Studium licencjackie

Kierunek: Metody ilościowe w ekonomii i systemy informacyjne
Specjalność:

Imię i nazwisko autora Jakub Durczok
Nr albumu 107950

Wpływ cech społeczno-ekonomicznych na wysokie zarobki

Praca licencjacka
pod kierunkiem naukowym
prof. dr hab. Tomasza Szapiro
Instytut Ekonometrii,
Zakład Wspomagania i Analizy Decyzji

Warszawa, 2023 r.

Spis treści

1 Wprowadzenie

Celem pracy jest wyróżnienie czynników determinujących wysoki poziom dochodów. W obliczeniach wykorzystano dane dotyczące populacji Stanów Zjednoczonych. Identyfikowanie jednostek charakteryzujących się wysokim poziomem dochodów jest kluczowe z perspektywy wielu działań i decyzji gospodarczych.

Pośrednie ustalenie czy dochód danej osoby jest wysoki, bez danych dotyczących stanu jej finansów może być przydatne w przypadku szeregu działań ekonomicznych takich jak: marketing ukierunkowany, planowanie inwestycji czy sprzedaż dóbr i usług luksusowych. Potrzeba identyfikacji osób o wysokich zarobkach wynika więc z ich dużego znaczenia ekonomicznego.

Dokładne określenie czy dochody danej osoby są wysokie, jest istotne dla celowanych kampanii marketingowych, które odgrywają ogromną rolę we współczesnym handlu. Zgodnie z badaniem ośrodka IHS Markit identyfikowanie grupy docelowej pozwala istotnie zwiększyć skuteczność i wydajność[1]. Reklamy kierowane mają średnio 5,3 razy wyższy współczynnik klikalności niż standardowe reklamy, które nie wykorzystują danych użytkowników. W przypadku wykorzystania danych do targetowania osób, które wcześniej zwróciły uwagę na dany produkt, współczynnik klikalności jest 10,8x wyższym.

Determinacja poziomu dochodów jest szczególnie ważna na rynku dóbr luksusowych, ponieważ pozwala firmom kierować swoje działania marketingowe do osób, które z większym prawdopodobieństwem będą mogły sobie pozwolić na produkty i usługi *premium*. Rynek dóbr luksusowych odnotował w ostatnich latach znaczny wzrost, napędzany przez kombinację takich czynników jak wzrost gospodarki światowej, rosnące wydatki konsumentów oraz powiększająca się klasa średnia na rynkach wschodzących. Według badania przeprowadzonego przez Deloitte, rynek dóbr luksusowych osiągnął składaną roczną stopę wzrostu na poziomie 5,2% w okresie od 2018 do 2021, dochodząc do łącznej sprzedaży na poziomie 305 mld dolarów[2].

Identyfikacja czynników decydujących o wysokich zarobkach ma istotne znaczenie poznawcze. Zrozumienie tych czynników może pomóc w podejmowaniu świadomych decyzji

---

[1] The Economic Value of Behavioural Targeting in Digital Advertising, https://datadrivenadvertising.eu/wp-content/uploads/2017/09/BehaviouralTargeting_FINAL.pdf, data dostępu: 24 lutego 2023

[2]Global Powers of Luxury Goods 2022, https://www.deloitte.com/content/dam/assets-shared/legacy/docs/analysis/2022/gx-global-powers-of-luxury-goods-report.pdf, data dostępu: 24 lutego 2023

dotyczących edukacji i wyboru ścieżki kariery, a także prowadzić do zwiększenia zarobków i zapewnienia bezpieczeństwa finansowego.

Osoby z wyższym poziomem wykształcenia wykazują istotnie wyższe zarobki. Mediana tygodniowych zarobków osób z tytułem doktora wynosiła w 2021 roku 1909 dolarów, podczas gdy osoby z nieukończonym wykształceniem średnim zarabiały 626 dolarów[3].

Dodatkowo, statystyki wskazują również wyraźną korelację pomiędzy poziomem wykształcenia a stopą bezrobocia. Osoby z wyższym poziomem wykształcenia są w mniejszym stopniu narażone na bezrobocie. Wśród osób z tytułem doktora bezrobocie wynosiło 1,5%, podczas gdy 8,3% osób z wykształceniem poniżej średniego pozostawało bez pracy.

Do realizacji celu pracy zostaną wykorzystane metody uczenia maszynowego, będącego rodzajem sztucznej inteligencji, która umożliwia systemom komputerowym doskonalenie na podstawie danych i poprawę ich wydajności w czasie bez wyraźnego programowania. Polega ono na wykorzystaniu algorytmów i modeli statystycznych do analizy dużych ilości danych oraz identyfikacji wzorców i spostrzeżeń. Modele są szkolone na danych historycznych, aby dokonać przewidywań lub decyzji dotyczących nowych danych. Uczenie maszynowe jest wykorzystywane w wielu sytuacjach, takich jak rozpoznawanie obrazów i mowy, przetwarzanie języka naturalnego, tworzenie systemów rekomendacji i wykrywanie oszustw.

Poruszany w pracy problem decyzyjny, został poddany analizie metodą wedle której stworzone zostały reguły decyzyjne, pozwalające na określenie prawdopodobieństwa przynależności jednostki do klasy wyższej. Reguły decyzyjne stworzone zostały w postaci dwóch drzew klasyfikacyjnych, lasu losowego oraz regresji logistycznej.

Wykorzystany został zbiór danych Adult Dataset publicznie dostępny w UCI Machine Learning Repository. Zawiera on informacje o cechach demograficznych, społeczno-ekonomicznych i zatrudnieniu poszczególnych osób. Zbiór danych został stworzony w 1994 roku i od tego czasu był szeroko wykorzystywany w badaniach nad zadaniami klasyfikacji i predykcji. Zawiera on ponad 48000 obserwacji oraz 14 atrybutów, w tym wiek, poziom wykształcenia, zawód i stan cywilny. Zmienną docelową w zbiorze danych jest dochód danej osoby przekraczający 50 000 dolarów rocznie. Zbiór danych Adult jest powszechnie używany do badania relacji pomiędzy czynnikami społeczno-demograficznymi a poziomem dochodu

[3] Education Pays, https://www.bls.gov/emp/chart-unemployment-earnings-education.htm, data dostępu: 24 lutego 2023

oraz do tworzenia modeli predykcyjnych do klasyfikacji poziomu dochodu osób na podstawie informacji demograficznych ich dotyczących.

Po niniejszym Wprowadzeniu Rozdział 2 poświęcony jest analizie klasy wyższej z perspektywy jej znaczenia ekonomicznego oraz determinantów przynależności do niej. Przedstawiono m.in. różne aspekty, takie jak rola klasy wyższej jako konsumentów, marketing ukierunkowany oraz wartość poznawcza demografii klasy wyższej. Omówiono również rolę klasy wyższej we wspieraniu innowacji i tworzeniu miejsc pracy. W Rozdziale 3 przedstawiono wybrane metody badawcze, które są wykorzystywane w badaniu. Rozdział 3 skupia się na uczeniu maszynowym i wyborze odpowiedniej metody badawczej, w tym drzewa klasyfikacyjnego, lasu losowego czy regresji logistycznej. W Rozdziale 4 przedstawiono determinanty przynależności do klasy wyższej oraz sposoby ich badania. Omówiono problemy decyzyjne oraz dane wykorzystywane w badaniu a także sam proces analizy danych. Pracę zamykają Uwagi Końcowe oraz Wykazy.

2 Ekonomiczne znaczenie klasy wyższej

Rozdział 2 omawia temat klasy wyższej i składa się z trzech podrozdziałów. W podrozdziale 2.1 omówiono pojęcie stratyfikacji społecznej względu na kryteria takie jak dochody, wykształcenie i zawód. Podrozdział 2.2 charakteryzuje ekonomiczne znaczenie klasy wyższej. W podrozdziale 2.3 analizowane są determinanty przynależności do klasy wyższej. Podrozdział 2.3.1 omawia determinanty ekonomiczne, takie jak dochody i edukację. W podrozdziale 2.3.2 autor przedstawia determinanty demograficzne, takie jak płeć, wiek, stan cywilny, kraj pochodzenia czy rasę.

2.1 Stratyfikacja społeczna

W większości społeczeństw jednostki i gospodarstwa domowe są pogrupowane w różne klasy społeczne na podstawie poziomu dochodów. Klasy te można definiować na różne sposoby, ale generalnie dzieli się je na *klasę wyższą, średnią i niższą*. Dokładne progi dochodów, które definiują te klasy, mogą się różnić w zależności od kraju, a nawet w obrębie jednego kraju, ale ogólnie rzecz biorąc, odzwierciedlają one podział dochodów i bogactwa w społeczeństwie.

Klasa wyższa składa się zazwyczaj z osób i gospodarstw domowych o najwyższych dochodach i największym bogactwie. Osoby te są często właścicielami firm, kadrą kierowniczą i profesjonalistami o wysokim poziomie wykształcenia i zaawansowanych stopniach naukowych. Mają one zazwyczaj do dyspozycji duży dochód i mogą sobie pozwolić na zakup luksusowych towarów i usług oraz inwestowanie w aktywa takie jak akcje i nieruchomości.

Klasa średnia składa się zazwyczaj z osób i gospodarstw domowych o umiarkowanych dochodach i pewnym poziomie zamożności. Osoby te są często menedżerami, specjalistami i właścicielami małych firm. Zazwyczaj mają komfortowy poziom życia i mogą sobie pozwolić na zakup domów, samochodów i innych dóbr konsumpcyjnych. Mają też zwykle pewien poziom oszczędności i inwestycji, ale mogą też charakteryzować się znacznym zadłużeniem.

Klasa niższa składa się zazwyczaj z osób i gospodarstw domowych o najniższych dochodach i najmniejszym majątku. Osoby te są często pracownikami o niskich kwalifikacjach, np. zatrudnionymi w handlu detalicznym, usługach lub pracy fizycznej. Mają ograniczony dochód rozporządzalny i mogą mieć trudności z zapewnieniem sobie podstawowych potrzeb, takich jak żywność, mieszkanie i opieka zdrowotna. Mają również ograniczony dostęp do edukacji, szkoleń i możliwości zatrudnienia, co może utrudniać im awans w hierarchii społecznej.

Warto zauważyć, że progi dochodowe określające te klasy mogą być dość płynne, a osoby i gospodarstwa domowe mogą z czasem przemieszczać się między klasami. Ponadto w ramach każdej z klas dochody i majątek mogą znacznie różnić się od siebie.

Co więcej, dochód nie jest jedynym czynnikiem, który określa klasę społeczną. Inne czynniki, takie jak poziom wykształcenia, zawód i majątek, również mogą być brane pod uwagę. Na przykład osoba o wysokim poziomie wykształcenia i pozycji zawodowej może być uznana za należącą do klasy wyższej, nawet jeśli jej dochód nie jest wyższy od przeciętnego.

Podsumowując, klasy społeczne są zazwyczaj definiowane w oparciu o poziom dochodów, przy czym najbardziej powszechne są klasa wyższa, średnia i niższa. Dokładne progi dochodów, które definiują te klasy mogą się różnić, ale ogólnie odzwierciedlają one rozkład dochodów i bogactwa w społeczeństwie. Na przynależność do klasy społecznej mogą wpływać również inne czynniki, takie jak poziom wykształcenia, zawód i zamożność.

2.2 Znaczenie ekonomiczne klasy wyższej

Klasa wyższa wywiera silny wpływ na gospodarkę. Ich wydatki konsumpcyjne stanowią siłę napędową wielu gałęzi handlu, w szczególności dóbr luksusowych. Dodatkowo z racji na wysokie zarobki wnosi znaczący wkład w projekty publiczne poprzez odprowadzanie danin do budżetu państwa i wspieranie organizacji pozarządowych. Przedstawiciele klasy wyższej często inwestują również swój majątek, czym przyczyniają się do rozwoju innowacji i tworzenia miejsc pracy.

2.2.1 Rola klasy wyższej jako konsumentów

Rola klasy wyższej jako konsumentów była przedmiotem licznych studiów i projektów badawczych, mających na celu zrozumienie wpływu ich zachowań konsumpcyjnych na gospodarkę i społeczeństwo.

Badania wykazały, że klasa wyższa ma tendencję do wydawania większej części swoich dochodów na dobra i usługi luksusowe, w porównaniu do innych grup dochodowych. Według badania przeprowadzonego przez U.S. Department of Labor Statistics wśród 20% najzamożniejszych gospodarstw domowych wydatki na dobra luksusowe stanowiły w 2010 roku 65%[4].

Zgodnie z badaniami przeprowadzonymi przez Oracle[5], przedstawiciele klasy wyższej w USA są sześć razy bardziej skłonni niż przeciętni Amerykanie do wydawania pieniędzy na ekskluzywną odzież, zarówno w sklepach, jak i w intrenecie. Luksusowe domy towarowe, sklepy z odzieżą do jogi i aktywną oraz sklepy z wyposażeniem kuchennym to ich ulubione miejsca zakupów, w których są pięciokrotnie bardziej skłonni do wydawania pieniędzy.

Przedstawiciele 1 procentu 14 razy częściej przeglądają strony internetowe firm zajmujących się wyposażeniem wnętrz. Jeśli chodzi o wystrój wnętrz, pięciokrotnie częściej robią zakupy w internecie, ale tylko 2,7 razy częściej w sklepie, w porównaniu do średniej dla USA[6].

---

[4] Income Inequality and Income-Class Consumption Patterns - Federal Reserve Bank of Cleveland, https://www.clevelandfed.org/publications/economic-commentary/2014/ec-201418-income-inequality-and-income-class-consumption-patterns, data dostępu: 24 lutego 2023

[5] Lifestyles of the top 1%: How America's elite live, shop, and play, https://blogs.oracle.com/advertising/post/lifestyles-of-the-top-1-how-americax27s-elite-live-shop-and-play, data dostępu: 24 lutego 2023

[6] Income before taxes: Annual expenditure means, shares, standard errors, and coefficients of variation, Consumer Expenditure Surveys, 2021, https://www.bls.gov/cex/tables/calendar-year/mean-item-share-average-standard-error/cu-income-before-taxes-2021.pdf, data dostępu: 24 lutego 2023

Na podstawie badania wydatków konsumenckich przeprowadzonego w 2021 roku przez Bureau of Labor Statistics (BLS) możemy uzyskać wgląd w aktualną rolę klasy wyższej jako konsumentów, w tym ich wzorce wydatków i preferencje.

Nieruchomości to jeden z największych obszarów wydatków klasy wyższej. Według raportu BLS, przeciętne gospodarstwo domowe wśród 20% najlepiej zarabiających wydało ponad 29 tysięcy dolarów na utrzymanie. Jest to znacznie więcej niż średnie wydatki mieszkaniowe dla gospodarstw domowych w niższych przedziałach dochodowych. Klasa wyższa wydaje również więcej na ulepszenia domu, wyposażenie i urządzenia.

Gospodarstwa należące do klasy wyższej wydały w 2021 roku przeciętnie 10 tysięcy dolarów na zakupy i utrzymanie samochodów. To ponad dwukrotnie więcej niż średnie wydatki na samochody gospodarstw domowych należących do 20 procent najmniej zarabiających.

Raport BLS pokazuje, że przeciętne gospodarstwo domowe należące do klasy wyższej wydało w 2021 roku ponad 13 tysięcy dolarów na jedzenie. Obejmuje to zarówno konsumpcję w domu, jak i poza nim, z wyższymi wydatkami na stołowanie się poza domem i gotową żywność. Klasa wyższa ma również tendencję do kupowania więcej żywności organicznej i specjalistycznej.

Według badania Consumer Expenditure Survey gospodarstwa klasy wyższej wydają średnio 9 tysięcy dolarów rocznie na opiekę zdrowotną, czyli ponad dwukrotnie więcej niż wynosi średnia krajowa.

W przypadku edukacji wydatki to średnio 5 tysięcy dolarów rocznie, również ponad dwukrotnie więcej od średniej krajowej.

Wpływ klasy wyższej na środowisko jest kolejnym ważnym aspektem, zwłaszcza biorąc pod uwagę wysoki poziom konsumpcji i podróży. Według raportu Oxfam, najbogatsze 10% światowej populacji jest odpowiedzialne za około 50% globalnej emisji gazów cieplarnianych, a najbogatszy 1% za około 15%[7]. Raport zauważa, że ta dysproporcja jest szczególnie wyraźna w krajach o wysokich dochodach, gdzie klasa wyższa ma tendencję do emitowania wyższego śladu węglowego niż inne grupy dochodowe.

Reasumując, zwyczaje wydatkowe klasy wyższej mają znaczący wpływ na gospodarkę i społeczeństwo. Ich wysokie poziomy wydatków na dobra i usługi luksusowe mogą napędzać

[7] Confronting carbon inequality, https://oxfamilibrary.openrepository.com/bitstream/handle/10546/621052/mb-confronting-carbon-inequality-210920-en.pdf, data dostępu: 24 lutego 2023

wzrost gospodarczy i innowacje, ale mogą również przyczyniać się do negatywnych zjawisk, takich jak degradacja środowiska. Znaczenie klasy wyższej jako klientów wynika z jej zdolności do kształtowania gospodarki i społeczeństwa oraz potencjalnych konsekwencji jej nawyków wydatkowych dla całego społeczeństwa.

2.2.2 Marketing kierunkowany

Marketing kierunkowany zmienił oblicze handlu. Wzrost znaczenia marketingu cyfrowego i analityki danych umożliwił dostarczanie konsumentom spersonalizowanych i istotnych komunikatów marketingowych, a korzyści płynące z marketingu ukierunkowanego wykraczają daleko poza samo zwiększenie sprzedaży.

Według badań przeprowadzonych przez Accenture, spersonalizowany marketing jest głównym motorem lojalności i zaangażowania klientów. W raporcie stwierdzono, że 91% konsumentów jest bardziej skłonnych do robienia zakupów wśród marek, które zapewniają spersonalizowane oferty i rekomendacje[8]. Oznacza to, że firmy, które wykorzystują marketing ukierunkowany do dostarczania spersonalizowanych wiadomości, mają większe szanse na zatrzymanie klientów.

Jedną z głównych zalet marketingu ukierunkowanego jest to, że pozwala firmom na optymalizację wydatków na reklamę. Dostarczając spersonalizowane wiadomości do konkretnych grup konsumentów, firmy mogą zminimalizować zmarnowane wydatki na reklamę i zmaksymalizować zwrot z inwestycji (ROI) swoich kampanii marketingowych. Jest to szczególnie ważne w erze cyfrowej, gdzie konsumenci są bombardowani reklamami na wielu platformach i urządzeniach.

Według raportu firmy eMarketer, światowe wydatki na reklamę cyfrową osiągnęły w 2021 roku 389,29 mld dolarów, przy czym większość tych wydatków zostało poświęconych na reklamę ukierunkowaną[9]. Firmy są skłonne inwestować w marketing ukierunkowany, ponieważ jest to sprawdzona metoda docierania do konkretnych odbiorców z odpowiednimi komunikatami.

Marketing ukierunkowany prowadzi do zwiększenia zaangażowania klientów i lojalności wobec marki. Dostarczając spersonalizowane wiadomości, firmy mogą pokazać

---

[8] Making It Personal, https://www.accenture.com/_acnmedia/pdf-83/accenture-making-personal.pdf, data dostępu: 24 lutego 2023

[9] Worldwide Digital Ad Spending Year-End Update, https://www.insiderintelligence.com/content/worldwide-digital-ad-spending-year-end-update, data dostępu: 24 lutego 2023

klientom, że rozumieją ich potrzeby i preferencje, co z czasem może pomóc w budowaniu zaufania i lojalności. Raport The State of Personalization 2021 wykazał, że 60% konsumentów jest bardziej lojalnych wobec marek, które dostarczają spersonalizowanych propozycji i rekomendacji[10].

Handel elektroniczny odegrał znaczącą rolę w rozwoju marketingu ukierunkowanego w ostatnich latach. Według raportu UNCTAD, globalna sprzedaż e-commerce osiągnęła 26,7 biliona dolarów w 2019 roku, przy czym e-commerce stanowi 16% całkowitej sprzedaży detalicznej[11]. Pandemia COVID-19 przyspieszyła przesunięcie w kierunku handlu elektronicznego, ponieważ konsumenci coraz częściej decydowali się na zakupy online z powodu wprowadzonych zasad dystansowania społecznego.

W wyniku tej zmiany przedsiębiorstwa musiały dostosować swoje strategie marketingowe, aby dotrzeć do konsumentów online. Ukierunkowany marketing stał się jeszcze ważniejszy w przestrzeni e-commerce, ponieważ firmy muszą konkurować z większą pulą konkurentów, aby uchwycić uwagę konsumentów online. Spersonalizowane wiadomości i oferty mogą pomóc firmom wyróżnić się z tłumu i zwiększyć prawdopodobieństwo sprzedaży.

Rozwój handlu elektronicznego ułatwił również firmom zbieranie i analizowanie danych o swoich klientach. Dzięki temu marketing ukierunkowany stał się jeszcze bardziej skuteczny, ponieważ firmy mogą wykorzystywać dane do identyfikacji konkretnych potrzeb i preferencji swoich klientów i dostarczać im odpowiednio spersonalizowane wiadomości. Na przykład, firma może wykorzystać dane do zidentyfikowania produktów, które dany klient kupił w przeszłości i wykorzystać te informacje do zarekomendowania mu dodatkowych produktów lub usług.

Według raportu firmy eMarketer[12], najpopularniejszymi kanałami reklamy ukierunkowanej w 2021 roku są media społecznościowe, reklama w wyszukiwarkach oraz reklama display. Kanały te oferują przedsiębiorstwom szereg narzędzi i funkcji do dostarczania spersonalizowanych komunikatów do konkretnych odbiorców.

---

[10] The State of Personalization 2021, https://segment.com/state-of-personalization-report-2021/, data dostępu: 24 lutego 2023

[11] Global E-Commerce Jumps to $26.7 Trillion, Covid-19 Boosts Online Retail Sales, https://unctad.org/news/global-e-commerce-jumps-267-trillion-covid-19-boosts-online-sales, data dostępu: 24 lutego 2023

[12] eMarketer to firma prowadząca badania rynkowe dotyczące marketingu cyfrowego.

Podsumowując, znaczenie ekonomiczne marketingu ukierunkowanego jest nie do przecenienia. Poprzez dostarczanie spersonalizowanych wiadomości do konkretnych odbiorców, firmy mogą zwiększyć zaangażowanie oraz lojalność klientów, przy jednoczesnej optymalizacji wydatków na reklamę i maksymalizacji ROI. Wzrost e-commerce uczynił ukierunkowany marketing jeszcze ważniejszym, ponieważ firmy muszą konkurować w zatłoczonej przestrzeni internetowej, aby przyciągnąć uwagę konsumentów.

2.2.3 Znaczenie klasy wyższej w finansowaniu przedsięwzięć publicznych

Dane przedstawione przez Tax Foundation podkreślają ekonomiczne znaczenie klasy wyższej w odniesieniu do finansów publicznych. Według danych dotyczących federalnego podatku dochodowego, górny 1% podatników w Stanach Zjednoczonych wygenerował 20,9% całkowitego dochodu kraju w 2018 r., ale zapłacił również nieproporcjonalną ilość federalnych podatków dochodowych. Konkretnie, górny 1% zapłacił 40,1% wszystkich federalnych podatków dochodowych w 2018 roku, co jest więcej niż całe dolne 90% podatników razem wzięte[13].

Co więcej, górny 1% podatników miał również znacznie wyższą średnią efektywną stawkę podatkową niż inne grupy dochodowe. W 2018 roku górny 1% zapłacił średnią efektywną stawkę podatkową w wysokości 25,4%, co jest ponad dwukrotnie wyższe niż średnia efektywna stawka podatkowa dla wszystkich podatników (11,8%). Oznacza to, że klasa wyższa wnosi do rządu federalnego znaczną ilość dochodów, które mogą być wykorzystane do finansowania ważnych programów i usług publicznych.

Ponadto, 50% najmniej zamożnych podatników wygenerowało zaledwie 11% całkowitego dochodu i płaciło tylko 2,9% całkowitego federalnego podatku dochodowego. Górny 1% podatników zapłacił więcej federalnego podatku dochodowego niż dolne 90% razem wzięte.

1% najzamożniejszych podatników odnotował wzrost swojego udziału w całkowitym dochodzie w ciągu ostatnich kilku dekad, z 11,3% w 1980 roku do 20,8% w 2018 roku. W tym samym okresie górna krańcowa stawka podatku dochodowego spadła z 70% do 37%.

Warto zauważyć, że poza podatkami dochodowymi osoby zamożne płacą wyższe podatki pośrednie, co wynika z ich wyższego poziomu konsumpcji dóbr i usług.

---

[13] Summary of the Latest Federal Income Tax Data, https://taxfoundation.org/publications/latest-federal-income-tax-data/, data dostępu: 24 lutego 2023

Dane przedstawione w artykule podkreślają znaczącą rolę, jaką klasa wyższa odgrywa w finansowaniu programów i usług publicznych w Stanach Zjednoczonych. Bez ich wkładu, znacznie trudniej byłoby utrzymać infrastrukturę, edukację i inne ważne usługi, z których korzystają wszyscy obywatele. Dodatkowo, dochody podatkowe generowane przez klasę wyższą mogą być wykorzystane do zmniejszenia deficytu i finansowania innych ważnych inicjatyw, takich jak badania naukowe i rozwój infrastruktury.

Działalność charytatywna polega na wspieraniu organizacji i osób w potrzebie. Choć wiele osób wspiera cele charytatywne, nie warto lekceważyć roli klasy wyższej.

Według opracowania National Bureau of Economic Research, klasa wyższa – zdefiniowana jako osoby z górnego 1% rozkładu dochodów – jest odpowiedzialna za znaczną część darowizn na cele charytatywne w Stanach Zjednoczonych. W dokumencie stwierdzono, że górny 1% darczyńców odpowiadał za prawie 20% wszystkich darowizn charytatywnych w 2016 roku[14].

Znaczenie klasy wyższej we wspieraniu celów charytatywnych nie ogranicza się tylko do wielkości ich darowizn. W artykule tym stwierdzono również, że klasa wyższa ma tendencję do przekazywania darowizn na inne rodzaje organizacji niż ogół populacji. Na przykład, są oni bardziej skłonni przekazywać datki na instytucje kulturalne, takie jak muzea i organizacje zajmujące się sztuką widowiskową, oraz na instytucje szkolnictwa wyższego. Ten rodzaj darowizn może być szczególnie ważny dla trwałości tych instytucji, które często mają problemy z zapewnieniem finansowania z innych źródeł.

Co więcej, klasa wyższa ma tendencję do bardziej strategicznego traktowania swoich datków, często wykorzystując je do wpływania na wyniki społeczne i polityczne. Na przykład mogą przekazywać datki organizacjom politycznym lub ośrodkom analitycznym, które są zgodne z ich ideologią lub preferencjami politycznymi. Może to mieć istotny wpływ na kierunek polityki publicznej i alokację zasobów w społeczeństwie.

Podsumowując, klasa wyższa odgrywa istotną rolę w finansowaniu wydatków publicznych oraz organizacji charytatywnych, będąc tym samym kluczową z pespektywy licznych przedsięwzięć społecznych. To w dużej mierze jej kontrybucjom zawdzięcza się w rozwiniętych społeczeństwach wysoko rozwinięty system edukacji, służbę zdrowia czy też infrastrukturę.

---

[14] Generosity Across the Income and Wealth Distributions, https://www.nber.org/system/files/working_papers/w27076/w27076.pdf, data dostępu: 24 lutego 2023

2.2.4 Rola klasy wyższej we wspieraniu innowacji oraz tworzeniu miejsc pracy

Klasa wyższa odgrywa znaczącą rolę w tworzeniu miejsc pracy, innowacji i inwestycji. Są one niezbędne w napędzaniu wzrostu gospodarczego i rozwoju w wielu krajach. Znaczenie klasy wyższej w tych obszarach można przypisać jej dostępowi do zasobów finansowych, wiedzy i kapitału społecznego.

Klasa wyższa odgrywa kluczową rolę w tworzeniu miejsc pracy poprzez inwestowanie w małe i średnie przedsiębiorstwa (MŚP) oraz startupy. Zgodnie z artykułem "The Importance of Angel Investors in Financing the Growth of Small and Medium-Sized Enterprises" autorstwa Velanda Ramadaniego, Tzw. inwestorzy-aniołowie zapewniają niezbędne finansowanie dla startupów i MŚP. Inwestorzy-aniołowie to osoby o wysokim statusie majątkowym, które zapewniają wsparcie finansowe i biznesowe przedsiębiorcom na wczesnych etapach ich przedsięwzięć. Zazwyczaj inwestują w firmy innowacyjne i mające duży potencjał wzrostu[15].

Inwestorzy-aniołowie zapewniają nie tylko wsparcie finansowe dla startupów i MŚP. Oferują również swoje doświadczenie, wiedzę biznesową i znajomość branży, aby pomóc przedsiębiorcom odnieść sukces. Wsparcie to jest cenne dla firm na wczesnym etapie rozwoju, które mogą nie mieć zasobów lub doświadczenia w poruszaniu się po skomplikowanym krajobrazie biznesowym. Zapewniając takie wsparcie, inwestorzy-aniołowie zwiększają prawdopodobieństwo sukcesu tych startupów, co z kolei prowadzi do tworzenia miejsc pracy.

Oprócz bezpośredniego finansowania startupów i MŚP, klasa wyższa inwestuje również w fundusze venture capital, które z kolei inwestują w startupy. Artykuł Velanda Ramadaniego przytacza również Europejskie Stowarzyszenie Handlowe Aniołów Biznesu, Funduszy Zalążkowych i Uczestników Rynku Wczesnego Etapu (EBAN), według którego w 2020 r. inwestycje venture capital w Europie wyniosły 39,4 mld euro, co stanowiło wzrost o 3,8% w stosunku do roku poprzedniego. Inwestycje te wspierały tworzenie nowych firm, produktów i usług, prowadząc do tworzenia miejsc pracy i wzrostu gospodarczego.

Klasa wyższa często stanowi znaczną część zespołów ścisłego kierownictwa w wielu firmach. W związku z tym ma ona znaczący wpływ na kulturę i kierunek rozwoju tych firm.

Co więcej, przedstawiciele klasy wyższej często inwestują w badania i rozwój (R&D), które są kluczowym czynnikiem napędzającym innowacje. R&D to proces tworzenia nowej wiedzy i rozwoju nowych produktów, usług lub procesów. Inwestycje klasy wyższej w badania

---

[15] The Importance Of Angel Investors In Financing The Growth Of Small And Medium Sized Enterprises, https://hrmars.com/papers_submitted/9176/the-importance-of-angel-investors-in-financing-the-growth-of-small-and-medium-sized-enterprises.pdf, data dostępu: 24 lutego 2023

i rozwój często skutkują tworzeniem nowych lub ulepszonych produktów i usług, które napędzają wzrost gospodarczy.

Klasa wyższa jest również kluczowa w napędzaniu inwestycji w wielu krajach. Inwestycje oznaczają alokację zasobów, takich jak kapitał finansowy, w celu stworzenia nowych lub rozszerzenia istniejących przedsiębiorstw, produktów lub usług. Inwestycje są kluczowe dla wzrostu gospodarczego, ponieważ tworzą miejsca pracy, zwiększają wydajność i napędzają innowacje.

Przedstawiciele klasy wyższej mają dostęp do znacznych zasobów finansowych, które można wykorzystać do inwestowania w przedsiębiorstwa i projekty. Ich inwestycje mogą przybierać różne formy, w tym inwestycje bezpośrednie, inwestycje w fundusze venture capital oraz inwestycje na rynkach publicznych.

Według artykułu "The Importance of Angel Investors in Financing the Growth of Small and Medium-Sized Enterprises," Inwestorzy aniołowie inwestują w startupy i MŚP, które mają potencjał do wysokiego wzrostu. Inwestycje te zapewniają kluczowe finansowanie, którego te firmy potrzebują, aby rosnąć i rozwijać się. Bez tego finansowania firmy te mogą nie być w stanie osiągnąć swojego potencjału, a tworzenie miejsc pracy i wzrost gospodarczy zostałyby stłumione.

Raport OECD "Financing High-Growth Firms: The Role of Angel Investors" podkreśla ważną rolę, jaką odgrywają inwestorzy-anioły w finansowaniu szybko rozwijających się firm. Jak zauważono w raporcie, inwestorzy-aniołowie mogą zapewnić kilka korzyści dla szybko rozwijających się firm, w tym finansowanie na wczesnym etapie, wiedzę branżową, sieci i powiązania oraz kapitał. Stanowią oni zazwyczaj etap pośredni pomiędzy finansowaniem ze środków własnych a wkładem funduszy venture capital. Inwestują zazwyczaj od 20 do 500 tysięcy dolarów[16].

Inwestycje klasy wyższej na rynkach publicznych zapewniają tym firmom kapitał, którego potrzebują do rozwoju i ekspansji. Co więcej, inwestycje klasy wyższej na rynkach publicznych mogą również stymulować wzrost gospodarczy poprzez zwiększenie zaufania konsumentów i inwestycji w całej gospodarce.

Podsumowując, klasa wyższa odgrywa znaczącą rolę w tworzeniu miejsc pracy, innowacji i inwestycji. Dzięki dostępowi do zasobów finansowych, wiedzy fachowej i kapitału

[16] Financing high-growth firms: The role of angel investors, https://www.oecd.org/sti/ind/49320041.pdf, data dostępu: 24 lutego 2023

społecznego mają oni dobrą pozycję do napędzania wzrostu gospodarczego i rozwoju w wielu krajach. Poprzez swoje inwestycje w małe i średnie przedsiębiorstwa, fundusze kapitałowe i rynki publiczne, klasa wyższa zapewnia niezbędne finansowanie, którego te firmy potrzebują, aby się rozwijać i tworzyć miejsca pracy.

2.3 Determinanty przynależności do klasy wyższej

Wpływ edukacji na zarobki jest tematem budzącym duże zainteresowanie decydentów i badaczy. Edukacja jest postrzegana jako katalizator wzrostu i rozwoju ekonomicznego, a związek między edukacją a zarobkami jest dobrze ugruntowany.

Badania konsekwentnie pokazują, że osoby z wyższym poziomem wykształcenia zarabiają na ogół więcej niż osoby słabo wykształcone. Średnia premia zarobkowa dla osób z wykształceniem wyższym (tj. pomaturalnym) w porównaniu do osób z wykształceniem średnim II stopnia wynosi w krajach OECD około 56%[17]. Oznacza to, że osoby z wykształceniem wyższym zarabiają średnio 56% więcej niż osoby z wykształceniem średnim II stopnia. Premia zarobkowa jest jeszcze wyższa w niektórych krajach, takich jak Stany Zjednoczone, gdzie dla osób z wykształceniem wyższym wynosi około 73%[18].

Zjawisko to nie ogranicza się do wykształcenia wyższego a dotyczy również niższych poziomów edukacji. Według Wharton Research Scholars (2016) osoby z dyplomem ukończenia szkoły średniej zarabiają średnio o 32% więcej niż osoby bez dyplomu[19]. Premia zarobkowa w przypadku dyplomu licencjata jest jeszcze wyższa–absolwenci zarabiają o 66% więcej niż osoby posiadające jedynie świadectwo ukończenia szkoły średniej.

Związek między wykształceniem a zarobkami nie jest ograniczony do konkretnego kraju czy regionu. Jest to zjawisko globalne, które obserwuje się zarówno w krajach rozwiniętych, jak i rozwijających się. Według OECD premia zarobkowa za wykształcenie wyższe wynosi od 35% na Węgrzech do 135% w Chile. W Stanach Zjednoczonych jest wyższa dla kobiet niż dla mężczyzn. Kobiety z wykształceniem wyższym zarabiają o 81% więcej niż osoby posiadające jedynie wykształcenie średnie, natomiast mężczyźni z wykształceniem wyższym zarabiają o 68% więcej.

Chociaż istnieje silna pozytywna korelacja między wykształceniem a zarobkami, istnieje kilka czynników, które mogą wpłynąć na związek między nimi. Jednym z nich jest kierunek studiów. Premia zarobkowa w przypadku wykształcenia wyższego różni się w

---

[17] Education at a Glance 2022, https://www.oecd-ilibrary.org/docserver/3197152b-en.pdf?, data dostępu: 24 lutego 2023

[18] Effect of Education on Wage Earning, https://smartech.gatech.edu/bitstream/handle/1853/60543/talley_wang_zaski_effect_of_education_on_wage_earning.pdf, data dostępu: 24 lutego 2023

[19] The Impact of Education on Wages: Analysis of an Education Reform in Turkey, https://repository.upenn.edu/cgi/viewcontent.cgi?article=1111&context=wharton_research_scholars, data dostępu: 24 lutego 2023

zależności od kierunku studiów. Według OECD osoby z wykształceniem inżynierskim lub informatycznym zarabiają więcej niż osoby z wykształceniem humanistycznym lub społecznym. W Stanach Zjednoczonych premia zarobkowa dla absolwentów inżynierii i informatyki wynosi około 93%, podczas gdy dla absolwentów nauk humanistycznych i społecznych wynosi około 33%.

Kolejnym czynnikiem, który może wpływać na związek między wykształceniem a zarobkami, jest płeć. Podczas gdy wykształcenie ma pozytywny wpływ na zarobki zarówno w przypadku mężczyzn, jak i kobiet, premia zarobkowa dla kobiet jest często wyższa niż dla mężczyzn. Louis Federal Reserve (2015), kobiety z tytułem licencjata zarabiają średnio o 77% więcej niż osoby posiadające jedynie świadectwo ukończenia szkoły średniej, natomiast mężczyźni z tytułem licencjata zarabiają o 63% więcej. Może to wynikać z faktu, że kobiety mają tendencję do zarabiania mniej niż mężczyźni na niższych poziomach edukacji, więc względny wzrost zarobków wraz z dodatkowym wykształceniem jest większy.

Kolejnym istotnym czynnikiem jest rasa. Według Rezerwy Federalnej w St. Louis premia zarobkowa za wykształcenie jest wyższa dla białych niż dla nie-białych w Stanach Zjednoczonych. Białe osoby z tytułem licencjata zarabiają średnio o 69% więcej niż osoby posiadające jedynie dyplom ukończenia szkoły średniej, natomiast osoby nie-białe z tytułem licencjata zarabiają o 49% więcej[20]. Przyczyny różnic rasowych w premii zarobkowej za wykształcenie są złożone i wieloaspektowe, ale mogą obejmować takie czynniki jak dyskryminacja, nierówności w dostępie do edukacji oraz różnice w rodzajach zawodów i branż, w których zatrudniane są osoby różnych ras.

Pochodzenie społeczne również nie jest bez znaczenia. Premia zarobkowa za tytuł licencjata jest wyższa dla osób, które pochodzą z rodzin o niskich dochodach, niż dla osób, które pochodzą z rodzin o wysokich dochodach. Sugeruje to, że edukacja może być szczególnie ważna dla osób ze środowisk o niskim statusie społecznym, które mogą napotkać na większe wyzwania ekonomiczne.

Warunki panujące na rynku pracy w danym regionie mogą również wpływać na relację pomiędzy wykształceniem a zarobkami. W regionach, w których istnieje duże zapotrzebowanie na wykwalifikowanych pracowników, osoby z wyższym poziomem wykształcenia

[20] The Demographics of Wealth, https://www.stlouisfed.org/-/media/project/frbstl/stlouisfed/Files/PDFs/HFS/essays/HFS-Essay-2-2015-Education-and-Wealth.pdf, data dostępu: 24 lutego 2023

prawdopodobnie będą zarabiać więcej niż osoby z niższym poziomem wykształcenia. I odwrotnie, w regionach, gdzie występuje nadwyżka wykwalifikowanych pracowników, premia zarobkowa za wykształcenie może być niższa.

Wykształcenie ma więc istotny wpływ na zarobki, a związek między nimi jest dobrze ugruntowany. Badania konsekwentnie pokazują, że osoby z wyższym poziomem wykształcenia zwykle zarabiają więcej niż osoby z niższym poziomem wykształcenia. Premia zarobkowa wynikająca z wykształcenia dotyczy zarówno mężczyzn, jak i kobiet i jest obserwowana zarówno w krajach rozwiniętych, jak i rozwijających się. Chociaż istnieje kilka czynników, które mogą wpływać na związek między wykształceniem a zarobkami, takich jak kierunek studiów, płeć, rasa i pochodzenie społeczno-ekonomiczne, edukacja pozostaje ważnym narzędziem wzrostu gospodarczego i rozwoju. Decydenci polityczni mogą wykorzystać te informacje do opracowania polityki promującej dostęp do edukacji i zachęcającej jednostki do osiągania wyższych poziomów wykształcenia.

Wybór odpowiedniej branży i ścieżki kariery jest kluczowy z perspektywy sukcesu zawodowego. Ważne jest, aby przy podejmowaniu tej decyzji wziąć pod uwagę różne czynniki, takie jak umiejętności, zainteresowania, perspektywy zawodowe i potencjalne dochody.

Jednym z najważniejszych czynników do rozważenia przy wyborze ścieżki kariery jest perspektywa zatrudnienia. Według Glassdoor Job Market Report dla Stanów Zjednoczonych, niektóre z najszybciej rozwijających się zawodów w 2021 roku to inżynier oprogramowania, analityk danych i specjalista opieki zdrowotnej. W tych dziedzinach występuje wysoki popyt na pracowników[21].

BLS Occupational Employment and Wage Statistics dostarcza bardziej szczegółowych informacji na temat perspektyw zawodowych. Na przykład BLS informuje, że zatrudnienie programistów ma wzrosnąć o 21% od 2019 do 2029 roku, znacznie szybciej niż średnia dla wszystkich zawodów. Podobnie, zatrudnienie pielęgniarzy i pielęgniarek ma wzrosnąć o 45% w tym samym okresie, ponownie znacznie szybciej niż średnia dla wszystkich zawodów. Statystyki te sugerują, że wybór kariery w tych branżach może gwarantować stabilne i obiecujące perspektywy zawodowe[22].

---

[21] United States Job Market Report, https://www.glassdoor.com/research/job-market-report-united-states/, data dostępu: 24 lutego 2023

[22] Occupational Employment and Wage Statistics, https://www.bls.gov/oes/current/oes_nat.htm#00-0000, data dostępu: 24 lutego 2023

Kolejną ważną kwestią przy wyborze ścieżki kariery są potencjalne dochody. Raport Tap Talent Salary Report 2022 dostarcza cennych informacji na temat średnich wynagrodzeń i premii dla różnych zawodów technologicznych w Europie. Na przykład raport pokazuje, że średnia pensja starszego inżyniera oprogramowania wynosi 75 000 euro, podczas gdy średnia pensja naukowca ds. danych wynosi 70 000 euro. Te liczby podkreślają potencjalny dochód, jakiego mogą oczekiwać osoby pracujące w tych branżach.

Podobnie, raport JobStreet Salary Report 2022 dostarcza informacji na temat średnich wynagrodzeń dla różnych zawodów w Malezji. Na przykład raport pokazuje, że średnie miesięczne wynagrodzenie dla programisty wynosi 5 141 ringgitów malezyjskich, podczas gdy średnie miesięczne wynagrodzenie dla kierownika ds. marketingu wynosi 3 811 ringgitów[23]. Te statystyki pokazują, że potencjalne dochody mogą się znacznie różnić w zależności od branży i stanowiska pracy.

Przywołana literatura wskazuje, że dochód nie powinien być jedynym czynnikiem branym pod uwagę przy wyborze ścieżki kariery. Warto również wziąć pod uwagę inne czynniki, takie jak osobiste zainteresowania, umiejętności i równowaga między życiem zawodowym a prywatnym. Na przykład, raport Glassdoor dotyczący rynku pracy w Stanach Zjednoczonych pokazuje, że niektóre z najwyżej ocenianych tytułów zawodowych pod względem ogólnego zadowolenia z pracy i równowagi między życiem zawodowym a prywatnym w 2021 roku obejmowały higienistkę stomatologiczną, laryngologa i terapeutę zajęciowego. Te tytuły zawodowe mogą nie mieć najwyższych potencjalnych dochodów, ale oferują inne korzyści, takie jak satysfakcja z pracy i zdrowa równowaga między życiem zawodowym a prywatnym.

Innym czynnikiem, który warto wziąć pod uwagę przy wyborze ścieżki kariery, jest poziom konkurencji w branży. Raport Tap.Talent Salary Report 2022 dostarcza danych na temat najbardziej pożądanych umiejętności dla różnych zawodów technologicznych w Europie. Na przykład, raport pokazuje, że najbardziej pożądaną umiejętnością dla naukowca zajmującego się danymi jest uczenie maszynowe, podczas gdy najbardziej pożądaną umiejętnością dla inżyniera oprogramowania jest Java. Osoby, które posiadają te pożądane umiejętności, prawdopodobnie będą miały przewagę konkurencyjną w poszukiwaniu pracy[24].

---

[23] Salary Report 2022, https://www.jobstreet.com.my/en/cms/employer/wp-content/themes/jobstreet-employer/assets/pdf/MY-SalaryReport-R3.5-25thJan2022-final.pdf, data dostępu: 24 lutego 2023

[24] Salary Report 2022, https://taptalent.eu/wp-content/uploads/2022/04/tap.talent-salary-report-2022-1.pdf, data dostępu: 24 lutego 2023

Podobnie, raport JobStreet Salary Report 2022 dostarcza informacji na temat czynników, które mogą wpływać na poziom wynagrodzeń w Malezji. Na przykład, raport pokazuje, że główne czynniki, które mogą zwiększyć wynagrodzenie danej osoby to lata doświadczenia, poziom wykształcenia i branża, w której pracują. Osoby, które mają wyższy poziom wykształcenia i pracują w branżach o dużym zapotrzebowaniu na wykwalifikowanych specjalistów, prawdopodobnie będą miały wyższe wynagrodzenie.

Innym ważnym czynnikiem przy wyborze ścieżki kariery jest poziom wykształcenia wymagany dla różnych stanowisk. BLS Occupational Employment and Wage Statistics dostarcza informacji na temat typowych poziomów wykształcenia wymaganych dla różnych zawodów. Na przykład, BLS podaje, że większość stanowisk programisty wymaga tytułu licencjata w dziedzinie informatyki lub pokrewnej, podczas gdy większość stanowisk pielęgniarki wymaga tytułu magistra pielęgniarstwa. Statystyki te podkreślają znaczenie rozważenia poziomu wymaganego wykształcenia przy wyborze ścieżki kariery.

Dodatkowo, ważne jest, aby rozważyć potencjał wzrostu i awansu w wybranej ścieżce kariery. Raport Glassdoor dotyczący rynku pracy w Stanach Zjednoczonych podkreśla niektóre z tytułów zawodowych o wysokim potencjale wzrostu i awansu w 2021 roku. Na przykład raport wymienia kierownika produktu, kierownika ds. inżynierii oprogramowania i architekta rozwiązań jako stanowiska o wysokim potencjale rozwoju kariery. Wybór kariery w branży, która oferuje możliwości awansu, może zapewnić osobom poczucie celu i motywację do dalszego rozwijania swoich umiejętności i wiedzy.

Jednym z najważniejszych demograficznych wyznaczników przynależności do klasy wyższej jest płeć. Badania konsekwentnie pokazują, że mężczyźni częściej niż kobiety należą do tej grupy społecznej. Według raportu Banku Rezerwy Federalnej w San Francisco (FRBSF), mężczyźni częściej niż kobiety znajdują się w górnym 1% zarabiających[25]. Z raportu wynika, że w 2015 roku mężczyźni stanowili 83% osób z górnego 1% zarabiających, podczas gdy kobiety stanowiły jedynie 17%. Raport zauważa ponadto, że udział kobiet w górnym 1% pozostawał stosunkowo stabilny w ciągu ostatniej dekady, oscylując między 15% a 17%.

Różnica w zarobkach mężczyzn i kobiet jest istotnym czynnikiem wyjaśniającym dysproporcje w przynależności do klasy wyższej. Według badania przeprowadzonego przez Economic Policy Institute (EPI), różnica w zarobkach kobiet i mężczyzn utrzymywała się przez

[25] Disappointing Facts about the Black-White Wage Gap, https://www.frbsf.org/wp-content/uploads/sites/4/el2017-26.pdf, data dostępu: 24 lutego 2023

lata. W 2019 roku kobiety zarobiły 82 centy na każdego dolara zarobionego przez mężczyzn. Badanie zauważa, że ta luka jest jeszcze bardziej znacząca dla kobiet czarnoskórych. Na przykład Afroamerykanki zarabiały tylko 62 centy na każdego dolara zarobionego przez białych mężczyzn[26].

Wiek jest kolejnym demograficznym wyznacznikiem przynależności do klasy wyższej. Według raportu FRBSF, wiek jest pozytywnie skorelowany z prawdopodobieństwem znalezienia się w górnym 1% zarabiających. W raporcie stwierdzono, że mediana wieku osób z górnego 1% zarabiających wynosi 54 lata, czyli jest znacznie wyższa niż mediana wieku w całej populacji. Raport zauważa ponadto, że prawdopodobieństwo znalezienia się w górnym 1% zarabiających wzrasta wraz z wiekiem. W 2015 roku osoby w wieku 65 lat i więcej stanowiły 9,1% najlepiej zarabiających, w porównaniu z zaledwie 3,1% populacji ogólnej.

Związek między wiekiem a przynależnością do klasy wyższej można wyjaśnić kilkoma czynnikami. Po pierwsze, osoby starsze miały więcej czasu na zgromadzenie bogactwa i aktywów. Po drugie, osoby starsze częściej zajmują wysokie stanowiska w organizacjach, które zwykle oferują wyższe wynagrodzenia i lepsze świadczenia. Wreszcie, osoby starsze mogą mieć dostęp do koneksji, które umożliwiają im dostęp do najlepiej płatnych miejsc pracy i możliwości inwestycyjnych.

Stan cywilny jest istotnym demograficznym wyznacznikiem przynależności do klasy wyższej. Badanie przeprowadzone przez Federal Reserve Bank of St. Louis wykazało, że żonaci mężczyźni znajdują się na szczycie drabiny płac, zarabiając więcej niż jakakolwiek inna grupa demograficzna. W 2016 roku mediana dochodów żonatych mężczyzn wyniosła 78 000 dolarów, w porównaniu z 44 000 dolarów dla samotnych mężczyzn i 42 000 dolarów dla samotnych kobiet. Badanie wykazało również, że żonaci mężczyźni częściej znajdują się w górnych 10% zarabiających niż jakakolwiek inna grupa demograficzna[27].

Kraj pochodzenia jest kolejną determinantą demograficzną, która wpływa na prawdopodobieństwo przynależności jednostki do klasy wyższej. Według raportu Banku Rezerwy Federalnej w San Francisco, imigranci w Stanach Zjednoczonych mają mniejsze szanse na przynależność do klasy wyższej niż osoby urodzone w danym kraju. Raport pokazuje,

---

[26] Black-white wage gaps expand with rising wage inequality, https://files.epi.org/pdf/101972.pdf, data dostępu: 24 lutego 2023

[27] Married Men Sit Atop the Wage Ladder, https://files.stlouisfed.org/files/htdocs/publications/economic-synopses/2018/09/14/married-men-sit-atop-the-wage-ladder.pdf, data dostępu: 24 lutego 2023

że mediana dochodów gospodarstw domowych imigrantów jest o 20% niższa niż gospodarstw domowych urodzonych w danym kraju, a wskaźnik ubóstwa wśród gospodarstw domowych imigrantów jest o 50% wyższy niż wśród gospodarstw domowych urodzonych w danym kraju. Raport sugeruje, że powodem tej różnicy jest fakt, że imigranci stają przed kilkoma wyzwaniami, takimi jak bariery językowe, dyskryminacja, brak dostępu do edukacji i środków finansowych. Jednakże zauważa również, że istnieją znaczne różnice pomiędzy grupami imigrantów, przy czym niektóre grupy mają wyższe dochody i niższe wskaźniki ubóstwa niż inne.

Rasa i pochodzenie etniczne również odgrywają znaczącą rolę w określaniu statusu społeczno-ekonomicznego. Historycznie rzecz biorąc, mniejszości rasowe i etniczne spotykały się z dyskryminacją w zakresie edukacji, zatrudnienia i dostępu do zasobów, co przyczyniło się do ich niższej reprezentacji w klasie wyższej. Poniższe podrozdziały opisują ten temat bardziej szczegółowo.

Czarnoskórzy Amerykanie historycznie napotykali na bariery w awansie z powodu instytucjonalnego rasizmu, dyskryminacji i uprzedzeń. W rezultacie są oni niedostatecznie reprezentowani w klasie wyższej. Według raportu EPI, czarnoskórzy Amerykanie stanowili tylko 1,6% najwyższego 1% osób uzyskujących dochody w 2015 roku, podczas gdy stanowili 12,7% całej populacji. Raport zauważa również, że mediana wartości netto czarnych gospodarstw domowych w 2013 roku wynosiła zaledwie 11 000 dolarów, w porównaniu do 141 900 dolarów dla białych gospodarstw domowych.

Według raportu „Changing America" czarnoskórzy pracownicy zarabiają tylko 74 centów na każdego dolara zarobionego przez białych pracowników[28]. Ponadto Czarnoskórzy pracownicy rzadziej zajmują wysokopłatne miejsca pracy i częściej pracują w niskopłatnych sektorach, takich jak handel detaliczny i hotelarstwo.

Latynosi również napotykają na przeszkody w awansie, choć ich sytuacja jest nieco lepsza niż czarnych Amerykanów. Według raportu EPI, Latynosi stanowili 7,4% najwyższego 1% osób uzyskujących dochody w 2015 roku, podczas gdy stanowili 17,6% całej populacji. Jednak ich mediana wartości netto w 2013 roku wynosiła zaledwie 13 700 dolarów, w porównaniu do 141 900 dolarów dla białych gospodarstw domowych.

[28] Changing America, https://web.archive.org/web/20120131063624/http://www.gpoaccess.gov/eop/ca/pdfs/ca.pdf, data dostępu: 24 lutego 2023

Badanie St. Louis Federal Reserve wykazało, że latynoscy pracownicy zarabiają 80 centów na każdego dolara zarobionego przez białych pracowników. Jest to dysproporcja nieco mniejsza niż w przypadku czarnych pracowników. Jednak latynoscy pracownicy częściej pracują w nisko płatnych sektorach, takich jak budownictwo i rolnictwo.

Azjatyccy Amerykanie są grupą rasową lub etniczną, która jest najszerzej reprezentowana w klasie wyższej. Według raportu EPI, stanowili oni 10,7% najwyższego 1% osób uzyskujących dochody w 2015 roku, podczas gdy reprezentowali jedynie 5,6% całej populacji. W raporcie zauważono, że Azjatyccy Amerykanie mają również najwyższą medianę dochodów gospodarstw domowych spośród wszystkich grup rasowych i etnicznych w USA.

Warto jednak zauważyć, że w obrębie społeczności azjatyckich Amerykanów istnieje znaczne zróżnicowanie. Na przykład, Hindusi i Chińczycy są częściej reprezentowani w klasie wyższej niż Wietnamczycy i Kambodżanie. Co więcej, społeczność azjatycko-amerykańska boryka się z własnymi unikalnymi wyzwaniami, takimi jak bariery językowe i dyskryminacja.

Przynależność do klasy wyższej jest determinowana przez złożony zestaw czynników demograficznych. Płeć, wiek, kraj pochodzenia, stan cywilny oraz rasa i pochodzenie etniczne odgrywają znaczącą rolę w określaniu statusu społeczno-ekonomicznego.

3 Wybrane metody badania

Rozdział 3 stanowi wprowadzenie do metod uczenia maszynowego, które zostanie wykorzystane w badaniu. Pokrótce opisane zostały w nim podejścia do uczenia maszynowego orasz jego historia, zastosowania i ograniczenia. Następnie omówione zostały metody uczenia maszynowego z nadzorem, które posłużyły do stworzenia modeli: drzewa decyzyjne, modele regresji oraz lasy losowe. Ostatecznie przedstawione zostały wybrane metryki oceny jakości tych modeli.

3.1 Uczenie maszynowe

Uczenie maszynowe to dziedzina sztucznej inteligencji, która w ostatnich latach zyskała znaczną uwagę i popularność. Zdolność maszyn do uczenia się i poprawy wydajności bez wyraźnego programowania otworzyły nowe możliwości w szerokim zakresie zastosowań, od rozpoznawania obrazów i przetwarzania języka naturalnego do tworzenia autonomicznych pojazdów.

Aby zrozumieć podstawy uczenia maszynowego, ważne jest, aby zacząć od podstaw sztucznej inteligencji. Sztuczna inteligencja (AI) odnosi się do zdolności maszyn do wykonywania zadań, które zazwyczaj wymagają ludzkiej inteligencji, takich jak rozpoznawanie mowy lub gra w szachy. Uczenie maszynowe to podzbiór AI, który obejmuje wykorzystanie technik statystycznych w celu umożliwienia maszynom uczenia się na podstawie danych i poprawy ich wydajności w czasie.

Istnieją trzy główne rodzaje uczenia maszynowego: uczenie nadzorowane, uczenie bez nadzoru i uczenie wzmacniające. Uczenie nadzorowane polega na trenowaniu modelu uczenia maszynowego na zbiorze oznaczonych danych, gdzie dla każdych danych wejściowych znane są dane wejściowe. Celem modelu jest nauczenie się mapowania od danych wejściowych do danych wyjściowych, tak aby mógł on przewidywać na nowych, niewidzianych danych. Typowe zastosowania uczenia nadzorowanego obejmują klasyfikację obrazów, rozpoznawanie mowy i modelowanie predykcyjne.

Z drugiej strony, uczenie bez nadzoru polega na trenowaniu modelu na nieoznakowanym zbiorze danych, gdzie celem jest odkrycie wzorców i struktury w danych. Może to obejmować techniki takie jak klasteryzacja, która grupuje podobne punkty danych razem, lub redukcja wymiarowości, która zmniejsza liczbę zmiennych w zbiorze danych, zachowując najważniejsze informacje.

Uczenie wzmacniające jest rodzajem uczenia maszynowego, które obejmuje doskonalenie modelu w celu podejmowania decyzji na podstawie nagród i kar. Model uczy się poprzez interakcję ze środowiskiem i otrzymuje informacje zwrotne w postaci nagród lub kar

na podstawie swoich działań. Celem jest nauczenie się zachowań maksymalizujących zyski w czasie, co może być wykorzystane do rozwiązania złożonych problemów decyzyjnych.

Istnieje kilka kluczowych wyzwań i rozważań, które warto wziąć pod uwagę podczas stosowania technik uczenia maszynowego. Jednym z najważniejszych jest kwestia stronniczości i sprawiedliwości, która może pojawić się, gdy dane treningowe nie są reprezentatywne dla populacji lub gdy model nie został zaprojektowany tak, aby uwzględniać pewne czynniki, takie jak rasa czy płeć. Innym wyzwaniem jest kwestia interpretowalności, gdzie złożony charakter niektórych modeli uczenia maszynowego może utrudniać zrozumienie, w jaki sposób dokonują one swoich przewidywań.

Aby sprostać tym wyzwaniom, badacze i praktycy w dziedzinie uczenia maszynowego opracowali szeroki zakres technik i narzędzi. Należą do nich metody wstępnego przetwarzania i czyszczenia danych, wyboru i oceny modeli oraz interpretowalności i uczciwości. Istnieje również wiele popularnych języków programowania i ram uczenia maszynowego, takich jak Python czy R, które ułatwiają tworzenie i wdrażanie modeli uczenia maszynowego.

3.1.1 Podejścia do uczenia maszynowego

W literaturze przedstawiono trzy podejścia do uczenia maszynowego istotne dla niniejszej pracy.

Uczenie nadzorowane, uczenie bez nadzoru i uczenie wzmacniające to trzy główne kategorie uczenia maszynowego. Różnią się one rodzajem danych używanych do szkolenia, celami procesu uczenia oraz metodami wykorzystywanymi do osiągnięcia tych celów. W Podrozdziale 3.1.2 zostaną one krótko opisane na podstawie publikacji „Introduction to Machine Learning" autorstwa Ethem Alpaydin[29].

Uczenie nadzorowane jest najbardziej powszechnym typem uczenia maszynowego i jest wykorzystywane w wielu sytuacjach. Polega ono na trenowaniu modelu na zbiorze oznaczonych danych, co oznacza, że każdy punkt danych w zbiorze jest powiązany z etykietą lub zmienną docelową, którą model próbuje przewidzieć. Celem uczenia nadzorowanego jest przedstawienie zależności pomiędzy cechami wejściowymi a zmienną docelową, tak aby model mógł przewidzieć zmienną docelową dla nowych danych wejściowych.

Istnieją dwa główne typy uczenia nadzorowanego: regresja i klasyfikacja. W regresji zmienna docelowa jest ciągła, a celem jest przewidzenie jej wartości. Na przykład, w procesie przewidywania cen mieszkań, zmienną docelową jest cena domu, a celem jest przewidzenie tej

[29] „Introduction to Machine Learning", Ethem Alpaydin, Massachusetts Institute of Technology, 2014

ceny na podstawie takich cech jak liczba sypialni, powierzchnia i lokalizacja. W klasyfikacji, zmienna docelowa jest kategoryczna, a celem jest przewidzenie klasy danych wejściowych. Na przykład, w procesie wykrywania spamu, zmienną docelową jest to, czy wiadomość e-mail jest spamem czy nie, a celem jest przewidzenie klasy na podstawie cech takich jak adres e-mail nadawcy, temat i treść wiadomości.

Aby wytrenować model uczenia nadzorowanego, potrzebujemy etykietowanego zbioru danych, który składa się z cech wejściowych i zmiennych docelowych. Następnie możemy użyć różnych algorytmów, aby nauczyć się mapowania pomiędzy cechami wejściowymi a zmienną docelową. Przykłady algorytmów uczenia nadzorowanego obejmują regresję liniową, regresję logistyczną, drzewa decyzyjne, maszyny wektorów wspierających czy sieci neuronowe.

W uczeniu bez nadzoru nie mamy żadnych oznaczonych danych, a celem jest odkrycie wzorców i struktury, które mogą być wykorzystane do grupowania podobnych punktów danych. Uczenie bez nadzoru znajduje zastosowanie w zadaniach takich jak segmentacja klientów, grupowanie obrazów i tekstów oraz wykrywanie anomalii.

Istnieją dwa główne rodzaje uczenia bez nadzoru: grupowanie (klasteryzacja) i redukcja wymiaru. W klasteryzacji celem jest pogrupowanie podobnych punktów danych na podstawie ich cech. Na przykład, w przypadku segmentacji klientów, celem jest pogrupowanie klientów w klastry na podstawie ich zachowań zakupowych. W redukcji wymiarowości celem jest zmniejszenie liczby cech w danych przy zachowaniu najważniejszych informacji. Jest to przydatne, gdy dane mają wiele cech, a niektóre z nich są zbędne lub nieistotne.

Aby wytrenować model uczenia bez nadzoru, potrzebujemy nieoznakowanego zbioru danych, który składa się tylko z cech wejściowych. Następnie możemy użyć różnych algorytmów do odkrycia wzorców i struktury w danych. Przykłady algorytmów uczenia bez nadzoru obejmują algorytm centroidów, grupowanie hierarchiczne, analizę głównych składowych i metodę redukcji wymiarów t-SNE.

Uczenie wzmacniające jest rodzajem uczenia maszynowego, które jest wykorzystywane do uczenia, jak podejmować decyzje w oparciu na informacjach zwrotnych ze środowiska. Agent wchodzi w interakcję ze środowiskiem poprzez podejmowanie działań i otrzymuje nagrody lub kary w zależności od wyniku tych działań. Celem uczenia wzmacniającego jest wykształcenie zachowań maksymalizujących skumulowany zysk w czasie.

Algorytmy uczenia wzmacniającego są wykorzystywane m.in. w robotyce, grach czy systemach sterowania.

Podsumowując, uczenie nadzorowane, uczenie bez nadzoru i uczenie wzmacniające to trzy podstawowe podejścia w uczeniu maszynowym. Każde z nich ma swoje mocne i słabe

strony, a wybór, które z nich warto zastosować, zależy od danego problemu. Uczenie nadzorowane jest stosowane, gdy dostępne są dane z etykietami, a celem jest przewidzenie zmiennej docelowej. Uczenie bez nadzoru jest stosowane, gdy nie ma dostępnych danych z etykietami, a celem jest znalezienie wzorców lub struktur w danych. Uczenie wzmacniające jest stosowane, gdy optymalne działanie nie jest znane z góry i może być wypracowane metodą prób i błędów. Poprzez zrozumienie tych trzech podejść, praktycy uczenia maszynowego mogą wybrać odpowiedni algorytm dla swojego konkretnego problemu i osiągnąć optymalne wyniki.

3.1.2 Historia uczenia maszynowego

Początki uczenia maszynowego sięgają lat 40. i 50. ubiegłego wieku, kiedy to naukowcy zaczęli badać koncepcję sztucznej inteligencji. Jednym z najwcześniejszych przykładów uczenia maszynowego było opracowanie perceptronu, rodzaju sztucznej sieci neuronowej, która może nauczyć się rozpoznawać wzorce w danych. Perceptron został wynaleziony przez Franka Rosenblatta w 1958 roku i był jednym z pierwszych algorytmów zdolnych do uczenia się na podstawie danych.

W latach 60. i 70. XX wieku naukowcy zaczęli opracowywać bardziej zaawansowane algorytmy uczenia maszynowego, w tym drzewa decyzyjne i sieci bayesowskie. Algorytmy te zostały zaprojektowane tak, aby umożliwić komputerom podejmowanie decyzji w oparciu o rozumowanie probabilistyczne i były wykorzystywane w wielu zastosowaniach, w tym w rozpoznawaniu mowy i przetwarzaniu języka naturalnego.

W 1967 roku wymyślono algorytm najbliższego sąsiada, który był początkiem podstawowego rozpoznawania wzorców. Algorytm ten został wykorzystany do wyznaczania tras i był jednym z najwcześniejszych algorytmów zastosowanych do znalezienia rozwiązania problemu komiwojażera, polegającego na znalezieniu najbardziej efektywnej trasy. Używając go, sprzedawca wpisuje wybrane miasto i wielokrotnie zleca programowi odwiedzenie najbliższych miast, aż wszystkie zostaną odwiedzone.

Pod koniec lat 70. i na początku lat 80. badania nad sztuczną inteligencją koncentrowały się głównie wokół logicznych, opartych na wiedzy podejść, a nie algorytmów, co spowodowało rozdźwięk między sztuczną inteligencją a uczeniem maszynowym. Branża została zreorganizowana jako odrębna dziedzina, a jej główny cel przeszedł od szkolenia w zakresie sztucznej inteligencji do rozwiązywania praktycznych problemów i świadczenia usług. W rezultacie, uwaga branży odeszła od podejść odziedziczonych po badaniach nad AI i skierowała się w stronę metod i taktyk stosowanych w teorii prawdopodobieństwa i statystyce.

W latach 80. XX wieku propagacja wsteczna stała się szeroko stosowaną techniką szkolenia sieci neuronowych. Propagacja wsteczna jest algorytmem uczenia nadzorowanego używanym do szkolenia sztucznych sieci neuronowych, w którym dane wejściowe są przekazywane dalej przez sieć, a wyjście jest porównywane z oczekiwanym wyjściem. Różnica między wyjściem a oczekiwanym wyjściem jest następnie wykorzystywana do dostosowania wag połączeń między neuronami w sieci.

Propagacja wsteczna pozwoliła na bardziej efektywne i dokładne wytrenowanie sieci neuronowych i odegrała kluczową rolę w rozwoju głębokiego uczenia. Głębokie sieci neuronowe mają wiele warstw połączonych ze sobą neuronów, a ich doskonalenie wymaga zastosowania wstecznej propagacji w celu dostosowania wag w każdej warstwie.

Obecnie, wsteczna propagacja jest podstawową techniką używaną w wielu algorytmach uczenia maszynowego i została wykorzystana w wielu zastosowaniach, w tym w rozpoznawaniu obrazów i mowy, przetwarzaniu języka naturalnego i robotyce. Rozwój backpropagacji w latach 80. był ważnym kamieniem milowym w historii uczenia maszynowego i utorował drogę dla wielu postępów w sztucznej inteligencji i głębokim uczeniu, które widzimy dzisiaj.

Algorytmy boostingowe zostały opracowane w latach 90. jako niezbędny rozwój dla ewolucji uczenia maszynowego. Są one używane do zmniejszenia stronniczości podczas uczenia nadzorowanego i obejmują algorytmy, które przekształcają słabych uczących się w silnych. Większość algorytmów boostingowych składa się z powtarzających się uczących się słabych klasyfikatorów, które następnie dodają się do końcowego silnego klasyfikatora.

W 1997 r. superkomputer Deep Blue firmy IBM pokonał mistrza świata w szachach Garry'ego Kasparowa. Deep Blue był wyspecjalizowanym systemem komputerowym przeznaczonym do gry w szachy i był w stanie ocenić do 200 milionów pozycji szachowych na sekundę.

Zwycięstwo Deep Blue nad Kasparowem było postrzegane jako ważny kamień milowy w rozwoju sztucznej inteligencji i uczenia maszynowego. Mecz pokazał, że komputery mogą doskonalić się w złożonych zadaniach, które wcześniej uważano za wyłączną domenę ludzkiej inteligencji. Wywołał on również debatę na temat przyszłości ludzkiej pracy i roli technologii w społeczeństwie.

W świecie biznesu uczenie maszynowe jest obecnie wykorzystywane na kilka sposobów, w tym do analizy danych sprzedażowych, personalizacji mobilnej w czasie rzeczywistym, wykrywania oszustw, rekomendacji produktów, systemów zarządzania uczeniem się, dynamicznego ustalania cen i przetwarzania języka naturalnego. Modele uczenia

maszynowego stały się wysoce adaptacyjne, stale się uczą i stają się coraz dokładniejsze. W połączeniu z nowymi technologiami obliczeniowymi i analityką biznesową, uczenie maszynowe może rozwiązywać różne problemy organizacyjne. Nowoczesne modele uczenia maszynowego mogą być wykorzystywane do tworzenia prognoz, począwszy od epidemii chorób, a skończywszy na wzroście i spadku wartości akcji.

3.1.3 Zastosowania

Uczenie maszynowe zyskuje coraz większą popularność. Technologia ta jest wykorzystywana w różnych dziedzinach, takich jak finanse, opieka zdrowotna, marketing i wiele innych, aby zautomatyzować zadania i podejmować lepsze decyzje.

Jednym z najbardziej obiecujących obszarów, w których stosuje się uczenie maszynowe, jest opieka zdrowotna. Zdolność modeli uczenia maszynowego do analizowania ogromnych ilości danych medycznych i dostarczania dokładnych diagnoz i zaleceń dotyczących leczenia rewolucjonizuje przemysł opieki zdrowotnej. Na przykład, algorytmy uczenia maszynowego są wykorzystywane do analizy obrazów medycznych, takich jak zdjęcia rentgenowskie, tomografia komputerowa i rezonans magnetyczny, w celu identyfikacji i diagnozowania chorób takich jak rak. Algorytmy uczenia maszynowego mogą wykrywać wzorce i struktury w danych, które mogą być wykorzystane do rozróżnienia zdrowej i chorej tkanki z dużą dokładnością.

Innym zastosowaniem uczenia maszynowego w opiece zdrowotnej jest medycyna spersonalizowana. Modele uczenia maszynowego mogą analizować dane genetyczne i medyczne pacjentów, aby przewidzieć skuteczność różnych terapii i pomóc lekarzom podejmować bardziej świadome decyzje. Może to prowadzić do skuteczniejszych terapii i mniejszej liczby skutków ubocznych.

Kolejną branżą, która jest przekształcana przez uczenie maszynowe, są finanse. Modele uczenia maszynowego są wykorzystywane do automatyzacji zadań takich jak wykrywanie oszustw, scoring kredytowy i handel akcjami. Na przykład, algorytmy uczenia maszynowego mogą analizować transakcje kart kredytowych i wykrywać oszustwa w czasie rzeczywistym. Algorytmy uczenia maszynowego przetwarzają duże ilości danych i wykrywają wzorce wskazujące na oszustwa, które mogą zostać przeoczone przez analityków.

Uczenie maszynowe jest również wykorzystywane w handlu akcjami do przewidywania trendów rynkowych i podejmowania bardziej świadomych decyzji inwestycyjnych. Modele uczenia maszynowego analizują historyczne dane rynkowe i identyfikują wzorce, które mogą być wykorzystane do przewidywania przyszłych trendów rynkowych.

Uczenie maszynowe przekształca branżę marketingową. Modele uczenia maszynowego mogą analizować dane klientów, takie jak historia zakupów, zachowanie podczas przeglądania stron internetowych i aktywność w mediach społecznościowych, aby przewidzieć preferencje i zachowania klientów. Może to pomóc sprzedawcom w personalizacji ich kampanii i oferowaniu klientom bardziej odpowiednich produktów i usług.

Algorytmy uczenia maszynowego analizują dane o klientach, aby przewidzieć, które produkty klient prawdopodobnie kupi i zaoferować spersonalizowane rekomendacje produktów.

Innym zastosowaniem uczenia maszynowego w marketingu jest tzw. „predictive lead scoring". Modele uczenia maszynowego mogą analizować dane o klientach i przewidywać, którzy potencjalni klienci (ang. *lead*) najprawdopodobniej zakupią produkt. Może to pomóc zespołom sprzedaży w ustaleniu priorytetów swoich działań i skupieniu się na klientach, którzy najprawdopodobniej wygenerują przychody.

Wpływ uczenia maszynowego jest dalekosiężny i przekształca różne branże. Pozwala na podejmowanie bardziej świadomych decyzji i poprawę ogólnej wydajności. Uczenie maszynowe może pomóc firmom podejmować lepsze decyzje, poprawić jakość produktów i usług oraz zwiększyć zadowolenie klientów.

Przyjęcie uczenia maszynowego rodzi jednak również obawy dotyczące prywatności czy stronniczości. O ograniczeniach uczenia maszynowego mowa w kolejnym rozdziale.

3.1.4 Ograniczenia uczenia maszynowego

Uczenie maszynowe okazało się ogromnym sukcesem w wielu dziedzinach zastosowań. Ma ono jednak również pewne ograniczenia, które warto wziąć pod uwagę.

Jednym z głównych ograniczeń uczenia maszynowego jest problem nadmiernego dopasowania. Nadmierne dopasowanie ma miejsce, gdy model jest zbyt złożony i uczy się szczegółów w danych treningowych zamiast podstawowych wzorców. W rezultacie, model osiąga dobre wyniki na danych treningowych, ale słabe na nowych danych. Ethem Alpaydin w swojej książce „Introduction to Machine Learning" wyjaśnia, że przeuczenie jest jednym z najpoważniejszych problemów w uczeniu maszynowym i występuje wtedy, gdy model jest zbyt złożony w stosunku do ilości dostępnych danych. Nadmiernemu dopasowaniu można zaradzić na przykład poprzez zastosowanie prostszych modeli, które mają mniej parametrów.

Kolejnym ograniczeniem uczenia maszynowego jest zależność od wysokiej jakości i reprezentatywnych danych. Jednak zbieranie i etykietowanie danych mogą być kosztowne i czasochłonne. Ponadto, jakość danych może znacząco wpłynąć na wydajność modelu, a stronniczość danych, taka jak stronniczość próbkowania lub stronniczość etykiet, może

prowadzić do nieobiektywnych przewidywań i decyzji. Dlatego kluczowe jest staranne przetwarzanie danych, aby zapewnić, że są one reprezentatywne i bezstronne.

Ważnym ograniczeniem uczenia maszynowego jest brak możliwości interpretacji i wyjaśnienia modeli. Wiele modeli uczenia maszynowego, takich jak sieci neuronowe, to tzw. „czarne skrzynki", co oznacza, że trudno jest zrozumieć, w jaki sposób dochodzą one do swoich przewidywań. Ten brak przejrzystości może stanowić istotną barierę dla przyjęcia systemów uczenia maszynowego w zastosowaniach, w których przejrzystość jest krytyczna, takich jak opieka zdrowotna lub finanse. Do interpretacji i wyjaśnienia decyzji modeli uczenia maszynowego można wykorzystać różne metody, takie jak analiza istotności cech czy wizualizacja modelu.

Innym ograniczeniem uczenia maszynowego jest kruchość modeli. Modele uczenia maszynowego mogą być wrażliwe na zmiany w danych wejściowych, a niewielkie perturbacje mogą skutkować drastycznie różnymi przewidywaniami. Ta kruchość może być problemem w zastosowaniach, w których solidność i niezawodność są krytyczne, takich jak samochody autonomiczne czy diagnostyka medyczna.

Wreszcie, modele uczenia maszynowego mogą utrwalać i wzmacniać istniejące uprzedzenia w danych. Na przykład, jeśli zbiór danych zawiera więcej mężczyzn niż kobiet ubiegających się o pracę, model uczenia maszynowego wyszkolony na tych danych może nauczyć się preferować kandydatów płci męskiej nad kandydatami płci żeńskiej, nawet jeśli płeć nie jest wyraźnie używana jako cecha. Może to prowadzić do dyskryminujących wyników i utrwalać społeczne uprzedzenia. Aby rozwiązać ten problem, badacze zaproponowali różne techniki, takie jak algorytmy usuwania błędów, powiększanie danych lub zbieranie bardziej zróżnicowanych i reprezentatywnych danych.

Podsumowując, uczenie maszynowe ma kilka ograniczeń, które warto wziąć pod uwagę przy projektowaniu skutecznych i godnych zaufania systemów. Nadmierne dopasowanie, zależność od wysokiej jakości danych, brak możliwości interpretacji i wyjaśnienia, kruchość i utrwalanie uprzedzeń to niektóre z głównych ograniczeń uczenia maszynowego. Ograniczenia te mogą mieć znaczący wpływ na wydajność, niezawodność i uczciwość systemów uczenia maszynowego, a ich rozwiązanie jest kluczowe dla zapewnienia, że uczenie maszynowe może być skutecznie i etycznie wykorzystywane w praktyce. Zrozumienie ograniczeń uczenia maszynowego jest jednak równie ważne jak zrozumienie jego mocnych stron. Będąc świadomym tych ograniczeń i opracowując odpowiednie techniki do ich rozwiązania, możemy wykorzystać uczenie maszynowe do rozwiązywania złożonych problemów i poprawy naszego życia.

3.2 Wybór metody

Uczenie nadzorowane to rodzaj uczenia maszynowego, w którym algorytm jest trenowany na zbiorze danych oznaczonych etykietami, aby nauczyć się wzorców i zależności między cechami wejściowymi a odpowiadającymi im etykietami wyjściowymi. Jest ono szeroko stosowane w różnych dziedzinach, w tym w finansach, opiece zdrowotnej i marketingu. Jednym z powszechnych zastosowań uczenia nadzorowanego jest przewidywanie dochodów na podstawie różnych czynników demograficznych i społeczno-ekonomicznych, takich jak wiek, wykształcenie, zawód i stan cywilny.

Zbiór danych „Adult Data" jest dobrze znany w środowisku uczenia maszynowego i zawiera informacje o osobach z różnych grup społecznych i wykonujących zróżnicowane zawody. Zbiór danych składa się z szeregu zmiennych objaśniających takich jak wiek, wykształcenie, stan cywilny, zawód, rasa, płeć, czy kraj pochodzenia. Atrybut dochód jest zmienną docelową i przyjmuje dwie wartości: <=50K i >50K, wskazujące, czy dana osoba zarabia odpowiednio mniej lub więcej niż 50 000 USD rocznie.

Zbiór danych jest odpowiednim kandydatem do uczenia nadzorowanego z kilku powodów. Po pierwsze, jest to duży i zróżnicowany zbiór danych ze znaczną liczbą atrybutów i obserwacji. Zapewnia to bogate źródło informacji dla algorytmów i pomaga zredukować nadmierne dopasowanie. Dodatkowo, zbiór danych zawiera zarówno atrybuty numeryczne, jak i kategoryczne, co jest powszechne w zastosowaniach w świecie rzeczywistym. Algorytmy uczenia nadzorowanego mogą obsługiwać oba typy danych albo poprzez konwersję atrybutów kategorycznych na wartości liczbowe, albo poprzez zastosowanie wyspecjalizowanych algorytmów dla danych kategorycznych.

Po drugie, zbiór danych dobrze nadaje się do przewidywania dochodu, ponieważ zawiera szereg czynników demograficznych i społeczno-ekonomicznych, o których wiadomo, że wpływają na poziom dochodu danej osoby. Wiek, wykształcenie i zawód są silnie skorelowane z dochodem, przy czym wyższy poziom wykształcenia i kwalifikacji zawodowych zazwyczaj skutkują wyższymi dochodami. Podobnie, stan cywilny i rasa również mogą odgrywać rolę w określaniu dochodu, przy czym osoby zamężne i należące do pewnych grup rasowych zazwyczaj mają wyższe dochody. Poprzez trenowanie algorytmu uczenia nadzorowanego na tym zbiorze danych, możemy zidentyfikować najważniejsze czynniki, które przyczyniające się do wysokich i dokonać dokładnych prognoz w oparciu o te czynniki.

Po trzecie, zbiór danych jest publicznie dostępny i był szeroko wykorzystywany w społeczności uczenia maszynowego, co czyni go wartościowym zbiorem wzorcowym do oceny

wydajności różnych algorytmów uczenia nadzorowanego. Wielu badaczy zastosowało różne algorytmy do zbioru danych, w tym drzewa decyzyjne, sieci neuronowe czy lasy losowe.

Wreszcie, przewidywanie dochodu jest powszechnym zadaniem w wielu sytuacjach praktycznych, w tym w ocenie ryzyka kredytowego, marketingu i tworzeniu polityki rządowej. Stosując uczenie nadzorowane do przewidywania dochodu na podstawie zbioru danych „Adult Data", możemy opracować modele, które mogą być wykorzystane do podejmowania bardziej świadomych decyzji. Na przykład, banki mogą używać modeli przewidywania dochodów do oceny zdolności kredytowej, podczas gdy rządy mogą używać ich do tworzenia polityki promującej równość ekonomiczną i zmniejszającej różnice w dochodach.

Podsumowując, zbiór danych „Adult Data" jest odpowiednim kandydatem do uczenia nadzorowanego w celu przewidywania dochodu, ponieważ jest to duży i zróżnicowany zbiór danych z szeregiem czynników demograficznych i społeczno-ekonomicznych, o których wiadomo, że wpływają na poziom dochodu. Zestaw danych jest dobrze dostosowany do przewidywania dochodu, ponieważ zawiera zarówno atrybuty liczbowe, jak i kategoryczne, i był szeroko stosowany w społeczności uczenia maszynowego jako zestaw danych wzorcowych do oceny wydajności różnych algorytmów uczenia nadzorowanego. Poprzez wykorzystanie uczenia nadzorowanego do przewidywania dochodu na podstawie tego zestawu danych, możemy opracować modele, które mogą być wykorzystane w rzeczywistych zastosowaniach do podejmowania bardziej świadomych decyzji, takich jak ocena ryzyka kredytowego, marketing i tworzenie polityki rządowej. Dodatkowo, modele przewidywania dochodu mogą pomóc w zmniejszeniu uprzedzeń i dyskryminacji w decyzjach o zatrudnieniu i awansie poprzez oparcie się na obiektywnych danych, a nie subiektywnych czynnikach.

3.2.1 Drzewo klasyfikacyjne

Drzewa klasyfikacyjne są popularną i efektywną techniką uczenia maszynowego, która znalazła szerokie zastosowanie w wielu dziedzinach, w tym w medycynie, finansach i inżynierii. Drzewo klasyfikacyjne to drzewo decyzyjne, które służy do przewidywania etykiety klasy obiektu. Główną ideą drzewa klasyfikacyjnego jest podział przestrzeni wejściowej na regiony, gdzie każdy region jest związany z etykietą klasy.

Drzewa klasyfikacyjne mają szeroki zakres zastosowań w różnych dziedzinach. Jednym z najczęstszych zastosowań drzew klasyfikacyjnych jest diagnostyka medyczna. Drzewa klasyfikacyjne mogą być używane do diagnozowania różnych stanów chorobowych na podstawie objawów, badań laboratoryjnych i innych testów diagnostycznych. Na przykład, drzewa klasyfikacyjne zostały wykorzystane do przewidywania ryzyka wystąpienia raka piersi

na podstawie różnych czynników, takich jak wiek, historia zdrowotna rodziny i czynniki hormonalne.

Ta metoda uczenia maszynowego była również wykorzystywane w finansach do przewidywania cen akcji i trendów rynkowych. Na przykład, drzewa klasyfikacyjne mogą być wykorzystywane do przewidywania ruchów na rynku akcji w oparciu o różne czynniki, takie jak cena towarów, inflacja i stopy procentowe. Ponadto, drzewa klasyfikacyjne są wykorzystywane w scoringu kredytowym do przewidywania zdolności kredytowej na podstawie różnych czynników, takich jak dochód, historia kredytowa i status zatrudnienia.

Drzewa klasyfikacyjne są również szeroko stosowane w inżynierii do diagnozowania błędów i kontroli jakości. Na przykład, mogą być wykorzystywane do wykrywania błędów w systemach przemysłowych na podstawie danych z czujników, parametrów procesu i innych powiązanych czynników.

Drzewa budowane są przy użyciu różnych algorytmów, w tym algorytmu CART (Classification and Regression Trees), będącego binarnym algorytmem rekurencyjnego podziału, który buduje drzewa decyzyjne dla problemów klasyfikacji i regresji. Został opracowany przez Breimana, Friedmana, Olshena i Stone'a w latach 80-tych[30].

Algorytm CART rekurencyjnie dzieli dane na dwie części w oparciu o najlepszy podział na każdym kroku. Najlepszy podział jest określany poprzez minimalizację wskaźnika Giniego, który jest miarą nieczystości danych. Indeks Giniego dla węzła jest zdefiniowany jako

$$GI(X) = 1 - \sum_{d \in C} \left(\frac{|X_d|}{|X|} \right)^2 \in [0; 1]$$

gdzie C to liczba klas, $|X_d|$ to liczba obserwacji należących do danej klasy, a $|X|$ to liczba obserwacji w danym podzbiorze. Węzeł jest czysty, jeśli wszystkie obiekty w węźle należą do tej samej klasy, a wskaźnik Giniego wynosi 0.

Aby wyznaczyć najlepszy podział, CART rozważa wszystkie możliwe podziały dla każdej cechy i wybiera podział, który minimalizuje sumę ważoną indeksów Giniego węzłów podrzędnych. Suma ważona obliczana jest jako:

$$GI_t(X) = 1 - \sum_{r,t \in C} \frac{|X_{tr}|}{|X|} * GI_{tr}(X)$$

[30] The Complete Guide to Decision Trees, https://www.explorium.ai/blog/the-complete-guide-to-decision-trees/, data dostępu: 3 marca 2023 roku

gdzie r to liczba klas powstałych w wyniku testu, t to numer testu, X_{tr} to liczba obserwacji dla danej klasy powstałych w wyniku testu, $|X|$ to ogólna liczba obserwacji w zbiorze, a $GI_{tr}(X)$ to wartość współczynnika Giniego dla każdej z klas powstałych w wyniku testu.

Po dokonaniu podziału, algorytm rekurencyjnie stosuje ten sam proces do węzłów potomnych, aż do spełnienia kryterium zatrzymania. Kryterium zatrzymania może być oparte na głębokości drzewa, minimalnej liczbie obiektów w węźle liścia lub maksymalnej liczbie węzłów liścia.

W problemach klasyfikacji, ostateczne drzewo decyzyjne składa się z węzłów liści, które odpowiadają przewidywanym etykietom klas. W problemach regresji, ostateczne drzewo decyzyjne składa się z węzłów liści, które odpowiadają przewidywanym wartościom liczbowym.

Jedną z zalet algorytmu CART jest to, że może on obsługiwać zarówno cechy kategoryczne, jak i ciągłe. Dla cech kategorycznych algorytm rozważa wszystkie możliwe podziały na podstawie kategorii. Dla cech ciągłych algorytm rozpatruje wszystkie możliwe progi podziału cechy.

Algorytm CART ma jednak pewne ograniczenia. Jednym z nich jest to, że może on nadmiernie dopasować się do danych, zwłaszcza jeśli drzewo jest zbyt głębokie lub jeśli istnieją nieistotne cechy. Nadmiernego dopasowania można uniknąć poprzez przycięcie drzewa lub zastosowanie innych technik.

Podsumowując, algorytm CART jest elastyczną metodą budowania drzew decyzyjnych dla problemów klasyfikacji i regresji. Wykorzystuje on wskaźnik Giniego do określenia najlepszego podziału i może obsługiwać zarówno cechy kategoryczne, jak i ciągłe. Może jednak nadmiernie dopasować dane i wymaga starannego dostrojenia parametrów, aby uniknąć przeuczenia.

Do stworzenia drzew klasyfikacyjnych w badaniu, posłuży funkcja rpart w R, która wykorzystuje algorytm Recursive Partitioning and Regression Trees (RPART), będący rozszerzeniem algorytmu CART. Algorytm RPART wykorzystuje to samo binarne podejście do partycjonowania rekursywnego, co CART, ale używa nieco innego kryterium podziału zwanego "parametrem złożoności" lub "minimalnym kosztem-kompleksowością" do określenia optymalnego drzewa.

Parametr złożoności jest parametrem dostrojenia, który równoważy kompromis między złożonością drzewa a jego dokładnością na danych treningowych. Algorytm RPART najpierw tworzy duże, nieokrojone drzewo, które jest zbyt dopasowane do danych, a następnie przycina drzewo poprzez usunięcie gałęzi, które nie poprawiają dokładności drzewa na danych

walidacyjnych. Optymalny parametr złożoności jest określany poprzez minimalizację funkcji kosztu, która penalizuje złożoność drzewa.

Funkcja rpart zapewnia również kilka innych kryteriów podziału, w tym wskaźnik Giniego, kryterium entropii krzyżowej i zasadę dwóch stron. Kryteria te mogą być określone za pomocą argumentu parms w funkcji. Domyślnie funkcja używa wskaźnika Giniego dla problemów klasyfikacji i sumy błędów kwadratowych dla problemów regresji.

Drzewa klasyfikacyjne mogą być oceniane przy użyciu różnych metryk opartych na macierzy błędów, w tym accuracy, precision, specificity, sensitivity czy recall. Oprócz tych metryk, do oceny drzew klasyfikacyjnych można również wykorzystać krzywą ROC, AUC i krzywą Lift. Krzywa ROC jest graficzną reprezentacją współczynnika prawdziwych pozytywów w stosunku do współczynnika fałszywych pozytywów przy różnych progach. Metody oceny jakości drzew decyzyjnych zostaną opisane szczegółowo w Podrozdziale 3.2.4.

3.2.2 Las losowy

Lasy losowe to rodzaj algorytmu uczenia maszynowego, który jest używany do klasyfikacji, regresji i innych zadań. Są one rodzajem metody uczenia zespołowego, co oznacza, że łączą wiele modeli, aby dokonać przewidywań. Lasy losowe są szczególnie przydatne dla dużych, złożonych zbiorów danych z wieloma cechami i stały się popularnym narzędziem dla naukowców zajmujących się danymi w różnych dziedzinach.

Las losowy to rodzaj algorytmu uczenia maszynowego, który jest używany zarówno do zadań regresji, jak i klasyfikacji. Jest to metoda uczenia zespołowego, która łączy wiele drzew decyzyjnych, aby stworzyć bardziej solidny i dokładny model.

Las losowy jest tworzony poprzez trenowanie wielu drzew decyzyjnych na różnych podzbiorach danych i cech, a następnie łączenie ich przewidywań. Każde drzewo decyzyjne w lesie jest trenowane na losowo wybranym podzbiorze danych, zwanym próbką bootstrapową oraz losowo wybranym podzbiorze cech. Ta losowość pomaga zmniejszyć przeuczenie i poprawić wydajność generalizacji modelu.

Aby dokonać przewidywania za pomocą lasu losowego, dane wejściowe są przepuszczane przez każde drzewo decyzyjne w lesie, a przewidywania są łączone przy użyciu głosowania większościowego (dla klasyfikacji) lub wartości średniej (dla regresji). Wyjściem lasu losowego jest pojedyncza prognoza, która reprezentuje decyzję wszystkich drzew w lesie.

Lasy losowe działają poprzez połączenie przewidywań wielu drzew decyzyjnych w celu stworzenia bardziej solidnego i dokładnego modelu. Każde drzewo decyzyjne w lesie jest trenowane na innym podzbiorze danych i cech, co pomaga zredukować przeuczenie i poprawić wydajność generalizacji.

Proces tworzenia lasu losowego można podzielić na trzy główne kroki: próbkowanie, dzielenie i łączenie.

Próbkowanie jest pierwszym krokiem w tworzeniu lasu losowego. Polega ono na wygenerowaniu wielu próbek bootstrapowych danych. Próbka bootstrapowa jest losowym podzbiorem oryginalnych danych, utworzonym przez losowanie ze zwracaniem. Oznacza to, że każda próbka zawiera niektóre dane unikatowe, ale może również zawierać duplikaty.

Tworząc wiele próbek bootstrapowych danych, możemy trenować wiele drzew decyzyjnych na różnych podzbiorach danych. Pomaga to zredukować nadmierne dopasowanie, ponieważ każde drzewo jest trenowane na innym zestawie przykładów.

Podział jest drugim krokiem w tworzeniu lasu losowego i polega na wytrenowaniu wielu drzew decyzyjnych na próbkach bootstrapowych. Każde drzewo decyzyjne jest trenowane przy użyciu innego losowego podzbioru cech. Pomaga to zredukować przeuczenie, ponieważ każde drzewo jest trenowane na innym podzbiorze dostępnych cech.

Podczas trenowania drzewa decyzyjnego, rekurencyjnie dzielimy dane na mniejsze podzbiory w oparciu o wartości cech. Celem jest stworzenie podzbiorów, które są jak najbardziej homogeniczne w odniesieniu do zmiennej docelowej. Proces dzielenia danych jest kontynuowany do momentu spełnienia kryterium zatrzymania, takiego jak osiągnięcie maksymalnej głębokości lub minimalnej liczby przykładów w węźle liścia.

Łączenie przewidywań wielu drzew decyzyjnych stanowi ostatni krok w tworzeniu lasu losowego jest połączenie przewidywań wielu drzew decyzyjnych. Aby dokonać przewidywania za pomocą lasu losowego, przepuszczamy dane wejściowe przez każde drzewo decyzyjne w lesie, a następnie łączymy przewidywania za pomocą głosowania większościowego (dla klasyfikacji) lub wartości średniej (dla regresji).

Wyjściem lasu losowego jest pojedyncza prognoza, która reprezentuje konsensus wszystkich drzew w lesie. Ponieważ każde drzewo w lesie zostało wytrenowane na innym podzbiorze danych i cech, połączone przewidywania są często bardziej dokładne i mniej podatne na przeuczenie niż przewidywania pojedynczego drzewa decyzyjnego.

Lasy losowe są solidne, gdyż są odporne na hałaśliwe lub nieistotne cechy w danych. Ponieważ każde drzewo decyzyjne w lesie jest trenowane na innym podzbiorze cech, drzewa są mniej podatne na wpływ nieistotnych lub odstających cech w danych. Może to skutkować bardziej dokładnym modelem, który jest mniej podatny na przeuczenie.

Lasy losowe są skalowalne do dużych zbiorów danych z wieloma cechami. Ponieważ drzewa mogą być trenowane równolegle na różnych podzbiorach danych, lasy losowe mogą

być efektywnie trenowane na dużych zbiorach danych z wieloma cechami. To czyni je popularnym wyborem dla naukowców pracujących z dużymi, złożonymi zbiorami danych.

Lasy losowe mogą być używane do różnych zadań uczenia maszynowego, w tym klasyfikacji, regresji i selekcji cech. Są one również wystarczająco elastyczne, aby obsługiwać zarówno dane kategoryczne, jak i ciągłe, co czyni je uniwersalnym narzędziem dla naukowców pracujących z różnymi typami danych.

Lasy losowe są stosunkowo łatwe w interpretacji w porównaniu do innych modeli uczenia maszynowego, takich jak sieci neuronowe. Ponieważ dane wyjściowe lasu losowego są oparte na konsensusie wielu drzew decyzyjnych, często możliwe jest określenie, które cechy są najważniejsze dla tworzenia prognoz. Może to być przydatne do zrozumienia wzorców leżących u podstaw danych oraz do wyjaśnienia wyników modelu.

Lasy losowe stały się popularnym narzędziem dla naukowców zajmujących się danymi w różnych dziedzinach.

Lasy losowe mogą być wykorzystywane do zadań klasyfikacji obrazów, takich jak identyfikacja obiektów na obrazach lub klasyfikowanie obrazów do różnych kategorii. Poprzez szkolenie wielu drzew decyzyjnych na różnych podzbiorach cech obrazu, las losowy może stworzyć dokładniejszy i solidniejszy model klasyfikacji obrazu.

Wykrywanie oszustw, takich jak identyfikacja fałszywych transakcji lub kont, jest kolejnym istotnym zastosowaniem. Poprzez trening lasu losowego na danych historycznych, możliwe jest zidentyfikowanie wzorców nieuczciwych zachowań i wykorzystanie tych informacji do oznaczenia podejrzanych transakcji lub kont w czasie rzeczywistym.

Działania marketingowe, takie jak przewidywanie zachowań klientów lub identyfikowanie segmentów klientów, to kolejne zastosowanie lasów losowych. Poprzez szkolenie lasu losowego na danych klientów, możliwe jest zidentyfikowanie wzorców zachowań, które mogą być wykorzystane do tworzenia ukierunkowanych kampanii marketingowych lub personalizacji doświadczeń klientów.

Lasy losowe mogą być wykorzystywane do zadań związanych z diagnostyką medyczną, takich jak identyfikacja pacjentów zagrożonych określonymi chorobami lub przewidywanie skuteczności różnych metod leczenia. Trenując las losowy na danych medycznych, można zidentyfikować wzorce objawów lub czynników ryzyka, które można wykorzystać do stawiania dokładniejszych diagnoz i zaleceń dotyczących leczenia.

Podczas korzystania z lasów losowych do zadań uczenia maszynowego istnieje kilka najlepszych praktyk, które mogą pomóc w zapewnieniu najlepszej możliwej wydajności:

Staranny dobór cech jest kluczem do stworzenia dokładnego i solidnego modelu lasu losowego. Wybierając najważniejsze cechy dla danego zadania, można stworzyć bardziej wydajny i dokładny model, który jest mniej podatny na przeuczenie. Niektóre popularne metody wyboru cech obejmują analizę korelacji, analizę głównych składowych (PCA) i miary istotności cech.

Lasy losowe posiadają kilka parametrów kontrolnych, które mogą być dostrojone w celu optymalizacji wydajności modelu. Niektóre popularne parametry obejmują liczbę drzew w lesie, głębokość drzew i wielkość próby bootstrapowej. Eksperymentując z różnymi kombinacjami parametrów można stworzyć model, który jest dobrze dostrojony do konkretnego zadania i osiąga najlepszą możliwą wydajność.

Walidacja krzyżowa jest ważną techniką oceny wydajności modelu lasu losowego. Poprzez podział danych na zbiory treningowe i testowe można ocenić, jak dobrze model sprawdza się na nowych danych. K-krotna walidacja krzyżowa jest powszechną techniką oceny wydajności modelu lasu losowego, gdzie dane są dzielone na K podzbiorów, a model jest trenowany i testowany na różnych podzbiorach danych.

Lasy losowe mogą mieć problemy z niezbilansowanymi zbiorami danych, gdzie jedna klasa jest znacznie bardziej powszechna niż inne. W takich przypadkach pomocne może być zrównoważenie zbioru danych poprzez nadpróbkowanie klasy mniejszościowej, niedopróbkowanie klasy większościowej lub użycie kombinacji obu technik. Inną opcją jest użycie technik takich jak uczenie wrażliwe na koszty lub ważone lasy losowe, które przypisują większe wagi klasie mniejszościowej, aby zapewnić, że nie zostanie ona pominięta przez model.

Interpretowalność jest istotną kwestią przy stosowaniu lasów losowych do zadań uczenia maszynowego. Poprzez zrozumienie, które cechy są najważniejsze dla przewidywań, możliwe jest uzyskanie wglądu w wzorce leżące u podstaw danych i zrozumienie wyników modelu.

Lasy losowe oferują istotne zalety. Istnieje jednak kilka wyzwań, które warto wziąć pod uwagę.

Lasy losowe mogą być intensywne obliczeniowo, szczególnie gdy mamy do czynienia z dużymi zbiorami danych z wieloma cechami. Trenowanie wielu drzew decyzyjnych na różnych podzbiorach danych wymaga znacznych zasobów obliczeniowych, co może stanowić barierę.

Chociaż lasy losowe są mniej podatne na przeuczanie niż pojedyncze drzewa decyzyjne, zjawisko to może ciągle występować. Ważne jest, aby starannie dostroić parametry kontrolne

modelu i użyć technik takich jak walidacja krzyżowa, aby upewnić się, że model dobrze sprawdza się na nowych danych.

Lasy losowe są stosunkowo łatwe w interpretacji w porównaniu z innymi modelami uczenia maszynowego. Jednak w niektórych przypadkach mogą stanowić wyzwanie. Na przykład, gdy występują silne interakcje pomiędzy cechami lub gdy cechy mają nieliniowe zależności ze zmienną wynikową, zrozumienie, w jaki sposób model dokonuje przewidywań, może być problemem.

Podsumowując, lasy losowe są techniką uczenia maszynowego, która zyskała na popularności w ostatnich latach. Poprzez trenowanie wielu drzew decyzyjnych na różnych podzbiorach danych i cech, lasy losowe mogą stworzyć bardziej dokładny i solidny model, który jest mniej podatny na przetrenowanie niż pojedyncze drzewa decyzyjne. Lasy losowe mają wiele zalet i mogą być stosowane w wielu zadaniach uczenia maszynowego, w tym klasyfikacji obrazów, wykrywaniu oszustw, marketingu czy diagnostyki medycznej. Jednakże, istnieje również kilka wyzwań, które warto rozważyć podczas korzystania z lasów losowych, w tym ich złożoność obliczeniowa, ryzyko przeuczenia oraz trudności w interpretacji. Poprzez stosowanie najlepszych praktyk w zakresie wyboru cech, dostrajania parametrów kontrolnych oraz walidacji krzyżowej. możliwe jest stworzenie dobrze dostrojonego i dokładnego modelu lasu losowego, który osiągnie wysoką wydajność.

3.2.3 Regresja

Analiza regresji jest metodą statystyczną stosowaną do badania zależności pomiędzy zmienną zależną a jedną lub więcej zmiennych niezależnych. Służy do modelowania związku między zmiennymi oraz do dokonywania przewidywań na podstawie tego związku. Analiza regresji jest silnym narzędziem w dziedzinie statystyki, ekonomii, finansów oraz nauk społecznych i wielu innych.

Zmienna zależna jest przewidywana, a zmienne niezależne są używane do tworzenia przewidywań. Celem analizy regresji jest znalezienie najlepiej dopasowanej linii lub krzywej, która opisuje związek między zmiennymi. Istnieje kilka rodzajów analizy regresji, z których każdy ma swój własny zestaw założeń i metod.

Prosta regresja liniowa jest najprostszą formą analizy regresji, która polega na przewidywaniu zmiennej zależnej na podstawie jednej zmiennej niezależnej. Związek między dwiema zmiennymi jest modelowany za pomocą linii prostej. Równanie dla prostej regresji liniowej to:

$$Y = \beta_0 + \beta_1 * X + \varepsilon,$$

gdzie Y jest zmienną zależną, X jest zmienną niezależną, β_0 jest wyrazem wolnym, β_1 jest nachyleniem, a ε składnikiem błędu. Nachylenie linii reprezentuje zmianę Y dla każdej jednostkowej zmiany X, natomiast wyraz wolny β_0 reprezentuje wartość Y, gdy X jest równe zero. Prosta regresja liniowa zakłada, że związek między dwiema zmiennymi jest liniowy, co oznacza, że nachylenie linii jest stałe dla wszystkich wartości X. Zakłada również, że składnik błędu ε jest ma rozkład normalny ze średnią równą zeru stałą wariancją.

Regresja liniowa wielokrotna to technika regresji polegająca na przewidywaniu zmiennej zależnej na podstawie dwóch lub więcej zmiennych niezależnych. Związek pomiędzy zmienną zależną a zmiennymi niezależnymi modelowany jest za pomocą równania liniowego. Równanie dla regresji liniowej wielokrotnej to:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_n * X_n + \varepsilon,$$

gdzie Y jest zmienną zależną, X_1, X_2, X_3, X_n są zmiennymi niezależnymi, β_0 jest wyrazem wolnym, β_1, β_2, $\beta_3, \dots$, β_n są nachyleniami, a ε jest składnikiem błędu. Wielokrotna regresja liniowa zakłada, że związek między zmienną zależną a zmiennymi niezależnymi jest liniowy, a składnik błędu ε ma rozkład normalny o średniej zero i stałej wariancji.

Regresja wielomianowa jest rodzajem analizy regresji, która polega na przewidywaniu zmiennej zależnej na podstawie wielomianowej funkcji jednej lub więcej zmiennych niezależnych. Związek między zmienną zależną a zmiennymi niezależnymi jest modelowany za pomocą równania wielomianowego. Równanie regresji wielomianowej to:

$$\mathrm{Y} = \beta_0 + \beta_1 * \mathrm{X}^2 + \beta_2 * \mathrm{X}^2 + \beta_3 * \mathrm{X}^3 + \beta_\mathrm{n} * \mathrm{X}^\mathrm{n} + \varepsilon,$$

gdzie Y to zmienna zależna, X to zmienna niezależna, β_0 to wyraz wolny, β_1, β_2, $\beta_3, \dots, \beta_n$ to nachylenia, a ε to składnik błędu. Regresja wielomianowa zakłada, że związek między zmienną zależną a zmienną niezależną nie ma postaci prostej, a składnik błędu ε ma rozkład normalny o średniej zero i stałej wariancji.

Regresja logistyczna jest rodzajem analizy regresji, która służy do modelowania prawdopodobieństwa wyniku binarnego (np. tak lub nie, sukces lub porażka) w oparciu o jedną lub więcej zmiennych niezależnych. Związek między zmiennymi niezależnymi a zmienną zależną jest modelowany za pomocą funkcji logistycznej. Równanie regresji logistycznej to:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_n * X_n)}},$$

gdzie P(Y=1) jest prawdopodobieństwem, że zmienna zależna jest równa 1 (tzn. zdarzenie zachodzi), X_1, X_2, X_3, X_n są zmiennymi niezależnymi, β_0 jest wyrazem wolnym, a β_1, β_2, $\beta_3, \dots, \beta_n$ są nachyleniami. Regresja logistyczna zakłada, że związek między zmiennymi

niezależnymi a zmienną zależną jest nieliniowy, a prawdopodobieństwo wystąpienia zdarzenia ma rozkład logistyczny.

Analiza regresji jest szeroko stosowana w ekonomii do modelowania zależności pomiędzy różnymi zmiennymi ekonomicznymi. Na przykład, analiza regresji może być wykorzystana do przewidywania wpływu zmian stóp procentowych na wydatki konsumentów lub do oszacowania popytu na dany produkt na podstawie jego ceny i innych czynników.

Technika ta jest używana w finansach do modelowania relacji pomiędzy różnymi zmiennymi finansowymi. Na przykład, może być używana do przewidywania cen akcji na podstawie finansów firmy, wskaźników ekonomicznych i innych czynników.

Analizę wykorzystuje się również w marketingu do modelowania relacji pomiędzy zmiennymi marketingowymi a zachowaniem konsumentów. Na przykład, analiza regresji może być używana do przewidywania wpływu reklamy na sprzedaż lub do oszacowania wrażliwości cenowej konsumentów.

W naukach społecznych analizę stosuje się do modelowania relacji pomiędzy różnymi zmiennymi społecznymi. Na przykład, może być wykorzystana do przewidywania wpływu wykształcenia na dochód lub do oszacowania wpływu czynników demograficznych na zachowania wyborcze.

Istnieje kilka technik stosowanych w analizie regresji w celu oszacowania parametrów modelu i przetestowania jego poprawności, w tym:

Regresja zwykła najmniejszych kwadratów (MNK) jest najczęściej stosowaną techniką szacowania parametrów modelu regresji liniowej. Metoda ta polega na znalezieniu linii, która minimalizuje sumę kwadratów różnic między obserwowanymi wartościami zmiennej zależnej a wartościami przewidywanymi.

Metoda największej wiarygodności (MLE) jest techniką stosowaną do estymacji parametrów modelu regresji, gdy znany jest rozkład składnika błędu. Polega ona na znalezieniu takich wartości parametrów, które maksymalizują prawdopodobieństwo obserwacji danego zbioru danych.

Analiza regresji to szeroko stosowana technika statystyczna, która ma jednak kilka poważnych ograniczeń. Po pierwsze, zakłada ona stały związek przyczynowo-skutkowy pomiędzy analizowanymi zmiennymi, co nie zawsze może być prawdą. W związku z tym szacunki dokonywane na podstawie równania regresji mogą dawać błędne i mylące wyniki oraz stwarzać problemy w przypadku ich ekstrapolacji.

Problem w przypadku regresji stanowić może również współliniowość. Jest istotnym ograniczeniem, zwłaszcza w przypadkach, gdy dwie lub więcej zmiennych niezależnych jest

ze sobą silnie skorelowanych. W takich sytuacjach model regresji może dawać niewiarygodne lub niestabilne oszacowania współczynników regresji, co może prowadzić do mylących wniosków.

Współliniowość może również powodować trudności w interpretacji współczynników regresji, ponieważ trudne staje się określenie niezależnego wpływu każdej zmiennej predykcyjnej na zmienną zależną. Ponadto, współliniowość może spowodować, że model regresji będzie nadawał zbyt dużą wagę niektórym zmiennym objaśniającym, ignorując inne, co może prowadzić do tendencyjnych oszacowań współczynników regresji.

Po trzecie, analiza regresji wiąże się z długim i skomplikowanym procesem obliczeń i analizy, który może być trudny do przeprowadzenia, zwłaszcza dla osób o ograniczonej wiedzy i zasobach statystycznych.

Analiza regresji nie nadaje się również do analizy zjawisk jakościowych, takich jak uczciwość czy przestępczość. W tych sytuacjach warto stosować alternatywne techniki.

Podsumowując, analiza regresji jest narzędziem pozwalającym na modelowanie związku między zmiennymi i dokonywanie prognoz na podstawie tego związku. Ma ona wiele zastosowań w różnych dziedzinach i może być wykorzystywana wraz z różnymi technikami do szacowania parametrów modelu i testowania jego poprawności. Ważne jest jednak, aby rozpoznać jej ograniczenia i w razie potrzeby używać jej odpowiednio w połączeniu z innymi metodami.

3.2.4 Metody oceny jakości modeli uczenia maszynowego

Budowanie modelu to tylko pierwszy krok w procesie uczenia maszynowego. Ocena jakości modelu jest równie ważna, gdyż pozwala na zapewnienie jego skuteczności w przypadku nowych danych.

Jednym z najczęstszych sposobów oceny modeli uczenia maszynowego jest użycie macierzy błędów. Prezentuje ona porównanie przewidywanych i rzeczywistych klas w problemie klasyfikacji. Na podstawie tej macierzy można uzyskać kilka metryk oceny, takich jak dokładność, współczynnik błędnej klasyfikacji, precyzja, negatywna wartość predykcyjna, czułość, specyficzność oraz miara F1.

Oprócz macierzy błędów istnieją inne metody oceny modeli uczenia maszynowego. Należą do nich wykresy ROC i Lift oraz AUC, które są powszechnie stosowane w problemach klasyfikacji. Dla problemów regresji często stosuje się metryki oceny takie jak resztowa suma kwadratów odchyleń, średni błąd bezwzględny czy średni błąd kwadratowy.

Metryki ewaluacyjne są kluczowym elementem procesu uczenia maszynowego, ponieważ pozwalają nam mierzyć jakość naszych modeli i podejmować świadome decyzje o

tym, których modeli użyć do konkretnego zastosowania. Poprzez staranny wybór odpowiedniej metryki oceny, odpowiednie szkolenie i testowanie modelu oraz unikanie nadmiernego dopasowania, możemy zbudować modele, które są dokładne, odporne i dobrze generalizują na nowe dane. Celem Podrozdziału 3.2.4 jest omówienie najważniejszych metryk oraz ich zastosowań.

Macierz błędów jest tabelą, która podsumowuje liczbę poprawnych i niepoprawnych przewidywań dokonanych przez model klasyfikacyjny na zestawie danych testowych[31]. Jest ona przydatna do oceny wydajności modelu uczenia maszynowego i stanowi podstawę do obliczenia istotnych metryk oceny. Dla ilustracji, rozważmy następujący przykład macierzy błędów dla problemu klasyfikacji binarnej:

Przykładowa macierz błędów

	Stan Pozytywny	**Stan Negatywny**
Klasyfikacja Pozytywna	80 Prawdziwie dodatnia (TP)	20 Fałszywie dodatnie (FP)
Klasyfikacja Negatywna	10 Fałszywie ujemna (FN)	90 Prawdziwie ujemna (TN)

Źr.: opr. wł. na podst. David M W Powers (2007)

Tabela 3.1. *Dla przykładowych danych model pozytywnie klasyfikuje 100 obserwacji, z czego 80 to klasyfikacje poprawne. Negatywnie zaklasyfikowanych jest również 100 obserwacji, z czego 90 poprawnie*.

W macierzy błędów przedstawionej w Tabeli 3.1. wyróżnić można cztery komórki, z których każda reprezentuje możliwy wynik przewidywań modelu klasyfikacyjnego. Komórki są zdefiniowane w następujący sposób: prawdziwie dodatnia (ang. *true positive* – TP): liczba obserwacji poprawnie przewidywanych jako pozytywne; fałszywie dodatnia (ang. *false positive* – FP): liczba przypadków błędnie przewidywanych jako pozytywne; fałszywie ujemna (ang. *false negative* – FN): liczba przypadków błędnie przewidzianych jako negatywne; prawdziwie ujemna (ang. true negative – TN): liczba przypadków prawidłowo przewidzianych jako negatywne.

Macierz błędów jest cennym narzędziem, pozwalającym na ocenę jakości modelu uczenia maszynowego. Na podstawie wartości w niej zawartych wyróżnić można następujące pochodne metryki oceny:

[31] „Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", David M W Powers, 2007

Dokładność (ang. *accuracy*) jest powszechnie stosowaną metryką, która mierzy odsetek poprawnych przewidywań dokonanych przez model. Oblicza się ją w następujący sposób:

$$\text{Dokładność} = (TP + TN) / (TP + TN + FP + FN)$$

W przykładowej macierzy błędów dokładność wynosi (80 + 90) / (80 + 90 + 20 + 10) = 85%. Dokładność jest przydatną metryką, gdy klasy w danych są zrównoważone. Jednak może być myląca w sytuacjach, w których klasy są niezrównoważone. Na przykład, jeśli 90% danych należy do klasy negatywnej, model, który zawsze przewiduje klasę negatywną, będzie miał wysoką dokładność, ale może nie być użyteczny w praktyce.

Współczynnik błędnej klasyfikacji (ang. *misclassification error rate*) to odsetek błędnych przewidywań dokonanych przez model. Oblicza się go w następujący sposób:

$$MER = (FP + FN) / (TP + TN + FP + FN)$$

W przykładowej macierzy błędów MER wynosiłby (20 + 10) / (80 + 90 + 20 + 10) = 15%. Współczynnik błędnej klasyfikacji jest uzupełnieniem dokładności (precision) i reprezentuje proporcję obserwacji, które model klasyfikuje błędnie.

Precyzja (ang. *precision*) mierzy odsetek klasyfikacji prawdziwie dodatnich wśród wszystkich obserwacji przewidywanych jako pozytywne. Oblicza się ją w następujący sposób:

$$\text{Precyzja} = TP / (TP + FP)$$

W przykładowej macierzy błędów precyzja wynosiłaby 80 / (80 + 20) = 0,8 lub 80%. Precyzja jest użyteczną metryką, gdy koszt klasyfikacji fałszywie pozytywnych jest wysoki. Na przykład w medycynie, fałszywa diagnoza może prowadzić do niepotrzebnego leczenia, więc precyzja byłaby ważną metryką do optymalizacji.

Negatywna wartość predykcyjna (ang. *negative predictive value*) mierzy proporcję prawdziwych negatywów wśród wszystkich obserwacji przewidywanych jako negatywne. Oblicza się ją w następujący sposób:

$$NPV = TN / (TN + FN)$$

W przykładowej macierzy błędów, NPV wynosiłoby 90 / (90 + 10) = 90%. NPV jest przydatną metryką, gdy koszt fałszywych negatywów jest wysoki.

Czułość (ang. *sensitivity*) mierzy odsetek prawdziwych pozytywów wśród wszystkich obserwacji, które są rzeczywiście pozytywne. Oblicza się ją w następujący sposób:

$$\text{Czułość} = TP / (TP + FN)$$

W przykładowej macierzy błędu czułość wynosiłaby 80 / (80 + 10) = 89%. Czułość jest przydatną metryką, gdy koszt klasyfikacji fałszywie negatywnych jest wysoki. Na przykład w

problemie diagnozowania raka, fałszywa klasyfikacja negatywna może zagrażać życiu, więc czułość byłaby ważną metryką do optymalizacji.

Specyficzność (ang. *specificity*) mierzy proporcję prawdziwych negatywów wśród wszystkich przypadków, które są rzeczywiście negatywne. Oblicza się ją w następujący sposób:

$$\text{Specificity} = TN / (TN + FP)$$

W przykładowej macierzy konfuzji swoistość wynosiłaby 90 / (90 + 20) = 82%. Swoistość jest przydatną metryką, gdy koszt fałszywych pozytywów jest wysoki.

Współczynnik fałszywych pozytywów (ang. *false positive rate*) jest proporcją negatywnych obserwacji, które są błędnie klasyfikowane jako pozytywne przez klasyfikator binarny. Innymi słowy, mierzy on, jak często klasyfikator błędnie identyfikuje instancję jako należącą do klasy pozytywnej, podczas gdy w rzeczywistości należy ona do klasy negatywnej. FPR oblicza się w następujący sposób:

$$FPR = FP / (FP + TN)$$

W przykładowej macierzy konfuzji wynik FPR wyniósłby 20 / (20 + 90) = 18%. Wysoki współczynnik fałszywych wyników pozytywnych może prowadzić do niepotrzebnych kosztów i zasobów wydatkowanych na rozwiązywanie nieistniejących problemów. Dlatego minimalizacja współczynnika fałszywych pozytywów jest ważna w wielu zastosowaniach, takich jak diagnostyka medyczna i wykrywanie oszustw.

Miara F1 jest ważoną średnią precyzji i czułości. Oblicza się ją w następujący sposób:

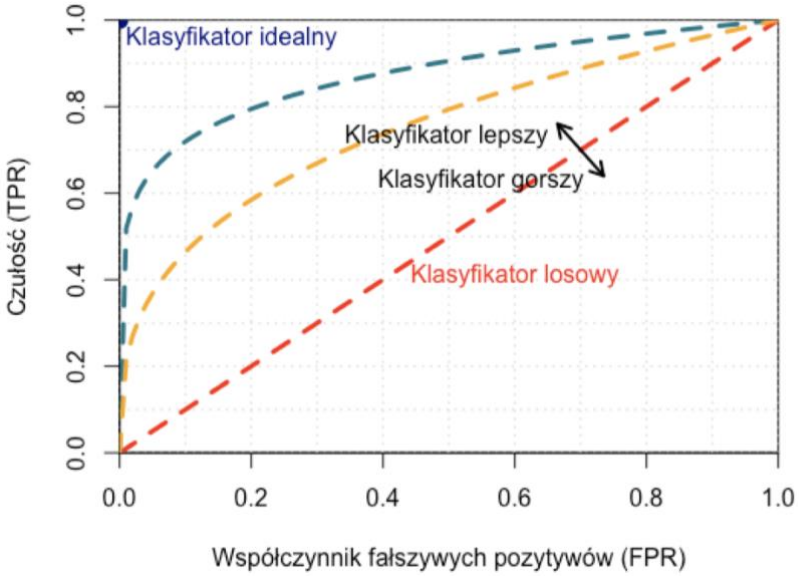
$$F1 = 2 * (\text{Precyzja} * \text{Czułość}) / (\text{Precyzja} + \text{Czułość})$$

W przykładowej macierzy konfuzji wynik F1 wyniósłby 2 * (0,8 * 0,89) / (0,8 + 0,89) = 84%. Wynik F1 jest przydatną metryką, gdy zarówno precyzja, jak i czułość są ważne. Zapewnia on równowagę między tymi dwoma metrykami i jest szczególnie przydatny, gdy klasy w danych są niezrównoważone.

Oprócz macierzy konfuzji istnieją metody oceny modeli uczenia maszynowego wykorzystujące wykresy. Wykresy te są powszechnie stosowane dla problemów klasyfikacji binarnej i obejmują ROC, AUC i krzywą Lift.

Krzywa ROC (ang. *Receiver Operating Characteristic*) jest wykresem czułości (ang. *sensitivity*) względem współczynnika fałszywych pozytywów (false positive rate) dla różnych wartości progowych modelu klasyfikacyjnego. Jest ona wykorzystywana do oceny wydajności binarnego modelu klasyfikacyjnego i stanowi graficzną reprezentację kompromisu pomiędzy czułością i specyficznością.

Przykładowe krzywe ROC



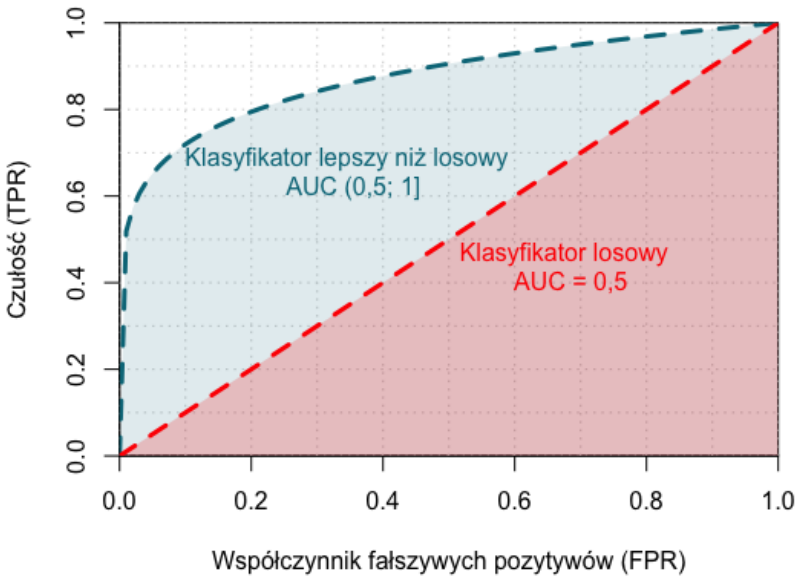
Źr.: opr. wł. na podst. Powers (2007)

Rysunek 3.1. *Rysunek przedstawia przykładowe krzywe ROC dla różnej jakości klasyfikatorów.*

Krzywą ROC tworzy się poprzez wykreślenie Czułości (TPR) na osi y oraz współczynnika fałszywych pozytywów (FPR) na osi x (por. Rysunek 3.1.).

Obszar pod krzywą ROC (ang. *area under curve*) jest metryką mierzącą ogólną wydajność binarnego modelu klasyfikacyjnego. Oblicza się go jako obszar pod krzywą ROC, a jego wartość waha się od 0,5 do 1. Model doskonały miałby AUC równe 1, natomiast model losowy miałby AUC równe 0.

Pola pod krzywymi ROC



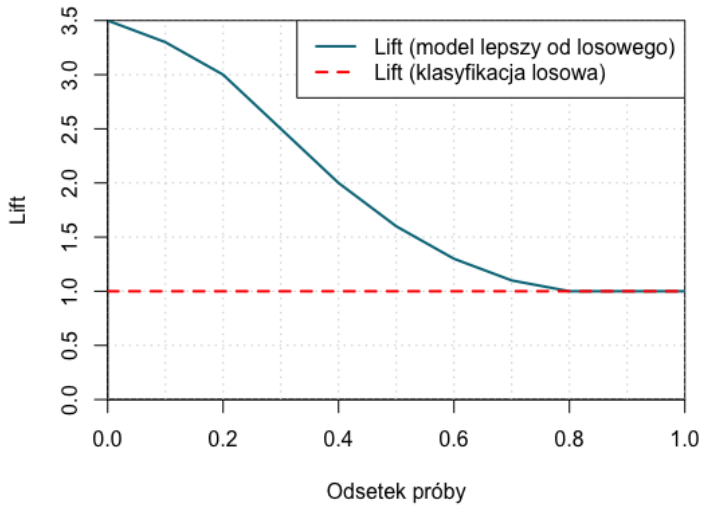
Źr.: opr. wł. na podst. Powers *(2007)*

Rysunek 3.2. *Rysunek przedstawia przykładowe powierzchnie pod krzywymi ROC.*

Pole pod krzywą ROC dla klasyfikatora losowego jest równe ½, podczas gdy dla klasyfikatorów lepszych niż losowe waha się w przedziale od ½ do 1 (por. Rysunek 3.2.).

Krzywa lift pokazuje poprawę modelu klasyfikacji binarnej w porównaniu z przypadkowym przypuszczeniem. Wykres jest tworzony przez podzielenie danych na centyle i obliczenie proporcji pozytywnych klasyfikacji w każdym centylu dla modelu i dla przypadkowego trafienia.

Przykładowa krzywa lift



Źr.: opr. wł. na podst. Powers *(2007)*

Rysunek 3.3. *Rysunek przedstawia przykładową skumulowaną krzywą lift, wykreśloną dla klasyfikatora lepszego niż losowy (kolor zielony).*

Krzywa lift dla poprawnego modelu jest monotonicznie malejąca i w postaci skumulowanej przyjmuje wartości od 1 do odwrotności udziału klasy pozytywnej w zbiorze (por. Rysunek 3.3.)

Krzywa lift jest przydatna w sytuacjach, gdzie koszt fałszywych pozytywów i fałszywych negatywów może być różny. Na przykład, w kampanii marketingowej, fałszywa klasyfikacja pozytywna może prowadzić do zmarnowania zasobów, podczas gdy fałszywa klasyfikacja negatywna może prowadzić do utraty możliwości dotarcia do zainteresowanego klienta. Wykres Lift może pomóc w określeniu optymalnego progu dla modelu klasyfikacyjnego, gdzie lift jest maksymalny.

Oprócz metryk oceny opartych na macierzy błędów oraz wykorzystujących wykresy, istnieją również metryki oceny specyficzne dla problemów regresji. Metryki te są używane do oceny wydajności modeli regresji, które przewidują ciągłe wartość liczbową.

Resztowa suma kwadratów odchyleń (ang. *residual sum of squares – RSS*) mierzy sumę różnic kwadratowych pomiędzy wartościami przewidywanymi a rzeczywistymi. Oblicza się ją w następujący sposób:

$$RSS = \sum(y_i - \hat{y}_i)^2,$$

gdzie $y_i$to wartość rzeczywista, $\hat{y}_i$ to wartość przewidywana, a suma obliczana jest poprzez zsumowanie wszystkich obserwacji w zbiorze testowym. We wzorze RSS różnica między wartością przewidywaną a rzeczywistą jest podniesiona do kwadratu, co oznacza, że większe różnice są karane bardziej niż mniejsze.

Średni błąd bezwzględny (ang. *mean average error – MAE*) mierzy średnią bezwzględną różnicę między wartościami przewidywanymi a rzeczywistymi. Jest on obliczany w następujący sposób:

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i|,$$

gdzie y_i jest wartością rzeczywistą, $\hat{y}_i$ jest wartością przewidywaną, a suma jest obliczana przy uwzględnieniu wszystkich obserwacji zbioru testowego. We wzorze MAE oblicza się bezwzględną różnicę między wartością przewidywaną a rzeczywistą, co oznacza, że znak różnicy jest ignorowany.

Średni błąd kwadratowy (ang. *Root Mean Squared Error*) jest podobny do MAE, ale uwzględnia pierwiastek kwadratowy średniej kwadratowej różnicy między wartościami przewidywanymi a rzeczywistymi. Oblicza się go w następujący sposób:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2},$$

gdzie y_i jest wartością rzeczywistą, $\hat{y}_i$ jest wartością przewidywaną, a suma jest obliczana przy uwzględnieniu wszystkich obserwacji zbioru testowego. Wzór RMSE bardziej penalizuje większe błędy niż mniejsze, ponieważ wykorzystywana jest różnica kwadratowa zamiast różnicy bezwzględnej. Pierwiastek kwadratowy jest wprowadzany na końcu, aby przywrócić RMSE do tej samej skali co zmienna docelowa.

4 Determinanty przynależności do klasy wyższej

Na tym etapie pracy przeprowadzone zostało badanie, mające na celu wyróżnienie determinantów przynależności do klasy wyższej. Podrozdział 4.1. zawiera opis zbioru danych wykorzystanego w badaniu. W Podrozdziale 4.2. przeprowadzana jest analiza danych oraz transformacja zmiennych. Podrozdział 4.3. porównuje modele klasyfikacyjne a Podrozdział 4.4. poddaje je ocenie.

4.1 Zbiór danych

UCI Machine Learning Repository to publiczne repozytorium danych, które zawiera dużą liczbę zbiorów danych przydatnych dla badaczy i praktyków uczenia maszynowego. Jednym ze zbiorów danych dostępnych w repozytorium UCI jest Adult dataset, który zawiera informacje o osobach ze spisu powszechnego w Stanach Zjednoczonych z 1994 roku.

Zbiór danych Adult zawiera ponad 48 000 obserwacji, z których każda odpowiada osobie fizycznej. Zbiór danych zawiera informacje o wieku każdej osoby, klasie zawodowej, edukacji, stanie cywilnym, zawodzie, związku, rasie, płci, zyskach kapitałowych, stratach kapitałowych, godzinach przepracowanych w tygodniu i dochodach. Pole dochodu jest zmienną docelową i jest podzielone na dwie kategorie: dochód mniejszy lub równy $50K rocznie i dochód większy niż $50K rocznie.

Zbiór danych Adult jest powszechnie używany w badaniach nad uczeniem maszynowym, aby przewidzieć dochód osoby na podstawie jej danych demograficznych i informacji o zatrudnieniu.

Jedną z kluczowych cech zbioru danych Adult jest jego wielkość. Z ponad 48 000 przypadków, zbiór danych jest wystarczająco duży, aby być reprezentatywnym dla populacji, a naukowcy mogą trenować modele uczenia maszynowego używając zbioru danych z ufnością. Jednakże, rozmiar zbioru danych może również stanowić wyzwanie. Szkolenie modeli uczenia maszynowego na dużych zbiorach danych może być kosztowne obliczeniowo i może wymagać znacznych zasobów obliczeniowych.

Inną cechą zbioru danych dla dorosłych jest różnorodność danych. Zbiór danych zawiera informacje o osobach z różnych grup demograficznych, w tym wieku, rasy i płci. Ta różnorodność jest ważna dla badań nad uczeniem maszynowym, ponieważ pozwala na zbadanie sposobów, w jaki różne czynniki demograficzne wpływają na dochód danej osoby.

Zbiór danych Adult zawiera również brakujące wartości, co może stanowić wyzwanie dla algorytmów uczenia maszynowego. Naukowcy muszą zdecydować, jak obsłużyć brakujące wartości w zbiorze danych albo przez imputację brakujących wartości, albo przez usunięcie obserwacji z brakującymi wartościami. Decyzja o sposobie postępowania z brakującymi

wartościami może wpłynąć na dokładność modeli uczenia maszynowego wytrenowanych na zbiorze danych.

Zbiór danych Adult został również wykorzystany w badaniach związanych z uczciwością w uczeniu maszynowym. Naukowcy badali sposoby, w jakie algorytmy uczenia maszynowego trenowane na zbiorze danych Adult mogą prowadzić do tendencyjnych prognoz opartych na czynnikach demograficznych, takich jak rasa i płeć. Poprzez identyfikację i zajęcie się tymi uprzedzeniami, badacze mogą poprawić uczciwość i dokładność modeli uczenia maszynowego trenowanych na zbiorze danych.

Podsumowując, zbiór danych Adult z UCI Machine Learning Repository jest cennym zasobem dla badaczy i praktyków uczenia maszynowego. Zbiór danych zawiera dużą liczbę obserwacji, zróżnicowane informacje demograficzne oraz zmienną docelową, która jest przydatna do przewidywania dochodów jednostki. Chociaż zbiór danych stanowi wyzwanie związane z jego wielkością i brakującymi wartościami, naukowcy wykorzystali go do zbadania ważnych pytań badawczych związanych z przewidywaniem dochodu i sprawiedliwością w uczeniu maszynowym. Zestaw danych Adult prawdopodobnie będzie nadal cennym źródłem dla badaczy w przyszłości, ponieważ algorytmy uczenia maszynowego stają się coraz ważniejsze w przewidywaniu i rozumieniu zjawisk społecznych.

Zbiór danych Adult z UCI Machine Learning Repository zawiera łącznie 15 zmiennych, w tym 14 atrybutów i jedną zmienną docelową. Zmienne te dostarczają informacji o cechach demograficznych i finansowych osób, wraz z poziomem ich dochodów.

Age jest ciągłą zmienną numeryczną, która przedstawia wiek respondenta w latach i przyjmuje wartości od 17 do 99.Zmienna workclass jest zmienną kategoryczną, która reprezentuje klasę zawodową. Zmienna fnlwgt jest ciągłą zmienną numeryczną, która reprezentuje ostateczną wagę, czyli liczbę osób w populacji, które dana osoba reprezentuje. Zmienna education jest zmienną kategoryczną, która reprezentuje najwyższy poziom wykształcenia, jaki dana osoba osiągnęła. Zmienna Education.num jest ciągłą zmienną numeryczną, która reprezentuje liczbę lat edukacji, jaką dana osoba ukończyła. Zmienna Marital.status jest zmienną kategoryczną, która reprezentuje stan cywilny. Zmienna occupation jest zmienną kategoryczną, która reprezentuje zawód osoby. Zmienna relationship jest zmienną kategoryczną, która reprezentuje status związku danej osoby. Zmienna race jest zmienną kategoryczną, która przedstawia rasę. Zmienna sex zmienną kategoryczną, która reprezentuje płeć. Różne kategorie w tej zmiennej to Mężczyzna i Kobieta. Zmienna capital.gain jest ciągłą zmienną numeryczną, która reprezentuje zyski kapitałowe, jakie dana osoba osiągnęła. Zmienna capital.loss jest ciągłą zmienną numeryczną, która reprezentuje straty kapitałowe,

które dana osoba zrealizowała. Zmienna hours.per.week jest ciągłą zmienną numeryczną, która reprezentuje liczbę godzin, które respondent przepracowuje w tygodniu. Zmienna country jest zmienną kategoryczną, która reprezentuje kraj pochodzenia danej osoby.

Dochód jest zmienną docelową i jest zmienną kategoryczną, która reprezentuje poziom dochodów jednostki. Różne kategorie w tej zmiennej obejmują „>50K" i „<=50K", wskazując, czy dana osoba zarabia więcej niż 50 000 USD rocznie, czy mniej.

Aby podsumować zmienne w zbiorze danych Adult, możemy stworzyć tabelę z nazwą zmiennej, typem danych i krótkim opisem:

Zmienne zawarte w zbiorze Adult Data

Nazwa	**Opis**	**Typ**
Age	wiek w latach	ciągła
workclass	rodzaj zatrudnienia	kategoryczna
fnlwgt	waga końcowa, liczba osób w populacji	ciągła
education	ukończony poziom edukacji	kategoryczna
education.num	liczba lat edukacji	ciągła
marital.status	stan cywilny	kategoryczna
occupation	zawód	kategoryczna
relationship	rola pełniona w rodzinie	kategoryczna
race	rasa	kategoryczna
sex	płeć	kategoryczna
capital.gain	zyski kapitałowe osiągnięte przez osobę	ciągła
capital.loss	straty kapitałowe poniesione przez osobę	ciągła
hours.per.week	liczba godzin pracy w tygodniu	ciągła
country	kraj pochodzenia	kategoryczna
salary	dochód	kategoryczna

Źr.: opr. wł

Tabela 4.1. *Tabela przedstawia listę zmiennych zawartych w zbiorze Adult Data z UCI Machine Learning Repository wraz z ich opisem oraz kategorią.*

Zbiór Adult Data zawiera 9 zmiennych kategorycznych i 6 ciągłych (por. Tabela 4.1.). Każda ze zmiennych odnosi się do cech społeczno-ekonomicznych respondentów.

4.2 Analiza danych

Tabela 4.2. przedstawia liczbę i odsetek brakujących wartości dla każdej zmiennej w zbiorze. Brakujące wartości są częstym problemem i mogą stanowić wyzwanie dla analizy danych i modelowania.

Braki danych w zbiorze Adult Data

Zmienna	Liczba braków	Odsetek braków
age	0	0
workclass	1836	5,64%
fnlwgt	0	0
education	0	0
education.num	0	0
marital.status	0	0
occupation	1843	5,66%
relationship	0	0
race	0	0
sex	0	0
capital.gain	0	0
capital.loss	0	0
hours.per.week	0	0
country	583	1,79%
salary	0	0

Źr.: opr. wł.

Tabela 4.2. *Jedynie 3 zmienne spośród 15 zawierają braki danych.*

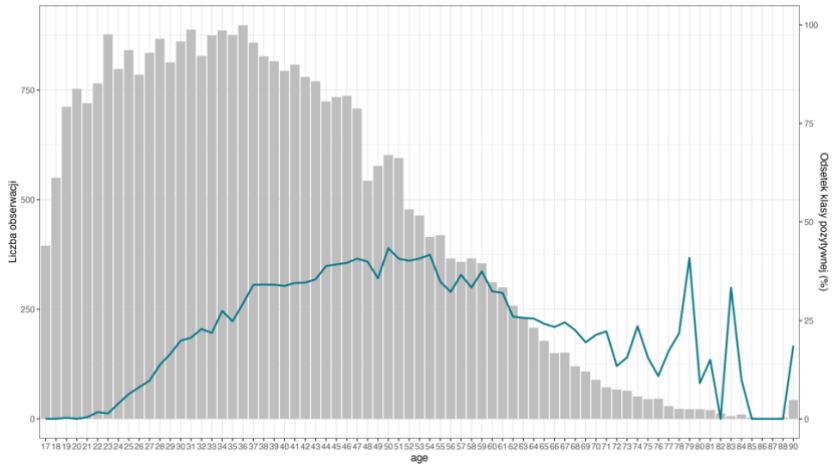
W zbiorze danych Adult, zmienne "workclass", "occupation" i "country" zawierają brakujące wartości. Dla "workclass", 1836 z 32561 obserwacji ma brakujące wartości, co odpowiada 5,64% obserwacji. Podobnie, dla "occupation", 1843 z 32561 obserwacji ma brakujące wartości, co odpowiada 5,66%. W przypadku zmiennej "country", 583 z 32561 obserwacji ma brakujące wartości, co odpowiada 1,79% obserwacji (por. Tabela 4.2.)

Kiedy zmienna ma brakujące wartości, badacze muszą zdecydować, jak poradzić sobie z tymi wartościami. Jednym z podejść jest usunięcie obserwacji, które zawierają brakujące wartości. Jednak to podejście może spowodować utratę cennych informacji i może wpłynąć na wyniki, jeśli brakujące wartości nie są przypadkowe. Innym podejściem jest imputacja brakujących wartości, która polega na zastąpieniu brakujących wartości wartościami szacowanymi na podstawie innych zmiennych w zbiorze danych. Na potrzeby badania opisanego w pracy, braki danych nie zostały usunięte. Sposób postępowania z nimi zostanie opisany na etapie analizy i transformacji zmiennych.

Celem przygotowania danych do procesu uczenia maszynowego, zmienne zostały poddane analizie oraz transformacji, które zostały w tym podrozdziale szczegółowo opisane.

Zmienna age opisująca wiek osoby wyrażony w latach przyjmuje w zbiorze danych wartości ciągłe od 17 do 90.

Zmienna *age* przed transformacją



Źr.: opr. wł.

Rysunek 4.1. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej age przed transformacją.*

Liczba obserwacji jest duża pomiędzy 17 a 60 rokiem życia, po czym zaczyna istotnie spadać. Od 60 roku życia w górę każda grupa stanowi mniej niż 1% obserwacji. Odsetek klasy pozytywnej, czyli osób zarabiających powyżej 50 tysięcy USD wzrasta do około 50 roku życia, po czym zaczyna spadać. Powyżej 70 roku życia występują duże fluktuacje, co wynika z małej liczby obserwacji dla tych grup (por. Rysunek 4.1.)

W celu uzyskania grup, których liczebności stanowią co najmniej 5% ogółu obserwacji, zmienna age została pogrupowana.

Wartości zmiennej *age* przed transformacją i po niej

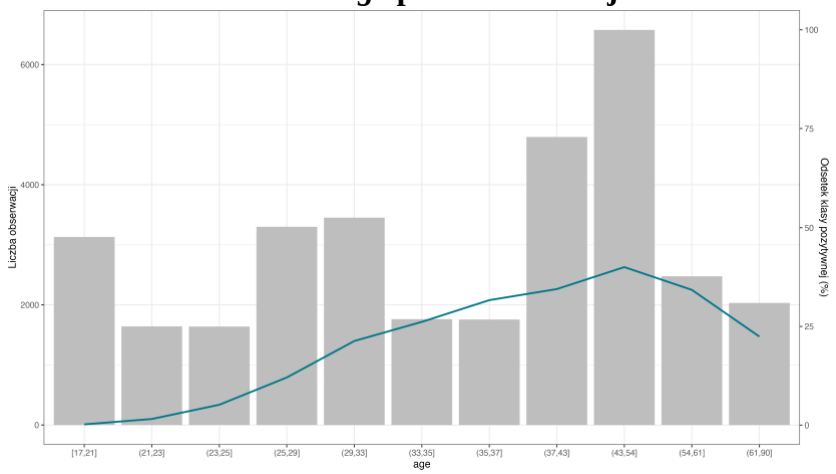
Zmienna	Wartości przed transformacją	Wartości po transformacji
age	Wartości całkowite of 17 do 90	(37,43], (43,54], (25,29], (35,37], (29,33], (21,23], (33,35], (23,25], (54,61], [17,21], (61,90]

Tabela 4.3. *Tabela przedstawia wartości zmiennej age przed transformacją i po niej.*

Do grupowania zmiennych wykorzystany został pakiet smbinning, który pozwala na optymalne grupowanie zmiennych objaśniających. W efekcie zmienna ciągła przetransformowana została na zmienną kategoryczną, przyjmującą 11 poziomów (por. Tabela 4.3.).

W wyniku transformacji danych, udało się osiągnąć zamierzone efekty. Każda grup stanowi co najmniej 5% obserwacji, a fluktuacje zostały usunięte.

Zmienna *age* po transformacji



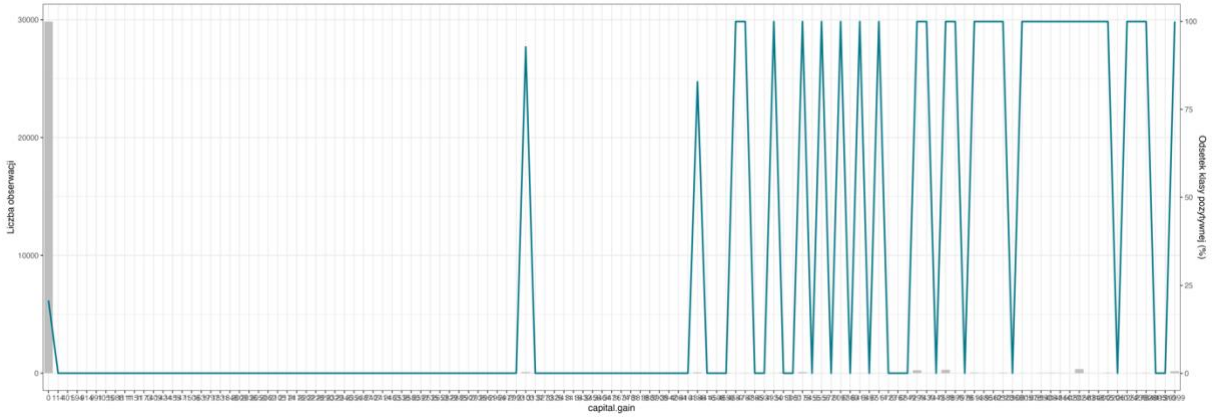
Źr.: opr. wł.

Rysunek 4.2. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej age po transformacji.*

Odsetek klasy pozytywnej jest najniższy dla grupy osób od 17 do 21 roku życia oraz stabilnie rośnie. Maksimum osiąga w przypadku osób pomiędzy 43 a 54 rokiem życia, po czym zaczyna spadać (por. Rysunek 4.2.).

Zmienna capital.gain określa poziom osiągniętych przez respondenta zysków kapitałowych. Przymuje ona wartości całkowite od 0 do 99999.

Zmienna *capital.gain* przed transformacją



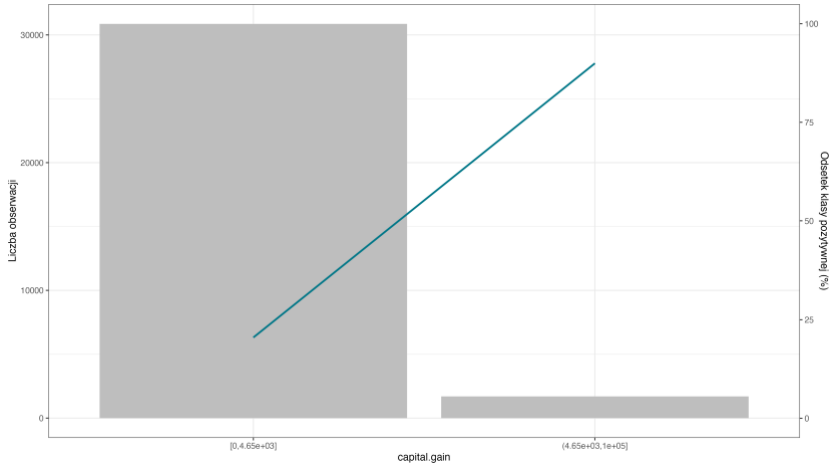
Źr.: opr. wł.

Rysunek 4.3. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej capital.gain przed transformacją.*

Ze względu na bardzo małą liczbę obserwacji w każdej z grup poza capital.gain = 0, w grupach o wyższych zyskach kapitałowych występują bardzo duże fluktuacje klasy pozytywnej (por. Rysunek 4.3.).

Chociaż zyski kapitałowe mogą być przydatne w przewidywaniu dochodu, istnieją istotne powody, dla których konieczne może jest usunięcie zmiennej ze zbioru danych. Dochody wynikają bezpośrednio z zysków kapitałowych a zyski kapitałowe są bezpośrednio związane z dochodami. Wysokie zyski kapitałowe wiążą się w bezpośredni sposób z wysokim poziomem dochodów.

Zmienna *capital.gain* po pomocniczej transformacji



Źr.: opr. wł.

Rysunek 4.4. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej capital.gain po pomocniczej transfromacji.*

W grupie jednostek, których zyski kapitałowe były <= 4650 USD, udział klasy pozytywnej wynosił mniej niż 20%. W grupie, której zyski kapitałowe były > 4650 USD udział klasy wzrastał do 90% (por. Rysunek 4.4.)

Przewidywanie dochodów na podstawie zysków kapitałowych byłoby co prawda trafne, lecz pozbawione sensu praktycznego. Obie zmienne, capital gain i salary, dostarczają informacji dotyczących finansów respondenta. Mając wgląd do zysków kapitałowych jednostki, z dużym prawdopodobieństwem mielibyśmy również informację dotyczącą jej dochodów. Celem modelu jest przewidywanie dochodów jednostki w sytuacjach, gdy informacja ta nie jest dostępna.

Wartości zmiennej *capital gain* przed transformacją i po niej

Zmienna	Wartości przed transformacją	Wartości po transformacji
capital.gain	Wartości całkowite of 0 do 99999	Zmienna usunięta

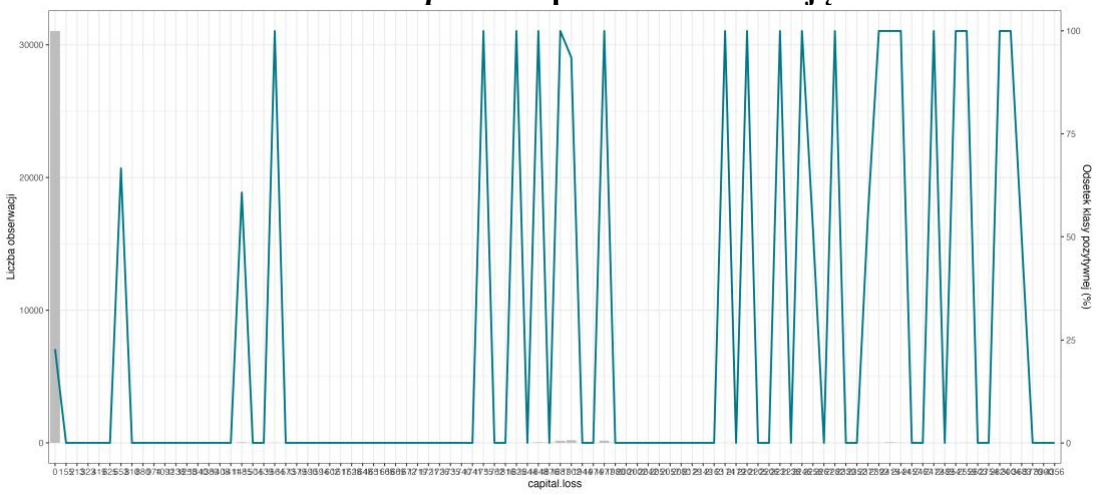
Źr.: opr. wł.

Tabela 4.4. *Tabela przedstawia wartości zmiennej capital.gain przed transformacją i po niej.*

Zmienna capital.gain nie została uwzględniona podczas tworzenia modeli (por. Tabela 4.4.)

Zmienna capital.loss określa poziom ponoszonych przez respondenta strat kapitałowych. Przymuje ona wartości całkowite od 0 do 4356.

Zmienna *capital.loss* przed transformacją



Źr.: opr. wł.

Rysunek 4.5. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej capital.loss przed transformacją.*

Ze względu na małą liczbę obserwacji w każdej z grup poza capital.loss = 0, w grupach o wyższych zyskach kapitałowych występują bardzo duże fluktuacje klasy pozytywnej (por. Rysunek 4.5.)

Tak jak zmienna capital.gain, zmienna capital.loss zawiera informacje, których uwzględnienie w modelu uczyniłoby go niepraktycznym.

Wartości zmiennej *capital.loss* przed transformacją i po niej

Zmienna	Wartości przed transformacją	Wartości po transformacji
capital.loss	Wartości całkowite of 0 do 4356	Zmienna usunięta

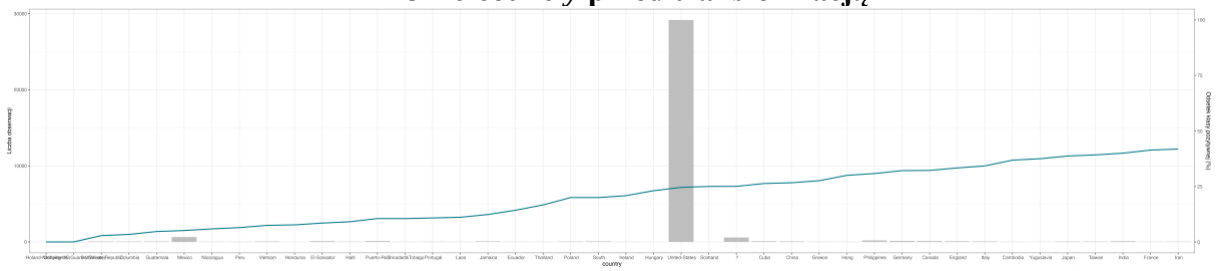
Źr.: opr. wł.

Tabela 4.5. *Tabela przedstawia wartości zmiennej capital.loss przed transformacją i po niej.*

Zmienna capital.loss nie została uwzględniona na dalszych etapach badania (por. Tabela 4.5.).

Zmienna country zawiera 42 unikalne wartości, które wskazują na kraj pochodzenia respondenta.

Zmienna *country* przed transformacją



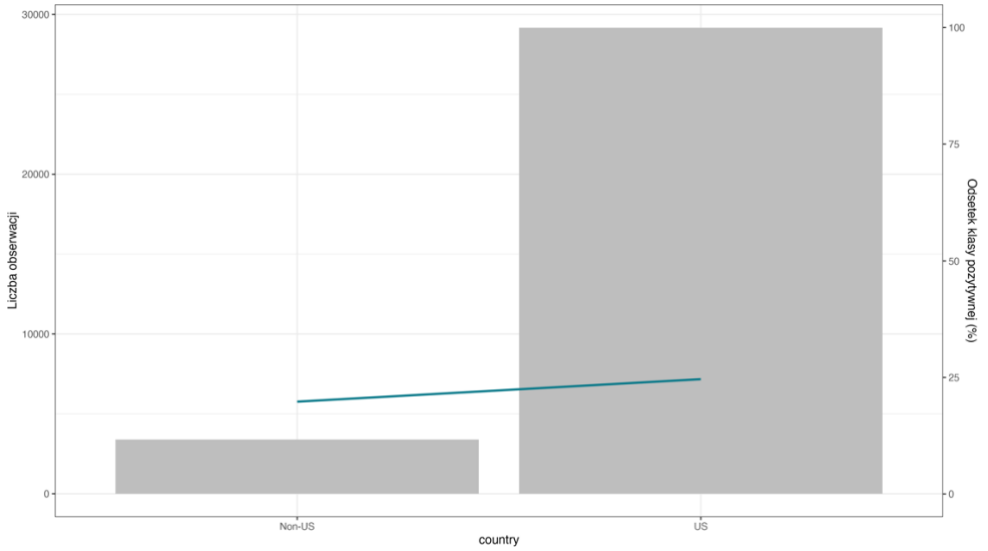
Źr.: opr. wł.

Rysunek 4.6. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej country przed transformacją.*

Z racji na fakt, iż zbór danych opisuje populację Stanów Zjednoczonych, grupa „United-States" zawiera 90% ogółu obserwacji. Liczebność żadnej z pozostałych grup nie przekracza 1% wszystkich obserwacji (por. Rysunek 4.6.).

Celem zagwarantowania w przypadku każdej z grup liczebności powyżej 5% oraz umożliwienia interpretacji merytorycznej, zmienna country została podzielona na 2 grupy: „US" i „non-US". Taki podział umożliwi porównanie poziomu dochodów z uwzględnieniem tła migracyjnego respondentów.

Zmienna *country* po transformacji



Źr.: opr. wł.

Rysunek 4.7. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej country po transformacji.*

Powstałe w wyniku transformacji grupy „Non-US", „US" liczą odpowiednio 10 i 90 procent obserwacji. Wśród respondentów należących do grupy „Non-US" odsetek klasy pozytywnej

wynosi 20%, a wśród osób respondentów należących do grupy „US” – 25% (por. Rysunek 4.7.).

Wartości zmiennej *country* przed transformacją i po niej

Zmienna	Wartości przed transformacją	Wartości po transformacji
country	United-States	US
	Cuba, Jamaica, India, ?, Mexico, South Korea, Puerto-Rico, Honduras, England, Canada, Germany, Iran, Philippines, Italy, Poland, Columbia, Cambodia, Thailand, Ecuador, Laos, Taiwan, Haiti, Portugal, Dominican-Republic, El-Salvador, France, Guatemala, China, Japan, Yugoslavia, Peru, Outlying-US(Guam-USVI-etc), Scotland, Trinadad&Tobago, Greece, Nicaragua, Vietnam, Hong, Ireland, Hungary, Holand-Netherlands	Non-US

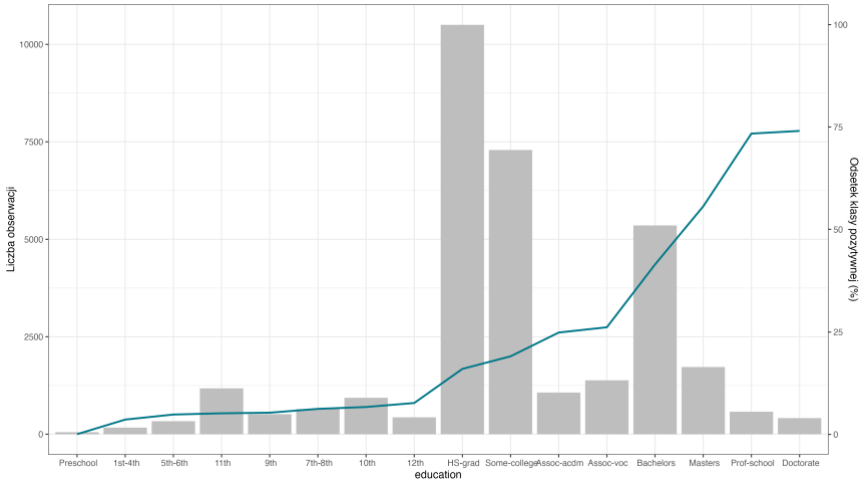
Źr.: opr. wł.

Tabela 4.6. *Tabela przedstawia wartości zmiennej country przed transformacją i po niej.*

Powstała w wyniku transformacji zmienna country przyjmuje tylko 2 poziomy wobec 42 przed transformacją (por. Tabela 4.6.).

Zmienna porządkowa education przyjmuje 16 wartości odpowiadających kolejnym etapom edukacji. Liczebności poszczególnych grup są bardzo zróżnicowane: od mniej niż 1% w przypadku osób, które nie ukończyły szkoły podstawowej, do 32% osób z wykształceniem średnim.

Zmienna *education* przed transformacją



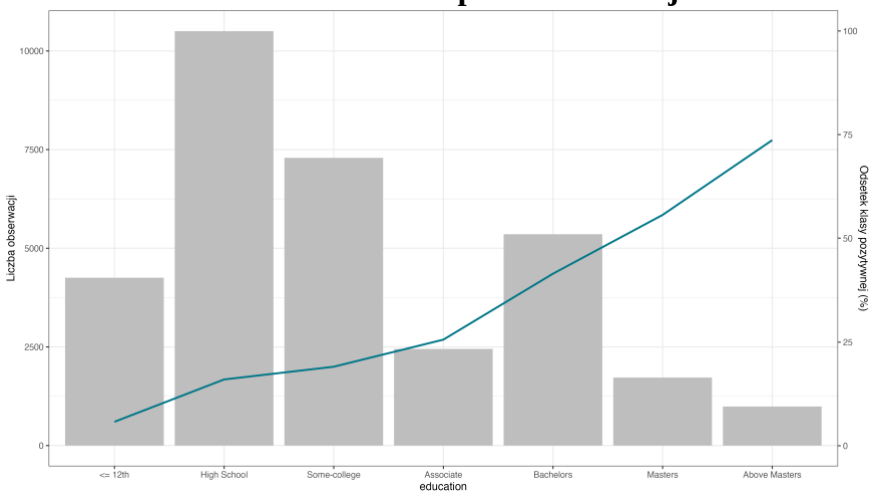
Źr.: opr. wł.

Rysunek 4.8. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej education przed transformacją.*

Odsetek respondentów zarabiających powyżej 50 tysięcy USD rośnie wraz z każdym kolejnym poziomem edukacji. Największe wzrosty zauważyć można w przypadku wykształcenia wyższego (por. Rys 4.8.). Potwierdza to tezę o edukacji będącej jednym z kluczowych determinantów wysokich zarobków.

Ze względu na małą liczebność niektórych kategorii, zmienna została poddana transformacji.

Zmienna *education* po transformacji



Źr.: opr. wł.

Rysunek 4.9. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej educaation przed transformacją.*

Powstałe w wyniku transformacji kolejne poziomy zmiennej education wskazują na wyraźnie dodatnią zależność pomiędzy wykształceniem a zarobkami (por. Rysunek 4.9.)

Wartości zmiennej *education* przed transformacją i po niej

Zmienna	Wartości przed transformacją	Wartości po transformacji
education	Preschool, 1^{st} - 4^{th}, 5^{th} - 6^{th}, 7^{th} - 8^{th}, 9^{th}, 10^{th}, 11^{th}, 12^{th}	<= 12th
	HS-grad	High School
	Some-college	Some-college
	Bachelors	Bachelors
	Masters	Masters
	Prof-school, Doctorate	Above Masters

Źr.: opr. wł.

Tabela 4.7. *Tabela przedstawia wartości zmiennej education przed transformacją i po niej.*

W wyniku transformacji połączone zostały wszystkie jednostki, dla których wartość zmiennej education jest mniejsza niż 12th oraz osoby z wykształceniem wyższym niż magisterskie (por. Tabela 4.7.)

Zmienna fnlwgt w zbiorze danych Adult z UCI Machine Learning Repository oznacza wagę końcową. Jest to waga przypisana do każdej obserwacji w zbiorze danych, która reprezentuje liczbę osób w populacji, którą dana obserwacja reprezentuje. Celem użycia zmiennej fnlwgt jest skorygowanie faktu, że zbiór danych nie był prostą próbą losową populacji, ale raczej złożoną próbą badawczą. Aby móc wnioskować o całej populacji na

podstawie próby badawczej, zmienna wagi końcowej jest używana do skorygowania faktu, że niektóre osoby są nadreprezentowane lub niedoreprezentowane w próbie.

To, czy uwzględnić zmienną "fnlwgt" w drzewie decyzyjnym, zależy od pytania badawczego i celu analizy. Jeśli celem analizy jest wnioskowanie o populacji, wówczas ważne jest uwzględnienie złożonego procesu badania i zmienna wagi końcowej powinna być włączona do analizy. W tym przypadku właściwe byłoby włączenie zmiennej "fnlwgt" do drzewa decyzyjnego w celu uwzględnienia złożonego procesu badania. Jednakże, jeśli celem analizy jest zbudowanie drzewa decyzyjnego dla celów predykcyjnych, a celem końcowym nie jest wnioskowanie o populacji, wówczas włączenie zmiennej "fnlwgt" do drzewa decyzyjnego może nie być konieczne. W tym przypadku ważniejsze byłoby skupienie się na innych zmiennych, które są bardziej predykcyjne dla interesującego nas wyniku.

Wartości zmiennej *fnlwgt* przed transformacją i po niej

Zmienna	Wartości przed transformacją	Wartości po transformacji
fnlwgt	wartości ciągłe od 12285 do 1484705	wartości ciągłe od 12285 do 1484705

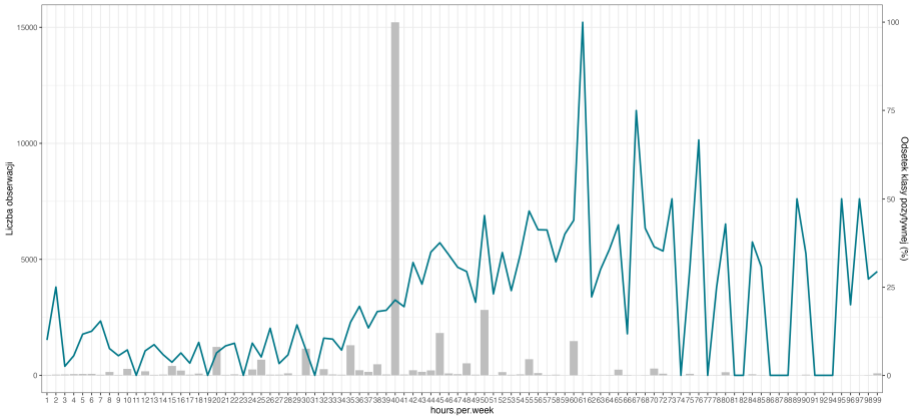
Źr.: opr. wł.

Tabela 4.8. *Tabela przedstawia wartości zmiennej fnlwgt przed transformacją i po niej.*

W trakcie badania stworzone zostaną zarówno modele wykorzystujące „fnlwgt", jak i te, które ją pomijają. Jak na razie, nie zostanie ona usunięta ze zbioru (por. Tabela 4.8.)

Zmienna hours.per.week oznacza liczbę godzin, które jednostka pracuje w tygodniu.

Zmienna *hours.per.week* przed transformacją



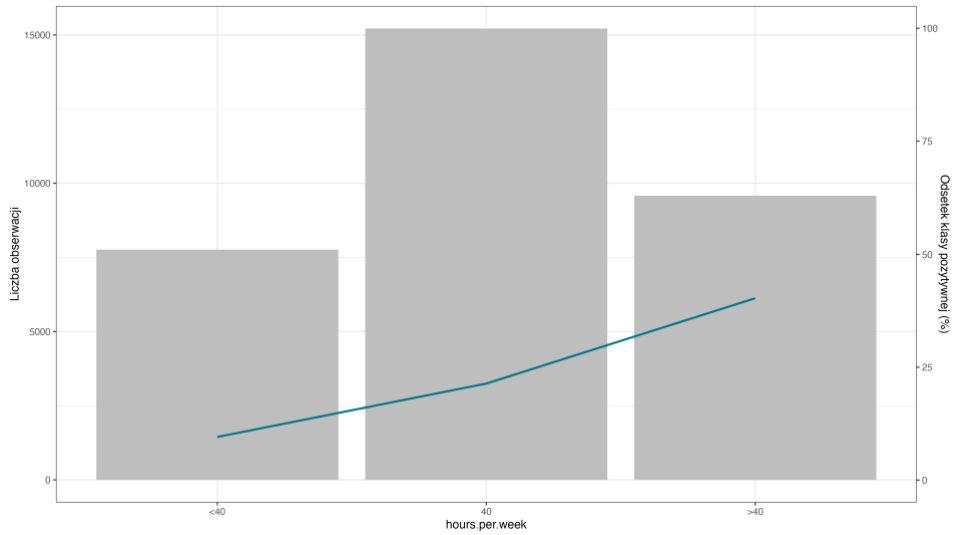
Źr.: opr. wł.

Rysunek 4.10. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej* hours.per.week *przed transformacją.*

Większość osób pracuje 40 godzin tygodniowo (8h dziennie), co stanowi przyjętą normę społeczną. Ze względu na niewielką liczbę obserwacji w każdej z klas, występują bardzo duże fluktuacje udziału klasy pozytywnej (por. Rysunek 4.10.)

Celem uzyskania dostatecznie licznych kategorii oraz zapewnienia monotoniczności udziału klasy pozytywnej, wartości zmiennej hours.per.week zostały pogrupowane.

Zmienna *hours.per.week* po transformacji



Źr.: opr. wł.

Rysunek 4.11. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej* hours.per.week *po transformacji.*

Zgodnie z oczekiwaniami, osoby pracujące na część etatu (<40 h) najrzadziej zarabiają powyżej 50 tysięcy USD, a dochód rośnie wraz z liczbą przepracowanych godzin. Wśród osób pracujących na więcej niż cały etat (>40 h) odsetek klasy pozytywnej rośnie do 40% (por. Rysunek 4.11.).

Wartości zmiennej *hours.per.week* przed transformacją i po niej

Zmienna	Wartości przed transformacją	Wartości po transformacji
hours.per.week	Wartości ciągłe od 1 do 39	<40
	40	40
	Wartości ciągłe od 41 do 99	>40

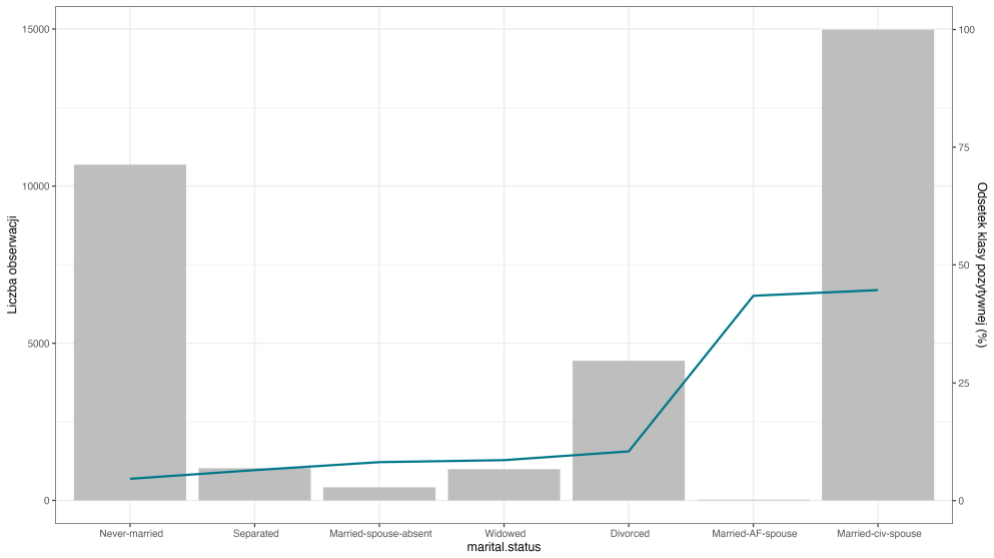
Źr.: opr. wł.

Tabela 4.9. *Tabela przedstawia wartości zmiennej hours.per.week przed transformacją i po niej.*

W wyniku transformacji liczba poziomów zmiennej hours.per.week zmalała z 99 do 3 (por. Tabela 4.9.)

Zmienna kategoryczna "marital.status" przyjmuje 7 różnych wartości i opisuje stan cywilny respondentów.

Zmienna *marital.status* przed transformacją



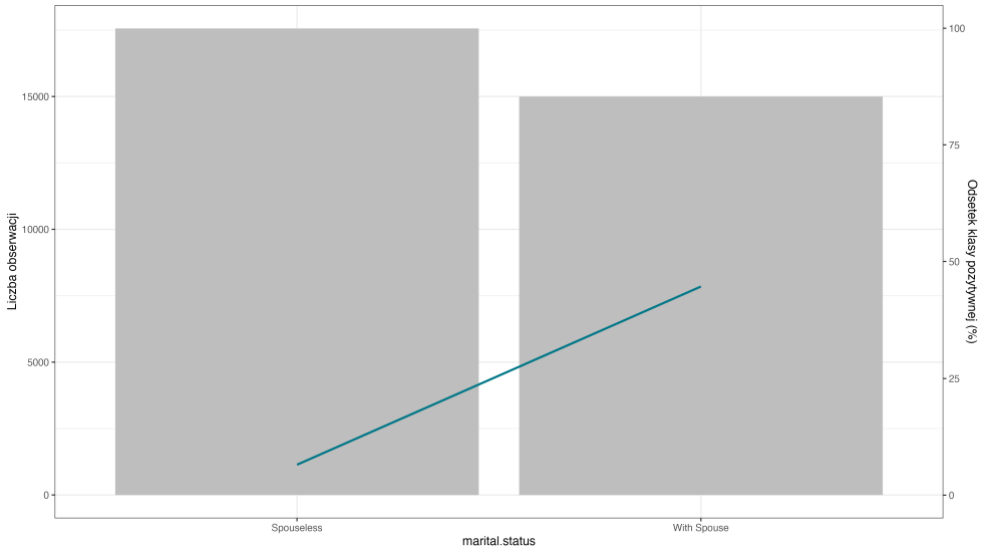
Źr.: opr. wł.

Rysunek 4.12. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej* marital.status *przed transformacją.*

Zauważyć można, że respondenci posiadający partnerów częściej niż osoby samotne zarabiają więcej niż 50 tysięcy USD (por. Rysunek 4.12.)

W celu uzyskania grup o dostatecznej liczebności, zmienna została poddana transformacji.

Zmienna *marital.status* po transformacji



Źr.: opr. wł.

Rysunek 4.13. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej* marital.status *po transformacji.*

Pogrupowane zmienne wyraźnie wskazują, że osoby żyjące z małżonkiem znacznie częściej zarabiają powyżej 50 tysięcy USD. Odsetek klasy pozytywnej dla kategorii „With Spouse" wynosi 46%, wobec jedynie 6% dla kategorii „Spouseless" (por. Rysunek 4.13.).

Wartości zmiennej *marital.status* przed transformacją i po niej

Zmienna	**Wartości przed transformacją**	**Wartości po transformacji**
marital.status	Widowed, Divorced, Separated, Married-spouse-absent, Never-married	Spouseless
	Married-AF-spouse, Married-civ-spouse	With Spouse

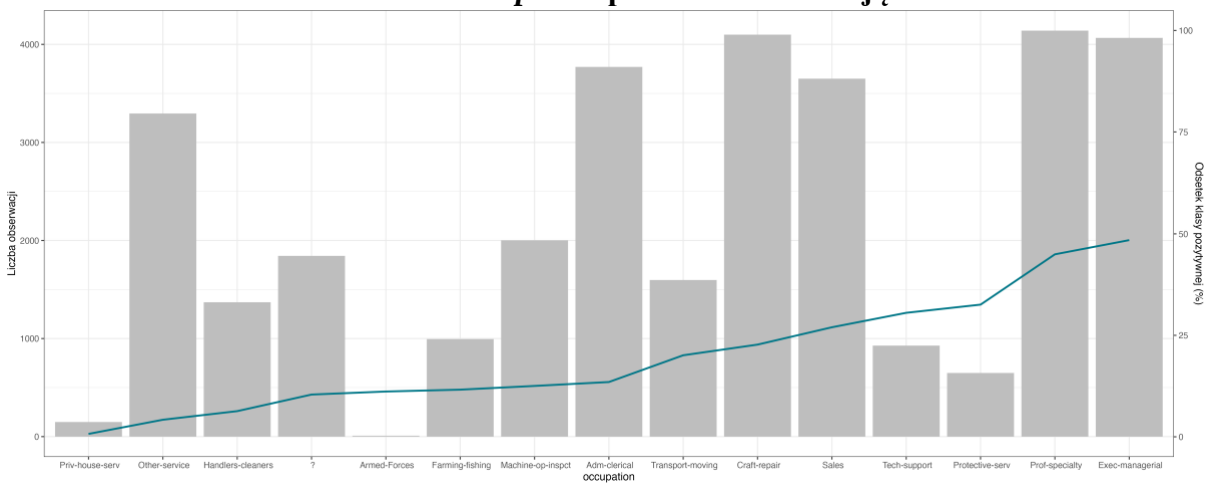
Źr.: opr. wł.

Tabela 4.10. *Tabela przedstawia wartości zmiennej marital.status przed transformacją i po niej*

W wyniku transformacji wszystkie osoby żyjące bez partnera zostały przypisane do kategorii „spouseless", podczas gdy osoby żyjące z partnerem zostały przypisane do kategporii „With Spouse" (por. Tabela 4.10.)

Zmienna occupation przyjmuje 15 wartości i opisuje zawód respondenta.

Zmienna *occupation* przed transformacją



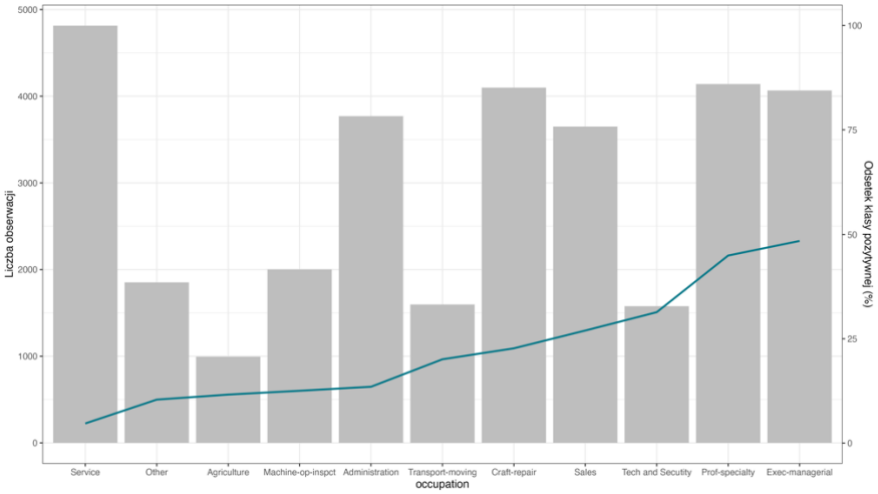
Źr.: opr. wł.

Rysunek 4.14. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej* occupation *przed transformacją.*

Odsetek klasy pozytywnej dla zmiennej „occupation", oznaczającej zawód respondenta, istotnie różni się między grupami. Poziom dochodu zależy od rodzaju zatrudnienia (por. Rysunek 4.14.).

Ze względu na dużą liczbę różnych kategorii a także niską liczebność części z nich, zmienne zostały pogrupowane.

Zmienna *occupation* po transformacji



Źr.: opr. wł.

Rysunek 4.15. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej* occupation *po transformacji.*

Najwięcej respondentów w klasie pozytywnej jest wśród specjalistów oraz menedżerów (odpowiednio 45% i 48%), najmniej wśród osób pracujących w usługach (4,6%) (por. Rysunek 4.15.)

Wartości zmiennej *occupation* przed transformacją i po niej

Zmienna	Wartości przed transformacją	Wartości po transformacji
occupation	Priv-house-serv, Handlers-cleaners, Other-service	Service
	Armed-forces, ?	Other
	Farming-fishing	Agriculture
	Machine-op-inspct	Machine-op-inspct
	Adm-clerical	Administration
	Transport-moving	Transport-moving
	Crafr-repair	Craft-repair
	Sales	Sales
	Tech-support, Protective-Serv	Tech and Security
	Prof-specialty	Prof-specialty
	Exec-managerial	Exec-managerial

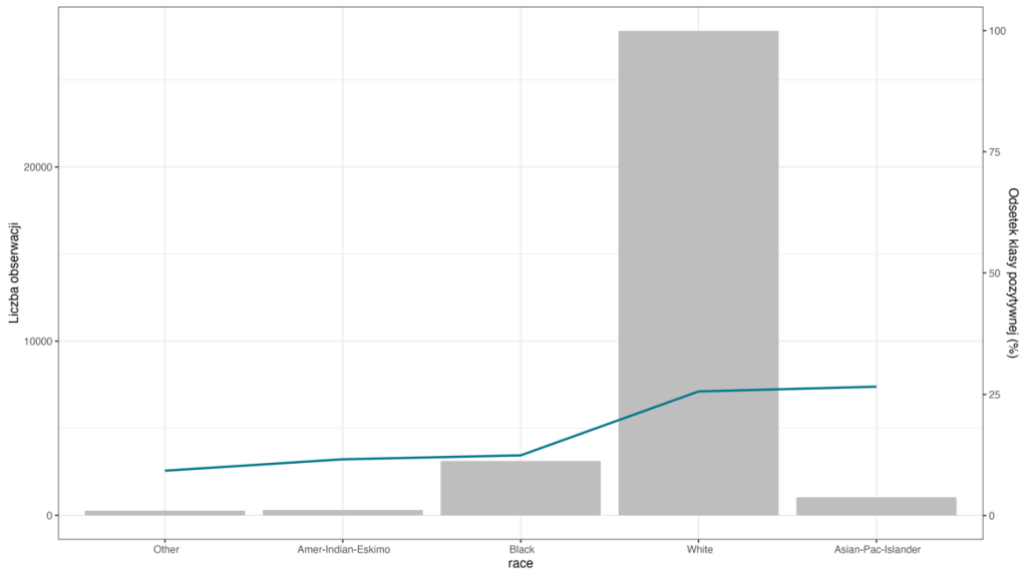
Źr.: opr. wł.

Tabela 4.11. *Tabela przedstawia wartości zmiennej occupation przed transformacją i po niej*

W wyniku transformacji zmiennych, ze względu za zbliżony udział klasy pozytywnej, zmienna Armed-forces oraz braki danych zostały przypisane do kategorii Other. Zawody usługowe zostały przypisane do klasy Service, a klasy Tech-support oraz Protective-Serv zostały pogrupowane do Tech and Security (por. Tabela 4.11.).

Zmienna race zawiera 5 kategorii nominalnych oznaczających rasę, do której należy respondent.

Zmienna *race* przed transformacją



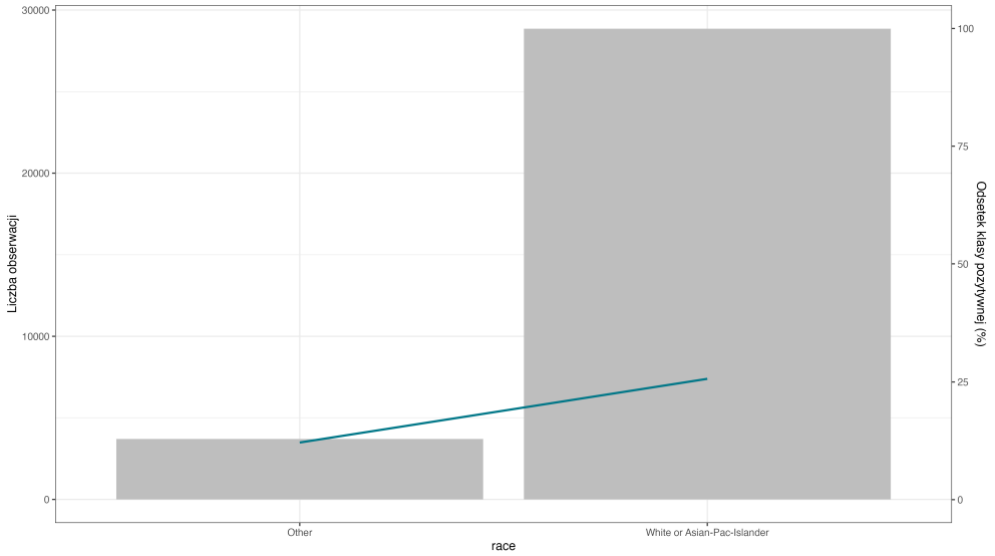
Źr.: opr. wł.

Rysunek 4.16. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej race przed transformacją.*

Aż 85% obserwacji należy do kategorii „White”. Istnieje zależność pomiędzy rasą a zarobkami. Osoby białe oraz pochdzące z rejonu Azji i Pacyfiku zarabiają więcej od przedstawicieli innych ras (por. Rysunek 4.16.)

Ze względu na zbliżone wartości odsetka klasy pomiędzy poszczególnymi kategoriami, a także małą liczebność części z nich zmienne zostały poddane transformacji.

Zmienna *race* po transformacji



Źr.: opr. wł.

Rysunek 4.17. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej race po transformacji.*

Wśród przedstawicieli kategorii „White or Asian-Pac-Islander" udział klasy pozytywnej wynosi 25%, a wśród przedstawicieli pozostałych ras 12%. (por. Rysunek 4.17.).

Wartości zmiennej *race* przed transformacją i po niej

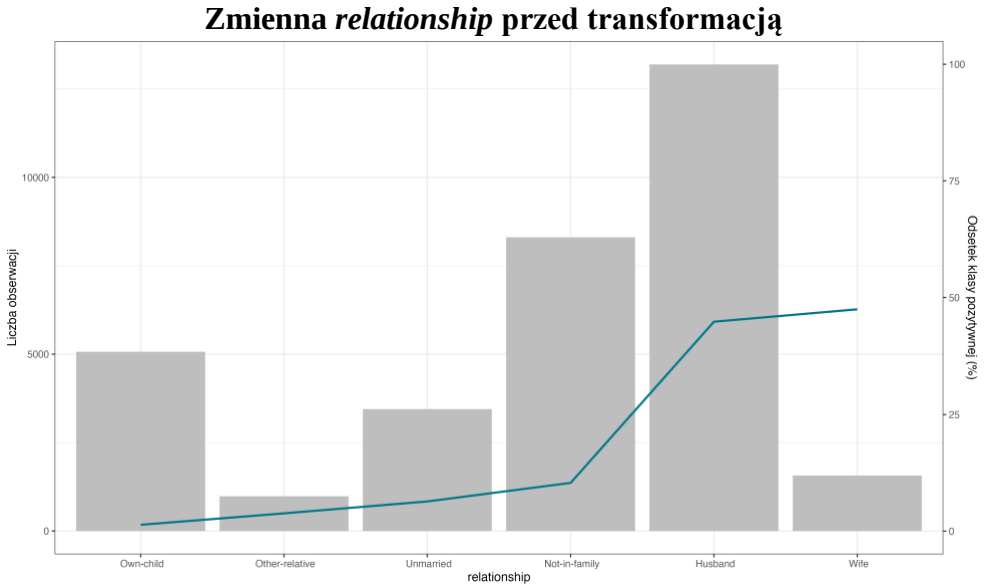
Zmienna	Wartości przed transformacją	Wartości po transformacji
race	White, Asian-Pac-Islander	White or Asian-Pac-Islander
	Black, Amer-Indian-Eskimo, Other	Other

Źr.: opr. wł.

Tabela 4.12. *Tabela przedstawia wartości zmiennej race przed transformacją i po niej*

Zmienne White oraz Asian-Pac-Islander zostały zmapowane do White or Asian-Pac-Islander a zmienne Black, Amer-Indian-Eskimo, Other do Other (por. Tabela 4.12.).

Zmienna kategoryczna "relationship" opisuje rodzaj relacji łączących daną osobę z innymi członkami jej gospodarstwa domowego. Przyjmuje ona 6 wartości.



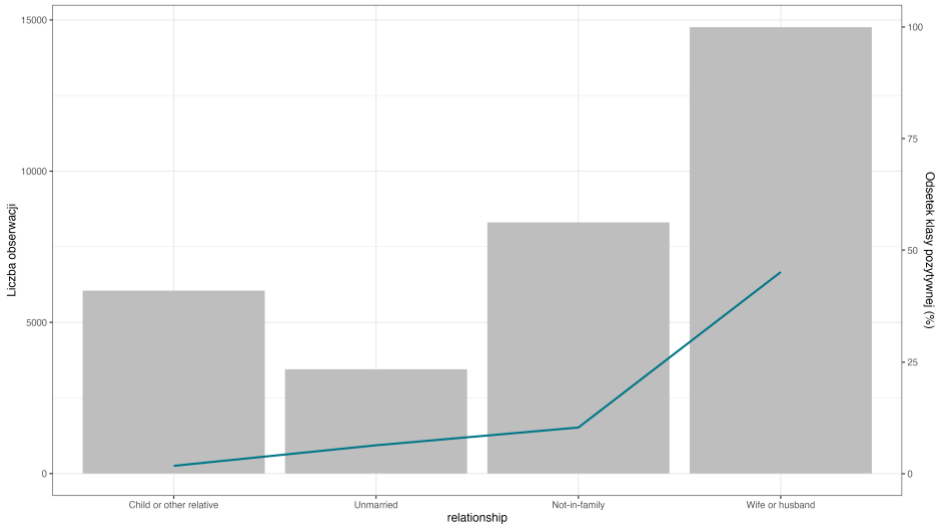
Źr.: opr. wł.

Rysunek 4.18. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej relationship przed transformacją.*

Ponownie, wśród osób będących w związku małżeńskim dochody najczęściej przekraczają 50 tysięcy USD. Wśród osób samotnych dzieje się to znacznie rzadziej. Najrzadziej klasę pozytywną zaobserwować można wśród osób pełniących rolę dziecka lub innego członka rodziny (por. Rysunek 4.18.).

Zmienne zostały pogrupowane ze względu na odsetek klasy pozytywnej oraz aspekty merytoryczne.

Zmienna *relationship* po transformacji



Źr.: opr. wł.

Rysunek 4.19. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej relationship po transformacji.*

I tak, wśród osób tworzących związek małżeński odsetek klasy pozytywnej wynosi 45%, podczas gdy wśród osób pełniących rolę dziecka lub innego członka rodziny 1,7% (por. Rysunek 4.19.)

Wartości zmiennej *relationship* przed transformacją i po niej

Zmienna	Wartości przed transformacją	Wartości po transformacji
relationship	Own-child, Other-relative	Child or other relative
	Unmarried	Unmarried
	Not-in-family	Not-in-family
	Husband, Wife	Wife or husband

Źr.: opr. wł.

Tabela 4.13. *Tabela przedstawia wartości zmiennej relationship przed transformacją i po niej*

W wyniku transformacji połączone zostały kategorie Own-child i Other-relative oraz Husband, Wife (por. Tabela 4.13.).

Zmienna salary to zmienna objaśniana, wskazująca na poziom zarobków respondenta. Dzieli ona respondentów na dwie grupy: zarabiających powyżej 50 tysięcy USD (klasa pozytywna) i poniżej.

Wartości zmiennej *salary* przed transformacją i po niej

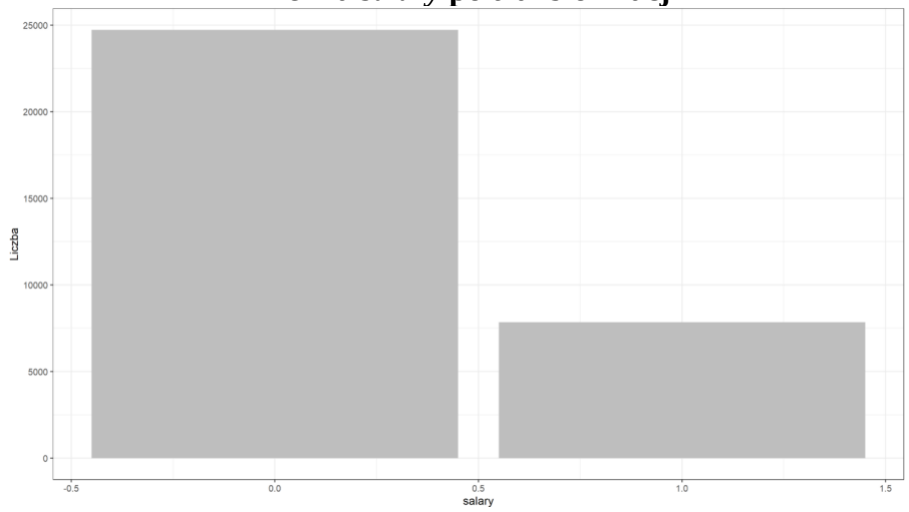
Zmienna	Wartości przed transformacją	Wartości po transformacji
salary	<=50K	0
	>50K	1

Źr.: opr. wł.

Tabela 4.14. *Tabela przedstawia wartości zmiennej salary przed transformacją i po niej*

Zmienna salary została poddana binaryzacji, aby lepiej przystosować ją do procesów uczenia maszynowego (por. Tabela 4.14.).

Zmienna *salary* po transformacji



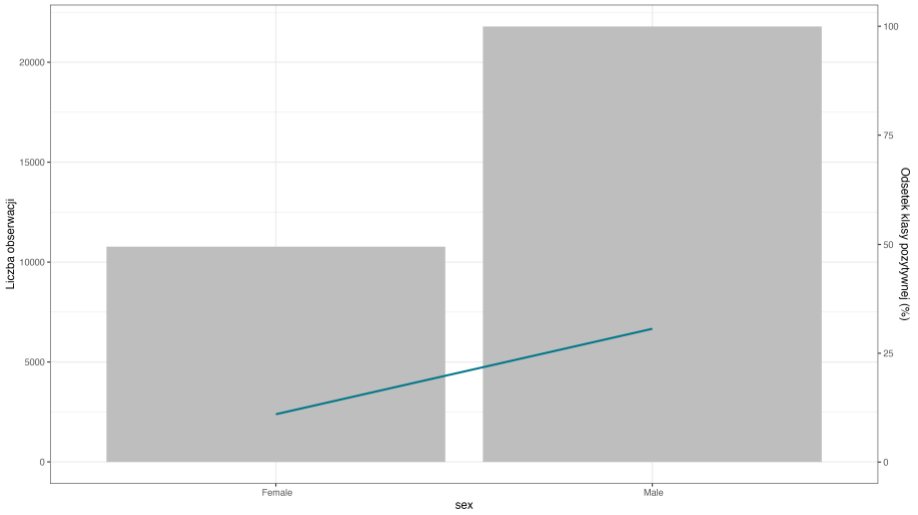
Źr.: opr. wł.

Rysunek 4.20. *Rysunek przedstawia liczbę obserwacji dla każdego z poziomów zmiennej salary po transformacji.*

Do klasy pozytywnej warto około 24% respondentów. Do klasy negatywnej około 76% (por. Rysunek 4.20.)

Zmienna sex wskazuje na płeć respondenta i przyjmuje dwie wartości: Male lub Female, oznaczające odpowiednio mężczyzn i kobiety.

Zmienna *sex*



Źr.: opr. wł.

Rysunek 4.21. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej sex.*

Zbiór danych składa się w 67% z mężczyzn oraz w 33% z kobiet. Wśród kobiet udział klasy pozytywnej wynosi 10%. W przypadku mężczyzn jest on istotnie wyższy i wynosi 30% (por. Rysunek 4.21.).

Wartości zmiennej *sex* przed transformacją i po niej

Zmienna	Wartości przed transformacją	Wartości po transformacji
Sex	Male	Male
	Female	Female

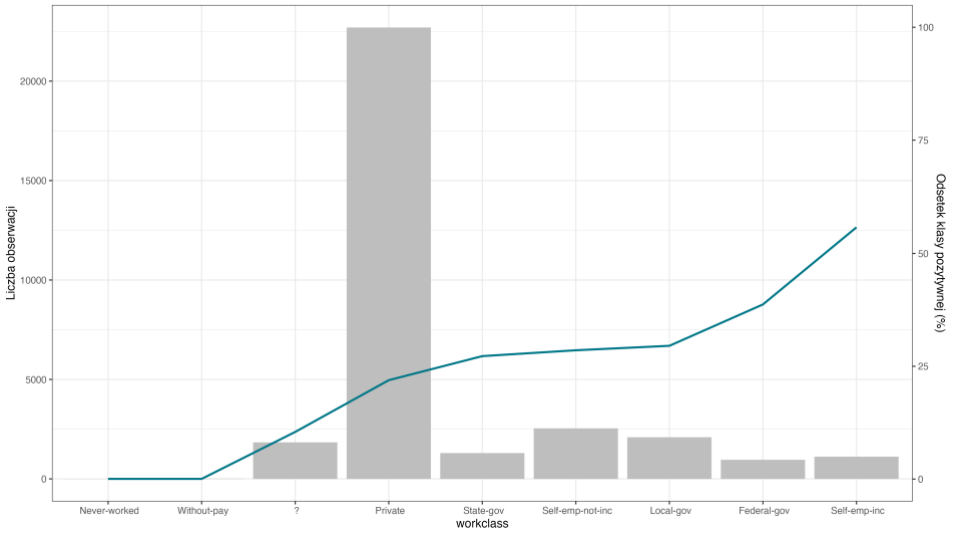
Źr.: opr. wł.

Tabela 4.15. *Tabela przedstawia wartości zmiennej sex przed transformacją i po niej*

Zmienna Sex nie została poddana żadnym transformacjom (por. Tabela 4.15.).

Zmienna workclass wskazuje na rodzaj zatrudnienia respondenta. 70% respondentów zatrudnionych jest w prywatnych przedsiębiorstwach (private). 8% respondentów prowadzi działalność gospodarczą nieposiadającą osobowości prawnej (self-emp-not-inc), a 3,4% prowadzi działalność posiadającą osobowość prawną. W sektorze publicznym (State-gov, Local-gov lub Federal-gov) pracuje w sumie 13% respondentów.

Zmienna *workclass* przed transformacją



Źr.: opr. wł.

Rysunek 4.22. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej workclass przed transformacją.*

Najwyższy udział klasy pozytywnej, 55%, występuje w przypadku respondentów prowadzących działalność gospodarczą, która posiada osobowość prawną (por. Rysunek 4.22.).

Wartości zmiennej *workclass* przed transformacją i po niej

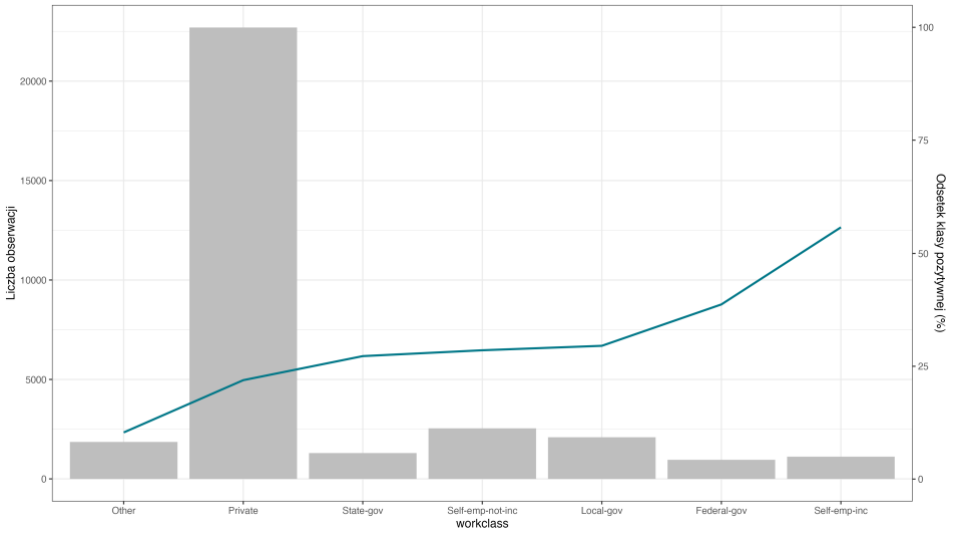
Zmienna	Wartości przed transformacją	Wartości po transformacji
workclass	State-gov	State-gov
	Self-emp-not-inc	Self-emp-not-inc
	Private	Private
	Federal-gov	Federal-gov
	Local-gov	Local-gov
	Self-emp-inc	Self-emp-inc
	Without-pay, Never-worked, ?	Other

Źr.: opr. wł.

Tabela 4.16. *Tabela przedstawia wartości zmiennej workclass przed transformacją i po niej*

Ze względów merytorycznych oraz z racji na zbliżony udział klasy pozytywnej, zmienne Without-pay, Never-worked oraz braki danych, zostały przyporządkowane do zmiennej Other. Pozostałe zmienne nie zostały poddane transformacji (por. Tabela 4.16.).

Zmienna *workclass* po transformacji



Źr.: opr. wł.

Rysunek 4.23. *Rysunek przedstawia liczbę obserwacji oraz odsetek klasy pozytywnej dla każdego z poziomów zmiennej workclass po transformacji.*

Zmienna po transformacji również wskazuje na istotną zależność pomiędzy rodzajem zatrudnienia a zarobkami (por. Rysunek 4.23.)

4.3 Porównanie modeli identyfikujących

Prognozowaniu wysokich zarobków posłużyło 5 modeli uczenia maszynowego, w tym: 2 drzewa decyzyjne, 2 modele regresji oraz las losowy. W podrozdziale 4.3 zostaną one omówione, a następnie poddane analizie, przy pomocy metryk opartych na macierzy błędów, krzywych ROC, AUC i Lift, a także metryk oceniających modele regresji. Ostatecznie, wybrany zostanie najskuteczniejszy model prognozujący.

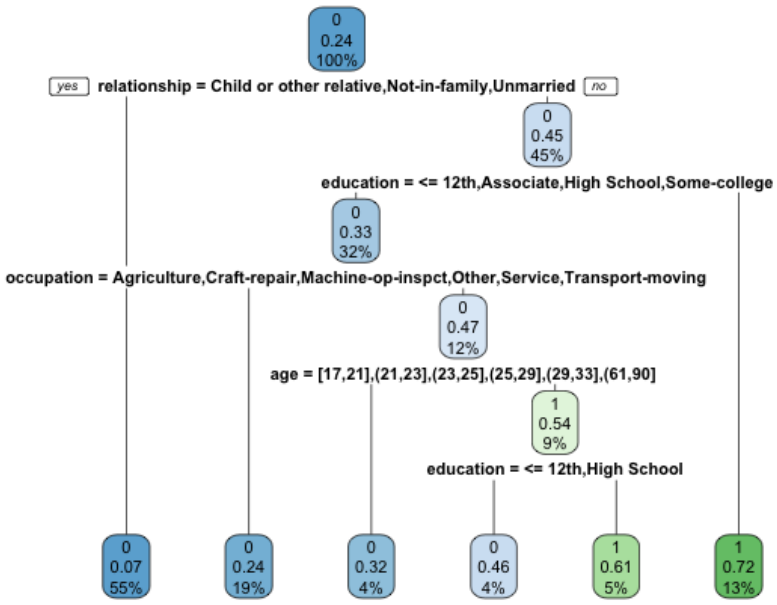
Drzewa decyzyjne zostały stworzone w R przy użyciu pakietu rpart(), którego działanie opisano na wcześniejszym etapie pracy. Podczas budowy drzew pod uwagę wzięte zostały 4 kombinacje:

1. Drzewo z domyślnymi ustawieniami, biorące pod uwagę wszystkie zmienne;
2. Drzewo z domyślnymi ustawieniami, biorące pod uwagę wszystkie zmienne, ale z dodaniem wag (fnlwgt);
3. Drzewo biorące pod uwagę wszystkie zmienne z minimalną liczebnością liścia = 500;
4. Drzewo biorące pod uwagę wszystkie zmienne z maksymalną głębokością = 2.

W procesie budowy okazało się, że konfiguracje 1, 2 i 3 generują identyczne drzewa decyzyjne. Z tego względu, ostatecznie pod uwagę wzięte zostały dwa drzewa:

1. „tree”: Drzewo z domyślnymi ustawieniami, biorące pod uwagę wszystkie zmienne.

Drzewo „tree”



Źr.: opr. wł.

Rysunek 4.24. *Rysunek przedstawia drzewo „tree” zbudowane z wykorzystaniem wszystkich danych przy domyślnych ustawieniach.*

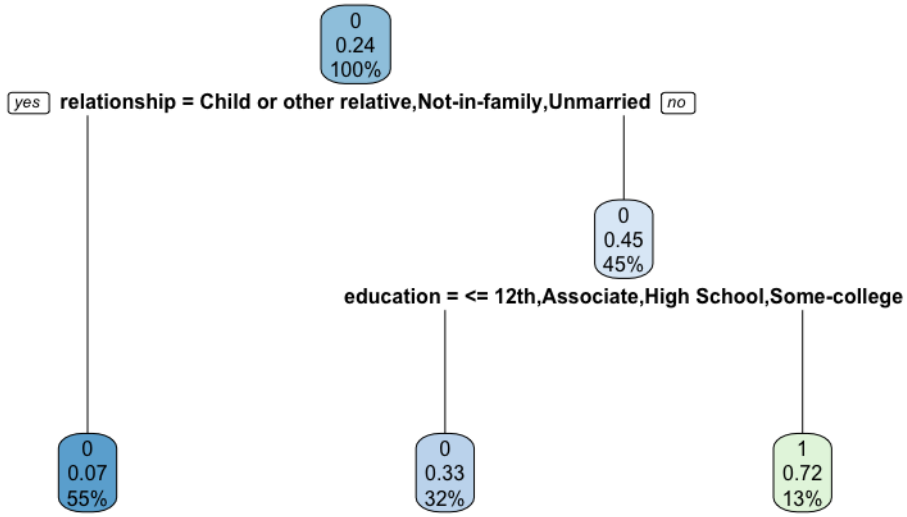
Drzewo decyzyjne wykonuje 5 podziałów, w wyniku których powstaje 6 węzłów końcowych przewidujących wynagrodzenie w zależności od zmiennych relationship, occupation, education oraz age (por. Rysunek 4.24.).

Węzeł główny drzewa pokazuje, że spośród wszystkich obserwacji, 24% osób charakteryzuje się wynagrodzeniem powyżej 50 tys. USD, podczas gdy 76% ma dochody na poziomie lub poniżej 50 tys. USD. Następnie drzewo dzieli się na podstawie zmiennej „relationship”, rozbijając zbiór danych na dwie gałęzie. Do pierwszej gałęzi przypisywani są respondenci, którzy w rodzinie mają status dziecka lub innego krewnego lub nie są częścią żadnej rodziny. Pierwsza gałąź zawiera 55% obserwacji i nie wymaga dalszego podziału, a model przewiduje wartość zmiennej salary „<=50K”. Druga gałąź zawiera 45% obserwacji i dzieli się na podstawie poziomu wykształcenia osób. W drugim podziale drzewo przewiduje wartość zmiennej salary „>50K” dla osób z poziomem wykształcenia „Above Masters”, „Bachelors”, „Masters” oraz „<=50K” dla osób z poziomem wykształcenia "<= 12", "Associate", "High School" lub "Some-college". Trzeci podział występuje w obrębie gałęzi dla osób z poziomem wykształcenia "<= 12", "Associate", "High School" lub "Some-college", w oparciu o zawód. Respondenci, dla których zmienna occupation przyjmuje wartości

Agriculture, Craft-repair, Machine-op-inspct, Other,Service, Transport-moving przyporządkowywani są do klasy negatywnej. Dla pozistałych obserwacji, czwarty podział odbywa się ze względu na wiek. Osoby w wieku od 17 do 33 roku życia oraz od 61 do 90 roku życia trafiają do grupy stanowiącej 4% ogółu obserwacji, dla której model przewiduje wartość zmiennej salary „<=50K”. Ostatni, piąty podział opiera się na poziomie wykształcenia osób. W wyniku tego podziału powstają dwa węzły końcowe, w których dla 4% badanych, poziom wykształcenia to "<= 12" lub "High School", przewidywana wartość zmiennej salary jest równa „<= 50K”, a dla pozostałych 5% „>50K”.

2. „tree_lim_depth”: Drzewo biorące pod uwagę wszystkie zmienne z maksymalną głębokością = 2

Drzewo „tree_lim_depth”



Źr.: opr. wł.

Rysunek 4.25. *Rysunek przedstawia drzewo „*tree_lim_depth*” zbudowane z wykorzystaniem wszystkich danych przy domyślnych ustawieniach.*

Drzewo decyzyjne wykonuje 2 podziały, w wyniku których powstają 3 węzły końcowe przewidujące wynagrodzenie w zależności od zmiennych relationship oraz education (por. Rysunek 4.25.).

Podziały w przypadku drzewa „Tree_lim_depth” są analogiczne, do podziałów występujących w drzewie „tree”. Początkowo drzewo dzieli się na podstawie zmiennej „relationship”, rozbijając zbiór danych na dwie gałęzie. Do pierwszej gałęzi przypisywani są respondenci, którzy w rodzinie mają status dziecka lub innego krewnego, nie są częścią żadnej rodziny lub są niezamężne. Pierwsza gałąź zawiera 55% obserwacji i nie wymaga dalszego podziału, a model przewiduje wartość zmiennej salary „<=50K”. Druga gałąź zawiera 45% obserwacji i dzieli się na podstawie poziomu wykształcenia osób. W drugim podziale drzewo przewiduje wartość zmiennej salary „>50K” dla osób z poziomem wykształcenia „Above Masters”, „Bachelors”, „Masters” oraz „<=50K” dla osób z poziomem wykształcenia "<= 12", "Associate", "High School" lub "Some-college".

Modele regresji zostały stworzone w R przy użyciu pakietów stats oraz MASS rpart(). Podczas budowy modeli pod uwagę wzięte zostały 2 konfiguracje:

1. Regresja logistyczna przy wykorzystaniu wszystkich zmiennych z wyjątkiem education.num, która skutkowała powstaniem błędnego modelu, co wynikało ze współliniowości.

Model „regresja_full”

	Zmienna	Współczynnik	p-value
1	(Intercept)	-8.44628454	0.00
2	age(21,23]	1.25836290	0.05
3	age(23,25]	2.01132095	0.00
4	age(25,29]	2.32202175	0.00
5	age(29,33]	2.84462998	0.00
6	age(33,35]	3.01539165	0.00
7	age(35,37]	3.25770221	0.00
8	age(37,43]	3.37099024	0.00
9	age(43,54]	3.55008420	0.00
10	age(54,61]	3.52266267	0.00
11	age(61,90]	3.01776455	0.00
12	workclassLocal-gov	-0.59635520	0.00
13	workclassOther	-0.96056880	0.33
14	workclassPrivate	-0.36217918	0.00
15	workclassSelf-emp-inc	-0.02184807	0.87
16	workclassSelf-emp-not-inc	-0.70070442	0.00
17	workclassState-gov	-0.80072949	0.00
18	educationAbove Masters	3.00619470	0.00
19	educationAssociate	1.32957586	0.00
20	educationBachelors	1.96925051	0.00
21	educationHigh School	0.85838007	0.00
22	educationMasters	2.28585159	0.00

	Zmienna	Współczynnik	p-value
22	educationMasters	2.28585159	0.00
23	educationSome-college	1.20104730	0.00
24	marital.statusWith Spouse	2.39979046	0.00
25	occupationAgriculture	-1.02931483	0.00
26	occupationCraft-repair	-0.12948827	0.15
27	occupationExec-managerial	0.65259306	0.00
28	occupationMachine-op-inspct	-0.46504876	0.00
29	occupationOther	0.19245955	0.85
30	occupationProf-specialty	0.45772030	0.00
31	occupationSales	0.18403719	0.05
32	occupationService	-0.92045733	0.00
33	occupationTech and Secutity	0.52914536	0.00
34	occupationTransport-moving	-0.25132440	0.02
35	relationshipNot-in-family	1.00767087	0.00
36	relationshipUnmarried	0.54022118	0.00
37	relationshipWife or husband	0.57533369	0.02
38	raceWhite or Asian-Pac-Islander	0.19184047	0.01
39	sexMale	0.07932219	0.18
40	hours.per.week40	0.45368823	0.00
41	hours.per.week>40	0.88460422	0.00
42	countryUS	0.36598687	0.00

Źr.: opr. wł.

Rysunek 4.26. *Rysunek przedstawia model regresji logistycznej „*regresja_full*” zbudowane z wykorzystaniem wszystkich danych przy domyślnych ustawieniach.*

Odczyt z Rysunku 4.26 przedstawia współczynniki przy zmiennych objaśniających wraz z odpowiadającymi im błędami standardowymi oraz wartościami z i p. Współczynniki zmiennych predykcyjnych reprezentują zmianę logarytmu stosunku prawdopodobieństw zmiennej wynikowej dla jednostkowej zmiany w odpowiedniej zmiennej predykcyjnej przy zachowaniu wszystkich innych zmiennych na stałym poziomie.

Na przykład, współczynnik zmiennej wiek sugeruje, że dla każdego wzrostu jednostki w kategorii wieku, prawdopodobieństwo przynależności do klasy pozytywnej wzrasta. Podobnie, współczynniki zmiennych dotyczących wykształcenia sugerują, że osoby z wyższym poziomem wykształcenia częściej zarabiają więcej niż 50 tysięcy USD. Wartości p zmiennych dotyczących wykształcenia są istotne, co wskazuje, że wykształcenie jest istotnym predyktorem wynagrodzenia.

Współczynniki zmiennych klasy zawodowej sugerują, że osoby pracujące w sektorze prywatnym charakteryzują się większym prawdopodobieństwem przynależenia do klasy pozytywnej. Współczynniki zmiennej marital status wskazują, że osoby posiadające małżonków częściej zarabiają więcej niż 50 tysięcy USD, tak jak osoby pracujące w zawodach kierowniczych, specjalistycznych i handlowych. Model regresji logistycznej dostarcza cennych spostrzeżeń na temat czynników wpływających na wysokość wynagrodzeń. Współczynniki modelu sugerują, że wiek, wykształcenie, klasa zawodowa, stan cywilny i zawód są istotnymi determinantami wynagrodzenia. Interpretacja współczynników dostarcza przydatnych

informacji dla zrozumienia związków pomiędzy zmiennymi objaśniającymi a zmienną objaśnianą. Wnioski z modelu wykorzystującego selekcję krokową są analogiczne. Wykształcenie, status związku, zawód, rodzaj wykonywanej pracy, liczba przepracowanych godzin oraz wiek wpływają istotnie na prawdopodobieństwo przynależności do klasy pozytywnej.

W celu stworzenia drugiego modelu regresji wykorzystany jest proces selekcji krokowej, polegający na iteracyjne usuwania predykatorów, które nie przyczyniają się znacząco do dopasowania modelu na podstawie kryterium AIC.

Początkowy model, regresja_full, zawiera wszystkie zmienne predykcyjne (age, workclass, fnlwgt, education, education.num, marital.status, occupation, relationship, race, sex, hours.per.week oraz country). W pierwszym kroku procesu usuwana jest zmienna sex, ponieważ jej usunięcie minimalizuje AIC. W drugim kroku nie są usuwane żadne zmienne, ponieważ model ze wszystkimi pozostałymi zmiennymi charakteryzuje się najniższą wartością AIC.

Model „regresja_step”

	Zmienna	Współczynnik	p-value		Zmienna	Współczynnik	p-value
1	(Intercept)	-8.43265962	0.00	23	educationSome-college	1.20078180	0.00
2	age(21,23]	1.25861088	0.05	24	marital.statusWith Spouse	2.40481229	0.00
3	age(23,25]	2.01025191	0.00	25	occupationAgriculture	-1.00527617	0.00
4	age(25,29]	2.31983745	0.00	26	occupationCraft-repair	-0.10048697	0.25
5	age(29,33]	2.84469593	0.00	27	occupationExec-managerial	0.66844787	0.00
6	age(33,35]	3.01660531	0.00	28	occupationMachine-op-inspct	-0.44175854	0.00
7	age(35,37]	3.25867205	0.00	29	occupationOther	0.21920189	0.82
8	age(37,43]	3.37082313	0.00	30	occupationProf-specialty	0.46969410	0.00
9	age(43,54]	3.55134242	0.00	31	occupationSales	0.20427783	0.03
10	age(54,61]	3.52423800	0.00	32	occupationService	-0.90299477	0.00
11	age(61,90]	3.02576242	0.00	33	occupationTech and Secutity	0.55294780	0.00
12	workclassLocal-gov	-0.60508830	0.00	34	occupationTransport-moving	-0.22291033	0.04
13	workclassOther	-0.97821209	0.32	35	relationshipNot-in-family	1.00762294	0.00
14	workclassPrivate	-0.36655022	0.00	36	relationshipUnmarried	0.51971265	0.00
15	workclassSelf-emp-inc	-0.02166538	0.88	37	relationshipWife or husband	0.59518432	0.02
16	workclassSelf-emp-not-inc	-0.70206872	0.00	38	raceWhite or Asian-Pac-Islander	0.19572250	0.01
17	workclassState-gov	-0.80371850	0.00	39	hours.per.week40	0.46607634	0.00
18	educationAbove Masters	3.01444159	0.00	40	hours.per.week>40	0.90101324	0.00
19	educationAssociate	1.32741828	0.00	41	countryUS	0.36608913	0.00
20	educationBachelors	1.97129990	0.00				
21	educationHigh School	0.85696058	0.00				
22	educationMasters	2.28731809	0.00				

Źr.: opr. wł.

Rysunek 4.27. *Rysunek przedstawia model regresji logistycznej „regresja_step” zbudowany w procesie selekcji krokowej.*

Po przeprowadzeniu selekcji krokowej otrzymano model ze zmiennymu age, workclass, education, marital.status, occupation, relationship, race, hours.per.week oraz country (por. Rysunek 4.27.).

Las losowy stworzony został przy wykorzystaniu pakietu „randomForest” w R. Pod uwagę zostało wziętych 5 różnych lasów losowych. Każdy z lasów składał się z 300 drzew losowych uwzględniających wszystkie zmienne objaśniające. Różniły się one liczbą zmiennych

losowanych na etapie budowy kolejnych węzłów. W rozpatrywanych lasach losowano odpowiednio od 1 do 5 zmiennych, a ich skuteczność mierzono przy pomocy metryki AUC.

Ostatecznie wybrany las wykorzystuje wszystkie zmienne, składa się z 300 drzew, a w procesie budowy drzewa losowana jest jedna zmienna w każdym węźle.

Źr.: opr. wł.

Rysunek 4.28. *Rysunek przedstawia średni spadek wartości współczynnika Giniego dla każdej ze zmiennych w modelu „las_losowy”.*

Wysoki średni spadek nieczystości Giniego wskazuje, że odpowiednia zmienna predykcyjna ma silny wpływ na zmienną wynikową i że usunięcie tej zmiennej z modelu prawdopodobnie zmniejszyłoby dokładność przewidywań. I odwrotnie, niski średni spadek nieczystości Giniego wskazuje, że odpowiednia zmienna przewidująca jest mniej istotna dla modelu.

W modelu lasu losowego, najbardziej istotna są zmienne relationship oraz marital.status. Znaczenia mają również wykształcenie, zawód oraz płeć. Mało istotne są rasa oraz kraj pochodzenia (por. Rysunek 4.28.).

4.4 Ocena modeli

W celu zbadania skuteczności modeli zbiór danych został podzielony na podzbiory treningowy oraz testowy w proporcji 70:30. Na podstawie zbioru testowego skuteczność modeli zmierzona została przy pomocy metryk: dokładność, współczynnik błędnych klasyfikacji (MER), precyzja, czułość, specyficzność, miary F1 oraz AUC (powierzchni pod krzywą ROC).

Porównanie jakości modeli klasyfikujących

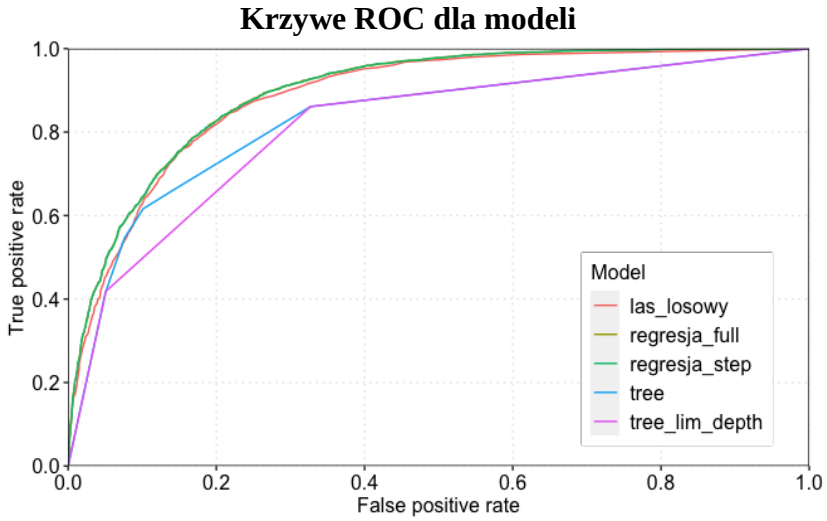
	regresja_full	regresja_step	las_losowy	tree	tree_lim_depth
dokładność	**0.8417281**	**0.8417281**	0.8127559	0.8317977	0.8197174
MER	**0.1582719**	**0.1582719**	0.1872441	0.1682023	0.1802826
precyzja	0.7130479	**0.7134831**	0.8068812	0.6985492	0.7261471
czułość	**0.5873216**	0.5864819	0.3052057	0.5457599	0.4185558
specyficzność	0.9237747	0.9240455	**0.9764419**	0.9240455	0.9490929
F1	**0.6441068**	0.6437788	0.4428876	0.612774	0.5310253
AUC	**0.8952236**	0.89518	0.8865518	0.8304776	0.8135073
Ranking	**1**	**2**	**3**	**4**	**4**

Źr.: opr. wł.

Tabela 4.17. *Tabela przedstawia porównanie jakości modeli klasyfikujących na podstawie wyżej opisanych metryk.*

W Tabeli 4.17. kolorem zielonym wyróżniono najlepsze najbardziej pożądane wartości w każdej z kategorii. Model regresja_full, wykorzystujący wszystkie zmienne okazał się najskuteczniejszy w przypadku 5 spośród 7 metryk, model regresja_step – w przypadku 3 z 7 metryk. W przypadku 1 metryki największą skutecznością charakteryzował się las losowy. Drzewa decyzyjne nie charakteryzowały się najwyższą skutecznością w przypadku żadnych analizowanych metryk.

Wykresy krzywych ROC potwierdzają wysoką skuteczność modeli regresji. Nieco niższą wydajnością charakteryzuje się las losowy. W przypadku zbioru Adult Data, drzewa decyzyjne w najmniej skuteczny sposób przewidują przynależność respondentów do klasy pozytywnej.

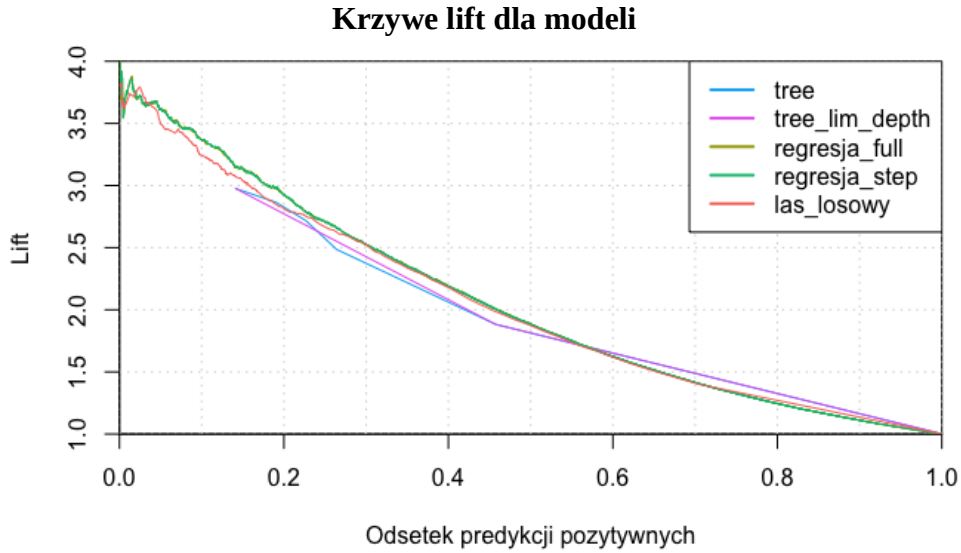


Źr.: opr. wł.

Rysunek 4.29. *Rysunek przedstawia wykresy krzywych ROC dla każdego z modeli.*

Warto jednak zauważyć, że każdy z modeli jest bardzo skuteczny. Zarówno modele regresji, jak i drzewa decyzyjne oraz lasy losowe dobrze przewidują przynależność do klasy pozytywnej, co potwierdzają korzystne wartości metryk służących do oceny jakości modeli (por. 4.29.).

Analizie jakości posłużyły również krzywe lift. Wysoka pozycja i monotoniczność krzywej są kluczowymi aspektami wskazującymi na dobry model. W przypadku krzywych dla modeli regresji zaobserwować można, że w większości przypadków krzywa położona jest najwyżej, a model ten poprawnie klasyfikuje obserwacje o najwyższym prawdopodobieństwie przynależności do klasy pozytywnej czterokrotnie częściej niż klasyfikator losowy. Las losowy jest również bardzo skutecznym klasyfikatorem, a najwyższe odpowiadające mu wartości krzywej lift sięgają około 3,5.



Źr.: opr. wł.

Rysunek 4.30. *Rysunek przedstawia wykresy krzywych lift dla każdego z modeli.*

Z Rysunku 4.30. wnioskować można, że modele regresji oraz las losowy nie charakteryzują się niestety pełną monotonicznością. Wynika to ze znacznie większej liczby punktów odcięcia niż w przypadku drzew decyzyjnych. Ponadto, fluktuacje krzywej lift dla modeli regresji oraz lasu losowego są stosunkowo niewielkie. Dlatego też nie stanowią czynnika dyskwalifikacyjnego. Wyniki oceny modeli pozwalają wnioskować, że modele regresji logistycznej najskuteczniej przewidują przynależność respondentów do klasy pozytywnej. Pozostałe modele dostarczają jednak istotną wartość poznawczą. Gdyż są znacznie łatwiejsze w interpretacji, dzięki czemu pozawalają na skuteczne wyróżnienie społeczno-ekonomicznych determinantów przynależności do klasy wyższej.

5 Uwagi końcowe

Celem badania była identyfikacja czynników społeczno-demograficznych, które determinują wysokie zarobki. Możliwość określenia dochodów jednostki na podstawie jej cech społecznych i demograficznych jest cenna zarówno w kontekście biznesowym, jak i w celach badawczych. Identyfikacja osób o wysokich zarobkach może być przydatna np. w marketingu ukierunkowanym, gdzie firmy specjalizujące się w sprzedaży produktów premium mogą kontaktować się z potencjalnie zamożnymi osobami w celu zwiększenia skuteczności swoich kampanii marketingowych. Identyfikacja czynników warunkujących przynależność do klasy wyższej ma również istotne znaczenie poznawcze, gdyż pozwala na weryfikację hipotez dotyczących utrzymujących się dysproporcji płacowych z uwzględnieniem płci, rasy i pochodzenia. Przedstawienie wyraźnej zależności między wykształceniem, rodzajem wykonywanego zawodu a dochodami może natomiast ułatwić podejmowanie decyzji osobom stojącym przed ważnymi wyborami życiowymi. Świadomość czynników determinujących wysokie zarobki może być również przydatna w doradztwie zawodowym lub akademickim.

Determinanty wysokich zarobków zostały zidentyfikowane przy użyciu metod uczenia maszynowego z nadzorem. Modele uczenia maszynowego, takie jak drzewa decyzyjne, modele regresyjne i las losowy, zostały zbudowane na podstawie oznaczonego zbioru danych Adult Data, który od lat jest z powodzeniem wykorzystywany przez badaczy i ekonomistów. Został on zbudowany w oparciu o badania przeprowadzone w Stanach Zjednoczonych w latach 90. XX wieku i zawiera szereg informacji na temat społecznych i ekonomicznych cech osób, w tym rodzaj wykonywanego zawodu, wykształcenie, płeć, stan cywilny oraz wiek. Modele zbudowane na tym zbiorze danych skutecznie przewidują przynależność jednostki do klasy pozytywnej, wykazując wysoką dokładność, co potwierdzają popularne metryki oceny modeli uczenia maszynowego.

Z badania wyniknęło, iż najważniejszymi czynnikami determinującymi przynależność do klasy wyższej są stan cywilny, wykształcenie i wiek, podczas gdy płeć, rasa i kraj pochodzenia są mniej istotne. Te ustalenia są zgodne z teorią ekonomiczną.

Istotnym ograniczeniem tego badania jest fakt, że zbiór danych stanowiących podstawę modeli uczenia maszynowego pochodzi z badania przeprowadzonego 30 lat temu w USA. Oznacza to, że nie odzwierciedla on istotnych zmian gospodarczych, które zaszły w ostatnich latach. Na przykład, można się spodziewać większego odsetka osób należących do klasy pozytywnej wśród informatyków w dzisiejszej gospodarce opartej na praktycznych umiejętnościach. Interesujące byłoby zbadanie obecnego wpływu wykształcenia na zarobki i ocena, czy w ostatnich latach nastąpiło zmniejszenie poziomu różnic w zarobkach według rasy,

płci lub pochodzenia. Dodatkowo, ekstrapolacja wyników tego badania na inne kraje, takie jak Polska, implikuje założenie podobnych korelacji, co może nie być trafne, gdyż zjawiska ekonomiczne różnią się w poszczególnych regionach.

Uchyleniu tych ograniczeń posłużyć może powtórzenie badania współcześnie w różnych krajach. Pozwoliłoby ono na identyfikację zmian, które zaszły na przestrzeni lat oraz zrozumienie różnic w zależnościach społeczno-ekonomicznych pomiędzy różnymi krajami. Dodatkowym obszarem rozwoju jest również rozszerzenie zbioru danych o dodatkowe zmienne i uwzględnienie takich czynników jak na przykład zarobki rodziców, przynależność religijna, kierunek studiów czy miejsce zamieszkania. Rozszerzenie badania o te czynniki może pozwolić na zwiększenie precyzji oszacowań, a także wartości poznawczej zbudowanych modeli. Współcześnie powszechnie dostępne są informacje dotyczące wysokości zarobków zależności od rodzaju wykonywanej pracy wykształcenia kierunku studiów miejsca zamieszkania rasy czy płci. Trudno jednak o zbiór danych uwzględniający wszystkie te czynniki jednocześnie, tak jak ma to miejsce w przypadku zbioru Adult Data.

Z tego względu, pomimo problemów związanych z ekstrapolacją wyników na różne kraje oraz wiekiem zbioru danych, zależności wykryte w pracy pozostają aktualne i zgodne również ze współczesną wiedzą ekonomiczną.

6 Bibliografia i przywołania internetowe

1. The Economic Value of Behavioural Targeting in Digital Advertising, https://datadrivenadvertising.eu/wp-content/uploads/2017/09/BehaviouralTargeting_FINAL.pdf, data dostępu: 24 lutego 2023
2. Global Powers of Luxury Goods 2022, https://www.deloitte.com/content/dam/assets-shared/legacy/docs/analysis/2022/gx-global-powers-of-luxury-goods-report.pdf, data dostępu: 24 lutego 2023
3. Education Pays, https://www.bls.gov/emp/chart-unemployment-earnings-education.htm, data dostępu: 24 lutego 2023
4. Income Inequality and Income-Class Consumption Patterns - Federal Reserve Bank of Cleveland, https://www.clevelandfed.org/publications/economic-commentary/2014/ec-201418-income-inequality-and-income-class-consumption-patterns, data dostępu: 24 lutego 2023
5. Lifestyles of the top 1%: How America's elite live, shop, and play, https://blogs.oracle.com/advertising/post/lifestyles-of-the-top-1-how-americax27s-elite-live-shop-and-play, data dostępu: 24 lutego 2023
6. Income before taxes: Annual expenditure means, shares, standard errors, and coefficients of variation, Consumer Expenditure Surveys, 2021, https://www.bls.gov/cex/tables/calendar-year/mean-item-share-average-standard-error/cu-income-before-taxes-2021.pdf, data dostępu: 24 lutego 2023
7. Gore, Alestig, Ratcliff, 2020, Confronting carbon inequality, Oxfam
8. Making It Personal, https://www.accenture.com/_acnmedia/pdf-83/accenture-making-personal.pdf, data dostępu: 24 lutego 2023
9. Cramer-Flood, 2021, Worldwide Digital Ad Spending Year-End Update, eMarketer
10. The State of Personalization 2021, https://segment.com/state-of-personalization-report-2021/, data dostępu: 24 lutego 2023
11. Global E-Commerce Jumps to $26.7 Trillion, Covid-19 Boosts Online Retail Sales, https://unctad.org/news/global-e-commerce-jumps-267-trillion-covid-19-boosts-online-sales, data dostępu: 24 lutego 2023
12. York, 2023, Summary of the Latest Federal Income Tax Data, Tax Foundation
13. Meer, Priday, 2020, Generosity Across the Income and Wealth Distributions, National Bureau of Economic Research

14. Ramadani, 2012, The Importance of Angel Investors In Financing The Growth Of Small And Medium Sized Enterprises, Human Resource Management Academic Research Society
15. Wilson, 2011, Financing high-growth firms: The role of angel investors, Organisation for Economic Co-operation and Development
16. Education at a Glance 2022, https://www.oecd-ilibrary.org/docserver/3197152b-en.pdf?, data dostępu: 24 lutego 2023
17. Talley, Weng, Zaski, 2022, Effect of Education on Wage Earning, Organisation for Economic Co-operation and Development
18. Mocan, 2014, The Impact of Education on Wages: Analysis of an Education Reform in Turkey, University of Pennsylvania
19. Boshara, Emmons, Noeth, 2015, The Demographics of Wealth, St. Louis Federal Reserve
20. United States Job Market Report, https://www.glassdoor.com/research/job-market-report-united-states/, data dostępu: 24 lutego 2023
21. Occupational Employment and Wage Statistics, https://www.bls.gov/oes/current/oes_nat.htm#00-0000, data dostępu: 24 lutego 2023
22. Salary Report 2022, https://www.jobstreet.com.my/en/cms/employer/wp-content/themes/jobstreet-employer/assets/pdf/MY-SalaryReport-R3.5-25thJan2022-final.pdf, data dostępu: 24 lutego 2023
23. Salary Report 2022, https://taptalent.eu/wp-content/uploads/2022/04/tap.talent-salary-report-2022-1.pdf, data dostępu: 24 lutego 2023
24. Daly, Hobijn, Pedtke, 2017, Disappointing Facts about the Black-White Wage Gap, Federal Reserve Bank of San Francisco
25. Wilson, Rodgers, 2016, Black-white wage gaps expand with rising wage inequality, Economic Policy Institute
26. Vandenbroucke, 2018, Married Men Sit Atop the Wage Ladder, St. Louis Federal Reserve
27. Changing America, https://web.archive.org/web/20120131063624/http://www.gpoaccess.gov/eop/ca/pdfs/ca.pdf, data dostępu: 24 lutego 2023
28. Alpaydin, 2014, Introduction to Machine Learning, Massachusetts Institute of Technology
29. The Complete Guide to Decision Trees, https://www.explorium.ai/blog/the-complete-guide-to-decision-trees/, data dostępu: 3 marca 2023 roku
30. Powers, 2007, Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", Flinders University

7 Spis tabel

8 Spis rysunków

9 Streszczenie

Celem pracy było wyróżnienie czynników wpływających na wysoki poziom dochodów przy wykorzystaniu informacji o populacji Stanów Zjednoczonych.

Wyróżnienie klasy wyższej jest istotne ekonomicznie, co wynika z jej znaczenia z perspektywy marketingu, finansów publicznych czy finansowania inwestycji. Znaczenie to zostało szeroko udokumentowane poprzez przytoczone badania.

Aby zrealizować cel pracy wykorzystano dane publicznie dostępne w zbiorze „Adult Dataset" z UCI Machine Learning Repository. Zawierają one informacje o cechach demograficznych, społeczno-ekonomicznych i finansowych respondentów. W celu wyróżnienia determinantów wysokich zarobków zastosowano techniki uczenia maszynowego z nadzorem: lasy losowe, regresję logistyczną oraz drzewa decyzyjne. Modelem najskuteczniej przewidującym przynależność do klasy wyższej okazała się regresja logistyczna.

Z badania wyniknęło, iż najważniejszymi czynnikami determinującymi przynależność do klasy wyższej są: stan cywilny, wykształcenie oraz wiek. Mniej istotne są z kolei płeć, rasa czy kraj pochodzenia.