czechitas

5th Lecture: Regression

Justina Ivanauskaite, Josef Svec, Martin Korytak

16.5.2023

Your team today

Justina

Data Science Lead,

MSD Animal Health

Josef

Data Scientist in

Workday

Martin

Data Scientist in

Workday

Today's structure

1. Linear Regression
2. Decision Tree Regression
3. Random Forest Regression
4. How variables are contributing to prediction?
5. Performance evaluation
6. Regression methods - summary

REGRESSION TASKS - INTRODUCTION

Regression vs Classification

Making predictions

Regression - predicts continuous variable

Classification - predicts categorical or discrete variable

Regression

Aim: Predict continuous numerical values.

Regression

Aim: Predict continuous numerical values.

Algorithms: Linear Regression, Polynomial Regression, Support Vector Regression, Decision Trees, Random Forests, Neural Networks, etc.

Regression

Aim: Predict continuous numerical values.

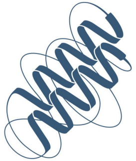
Algorithms: Linear Regression, Polynomial Regression, Support Vector Regression, Decision Trees, Random Forests, Neural Networks, etc.

Evaluation Metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared, etc.

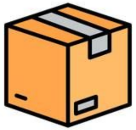
Regression Use Cases

Predicting **house selling prices** based on location, n. of bedrooms, ..

Predicting **flight delays** based on location, airline, weather conditions, ..



Predicting **amount of synthesized protein** based on temperature, pH, feeding of cells,...

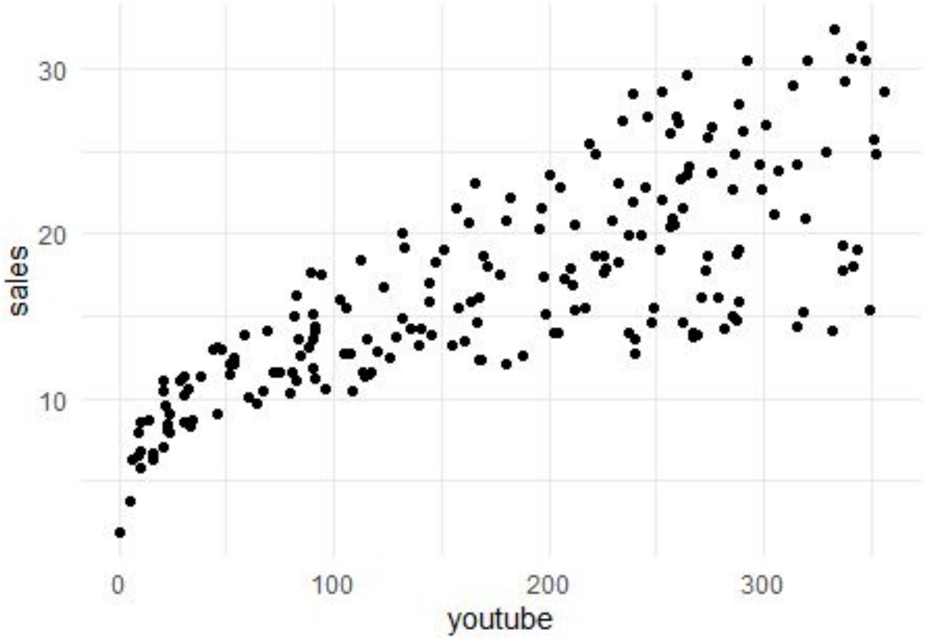


Predicting **number of sold units** based on location, material, colors, ..

LINEAR REGRESSION

Linear Regression

youtube	sales
276.12	26.52
53.40	12.48
20.64	11.16
181.80	22.20
216.96	15.48
10.44	8.64
69.00	14.16
144.24	15.84
10.32	5.76

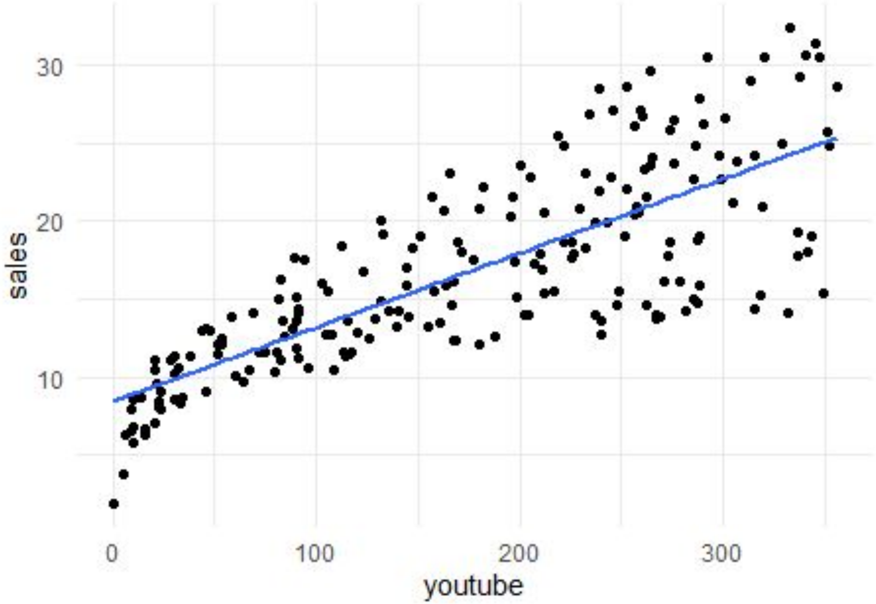


czechitas

Linear Regression

youtube	sales
276.12	26.52
53.40	12.48
20.64	11.16
181.80	22.20
216.96	15.48
10.44	8.64
69.00	14.16
144.24	15.84
10.32	5.76
239.76	?
79.32	?
257.64	?

$$y = \beta_0 + \beta_1 x + \varepsilon$$

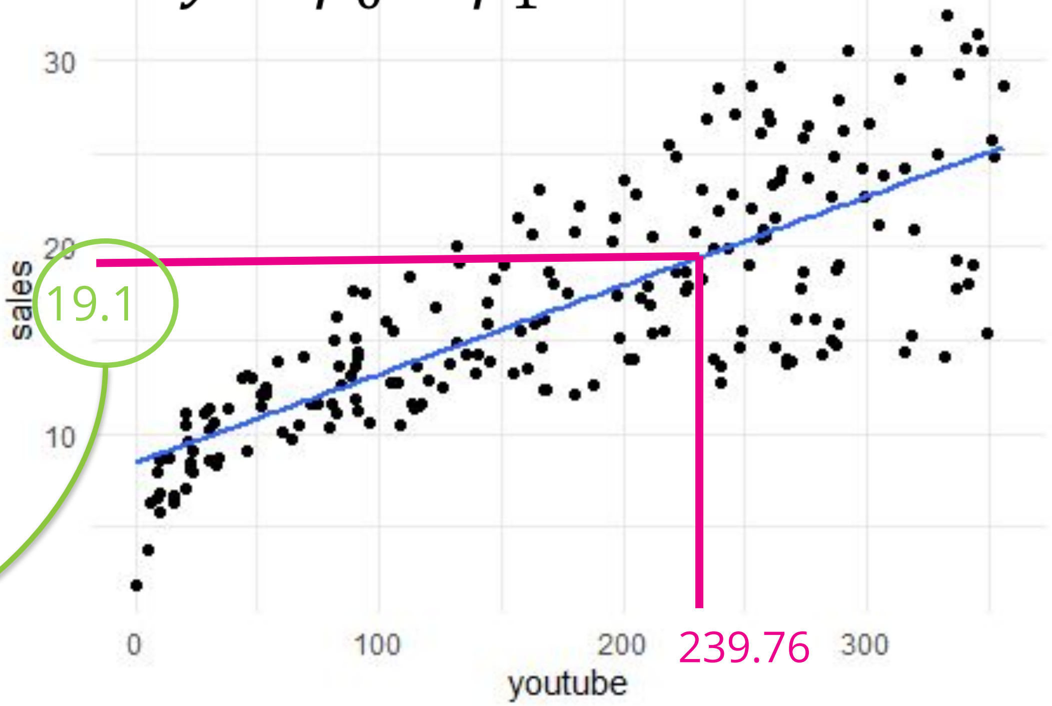


czechitas

Linear Regression

youtube	sales
276.12	26.52
53.40	12.48
20.64	11.16
181.80	22.20
216.96	15.48
10.44	8.64
69.00	14.16
144.24	15.84
10.32	5.76
239.76	?
79.32	?
257.64	?

$$y = \beta_0 + \beta_1 x + \varepsilon$$



Splitting data into training and testing

Split data into train test

Use majority of data for training the model and minority for evaluation of performance

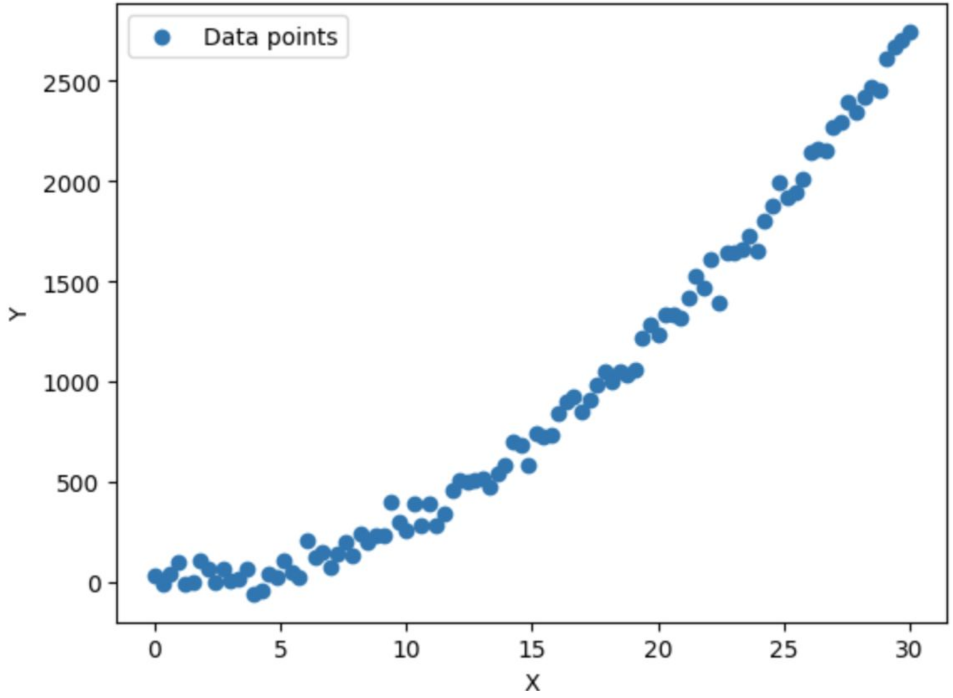
Can the model generalize?

Evaluate model on unseen data

Rule of thumb - 70%/30% or 80%/20%

Polynomial regression

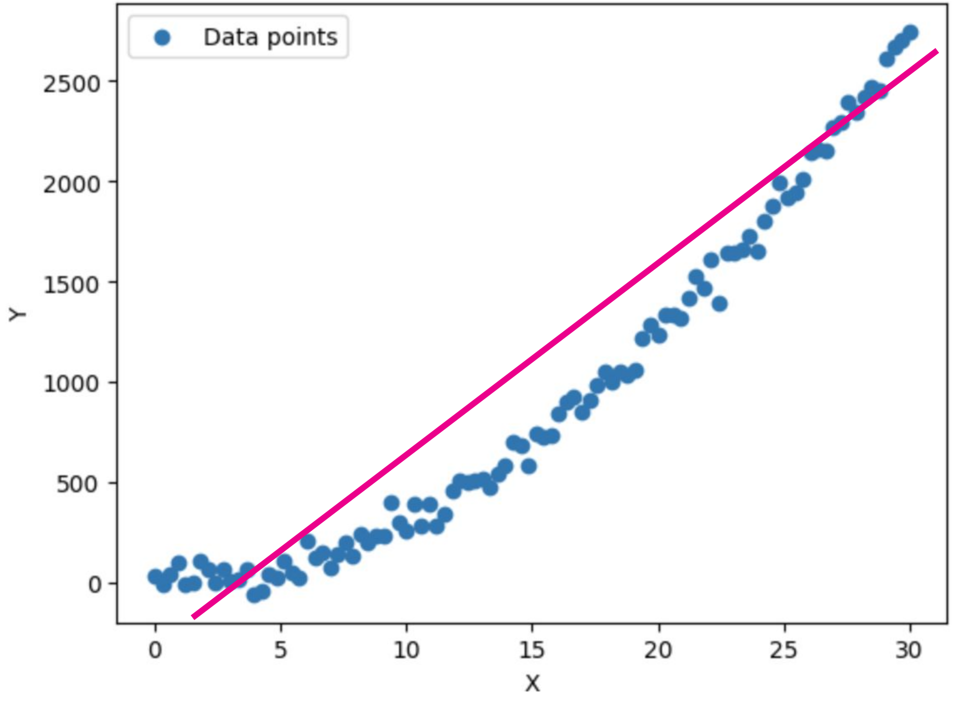
Can you draw straight line through this data?



czechitas

Can you draw straight line through this data?

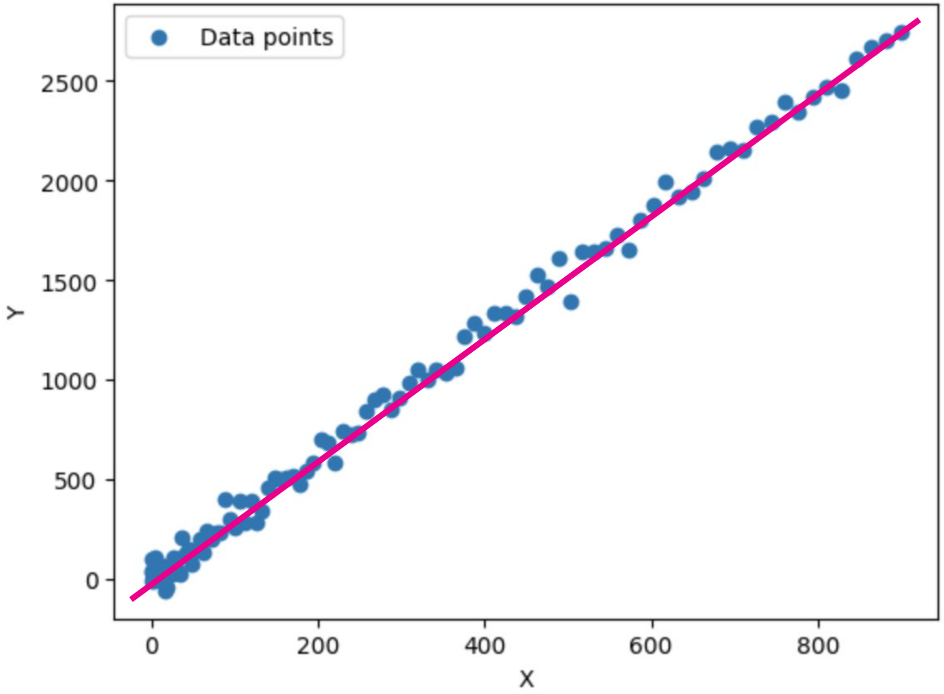
y = b1x +b0



czechitas

Now you can

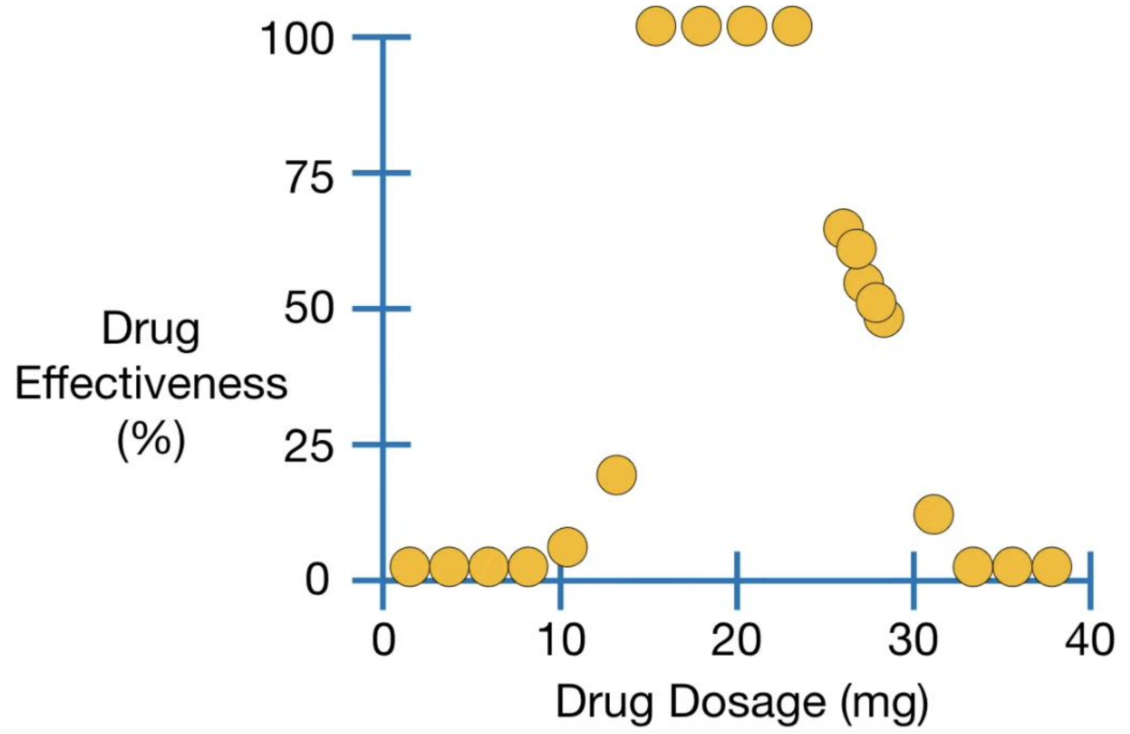
$y = b1x^2 + b0$



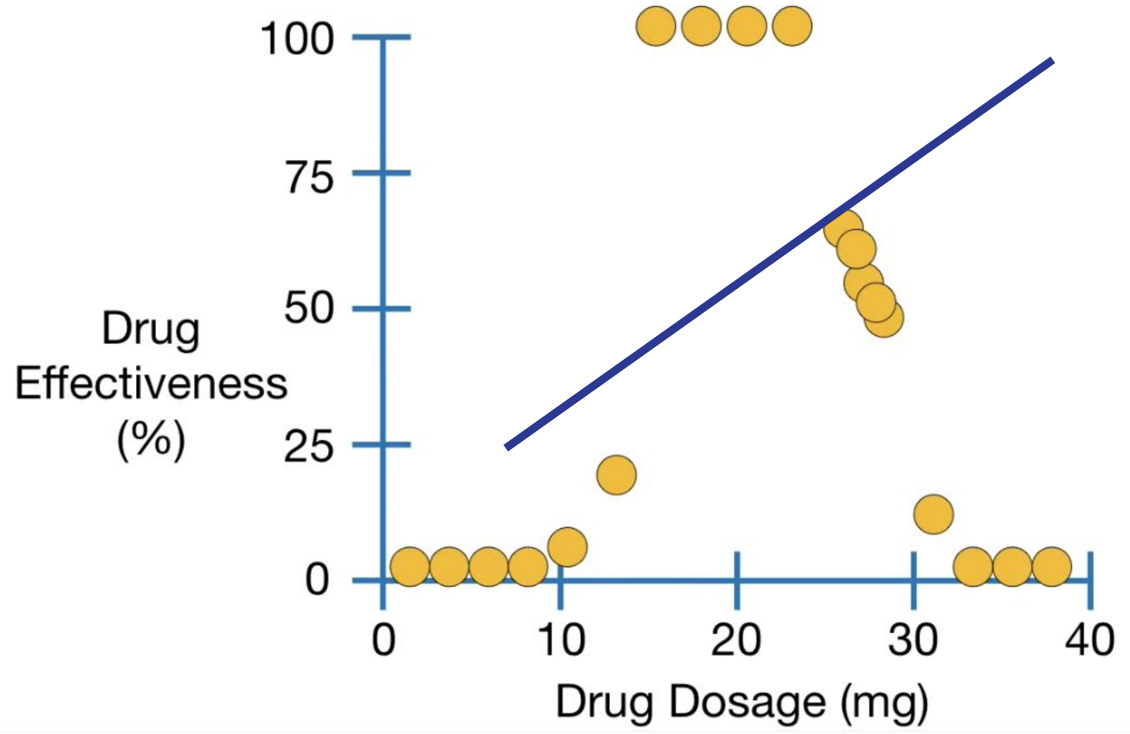
czechitas

DECISION TREE REGRESSION

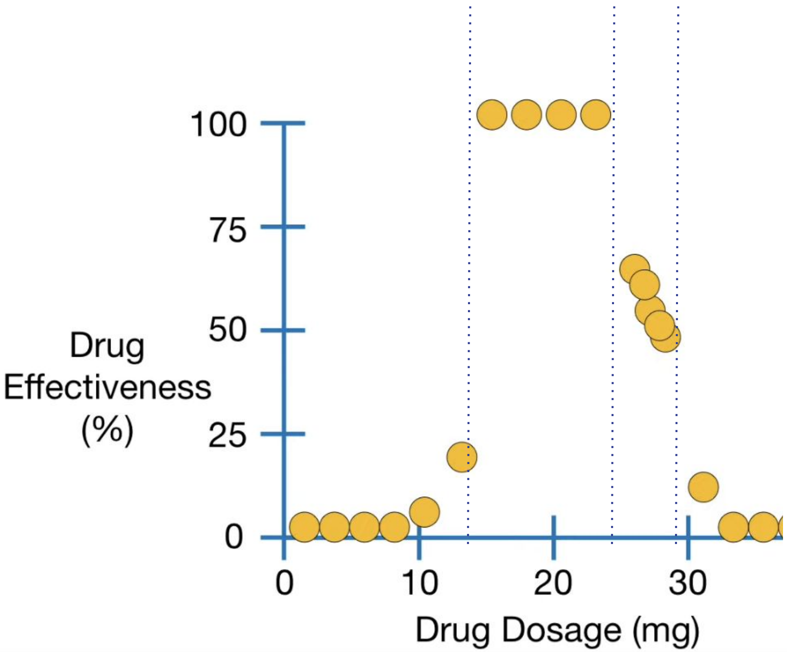
Non-linear relationship

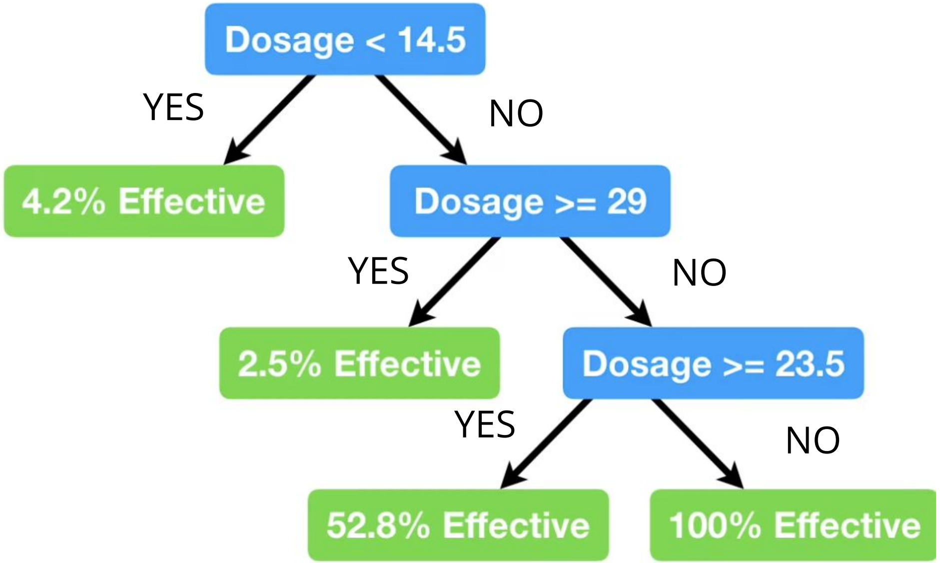


Non-linear relationship



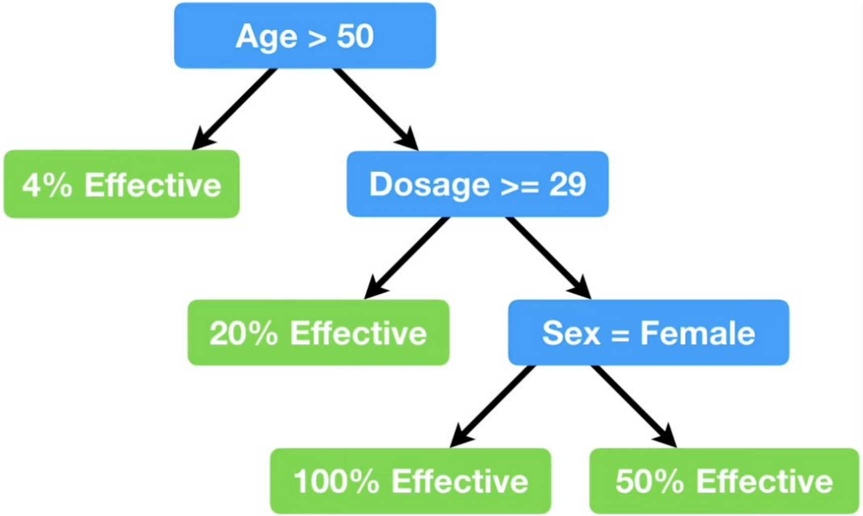
Regression tree example





czechitas

Regression tree example - multiple variables



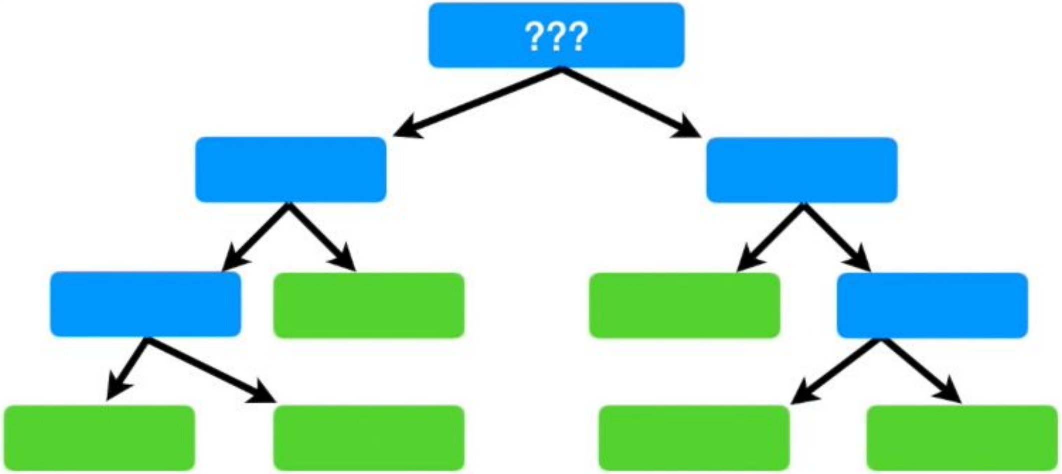
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

czechitas

Building a tree - predicting effectiveness by dosage

Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc...	etc...

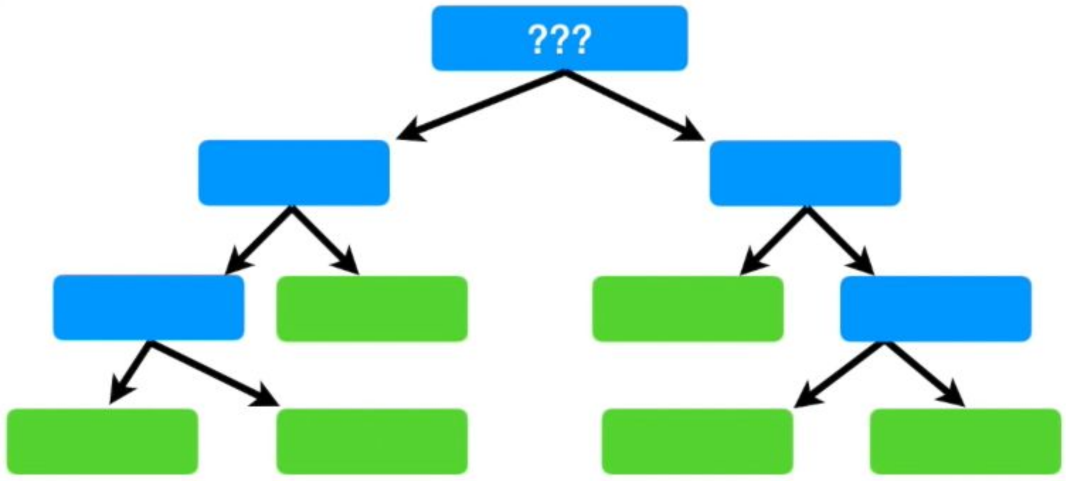
What condition do we start with?



Building a tree - predicting effectiveness by dosage

What condition do we start with?

We try all possible thresholds, and **see which threshold gives us the lowest prediction error.**



Building a tree - predicting effectiveness by dosage

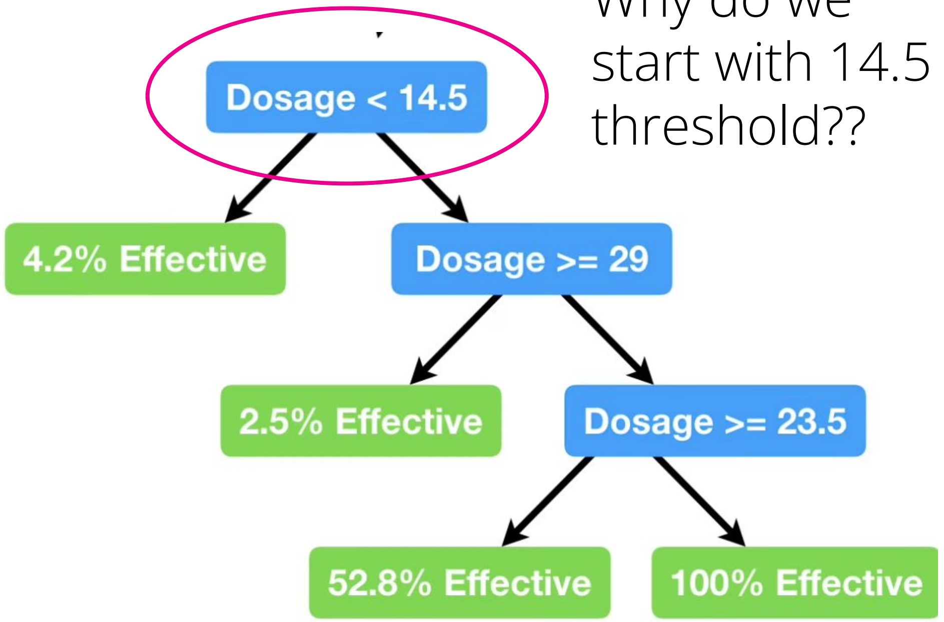
What condition do we start with?

We try all possible thresholds, and **see which threshold gives us the lowest prediction error.**

Error = Predicted value - Observed value

Building a tree - predicting effectiveness by dosage

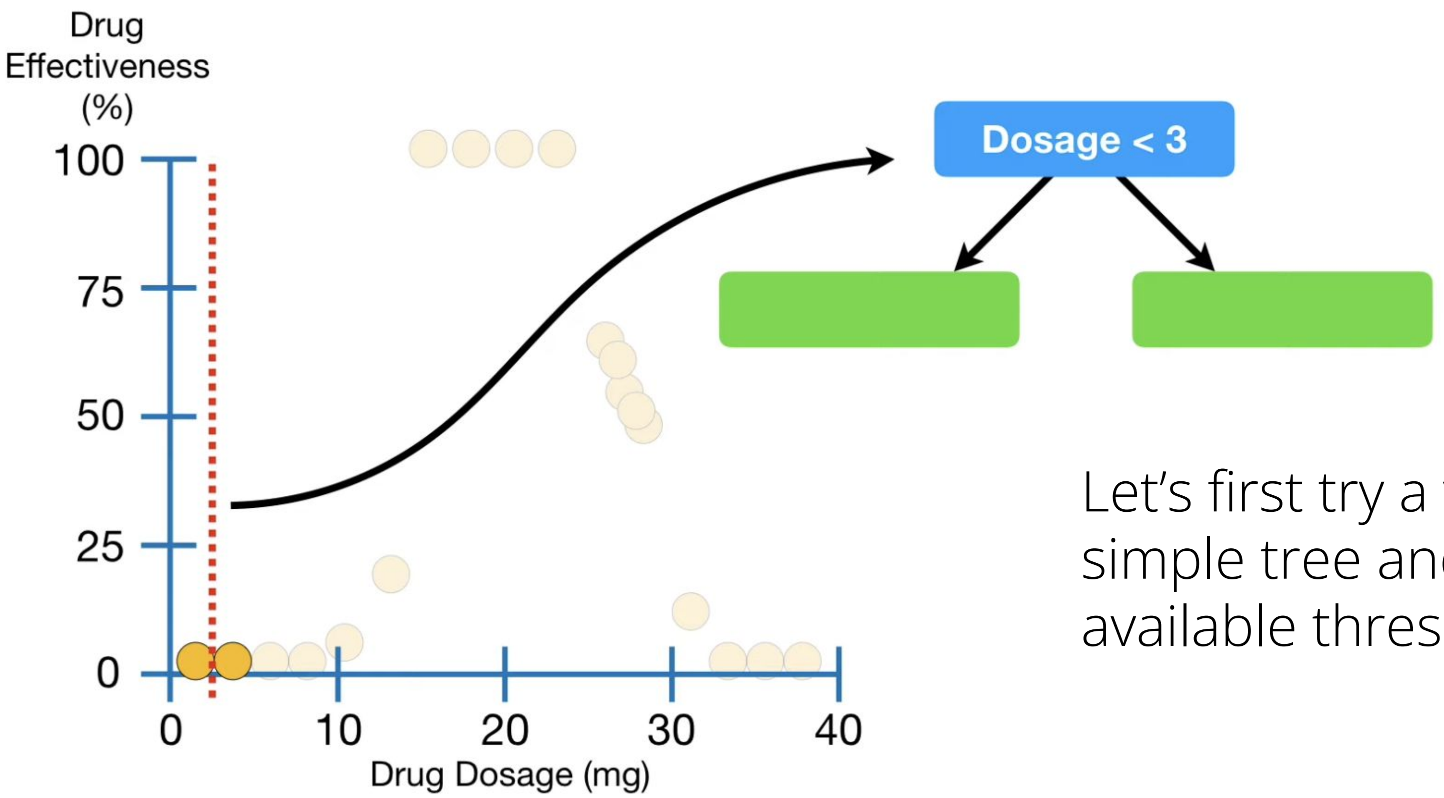
Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…



Why do we start with 14.5 threshold??

czechitas

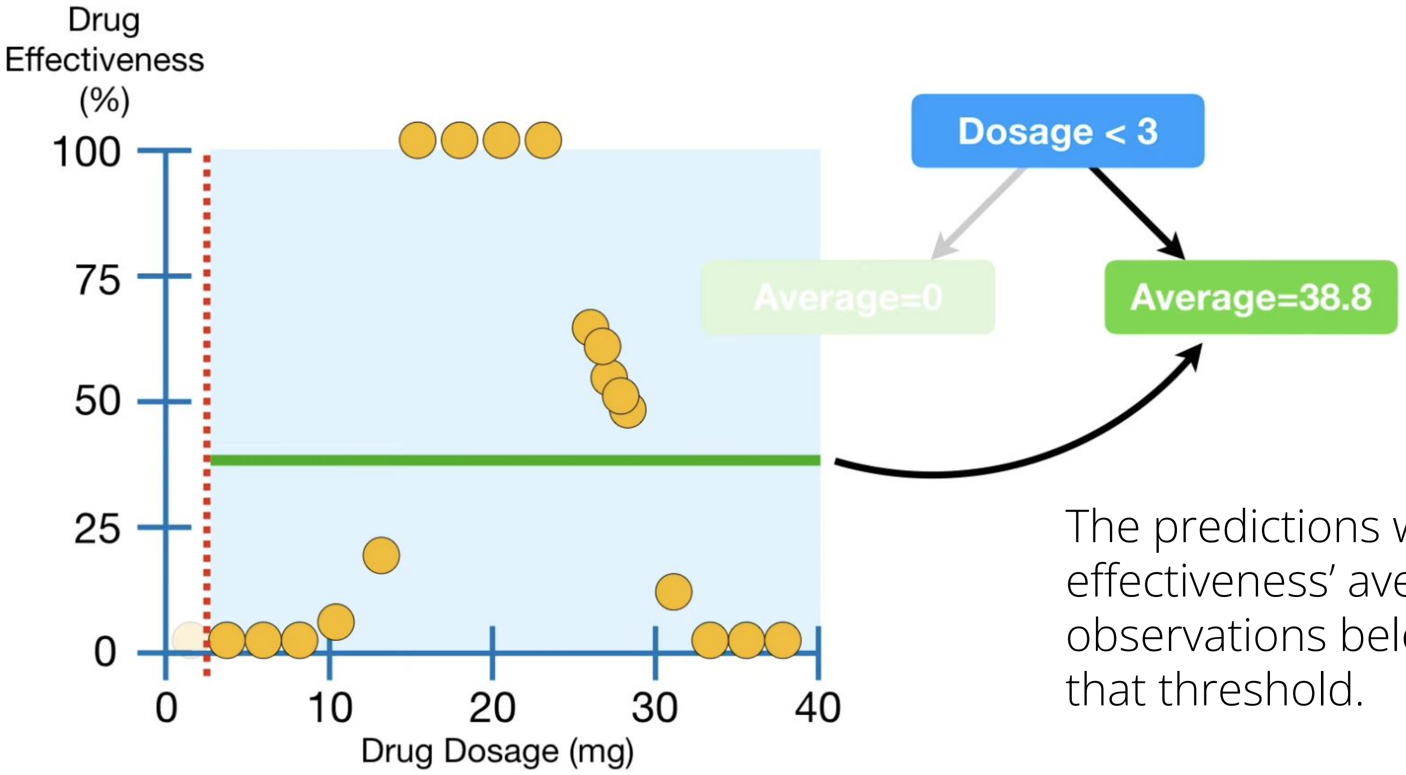
Building a tree - predicting effectiveness by dosage



Let's first try a very simple tree and the first available threshold.

czechitas

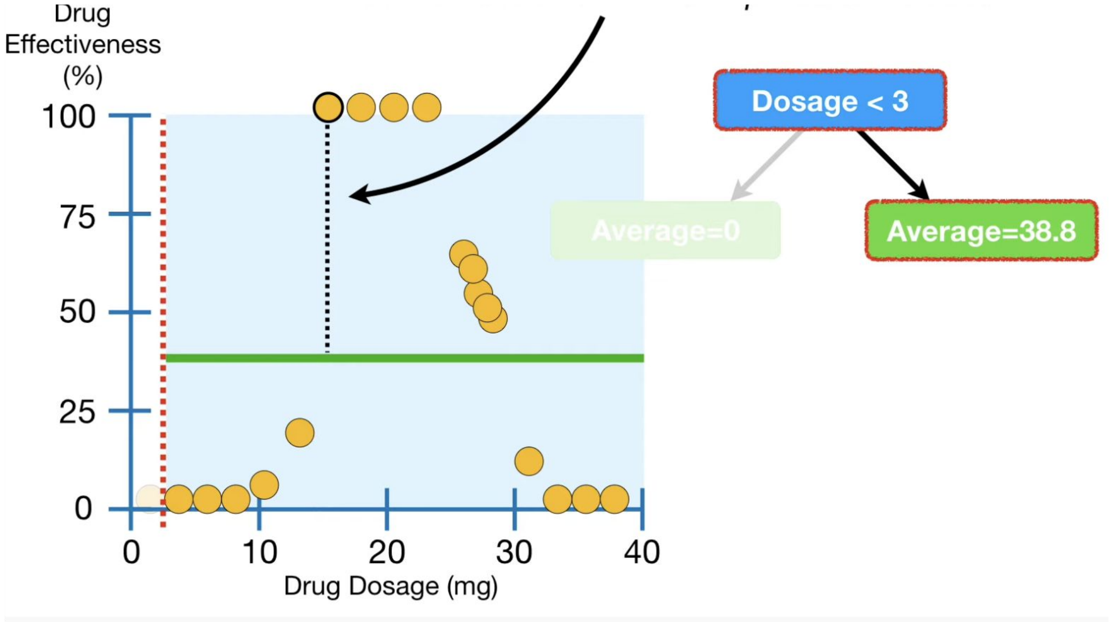
Building a tree - predicting effectiveness by dosage



The predictions will be just the effectiveness' averages of the observations below and above that threshold.

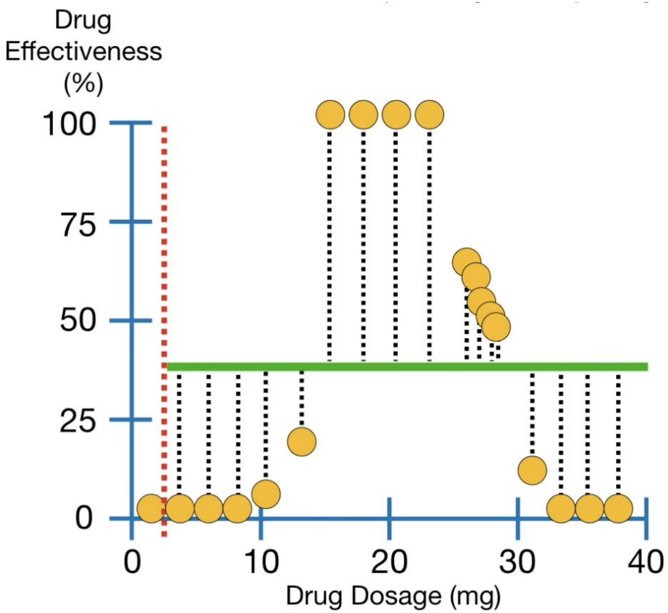
Building a tree - predicting effectiveness by dosage

Then we will look at the difference between prediction (the average) and the observed value - **the residual** (the error).



czechitas

Building a tree - predicting effectiveness by dosage



We want to calculate the sum of all the residuals.

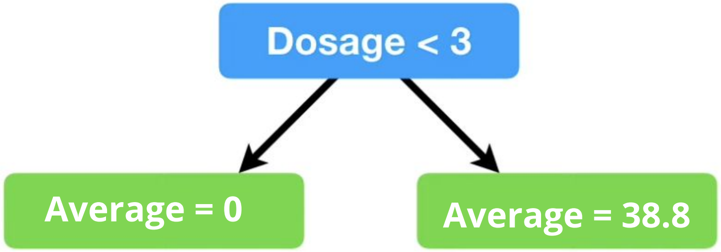
$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$
$$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (100 - 38.8)^2$$
$$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$$

= 27 469

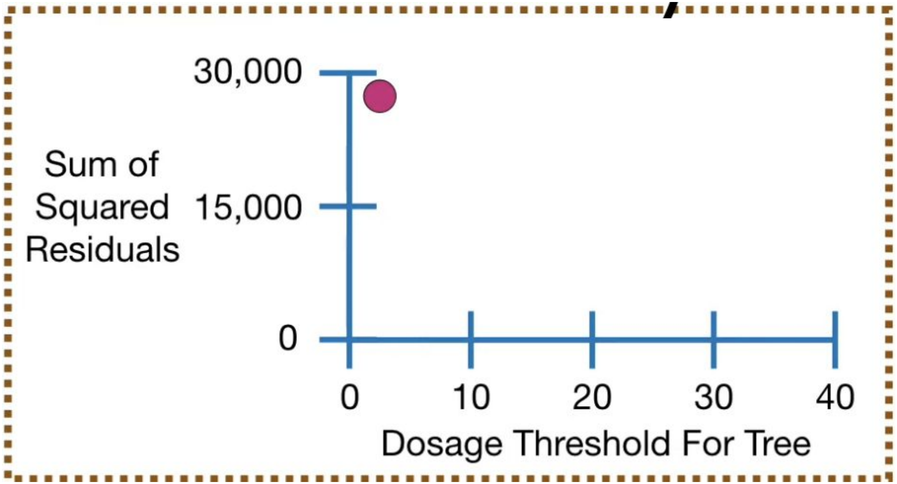
czechitas

Building a tree - predicting effectiveness by dosage

We can note down the threshold we used and the sum of squared residuals into the plot below.



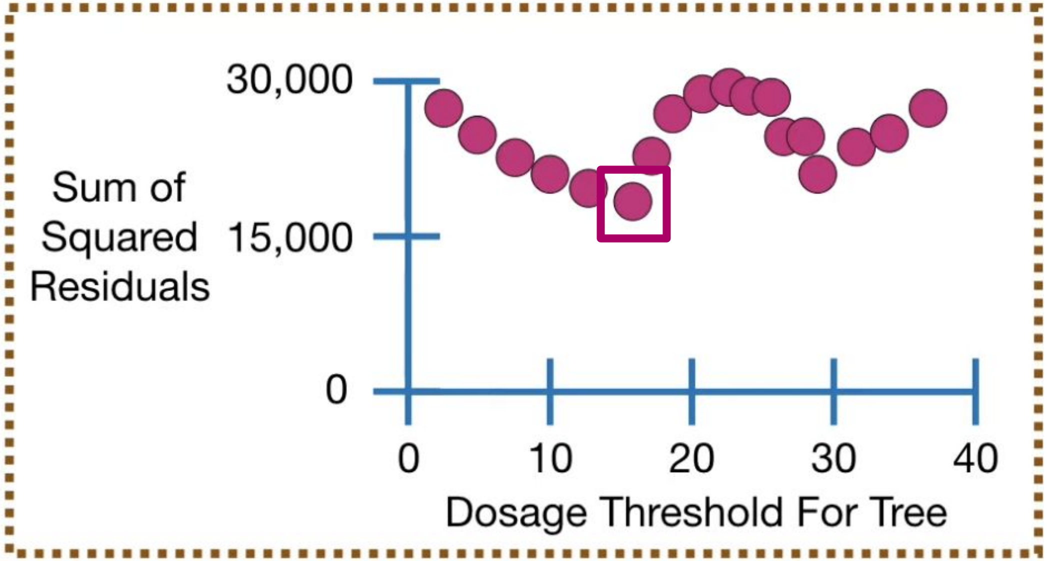
Sum of Squared Residuals = 27 469



czechitas

Building a tree - predicting effectiveness by dosage

The same way, we can try to use many different thresholds and always calculate the sum of squared residuals.

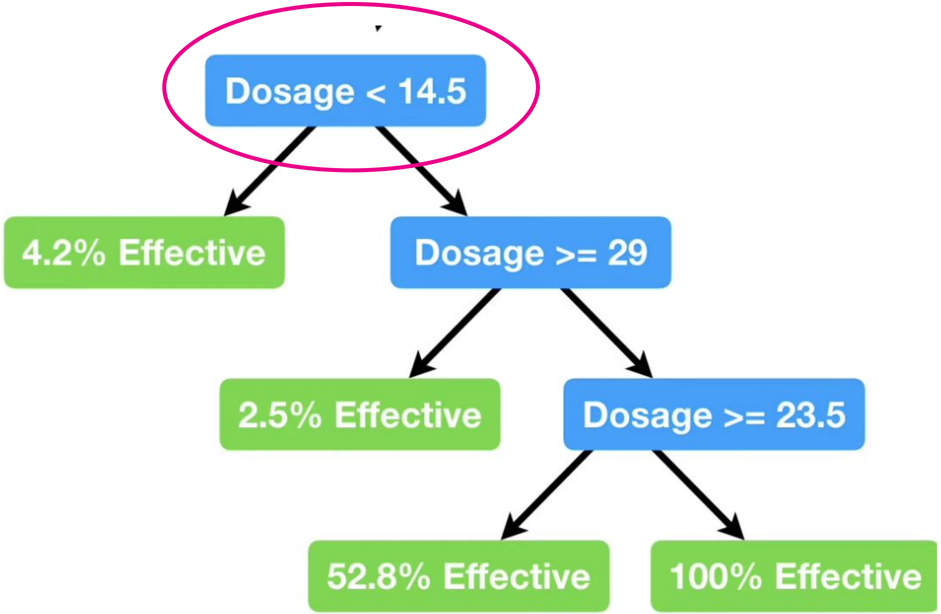


The dosage of **14.5** had the smallest sum of squared residuals.

Building a tree - predicting effectiveness by dosage

The dosage of **14.5** had the smallest sum of squared residuals.

→ This is the first condition in our tree.

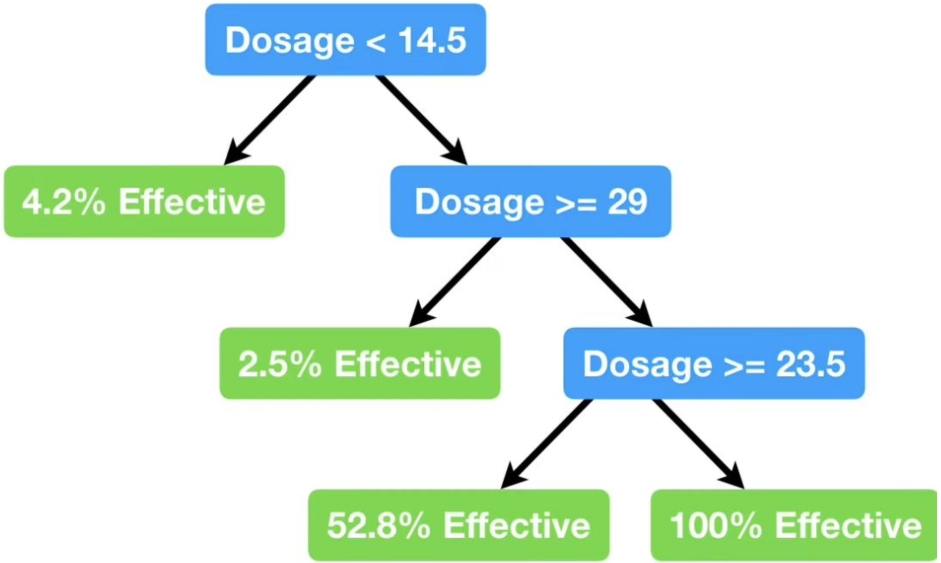


czechitas

Building a tree - predicting effectiveness by dosage

Then we continue testing other thresholds to get further conditions.

However, we should set some minimum number of observations in the leaf node.



czechitas

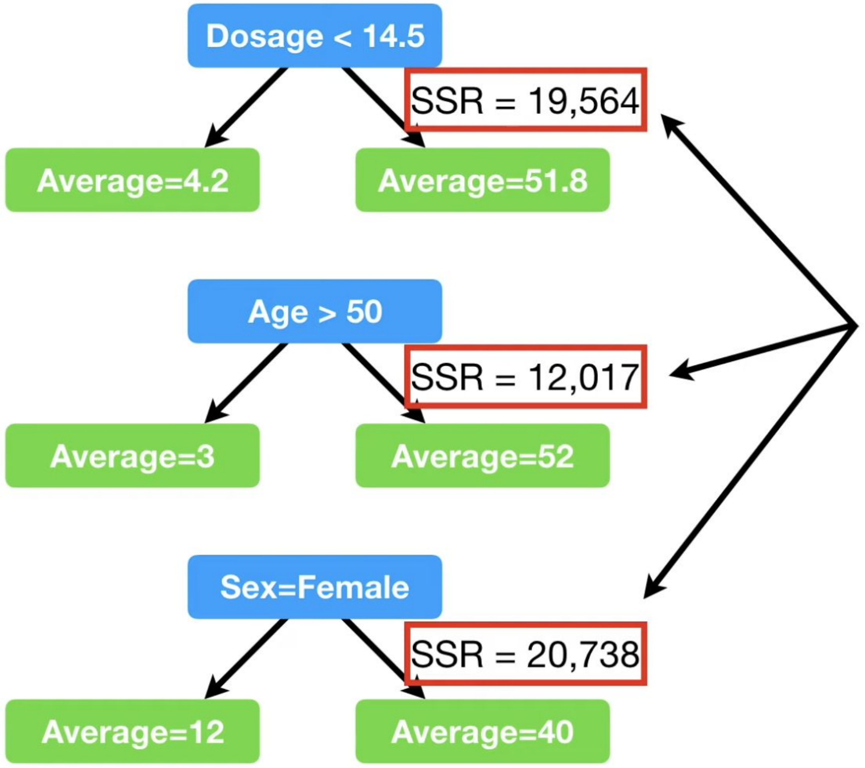
Predicting effectiveness using multiple variables

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc...	etc...	etc...	etc...

Usually we have more than one explanatory variables.

In that case, we calculate **sum of squared residuals for every threshold of every variable**, and we choose the one with the lowest value.

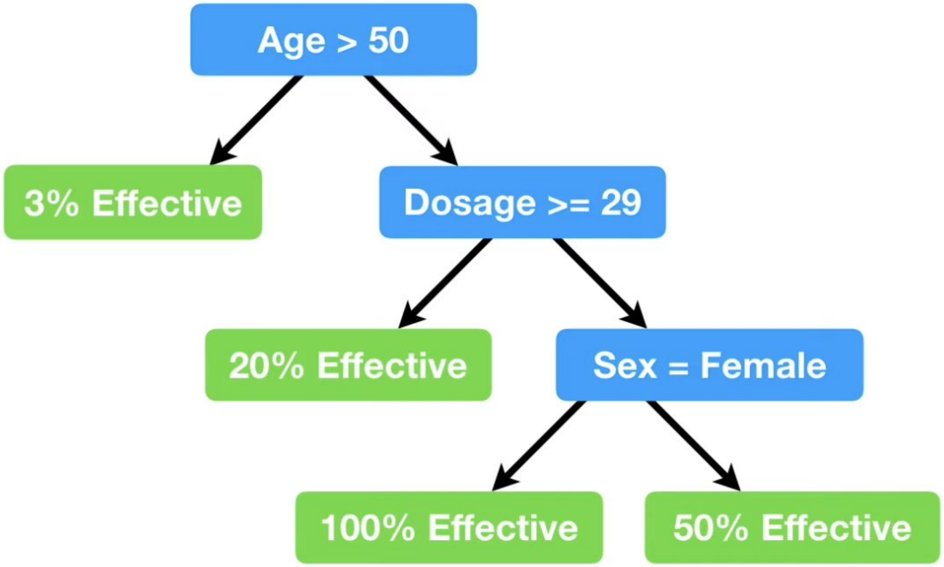
Predicting effectiveness using multiple variables



We select the candidate with the lowest SSR as the root node of our tree:

Age > 50

Predicting effectiveness using multiple variables



We select the candidate with the lowest SSR as the root node of our tree:

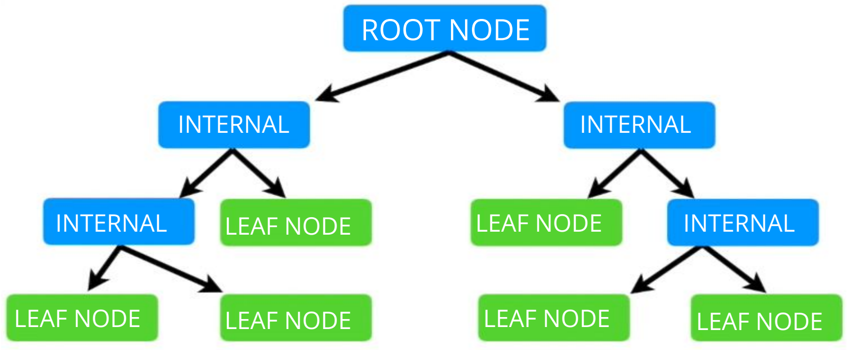
Age > 50

Then we continue the same way, and we may end up with a tree like this.

czechitas

Regression Tree - Summary

Regression tree follows a set of if-else conditions to predict a continuous variable.



Specific conditions are chosen based on the lowest sum of squared residuals.

Decision Tree: Pros and Cons

Advantages

- **Interpretable and easy to understand:** Decision trees provide a transparent and intuitive representation of the decision-making process.
- **Handling both numerical and categorical data:** Decision trees can handle both numerical and categorical features without requiring extensive data preprocessing.
- Feature importance estimation
- Robustness to outliers and missing data

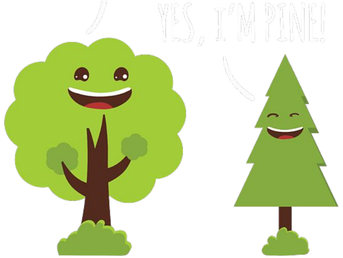
czechitas

Decision Tree: Pros and Cons

Disadvantages

- **Overfitting:** Decision trees are prone to overfitting, particularly when the tree becomes too deep or complex.
- **Biased towards features with high cardinality:** Decision trees tend to favor features with high cardinality (many unique values) because they can potentially provide more splits and finer partitions.
- **Data imbalance:** They may prioritize the majority class and struggle to accurately predict the minority class.
- **High variance:** They can produce different trees and predictions when trained on different subsets of the data.

Quiz 1 Which statement is false?



1. When decision tree is trained, all thresholds per each explanatory variable are tested as a potential condition.

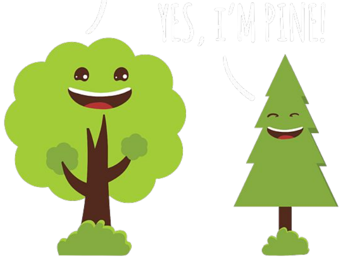
2. Decision tree minimizes the sum of squared residuals when it selects the decision conditions.

3. Decision tree maximizes the sum of squared residuals when it selects the decision conditions.

4. It is good to control the minimum number of observations in the leaf node when specifying the model.

Quiz 1

Which statement is false?



1. When decision tree is trained, all thresholds per each explanatory variable are tested as a potential condition.

2. Decision tree minimizes the sum of squared residuals when it selects the decision conditions.

3. **Decision tree maximizes the sum of squared residuals when it selects the decision conditions.**

4. It is good to control the minimum number of observations in the leaf node when specifying the model.

czechitas

RANDOM FOREST REGRESSION

Regression Tree: The starting algorithm

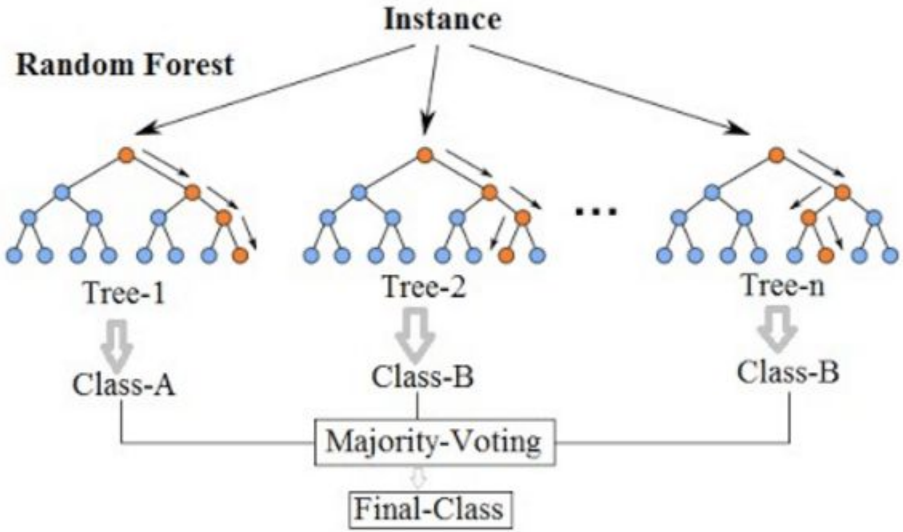
- Decision tree is a simple algorithm, it has **low bias** but **high variance**
- Usually, a large number of decision trees is combined to reduce the variance
- Combining trees is known as an **‘ensemble method’**
- One of the most widely used ensemble methods is Random Forest

What is Random Forest

- A widely used algorithm, an **ensemble method** that builds **multiple decision trees**
- The forest can use for example 50, 100, 200, 500, ... trees → depends on the dataset size
- Can be used **both for classification and regression**

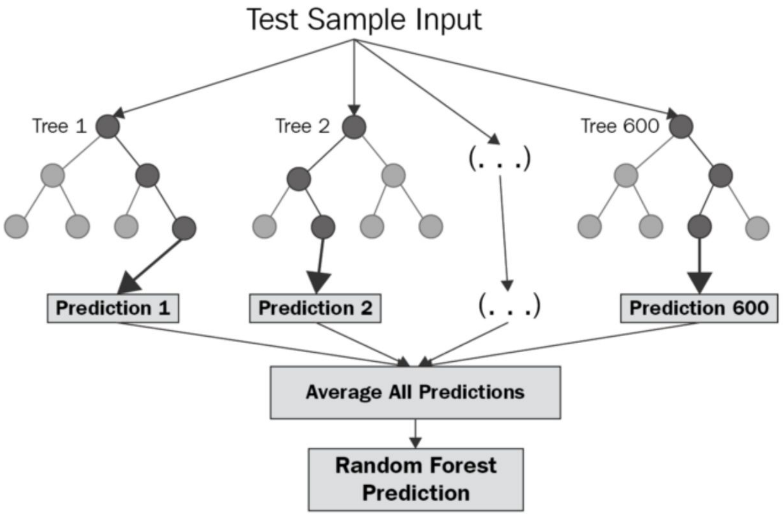
What is Random Forest

In classification, the prediction is **the majority vote of all decision trees' predictions.**



What is Random Forest

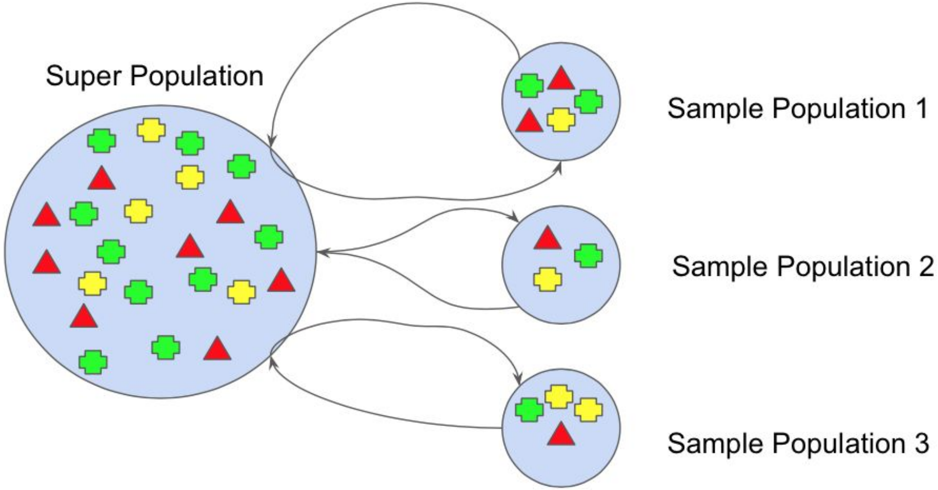
In regression, the prediction is **the average of all decision trees' predictions.**



czechitas

Bootstrapping

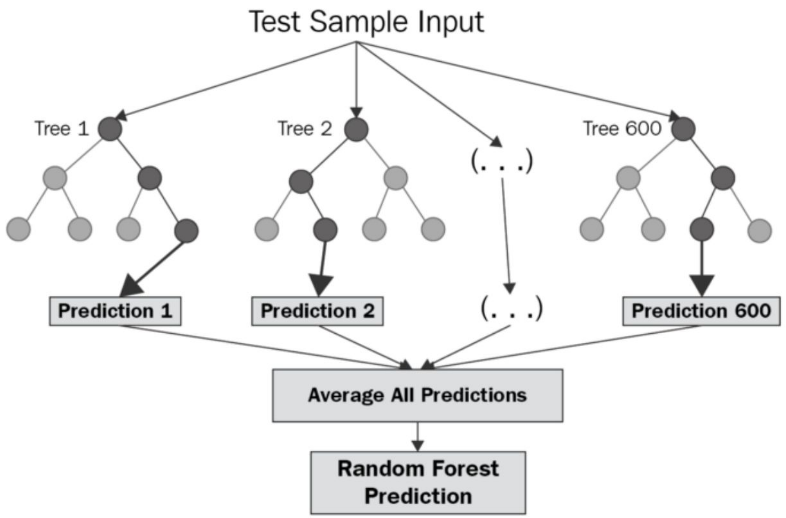
Bootstrapping is a resampling technique used in statistics and machine learning to create **multiple datasets with replacement** from a given dataset.



czechitas

Random Forest - How it works

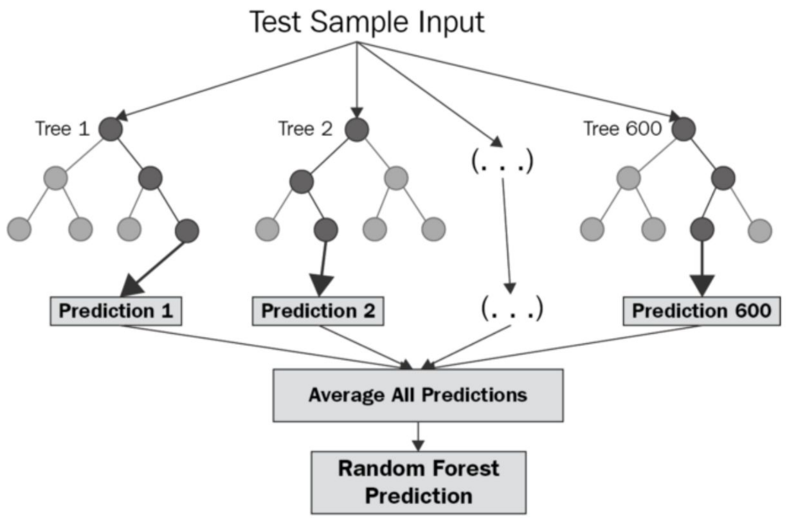
- Each tree is build based on a random sample from training data.



Random Forest - How it works

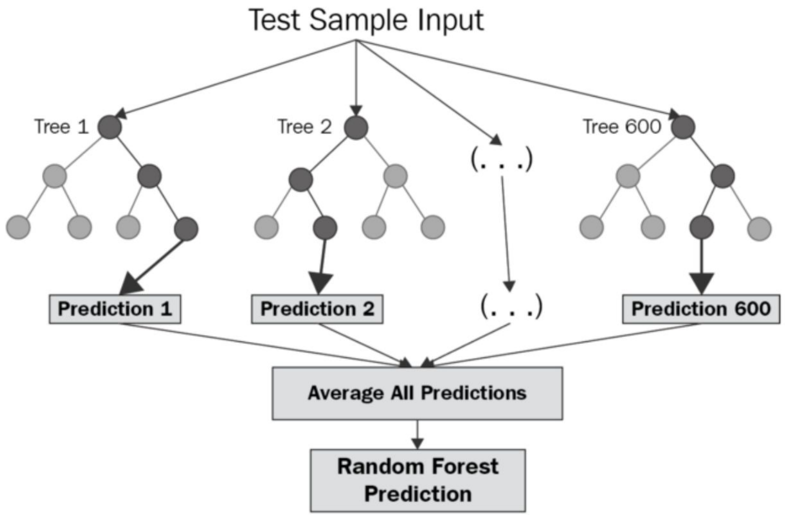
- Each tree is build based on a **random sample from training data**.

- The algorithm **randomly selects a subset of explanatory variables** for each split in each decision tree.



Random Forest - How it works

- Each tree is build based on a **random sample from training data**.
- The algorithm **randomly selects a subset of explanatory variables** for each split in each decision tree.
- **Trees' predictions are averaged** to get the final output.



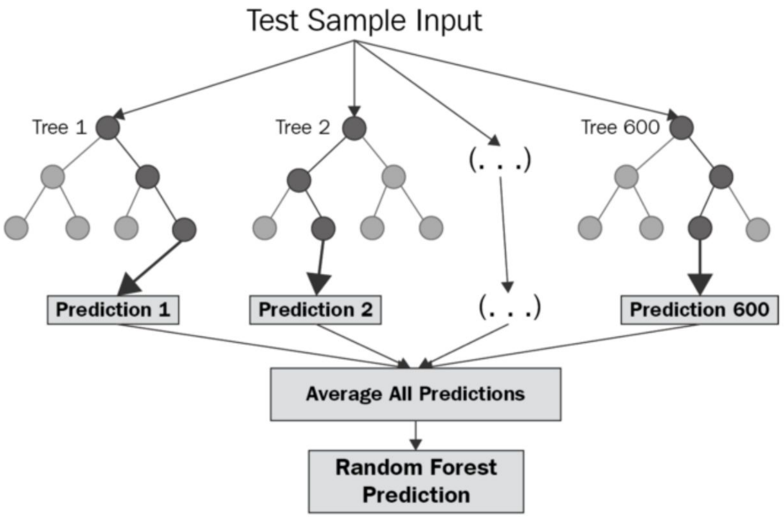
czechitas

Random Forest - Hyperparameters tuning

a) Specify the **maximum depth** of the trees

a) Increase or decrease the **number of trees**

a) Specify the **maximum number of features** to be included at each node split



czechitas

Random Forest: Pros and Cons

Advantages

- **One of the most accurate** learning algorithms available
- **Efficient** with large datasets
- It **can handle thousands of input variables** without variable deletion
- Gives us **variable importance**
- **Maintains accuracy** when a large proportion of the data is **missing**

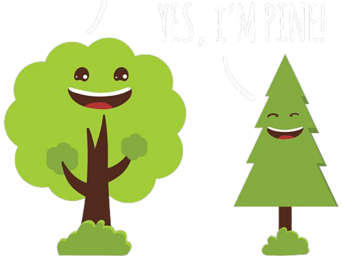
Random Forest: Pros and Cons

Disadvantages

- May not get good results for **small data or low-dimensional data** (data with few features) - the randomness becomes greatly reduced
- Black box model

Quiz 2

Which statement is false?



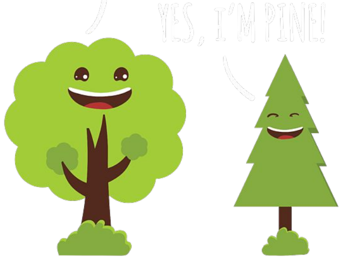
1. All explanatory variables are used in each decision tree in the random forest.

2. Random forest has a lower variance than a single decision tree.

3. Number of trees in a random forest can be as high as 1000, if we have enough data.

Quiz 2

Which statement is false?



1. **All explanatory variables are used in each decision tree in the random forest.**

2. Random forest has a lower variance than a single decision tree.

3. Number of trees in a random forest can be as high as 1000, if we have enough data.

Bagging vs Boosting

Bagging (Bootstrap Aggregating)

- Bagging involves creating multiple base models, each trained on a random subset of the training data, obtained through **bootstrapping** (sampling with replacement).
- The base models are typically trained **independently** and **in parallel**
- During prediction, the final prediction is made by **aggregating the predictions** of all the base models, such as by taking the average (for regression) or majority voting (for classification).
- Bagging helps **reduce variance** and overfitting by creating diverse base models that have different sources of randomness.

Bagging vs Boosting

Boosting

- Boosting involves creating an ensemble of base models **sequentially**, where each subsequent model is trained to correct the mistakes made by the previous models.
- During prediction, the final prediction is made by **combining the predictions** of all the base models, with each model's contribution weighted based on its performance.
- Boosting helps **reduce bias** and improve the overall performance by iteratively refining the ensemble to focus on the challenging instances in the dataset.

Decoding black boxes:
Partial dependence plots & Shapley values

How variables are contributing to final prediction in different algorithms: linear methods

Linear regression ?

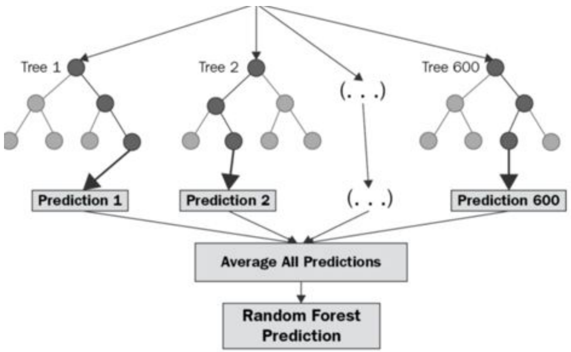
Straightforward – just look to the estimated formula!!!

$$crime\,rate = \mathbf{29.4} + \mathbf{2.86}\ unemployment\,rate + \varepsilon$$

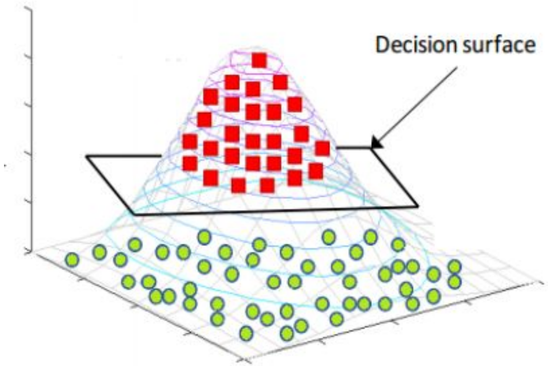
And often linear models are frowned upon because of being too simple - linear!

How variables are contributing to final prediction in different algorithms: non-linear methods

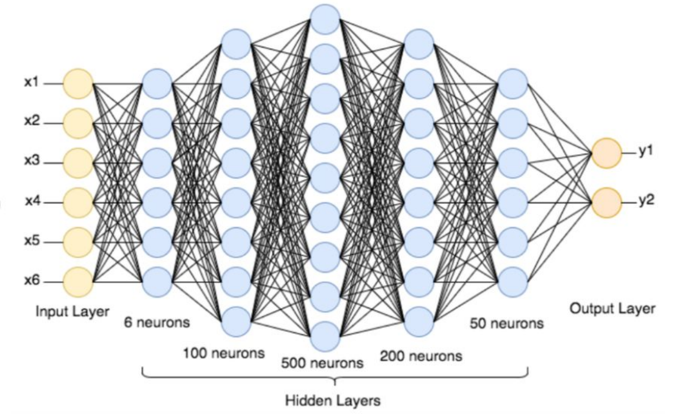
Random forests



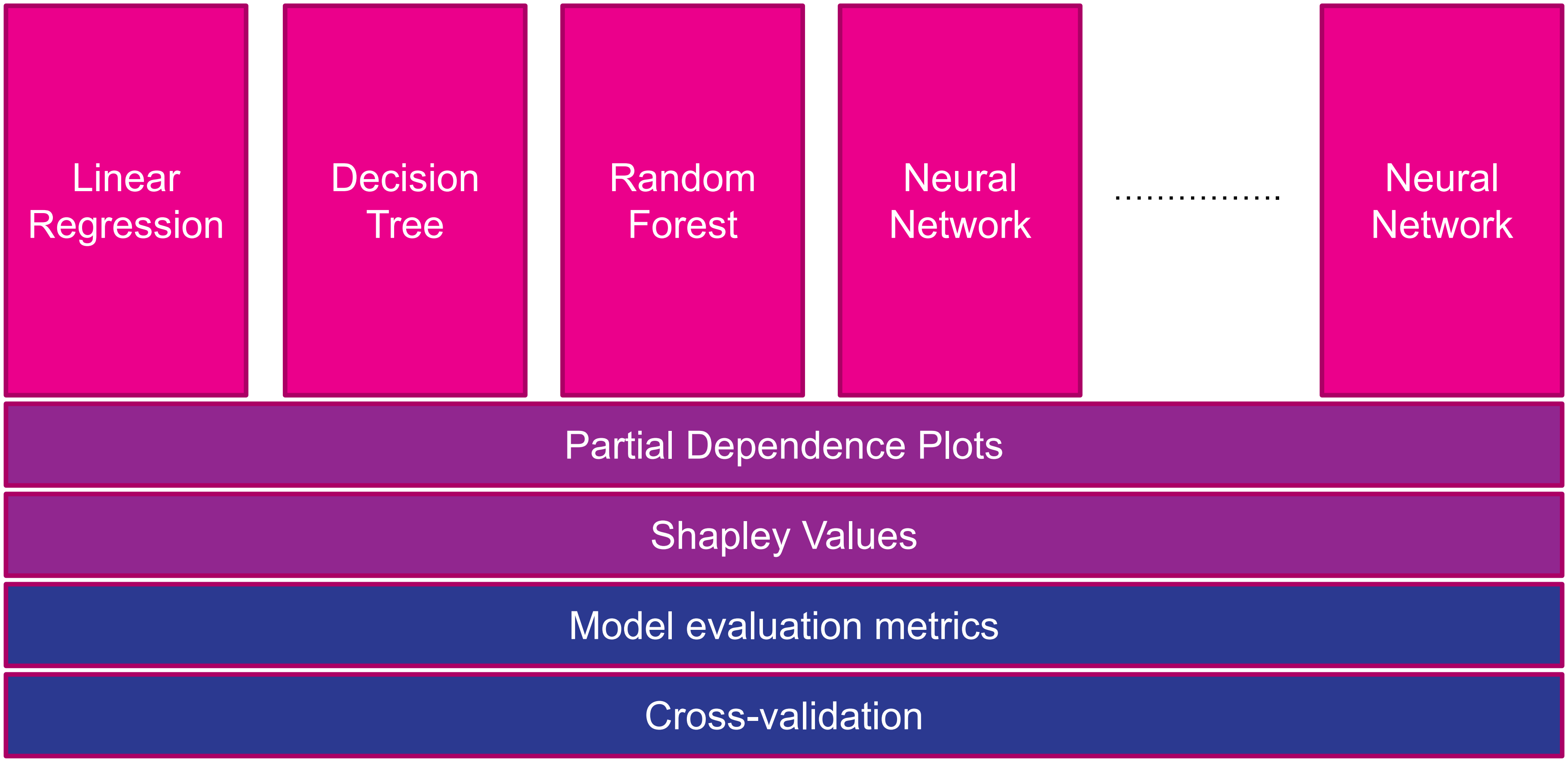
Support vector machine



Neural networks



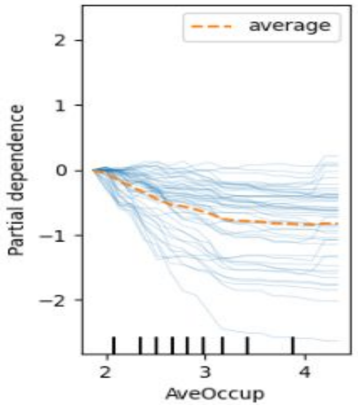
Conseptual intro

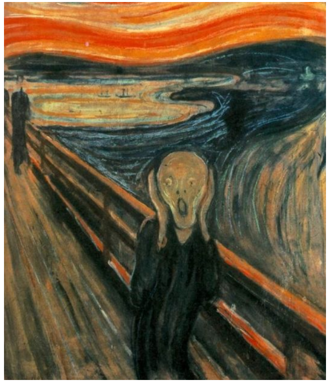


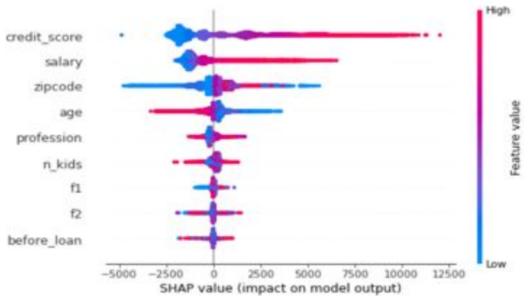
Model agnostic ways to decode black box: Partial dependence plots, Shapley values

Partial dependence plot shows the marginal effect one or two features have on the predicted outcome of a machine learning model

Shapley value is the average marginal contribution of a feature value across all possible coalitions







Describes relationship between model inputs and outputs

High level steps

- Understand your dataset
  - What data means
  - How data looks like (plot)
- Build some model (of your choise)
- Apply functions to plot partial dependence plots
- Analyze

Example dataset Boston housing prices

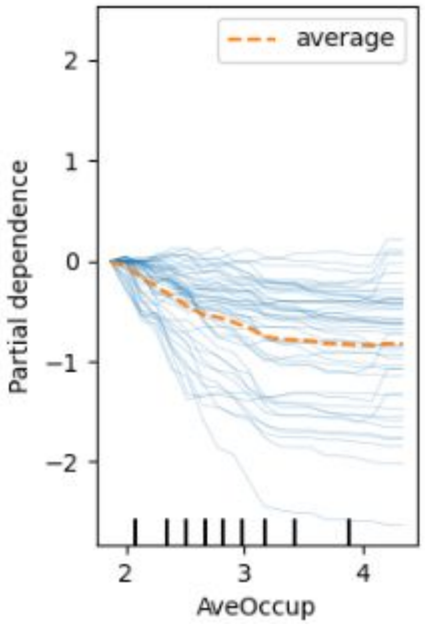
lon	lat	cmedv	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat
-70.9550	42.2550	24.0	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98
-70.9500	42.2875	21.6	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14
-70.9360	42.2830	34.7	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

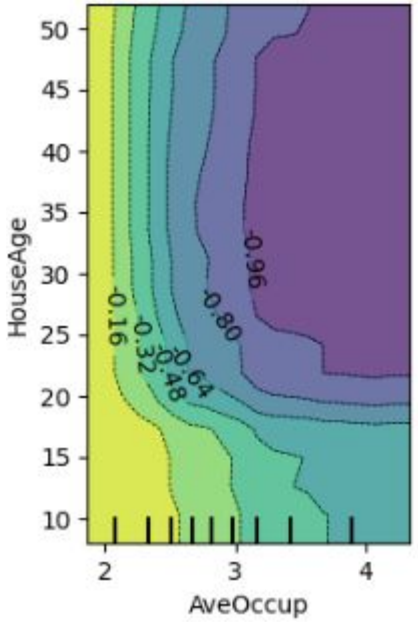
- https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

Partial dependence plots

One way – one variable impact to predicted values



Two way – two variables impact to predicted values

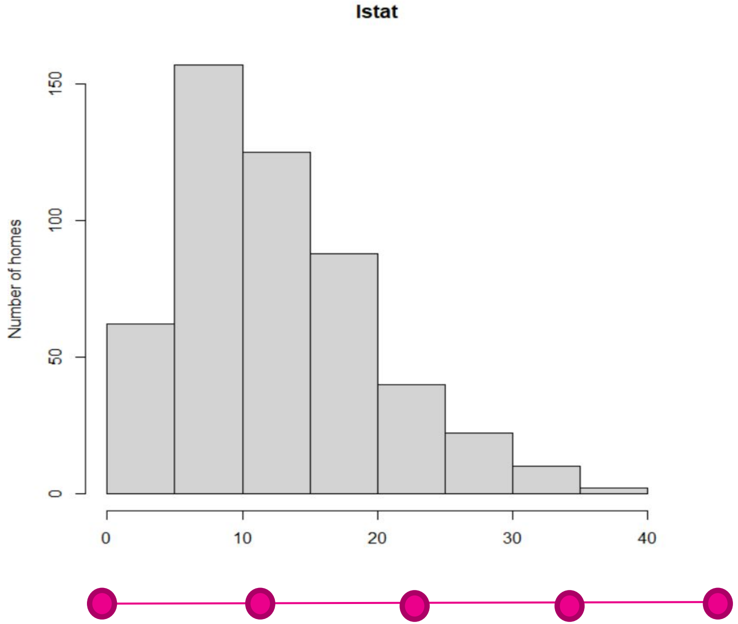


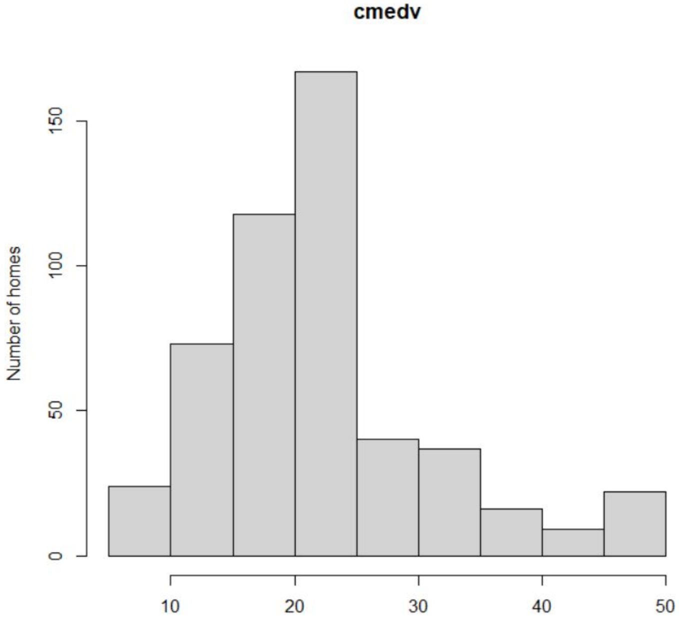
lon	lat	cmedv	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat
-70.9550	42.2550	24.0	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98
-70.9500	42.2875	21.6	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14
-70.9360	42.2830	34.7	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03
-70.9280	42.2930	33.4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94
-70.9220	42.2980	36.2	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33
-70.9165	42.3040	28.7	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21
-70.9360	42.2970	22.9	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43
-70.9375	42.3100	22.1	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15
-70.9330	42.3120	16.5	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93
-70.9290	42.3160	18.9	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10
-70.9350	42.3160	15.0	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45
-70.9440	42.3170	18.9	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27
-70.9510	42.3060	21.7	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71
-70.9645	42.2920	20.4	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26
-70.9720	42.2870	18.2	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26
													21.0	395.62	8.47
													21.0	386.85	6.58
													21.0	386.75	14.67
													21.0	288.99	11.69

Proportion of population that is lower status = 1/2 (proportion of adults without, some high school education and proportion of male workers classified as laborers). The logarithmic specification implies that socioeconomic status distinctions mean more in the upper brackets of society than in the lower classes. Source: 1970 U. S. Census

Creating grid

- Software can automatically choose some grid
- You can specify over how many equidistant points you want to analyze
- You can pass your own grid



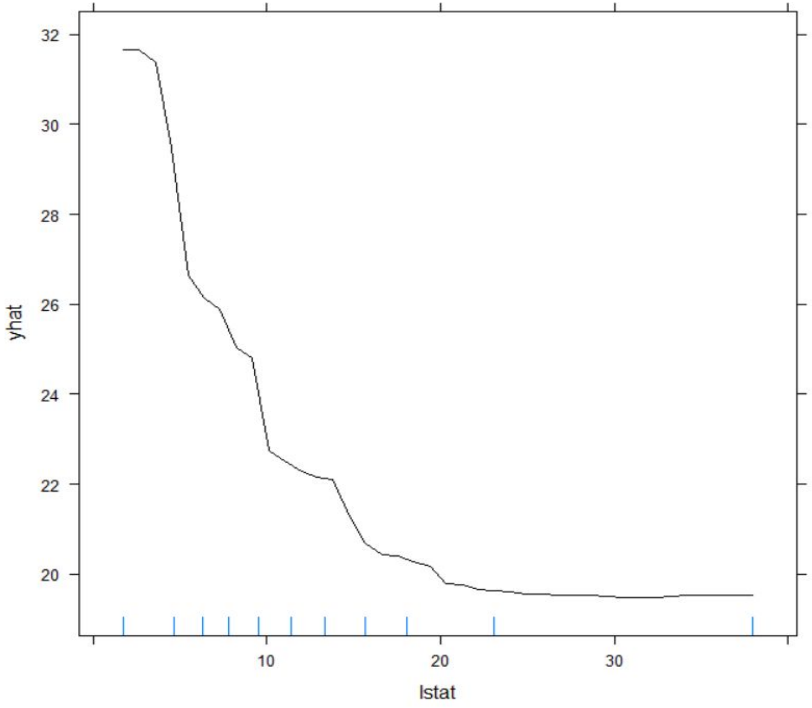


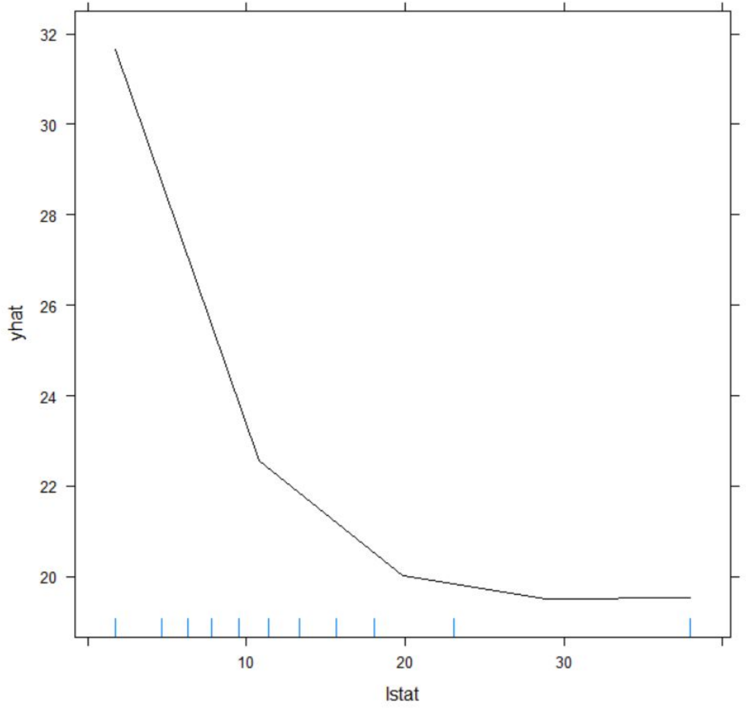
lon	lat	cmedv	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat
-70.9550	42.2550	24.0	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98
-70.9500	42.2875	21.6	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14
-70.9360	42.2830	34.7	0.02729	0.0										392.83	4.03
-70.9280	42.2930	33.4	0.03237	0.0										394.63	2.94
-70.9220	42.2980	36.2	0.06905	0.0										396.90	5.33
-70.9165	42.3040	28.7	0.02985	0.0										394.12	5.21
-70.9360	42.2970	22.9	0.08829	12.5										395.60	12.43
-70.9375	42.3100	22.1	0.14455	12.5										396.90	19.15
-70.9330	42.3120	16.5	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93
-70.9290	42.3160	18.9	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10
-70.9350	42.3160	15.0	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45
-70.9440	42.3170	18.9	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27
-70.9510	42.3060	21.7	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71
-70.9645	42.2920	20.4	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26
-70.9720	42.2870	18.2	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26
-70.9765	42.2940	19.9	0.62739	0.0	8.14	0	0.5380	5.834	56.5	4.4986	4	307	21.0	395.62	8.47
-70.9870	42.2985	23.1	1.05393	0.0	8.14	0	0.5380	5.935	29.3	4.4986	4	307	21.0	386.85	6.58
-70.9780	42.2850	17.5	0.78420	0.0	8.14	0	0.5380	5.990	81.7	4.2579	4	307	21.0	386.75	14.67
-70.9925	42.2825	20.2	0.80271	0.0	8.14	0	0.5380	5.456	36.6	3.7965	4	307	21.0	288.99	11.69

1. Set value of lstat instead of real value to value from grid
2. Use model to calculate predicted cmedv
3. Repeat for all rows in data
4. Calculate mean predicted cmedv over 1 value from grid
5. Repeat for other values from grid

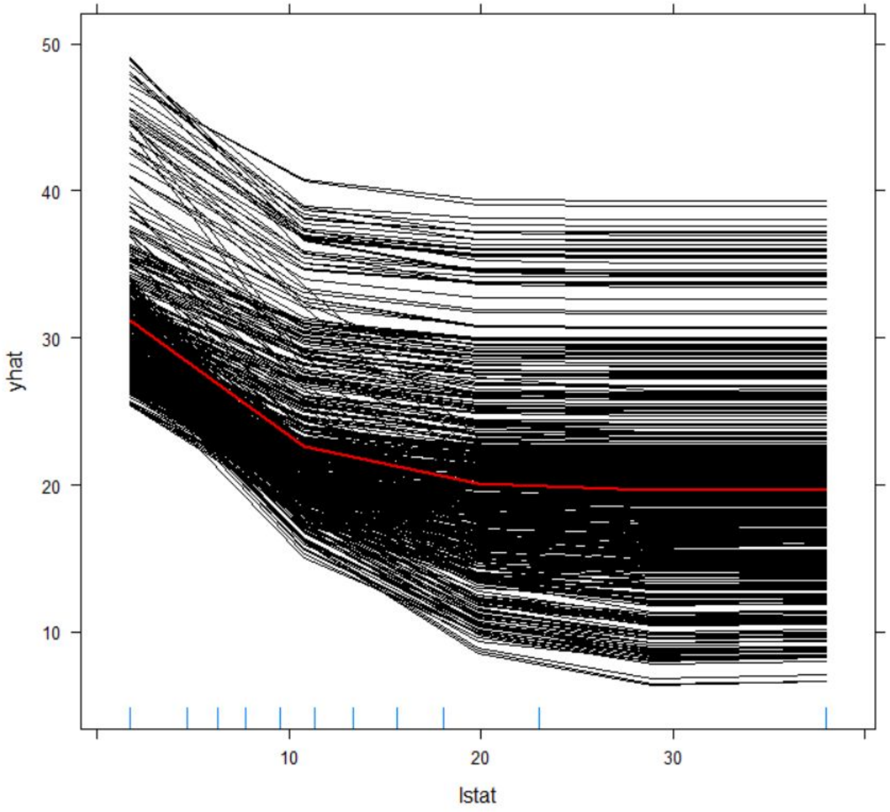
Partial dependence plots – one way – housing dataset example

What is different in these graphs?





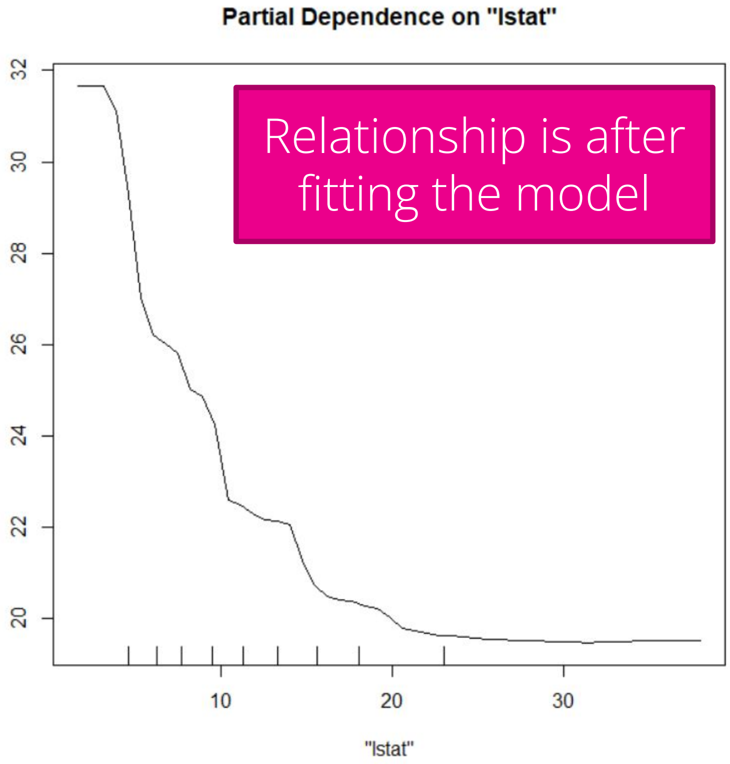
Partial dependence plots – one way – housing dataset example

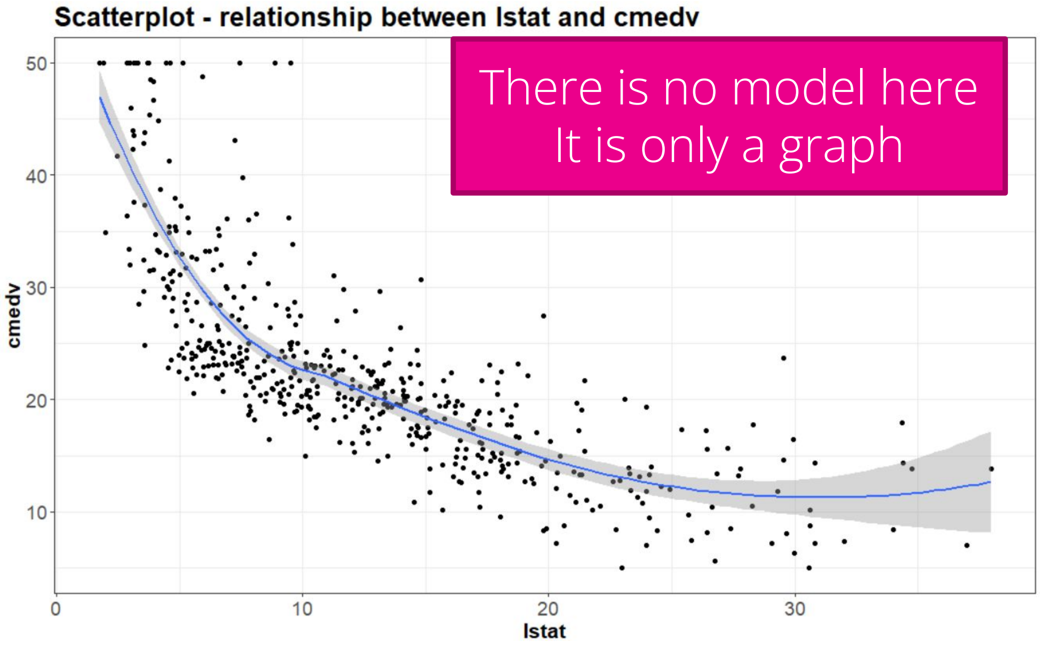


Compute individual conditional expectation (ICE) curves

There are as many as rows in the data used to train the model!

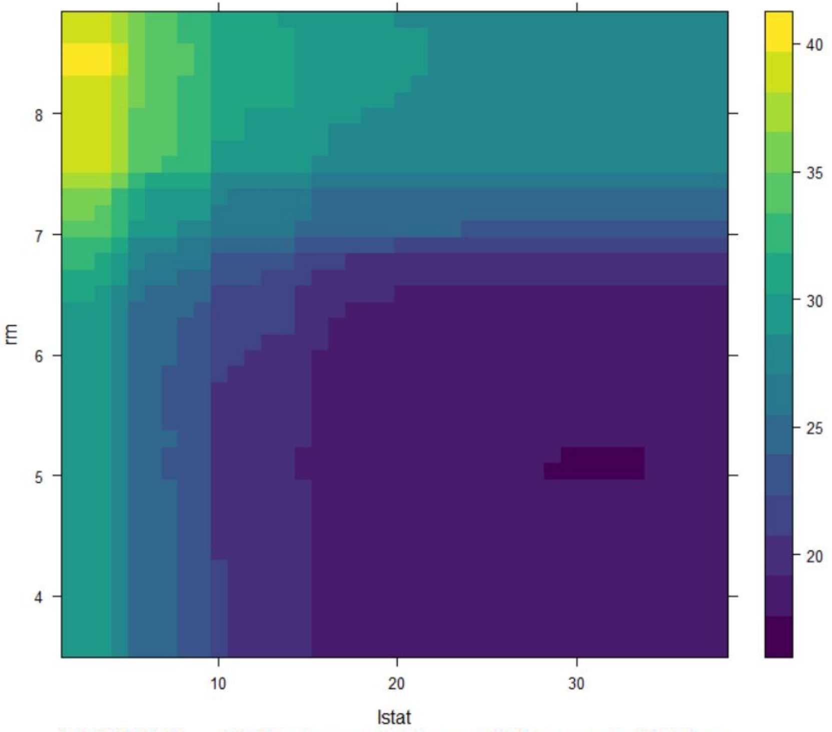
What is the difference between simple scatterplot and partial dependence graph?





Partial dependence plots – two way – good to analyze interactions

RM – average number of rooms per dwelling

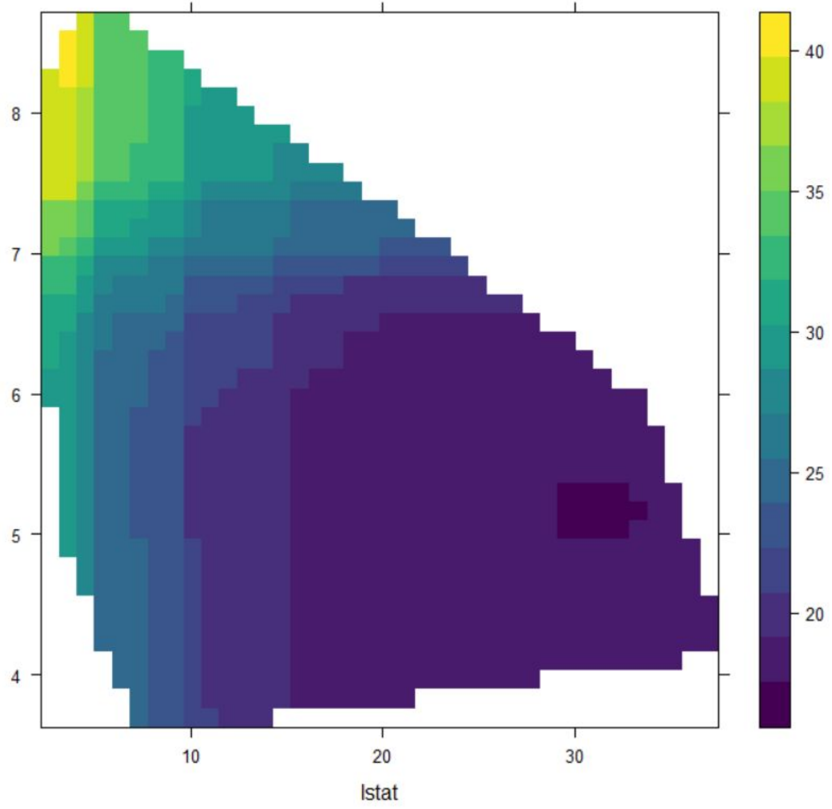


LSTAT - % lower status of the population

Logical indicating whether or not to restrict the values of the first two variables in pred.var to lie within the convex hull of their training values; this affects pred.grid. This helps reduce the risk of interpreting the partial dependence plot outside the region of the data (i.e., extrapolating).Default is FALSE.

Partial dependence plots – two way – good to analyze interactions

RM – average number of rooms per dwelling



LSTAT - % lower status of the population

Logical indicating whether or not to restrict the values of the first two variables in pred.var to lie within the convex chull of their training values; this affects pred.grid. This helps reduce the risk of interpreting the partial dependence plot outside the region of the data (i.e., extrapolating).Default is FALSE.

Disadvantages of partial dependence plots

1. Assumes independence of variables

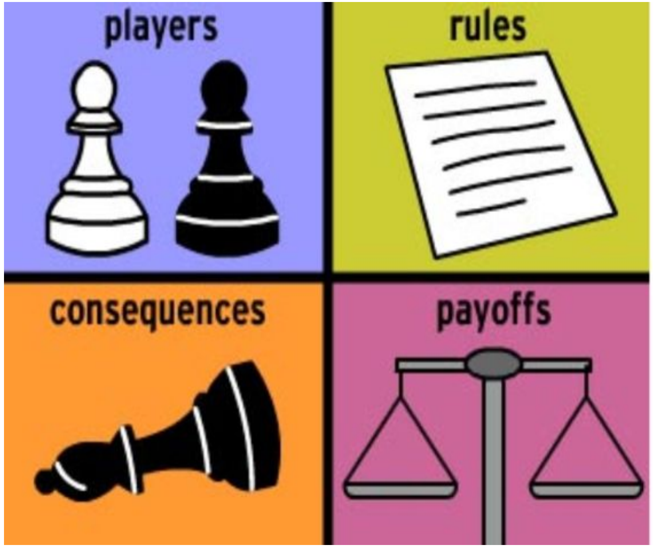
2. Artificial dataset created with values which realistically might not exist

3. Why choosing average (could be corrected by choosing median or adding confidence interval over the ranges)

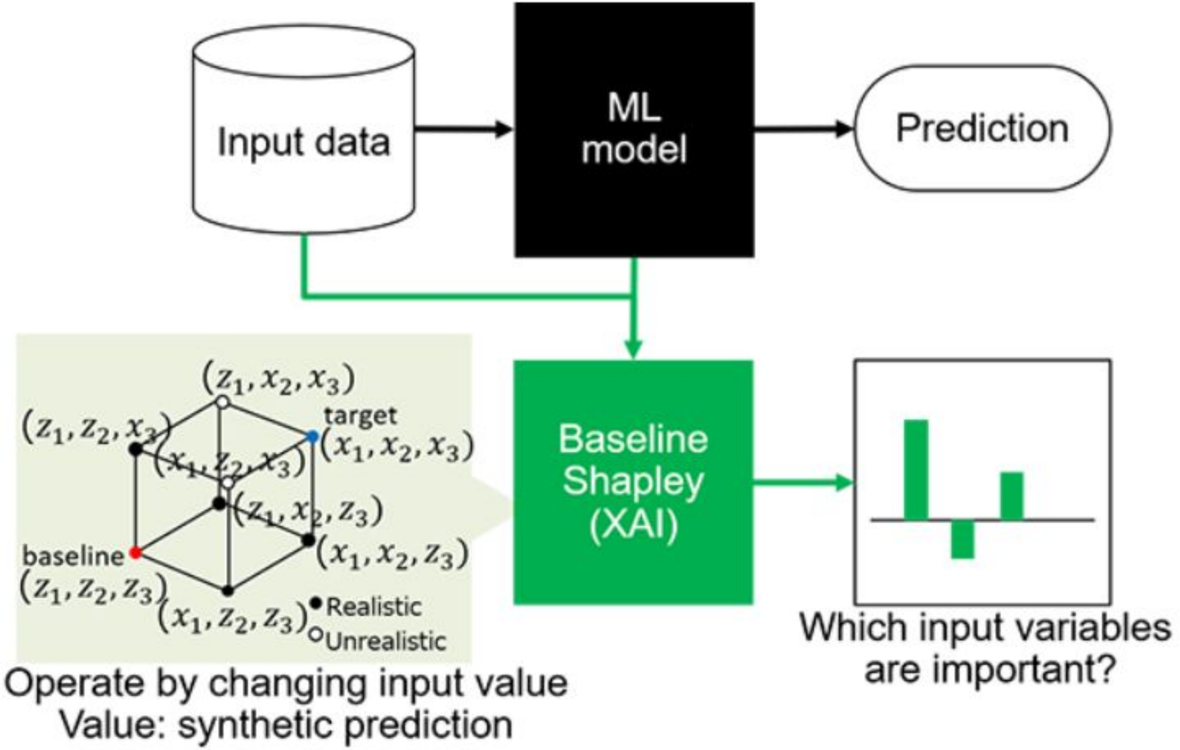
Shapley values

In cooperative game theory, the Shapley value is the average marginal contribution of a player across all possible coalitions in a game (Shapley, 1951)

- Game = prediction task for a single observation x0x0
- Players = the feature values of x0x0 that collaborate to receive the gain or payout
- Payout = prediction for x0x0 minus the average prediction for all training observations (i.e., baseline)



Shapley values



https://www.hitachi.com/rd/sc/aiblog/042/index.html

lon	lat	cmedv	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat
-70.9550	42.2550	24.0	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98
-70.9500	42.2875	21.6	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14
-70.9360	42.2830	34.7	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03
-70.9280	42.2930	33.4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94
-70.9220	42.2980	36.2	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33
-70.9165	42.3040	28.7	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21
-70.9360	42.2970	22.9	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43
-70.9375	42.3100	22.1	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15
-70.9330	42.3120	16.5	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93
-70.9290	42.3160	18.9	0.17004	12.5	7.87	0	0.5240							386.71	17.10
-70.9350	42.3160	15.0	0.22489	12.5	7.87	0	0.5240							392.52	20.45
-70.9440	42.3170	18.9	0.11747	12.5	7.87	0	0.5240							396.90	13.27
-70.9510	42.3060	21.7	0.09378	12.5	7.87	0	0.5240							390.50	15.71
-70.9645	42.2920	20.4	0.62976	0.0	8.14	0	0.5380	[illegible]	[illegible]	[illegible]	4	[illegible]	[illegible]	396.90	8.26
-70.9720	42.2870	18.2	0.63796	0.0	8.14	0	0.5380							380.02	10.26
-70.9765	42.2940	19.9	0.62739	0.0	8.14	0	0.5380							395.62	8.47
-70.9870	42.2985	23.1	1.05393	0.0	8.14	0	0.5380							386.85	6.58
-70.9780	42.2850	17.5	0.78420	0.0	8.14	0	0.5380							386.75	14.67
-70.9925	42.2825	20.2	0.80271	0.0	8.14	0	0.5380	[illegible]	[illegible]	[illegible]	4	[illegible]	[illegible]	288.99	11.69

How does the value of 19.5 of lstat influence prediction of that row compared to average prediction?

Average prediction = mean(fitted value for each row)

Shapley values

The main point of interpreting Shapley Value is to know that they

1. Work with model predictions
2. Represent the difference between one particular prediction (one home) versus average prediction across all homes

It is possible to manipulate the sample which is used for obtaining the average prediction:

1. In standard case Shapley value describes the improvement compared to other homes if the variable change is implemented
2. If we choose as a reference sample alternative strategies for this same home, Shapley value describes the improvement of variable change compared to other possibilities for the same home

Pros and cons of Shapley values

Advantages

- Independent of statistical model used for prescriptions
- Mathematically sound and recognized in the data science community
- Fairly distributes prediction between features

Disadvantages

- Always uses all features
- Might still be a bit complex to explain
- Works with model predictions, so confidence bounds apply

czechitas

REGRESSION PERFORMANCE

Regression Performance: MAE

How to best compare observed values vs. predicted values?

PersonID	Gender	Years Education	Age	Income - Observed	Income - Predicted
2343	F	17	35	63 000	65 200
1213	M	15	32	35 000	37 300
4533	M	15	53	40 000	38 900
4563	M	19	51	100 000	91 450
7453	M	13	32	34 000	35 600
…	…	…	…	…	…

Dataset: 750 individuals

$$\frac{\sum(|y_{predicted} - y_{actual}|)}{n}$$

czechitas

Regression Performance: RMSE

How to best compare observed values vs. predicted values?

PersonID	Gender	Years Education	Age	Income - Observed	Income - Predicted
2343	F	17	35	63 000	65 200
1213	M	15	32	35 000	37 300
4533	M	15	53	40 000	38 900
4563	M	19	51	100 000	91 450
7453	M	13	32	34 000	35 600
…	…	…	…	…	…

Dataset: 750 individuals

$$Root\ Mean\ Squared\ Error\ (RMSE) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Predicted_i - Observed_i)^2}$$

n – number of observations

$$RMSE = \sqrt{\frac{(65\,200 - 63\,000)^2 + (\,37\,300\ - 35\,000)^2 + \ \ldots}{750}}$$

czechitas

Regression Performance: MAPE

How to best compare observed values vs. predicted values?

PersonID	Gender	Years Education	Age	Income - Observed	Income - Predicted
2343	F	17	35	63 000	65 200
1213	M	15	32	35 000	37 300
4533	M	15	53	40 000	38 900
4563	M	19	51	100 000	91 450
7453	M	13	32	34 000	35 600
…	…	…	…	…	…

Dataset: 750 individuals

$$Mean\ Absolute\ Percentage\ Error\ (MAPE) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Observed_i - Predicted_i}{Observed_i}\right|$$

n – number of observations

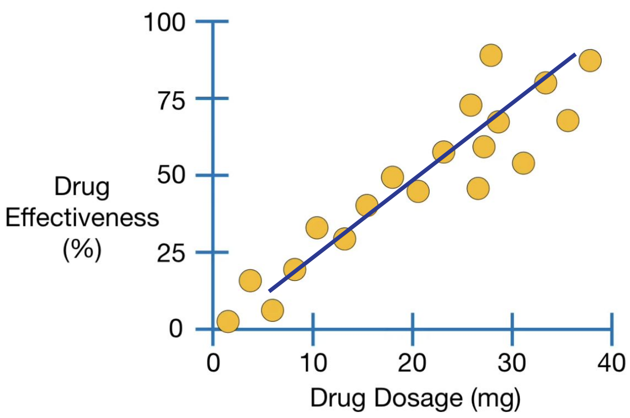
$$MAPE = \frac{1}{750}\left(\left|\frac{63\,000 - 65\,200}{63\,000}\right| + \left|\frac{35\,000 - 37\,300}{35\,000}\right| + \cdots\right)$$

czechitas

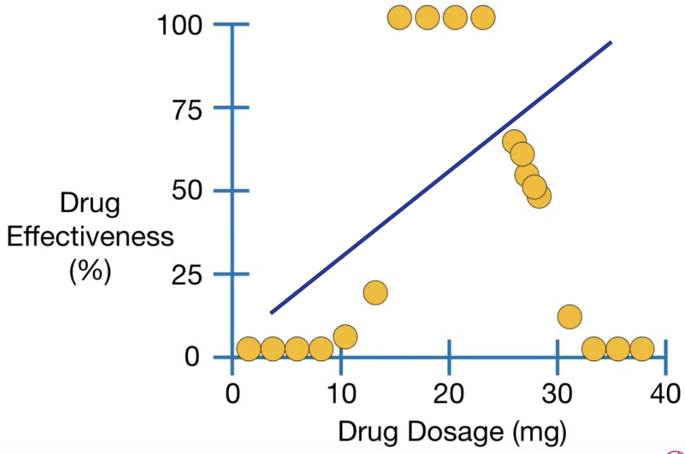
Regression Performance: Benchmarks

Compare RMSE/MAPE of your favorite model with:

a) **a simple model:** Linear Regression (is non-linearity better than linearity?)



vs.



Regression Performance: Benchmarks

Compare RMSE/MAPE of your favorite model with:

b) **multiple other models:** try 2-3 methods and see which has the lowest prediction errors

TRAIN/TEST DATA SPLIT

CustomerID	Education	Age	Income
2343	17	35	50 000
1213	15	32	35 000
4533	15	53	40 000
4563	19	51	100 000
7554	18	28	50 000
6465	13	25	27 500
7453	13	32	34 000
6775	18	43	72 000
4643	19	47	??
6886	19	37	??
8668	21	39	??
8765	23	46	??
9797	12	29	??

TRAIN DATA
(~60-80% of available data with income values)

TEST DATA

czechitas

TRAIN/TEST DATA SPLIT

1. **Find the model using the training data**
2. Calculate model performance
3. **Use the same model to predict with testing data**
4. Calculate model performance

→ We want both training and testing model performance to be similarly good.

TRAIN/TEST DATA SPLIT

1. Find the model using the training data

$$Income = 15000 + 1500\ MaleGender + 2100\ Education + 560\ Age\ + \varepsilon$$

2. Calculate error MAPE = 1.2%

3. Using the same model, predict on testing data

PersonID	Gender	Years Education	Age	Income
8112	F	17	35	???

$$Income = 15000 + 1500\ * 0 + 2100 * 17 + 560\ * 35 = \mathbf{65\,200}$$

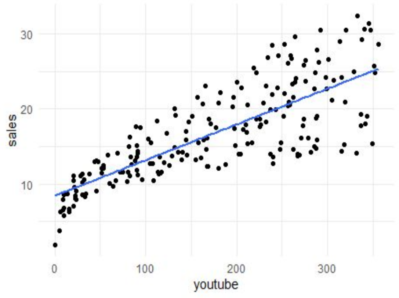
4. Calculate error MAPE = 1.3%

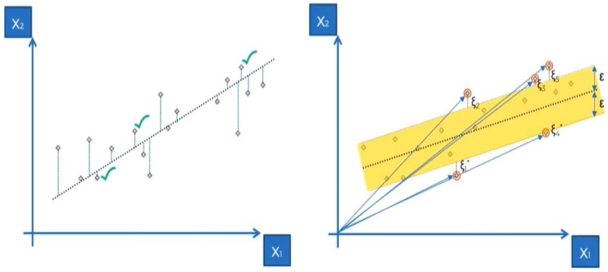
czechitas

REGRESSION ALGORITHMS - SUMMARY

Regression methods we covered

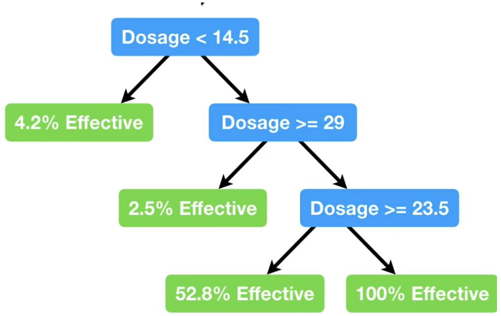
Linear regression

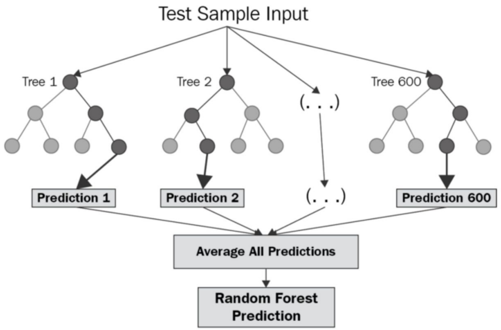




Support Vector Regression

Regression Trees





Random Forest

czechitas

Regression methods - Summary

- Used for predicting **continuous variables**
- **Linear model** used rather as a **benchmark**
- In practice, we would mostly use **ensemble methods** such as random forest
- Performance metrics include **RMSE** or **MAPE**
- Black box models can be made explainable using **Partial dependence plots** or **Shapley values**

Next lecture: Classification


czechitas
Thank you for your attention.
We are looking forward to the next lecture!

Sources

https://www.keboola.com/blog/random-forest-regression

https://www.youtube.com/watch?v=g9c66TUylZ4

https://builtin.com/data-science/random-forest-python

https://www.rebellionresearch.com/what-are-the-disadvantages-of-random-forest

https://www.youtube.com/watch?v=DBApaR2mTg0&ab_channel=AninditaDas

https://scikit-learn.org/stable/modules/partial_dependence.html

https://scikit-learn.org/stable/modules/tree.html

https://www.youtube.com/watch?v=NBg7YirBTN8&ab_channel=ritvikmath

czechitas