UNIWERSYTET ŚLĄSKI
W KATOWICACH

Wydział Nauk Ścisłych i
Technicznych

Konrad Matuszewski

354046

Badanie możliwości poprawy jakości modeli regresji uczenia maszynowego dla problemu predykcji wartości na rynku nieruchomości

PRACA DYPOLOMOWA

MAGISTERSKA

dr hab. inż. Mariusz Boryczka

Sosnowiec 2024

Spis treści

Wstęp

Rynek nieruchomości stanowi ciekawy temat badawczy, który w ostatnich latach bardzo zyskał na znaczeniu. W dzisiejszych czasach analiza i predykcja cen nieruchomości mają ogromne znaczenie dla inwestorów, pośredników oraz osób prywatnych, pragnących oszacować wartość nieruchomości na podstawie różnorodnych czynników. Zastosowanie technik regresji w dziedzinie uczenia maszynowego otwiera wiele możliwości poprawy jakości modeli predykcyjnych, które pomagają w rozwiązaniu problemu prognozy cen na rynku nieruchomości.

W ostatnich latach rynek nieruchomości nabrał dużego znaczenia ze względu na dynamiczny rozwój gospodarczy, wzrost liczby transakcji oraz rosnące zainteresowanie inwestorów w tej dziedzinie. Zmieniające się trendy i preferencje klientów, połączone z technologicznymi zaawansowaniami sprawiły, że analiza danych stała się kluczowym narzędziem dla wszystkich uczestników rynku nieruchomości.

Badanie wpływu różnych czynników, takich jak odległość od sklepów czy wskaźnik przestępczości, na cenę mieszkania jest szczególnie istotne, ponieważ pozwala lepiej zrozumieć mechanizmy kształtujące cenę nieruchomości. Włączenie dodatkowych cech, pozyskanych za pomocą zaawansowanych metod, takich jak automatycznego pobierania danych z internetu (*ang. web scraping*), pozwala na bardziej kompleksową analizę i zwiększa trafność predykcji cen.

Celem pracy było zbadanie jak na cenę mieszkania wypływają takie czynniki jak odległość od sklepów czy wskaźnik przestępczości. Do realizacji celu zostały wykorzystane metody regresji w dziedzinie uczenia maszynowego. Wartością ostatecznych wyników jest dostarczenie istotnych informacji dla inwestorów działających na tym wyjątkowo konkurencyjnym rynku. Ponadto, wykorzystanie popularnych narzędzi programistycznych, takich jak Pandas, NumPy i Matplotlib, pozwoliło na dokładną analizę danych, implementację zaawansowanych modeli regresji oraz wizualizację wyników. To stanowi solidne podstawy do uzyskania wiarygodnych i wartościowych ustaleń w dziedzinie rynku nieruchomości.

Praca koncentruje się na wykorzystaniu narzędzi uczenia maszynowego oraz bibliotek języka Python w celu stworzenia modeli predykcyjnych, które uwzględniają różnorodne czynniki wpływające na cenę nieruchomości. Celem jest nie tylko

osiągnięcie wysokiej jakości predykcji, ale również ukazanie korzyści związanych z uwzględnieniem dodatkowych danych przy prognozowaniu cen na rynku nieruchomości.

1. Uczenie maszynowe, regresja i predykcja

W dzisiejszym złożonym środowisku rynku nieruchomości, dokładne przewidywanie cen mieszkań i innych nieruchomości jest kluczowe dla wielu decyzji finansowych, inwestycyjnych i planistycznych. Rozwój technologii i dostępność dużej ilości danych umożliwiają zastosowanie zaawansowanych metod analizy, takich jak uczenie maszynowe i modele regresji, w celu lepszego zrozumienia wpływu różnych czynników na ceny nieruchomości. W niniejszym rozdziale zostaną wprowadzone kluczowe pojęcia oraz omówione zostanie, w jaki sposób mają one zastosowanie w analizie rynku nieruchomości.

1.1. Uczenie maszynowe

Uczenie maszynowe stanowi dynamiczną dziedzinę sztucznej inteligencji, która umożliwia komputerom zdobywanie wiedzy poprzez analizę danych. Przeciwnie do tradycyjnych, statycznych algorytmów programowania, uczenie maszynowe pozwala systemom uczącym się dostosowywać się do wzorców w danych i doskonalić swoje działania na podstawie doświadczenia. W kontekście analizy rynku nieruchomości, techniki uczenia maszynowego mogą być stosowane do tworzenia modeli prognostycznych, które wykorzystują dane historyczne, aby przewidywać przyszłe ceny nieruchomości.

Analizujac algorytmu uczenia maszynowego, można zwrócić uwagę że algorytmy w uczeniu maszynowym to specjalne rodzaje algorytmów, które różnią się od tradycyjnych algorytmów w tym, że nie są zaprogramowane wprost do wykonywania określonych kroków. Zamiast tego, algorytmy uczące tworzą swoje instrukcje na podstawie danych uczących. Proces ten nazywany jest uczeniem, a efektem tego procesu jest stworzenie modelu danych, który zawiera listę instrukcji lub reguł, pomagających mu wykonywać określone zadania lub analizować dane. To podejście umożliwia maszynom uczenie się na podstawie doświadczenia i dostosowywanie się do nowych sytuacji [11].

Istotnym elemntem uczenia maszynowego są dane. Zbiór uczący jest zestawem informacji przekazanym do algorytmu w celu uczenia algorytmów komputerowyc.

Zbiór danych zawiera obiekty lub przykłady pochodzące z określonej dziedziny lub tematu. Każdy z tych obiektów jest opisany za pomocą różnych cech, które pozwalają na charakteryzowanie tych obiektów. Poprzez analizę danych, algorytmy uczące się potrafią wyodrębnić wzorce, reguły i zależności, które pozwalają na podejmowanie decyzji lub wykonywanie predykcji.

Kolejnym istotym pojęciem występującym przy uczeniu maszynowym jest próbka i atrybut.

- Próbka: jest opisem pojedynczego obiektu w zbiorze danych. W przypadku zbioru uczącego, przyjmującego postać tabeli, instancjami są kolejne wiersze.
- Atrybut: to cecha użyta do opisania obiektu. Na przykład ,w przypadku zbioru uczącego w postaci tabeli, atrybutami są kolejne kolumny.

W kontekście uczenia maszynowego, analizowanie próbek i ich atrybutów jest kluczowym etapem, ponieważ algorytmy uczące wykorzystują te dane do wyodrębniania wzorców, tworzenia modeli i podejmowania decyzji. Zrozumienie, jak próbki i atrybuty są używane w procesie uczenia maszynowego, jest fundamentalne dla skutecznego tworzenia i oceny modeli predykcyjnych. Na rys.1 znajduje się zdjęcie przedstawiające schemat procesu uczenia maszynowego, który zawiera następujące etapy:

- Pozyskiwnaie danych wejściowych: początkowy etap, w którym dane są gromadzone i przygotowywane do analizy. Wprowadzenie właściwych danych jest kluczowe, ponieważ jakość danych ma wpływ na skuteczność modelu uczenia maszynowego.
- Przygotowanie danych: W tym etapie dane są przetwarzane i przygotowywane do analizy. To może obejmować oczyszczanie danych (usuwanie błędów i brakujących wartości), przekształcanie danych do odpowiednich formatów, oraz normalizację danych, aby były spójne i gotowe do dalszej obróbki.
- Analiza danych: W tej fazie model uczenia maszynowego stara się zrozumieć strukturę danych oraz zależności między różnymi cechami. Może to obejmować stosowanie technik statystycznych, wizualizacji danych oraz eksploracyjnej analizy danych (EDA - *ang. Exploratory Data Analysis*), aby odkryć istotne wzorce, tendencje i anomalie w zestawie danych. Analiza danych pomaga w

zrozumieniu, jakie informacje zawierają dane i jakie pytania można na ich podstawie zadać.

- Testowanie modelu: Bazując na przeprowadzonej analizie, model uczenia maszynowego jest trenowany, aby identyfikować i wykorzystywać znalezione wzorce. W trakcie procesu trenowania model dostosowuje swoje parametry w taki sposób, aby jak najlepiej odwzorować zależności w danych.

- Przewidywanie: Model wykorzystuje znalezione wzorce, aby dokonywać prognoz dla nowych, wcześniej niewidzianych danych. Na podstawie tych prognoz może podejmować decyzje lub wyciągać wnioski.

- Decyzyjność: Na podstawie wykonanych prognoz, system podejmuje konkretne decyzje i podejmuje odpowiednie działania. Przykładowo, jest w stanie ocenić, czy otrzymany e-mail jest spamem, czy wartościową wiadomością, a następnie kierować go do odpowiedniego folderu.

Rozróżniane są cztery główne podejścia w uczeniu maszynowym, a wybór zależy od oczekiwanego celu analizy lub właściwości zbioru danych. Poniżej przedstawiane są te cztery główne metody uczenia maszynowego, a każda z nich jest dostosowywana do konkretnych potrzeb:

- Nadzorowane uczenie maszynowe: W przypadku nadzorowanego uczenia maszynowego, model jest trenowany na danych, które posiadają etykiety lub odpowiedzi. Model jest „nadzorowany” przez poprawne wyniki i jest używany do klasyfikacji i regresji, czyli przewidywania etykiet lub wartości wynikowych na podstawie dostarczonych danych uczących.

- Nienadzorowane uczenie maszynowe: To podejście koncentruje się na analizie danych, które nie posiadają etykiet lub odpowiedzi. Głównym celem jest odkrywanie wzorców, grupowanie danych, redukcja wymiarów lub analiza skupień, co pozwala na zrozumienie struktury danych i wykrywanie ukrytych zależności.

- Częściowo nadzorowane uczenie maszynowe: W przypadku częściowo nadzorowanego uczenia maszynowego tylko część danych jest opatrzona etykietami,

podczas gdy pozostałe dane są bez etykiet. Jest to wykorzystywane, gdy dostępne są jedynie częściowe informacje zwrotne lub etykiety, co sprawia, że trenowanie modelu jest bardziej wymagające.

- Uczenie ze wzmocnieniem: Ten rodzaj uczenia skupia się na rozwijaniu umiejętności agenta (np. algorytmu) poprzez interakcję z otoczeniem. Agent podejmuje działania w środowisku i otrzymuje nagrody lub kary w zależności od wyników tych działań, dążąc do maksymalizacji zdobywanych nagród i osiągnięcia określonych celów. W takim przypadku model musi samodzielnie próbować zrozumieć ukryte wzorce i relacje w nieetykietowanych danych, co może być trudniejsze niż w przypadku uczenia maszynowego z pełnym nadzorem, gdzie każdy przykład treningowy jest opatrzony etykietą.

Różnorodność podejść w uczeniu maszynowym ma wpływ na sposób, w jaki modele są trenowane i wykorzystywane do rozwiązywania różnych rodzajów problemów. Wybór konkretnego podejścia zależy od celu analizy oraz właściwości dostępnych danych. Oceniając zdolność modelu do rozwiązywania problemów, szczególną uwagę należy zwrócić na jego skuteczność na zbiorze testowym, co pozwala na naukową ocenę jego zdolności do generalizacji na nowe dane.

W ramach procesu uczenia maszynowego, dane zazwyczaj są podzielone na trzy główne zbiory: zbiór treningowy (*ang. training set*), zbiór walidacyjny (*ang. validation set*) i zbiór testowy (*ang. test set*).

Zbiór treningowy reprezentuje część dostępnych danych, która służy jako podstawa do uczenia modelu. W trakcie tego procesu model analizuje dane treningowe, identyfikuje istotne wzorce oraz optymalizuje swoje parametry tak, aby jak najlepiej odwzorować charakterystyki tych danych.

Zbiór walidacyjny pełni funkcję dostosowania parametrów modelu, a jego istnienie pomaga uniknąć nadmiernego dopasowania do danych treningowych, co może wpłynąć na ogólną skuteczność modelu.

Z kolei zbiór testowy stanowi niezależny zbiór danych, który nie uczestniczył w procesie uczenia modelu. Jego główną rolą jest ocena zdolności modelu do generalizacji wiedzy na nowe, niewidziane wcześniej dane. Model jest testowany na danych testowych, a wyniki pomagają określić, jak skutecznie jest on w stanie dokonywać prognoz na danych pochodzących z rzeczywistego świata.

Ocena modelu na zbiorze testowym pozwala na oszacowanie jego zdolności do rozwiązywania problemów poza zbiorem treningowym. Skuteczność modelu na danych testowych dostarcza istotnych informacji na temat jego ogólnej jakości i zdolności generalizacji. Jeśli model osiąga wysoką skuteczność na zbiorze testowym, można wnioskować, że jest w stanie efektywnie rozwiązywać zadania na nowych danych. Natomiast niska skuteczność na danych testowych może wskazywać na problemy, takie jak nadmierne dopasowanie do danych treningowych (przeuczenie) lub inne niepożądane zachowania.

Podsumowując, podział na zbiór treningowy i testowy jest kluczowym aspektem procesu uczenia maszynowego, umożliwiającym naukową ocenę skuteczności modelu na nowych danych, co ma znaczenie dla praktycznej jego użyteczności.

Rysunek 1: Proces działania uczenia maszynowego

Źródło: https://www.qtravel.ai/pl/blog/uczenie-maszynowe-definicja-rodzaje-i-przyklady-zastosowania/

1.2. Model regresji

Model regresji jest fundamentalnym narzędziem statystycznym wykorzystywanym do badania zależności między różnymi zmiennymi. Jego wszechstronność i elastyczność czynią go niezastąpionym narzędziem w analizie danych, które może być dostosowywane do różnych problemów w różnych dziedzinach, takich jak ekonomia, medycyna, marketing, finanse, inżynieria oraz nauki społeczne, zgodnie z informacjami zawartymi w publikacji Andreasa Müllera i Sarah Guido [6].

Przy użyciu tych modeli można stworzyć funkcje matematyczne, które opisują relacje między zmiennymi i umożliwiają prognozowanie cen nieruchomości na podstawie dostępnych informacji. W ogólnym kontekście, regresja może przyjmować różne formy, ale dwie podstawowe kategorie to:

- Regresja liniowa: W przypadku regresji liniowej zakłada się, że istnieje liniowa zależność między zmiennymi niezależnymi a zmienną zależną.
- Regresja nieliniowa: W przypadku regresji nieliniowej zakłada się, że zależność między zmiennymi niezależnymi a zmienną zależną nie jest liniowa, co znajduje potwierdzenie w literaturze.

W kontekście analizy rynku nieruchomości, modele regresji pozwalają zrozumieć, jak różnorodne czynniki, takie jak lokalizacja, metraż, liczba pomieszczeń czy dostępność usług, wpływają na cenę nieruchomości.

1.3. Predykcja wartości na rynku nieruchomości

Problem predykcji wartości na rynku nieruchomości polega na opracowaniu modelu zdolnego do przewidywania cen nieruchomości na podstawie różnych zmiennych, takich jak cechy nieruchomości i zmiennych otoczenia. Dla analizy rynku nieruchomości, cel ten jest kluczowy, umożliwiając inwestorom, deweloperom i planistom dokonywanie informowanych decyzji na podstawie przewidywanych wartości nieruchomości. W tej części zostaną omówione szczegóły problemu predykcji wartości na rynku nieruchomości, istotne czynniki wpływające na ceny oraz znaczenie skutecznych modeli predykcyjnych.

2. Historia rynku nieruchomości w USA

Kształtowanie się cen na rynku nieruchomości to proces wielowymiarowy, podlegający wpływom różnorodnych czynników, które oddziałują na tę dynamiczną dziedzinę. Historia rynku nieruchomości rysuje się jako kompleksowy obraz zmian, odzwierciedlający nie tylko przemiany ekonomiczne, lecz także społeczne, polityczne i technologiczne trendy. W miarę upływu lat, rynek nieruchomości przechodził przez różne fazy, z każdą niosącą ze sobą charakterystyczne wyzwania i możliwości.

Wspomnienie kryzysu finansowego w latach 2007-2008 jest jednym z najważniejszych momentów w historii rynku nieruchomości. Ten globalny kryzys miał ogromny wpływ na sektor, wywołując spadki cen i niewypłacalności kredytobiorców, co miało długotrwałe skutki. Okres bezpośrednio po kryzysie charakteryzował się obniżającymi się cenami nieruchomości, co stwarzało okazje dla potencjalnych nabywców.

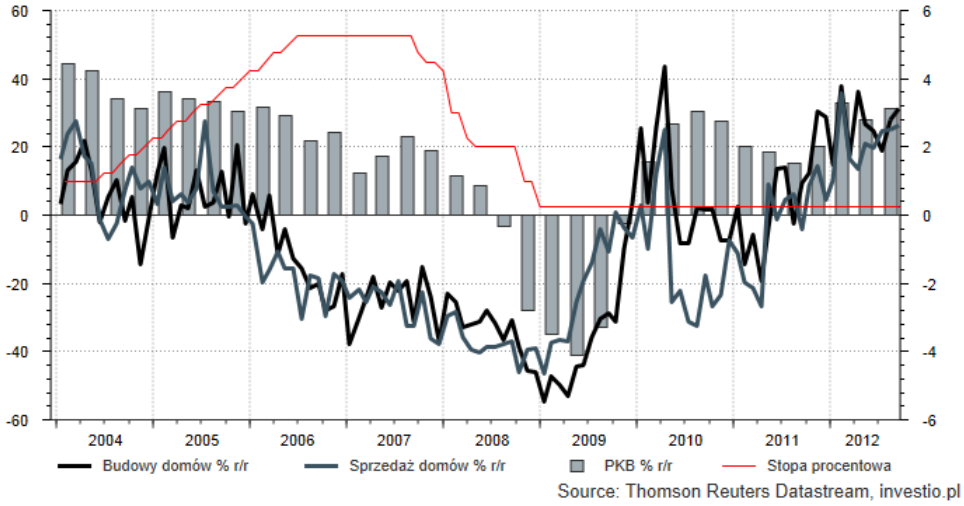
Następnie, proces odbudowy po kryzysie wprowadził nowe trendy na rynek. W miarę ożywienia gospodarki, ceny nieruchomości zaczęły rosnąć, wspomagane niskimi stopami procentowymi i rosnącym popytem. Wzrost cen nieruchomości nie był jednak jednolity i różnice w lokalizacji, infrastrukturze oraz czynnikach demograficznych przyczyniły się do zróżnicowanego tempa wzrostu w różnych regionach[7].

2.1. Wielki kryzys (2006-2009)

Wraz z malejącą liczbą sprzedanych domów i rozpoczętych budów drastycznie malały zyski przedsiębiorstw budowlanych, pośredników, sprzedaż maszyn, materiałów budowlanych i komponentów, a i trzeba także pamiętać o wpływach do budżetu państwa, zależnych w stosunkowo dużej mierze od wielkości wszystkich elementów, związanych z rynkiem nieruchomości.

Za kryzys na rynku nieruchomości często obwinia się Federalną Rezerwę (Fed) i quasi-rządowe agencje, takie jak Fannie Mae (FNMA) i Freddie Mac (FHLMC). Rząd także jest krytykowany za wspieranie zakupu nieruchomości przez osoby o słabej zdolności kredytowej. Zwolennicy ekonomii wolnorynkowej uważają, że nadmiar regulacji rządowych przyczynił się do globalnych problemów na rynkach finansowych. Istotnym problemem było również nierówne tempo wzrostu między liczbą

budynków w trakcie budowy a sprzedażą nowo wybudowanych domów. W 2006 roku proporcja ta wynosiła 80% do 75%. W miarę spadku sprzedaży nieruchomości i liczby rozpoczętych budów, straty odnotowałyprzedsiębiorstwa budowlane, pośrednicy, transport oraz doststawcy maszyn, komponentów i materiałów budowlanych. Trzeba również wziąć pod uwagę, że budżet państwa częściowo zależy od dochodów generowanych przez rynek nieruchomości. Ograniczenie tej aktywności miało wpływ na różne elementy gospodarki i wpłynęło na dalszy rozwój kryzysu. Rysunek 2 przedstawia szczegółowe fluktuacje cen nieruchomości w tym okresie, ilustrując skutki kryzysu na rynek nieruchomości [2].



Rysunek 2: Zrzut ekranu przedstawiający wachania cen w trakcie kryzysu
Źródło: https://analizy.investio.pl/rynek-nieruchomosci-w-usa-cz-2-wskaznik-wyprzedzajacy-dla-gospodarki/

2.2. Po kryzysie finansowym (2010-2012)

W okresie bezpośrednio po kryzysie finansowym w 2008 roku, rynek nieruchomości w USA zmagał się z problemami związanymi z wysoką liczbą niewypłacalności kredytobiorców, co przyczyniło się do spadku cen nieruchomości. W wyniku tego spadku, wiele osób dostrzegało możliwość zakupu nieruchomości po atrakcyjnych cenach. Sam kryzyz był uwarunkowany wieloma czynnikami, takimi jak [7]:

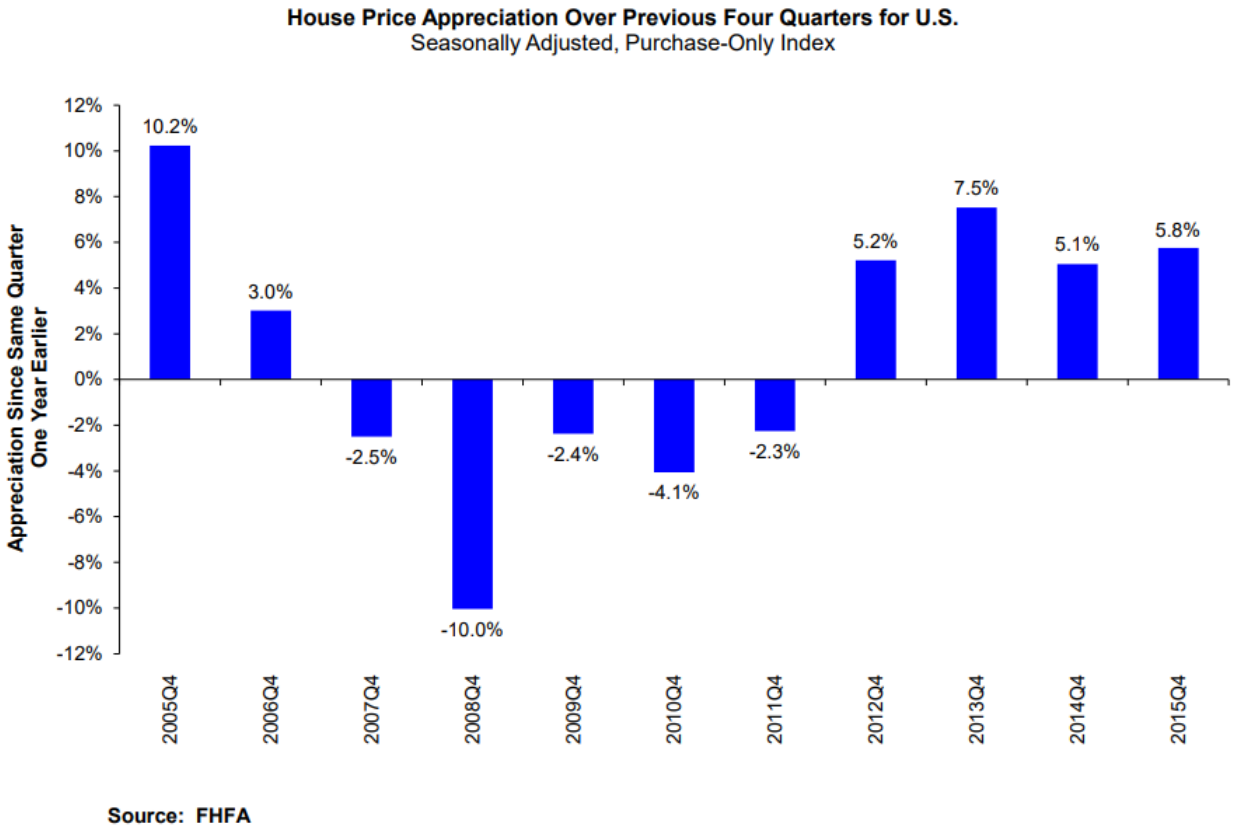
- utrzymywanie niskiego poziomu stóp procentowych,
- rozwój i wzrost znaczenia rynków finansowych,

- sponsorowanie przez rząd prywatnych instytucji,
- luzowanie kryteriów standardów kredytowych,
- sytuacja na rynku nieruchomości w USA,
- model funkcjonowania banków, w którym wykorzystywały one sekuryty-zację aktywów i kredytowe instrumenty pochodne,
- rozwój inżynierii finansowej i powstanie skomplikowanych, trudnych do wyceny instrumentów finansowych,
- model finansowania, polegający na finansowanie długoterminowych aktywów z użyciem krótkoterminowych pożyczek.

Nalezy pamiętać, że czynników moze być znacznie więcej oraz mogą być one uzaleznione od położenia geograficznego lub czasu.

2.3. Odzyskiwanie po kryzysie (2013-2015)

Po recesji z 2008 roku, gospodarka Stanów Zjednoczonych wkroczyła w fazę regeneracji, napędzaną przez politykę luzowania ilościowego (quantitative easing) prowadzoną przez Rezerwę Federalną. Wpompowanie przez Fed bilionów dolarów w system finansowy miało na celu obniżenie stóp procentowych, co z kolei pobudziło inwestycje i konsumpcję. Ożywienie gospodarcze przyczyniło się do wznowienia aktywności na rynku nieruchomości, co było widoczne w rosnących cenach oraz zwiększonym popycie, szczególnie w obszarach miejskich o wysokim statusie i ograniczonej przestrzeni, takich jak Nowy Jork i San Francisco. Dane z Federal Housing Finance Agency pokazały średnioroczny wzrost cen domów na poziomie 5-7% w latach 2013-2015 [15], sygnalizując odrodzenie sektora nieruchomości co zostało zaprezentowane na Rysunku 3.



Rysunek 3: Wzrost cen nieruchomości w ciągu ostatnich czterech kwartałów w Stanach Zjednoczonych

2.4. Wzrost cen i napięcie rynkowe (2016-2019)

W latach 2015-2019 rynek mieszkaniowy w Stanach Zjednoczonych charakteryzował się dynamicznym wzrostem cen, co było rezultatem szeregu czynników, w tym ograniczonej podaży nieruchomości i silnego popytu. Ten okres charakteryzował się także niskimi stopami procentowymi, co sprzyjało aktywności kredytowej i inwestycyjnej na rynku nieruchomości. Budownictwo jednorodzinne znacząco wzrosło co wskazuje na potencjalny dalszy wzrost w kontekście spadających stóp procentowych i zachęt dla budownictwa, które mogły przyciągnąć potencjalnych nabywców z powrotem na rynek mieszkaniowy

Problem dostępności mieszkań stał się na tyle poważny, że rząd rozważał różno-

rodne inicjatywy mające na celu stymulowanie budownictwa i wsparcie dla pierwszorazowych nabywców domów. W tym kontekście pojawiały się propozycje, takie jak rozszerzenie kluczowych programów mieszkaniowych czy wprowadzenie ulg podatkowych dla osób o niższych dochodach kupujących swoje pierwsze mieszkanie.

Ze względu na długoterminową politykę rządu, mającą na celu wyjście z kryzysu poprzez utrzymanie niskich stóp procentowych, ceny mieszkań stopniowo rosły. To jednak nie doprowadziło do kolejnego kryzysu, ponieważ zapotrzebowanie na mieszkania pozostało wysokie, a analitycy rynkowi zaczęli podkreślać znaczenie zrównoważonego rozwoju rynku oraz jego długoterminowej stabilności. Warto ndmienić że stopy procentowe po 2016 roku stopniowo rosyły jednak wzrost ten był stosunkowo mały i osiągnał trochę ponad 2% w 2019 roku.

2.5. Pandemia COVID-19 i zmiany rynkowe (2020-2021)

Pandemia COVID-19, która rozpoczęła się na początku 2020 roku, zainicjowała globalne perturbacje w wielu sektorach gospodarki, w tym w branży nieruchomości. W pierwszych miesiącach pandemii, zgodnie z danymi Federal Housing Finance Agency, rynek nieruchomości doświadczył spadku aktywności transakcyjnej, spowodowanego restrykcjami dotyczącymi przemieszczania się i niepewnością ekonomiczną.

Jednakże, w miarę adaptacji do nowej rzeczywistości, rynek zaczął wykazywać niespodziewaną odporność. Wprowadzenie pracy zdalnej jako standardu dla wielu sektorów przyczyniło się do przewartościowania potrzeb mieszkaniowych. W drugiej połowie 2020 roku obserwowano zwiększony popyt na nieruchomości o większej powierzchni i z dodatkowymi pomieszczeniami mogącymi służyć jako biuro domowe. To zjawisko, w połączeniu z rekordowo niskimi stopami procentowymi, doprowadziło do gwałtownego wzrostu cen, szczególnie w przedmieściach i obszarach miejskich oferujących lepszą jakość życia.

Badania przeprowadzone przez Urban Land Institute wskazują, że pandemia przyspieszyła również trendy takie jak suburbanizacja i migracja do „miast drugiego rzutu", które oferują więcej przestrzeni i lepszy stosunek ceny do jakości życia. To z kolei doprowadziło do rosnącego napięcia rynkowego w tych lokalizacjach, zwiększając wycenę nieruchomości i wywołując dyskusje na temat długoterminowej zrównoważoności wzrostu cen.

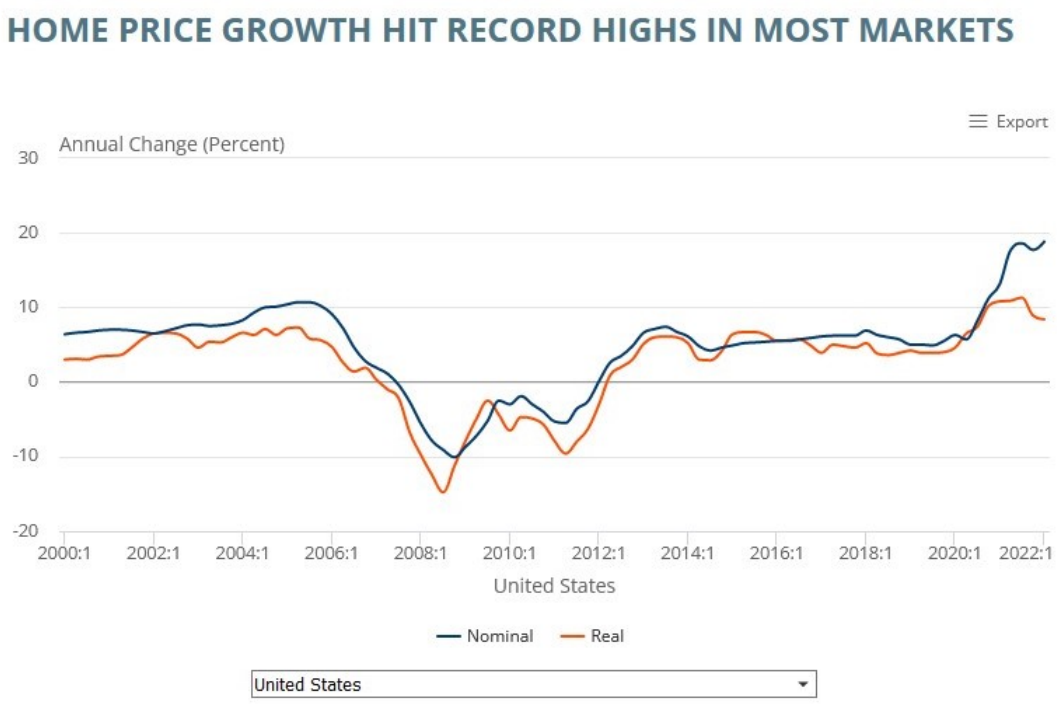
Dodatkowo, według danych Freddie Mac, deficyt mieszkań w USA został dodatkowo spotęgowany przez pandemię, osiągając poziom 3.8 miliona jednostek mieszkaniowych na koniec 2020 roku. Zjawisko to, w połączeniu z ograniczeniami w budownictwie spowodowanymi przez zakłócenia w łańcuchach dostaw i wyzwania związane z utrzymaniem protokołów zdrowotnych na placach budowy, zwiastuje potencjalne wyzwanie w sprostaniu rosnącemu popytowi.

2.6. Obecnie (2022-2023)

W ostatnich latach obserwujemy kontynuację różnorodnych trendów na rynku nieruchomości w USA, które są silnie zależne od lokalnych warunków. W miastach o intensywnym popycie, takich jak San Francisco czy Nowy Jork, ceny nieruchomości utrzymują wysoki poziom, co odzwierciedla stałą konkurencję i ograniczoną podaż. Z drugiej strony, w niektórych obszarach, szczególnie tych mniej zaludnionych lub mniej pożądanych ze względów ekonomicznych czy klimatycznych, ceny mogą wykazywać stabilizację lub nawet tendencje spadkowe.

Polityka rządowa, w tym decyzje dotyczące stóp procentowych przez Rezerwę Federalną oraz zmiany w prawodawstwie podatkowym i mieszkaniowym, nadal wpływa na decyzje inwestycyjne i akcesyjne na rynku nieruchomości. Również zmiany demograficzne, takie jak starzenie się społeczeństwa i migracje wewnętrzne, odgrywają kluczową rolę w kształtowaniu popytu na różne typy nieruchomości.

Załączony Rysunek 3 pokazuje, że wzrost cen nieruchomości w ostatnich latach osiągnął rekordowe poziomy w większości rynków, co może być wynikiem kombinacji niskich stóp procentowych, ograniczonej podaży nowych nieruchomości, a także zmiany preferencji konsumentów, którzy poszukują większej przestrzeni mieszkaniowej w odpowiedzi na pandemię COVID-19 i rosnącą popularność pracy zdalnej. Wzrost ten jest wyzwaniem dla dostępności mieszkaniowej, zwłaszcza dla pierwszorazowych nabywców i osób o niższych dochodach, co może wymagać odpowiedzi politycznej w celu zapewnienia większej równowagi na Rysunku 4.



Rysunek 4: Zrzut ekranu przedstawiający wachania cen nieruchomości
Źródło: Harvard Joint Center for Housing Studies

2.7. Czynniki wpyływajacena na ceny nieruchomości

Istnieje szereg kryteriów wpływających na cenę mieszkania, takich jak:

- lokalizacja — centrum dużych miast daje możliwość uzyskanie atrakcyjnych cen. Należy jednak pamiętać że w przypadku lokalizacji mozemy rozpatrzeć szereg innych zmiennych takich jak dostę do komunikacji miejskiej jakść infrastruktury dzielnicy lub dostęp do atrakcji miejskich,
- kondygnacja — parter oraz ostatnie piętro są najmniej pożądane ze wzgledu na oabwy na skutek kradziezy których najłatwiej dokonać na wyżej wymienionych kdygnachjach,
- stan techniczny — mieszkania w dobrym stanie technicznym, wolne od poważnych uszkodzeń i wymagające niewielkich nakładów na remonty, zazwyczaj są bardziej atrakcyjne dla potencjalnych nabywców. Taki stan obiektu może

podnieść cenę, ponieważ nowy właściciel unika natychmiastowych kosztów naprawczych i modernizacyjnych,

- rozkład pomieszczeń — funkcjonalny rozkład pomieszczeń, który sprzyja wygodnemu użytkowaniu przestrzeni, może przyciągnąć większą liczbę potencjalnych klientów. Mieszkania z dobrze zaprojektowanymi układami pomieszczeń, które spełniają potrzeby osób zamieszkujących nieruchomość, mogą być warte więcej,

- metraż — zazwyczaj większa powierzchnia mieszkania oznacza wyższą cenę. Większa przestrzeń jest postrzegana jako luksus, co może przekładać się na wyższą wartość nieruchomości. Zaobserwowano jednak że wraz ze wzrostem metrażu mieszkania cena za m^2 spada w stosununku do mieszkań o niższej powierzchni,

- atrakcyjności wykończenia — wykończenie mieszkania, takie jak jakość użytych materiałów, nowoczesność wyposażenia i estetyka wnętrza, może znacząco wpłynąć na odbiór nieruchomości. Mieszkania z wysokiej jakości wykończeniem i modnym designem są często bardziej poszukiwane, co może prowadzić do wyższej ceny sprzedaży.

W kontekście głównego celu pracy, którym jest poprawa jakości modeli regresji uczenia maszynowego dla problemu predykcji wartości na rynku nieruchomości, rozpatrzenie i zrozumienie czynników wpływających na ceny nieruchomości staje się kluczowe. Wspomniane kryteria, takie jak lokalizacja, kondygnacja, stan techniczny, rozkład pomieszczeń, metraż oraz atrakcyjność wykończenia, są nie tylko istotnymi zmiennymi, które potencjalni nabywcy biorą pod uwagę, ale również zmiennymi wejściowymi dla modeli predykcyjnych. Dążenie do tego, by wzbogacić modele regresyjne o nowe parametry, które mogą zawierać bardziej szczegółowe informacje o tych czynnikach, zwiększy precyzję i trafność przewidywań.

Przykładowo, dodanie parametrów odzwierciedlających dokładniejsze dane dotyczące lokalizacji, takie jak odległość od centrów biznesowych lub ocena infrastruktury dzielnicy, może pomóc w lepszym modelowaniu wpływu lokalizacji na wartość nieruchomości. Podobnie, szczegółowa analiza danych dotyczących kondygnacji, z uwzględnieniem specyficznych cech danego budynku, może dostarczyć modelowi bardziej złożonego obrazu, jak kondygnacja wpływa na cenę. Rozszerzenie danych o

szczegóły dotyczące stanu technicznego, wykończenia, czy metrażu pomoże w stworzeniu bardziej zróżnicowanego i adekwatnego do rzeczywistości modelu.

Wzmocnienie modeli regresji poprzez dokładniejsze uwzględnienie tych czynników pomoże nie tylko w trafniejszym określeniu wartości nieruchomości, ale także w lepszym zrozumieniu dynamiki rynku. Pozwoli to na bardziej efektywne reagowanie na zmieniające się warunki rynkowe oraz na budowanie modeli, które będą mogły być stosowane w różnych scenariuszach ekonomicznych i geograficznych, co jest szczególnie ważne w świetle bieżących zmian, jakie przynosi dynamicznie rozwijający się rynek nieruchomości.

3. Metodyka badań

Założeniem pracy było przeprowadzenie kompleksowej analizy i predykcji cen nieruchomości na rynku lokalnym, uwzględniając różnorodne czynniki wpływające na wartość nieruchomości. Celem było opracowanie skutecznych modeli regresji w dziedzinie uczenia maszynowego, które pozwolą na oszacowanie cen mieszkań i domów na podstawie danych dostępnych dla danego obszaru geograficznego.

Główne założenia pracy obejmują:

- Skompletowanie odpowiednich danych: Pierwszym krokiem było zebranie kompletnego zestawu danych dotyczących cen nieruchomości oraz różnorodnych czynników wpływających na ich wartość. Dane te obejmuje informacje o lokalizacji, wielkości, stanu technicznego, odległości od różnych udogodnień, wskaźników ekonomicznych i społecznych itp.

- Wybór odpowiednich metod analizy danych: W pracy zostaną zastosowane zaawansowane techniki analizy danych i uczenia maszynowego, w tym regresja liniowa, regresja wielomianowa, oraz inne zaawansowane modele regresyjne, w celu zrozumienia zależności między danymi a cenami nieruchomości.

- Wzbogacenie zbioru danych o dodatkowe cechy: Dodatkowe cechy, takie jak informacje o infrastrukturze, dostępności komunikacji, atrakcyjności okolicy, zostaną pozyskane przy użyciu metod takich jak „web scraping" (p obieranie danych z internetu), aby zwiększyć kompletność danych i polepszyć jakość modeli predykcyjnych.

- Walidacja modeli predykcyjnych: Zastosowane modele zostaną poddane gruntownej walidacji za pomocą odpowiednich metryk oceny jakości predykcji. Jest to kluczowe dla określenia trafności i użyteczności opracowanych modeli.

- Analiza wyników i wnioski: Po przeprowadzeniu analizy danych i predykcji cen nieruchomości, będą wyciągnięte wnioski dotyczące wpływu poszczególnych.

Wartościowe ustalenia z pracy mogą przyczynić się do lepszego zrozumienia rynku nieruchomości oraz wspomóc decyzje inwestycyjne i zarządcze na tym rynku.

Przeprowadzenie tych badań pozwoli na lepsze zrozumienie mechanizmów kształtujących ceny nieruchomości oraz na wypracowanie praktycznych narzędzi predykcyjnych, które mogą znaleźć zastosowanie w realnym środowisku rynku nieruchomości.

3.1. Doskonalenie modeli regresji w problemie predykcji wartości na rynku nieruchomości

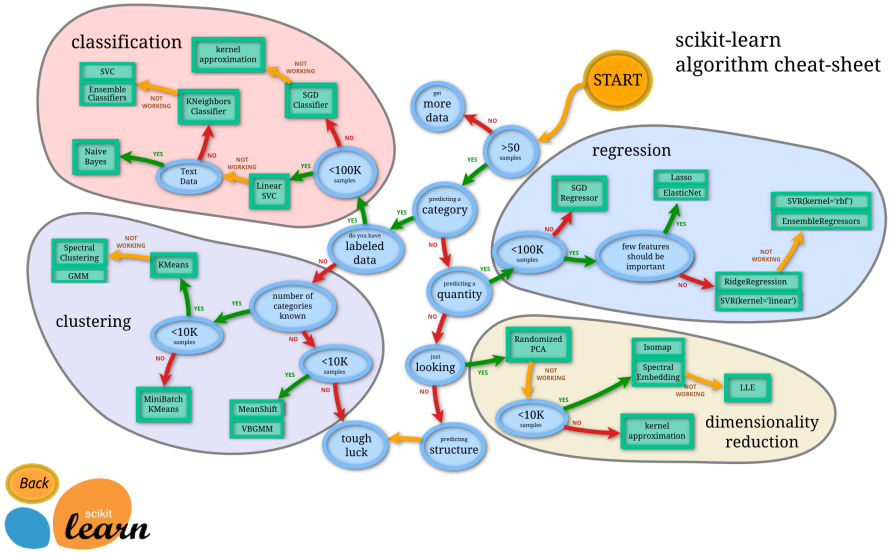
Przedmiotem niniejszej pracy było doskonalenie modeli regresji w kontekście prognozowania wartości na rynku nieruchomości. Proces precyzyjnej predykcji cen nieruchomości jest kluczowym aspektem dla wielu dziedzin, w tym dla sektora nieruchomości, finansów czy planowania urbanistycznego. W celu poprawy skuteczności tych prognoz, w pracy zaproponowano nowe podejście, które uwzględnia zarówno tradycyjne, jak i nowoczesne atrybuty.

W niniejszym rozdziale zostały przedstawione oryginalne pomysły dotyczące nowego spojrzenia na problem doskonalenia modeli regresji w kontekście rynku nieruchomości. Analiza ta obejmie zarówno klasyczne metody modelowania, jak i eksperymentalne podejścia, mające na celu zidentyfikowanie najbardziej efektywnych strategii prognozowania cen. Wprowadzenie nowych atrybutów, ich standaryzacja oraz analiza porównawcza różnych modeli regresji są kluczowymi elementami tego badania.

3.2. Wybór Modeli Uczenia Maszynowego

W poprzednich rozdziałach analizowano historię rynku nieruchomości w Stanach Zjednoczonych, przyglądając się zarówno bieżącym tendencjom, jak i czynnikom kształtującym jego rozwój. W niniejszym rozdziale skupimy się na drugim etapie naszego badania, który obejmuje wybór modeli uczenia maszynowego oraz dostarczenie im konkretnie spreparowanych danych.

Aby przewidzieć wartość nieruchomości, należy dokładnie wybrać model uczenia maszynowego. Wybór odpowiednich modeli jest kluczowy dla skutecznego przewidywania cen nieruchomości. Korzystając z schematu zamieszczonego na rys.4 dla zbioru danych wykorzystywanego w badaniu drogą elmininacji zostały wybrane dwa modele dla naszego przypadku.



Rysunek 5: Schemat wyboru modeli na podstawie parametrów zbioru danych. Źródło: https://scikit-learn.org

Wybrane modele to:

1. Lasso.
2. ElasticNet.

Z obu modeli przeprowadzona została szczegółowa analiza, uwzględniająca różnorodne aspekty ich działania. Następnie wyniki tych analiz zostały porównane, mając na celu wyłonienie modelu, który najlepiej sprawdza się w prognozowaniu cen nieruchomości.

Analiza modelu Lasso

Model Lasso (ang. *Least Absolute Shrinkage and Selection Operator*) stanowi kluczowy element w uczeniu maszynowym, zwłaszcza w kontekście regresji liniowej, gdzie selekcja cech i regularyzacja odgrywają ważną rolę. Jego unikalne właściwości, takie jak zdolność do redukcji współczynników nieistotnych cech do zera, sprawiają,

że model ten jest szczególnie przydatny w sytuacjach, gdzie analizowany zestaw danych zawiera dużą liczbę cech. Jedną z głównych zalet modelu Lasso jest jego zdolność do selekcji cech. Poprzez zastosowanie kary na absolutne wartości współczynników regresji, model Lasso redukuje wpływ mniej istotnych cech, co skutkuje prostszym i bardziej zrozumiałym modelem. Ponadto, dzięki tej technice, model Lasso skutecznie zapobiega nadmiernemu dopasowaniu, co jest kluczowe dla zachowania zdolności generalizacji modelu do nowych danych. Modele utworzone przy użyciu tego modlu charakteryzują się prostotą interpretacji. Skoncentrowanie się na istotnych cechach umożliwia lepsze zrozumienie mechanizmów stojących za danymi. Dodatkowo, model Lasso wykazuje się dużą efektywnością w przetwarzaniu dużych zestawów danych, co jest istotne w erze rosnącej ilości danych i złożoności problemów. Celem jest pokazanie wpływu różnych konfiguracji danych na wyniki modelu.

Analiza modelu ElasticNet

Model ElasticNet stanowi cenne rozszerzenie w dziedzinie uczenia maszynowego, łącząc cechy modeli Lasso i Ridge (ang. *Ridge Regression*). Jest to technika regresji liniowej, która wykorzystuje zarówno regularyzację L1 (charakterystyczną dla Lasso), jak i L2 (typową dla Ridge), aby skutecznie radzić sobie z różnymi problemami występującymi w analizie danych [10].

ElasticNet, podobnie jak Lasso, ma zdolność do redukowania współczynników nieistotnych cech do zera, co ułatwia selekcję cech. Jednakże, w odróżnieniu od Lasso, model ElasticNet stosuje także kary za względnie duże wartości współczynników (regularyzacja L2), co pomaga w sytuacjach, gdy zestaw danych zawiera cechy skorelowane. Ta kombinacja regularyzacji sprawia, że ElasticNet jest szczególnie efektywny w przypadkach, gdy występuje wielokoliniowość (wysoka korelacja między zmiennymi niezależnymi) w zestawie danych.

Kluczową zaletą ElasticNet jest jego elastyczność i zdolność do równoczesnego wykorzystania zalet obu rodzajów regularyzacji. Dzięki temu model ten jest w stanie skutecznie radzić sobie w różnorodnych scenariuszach analizy danych, szczególnie tam, gdzie Lasso mogłoby być zbyt restrykcyjne w eliminacji cech, a Ridge nie radziłoby sobie z redukcją liczby cech. Model ElasticNet jest więc rozwiązaniem bardziej wszechstronnym, zachowującym równowagę między selekcją cech a unikaniem problemów związanych z wielokoliniowością.

Podobnie jak Lasso, ElasticNet zapobiega nadmiernemu dopasowaniu (overfitting) i zachowuje zdolność generalizacji modelu do nowych danych. Jest to szczególnie ważne w kontekście uczenia maszynowego, gdzie modele muszą być zdolne do efektywnego przewidywania na podstawie nieznanych wcześniej danych. ElasticNet, dzięki swojej elastyczności i równoważeniu między L1 a L2, oferuje ulepszone możliwości dostosowywania modelu do specyfiki danych, co przekłada się na lepsze wyniki w wielu zastosowaniach.

3.3. Technologie i biblioteki

Praca została opracowana w języku Python, który stanowi często stosowaną platformę programistyczną, szczególnie cenioną w kontekście analizy danych oraz przetwarzania informacji. Python wykazuje się szeregiem właściwości, które uzasadniają jego popularność w dziedzinie analizy danych. Przede wszystkim jest to język o otwartym źródle, co implikuje dostępność jego kodu źródłowego oraz bogate wsparcie ze strony rozwijającej się społeczności programistycznej. Co więcej, Python wykazuje się prostotą w nauce i zrozumieniu, co przyczynia się do jego atrakcyjności dla początkujących programistów.

Python przyciąga uwagę w dziedzinie analizy danych z uwagi na dostępność specjalistycznych bibliotek i narzędzi, takich jak NumPy, pandas, matplotlib, seaborn, scikit-learn, TensorFlow oraz wiele innych. Wspomniane biblioteki ułatwiają manipulację danymi, przeprowadzanie analiz statystycznych, wizualizację danych oraz implementację algorytmów uczenia maszynowego, co stanowi nieodzowny element procesu analizy danych [12].

JupyterHub, jak wcześniej wspomniano, stanowi nieocenione narzędzie, pozwalające na efektywne zarządzanie wieloma instancjami serwera Jupyter Notebook na jednym serwerze lub klastrze. W aspekcie analizy danych, JupyterHub jest przydatnym środowiskiem, które umożliwia tworzenie interaktywnych notatników służących do eksploracji danych, udostępnianie wyników analizy oraz efektywną współpracę nad projektami zespołowymi. Dzięki JupyterHub, wiele osób ma możliwość współdzielenia tych samych zasobów obliczeniowych, co sprzyja wspólnej pracy nad projektami analizy danych.

3.3.1. Scikit-Learn

Scikit-Learn, często nazywany skrótowo sklearn, to popularna biblioteka do uczenia maszynowego w języku Python. Jest to otwarte oprogramowanie (open-source) i stanowi część ekosystemu narzędzi i bibliotek dostępnych w Pythonie do analizy danych, uczenia maszynowego i przetwarzania danych. Scikit-Learn oferuje wiele algorytmów i narzędzi do klasyfikacji, regresji, grupowania, redukcji wymiarów, selekcji cech i oceny modeli uczenia maszynowego [13].

3.3.2. OSMPythonTools

OSMPythonTools to biblioteka w języku Python, która dostarcza narzędzi do pracy z danymi geoprzestrzennymi zawartymi w bazie OpenStreetMap (OSM). Projekt OpenStreetMap to inicjatywa tworzenia darmowej, otwartej mapy świata, która jest tworzona i rozwijana przez społeczność użytkowników. OSM zawiera obszerne informacje geograficzne, takie jak drogi, budynki, granice administracyjne, rzeki, miejsca publiczne i wiele innych, co czyni ją cennym źródłem danych geoprzestrzennych.

W kontekście pracy magisterskiej, OSMPythonTools jest wykorzystywane do wykonywania różnych operacji na danych geoprzestrzennych z bazy OpenStreetMap. Obejmuje to wyszukiwanie i pobieranie danych geograficznych oraz przeprowadzanie analizy obszarów o określonych cechach. W tym przypadku konkretny kod wykorzystuje bibliotekę do przeszukiwania obszarów wokół podanych współrzędnych geograficznych, a następnie analizuje te obszary w poszukiwaniu określonych cezch, takich jak lasy, parki lub siłownie.

3.3.3. GeoPy's

Biblioteka geopy.distance to narzędzie w języku Python, które dostarcza funkcjonalności do obliczeń odległości geograficznych na podstawie współrzędnych geograficznych. Jest używane w kontekście naszej pracy magisterskiej do pomiaru odległości między różnymi punktami geograficznymi, takimi jak sklepy a obszary zielone lub sklepy a siłownie.

Biblioteka geopy.distance jest cennym narzędziem do wykonywania obliczeń geograficznych i umożliwia nam dokładne pomiar odległości między różnymi punktami na Ziemi. W kontekście pracy magisterskiej, ta funkcjonalność jest wykorzystywana do analizy odległosci różnych obiektów geoprzestrzennych w okolicy podanych

współrzędnych geograficznych. Wynik obliczeń za pomocą metody geopy.distance jest następnie zapisywany jako dodatkowa cecha w zbiorze danych dotyczących nieruchomości. Ta odległość stanowi istotną informację, która może być wykorzystywana w analizie oraz w procesie trenowania modeli regresji. Dzięki niej można dokładnie określić, jak daleko od nieruchomości znajduje się najbliższy punkt docelowy, co jest kluczowe dla naszej analizy rynku nieruchomości. Należ jednak podkreslić iż odległość liczona miedzy dwoma punktami jest obliczana w lini prostej a nie na podstawie rzeczywistej drogi do celu.

Geopy zawiera klasy geokodera dla usług geokodowania, takie jak OpenStreetMap Nominatim, Google Geocoding API (V3) i wiele innych. Pełna lista dostępnych usług geokodowania jest dostępna w sekcji dokumentacji Geocoderów. Klasy geokodera znajdują się w module geopy.geocoders [8].

3.3.4. Pandas

Pandas to biblioteka zarządzania danymi w języku Python, która umożliwia pracę z danymi uporządkowanymi, przypominającą strukturę danych dataFrame znanej z innych systemów obliczeniowych, takich jak język R. Pandas oferuje wiele funkcji, które pozwalają na łatwe pobieranie, przetwarzanie i analizowanie danych z różnych źródeł. W pracy magisterskiej, biblioteka Pandas jest wykorzystywana do zarządzania danymi, co obejmuje odczyt, dodawanie nowych rekordów do zbioru danych, a także przeprowadzanie różnych typów predykcji przy użyciu różnych modeli.

3.3.5. requests

Biblioteka „requests"to narzędzie w języku Python, które umożliwia wykonywanie zapytań HTTP do różnych serwerów i uzyskiwanie odpowiedzi. Jest często używana do komunikacji z serwisami internetowymi, pobierania danych z sieci i wykonywania różnych operacji związanych z przesyłaniem danych. W kontekście pracy magisterskiej, biblioteka ‘requests‘ może być wykorzystywana do interakcji z serwisami internetowymi, w tym do tworzenia zapytań HTTP i pobierania odpowiedzi.

W pracy magisterskiej biblioteka ‘requests‘ wykorzystywana jesy do komunikacji z serwerem OSRM (Open Source Routing Machine), który umożliwia obliczanie najkrótszej trasy pomiędzy dwoma punktami na mapie, w tym tras dla pieszych użytkowników. Wykonujesz zapytanie HTTP za pomocą biblioteki ‘requests‘ do ser-

wera OSRM, przesyłając współrzędne początkowe i końcowe, aby uzyskać informacje o najkrótszej trasie pomiędzy tymi punktami [14].

4. Przebieg badań

Celem tego rozdziału jest szczegółowe przedstawienie procesu obróbki danych, niezbędnego w kontekście badania dotyczącego analizy rynku nieruchomosci. Obróbka danych ma kluczowe znaczenie dla zapewnienia ich wiarygodności, poprawności oraz reprezentatywności. W tym rozdziale opisano metodykę obróbki i analizy danych pozyskanych do zbioru `kc_house_data.csv`, skupiając się na zapewnieniu jakości i spójności danych.

4.1. Przygotowanie i opracowanie danych

W celu dostarczenia modelom uczenia maszynowego odpowiednich danych, przeprowadza się proces przygotowania danych. Dane o nieruchomościach, które zostały rozszerzone o informacje dotyczące najbliższych obiektów, zawierają teraz kompleksowy zestaw cech. W procesie przygotowania danych stosujemy następujące kroki:

1. Usuwanie niepłnych danych: Sprawdzić, czy dane zawierają brakujące wartości oraz podjąć odpowiednie działania, takie jak uzupełnianie brakujących danych lub usuwanie rekordów z brakującymi danymi.

2. Kodowanie zmiennych kategorycznych: Przy obecności zmiennych kategorycznych, stosowane są odpowiednie metody kodowania, takie jak kodowanie one-hot. Polega ono na przekształceniu każdej unikalnej kategorii w osobną kolumnę, gdzie obecność danej kategorii oznacza się za pomocą 1, natomiast pozostałe kolumny przyjmują wartość 0. W ten sposób zmienną kategoryczną zamienia się na formę numeryczną, umożliwiającą efektywne jej wykorzystanie w analizie danych i modelowaniu.

3. Skalowanie cech: Dokonuje się skalowania cech, aby uzyskać takie same skale wartości, co przyczynia się do efektywnego działania niektórych modeli, między innymi regresji liniowej.

4. Podział na zbiór treningowy i testowy: Dane są dzielone na zbiór treningowy i testowy w celu oceny skuteczności modeli na danych, które nie były uwzględniane podczas treningu.

5. Inżynieria cech (jeśli konieczna): W pewnych sytuacjach korzysta się z inżynierii cech, celem utworzenia nowych atrybutów lub przekształcenia istniejących, co może wpłynąć na zdolność modeli do wykrywania wzorców.

Odpowiednio przygotowane dane umożliwiają przystąpienie do procesu trenowania i oceny wybranych modeli, co zostało szczegółowo opisane w kolejnych rozdziałach.

4.1.1. Przystosowanie danych dla modeli uczenia maszynowego

W celu dostarczenia modelom uczenia maszynowego odpowiednich danych, przeprowadza się proces przygotowania danych. Dane o nieruchomościach, które zostały rozszerzone o informacje dotyczące najbliższych obiektów, zawierają teraz kompleksowy zestaw cech. W procesie przygotowania danych stosujemy następujące kroki:

1. Usuwanie niepłnych danych: Sprawdzić, czy dane zawierają brakujące wartości oraz podjąć odpowiednie działania, takie jak uzupełnianie brakujących danych lub usuwanie rekordów z brakującymi danymi.

2. Kodowanie zmiennych kategorycznych: Przy obecności zmiennych kategorycznych, stosowane są odpowiednie metody kodowania, takie jak kodowanie one-hot. Polega ono na przekształceniu każdej unikalnej kategorii w osobną kolumnę, gdzie obecność danej kategorii oznacza się za pomocą 1, natomiast pozostałe kolumny przyjmują wartość 0. W ten sposób zmienną kategoryczną zamienia się na formę numeryczną, umożliwiającą efektywne jej wykorzystanie w analizie danych i modelowaniu.

3. Skalowanie cech: Dokonuje się skalowania cech, aby uzyskać takie same skale wartości, co przyczynia się do efektywnego działania niektórych modeli, między innymi regresji liniowej.

4. Podział na zbiór treningowy i testowy: Dane są dzielone na zbiór treningowy i testowy w celu oceny skuteczności modeli na danych, które nie były uwzględniane podczas treningu.

5. Inżynieria cech (jeśli konieczna): W pewnych sytuacjach korzysta się z inżynierii cech, celem utworzenia nowych atrybutów lub przekształcenia istniejących, co może wpłynąć na zdolność modeli do wykrywania wzorców.

Odpowiednio przygotowane dane umożliwiają przystąpienie do procesu trenowania i oceny wybranych modeli, co zostało szczegółowo opisane w kolejnych rozdziałach.

4.1.2. Pozyskiwanie nowych danych

W rozdziale tym przedstawiono szczegółowy opis pozyskanych danych do pracy magisterskiej. Dane zostały pozyskane za pomocą specjalnie opracowanego kodu, który wykorzystuje różne źródła informacji, takie jak OpenStreetMap i usługi geolokalizacyjne. Pozyskane dane dotyczą sześciu kategorii obiektów: sklepów spożywczych, przystanków autobusowych, szkół, restauracji przychodni medycznych oraz liczby atrakcji. W celu pozyskania danych zastosowano następujący proces:

1. Wczytanie danych źródłowych z pliku CSV:
    - Wczytano dane geograficzne badanych obiektów, zakładając, że plik CSV zawiera kolumny „Lat"i „Long"reprezentujące współrzędne geograficzne tych obiektów.
2. Definicja funkcji znajdującej najbliższy przystanek autobusowy:
    - Przy użyciu API Overpass i zapytania Overpass, wyszukano przystanki autobusowe w określonym promieniu od współrzędnych badanego obiektu.
    - Znaleziono najbliższy przystanek autobusowy na podstawie minimalnej odległości w linii prostej.
3. Dodanie kolumn do istniejącej ramki danych:
    - Do istniejącej ramki danych dodano kolumny zawierające informacje o nazwie najbliższego przystanku autobusowego, jego współrzędnych oraz odległości w linii prostej od badanego obiektu.
4. Przetwarzanie danych w ramce danych:

- Dla każdego rekordu w ramce danych, wykorzystano funkcję znajdującą najbliższy przystanek autobusowy.
- Wyniki tej operacji zostały zapisane w odpowiednich kolumnach ramki danych.

5. Zapis wyników do nowego pliku CSV:
   - Po przetworzeniu danych, wyniki zostały zapisane do nowego pliku CSV.

Rysunek 6: Schemat działania algorytmu pozyskania nowych danych

Dla każdej z sześciu poszukiwanych kategorii obiektów, opracowano odrębną funkcję. Każda z tych funkcji ma za zadanie indywidualnie obliczyć trzy kluczowe parametry: nazwę obiektu, współrzędne geograficzne obiektu oraz odległość w linii prostej od badanego obiektu. Ten schemat działania został zastosowany osobno dla każdej z kategorii obiektów w ramach pracy magisterskiej. Podział na osobne funkcje dla każdej z sześciu kategorii obiektów został przeprowadzony w celu efektywnego zarządzania dużą ilością danych. Baza danych zawiera aż 23 000 rekordów, co sprawia, że operowanie na jednym ogólnym bloku kodu byłoby niewygodne i

mniej efektywne. Dzięki podziałowi na osobne funkcje możliwe jest weryfikowanie i przetwarzanie mniejszych porcji nowo przybyłych danych, co ułatwia kontrolę nad procesem i pozwala na uniknięcie problemów z wydajnością, które mogłyby pojawić się w przypadku przetwarzania wszystkich rekordów naraz.

Dodatkowo, podział na osobne funkcje umożliwia elastyczne zarządzanie procesem pozyskiwania danych. W przypadku wystąpienia zbyt wielu zapytań lub problemów technicznych, proces dla danej kategorii obiektów może zostać zatrzymany i wznowiony od ostatniego poprawnie zakończonego etapu, co minimalizuje ryzyko utraty danych i ułatwia administrację całym procesem. Dzięki takiemu podejściu możliwe jest efektywne i bezpieczne przetwarzanie dużych ilości danych w ramach pracy magisterskiej.

Tabela 2: Opis nowych atrybutów w zbiorze danych

Nazwa	Opis
nearest_grocery_name	Nazwa najbliższego sklepu spożywczego
nearest_grocery_address	Adres najbliższego sklepu spożywczego
distance_to_nearest_grocery	Odległość w linii prostej do najbliższego sklepu spożywczego
bus_stop_name	Nazwa najbliższego przystanku autobusowego
bus_stop_coords	Współrzędne przystanku autobusowego
distance_to_bus_stop_meters	Odległość w linii prostej do przystanku autobusowego
real_distance_bus	Długość najkrótszej drogi do przystanku autobusowego
school_name	Nazwa najbliższej szkoły
school_coords	Współrzędne szkoły
distance_to_school_meters	Odległość w linii prostej do szkoły
real_distance_school	Długość najkrótszej drogi do szkoły
restaurant_name	Nazwa najbliższej restauracji
restaurant_coords	Współrzędne restauracji
distance_to_restaurant_meters	Odległość w linii prostej do restauracji
real_restaurant	Długość najkrótszej drogi do restauracji
clinic_name	Nazwa najbliższej przychodni
clinic_coords	Współrzędne przychodni
distance_to_clinic_meters	Odległość w linii prostej do przychodni
real_clinic	Długość najkrótszej drogi do przychodni
liczba_atrakcji	Suma atrakcji z tagiem POI

4.1.3. Metody pozyskania nowych danych

W kontekście rosnącej liczby aplikacji związanych z analizą danych geograficznych, OpenStreetMap (OSM) staje się bardzo przydatnym narzędziem umożliwiającym wzbogacanie istniejących baz danych o aktualne informacje geograficzne. W niniejszej pracy OSM jest stosowane w celu uzupełnienia bazy danych mieszkań o informacje dotyczące najbliższych sklepów i przystanków. Ponadto OSM jest klu-

czowym narzędziem do realizacji procesu geokodowania, czyli przekształcania adresów na współrzędne geograficzne. Dzięki wykorzystaniu OSM, możliwe jest automatyczne przekształcanie adresów nieruchomości na współrzędne, co znacząco ułatwia porównywanie ich lokalizacji oraz dodawanie ich do naszej bazy danych. Pierwszym etapem pracy jest rozszerzenie zbioru danych o nowe parametry, które potencjalnie mogą wpłynąć na predykcję wartości. W tym celu zostały utworzone algorytmy, które wyszukują dodatkowe informacje i je agregują. W programie występują następujące funkcje:

- find_nearest_bus_stop
- find_nearest_school
- find_nearest_restaurant
- find_nearest_clinic
- find_nearest_grocery_store
- find_count_points

Każda z funkcje jest odpowidzialna za zlokalizowanie innego rodzaju usługi istotnej dla rynku nieruchomości. Ogólny schemat działania funkcje wyszukiwania najbliższych obiektów jest bardzo zbliżony do siebie. Algorytmy 1 i 2 reprezentują funkcję find_neares_transport_stop.

Algorytm 1 Znajdowanie najbliższego obiektu

Wejście: Dane geograficzne obiektów z pliku CSV
Wyjście: Ramka danych z informacjami o najbliższych obiektach
for Każdy rekord w ramce danych row **do**
$lat \leftarrow row[\text{"}lat\text{"}]$
$lng \leftarrow row[\text{"}long\text{"}]$
$(object\_name, restaurant\_coords, distance) \leftarrow$ Znajdź najbliższy obiekt(lat, lng)
$row[\text{"}object\_name\text{"}] \leftarrow restaurant\_name$
$row[\text{"}object\_coords\text{"}] \leftarrow restaurant\_coords$
$row[\text{"}object\_to\_object\_meters\text{"}] \leftarrow distance$
end for
$df \rightarrow$ Zapisz wyniki do nowego pliku CSV "wyniki.csv"

Algorytmy 1 i 2 można podzielić na dwie główne sekcje:

Algorytm 2 Procedura znajdowania najbliższego obiektu

Wejście: Współrzędne geograficzne obiektu (lat, lng)
Wyjście: Najbliższy obiekt
$api \leftarrow$ Inicjalizuj API Overpass
$overpass\_query \leftarrow$ Zbuduj zapytanie Overpass na podstawie współrzędnych obiektu
$result \leftarrow$ Wykonaj zapytanie Overpass
if $result$ jest różne od $None$ **then**
$nearest\_object, min\_distance \leftarrow$ Znajdź najbliższy obiekt na podstawie odległości w linii prostej
for Każdy obiekt w $result.nodes()$ **do**
$distance \leftarrow$ Oblicz odległość w linii prostej między nieruchomością a obiektem
if $distance$ jest mniejsza od $min\_distance$ **then**
$min\_distance \leftarrow distance$
$nearest\_object \leftarrow$ Aktualny obiekt
end if
end for
$object\_name \leftarrow$ Pobierz nazwę najbliższego obiktu
$object\_coords \leftarrow$ Zbuduj ciąg znaków zawierający współrzędne obiektu
Zwróć $object\_name, object\_coords, min\_distance$
end if

- Wyszukiwanie najbliższych obiektów:

  Algorytmy rozpoczynają działanie od określenia rodzajów poszukiwanych obiektów, takich jak przystanki autobusowe lub stacje metra. Następnie przeszukują okolicę nieruchomości o podanych współrzędnych geograficznych (szerokość i długość geograficzna) w promieniu 1000 metrów, używając usług OSM. Jeśli zostanie znaleziony przynajmniej jeden obiekt danego rodzaju w zadanym promieniu, metoda zwraca informacje o najbliższym obiekcie, takie jak nazwa i adres.

- Rozszerzenie danych nieruchomości:

  Jeśli w wyniku wyszukiwania zostanie znaleziony najbliższy obiekt, dane dotyczące nieruchomości zostają rozszerzone o trzy dodatkowe parametry:

  - Nazwa najbliższego obiektu: Nazwa obiektu, na przykład nazwa przystanku autobusowego.
  - Adres najbliższego obiektu: Adres obiektu, co ułatwia identyfikację lokalizacji.
  - Odległość do najbliższego obiektu: Obliczona odległość od nieruchomości

do najbliższego obiektu, przy użyciu funkcji distance_matrix z OSM. Odległość ta jest mierzona w metrach.

4.1.4. Opis zbioru danych

W celu przeprowadzenia badania, kluczowym etapem było pozyskanie odpowiedniego zbioru danych, który umożliwiłby analizę rynku nieruchomości. Wybranym zbiorem danych jest `kc_house_data`, który zawiera istotne informacje dotyczące sprzedaży nieruchomości w King County, w stanie Waszyngton, USA.

Zbiór `kc_house_data` został pozyskany z platformy Kaggle, która zapewnia wysoką jakość danych oraz reprezentatywność dla badanego obszaru. Zawiera on rozmaite atrybuty, takie jak lokalizacja nieruchomości, jej powierzchnia, liczba sypialni i łazienek, rok budowy, a także cenę sprzedaży, która było stanowiła zmienną celu w analizie.

Dzięki temu zbiorowi danych możliwe było przeprowadzenie wszechstronnych analiz i predykcji dotyczących rynku nieruchomości. Wybór tego zbioru danych pozwala na badanie zależności między różnymi cechami nieruchomości a jej ceną, identyfikację kluczowych czynników wpływających na wartość nieruchomości oraz konstruowanie skutecznych modeli predykcyjnych.

Podczas analizy `kc_house_data` dążono do uzyskania rzetelnych i wartościowych wniosków, które pozwolą na lepsze zrozumienie rynku nieruchomości w badanym regionie oraz dostarczą istotnych informacji dla decydentów i inwestorów.

Zestaw danych zawiera 21613 obserwacji z 19 cechami plus cena domu. W tabeli nr 1 podano nazwy kolumn:

Tabela 1: Opis cech w zbiorze danych

Nazwa kolumny	Opis
ID	Numer id w zbiorze danych
date	Data sprzedaży nieruchomości (String)
price	Cena sprzedanej nieruchomości
bedrooms	Liczba sypialni
bathrooms	Liczba łazienek
sqft_living	Powierzchnia salonu
sqrt_log	Powierzchnia działki
floors	Suma pięter w nieruchomości
waterfront	Czy dom ma widok na nabrzeże (1: tak, 0: nie)
view	Zakres od 0 do 5 (w 90% wartości równa 0)
condition	Stan nieruchomości
grade	ocena domu
sqft_above	Powierzchnia domu bez piwnicy
sqft_basement	Powierzchnia piwnicy
yr_built	Rok budowy
yr_renovated	Rok remontu domu
zipcode	Kod pocztowy domu
lat	Współrzędna szerokości geograficznej
long	Współrzędna długości geograficznej
sqft_living15	Powierzchnia salonu w 2015 r. (oznacza pewne remonty)
sqrt_lot15	Powierzchnia działki w 2015 r. (zawiera pewne renowacje)

Źródło: https://www.kaggle.com/code/lashkingl/eda-kc-house-data

4.1.5. Usuwanie duplikatów i czyszczenie danych

W procesie analizy danych, szczególnie w kontekście badań naukowych, kluczowe jest zapewnienie wiarygodności i czystości zbioru danych. Jak wskazują McHugh (2012) [4] jednym z fundamentalnych etapów wstępnego przetwarzania danych jest eliminacja duplikatów, mająca na celu zapobieganie błędom analitycznym i skrzywieniom w interpretacji wyników.

Biblioteka Pandas w Pythonie, będąca standardowym narzędziem w analizie danych, oferuje funkcję drop_duplicates(), która automatycznie identyfikuje i usuwa duplikaty w zbiorze danych. Jak definiuje Wes McKinney, twórca Pandas, drop_duplicates() porównuje wiersze w DataFrame, umożliwiając wykrycie i eliminację powtarzających się rekordów. Domyślnie, funkcja zachowuje pierwszy napotkany wiersz, eliminując pozostałe kopie, co jest poprawne w kontekście przetwarzania danych.

Eliminacja duplikatów jest niezbędna dla zapewnienia jakości danych, co jest kluczowe w modelowaniu statystycznym i analizie danych. Duplikaty mogą prowadzić do błędnych wniosków statystycznych.

Kolejnym etapem była eliminacja wierszy z brakującymi wartościami. Ten krok jest kluczowy, aby zapewnić kompletność danych, co jest istotne dla dokładności przeprowadzonych analiz.

4.1.6. Z-score i wykrywanie wartości odstających

Po zidentyfikowaniu i usunięciu outlierów z kolumny „price", przeprowadzono kolejny kluczowy etap obróbki danych. W ramach tego etapu, wartości cen zostały zaokrąglone do najbliższych setek, co zwiększa czytelność i praktyczność danych, jednocześnie redukując wpływ drobnych, potencjalnie mylących różnic. Powołując sie na prace How to Detect and Handle Outliers [3] w której zwrócono uwagę na to, że nawet drobne błędy w zaokrąglaniu mogą mieć znaczący wpływ na analizę danych, szczególnie w dużych zbiorach, gdzie nawet niewielkie niedokładności mogą się kumulować i prowadzić do błędnych wniosków, zatem zaokrąglanie danych może znacząco wpłynąć na jakość analizy statystycznej, zwłaszcza w dużych zbiorach danych, takich jak rynek nieruchomości.

Nowa kolumna „P" została utworzona w celu zapisania tych przetworzonych wartości, zapewniając tym samym, że oryginalna kolumna „price" pozostała nienaruszona dla ewentualnych dalszych analiz. Kolumna „P" stanowi więc oczyszczoną i ustandaryzowaną wersję danych cenowych, gotową do dalszych analiz statystycznych i modelowania predykcyjnego. Fragment kodu odpowiedzialnego za to przetwarzanie został przedstawiony poniżej:

```
dj = []
for i in df2.price:
    if i in set(outlier):
        dj.append(0.0)
    else:
        dj.append(round(i, -5))  # Zaokrąglenie do najbliższych setek
df2['P'] = dj
df2 = df2[df2['P'] != 0.0]
```

Usunięcie obserwacji, w których „P" wynosi 0.0, było niezbędne do zapewnienia, że wartości odstające nie będą miały wpływu na dalsze etapy analizy.

4.1.7. Przetwarzanie kolumny „price" i tworzenie kolumny „P"

Po zidentyfikowaniu i usunięciu outlierów z kolumny „price", przeprowadzono kolejny kluczowy etap obróbki danych. W ramach tego etapu wartości cen zostały zaokrąglone do najbliższych setek, co zwiększa czytelność i praktyczność danych, jednocześnie redukując wpływ drobnych, potencjalnie mylących różnic. Powołując się na pracę opublikowaną w artykule „Statistical Analysis of Rounded Data: Measurement Errors vs Rounding Errors"[9], autorzy wykazali, że drobne błędy w zaokrąglaniu mogą mieć znaczący wpływ na analizę danych, zwłaszcza w dużych zbiorach, gdzie nawet niewielkie niedokładności mogą się kumulować i prowadzić do błędnych wniosków. Zatem zaokrąglanie danych może znacząco wpłynąć na jakość analizy statystycznej, zwłaszcza w dużych zbiorach danych, takich jak rynek nieruchomości.

Nowa kolumna „P" została utworzona w celu zapisania tych przetworzonych wartości, zapewniając tym samym, że oryginalna kolumna „price" pozostała nienaruszona dla ewentualnych dalszych analiz. Kolumna „P" stanowi więc oczyszczoną i ustandaryzowaną wersję danych cenowych, gotową do dalszych analiz statystycznych i modelowania predykcyjnego. Fragment kodu odpowiedzialnego za to przetwarzanie został przedstawiony poniżej:

```
dj = []
for i in df2.price:
    if i in set(outlier):
        dj.append(0.0)
    else:
        dj.append(round(i, -5))  # Zaokrąglenie do najbliższych setek
df2['P'] = dj
df2 = df2[df2['P'] != 0.0]
```

Usunięcie obserwacji, w których „P" wynosi 0.0, było niezbędne do zapewnienia, że wartości odstające nie będą miały wpływu na dalsze etapy analizy.

4.1.8. Konwersja typów danych

Ostatni etap procesu obróbki danych dotyczył konwersji typów danych, a w szczególności przekształcenia kolumny 'date' na typ daty/czasu. Ta konwersja jest niezbędna z kilku powodów:

- Ułatwienie analizy czasowej: Dane czasowe często są zapisywane w różnych formatach, co może utrudniać ich analizę. Przekształcenie kolumny „date"na jednolity format daty/czasu jest kluczowe dla ujednolicenia danych i ułatwienia analizy trendów czasowych. Spójny format daty/czasu umożliwia lepsze wykorzystanie funkcji i technik analitycznych specyficznych dla danych czasowych.

- Zgodność z narzędziami analizy danych: Wiele narzędzi i bibliotek analizy danych, takich jak Pandas w Pythonie, wymaga, aby dane czasowe były w odpowiednim formacie, aby można było efektywnie korzystać z funkcji dotyczących analizy serii czasowych. Konwersja typu danych zwiększa kompatybilność danych z tymi narzędziami.

- Poprawa dokładności modeli statystycznych i predykcyjnych: Niektóre modele statystyczne i algorytmy uczenia maszynowego są specjalnie zaprojektowane do pracy z danymi czasowymi. Dlatego konwersja danych do odpowiedniego formatu daty/czasu jest kluczowa dla dokładności tych modeli.Precyzyjna analiza danych czasowych jest niezbędna do efektywnego prognozowania i modelowania.

W naszym badaniu, kolumna „date" została przekonwertowana na typ daty/czasu w Pythonie za pomocą funkcji `pd.to_datetime`, co umożliwiło dokładną i efektywną analizę serii czasowych:

```
df2['date'] = pd.to_datetime(df2['date'])
```

Ta operacja zapewniła, że dane czasowe są reprezentowane w sposób spójny i dokładny, co jest niezbędne dla dalszej analizy trendów i wzorców w danych.

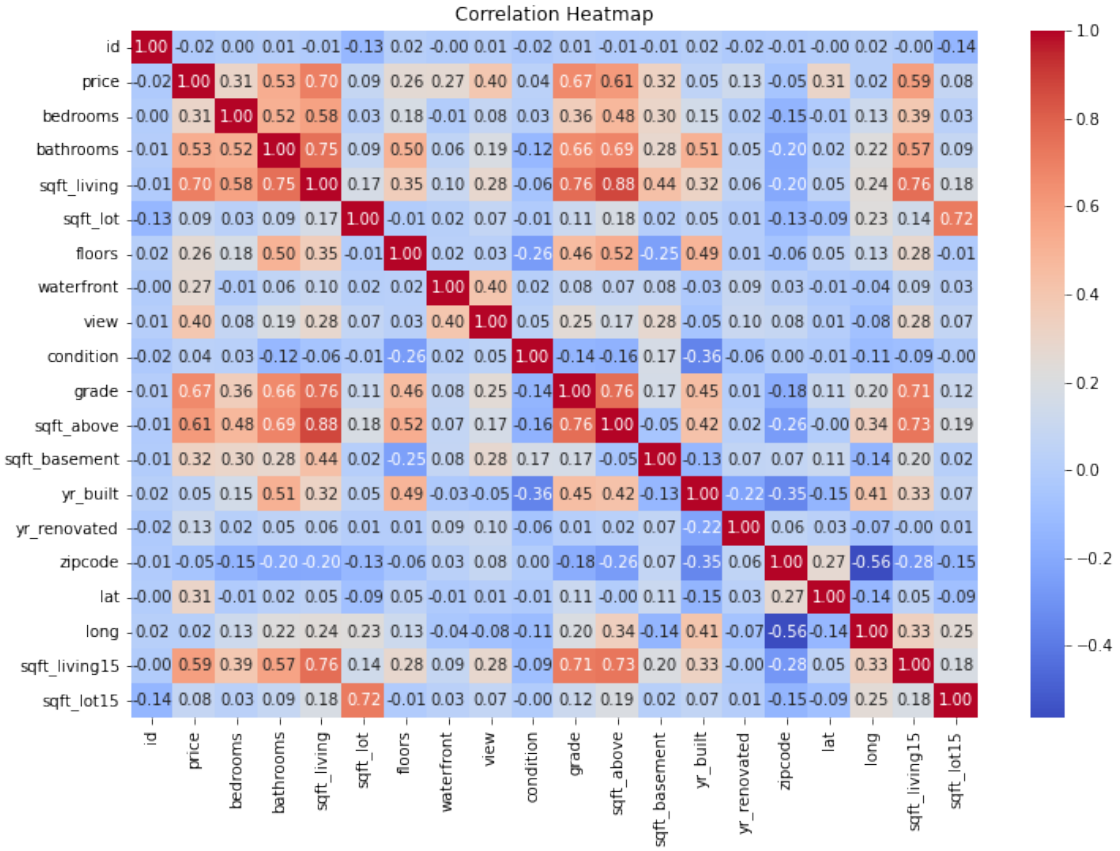
4.2. Analiza korelacji danych

Analiza korelacyjna jest jednym z fundamentalnych narzędzi w statystyce pozwalającym na ocenę zależności między zmiennymi. Mapa korelacji, czyli graficzna reprezentacja współczynników korelacji Pearsona, jest szczególnie użyteczna do wizualnego przedstawienia siły i kierunku zależności między parami zmiennych w badaniu. Współczynnik korelacji Pearsona, przyjmujący wartości od -1 do 1, wskazuje na siłę i kierunek liniowej zależności między zmiennymi; wartości bliskie 1 lub -1 oznaczają silną korelację dodatnią lub ujemną, podczas gdy wartość bliska 0 oznacza brak liniowej zależności .

Artykuł autorstwa Gogtay i Thatte „Usefulness of Correlation Analysis"omawia zastosowanie współczynnika korelacji Pearsona w analizie statystycznej. Praca ta zawiera informacje na temat interpretacji współczynnika korelacji Pearsona, wskazując na jego zastosowanie do oceny siły i kierunku zależności między zmiennymi. Artykuł podkreśla, że współczynnik korelacji Pearsona jest przydatny w ocenie liniowej zależności między zmiennymi, ale nie dostarcza informacji o przyczynowości tej zależności.

4.2.1. Mapa korelacji standardowych parametrów nieruchomości

W tej sekcji skupiono się na analizie korelacji standardowych parametrów rynku nieruchomości takich jak powierzchnia, liczba pokoi, rok budowy itp. Rozumienie tych zależności jest niezbędne do wyceny i predykcji cen na rynku nieruchomości.



Rysunek 7: Mapa korelacji standardowych parametrów nieruchomości

W analizie rynku nieruchomości, korelacje między różnymi parametrami odgrywają istotną rolę w zrozumieniu czynników wpływających na wartość nieruchomości. Mapa korelacji ujawnia, że cena nieruchomości ma tendencję do wzrostu wraz ze wzrostem jej powierzchni, co odzwierciedla powszechną praktykę wyceny nieruchomości, gdzie większe metraże są zwykle bardziej cenione. Dodatkowo, liczba pokoi w nieruchomości ma istotny wpływ na jej cenę, co może być związane nie tylko z samą powierzchnią, ale również z zapotrzebowaniem na przestrzeń mieszkalną i możliwości adaptacyjne nieruchomości.

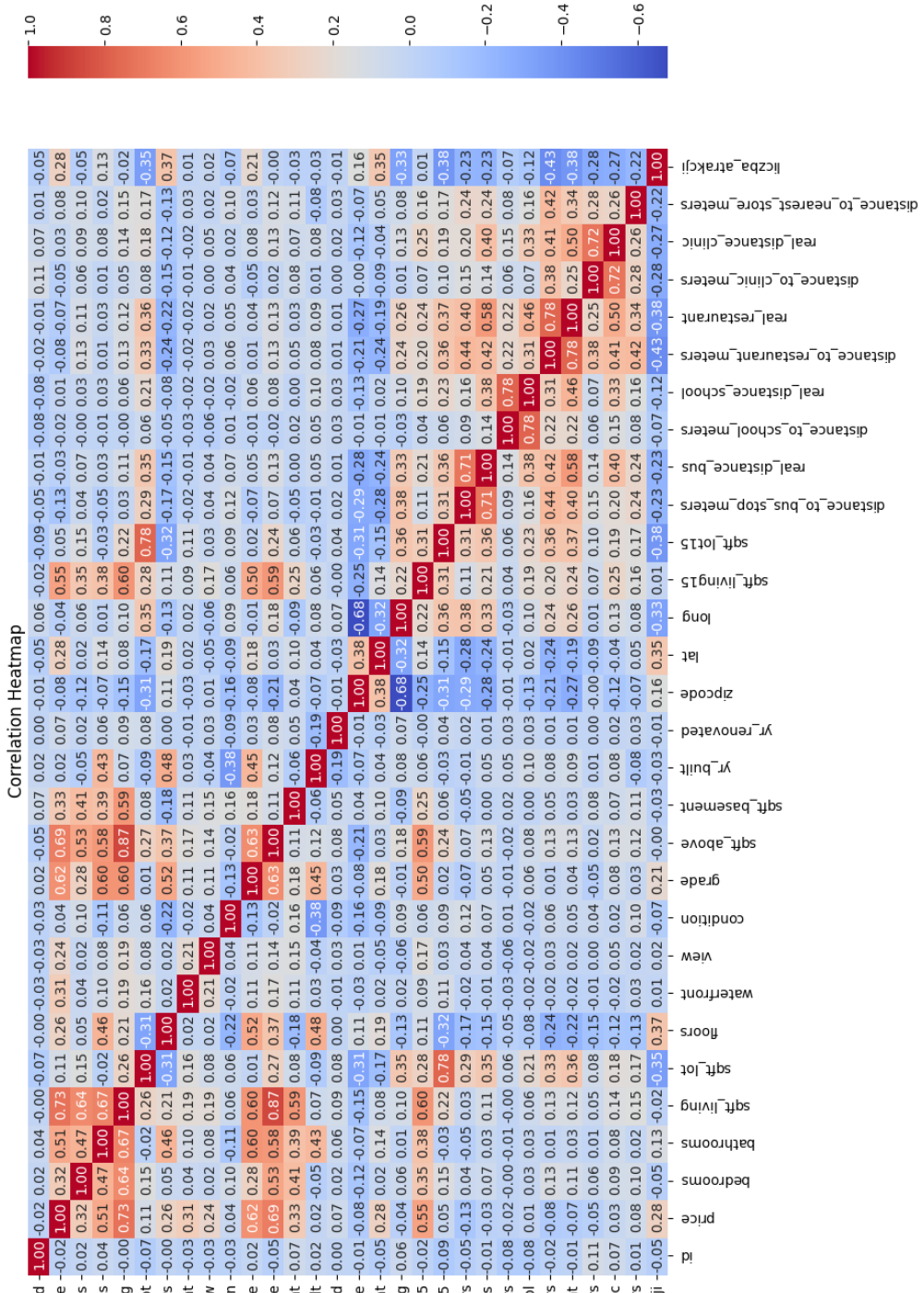
Podczas analizy wieku budynku, nowoczesne konstrukcje zazwyczaj cechuje wyższa cena, co wynika m.in. z ich nowoczesnej architektury, niższej potrzeby remontów oraz ogólnej atrakcyjności dla potencjalnych nabywców. Jakość wykonania, oceniana na podstawie stanu technicznego i stopnia wykończenia, również odgrywa kluczową rolę, choć może być trudniejsza do jednoznacznego określenia bez konkretnych danych liczbowych.

Nie można także ignorować wpływu lokalizacji, która choć nie jest zawsze bezpośrednio mierzona, ma kluczowe znaczenie dla ceny nieruchomości. Korelacje cen z kodami pocztowymi lub współrzędnymi geograficznymi mogą sugerować, że nieruchomości w preferowanych lokalizacjach osiągają wyższe ceny ze względu na dostępność usług, lepsze szkoły czy pożądane otoczenie.

Należy jednak pamiętać, że mapy korelacji mogą jedynie wskazywać na istnienie liniowej zależności między zmiennymi, a nie na bezpośrednie przyczynowo-skutkowe związki. Wartości te są istotnym wskaźnikiem dla osób zajmujących się wyceną i analizą rynku nieruchomości, ale powinny być uzupełnione dalszymi analizami statystycznymi, aby zapewnić dokładne i wiarygodne wnioski.

4.2.2. Mapa korelacji z dodatkowymi parametrami odległości od usług

Nowe parametry zostały wprowadzone do analizy w celu zbadania, w jaki sposób odległość od kluczowych usług publicznych i komercyjnych wpływa na ceny nieruchomości. Te parametry obejmują odległość do szkół, restauracji, klinik i innych ważnych punktów usługowych.



Rysunek 8: Mapa korelacji standardowych parametrów nieruchomości

Analizując dostarczoną mapę korelacji z dodatkowymi parametrami, można zauważyć, że nie wszystkie z dodatkowych czynników odległościowych wpływają w sposób znaczący na cenę nieruchomości. W szczególności, parametry takie jak odległość do transportu publicznego, szkół, restauracji czy klinik nie wykazują silnej korelacji z ceną. To sugeruje, że w tym konkretnym zestawie danych, te aspekty lokalizacji mogą być mniej istotne dla potencjalnych nabywców. W kontrastowej analizie, interesująco, to liczba atrakcji turystycznych w najbliższej okolicy prezentuje największą korelację z ceną nieruchomości. Może to wskazywać, że właśnie obecność atrakcji, a nie codzienne udogodnienia, stanowi główny czynnik przyciągający kupujących i wpływający na wycenę nieruchomości w analizowanym zbiorze danych.

Podsumowując, na dostarczonej mapie korelacji, dodatkowe parametry odległości od usług nie wykazują silnych liniowych zależności z ceną nieruchomości. Oznacza to, że w analizowanym zbiorze danych te czynniki mogą nie odgrywać kluczowej roli w określaniu wartości nieruchomości, lub też związki te mogą być nieliniowe lub zależne od innych zmiennych nieuwzględnionych bezpośrednio w analizie. Interpretacja wyników powinna zatem być przeprowadzona z ostrożnością i najlepiej zostać uzupełniona o dodatkowe badania i analizy statystyczne.

4.3. Analiza zachowania modeli uczenia maszynowego dla różnych konfiguracji parametrów

Celem tego rozdziału jest zilustrowanie wpływu różnych kombinacji parametrów na efektywność i dokładność modeli uczenia maszynowego. W zakresie uczenia maszynowego, badane parametry mogą obejmować aspekty takie jak współczynniki uczenia, struktury modeli, wielkość partii danych, czy metody regularyzacji. Dobór tych parametrów odgrywa kluczową rolę w procesie uczenia, wpływa na zdolność modelu do generalizacji i determinuje jego efektywność w praktycznych zastosowaniach. Szczególna uwaga poświęcona zostaje analizie, w jaki sposób te zmienne wpływają na wyniki modeli, co jest niezbędne do zrozumienia ich optymalnego zastosowania i dostosowania do specyficznych potrzeb badawczych

Przedstawiona analiza ma na celu odpowiedźieć na pytania dotyczące optymalnych ustawień parametrów, które mogą poprawić skuteczność modeli uczenia

maszynowego w konkretnych scenariuszach. Poprzez systematyczną ocenę różnych konfiguracji parametrów, można uzyskać głębsze zrozumienie, jakie ustawienia mają kluczowe znaczenie dla osiągania lepszych wyników w danym zadaniu.

W kolejnych podrozialach zostana zaprezentowane wyniki anailiz dla modelu Lasso, ElasticNet oraz ExtraTreesRegressor. Dla każdego z modeli zostały przygotowane nastepujące przypadki:

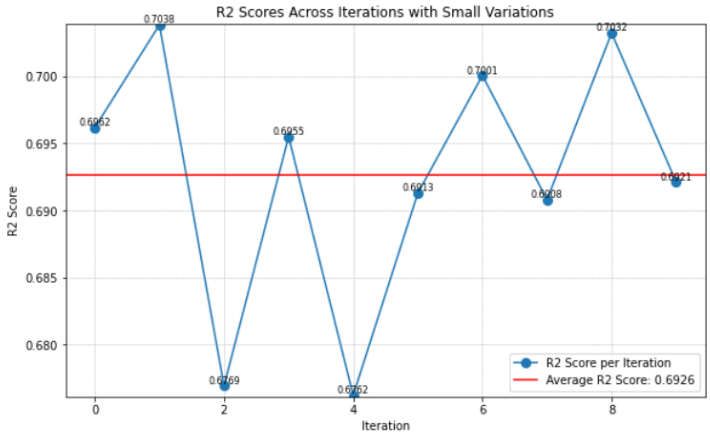
- Analiza dla modelu z orginalnym zbiorem danych i z użyciem wszystkich nowych parametrów.
- Analiza dla modelu z parametrami dotakowymimi rzeczywistej odległości i w lini prostej do obiektu .
- Analiza dla modelu z poszczególnymi parametrami.

4.4. Analiza modelu Lasso

Rozdiał skupia się na analizie wyników uzyskanych przy użyciu modelu Lasso, który jest istotnym narzędziem w dziedzinie uczenia maszynowego, wykorzystywanym do selekcji zmiennych oraz regularyzacji modeli. Model Lasso, stosujący regularyzację L1, pozwala na eliminację nieistotnych cech, co jest szczególnie przydatne w przypadku zbiorów danych o wysokiej wymiarowości. Celem tej analizy było zrozumienie, jak model Lasso radzi sobie w różnych konfiguracjach danych, począwszy od oryginalnego, nieprzetworzonego zbioru, aż po zbiory z wykorzystaniem wszystkich nowych parametrów.

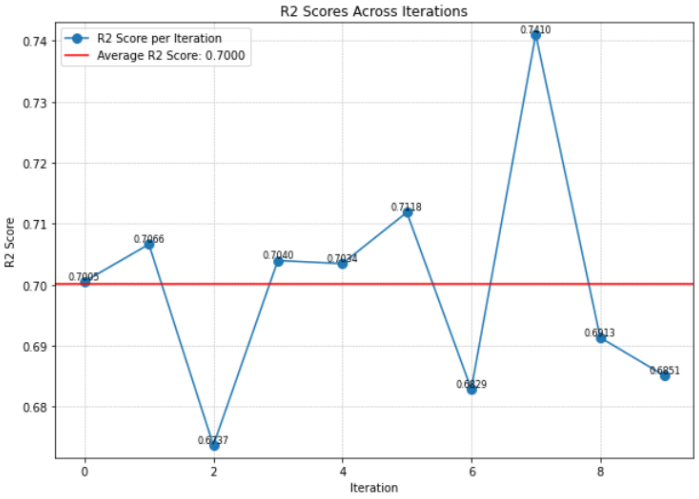
4.4.1. Analiza wyników modelu Lasso dla orginalnego zbioru danych i z użyciem wszystkich nowych parametrów

Oryginalny, nieprzetworzony zbiór danych stanowi realistyczne środowisko do testowania modelu Lasso, uwzględniając naturalne charakterystyki środowiskowe. W tej konfiguracji, model osiągnął średni wynik R^2 równy 0.6926. Wynik ten pozwala ocenić, jak model radzi sobie z typowymi wyzwaniami danych rzeczywistych, takimi jak zanieczyszczenia czy brakujące wartości, i służy jako podstawa do oceny jego wydajności. Rysunek 9 reprezetuje wyniki analizy modelu Lasso na oryginalnym zbiorze danych wykonany 10 razy dla różnego zestwu danych.



Rysunek 9: wyniki analizy modelu Lasso na oryginalnym zbiorze danych

Następnym krokiem analizy jest zastosowanie modelu *Lasso* na zbiorze danych rozszerzonym o nowe parametry. W tej konfiguracji średni wynik R^2 wzrósł nieznacznie do 0.7000. To zwiększenie, choć niewielkie, może wskazywać na korzystny wpływ nowych parametrów na zdolność modelu do przetwarzania i interpretacji danych. Rysunek 10 reprezentuje wyniki analizy modelu *Lasso* z nowymi argumentami wykonane 10 razy dla różnych zestawów danych.



Rysunek 10: Wyniki analizy modelu *Lasso* na zbiorze danych z nowymi argumentami

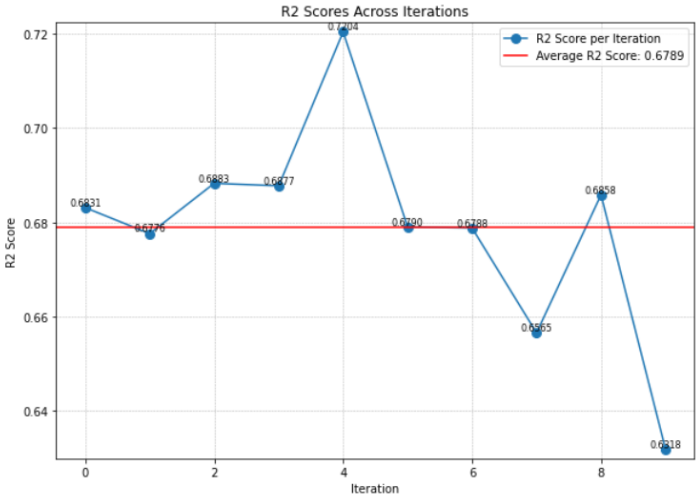
Porównanie wyników modelu Lasso na oryginalnym zbiorze danych z wynikami po wprowadzeniu nowych parametrów ukazuje subtelne różnice w skuteczności modelu. Niewielki wzrost wyniku R2 może sugerować, że model jest w stanie skorzystać z dodatkowych informacji zawartych w nowych parametrach, co pozytywnie wpływa na jego zdolność do przetwarzania danych. Wyniki te podkreślają, że nawet niewielkie modyfikacje w zestawie danych mogą mieć wpływ na działanie modelu, co jest ważne przy optymalizacji i dostosowywaniu modelu do specyficznych warunków.

4.4.2. Analiza wyników modelu Lasso dla odełegości rzeczywistych i w lini prostej

Na podstawie uzyskanych wyników zostały podjęte decyzje dotyczące potencjalnych dalszych eksperymentów. Identyfikacja korzystnych rodzajów danych i ich wpływ na jakość predykcji umożliwia skoncentrowanie się na najbardziej obiecujących obszarach badawczych. W dalszej części podrozdziału zostanie ocenione, który rodzaj danych może być bardziej perspektywiczny dla przyszłych badań oraz zidentyfikowane zostaną aspekty wymagające dalszej eksploracji.

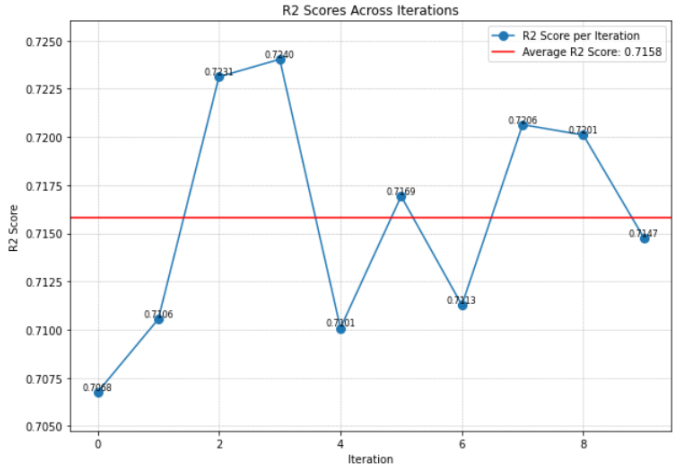
Parametry okreslajace drogę w lini prostej do obiektu

Średni wynik współczynnika determinacji R^2 dla modelu bez zaokrąglenia wynosi 0.6833, a liczba odstających wartości to 19. Porównując wyniki, można zaobserwować spadek jakości modelu względem wersji, gdzie w zbiór argumentów zawarte były wszystkie nowo pozyskane dane. Rysunek 11 przedstawia jakość modelu dla argumentów w linii prostej.



Rysunek 11: Jakość modelu dla argumentów w lini prostej bez zaokrąglenia

Średni wynik współczynnika determinacji R^2 dla modelu z zaokrąglonymi wartościami wynosi 0.7165, co sugeruje, że dane z zaokrąglonymi wartościami znacznie lepiej odpowiadają modelowi. Stabilność wyników współczynnika determinacji, oscylujących między 0.7068 a 0.7240, jest zauważalnie wyższa w przypadku modelu zaokrąglonego. Niemniej jednak, obserwuje się wzrost liczby odstających wartości na poziomie 406. Rysunek 12 przedstawia jakość modelu dla argumentów w linii prostej wraz z zaokrągleniami.



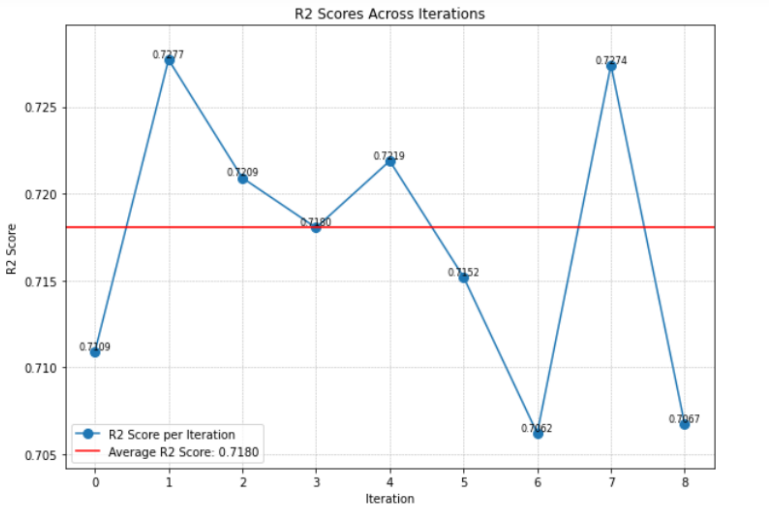
Rysunek 12: Jakość modelu dla argumentów w lini prostej z zaokrągleniem

Warto zaznaczyć, że model zaokrąglony wykazuje tendencję do uzyskiwania wyższych średnich wyników R^2, ale jednocześnie charakteryzuje się większą liczbą odsta-

jących wartości. To sugeruje, że procedura zaokrąglania danych wpływa na ogólną stabilność modelu, choć jednocześnie może wprowadzać pewne zakłócenia, co wymaga dalszej analizy w kontekście rzeczywistych odległości.

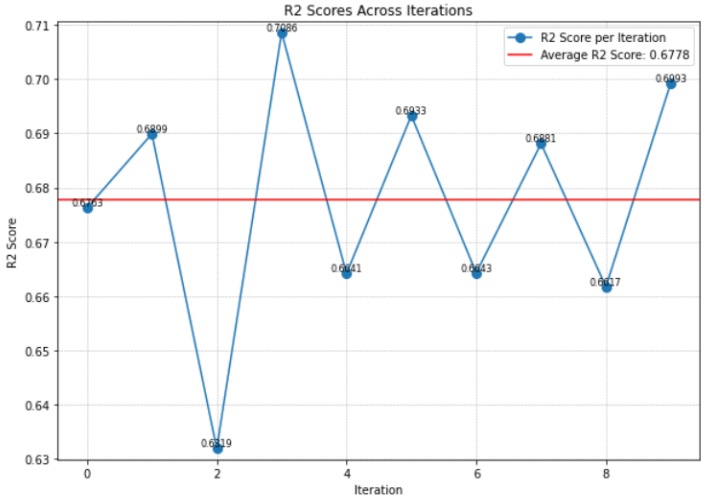
Parametry okreslajace rzeczywistą drogę do obiektu

Średni wynik współczynnika determinacji R^2 dla modelu z argumentami określajacymi rzeczywistą drogę do obiektu wynosi 0.6778. Wartości te oscylują między 0.6319 a 0.7086, z liczbą odstających wartości wynoszącą 19. Warto zauważyć, że spadek jakości modelu jest zauważalny w porównaniu do wersji, gdzie zbiór argumentów zawierał wszystkie nowo pozyskane dane. Rysunek 13 przedstawia jakość modelu dla argumentów z rzeczywistą drogą do obiektów.



Rysunek 13: Jakość modelu dla argumentów w lini prostej bez zaokrąglenia

Średni wynik współczynnika determinacji R^2 dla modelu z argumentami określającymi zaokrąglone wartosci wynosi 0.7180. Stabilność wyników (R2R2 oscylujące między 0.6508 a 0.7114) jest zauważalnie wyższa niż w przypadku modelu bez zaokrąglania. Jednakże, liczba odstających wartości wzrosła do 72. Rysunek 14 przedstawia jakość modelu dla argumentów z rzeczywistą drogą do obiektów wraz z zaokrągleniem.



Rysunek 14: Jakość modelu dla argumentów w lini prostej bez zaokrąglenia

Analizując oba przypadki, mozna zauważyć, że model z zaokrąglonymi danymi wykazuje lepszą ogólną stabilność (R2R2 średnio 0.7180), ale jednocześnie charakteryzuje się większą liczbą odstających wartości (72 przypadki). To sugeruje, że procedura zaokrąglania danych wpływa na poprawę stabilności modelu, ale może jednocześnie wprowadzać pewne zakłócenia, co wymaga dalszej analizy w kontekście rzeczywistych odległości.

Wnioski

Analizując modele oparte na parametrach określających drogę w linii prostej do obiektu, można zauważyć, że model z zaokrąglonymi danymi osiąga najwyższą średnią wartość współczynnika determinacji R^2 równą 0.7165. Dane z zaokrąglonymi wartościami lepiej odpowiadają temu modelowi, jednakże wiąże się to z większą liczbą odstających wartości, co potencjalnie może wprowadzać zakłócenia. Podobnie, dla parametrów określających rzeczywistą drogę do obiektu, model z zaokrąglonymi danymi uzyskuje wyższą średnią wartość współczynnika determinacji (0.7180) niż model bez zaokrąglania (0.6778). Niemniej jednak, model z zaokrąglonymi danymi charakteryzuje się również wyższą liczbą odstających wartości (72), co potencjalnie wprowadza pewne błędy do modelu. Te obserwacje są zgodne z danymi przedstawionymi w Tabeli 2, gdzie modele z zaokrąglonymi danymi mają wyższe średnie wartości R^2, ale jednocześnie większą liczbę odstających przypadków. W kolejnych etapach pracy będzie używana zaokrąglona wersja Modelu 4 (zaokrąglone

dane dla argumentów z rzewywitymi odległościami), który uzyskał najwyższą średnią wartość współczynnika determinacji R^2. Tabela 2 podsumowuje wyniki uzyskane dla każdego modelu.

Tabela 2: Tabela wyników dla odległości rzeczywistych i w linii prostej z zaokrąglonymi wartościami do 4 miejsc po kropce

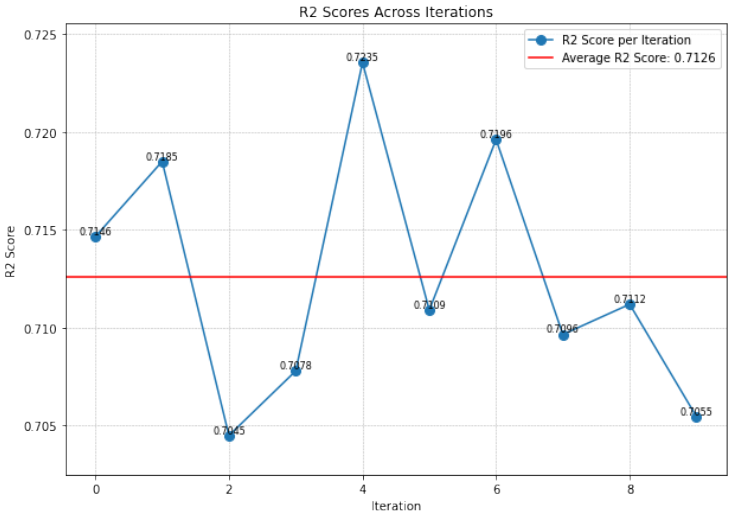
Model 1	**Model 2**	**Model 3**	**Model 4**
0.6789	0.7158	0.6778	0.7180
0.6831	0.7109	0.6763	0.7109
0.6776	0.7106	0.6899	0.7277
0.6883	0.7231	0.6319	0.7209
0.6877	0.7240	0.7086	0.7180
0.7204	0.7101	0.6641	0.7219
0.6790	0.7169	0.6933	0.7152
0.6788	0.7113	0.6643	0.7062
0.6565	0.7206	0.6881	0.7274
0.6858	0.7201	0.6617	0.7067
0.6318	0.7147	0.6993	0.7254

4.4.3. Analiza wyników modelu Lasso z wybranymi parametrami

W kontekście analizy danych odległości do obiektu w badanym obszarze, wybór odpowiednich wskaźników jest kluczowy dla skutecznej prognozy. Na podstawie wcześniejszych obserwacji stwierdzono, że uwzględnienie dokładnych informacji na temat odległości może istotnie poprawić jakość modelu prognostycznego. Zatem, aby zwiększyć precyzję prognoz, postanowiono skorzystać z danych wskazujących na rzeczywistą odległość do obiektu wraz z zaokrąglonymi wartościami.

Parametr określający odległość do najbliższego przystanku

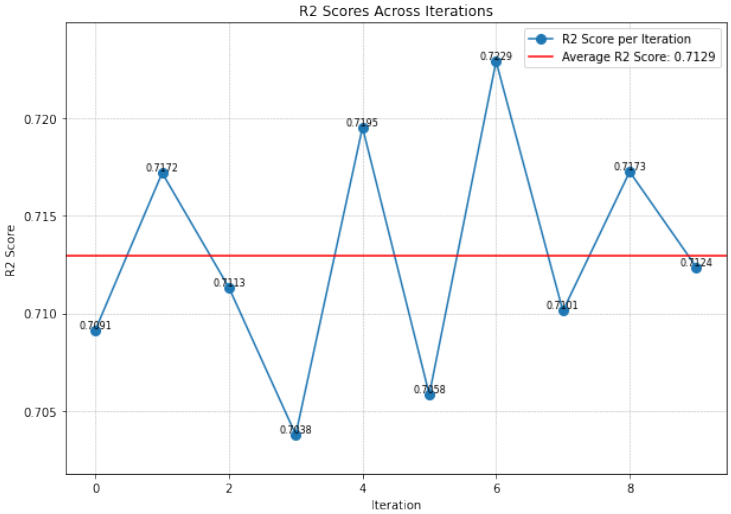
Średni wynik współczynnika determinacji (R2) z użyciem parametru informującego o najbliższym przystanku autobusowym wynosi 0.7125, a liczba odstających wartości to 406. Rysunek 15 przedstawia jakość modelu dla argumentów z parametrem określającym odległość do najbliższego przystanku.



Rysunek 15: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do przystanku

Parametr okrelający odległość do najblizszego sklepu

Średni wynik współczynnika determinacji (R2) z uzyciem prametru informujacym o najbliższym sklepie wynosi 0.7129, a liczba odstających wartości to 406. Rysunek 16 przedstawia jakość modelu dla argumentów z parametrem określającym odległość do najbliższego sklepu.

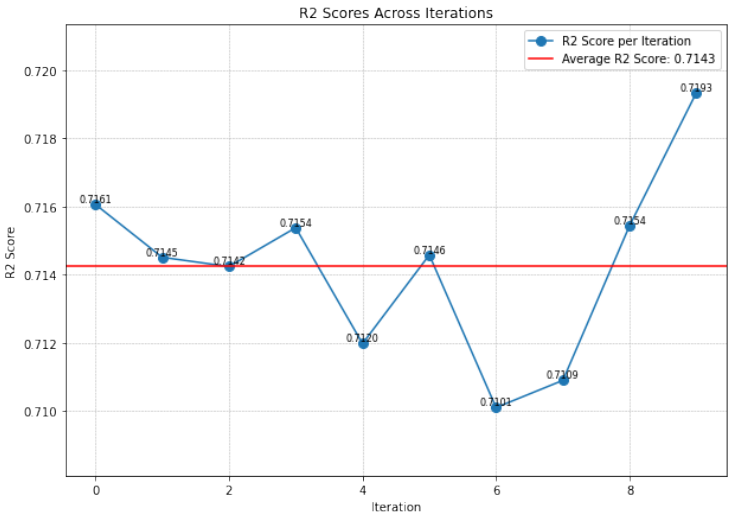


Rysunek 16: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do sklepu

Parametr okrelający odległość do najblizszej szkoły

Średni wynik współczynnika determinacji (R2) z uzyciem prametru informuja-

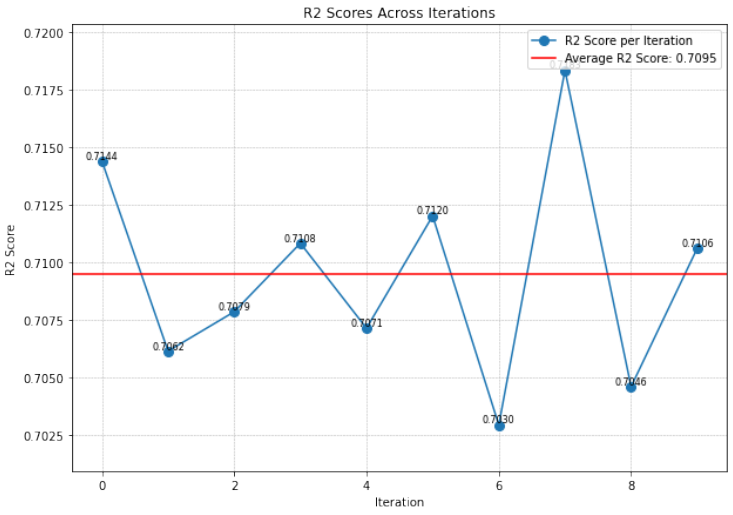
cym o najbliższej szkole 0.7142, a liczba odstających wartości to 406. Rysunek 17 przedstawia jakość modelu dla argumentów z parametrem określającym odległość do najbliższej szkoły.



Rysunek 17: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do szkoły

Parametr okrelający odległość do najblizszej kliniki

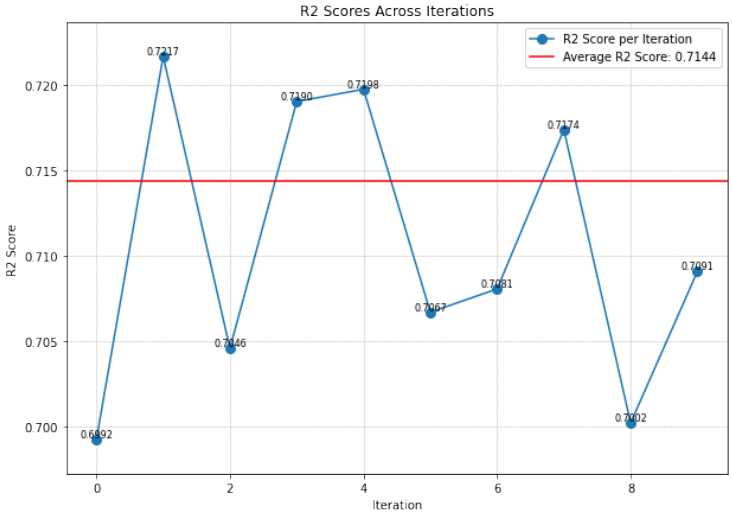
Średni wynik współczynnika determinacji (R2) z uzyciem prametru informujacym o najbliższej klinice wynosi 0.7094, a liczba odstających wartości to 406. Rysunek 18 przedstawia jakość modelu dla argumentów z parametrem określającym odległość do najbliższej kliniki.



Rysunek 18: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do przychodni

Parametr okrelający odległość do najblizszej restauracji

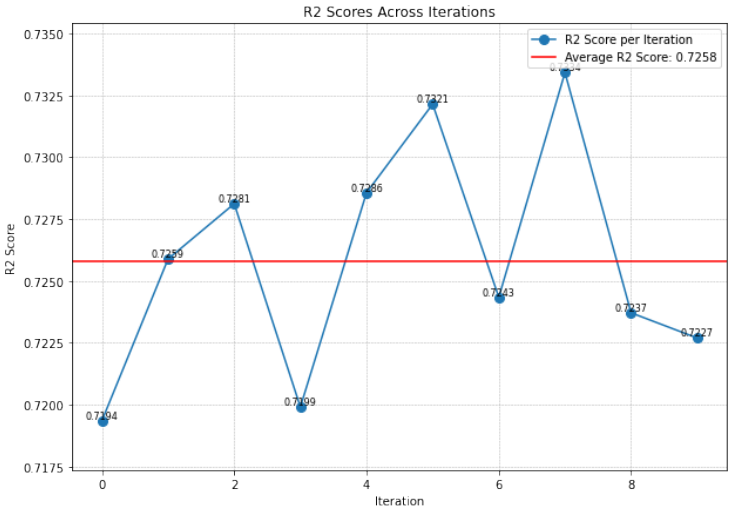
Średni wynik współczynnika determinacji (R2) z uzyciem prametru informujacym o najbliższej restauracji wynosi 7143, a liczba odstających wartości to 406. Rysunek 19 przedstawia jakość modelu dla argumentów z parametrem określającym odległość do najbliższej restauracji.



Rysunek 19: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do restauracji

Parametr okrelający liczbę atrakcji w okolicy

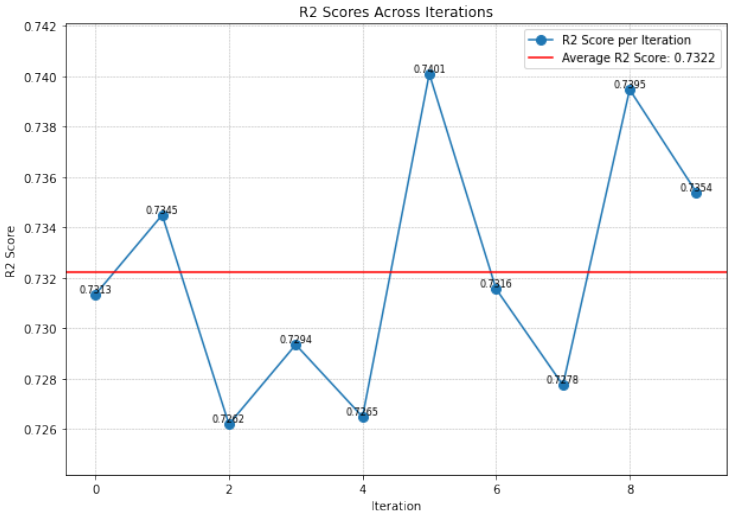
Średni wynik współczynnika determinacji (R2) z uzyciem prametru informujacym o liczbie atekacji w promieniu 1km wynosi 0.7258, a liczba odstających wartości to 406. Rysunek 20 przedstawia jakość modelu dla argumentów z parametrem określający liczbę atrakcji w okolicy.



Rysunek 20: Jakość modelu dla argumentów w lini prostej bez zaokrąglenia

Parametr okrelający liczbę atrakcji w okolicy z dodatkową kategoryzacją

Średni wynik współczynnika determinacji R^2 z uzyciem prametru informujacym o liczbie atekacji w promieniu 1km wraz z dodatkową kategoryzacją wynosi 0.7322, a liczba odstających wartości to 406. Rysunek 21 przedstawia jakość modelu dla argumentów z parametrem określający liczbę atrakcji w okolicy z dodatkową kolumną kategoyzującą. Rysunek 21 przedstawia jakość modelu dla argumentów z parametrem określający liczbę atrakcji w okolicy z zaokrągleniem.



Rysunek 21: Jakość modelu dla argumentów w lini prostej z zaokrąglenia

Wnioski

Analiza wyników modelu Lasso z wybranymi parametrami dostarcza istotnych informacji dotyczących skuteczności modelu w prognozowaniu cen nieruchomości. Badanie to skupiło się na ocenie, jak różne parametry, takie jak odległość do najbliższego przystanku, sklepu, szkoły, kliniki, restauracji oraz liczby atrakcji w okolicy i ich kategoryzacji, wpływają na jakość prognoz modelu.

Wyniki pokazują, że najwyższą średnią wartość współczynnika determinacji R^2 osiągnięto dla parametru określającego liczbę atrakcji w okolicy z dodatkową kategoryzacją, gdzie średnia wartość R^2 wyniosła 0.7322. Ten wynik sugeruje, że ten parametr miał największy wpływ na poprawę jakości modelu Lasso w kontekście prognozowania cen nieruchomości.

Parametr określający liczbę atrakcji w okolicy bez dodatkowej kategoryzacji również wykazał wysoką efektywność, z wynikiem R^2 na poziomie 0.7258. To wskazuje na to, że oba parametry związane z atrakcjami w okolicy są istotnymi wskaźnikami w modelu.

Z kolei najniższą średnią wartość współczynnika determinacji R^2 odnotowano dla parametru określającego odległość do najbliższej kliniki, gdzie średnia wartość R^2 wyniosła 0.7094. Mimo że jest to najniższy wynik wśród analizowanych parametrów, wciąż przyczynia się on do poprawy modelu w stosunku do wersji bazowej.

Podobnie jak w przypadku modelu ElasticNet, dla wszystkich rozpatrywanych parametrów w modelu Lasso odnotowano dużą liczbę odstających wartości (406), co może sugerować potrzebę dalszej optymalizacji modelu oraz dokładniejszej analizy danych wejściowych.

Podsumowując, przeprowadzona analiza wykazała, że różne parametry mają zróżnicowany wpływ na jakość modelu Lasso w kontekście prognozowania cen nieruchomości. W szczególności parametry związane z liczbą atrakcji w okolicy, zarówno z i bez dodatkowej kategoryzacji, wykazały największy potencjał w poprawie skuteczności modelu. Wyniki te mogą stanowić cenną wskazówkę dla dalszego rozwoju i optymalizacji modeli predykcyjnych w analizie rynku nieruchomości.

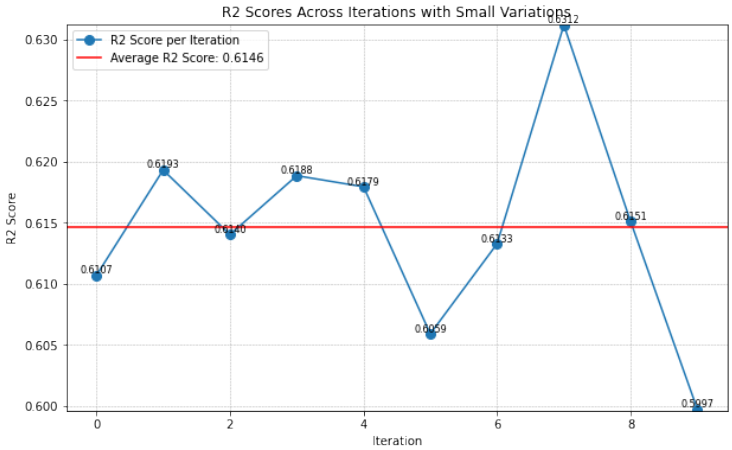
4.5. Model ElasticNet

W celu zilustrowania wpływu różnych konfiguracji danych na wyniki modelu ElasticNet, przeprowadzone zostały analizy porównawcze z wykorzystaniem różnych zestawów danych. Celem jest zrozumienie, w jaki sposób różne kombinacje regularyzacji L1 i L2 wpływają na wydajność modelu w kontekście konkretnej analizy danych. ElasticNet, dzięki swojej wszechstronności, jest potężnym narzędziem w arsenale każdego analityka danych i stanowi kluczowy element w nowoczesnym uczeniu maszynowym.

4.5.1. Analiza wyników modelu ElasticNet dla zbioru bazowego i z nowmymi parametrami

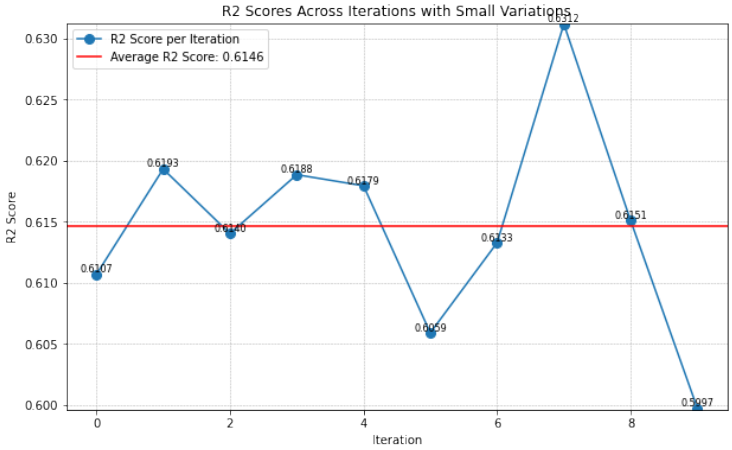
W rozdziale tym przeprowadzono analizę, mającą na celu sprawdzenie, czy dodanie nowych danych do zbioru danych wpływa na poprawę jakości modelu ElasticNet w porównaniu do zbioru podstawowego. Badanie to pozwoli na ocenę skuteczności modelu przy zastosowaniu różnych reprezentacji odległości i umożliwi wyciągnięcie wniosków o wpływie tych danych na dokładność prognoz cen nieruchomości.

Wyniki dla modelu ElasticNet bez parametrów wskazują na średnią wartość współczynnika determinacji (R2) wynoszącą 0.6146 na przestrzeni 10 iteracji. Ta wartość sugeruje, że model ma ograniczoną zdolność do wyjaśniania zmienności danych, co może świadczyć o potrzebie dodatkowych czynników w celu poprawy dokładności prognozowania cen nieruchomości. Na rys. 22 przedstawiono reprezentację wyników dla każdej z 10 iteracji, ilustrującą wydajność modelu w różnych warunkach testowych.



Rysunek 22: Jakość modelu ElasticNet dla bazowego zestawu danych

W kolejnym etapie analizy, wyniki dla modelu ElasticNet z wykorzystaniem wszystkich nowych parametrów wskazują na wyższą średnią wartość R2, osiągając 0.6640. Ten wynik stanowi znaczące polepszenie w porównaniu do modelu bez dodatkowych parametrów, co wskazuje na efektywne wykorzystanie nowych czynników w prognozowaniu cen nieruchomości. Liczba odstających wartości dla tego modelu wynosi 19, co może sugerować potrzebę dalszej analizy i optymalizacji. W celu zilustrowania tych wyników, dołączona na rys. Na rys. 23 widnieje reprezentacja wyników dla każdej z 10 iteracji.



Rysunek 23: Jakość modelu ElasticNet dla zbirouz z nowymi parametrami

Wnioski

Z przeprowadzonej analizy wynika, że włączenie nowych parametrów do modelu

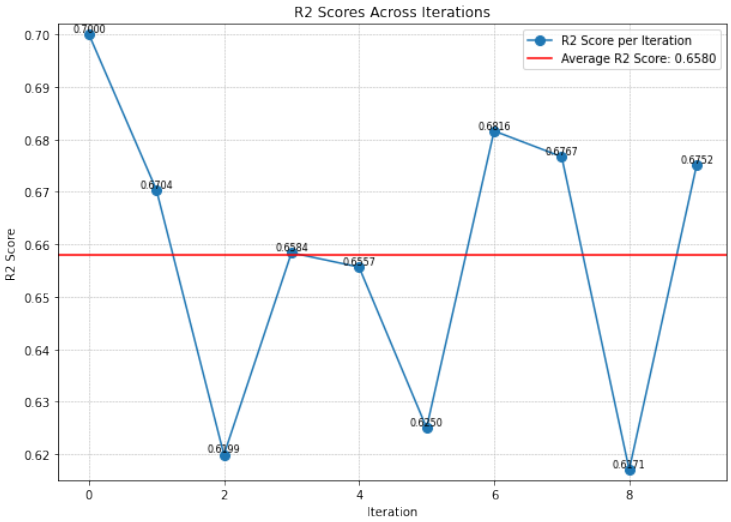
ElasticNet znacząco wpłynęło na poprawę jego skuteczności. Porównując wyniki dla zestawu bazowego i zestawu z nowymi parametrami, zaobserwowano wzrost średniej wartości współczynnika determinacji R^2 z 0.6146 do 0.6640. Ten wzrost wskazuje na lepszą zdolność modelu do prognozowania cen nieruchomości po zastosowaniu dodatkowych danych. Wyniki te podkreślają znaczenie właściwego doboru parametrów dla optymalizacji modeli predykcyjnych w analizie rynku nieruchomości.

4.5.2. Analiza wyników modelu ElasticNet dla odełegości rzeczywistych i w lini prostej

W tej części pracy pokazana została analiza wyników modelu ElasticNet, przy uwzględnieniu dwóch rodzajów danych odległościowych: rzeczywistych i w linii prostej. Ocena ta pozwala na zrozumienie, które podejście do reprezentacji odległości lepiej przyczynia się do skuteczności modelu w kontekście prognozowania cen nieruchomości. Przeanalizowanie tych wyników umożliwia podjęcie decyzji o kierunkach dalszych eksperymentów i wskazuje, które metody odległościowe mają większy potencjał w kontekście przyszłych badań. Ponadto, zidentyfikowane zostały obszary, które mogą wymagać głębszej analizy i eksploracji

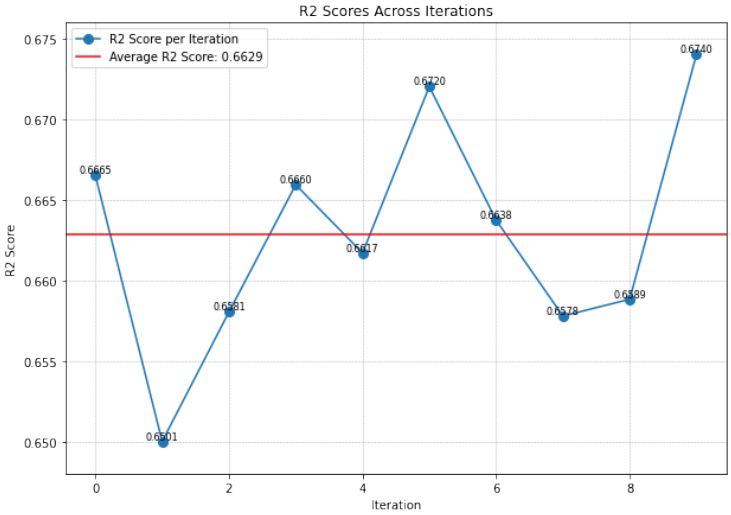
Parametry okreslajace drogę w lini prostej do obiektu

Średni wynik współczynnika determinacji R^2 dla modelu bez zaokrąglenia wynosi 0.6579, a liczba odstających wartości to 19. Porównując wyniki, można zaobserwować spadek jakości modelu względem wersji, gdzie w zbiorze argumentów zawarte były wszystkie nowo pozyskane dane. Rysunek 24 przedstawia jakość modelu dla argumentów z parametrem o okreslajacym drogę w lini prostej do obiektu bez zaokrąglenia.



Rysunek 24: Jakość modelu dla argumentów w lini prostej bez zaokrąglenia

Średni wynik współczynnika determinacji R^2 dla modelu bez zaokrąglenia wynosi 0.6628, a liczba odstających wartości to 406. Porównując wyniki, można zaobserwować spadek jakości modelu względem wersji, gdzie w zbiór argumentów zawarte były wszystkie nowo pozyskane dane. Rysunek 25 przedstawia jakość modelu dla argumentów z parametrem o okreslajacym drogę w lini prostej do obiektu z zaokrągleniem.



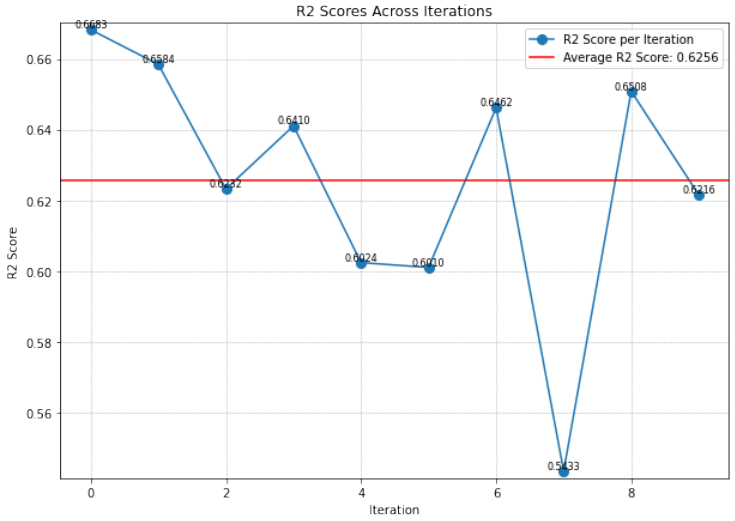
Rysunek 25: Jakość modelu dla argumentów w lini prostej z zaokrągleniem

Warto zuważyć że porawa modelu wzgledem mdelu Lasso jest mniejsza jednak oba przypadki są zbliżone. W obu modelach nastąpiła poprawa jakosci modelu oraz

zwrost wartosci odstajcych w przypadku zaokrąglania wartości.

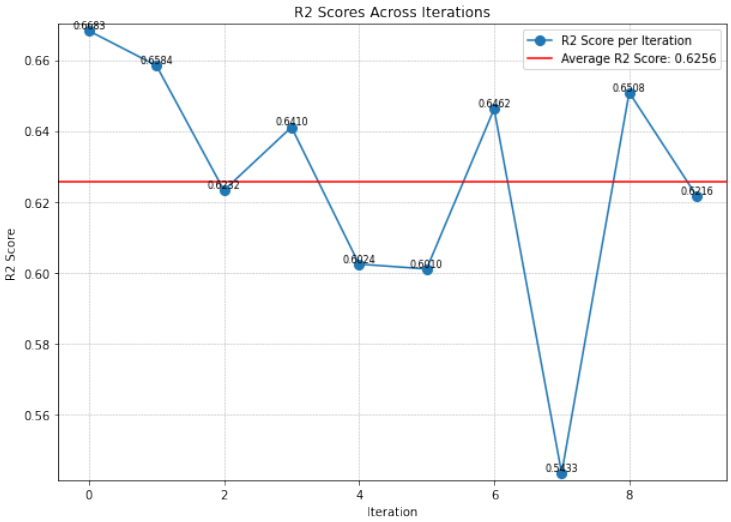
Parametry okreslajace rzeczywistą drogę do obiektu

Średni wynik współczynnika determinacji R^2 dla modelu z argumentami określajacymi rzeczywistą drogę do obiektu wynosi 0.62561. Wartości te oscylują między 0.5433 a 0.6682, z liczbą odstających wartości wynoszącą 19. Warto zauważyć, że model radzi sobie znacznie gorzej wzgledem wcześnieszych ekspertów. Rysunek 26 przedstawia jakość modelu dla argumentów z parametrem okreslajacy drogę rzeczywistą do obiektu bez zaokrąglenia.



Rysunek 26: Jakość modelu dla argumentów okreslajacy rzeczywistą drogę bez zaokrąglenia

Średni wynik współczynnika determinacjiR^2) dla modelu z argumentami określającymi zaokrąglone wartosci wynosi 0.6691. Stabilność wyników R^2 oscylujące między 0.6653 a 0.6672) jest zauważalnie wyższa niż w przypadku modelu bez zaokrąglania. Liczba odstających wartości wynosi 406. Rysunek 27 przedstawia jakość modelu dla argumentów z parametrem o okreslajacym rzeczywistą drogę do obiektu z zaokrągleniem.



Rysunek 27: Jakość modelu dla argumentów w lini prostej z zaokrągleniem

Porównując wyniki z Tabelą 2, można zauważyć, że model ElasticNet radzi sobie znacznie słabiej z nowymi danymi w stosunku do modelu Lasso. Choć obserwuje się pewną poprawę w wynikach modelu ElasticNet względem wersji bez dodatkowych argumentów, to jednak ta poprawa jest mniej widoczna. Najwieksze straty wzgledem modelu Lasso mozemy zaobserwować w modelach które bazowały na danych z zaokrąglonymi odległosciami do obiektów. W Tabeli 3, przedstawiono szczegółowe wyniki dla różnych modeli. Jak widać, Model 4 wykazuje najwyższą wartość współczynnika determinacji R^2 w większości przypadków.

Tabela 3: Tabela wyników dla odległości rzeczywistych i w linii prostej z zaokrąglonymi wartościami do 4 miejsc po kropce

Model 1	**Model 2**	**Model 3**	**Model 4**
0.6579	0.6628	0.6256	0.6691
0.7000	0.6665	0.6682	0.6653
0.6703	0.6500	0.6584	0.6748
0.6199	0.6580	0.6231	0.6675
0.6583	0.6659	0.6410	0.6705
0.6556	0.6617	0.6023	0.6629
0.6250	0.6720	0.6009	0.6748
0.6815	0.6637	0.6461	0.6672
0.6766	0.6578	0.5433	0.6732
0.6170	0.6588	0.6508	0.6626
0.6751	0.6740	0.6215	0.6725

Wnioski

Analiza wyników modelu ElasticNet wskazuje, że Model 4 (z parametrami określającymi zaokrągloną rzeczywistą drogę do obiektu) osiąga najwyższą średnią wartość R2 (0.6691), co sugeruje, że jest to najskuteczniejsze podejście w kontekście prognozowania cen nieruchomości. Modele z parametrami określającymi drogę w linii prostej (Model 1 i Model 2) również wykazują poprawę w stosunku do modelu bazowego, ale wyniki są mniej efektywne niż w przypadku Modelu 4. Model 3, z parametrami rzeczywistej drogi bez zaokrąglenia, prezentuje najniższe wartości R2, co wskazuje na jego ograniczoną skuteczność w porównaniu z innymi modelami.

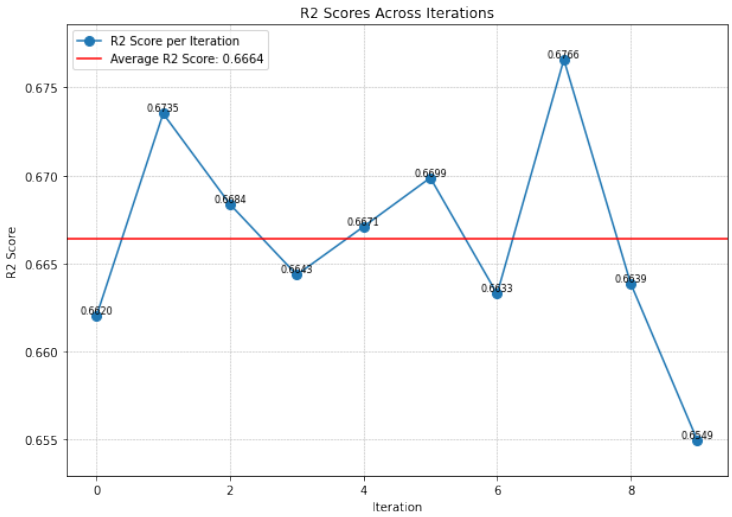
4.5.3. Analiza wyników modelu ElasticNet z wybranymi parametrami

W poprzednich badaniach zauważono pewne ograniczenia efektywności modelu ElasticNet w kontekście prognozowania nowo pozyskanych danych. Istotne jest teraz porównanie wyników tego modelu z rezultatami uzyskanymi za pomocą metody Lasso, która stanowi jedną z podstawowych technik regresji liniowej.

Analiza tych dwóch modeli pozwoli na lepsze zrozumienie, który z nich radzi sobie lepiej w przypadku nowych danych. Wartości parametrów oraz wyniki prognozowania zostaną szczegółowo przeanalizowane, a ewentualne różnice w skuteczności modeli zostaną uwzględnione. Celem tego porównania jest identyfikacja, czy jedna z metod regresji liniowej, tj. Lasso lub ElasticNet, wykazuje przewagę w kontekście konkretnego zbioru danych, zwłaszcza mając na uwadze wcześniejsze obserwacje wskazujące na pewne ograniczenia modelu ElasticNet.

Parametr określający odległość do najbliższego przystanku

Średni wynik współczynnika determinacji R^2 z użyciem parametru informującego o najbliższym przystanku autobusowym wynosi 0.6663, a liczba odstających wartości to 406. Rysunek 28 przedstawia jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do przystanku.

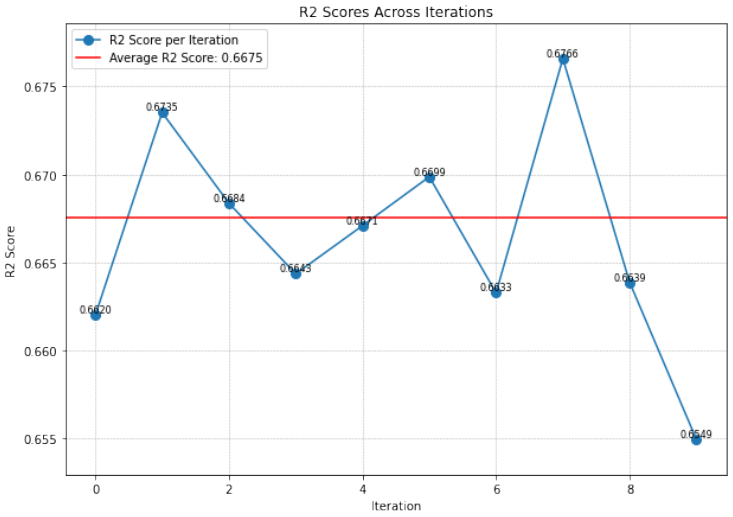


Rysunek 28: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do przystanku

Parametr określający odległość do najbliższego sklepu

Średni wynik współczynnika determinacji R^2 z użyciem parametru informującego

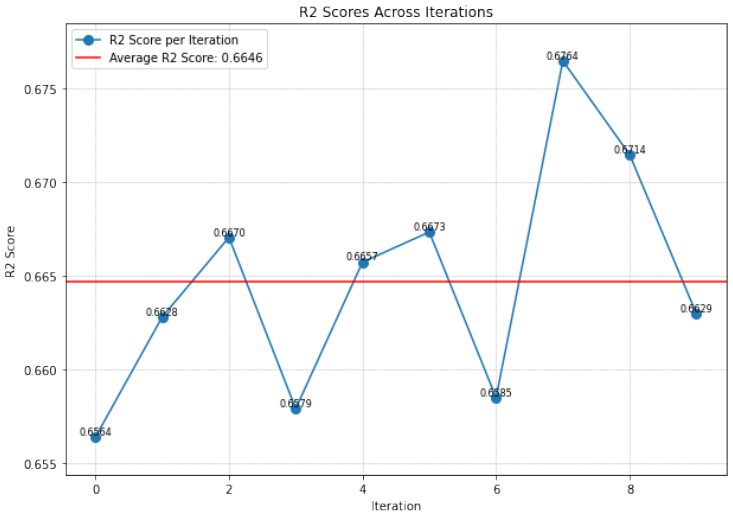
o najbliższym szklepie wynosi 0.6675, a liczba odstających wartości to 406. Rysunek 29 przedstawia jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do sklepu.



Rysunek 29: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do sklepu

Parametr określający odległość do najbliższej szkoły

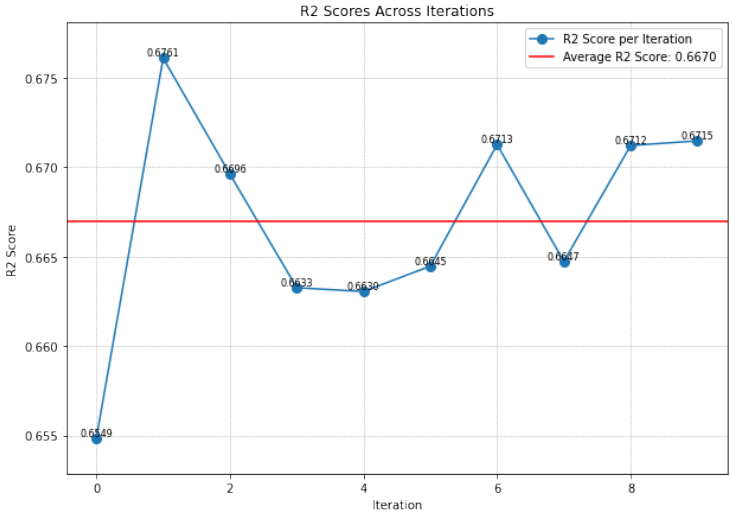
Średni wynik współczynnika determinacji R^2 z użyciem parametru informującego o najbliższej szkole wynosi 0.6646, a liczba odstających wartości to 406. Rysunek 30 przedstawia jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do skoły.



Rysunek 30: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do skoły

Parametr określający odległość do najbliższej kliniki

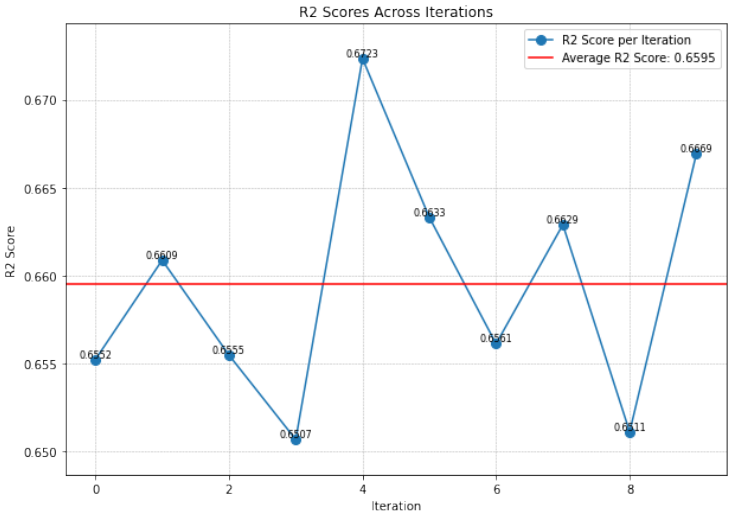
Średni wynik współczynnika determinacji R^2 z użyciem parametru informującego o najbliższej klinice wynosi 0.6670, a liczba odstających wartości to 406. Rysunek 31 przedstawia jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do kliniki.



Rysunek 31: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do kliniki

Parametr określający odległość do najbliższej restauracji

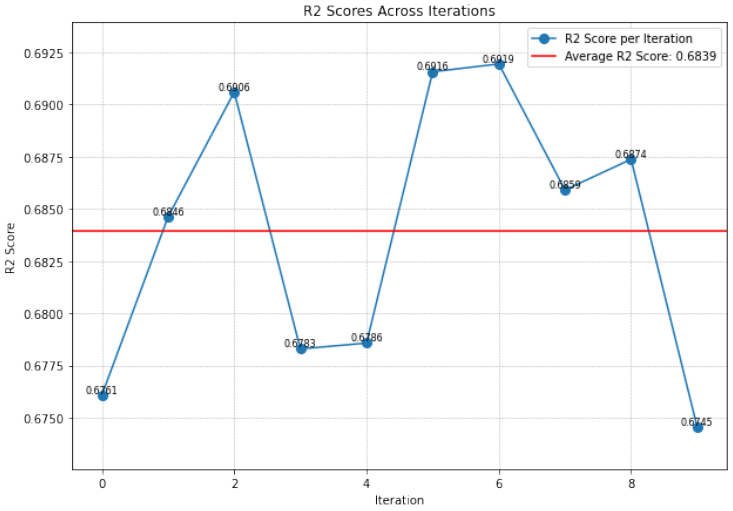
Średni wynik współczynnika determinacji R^2 z użyciem parametru informującego o najbliższej restauracji wynosi 0.6594, a liczba odstających wartości to 406. Rysunek 32 przedstawia jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do restauracji.



Rysunek 32: Jakość modelu dla argumentu orkeslajćego rzeczywistą odległoś do restauracji

Parametr okrelający liczbę atrakcji w okolicy

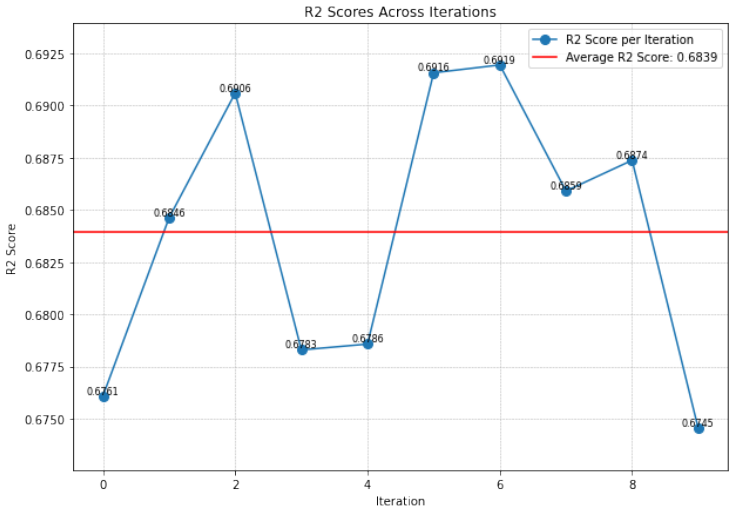
Średni wynik współczynnika determinacji R^2 z użyciem parametru informującego o licizbie atrakcji w promieńu 1km wynosi 0.6839, a liczba odstających wartości to 406. Rysunek 33 przedstawia jakość modelu z argumentem orkeslajćego liczbę atrakcji w okolicy.



Rysunek 33: Jakość modelu dla argumentu orkeslajćego liczbę atrakcji w okolicy

Parametr okrelający liczbę atrakcji w okolicy z dodatkową kategoryzacją

Średni wynik współczynnika determinacji R^2 z użyciem parametru informującego o licizbie atrakcji wraz z dodatkową kolumną kategoryzującą dane w promieńu 1km wynosi 0.6921, a liczba odstających wartości to 406. Rysunek 34 przedstawia jakość modelu z argumentem orkeslajćego liczbę atrakcji w okolicy.



Rysunek 34: Jakość modelu dla argumentu orkeslajćego liczbę atrakcji w okolicy z dodatkową kategoryzacją

Wnioski

Analiza wyników modelu ElasticNet z wybranymi parametrami ujawnia cenne informacje odnośnie do efektywności tego modelu w prognozowaniu cen nieruchomości. W badaniu skupiono się na ocenie wpływu różnych parametrów, takich jak odległość do najbliższego przystanku, sklepu, szkoły, kliniki, restauracji oraz liczby atrakcji w okolicy i ich kategoryzacji, na jakość prognoz modelu.

Wyniki wskazują, że najwyższą średnią wartość współczynnika determinacji R^2 , a tym samym najlepszą skuteczność w modelu, osiągnięto dla parametru określającego liczbę atrakcji w okolicy z dodatkową kategoryzacją, gdzie średnia wartość R^2 wyniosła 0.6921. Ten wynik sugeruje, że uwzględnienie zarówno liczby atrakcji, jak i ich kategoryzacji, znacząco przyczynia się do poprawy dokładności prognozowania cen nieruchomości. Z kolei najniższą średnią wartość współczynnika determinacji R^2 odnotowano dla parametru określającego odległość do najbliższej restauracji, gdzie

średnia wartość R^2 wyniosła 0.6594. Choć ten wynik wciąż przyczynia się do poprawy modelu w stosunku do wersji bazowej, jest on niższy w porównaniu do innych rozpatrywanych parametrów.

Analiza wykazała również, że wszystkie rozpatrywane parametry charakteryzowały się dużą liczbą odstających wartości, co może wskazywać na konieczność dalszej optymalizacji modelu oraz dokładniejszej analizy danych wejściowych. Dalsze badania nad wykorzystaniem dodatkowych parametrów, zwłaszcza tych związanych z atrakcjami w okolicy i ich kategoryzacją, mogą przyczynić się do zwiększenia dokładności prognozowania cen nieruchomości. Podsumowując, przeprowadzona analiza dostarcza istotnych wskazówek dotyczących skuteczności różnych parametrów w modelu ElasticNet. W szczególności wyniki wskazują na znaczącą wartość dodania parametrów związanych z atrakcjami w okolicy i ich kategoryzacją, co może stanowić ważny kierunek dla dalszych badań i rozwoju modeli predykcyjnych w analizie rynku nieruchomości.

4.6. Analiza wyników

W pracy przeprowadzono szczegółową analizę dwóch modeli regresji: Lasso i ElasticNet, przy użyciu różnych zestawów danych. Analiza ta miała na celu zrozumienie, jak różne konfiguracje danych wpływają na wydajność obu modeli. Poniżej przedstawiono wnioski z tej analizy, uwzględniając różne aspekty porównania.

Analiza porównawcza modeli Lasso i ElasticNet, przeprowadzona na różnych konfiguracjach danych, dostarcza cennych wskazówek dotyczących ich efektywności i odpowiednich zastosowań. Ogólnie rzecz biorąc, model Lasso wykazuje lepszą skuteczność w większości przypadków, co sugeruje, że jest on bardziej odpowiedni do zadań, w których kluczowe jest wybieranie najbardziej istotnych cech oraz efektywna regularyzacja. ElasticNet, chociaż ogólnie mniej skuteczny niż Lasso, może nadal być wartościowym narzędziem w niektórych scenariuszach, zwłaszcza gdy równowaga pomiędzy regularyzacjami L1 i L2 jest szczególnie ważna.

Tabela 4 przedstawia porównanie wyników modeli Lasso i ElasticNet w różnych konfiguracjach danych, z uwzględnieniem średnich wartości współczynnika determinacji R^2 oraz liczby odstających wartości.

Tabela 4: Porównanie wyników modeli Lasso i ElasticNet dla różnych konfiguracji danych

Konfiguracja danych	**Model Lasso**	**Model ElasticNet**
Oryginalny zbiór danych	0.6926	0.6146
Zbiór z nowymi parametrami	0.7000	0.6640
Odległość w linii prostej bez zaokrąglenia	0.6833	0.6579
Odległość w linii prostej z zaokrągleniem	0.7165	0.6628
Rzeczywista droga do obiektu	0.6778	0.6256
Rzeczywista droga z zaokrągleniem	0.7180	0.6691
Odległość do przystanku	0.7125	0.6663
Odległość do sklepu	0.7129	0.6675
Odległość do szkoły	0.7142	0.6646
Odległość do kliniki	0.7094	0.6670
Odległość do restauracji	0.7143	0.6594
Liczba atrakcji w okolicy	0.7258	0.6839
Liczba atrakcji z kategoryzacją	0.7322	0.6921

Analizując wyniki dla oryginalnego zbioru danych i zestawu z nowymi parametrami, widoczna jest przewaga modelu Lasso nad ElasticNet. W obu przypadkach, Lasso osiąga lepsze wyniki (R^2: 0.6926 i 0.7000) niż ElasticNet (R^2: 0.6146 i 0.6640). Ta konsystencja w wyższej skuteczności Lasso może wskazywać na jego lepszą zdolność do radzenia sobie z zanieczyszczeniami i brakującymi wartościami w danych. Wnioski te sugerują, że model Lasso jest bardziej elastyczny i lepiej dostosowuje się do różnorodnych zestawów danych, co jest kluczowe w praktycznych zastosowaniach analizy danych. Zauważalny wzrost wyniku R^2 przy przejściu od oryginalnego zbioru do zbioru z nowymi parametrami w modelu Lasso wskazuje na jego zdolność do efektywnego wykorzystania dodatkowych informacji, co jest ważne w kontekście rozwijających się zastosowań uczenia maszynowego.

Porównując wyniki dla odległości rzeczywistych i w linii prostej, model Lasso ponownie wykazuje wyższą skuteczność w stosunku do modelu ElasticNet. Dla danych bez zaokrąglenia Lasso osiąga $R^2 = 0.6833$, a z zaokrągleniem nawet $R^2 = 0.7165$, podczas gdy ElasticNet osiąga odpowiednio $R^2 = 0.6579$ i 0.6628. Wyższe wyniki dla modelu Lasso mogą wynikać z lepszego dostosowania do specyfiki danych odległościowych. Zauważalne jest, że proces zaokrąglania danych znacząco wpływa na wyniki obu modeli, co może świadczyć o wrażliwości modeli na drobne zmiany w danych. W przypadku modelu Lasso, zaokrąglenie danych prowadzi do zwiększenia średniej wartości R^2, co może sugerować, że model lepiej radzi sobie z mniejszą szcze-

gółowością danych. W przypadku modelu ElasticNet, mimo że zaokrąglenie również prowadzi do wzrostu R^2, to jednak w mniejszym stopniu niż w modelu Lasso.

Analizując wyniki modeli Lasso i ElasticNet z wybranymi parametrami, takimi jak odległość do różnych obiektów i liczba atrakcji w okolicy, można zaobserwować, że model Lasso konsekwentnie osiąga wyższe wyniki niż ElasticNet. Szczególnie warto zwrócić uwagę na parametry związane z liczbą atrakcji w okolicy, zarówno bez jak i z dodatkową kategoryzacją, gdzie Lasso osiąga R^2: 0.7258 i 0.7322, w porównaniu do R^2 równego 0.6839 i 0.6921 dla ElasticNet. Wyniki te pokazują, że Lasso jest bardziej skuteczne w wykorzystywaniu szczegółowych informacji o lokalizacji, co może być kluczowe w prognozowaniu cen nieruchomości. ElasticNet, mimo że również poprawia swoje wyniki przy dodaniu nowych parametrów, to jednak w mniejszym stopniu niż Lasso. Może to wskazywać, że ElasticNet jest mniej skuteczny w sytuacjach wymagających szczegółowej analizy danych przestrzennych. Ogólnie, wnioski te podkreślają wyższą skuteczność modelu Lasso w analizie danych złożonych i szczegółowych, co jest istotne w kontekście rosnącej złożoności problemów analizy danych.

Analiza wyników dla różnych konfiguracji danych ujawnia, że model Lasso generalnie przewyższa model ElasticNet pod względem skuteczności, wyrażonej w wyższych wartościach współczynnika determinacji R^2. Zarówno w przypadku oryginalnego zbioru danych, jak i zbioru z nowymi parametrami, Lasso konsekwentnie osiągał lepsze wyniki niż ElasticNet. To wskazuje na większą zdolność modelu Lasso do radzenia sobie z zanieczyszczeniami i brakującymi wartościami, co jest kluczowe w realistycznych scenariuszach analizy danych.

Szczególnie istotne jest, że model Lasso wykazał się lepszą skutecznością w analizie danych przestrzennych, zarówno w przypadku odległości rzeczywistych, jak i w linii prostej, co sugeruje jego przewagę w sytuacjach wymagających dokładnej analizy lokalizacji i odległości. Proces zaokrąglania danych miał znaczący wpływ na wyniki obu modeli, ale Lasso wydaje się lepiej radzić sobie z takimi zmianami.

Co więcej, w analizie wyników z wybranymi parametrami, takimi jak odległość do różnych obiektów i liczba atrakcji w okolicy, model Lasso ponownie wykazał się większą skutecznością. Szczególnie efektywne okazało się uwzględnienie parametrów związanych z liczbą atrakcji w okolicy, zarówno z i bez dodatkowej kategoryzacji. To wskazuje na zdolność Lasso do efektywnego wykorzystania szczegółowych informacji o otoczeniu, co jest istotne w kontekście prognozowania cen nieruchomości.

Biorąc pod uwagę potencjalne zmiany w lokalizacjach obiektów od roku 2015, a także fakt, że dane zostały zebrane z OpenStreetMap, istotne jest podkreślenie znaczenia parametru liczby atrakcji w okolicy. Ten parametr wydaje się być lepszym wskaźnikiem zagęszczenia obiektów atrakcyjnych w danym obszarze niż odległość do najbliższej atrakcji. W kontekście rynku nieruchomości, liczba atrakcji w okolicy może odzwierciedlać atrakcyjność lokalizacji, co przekłada się na wyższe ceny nieruchomości. Koncentracja atrakcji w danym obszarze może być lepszym wskaźnikiem wartości nieruchomości niż odległość do pojedynczych obiektów, odzwierciedlając ogólną atrakcyjność i dynamikę danego miejsca.

Dodatkowo, parametr liczby atrakcji w okolicy może być bardziej stabilnym wskaźnikiem niż odległość do poszczególnych obiektów, które mogą ulec zmianie w czasie. W analizie danych nieruchomości, uwzględnienie kompleksowych wskaźników, takich jak liczba atrakcji, może lepiej odzwierciedlać rzeczywistą wartość nieruchomości i jej potencjał inwestycyjny.

Podsumowując, wyniki analizy modeli Lasso i ElasticNet wskazują na wyższą skuteczność Lasso w różnych scenariuszach analizy danych, szczególnie w kontekście danych przestrzennych i skomplikowanych zestawów danych. ElasticNet, mimo że także jest użyteczny, wydaje się mniej skuteczny w porównaniu do Lasso, zwłaszcza w przypadkach wymagających szczegółowej analizy przestrzennej i lokalizacyjnej. Szczególną uwagę należy zwrócić na parametry związane z liczbą atrakcji w okolicy, gdyż mogą one znacząco wpływać na jakość prognozowania cen nieruchomości, zwłaszcza w kontekście oceny atrakcyjności lokalizacji i dynamiki zmian w przestrzeni miejskiej.

Zakończenie

W pracy dokonano analizy różnych modeli uczenia maszynowego z zastosowaniem do prognozowania cen nieruchomości. Szczególną uwagę zwrócono na modele Lasso i ElasticNet, jednakże we wstępnych badaniach dodatkowo przetestowano również dwa inne modele: GradientBoostingRegressor i ExtraTreesRegressor. Wyniki uzyskane na podstawowych danych były bardzo obiecujące, gdzie model ExtraTreesRegressor uzyskał wynik nieco lepszy niż GradientBoostingRegressor , jednak oba modele były znacząco lepsze niż modele Lasso i ElasticNet. Niestety, po wzbogaceniu zbioru danych o nowe parametry, jakość obu modeli uległa obniżeniu. Obserwacje te wskazują na istotny potencjał tych modeli, jednak ze względu na sposób doboru algorytmów do tego badania nie badano ich głebiej w niniejszej pracy. Modele te mogą zostać dokładniej zbadany w przyszłych badaniach.

Modele bazowe, takie jak GradientBoostingRegressor i ExtraTreesRegressor, wykazały się dużym potencjałem, co sugeruje możliwość osiągnięcia lepszych wyników przy dalszej optymalizacji i dostosowaniu ich do bardziej skomplikowanych zestawów danych. W przyszłych badaniach warto pochylić się nad tymi modelami, aby zbadać ich pełne możliwości i efektywność w różnych scenariuszach.

W dalszej perspektywie, istotne będzie również rozważenie zastosowania innych modeli uczenia maszynowego, aby ocenić ich skuteczność w kontekście prognozowania cen nieruchomości. Ponadto, pozyskanie aktualniejszych danych rynkowych może znacząco przyczynić się do zwiększenia dokładności i wiarygodności prognoz. Warto również rozważyć zastosowanie zaawansowanych technik analizy danych, takich jak uczenie głębokie, które mogą odkryć nowe, bardziej złożone wzorce w danych.

Podsumowując, niniejsza praca wskazuje na znaczący potencjał uczenia maszynowego w analizie rynku nieruchomości, ale jednocześnie podkreśla potrzebę dalszych badań, eksperymentowania i optymalizacji metod oraz modeli, aby w pełni wykorzystać możliwości oferowane przez te narzędzia w prognozowaniu cen nieruchomości.

Literatura

[1] Burinskienė, M., Rudzkienė, V., & Venckauskaitė, J. (2019). Models of factors influencing the real estate price. ResearchGate https://www.researchgate.net/publication/330482838_Models_of_factors_influencing_the_real_estate_price.

[2] Hanna Kołodziejczyk "Geneza kryzysu hipotecznego w USA z perspektywy dekady". https://pressto.amu.edu.pl/index.php/rpeis/article/view/6852/6850.

[3] Iglewicz, B., & Hoaglin, D. C. (1993). "How to Detect and Handle Outliers". ASQC Quality Press. 9-16.

[4] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, 22(3), 276-282.

[5] McKinney, W. (2010). Data Structures for Statistical Computing in Python. W: Proceedings of the 9th Python in Science Conference.

[6] Müller, A.C., & Guido, S. (2016). Introduction to Machine Learning with Python. O'Reilly Media, Inc. ISBN: 9781449369897.

[7] Richter, R. Rynek nieruchomości w USA: Cz. 2 - Wskaźnik wyprzedzający dla gospodarki. https://analizy.investio.pl/rynek-\nieruchomosci-w-usa-cz-2-wskaznik-wyprzedzajacy-dla-gospodarki/

[8] Rob Story. „Geopy: Python Geocoding Toolbox."2022. https://pypi.org/project/geopy/.

[9] V. G. Ushakov, N. G. Ushakov, "Statistical Analysis of Rounded Data: Measurement Errors vs Rounding Errors,"Proceedings of the XXXII International Seminar on Stability Problems for Stochastic Models, Svetlogorsk, Russia, June 12–18, 2016.

[10] Praktyczny samouczek dotyczący regresji ElasticNet. Fikiri.net. https://fikiri.net/praktyczny-samouczek-dotyczacy-regresji-elasticnet/.

[11] "Uczenie maszynowe."Akademia Górniczo-Hutnicza. Dostępne na: https://ai.ia.agh.edu.pl/_media/pl:dydaktyka:mbn:uczenie_maszynowe.pdf.

[12] https://coderslab.pl/pl/blog/dlaczego-python-jest-tak-powszechny-w-badaniach-naukowych

[13] https://jakbadacdane.pl/dlaczego-warto-zainteresowac-sie-scikit-learn/ https://hashdork.com/pl/scikit-learn/.

[14] "Python Requests - biblioteka do wykonywania zapytań HTTP. https://analityk.edu.pl/python-requests-biblioteka-do-wykonywania-zapytan-http.

[15] Stefanie Johnson, Corinne Russell „FEDERAL HOUSING FINANCE AGENCY. (2019, November 26). U.S. House Prices Rise 1.4 Percent in Fourth Quarter 18 Consecutive Quarterly Increases" https://www.fhfa.gov/AboutUs/Reports/ReportDocuments/HPI4Q2015_2252016.pdf

[16] https://pl.tradingeconomics.com/united-states/interest-rate

[17] Stefanie Johnson, Raffi Williams „FEDERAL HOUSING FINANCE AGENCY. (2019, November 26). News Release: U.S. House Prices Rise 1.1 Percent in Third Quarter; Up 4.9 Percent from Last Year. For Immediate". https://www.fhfa.gov/AboutUs/Reports/ReportDocuments/HPI_2021Q4.pdf

[18] Canadian real estate proves resilient with opportunity amid accelerated change

[19] https://forsal.pl/nieruchomosci/artykuly/8172042, boom-na-rynkach-mieszkaniowych-przyczyny.html

Spis rysunków