AKADEMIA GÓRNICZO – HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ ZARZĄDZANIA

PRACA DYPLOMOWA

licencjacka

„Zastosowanie algorytmu regresyjnych lasów losowych do analizy i prognozowania cen mieszkań w polskich powiatach”

Sylwia Kołodziej

Informatyka i Ekonometria

studia stacjonarne

promotor: dr Łukasz Lach

Kraków 2020

OŚWIADCZENIE AUTORA PRACY DYPLOMOWEJ

„Oświadczam, że dokumentacja pracy dyplomowej nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz. U. z 2006 r. Nr 90 poz. 631 z późniejszymi zmianami) oraz dóbr osobistych chronionych prawem cywilnym. Nie zawiera ona również danych i informacji, które uzyskałam w sposób niedozwolony. Wersja elektroniczna pracy dołączona przeze mnie na nośniku CD/DVD jest w pełni zgodna z wydrukiem pracy.

Zaświadczam także, że niniejsza praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów wyższej uczelni lub tytułów zawodowych."

……. …….

data podpis autora

OŚWIADCZENIE PROMOTORA PRACY

„Zatwierdzam do rejestracji i dopuszczam do obrony."

……. …….

data podpis promotora

Spis treści

1. Wprowadzenie 5

1.1. Rynek nieruchomości w Polsce 5

1.2. Rynek nieruchomości mieszkaniowych 8

1.3. Uzasadnienie podjęcia tematu 10

1.4. Cel pracy 11

2. Wstęp do studium przypadku 12

2.1. Dobór zmiennych 12

2.2. Charakterystyka zbioru danych 16

3. Metodologia 21

3.1. Drzewa decyzyjne 21

3.2. Metody agregacji drzew 30

3.3. Lasy losowe 33

4. Główne hipotezy badawcze 36

5. Badanie empiryczne 38

5.1. Pierwsze podejście do budowy lasu 38

5.2. Poszukiwanie optymalnych parametrów 40

5.3. Badanie istotności zmiennych objaśniających 45

5.4. Wyniki algorytmu a podział zbioru danych 48

5.5. Lasy losowe a regresja liniowa 54

6. Zakończenie 58

6.1. Podsumowanie wyników 58

6.2. Wnioski i uwagi końcowe 60

7. Literatura 62

8. Spis ilustracji 63

9. Spis tabel 63

10. Spis równań 64

11. Wykaz załączników 64

Streszczenie

Zakres pracy obejmuje wykorzystanie metod wielowymiarowej analizy danych do badań rynku nieruchomości mieszkaniowych w Polsce. W tym celu przeprowadzono analizę sił oddziaływania 21 determinant o charakterze społeczno-gospodarczym, demograficznym oraz polityczno-prawnym na wysokość średniej ceny za $1m^2$ lokali mieszkalnych w powiatach kraju. Badanie oparto na danych z roku 2018.

W niniejszej pracy zastosowano nowoczesne narzędzia analizy danych, służące do rozwiązywania problemów klasyfikacji i regresji. Na przykładzie polskiego rynku mieszkań wyjaśniono sposób działania algorytmu drzew decyzyjnych oraz wprowadzono pojęcie baggingu. Właściwe badanie empiryczne oparto na eksploracji danych z wykorzystaniem regresyjnych lasów losowych. Zbudowano modele regresji o możliwie najwyższej precyzji oraz minimalnym błędzie prognozy. Odkryto szereg wad i zalet techniki lasów, a także porównano jej efektywność z klasyczną regresją liniową. Dzięki przeprowadzonym obliczeniom znaleziono główne źródła nierówności regionalnych, które można zaobserwować na rodzimym rynku nieruchomości. Odkryto także, iż zasadne i skuteczne analizowanie tego sektora krajowej gospodarki wymaga osobnego spojrzenia na miasta o własnych prawach oraz na pozostałe powiaty Polski. Opracowanie zawiera dowody na to, iż w obu grupach jednostek administracyjnych wysokość cen mieszkań jest kształtowana w inny sposób.

Słowa kluczowe

rynek nieruchomości, ceny mieszkań, powiaty w Polsce, klasyfikacja, regresja, drzewa decyzyjne, bagging, lasy losowe

1. Wprowadzenie

1.1. Rynek nieruchomości w Polsce

W dobie fascynacji podróżami i poznawaniem świata oraz wzrastającej mobilności ludności mieszkanie lub dom niezmiennie pozostają jednymi z dóbr, zapewniających człowiekowi zaspokojenie podstawowych potrzeb. Wśród nich znajduje się poczucie bezpieczeństwa, schronienia i przynależności, a także potrzeba niezależności od środowiska zewnętrznego. Odpowiednie warunki mieszkaniowe są ponadto fundamentem rozwoju rodziny i relacji międzyludzkich. Ciężko wyobrazić sobie jak wyglądałoby życie gospodarstw domowych, gdyby nie posiadały one „dachu nad głową". Dom czy mieszkanie pełnią istotne funkcje nie tylko jako dobra materialne. Z psychologicznego punktu widzenia posiadanie własnego lokum jest dla ludzi ważnym celem, a nawet marzeniem, którego spełnienie daje poczucie dumy i satysfakcji.

Wybór miejsca zamieszkania ma ponadto ogromny wpływ na postać systemu społeczno-gospodarczego zespołów aglomeracyjnych. Chęć przyciągania do siebie nowych mieszkańców wzbudza konkurencję między jednostkami administracyjnymi, które kładą nacisk na proces pozyskiwania kolejnych zasobów mieszkaniowych, a co za tym idzie – nowych mieszkańców. Bezpośrednią korzyścią, jaką odnoszą budżety dzięki takiemu napływowi jest wzrost dochodów – w tym przypadku szczególnie z tytułu podatku od nieruchomości. Liczy się także wzrost zasobów kapitału ludzkiego. Z kolei dla młodych ludzi polityka mieszkaniowa oraz poziom cen stanowią jedne z najistotniejszych czynników, decydujących o wyborze miejsca do życia. Niełatwo jest im znaleźć „cztery ściany", które jednocześnie będą spełniały wszystkie stawiane wymagania- także te cenowe.

Od kilkunastu lat rynek nieruchomości w Polsce prężnie się rozwija. Mimo rosnących cen zainteresowanie pozostaje na wysokim poziomie. Wynika to z niezaspokojenia popytu na mieszkania przez wiele wcześniejszych lat. Kraj boryka się z niedoborem, który potwierdza fakt posiadania najniższej w Europie liczby mieszkań na 1000 mieszkańców[1]. Z drugiej strony - od momentu przystąpienia Polski do Unii Europejskiej zyskała ona na atrakcyjności dla

[1] Eurostat Database, *Income and Living Conditions*, https://ec.europa.eu/eurostat/web/income-and-living-conditions/data/database, [02.03.2020].

potencjalnych inwestorów. Rok 2019 był rekordowy nie tylko pod względem liczby mieszkań sprzedanych przez deweloperów, ale i liczby udzielonych pozwoleń na budowę. Popyt inwestycyjny jest napędzany przez wzrost dochodów gospodarstw domowych, wspomniany „głód mieszkań" i poprawę poziomu zamożności Polaków. W porównaniu do niskich stóp procentowych relatywnie wysokie stopy zwrotu z wynajmu także zachęcają do nabywania nieruchomości. Co ciekawe, aż w 25 krajach europejskich kupowanych jest mniej nowych mieszkań niż w Polsce. Popularność zakupów od deweloperów jest tutaj wyższa niż chociażby w Niemczech, Włoszech, Hiszpanii, Francji, Wielkiej Brytanii czy Danii[2]. Co zadziwiające - szybko rosnące ceny mieszkań nie zniechęcają zainteresowanych, którzy w obawie przed kolejnymi podwyżkami nie chcą odkładać momentu zakupu na później. W ten oto sposób cały mechanizm rynku nieruchomości nabiera rozpędu. Według Indeksu Cen Domów (House Price Index) w 2019 roku Polska zajęła 9 miejsce w Unii Europejskiej pod względem wzrostu cen mieszkań. Gdy dla całej wspólnoty odnotowano ich średni wzrost o 4,2%, w naszym kraju mieszkania podrożały aż o 8,2% w stosunku do roku poprzedniego[3]. Jakie zmiany spowodowały tak duży skok? Skąd bierze się zróżnicowanie przestrzenne regionów Polski pod względem cen lokali mieszkalnych? Dlaczego w polskich miastach ceny różnią się aż tak bardzo? Przykładowo, gdy w Sopocie 1m^2 mieszkania zakupionego na rynku pierwotnym kosztuje średnio blisko 12 tysięcy złotych, w Warszawie jest to około 9 tysięcy, a w Opolu zaledwie 4,5. Czy różnorodność stawek za mieszkania dotyczy tylko dużych miast, podczas gdy dla pozostałych obszarów kraj jest pod tym względem stosunkowo jednorodny? Warto także zwrócić uwagę na niepewność i zagrożenie zachwianiem sytuacji na rynku nieruchomości, związane z panującą obecnie pandemią wirusa SARS-CoV-2. W jaki sposób koronawirus „uderzy" w ten sektor gospodarki i jakich skutków należy się spodziewać? Można doszukiwać się w tym przypadku subiektywnych podejrzeń i stawiać przeróżne hipotezy. Lepszym sposobem na znalezienie odpowiedzi na te i inne pytania jest jednak posłużenie się metodami ekonometrycznymi, które umożliwiają analizę determinant regionalnego zróżnicowania cen na rynku mieszkaniowym w Polsce i dają konkretne, mierzalne i zweryfikowane odpowiedzi.

[2] *Polski boom na nowe mieszkania. Europa w tyle,* https://nieruchomosci.dziennik.pl/news/artykuly/6451997,polska-boom-nowe-mieszkania-europa-rynek-nieruchomosci.html, [16.03.2020].

[3] *Rynek mieszkaniowy w UE. Polska w pierwszej dziesiątce państw z największym wzrostem cen*, https://forsal.pl/nieruchomosci/mieszkania/artykuly/1433912,ceny-mieszkan-w-ue-ceny-mieszkan-w-polsce-eurostat.html, [16.03.2020].

Przejście do części badawczej wymaga wprowadzenia w „świat nieruchomości” i próby zdefiniowania jego roli i wpływu na funkcjonowanie gospodarki.

1.2. Rynek nieruchomości mieszkaniowych

Nieruchomości kryją za sobą łańcuch zależności, który w istotny sposób obejmuje wiele podmiotów i odgrywa ważną rolę w kształtowaniu ogólnej sytuacji społeczno-gospodarczej w danym kraju. Od innych dóbr masowych odróżniają je przede wszystkim dwie cechy: nieprzenaszalność oraz różnorodność praw. *„Ich problematyka ekonomiczna zajmuje szczególne miejsce na tle innych obszarów, co wynika m.in. z roli jako składnika majątku narodowego, z rynkowego charakteru czy specyficznych cech jako towaru czy obiektu inwestowania”*[4]. W pojęciu ogólnym wyróżnia się trzy rodzaje nieruchomości:

1) **gruntowe**, czyli powierzchnie ziemskie stanowiące odrębny przedmiot własności,
2) **budynkowe**, czyli budynki będące odrębnym od gruntu przedmiotem własności,

a także

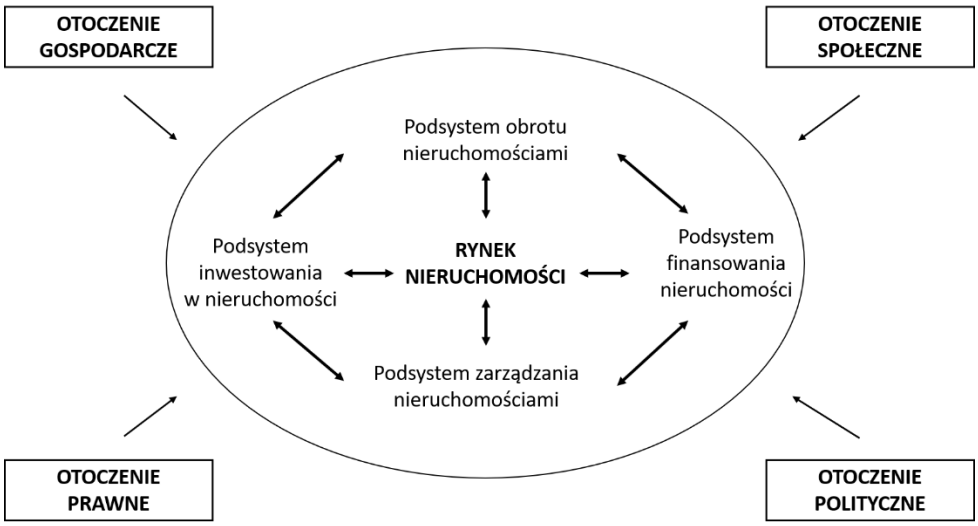
3) **lokalowe**, czyli części budynków stanowiące odrębny przedmiot własności.[5]

Samo miejsce obrotu nimi, zwane rynkiem nieruchomości sformułowane jest w literaturze jako ,,*ukształtowany zbiór wzajemnie ze sobą powiązanych uczestników procesów oferowania i wymiany nieruchomości polegających na przekazywaniu praw własności i praw do korzystania z nieruchomości, oraz warunków, w których realizowane są te procesy”*[6]. W nieprawidłowościach struktury tego rynku słusznie upatruje się źródła wielu problemów gospodarczych. To właśnie pęknięcie tzw. „bańki na rynku nieruchomości” doprowadziło w 2007 roku do kryzysu finansowego w Stanach Zjednoczonych, który spowodował pogorszenie globalnej sytuacji finansowej i załamanie wielu gospodarek na całym świecie. Wieloletnie obserwacje pozwalają zauważyć dodatnią korelację między kondycją gospodarki i rynku nieruchomości w danym państwie. Oznacza to, że jeśli poprawiają się parametry makroekonomiczne zaobserwować można polepszenie się statystyk opisujących

[4] Olszowy Joanna, *Wybrane aspekty wpływające na rynek nieruchomości w Polsce*, Społeczna Akademia Nauk, Wydział Zamiejscowy w Tarnobrzegu, s. 82.
[5] *Ustawa z dnia 23 kwietnia 1964 r. - Kodeks cywilny*, Art. 46. § 1, http://prawo.sejm.gov.pl/isap.nsf /download.xsp/WDU19640160093/O/D19640093.pdf, [06.03.2020].
[6] Kałkowski Leszek, *Rynek nieruchomości w Polsce*, TWIGGER, Warszawa 2003, s. 20-21.

kondycję sektora nieruchomości[7]. Wynika to z faktu, iż rynek ten jako system wiąże otoczenie tworzone przez struktury polityczne, prawne, gospodarcze i ekonomiczne, co można zaobserwować na poniższym schemacie.

Rysunek 1: Rynek nieruchomości jako system



Źródło - Bryx M. Rynek nieruchomości. System i funkcjonowanie, Poltext, Warszawa 2006, rys. 4.2 s. 88

Przedstawiony na rysunku system jest niezwykle dynamiczny. Zmieniają się preferencje konsumentów, uregulowania prawne, cele interwencji państwa, a także działania deweloperów w zakresie sprzedaży i wynajmu. Obok dynamiki zmian, ważną cechą jest niejednolitość rynku nieruchomości. Ciężko jest badać go jako ogół, dlatego przy analizie procesów w nim zachodzących wybiera się konkretny rodzaj. Wśród podziałów ze względu na obszar przestrzenny (rynek lokalny, regionalny, krajowy czy międzynarodowy) czy kryterium nabywanych praw (rynek najmu i lokat) wydziela się także podział ze względu na funkcję. Na tym etapie należy wprowadzić pojęcie **rynku nieruchomości mieszkaniowych**. Co go wyróżnia i sprawia, że staje się odrębnym segmentem całości? Odpowiedzią na to pytanie jest przede wszystkim jego **społeczna funkcja**. Swoim zasięgiem nie obejmuje on nieruchomości komercyjnych (biura, hotele), rolnych (użytki rolne, pastwiska) ani przemysłowych (fabryki, magazyny), których przeznaczenie jest inne. Stanowią one głównie pośrednie źródło dochodu i narzędzie do lokowania kapitału oraz pomnażania zysku. Rynek mieszkaniowy obejmuje

[7] Olszowy Joanna, *Wybrane aspekty wpływające na rynek nieruchomości w Polsce*, Społeczna Akademia Nauk, Wydział Zamiejscowy w Tarnobrzegu, s. 86.

natomiast – co jest intuicyjne - mieszkania oraz domy i jest ściśle skoncentrowany na dostarczaniu wartości użytkowej [8].

Ogólnie rzecz ujmując, zakres badania zostanie zawężony i obejmować będzie analizę zróżnicowania cen tych nieruchomości, które pełnią **funkcję mieszkalną**. Główny problem badawczy oraz najważniejsze czynniki, determinujące kondycję rynku mieszkaniowego przedstawione są w dalszej części pracy. W następnym podrozdziale zaprezentowano kierunek, w jakim zmierza całokształt analizy, a także cel i oczekiwania wobec włożonego w nią wysiłku.

1.3. Uzasadnienie podjęcia tematu

Od kilku lat temat mieszkań w Polsce zyskuje na popularności, będąc tym samym przedmiotem dyskusji w wielu środowiskach. Dla młodych osób konfrontacja z tym rynkiem i zakup własnego mieszkania jest jednym z pierwszych poważnych wyzwań w dorosłym życiu. Nic więc dziwnego, że wymienia się poglądami o obecnej sytuacji, a także podejmuje próby przewidywania przyszłości sfery mieszkaniowej. Coraz częściej pojawia się także pytanie o to „kiedy ta bańka pęknie". Ze względu na świadomość wpływu tego sektora gospodarki na ogół sytuacji w kraju postanowiono poddać go dokładniejszej analizie. Zastanawiające są przede wszystkim nierówności cenowe na przestrzeni całego kraju, które są wyraźnie widoczne na przekroju powiatów Polski. Z punktu widzenia młodego Polaka zainteresowanie tematem w dużej mierze wynika z czystej ciekawości co do źródeł tych nierówności oraz z chęci zweryfikowania przewidywań, dostarczanych przez media. Patrząc zaś na ogół problemu z punktu widzenia ekonometryka wielkości związane z rynkiem mieszkaniowym w regionach Polski wydają się być nieregularne i na pierwszy rzut oka ciężko dostrzec jednoznaczne, silne zależności. Zmierzenie się z tak nierównomiernymi oraz dynamicznymi danymi ma szansę stać się ciekawym wyzwaniem zarówno pod względem metodologicznym jak i naukowo-badawczym.

[8] Kucharska-Stasiak Ewa, *Ekonomiczny wymiar nieruchomości,* Wydawnictwo Naukowe PWN Warszawa 2006, s. 38-55.

1.4. Cel pracy

Zastanawiający się nad tym dlaczego w danym regionie ceny lokali mieszkalnych są tak niskie, czy częściej - zadziwiająco wysokie, zazwyczaj analizują czynniki takie jak poziom wynagrodzeń, rozwój gospodarczy czy szeroko pojęty poziom dobrobytu. Jednak czy zawsze okazuje się, że drogie mieszkania są tam, gdzie dochody pozostają na wysokim poziomie, a stopa bezrobocia jest niska? Z drugiej strony, czy faktycznie najtaniej dom czy mieszkanie można kupić w oddali od miasta, gdzie gęstość zaludnienia i liczba miejsc pracy są dużo niższe?

Praca ma na celu **zbadanie wpływu oraz siły oddziaływania** 21 determinant, które podejrzewane są o stymulowanie średniej ceny za $1m^2$ lokali mieszkalnych w powiatach Polski. Ma to pokazać **realne zależności** i pomóc zaobserwować pewne prawidłowości w tym segmencie krajowej gospodarki. Dążeniem pośrednim badania jest ukształtowanie na zebranym zbiorze danych jak najlepszych **zdolności prognostycznych**, dotyczących cen na rynku nieruchomości, aby dzięki temu możliwe było ich przewidywanie na podstawie zmian wartości kluczowych czynników. Wykorzystanie do tego nowatorskiej techniki eksploracji danych, jaką jest algorytm regresyjnych lasów losowych, ma pokazać **sposób działania i możliwości**, jakie otwiera przed analitykami ta technika.

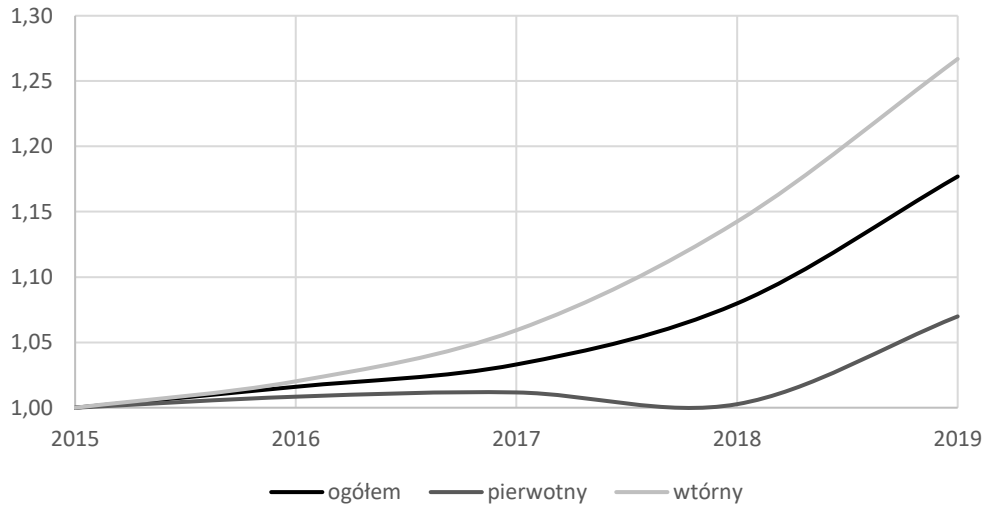
Dodatkowo, uzyskane wnioski mogą być przydatne dla samorządów powiatów, chcących nie tylko poznać drogi wpływu na notowania na rynku nieruchomości, ale i dzięki przewidywaniom jak najlepiej dostosować zarządzanie obszarem do prawdopodobnej przyszłej sytuacji. Dążeniem autora jest także wypracowanie wartości badania dla potencjalnych nabywców oraz podmiotów inwestujących i sprzedających lokale w obrocie rynkowym. Celem w najogólniejszym pojęciu jest taka eksploracja przekonań i przypuszczeń, aby dzięki narzędziom ekonometrycznym wykształcić z nich zweryfikowane, realne prawidłowości.

2. Wstęp do studium przypadku

2.1. Dobór zmiennych

Do tej pory w niniejszej pracy o rynku nieruchomości mieszkaniowych mówiono jedynie w ujęciu teoretycznym. Wprowadzenie w tematykę umożliwiło przejście na tym etapie do zdefiniowania i zgromadzenia odpowiednich zmiennych, kształtujących postać tego rynku. Swoboda wyboru była ograniczona jedynie przez dostępność, spójność i wiarygodność danych. W Polsce ich gromadzeniem zajmuje się Główny Urząd Statystyczny i to właśnie z zasobów jego Banku Danych Lokalnych zdecydowano skorzystać, aby pozyskać bazę badawczą. Jako miarę opisującą zróżnicowanie przestrzenne polskiego rynku mieszkań przyjęto **średnią cenę za 1m^2 lokali mieszkalnych** sprzedanych w ramach transakcji rynkowych. Dane dotyczą roku 2018 i swym zakresem obejmują wszystkie 380 powiatów. Fakt, że wartość ta jest godna uwagi potwierdza poniższy wykres, przedstawiający zmianę wysokości cen mieszkań w stosunku do roku 2015. Okazuje się, że na rynku wtórnym, obejmującym lokale z odsprzedaży, ceny w przeciągu 4 lat wzrosły aż o 27%. Jeśli chodzi o nowe mieszkania, a zatem rynek pierwotny, wzrost nominalny wyniósł około 7%, przy czym ceny mieszkań oddanych do użytkowania były wysokie już 5 lat temu.

Rysunek 2: Zmiana ceny za 1m^2 mieszkań w latach 2015-2019 w stosunku do roku 2015



Źródło: Opracowanie własne na podstawie danych GUS, https://bdl.stat.gov.pl/BDL/metadane/cechy/3788?back=True

Biorąc pod uwagę dynamikę i otwartość systemu nieruchomości do analizy wybrano determinanty z otoczenia gospodarczego, społecznego, demograficznego, polityczno-prawnego, a także bezpośrednio związane ze strukturą rynku nieruchomości (wewnętrzne). Zakres pracy nie został ograniczony do rynku wtórnego czy pierwotnego, ale badano ogół rynku nieruchomości. W Tabeli nr 1 zaprezentowano i scharakteryzowano wszystkie zmienne, które najpewniej kształtują ceny nieruchomości w powiatach Polski. Warto przy ich przeglądzie zwrócić uwagę na jednostki, które zostały wyrażone tak, aby nadać porównywalność powiatom o różnej powierzchni czy liczbie ludności. Dla zachowania przejrzystości jedynym przedstawicielem bazy danych w tabeli jest **Sopot** - miasto na prawach powiatu, w którym za metr kwadratowy mieszkań w 2018 roku płacono najwięcej, bo prawie 12 tysięcy złotych. Tuż za nim na podium znalazły się Warszawa oraz powiat tatrzański (z siedzibą w Zakopanem).

Tabela 1: Czynniki determinujące ceny lokali mieszkalnych w powiatach Polski

	Nazwa czynnika, (skrót)*	**Opis**	**Jednostka**	**Wartość (Sopot)**
gospodarcze	**Wynagrodzenie brutto, (wyn)**	Przeciętne miesięczne wynagrodzenie brutto w danym powiecie.	zł	5700,7
	Dochody budżetów, (dochBud)	Dochody jednostek samorządu terytorialnego w przeliczeniu na mieszkańca.	zł/os.	9063,8
	Wydatki budżetów, (wydBud)	Wydatki jednostek samorządu terytorialnego w przeliczeniu na mieszkańca.	zł/os.	8974,8
	Liczba przedsiębiorstw, (przeds)	Wielkość obrazująca sytuację gospodarczą danego powiatu. Jest to liczba podmiotów gospodarki narodowej wpisanych do rejestru REGON w przeliczeniu na 1000 mieszkańców.	jedn./1000os.	239,0
	Liczba nowych przedsiębiorstw, (nowePrzeds)	Miara rozwoju przedsiębiorczości powiatu, podana jako liczba nowo zarejestrowanych firm w przeliczeniu na 1000 mieszkańców.	jedn./1000os.	19,0

demograficzne	**Współczynnik urbanizacji, (wspUrb)**	Procentowy udział ludności zamieszkującej miasta w łącznej liczbie ludności danego powiatu. Dla miast na prawach powiatu zawsze wynosi 100%.	%	100,0
	Stopa bezrobocia, (stBez)	Stosunek liczby zarejestrowanych bezrobotnych do liczby ludności aktywnej zawodowo.	%	1,8
	Gęstość zaludnienia, (gestZal)	Liczba ludności przypadająca na 1km^2 powierzchni.	os./km^2	2086,0
	Ludność w wieku produkcyjnym, (prod)	Procentowy udział mężczyzn w wieku 18-64 i kobiet w wieku 18-59 w łącznej liczbie ludności danego powiatu.	%	55,8
	Przyrost naturalny, (przyrNat)	Współczynnik przyrostu naturalnego, obliczony jako stosunek różnicy pomiędzy urodzeniami żywymi a zgonami do całkowitej liczby ludności. Podany w przeliczeniu na 1000 mieszkańców.	os./1000os.	-5,8
	Saldo migracji, (salMig)	Współczynnik salda migracji obliczony jako stosunek różnicy liczby imigrantów i emigrantów na danym terenie do łącznej liczby ludności. Oznacza ubytek/przyrost ludności na skutek migracji przypadający na 1000 mieszkańców danego powiatu.	os./1000os.	-3,1
społeczno-polityczne	**Dzieci objęte programem 500+, (piecset)**	Procentowy udział liczby dzieci, otrzymujących świadczenie z programu „Rodzina 500+" w łącznej liczbie dzieci w danym powiecie. Kryterium dochodowe na pierwsze dziecko w 2018 r. wynosiło 800zł dochodu netto miesięcznie na członka rodziny.	%	38,8
	Zawarte małżeństwa, (malz)	Liczba nowych małżeństw, zawartych w ciągu roku podana przez Urząd Stanu Cywilnego. Wartość została przeliczona na 1000 mieszkańców.	jedn./1000os.	5,0

	Liczba lekarzy, (lek)	Miara dostępności ochrony zdrowia w danym powiecie. Średnia dla Polski wynosiła w 2018 r. 6 lekarzy na 1000 mieszkańców.	os./1000os.	12,0
	Liczba studentów, (stud)	Miara działalności szkolnictwa wyższego w danym powiecie. Średnio w Polsce jest to 32 studentów na 1000 osób.	os./1000os.	60,0
wewnętrzne	**Liczba osób na mieszkanie, (osNAmieszk)**	Miara wielkości zapotrzebowania na rynku nieruchomości, określająca przeciętną liczbę osób, przypadających na jedno mieszkanie.	os./mieszk.	2,0
	Cena gruntów, (grunt)	Miara średnia dla transakcji kupna oraz sprzedaży nieruchomości gruntowych przeznaczonych pod zabudowę mieszkaniową. Dane dotyczące gruntów dostępne są na poziomie województw z uwzględnieniem podziału na dwie wartości: dla miast na prawach powiatu oraz pozostałych powiatów.	zł/m^2	219,4
	Powierzchnia mieszkań, (powNAos)	Miara podaży na rynku nieruchomości, określająca przeciętną powierzchnię użytkową mieszkań w przeliczeniu na osobę.	m^2/os.	34,0
	Powierzchnia nowych mieszkań, (powNowych)	Powierzchnia użytkowa mieszkań oddanych do użytkowania w 2018 r., przypadająca na 1000 osób.	m^2/1000os.	193,5
	Powierzchnia mieszkań sprzedanych, (powSprz)	Miara wielkości sprzedaży na rynku nieruchomości w danym powiecie. Informuje o tym ile m^2 lokali mieszkalnych zostało sprzedanych w ramach transakcji rynkowych w przeliczeniu na 1000 mieszkańców.	m^2/1000os.	690,2
	Udział rynku pierwotnego, (pierwWogolem)	Procentowy udział liczby mieszkań sprzedanych na rynku pierwotnym w sprzedaży ogółem. Informuje o tym ile % transakcji dotyczyło nowych lokali mieszkalnych.	%	22,0

** Skrót nazwy zmiennej, stosowany do obliczeń w środowisku RStudio*

Tak skomponowany zasób **21 zmiennych objaśniających** stanowi podstawę do stawiania kolejnych kroków w drodze do uzyskania satysfakcjonujących wyników badawczych. Wpływ niektórych z nich, jak np. liczby dzieci objętych świadczeniem z programu „Rodzina 500+" może budzić wątpliwości. Wprowadzenie determinant tego typu ma charakter eksperymentalny i daje szansę odkrycia pewnych zaskakujących i nietypowych współzależności.

2.2. Charakterystyka zbioru danych

Analiza wstępna to etap, który wymaga spojrzenia na **statystyki opisowe** danych. Ich podstawowe właściwości wraz z komentarzem i wyszczególnieniem powiatów ze skrajnymi wartościami danej zmiennej zamieszczono w Tabeli nr 2.

Tabela 2: Statystyki opisowe zbioru danych

Zmienna	Statystyka						
	Minimum	**Maksimum**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Współczynnik skośności**	**Kurtoza**
Cena za 1m² lokali mieszkalnych	1543,0 prudnicki	11763,0 m. Sopot	3143,5	2896,0	1105,5	2,8	13,3
Najdroższe mieszkania znajdują się w województwach: mazowieckim, małopolskim oraz pomorskim. Na większości obszaru Polski średnia cena lokali mieszkalnych wynosi jednak mniej niż 3143,5 zł/m².							
Wynagrodzenie brutto	3183,3 kępiński	8121,1 m. Jastrzębie Zdrój	4142,1	4017,2	561,2	2,7	12,7
Większość Polaków zarabia mniej niż średnie 4142,1 zł brutto. Najwyższymi wynagrodzeniami mogą pochwalić się mieszkańcy Jastrzębia-Zdroju, powiatu lubińskiego, Warszawy oraz powiatu bełchatowskiego.							
Dochody budżetów	616,6 grudziądzki	9609,7 m. Warszawa	2027,9	1153,6	2003,7	1,9	2,1
Najwięcej środków w przeliczeniu na mieszkańca otrzymują miasta na własnych prawach. Prawostronna skośność rozkładu tej cechy wynika z faktu, iż we wszystkich pozostałych powiatach kraju dochody są niższe od średniej krajowej.							

Wydatki budżetów	518,6 grudziądzki	9393,6 m. Warszawa	2082,7	1181,4	2064,4	1,9	1,9
Także wydatki na mieszkańca najwyższe są w miastach na prawach powiatu. Bardzo silna zmienność świadczy o dużym zróżnicowaniu struktury budżetów jednostek administracyjnych.							
Liczba przedsiębiorstw	52,4 chełmski	244,0 m. Warszawa	96,5	92,0	28,0	1,7	4,5
Najwięcej podmiotów, zarejestrowanych w rejestrze REGON, występuje w województwach: mazowieckim, wielkopolskim oraz śląskim. Prawostronna skośność rozkładu informuje o tym, iż w większości powiatów Polski jest to mniej niż 97 przedsiębiorstw różnej wielkości na 1000 osób. Zmienna ta jest najsilniej skorelowana ze zmienną objaśnianą (współczynnik korelacji liniowej Pearsona wynosi 0,69).							
Liczba nowych przedsiębiorstw	4,5 strzelecki	21,1 m. Warszawa	8,9	8,4	2,3	1,6	3,8
Jednostkami o najbardziej dynamicznym rozwoju przedsiębiorczości są Warszawa, Sopot oraz Wrocław. W 2018 roku w Polsce rejestrowanych było średnio około 9 nowych firm w przeliczeniu na 1000 mieszkańców. Liczba powstających przedsiębiorstw również jest silnie powiązana z ceną mieszkań (korelacja na poziomie 0,68).							
Współczynnik urbanizacji	0,0 skierniewicki, przemyski, suwalski	100,0 miasta na prawach powiatu	50,6	45,8	27,4	0,6	-0,6
Procentowy udział mieszkańców miast w ogólnej liczbie ludności w Polsce to około 50,6%. Niska, lecz dodatnia korelacja tego czynnika ze zmienną zależną świadczy o tym, że im więcej osób mieszka w mieście, tym nieruchomości mieszkaniowe w danym powiecie są droższe.							
Stopa bezrobocia	1,2 m. Poznań	24,3 szydłowiecki	7,8	6,9	4,1	0,9	0,6
Wysoka zmienność stopy bezrobocia świadczy o niejednorodności kraju pod względem natężenia tego problemu. Najwyższą wartość wskaźnika odnotowano w 2018 roku w województwach: warmińsko-mazurskim, kujawsko-pomorskim i podkarpackim.							
Gęstość zaludnienia	19,0 bieszczadzki	3757,0 m. Świętochłowice	369,5	90,5	654,3	2,5	5,7
W kraju obserwuje się ogromne zróżnicowanie gęstości zaludnienia (współczynnik zmienności równy 177%). Najwięcej osób na 1km^2 przypada w miastach na prawach powiatu, zaś najmniej w rejonie Bieszczad oraz Suwałk.							

Ludność w wieku produkcyjnym	55,8 m. Sopot	64,4 niżański	61,1	61,2	1,3	-0,4	0,4
Ze względu na demograficzny charakter czynnik ten wykazuje bardzo niski poziom zróżnicowania na terenie Polski (współczynnik zmienności to zaledwie 2,2%). Ku zaskoczeniu, im wyższy jest odsetek ludności w wieku produkcyjnym, tym średnia cena lokali mieszkalnych jest niższa (ujemna, umiarkowana korelacja).							
Przyrost naturalny	-10,6 hajnowski	9,4 kartuski	-1,1	-1,3	2,6	0,2	0,6
Średni wskaźnik przyrostu naturalnego w Polsce był ujemny. Oznacza to, iż w 2018 roku więcej osób zmarło, niż się urodziło. Bardzo wysoki współczynnik zmienności tej cechy sugeruje niejednorodność kraju pod względem różnicy urodzeń oraz zgonów.							
Saldo migracji	-8,1 kolneński	28,1 wrocławski	-1,2	-2,0	4,1	2,3	9,2
Jest to kolejna determinanta o bardzo wysokiej zmienności (ponad 341,7%). Co ciekawe, w 2018 roku najwięcej osób migrowało do powiatów, otaczających duże miasta, tj. do powiatu wrocławskiego, poznańskiego oraz gdańskiego. Jest to zjawisko suburbanizacji, czyli rozwoju stref podmiejskich.							
Dzieci objęte programem „500+”	33,8 m. Świnoujście	73,0 kolneński	54,9	55,7	7,1	-0,1	-0,5
Średnio ponad połowa osób niepełnoletnich w Polsce była objęta w 2018 r. świadczeniem z tytułu programu „Rodzina 500+”. Zmienna ta cechuje się ujemną, słabą korelacją ze średnią ceną za 1m^2 lokali mieszkalnych.							
Zawarte małżeństwa	3,6 jeleniogórski	6,4 limanowski	4,9	4,9	0,4	0,2	0,1
Liczba zawartych związków małżeńskich w przeliczeniu na 1000 mieszkańców jest bardzo podobna w całym kraju. Najwięcej ślubów odbyło się w trzech powiatach województwa małopolskiego: limanowskim, nowosądeckim i myślenickim.							
Liczba lekarzy	0,0 łomżyński	20,0 m. Lublin	4,0	3,2	3,0	2,2	6,4
Pod względem liczby lekarzy w przeliczeniu na 1000 mieszkańców przodują duże miasta, a w szczególności Lublin, Katowice, Rzeszów oraz Kraków. Wartość współczynnika asymetrii sugeruje, iż w większości powiatów na 1 osobę przypada mniej niż 4 specjalistów.							
Liczba studentów	0,0 lipski, bielski	198,0 m. Rzeszów	11,0	0,0	31,0	4,0	17,4

Polskie powiaty są bardzo zróżnicowane pod względem liczby studentów. Aż w 227 z nich nie ma ani jednej uczelni wyższej. Najwięcej osób w przeliczeniu na 1000 mieszkańców studiuje w Rzeszowie, Poznaniu, Katowicach, Lublinie , Krakowie oraz Wrocławiu. Dodatni współczynnik korelacji sugeruje, że mieszkania drożeją wraz ze wzrostem udziału studentów w ogóle ludności.

Liczba osób na mieszkanie	1,8 m. Warszawa	4,0 nowosądecki	2,9	2,9	0,4	-0,2	0,0

W Polsce na jeden lokal mieszkalny o dowolnej wielkości przypada średnio około 3 osoby. W większości jednostek administracyjnych chętnych jest jednak więcej. Ujemna korelacja wskazuje, że wraz ze wzrostem popytu na mieszkania ich ceny spadają.

Cena gruntów	8,0 powiaty woj. kujawsko-pomorskiego	219,4 miasta woj. pomorskiego	61,8	61,3	41,9	1,3	2,3

Najdroższe grunty pod zabudowę mieszkaniową występują nad morzem (woj. pomorskie), zaś najtańsze w rejonie Bydgoszcza i Torunia. Średnia cena nieruchomości tego typu w Polsce to około 62zł/m^2, czyli za działkę o wielkości 10 arów należy zapłacić średnio 62 tysiące złotych.

Powierzchnia mieszkań	22,5 kwidzyński	42,5 warszawski zachodni	27,7	27,3	3,0	1,1	2,4

Na rodzimym rynku nieruchomości oferowane jest średnio niecałe 28m^2 lokali mieszkalnych na 1 osobę. Największą podażą mieszkań (ze względu na powierzchnię) cechują się powiaty: warszawski zachodni, piaseczyński oraz wrocławski.

Powierzchnia nowych mieszkań	60,5 m. Bytom	1885,3 wrocławski	371,1	325,4	217,2	2,3	8,8

Powierzchnia użytkowa mieszkań oddanych do użytkowania w 2018 roku wynosiła średnio 371,1m^2 na 1000 osób. Najwięcej lokali przekazano do sprzedaży w otoczeniu dużych miast (w powiatach: wrocławskim, poznańskim oraz warszawskim zachodnim).

Powierzchnia mieszkań sprzedanych	0,0 skierniewicki	1090,5 m. Gdańsk	189,2	142,7	178,0	1,9	4,3

W różnych regionach kraju obserwuje się duże zróżnicowanie co do wielkości sprzedaży na rynku nieruchomości. Przeniesienie praw własności do największej powierzchni lokali w przeliczeniu na 1000 osób nastąpiło w 2018 roku w Gdańsku, Wrocławiu, Krakowie, Poznaniu oraz Warszawie. Zmienna ta jest silnie i dodatnio skorelowana z ceną mieszkań, co świadczy o istotnej współzależności obu wielkości.

Udział rynku pierwotnego	0,0	0,8	0,2	0,2	0,2	0,7	-0,6
	kielecki, sandomierski	kolbuszowski					
W Polsce średnio 20% lokali mieszkalnych, będących przedmiotem obrotu na rynku,. pochodziło w 2018 roku bezpośrednio od deweloperów. Współczynnik korelacji sugeruje, że im większy udział nowych mieszkań w ogólnej sprzedaży, tym ich średnia cena za 1m^2 jest wyższa. Aż w 88 powiatach rynek pierwotny nie miał żadnego udziału, co oznacza, że transakcje dotyczyły wyłącznie odsprzedaży z rąk wcześniejszego właściciela. Spośród dużych miast najwięcej nowych lokali sprzedano w Krakowie (70%), Gdańsku (65%) i w Poznaniu (62%).							

Przeanalizowanie powyższych wartości oraz zapoznanie się z komentarzami umożliwia zyskanie poglądu na ogół sytuacji w Polsce oraz jej zróżnicowanie przestrzenne. Stanowi to bazę wiedzy do zasadnej i świadomej „manipulacji" tymi wielkościami w celu uzyskania jak największej liczby informacji oraz wartościowych wniosków z badania. W jaki sposób podejść do wspomnianej „manipulacji" i o co tak naprawdę chodzi w tym zabiegu? Czy posiadając dane w takiej postaci i tylko dla jednego punktu czasowego (przekrojowe) da się wyjaśnić zróżnicowanie poziomu cen lokali mieszkalnych w różnych powiatach kraju?

Z pomocą przychodzą techniki statystyczne oraz narzędzia matematyczne, które umożliwiają proces tzw. **eksploracji danych** (ang. data mining). Pojęcie to, podobnie jak kilka innych narzędzi ekonometrycznych, zostanie przedstawione w następnym rozdziale.

3. Metodologia

Badania empiryczne, przeprowadzone w niniejszej pracy badawczej oparte są na współczesnych technikach wielowymiarowej analizy danych – **baggingu, algorytmie drzew decyzyjnych oraz na połączeniu obu, czyli lasach losowych**. Wszystkie z wymienionych znajdują się wśród metod, które umożliwiają skuteczną eksplorację danych, to znaczy pozyskiwanie z nich nowej wiedzy. Popularność tego zabiegu rodzi się z konieczności radzenia sobie z ogromnymi zbiorami przy jednoczesnym wzroście wymagań co do dokładności i nieomylności wyników. Okazuje się, że tradycyjne techniki ekonometryczne, jak chociażby klasyczna regresja liniowa często nie są w stanie rozwiązać problemów, przed którymi stają analitycy. W takich sytuacjach wykorzystuje się moc obliczeniową urządzeń oraz potencjał uczenia maszynowego. Umożliwiają one proces odkrywania reguł i zależności, a więc „drążenie" i odszukiwanie tego, co do tej pory pozostawało nieznane.

W kolejnych podrozdziałach zaprezentowano metody zastosowane w pracy w celu zdobycia informacji o zależnościach, które zachodzą na rynku mieszkaniowym w Polsce.

3.1. Drzewa decyzyjne

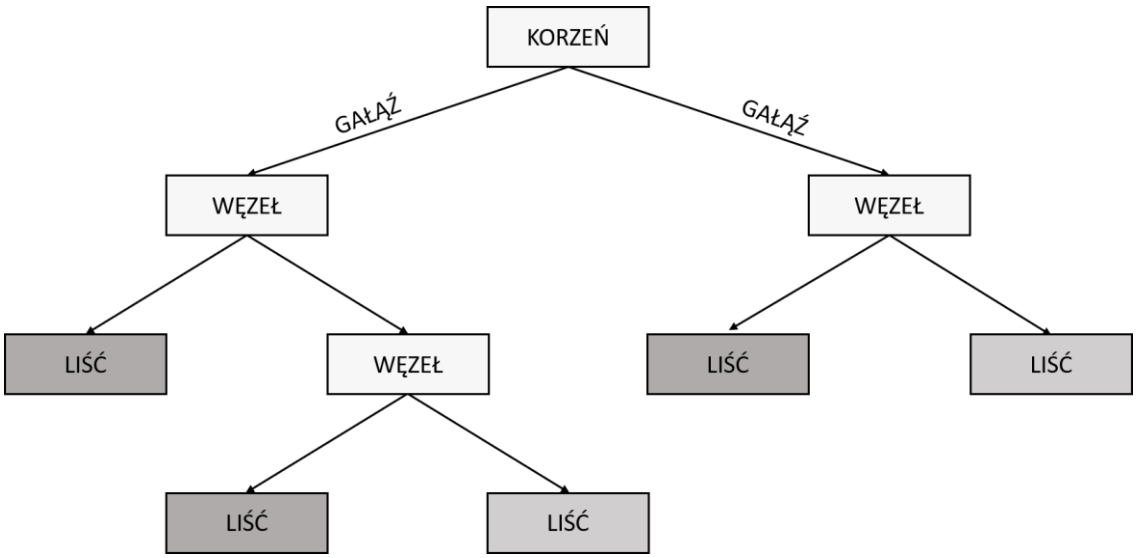
Od 1984 roku drzewa decyzyjne są jedną z najpopularniejszych i najczęściej stosowanych technik wielowymiarowej analizy danych. To właśnie wtedy wydana została książka „Classification and regression trees", w której **Leo Breiman**, Jerome Friedman, Charles J. Stone i Richard Olshen opracowali fundamentalne właściwości metod z grupy **CART (Classification and Regression Trees)**. Największy wkład w dziedzinę uczenia maszynowego miał pierwszy z autorów. Wyniki jego badań stały się uzupełnieniem luki między statystyką a informatyką, co otworzyło przed badaczami szeroki wachlarz nowych możliwości. Oprócz opracowania metody CART Breiman w znacznym stopniu przyczynił się do spopularyzowania pojęć takich jak bagging i lasy losowe[9]. W pierwszej kolejności należy jednak zapoznać się

[9] Bickel Peter, Jordan Michael, Rice John, *In Memoriam - Leo Breiman*, https://senate.universityofcalifornia.edu/_files/inmemoriam/html/leobreiman.htm, [14.03.2020].

z drzewami decyzyjnymi. Nie da się ukryć, że nazwa - nawiązująca jednocześnie do przyrody i sztucznej inteligencji – może być zastanawiająca i wzbudzać zaciekawienie.

Najprościej rzecz ujmując, algorytm **drzew decyzyjnych** jest graficzną metodą wspomagania procesów decyzyjnych. Elementy składowe pojedynczego drzewa to **korzeń** (zbiór wyjściowy), **gałęzie** (ścieżki, na których zbiór ulega podziałowi), **węzły** (będące punktami, w których podejmowana jest decyzja) oraz **liście**, czyli węzły końcowe (podzbiory, będące wynikiem obliczeń).

Rysunek 3: Elementy drzewa decyzyjnego



Źródło: Opracowanie własne

Model, prezentowany przez drzewo, powstaje na skutek podziału przestrzeni wielu zmiennych $x_1,\ldots,x_m$, w której znajdują się wszystkie analizowane obiekty na K rozłącznych podzbiorów $(P_1,\ldots,P_k)$ tej przestrzeni[10]. W zależności od postaci zmiennej zależnej y model taki może mieć charakter klasyfikacyjny lub regresyjny.

W pierwszym przypadku zmienna y jest przedstawiona na skali nominalnej, co oznacza, że obiektom nadawana jest pewna etykieta według określonych zasad. Przykładowo, powiaty Polski można podzielić na drogie i tanie pod względem cen mieszkań, przyjmując, że do pierwszej kategorii zaliczane są te jednostki, w których 1m^2 kosztuje więcej niż 4,5 tysiąca złotych. Zmienna zależna przyjmie więc jedną z dwóch kategorii, określających nieruchomości

[10] Gatnar Eugeniusz, Walesiak Marek, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych,* Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2004, s. 123.

(„drogie” lub „tanie”). Graficzną interpretacją modelu tego typu jest **drzewo klasyfikujące**, którego parametry obliczane są według zasady „głosowania”. Oznacza to, że w podzbiorze P_k zmienna y przyjmuje najczęściej występującą wartość:

$$\mathrm{a_k} = \arg\max_{\mathrm{t}} \{ \mathrm{p(t|k)} \}, \tag{1}$$

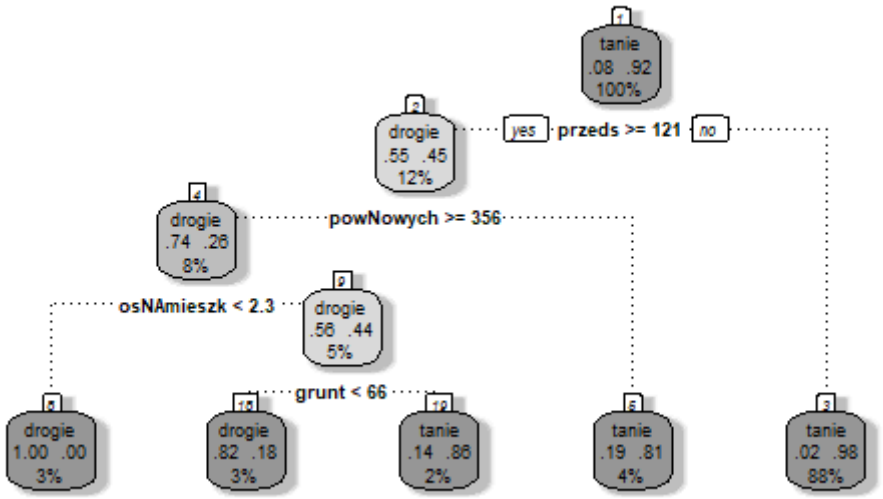
gdzie $p(t|k)$ jest prawdopodobieństwem tego, że obserwacja k z podzbioru P_k należy do kategorii t. Procedurę rekurencyjnego podziału, wykorzystywanego do klasyfikacji, można sprowadzić do następujących kroków:

1. Sprawdzenie warunku początkowego, jakim jest posiadanie wielowymiarowej przestrzeni zmiennych X, dla której wartości zmiennej objaśnianej y nie są jednorodne.
2. Dokonanie wszystkich możliwych podziałów zbioru X na rozłączne podzbiory $(P_1,\ldots,P_k)$ przy uwzględnieniu wartości kolejnych zmiennych $x_1,\ldots,x_m$.
3. Wybór najlepszego podziału zgodnie z przyjętym kryterium.
4. Podzielenie zbioru X w optymalny sposób.
5. Powtórzenie kroków 1-4 dla każdego z podzbiorów $(P_1,\ldots,P_k)$.
6. Warunkiem zakończenia jest osiągnięcie jednorodności obiektów w uzyskanych podzbiorach. Można także określić kryterium stopu, czyli minimalną liczność obiektów w węzłach końcowych.[11]

Aby zobrazować wyżej opisany mechanizm klasyfikacji posłużono się zbiorem danych przygotowanym na potrzeby niniejszej pracy. Warto zaobserwować w jaki sposób powiaty Polski są przydzielane do odpowiednich klas, określających poziom cen na rynku nieruchomości. Dla pierwszego drzewa (Rysunek 5) minimalna liczba powiatów w węźle końcowym wynosi 5:

[11] Gatnar Eugeniusz, Walesiak Marek, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych,* Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2004, s. 108.

Rysunek 4: Drzewo klasyfikujące o minimalnej liczbie obiektów w węźle równej 15



Źródło: Opracowanie własne w środowisku RStudio

Przedstawioną ilustrację można interpretować w następujący sposób:

- najwyżej znajdujący się element to **korzeń drzewa**, w którym skupiona jest cała próba. W jej skład wchodzi 380 polskich powiatów. Procent (tutaj 100%) oznacza odsetek całkowitej populacji, który zakwalifikowany jest do danego podzbioru. Liczby powyżej (tu .08 i .92) oznaczają natomiast, że odpowiednio 8% i 92% obserwacji, zawartych w węźle przechodzi w stronę lewej i prawej gałęzi,

- zaprezentowane drzewo klasyfikujące składa się z **4 węzłów** (kształtów, w których podejmowana jest decyzja o kierunku przesunięcia elementów), oznaczonych odpowiednią etykietą (powiaty „drogie” i „tanie”),

- pod każdym z węzłów znajduje się **kryterium podziału**, np. *przeds>=121*, co oznacza, że powiaty w tym punkcie dzielone są na dwie klasy względem liczby przedsiębiorstw na 1000 mieszkańców,

- podgrupa, która **spełnia** dany warunek przesuwana jest zawsze **w lewą stronę**,

- **istotność zmiennych**, stosowanych w kolejnych podziałach jest coraz mniejsza. Oznacza to, że im wyżej na grafice znajduje się dana determinanta tym jest ona ważniejsza z punktu widzenia skuteczności algorytmu. Na sposobach pomiaru istotności skupiono się w rozdziale 5,

- nad każdym z kształtów znajduje się cyfra, oznaczająca **numer gałęzi** (1, 2, 3, 4, 5, 9, 15 i 19). Dla rozwiania wątpliwości, braki w numeracji wynikają z dostrajania drzewa, w trakcie którego algorytm usuwa niektóre gałęzie i poszukuje lepszych rozwiązań,

- drzewo składa się z **5 liści**, czyli w wyniku zwraca 5 podprób. Każda z nich zawiera co najmniej ustalone minimum obiektów (5),

- liście i węzły stanowią razem **9 wierzchołków** graficznej prezentacji modelu.

Model drzewa klasyfikującego jest źródłem **reguł decyzyjnych**, które umożliwiają scharakteryzowanie powstałych klastrów. Reguły te są podstawą podziału zbioru i mają postać koniunkcji o przykładowej postaci:

$$(x_1 < w_1) \wedge (x_2 < w_2) \wedge \dots (x_m < w_m), \tag{2}$$

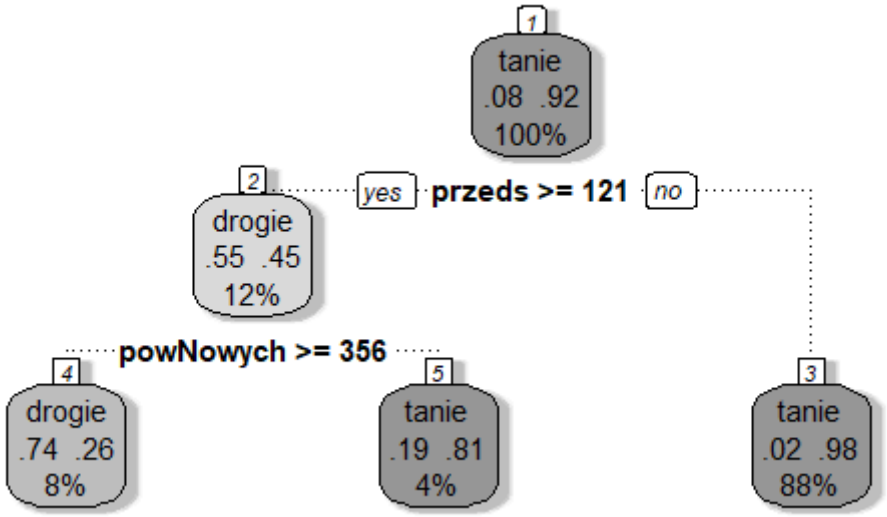
gdzie $w_1, \dots, w_m$ są wartościami kolejnych zmiennych objaśniających.

Dla drzewa z Rysunku nr 5 reguła taka może brzmieć następująco:

„Jeżeli liczba przedsiębiorstw na 1000 mieszkańców jest większa lub równa 121 i powierzchnia nowych mieszkań na 1000 mieszkańców wynosi mniej niż 356m^2 to lokale mieszkalne w takim powiecie są tanie."

W analizowanym przykładzie według takiej reguły do węzła z numerem 6 trafiło 16 powiatów, wśród których znalazły się: Łódź, Legnica, Katowice czy powiat gryficki (woj. zachodniopomorskie). Warto być świadomym, iż 3 obiekty zostały zaklasyfikowane błędnie i gdyby był możliwy dalszy podział trafiłyby zapewne do odrębnej podpróby. Zanim nastąpi przejście do modeli regresyjnych warto zobrazować skutki „przycięcia" drzewa. Zabieg ten to nic innego jak zamienienie niektórych wierzchołków na końcowe liście, co skutkuje zmniejszeniem głębokości drzewa. Sztuka polega natomiast na znalezieniu optimum między wielkością drzewa a precyzją klasyfikacji. Założono, że wcześniejszemu przykładowi należy narzucić maksymalną głębokość równą 5 wierzchołków (Rysunek 6).

Rysunek 5: Drzewo klasyfikujące po przycięciu do 5 węzłów



Źródło: Opracowanie własne w środowisku RStudio

Łatwo zauważyć, że już rozstrzygnięcie drugiego kryterium decyzyjnego prowadzi do uzyskania trzech finalnych klas. Innym kluczowym aspektem jest oczywiście pomiar skuteczności drzew. Na potrzeby jej weryfikacji podzielono zbiór wyjściowy na uczący (70% obserwacji) i testowy (30%). Aby upewnić się co do zasadności takiego podziału sprawdzono czy odsetek powiatów drogich i tanich w obu podzbiorach jest zbliżony. Okazało się, że proporcje zostały zachowane wręcz idealnie – zarówno w próbie uczącej jak i testowej obiekty „tanie” oraz „drogie” stanowiły odpowiednio 92% i 8% próby. Po tych ustaleniach na 101 powiatach testowych przeprowadzono prognozowanie. Wyuczony model drzewa decyzyjnego przydzielał obiekty z próby testowej do odpowiednich klas. Tabela nr 3 jest **macierzą pomyłek**, służącą do oceny jakości klasyfikacji:

Tabela 3: Macierz pomyłek dla klasyfikacji drzewa decyzyjnego

		Wartość przewidywana	
Wartość rzeczywista		tanie	drogie
	tanie	**92**	**1**
	drogie	**5**	**3**

Źródło: Opracowanie własne

Co ciekawe – model okazał się bardzo skuteczny. Obliczona jako odsetek poprawnie zaklasyfikowanych danych dokładność klasyfikacji wyniosła w przybliżeniu aż **94,1%**, co daje podstawę do twierdzenia o wysokiej efektywności klasyfikatorów. Na tym można zakończyć analizę przypadku, w którym zmienna y wyrażona jest jakościowo. W przeciwnym razie, czyli dla ilościowej zmiennej zależnej, stosuje się model regresji, którego graficzna prezentacja wygląda nieco inaczej.

Drzewa regresyjne to drugi rodzaj metod z grupy CART. Jak już wspomniano, y w modelu przez nie prezentowanym wyrażone jest ilościowo (na jednej ze skal mocnych). Na potrzeby modelu regresji zbiór danych przywrócono do pierwotnej postaci, w której zmienną zależną jest cena za 1m^2 lokali mieszkalnych w danym powiecie (wartości z przedziału [1542;11763] zł). Schemat kroków algorytmu jest taki sam jak dla problemu klasyfikacji. Różnica dotyczy natomiast sposobu pozyskiwania parametrów a_k, które w tym przypadku uzyskiwane są zgodnie z zasadą uśredniania:

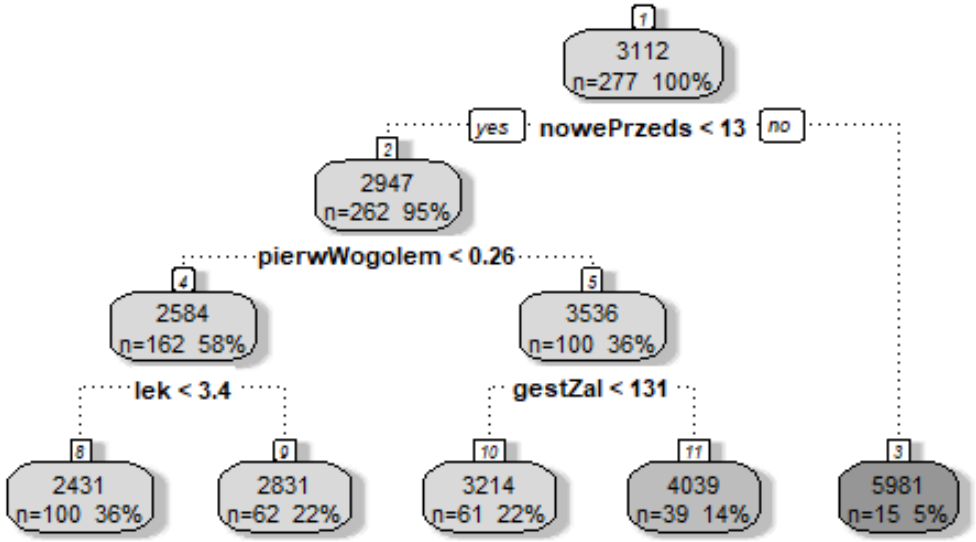
$$\mathrm{a_k} = \frac{1}{\mathrm{L(k)}} \sum_{\mathrm{i=1}}^{\mathrm{L(k)}} \mathrm{y_i} \tag{3}$$

,gdzie $L(k)$ jest liczbą obserwacji w k-tym podzbiorze[12]. Oznacza to, że estymatorem regresji na podstawie drzewa jest średnia wartości w liściu (podpróbie końcowej).

Na potrzeby oceny jakości prognoz, tak samo jak poprzednio, zbiór wyjściowy podzielono na uczący i testowy. Dla zachowania przejrzystości drzewo przycięto, ustalając, że w liściach ma znajdować się minimum 15 jednostek. Wynik, uzyskany dla 277 treningowych powiatów przedstawia poniższa grafika:

[12] Gatnar Eugeniusz, Walesiak Marek, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych,* Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2004, s. 108.

Rysunek 5: Drzewo regresyjne, zbudowane na zbiorze uczącym



Źródło: Opracowanie własne w środowisku RStudio

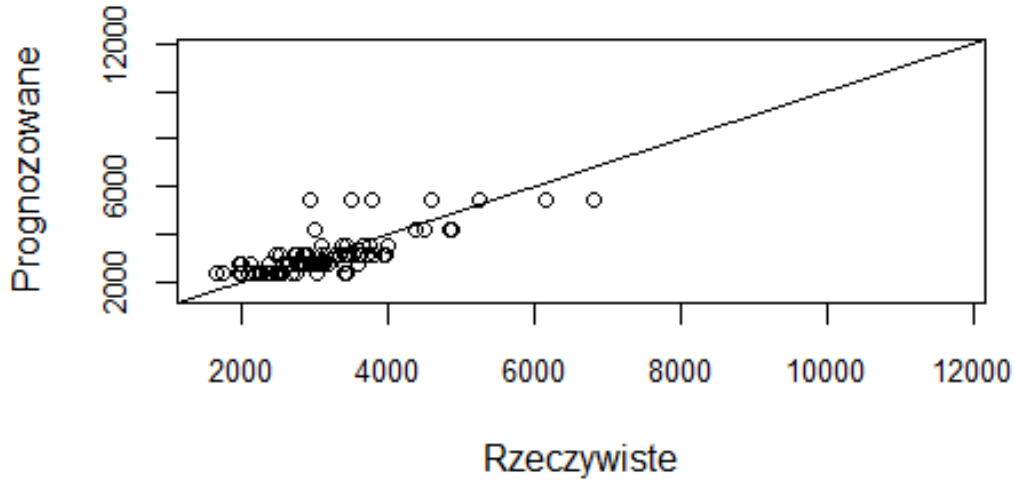
Już na pierwszy rzut oka widoczne są pewne różnice w stosunku do drzewa klasyfikującego. Po pierwsze – zamiast etykiet węzły oznaczone są konkretnymi wartościami (np. dla korzenia to 3112). Liczba ta to średnia wartości zmiennej zależnej dla obserwacji, które trafiły do danego podzbioru. Po drugie – reguły decyzyjne przyjmują nieco inną formę. Jedna z nich dla drzewa z Rysunku nr 5 brzmi:

„Jeżeli liczba nowych przedsiębiorstw na 1000 mieszkańców jest niższa niż 13, sprzedaż nowych mieszkań na rynku pierwotnym stanowi mniej niż 26% oraz na 10000 osób przypada co najmniej 34 lekarzy to w danym powiecie za 1m^2 mieszkania należy zapłacić około 2831zł."

Koniunkcja ta jest prawdziwa dla podzbioru 62 powiatów, co stanowi ponad 22% korzenia. Znalazły się tutaj m.in. Jelenia Góra, Krosno, Zabrze oraz powiaty: sandomierski, sanocki i opolski. Widać jeszcze, że zmieniła się hierarchia istotności zmiennych. Tym razem to liczba nowych przedsiębiorstw na 1000 mieszkańców ma dominującą wagę przy ustalaniu struktury podziału powiatów względem cen nieruchomości. Dla przypomnienia – temat pomiaru istotności predyktorów został poruszony w rozdziale o tytule „Badanie empiryczne".

Mając wyodrębnioną próbę testową można przejść do wykonania prognozy, której wyniki ilustruje poniższy wykres:

Rysunek 6: Wartości rzeczywiste i prognozowane w modelu drzewa regresyjnego



Źródło: Opracowanie własne w środowisku RStudio

Aby zbadać jakość przewidywania modelu nie można w problemie regresji posłużyć się macierzą błędów. Jedną z miar, która zastępuje to narzędzie jest tzw. błąd $\boldsymbol{MAPE}$ (Mean Absolute Percentage Error), czyli **średni bezwzględny błąd procentowy**. Oblicza się go według wzoru:

$$\mathrm{MAPE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left|\frac{\mathrm{y_i}-\mathrm{y_i^p}}{\mathrm{y_i}}\right| \cdot 100\%, \tag{4}$$

gdzie y_i oznacza rzeczywiste, a y_i^p prognozowane wartości zmiennej zależnej y. Obliczony dla próby testowej błąd prognozy modelu wyniósł ponad **15,4%**, co daje podstawy do twierdzenia, iż model nie jest najlepszej jakości. Aby to potwierdzić należałoby jednak zweryfikować problem dokładniej. Drugim podejściem do oceny skuteczności prognostycznej jest pomiar błędu $\boldsymbol{RMSE}$ (Root Mean Square Error), zwanego **pierwiastkiem błędu średniokwadratowego**. Formułę obliczania tej wielkości można zapisać następująco[13]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y_i}-\mathrm{y_i^p}\right)^2}, \qquad \mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{5}$$

[13] *Metody prognozowania: Jakość prognoz*, https://m6.pk.edu.pl/materialy/mp/MP_02_bledy_prognozy.pdf, [12.04.2020].

Dla analizowanego przykładu błąd wyniósł około **670,8 zł**, co oznacza, że o taką kwotę wartości estymowane przez drzewo regresyjne różnią się średnio od wartości rzeczywistych. Niestety, wielkość ta nie wydaje się satysfakcjonująca i potwierdza wcześniejsze obawy co do stosunkowo słabej jakości modelu.

Czy da się ją w takim wypadku w jakiś sposób poprawić? Co można zrobić, aby zminimalizować wartości błędów, a co za tym idzie – zmaksymalizować precyzję? Należy zwrócić uwagę na kluczowy problem, jakim jest fakt, iż algorytm w losowy sposób przydziela obiekty do próby uczącej i testowej. Błędne jest zatem wysuwanie wniosków co do jakości przewidywań po jednokrotnym wykonaniu obliczeń. Każdorazowe powtórzenie algorytmu skutkuje utworzeniem nieco innego drzewa. Przykładowo – drugie podejście do omawianego przykładu dało gorsze wyniki błędów ($MAPE$ = 17,4% i $RMSE$ = 890,1 zł), a trzecie – nieco lepsze ($MAPE$ = 15,1% i $RMSE$ = 609,7 zł). Do pomysłów udoskonalenia tej klasycznej metody z dziedziny eksploracji danych sięgnięto w dalszej części opracowania.

3.2. Metody agregacji drzew

Nawiązując do wcześniejszych rozważań, drzewa regresyjne i klasyfikacyjne, wśród wielu zalet, posiadają także pewną istotną wadę – brak stabilności. Oznacza to, że nawet niewielka zmiana wartości w zbiorze uczącym może spowodować znaczną modyfikację wyników. Konsekwencją tej wady jest niska trafność prognostyczna pojedynczych instancji drzew. Nie stanowi to jednak problemu nie do rozwiązania. Aby wyeliminować tę niedoskonałość można połączyć kilka modeli w jeden zagregowany zespół. Technika ta, zwana **agregacją**, opiera się na 2 prostych krokach:

1. Wyodrębnienie N zbiorów uczących i zbudowanie dla nich modeli w postaci drzew $D_1(x),\ D_2(x)\ , \dots , D_n(x)$.
2. Połączenie uzyskanych klasyfikatorów w model zagregowany $D^*(x)$[14].

[14] Gatnar Eugeniusz, Walesiak Marek, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych,* Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2004, s. 124.

Agregacja może mieć zarówno **deterministyczny**, jak i **stochastyczny** charakter. Do pierwszej grupy zalicza się tzw. **boosting**. Pojęcie to oznacza rodzinę algorytmów uczenia nadzorowanego, w których stosowany jest system wag obiektów. Model zagregowany powstaje poprzez budowanie prób uczących, które zawierają te same obiekty co zbiór bazowy, jednak ich wagi są w odpowiedni sposób modyfikowane. Istotą działania boostingu jest więc iteracyjne doskonalenie reguł decyzyjnych na tym samym podzbiorze danych. W każdej iteracji reguła, która jest zbyt słaba, aby stanowić źródło samodzielnej decyzji zostaje zamieniona na mocniejszą. Dzieje się to poprzez przyłożenie wyższej wagi do błędnie zaklasyfikowanych danych, aby algorytm mógł zwrócić na nie większą uwagę. Zaletą boostingu jest dążenie do maksymalnej precyzji, co pozwala uzyskać możliwie najmocniejsze reguły decyzyjne. Najczęściej stosowane algorytmy z rodziny agregacji deterministycznej to:

1. **Adaptive Boosting (AdaBoost)** – wykorzystujący średnią ważoną. Każdej z n obserwacji przydziela wagę $w_i = \frac{1}{n}$. W kolejnych iteracjach poprawnie zaklasyfikowanym obserwacjom zwiększa wagę o jednostkę, a błędnie przydzielonym – analogicznie zmniejsza.
2. **Gradient Boosting (GMB)** – używający funkcji gradientowej z pochodną pierwszego rzędu. Od AdaBoost różni się sposobem uwzględniania obserwacji niepoprawnie zaklasyfikowanych. W każdej iteracji wartość pochodnej jest minimalizowana, a co za tym idzie – funkcja zmierza do minimum. Ryzyko tej metody, związane z nieodpowiednim doborem parametrów, może prowadzić do nadmiernego dopasowania.
3. **Extreme Gradient Boosting (XGBoost)** – będący zaawansowaną wersją GBM. Przewaga tej techniki polega na wyeliminowaniu problemu przetrenowania modelu. Jest to możliwe dzięki zastosowaniu operacji takich jak regularyzacja, liczenie równoległe czy walidacja krzyżowa.

Nie da się ukryć, że o wyżej przedstawionych algorytmach można by było napisać obszerne opracowanie. Ze względu na to, że agregacja deterministyczna nie została zastosowana w późniejszych obliczeniach zdecydowano poświęcić jej jedynie kilka słów wstępnych.

Na potrzeby właściwego studium przypadku ważne jest jednak wyczerpujące omówienie i zrozumienie zasady działania algorytmów o charakterze **stochastycznym**. Najczęściej stosowanym spośród nich sposobem łączenia modeli jest agregacja bootstrapowa, określana jako **bagging**. Na taki pomysł rozwiązania problemu niestabilności drzew wpadł Leo Breiman w 1996 roku. W odróżnieniu od boostingu, metoda ta bazuje na budowaniu drzewa na nowym podzbiorze danych w każdej iteracji. Podzbiory te, zwane uczącymi, powstają na skutek losowania ze zwracaniem z wyjściowego zbioru danych. Część obserwacji, która nie trafia do próby uczącej tworzy tzw. zbiór walidacyjny (out of bag). W agregacji bootstrapowej obowiązuje zasada tworzenia podzbiorów o takim samym rozmiarze. Każdy z nich liczy po n obiektów, których waga wynosi początkowo $w_i = \frac{1}{n}$.

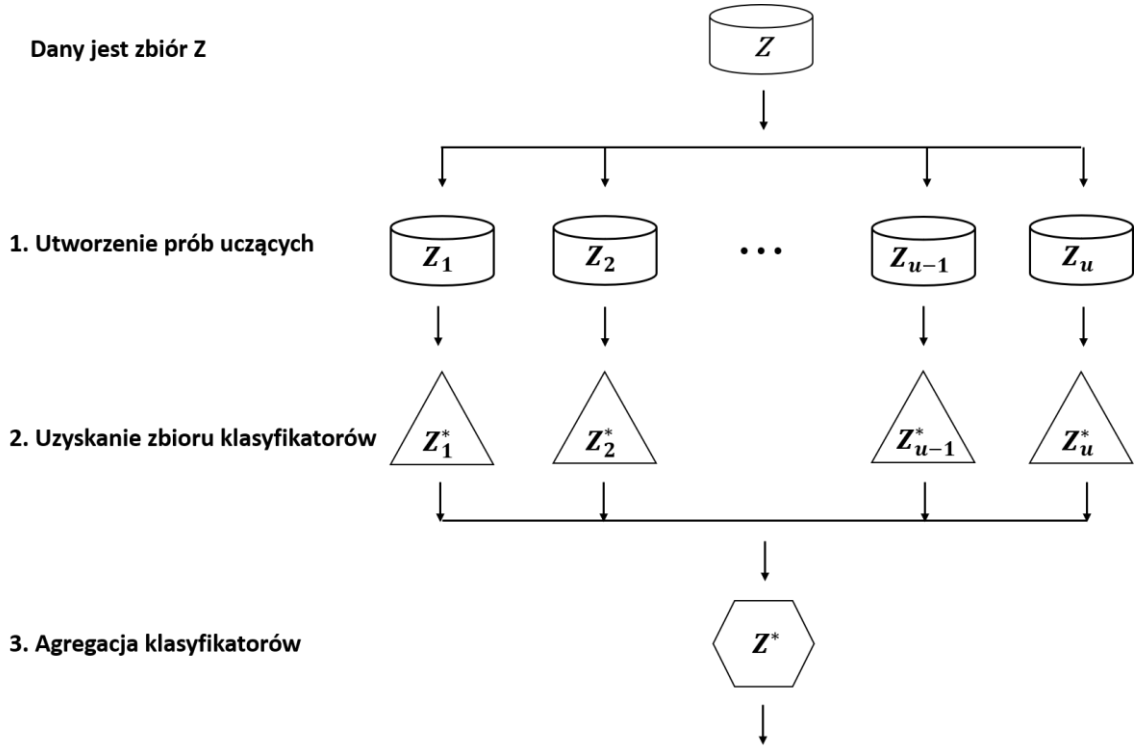
Działanie baggingu można najogólniej przedstawić w kilku punktach[15]:

1. Ustalenie liczby U prób uczących $Z_1, \dots, Z_u$ dla zbioru bazowego Z.
2. Dla $u \in \{1, \dots, U\}$:
   - losowanie ze zwracaniem n obiektów ze zbioru bazowego,
   - utworzenie przez wylosowane obiekty próby Z_u, przy czym każda z prób jest takiej samej liczności jak zbiór Z,
   - zbudowanie drzewa $D_u(x)$ dla każdej próby uczącej i dokonanie predykcji zgodnie z przyjętym modelem.
3. Wybór klasyfikatorów dla każdej obserwacji poprzez uśrednienie wszystkich predyktorów $Z_1^*, \dots, Z_n^*$.
4. Utworzenie modelu zagregowanego Z^*.

Algorytm zilustrowano na Rysunku nr 8:

[15] Gatnar Eugeniusz, Walesiak Marek, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych,* Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2004, s. 125.

Rysunek 7: Ilustracja działania baggingu



Źródło: Opracowanie własne na podstawie: Gatnar Eugeniusz, Walesiak Marek, Metody statystycznej analizy wielowymiarowej w badaniach marketingowych, *Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2004, s. 103-135.*

Agregacja bootstrapowa jest wykorzystywana do zwiększania wiarygodności współczynników w niemalże wszystkich metodach uczenia nadzorowanego. Zabieg ten skutecznie redukuje wariancję, poprawiając jednocześnie precyzję modelu. Nic więc dziwnego, że coraz częściej znajduje zastosowanie przy rozwiązywaniu problemów regresji i klasyfikacji. Bagging jest także ściśle powiązany z algorytmem lasów losowych, o którym mowa w następnym podrozdziale.

3.3. Lasy losowe

Po pięciu latach zbierania doświadczeń, w 2001 roku, Leo Breiman zaobserwował pewną znaczącą **wadę** działania baggingu. Zauważył on, że problem pojawia się, gdy niektóre zmienne objaśniające w zbiorze danych są bardzo silnie skorelowane ze zmienną zależną. W takim przypadku istnieje ryzyko, że każde kolejne drzewo będzie dokonywać pierwszego podziału ze względu na ten sam predyktor. Na skutek tego, mimo utworzenia wielu drzew, ich struktura

będzie bardzo zbliżona, a co za tym idzie – model nie będzie radził sobie z uogólnianiem wyników poza pierwotnym zbiorem danych. Innym problemem dla baggingu jest zmierzenie się z dużą liczbą zmiennych objaśniających. W efekcie tego, uzyskanie ostatecznego wyniku jest czasochłonne i skomplikowane. Jaka więc była odpowiedź amerykańskiego statystyka na te niedociągnięcia?

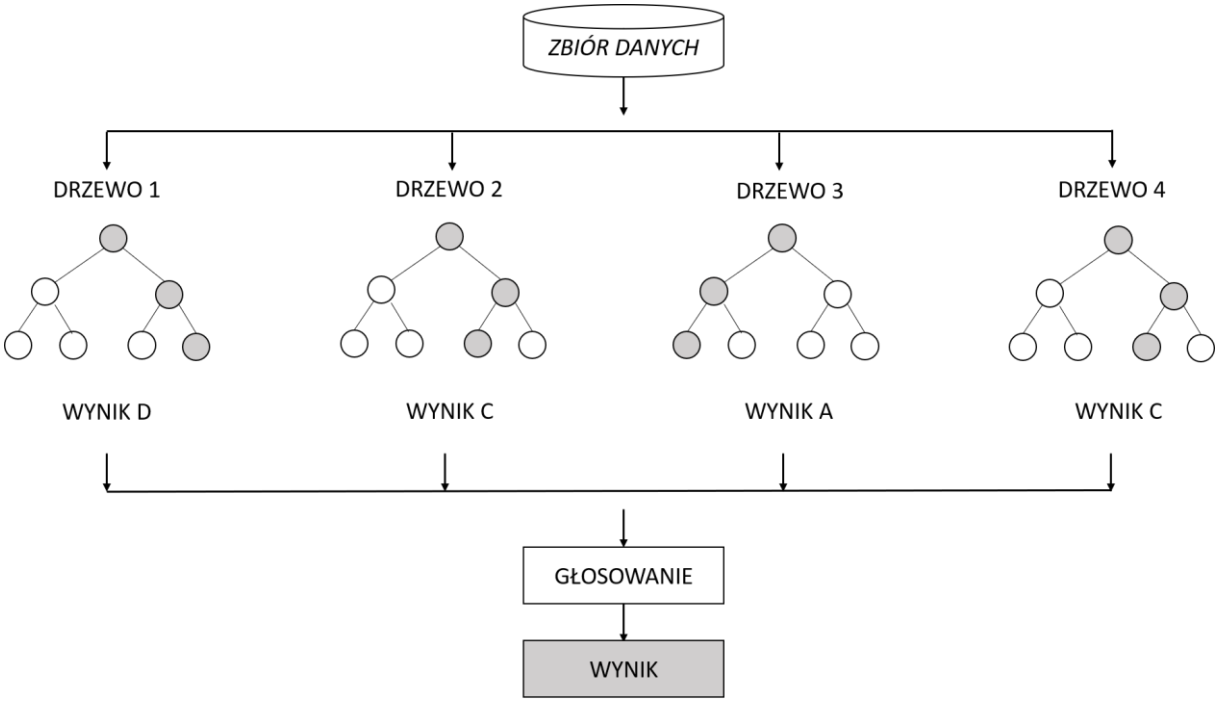
Prawie 20 lat temu Breiman zaproponował jedną z najbardziej wszechstronnych metod uczenia maszynowego, którą wdzięcznie nazwał **lasami losowymi**. Algorytm ten, podobnie jak wcześniej opisane, skutecznie radzi sobie z problemami regresji i klasyfikacji. Sposób działania lasów do złudzenia przypomina agregację bootstrapową. Próby uczące tworzone są analogicznie jak w przypadku baggingu – przez losowanie ze zwracaniem. Zasadnicza różnica polega na tym, iż do konstruowania kolejnych modeli drzew decyzyjnych nie są wykorzystywane wszystkie zmienne niezależne, a jedynie część z nich. Ich wybór odbywa się na zasadzie losowania. Można natomiast narzucić algorytmowi liczbę predyktorów, stosowanych w kolejnych podziałach. Innym parametrem, który można modyfikować jest liczba drzew w lesie. Jak te wartości mają się do jakości wyników? Odpowiedź na to pytanie zawarta jest w 5 rozdziale pracy.

Podsumowując, w algorytmie lasów losowych można wyróżnić 5 kroków:

1. Ustalenie liczby U prób uczących $Z_1, \dots, Z_u$ dla zbioru bazowego Z (a tym samym liczby drzew w lesie).
2. Ustalenie liczby S zmiennych objaśniających $x_1 \dots, x_s$, wykorzystywanych do utworzenia drzewa dla każdej próby.
3. Dla $u \in \{1, \dots, U\}$:
   - losowanie ze zwracaniem n obiektów ze zbioru bazowego,
   - utworzenie przez wylosowane obiekty próby Z_u, przy czym każda z prób jest takiej samej liczności jak zbiór Z,
   - zbudowanie drzewa $D_u(x)$ dla każdej próby uczącej i dokonanie predykcji zgodnie z przyjętym modelem.
4. Wybór klasyfikatorów dla każdej obserwacji poprzez uśrednienie wszystkich predyktorów $Z_1^*, \dots, Z_n^*$ (regresja – „uśrednianie", klasyfikacja – „głosowanie")
5. Utworzenie modelu zagregowanego Z^*.

Sposób uzyskiwania ostatecznego wyniku predykcji dla całego modelu lasu losowego przedstawiono na poniższym schemacie:

Rysunek 8: Schemat predykcji lasu losowego



Źródło: Opracowanie własne na podstawie https://predictivesolutions.pl/jak-udoskonalic-algorytm-drzew decyzyjnych

Wynikiem głosowania jest (w problemach klasyfikacyjnych) wartość przewidziana przez największą liczbę pojedynczych drzew lub (w problemach regresyjnych) średnia wartość dla wszystkich drzew w lesie[16]. Cały fenomen techniki lasów losowych polega na tym, iż rozwiązuje ona problem braku stabilności modeli predykcyjnych. Ceną, jaką trzeba za to zapłacić jest dłuższy czas obliczeń niż w przypadku klasycznego drzewa oraz trudniejsze w interpretacji i bardziej rozbudowane reguły decyzyjne. Zachwyt nad satysfakcjonującymi wynikami może być jednak wart takiego poświęcenia. W rozdziale 5 okaże się jak skuteczne są lasy losowe na przykładzie badania rynku nieruchomości w powiatach Polski. Czy dzięki nim udało się osiągnąć cel pracy i odpowiedzieć na nurtujące pytania? A może obiecujący algorytm okazał się jedynie rozczarowaniem?

[16] Grabowska Ewa, *Jak udoskonalić algorytm drzew decyzyjnych?,* https://predictivesolutions.pl/jak-udoskonalic-algorytm-drzew-decyzyjnych, [13.04.2020].

4. Główne hipotezy badawcze

Już pierwsze spojrzenie na zestaw danych prowadzi do pojawienia się pewnych domysłów co do zależności zachodzących na rynku mieszkaniowym. Ich formalna weryfikacja w toku badania wymaga sformułowania zestawu hipotez badawczych. Dla przypomnienia – problem przewodni dotyczy wyjaśnienia powodów nierówności cen mieszkań na przekroju powiatów Polski. Do analizy głównych podejrzeń wykorzystany został (znany do tej pory jedynie w teorii) algorytm regresyjnych lasów losowych.

Hipoteza 1. Na wysokość średniej ceny 1m^2 lokali mieszkalnych w poszczególnych powiatach statystycznie największy wpływ mają: liczba przedsiębiorstw (zarówno wszystkich jak i nowo powstałych), sprzedana powierzchnia mieszkań oraz odsetek transakcji na rynku pierwotnym.

Zmienne te są najsilniej skorelowane ze zmienną zależną. Aby zbadać słuszność tej hipotezy posłużono się parametrami, określającymi wagi poszczególnych predyktorów dla algorytmu lasów losowych. Czynniki o dominujących wartościach są jako pierwsze brane pod uwagę przy uzgadnianiu struktury kolejnych gałęzi drzew decyzyjnych.–

Hipoteza 2. Niewielką rolę w kształtowaniu struktury rynku mieszkaniowego odgrywają: odsetek dzieci objętych programem „500+", liczba zawartych małżeństw oraz liczba studentów.

Są to jedynie subiektywne spostrzeżenia, które powstały na bazie literatury (dotyczącej rynku nieruchomości) oraz artykułów internetowych. Eksperymentalne wprowadzenie tych czynników do zbioru danych wynika z zaciekawienia.

Hipoteza 3. Im więcej drzew w lesie losowym, tym jakość modelu jest wyższa (niższy błąd prognozy).

Wydaje się, że im więcej pojedynczych klasyfikatorów tym lepiej. Mimo tego, po przekroczeniu ich określonej liczności dalszy rozrost lasu może jedynie zwiększać złożoność obliczeń (bez istotnej poprawy jakości modelu). Należy zweryfikować czy wraz ze wzrostem liczby tworzonych drzew błąd prognozy zbliża się do pewnej granicznej wartości.

Hipoteza 4. Dzięki możliwościom, jakie daje zastosowanie techniki lasów losowych, możliwe jest uzyskanie skuteczniejszego modelu niż w przypadku regresji liniowej.

Klasyczna metoda najmniejszych kwadratów wymaga, aby zbiór danych spełniał wiele założeń. Tylko wtedy estymator KMNK jest zgodny, nieobciążony i najefektywniejszy w klasie liniowych estymatorów wektora parametrów. Nowatorska technika eksploracji danych jest natomiast całkowicie wolna od wstępnych restrykcji. Wydaje się, że błędy prognozy dla modelu lasów losowych będą niższe niż w tradycyjnej metodzie.

Hipoteza 5. Lepszym podejściem do analizy rynku mieszkań jest utworzenie dwóch lasów losowych – osobno dla 66 miast i 314 pozostałych powiatów.

Jako „lepsze" podejście uznaje się to, dzięki któremu możliwe będzie uzyskanie modeli o mniejszym błędzie prognozy. Jakość przewidywań może poprawić się po podzieleniu pierwotnej bazy danych na dwie i skonstruowaniu dla nich oddzielnych modeli regresyjnych. Istnieje jednak ryzyko pojawienia się problemu z wielkością zbiorów. Czy próba, zawierająca tylko 66 obiektów (miasta) nie okaże się zbyt mała i czy będzie dla niej możliwa skuteczna implementacja algorytmu?

Hipoteza 6. **Lasy losowe są techniką analityczną, charakteryzującą się wysoką odpornością na występowanie w zbiorze danych obserwacji odstających.**

Słuszność tego stwierdzenia można sprawdzić usuwając doświadczalnie te powiaty, które wyróżniają się skrajnie niskimi lub wysokimi cenami mieszkań. Porównanie błędów prognozy po takiej modyfikacji pozwoli ocenić czy występowanie w próbie odległych obiektów istotnie zaburza dokładność predykcji.

Podejrzenia, wypunktowane wyżej, zostaną potwierdzone bądź obalone w ostatniej części pracy. W następnej sekcji zaprezentowano wykonane w tym celu obliczenia.

5. Badanie empiryczne

5.1. Pierwsze podejście do budowy lasu

Zbiór danych, opisujący rynek nieruchomości w powiatach Polski, scharakteryzowano w podrozdziale 2.2. Zmienna objaśniana, czyli średnia cena 1m^2 lokali mieszkalnych, wyrażona została ilościowo (w polskich złotych), co nakierowuje badanie w stronę zagadnień regresyjnych. Poniżej zwizualizowano zróżnicowanie przestrzenne cen mieszkań w 2018 roku w poszczególnych województwach. Mapa ta będzie przydatnym punktem odniesienia w trakcie interpretacji wyników.

Rysunek 9: Ceny mieszkań według województw Polski w roku 2018

Źródło: Opracowanie własne na podstawie danych: https://bdl.stat.gov.pl/BDL/dane/podgrup/tablica

Pierwszym krokiem, podjętym w celu implementacji algorytmu lasów losowych było podzielenie próby badawczej zgodnie z ideą trenowania modelu. Polega ona na wyuczeniu modelu na **części uczącej** oraz testowaniu go na **części testowej**, która nie była wykorzystywana w procesie nauki. Wektory wartości testowych zawierają informacje o realnym wyniku, natomiast nauczony model dostarcza przewidywania[17]. Taka para umożliwia wyznaczenie ocen, takich jak średni bezwzględny błąd procentowy ($MAPE$) czy pierwiastek błędu średniokwadratowego ($RMSE$) dla prognozy. Dla powiatów Polski podział na zbiór treningowy i testowy zrealizowano wstępnie w stosunku 70% do 30% (obserwacje przydzielane do prób w sposób losowy). Pierwszy utworzony las losowy składał się z 1000 drzew. Miał on całkowitą swobodę co do wyboru zmiennych objaśniających do każdego kolejnego drzewa, co wynika z faktu, iż nie narzucono ich maksymalnej liczby, możliwej przy dokonywaniu podziału. Aby mieć świadomość co do stopnia dopasowania modelu do danych empirycznych należy kontrolować wartość **współczynnika determinacji** $\mathbf{R^2}$. Im jest ona wyższa, tym model w większym stopniu objaśnia zmienność średniej ceny mieszkań w powiatach. Miara ta przyjmuje wartości z przedziału $< 0{,}1 >$. Częściej jednak stopień dopasowania wyraża się procentowo, według wzoru[18]:

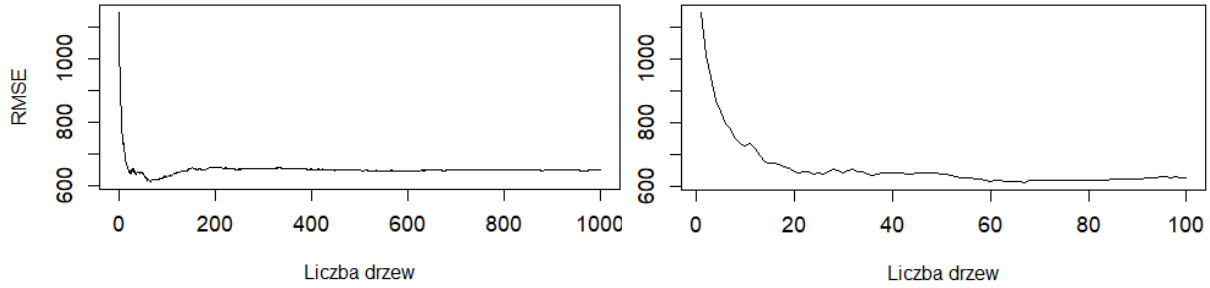
$$\mathrm{R}^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\mathrm{y}}-\bar{\mathrm{y}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}-\bar{\mathrm{y}})^2} \cdot 100\% \,, \tag{6}$$

gdzie y jest rzeczywistą wartością zmiennej objaśnianej, $\hat{y}$ wartością teoretyczną (na podstawie modelu), a $\bar{y}$ to średnia arytmetyczna rzeczywistych wartości zmiennej objaśnianej. Dla utworzonego lasu losowego R^2 wyniósł **70,1%**. Na tej podstawie można stwierdzić, że model zadowalająco wyjaśnia mierzone zjawisko, a jego poziom dopasowania jest wysoki. W następnej kolejności warto przeanalizować zależność błędu prognozy od liczby drzew w lesie losowym (Rysunek 11).

[17] Blachnik Marcin, Testowanie modeli predykcyjnych, http://mblachnik.pl/lib/exe/fetch.php/dydaktyka/zajecia/ai/lab/rm/algorytmy_testowania_rm.pdf, [19.03.2020].

[18] Kufel Tadeusz, *Ekonometria – rozwiązywanie problemów z wykorzystaniem programu GRETL*, Wydawnictwo Naukowe PWN, Warszawa 2013, s. 58.

Rysunek 10: Liczba drzew w lesie losowym a jakość prognoz.



Źródło: Opracowanie własne w środowisku RStudio

Leo Breiman udowodnił, że algorytm lasów losowych nie jest nadmiernie narażony na problem przetrenowania, czyli nadmiernego dopasowania. Mimo tego, zwracał uwagę na to, iż zbyt duża liczba drzew jest powodem obciążenia obliczeniowego, związanego z budową i szkoleniem modelu[19]. Łatwo zauważyć, że po skonstruowaniu około 200 pojedynczych modeli, dodawanie każdego kolejnego nie ma znaczącego wpływu na zmniejszenie wielkości błędu. Co więcej – minimalną wartość $RMSE$ równą **609,9 zł** las osiąga, gdy składa się już z zaledwie **67 drzew**. Ponadto, słabiej zagęszczony model ma wyższy współczynnik R^2 (74%).

Na tej podstawie **obalono Hipotezę 3**, która stanowiła, iż „im więcej drzew w lesie losowym, tym jakość modelu jest wyższa (niższy błąd prognozy)". Można przypuszczać, że błąd ten zbiega do swojej granicy i dla liczby drzew równej około 200 jest jej bardzo bliski. Skoro to nie ten parametr ma decydujący wpływ na optymalizację wyników to gdzie szukać takiej możliwości? Jakich modyfikacji dokonać, aby zmaksymalizować precyzję lasu losowego? O wpływie innych czynników na badaną technikę przeczytać można w kolejnych podrozdziałach.

5.2. Poszukiwanie optymalnych parametrów

Wiele modeli ekonometrycznych posiada ważne parametry, których nie można bezpośrednio ustalić na podstawie danych. Ze względu na to, że nie ma analitycznych wzorów, pozwalających obliczyć ich wartości, parametry te nazywane są dostrajającymi (ang. tuning parameter). Atrybuty te kontrolują złożoność modelu, dlatego słabo dobrane mogą

[19] Kuhn Max, Johnson Kjell, *Applied Predictive Modeling,* Springer 2013, s. 200.

powodować jego nadmierne dopasowanie. Istnieją różne podejścia do poszukiwania optymalnych wielkości. Jedno z nich, które można zastosować do niemalże wszystkich zagadnień, polega na zdefiniowaniu zestawu kandydatów, wygenerowaniu dla nich oszacowań skuteczności modelu, a następnie wybraniu najlepszego ustawienia[20].

Wiadomo już, że sposób poprawy jakości modelu w metodzie lasów losowych nie jest oczywisty. Niekoniecznie pomoże to, co jako pierwsze przychodzi na myśl, czyli rozbudowa lasu. Jakie inne atrybuty algorytmu dostrajać, aby zmierzać ku najlepszemu rozwiązaniu? W poszukiwaniu odpowiedzi na to pytanie skonstruowano ramkę danych, zawierającą **5670 kombinacji**, określających sposób utworzenia lasu losowego. Każdą z możliwości określa wektor wartości czterech parametrów. Są to:

1. **Liczba zmiennych objaśniających**, będących kandydatami do podziału każdego węzła w drzewach decyzyjnych. Uwzględnione wartości to liczby całkowite z przedziału $< 1{,}21 >$ (górne ograniczenie oznacza próbkowanie do podziału wszystkich predyktorów ze zbioru danych). W swoich sugestiach, Leo Breiman wskazuje, że liczba ta powinna wynosić około $\frac{1}{3}$ wszystkich dostępnych zmiennych niezależnych, czyli w tym wypadku 7.
2. **Minimalna liczność obserwacji w węzłach końcowych**. Od tego parametru zależy głębokość drzew. Im więcej obserwacji ma znaleźć się w liściach, tym struktura drzewa jest bardziej płaska (mniejsza głębokość). Do tworzenia kombinacji wprowadzono sekwencję sześciu wartości od 1 do 51, zwiększanych co 10.
3. **Proporcja podziału próby.** Wielkość ta określa, w jakim stosunku podzielić obserwacje na uczące i testowe, aby możliwe było utworzenie optymalnego modelu. Rozważono trzy rozmiary próby uczącej, takie jak: 60%, 70% i 80% całkowitej próby.
4. **Liczba tworzonych drzew.** Mimo, iż modyfikowanie wyłącznie tego parametru okazało się mało skuteczne, należy sprawdzić łączny wpływ wszystkich czynników, uwzględniając także stopień zagęszczenia lasu. Na podstawie wcześniejszych doświadczeń uznano, że sensowne będzie przeanalizowanie jakości modelu dla wektora piętnastu wartości liczby drzew (60,70,80,…,200).

[20] Kuhn Max, Johnson Kjell, *Applied Predictive Modeling,* Springer 2013, s. 64.

Dla każdej z blisko sześciu tysięcy kombinacji tworzony był regresyjny las losowy. Wszystkim modelom obliczono wartość błędu prognozy. Zabezpieczenie przed losowością algorytmu stanowiło ustawienie identycznego ziarna generatora dla każdego z lasów. Zabieg ten zapewnia porównywalność wyników, a co za tym idzie – można bez obaw wnioskować co do wpływu wartości poszczególnych parametrów na efektywność narzędzia. Tabela nr 4 przedstawia ranking 10 najlepszych rozwiązań. Lasy losowe uporządkowano rosnąco według wartości pierwiastka błędu średniokwadratowego (ostatnia kolumna).

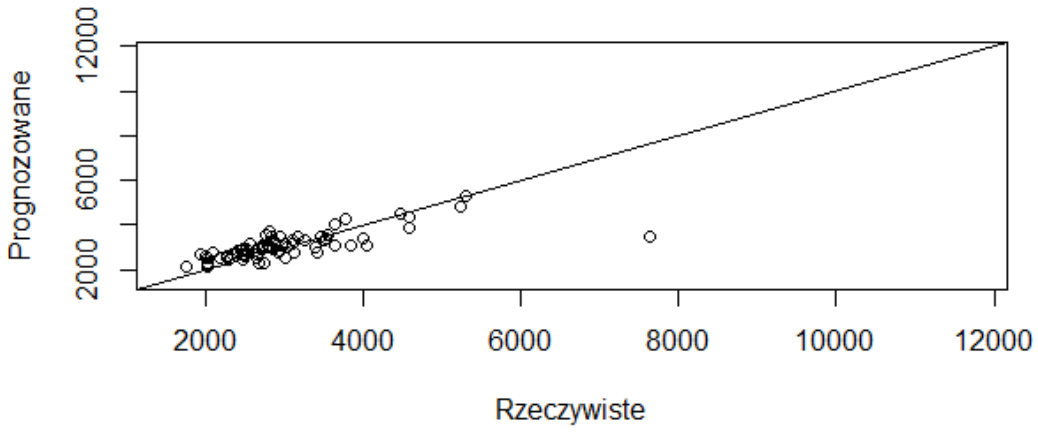
Tabela 4: Najlepsze modele lasów losowych według wartości 4 parametrów

Lp.	Zmienne objaśniające	Minimum w liściu	Udział zbioru uczącego	Liczba drzew	$RMSE$
1.	8	1	0,8	100	**602,9**
2.	8	1	0,8	110	**603,8**
3.	8	1	0,8	80	**606,2**
4.	8	1	0,8	90	**607,2**
5.	8	1	0,8	120	**609,3**
6.	8	1	0,8	70	**609,9**
7.	11	1	0,6	120	**610,6**
8.	11	1	0,8	70	**611,2**
9.	11	1	0,8	120	**611,8**
10.	12	1	0,6	100	**612,1**

Źródło: Opracowanie własne

Najniższy uzyskany błąd prognozy wyniósł 602,9 zł, natomiast najwyższy 781,5 zł. Rozstęp jest ogromny. Oznacza to, że dobór parametrów ma **decydujący wpływ** na wyniki pracy z lasami losowymi. Trafność przewidywania lasu z pozycji nr 1 ilustruje poniższy wykres:

Rysunek 11: Trafność prognostyczna najlepszego modelu lasu losowego.

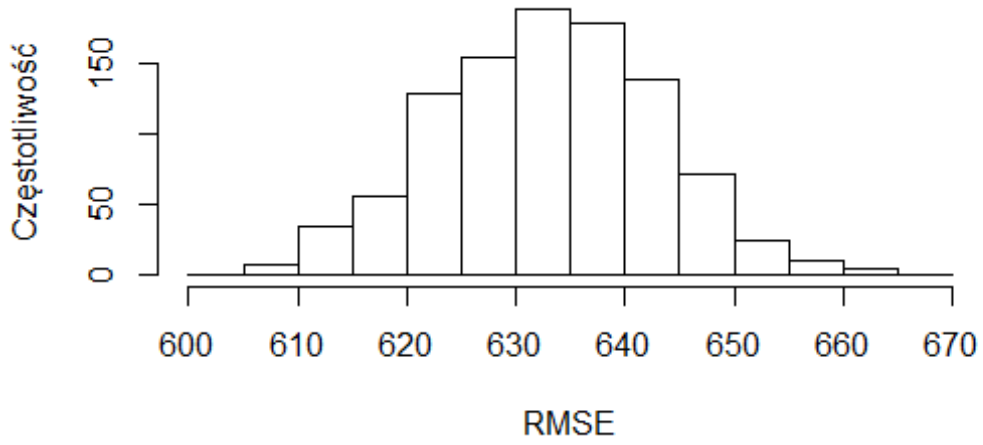


Źródło: Opracowanie własne w środowisku RStudio

Patrząc na sześć najdokładniejszych modeli lasów losowych widać, że liczba zmiennych objaśniających, według których najpewniej powinien być dzielony zbiór wynosi 8. Stanowi to 38% dostępnych dla algorytmu predyktorów. Wskazówki Breimana, który uważa, że parametr ten powinien przyjmować wartość bliską 33,(3)% całego zbioru zmiennych objaśniających faktycznie są zasadne. Wyniki są najlepsze, gdy kolejne drzewa decyzyjne powstają wskutek podziału względem najbardziej istotnych zmiennych. Kolejny wniosek – **im głębsze drzewa, tym las obarczony jest niższym błędem**. Wynika to z faktu, iż aż 110 najskuteczniejszych lasów tworzonych było z minimalną licznością obserwacji w węźle końcowym równą 1. Być może niektóre powiaty „odstawały" od pozostałych i przydzielenie ich do odrębnej klasy w znacznym stopniu obniżyło wartość błędu. Na podstawie powyższej tabeli wolno także twierdzić, że zbiór danych najkorzystniej będzie podzielić na uczący i testowy w stosunku 80% do 20%. Najwidoczniej las losowy potrzebuje **dużego zbioru** w procesie uczenia i to właśnie wtedy wykazuje się najwyższą precyzją. Żeby wysnuć najważniejszy wniosek należy raz jeszcze przyjrzeć się sześciu „topowym" lasom losowym. Okazuje się, że różnią się one wartością tylko jednego atrybutu, a mianowicie liczbą drzew. Informuje to o tym, że dostrojenie tego parametru (w połączeniu z wpływem pozostałych) w istotny sposób stymuluje wyniki algorytmu. Spostrzeżenia te świadczą o wielowymiarowości zagadnienia. Aby w poprawny sposób eksplorować dane przy użyciu techniki lasów należy zdawać sobie sprawę z konieczności skupienia uwagi na wielu elementach, które wymagają dopasowania do danych empirycznych.

Kolejną wątpliwością, która się nasuwa, jest losowość analizowanej metody. Czy skonstruowanie lasu losowego z takimi samymi parametrami, ale na nieco innym zbiorze treningowym zapewni stabilność wyników? Aby to zweryfikować utworzono 1000-krotnie najlepszy model lasu (wykorzystując optymalny wektor parametrów z pierwszego wiersza Tabeli nr 4). Ziarno generatora w tym wypadku było całkowicie losowe, co oznacza, że każdy kolejny las przydzielał do próby uczącej inne powiaty. **Histogram**, znajdujący się poniżej, przedstawia rozkład empiryczny błędu $RMSE$ dla przeprowadzonej symulacji:

Rysunek 12: Histogram błędu prognozy dla 1000 lasów losowych

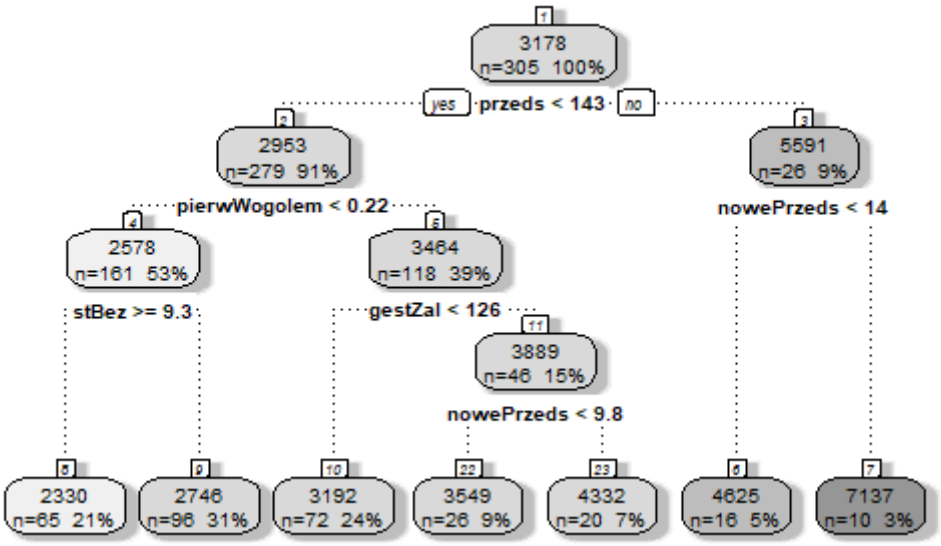


Źródło: Opracowanie własne w środowisku RStudio

Tym razem najniższy uzyskany błąd prognozy wyniósł 604,5 zł. Ważnym aspektem jest przetestowanie w tym momencie normalności rozkładu obliczanej cechy. Jedno z narzędzi, które do tego służą, to test Shapiro-Wilka. Hipoteza zerowa stanowi w nim, że próba badawcza pochodzi z populacji o rozkładzie zbliżonym do krzywej Gaussa. Uzyskana dla rozważanego przykładu wartość p-value wyniosła 44%, a ponieważ jest ona wyższa niż przyjęty poziom istotności 5% można przyjąć, że rozkład błędu prognozy jest **rozkładem normalnym**. Na tej podstawie klaruje się stwierdzenie, że lasy losowe dają akceptowalnie stabilne wyniki. Lepiej oczywiście – jeśli tylko się da – zabezpieczyć się przed niechcianą losowością. Pokazanie dwóch podejść (stochastycznego i deterministycznego) miało na celu uświadomienie różnic, wynikających z wyboru ścieżki wielokrotnego powtarzania algorytmu.

Niech dopełnieniem tego podrozdziału będzie zaprezentowanie jednego z regresyjnych drzew decyzyjnych, „rosnących” w najlepszym lesie losowym. Zbiór uczący ulega podziałowi w następujący sposób:

Rysunek 13: Losowe drzewo decyzyjne z lasu o optymalnych parametrach



Źródło: Opracowanie własne w środowisku RStudio

Jest to najwyższy czas, aby zastanowić się nad sposobem wyznaczania wag poszczególnych zmiennych objaśniających. Na jakiej podstawie algorytm uznaje liczbę przedsiębiorstw za najistotniejszy czynnik, determinujący postać kolejnych gałęzi drzewa?

5.3. Badanie istotności zmiennych objaśniających

Dwie pierwsze, sformułowane w rozdziale 4 hipotezy badawcze odnoszą się do siły oddziaływania pewnych determinant na ceny nieruchomości w powiatach Polski. Złożony charakter badanego algorytmu, przypominający „czarną skrzynkę", utrudnia zrozumienie związku między predyktorami a odpowiedzią. Leo Breiman zaproponował jednak dwie miary, dzięki którym możliwe jest określenie istotności poszczególnych zmiennych objaśniających w modelu lasu losowego.

Najbardziej miarodajnym środkiem pomiaru wag jest **średni spadek dokładności** modelu, gdy wartości danej zmiennej ulegną losowej permutacji (oznaczany jako $\%IncMSE$). Wyraża się go jako procentową zmianę błędu średniokwadratowego na skutek „tasowania" wartości i-tej kolumny na wszystkie możliwe sposoby. Dla każdego drzewa obliczany jest bazowy błąd prognozy na zbiorze testowym (MSE_0). To samo dzieje się po permutacji i-tej zmiennej predykcyjnej (MSE_i) w każdym drzewie. Różnica między tymi wartościami jest

sumowana w całym lesie, a następnie uśredniana dla wszystkich drzew i normalizowana przez odchylenie standardowe różnic. Średni spadek dokładności wyraża się wzorem:

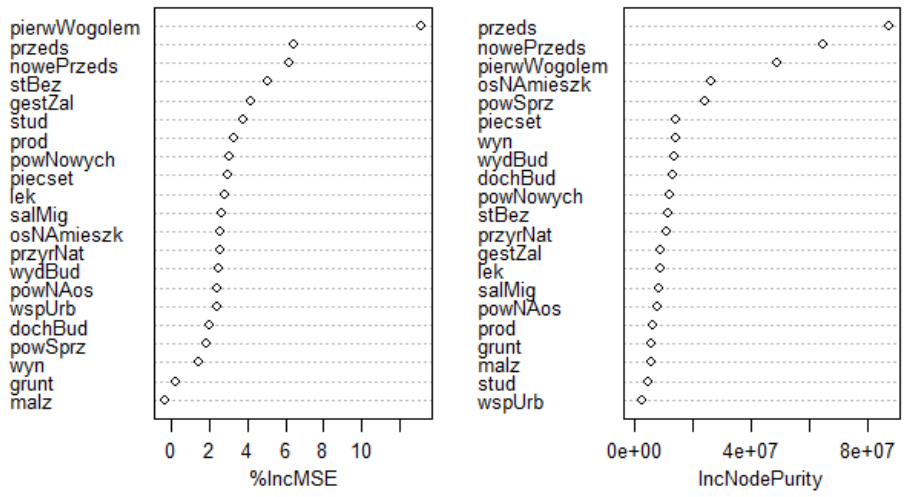
$$\%\mathrm{IncMSE_i} = \frac{(\mathrm{MSE_i} - \mathrm{MSE_0})}{\mathrm{MSE_0}} \cdot 100\% \qquad (7)$$

Im wyższa wartość, tym dana zmienna lasu losowego jest ważniejsza.

Innym podejściem jest pomiar **poprawy jednorodności węzłów** ($IncNodePurity$), która zachodzi na skutek wystąpienia zmiennej objaśniającej w lesie losowym. Celem algorytmu jest znalezienie takiego podziału, żeby wariancja między węzłami była wysoka, a wewnątrz nich - możliwie najniższa. Dla każdego predyktora wartość $IncNodePurity$ obliczana jest (tam, gdzie jest on brany do podziału) jako całkowity wzrost jednorodności węzłów, uśredniony dla wszystkich drzew (resztkowa suma kwadratów)[21]. Analogicznie jak poprzednio – im wyższa wartość $IncNodePurity_i$, tym istotniejszy jest i-ty predyktor. Breiman potwierdził, że miernik ten jest niestety dość obciążony i powinien być używany tylko wtedy, gdy czasochłonność obliczenia $\%IncMSE$ jest nieakceptowalna.

Mając już świadomość co do opisanych możliwości, można przejść do sprawdzenia siły oddziaływania poszczególnych czynników na ceny mieszkań w Polsce. Poniższy wykres przedstawia wyniki obliczeń, wykonanych dla 21 determinant:

Rysunek 14: Istotność zmiennych objaśniających mierzona jako średni spadek dokładności ($\%IncMSE$) oraz poprawa jednorodności węzłów ($IncNodePurity$).



Źródło: Opracowanie własne w środowisku RStudio.

[21] Kuhn Max, Johnson Kjell, *Applied Predictive Modeling,* Springer 2013, s. 202.

Ranking 8 najbardziej wpływowych zmiennych znajduje się w Tabeli nr 5. Kolejność według obu mierników nie jest taka sama, więc uporządkowanie występuje w dwóch wariantach.

Tabela 5: Ranking 8 najistotniejszych zmiennych objaśniających w lesie losowym

Lp.	Nazwa	$\%IncMSE$	Lp.	Nazwa	$IncNodePurity$
1.	Udział rynku pierwotnego	13,1	1.	Liczba przedsiębiorstw	87497141
2.	Liczba przedsiębiorstw	6,4	2.	Liczba nowych przedsiębiorstw	64710455
3.	Liczba nowych przedsiębiorstw	6,2	3.	Udział rynku pierwotnego	48924565
4.	Stopa bezrobocia	5,1	4.	Liczba osób na mieszkanie	26452384
5.	Gęstość zaludnienia	4,2	5.	Powierzchnia mieszkań sprzedanych	12486103
6.	Liczba studentów	3,7	6.	Dzieci objęte programem 500+	14522274
7.	Ludność w wieku produkcyjnym	3,3	7.	Wynagrodzenie brutto	14176174
8.	Powierzchnia nowych mieszkań	3,1	8.	Wydatki budżetów	13628646
…	…	…	…	…	…
21.	Zawarte małżeństwa	-0,3	21.	Współczynnik urbanizacji	2996068

Nietrudno zauważyć, że inne czynniki wpływają na dokładność modelu, a inne na jednorodność węzłów. Pierwsza miara jest uznawana za mającą wyższą wartość informacyjną, dlatego to właśnie ją należy traktować nadrzędnie. Uzyskane uszeregowanie jest niewątpliwie zgodne z mechanizmami, które zachodzą na rynku nieruchomości. Na podium w przypadku obu miar znalazły się **te same determinanty, lecz w innej kolejności**. Jest to bardzo cenna informacja z punktu widzenia prowadzonej analizy. Losowa permutacja zmiennej „udział rynku pierwotnego" powoduje wzrost błędu prognozy MSE o 13,1%. Oznacza to, że wartości tego czynnika mają znaczący wpływ na precyzję wyników lasu losowego. Rzeczywiście średnia cena lokali jest tym wyższa, im więcej nowych mieszkań jest przedmiotem transakcji rynkowych. Są one z reguły droższe niż te oferowane przez właścicieli w formie odsprzedaży. Oba sposoby pomiaru wskazały także na silny wpływ liczby przedsiębiorstw (ogółu i nowych) na badaną zmienną objaśnianą. Ma to ścisły związek z dostępnością miejsc pracy oraz poziomem rozwoju gospodarczego danego powiatu. Co zaskakujące, na 6 pozycji w rankingu według miernika jednorodności węzłów ($IncNodePurity$) i na 9 według $\%IncMSE$ znalazł się odsetek dzieci, którym w danym powiecie przyznano świadczenie „500+". Jest to bardzo ciekawa obserwacja, która pokazuje jak sprzedawcy dostosowują ceny mieszkań do wzrostu dobrobytu rodzin. Z pewnością wpływowe są także: stopa bezrobocia (lepsza kondycja rynku

pracy daje podstawy do podwyżki cen), gęstość zaludnienia (w słabiej zaludnionych regionach powstaje mniej luksusowych inwestycji deweloperskich), liczba studentów (powodująca wzrost popytu głównie na nieduże stancje), odsetek ludności w wieku produkcyjnym (będącej głównym nabywcą) oraz powierzchnia nowych mieszkań (oddanych do użytkowania, prawdopodobnie w stanie determinującym wysoką cenę za 1m^2). Uwagę przyciąga jeszcze jedyna ujemna wartość $\%IncMSE$ dla liczby zawartych małżeństw. Jest to równoznaczne z faktem, że zmienna ta jest na tyle zbędna w modelu, iż nawet jej losowe wartości wpływają lepiej na jakość prognoz. Gdy ulega ona losowej permutacji wartość błędu MSE maleje o 0,3%. Należy zwrócić także uwagę na to, że pod względem poprawy jednorodności węzłów współczynnik urbanizacji jest na ostatnim miejscu. Jest tak, ponieważ we wszystkich miastach na prawach powiatu odsetek ludności miejskiej wynosi 100%. Można by się na tym etapie „zbuntować" i nie zgodzić z tym, że z perspektywy całego kraju czynnik ten nie gra roli w różnicowaniu rynku nieruchomości. Co można zrobić, aby zwiększyć istotność tej zmiennej? Nasuwa się pewien pomysł. A gdyby spróbować powtórzyć wcześniejsze kroki dla dwóch rozłącznych zbiorów danych? Pierwszy z nich to tylko i wyłącznie miasta, a drugi – pozostałe powiaty. Rezultaty tego działania opisane zostały w następnym podrozdziale.

5.4. Wyniki algorytmu a podział zbioru danych

W celu zweryfikowania prawidłowości, zachodzących w różnych segmentach polskiego rynku nieruchomości zdecydowano zmodyfikować pierwotną bazę danych. Tym razem powstały **dwa regresyjne lasy losowe**. Pierwszy z nich tworzony jest dla zbioru 66 miast Polski. Drugi zaś dzieli na odpowiednie klasy co do wysokości cen mieszkań 314 pozostałych powiatów. Przyjęte oznaczenia znajdują się w pierwszej kolumnie Tabeli nr 6.

Tabela 6: Opis badanych lasów losowych

Symbol lasu losowego	Uwzględnione obserwacje	Liczba obserwacji w zbiorze danych	Przykładowe powiaty	Średnia cena 1m^2 mieszkań [zł]	Współczynnik zmienności średniej ceny [%]
Las W	Wszystkie powiaty	380	bieszczadzki, miasto Rybnik	3143,5	35,2%
Las M	Miasta na prawach powiatu	66	miasto Rybnik, miasto Gdańsk	3942,2	41,1%
Las P	Pozostałe powiaty (bez miast)	314	bieszczadzki, stalowowolski	2975,6	29,4%

Oto najbardziej nurtujące pytania, na które poszukiwano odpowiedzi w tej części analizy:

1. Czy dokładność prognostyczna każdego z dwóch nowych lasów losowych jest wyższa niż dla modelu, który powstał dla wszystkich obserwacji?
2. Czy lepiej badać rynek nieruchomości osobno w miastach na prawach powiatu i w pozostałych jednostkach samorządu terytorialnego w Polsce?
3. Jak zmieniła się istotność zmiennych objaśniających w lasach losowych dla dwóch zbiorów danych?
4. Czy inne czynniki determinują średnią cenę mieszkań w miastach liczących ponad 100 tysięcy mieszkańców, a inne poza nimi?

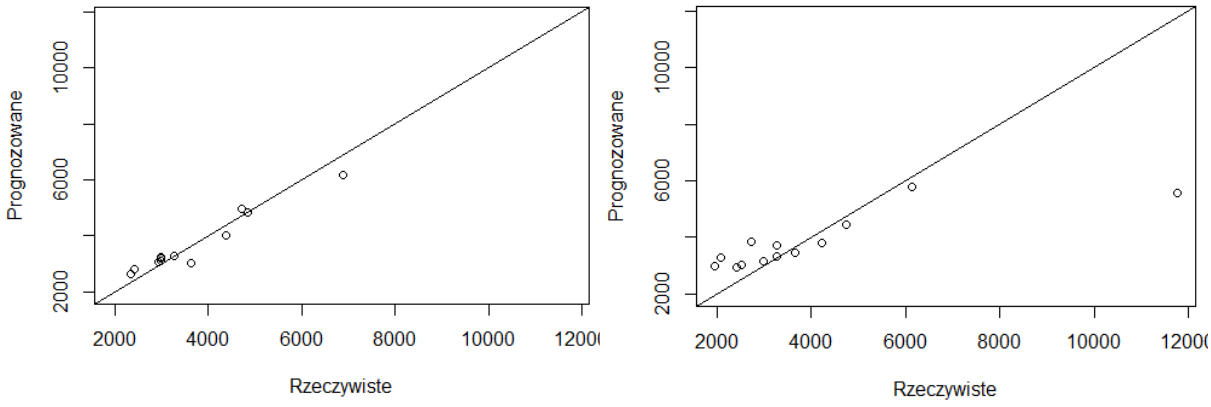
Regresyjne lasy losowe M i P, tak jak las W we wcześniejszej części pracy, przeszły etap **dostrajania parametrów**. Dla każdego z nich analizowano w tym celu 5670 różnych kombinacji. Dzięki temu możliwe jest porównanie oceny jakości prognoz, jaką można uzyskać dla lasów losowych bazujących na trzech różnych próbach. Niżej zaprezentowano najniższe wartości średniego błędu procentowego. $MAPE$ pozwala na porównanie dokładności przewidywań różnych modeli. Tabela przedstawia tym samym skuteczność pięciu najlepszych lasów regresyjnych (spośród blisko sześciu tysięcy), skonstruowanych odpowiednio dla 380, 66 i 314 obiektów.

Tabela 7: Porównanie jakości prognoz trzech regresyjnych lasów losowych

$MAPE$ [%]		
Las W	**Las M**	**Las P**
11,86	7,82	11,11
11,86	7,91	11,11
11,87	7,96	11,11
11,87	8,04	11,13
11,89	8,11	11,13

Jak się okazało, najwyższy poziom dokładności prognoz wykazuje model lasu losowego M. Wartości przewidywane odbiegają w najlepszym wypadku średnio o 7,8% od rzeczywistych. Inna miara oceny jakości prognoz – $RMSE$ wynosi tutaj minimalnie 344,4 zł, czyli aż o 258,5 zł mniej niż w najlepszym lesie losowym W. Mimo wysokiej zmienności cen mieszkań (o współczynniku, który wynosi ponad 41%) dokładność prognostyczna, możliwa do uzyskania na takim zbiorze danych jest najwyższa. Liczebność zbioru miast Polski najwidoczniej nie była na tyle mała, aby uniemożliwić algorytmowi poprawę jakości rozwiązań. Wadą takiego podejścia jest natomiast niska stabilność wyników. Ze względu na to, że wzrasta ona wraz z rozmiarem próby badawczej, jej rozmiar w tym przypadku jest dla algorytmu dużym ograniczeniem. Trafność prognostyczną najlepszego oraz najgorszego ($MAPE$ = 22,4%) modelu regresyjnego dla miast na prawach powiatu ilustrują poniższe wykresy:

Rysunek 15: Trafność prognostyczna najlepszego i najgorszego modelu lasu losowego dla miast na prawach powiatu



Źródło: Opracowanie własne w środowisku RStudio

Wspomniano już, że ceny mieszkań w miastach wykazują się wysokim stopniem rozproszenia, czyli są niejednorodne. Można się więc spodziewać, że wśród obserwacji,

będących częścią lasu M, znajdują się takie, których średnia cena lokali jest bardzo oddalona od pozostałych. Oznacza to prawdopodobne występowanie **obserwacji odstających** w tym modelu regresji. Po przeprowadzeniu testu Grubbsa okazało się, że dwa obiekty są tzw. outlierami pod względem średniej ceny za $1m^2$ lokali mieszkalnych. Są to wyjątkowo drogie miasta, a mianowicie Sopot (11763 zł) oraz Warszawa (8769 zł). Czy algorytm lasów losowych jest odporny na występowanie takiej wady w zestawie danych? Sprawdzono czy dokładność przewidywań wzrośnie po usunięciu obu obserwacji odstających. Co zaskakujące – nie udało się po tym zabiegu uzyskać lepszego niż poprzednio modelu. Potwierdzona zostaje tym samym prawdziwość spotykanego w literaturze stwierdzenia o odporności lasów losowych na problem oddalonych obiektów. Można to wyjaśnić w prosty sposób. Jeżeli model nie jest ograniczony co do minimalnej liczby obserwacji w węźle końcowym, to może przydzielić oddalony obiekt do osobnego liścia w każdym drzewie. Dzięki temu precyzja modelu nie zostaje zaburzona. Tym akcentem można zakończyć interpretację lasu losowego M.

Jak natomiast zadziałał algorytm, bazując na zbiorze powiatów z wyłączeniem miast? Okazuje się, że także i w tym podejściu możliwa do uzyskania jakość prognoz była wyższa niż dla pierwotnego modelu. Zarówno wartość średniego bezwzględnego błędu procentowego (11,1%), jak i pierwiastka błędu średniokwadratowego (384,1 zł) były dla najlepszego z lasów P niższe niż w przypadku bazowania na wszystkich polskich powiatach. Zwiększenie jednorodności średniej ceny mieszkań w zbiorze danych po wyeliminowaniu miast poskutkowała poprawą stabilności wyników. Poniżej zaprezentowano zestawienie pięciu przykładowych wartości, przewidywanych przez model lasu P (o optymalnych parametrach), z odpowiadającymi im wartościami rzeczywistymi.

Tabela 8: Wartości rzeczywiste i prognozowane przez optymalny model lasu losowego dla powiatów Polski bez miast

Nazwa powiatu	Wartość rzeczywista [zł]	Wartość prognozowana [zł]	$MAPE$ [%]
bieszczadzki	2130,0	2252,4	5,8
cieszyński	3572,0	2989,3	16,3
krakowski	4579,0	3922,3	14,3
ostrowski	3206,0	3425,5	6,9
przeworski	2485,0	2612,2	5,1

Ostatnim, lecz równie ważnym etapem prowadzonej eksploracji danych, jest analiza istotności zmiennych objaśniających w lasach W, M oraz P. Z punktu widzenia badania rynku nieruchomości w Polsce da to pogląd na zróżnicowanie wpływowych czynników w różnych grupach jednostek administracyjnych. W podrozdziale 5.3 pisano już o tym, że najsłuszniejszym sposobem określenia wag zmiennych jest pomiar średniego spadku dokładności ($\%IncMSE$). W Tabeli nr 9 zestawiono najważniejsze determinanty kondycji rynku mieszkaniowego w trzech różnych modelach regresyjnych. Są to wyniki uzyskane dla lasów losowych po dostrojeniu i optymalizacji parametrów, takich jak liczba drzew, liczba zmiennych objaśniających (będących kandydatami w każdym podziale), minimum obserwacji w węzłach końcowych oraz proporcja podziału próby na zbiór uczący i testowy.

Tabela 9: Istotność zmiennych objaśniających w trzech lasach losowych: dla całego zbioru danych, dla miast na prawach powiatu oraz dla powiatów z wyłączeniem miast na prawach powiatu

	Las W		**Las M**		**Las P**	
Lp.	**Nazwa**	$\%IncMSE$	**Nazwa**	$\%IncMSE$	**Nazwa**	$\%IncMSE$
1.	Udział rynku pierwotnego	13,1	Liczba studentów	5,4	Udział rynku pierwotnego	9,6
2.	Liczba przedsiębiorstw	6,4	Liczba nowych przedsiębiorstw	4,9	Liczba nowych przedsiębiorstw	4,5
3.	Liczba nowych przedsiębiorstw	6,2	Udział rynku pierwotnego	4,5	Liczba przedsiębiorstw	4,3
4.	Stopa bezrobocia	5,1	Powierzchnia mieszkań sprzedanych	4,3	Stopa bezrobocia	3,9
5.	Gęstość zaludnienia	4,2	Wynagrodzenia brutto	2,3	Saldo migracji	3,4
6.	Liczba studentów	3,7	Liczba osób na mieszkanie	2,2	Gęstość zaludnienia	3,4
7.	Ludność w wieku produkcyjnym	3,3	Dochody budżetów	1,5	Powierzchnia mieszkań	2,8
8.	Powierzchnia nowych mieszkań	3,1	Liczba lekarzy	1,3	Przyrost naturalny	2,7
…	…	…			…	…
21.	Zawarte małżeństwa	-0,3	Dzieci objęte programem 500+	-1,3	Zawarte małżeństwa	0,2

Wyniki, zamieszczone w powyższej tabeli, rzucają nowe światło na polski rynek mieszkaniowy. Lokalne rynki nieruchomości reagują na wiele mikroekonomicznych

i makroekonomicznych impulsów. Przełomowym wnioskiem, nasuwającym się na tym etapie pracy jest fakt, iż **inne czynniki kształtują ceny mieszkań w dużych aglomeracjach, a inne na pozostałym obszarze kraju**. Co ciekawe – największy wpływ na kondycję rynku nieruchomości w miastach na prawach powiatu ma liczba studentów na 1000 mieszkańców. Losowa permutacja tej zmiennej w lesie M powoduje wzrost błędu średniokwadratowego (MSE) o ponad 5,4%. Wiadomo, że wrzesień i październik jest okresem „żniw" dla najemców i sprzedawców mieszkań w miastach akademickich. Pomimo remontów i stopniowej poprawy warunków uczelniane akademiki wciąż cechują się niższym standardem niż lokale prywatne. Napędza to popyt na mieszkania, a co za tym idzie – rynek nieruchomości w danym mieście staje się rynkiem sprzedawcy, który to może sobie pozwolić na dyktowanie wysokich cen. Kolejny bodziec w miastach to liczba nowych firm, świadcząca o wzroście inwestycji i zatrudnienia. Ważne okazały się także zmienne związane wprost ze strukturą rynku nieruchomości, tj. odsetek sprzedaży nowych mieszkań oraz ich łączna powierzchnia, a także liczba osób, przypadających na jeden lokal. Obserwowany w Polsce od lat 80-tych „głód mieszkań" nadal nie jest zaspokojony, co może być skutkiem nieodpowiedniego podejścia, zakładającego nadanie priorytetu luksusowym apartamentowcom zamiast taniemu budownictwu czynszowemu. Ponadto średnia cena $1m^2$ zależy od wysokości płac w danym mieście. Na podstawie zebranej bazy danych obliczono, że aby kupić lokal o przeciętnym standardzie i powierzchni $50m^2$ w jednej z 66 aglomeracji Polak musi przeznaczyć średnio 53 pensje netto. Patrząc na to bardziej realistycznie – odkładając połowę zarobionej kwoty na zakup będzie go stać dopiero po prawie 9 latach. Jak ma się natomiast liczba lekarzy do ceny mieszkań w miastach? Okazuje się, że – o dziwo – zmienna ta jest w tym wypadku istotna. Można to ściśle powiązać z intensywnością inwestycji na różne cele prorozwojowe aglomeracji, co zwiększa jej atrakcyjność dla przyszłych nabywców. Na uwagę zasługuje jeszcze fakt, iż odsetek dzieci, objętych świadczeniem z tytułu programu „Rodzina 500+" nie ma żadnego znaczenia w kształtowaniu miejskich cen nieruchomości. Ogólnie rzecz ujmując, w kształtowaniu miejskich rynków nieruchomości czynniki społeczno-demograficzne okazały się być zdominowane przez te o charakterze gospodarczym oraz endogenicznym.

Nieco inaczej sytuacja wygląda w pozostałych 314 powiatach kraju. Tutaj liczba studentów nie odgrywa już dominującej roli w kształtowaniu struktury rynku nieruchomości. Największy wpływ na poziom cen ma natomiast odsetek transakcji, których przedmiotem są

mieszkania w stanie deweloperskim. W niektórych częściach Polski, m.in. w powiatach górowskim, grudziądzkim czy łobeskim w 2018 roku sprzedaż odbywała się wyłącznie na rynku wtórnym. Jednocześnie ceny mieszkań były tam bardzo niskie i wynosiły zaledwie około 1800 zł/m^2. Można zauważyć, że poza obszarem największych miast waga czynników demograficznych jest wyższa. Kolejno na miejscu 4, 5 i 6 w rankingu istotności znalazły się: stopa bezrobocia, saldo migracji oraz gęstość zaludnienia. Pierwszy z wymienionych czynników wykazuje ujemną korelację ze stanem koniunktury mieszkaniowej. W latach 2013-2018 spadek bezrobocia o jeden punkt procentowy odpowiadał podwyżce cen mieszkań o około 5,5%[22]. Jeśli chodzi o aspekt migracji, coraz więcej osób w poszukiwaniu spokoju osiedla się w małych miastach. Mający tego świadomość deweloperzy inwestują na tych terenach, ponieważ zainteresowanie nieruchomościami daje im szansę na duże zyski. Powiatem, do którego dwa lata temu migrowało najwięcej osób (aż 28 na 1000 mieszkańców) był powiat wrocławski. Obejmuje on obszar wokół Wrocławia, zaś samo miasto posiada swoje własne prawa. Wysokie ceny mieszkań są tu spowodowane napływem ludności i tworzeniem się tzw. strefy podmiejskiej z rozwiniętą infrastrukturą. Na podstawie lasu losowego P wyraźnie zaznaczona została także współzależność stanu ludności (odzwierciedlonego przez gęstość zaludnienia i przyrost naturalny) i kondycji rynku mieszkaniowego w rodzimych powiatach.

Ostatnią, pozostałą do wyjaśnienia kwestią jest rozstrzygnięcie czy metoda lasów losowych może konkurować pod względem skuteczności i precyzji z klasyczną regresją liniową. Przechodząc do następnego podrozdziału przystąpiono więc do porównania obu technik.

5.5. Lasy losowe a regresja liniowa

Po wykonaniu wszystkich wcześniejszych obliczeń wydaje się, że algorytm regresyjnych lasów losowych jest niezwykle efektywny. Warto jednak odnieść się przy ocenie jego skuteczności do jednej z metod o podobnych zastosowaniach. Najpopularniejszym narzędziem, służącym do prognozowania statystycznego jest **klasyczna metoda**

[22] Hońdo Tomasz, *Spadające bezrobocie a rekordowe ceny mieszkań,* https://www.forbes.pl/gospodarka/stopa-bezrobocia-a-ceny-mieszkan/1cz8btm, [29.04.2020]

najmniejszych kwadratów. Jej idea polega na znalezieniu takich wartości wektora parametrów, dla którego suma kwadratów reszt (czyli różnic pomiędzy wartościami teoretycznymi a rzeczywistymi) jest minimalna. Aby wyestymowany wektor parametrów był nieobciążony, zgodny i najefektywniejszy w klasie liniowych estymatorów model powinien niestety spełniać wiele wstępnych restrykcji. Po pierwsze - pomiędzy zmiennymi objaśniającymi a zmienną objaśnianą musi występować zależność liniowa. Po drugie - rozważane czynniki powinny być nielosowe i słabo skorelowane ze sobą oraz ze składnikiem losowym modelu. Po trzecie - wymagane jest w miarę równomierne rozproszenie punktów dookoła obliczonej linii regresji (jednorodność wariancji). Co więcej, rozkład reszt modelu powinien być normalny. Fundamentalnym problemem, na który model regresji liniowej nie jest odporny, jest ponadto występowanie w badanej próbie obserwacji odstających. Wyniki, uzyskane na zakłóconym zbiorze (z niektórymi resztami istotnie niższymi lub wyższymi od pozostałych) są niewiarygodne i zaburzone[23].

Główną wadą klasycznej regresji liniowej jest więc czasochłonność przygotowywania danych, wymagana do prawidłowego przeprowadzenia analizy. Mający tego świadomość badacze coraz częściej korzystają z metody lasów losowych, która wykazuje się elastycznością w stosunku do próby, a jej wyniki są odporne na braki w danych i występowanie nietypowych obiektów. Czy jednak nowatorski algorytm eksploracji danych może być skuteczniejszy od tradycyjnej techniki z przełomu XVIII i XIX wieku? Aby odpowiedzieć na to pytanie ponownie posłużono się bazą danych, dotyczącą rynku nieruchomości w Polsce z 2018 roku. W Tabeli nr 6 (podrozdział 5.4) zdefiniowano trzy zbiory, na podstawie których tworzone były lasy losowe: W (wszystkie powiaty), M (miasta z własnymi prawami) oraz P (powiaty bez miast). Na podstawie analogicznych prób zweryfikowano jakość prognoz, uzyskaną po zastosowaniu klasycznej metody najmniejszych kwadratów. Liczba zmiennych objaśniających, wchodzących w skład modelu regresji odpowiadała liczbie zmiennych, wskazanych przez optymalny model lasu losowego. Przykładowo, dla zbioru ze wszystkimi obserwacjami wartość tego parametru po dostrojeniu wynosiła 8. Model regresji tworzony był więc z uwzględnieniem na wstępie ośmiu najistotniejszych predyktorów, wskazanych przez najlepszy las losowy. Aby jeszcze bardziej rozszerzyć zakres analizy prognozowanie za pomocą KMNK przeprowadzono zarówno

[23] Mućk Jakub, *Wprowadzenie do modelowania ekonometrycznego*, Katedra Ekonomii Ilościowej Szkoły Głównej Handlowej w Warszawie, http://web.sgh.waw.pl/~jmuck/Ekonometria/EkonometriaPrezentacja1.pdf, [29.04.2020].

dla zbiorów danych bez żadnych korekt jak i po odpowiednim przygotowaniu modelu (tak, aby zawierał on tylko istotne statystycznie zmienne objaśniające, a próba była oczyszczona z obserwacji odstających). Jakość prognoz regresji sprawdzano na tych samych zbiorach testowych jak w przypadku odpowiadających im lasów. Rezultaty, mierzone wartością średniego bezwzględnego błędu procentowego ($MAPE$) zebrano w poniższej tabeli.

Tabela 10: Porównanie błędów prognozy lasów losowych oraz regresji liniowej

Najlepsze lasy losowe			
	Las W	Las M	Las P
$\boldsymbol{MAPE}$ **[%]**	**11,86**	**7,82**	**11,11**
Optymalna liczba zmiennych objaśniających	8	4	9
Klasyczna metoda najmniejszych kwadratów przed oczyszczeniem danych			
	Zbiór W	Zbiór M	Zbiór P
Liczba obserwacji	380	66	314
Liczba obserwacji odstających	9	1	6
Liczba zmiennych objaśniających	8	4	9
Liczba nieistotnych zmiennych	4	3	3
Normalność rozkładu reszt	NIE	NIE	NIE
$\boldsymbol{MAPE}$ **[%]**	**14,45**	**13,13**	**13,85**
Klasyczna metoda najmniejszych kwadratów po oczyszczeniu danych			
	Zbiór W	Zbiór M	Zbiór P
Liczba obserwacji	380	66	314
Liczba obserwacji odstających	0	0	0
Liczba zmiennych objaśniających	5	3	7
Liczba nieistotnych zmiennych	0	0	0
Normalność rozkładu reszt	NIE	NIE	TAK
$\boldsymbol{MAPE}$ **[%]**	**13,33**	**9,99**	**12,83**

Dla wyjaśnienia, przy ustalaniu liczby obserwacji odstających tym razem zastosowano **zasadę trzech sigm**. Zgodnie z nią tzw. outlierem jest obiekt, którego wartość zmiennej objaśnianej jest oddalona od średniej w populacji o więcej niż wartość trzech odchyleń standardowych. Istotność zmiennych w modelu regresji była natomiast sprawdzana za pomocą testu t-Studenta. Przyjęty poziom istotności to 5%. Predyktory nieistotne eliminowano z modelu stosując metodę krokową wsteczną. Proces ten polega na usuwaniu

kolejno najmniej ważnej zmiennej i estymowaniu modelu ponownie aż do momentu, kiedy wszystkie zmienne niezależne są istotne statystycznie. Jak się okazało – w przypadku każdego zbioru danych **dokładność prognoz techniki lasów losowych była wyższa niż w przypadku klasycznej regresji liniowej**. Potwierdza to wysoką skuteczność tej metody oraz usprawiedliwia zachwyt, z jakim spotyka się ona w ostatnich latach. Widać, że wartości błędu $MAPE$ malały po oczyszczeniu zbioru danych. Mimo tego, tylko w jednym przypadku (dla zbioru P) udało się uzyskać reszty o rozkładzie normalnym. Tradycyjna regresja liniowa jest więc wymagająca względem zbioru danych, którego odpowiednie przygotowanie wymaga sporo wiedzy, czasu i zaangażowania.

6. Zakończenie

6.1. Podsumowanie wyników

Dzięki przeprowadzonemu w pracy badaniu odkryto **wiele zalet** algorytmu regresyjnych lasów losowych. Jedną z nich jest wykazany brak konieczności wstępnej modyfikacji i przekształcania próby badawczej. Nie przeszkadza to w osiąganiu satysfakcjonujących rozwiązań o niskim błędzie i umiarkowanej wariancji. Ponadto technika ta, bazując na podzbiorach, doskonale radzi sobie z wielowymiarowymi bazami danych. Leo Breiman często podkreślał, iż praktyka dużych zbiorów to decydujący czynnik, wpływający na efektywność algorytmów z grupy CART. Kolejną mocną stroną lasów losowych jest szybsze działanie niż w przypadku trenowania setek pojedynczych drzew decyzyjnych. Jest to związane z faktem, iż kandydatami do podziału kolejnych węzłów drzew w lasach jest tylko określona część zmiennych objaśniających. Nie można także pominąć cech szczególnych analizowanego narzędzia, dzięki którym dominuje ono nad klasyczną metodą najmniejszych kwadratów. Przewaga ta ujawnia się jako odporność na występowanie obserwacji odstających, a także obojętność wobec zależności nieliniowych. Z drugiej zaś strony, algorytm działa na zasadzie tzw. „czarnej skrzynki", co utrudnia śledzenie jego kolejnych kroków oraz interpretację wyników. Bardzo rozbudowane lasy losowe są niestety złożone obliczeniowo i mogą wymagać wiele czasu i pamięci urządzenia. Bez wątpienia istotną **wadą** jest także konieczność dostrojenia hiperparametrów. W przeprowadzonej analizie pokazano jak bardzo etap ten wpływa na końcowe rezultaty i chroni przed nadmiernym dopasowaniem modelu. Zastosowanie techniki lasów losowych to mimo wszystko okazja do zapoznania się z nowatorskim podejściem do eksploracji danych. Taki eksperyment umożliwia nie tylko dotarcie do ważnych wniosków, ale i odkrycie ciekawych i zaskakujących zależności.

W rozdziale 4 sformułowany był zestaw sześciu przewodnich hipotez badawczych. Po wykonaniu wszystkich niezbędnych obliczeń czas na **zweryfikowanie** ich prawidłowości. Czy początkowe przypuszczenia były słuszne? Rezultat badań wraz z odpowiednimi wyjaśnieniami zaprezentowano w Tabeli nr 11.

Tabela 11: Weryfikacja hipotez badawczych

Lp.	Hipoteza badawcza	Weryfikacja	Uzasadnienie
1.	„Na wysokość średniej ceny 1m^2 lokali mieszkalnych w poszczególnych powiatach statystycznie największy wpływ mają: liczba przedsiębiorstw (zarówno wszystkich jak i nowo powstałych), powierzchnia nowych mieszkań oraz odsetek transakcji na rynku pierwotnym.”	**PRAWDA**	W przypadku badania trzech zbiorów danych (w lasach losowych: W,M oraz P) czynniki te znalazły się w pierwszej dziesiątce rankingu istotności zmiennych objaśniających.
2.	„Niewielką rolę w kształtowaniu struktury rynku mieszkaniowego odgrywają: odsetek dzieci objętych programem „500+”, liczba zawartych małżeństw oraz liczba studentów.”	**FAŁSZ**	Liczba studentów ma decydujący wpływ na wysokość cen mieszkań w dużych miastach Polski.
3.	„Im więcej drzew w lesie losowym, tym jakość modelu jest wyższa (niższy błąd prognozy).”	**FAŁSZ**	Wraz ze wzrostem zagęszczenia lasu błąd prognozy zbiega do granicznej wartości. W pewnym momencie dalszy rozrost powoduje jedynie nadmierne obciążenie obliczeniowe.
4.	„Dzięki możliwościom, jakie daje zastosowanie techniki lasów losowych, możliwe jest uzyskanie skuteczniejszego modelu niż w przypadku regresji liniowej.”	**PRAWDA**	Elastyczność wobec zbioru danych oraz odporność na wiele problemów umożliwiają lasom losowym przewidywanie z wysoką dokładnością.
5.	„Lepszym podejściem do analizy rynku mieszkań jest utworzenie dwóch lasów losowych – osobno dla 66 miast i 314 pozostałych powiatów.”	**PRAWDA**	Utworzenie dwóch lasów losowych na rozłącznych zbiorach miast i pozostałych powiatów umożliwiło uzyskanie modeli wyższej jakości niż model ogólny Polski.
6.	„Lasy losowe są techniką analityczną, charakteryzującą się wysoką odpornością na występowanie w zbiorze danych obserwacji odstających.”	**PRAWDA**	Algorytm skutecznie radzi sobie z problemem outlierów, zasadniczo przydzielając je do osobnych liści drzew. Dzięki temu model nie jest zaburzony.

Wątpliwości co do prawdziwości hipotez zostały rozwiane dzięki budowaniu lasów losowych, korzystających z drzew o wcześniej dobranych parametrach. Drzewa te obejmowały swym zasięgiem wiele obserwacji (powiatów Polski) i czynników z różnych dziedzin życia (zmienne objaśniające) oraz wspólnie głosowały nad wynikiem. Aby uprościć i utrwalić działanie algorytmu postanowiono je zobrazować w następujący sposób. Można zapytać najlepszego eksperta (drzewo decyzyjne) o opinię na dany temat. Można też znaleźć **wielu ekspertów**, którzy mają **inną wiedzę** (podzbiór zmiennych objaśniających) i **inne doświadczenia życiowe** (zbiór uczący) i zwrócić się do nich o pomoc. Następnie na bazie wszystkich tych opinii należy wybrać tę, która najczęściej się powtarza lub uśrednić wskazywane wyniki[24]. Wybór oparty na kompetencjach i doświadczeniu grupy zapytanych będzie z pewnością trafniejszy niż ten dokonany po konsultacji tylko z jedną, przypadkową osobą.

6.2. Wnioski i uwagi końcowe

W ostatnich latach jednym z najważniejszych kierunków współczesnych badań ekonomicznych stało się **zróżnicowanie regionalne**. Procesy globalizacyjne prowadzą do dysproporcji pomiędzy poziomem rozwoju gospodarczego jednostek administracyjnych także w Polsce. Doskonale widać to na przykładzie rynku nieruchomości mieszkaniowych, a stanowi on przecież bardzo ważny fragment krajowej gospodarki. W literaturze ciężko znaleźć jednoznaczne określenie źródeł tak dużych różnic cen mieszkań w poszczególnych częściach Polski. Wielu autorów wskazuje, że nierówności te powiązane są głównie z położeniem geograficznym i lokalizacją. Twórcy artykułów internetowych i prasowych wyliczają natomiast wiele innych powodów drożejących mieszkań – od wzrostu cen gruntów, poprzez spadek bezrobocia, aż po wzrost zarobków Polaków.

Wyniki przeprowadzonej analizy dostarczyły dowodów na to, że zróżnicowanie cen lokali mieszkalnych ma **różne źródła** w dużych miastach oraz na pozostałym obszarze kraju. Udało się także sporządzić ranking, który wyjaśnia jakie czynniki najsilniej wpływają na

[24] Melniczuk Damian, *Jak nie zabłądzić w lesie… losowym?,* https://jakbadacdane.pl/jak-nie-zabladzic-w-lesie-losowym/, [02.05.2020].

kondycję rynku nieruchomości. Pokazało to realne zależności w tym segmencie krajowej gospodarki. Sformułowanie wielu wartościowych wniosków było możliwe dzięki zastosowaniu techniki lasów losowych. Okazało się, że narzędzie to bardzo dobrze radzi sobie z rozwiązywaniem nie tylko problemów klasyfikacyjnych, ale i regresyjnych. Nic więc dziwnego, że jest ono **powszechnie stosowane** w wielu dziedzinach życia, m.in. w diagnostyce medycznej, kryminalistyce, przy ocenie zdolności kredytowej przez banki czy do szacowania wydajności produkcji. W przeprowadzonej analizie wypracowano zadowalające zdolności prognostyczne modeli regresyjnych lasów losowych. Mogą one posłużyć jako narzędzie do przewidywania przyszłych cen nieruchomości przy założonych zmianach wartości pozostałych czynników.

Przygotowane opracowanie umożliwiło poznanie metod z grupy CART, takich jak drzewa decyzyjne, bagging oraz lasy losowe. Dogłębna analiza ostatniej z wymienionych technik pozwoliła na zidentyfikowanie jej zalet oraz wad. Biorąc pod uwagę wszystkie opisane wyżej aspekty - **cel pracy został osiągnięty**. Nie musi to oznaczać jednak wyczerpania możliwości. Zapoczątkowane obserwacje można **kontynuować** w celu zdiagnozowania szerszego zbioru czynników, które silnie wpływają na kondycję lokalnych rynków nieruchomości. Nie ma także przeszkód do analogicznego postępowania w celu analizy sytuacji na przekroju innych jednostek administracyjnych, np. województw czy gmin. W obecnym czasie niepewności i ryzyka, jakie pojawiły się także i w tej sferze gospodarki w związku z panującą pandemią koronawirusa, należałoby nadzorować i analizować zmieniającą się sytuację. Warto zabiegać o świadomość tego, co można zrobić, aby zminimalizować negatywne skutki i zapobiec załamaniu się kondycji tego rynku. Niezwykle ciekawe byłoby także wyjście poza obszar Polski i przeprowadzenie badania, dotyczącego zróżnicowania kondycji sektora nieruchomości na przykład w krajach Unii Europejskiej, w stanach USA czy w najbardziej rozwiniętych państwach świata. Jedynym ograniczeniem jest tak naprawdę dostępność i kompletność danych. Zagłębienie się w tematykę rynku nieruchomości przy użyciu narzędzia lasów losowych było rozwijającą przygodą, która umożliwiła z jednej strony poznanie nowatorskich możliwości analizy danych, a z drugiej – zrozumienie mechanizmów, zachodzących na polskim rynku nieruchomości. Trud, włożony w opracowanie bazy danych oraz przygotowanie obliczeń, zaowocował zbiorem konkretnych i zaskakujących konkluzji.

7. Literatura

1. Bickel Peter, Jordan Michael, Rice John, *In Memoriam – Leo Breiman*, https://senate.universityofcalifornia.edu/_files/inmemoriam/html/leobreiman.htm, [14.03.2020].
2. Blachnik Marcin, *Testowanie modeli predykcyjnych*, http://mblachnik.pl/lib/exe/fetch.php/dydaktyka/zajecia/ai/lab/rm/algorytmy_testowania_rm.pdf, [19.03.2020].
3. Bryx Marek, *Rynek nieruchomości. System i funkcjonowanie*, Poltext, Warszawa 2006, s. 88.
4. Hońdo Tomasz, *Spadające bezrobocie a rekordowe ceny mieszkań*, https://www.forbes.pl/gospodarka/stopa-bezrobocia-a-ceny-mieszkan/1cz8btm, [29.04.2020].
5. Gatnar Eugeniusz, Walesiak Marek, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław 2004, s. 103-135.
6. Grabowska Ewa, *Jak udoskonalić algorytm drzew decyzyjnych?*, https://predictivesolutions.pl/jak-udoskonalic-algorytm-drzew-decyzyjnych, [13.04.2020].
7. Kałkowski Leszek, *Rynek nieruchomości w Polsce*, TWIGGER, Warszawa 2003, s. 20-21.
8. Kufel Tadeusz, *Ekonometria – rozwiązywanie problemów z wykorzystaniem programu GRETL*, Wydawnictwo Naukowe PWN, Warszawa 2013, s. 58.
9. Kuhn Max, Johnson Kjell, *Applied Predictive Modeling,* Springer 2013, s. 173-203.
10. Olszowy Joanna, *Wybrane aspekty wpływające na rynek nieruchomości w Polsce*, Społeczna Akademia Nauk, Wydział Zamiejscowy w Tarnobrzegu, s. 82.
11. Melniczuk Damian, *Jak nie zabłądzić w lesie… losowym?*, https://jakbadacdane.pl/jak-nie-zabladzic-w-lesie-losowym/, [02.05.2020].
12. Mućk Jakub, *Wprowadzenie do modelowania ekonometrycznego*, Katedra Ekonomii Ilościowej Szkoły Głównej Handlowej w Warszawie, http://web.sgh.waw.pl/~jmuck/Ekonometria/EkonometriaPrezentacja1.pdf, [29.04.2020].
13. Eurostat Database, Income and Living Conditions, https://ec.europa.eu/eurostat/web/income-and-living-conditions/data/database, [02.03.2020].
14. *Metody prognozowania: Jakość prognoz*, https://m6.pk.edu.pl/materialy/mp/MP_02_bledy_prognozy.pdf, [12.04.2020].
15. *Polski boom na nowe mieszkania. Europa w tyle,* https://nieruchomosci.dziennik.pl/news/artykuly/6451997,polska-boom-nowe-mieszkania-europa-rynek-nieruchomosci.html, [16.03.2020].
16. *Rynek mieszkaniowy w UE. Polska w pierwszej dziesiątce państw z największym wzrostem cen*, https://forsal.pl/nieruchomosci/mieszkania/artykuly/1433912,ceny-mieszkan-w-ue-ceny-mieszkan-w-polsce-eurostat.html, [16.03.2020].
17. *Ustawa z dnia 23 kwietnia 1964 r. – Kodeks cywilny*, Art. 46. § 1, http://prawo.sejm.gov.pl/isap.nsf/download.xsp/WDU19640160093/O/D19640093.pdf, [06.03.2020].

8. Spis ilustracji

Rysunek 1: Rynek nieruchomości jako system 9
Rysunek 2: Zmiana ceny za 1m^2 mieszkań w latach 2015-2019 w stosunku do roku 2015 12
Rysunek 3: Elementy drzewa decyzyjnego 22
Rysunek 4: Drzewo klasyfikujące o minimalnej liczbie obiektów w węźle równej 15 24
Rysunek 5: Drzewo klasyfikujące po przycięciu do 5 węzłów 26
Rysunek 6: Wartości rzeczywiste i prognozowane w modelu drzewa regresyjnego 29
Rysunek 7: Ilustracja działania baggingu 33
Rysunek 8: Schemat predykcji lasu losowego 35
Rysunek 9: Ceny mieszkań według województw Polski w roku 2018 38
Rysunek 10: Liczba drzew w lesie losowym a jakość prognoz 40
Rysunek 11: Trafność prognostyczna najlepszego modelu lasu losowego. 42
Rysunek 12: Histogram błędu prognozy dla 1000 lasów losowych 44
Rysunek 13: Losowe drzewo decyzyjne z lasu o optymalnych parametrach 45
Rysunek 14: Istotność zmiennych objaśniających mierzona jako średni spadek dokładności ($\%IncMSE$) oraz poprawa jednorodności węzłów ($IncNodePurity$). 46
Rysunek 15: Trafność prognostyczna najlepszego i najgorszego modelu lasu losowego dla miast na prawach powiatu 50

9. Spis tabel

Tabela 1: Czynniki determinujące ceny lokali mieszkalnych w powiatach Polski 13
Tabela 2: Statystyki opisowe zbioru danych 16
Tabela 3: Macierz pomyłek dla klasyfikacji drzewa decyzyjnego 26
Tabela 4: Najlepsze modele lasów losowych według wartości 4 parametrów 42
Tabela 5: Ranking 8 najistotniejszych zmiennych objaśniających w lesie losowym 47
Tabela 6: Opis badanych lasów losowych 49
Tabela 7: Porównanie jakości prognoz trzech regresyjnych lasów losowych 50
Tabela 8: Wartości rzeczywiste i prognozowane przez optymalny model lasu losowego dla powiatów Polski bez miast 51
Tabela 9: Istotność zmiennych objaśniających w trzech lasach losowych: dla całego zbioru danych, dla miast na prawach powiatu oraz dla powiatów z wyłączeniem miast na prawach powiatu 52
Tabela 10: Porównanie błędów prognozy lasów losowych oraz regresji liniowej 56
Tabela 11: Weryfikacja hipotez badawczych 59

10. Spis równań

Równanie 1: Obliczanie parametrów modelu drzewa klasyfikującego – zasada głosowania.....23
Równanie 2: Reguła decyzyjna jako koniunkcja 25
Równanie 3: Obliczanie parametrów modelu drzewa regresyjnego – zasada uśredniania.....27
Równanie 4: Średni bezwzględny błąd procentowy ($MAPE$) 29
Równanie 5: Pierwiastek błędu średniokwadratowego ($RMSE$)..... 29
Równanie 6: Współczynnik determinacji R^2.....39
Równanie 7: Średni spadek dokładności ($\%IncMSE$).....46

11. Wykaz załączników

(Załączniki dostępne są na nośniku elektronicznym)

Załącznik 1: Plik programu RStudio z kodem źródłowym pracy (*Praca_licencjacka_Kolodziej_Sylwia.R).*
Załącznik 2: Plik programu MS Excel z danymi dot. wszystkich powiatów Polski (*ceny_mieszkan.csv).*
Załącznik 3: Plik programu MS Excel z danymi dot. miast powiatowych (*ceny_mieszkan_miasta.csv).*
Załącznik 4: Plik programu MS Excel z danymi dot. reszty powiatów (*ceny_mieszkan_bez_miast.csv).*