia zarządzania ryzykiem operacyjnym, reputacyjnym i finansowym.

# II. Omówienie danych wykorzystanych do przeprowadzenia analizy.

## II.1 Źródła danych wykorzystane do przeprowadzenia analizy.

Wykorzystanie danych o wysokiej jakości i pokryciu jest kluczowym wymaganiem do zbudowania modelu predykcyjnego, który będzie charakteryzował się zadowalającą jakością predykcji i będzie nadawał się do rozwiązania realnego problemu biznesowego. Uwzględnienie danych dobrze wyjaśniających zmienność badanego zjawiska jest niezbędnym warunkiem do rozróżniania poszczególnych klas zmiennej celu lub skutecznej predykcji na temat wartości zmiennej ciągłej. Proces budowy modelu nawet jeżeli zostanie przeprowadzony poprawnie ale przy wykorzystaniu danych o niskiej jakości, które nie tłumaczą modelowanego zjawiska w zadowalającym stopniu może z dużym prawdopodobieństwem zakończyć się porażką. W celu uniknięcia takiego scenariusza należy możliwie szybko ocenić potencjał zgromadzonych danych. Czynność ta została wykonana w celu weryfikacji, czy dostępne informacje pozwolą na zbudowanie modelu predykcyjnego, którego celem jest prognoza bankructwa przedsiębiorstw.

Głównym źródłem danych wykorzystanym do pobierania informacji na temat podmiotów jest wywiadownia gospodarcza EMIS, do której dzięki porozumieniu z dostawcą mają dostęp studenci Szkoły Głównej Handlowej. Firma dostarcza informacji gospodarczych dla 147 państw z obszaru rynków rozwiniętych i rozwijających się. Składowane informacje są przechowywane w wygodnych formatach, zarówno w postaci tabelarycznej jak i plikach w formacie XML i JSON. Użytkownicy mają możliwość składania zapytań przez wyszukiwarkę lub bezpośrednio przez udostępnione API. Baza danych zawiera informacje na temat danych rejestrowych. poszerzonych dodatkowo o informacje specyficzne dla firmy, w tym dane na temat liczby zatrudnionych osób oraz importowanych i eksportowanych dóbr. Oprócz tego udostępniane są informacje prasowe na temat firm oraz dane finansowe pozyskiwane z publikowanych przez firmy sprawozdań finansowych. Dane finansowe są poszerzane o dodatkowe wskaźniki opisujące kondycję finansową podmiotu w dostępnych latach finansowych oraz w ograniczonym zakresie o wskaźniki przedstawiające działalność firmy w porównaniu do całej branży[[12]](#footnote-12).

Omawiana wywiadownia gospodarcza jest szczególnie istotna z punktu widzenia udostępnianych danych finansowych, które są agregowane i udostępniane do użycia w przystępnej formie, w postaci tabelarycznej. Jak pokazują badania na temat przyczyn upadłości przedsiębiorstw oraz poprzednie podobne zrealizowane projekty informacje na ten temat są kluczowe z punktu widzenia predykcji niewypłacalności. Informacje z EMIS zostały pozyskane dzięki udostępnionemu API, które poprzez wysyłanie odpowiednio skonstruowanych zapytań pozwala pobierać dane rejestrowe i generalne informacje na temat firmy oraz dane finansowe za poszczególne lata finansowe (jeśli są takowe dostępne).

Kolejnym źródłem danych wykorzystanym do przeprowadzenia analizy są zasoby Głównego Urzędu Statystycznego, który udostępnia możliwość pozyskiwania danych z 3 baz poprzez wysyłanie zapytań do oddzielnych API. Na potrzeby budowanego narzędzia wykorzystano dane z bazy REGON i bazy jednostek regionalnych. Z bazy REGON pozyskano informacje rejestrowe, które agregują dane z baz CEIDG oraz KRS. Z bazy danych jednostek lokalnych pozyskano dane globalne agregowane na poziomie jednostek administracyjnych, które zawierają między innymi informacje na temat poziomu rozwoju i zamożności poszczególnych jednostek administracyjnych oraz regionów w Polsce, które można dołączyć do badanych podmiotów na podstawie adresu prowadzonych przez nich działalności gospodarczej, które są wymagane do zarejestrowania firmy.

W następnej kolejności wykorzystano dane z rejestru podatników VAT (tzw. białej listy podatników VAT) udostępnionego przedsiębiorcom przez Ministerstwo Finansów. Dane zawierają informacje na temat statusu podatnika VAT badanych firm z możliwością sprawdzenia go w dowolnym dniu począwszy od 01.01.2015 r. aż do aktualnej daty. Dzięki funkcjonalności sprawdzenia statusu podatnika VAT i pozostałych udostępnianych informacji w konkretnym dniu w przeszłości możliwe jest sprawdzenie informacji dla podmiotów, które zakończyły już działalność. Dla takich podmiotów zapytanie składano na zadany okres czasu poprzedzający zakończenie działalności. W bazie danych podawane są także podstawowe dane rejestrowe oraz ogólne informacje na temat płatności podatków, w tym lista dozwolonych rachunków wykorzystywanych do zlecania i otrzymywania płatności oraz przyczyny odmowy rejestracji, zawieszenia lub wykreślenia z rejestrów. Składowane są także dane odnośnie rachunków wirtualnych. Ograniczeniem liczbowym udostępnionego API jest możliwość pobierania danych dla 300 podmiotów dziennie z jednego adresu IP. Możliwe jest jednak pozyskanie danych bezpośrednio ze strony internetowej poprzez zastosowanie tzw. crawlera – skryptu automatyzującego pobieranie danych, który imitując działanie człowieka wyszukuje oraz zapisuje sprawdzone informacje na temat poszczególnych firm. W celu ominięcia ograniczenia i jednorazowego pobrania danych dla większej liczby podmiotów wykorzystanych do analizy zastosowano metodę web-scrapingu do pobrania informacji bezpośrednio z udostępnionej strony.

Ostatnim wykorzystanym źródłem danych jest baza danych Urzędu Regulacji Energetyki (URE), która zawiera informacje na temat zezwoleń na prowadzenie działalności, licencji i koncesji. Prowadzony jest rejestr aktywnych, wygasłych koncesji oraz informacje na temat odmowy ich udzielenia lub cofnięcia. Dane są zagregowane dla wszystkich podmiotów dostępnych w Polsce, które od początku istnienia rejestru otrzymały jakiekolwiek zezwolenia tego rodzaju. Po pobraniu danych w formacie XML i ich odpowiednim przetworzeniu możliwe jest pozyskanie informacji na ten temat dla konkretnych podmiotów poprzez użycie ich identyfikatorów NIP, REGON lub KRS.

Zebranie danych rozpoczęto od pobrania podstawowych informacji – metadanych - na temat dostępnych podmiotów w poszczególnych bazach. W pierwszej kolejności w bazie danych EMIS zidentyfikowano liczbę i status operacyjny firm, dla których odłożono informacje. Po identyfikacji podmiotów prowadzących działalność operacyjną w dniu rozpoczęcia gromadzenia danych oraz firm zamkniętych i zlikwidowanych w ostatnich latach stworzona została początkowa baza identyfikatorów działalności gospodarczych w poszczególnych statusach. Kolejnym krokiem było pobranie szczegółowych danych w dostępnych źródłach dla poszczególnych podmiotów posługując się ich identyfikatorami do utworzenia i wysłania zapytań do udostępnionych API w przypadku baz danych EMIS, GUS i URE lub do pozyskiwania informacji bezpośrednio ze strony internetowej w przypadku rejestru podatników VAT. Wykonano również niezbędny krok do określenia daty zamknięcia lub likwidacji podmiotów, które na dzień zebrania danych zakończyły prowadzoną działalność z dowolnych przyczyn. W tym celu wykorzystano dane pobrane z bazy GUS, które udostępniane są w najbardziej przystępnej formie i ich pokrycie jest najbardziej zadowalające. Dzięki temu możliwe jest określenie przewidywanej zmiennej celu jako zamknięcia firmy w określonym horyzoncie czasowym. Firmy zamknięte zaliczono do zbioru obserwacji pozytywnych, natomiast firmy prowadzące operacyjną jako obserwacje negatywne. Zgromadzone zmienne niezależne przypisane poszczególnym podmiotom są zatem interpretowane relatywnie w rozważanym horyzoncie czasowym do (potencjalnego) zakończenia działalności.

Wybór okresu, w którym przeprowadzona ma zostać predykcja zamknięcia (bankructwa) przedsiębiorstwa jest kluczowy z punktu widzenia poprawnego funkcjonowania narzędzia. Zbyt krótki okres czasu mimo wysokiej jakości predykcji, kiedy dane z dużym prawdopodobieństwem będą wskazywać możliwość wystąpienia niewypłacalności lub innych przyczyn poprzedzających zamknięcie firmy może okazać się niewystarczający dla osób wykonujących analizę do podjęcia działań, które przygotują firmę do odpowiedniej reakcji na zakończenie działalności własnej lub pozostałych podmiotów w ich otoczeniu gospodarczym. Z drugiej strony przewidywanie zamknięcia działalności w zbyt odległym horyzoncie czasowym powinno odznaczać się gorszą jakością prognozy z powodu dużej ilości czasu pozostałego do momentu realizacji prognozy i większej niepewności co do losu firmy. W tym okresie podmiot, którego dotyczy predykcja może zmienić swoją sytuację (zarówno na lepsze jak i na gorsze), przez co predykcja będzie inna od stanu faktycznego. Na takie zachowanie modeli predykcyjnych i pogorszenie ich skuteczności wraz z wydłużeniem rozważanego okresu wskazują dotychczasowe podobne analizy wykonane w przeszłości.

Były one wykonywane zazwyczaj przeprowadzone w okresie 12 i 24 miesięcy od bieżącej daty. Modelowanym zjawiskiem była zatem binarna zmienna celu, która opisuje czy firma zakończy działalność, czy będzie ją kontynuować odpowiednio za 12 i 24 miesięcy od dnia wykonania prognozy. Zastosowanie właśnie takich interwałów jest częściowo spowodowane ograniczeniami technicznymi z powodu okresów, w których są udostępniane sprawozdania finansowe i dane na temat kondycji finansowej przedsiębiorstw. W podobnych realizowanych analizach w pierwszej kolejności przyjmowano badany okres 12 miesięcy od bieżącej daty lub od daty aktualności najnowszych danych finansowych. Jeśli wyniki okazywały się zadowalające sprawdzano kolejny, dłuższy horyzont czasowy.

Autor zastosował podobne podejście i za modelowaną zmienną przyjęto zamknięcie działalności w okresie do 12 miesięcy od dnia pobrania aktualnie dostępnych danych i wykonania prognozy zamknięcia firmy. W opinii autora jest to w większości przypadków wystarczająco długi okres, który pozwoli firmom znacząco ograniczyć ryzyko związane z zamknięciem własnej działalności z powodu niewypłacalności lub działalności innego podmiotu w ich otoczeniu gospodarczym, czy też wykorzystać możliwości które powstają dzięki zamknięciu działalności podmiotów konkurencyjnych. Wybrany interwał czasowy stanowi swego rodzaju kompromis pomiędzy jakością predykcji a ilością czasu pozostałego do zamknięcia działalności[[13]](#footnote-13).

## II.2 Struktura i przekształcenia zebranych danych.

Po pozyskaniu danych z poszczególnych wykorzystanych źródeł została przeprowadzona rekoncyliacja na podstawie unikalnych identyfikatorów nadawanym podmiotom prowadzącym działalność gospodarczą. Posłużono się do tego celu numerami NIP nadawanymi płatnikom podatków. W przypadku nielicznych sytuacji, gdy z powodu braku danych lub ich niespójności niemożliwe okazało się złączenie danych przy użyciu NIP wykorzystano numery REGON, które są nadawane podmiotom prowadzącym działalność gospodarczą w bazie danych GUS. Podmioty prowadzące w dalszym ciągu działalność gospodarczą (obserwacje negatywne) zostały odróżnione od podmiotów zamkniętych (obserwacje pozytywne) poprzez wykorzystanie danych na temat ich statusu operacyjnego dostępnego w EMIS oraz dat zamknięcia działalności i wykreślenia z rejestrów dostępnych w zasobach GUS. W pierwszej kolejności status operacyjny badanych przedsiębiorstw był weryfikowany na podstawie danych dostępnych w EMIS dla wszystkich podmiotów dostępnych w bazie danych. Jeśli status zamkniętej działalności wystawiony przez EMIS był potwierdzony niepustą datą zakończenia lub likwidacji działalności bądź wykreślenia podmiotu z rejestru REGON obserwacja otrzymywała status zamkniętej działalności gospodarczej. Jako datę upadłości przyjęto datę najwcześniejszego z wyżej wymienionych zdarzeń. Zebrane informacje zostały złączone w zbiór danych zawierający informacje ze wszystkich wykorzystanych źródeł[[14]](#footnote-14).

Łącznie zebrano dane dla 54 664 i 14 814 podmiotów odpowiednio zamkniętych i zlikwidowanych, które były dostępne w bazie danych EMIS i na dzień rozpoczęcia pobierania danych (15.04.2010 r.) ich status operacyjny wskazywał, że są zamknięte. Liczba obserwacji pozytywnych wyniosła zatem 69 478. W celu uzupełnienia populacji o obserwacje negatywne i porównania cech, którymi charakteryzowały się dwie klasy podmiotów pobrano również informacje dla aktywnych firm. Łącznie zebrano dane dla 177 427 firm, które na dzień rozpoczęcia pobierania danych pozostawały w statusie operacyjnym i kontynuowały działalność gospodarczą. Łącznie zatem pobrano dane dla 247 175 firm.

Z powodu dużego wolumenu pobranych danych niemożliwe było sprawdzenie ich kompletności i weryfikacji poprawności formatu dla każdego podmiotu osobno. Z tej przyczyny ze zbioru w automatyczny sposób odrzucono podmioty, dla których dane z przynajmniej jednego źródła spośród EMIS i GUS zostały pobrane w nieprawidłowym formacie lub stwierdzono ich brak. Z powodu mniejszej wagi z punktu widzenia liczby wykorzystanych zmiennych, ich istotności oraz pokrycia danych podjęto decyzję o zachowaniu obserwacji, dla których nie udało się pobrać informacji dostępnych w zasobach wykazu podatników VAT i URE. Główną przyczyną błędów podczas procesowania danych były braki niektórych punktów danych w poszczególnych wartościach zagnieżdżonych plików JSON, lub błędne typy danych. W celu minimalizacji utraty informacji w przypadku występowania przeszkód, które obejmowały zasięgiem dużą liczbę podmiotów zostały one rozwiązywane, dzięki czemu utrata danych była na możliwie niskim poziomie. W sytuacji, gdy błędy obejmowały pojedyncze pliki obserwacje, których one dotyczyły były odrzucane z powodu dużego nakładu czasu potrzebnego do rozwiązania problemów. W niektórych przypadkach też problemy nie były możliwe do rozwiązania. Decyzję taką podjęto ze świadomością, że zebrana liczba obserwacji jest wystarczająca do przeprowadzenia analizy.

Ostatecznie z powodu braku danych lub błędów podczas pobierania danych odrzucono odpowiednio 457 i 13 podmiotów zamkniętych oraz aktywnych. Do dalszej analizy i transformacji zbioru danych wykorzystano zatem odpowiednio 176 970 i 69 465 podmiotów, które zakończyły prowadzoną działalność i kontynuowały ją w dniu rozpoczęcia pobierania danych.

Następnym krokiem było wykonanie transformacji zbioru danych w celu dodania do niego nowych zmiennych utworzonych na podstawie surowych danych w celu maksymalnego wykorzystania pobranych informacji. Do zbioru danych dołączono zmienne opisujące rodzaj prowadzonej działalności głównej oraz działalności dodatkowych (jeśli firma takie deklarowała) według kodów PKD i ich ryzyka z punktu widzenia podatkowego, wielkość kapitału zakładowego, wiek podmiotu (okres czasu prowadzenia działalności), reprezentację podmiotów, adres rejestracji pod adresem lokalu mieszkalnego lub biura wirtualnego, częstotliwość zmiany nazw w przeszłości, informacje odnośnie posiadanych oddziałów i placówek, informacje na temat danych kontaktowych i posiadanych stron internetowych, importu i eksportu towarów, informacji odnośnie posiadanych koncesji i zezwoleń.

Szczególnie istotną częścią transformacji danych było dodanie do zbioru wskaźników finansowych wykorzystanych w modelach, które zostały wykorzystane przez innych autorów podobnych projektów. Do atrybutów na podstawie pobranych danych finansowych dodano zmienne zbudowane na podstawie wskaźników z modeli Edwarda Altmana, Tomasza Korola i Błażeja Prusaka. Opis i pochodzenie poszczególnych atrybutów przedstawiono w części pracy poświęconej analizie eksploracyjnej danych[[15]](#footnote-15).

W celu zapewnienia równowagi proporcji liczebności obserwacji z poszczególnych klas ze zbioru podmiotów prowadzących działalność (obserwacje negatywne) na dzień wykonania analizy wybrano liczbę przedsiębiorstw zbliżoną do liczby firm, które zakończyły działalność gospodarczą (obserwacje pozytywne). Zabieg wykonano w celu wyrównania wag poszczególnych klas w badanej populacji. Niezbalansowane klasy zmiennej celu mogą stanowić problem w przypadku budowy modelu klasyfikacyjnego, ponieważ większość algorytmów uczenia maszynowego radzi sobie najlepiej z rozróżnianiem poszczególnych klas kiedy ich proporcje są do siebie stosunkowo zbliżone. Wynika to z faktu, że większa część algorytmów klasyfikacyjnych została zaprojektowana w ten sposób, żeby maksymalizować ogólną skuteczność i minimalizować przyjęte kryterium błędu a nie maksymalizować skuteczność przewidywania obserwacji pozytywnych, co byłoby właściwym kryterium oceny przydatności modelu w kontekście prowadzenia działalności biznesowej. W przypadku pozostawienia proporcji klas w początkowym stadium przygotowania danych ogólna skuteczność zbudowanego modelu prawdopodobnie byłaby wyższa. Zazwyczaj natomiast ceną zmniejszenia liczebności przeważającej klasy jest spadek skuteczności predykcji obserwacji pozytywnych, czyli upadłości przedsiębiorstw – co jest zasadniczym celem projektu. Drugą negatywną konsekwencją zmniejszenia próby jest utrata informacji, których dostarczają usuwane obserwacje. Nowoczesne rozwiązania optymalizujące wyrównanie liczebności klas pozwalają jednak zoptymalizować ten proces tak, żeby utrata informacji była możliwie najniższa[[16]](#footnote-16).

W tym wypadku z powodu dużej proporcji liczby podmiotów w stanie operacyjnym (71,71%) do podmiotów z zakończoną działalnością (28,29%) zmniejszono rozmiar próby obserwacji negatywnych – klasy o większym udziale w zebranej populacji generalnej o liczebności ponad 245 000 podmiotów. W celu wyboru optymalnej próby odpowiadającej rozkładom w grupie podmiotów zlikwidowanych i zamkniętych przeprowadzono grupowanie przedsiębiorstw na podstawie rodzaju prowadzonej działalności oraz formy prawnej firmy. Na tej podstawie przeprowadzono warstwowanie i wykonano redukcję próby poprzez zastosowanie algorytmu skondensowanych najbliższych sąsiadów (ang. Condensed Nearest Neighbors). Jest to metoda zmniejszania rozmiaru próby, która poszukuje takich podzbiorów próbek z generalnej populacji, które nie powodują utraty wydajności modelu. Do analizy wybierany jest podzbiór, który używany jako przechowywany zbiór odniesień dla reguły najbliższych sąsiadów pozwala na prawidłową klasyfikację wszystkich pozostałych obserwacji w wybranym podzbiorze. Wybrana metoda minimalizuje utratę informacji spowodowaną zmniejszeniem próby oraz zapewnia zbliżone rozkłady zmiennych, według których wykonywane jest warstwowanie. Celem jest zatem zapewnienie podobnych kluczowych cech w obu klasach, dzięki czemu możliwe jest zbadanie wpływu pozostałych zmiennych i sprawdzenie jaki one wpływ mają na badane zjawisko[[17]](#footnote-17).

Ostatecznie, po odrzuceniu ze zbioru przedsiębiorstw, dla których braki danych były zbyt duże aby możliwe było skuteczne zastosowanie zebranych dla nich punktów danych oraz po zmniejszeniu udziału przedsiębiorstw prowadzących w dalszym ciągu działalność gospodarczą wyżej omówioną metodą w ostatecznym zbiorze danych znalazło się 52 190 podmiotów. Główną przyczyną tak znacznego okrojenia danych było usunięcie obserwacji, dla których wystąpiły braki w przypadku danych finansowych, lub nie udało się ich zebrać. W większości przypadków wynikało to z formy prawnej firm i opublikowania okrojonych sprawozdań finansowych. W niektórych przypadkach w EMIS nie były dostępne sprawozdania finansowe za ostatni rok przed rozpoczęciem analizy, które powinny być wykorzystane do utworzenia zmiennych. Na badaną populację złożyło się 33 455 (64,1%) obserwacji negatywnych oraz 18 735 (35,9%) obserwacji pozytywnych co zapewnia wystarczającą liczbę obserwacji oraz zbalansowanie klas zmiennej celu do przeprowadzenia analizy.

Po zakończeniu procesu wyboru odpowiedniej próby nastąpiła selekcja zmiennych w otrzymanym końcowym zbiorze danych. W pierwszej kolejności ze zbioru odrzucone zostały zmienne z dużym udziałem braków danych. Usunięto atrybuty, dla których pokrycie zebranych danych nie było w opinii autora wystarczające. Jako próg odcięcia dla maksymalnego dopuszczalnego udziału braków danych przyjęto wartość 30%. Zmienne z ich większym udziałem zostały usunięte ze zbioru. Zabieg ten wykonano w przypadku atrybutów numerycznych, gdzie imputacja braków dowolną metodą nie byłaby uzasadniona, ponieważ występowały one głównie w przypadku zmiennych utworzonych na podstawie danych finansowych pobranych ze sprawozdań. W tym wypadku próba imputacji byłaby w opinii autora nieuzasadniona. Jest tak dlatego, że istnieje wysokie prawdopodobieństwo, że wprowadzone dane nie oddawałaby rzeczywistych wartości dla brakujących obserwacji. W przypadku cech kategorycznych braki danych potraktowano jako oddzielne kategorie i utworzono dla takich obserwacji osobną klasę reprezentującą brak danych w celu uniknięcia utraty informacji i zmaksymalizowania potencjału zgromadzonych danych[[18]](#footnote-18).

W dalszej kolejności przeprowadzono zgrupowanie klas o niskiej reprezentacji dla zmiennych kategorycznych i wartości takie złączono w osobną kategorię. Jako krytyczną wartość minimalną wskazującą na konieczność przeprowadzenia agregacji przyjęto próg odcięcia o wartości 3% liczebności frakcji. Kategorie o mniejszym udziale zostały złączone i nadano im oddzielną kategorię reprezentującą wartości, które nie były wystarczająco licznie reprezentowane, żeby utworzyć dla nich samodzielne klasy. Zastosowanie takiego rozwiązania pozwala na zredukowanie rozmiarów zbioru wykorzystanego do analizy powstałego w procesie kodowania zero-jedynkowego zmiennych i skrócenie czasu maszynowego koniecznego do wykonania obliczeń.

Kolejnym krokiem poprzedzającym przeprowadzanie analizy eksploracyjnej było zidentyfikowanie wartości odstających w zbiorze danych. Celem takiego zabiegu było wykrycie ewentualnych anomalii, poprawa jakości danych, ale też ich lepsze zrozumienie. Obecność obserwacji odstających może nieść ze sobą pewne problemy. Przede wszystkim w przypadku zmiennych ciągłych w znacznym stopniu może zmienić się średnia z badanej próby, podczas gdy mediana zmienia się jedynie w stopniu wynikającym z dodania dodatkowej obserwacji – zmiana następuje niezależnie od jej wartości, jedynie od tego, czy nowa obserwacja znajduje się powyżej czy poniżej mediany. Przez to metody parametryczne, wykorzystujące średnią lub kowariancję mogą podlegać zaburzeniom związanym z występowaniem obserwacji odstających, przez co model zbudowany na takich danych może w gorszym stopniu tłumaczyć rzeczywistość lub robić to w sposób nieprawidłowy.

W tym celu dla zmiennych ciągłych wygenerowano wykresy rozrzutu w celu identyfikacji takich obserwacji metodą wizualną. Zastosowano także 2 metody numeryczne - statystykę Z-score oraz metodę detekcji przy wykorzystaniu rozstępu międzykwartylowego. Test Z-score sprawdza o ile odchyleń standardowych pomiar dla konkretnej obserwacji odbiega od średniej. Jest wykorzystywany do określenia dokładnej lokalizacji każdej z obserwacji w rozkładzie. Według literatury przedmiotu przyjmuje się, że obserwacje, dla których wynik statystyki przekracza wartość 3 są traktowane jako obserwacje odstające. W przypadku metody rozstępu międzykwartylowego (wartości znajdujące się pomiędzy pierwszym a trzecim kwartylem) przyjmuje się, że obserwacje, których wartość przekracza o 1,5 (lub 3 w przypadku decyzji analityka lub specyfiki danych) dolną lub górną wartość przedziału międzykwartylowego są traktowane jako obserwacje odstające.

Po wykonaniu analizy rozkładów zidentyfikowane zostały wartości odstające. Pomiary odstające, które nie mogą występować w rzeczywistości, czyli najprawdopodobniej powstały na skutek błędów podczas wprowadzania lub przetwarzania danych (np. daty upadłości zarejestrowane w przyszłości, ujemne przychody, kapitał zakładowy poniżej ustawowej wartości minimalnej czy ujemna liczba pracowników) zostały usunięte ze zbioru. W przypadku braku racjonalnego uzasadnienia do usunięcia takich punktów danych z powodu występowania ich w rzeczywistości dane obserwacje pozostawiono w zbiorze. W celu lepszej prezentacji graficznej danych podjęto decyzje o wyłączeniu obserwacji odstających wykrytych metodą Z-score podczas przeprowadzania analizy eksploracyjnej. Również z powodu braku zasadnej podstawy biznesowej podjęto decyzję o uwzględnieniu obserwacji odstających podczas przeprowadzania analizy właściwej[[19]](#footnote-19).

Końcowym krokiem poprzedzającym analizę eksploracyjną było wykonanie analizy współliniowości w celu odrzucenia ze zbioru atrybutów, które tłumaczą zmienną celu w taki sam sposób. Współliniowość opisuje sytuację, w której co najmniej dwie zmienne predykcyjne w modelu statystycznym są powiązane liniowo. Współliniowość może obejmować również liniową relację co najmniej jednej zmiennej predykcyjnej ze zmienną niezależną. Wiele metod statystycznych jest wrażliwych na występowanie współliniowości. Z tego powodu oszacowania parametrów mogą być niestabilne, odchylenia i błędy standardowe będą zawyżone, przez co wnioskowanie statystyczne może być nieprawidłowe. Dodatkowo, nawet w przypadku mniej czułych metod w przypadku współliniowości występować będą dwa pozostałe kluczowe problemy. Mianowicie, efekty zmiennych nie będą mogły zostać rozdzielone i zinterpretowane samodzielnie a ekstrapolacja oszacowań parametrów może być z dużym prawdopodobieństwem obarczona błędem[[20]](#footnote-20).

W celu określenia, czy istnieje zjawisko współliniowości pomiędzy zmiennymi wykorzystano miarę statystyczną wskaźnika inflacji wariancji (ang. Variance Inflation Factor). Spełnione zostało założenie odnośnie liczebności próby, która powinna być większa od liczby szacowanych parametrów. Według literatury przedmiotu wartością krytyczną, która wskazuje na oznaki współliniowości jest VIF powyżej 10, lub 5 według niektórych autorów. Na potrzeby analizy ze zbioru danych odrzucono zmienne predykcyjne o największym liniowym powiązaniu ze zmienną celu, tzn. atrybuty dla których wartość VIF przekroczyła wartość 10. Zastosowano bardziej łagodne kryterium, ponieważ na okrojonym zbiorze danych w kolejnej fazie budowy modelu przeprowadzona została analiza i wybór zmiennych, które najlepiej tłumaczą badaną zmienną. Do wyboru najbardziej wartościowych zmiennych wykorzystano metody, które biorą pod uwagę zjawisko współliniowości pomiędzy zmiennymi. Z tego powodu również na tym etapie świadomie nie wykonano badania zjawiska współliniowości pomiędzy zmiennymi predykcyjnymi.

Wstępne sprawdzenie jak kształtuje się zjawisko współliniowości pozwala jednak na nabycie intuicji co do wykorzystywanych danych oraz do redukcji zbioru danych. Zastosowana procedura pozwala na ograniczenie wymiarów tabeli użytej do analizy i zmniejszenia ilości czasu maszynowego potrzebnego do wykonania obliczeń[[21]](#footnote-21).

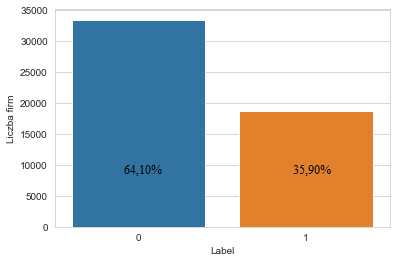
## II.3 Analiza eksploracyjna wykorzystanych danych.

Ostatecznie do analizy wykorzystano 52 190 obserwacji oraz 77 utworzonych zmiennych, z których zbudowano jedną tabelę. W badanej populacji znalazło się 33 455 (64,1%) podmiotów, które zakończyły działalność gospodarczą oraz 18 735 (35,9%) firm, które na dzień rozpoczęcia zbierania danych w dalszym ciągu ją prowadziły.

Analizę eksploracyjną rozpoczęto od sprawdzenia rozkładu zmiennej celu – nominalnej zmiennej „Label” o dwóch kategoriach:

1. 0 – dla firm, które nie zakończyły działalności gospodarczej w badanym okresie i na dzień pobrania danych prowadziły ją dalej (obserwacje pozytywne) – 64,1%,
2. 1 – dla firm, które zakończyły działalność gospodarczą w badanym okresie (35,9%).

Wykres 2. Liczba i udział firm aktywnych i zamkniętych w badanej populacji.



Źródło: Opracowanie własne na podstawie danych zebranych danych.

Z wykresu można odczytać, że udział obserwacji pozytywnych w populacji wykorzystanej do budowy modelu jest o niecałe 79% większy niż udział obserwacji negatywnych.

W dalszej kolejności przeprowadzono analizę eksploracyjną zmiennych objaśniających badane zjawisko bankructwa firm. Z początkowych 78 dostępnych zmiennych przeprowadzono proces wyboru najbardziej odpowiednich zmiennych do przeprowadzenia analizy. Do analizy wykorzystano 56 zmiennych numeryczne oraz 22 zmienne kategoryczne. Atrybutom przyporządkowano intuicyjne nazwy, w celu ułatwienia analizy eksploracyjnej oraz późniejszej interpretacji wyników. Wyjątkiem są niektóre atrybuty utworzone na podstawie danych finansowych, gdzie z powodu konieczności przyjęcia długich nazw zdecydowano się użyć krótkich łańcuchów znakowych. Po odrzuceniu ze zbioru danych obserwacji, gdzie wystąpiły braki danych dla zmiennych numerycznych oraz po przekształceniu zmiennych nominalnych i utworzeniu dla nich osobnych kategorii reprezentujących braki danych, w tabeli nie znalazły się żadne braki danych.

Wykorzystane zmienne zostały szerzej opisane w poniższym podsumowaniu. W podsumowaniu podczas opisywania kategorii atrybutów podano bardziej rozbudowany opis, do analizy właściwej zastosowano skrócone wartości w celu zmniejszenia wolumenu przetwarzanych informacji i przyspieszenia obliczeń. Z powodu odmiennych właściwości zmienne ciągłe i nominalne przedstawiono w oddzielnych tabelach. Poniższa tabela zawiera podsumowanie na temat zmiennych nominalnych.

Tabela 1. Zbiór dostępnych zmiennych nominalnych.

|  |  |  |
| --- | --- | --- |
| **Zmienne nominalne** | | |
| **Nazwa** | **Opis zmiennej** | **Kategorie zmiennej (udział procentowy)** |
| Voivodeship | Województwo, w którym znajduje się siedziba przedsiębiorstwa | Mazowieckie – 26,05%  Wielkopolskie – 10,20%  Śląskie – 10,19%  Inne – 9,79%  Dolnośląskie – 8,54%  Małopolskie – 8,29%  Pomorskie – 7,51%  Łódzkie – 4,90%  Zachodniopomorskie – 4,30%  Kujawsko–Pomorskie – 4,08%  Podkarpackie – 3,15%  Lubelskie – 3,00% |
| EMISLegalForm | Forma prawna zarejestrowania przedsiębiorstwa pobrana z EMIS według kodeksu klasyfikacji działalności gospodarczej | Spółka z ograniczoną odpowiedzialnością – 72,56%  Spółka komandytowa – 7,21%  Spółka jawna – 7,13%  Spółdzielnia – 5,16%  Spółka akcyjna – 4,61%  Inne – 3,32% |
| GUSLegalForm | Forma prawna zarejestrowania przedsiębiorstwa pobrana z GUS według kodeksu klasyfikacji działalności gospodarczej | Spółka z ograniczoną odpowiedzialnością –72,35%  Spółka komandytowa – 7,77%  Spółka jawna – 6,78%  Spółdzielnia – 5,14%  Spółka akcyjna – 4,64%  Inne – 3,32% |
| GUSSpecialLegalForm | Rodzaj formy prawnej, według której zostało zarejestrowane przedsiębiorstwo według klasyfikacji GUS | Osoba prawna – 83,90%  Jednostka organizacyjna niemająca osobowości prawnej – 16,10% |
| GUSOwnershipForm | Forma własności przedsiębiorstwa według klasyfikacji GUS | Własność krajowych osób fizycznych – 52,02%  Własność prywatna krajowa pozostała – 18,32%  Własność mieszana w sektorze prywatnym – 11,58%  Własność zagraniczna – 11,10%  Inne – 3,53%  Brak danych – 3,44% |
| MainPKD | Główna branża, w której przedsiębiorstwo prowadzi działalność według kodu PKD. Branża została zidentyfikowana na podstawie kodu a następnie przyporządkowana do konkretnego segmentu, na podstawie klasyfikacji działalności gospodarczej w Polsce | Handel hurtowy – 23,12%  Przetwórstwo przemysłowe – 17,58%  Działalność naukowa lub techniczna – 11,38%  Budownictwo – 10,62%  Nieruchomości – 7,65%  Administracja, łączność, media – 6,85%  Usługi pozostałe – 6,58%  Informacja i komunikacja – 5,42%  Transport, działalność magazynowa i logistyczna – 3,71%  Brak danych – 3,66%  Finanse i ubezpieczenia – 3,44% |
| MainNAICSCodes | Główna branża, w której przedsiębiorstwo prowadzi działalność według kodu NAICS (ang. North American Industry Classification System). Branża została zidentyfikowana na podstawie kodu NAICS a następnie przyporządkowana do konkretnego segmentu, na podstawie klasyfikacji kodów NAICS | Handel detaliczny – 28,65%  Produkcja przemysłowa – 28,38%  Usługi transportowe i logistyczne – 27,40%  Usługi – 5,96%  Handel hurtowy – 3,63%  Finanse, ubezpieczenia, nieruchomości – 3,52%  Inne – 2,46% |
| RiskyMainPKD | Zmienna binarna wskazująca, czy główna działalność przedsiębiorstwa definiowana kodem PKD jest ryzykowna w kontekście płatności podatku VAT i istnieje podwyższone ryzyko nadużycia | Nie – 85,61%  Tak – 10,82%  Brak danych – 3,57% |
| RiskySecondaryPKDs | Zmienna wskazująca, czy wśród dodatkowych działalności przedsiębiorstwa jest co najmniej 1 kod PKD ryzykowny w kontekście płatności podatku VAT i istnieje podwyższone ryzyko nadużycia | Nie – 68,59%  Tak – 27,85%  Brak danych – 3,57% |
| NoWebsite | Zmienna wskazująca, czy przedsiębiorstwo posiada stronę internetową (na podstawie danych GUS lub EMIS) | Nie (firma posiada stronę internetową) – 51,22%  Tak (brak strony internetowej) – 48,78% |
| PublicMail | Zmienna wskazująca, czy główny adres e-mail przedsiębiorstwa znajduje się w domenie publicznej (na podstawie danych GUS lub EMIS) | Nie – 87,88%  Tak – 12,12% |
| NoMail | Zmienna wskazująca, czy w bazach danych GUS lub EMIS jest dostępny adres e-mail przedsiębiorstwa | Nie (adres dostępny) – 61,48%  Tak (brak adresu) – 38,52% |
| NoFax | Zmienna wskazująca, czy w bazach danych GUS lub EMIS jest dostępny numer faksu przedsiębiorstwa | Nie (adres dostępny) – 58,12%  Tak (brak adresu) – 41,88% |
| NoPhone | Zmienna wskazująca, czy w bazach danych GUS lub EMIS jest dostępny numer telefonu przedsiębiorstwa | Nie (numer dostępny) – 67,63%  Tak (brak numeru) – 32,37% |
| NoDescription | Zmienna wskazująca, czy w bazie danych EMIS jest dostępny szczegółowy opis przedsiębiorstwa i jego działalności | Tak (opis niedostępny) – 84,91%  Nie (opis dostępny) – 15,09% |
| AdresVirtualOffice | Zmienna wskazująca, czy przedsiębiorstwo jest zarejestrowane pod adresem biura wirtualnego | Nie – 91,36%  Tak – 5,08%  Brak danych – 3,57% |
| AdresFlat | Zmienna wskazująca, czy przedsiębiorstwo jest zarejestrowane pod adresem lokalu mieszkalnego | Nie – 75,89%  Tak – 20,54%  Brak danych – 3,57% |
| CAACImport | Zmienna wskazująca, czy w bazie danych CAAC znajdują się dane na temat importu towarów przez przedsiębiorstwo | Nie – 87,55%  Tak – 8,88%  Brak danych – 3,57% |
| CAACExport | Zmienna wskazująca, czy w bazie danych CAAC znajdują się dane na temat eksportu towarów przez przedsiębiorstwo | Nie – 89,25%  Tak – 7,19%  Brak danych – 3,57% |
| EntityListedInVATRegistry | Zmienna wskazująca, czy przedsiębiorstwo jest zarejestrowane w rejestrze podatników VAT prowadzonym przez Ministerstwo Finansów | Tak – 91,97%  Nie – 8,03% |
| VirtualAccountsPresence | Zmienna wskazująca, czy przedsiębiorstwo posiada rachunki wirtualne | Nie – 89,18%  Brak danych – 8,03%  Tak – 2,79% |
| RiskyRemovalBasis | Zmienna wskazująca, czy przedsiębiorstwo zostało kiedykolwiek wykreślone z rejestru podatników VAT, a jeśli tak czy powód wykreślenia był spowodowany wystąpieniem nadużycia lub przestępstwa określone na podstawie artykułu ustawy podanego jako przyczynę wykreślenia | Nie wykreślone – 83,05%  Brak danych – 8,03%  Nadużycie lub przestępstwo – 4,66%  Przyczyna naturalna – 4,25% |

Poniższa tabela zawiera podsumowanie na temat zmiennych ciągłych. Statystyki opisowe zostały wygenerowane na zbiorze danych, z którego usunięto obserwacje odstające przy użyciu opisanej wcześniej metody Z-score.

Tabela 2. Zbiór dostępnych zmiennych ciągłych.

|  |  |  |
| --- | --- | --- |
| **Zmienne ciągłe** | | |
| **Nazwa** | **Opis zmiennej** | **Statystyki opisowe** |
| AgeDays | Liczba dni od daty rozpoczęcia działalności gospodarczej do daty zakończenia działalności lub daty pobrania danych dla przedsiębiorstwa | Średnia: 5681,99; mediana: 5007,0; odchylenie standardowe: 3748,69 |
| LocalBranches | Liczba oddziałów i lokalnych jednostek przedsiębiorstwa | Średnia: 0,13; mediana: 0,0; odchylenie standardowe: 0,84 |
| SecondaryPKDCount | Liczba kodów PKD wskazujących na dodatkowe działalności gospodarcze prowadzone przez przedsiębiorstwo | Średnia: 9,08; mediana: 6,0; odchylenie standardowe: 15,08 |
| ActiveLicenses | Liczba aktywnych licencji, koncesji lub zezwoleń posiadanych przez przedsiębiorstwo | Średnia: 0,0; mediana: 0,0; odchylenie standardowe: 0,0 |
| RevertedLicenses | Liczba wycofanych licencji, koncesji lub zezwoleń przedsiębiorstwa | Średnia: 0,0; mediana: 0,0; odchylenie standardowe: 0,0 |
| ExpiredLicenses | Liczba wygasłych licencji, koncesji lub zezwoleń przedsiębiorstwa | Średnia: 0,0; mediana: 0,0; odchylenie standardowe: 0,0 |
| DeclaredAccountsCount | Liczba rachunków bankowych zadeklarowanych przez przedsiębiorstwo w rejestrze podatników VAT | Średnia: 2,19; mediana: 1,0; odchylenie standardowe: 3,38 |
| RepresentationCount | Liczba osób uprawnionych do reprezentacji przedsiębiorstwa | Średnia: 0,03; mediana: 0,0; odchylenie standardowe: 0,16 |
| NumberOfEmployees | Liczba pracowników przedsiębiorstwa | Średnia: 74,32; mediana: 25,0; odchylenie standardowe: 117,32 |
| ExecutivesCount | Liczba pracowników osób na stanowiskach kierowniczych w przedsiębiorstwie | Średnia: 1,75; mediana: 1,0; odchylenie standardowe: 1,04 |
| OwnersCount | Liczba osób fizycznych lub osób prawnych będących właścicielami przedsiębiorstwa | Średnia: 1,42; mediana: 1,0; odchylenie standardowe: 1,17 |
| AffiliatesCount | Liczba podmiotów partnerskich, filii lub oddziałów przedsiębiorstwa | Średnia: 0,16; mediana: 0,0; odchylenie standardowe: 0,51 |
| ExternalIdsOthers | Liczba dodatkowych identyfikatorów przedsiębiorstwa w bazie danych EMIS poza numerami NIP, REGON i KRS | Średnia: 1,07; mediana: 1,0; odchylenie standardowe: 0,46 |
| RegisteredCapitalValue | Zarejestrowany kapitał zakładowy przedsiębiorstwa | Średnia: 3770845,0; mediana: 70000,0; odchylenie standardowe: 17656052,0 |
| AuditDaysAgo | Liczba dni od ostatniego zarejestrowanego audytu w bazie danych EMIS | Średnia: 245,51; mediana: 0,0; odchylenie standardowe: 816,91 |
| PreviousNamesCount | Liczba zmian nazwy przedsiębiorstwa w przeszłości | Średnia: 0,43; mediana: 0,0; odchylenie standardowe: 0,63 |
| PreviousNameChangeYearsAgo | Liczba lat od poprzedniej zmiany nazwy przez przedsiębiorstwo | Średnia: 2,15; mediana: 0,0; odchylenie standardowe: 3,88 |
| LatestMarketCapitalization | Kapitalizacja rynkowa przedsiębiorstwa w przypadku gdy prowadzi działalność w formie prawnej spółki akcyjnej | Średnia: 432902,66; mediana: 0,0; odchylenie standardowe: 11646691,0 |
| DividendSum | Suma wypłaconych dywidend przypadających na jedną akcję, które zostały wypłacone przez przedsiębiorstwo (jeśli takowe wystąpiły) | Średnia: 0,0; mediana: 0,0; odchylenie standardowe: 0,02 |
| NetSalesRevenue | Przychody netto ze sprzedaży przedsiębiorstwa według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 18572,75; mediana: 718,41; odchylenie standardowe: 74083,33 |
| OperatingProfitEBIT | Przychody z działalności operacyjnej przedsiębiorstwa przed odliczeniem podatków i odsetek według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 871,41; mediana: 0,0; odchylenie standardowe: 6026,59 |
| EmployeeBenefitExpense | Wynagrodzenia i świadczenia dla pracowników według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 1992,0; mediana: 92,0; odchylenie standardowe: 6205,54 |
| TotalAssets | Aktywa razem według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 23225,88; mediana: 1091,58; odchylenie standardowe: 132002,39 |
| NetProfitLossForThePeriod | Zysk netto przedsiębiorstwa według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 706,76; mediana: 3,79; odchylenie standardowe: 5416,29 |
| PropertyPlantAndEquipment | Środki trwałe przedsiębiorstwa w roku finansowym poprzedzającym likwidację działalności lub pobranie danych | Średnia: 5397,97; mediana: 9,0; odchylenie standardowe: 29804,2 |
| CashandCashEquivalents | Środki pieniężne i inne aktywa pieniężne według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 1565,08; mediana: 32,65; odchylenie standardowe: 8288,68 |
| TotalEquity | Całkowity kapitał własny według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 8358,12; mediana: 219,81; odchylenie standardowe: 44705,75 |
| IssuedCapital | Kapitał podstawowy według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 3518,62; mediana: 51,0; odchylenie standardowe: 21456,44 |
| WorkingCapital | Kapitał obrotowy według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy sprawozdania finansowego | Średnia: 1762,87; mediana: 89,21; odchylenie standardowe: 16911,71 |
| RetainedEarnings | Zatrzymane dochody według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -1132,75; mediana: 0,0; odchylenie standardowe: 16278,61 |
| TotalLiabilities | Zobowiązania razem według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 13608,12; mediana: 575,98; odchylenie standardowe: 91192,66 |
| CurrentLiabilities | Zobowiązania krótkoterminowe według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 6102,26; mediana: 335,25; odchylenie standardowe: 24324,95 |
| NonCurrentLiabilities | Zobowiązania długoterminowe według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 3797,89; mediana: 0,0; odchylenie standardowe: 26045,46 |
| ProfitBeforeIncomeTax | Przychody razem według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 876,36; mediana: 5,0; odchylenie standardowe: 6330,62 |
| IncomeTax | Podatek dochodowy według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -124,13; mediana: 0,0; odchylenie standardowe: 1015,79 |
| DepreciationImpairment | Utrata wartości amortyzacji według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 358,75; mediana: 0,0; odchylenie standardowe: 2155,69 |
| DepreciationAmortization | Deprecjacja wartości aktywów i amortyzacja według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -471,98; mediana: -5,0; odchylenie standardowe: 2179,43 |
| CurrentAssets | Aktywa obrotowe według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 7886,09; mediana: 602,91; odchylenie standardowe: 30414,29 |
| ROE | Wskaźnik rentowności kapitału własnego według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 0,02; mediana: 0,01; odchylenie standardowe: 1,44 |
| ROA | Wskaźnik rentowności aktywów według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -0,53; mediana: 0,0; odchylenie standardowe: 38,24 |
| ROS | Wskaźnik rentowności sprzedaży według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -3,91; mediana: 0,0; odchylenie standardowe: 66,6 |
| BruttoMargin | Marża brutto ze sprzedaży według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy (OperatingProfitEBIT / TotalAssets) | Średnia: -29,81; mediana: 0,0; odchylenie standardowe: 427,1 |
| RevenueToCash | Wskaźnik obrotów przedsiębiorstwa do środków pieniężnych i innych aktywów pieniężnych w roku finansowym poprzedzającym likwidację działalności lub pobranie danych | Średnia: 5,94; mediana: 0,0; odchylenie standardowe: 3,82 |
| RevenueToWages | Wskaźnik obrotów przedsiębiorstwa do wydatków na wynagrodzenia i świadczeń dla pracowników w roku finansowym poprzedzającym likwidację działalności lub pobranie danych | Średnia: 0,81; mediana: 0,0; odchylenie standardowe: 0,27 |
| A1 | WorkingCapital / TotalAssets – wskaźnik wartości aktywów bieżących w stosunku do wielkości przedsiębiorstwa według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -15,18; mediana: 0,28; odchylenie standardowe: 459,03 |
| A2 | RetainedEarnings / TotalAssets – wskaźnik skumulowanej rentowności, która odzwierciedla siłę sprzedażową przedsiębiorstwa i możliwości generowania przychodów, według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -15,93; mediana: 0,0; odchylenie standardowe: 309,3 |
| A3 | OperatingProfitEBIT / TotalAssets – wskaźnik efektywności operacyjnej, który mierzy wpływ zysków operacyjnych na rentowność długoterminową według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -0,44; mediana: 0,0; odchylenie standardowe: 30,12 |
| A4 | TotalEquity / TotalLiabilities – wskaźnik dodający wymiar rynkowy, który uwzględnia wahania ceny papierów wartościowych przedsiębiorstwa według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 95,71; mediana: 0,89; odchylenie standardowe: 1157,29 |
| A5 | NetSalesRevenue / TotalAssets – wskaźnik obrotu do wartości aktywów przedsiębiorstwa, który zazwyczaj różni się znacznie w zależności od branży, według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 2,65; mediana: 0,83; odchylenie standardowe: 14,66 |
| P3 | (NetSalesRevenue - DepreciationImpairment) / TotalLiabilities – wskaźnik obrotu do wartości zobowiązań przedsiębiorstwa według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 13,37; mediana: 1,82; odchylenie standardowe: 124,83 |
| X8 | CurrentAssets / CurrentLiabilities – wskaźnik wartości aktywów krótkoterminowych do zobowiązań krótkoterminowych według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 57,94; mediana: 1,86; odchylenie standardowe: 474,59 |
| X9 | CurrentLiabilities / TotalAssets – wskaźnik wartości zobowiązań krótkoterminowych do aktywów razem według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 16,19; mediana: 0,29; odchylenie standardowe: 459,14 |
| X10 | ProfitBeforeIncomeTax / CurrentLiabilities – wskaźnik wartości przychodów przed opodatkowaniem do zobowiązań krótkoterminowych według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -0,69; mediana: 0,03; odchylenie standardowe: 89,7 |
| X11 | TotalAssets / TotalLiabilities – wskaźnik wartości aktywów razem do wartości zobowiązań według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: 96,26; mediana: 1,89; odchylenie standardowe: 1154,24 |
| X13 | ProfitBeforeIncomeTax / NetSalesRevenue – wskaźnik wartości przychodów przed opodatkowaniem do przychodów ze sprzedaży netto według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -38,6; mediana: 0,01; odchylenie standardowe: 664,52 |
| X14 | ProfitBeforeIncomeTax / TotalAssets – wskaźnik wartości przychodów przed opodatkowaniem do aktywów razem według danych z roku finansowego poprzedzającego pobranie danych lub likwidację firmy | Średnia: -0,52; mediana: 0,01; odchylenie standardowe: 34,42 |

Wykorzystane zmienne A1, A2, A3, A4 oraz A5 pochodzą modelu Z-score do predykcji upadłości przedsiębiorstw zbudowanego przez E. Altmana w 1968 roku[[22]](#footnote-22).

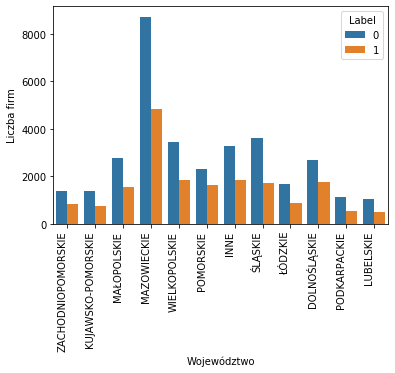
Wykorzystana zmienna P3 pochodzi z modelu zbudowanego przez T. Prusaka w 2005 roku[[23]](#footnote-23).

Wykorzystane zmienne X8, X9, X10, X11, X12, X13 oraz X14 pochodzą z modeli zbudowanych przez T. Korola[[24]](#footnote-24).

Atrybuty zostały utworzone w możliwie najdokładniejszy sposób, biorąc pod uwagę różnice pomiędzy standardem składania sprawozdań finansowych w poszczególnych państwach, w których budowano poszczególne modele.

W dalszej kolejności przeprowadzono analizę wizualną relacji pomiędzy zmiennymi predykcyjnymi a zmienną celu. W tym celu w automatyczny sposób wygenerowano odpowiednie wykresy. W przypadku zmiennych nominalnych wykorzystano podzielone, oddzielne histogramy dla obu klas zmiennej objaśnianej. Dla zmiennych numerycznych wykorzystano wykresy pudełkowe i rozrzutu dla obu klas zmiennej objaśnianej. Przykładowe wykresy dla pojedynczych zmiennych o obu typach znajdują się poniżej. Proces ten powtórzono dla wszystkich badanych zmiennych.

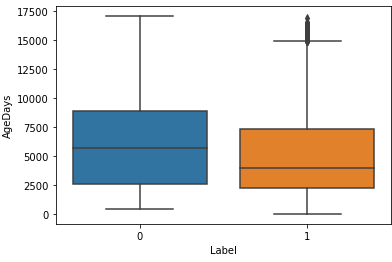
Wykres 3. Liczba firm według województw zgrupowane według klas zmiennej celu – przykładowy histogram dla zmiennych nominalnych.



Źródło: Opracowanie własne na podstawie danych zebranych danych.

Z wykresu można odczytać liczbę podmiotów zamkniętych i aktywnych, które zostały zarejestrowane w poszczególnych województwach. Dzięki zastosowaniu wykresu o takim typie możliwe jest nabycie intuicji co do danych, udziału poszczególnych kategorii oraz tego, jak poszczególne wartości są powiązane ze zmienną celu. Z wykresu można odczytać, że udział firm, które zakończyły działalność nie jest stały w poszczególnych województwach. Dla przykładu, udział bankrutów w województwie pomorskim jest znacznie większy niż w województwie mazowieckim. Zatem, na podstawie takiej analizy wizualnej można podejrzewać, że ten czynnik będzie miał znaczenie podczas budowy modelu i zmienna, która zawiera te informacje będzie uwzględniona jako atrybut istotny statystycznie. Analogiczną analizę przeprowadzono dla pozostałych zmiennych nominalnych.

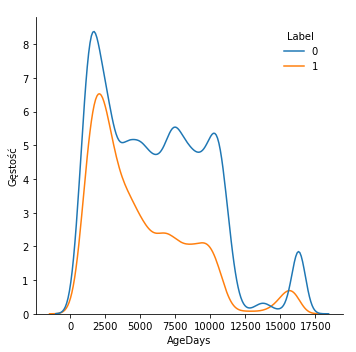
Wykres 4. Rozkłady wieku firm zgrupowane według klas zmiennej celu - przykładowy wykres pudełkowy dla zmiennych numerycznych.



Źródło: Opracowanie własne na podstawie danych zebranych danych.

Z wykresu można odczytać rozkłady wieku (liczonego w dniach) podmiotów zamkniętych i aktywnych dla poszczególnych klas zmiennej celu. Poprzez zastosowanie wykresu pudełkowego wykreślanego dla obu kategorii przyjmowanych przez zmienną objaśnianą możliwe jest porównanie rozkładów w obu grupach. Dzięki temu można odczytać, że rozkłady różnią się, co oznacza, że na prawdopodobieństwo upadłości podmiotu wpływa to, jak długo firma prowadzi działalność gospodarczą. Relacja ze zmienną celu wskazuje na tendencję, że im dłużej przedsiębiorstwo prowadzi działalność tym większa jest szansa na jej kontynuację. Zatem, porównanie rozkładów można podejrzewać, że czynnik wieku firmy będzie miał znaczenie z perspektywy budowy modelu a badana zmienna okaże się istotna statystycznie. Analogiczną analizę przeprowadzono dla pozostałych zmiennych numerycznych.

Wykres 5. Rozkłady wieku firm zgrupowane według klas zmiennej celu - przykładowy wykres gęstości dla zmiennych numerycznych.



Źródło: Opracowanie własne na podstawie danych zebranych danych.

Analiza wykresu 5 potwierdza wnioski odczytane z poprzedniego wykresu. Zależność pomiędzy wiekiem podmiotu a szansą na zakończenie przez niego działalności gospodarczej jest widoczna. Wskazuje to, że wiek jest istotnym czynnikiem, który powinien być brany pod uwagę podczas predykcji bankructwa przedsiębiorstwa. Dla pozostałych zmiennych numerycznych przeprowadzono analogiczną ocenę ich wpływu na zmienną celu.

## II.4 Wybór zmiennych najlepiej wyjaśniających zjawisko zakończenia działalności gospodarczej.

Ostatnim krokiem przed rozpoczęciem wyboru zmiennych o najlepszych wartościach predykcyjnych jest odpowiednie przygotowanie danych. Większość metod uczenia maszynowego wymaga lub jest bardziej efektywna, jeśli atrybuty danych są podane w tej samej skali. W celu transformacji zmiennych numerycznych skorzystano z metody standaryzacji. Standaryzacja odnosi się do przesunięcia rozkładu każdej zmiennej numerycznej tak, aby jej średnia wynosiła 0 a odchylenie standardowe 1. Metoda ta przynosi dobre efekty w przypadku korzystania z modeli, które wykorzystują rozkłady zmiennych objaśniających do wykonania predykcji[[25]](#footnote-25).

Transformacji poddano również zmienne nominalne, które zostały odpowiednio zakodowane według przyjmowanych wartości. Większość algorytmów uczenia maszynowego nie może zostać zaimplementowana na danych, które występują w tradycyjnej formie jako wartości nominalne i wymaga konwersji typów do postaci numerycznej. Jest to spowodowane zazwyczaj ograniczeniem w efektywności implementacji a nie naturą algorytmów. W celu konwersji zmiennej nominalnej na typ numeryczny w pierwszej kolejności następuje jej zakodowanie. Do każdej z unikalnych kategorii zostaje przypisana inna liczba całkowita. Następnie dla każdej z unikalnych liczb całkowitych dodawana jest nowa zmienna, która przyjmuje wartość 1 dla obserwacji, które należą do danej klasy lub 0, gdy należą do dowolnej z pozostałych klas. Pierwotna zmienna z zakodowanymi wszystkimi możliwymi wartościami zostaje usunięta. Proces ten prowadzi do zwiększenia rozmiarów wykorzystanych danych i spowalnia obliczenia, skutkuje jednak poprawą wyników modelu. Ostatecznie, bo zakodowaniu zmiennych nominalnych zbiór danych liczył 150 unikalnych zmiennych. Zbiór danych powiększył zatem swoją objętość o prawie połowę[[26]](#footnote-26).

Po przeprowadzeniu analizy eksploracyjnej i transformacji zebranych danych kolejnym wykonanym krokiem było zweryfikowanie istotności i jakości zmiennych z perspektywy budowy modelu. Czynność ta była niezbędna z powodu dużej liczby zmiennych w zbiorze danych. Konieczność pracy na dużych zbiorach danych jest coraz bardziej powszechnym zjawiskiem podczas rozwiązywania problemów biznesowych przy wykorzystaniu metod uczenia maszynowego. Aby wydobyć informacje z dużych ilości danych (zwłaszcza w kontekście dużej liczby zmiennych) i ograniczyć ich rozmiar przydatne okazują się procedury statystyczne. Wynika to z faktu, ze zazwyczaj nie wszystkie dostępne atrybuty są potrzebne z punktu widzenia budowy efektywnego modelu. Często zbyt duża liczba zmiennych wpływa wręcz negatywnie na wyniki, na przykład w sytuacji gdy w modelu uwzględnione są atrybuty nieskorelowane ze zmienną celu bądź powiązane liniowo. Redukcja liczby wykorzystanych zmiennych nie tylko pomaga w szybszym trenowaniu modelu ale też zmniejsza jego złożoność, poprawia interpretowalność i poprawia jakość predykcji.

W celu wyboru istotnych zmiennych i redukcji zbioru danych wykorzystano algorytm Boruta. Metoda ta bazuje na algorytmie klasyfikacyjnych lasów losowych. Jej celem jest znalezienie wszystkich ważnych cech, które występują w zbiorze danych w odniesieniu do zmiennej celu. Technika ta dobrze sprawdza się podczas wyboru istotnych atrybutów do rozwiazywania problemów klasyfikacyjnych. Dodatkowym atutem jest czytelność zwracanych wyników. Zaletą jest długi czas wykonywania obliczeń w przypadku wielowymiarowego zbioru danych.

Sposób działania algorytmu jest następujący. Działa on w sposób iteracyjny. W pierwszej kolejności powiela on zbiór danych i sortuje w losowy sposób wartości w każdej kolumnie reprezentującej zmienną objaśnianą. Następnie trenuje klasyfikator, taki jak np. klasyfikacyjny las losowy na otrzymanym zbiorze danych. Robiąc to, dla każdej ze zmiennych wyliczana jest jej średnia istotność z perspektywy powiązania ze zmienną objaśnianą - im jest ona wyższa, tym dana zmienna jest bardziej istotna.

Następnie algorytm weryfikuje każdą zmienną w rzeczywistej postaci (bez losowego sortowania), czy jest bardziej istotna w porównaniu do wartości posortowanych w losowy sposób. Jeśli tak jest, informacja jest zapisywana jako wektor i rejestrowana jako pojedynczy zapis, nazywany trafieniem. Następnie algorytm przechodzi do kolejnej iteracji. Po wykonaniu zdefiniowanej wcześniej liczby iteracji metoda zwraca tabelę trafień.

W każdej iteracji algorytm porównuje istotność zmiennej na posortowanych losowo danych z ich oryginalną kolejnością, żeby sprawdzić, czy te drugie działały lepiej niż poprzednie. Jeśli tak, algorytm oznaczy tę zmienną jako ważną. Odbywa się to poprzez porównanie, ile razy lepiej dana zmienna wypadła w porównaniu do odpowiednika losowej kolejności. Jeśli dla danej zmiennej nie zostało spełnione kryterium akceptacji jest ona odrzucana i usuwana ze zbioru danych. Za kryterium akceptacji zazwyczaj uznaje się zarejestrowanie określonej liczby trafień lub przekroczenie progu istotności. Po wykonaniu zadanej liczby iteracji - albo po odrzuceniu lub stwierdzeniu istotności wszystkich zmiennych – zwracane są wyniki. Algorytm jako wynik zwraca tabelę, w której zawarte są informacje na temat średniego, minimalnego i maksymalnego wyniku oraz liczby trafień dla każdej ze zmiennych[[27]](#footnote-27).

W ten właśnie sposób przetestowano istotność wszystkich początkowych zmiennych. Za kryterium akceptacji uznano przekroczenie progu wartości standardowo akceptowanej wartości 3,00 co oznaczałoby, że badana zmienna jest 3 raz bardziej istotna niż jej odpowiednik w losowej postaci. Wynik ten został osiągnięty dla 94 z nich, co oznacza, że ze startowego zbioru odrzucono 56 atrybuty. Opis oraz poszczególne kategorie dla każdej z nich przedstawiono w poprzednim podrozdziale pracy. Istotność zaakceptowanych zmiennych przedstawia poniższy wykres.

Wykres 6. Wpływ zmiennych zaakceptowanych przez algorytm Boruta na zmienną celu.

Obraz zawierający stół

Opis wygenerowany automatycznie

Z powodu dużej liczby zatwierdzonych zmiennych i poprawy czytelności poprzez dodanie wykresu w możliwie najwyższej jakości i rozdzielczości jego drugą część przy zachowaniu tej samej skali poziomej umieszczono na następnej stronie.

Obraz zawierający stół

Opis wygenerowany automatycznie

Źródło: Opracowanie własne na podstawie wyników istotności zmiennych na podstawie zastosowanego algorytmu.

Z wykresu 6 można wyciągnąć wniosek, że w kontekście przewidywania upadłości przedsiębiorstw najbardziej istotne okazały się dwa rodzaje zmiennych.

W pierwszej kolejności badane zjawisko tłumaczą najlepiej atrybuty opisujące w pośredni sposób wielkość przedsiębiorstwa i skalę prowadzonej przez nie działalności, czyli na przykład czynniki takie jak liczba rachunków bankowych, liczba pracowników na stanowiskach kierowniczych i właścicieli, wysokość kapitału własnego, wydatki na wynagrodzenia, czy to w jak wielu branżach aktywna jest firma. Nie bez znaczenia jest też długość prowadzonej działalności oraz osobowość prawna podmiotu, czas od ostatniego audytu, czy to z jakiej formy prowadzenia działalności korzysta dane przedsiębiorstwo. Można przypuszczać, że podmioty bardziej dojrzałe, zatrudniające wiele osób, aktywne w wielu branżach, które są w mniejszym stopniu zależne od panującej koniunktury w konkretnych segmentach rynku będą mniej narażone na bankructwo w krótkim horyzoncie czasowym. Zależności te wydają się intuicyjne i są zgodne z wnioskami wyciągniętymi na etapie przeprowadzania analizy eksploracyjnej, która wykazała, że rozkłady tych atrybutów różnią się pomiędzy podmiotami, które zbankrutowały i tymi, które kontynuowały działalność. Występowanie korelacji pomiędzy tymi czynnikami potwierdziły wyniki zwrócone przez algorytm.

Jako drugą grupę istotnych zmiennych można zidentyfikować wskaźniki finansowe utworzone na podstawie analizy literatury i podobnych projektów oraz zmienne wskazujące na wartości poszczególne wartości ze sprawozdania finansowego za okres przypadający na ostatni rok finansowy przed upadłością lub pobraniem danych. Warto zwrócić uwagę, że wszystkie wskaźniki finansowe utworzone na podstawie modeli Altmana, Prusaka i Korola zostały zaakceptowane przez algorytm. Duża liczba pozostałych czynników przedstawiających kondycję finansową firm również została zatwierdzona. Wydaje się to intuicyjne i również zgodne z wnioskami wyciągniętymi na etapie analizy eksploracyjnej. Zakończenie działalności przez przedsiębiorstwo zazwyczaj poprzedzają niesatysfakcjonujące wyniki finansowe. Podczas przeprowadzenia wielu projektów o podobnej charakterystyce wykorzystano tylko dane ze sprawozdań finansowych i modele te okazywały się być bardzo skuteczne, a przy tym nieskomplikowane i łatwe w interpretacji. Nie inaczej jest w tym wypadku, istotność tych czynników nie jest zaskakująca. Choć część z nich, jak na przykład wielkość przychodów ze sprzedaży, kapitału zakładowego są powiązane z pierwszą zidentyfikowaną grupą zmiennych, ponieważ wskazują na wielkość przedsiębiorstwa, to wykorzystane wskaźniki finansowe są wartościami relatywnymi i dobrze będą przedstawiać sytuację badanego podmiotu bez względu na skalę prowadzonej działalności.

Do budowy modelu wykorzystano zatem w pierwszej kolejności wszystkie zmienne zatwierdzone przez model. W celu porównania wyników i próby zbudowania możliwie najlepszego klasyfikatora przetestowano również wybrane algorytmy na całym startowym zbiorze liczącym 150 zmiennych. Uzyskane wyniki i ich interpretację przedstawiono w następnej części pracy.

# III. Analiza właściwa. Budowa i ocena jakości modeli klasyfikacyjnych.

## III.1 Model klasyfikacyjny zbudowany na podstawie algorytmu Gradient Boosting Machine.

Do zbudowania modelu klasyfikacyjnego wykorzystano 3 algorytmy uczenia maszynowego. Zastosowane algorytmy według literatury przedmiotu oraz doświadczenia biznesowego autora bardzo dobrze sprawdzają się do rozwiązywania problemów klasyfikacyjnych w przypadku, gdy w zbiorze danych znajduje się duża liczba zmiennych kategorycznych i numerycznych oraz gdy liczba obserwacji wielokrotnie przekracza liczbę atrybutów.

W pierwszej kolejności zbiór danych podzielono na część treningową oraz testową według proporcji 80% (41 752 obserwacji) do 20% (10 438 obserwacji). Podzbiór treningowy wykorzystywany jest do wytrenowania modelu, który na podstawie dostarczonych danych ma nauczyć się wykrywać powiązania między danymi i wpływ zmiennych objaśniających na zmienną celu. Podzbiór testowy służy do zapewnienia obiektywnej oceny dopasowania modelu do danych, poprzez wykorzystanie próbki danych, która nie została włączona do części treningowej i model nie mógł z niej skorzystać w celu uczenia się.

W celu zapewnienia odpowiednich warunków do uczenia i walidacji modelu w obu zbiorach zapewniono zbliżony rozkład zmiennej celu oraz rozkładu czasu, który upłynął od bankructwa firm. Rozkład czasu został wyliczony z maksymalną możliwą dokładnością liczoną w dniach. Zabieg ten zastosowano w celu uniknięcia różnic pomiędzy wynikami osiągniętymi na zbiorze treningowym i testowym, które mogą być wynikiem nierównomiernego rozkładu zmiennej celu w wyniku np. losowego podziału zbiorów.

Dodatkowo, zdefiniowana została początkowa wartość ziarna (ang. seed) w celu zapewnienia, że za każdym razem generowany jest ten sam zestaw liczb losowych i do zbiorów treningowego i testowego za każdym razem włączono te same obserwacje. Takie postępowanie usuwa niepewność dotyczącą wyników modelowania i pomaga dokładnie ocenić jakość modelu oraz porównać go z pozostałymi klasyfikatorami.

W pierwszej kolejności zastosowano algorytm wzmocnionego drzewa decyzyjnego (ang. Gradient Boosting Machine). Jest to algorytm, który w ostatnich latach w środowisku uczenia maszynowego cieszy się dużą popularnością z powodu efektywności, stosunkowo szybkiego tempa wykonywania kalkulacji, wysoką interpretowalnością w porównaniu z algorytmami o podobnej skuteczności oraz braku konieczności odpowiedniego przeprocesowania danych dostarczanych do modelowania. Do wad tej metody zaliczają się niska skalowalność (wysokie zużycie zasobów) oraz podatność na przeuczenie. Algorytm jest zbudowany w oparciu o strukturę o drzewiastym kształcie.

Proces wykonania klasyfikacji wykonywany jest w sposób iteracyjny. Rozpoczyna się on od utworzenia korzenia drzewa aż do osiągnięcia poziomu liści, czyli tzw. klas terminalnych. W poszczególnych iteracjach tworzone są węzły, które powstają dzięki odpowiednio dobranym zmiennym tłumaczącym możliwie najlepiej badane zjawisko. Wykorzystane atrybuty są dodawane w kolejnych iteracjach w taki sposób, żeby maksymalizować zysk informacji dla każdego węzła. Proces jest powtarzany aż do momentu osiągnięcia żądanego stopnia rozbudowania drzewa, który może być definiowany np. maksymalną dopuszczalną głębokością lub minimalną liczbą obserwacji w każdym liściu. Dodatkowym elementem dodanym do algorytmu, który wpływa na znaczną poprawę osiąganych wyników jest tzw. wzmocnienie (ang. boosting). Zastosowana metoda polega na losowaniu nowych ciągów uczących, które służą do trenowania kolejnych wersji klasyfikatorów zbudowanych przy użyciu wylosowanych parametrów. Oznacza to, że algorytm tworzy sekwencję klasyfikatorów. Każde nowe utworzone drzewo bierze pod uwagę i uczy się kompensować błąd pozostawiony przez poprzedni model. Zbudowana w ten sposób duża liczba modeli o stosunkowo niskiej jakości skorygowana o wzajemny wpływ (wzmocniona) składa się na bardzo dokładny klasyfikator[[28]](#footnote-28).

Z powodu stosunkowo krótkiego czasu oczekiwania na wstępne wyniki zwrócone przez model algorytm ten zastosowano w pierwszej kolejności, żeby ocenić potencjał zgromadzonych danych i zdecydować, czy możliwe jest zbudowanie na nich klasyfikatora o odpowiedniej jakości.

Za podstawowy wskaźnik oceny jakości modelu zastosowano miarę wartości AUC (ang. area under the curve) – pola pod krzywą ROC, która jest podstawowym narzędziem wykorzystywanym do oceny jakości modeli klasyfikacyjnych. Jako drugą miarę do oceny modelu zastosowano zdolność klasyfikatora do rozpoznawania obserwacji pozytywnych wyrażoną odsetkiem wykrytych upadłych firm, które zostały zaklasyfikowane jako 75% próby o prawdopodobieństwie przynależności do klasy reprezentującej podmioty, które zakończyły działalność gospodarczą. W kontekście biznesowym podstawowym wymaganiem do możliwości zastosowania budowanego narzędzia i wniesienia wartości dodanej jest zapewnienie wysokiej zdolności do wykrywania obserwacji pozytywnych, czyli firm zagrożonych upadłością. Jest to podstawowe kryterium, które będzie dostarczało wartości osobom korzystającym z budowanego narzędzia. Skupienie możliwie wysokiego odsetka bankrutów w jak najmniejszej części populacji pozwala na przykład na bardziej dokładną weryfikację podmiotów wskazanych jako zagrożone lub oszczędzenie czasu poświęconego na ocenę wiarygodności badanej populacji poprzez zawężenie jej do podmiotów o wysokiej ocenie ryzyka. Obydwa wskaźniki zostały wyliczone na zbiorach treningowym i testowym w celu weryfikacji, czy model nie jest przetrenowany lub niedotrenowany. Analogiczny schemat zastosowano później do oceny modeli wytrenowanych przy użyciu pozostałych algorytmów.

Po zbudowaniu początkowego modelu o standardowych wartościach parametrów bez próby ich optymalizacji okazało się, że model cechuje się zadowalającą jakością predykcji i zgromadzone dane i zmienne posiadają wystarczający potencjał do przeprowadzenia analizy.

W związku z tym rozpoczęto tworzenie modeli w sposób iteracyjny w celu znalezienia optymalnych wartości parametrów takich jak:

* maksymalna głębokość drzew decyzyjnych (zakres wartości od 1 do 3 z interwałem 1),
* zwiększenie szybkości uczenia się – boostingu (zakres wartości od 0,01 do 0,25 z interwałem 0,01),
* liczba drzew decyzyjnych zbudowanych podczas trenowania algorytmu (zakres wartości od 50 do 130 z interwałem 10)[[29]](#footnote-29).

W tym celu zastosowano metodę przeszukiwania za pomocą siatki (ang. grid search), czyli przeszukiwania możliwej przestrzeni wartości parametrów w pętli od zdefiniowanych minimalnej do maksymalnej wartości każdego parametru, tak żeby przetestowana została każda możliwa kombinacja wartości parametrów. Modele tworzono zarówno przy użyciu wyłącznie zmiennych zatwierdzonych przez algorytm Boruta oraz na pełnym zbiorze. Za najlepszy model uznano klasyfikator o najwyższych wartościach przyjętych kryteriów. Analogiczny schemat zastosowano w przypadku pozostałych wykorzystanych algorytmów.

Model o najlepszych wartościach kryteriów wykorzystanych do oceny został zbudowany przy wykorzystaniu zmiennych zaakceptowanych przez algorytm Boruta. Najlepsze wyniki osiągnięto dzięki zastosowaniu następujących wartości parametrów:

* maksymalna głębokość drzew decyzyjnych: 1,
* zwiększenie szybkości uczenia się: 0,01,
* liczba drzew decyzyjnych zbudowanych podczas trenowania algorytmu: 100.

W celu porównania najlepszych modeli zbudowanych na podstawie poszczególnych algorytm utworzono tabele kontyngencji oraz wykreślono krzywe ROC oraz krzywe przedstawiające skuteczność modelu w kwestii predykcji obserwacji pozytywnych. Podsumowanie wyników osiągniętych przez zbudowany model zostało przedstawione w poniższej tabeli[[30]](#footnote-30).

Tabela 3. Wyniki osiągnięte przez model Gradient Boosting Machine.

|  |  |  |  |
| --- | --- | --- | --- |
| Zbiór treningowy | | Zbiór testowy | |
| AUC | Wskaźnik 75% | AUC | Wskaźnik 75% |
| 0,932 | 0,982 | 0,930 | 0,981 |

Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

Z tabeli można odczytać, że wartości wskaźników na obu zbiorach są do siebie bardzo zbliżone, co oznacza, że model nie jest przetrenowany i niedotrenowany i zachowuje wysoką jakość predykcji również na danych wykorzystanych do jego walidacji.

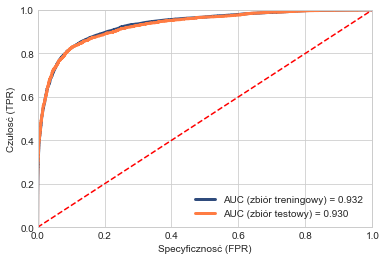
Tabela 4. Macierz błędów dla modelu Gradient Boosting Machine na zbiorze testowym.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predykcja | |
| Klasa pozytywna | Klasa negatywna |
| Rzeczywistość | Klasa pozytywna | 2 789 (26,72%) | 958 (9,18%) |
| Klasa negatywna | 381 (3,65%) | 6 310 (60,45%) |

Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

Z tabeli można odczytać, że model nie rozpoznał prawidłowo 958 spośród 3 747 firm, które upadły. Model prawidłowo zaklasyfikował zatem 74,43% podmiotów, które zbankrutowały. Jest to podstawowy wskaźnik do oceny jego jakości w kontekście omawianego problemu biznesowego. Należy jednak porównać te wartości z wynikami osiąganymi przez pozostałe algorytmy.

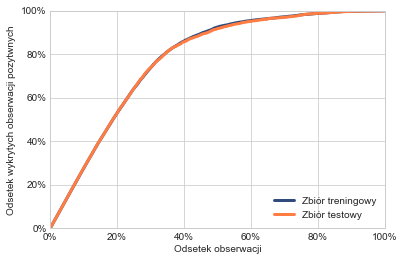
Wykres 7. Krzywe ROC dla modelu Gradient Boosting Machine.



Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

Z wykresu można odczytać, że model zachowuje się bardzo podobnie na obu wykorzystanych zbiorach i osiągane wyniki są stabilne. Podobnie wygląda zdolność modelu do wykrywania obserwacji pozytywnych.

Wykres 8. Zdolność modelu Gradient Boosting Machine do wykrywania obserwacji pozytywnych.



Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

## III.2 Model klasyfikacyjny zbudowany na podstawie algorytmu klasyfikacyjnego lasu losowego.

Jako drugi algorytm klasyfikacyjny wykorzystano klasyfikacyjny las losowy (ang. Random Forest lub Decision Forest). Metoda ta podobnie jak Gradient Boosting Machine opiera się na budowaniu grup drzew decyzyjnych. Las losowy w istocie składa się z dużej liczby pojedynczych drzew decyzyjnych o niskiej złożoności, które działają wspólnie w celu poprawy wyniku. Każde pojedyncze drzewo w losowym lesie generuje prognozę na temat wartości zmiennej celu dla danej obserwacji a klasa z największą liczbą głosów staje się główną prognozą modelu. Model bazuje w pewien sposób na założeniu, że wystarczająco duża liczba niezależnych modeli o stosunkowo niskiej jakości osiągnie łącznie wyższą jakość predykcji niż pojedynczy model o wysokim dopasowaniu do danych. Zastosowana metoda pozwala osiągnąć wysoką dokładność, ceną jest jednak gorsza interpretowalność w porównaniu do jednego drzewa decyzyjnego.

W celu zapewnienia niezależności pojedynczych zbudowanych drzew decyzyjnych stosowane są 2 techniki. Po pierwsze, dane wejściowe, na których trenowane są modele różnią się. Drzewa decyzyjne są bardzo wrażliwe na różnice w zbiorze treningowym. Korzystając z tej zależności algorytm dla każdego drzewa losuje ze zwracaniem inną próbę, na której budowany jest model. Metoda ta nazywana jest bootstrapowaniem. Po drugie, parametry na których budowane są pojedyncze drzewa są losowe. Należą do nich czynniki takie jak np. maksymalna głębokość drzewa, liczba wykorzystanych obserwacji, liczba wykorzystanych zmiennych oraz same zmienne, które zostały uwzględnione. Dzięki losowemu wyborowi tych parametrów pojedyncze modele będą różnić się od siebie. Następnie z wykorzystaniem mechanizmu maksymalizacji uzysku informacji (np. entropia) z danych treningowych wybierane są kolejne atrybuty do budowania reguł decyzyjnych i podziału obserwacji. Cały proces powtarzany jest aż do wytrenowania zdefiniowanej liczby drzew decyzyjnych[[31]](#footnote-31).

Podobnie jak w przypadku poprzedniego algorytmu modele budowano w iteracyjny sposób w celu znalezienia optymalnych wartości parametrów przy wykorzystaniu metody przeszukiwania za pomocą siatki. Optymalizacja objęła następujące parametry:

* liczba drzew decyzyjnych zbudowanych podczas trenowania algorytmu (zakres wartości od 20 do 50 z interwałem 1),
* maksymalna głębokość drzew decyzyjnych (zakres wartości od 1 do 3 z interwałem 1),
* minimalna liczba obserwacji wymaganych do podziału węzła (zakres wartości od 5 do 50 z interwałem 5)[[32]](#footnote-32).

Model o najlepszych wartościach kryteriów wykorzystanych do oceny został zbudowany przy wykorzystaniu zmiennych zaakceptowanych przez algorytm Boruta. Najlepsze wyniki osiągnięto dzięki zastosowaniu następujących wartości parametrów:

* liczba drzew decyzyjnych zbudowanych podczas trenowania algorytmu: 23,
* maksymalna głębokość drzew decyzyjnych: 1,
* minimalna liczba obserwacji wymaganych do podziału węzła: 5.

Analogicznie jak w przypadku poprzedniego algorytmu utworzono tabele kontyngencji dla macierzy błędów oraz wykreślono krzywe ROC oraz krzywe przedstawiające skuteczność modelu w kwestii predykcji obserwacji pozytywnych. Podsumowanie wyników osiągniętych przez zbudowany model zostało przedstawione w poniższej tabeli.

Tabela 5. Wyniki osiągnięte przez model klasyfikacyjnego lasu losowego.

|  |  |  |  |
| --- | --- | --- | --- |
| Zbiór treningowy | | Zbiór testowy | |
| AUC | Wskaźnik 75% | AUC | Wskaźnik 75% |
| 0,897 | 0,973 | 0,895 | 0,972 |

Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

Z tabeli można odczytać, że mimo osiągnięcia wysokich wartości kryteriów do oceny jakości modelu, to w porównaniu z klasyfikatorem zbudowanym z wykorzystaniem algorytmu Gradient Boosting Machine wyniki okazały się gorsze. Pole pod krzywą ROC jest znacznie mniejsze, natomiast zdolność modelu do wykrywania obserwacji pozytywnych pogorszyła się jedynie nieznacznie. Różnice nie są jednak znaczące i na tym etapie nie można jeszcze odrzucić zbudowanego modelu na korzyść poprzedniego klasyfikatora.

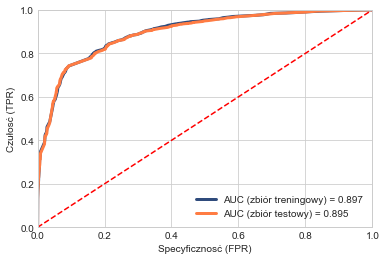
Tabela 6. Macierz błędów dla modelu klasyfikacyjnego lasu losowego na zbiorze testowym.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predykcja | |
| Klasa pozytywna | Klasa negatywna |
| Rzeczywistość | Klasa pozytywna | 3 027 (29,00%) | 720 (6,90%) |
| Klasa negatywna | 1 208 (11,57%) | 5 483 (52,23%) |

Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

Z tabeli można odczytać, że model nie rozpoznał prawidłowo 720 spośród 3 747 firm, które upadły, co jest wynikiem znacznie lepszym w porównaniu do poprzedniego modelu. Model prawidłowo zaklasyfikował 80,78% podmiotów, które zbankrutowały czyli o ponad 6 punktów procentowych więcej niż zbudowany wcześniej model. Zwraca jednak uwagę większa liczba podmiotów zaklasyfikowanych przez model do klasy pozytywnej. Znacznie zwiększył się odsetek podmiotów, które zostały zakwalifikowane do niej nieprawidłowo. Wraz ze wzrostem liczby prawidłowo klasyfikowanych obserwacji pozytywnych spadła ogólna skuteczność modelu. Zbudowany model można zatem ocenić jako bardziej konserwatywny w stosunku do poprzedniego w kwestii podejścia do ryzyka i wyznaczenia progu odcięcia. Nie da się jednoznacznie porównać, który z nich jest lepszy bez kontekstu biznesowego i podejścia do ryzyka decydentów, którzy będą korzystać z końcowych rezultatów.

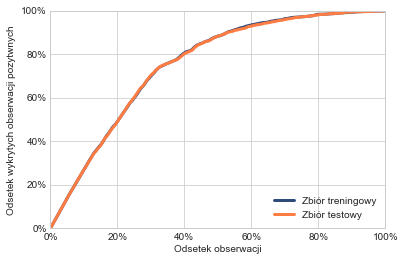
Wykres 9. Krzywe ROC dla modelu klasyfikacyjnego lasu losowego.



Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

Z wykresu można odczytać, że model zachowuje się bardzo podobnie na obu wykorzystanych zbiorach i osiągane wyniki są stabilne. Podobnie wygląda zdolność modelu do wykrywania obserwacji pozytywnych. Zdolność do wykrywania obserwacji pozytywnych jest jednak zauważalnie niższa w porównaniu do modelu zbudowanego wcześniej.

Wykres 10. Zdolność modelu klasyfikacyjnego lasu losowego do wykrywania obserwacji pozytywnych.



Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

## III.3 Model klasyfikacyjny zbudowany na podstawie algorytmu sieci neuronowej.

Ostatnim wykorzystanym algorytmem była sieć neuronowa. Z doświadczenia biznesowego autora jest to metoda, która ze wszystkich trzech wykorzystanych metod sprawdza się najlepiej do rozwiązywania podobnych problemów, pod warunkiem zapewnienia dużej liczby obserwacji oraz zapewniania wystarczającej mocy obliczeniowej i czasu do wytrenowania klasyfikatora. Z tego powodu sieci neuronowe zostały użyte jako ostatnie, po sprawdzeniu początkowych wyników osiągniętych dzięki wykorzystaniu algorytmów, które nie potrzebują tak znacznej mocy obliczeniowej i czasu oczekiwania na wyniki. Należy jednak pamiętać, że w przypadku implementacji biznesowej narzędzia czas potrzebny do trenowania modelu nie będzie istotny z perspektywy generowania wyników. Czas potrzebny do zastosowania już zbudowanego modelu jest zbliżony dla każdego typu algorytmu.

Sieci neuronowe to algorytm, który znajduje zastosowanie do rozwiązywania problemów klasyfikacji dwu i wieloklasowej. Może być również wykorzystany w przypadku problemów regresyjnych. Jest to najbardziej skomplikowany z zastosowanych algorytmów. Inspirowany jest działaniem ludzkiego mózgu i strukturami genetycznymi.

Sieci neuronowe składają się z wielu warstw neuronów połączonych ze sobą synapsami, które komunikują się ze sobą pomiędzy poszczególnymi warstwami i uczą się na podstawie dostarczonych informacji. Warstwy skupiają grupy neuronów. Każdy neuron jest połączony z innym neuronem warstwy poprzedzającej. Każdy neuron przeprowadza własne proste obliczenia, a sieć, którą tworzą wszystkie neurony, zwielokrotnia potencjał tych obliczeń. Najważniejszym elementem sieci neuronowej jest perceptron, czyli neuron przetwarzający, który skupia informacje z warstwy wejściowej. Perceptron posiada kilka wejść, do których przypisane są wagi (np. waga wpływu zmiennej objaśniającej na zmienną celu), określające jak duży wpływ ma wielkość z danego wejścia na wynik końcowy. W sieci znajduje się wiele neuronów, które posiadają dowolną liczbę wejść i wyjść.

Wartości zmiennych są zatem przekazywane kolejno pomiędzy poszczególnymi warstwami sieci, gdzie zachodzi proces uczenia się sieci. W każdej z warstw wykonywane są operacje na zmiennych, aż do osiągnięcia wartości wynikowej w ostatniej warstwie wyjściowej, która agreguje zebrane informacje, gromadzi wnioski i dostarcza wyniki analizy[[33]](#footnote-33).

Podobnie jak w przypadku pozostałych algorytmów modele budowano w iteracyjny sposób w celu znalezienia optymalnych wartości hiper parametrów. Z powodu długiego czasu wykonywania obliczeń oraz o wiele bardziej obszernej możliwej przestrzeni rozwiązań zrezygnowano jednak z przeszukiwania za pomocą metody siatki. Zamiast tego parametry były losowane podczas każdej iteracji. Łącznie wykonano po 2 000 iteracji na pełnym zbiorze danych oraz zbiorze okrojonym o zmienne odrzucone przez algorytm Boruta w celu znalezienia możliwie najlepszych wartości hiper parametrów. Optymalizacja objęła następujące parametry:

* zwiększenie szybkości uczenia się (ang. learning rate, zakres wartości od 0,01 do 0,15 z interwałem 0,01),
* liczba obserwacji wykorzystanych w pojedynczej iteracji do propagacji modelu (ang. batch size. zakres wartości od 20 do 110 z interwałem 10),
* liczba iteracji do wykonania na całym zbiorze danych (ang. epochs, zakres wartości od 50 do 300 z interwałem 25),
* liczba warstw ukrytych (zakres wartości od 1 do 6 z interwałem 1),
* szansa na odrzucenie neuronów w warstwie w celu uniknięcia przetrenowania modelu (ang. dropout, zakres wartości od 0,1 do 0,4 z interwałem 0,05),
* metoda optymalizacji wykorzystanej funkcji straty (zakres wartości: Adam, SGD),
* wykorzystana funkcja straty: (zakres wartości: binarna entropia krzyżowa (ang. binary crossentropy), średni błąd bezwzględny (ang. mean absolute error)[[34]](#footnote-34).

Model o najlepszych wartościach kryteriów wykorzystanych do oceny został zbudowany przy wykorzystaniu zmiennych zaakceptowanych przez algorytm Boruta. Najlepsze wyniki osiągnięto dzięki zastosowaniu następujących wartości parametrów:

* zwiększenie szybkości uczenia się: 0,05,
* liczba obserwacji wykorzystanych w pojedynczej iteracji do propagacji modelu: 40,
* liczba iteracji do wykonania na całym zbiorze danych: 100,
* liczba warstw ukrytych: 4,
* szansa na odrzucenie neuronów w warstwie: 0,1,
* metoda optymalizacji funkcji straty: Adam,
* funkcja straty: binarna entropia krzyżowa.

Analogicznie jak w przypadku poprzedniego algorytmu utworzono tabele kontyngencji oraz wykreślono krzywe ROC i krzywe przedstawiające skuteczność modelu w kwestii predykcji obserwacji pozytywnych. Podsumowanie wyników osiągniętych przez zbudowany model zostało przedstawione w poniższej tabeli.

Tabela 7. Wyniki osiągnięte przez model sieci neuronowych.

|  |  |  |  |
| --- | --- | --- | --- |
| Zbiór treningowy | | Zbiór testowy | |
| AUC | Wskaźnik 75% | AUC | Wskaźnik 75% |
| 0,955 | 0,995 | 0,941 | 0,991 |

Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

Z tabeli można odczytać, że mimo niewielkiego przetrenowania biorąc pod uwagę wszystkie 4 kryteria model osiągnął bardzo wysokie wyniki, lepsze w porównaniu do poprzednich 2 zbudowanych klasyfikatorów. Należy zwrócić jednak uwagę na różnice w wartościach wskaźnika AUC na obu wykorzystanych zbiorach. Model dopasowuje się zbytnio do danych treningowych i jego wyniki spadają na danych walidacyjnych. Wciąż są jednak najlepsze ze wszystkich 3 modeli na obu wykorzystanych zbiorach. To zbiór testowy powinien być najważniejszym wyznacznikiem, co oznacza, że klasyfikator zbudowany z użyciem sieci neuronowych okazał się najbardziej skuteczny.

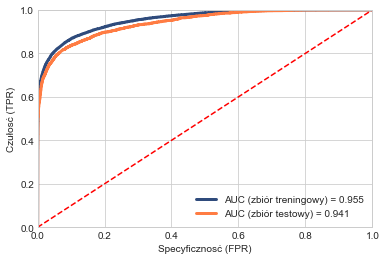
Tabela 8. Macierz błędów dla modelu sieci neuronowych na zbiorze testowym.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predykcja | |
| Klasa pozytywna | Klasa negatywna |
| Rzeczywistość | Klasa pozytywna | 2 966 (28,42%) | 781 (7,48%) |
| Klasa negatywna | 378 (3,62%) | 6 313 (60,48%) |

Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

Z tabeli można odczytać, że model nie rozpoznał prawidłowo 781 spośród 3 747 firm, które upadły, co jest wynikiem znacznie lepszym w porównaniu do klasyfikatora Gradient Boosting Machine ale jednocześnie gorszym poprzedniego modelu lasu losowego. Model prawidłowo zaklasyfikował 79,16% podmiotów, które zbankrutowały czyli o 1,62 punktu procentowego mniej niż najlepszy pod tym względem las losowy. Należy jednak zwrócić uwagę, że ogólna skuteczność modelu jest zdecydowanie wyższa w porównaniu do lasu losowego, który fałszywie klasyfikował 11,57% obserwacji jako pozytywnie. Sieć neuronowa popełniła ten błąd dla 3,62% obserwacji co jest najlepszym wynikiem ze wszystkich modeli. Różnica ta jest częściowo spowodowana rozbieżnościami w wartościach progów odcięcia wymaganych do zaklasyfikowania obserwacji do klasy pozytywnej. Nie można jednak jednoznacznie stwierdzić, że jeden bądź drugi model jest lepszy od konkurenta, charakteryzują się innymi właściwościami i pod uwagę należy wziąć również rodzaj problemu biznesowego, do którego ma zostać wykorzystany końcowy model. Priorytetem jest jak najbardziej trafna predykcja klasy pozytywnej.

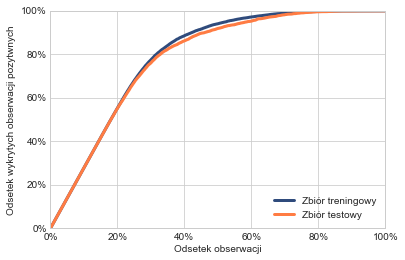
Wykres 11. Krzywe ROC dla modelu sieci neuronowych



Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

Z wykresu można odczytać podobne wnioski jak z porównania wartości AUC na zbiorze treningowym i testowym. Widoczne jest przetrenowanie. Różnice nie są jednak znaczące, ponadto wyniki osiągnięte na zbiorze testowym są lepsze w porównaniu do pozostałych modeli. Podobnie wnioski można wyciągnąć z wykresu przedstawiającego zdolność do wykrywania obserwacji pozytywnych. W tym wypadku również można zaobserwować zbytnie dopasowanie do danych treningowych i mniejszą zdolność do generalizowania. Różnica ta jest jednak nieco mniej widoczna w porównaniu do rozbieżności pomiędzy krzywymi ROC. Mimo tej różnicy, biorąc pod uwagę drugi wskaźnik wyniki również są najbardziej satysfakcjonujące na obu wykorzystanych zbiorach danych.

Wykres 12. Zdolność modelu sieci neuronowych do wykrywania obserwacji pozytywnych.



Źródło: Opracowanie własne na podstawie wyników modelu zbudowanego na podstawie zastosowanego algorytmu.

## III.4 Porównanie i ocena jakości zbudowanych modeli klasyfikacyjnych

W celu porównania wyników osiągniętych przez poszczególne modele i wyboru najlepszego klasyfikatora zagregowano wartości metryk wykorzystanych do ich oceny. Pod uwagę wzięto wartości osiągnięte na zbiorze testowym.

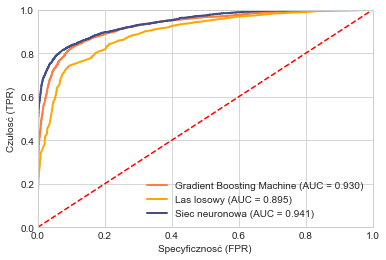
Tabela 9. Porównanie wyników zbudowanych modeli na zbiorze testowym.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Gradient Boosting Machine | Las losowy | Sieć neuronowa |
| AUC | 0,930 | 0,895 | 0,941 |
| Wskaźnik 75% | 0,981 | 0,972 | 0,991 |

Źródło: Opracowanie własne na podstawie wyników zbudowanych modeli.

Z tabeli można odczytać, że biorąc pod uwagę oba wskaźniki wykorzystane do oceny zbudowanych modeli sieć neuronowa wykazała się największą jakością predykcji na zbiorze testowym. Podobne wnioski można wyciągnąć z analizy krzywych ROC oraz krzywych przedstawiających zdolność modelu do wykrywania obserwacji pozytywnych, czyli podstawowego zadania modelu w kontekście implementacji biznesowej budowanego narzędzia.

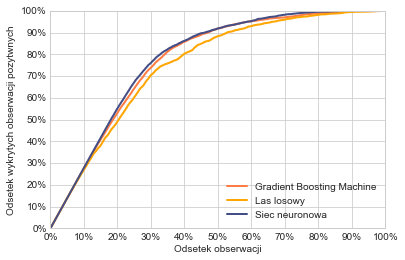
Wykres 13. Porównanie wyników zbudowanych modeli na zbiorze testowym - krzywe ROC.



Źródło: Opracowanie własne na podstawie wyników zbudowanych modeli.

Z wykresu można odczytać, że pomiędzy klasyfikatorami zbudowanymi na podstawie algorytmów sieci neuronowej i Gradient Boosting Machine a lasem losowym jest znaczna różnica jeśli chodzi o ogólną skuteczność predykcji. Biorąc pod uwagę analityczne wskaźniki oceny jakości modelu to sieć neuronowa osiągnęła najlepsze wyniki. Należy jednak pamiętać o kontekście biznesowym.

Wykres 14. Porównanie wyników zbudowanych modeli na zbiorze testowym - zdolność do wykrywania obserwacji pozytywnych.



Źródło: Opracowanie własne na podstawie wyników zbudowanych modeli.

Z wykresu można odczytać, że również w kwestii wykrywania obserwacji pozytywnych i zdolności do gromadzenia upadłych przedsiębiorstw w jak najmniejszej części badanej populacji o największej ocenie ryzyka model zbudowany na podstawie sieci neuronowej dostarcza najlepszych rezultatów. Oceny ryzyka nadawane przez ten model sprawdzają się lepiej w porównaniu z pozostałymi klasyfikatorami i statystycznie podmioty o przyznanej wysokiej wartości częściej kończyły działalność gospodarczą. Jest to kluczowe kryterium do zastosowania danego klasyfikatora, oprócz wysokich wartości miar stricte analitycznych powinien dostarczać wymiernych rezultatów, które można przełożyć na zastosowanie biznesowe. Model zbudowany na podstawie algorytmu sieci neuronowej spełnia oba te kryteria i to on powinien zostać wykorzystany do implementacji narzędzia.

## III.5 Interpretacja uzyskanych wyników i wyciągnięcie wniosków.

Uzyskane wyniki są w opinii autora satysfakcjonujące i uzasadniają prowadzenie dalszych prac nad wdrożeniem stworzonego narzędzia do rozwiązywania problemów, które mają miejsce w rzeczywistym świecie. Skuteczność zbudowanego modelu jest na tyle wysoka, że możliwe jest wykorzystanie go w praktyce i jest ona porównywalna z analogicznymi klasyfikatorami zbudowanymi w przeszłości podczas podobnych analiz. Wydaje się, że w kontekście prezentacji wyników modeli optymalnym rozwiązaniem jest prezentowanie użytkownikowi prognoz zwracanych przez wszystkie trzy algorytmy. Jest to korzystny zabieg w celu zwiększenia zaufania użytkownika, ponieważ różnice pomiędzy prognozami nie są znaczne, a ich zbieżność powinna budować zaufanie wśród osób, które nie posiadają wiedzy niezbędnej do ich interpretacji.

Powszechnie rozpoznawany i wykorzystywany model Z-score zbudowany przez E. Altmana na przestrzeni 30 lat testów wykazywał się skutecznością przewidywania upadłości firmy rok przed domniemanym zdarzeniem na poziomie 80-90%, podczas gdy błąd typu drugiego (nieprawidłowe zaklasyfikowanie firmy jako bankruta) utrzymywał się na poziomie 10-15%[[35]](#footnote-35).

Dla porównania, model zbudowany z użyciem algorytmu sieci neuronowej osiągnął skuteczność przewidywania bankructwa na poziomie 79,16% przy jednoczesnym błędzie predykcji typu drugiego na poziomie 3,62%. Klasyfikator wytrenowany na podstawie algorytmu lasu losowego osiągnął natomiast statystyki na poziomie 80,78% oraz 11,57%. Należy jednak pamiętać, że modele zbudowane z użyciem metod uczenia maszynowego są znacznie trudniej interpretowalne i bardziej rozbudowane w porównaniu z wieloma aktualnie wykorzystywanymi narzędziami stworzonymi w oparciu o prostsze metody, na przykład z wykorzystaniem regresji logistycznej, czy zwykłego równania arytmetycznego jak ma to miejsce w przypadku modelu E. Altmana. Należy pamiętać, że modele obecnie wykorzystywane do predykcji bankructwa firm (na przykład porównywany model E. Altmana) zostały przetestowane w bardzo szerokim horyzoncie czasowym i nie można stwierdzić jak będzie wyglądała skuteczność narzędzia w nadchodzących latach. Powinno ono być testowane i monitorowane.

Prawdopodobnym scenariuszem jest także konieczność wykonania analizy ponownie, np. w cyklicznych interwałach czasowych w celu jego modernizacji i dostosowania do wzorców danych, które mogą zmieniać się w zależności od czynników makroekonomicznych, takich jak choćby cykliczność gospodarki lub globalne szoki ekonomiczne, jak na przykład obecny spowodowany pandemią wirusa COVID-19. Nakład pracy i czasu potrzebny do wykonania takiej analizy podobnie będzie jednak znacznie niższy, ponieważ stworzony został już proces do pobierania i procesowania danych i całe rozwiązanie jest odtwarzalne i łatwe w utrzymaniu.

Próba zbudowania narzędzia zakończyła się powodzeniem w pierwszym etapie. Osiągnięto zadowalające wyniki i udało się wykazać, że możliwe jest zbudowanie modeli o wystarczającej skuteczności do predykcji bankructwa. W celu implementacji biznesowej wskazane są jednak dalsze prace. Poprawione powinny zostać szczególnie proces pobierania niezbędnych danych i ich oczyszczania oraz końcowy produkt dostarczony użytkownikom. Udostępnienie możliwości predykcji bankructwa firm może okazać się niewystarczające z powodu niskiej efektowności narzędzia z jedynie tą funkcją. Wskazane jest dodanie do niego wizualizacji prezentujących np. zastosowane modele, wpływ uwzględnionych czynników ryzyka czy zmianę ryzyka kontrahentów w czasie. Główna część niezbędna do jego stworzenia została jednak dostarczona i już na obecnym etapie prac możliwe jest wyciągnięcie korzyści biznesowych z jego zastosowania.

# Zakończenie.

Mimo początkowych obiecujących wyników nie można na tym etapie jednoznacznie stwierdzić, czy cel pracy, jakim było zbudowanie narzędzia do autonomicznej predykcji przedsiębiorstw na masową skalę został zrealizowany. Początkowe wyniki są satysfakcjonujące, natomiast konieczne jest sprawdzenie stabilności osiąganych rezultatów w czasie i przetestowanie narzędzia w dłuższym horyzoncie czasowym. Wydaje się jednak, że zbudowanie w pełni autonomicznego narzędzia, które będzie funkcjonować bez działania użytkownika, który będzie je obsługiwał jest na tym etapie prac trudne do zrealizowania. Jak okazało się, głównym wyzwaniem nie jest zbudowanie modelu o satysfakcjonującej jakości predykcji ale zapewnienie odpowiedniej jakości danych wejściowych o właściwym pokryciu.

Podczas budowy narzędzia okazało się, że istotnym problemem jest występowanie dużej liczby braków danych. W celu maksymalizacji potencjału modeli predykcyjnych wskazane jest dopracowanie procesu pozyskiwania informacji lub dodanie możliwości imputacji przez użytkownika. Na polskim rynku funkcjonują podmioty, które na bieżąco dostarczają odpłatnie dane rejestrowe i finansowe w przystępnej formie na temat firm, dla których składane jest zapytanie. Dużą zaletą wykorzystania tego rozwiązania jest to, że dostawcy pracują na sprawozdaniach finansowych w polskim standardzie oraz zazwyczaj możliwe jest pozyskanie danych na konkretny rok finansowy. Wiąże się to jednak z ponoszeniem kosztów za każde zapytanie. W celu poprawy jakości i pokrycia danych możliwe jest pozyskiwanie ich w ten sposób lub dodanie do narzędzia możliwości ręcznego wprowadzenia brakujących danych na przykład przez dedykowany formularz. W takim wypadku możliwe jest podjęcie próby budowy prostszego modelu z wykorzystaniem mniejszej liczby zmiennych w celu ograniczenia konieczności imputacji. Model taki byłby łatwiej interpretowalny oraz skróciłby się czas jego trenowania. Wysiłek potrzebny do przeprowadzenia weryfikacji kontrahenta przy zastosowaniu tego scenariusza zostaje jednak i tak zminimalizowany a wyniki są wzbogacone o informacje, które trudno jest pozyskać w standardowy sposób a dostarczają wartości dodanej użytkownikowi.

Wydaje się jednak, że dużą wartością biznesową wnosi możliwość automatycznego zebrania informacji dla weryfikowanych firm i połączenia różnych źródeł danych, które dostarczają kompletnego obrazu działalności danego przedsiębiorstwa. Osoby przeprowadzające analizę kontrahenta docenią takie ułatwienie, ponieważ będą mogły poświęcić więcej czasu na interpretację otrzymanych wyników zamiast w sposób manualny pozyskiwać i agregować potrzebne do tego celu dane.

W celu wdrożenia zbudowanego narzędzia wskazane jest udostępnienie go w formie online końcowym użytkownikom, na przykład w postaci aplikacji webowej, do której będą składane zapytania na temat firm za pomocą identyfikatorów NIP, REGON lub KRS. Po zebraniu i przeprocesowaniu danych oraz uruchomieniu modelu użytkownikowi będą prezentowane w formie panelu z danymi informacje na temat weryfikowanych firm. Oprócz tego użytkownik otrzyma oceny ryzyka badanych przedsiębiorstw w kontekście możliwości ich bankructwa wraz z binarną predykcją, czy firma upadnie w rozważanym horyzoncie czasowym. Wskazane jest również dodanie do panelu interpretacji czynników ryzyka wykorzystanych do budowy modelu i omówienie jak wpływają one na ryzyko bankructwa. Dodatkową wartość wniesie także dodanie możliwości przeprowadzenia cyklicznej weryfikacji (np. kwartalnej) w celu monitorowania ryzyka w czasie oraz dodanie systemu wczesnego ostrzegania uruchamianego w przypadku, gdy sytuacja danej firmy będzie stopniowo pogarszać się w czasie wykonywania kolejnych analiz.

# Bibliografia.

## I. Wydawnictwa zwarte.

Sozański T., *Teoria gier i decyzji dla socjologów i psychologów*, 2005,

Lowrance W., *Of Acceptable Risk: Science and the Determination of Safety*, William Kaufman In, Los Altos 1976,

Kościelniak H., *An improvement of information processes in enterprises—The analysis of sales profitability in the manufacturing company using ERP systems,* Polish Journal of Management Studies 2014,

Mach M., *Dynamiczna analiza barier wejścia na rynek jako zadanie dla narzędzia inteligentnego*, 2003,

Andrew B., Salemat A., *Systematic Literature Review of Missing Data Imputation Techniques for Effort Prediction*, 2012,

Cabras S. Morales J., *Extreme value analysis within a parametric outlier detection framework. Applied Stochastic Models in Business and Industry*, Department of Mathematics, University of Cagliari, 2007,

Altman E., *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*, 1968,

Altman E., *Predicting Financial Distress of Companies*, 2000

## II. Artykuły.

1. Horváthová J., Mokrišová M., *Risk of Bankruptcy, Its Determinants and Models*, 2018,
2. Chen S., Xu S., *Comparative study of bankruptcy prediction model*, 2018,
3. Vignan N., Jawaharlal N., *An Exhaustive Literature Review on Class Imbalance Problem*, 2013,
4. Hart P., *The condensed nearest neighbor rule, IEEE Transactions on Information Theory*, 1968,
5. Dormann C. F., *Collinearity: a review of methods to deal with it and a simulation study evaluating their performance*, 2013,
6. Dudek H., *Wykrywanie współliniowości za pomocą scentrowanych, niescentrowanych oraz uogólnionych czynników inflacji wariancji*, 2005,
7. Prusak B., *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstw*, 2005,
8. Korol T., *Systemy Ostrzegania Przedsiębiorstw Przed Ryzykiem Upadłości*, 2010,
9. Korol T., *Early warning models against bankruptcy risk for Central European and Latin American enterprises*, 2013,
10. Patro K., Kumar K., *Normalization: A Preprocessing Stage*, 2015,
11. Potdar K., Pardawala T. S., Pai C. D., *A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers*, 2017,
12. Bhalaji N., Sundharakumar Kb., Chithra S., *Empirical study of feature selection methods over classification algorithms*, 2016

## III. Akty prawne.

Ustawa *Kodeks cywilny z dn. 23.04.1964 r.* art. 55.1 Dz.U.2019.0.1145,

Ustawa *Prawo upadłościowe z dn. 28.02.2003 r.* art. 5-6, 9-11 Dz.U. 2003 poz. 1228 z późn. zm.,

Ustawa *Kodeks spółek handlowych z dnia 15.02.2000 r.* art. 266, 271, 282 Dz.U. 2000 94 poz. 1037,

Ustawa *O podatku od towarów i usług z dnia 11.03.2004 r.* art.. 3, 109 Dz.U.2020.0.106,

Ustawa *O przeciwdziałaniu praniu pieniędzy oraz finansowaniu terroryzmu z dnia 1.03.2018 r.* art.. 72, 157 Dz. U. 2018 poz. 723

## IV. Strony internetowe.

1. www.lawyerline.pl/blog/likwidacja-spolki-z-o-o-czy-upadlosc (data odsłony 18.07.2020 r.),
2. www.coig.com.pl (data odsłony 19.01.2021 r.),
3. www.emis.com (data odsłony 18.08.2020 r.),
4. www.towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab (data odsłony 5.01.2021 r.),
5. www.mateuszgrzyb.pl/wybor-odpowiedniego-algorytmu-czesc-2-algorytmy-klasyfikacyjne (data odsłony 5.01.2021 r.),
6. www.towardsdatascience.com/understanding-random-forest-58381e0602d2 (data odsłony 6.01.2021 r.),
7. www.lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html (data odsłony 6.01.2021 r.),
8. www.scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (data odsłony 8.01.2021 r.),
9. www.sztucznainteligencja.org.pl/definicja/sieci-neuronowe (data odsłony 6.01.2021 r.),
10. www.keras.io/api/models/sequential/ (data odsłony 22.01.2021 r.)

# Spis wykresów.

[Wykres 1. Liczba upadłości i likwidacji firm w latach 2000-2019. - 12 -](#_Toc62568160)

[Wykres 2. Liczba i udział firm aktywnych i zamkniętych w badanej populacji. - 30 -](#_Toc62568161)

[Wykres 3. Liczba firm według województw zgrupowane według klas zmiennej celu – przykładowy histogram dla zmiennych nominalnych. - 40 -](#_Toc62568162)

[Wykres 4. Rozkłady wieku firm zgrupowane według klas zmiennej celu - przykładowy wykres pudełkowy dla zmiennych numerycznych. - 41 -](#_Toc62568163)

[Wykres 5. Rozkłady wieku firm zgrupowane według klas zmiennej celu - przykładowy wykres gęstości dla zmiennych numerycznych. - 42 -](#_Toc62568164)

[Wykres 6. Wpływ zmiennych zaakceptowanych przez algorytm Boruta na zmienną celu. - 45 -](#_Toc62568165)

[Wykres 7. Krzywe ROC dla modelu Gradient Boosting Machine. - 52 -](#_Toc62568166)

[Wykres 8. Zdolność modelu Gradient Boosting Machine do wykrywania obserwacji pozytywnych. - 52 -](#_Toc62568167)

[Wykres 9. Krzywe ROC dla modelu klasyfikacyjnego lasu losowego. - 56 -](#_Toc62568168)

[Wykres 10. Zdolność modelu klasyfikacyjnego lasu losowego do wykrywania obserwacji pozytywnych. - 56 -](#_Toc62568169)

[Wykres 11. Krzywe ROC dla modelu sieci neuronowych - 60 -](#_Toc62568170)

[Wykres 12. Zdolność modelu sieci neuronowych do wykrywania obserwacji pozytywnych. - 61 -](#_Toc62568171)

[Wykres 13. Porównanie wyników zbudowanych modeli na zbiorze testowym - krzywe ROC. - 62 -](#_Toc62568172)

[Wykres 14. Porównanie wyników zbudowanych modeli na zbiorze testowym - zdolność do wykrywania obserwacji pozytywnych. - 62 -](#_Toc62568173)

# Spis tabel.

[Tabela 1. Zbiór dostępnych zmiennych nominalnych. - 31 -](#_Toc62568176)

[Tabela 2. Zbiór dostępnych zmiennych ciągłych. - 34 -](#_Toc62568177)

[Tabela 3. Wyniki osiągnięte przez model Gradient Boosting Machine. - 51 -](#_Toc62568178)

[Tabela 4. Macierz błędów dla modelu Gradient Boosting Machine na zbiorze testowym. - 51 -](#_Toc62568179)

[Tabela 5. Wyniki osiągnięte przez model klasyfikacyjnego lasu losowego. - 54 -](#_Toc62568180)

[Tabela 6. Macierz błędów dla modelu klasyfikacyjnego lasu losowego na zbiorze testowym. - 55 -](#_Toc62568181)

[Tabela 7. Wyniki osiągnięte przez model sieci neuronowych. - 59 -](#_Toc62568182)

[Tabela 8. Macierz błędów dla modelu sieci neuronowych na zbiorze testowym. - 59 -](#_Toc62568183)

[Tabela 9. Porównanie wyników zbudowanych modeli na zbiorze testowym. - 61 -](#_Toc62568184)

# Streszczenie.

Praca porusza temat predykcji upadłości przedsiębiorstw dzięki możliwości zastosowania algorytmów uczenia maszynowego. W obecnych warunkach biznesowych, po zebraniu danych historycznych, które zawierają informacje wskazujące na czynniki determinujące upadłość możliwe jest przeprowadzenie takiej prognozy w określonym, konkretnym horyzoncie czasowym – w tym wypadku 12 miesięcy od wykonania analizy. Posiadanie wiedzy na temat czynników, które wpływają na bankructwo firmy pozwala osobom kierującym własną firmą podejmować trafniejsze decyzje biznesowe a pozostałym uczestnikom rynku na trafniejsze wybieranie partnerów biznesowych.

Biorąc pod uwagę, że podmioty prowadzące działalność gospodarczą są zobowiązane przez polskie prawo do weryfikacji wiarygodności klientów oraz kontrahentów i w tym celu mogą korzystać ze źródeł, które umożliwiają bezpłatne pobieranie danych na temat innych firm na masową skalę możliwe staje się zbudowanie narzędzia, które będzie ten proces automatyzować. Do tego potrzebne jest stworzenie procesu automatycznego pobierania danych, ich rekoncyliacji i przetwarzania a następnie zbudowania modelu predykcyjnego, który będzie oceniał ryzyko upadłości badanych przedsiębiorstw. Praca opisuje proces budowy takiego narzędzia, szczegółowe uzasadnienie biznesowe do jego stworzenia, korzyści z jego wykorzystania oraz możliwości implementacji.

W pracy opisane zostały pojęcie ryzyka bankructwa lub likwidacji firm, skala zjawiska upadłości firm w Polsce na przestrzeni lat, regulacje którymi są objęte polskie firmy w kontekście weryfikacji swoich partnerów biznesowych i potencjalne korzyści płynące z automatyzacji tej czynności. Następnie opisane zostały wykorzystane źródła danych oraz czynniki ryzyka upadłości, które zostały uwzględnione w zbudowanych modelu klasyfikacyjnym jako zmienne predykcyjne. W dalszej kolejności przedstawiono proces budowy modeli predykcyjnych opartych na algorytmach uczenia maszynowego i wybór najlepszego z nich.

Pracę zamyka wyciągnięcie wniosków z przeprowadzonej analizy, interpretacja wyników i omówienie przykładowej implementacji i możliwości wykorzystania zbudowanego narzędzia do rozwiązywania problemów biznesowych, przed którymi stają polscy przedsiębiorcy i podmioty w ich otoczeniu gospodarczym.

Załącznik nr 2 do Zasad weryfikacji samodzielności przygotowania pracy dyplomowej oraz elektronicznej archiwizacji prac dyplomowych

**OŚWIADCZENIE AUTORA PRACY DYPLOMOWEJ MAGISTERSKIEJ**

**pod tytułem** ................................................................................................................................

......................................................................................................................................................

**napisanej przez**: ..........................................................**nr albumu** ............................................

**pod kierunkiem** .........................................................................................................................

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została

napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny

z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca dyplomowa nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy dyplomowej jest identyczna z załączoną wersją elektroniczną.

Wyrażam zgodę na poddanie pracy dyplomowej kontroli, w tym za pomocą programu wychwytującego znamiona pracy niesamodzielnej, zwanego dalej programem, oraz na umieszczenie tekstu pracy dyplomowej w bazie porównawczej programu, w celu chronienia go przed nieuprawnionym wykorzystaniem, a także przekazanie pracy do Ogólnopolskiego Repozytorium Prac Dyplomowych.

Wyrażam także zgodę na przetwarzanie przez Szkołę Główną Handlową w Warszawie moich danych osobowych umieszczonych w pracy dyplomowej w zakresie niezbędnym do jej kontroli za pomocą programu oraz w zakresie niezbędnym do jej archiwizacji i nieodpłatnego udostępniania na zasadach określonych w zarządzeniu.

…………………………………….…………………………………..

26.01.2021 r.

1. T. Sozański, *Teoria gier i decyzji dla socjologów i psychologów*, 2005, s. 8. [↑](#footnote-ref-1)
2. W. Lowrance, *Of Acceptable Risk:* *Science and the Determination of Safety*,1976.. Los Altos: William Kaufman Inc. [↑](#footnote-ref-2)
3. H. Kościelniak, *An improvement of information processes in enterprises—The analysis of sales profitability in the manufacturing company using ERP systems* 2014, Polish Journal of Management Studies 10, s. 65–71. [↑](#footnote-ref-3)
4. *Ustawa Kodeks cywilny z dnia 23.04.1964 r.* art. 55.1 Dz.U.2019.0.1145 [↑](#footnote-ref-4)
5. *Ustawa Prawo upadłościowe z dn. 28.02.2003 r*. art. 5-6, 9-11 Dz.U. 2003 poz. 1228 z późn. zm. [↑](#footnote-ref-5)
6. *Ustawa Kodeks spółek handlowych z dnia 15.02.2000 r.* art. 266, 271, 282 Dz.U. 2000 94 poz. 1037 [↑](#footnote-ref-6)
7. www.lawyerline.pl/blog/likwidacja-spolki-z-o-o-czy-upadlosc (data odsłony 18.07.2020 r.) [↑](#footnote-ref-7)
8. [www.coig.com.pl](http://www.stooq.pl) (data odsłony 19.01.2021 r.) [↑](#footnote-ref-8)
9. Ustawa *O podatku od towarów i usług z dnia 11.03.2004 r.* art.. 3, 109 Dz.U.2020.0.106 [↑](#footnote-ref-9)
10. Ustawa *O przeciwdziałaniu praniu pieniędzy oraz finansowaniu terroryzmu z dnia 1.03.2018 r.* art.. 72, 157 Dz. U. 2018 poz. 723 [↑](#footnote-ref-10)
11. M. Mach, *Dynamiczna analiza barier wejścia na rynek jako zadanie dla narzędzia inteligentnego*, 2003. [↑](#footnote-ref-11)
12. www.emis.com (data odsłony 18.08.2020 r.) [↑](#footnote-ref-12)
13. J. Horváthová, M. Mokrišová, *Risk of Bankruptcy, Its Determinants and Models*, 2018, s. 2-5. [↑](#footnote-ref-13)
14. W celu zbudowania modelu klasyfikacyjnego do danych wejściowych zostały włączone podmioty upadłe i te, które kontynuują działalność. Firmy, które na moment pobrania danych prowadziły w aktywny sposób działalność gospodarczą traktowane są jako obserwacje negatywne – dla tych firm zdarzenie upadłości nie zmaterializowało się (nie zakończyło się „sukcesem” tylko „porażką”). Są zatem one traktowane jako obserwacje negatywne i tak też nazywane według przyjętej konwencji. Analogicznie - firmy, które zbankrutowały są traktowane jako obserwacje pozytywne. [↑](#footnote-ref-14)
15. S. Chen, S. Xu*, Comparative study of bankruptcy prediction model*, 2018, s. 10-18. [↑](#footnote-ref-15)
16. N. Vignan, N. Jawaharlal, *An Exhaustive Literature Review on Class Imbalance Problem*, 2013, s. 3 [↑](#footnote-ref-16)
17. P. Hart, *The condensed nearest neighbor rule*, IEEE Transactions on Information Theory, 1968, s. 1-2 [↑](#footnote-ref-17)
18. B. Andrew, A. Salemat, *Systematic Literature Review of Missing Data Imputation Techniques for Effort Prediction*, 2012, s. 223-225 [↑](#footnote-ref-18)
19. S. Cabras, J. Morales, *Extreme value analysis within a parametric outlier detection framework. Applied Stochastic Models in Business and Industry*, Department of Mathematics, University of Cagliari, 2007, s. 158 – 164. [↑](#footnote-ref-19)
20. C. F. Dormann, *Collinearity: a review of methods to deal with it and a simulation study evaluating their performance*, 2013, s. 27-32. [↑](#footnote-ref-20)
21. H. Dudek, *Wykrywanie współliniowości za pomocą scentrowanych, niescentrowanych oraz uogólnionych czynników inflacji wariancji*, 2005, s. 1-4. [↑](#footnote-ref-21)
22. E. Altman, F*inancial ratios, discriminant analysis and the prediction of corporate bankruptcy*, 1968, s. 589-609. [↑](#footnote-ref-22)
23. B. Prusak, *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstw*, 2005. [↑](#footnote-ref-23)
24. T. Korol, *Systemy Ostrzegania Przedsiębiorstw Przed Ryzykiem Upadłości*, 2010, T. Korol, *Early warning models against bankruptcy risk for Central European and Latin American enterprises*, 2013, s. 22-30. [↑](#footnote-ref-24)
25. K. Patro, K. Kumar, *Normalization: A Preprocessing Stage*, 2015, s. 1-3. [↑](#footnote-ref-25)
26. K. Potdar, T. S. Pardawala, C. D. Pai, *A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers*, 2017, s. 7-10. [↑](#footnote-ref-26)
27. N. Bhalaji, Kb. Sundharakumar, S. Chithra, *Empirical study of feature selection methods over classification algorithms*, 2016, s. 5-19. [↑](#footnote-ref-27)
28. www.towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab (data odsłony 5.01.2021 r.),

    www.mateuszgrzyb.pl/wybor-odpowiedniego-algorytmu-czesc-2-algorytmy-klasyfikacyjne (data odsłony 5.01.2021 r.) [↑](#footnote-ref-28)
29. www.lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html (data odsłony 6.01.2021 r.) [↑](#footnote-ref-29)
30. Wskaźnik 75% reprezentuje odsetek wykrytych wszystkich obserwacji pozytywnych, w populacji 75% podmiotów wskazanych przez model jako najbardziej zagrożone ryzykiem upadłości. [↑](#footnote-ref-30)
31. www.towardsdatascience.com/understanding-random-forest-58381e0602d2 (data odsłony 8.01.2021 r.) [↑](#footnote-ref-31)
32. www.scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (data odsłony 6.01.2021 r.) [↑](#footnote-ref-32)
33. www.sztucznainteligencja.org.pl/definicja/sieci-neuronowe (data odsłony 6.01.2021 r.) [↑](#footnote-ref-33)
34. www.keras.io/api/models/sequential/ (data odsłony 22.01.2021 r.) [↑](#footnote-ref-34)
35. E. Altman, *Predicting Financial Distress of Companies*, 2000, s. 15-22 [↑](#footnote-ref-35)