

Regresja liniowa (bardzo mocno ugruntowana w teorii)

Cena = wps1\*zmienna1 + wsp2\*zm2 + wyraz wolny

Ważna jest teoria! żeby model dobrze opisywał rzeczywistość

**Intercept** - wyraz wolny -> wyjściowa cena mieszkania zakładając, że dla wszystkich innych zmiennych dana obserwacja (dany współczynnik) przyjmuje wartość 0

Prosta regresja liniowa

* jedna zmienna objaśniająca

Wielokrotna/Wielowymiarowa Regresja Liniowa (MLR - Multiple Linear Regression)

* wiele zmiennych objaśniających

Dodając poszczególne zmienne jesteśmy w stanie ocenić

* w jaką interakcję zmienne wchodzą z ceną,
* jak cena jest powiązana z tymi zmiennymi
* jaki wpływ mają zmienne na cenę końcową
* ważne: jeżeli już widzimy wyraźną dominację zmiennej w cenie to mówimy, że

-> coś jest silnie skorelowane z ceną (powstrzymując się od interpretacji na + lub - ; po prostu jest silna korelacja)

## 

## Porównywanie modeli liniowej regresji wielorakiej

Liniowa regresja wieloraka daje możliwość jednoczesnej analizy wielu zmiennych niezależnych. Pojawia się więc problem wyboru optymalnego modelu. W natłoku informacji jakie niesie zbyt duży model istnieje możliwość zagubienia ważnych informacji. Zbyt mały może pominąć te cechy, które w wiarygodny sposób mogłyby opisać badane zjawisko. Bowiem nie liczba zmiennych w modelu, ale ich jakość decyduje o jakości modelu. W wyborze zmiennych niezależnych niezbędna jest wiedza i doświadczenie związane z badanym zjawiskiem. Należy pamiętać, by w modelu znajdowały się zmienne silnie skorelowane ze zmienną zależną i słabo skorelowane między sobą.

Nie istnieje jedna prosta reguła statystyczna, która decydowałaby o liczbie zmiennych niezbędnych w modelu. Najczęściej w porównaniu posługujemy się miarami dopasowania modelu takimi jak: $R_{adj}^2$ - poprawiona wartość współczynnika determinacji wielorakiej (im wyższa wartość tym lepiej dopasowany model), $SE_e$ - błąd standardowy estymacji (im niższa wartość tym lepiej dopasowany model). W tym celu można również wykorzystać test F oparty o współczynnik determinacji wielorakiej $R^2$. Test ten służy do weryfikacji hipotezy, że dopasowanie obu porównywanych modeli jest tak samo dobre.

**[krok 1]** Zbudowanie modelu z wszystkich zmiennych.

**[krok 2]** Usunięcie jednej zmiennej z modelu. Usuwana zmienna to ta, która ze statystycznego punktu widzenia wnosi do aktualnego modelu najmniej informacji.

**[krok 3]** Porównanie modelu pełnego i zredukowanego.

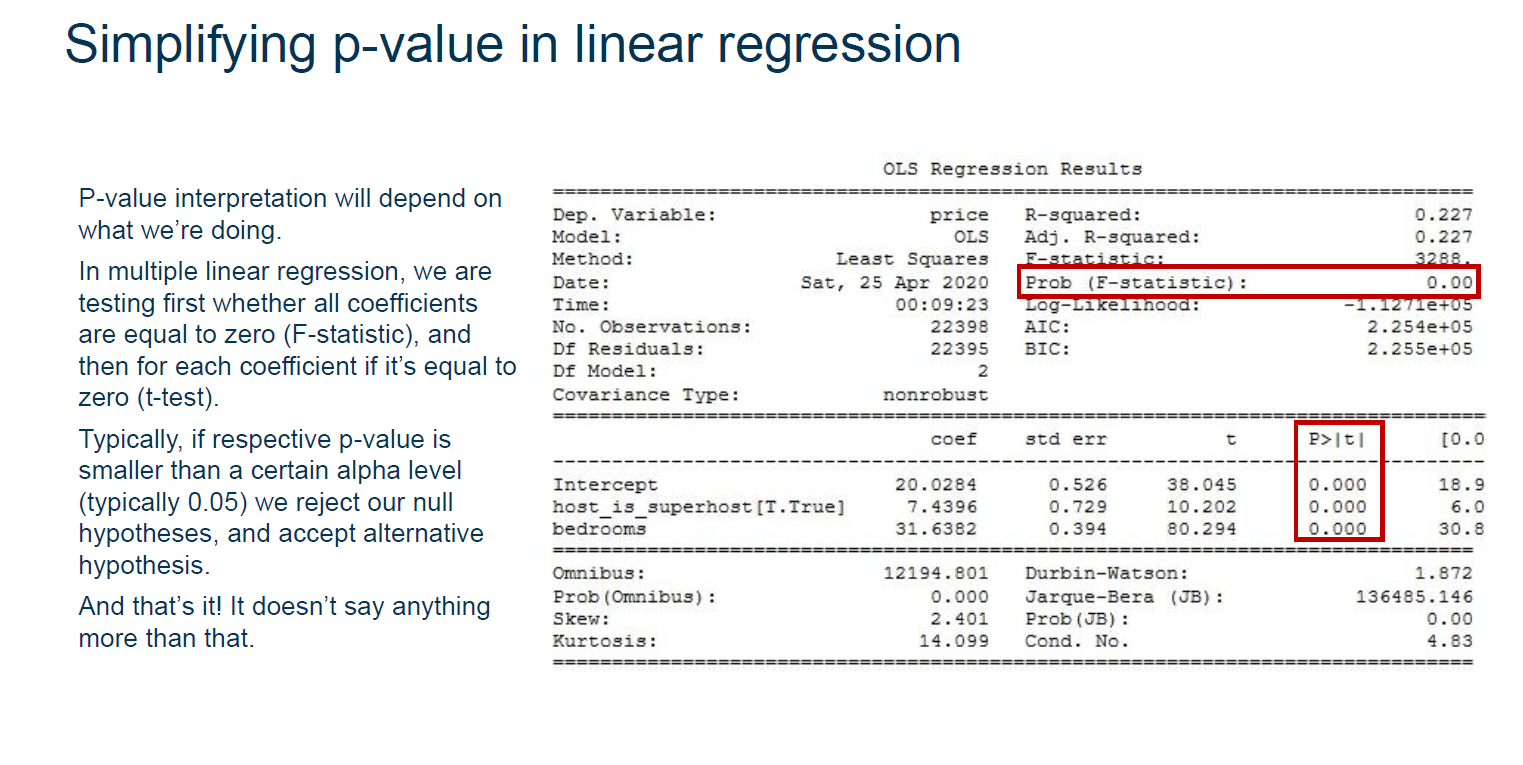
**[krok 4]** Usunięcie kolejnej zmiennej z modelu. Usuwana zmienna to ta, która ze statystycznego punktu widzenia wnosi do aktualnego modelu najmniej informacji.

**[krok 5]** Porównanie modelu wcześniejszego i nowo zredukowanego.

**[…]**

W ten sposób powstaje wiele, coraz mniejszych modeli. Ostatni model zawiera tylko 1 zmienną niezależną.

<http://manuals.pqstat.pl/statpqpl:wielowympl:porownpl>



P-Value w przypadku wielokrotnej regresji liniowej ma przyjmować wartość 0 (lub poniżej 0,05), dla:

* wszystkich współczynników (coefficient) łącznie (F-statistics)
* a następnie dla każdego współczynnika (P>|t|)
* daje to podstawy do odrzucenia hipotezy zerowej i zaakceptowania hipotezy alternatywnej

P-Value

* konstrukt bardzo specyficzny, często używany w świecie naukowym (statystyka częstości / Frequentist inference/Statistics)
  + Wnioskowanie częstościowe, NHST, statystyka częstościowa:
    - zakłada że istnieje zjawisko, a my tylko losowo pobieramy dane z tego zjawiska
    - przykład: istnieje nieskończenie wiele mieszkań w Berlinie, a my bierzemy sobie ceny tych mieszkań i tworząc jakąś próbkę, chcąc na tej podstawie być w stanie powiedzieć coś o całości mieszkań
    - uważane za podejście prehistoryczne
* ma sens wtedy kiedy testujemy hipotezy statystyczne:
  + zakładając, że hipoteza zerowa jest prawdziwa
    - [HIPOTEZA ZEROWA: mówi że nie ma żadnego związku między X a Y]
    - a mimo wszystko chcemy zrobić regresję liniową
  + to jakie jest prawdopodobieństwo tego, że otrzymaliśmy wyniki równie skrajne jak te które otrzymaliśmy bądź bardziej skrajne
  + p-value = prawdopodobieństwo takiego zdarzenia
* wartość P pomaga poczuć czy między danymi jakie widzimy jest jakaś korelacja, czy też nic się nie dzieje.

P-Value

* test który mówi czy możemy odrzucić hipotezę zerową mówiącą o tym, że wartość wszystkich współczynników razem wziętych powinna wynosić zero

Jeżeli danych jest mało to widać że zachodzą powiązania między x i y

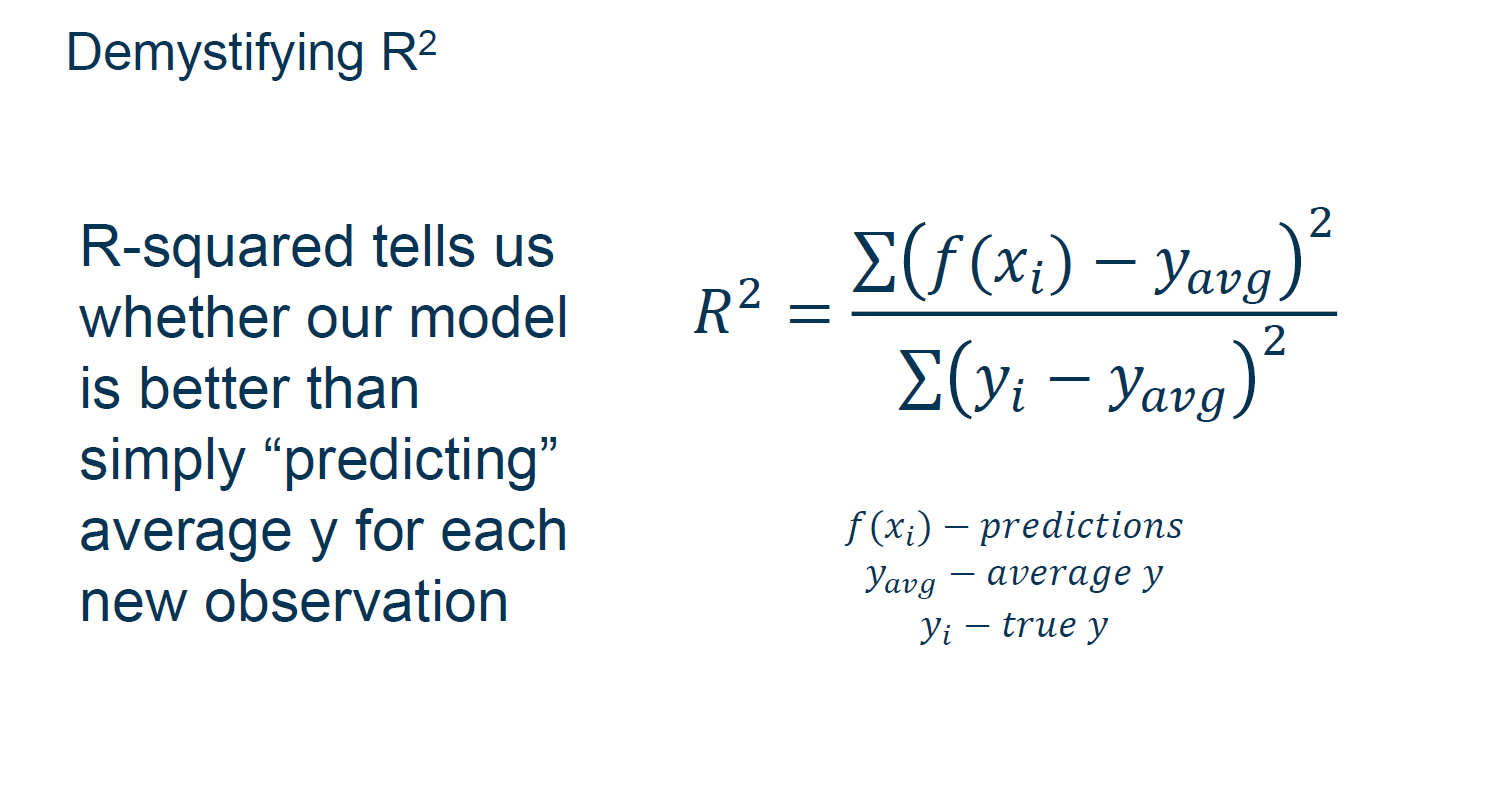
Jeżeli danych jest dużo widać, że powiązań jest coraz mniej

Statystyka t [kolumna P>|t|]

* zakładając że nie ma żadnego związku między zmienną x (przykładowo liczba łóżek) a ceną Y
* jakie jest prawdopodobieństwo, że otrzymalibyśmy określoną (wskazaną w wyniku) wartość współczynnika (wartość podaną lub bardziej skrajną)
  + czy też otrzymalibyśmy taką wartość współczynnika
* odpowiedź w kolumnie P>|t|

P>|t|

* test, który odpowiada na pytanie czy możemy odrzucić hipotezę zerową mówiącą o tym, że prawdziwa wartość współczynnika wynosi 0

****

**R2**

miara mówiąca o ile mocniejszy jest nasz model w porównaniu do modelu, który prognozuje na każde zapytanie średnią (średnią odp)

(wartości z przedziału [-1; +1], gdzie 1 = idealnie lepszy niż średnie prognozowanie)

Każda nowa zmienne dodawana do modelu będzie nam poprawiała R2

R2 jest monotonicznie niemalejący względem zmiennych, które dorzucamy do modelu (im więcej dorzucimy tym większy będzie R2)

**Dopasowane R2**

ma karę za ilość zmiennych

**AIC** - kryterium Akaike’a (Akaike information criterion)

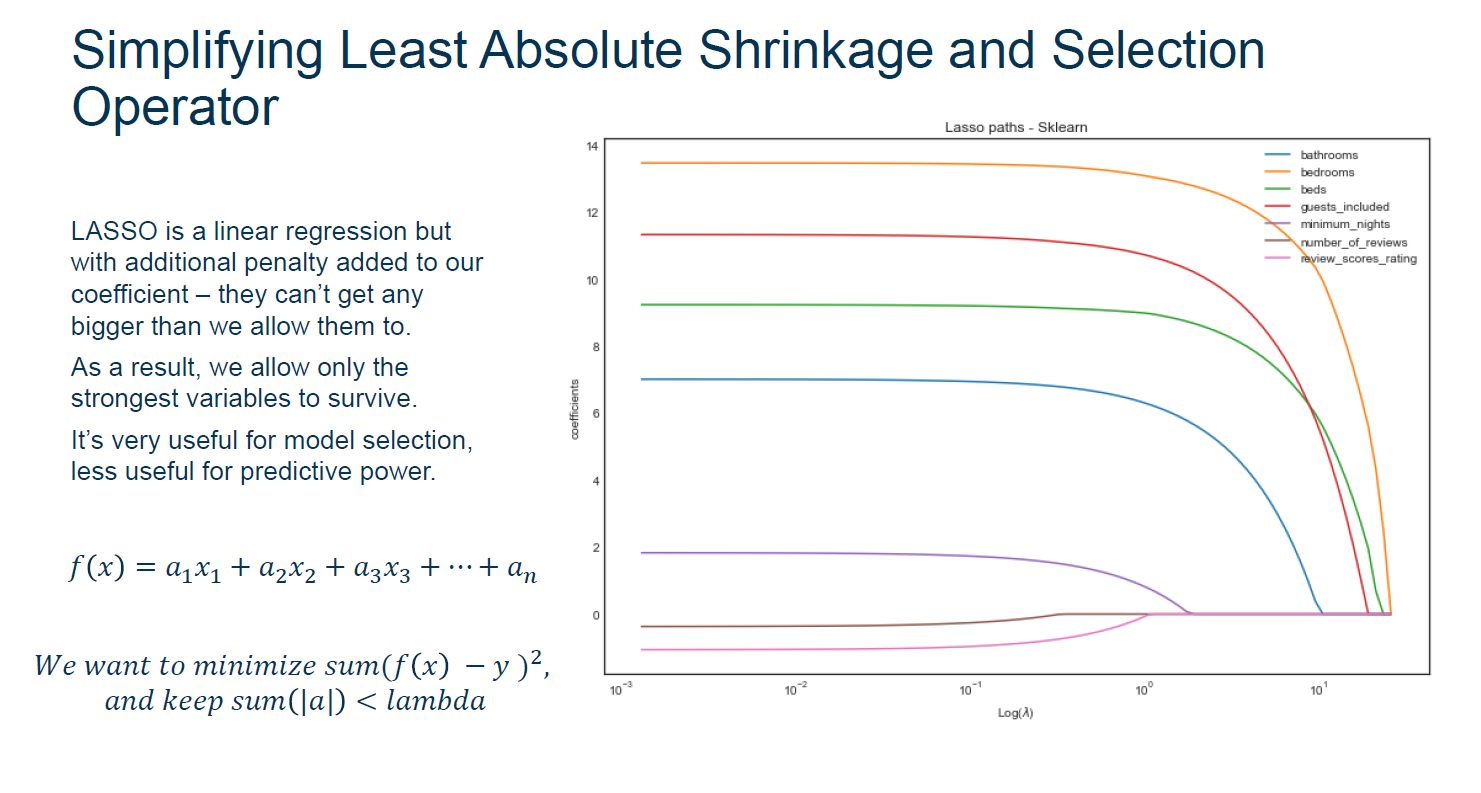
**BIC** - kryterium Bayes’a (Bayesian information criterion)

* pomocne w ocenie modeli wyspecyfikowanych na tych samych danych, ale które biorą pod uwagę inne zmienne
* im mniejsza wartość tym lepsza = model ma lepszą wartość prognostyczną
* można sprawdzać wyrzucając/podmieniając zestaw wartości

**Lasso**

* stosujemy kiedy mamy dużo zmiennych to do końca nie wiemy, których zmiennych chcemy użyć
* dlatego eksplorujemy, które zmienne są powiązane ze zmienną objaśnianą ( w naszym przykładzie ceną)
* Lasso -> nowsza metoda niż zwykłą regresja liniowa (uważana za “prehistoryczną”), czyli:
  + robimy regresję liniową, ale dodając dodatkowy element:
  + karę na współczynniki (w normalnej regresji liniowej współczynniki mogą przyjmować dowolne wartości (może być 20, może być i 200)
  + stawiamy taki model optymalizacji, gdzie:
    - robimy regresję liniową, ale
    - suma wartości bezwzględnych współczynników musi być mniejsza niż parametr LAMBDA.
    - im mniejsza lambda tym mniejsze wartości
    - im mniejsza jest lambda tym więcej współczynników będzie nam się zerowało - czyli będą nam wypadały z modelu

Lasso: pokazuje jak cena mieszkania jest powiązana z różnymi zmiennymi w zależności od parametru Lambda - im mniejsza Lambda tym suma modułów współczynników musi być mniejsza. Na tej podstawie możemy wybrać współczynniki, które najsilniej oddziałują na zmienną objaśnianą.



Zostaje kwestia: jak dobrać parametr Lambda.

Walidacja krzyżowa:

* + na zbiorze treningowym uczymy model
  + na walidacyjnym, walidujemy sprawdzamy jaki jest błąd
    - wartość parametru lambda, która nam minimalizuje błąd na zbiorze walidacyjnym jest wartością optymalną i taki model powinniśmy przyjąć

ALE:

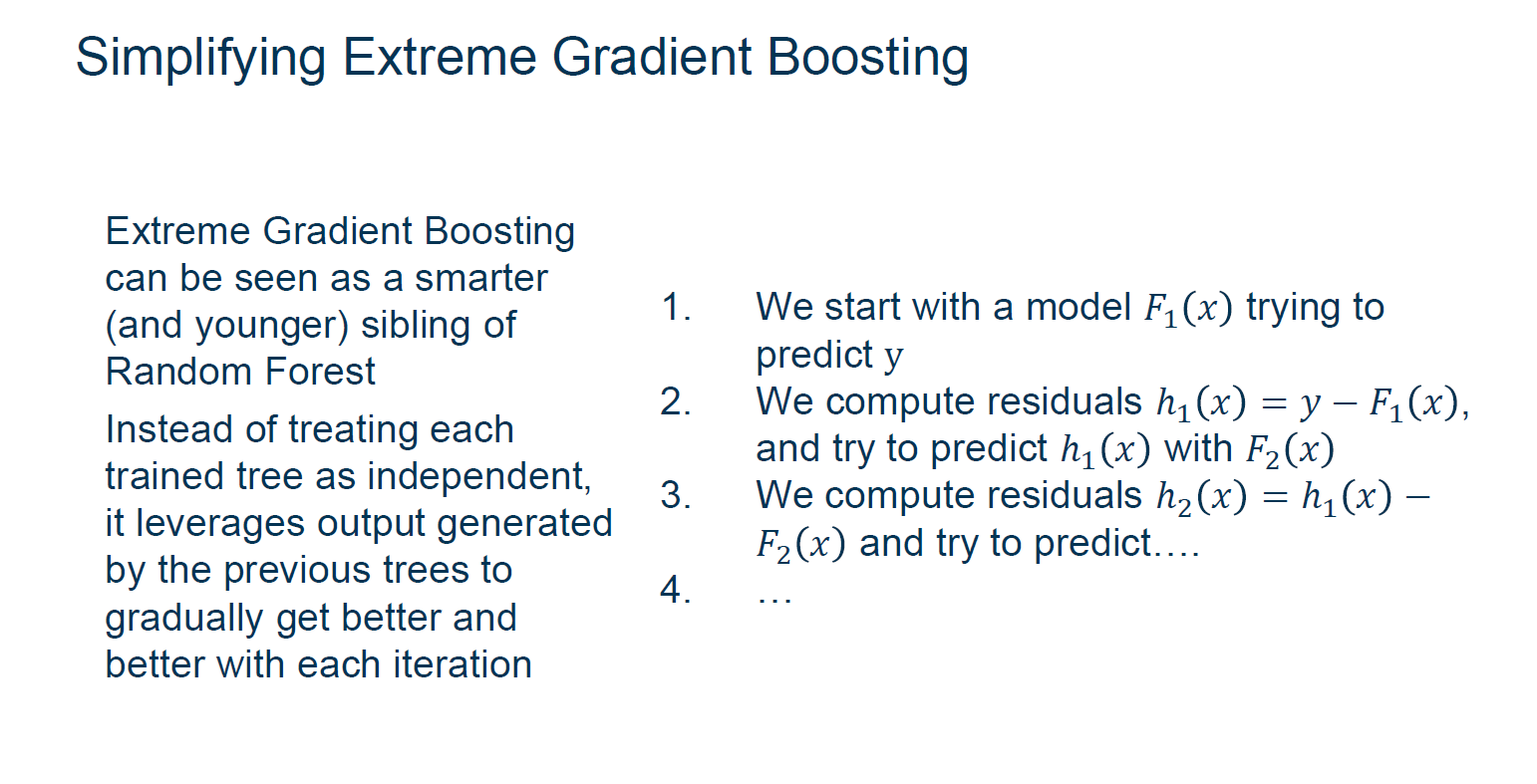
* + - jeżeli w wyniku przesunięcia Lambdy dostaniemy czytelniejszy model, to warto to zrobić

W Lasso wrzucamy wszystkie zmienne i stosując walidację krzyżową dobieramy optymalną Lambdę, żeby zobaczyć które zmienne odpadają najpóźniej.

**XGBoost**

Lasy losowe mają swoje plusy, ale Extreme Gradient Boosting jest szybszy.

Nie szkolimy wszystkich drzew jednocześnie. Szkolimy drzewa sekwencyjnie: pierwsze nasze drzewo stara się prognozować cenę. Kiedy to zrobimy liczymy rezydua - błędy. Następnie odejmujemy od wektora cen wektor predykcji i dostajemy wektor błędu. Mając wektor błędów dla pierwszego drzewa uczymy drugie drzewo prognozować błędy. Mając wyliczone prognozy błędów, tworzymy trzecie drzewo, które stara się prognozować błędy błędów...

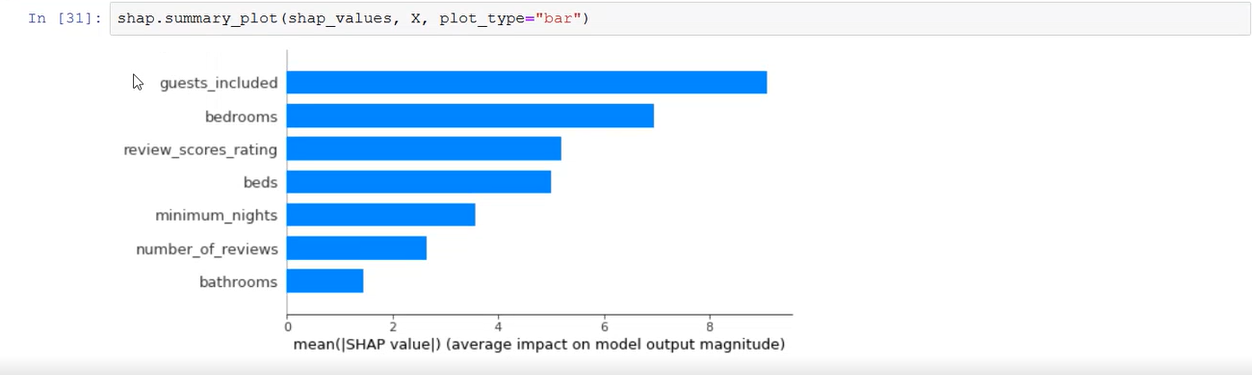


Ciężko jest wytłumaczyć, dlaczego model dał taką a nie inną odpowiedź. Model jest trudno interpretowany. Są sposoby wizualizacji **shapley values**

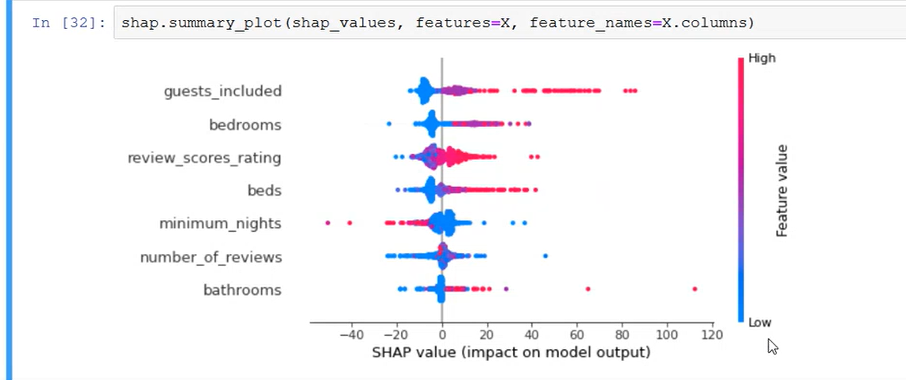
****

Pokazuje jak dla poszczególnych obserwacji cena będzie się różniła od ceny bazowej

* gdzie cena bazowa to cena dla średniego mieszkania [base value] (która ma średnią ilość sypialni, łazienek itd...)
* jak ta cena się odsunie od tej średniej ceny i dlaczego?
* dla poszczególnych obserwacji shap\_value[0, <- czyli dla mieszkania nr 0
  + to co nas pcha w prawo to zwiększa cenę
  + to co nas pcha w lewo to zmniejsza cenę

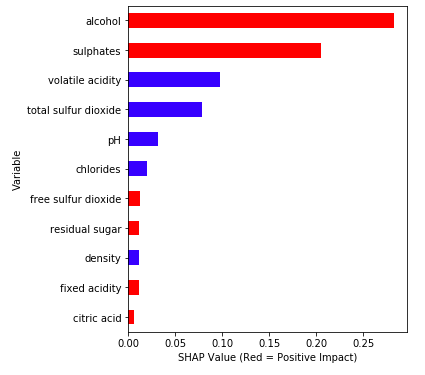


* **które zmienne mają największy wpływ na cenę**

****

* This plot is made of all the dots in the train data. It demonstrates the following information:
  + Feature importance: Variables are ranked in descending order.
  + Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
  + Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation.
  + Correlation: A high level of the “guests\_included” content has a high and positive impact on the price. The “high” comes from the red color, and the “positive” impact is shown on the X-axis. Similarly, we will say the “minimum\_nights” is negatively correlated with the target variable.
  + Simplified plot: I made the following simplified version for easier interpretation. It highlights the correlations in colors. The red color means a feature is positively correlated with the target variable. The python code is available in the end of the article for readers who are interested.

[przykład uproszczenia/dla innego modelu:]

****

Źródło:

<https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d>

[kod do uproszczenia na samym końcu: *How to Generate the Simplified Version?*]

This shows the Shap values on the x-axis. Here, all the values on the left represent the observations that shift the predicted value in the negative direction while the points on the right contribute to shifting the prediction in a positive direction. All the features are on the left y-axis.

Źródło:

<https://www.analyticsvidhya.com/blog/2019/11/shapley-value-machine-learning-interpretability-game-theory/>