Przewidywanie cen domów

Warsztaty z technik uczenia maszynowego

Piotr Duperas, Aleksandra Mach, Szymon Pawłowski, Jędrzej Wicha

Spis treści

1 Opis projektu

Temat projektu został znaleziony na platformie *kaggle.com*,
link: https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

Zadanie polega na dość dokładnej analizie zależności cen domów w mieście Ames w stanie Iowa od ich różnych cech – zaczynając od typowych jak powierzchnia domu, liczba pokoi, okolica poprzez mniej oczywiste jak np. powierzchnie poszczególnych pomieszczeń, materiały wykorzystywane do budowy fundamentów i ich jakość, stopień wykończenia, a kończąc na dość nieintuicyjnych cechach takich jak liczba kominków, rodzaj dachu, rok wybudowania garażu.

2 Eksploracja danych

Początkowym etapem pracy było dokładne zapoznanie się z danymi oraz ich przygotowanie do dalszej pracy, czyli do budowy modeli. Trzeba było zrozumieć każdą z 79 kolumn, czyli predyktorów, podzielić je na kategoryczne, ciągłe, jakościowe, zmienić zmienne kategoryczne na liczbowe, zestandaryzować odpowiednie zmienne, zweryfikować, czy mamy do czynienia z outlierami oraz uzupełnić braki danych. Etap pracy z samymi danymi zamknęliśmy w jednej funkcji o nazwie *preprocess*.

2.1 Zmienne kategoryczne

Kolumny, w których wyróżniamy po kilka (ewentualnie kilkanaście) różnych kategorii/grup obserwacji.

- **MSSubClass** – typ budynku, uwzględniający ilość pięter, wiek itp.
  Ma 16 kategorii, które zostały zamienione na wartości liczbowe.
- **MSZoning** – typ przestrzeni w jakiej znajduje się budynek, np. przestrzeń przemysłowa, rolnicza, osiedlowa itp.
  Ma 8 kategorii, które zostały zamienione na wartości liczbowe.
- **Street** – typ ulicy.
  Ma tylko dwie kategorie: Gravel i Paved, które zostały zamienione na wartości liczbowe 0, 1.
- **Alley** – typ alei.
  Ma trzy kategorie: Gravel, Paved, No alley access, które zostały zamienione na wartości liczbowe 0, 1, 2.
- **LotShape** – kształt posesji.
  Ma 4 kategorie, które zostały zamienione na wartości liczbowe.
- **LandContour** – typ przekroju posesji, np. płaska, ze spadkiem, z wgłębieniem.
  Ma 4 kategorie, które zostały zamienione na wartości liczbowe.
- **LotConfig** – typ umieszczenia posesji względem ulic i innych posesji, np. na rogu ulicy, sąsiadująca z 3 innymi posesjami itp.
  Ma 5 kategorii, które zostały zamienione na wartości liczbowe.

- **Neighborhood** – nazwa osiedla.
  Ma aż 25 kategorii, które zostały zamienione na wartości liczbowe.

- **Condition1, Condition2** – dodatkowe cechy, np. bliskość dużych ulic, parków itp.
  Mają 9 kategorii, które zostały zamienione na wartości liczbowe 0,1.

- **BldgType** – typ budynku.
  Ma 5 kategorii, które zostały zamienione na wartości liczbowe.

- **HouseStyle** – styl budynku odnoszący się głównie do ilości pięter.
  Ma 8 kategorii, które zostały zamienione na wartości liczbowe.

- **CentralAir** – obecność klimatyzacji.
  Ma tylko dwie kategorie: Yes, No, które zostały zamienione na wartości liczbowe 0,1.

- **Electrical** – rodzaj układu elektrycznego.
  Ma 5 kategorii: SBrkr, FuseA, FuseF, FuseP, Mix, które zostały zamienione na wartości liczbowe 0,1,2,3,4.

- **GarageType** – typ garażu, tzn. jego lokalizacja w domu.
  Ma braki danych, które z opisu danych oznaczają brak garażu – nadajemy brakom kategorię 0, pozostałe są pogrupowane w kategorie (od 1 do 6): Detchd (odłączony od domu), CarPort (wiata samochodowa), BuiltIn (wbudowany – część domu), Basment (garaż w piwnicy), Attchd (przyłączony do domu), 2Types (2 lub więcej typów garażu).

- **PavedDrive** – czy posiada podjazd,
  Ma 3 kategorie, tak, nie, częściowy, które zostały zamienione na wartości liczbowe.

- **MiscFeature** – Czy posiada udogodnienia, wśród nich są winda, drugi garaż, szopa, boisko do tennisa. Brak danych oznacza brak zanotowanych udogodnień

- **SaleType** Rodzaj wystawionej oferty sprzedaży – .
  Pogrupowane w kategorie (od 1 do 10): WD (Warranty Deed - Conventional), CWD (Warranty Deed - Cash), VWD (Warranty Deed - VA Loan), New (Home just constructed and sold), COD (Court Officer Deed/Estate), Con (Contract 15percent Down payment regular terms), ConLw (Contract Low Down payment and low interest), ConLI Contract Low Interest), ConLD (Contract Low Down), Oth (Other),

- **SaleCondition** jakość wystawionej oferty sprzedaży – .
  Pogrupowane w kategorie (od 1 do 6): Normal (Normal Sale), Abnorml (Abnormal Sale - trade, foreclosure, short sale), AdjLand (Adjoining Land Purchase), Alloca (Allocation - two linked properties with separate deeds, typically condo with a garage unit), Family (Sale between family members), Partial (Home was not completed when last assessed (associated with New Homes)),

2.2 Zmienne porządkowe/jakościowe

Kolumny, w których także mamy po kilka różnych kategorii/grup obserwacji, ale grupy te mają konkretny porządek (np. od najgorszej jakości do najlepszej).

- **Utilities** – dostępne media.
  Ma 4 kategorie porządkowe (od jedynie elektryczności, po wszystkie udogodnienia, czyli elektryczność, gaz, wodę, kanalizację – każdy kolejny poziom to kolejne udogodnienie dostępne), którym nadajemy wartości liczbowe od 0 do 3.

- **LandSlope** – pochylenie posesji.
  Ma 3 kategorie porządkowe (od drobnego pochylenia, do mocnego), którym nadajemy wartości liczbowe od 0 do 2.

- **OverallQual** – jakość materiałów i wykończenia.
  Ma 10 kategorii porządkowych (od najniższej jakości do najwyższej), którym nadajemy wartości liczbowe od 1 do 10.

- **OverallCond** – stan domu.
  Ma 10 kategorii porządkowych (od najniższej jakości do najwyższej), którym nadajemy wartości liczbowe od 1 do 10.

- **YearBuilt** – rok budowy.
  Pozostawiliśmy orygianalne wartości.

- **YearRemodAdd** – rok przebudowy, jeśli natomiast jej nie było, to taki sam jak rok budowy.
  Pozostawiliśmy orygianalne wartości.

- **KitchenQual** – jakość kuchni.
  Ma 5 kategorii porządkowych (od najniższej jakości wykonania kuchni do najwyższej), którym nadajemy wartości liczbowe od 0 do 5: Po (Poor), Fa (Fair), TA (Typical/Average), Gd (Good), Ex (Excellent).

- **Functional** – funkcjonalność domu.
  Ma 8 kategorii porządkowych (od najgorszej funkcjonalności do najlepszej), którym nadajemy wartości liczbowe od 0 do 7: Sal, Sev, Maj2, Maj1, Mod, Min2, Min1, Typ.

- **FireplaceQu** – jakość kominka.
  Ma braki danych, które z opisu danych oznaczają brak kominka – nadajemy brakom kategorię 0, pozostałe są uporządkowane od najniższej jakości do najwyższej (od 1 do 5): Po (Poor), Fa (Fair), TA (Typical/Average), Gd (Good), Ex (Excellent).

- **GarageFinish** – stopień wykończenia garażu.
  Ma braki danych, które z opisu danych oznaczają brak garażu – nadajemy brakom kategorię 0, pozostałe są uporządkowane od najmniej wykończonego do najbardziej (kategorie 1-3): Unf (niewykończony), RFn (lekko wykończony), Fin (wykończony).

- **TotRmsAbvGrd** – liczba pokoi jakości dobrej (powyżej normy).
  Od 2 do 14, im więcej tym lepiej.

- **HalfBath** – liczba niepełnych łazienek wykończonych w dobrej jakości.
  Ma wartości liczbowe 0,1,2 i im więcej, tym lepiej.

- **FullBath** – liczba pełnych łazienek wykończonych w dobrej jakości.
  Ma wartości liczbowe 0,1,2,3 i im więcej, tym lepiej.

- **BsmtHalfBath** – liczba niepełnych łazienek w piwnicy wykończonych w dobrej jakości. Ma wartości liczbowe 0,1,2 i im więcej, tym lepiej.
- **BsmtHalfBath** – liczba niepełnych łazienek w piwnicy wykończonych w dobrej jakości. Ma wartości liczbowe 0,1,2,3 i im więcej, tym lepiej.
- **GarageCars** – pojemność garaża w liczbie aut. Ma wartości liczbowe 0,1,2,3 i im więcej, tym lepiej.
- **GarageQual** – jakość garażu. Ma braki danych, które z opisu danych oznaczają brak garażu – nadajemy brakom kategorię 0, pozostałe są uporządkowane od najniższej jakości do najwyższej (od 1 do 5): Po (Poor), Fa (Fair), TA (Typical/Average), Gd (Good), Ex (Excellent).
- **GarageCond** – stan garażu. Ma braki danych, które z opisu danych oznaczają brak garażu – nadajemy brakom kategorię 0, pozostałe są uporządkowane od najniższej jakości do najwyższej (od 1 do 5): Po (Poor), Fa (Fair), TA (Typical/Average), Gd (Good), Ex (Excellent).
- **PoolQC** – jakość basenu. Ma braki danych, które z opisu danych oznaczają brak basenu – nadajemy brakom kategorię 0, pozostałe są uporządkowane od najniższej jakości do najwyższej (od 1 do 5): Po (Poor), Fa (Fair), TA (Typical/Average), Gd (Good), Ex (Excellent).
- **Fence** – jakość ogrodzenia. Ma braki danych, które z opisu danych oznaczają brak ogrodzenia – nadajemy brakom kategorię 0, pozostałe są uporządkowane od najniższej jakości do najwyższej (od 1 do 4): MnWw (Minimum Wood/Wire), GdWo (Good Wood), MnPrv (Minimum Privacy), GdPrv (Good Privacy).

2.3 Zmienne ciągłe

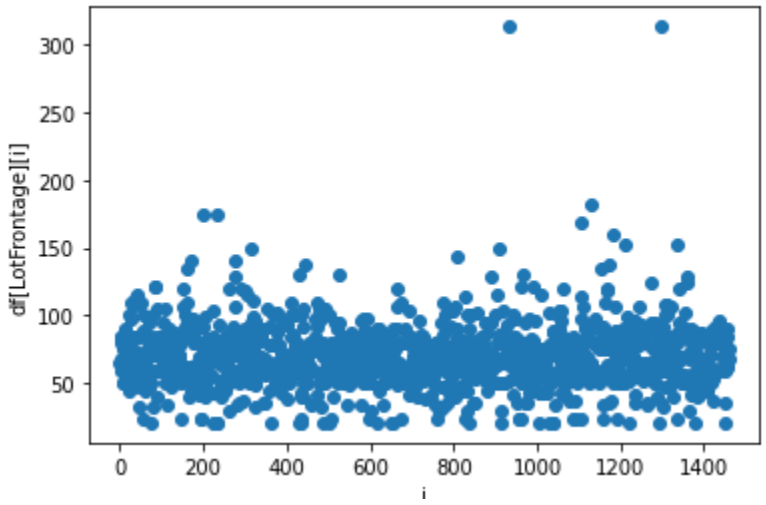
- **LotFrontage** – długość ulicy połączonej z posesją. Zastosowano standaryzację.
- **LotArea** – powierzchnia posesji w stopach kwadratowych. Zastosowano standaryzację.
- **1stFlrSF** – powierzchnia pierwszego piętra. Zastosowano standaryzację.
- **1stFlrSF** – powierzchnia pierwszego piętra. Zastosowano standaryzację.
- **1stFlrSF** – powierzchnia pierwszego piętra. Zastosowano standaryzację.
- **2ndflrSF** – powierzchnia drugiego piętra. Zastosowano standaryzację.
- **LowQualFinSF** – powierzchnia złej jakości. Zastosowano standaryzację.
- **GrLivAre** – powierzchnia dobrej jakości (jakości ponad normę). Zastosowano standaryzację.
- **WoodDeckSF** – drewniana powierzchnia. Zastosowano standaryzację.
- **OpenPorchSF** – powierzchnia werandy na otwartej przestrzeni. Zastosowano standaryzację.

- **EnclosedPorch** – powierzchnia werandy na zamkniętej przestrzeni. Zastosowano standaryzację.
- **3SsnPorch** – powierzchnia werandy przystosowanej do użytku przez 3 pory roku. Zastosowano standaryzację.
- **ScreenPorch** – powierzchnia werandy odgrodzonej szybą. Zastosowano standaryzację.
- **PoolArea** – Powierzchnia basenu. Zastosowano standaryzację.

Zmienne podejrzane o posiadanie outlierów:

- **LotFrontage**

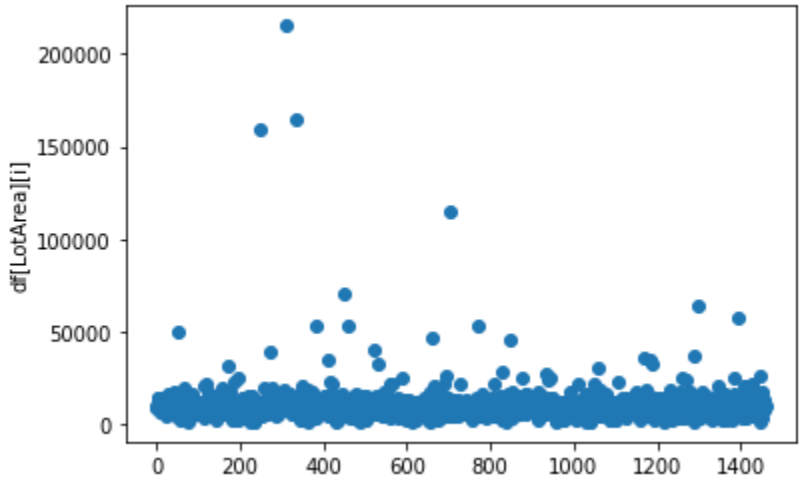
Rysunek 1: Wykres rozproszenia zmiennej LotFrontage



Widzimy tu dwie obserwacje znacząco wystających ponad chmurę punktów. Aby je wyeliminować, usunęliśmy rekordy, które są odchylone od średniej o ponad 4 odchylenia standardowe.

- **LotArea**

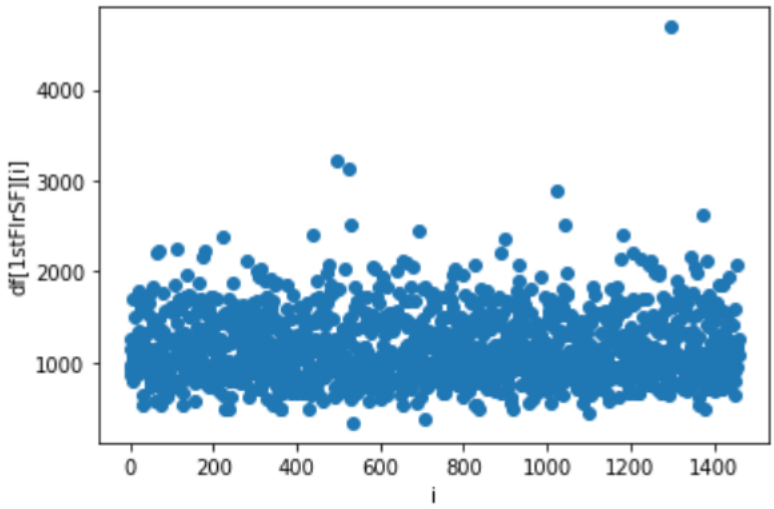
Rysunek 2: Wykres rozproszenia zmiennej LotArea



Widzimy tu kilka obserwacji wystających ponad chmurę punktów. Aby je wyeliminować, usunęliśmy rekordy, które są odchylone od średniej o ponad 4 odchylenia standardowe.

- **1stFlrSF**

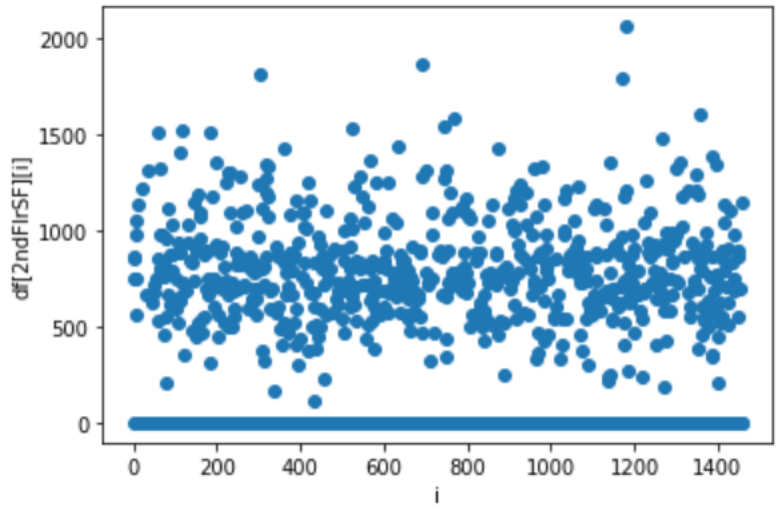
Rysunek 3: Wykres rozproszenia zmiennej 1stFlrSF



Widzimy tu kilka obserwacji wystających ponad chmurę punktów, w tym jedną wyróżniającą się. Jednak ponieważ zmienna ta opisuje powierzchnię pierwszego piętra, to zakładamy, że możliwe, że istnieje taki dom, który faktycznie może mieć tak duże pierwsze piętro (i też może to znacząco wpływać na jego cenę), więc nie uznajemy tego za obserwację odstającą.

- **2ndFlrSF**

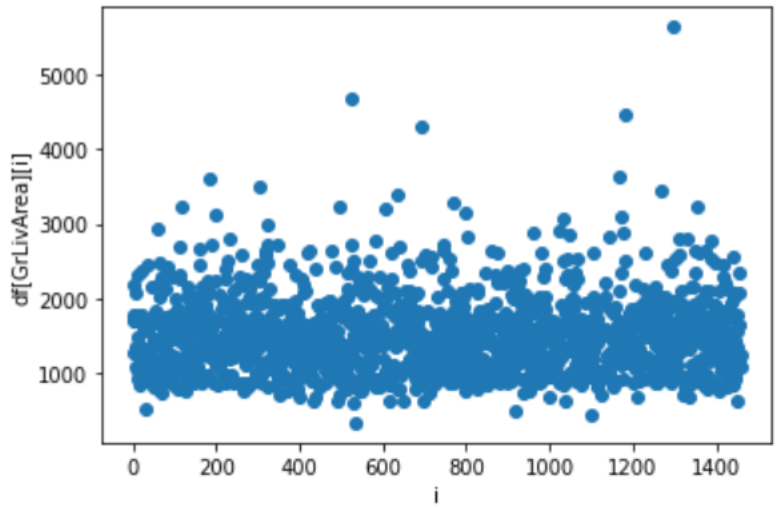
Rysunek 4: Wykres rozproszenia zmiennej 2ndFlrSF



Zakładamy, że zmienna 2ndFlrSF nie ma obserwacji odstających.

- **GrLivAre**

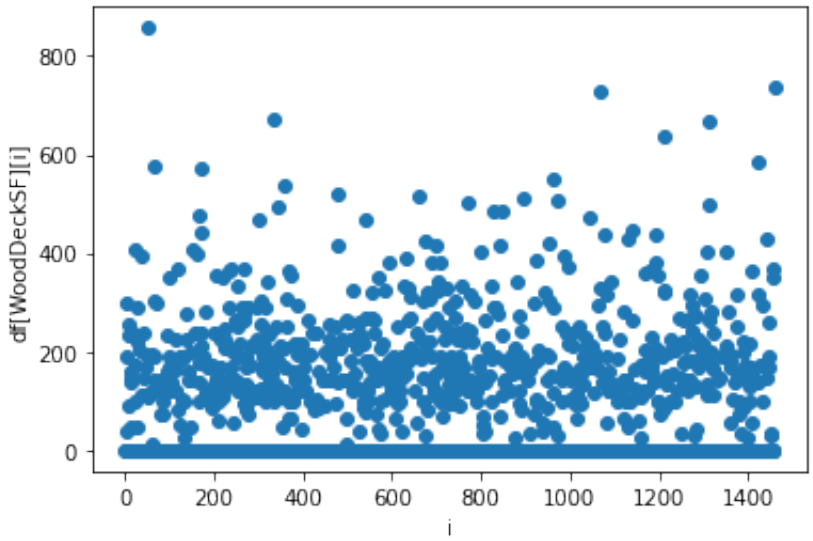
Rysunek 5: Wykres rozproszenia zmiennej GrLivAre



Ponownie ze względu na to, że zmienna opisuje powierzchnię domu w jakości powyżej przeciętnej, to uznajemy, że istnieją domy, w których ta powierzchnia może trochę odstawać od reszty, więc też tutaj nie wyróżniamy obserwacji odstających.

- **WoodDeckSF**

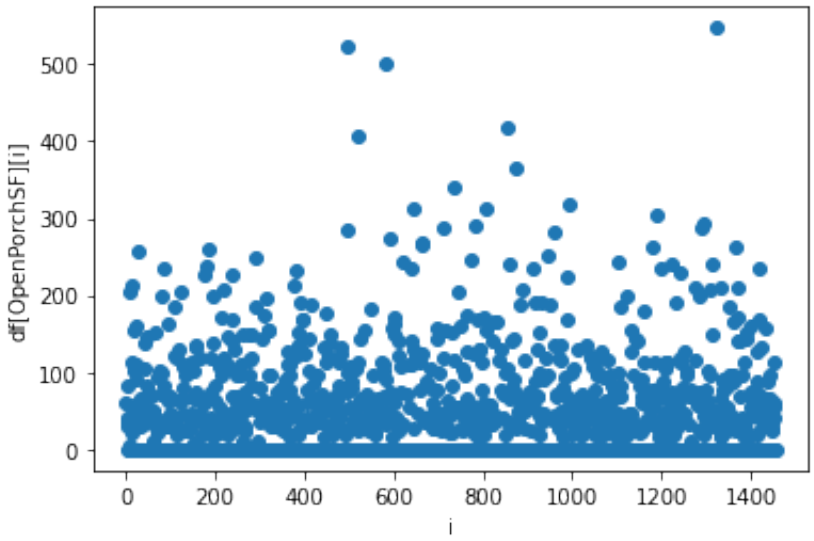
Rysunek 6: Wykres rozproszenia zmiennej WoodDeckSF



Widzimy tu kilka obserwacji wystających ponad chmurę punktów. Aby je wyeliminować, usunęliśmy rekordy, które są odchylone od średniej o ponad 5 odchylenia standardowe.

- **OpenPorchSF**

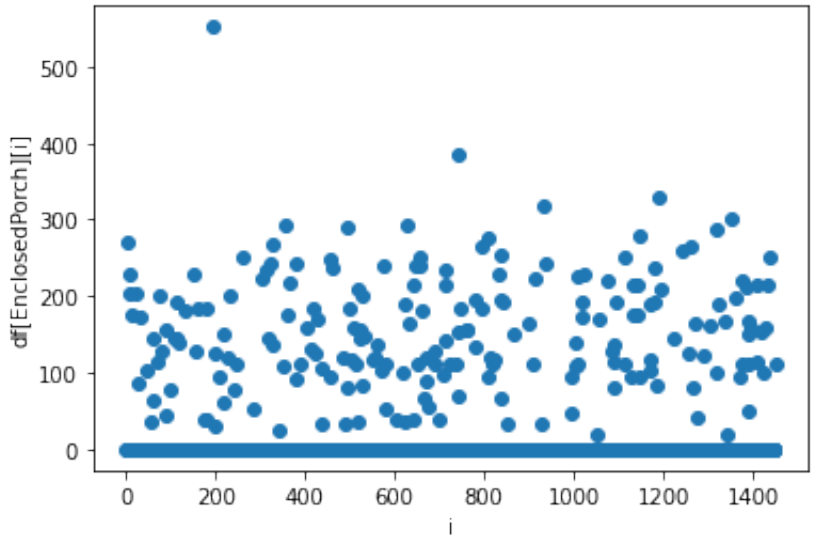
Rysunek 7: Wykres rozproszenia zmiennej OpenPorchSF



Widzimy tu kilka obserwacji wystających ponad chmurę punktów. Aby je wyeliminować, usunęliśmy rekordy, które są odchylone od średniej o ponad 5 odchylenia standardowe.

- **EnclosedPorch**

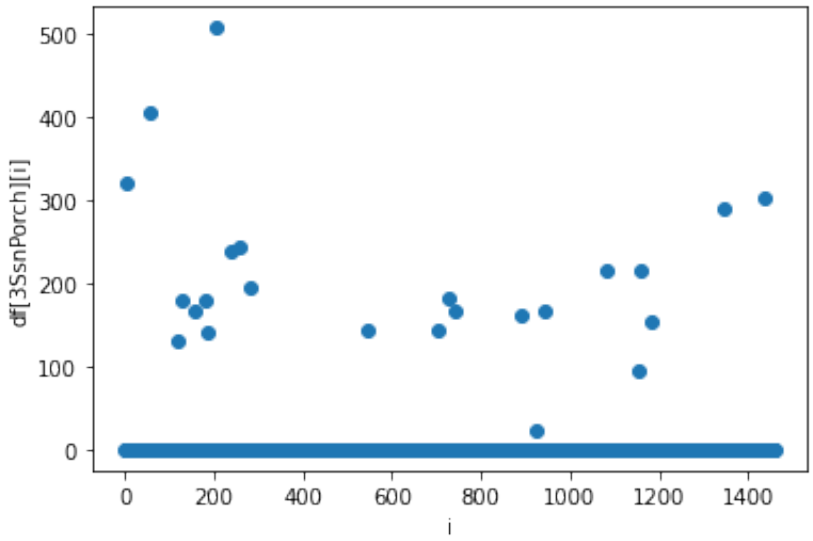
Rysunek 8: Wykres rozproszenia zmiennej EnclosedPorch



Widzimy tu kilka obserwacji wystających ponad chmurę punktów. Aby je wyeliminować, usunęliśmy rekordy, które są odchylone od średniej o ponad 5 odchylenia standardowe.

- **3SsnPorch**

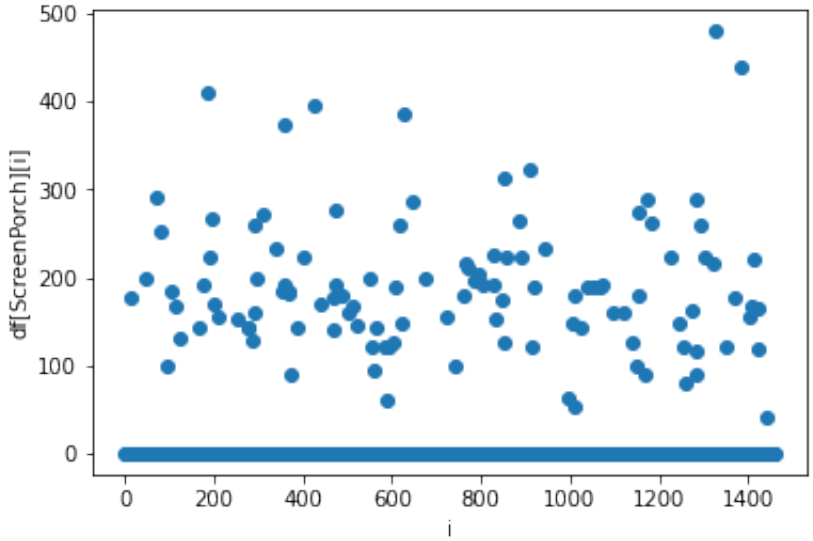
Rysunek 9: Wykres rozproszenia zmiennej 3SsnPorch



Dla tej zmiennej nie ma zbyt wiele danych stąd usuwanie obserwacji odstających nie ma zbyt dużych podstaw.

- **ScreenPorch**

Rysunek 10: Wykres rozproszenia zmiennej ScreenPorch



Zmienna nie posiada wyraźnych obserwacji odstających

3 Budowa modeli

Mieliśmy dostępne na stronie dwa zbiory danych – train.csv oraz test.csv, przy czym tylko pierwszy z nich zawierał także zmienną objaśnianą – cenę domów, wobec tego na nim budowaliśmy i testowaliśmy modele. W tym celu podzieliśmy go na część treningową – 70% obserwacji i testową – 30% obserwacji.

Do wyboru najlepszych parametrów dla każdego modelu używaliśmy funkcji *GridSearchCV*, która przyjmuje parametr *param_grid* – słownik parametrów, z których chcemy wybrać najlepszą kombinację parametrów dla danego modelu i zwraca ten model z dobranymi parametrami.

Aby porównywać otrzymywane rezultaty korzystaliśmy z metody .score dostępnej dla każdego z rozpatrywanych modeli. Metoda ta zwraca współczynnik determinacji R^2 zdefiniowany w następujący sposób:

$$R^2 = \frac{\text{SSR}}{\text{SST}},$$

gdzie

$$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2, \quad \text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

gdzie Y_i – i-ta obserwacja zmiennej Y, $\hat{Y}_i$ – predykcja i-tej obserwacji na podstawie dopasowanego modelu, $\bar{Y}$ – średnia obserwacji $Y_1, ..., Y_n$.

3.1 Las losowy

3.1.1 Opis modelu

Las losowy (ang. Random Forest) to metoda komitetów/zespołów w uczeniu maszynowym, która polega na konstruowaniu wielu drzew decyzyjnych regresyjnych, a następnie uśrednianiu ich wyniku. Drzewa decyzyjne są najczęściej wybieraną metodą bootstrapu, tzn. ustalamy liczbę B próbek bootsrapowych zbioru treningowego (losowanie ze zwracaniem), dla każdej próbki zbioru, nazwijmy ją D_i budujemy drzewo decyzyjne, które daje wynik M_i. Wtedy ostateczną predykcją będzie

$$\hat{Y} = \frac{1}{B}\sum_{i=1}^{B} M_i.$$

3.1.2 Wybór najlepszych parametrów

Dla funkcji *RandomForestRegressor* wybieraliśmy najlepsze z następujących parametrów:

- n_estimators – liczba drzew w lesie losowym, z przedziału [10,1010] co 50,
- min_samples_split – minimalna liczba obserwacji potrzebna, aby rozdzielić liść drzewa na grupy, z przedziału [2,20] co 1.

Najlepsze parametry: n_estimators = 110, min_samples_split = 7.

3.1.3 Wyniki

$R^2 \approx 0.878$

3.2 Bagging

3.2.1 Opis modelu

Bagging to metoda komitetów/zespołów w uczeniu maszynowym, która polega na dopasowywaniu składowych modeli na losowych podzbiorach oryginalnego zbioru danych, a następnie uśrednianiu ich wyników w celu osiągnięcia końcowego rezultatu. Domyślnie, jako składowe modele używane są drzewa decyzyjne.

3.2.2 Wybór najlepszych parametrów

Dla funkcji *BaggingRegressor* wybieraliśmy najlepsze z następujących parametrów:

- n_estimators – liczba drzew decyzyjnych, z przedziału [10,150] co 10,
- max_samples – ilość próbek ze zbioru do trenowania pojedynczego modelu, z przedziału [30,90] co 3.

Najlepsze parametry: n_estimators = 70, max_samples = 87.

3.2.3 Wyniki

$R^2 \approx 0.834$

3.3 Regresja liniowa

3.3.1 Opis modelu

Regresja liniowa jest jedną z najbardziej standardowych metod używanych w teorii uczenia maszynowego. Zakłada ona, że pomiędzy zmienną objaśnianą (odpowiedzią) $\mathbf{Y}$ a zmiennymi objaśniającymi (predyktorami) $\mathbf{X}_i$. Zależność ta jest modelowana z uwzględnieniem błędu losowego ε. Zatem model liniowy ma postać:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

gdzie $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$ – macierz, której kolumny są predyktorami, $\boldsymbol{\beta} \in \mathbb{R}^p$ - wektor współczynników kombinacji liniowej, a $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ – wektor niezależnych zmiennych losowych. Zadaniem, które rozwiązujemy jest przewidywanie zmiennej $\mathbf{Y}$ na podstawie macierzy $\mathbf{X}$ poprzez znajdowanie współczynników $\boldsymbol{\beta}$, które zapewniają jak najlepsze dopasowanie modelu. Najbardziej standardowym podejściem (używanym także w pakiecie sklearn, z którego korzystamy) jest metoda najmniejszych kwadratów, która minimalizuje sumę kwadratów błędów popełnianych przy predykcji każdego Y_i dla $i = 1, ..., n$, tzn.

$$\boldsymbol{\beta}_{MNK} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \, \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij} b_j \right)^2 .$$

3.3.2 Wybór najlepszych parametrów

Funkcja *LinearRegression* nie posiada parametrów, z których można byłoby wybierać najlepsze. Wobec tego zostawiamy ją z domyślnymi parametrami.

3.3.3 Wyniki

$R^2 \approx 0.807$

3.4 Regresja bayesowska

3.4.1 Opis modelu

Bayesowka regresja liniowa jest statystycznym podejściem do regresji liniowej w którym wykorzystywane jest wnioskowanie Bayesowskie. Zakładamy, że błędy są od siebie niezależne i mają rozkład normalny. Dodatkowo jako rozkład apriori dla parametrów alpha i lambda bierzemy rozkład Gamma. Razem z liczbą iteracji poddajemy je testom, dzięki którym możemy wybrać ich najlepsze wartości.

3.4.2 Wybór najlepszych parametrów

- n_iter – maksymalna liczba iteracji, ze zbioru [1, 5, 10, 20, 30, 50, 100, 300],
- alpha_1 – parametr kształtu dla rozkładu Gamma apriori dla parametrem alfa, ze zbioru [1e-8, 1e-6, 1e-4].
- alpha_2 – odwrócony parametr skali dla rozkładu Gamma apriori dla parametrem alfa, ze zbioru [1e-8, 1e-6, 1e-4, 1e-2, 1].
- lambda_1 – parametr kształtu dla rozkładu Gamma apriori dla parametrem alfa, ze zbioru [1e-8, 1e-6, 1e-4, 1e-2, 1].
- lambda_2 –odwrócony parametr skali dla rozkładu Gamma apriori dla parametrem alfa, ze zbioru [1e-8, 1e-6, 1e-4].

Najlepsze parametry: n_iter = 10, alpha_1 = 1e-08, alpha_2 = 1, lambda_1 = 1, lambda_2 = 1e-08

3.5 Regresja ARD

3.5.1 Opis modelu

Regresja ARD (Automatic Relevance Determination) działa podobnie do regresji Bayesowskiejm jednak w przeciwieństwie do niej zamiast używać zwykłej metody najmniejszych kwadratów skaluje otrzymane współczynniki w stronę zera co zapewnia ich większą stabilność. Podobnie jak w przypadku regresji Bayesowkiej zajmujemy się parametrami rozkładu apriori Gamma oraz liczbą iteracji.

3.5.2 Wybór najlepszych parametrów

- n_iter – maksymalna liczba iteracji, ze zbioru [1, 5, 10, 30, 50, 100],
- alpha_1 – parametr kształtu dla rozkładu Gamma apriori dla parametrem alfa, ze zbioru [1e-8, 1e-6, 1e-4, 1e-2, 1].
- alpha_2 – odwrócony parametr skali dla rozkładu Gamma apriori dla parametrem alfa, ze zbioru [1e-6, 1e-4, 1e-2, 1].
- lambda_1 – parametr kształtu dla rozkładu Gamma apriori dla parametrem alfa, ze zbioru [1e-6, 1e-4, 1e-2, 1].
- lambda_2 –odwrócony parametr skali dla rozkładu Gamma apriori dla parametrem alfa, ze zbioru [1e-6, 1e-4, 1e-2, 1].

Najlepsze parametry: n_iter = 10, alpha_1 = 1e-08, alpha_2 = 1, lambda_1 = 1e-2, lambda_2 = 1e-04

3.6 Boosting

3.6.1 Opis modelu

Podstawowym algorytmem do implementacji boostingu jest AdaBoost (ang. Adaptive Boosting). Podobnie jak las losowy jest to metoda komitetów, czyli polega na iteracyjnym budowaniu wielu modeli. Za każdym razem gdy dopasowywany jest model, algorytm przypisuje większe wagi źle sklasyfikowanym obserwacjom, dzięki czemu kolejne modele "zwracają większą uwagę" na te obserwacje. Wagi rekordów są wybierane w taki sposób, aby błąd na zbiorze treningowym malał z szybkością wykładniczą. Istotną cechą boostingu jest to, że przy przypisywaniu wag zapominamy o poprzednim modelu, więc kolejne modele są od siebie niezależne, a to skutkuje większą zmienniczością modelu.

3.6.2 Wybór najlepszych parametrów

Najczęściej wykorzystywanym modelem w algorytmie AdaBoost jest drzewko decyzyjne i dlatego też funkcja *AdaBoostRegressor* ma ustawiony model *DecisionTreeRegressor* jako domyślny. Oprócz drzewek spróbowaliśmy do niego dopasować także kilka innych modeli. Dla każdego przypadku dobieraliśmy najlepszą liczbę modeli z przedziału [10, 1010] co 10.

- Drzewka decyzyjne: n_estimators = 800,
- Regresja liniowa: n_estimators = 10,
- Regresja bayesowska: n_estimators = 460,
- Regresja grzbietowa: n_estimators = 10,
- Regresja ARD: n_estimators = 10.

3.6.3 Wyniki

- Drzewka decyzyjne: $R^2 \approx 0.787$,
- Regresja liniowa: $R^2 \approx 0.633$,
- Regresja bayesowska: $R^2 \approx 0.662$,
- Regresja grzbietowa: $R^2 \approx 0.653$,
- Regresja ARD: $R^2 \approx 0.674$.

4 Podsumowanie

Najlepszym modelem na podstawie kryterium R^2 okazał się być las losowy, dla którego otrzymaliśmy bardzo wysoki wynik $R^2 \approx 0.878$. Oznacza to, że nasz model z wysoką dokładnością potrafi przewidywać ceny domów w zadanym problemie.