NATIONAL RESEARCH UNIVERSITY

School of Data Analysis and Artificial Intelligence Department of Computer Science

DATA SCIENCE FOR BUSINESS

Lecture 4. Data Science in Retail. Forecasting with regression.

Moscow, May 15th, 2020.

lzhukov@hse.ru

WHAT IS RETAIL?

Retail is the sale of consumer goods (or services) through a distribution channel (store, catalogue, online) directly to the consumer

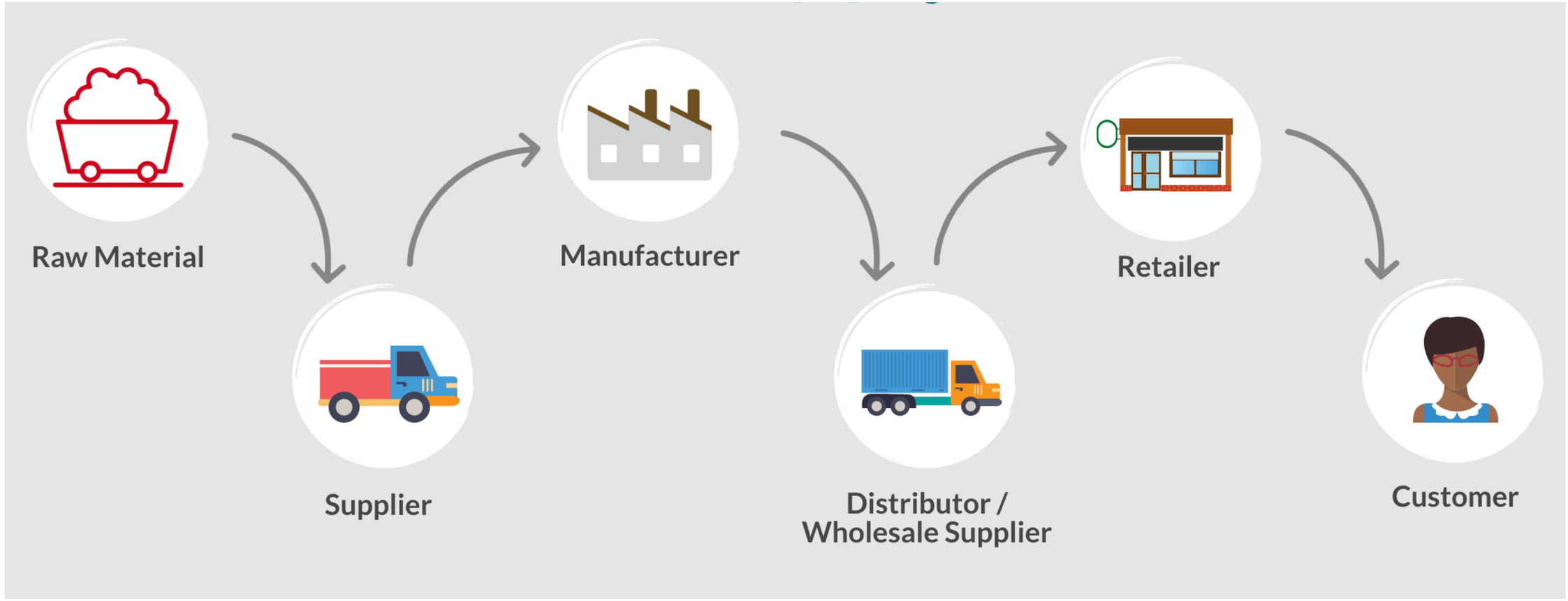
Some retail segments

- Grocery/food retailor & mass merchants
- Fashion/apparel and department stores
- Specialty retail
- Restaurants, cafes, and fast food

RETAIL SUPPLY CHAIN

Presentation subtitle

Supply ➡ ⬅ Demand



DATA DRIVEN DECISION MAKING

Examples of DS topics

Operations (supply side)
[buying, logistics, sales]

- Demand forecast
- Sales forecast
- Buying volumes / inventory management
- Store allocation optimization
- Price optimization / price elasticity
- Mark down / promotion effectivness

Customer (demand side)
[marketing]

- Personalized marketing
- Recommendation engines, next best offer
- Market basket analysis
- Cross-selling and up-selling
- Propensity to buy
- Loyalty program optimization
- Customer sentiment analysis

prediction → **optimization** → **cost / revenue**

DATA DRIVEN DECISION MAKING

Customer and SKU level data analysis

Sales data

Time	Store	SKU	Units	Dollars
Week	Region	Category		
Month	Age	Model		
Quarter	Size	Color		
Year	“Same” status	Size		

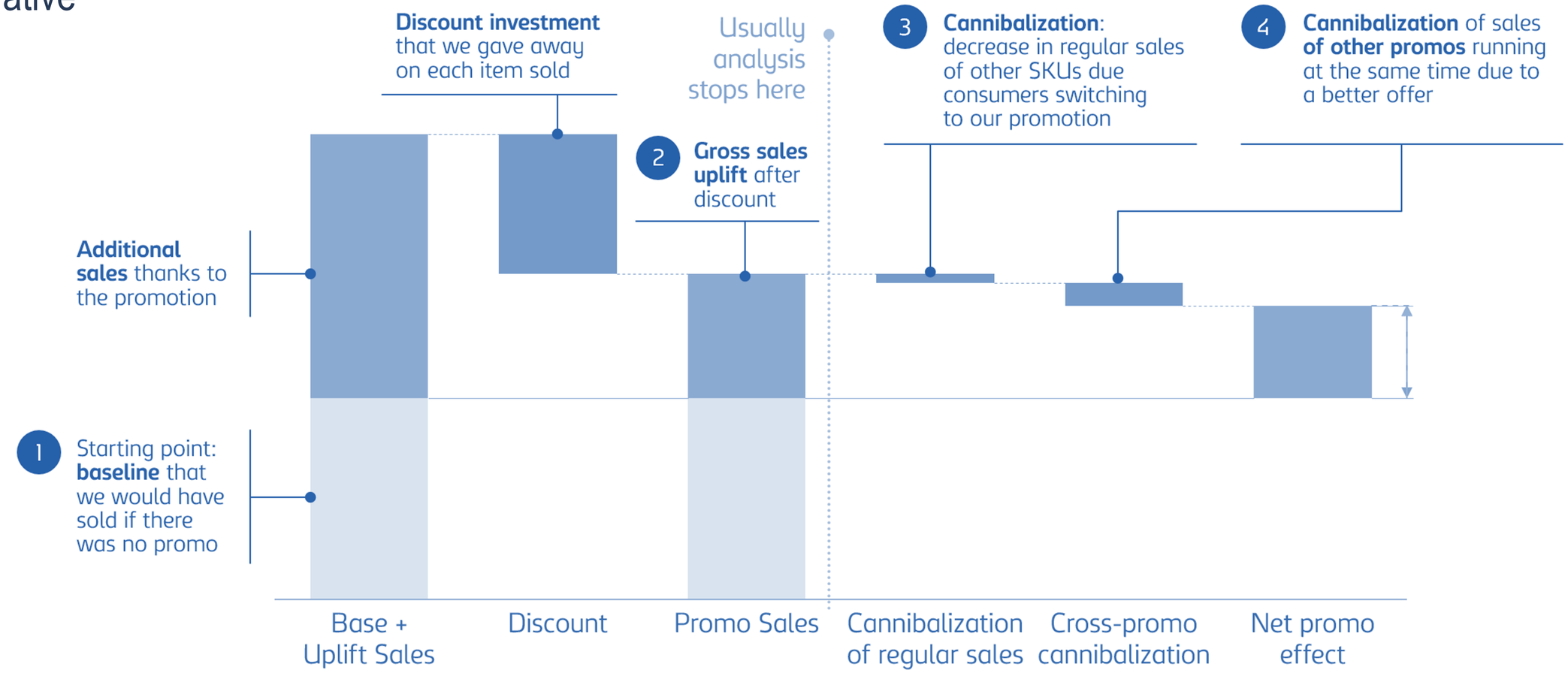
Customer data (CRM)

Customer ID	Date	SKU	Store	Units	Dollars
Demos					

Promo data, marketing data, external data (economics, geographical, population, brands)

PROMO EFFECTIVENESS

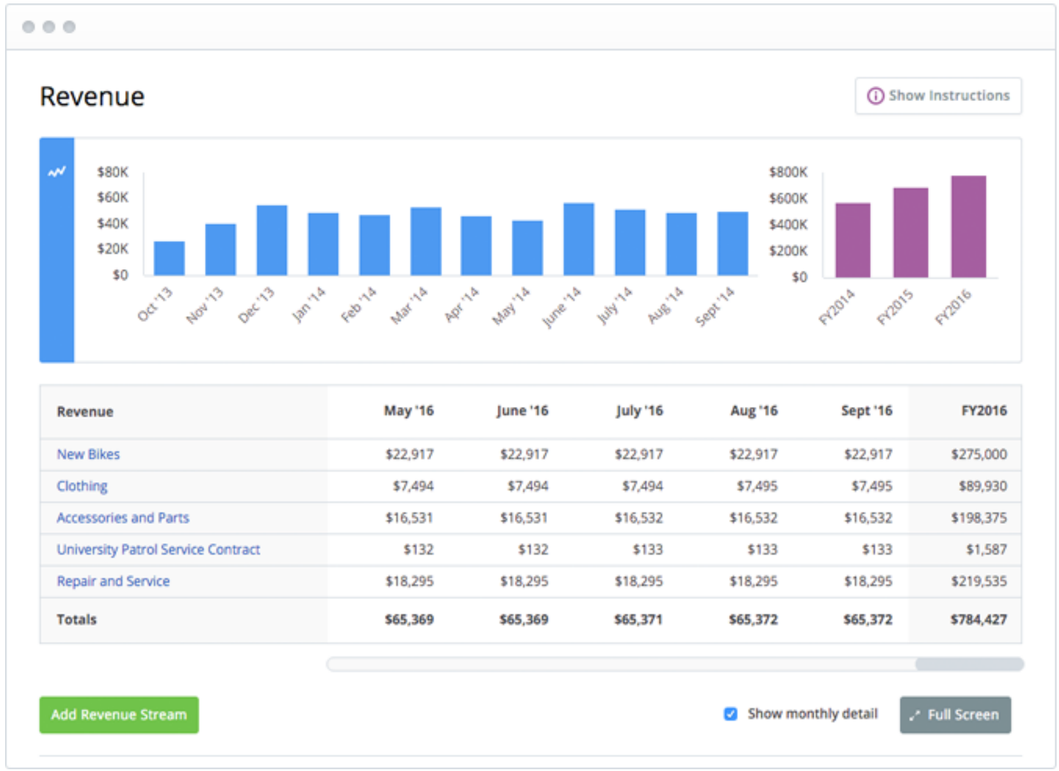
Illustrative



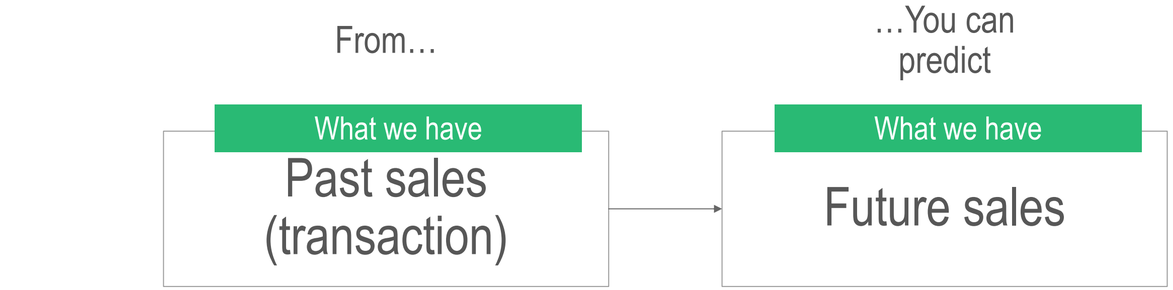
School of Data Analysis and Artificial Intelligence
Department of Computer Science

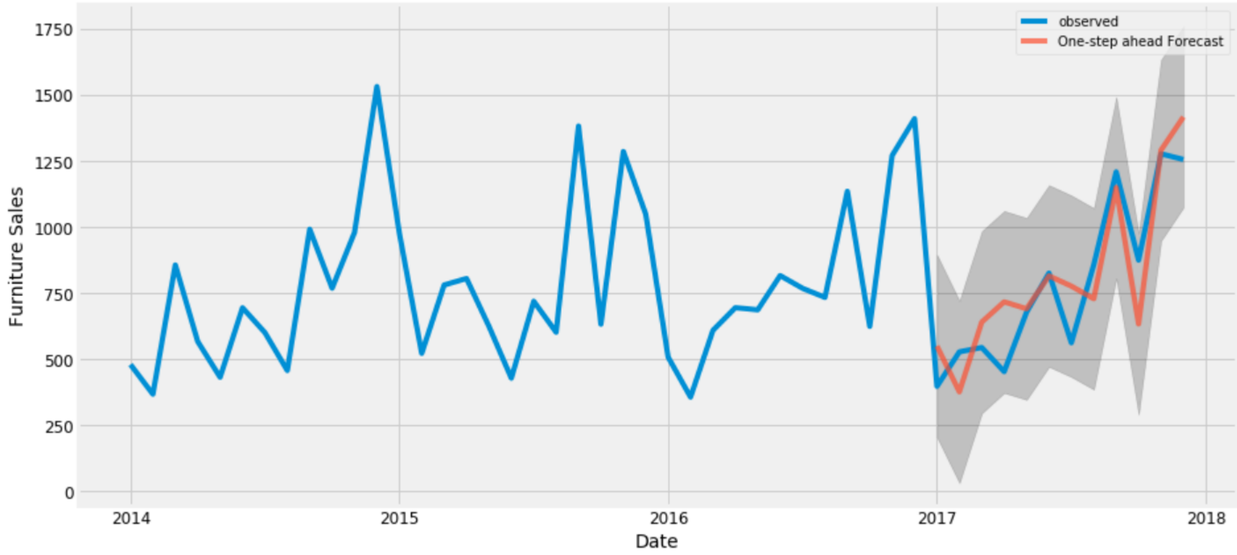
SALES FORECASTING

Estimating the future sales



Revenue	May '16	June '16	July '16	Aug '16	Sept '16	FY2016
New Bikes	$22,917	$22,917	$22,917	$22,917	$22,917	$275,000
Clothing	$7,494	$7,494	$7,494	$7,495	$7,495	$89,930
Accessories and Parts	$16,531	$16,531	$16,532	$16,532	$16,532	$198,375
University Patrol Service Contract	$132	$132	$133	$133	$133	$1,587
Repair and Service	$18,295	$18,295	$18,295	$18,295	$18,295	$219,535
Totals	**$65,369**	**$65,369**	**$65,371**	**$65,372**	**$65,372**	**$784,427**





FORECASTING METHODS

Time series forecasting / signal extrapolation

- Signal history
- Few external factors
- Structured (trend, cyclicality) signal

$y(t+1) = f(y(t), y(t-1), y(t-2), \ldots)$

- Moving average
- Exponential smoothing
- ARIMA

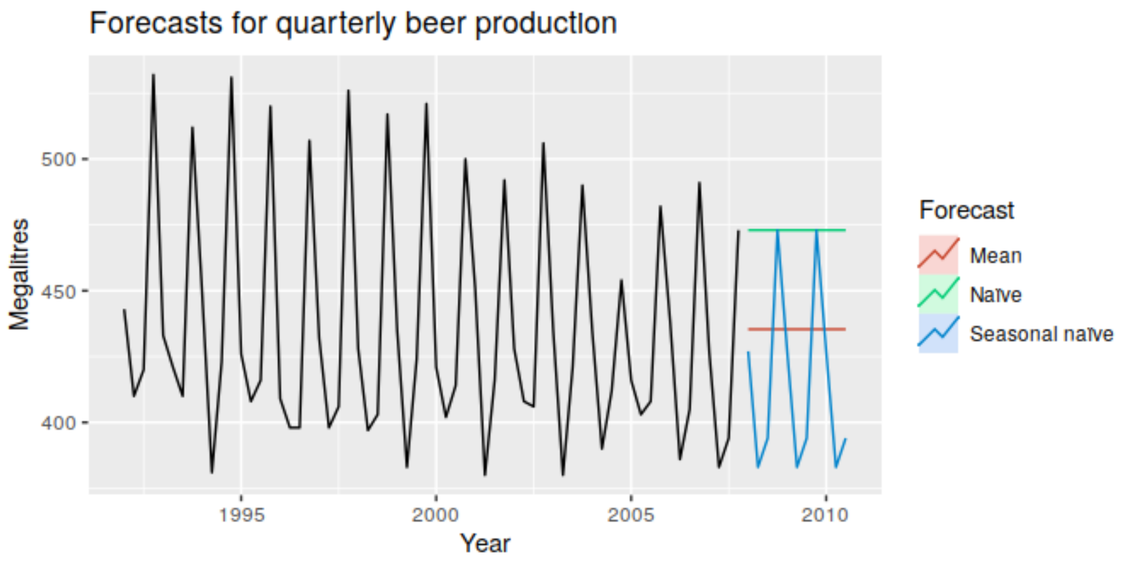
Point matching / regression

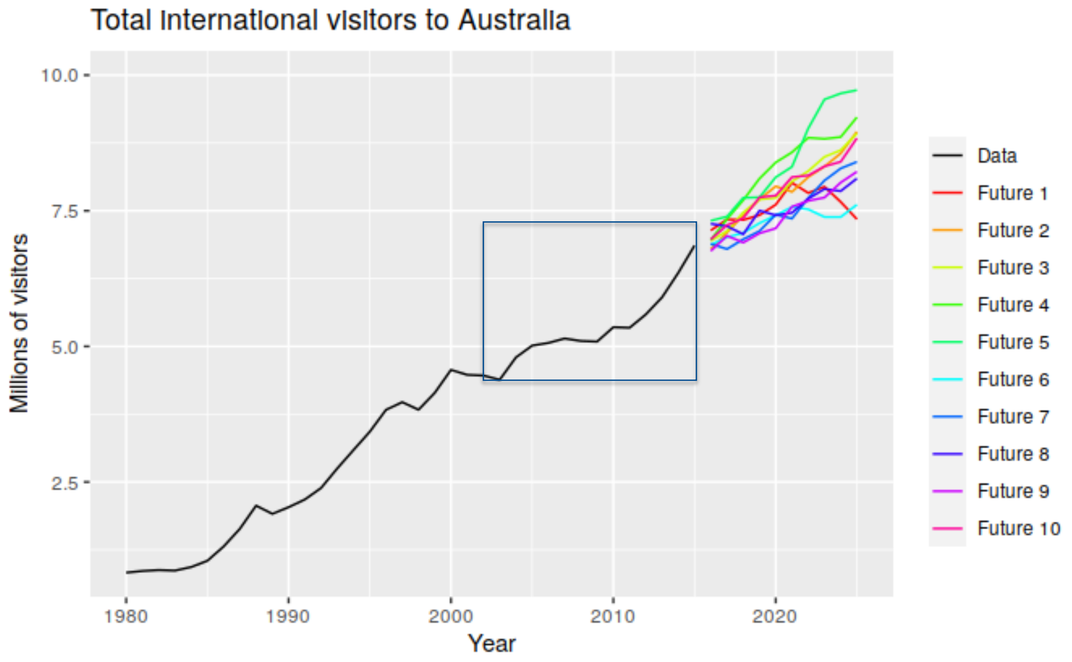
- History of comparable signals
- Many explanatory factors
- Large datasets

$y(t) = f(x1(t), x2(t), \ldots))$
$y(t+h) = f(x1(t), x2(t), \ldots)$

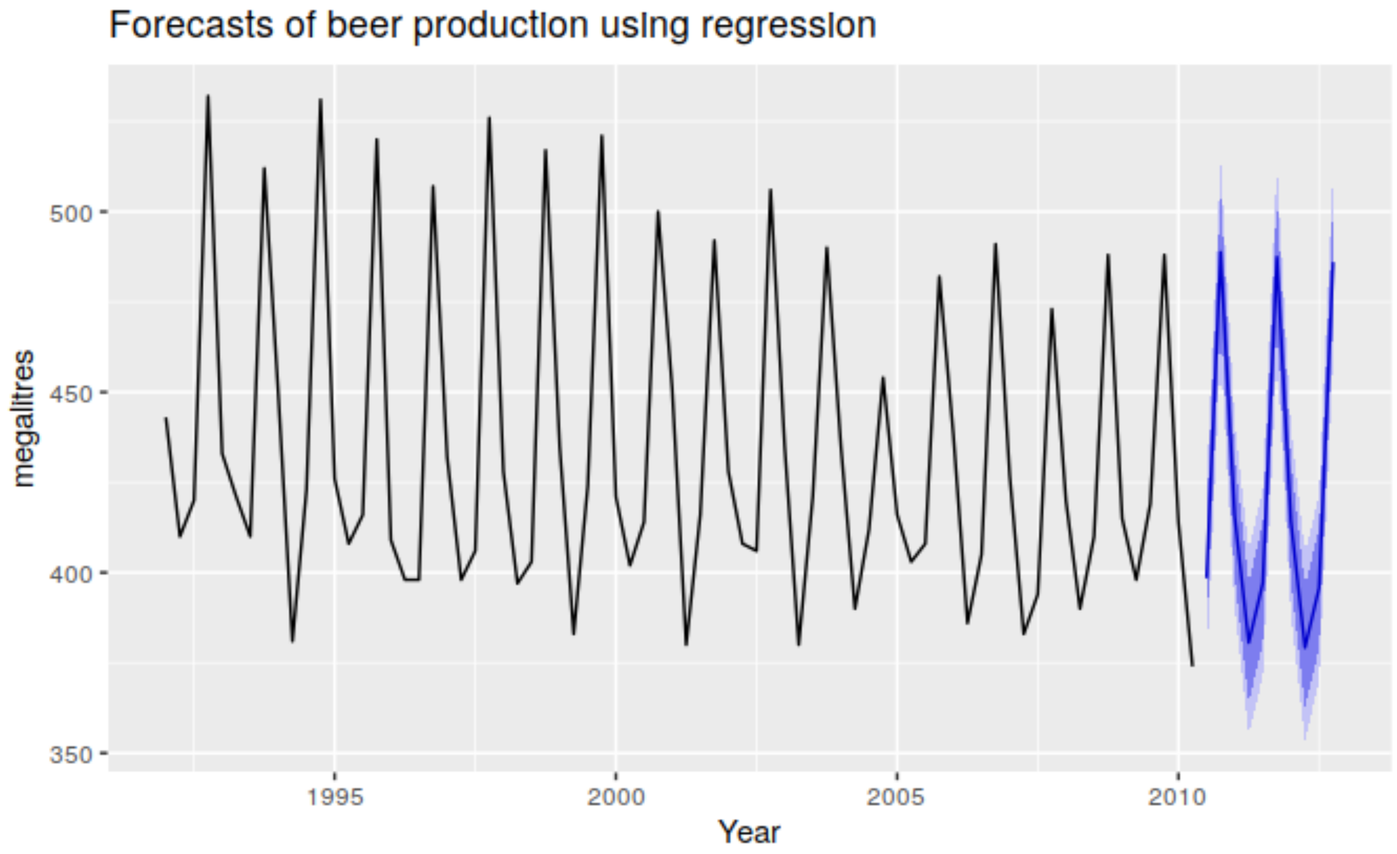
ML algorithms
- GLM
- Random forest
- Gradient boosting

TIME SERIES FORECASTING



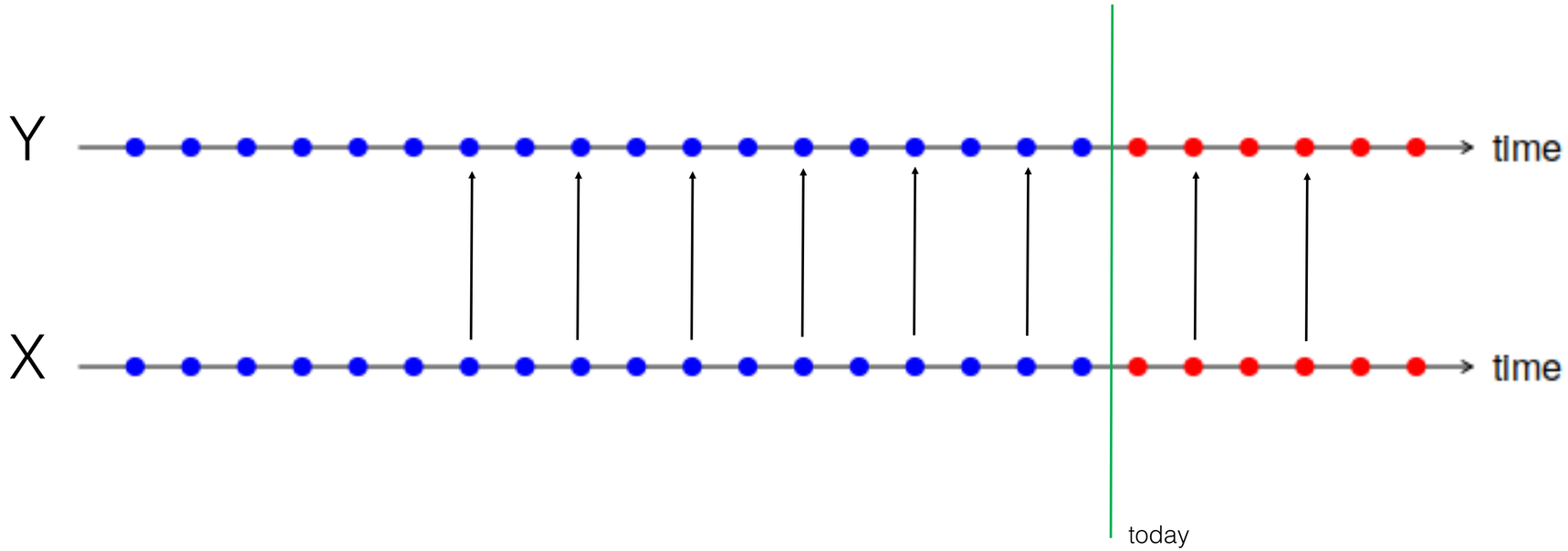


FORECASTING WITH REGRESSION



FORECASTING WITH REGRESSION

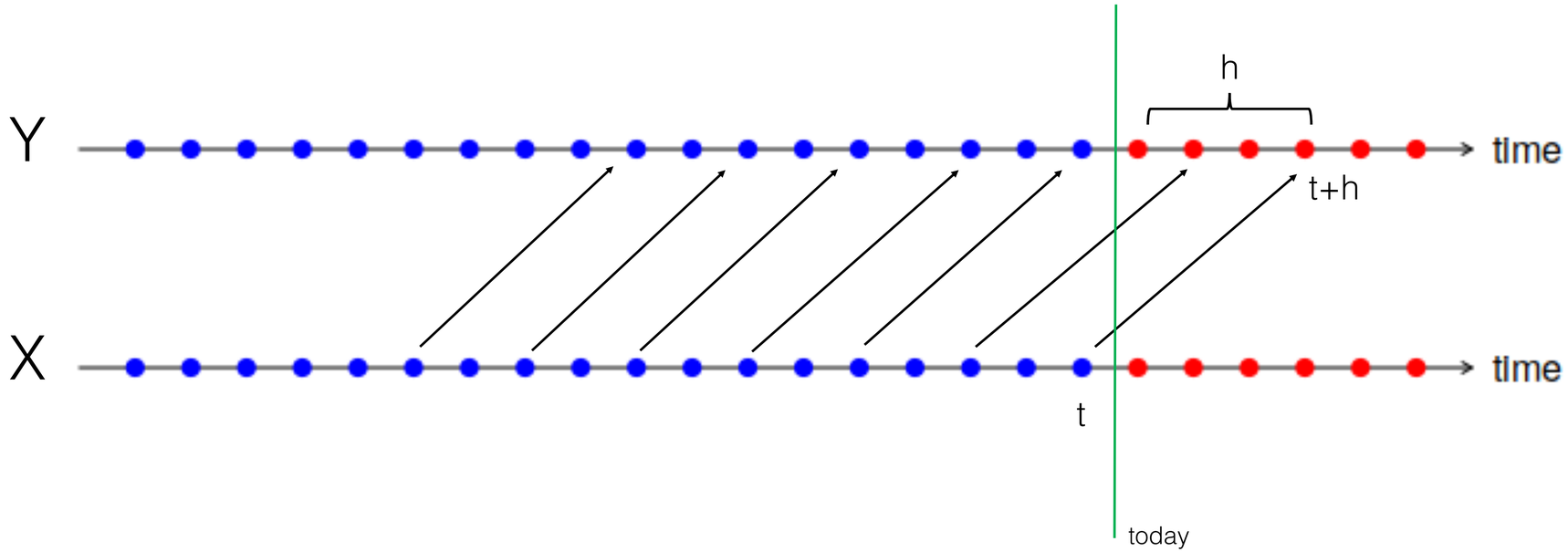
Direct modeling



Regression: Y(t) = f(x1(t), x2(t), x3(t), x4, x5, x6)

FORECASTING WITH REGRESSION

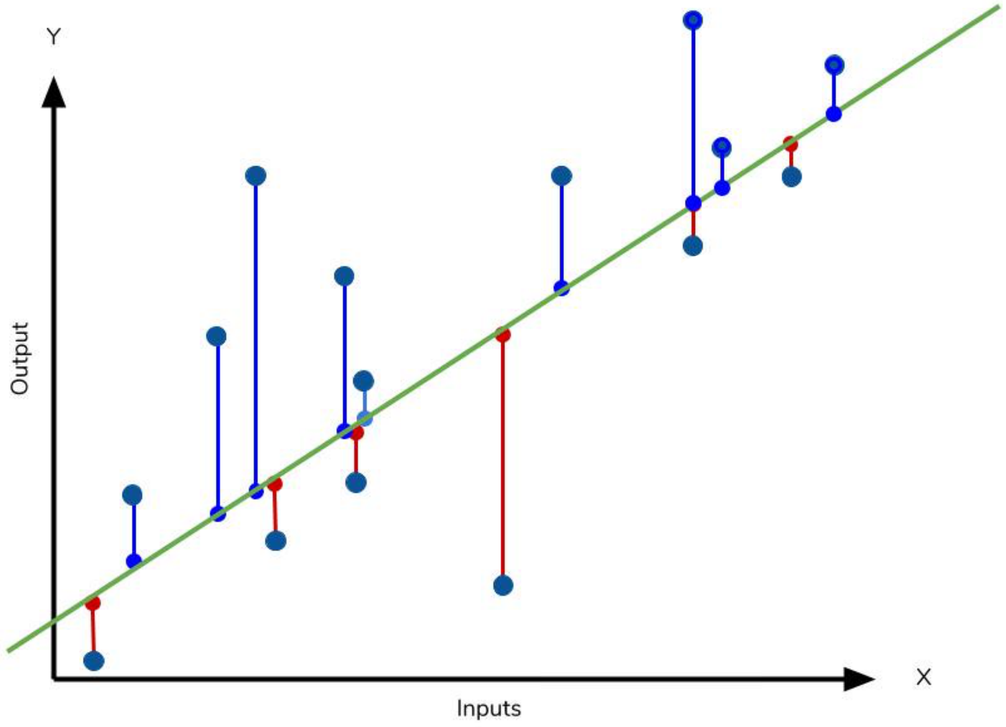
Modeling with time lag



Regression: Y(t+h) = f(x1(t), x2(t), x3(t), x4, x5, x6)

REGRESSION EVALUATION

Quality metrics



Standard quality metrics

Mean absolute error: $$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

Mean squared error: $$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

Root mean squared error: $$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

R-squared: $$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

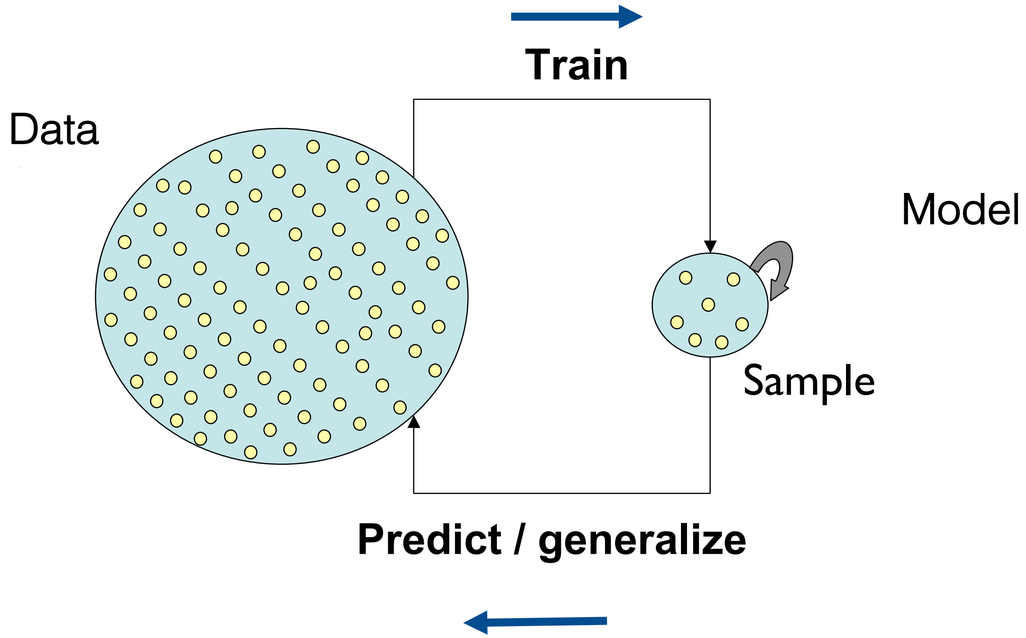
Where,

$\hat{y}$ – *predicted value of* y

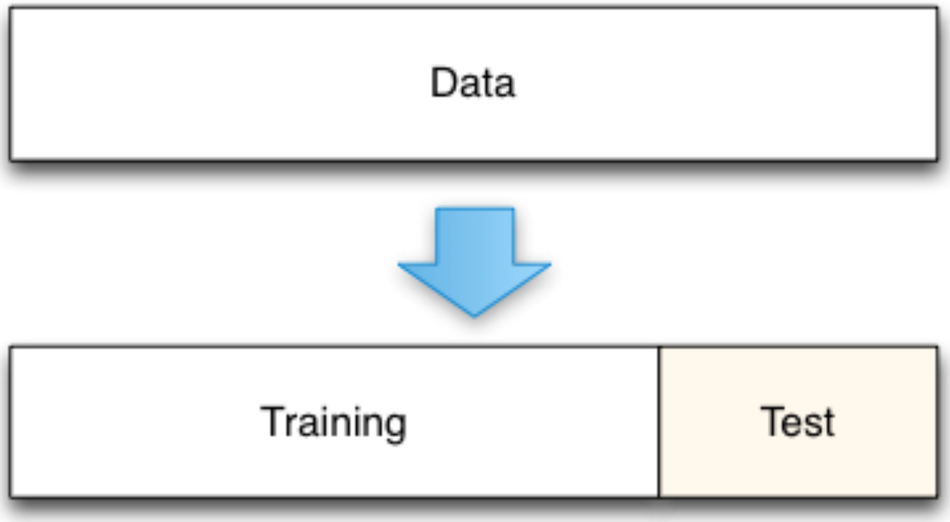
$\bar{y}$ – *mean value of* y

TRAINING AND TESTING

Learning on data



Train & Test split

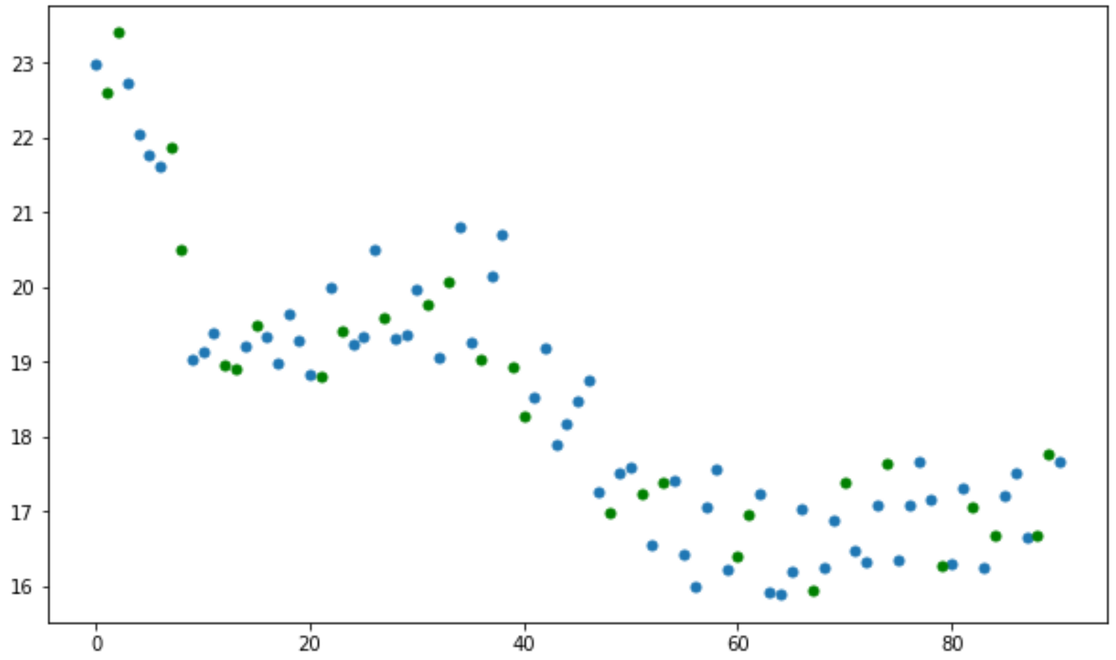


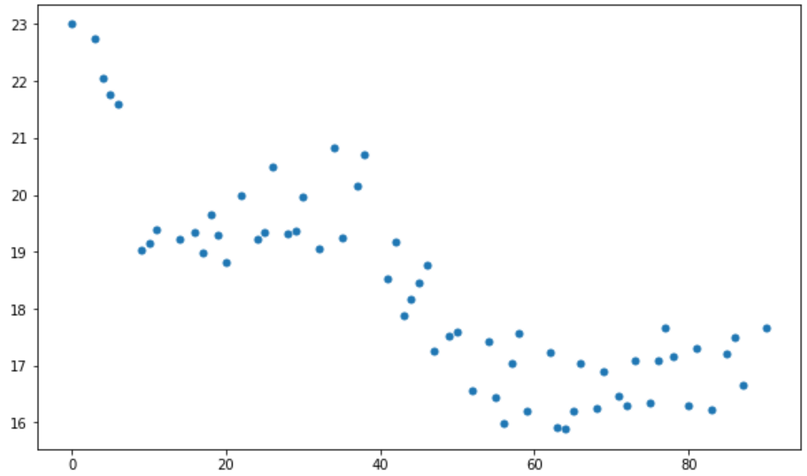
- Build the model
- TRAINING ERROR

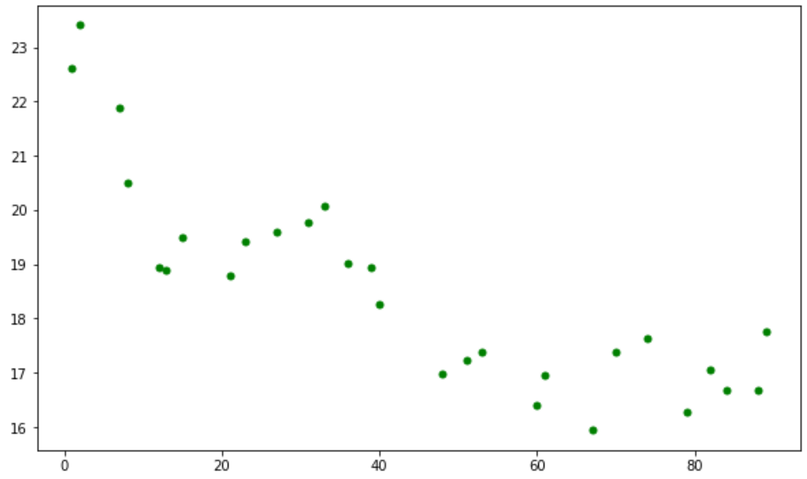
- Test the model
- TESTING ERROR

REGRESSION

Modeling

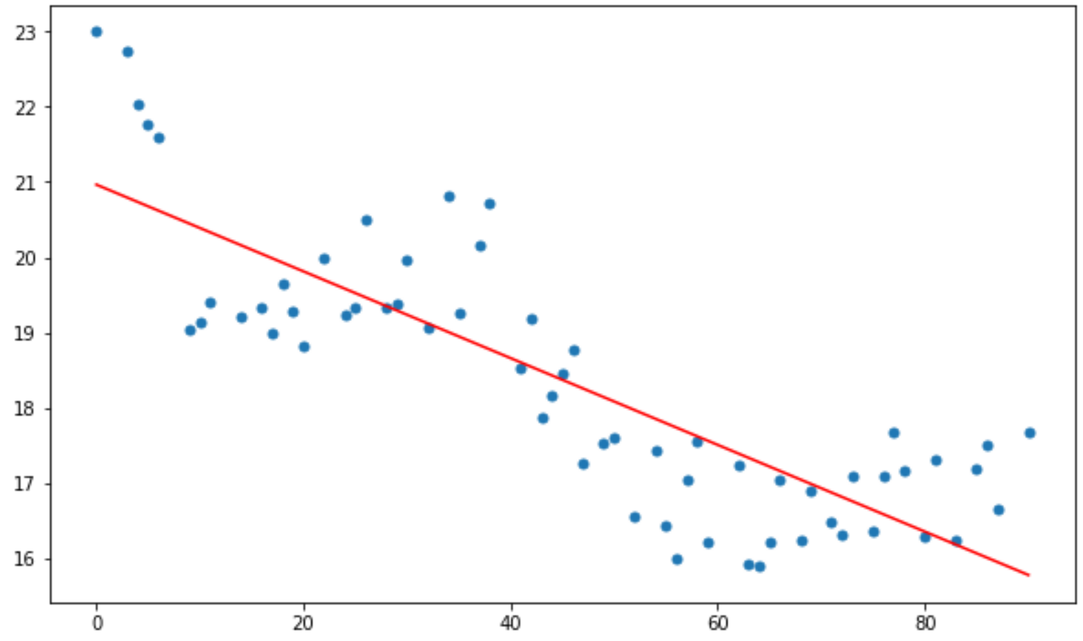




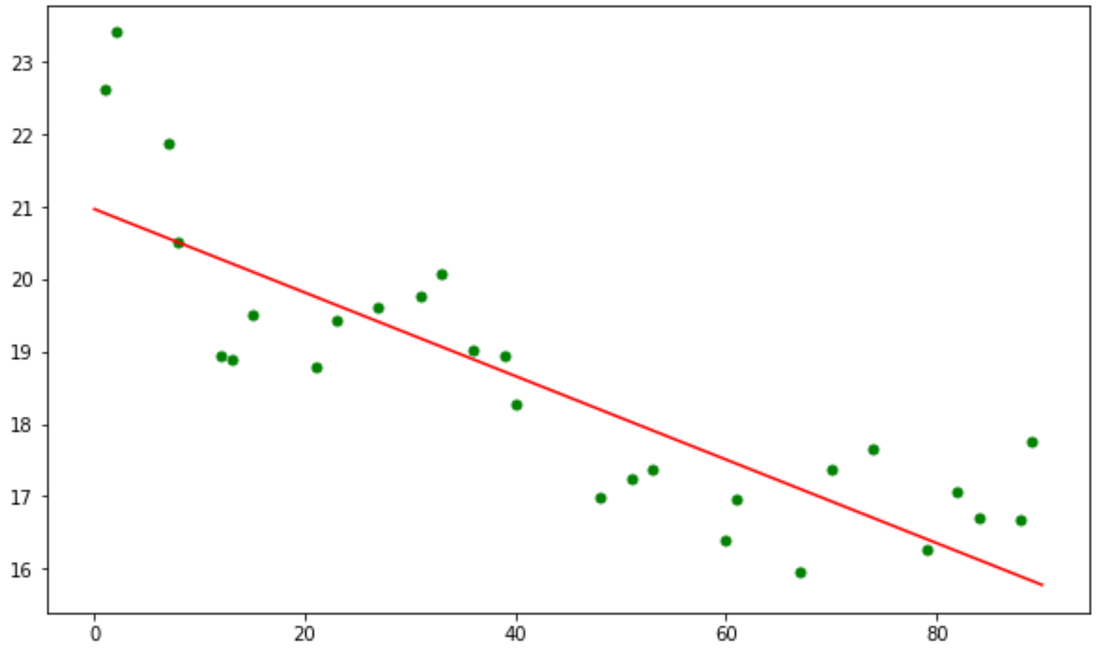


LINEAR REGRESSION

Modeling



Train error: 0.9733

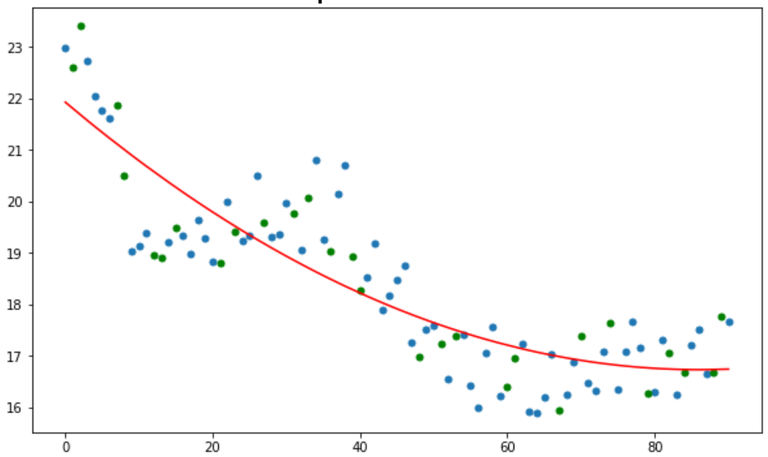


Test error: 1.0222

POLYNOMIAL REGRESSION

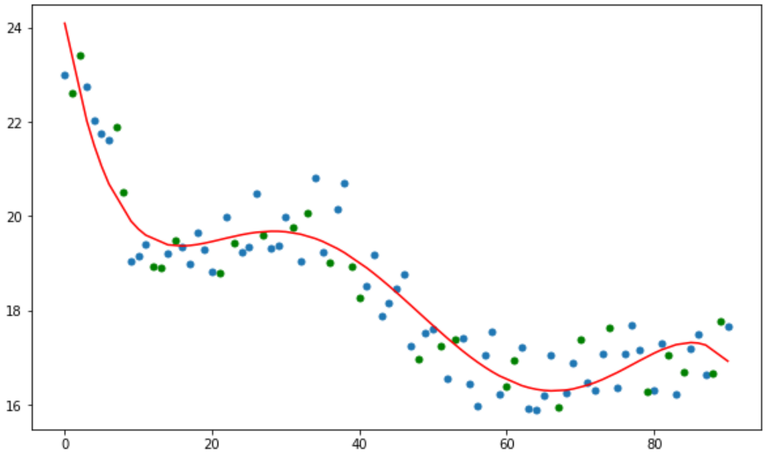
Modeling

p=2



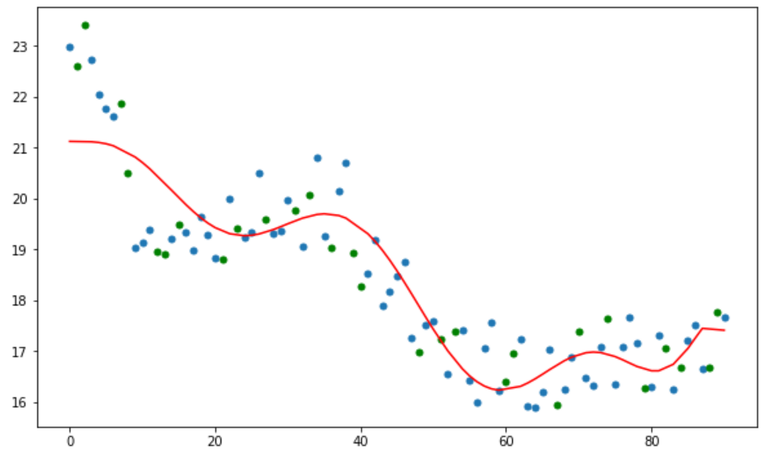
Train error: 0.8792
Test error: 0.8319

p=5



Train error: 0.5989
Test error: 0.6242

p=10

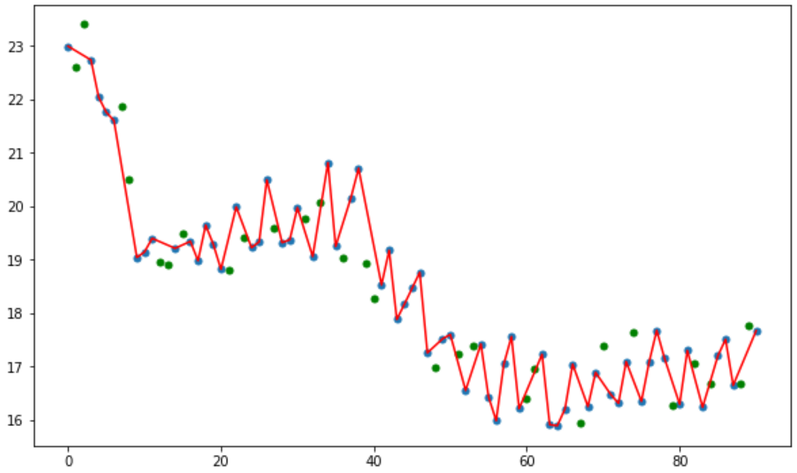


Train error: 0.7399
Test error: 0.8418

KNN REGRESSION

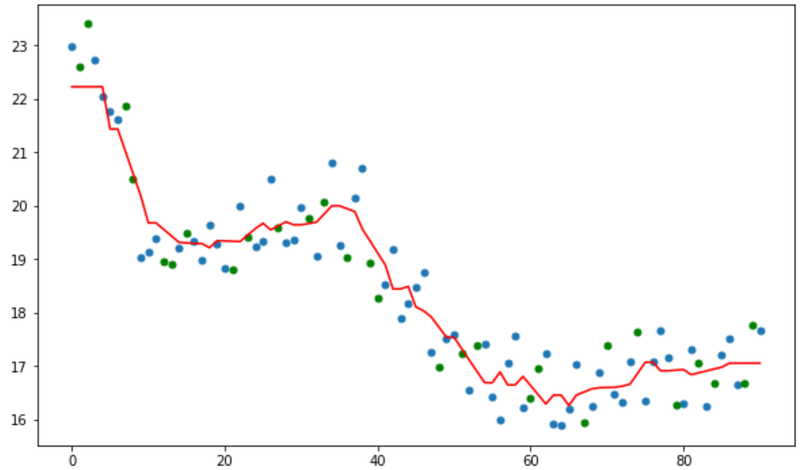
Modeling

k=1



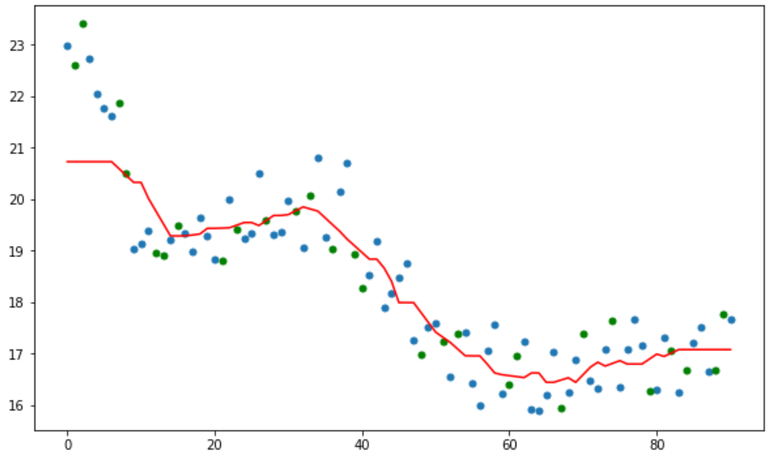
Train error: 0.0
Test error: 0.7574

k=5



Train error: 0.5468
Test error: 0.6248

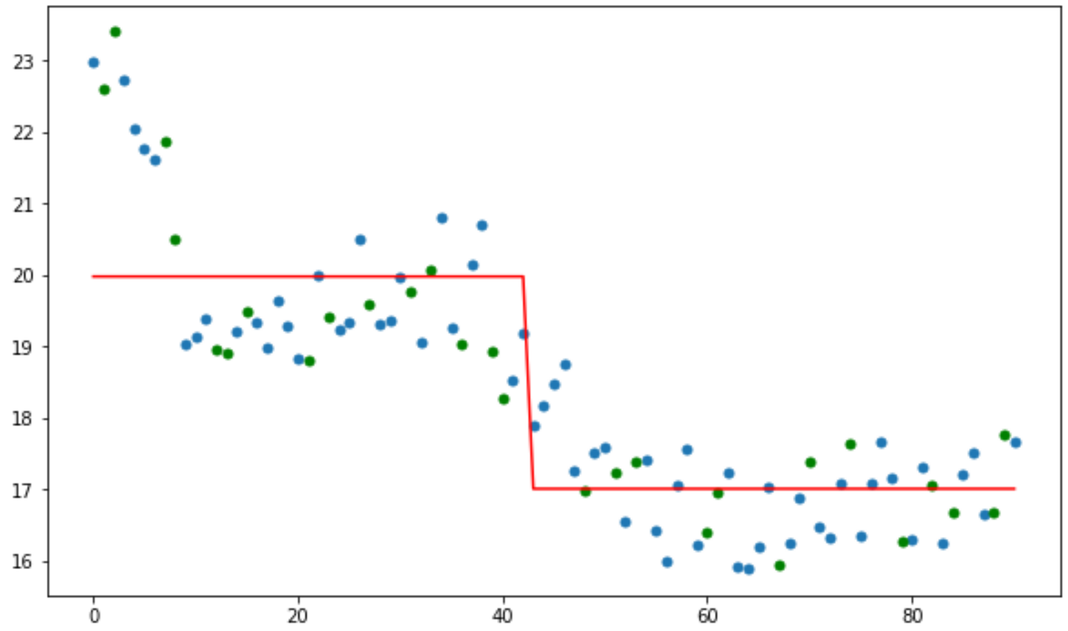
k=10

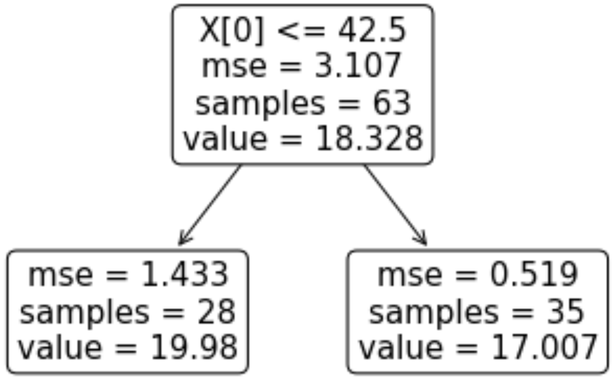


Train error: 0.7399
Test error: 0.8241

REGRESSION TREES

Modeling

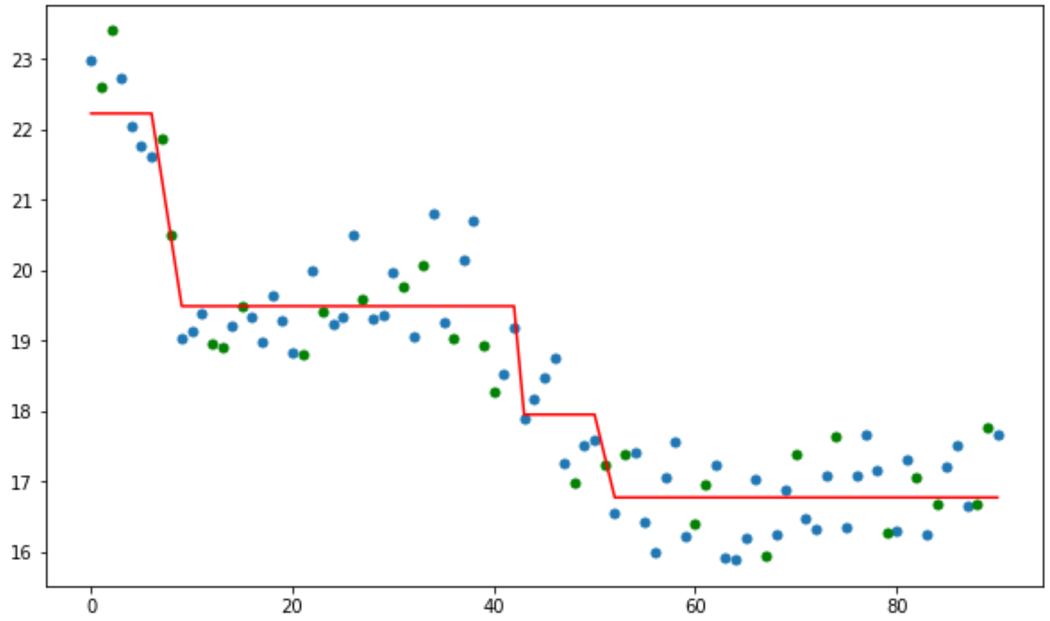


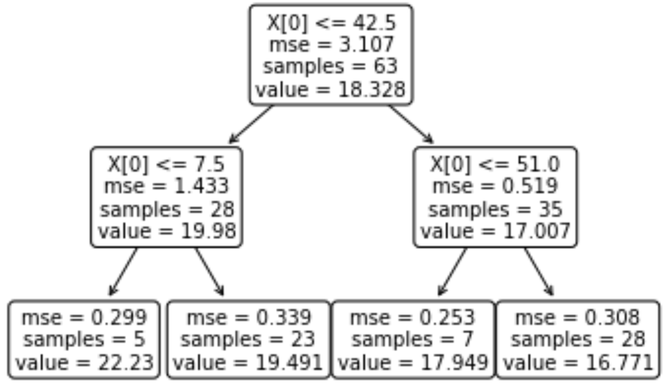


Train error: 0.9617
Test error: 1.1243

REGRESSION TREES

Modeling

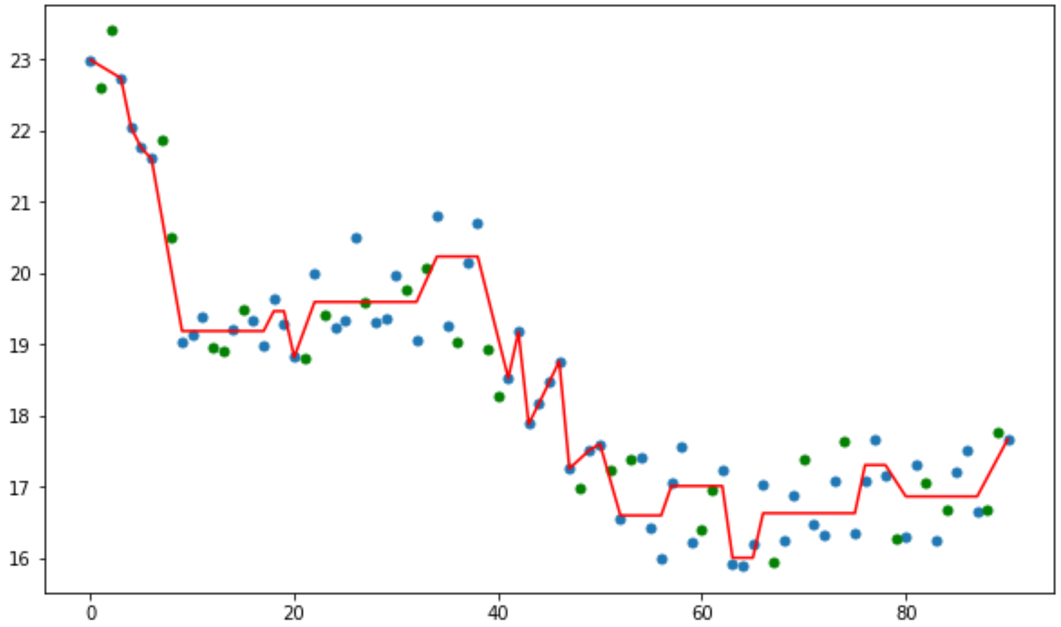


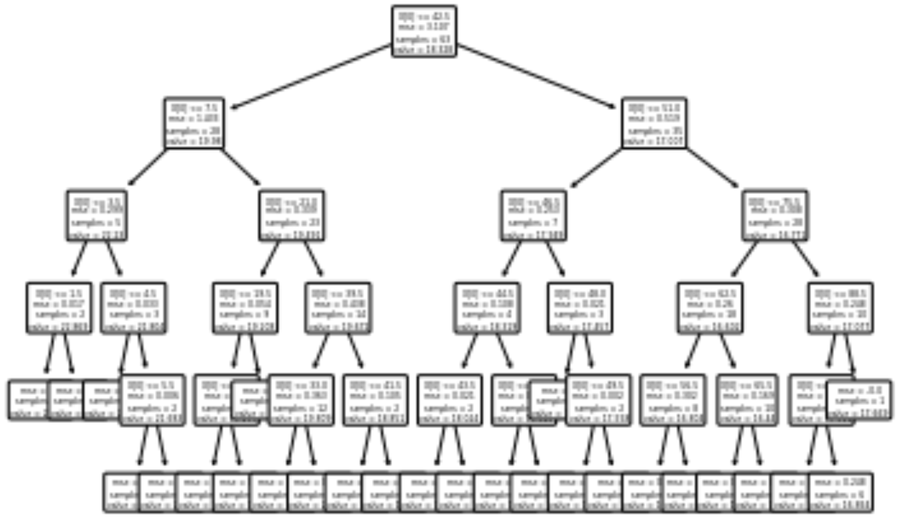


Train error: 0.5589
Test error: 0.6371

REGRESSION TREES

Modeling

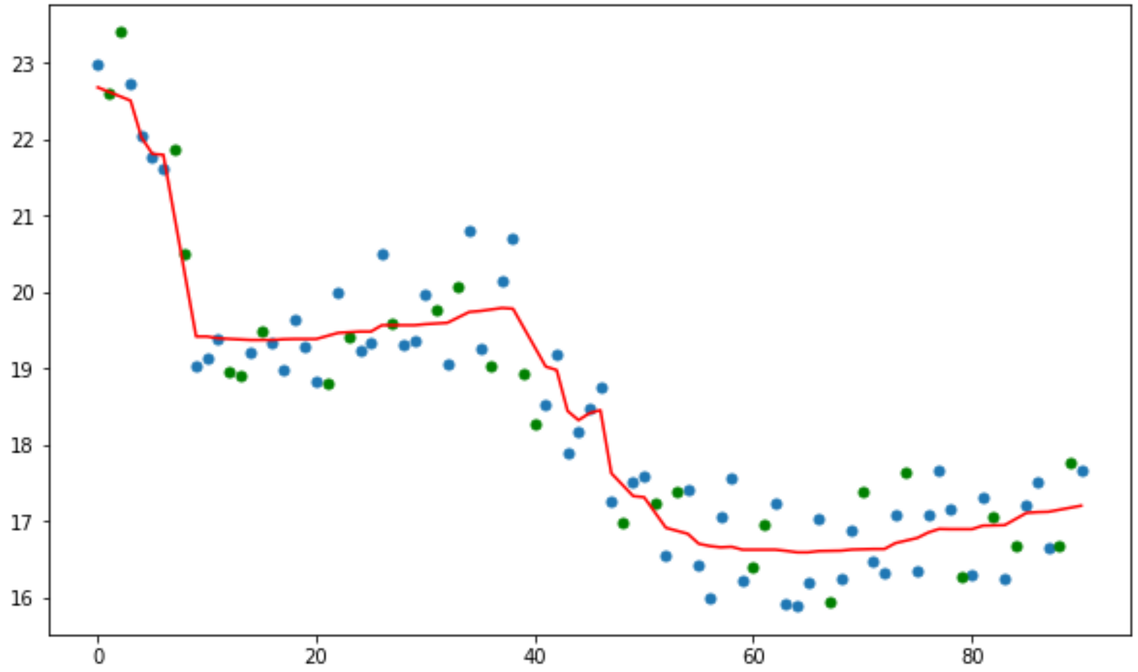


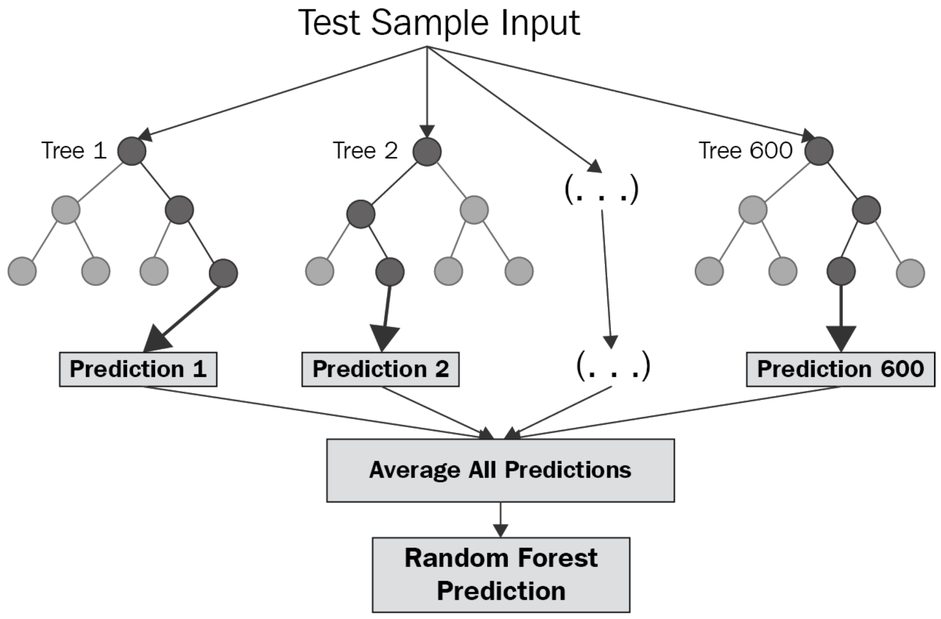


Train error: 0.3586
Test error: 0.6142

ENSEMBLE METHOD: RANDOM FOREST REGRESSION

Modeling



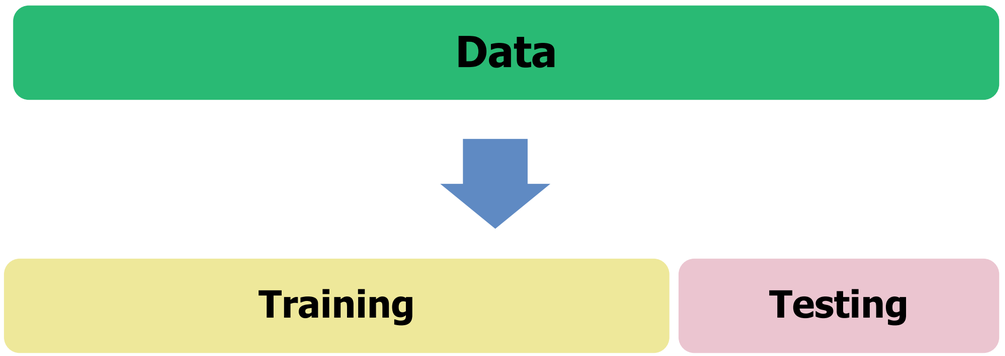


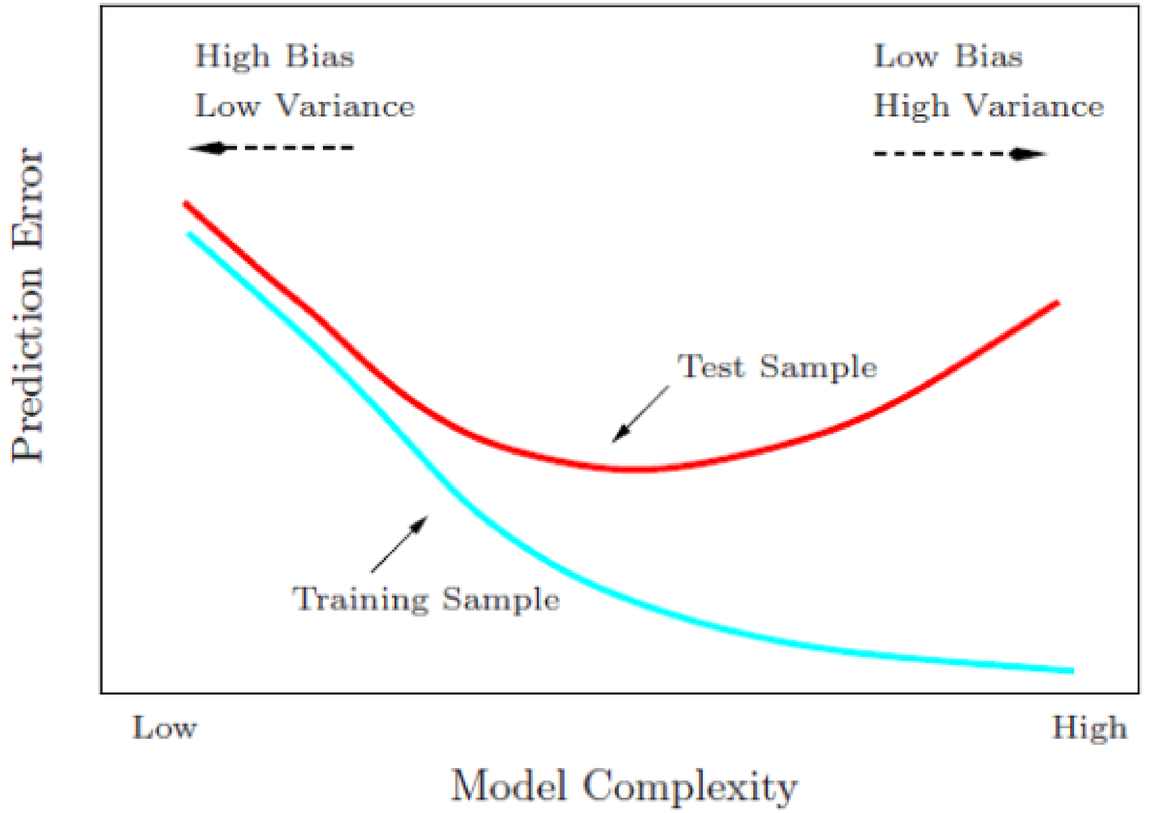
Train error: 0.4538
Test error: 0.5178

MODEL TRAINING AND TESTING

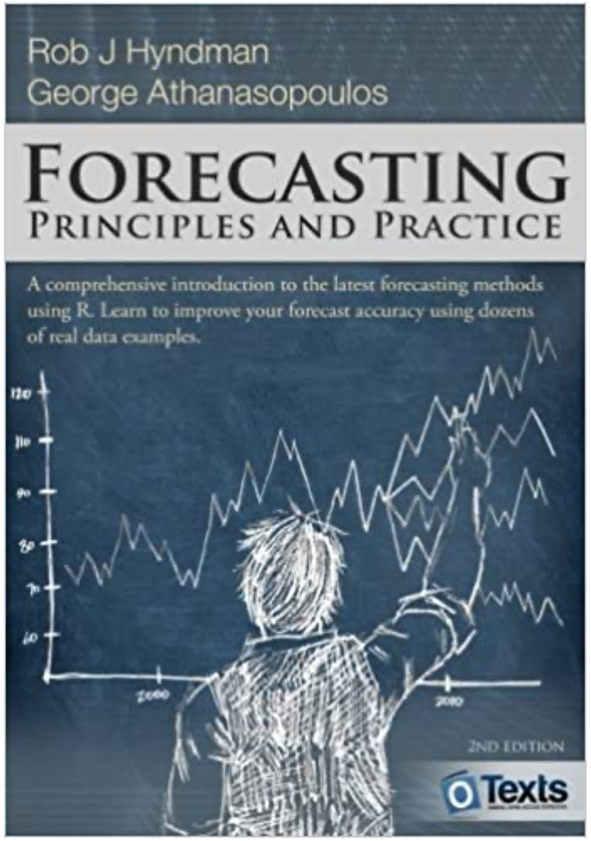
Presentation subtitle

Thematic title of the main text





ONE MORE BOOK



NATIONAL RESEARCH
UNIVERSITY