An introduction to Statistical Learning - chapter 2

Lila Gmerek / Week 1 / 29.04.2019

Links

- Book http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf

- Github: https://github.com/erg0-0/ML-study-group-pl-fc

- Link do pliku z podzialem rozdzialow: https://docs.google.com/document/d/10DWwfxazT9RiKX0ZV7xAcYVtZspaalqn8S6lCgzmMdE/edit

Agenda:

1. What is statistical learning? (*covered elsewhere)*

2. Why estimate f?

    - Prediction
    - Interception

3. How do we estimate f? (*covered elsewhere)*

X

te rzeczy znamy, są w naszym datasecie

INPUT VARIABLES / zmienne wejściowe

PREDICTORS / predykator

INDEPENDENT VARIABLES / zmienne niezależne

FEATURES / cechy

VARIABLES / zmienne

Y

te rzeczy znamy tylko dla przeszłości
oczekujemy, że algorytm „powie jak będzie też później”

OUTPUT VARIABLE / zmienne wyjściowe

RESPONSE / odpowiedź

DEPENDENT VARIABLE / zmienne zależne

What is statistical learning?

GENERAL NOTES / prez.Mateusz

Cel analizy: jak poprawić sprzedaż danego konkretnego produktu?

Input data / Zmienne wejściowe (X): Advertising dataset - dane nt. reklam

- [rynek] 200 markets
- [budzet na reklame per rynek] advertising budget for product in each of the market
- [medium reklamowe] 3 different media (X1: telewizja, X2: radio, X3: gazeta)

Output varible / dane=zmiene wyjściowe:

UWAGA! to są te dane które chcemy przewidzieć w przyszłości, dlatego właśnie należą do zmiennych typu OUTPUT:

- [wynik sprzedazy]: sales results of this product

Założenia:

1. Klient nie jest w stanie bezpośrednio podnieść sprzedaży produktu.

2. Klient może zmieniać wydatki/budżet przeznaczany na media reklamowe.

3. Jeżeli ustalimy zależność między wydatkami na konkretne medium reklamowe, a wzrostem lub spadkiem sprzedaży - możemy powiedzieć klientowi jakie kroki podjąć, żeby zwiększyć sprzedaż na danym rynku pośrednio.

GENERAL NOTES
/ prez.Mateusz

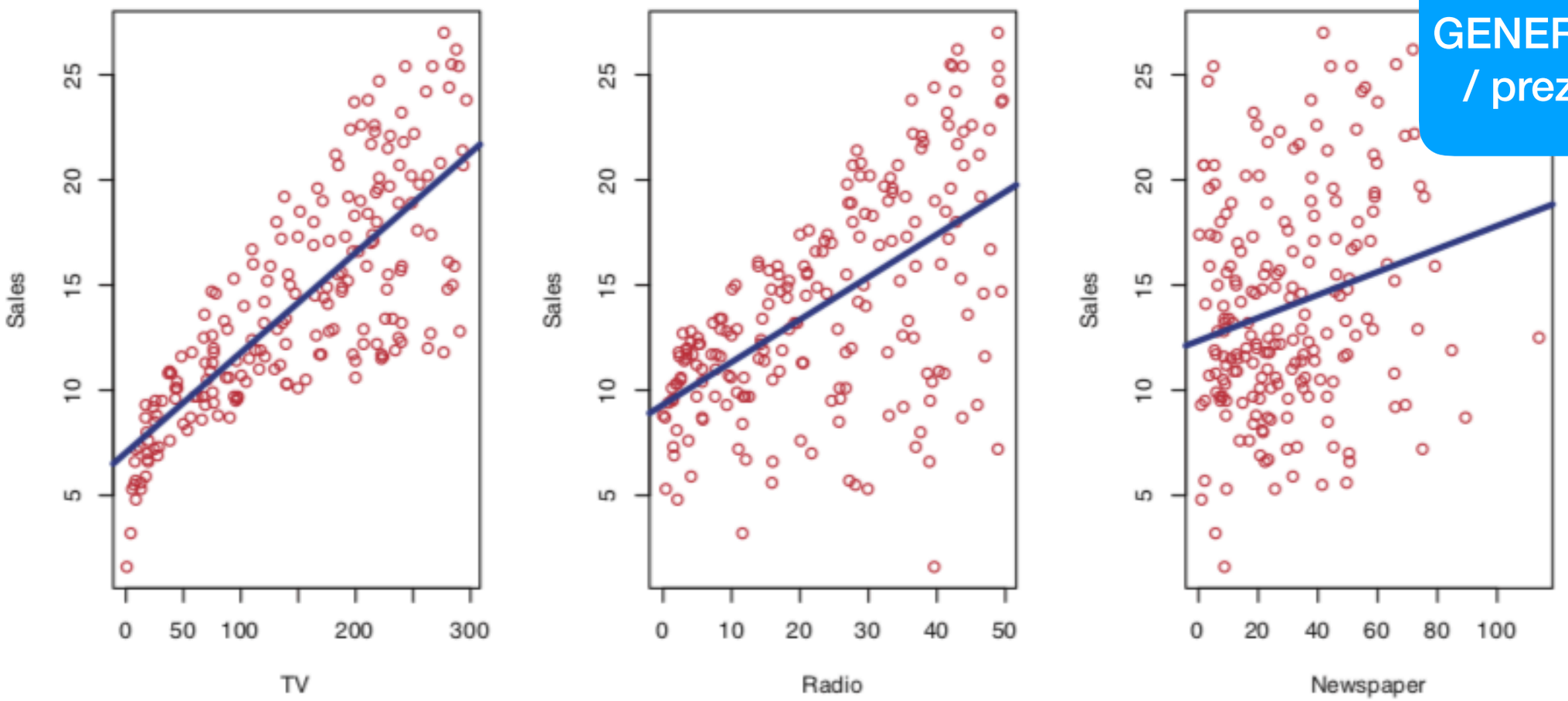

FIGURE 2.1. *The* `Advertising` *data set. The plot displays* `sales`*, in thousands of units, as a function of* `TV`*,* `radio`*, and* `newspaper` *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* `sales` *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* `sales` *using* `TV`*,* `radio`*, and* `newspaper`*, respectively.*

GENERAL NOTES / prez.Mateusz

$$Y = f(X) + \epsilon. \tag{2.1}$$

Here f is some fixed but unknown function of $X_1, \ldots, X_p$, and ϵ is a random *error term*, which is independent of X and has mean zero. In this formulation, f represents the *systematic* information that X provides about Y.

error term

systematic

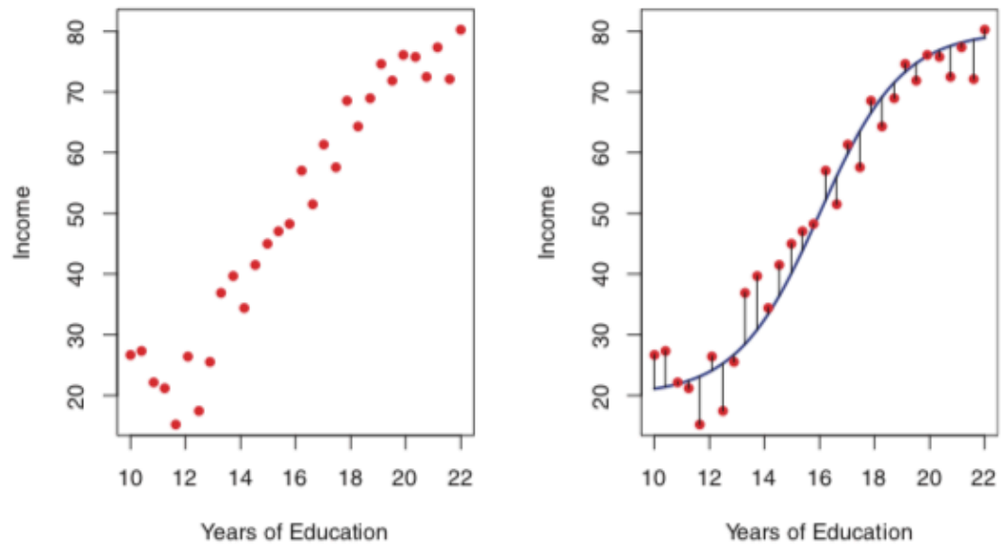

FIGURE 2.2. *The* `Income` *data set.* Left: *The red dots are the observed values of* `income` *(in tens of thousands of dollars) and* `years of education` *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* `income` *and* `years of education`*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

NA POCZĄTKU NIE ZNASZ TEJ FUNKCJI X.
NP. PRZEWIDUJĄC CENE MIESZKAŃ W GDANSKU, NIE WIESZ JAKA JEST ZALEŻNOŚĆ,
PONADTO WE WZORZE ZAWSZE JEST TEZ JAKIS ‚E'(ERROR TERM).
NP. NAJPEWNIEJ NIE UWZGLEDNISZ W MODEU RYZYKA TORNADA, BO PO CO,
ALE GDYBY SIE POJAWILO A CENY BY SPADLY MIESCILOBY SIE TO W RAMACH
TEGO ERROR TERMU :)... "WYOLBRZYMIAJĄC".
W „NORMALNYCH" MODELACH JEST TO JEDNAK NAJCZĘSCIEJ MIARA DOKŁADNOŚCI
MODELU - CZYLI JAK BARDZO PEWNI JESTESMY, ŻE MODEL SIĘ POMYLI.

Ponieważ nie znamy funkcji X,
trzeba ją oszacować na podstawie danych = obserwacji.
Na drugim wykresie jest zasymulowana jako niebieska krzywa.
Linie poziome (odleglosci obserwacji od krzywej) reprezentuja error.

Uwaga - gdyby policzyć średnią ze wszystkich errorów - wyjdzie ZERO!
Czy to znaczy ze model się myli? nie, to znaczy ze trzeba coś policzyć ;-) —> podnieść error term do kwadratu i dopiero wtedy liczyć średnią===> stąd MEAN SQUARED ERROR / mse = popularna miara jakości modeli regresyjnych.

W praktyce nigdy nie trzeba liczyć MSE za pomocą wzoru, ale dobrze go rozumieć. Python ma wbudowaną funkcję, która liczy to automatycznie.

Why estimate f?

Dlaczego zależy mi na tym, żeby dobrze oszacować funkcję f?

2 powody:

PREDICTION - przewidywanie

INTERFERENCE - WNIOSKOWANIE

Prediction

$$\hat{Y} = \hat{f}(X),$$

^ oznacza wartość nieznaną
- predykowaną

Najczęściej mamy jakieś dane na początek (X), ale nie wiadomo jaka jest predykcja (Y z daszkiem) oraz nie wiadomo w jaki sposób ją uzyskać (f z daszkiem).
W przypadku problemów natury predykcyjnej, nie jest to jednak szczególnie istotne dla klienta jak wygląda ta funkcja (tzw. black box), ponieważ interesuje go ostateczny rezultat.

- przykład: koncern farmaceutyczny (lub po prostu laborant) chce się dowiedzieć na podstawie próbki krwi czy u pacjenta pojawi się ciężka reakcja alergiczna na lek, nad którym pracuje. (X1, X2, X3… Xn to cechy badania krwi; Y to zmienna kodujaca ryzyko wystąpienia alergii. To w jaki sposób przebiega funkcja f z punktu widzenia widzenia laboranta/pacjenta ma drugorzędne znaczenie. W przypadku pacjentów, u których ryzyko jest wysokie (zmienna zależna zgodnie z predykcją naszego modelu Y koduje wysokie prawdopodobieństwo wystąpienia alergii na lek) jest to bardzo istotne - lek nie zostanie podany, a oni unikną cierpienia!

Prediction

$$\hat{Y} = \hat{f}(X),$$

^ oznacza wartość nieznaną
- predykowaną

Od czego zależy dokładność predykcji Y ^?
Oczywiście predykcje nigdy nie są idealne. Nie da się przewidzieć wszystkiego ze 100% pewnością. To czego nie da się przewidzieć okresla się za pomocą „error” - miary błędu.

Kluczowe są dwa elementy:

1. **reducible error (redukowalny błąd) -** to błąd naszego modelu predykcyjnego (czyli f^, funkcji wybranej przez nas do zamodelowania przyszłych wartości Y), który można zredukować za pomocą różnych metod statystycznych. W przypadku przykładu z pacjentem np. model osiąga dokładność (*accuracy*) na poziomie 0.75 (czyli 75%). W przypadku badań medycznych to naprawdę nie wystarczająco, i może to kosztować czyjeś życie. Chcąc osiągnąć accuracy modelu na poziomie 99,8%, będziemy redukować ten błąd poprzez wybór lepszych parametrów modelu, albo wręcz wybór innego modelu lub innych metod statystycznych, które pozwolą na znaczne podniesienie *accuracy* (dokładności) naszej predykcji.

2. irreducible error (nieredukowalny błąd) - zła wiadomość jest taka - nigdy nie uda nam się wyeliminować wszystkich błędów w predykcji! Nie ważne jak dobra będzie moja predykcja (czyli Y z ^), i tak będzie zawierała w sobie błąd nieredukowalny! Wartość e błędu jest zawsze większa od zera. Dlaczego? Na przykładzie pacjenta - predykcja Y może mieć accuracy na poziomie 99,9%, wskazując na podstawie próbki krwi, że pacjent nie będzie miał reakcji alergicznej. Okazuje się jednak, że pacjent ma tę reakcję! Jak to się stało? Mógł przykładowo brać inne leki, które wchodziły w interferencje z nowym lekiem - ta informacja była istotna z punktu widzenia predykcji. Ponieważ danych wejściowych X, nie było żadnych pomiarów dotyczących pozostałych leków, **te dane nie zostały zmierzone, ani uwzględnione w predykcji.** Nie wiedząc o istnieniu tej zależności, nie mogłam zredukować tego błędu w mojej predykcji. Inne przykłady błędu tego rodzaju? Lek wywołał alergię z powodu błędu w jego produkcji i złej dawki; błędu pielęgniarki która podała za dużo leku; złego samopoczucia pacjenta itd. Ten błąd to samo życie.

Prediction

$$\hat{Y} = \hat{f}(X),$$

^ oznacza wartość nieznaną
- predykowaną

Kluczowe są dwa elementy:

1. **reducible error (redukowalny błąd)**
2. **irreducible error (nieredukowalny błąd)**

Zależność matematyczna można wyrazić następująco:

Lewa strona równania - średnia (*expected value,* wartość oczekiwana) z podniesionej do kwadratu różnicy między prawdziwa i przewidywaną wartością Y
równa się wariancji związanej z błędem nieredukowalnym

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}, \end{aligned}$$

Statistical learning skupia się na technikach oszacowania f oraz zminimalizowania błędu redukowalnego. Niemniej, należy mieć na uwace, że błąd nieredukowalny zawsze będzie stanowił górną granicę dokładności (accuracy) predycji dla Y. W praktyce ta granica zawsze pozostaje nieznana.

Inference - wnioskowanie

W niektórych sytuacjach predykcja nie ma znaczenia. Przedmiotem zainteresowania klienta, jest zrozumienie zachodzących zmian oraz wyciągnięcie z nich wniosków dla swoich dalszych działań.

- przykład: reklamy w które medium (TV, radio, gazeta) przyczyniają się do większej sprzedaży produktu? Czy różne rynki różnią się pod tym kątem i należałoby zastosować inne strategie? Czy istnieje w ogóle zależność między wydatkami na reklamę w tych mediach czy nie? A może nie ma zależności i nie warto wydawać na to pieniędzy?

Większe znaczenie ma ZROZUMIENIE ZALEŻNOŚCI, czyli funkcji f^, a nie Y.

$$\hat{Y} = \hat{f}(X),$$

Pytania, które można postawić w obliczu takiego problemu:

- która zmienna niezależna ma wpływ na zmienną zależną? (*identyfikacja najważniejszych dla predykcji featerów)*
- *jaka jest zależność między predykatorami (zm.niezależnymi) a odpowiedzią (zm.zależnymi)? (np. korelacja?)*
- *Czy zależność f między Y a predykatorami X1, X2,… Xn może być wyrażona za pomocą funkcji liniowej, czy jest to zależność bardziej skomplikowana?*

Inference - wnioskowanie

$$\hat{Y} = \hat{f}(X),$$

Większe znaczenie ma ZROZUMIENIE ZALEŻNOŚCI, czyli funkcji f^, a nie Y.

Przykłady, które pojawią się w tej książce:

- [PREDYKCJA] klient to firma, która chce przeprowadzić kampanię marketingową bezpośrednią. Chce ustalić, którzy klienci odpowiedzą pozytywnie na mailing, bazując na danych demograficznych dostępnych na ich temat w bazie danych. (predykatory X: dane demograficzne klientów; odpowiedź Y: odpowiedź pozytywna/negatywna na mailing=zakup produktu) .
- [WNIOSKOWANIE] firma chce zbadać swoje kanały medialne, w których zamieszcza reklamy. Stawia pytania takie jak: reklamy w których mediach wpływają na sprzedaż? Reklamy w których mediach przyczyniają się do największego podniesienia wyników w sprzedaży? Jak wysoki wzrost w sprzedaży jest związanych z jak wysokim wzrostem finansowania reklam w telewizji?
- [WNIOSKOWANIE] firma chce zamodelować markę produktu , który klient może chcieć zakupić, bazując na predykatorach (X) takich jak cena, lokalizacja sklepu, cena u konkurencji itd. Pytanie, które stawia to: jak każda z tych zmiennych wpływa na prawdopodobieństwo zakupu? Jaki wpływ na sprzedaż może mieć zmiana ceny produktu?
- [WNIOSKOWANIE] nieruchomości - jaka wpływają na wartość domów czynniki takie jak wskaźnik przestępstw, *zoning* (tzw. tworzenie się gett itp.), odległość od rzeki, jakość powietrza, szkoły, średni dochód danej mieszkańców danej dzielnicy, wielkość domów itp. Jak każda z tych zmiennych wpływa na ceny, czyli o ile więcej dom będzie wart jeśli będzie mieć widok na rzekę?
- [PREDYKCJA] nieruchomosci - jaka będzie wartość domu, któr ma dane charakterystyki? Czy ten dom jest wyceniony zbyt tanio czy zbyt drogo?

Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f may be appro- priate. For example, linear models allow for relatively simple and inter- pretable inference, but may not yield as accurate predictions as some other approaches. In contrast, some of the highly non-linear approaches that we discuss in the later chapters of this book can potentially provide quite accu- rate predictions for Y , but this comes at the expense of a less interpretable model for which inference is more challenging.

The Trade-Off Between Prediction Accuracy and Model Interpretability

Nie można mieć wszystkiego.
Albo model jest łatwy w interpretacji, ale bardzo restrykcyjny;
albo jest trudny w interpretacji, ale dużo bardziej elastyczny.

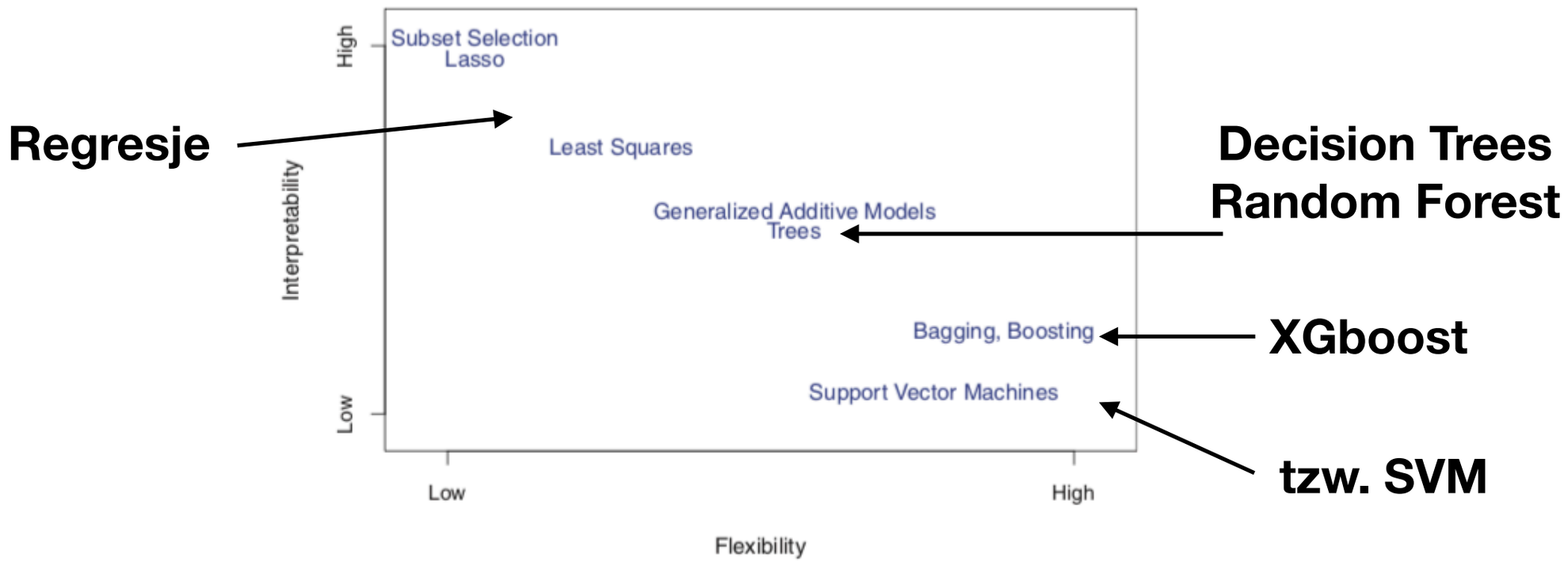

FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

The Trade-Off Between Prediction Accuracy and Model Interpretability

Po co mi w ogóle model nieelastyczny (restryktywny)?
Dobrze sprawdzają się w przypadku problemów opartych na wnioskowaniu lub badaniu zależności różnych zmiennych X na Y. Łatwa w interpretacji regresja pomaga szybko to ocenić. Po co zatem komplikować sobie życie i tracić czas na skomplikowane modele, zwłaszcza, że ich rezultaty mogą być dużo bardzeij niejednoznaczne.

Przy okazji, łatwiej wytłumaczyć taki łatwiejszy model szefowi lub klientowi, którzy za te nasze statystyki płacą - co nierzadko przemawia za ich zastosowaniem.

Czy zawsze bardziej elastyczne modele wpływają na lepszą predykcję?

Istnieje pewien paradoks - czasami lepiej zastosować bardziej restrykcyjny model aby uzyskać lepszą predykcję. Okazuje się, że bardziej skomplikowane modele (boosting czy nawet te popularne sieci neuronowe i cały deep learning) dużo częściej cierpią na problem z tzw. overfittingiem (czyli przetrenowaniem modelu). W rezultacie, fantastycznie radzą sobie na danych treningowych, ale dużo gorzej na danych na których są testowane.