NATIONAL RESEARCH UNIVERSITY

School of Data Analysis and Artificial Intelligence Department of Computer Science

DATA SCIENCE FOR BUSINESS

Lecture 3. Data Science in Retail.

Moscow, April 22nd, 2022.

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 CPG/FMCG
- Fashion/apparel and department stores
- Specialty retail
- Restaurants, cafes, and fast food

CPG/FMCG

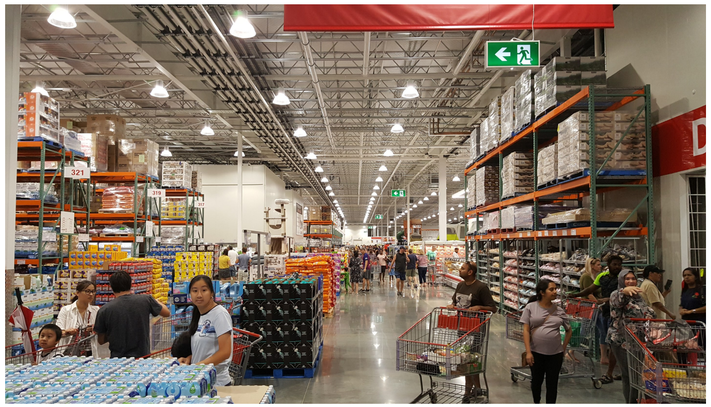
CPG – consumer packaged goods
FMCG – fast-moving consumer goods

Consumer:

- Frequent /regular purchases (daily consumption)
- Low price
- Short shelf life
- Low customer engagement (no effort to choose)

Retailer:

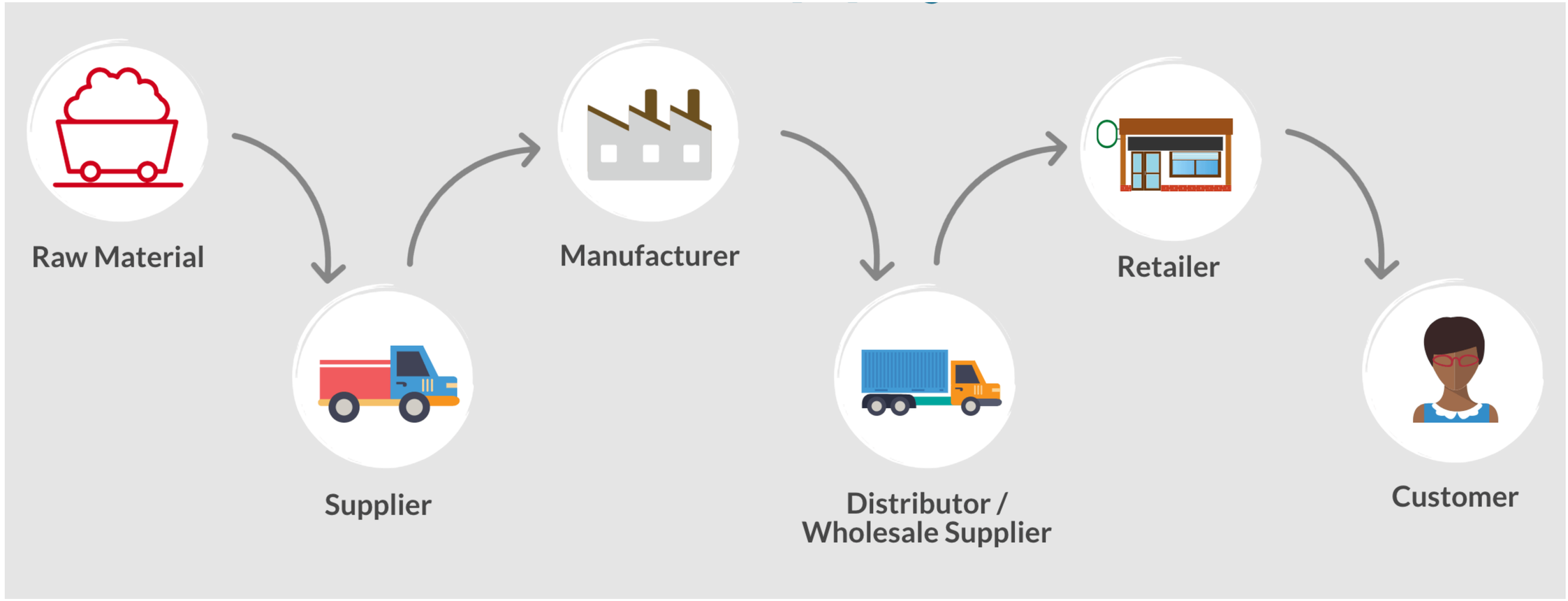
- High volume
- Low margins
- Extensive distribution
- High inventory turnover



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 effectiveness
- Supply chain optimization

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 IN RETAIL

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		

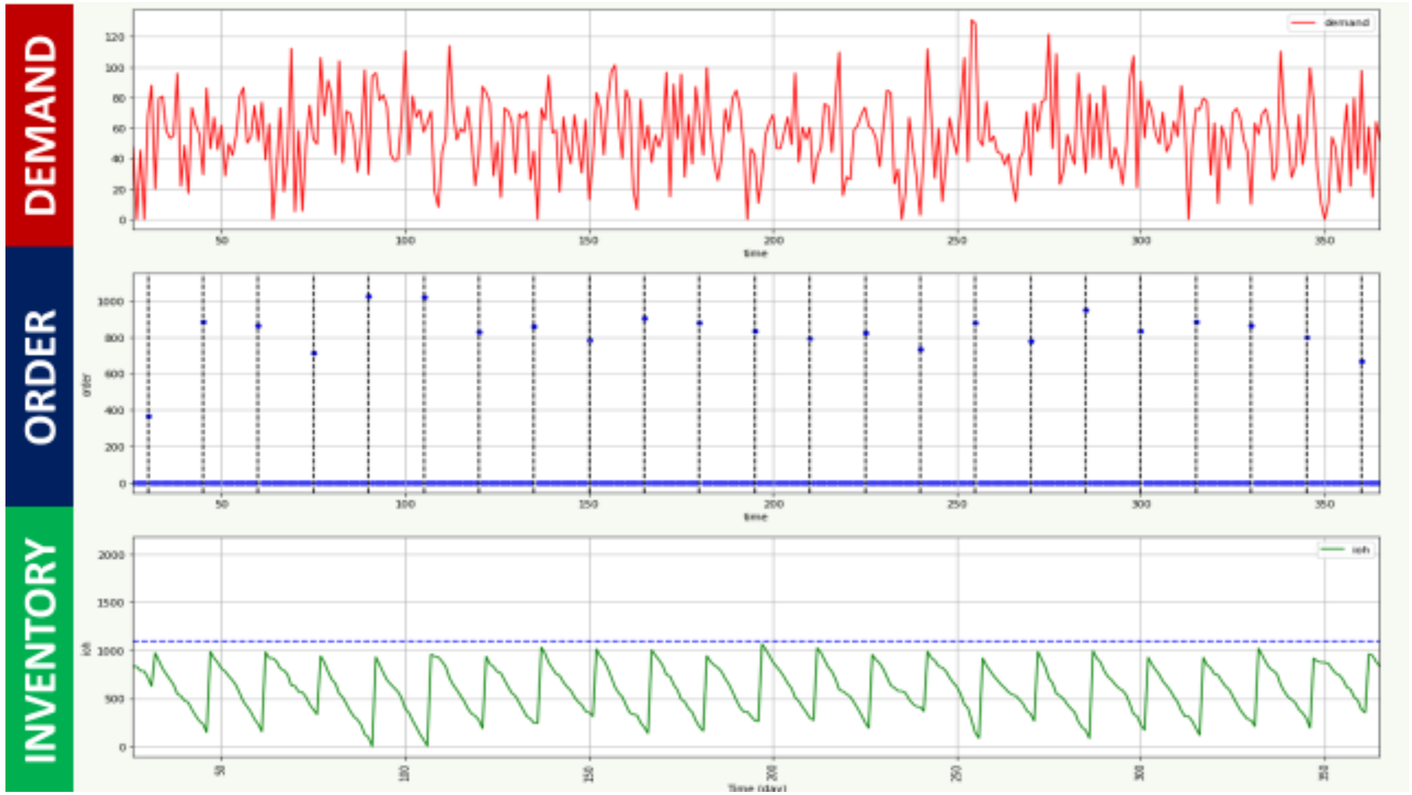
Customer data (CRM)

Customer ID	Date	SKU	Store	Units	Dollars
Demos					

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

INVENTORY MANAGEMENT

Illustrative

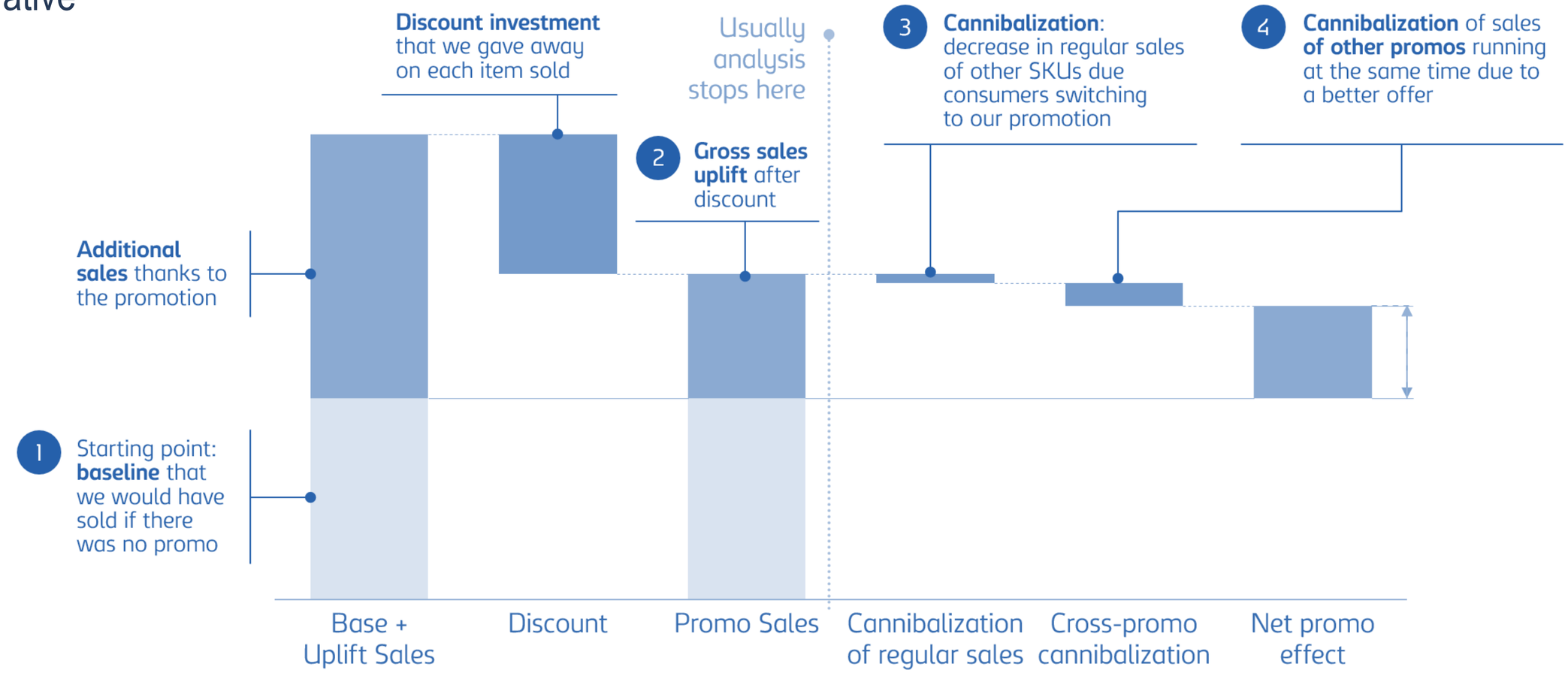


Develop replenishment policy:

- Satisfy store demand
- Minimize :
- Ordering/shipping costs
- Holding costs
- Shortage costs

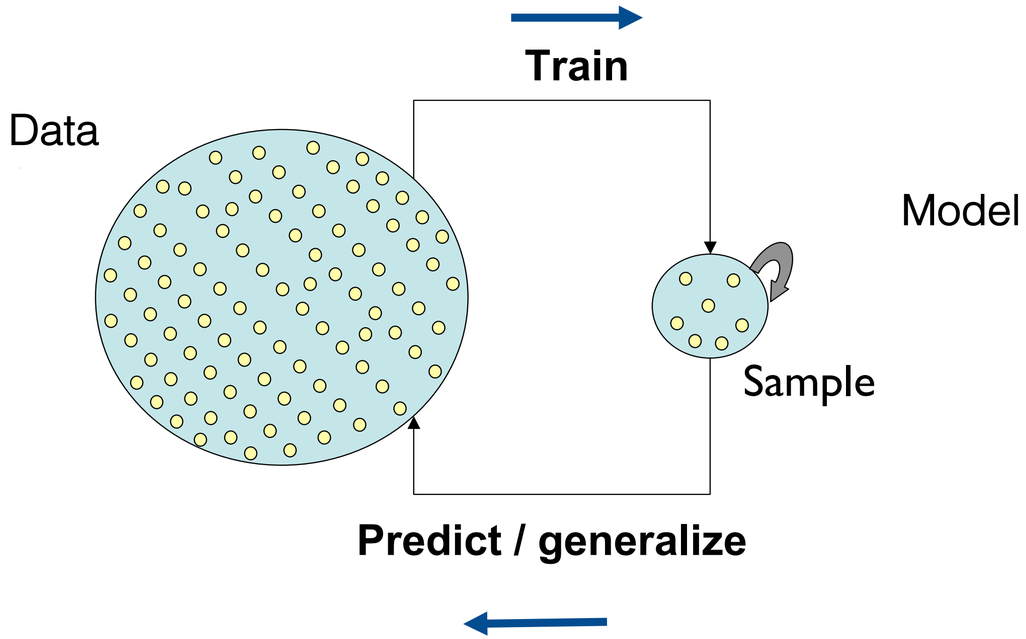
PROMO EFFECTIVENESS

Illustrative

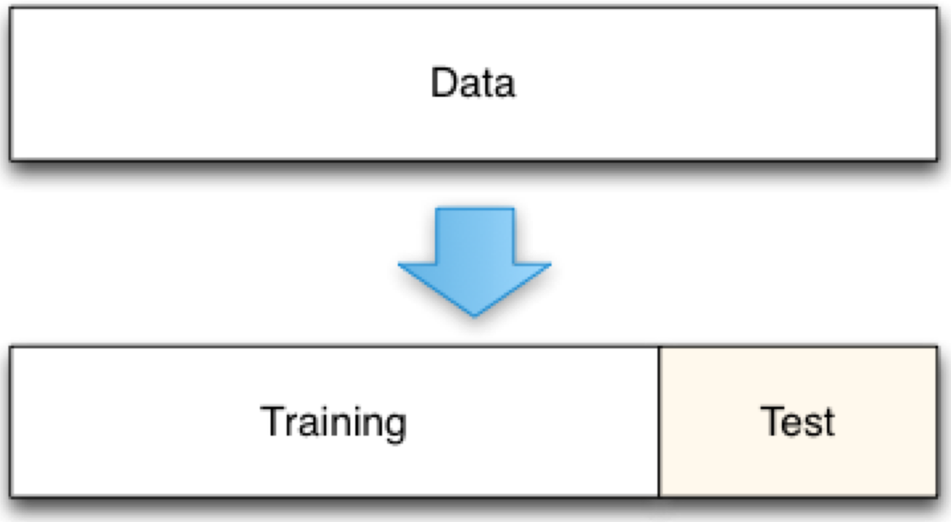


TRAINING AND TESTING

Learning on data



Train & Test split

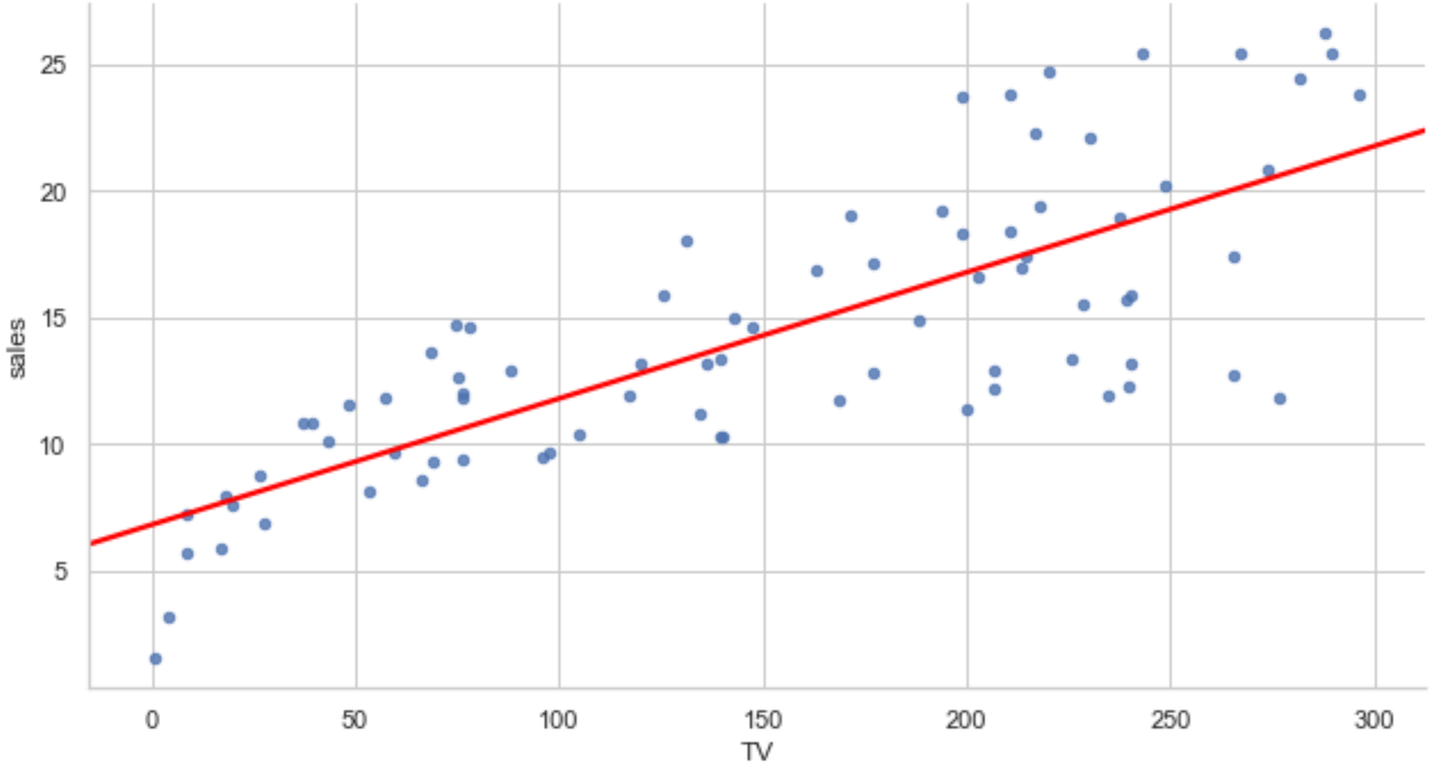


- Build the model
- TRAINING ERROR

- Test the model
- TESTING ERROR

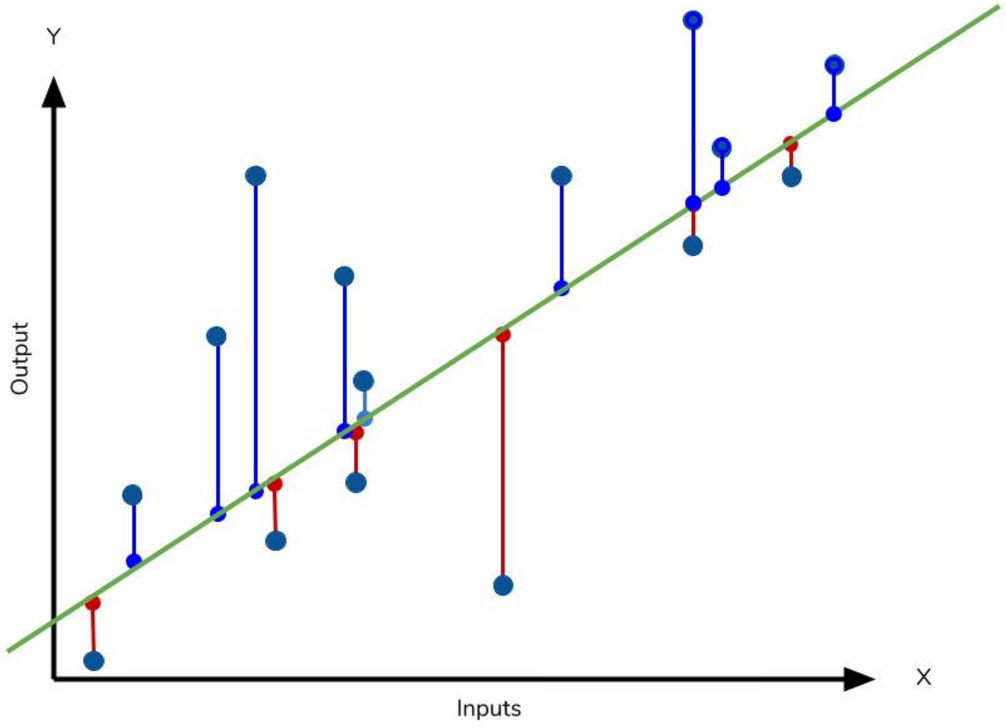
REGRESSION

Quality metrics



REGRESSION EVALUATION

Error metrics/loss function



Standard quality metrics

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

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

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

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

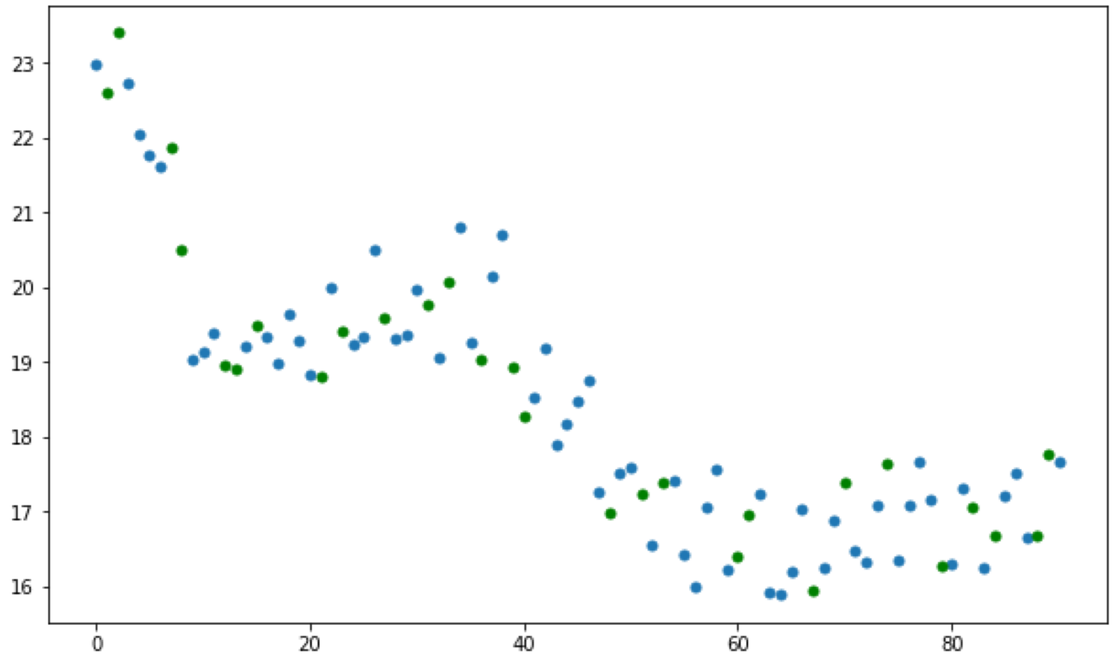
Where,
$\hat{y}$ – predicted value of y
$\bar{y}$ – mean value of y

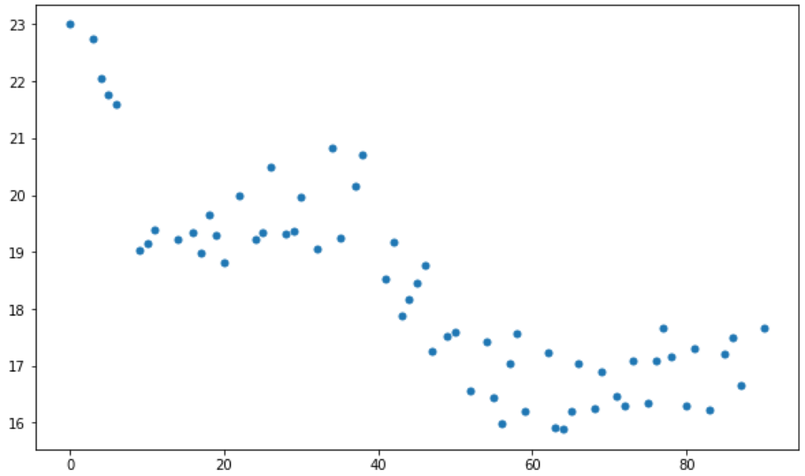
Baseline model (model= mean value), R^2 = 0

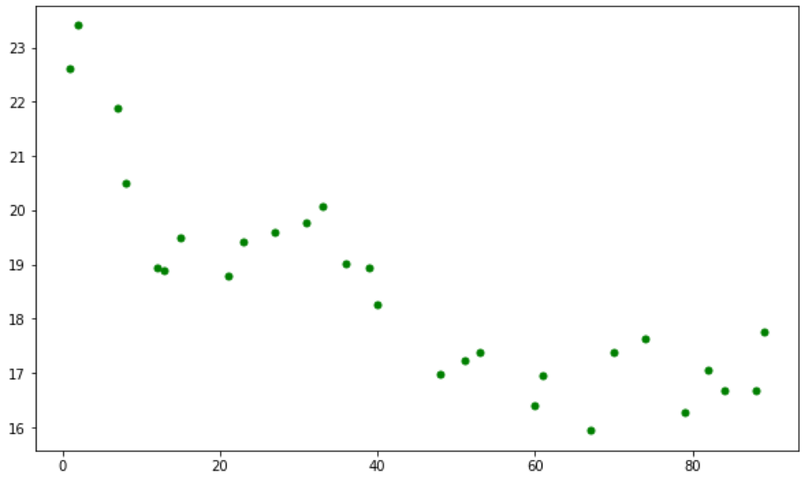
Perfect model (residual =0), R^2 = 1

REGRESSION

Modeling

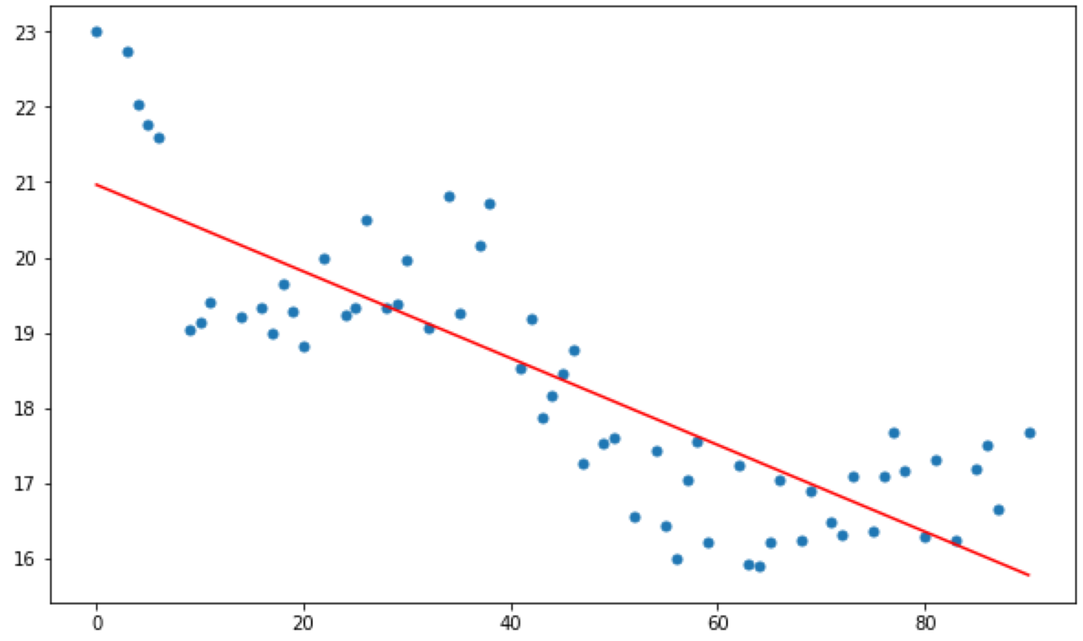




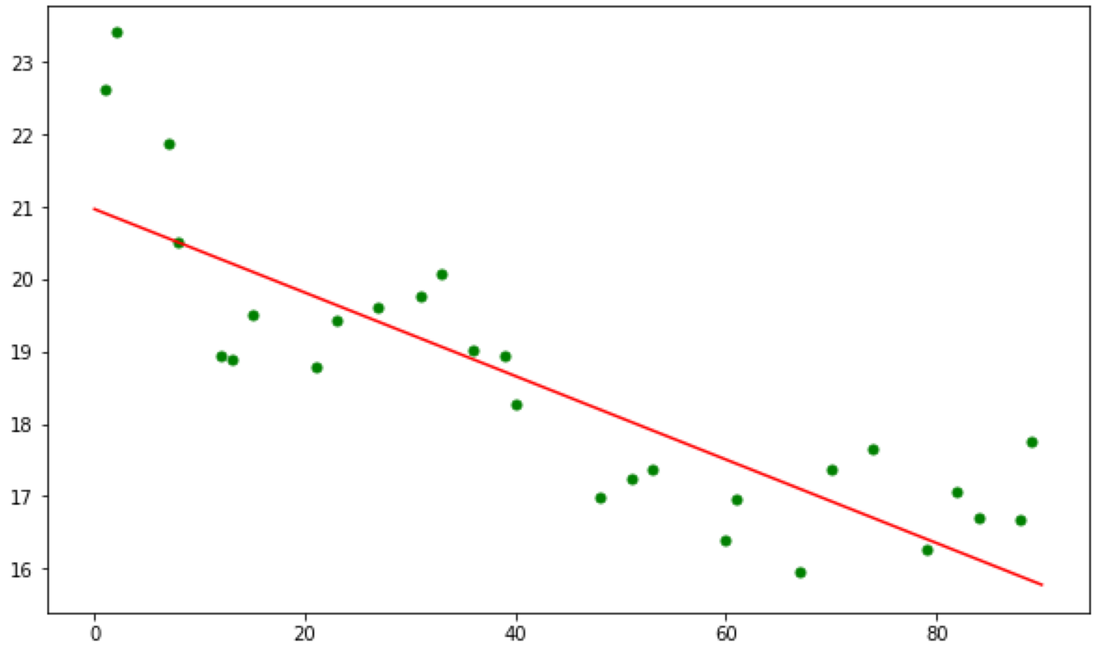


LINEAR REGRESSION

Modeling



Train error: 0.966



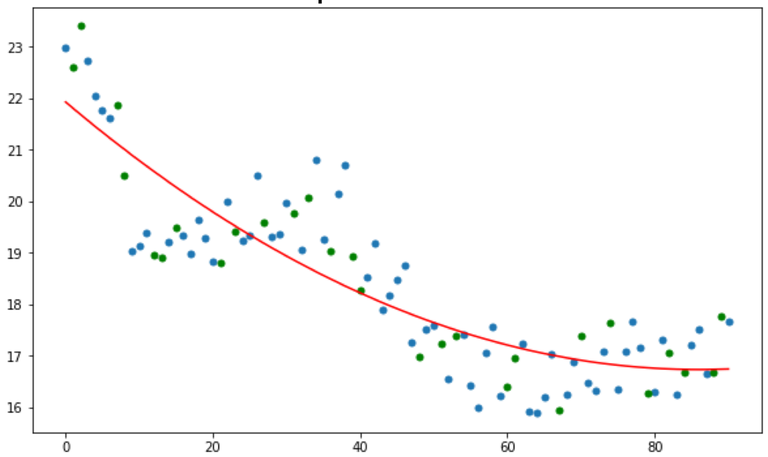
Test error: 1.002

POLYNOMIAL REGRESSION

Modeling

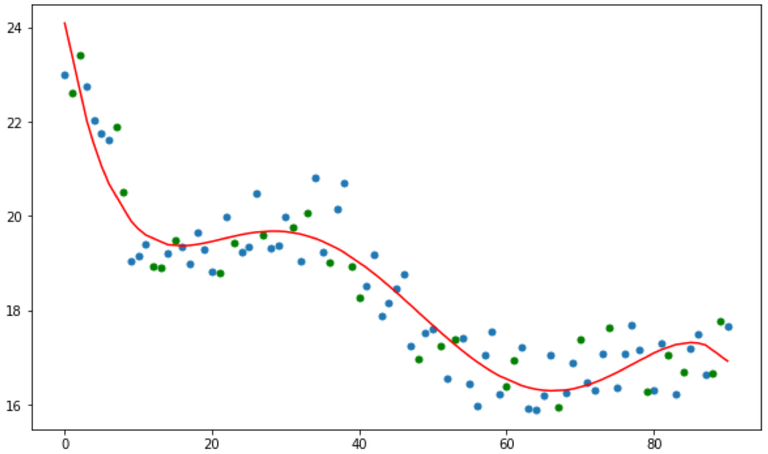
parametric method that estimates the relationship as an n-th degree polynomial

p=2



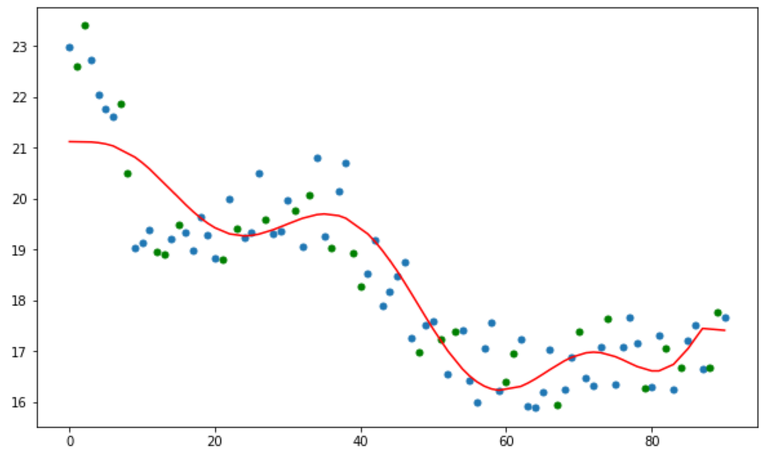
Train error: 0.865
Test error: 0.869

p=5



Train error: 0.583
Test error: 0.682

p=10



Train error: 0.761
Test error: 0.879

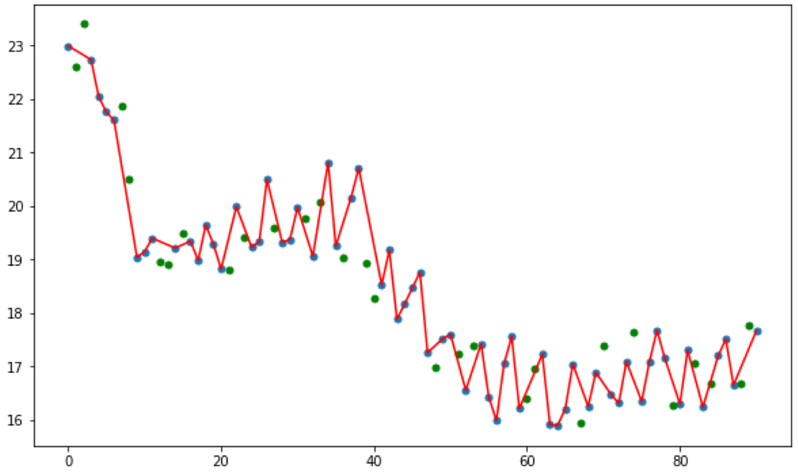
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \varepsilon_i \ (i = 1, 2, \ldots, n)$$

KNN REGRESSION

Modeling

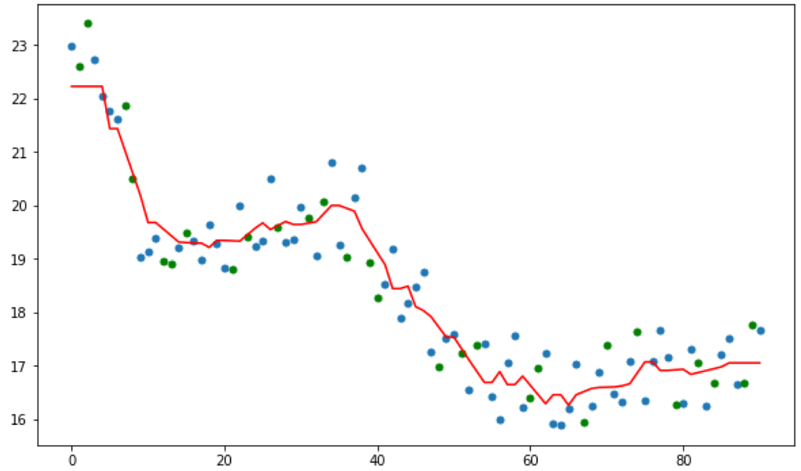
non-parametric method that approximates continuous outcome by averaging the observations in the same neighborhood

k=1



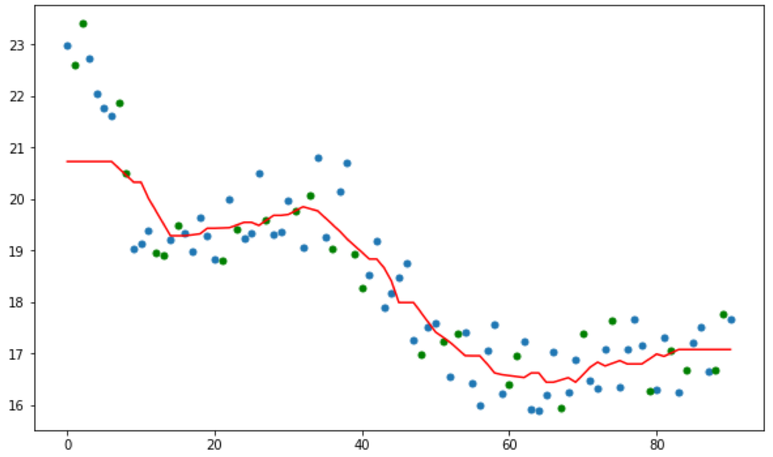
Train error: 0.0
Test error: 0.7574

k=5



Train error: 0.5468
Test error: 0.6248

k=10

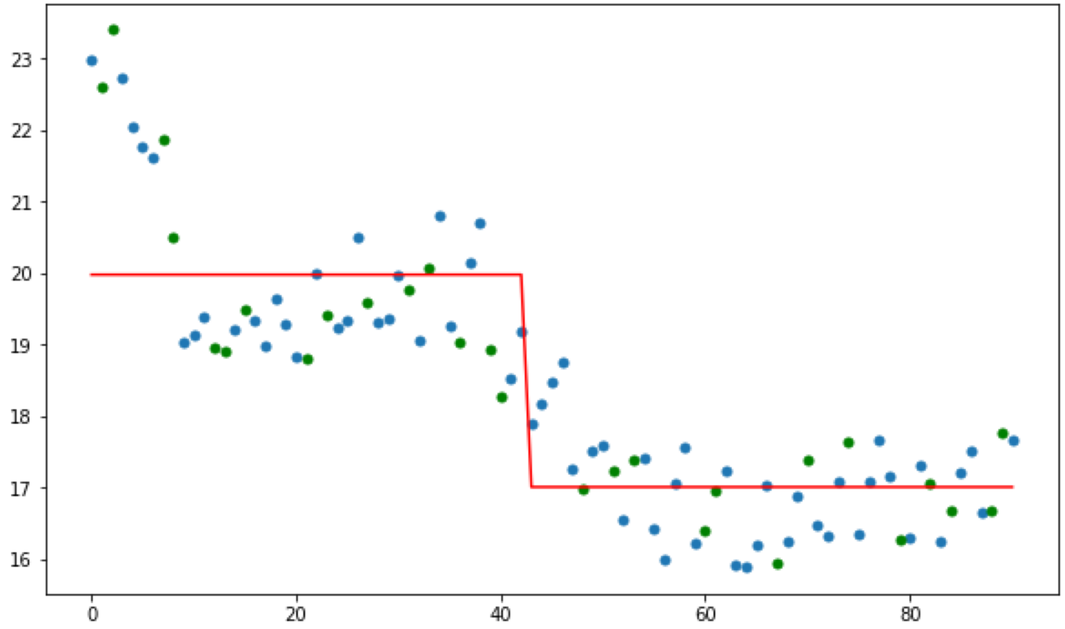


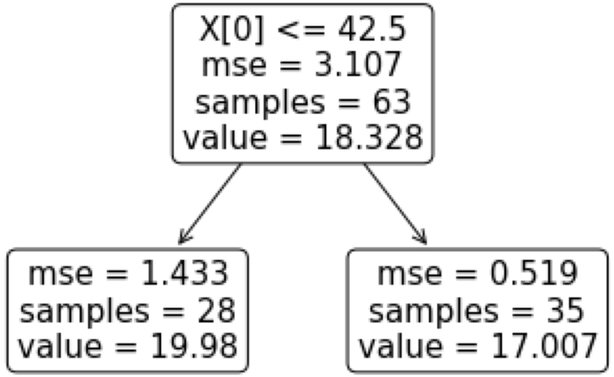
Train error: 0.7399
Test error: 0.8241

REGRESSION TREES

Modeling

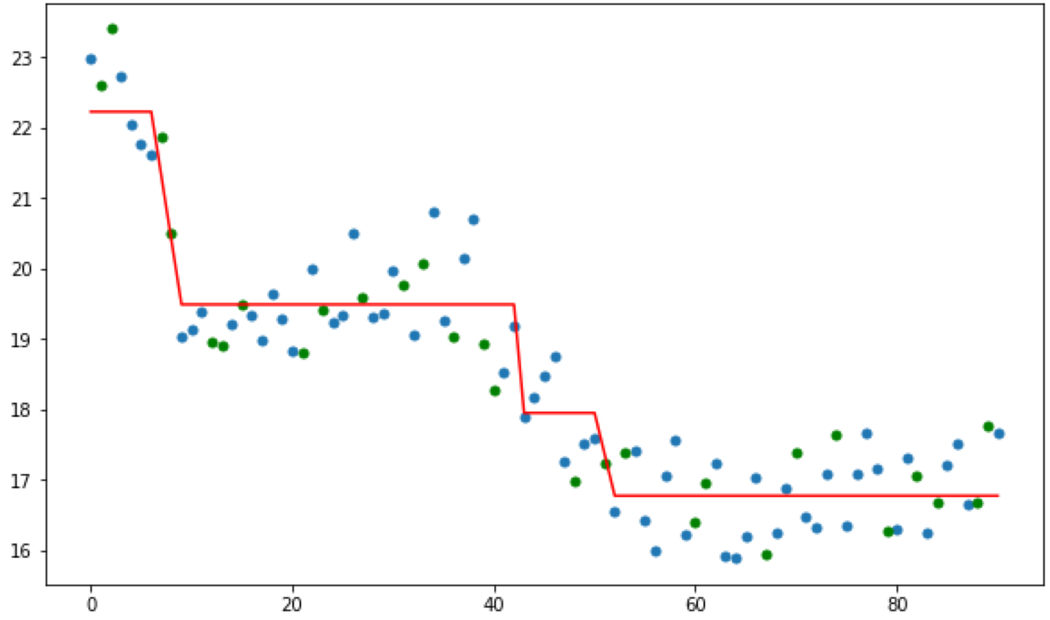
decision trees where the target variable can take continuous values (typically real numbers)

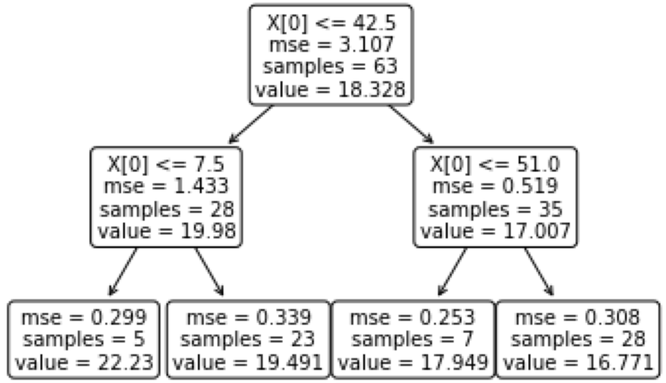




REGRESSION TREES

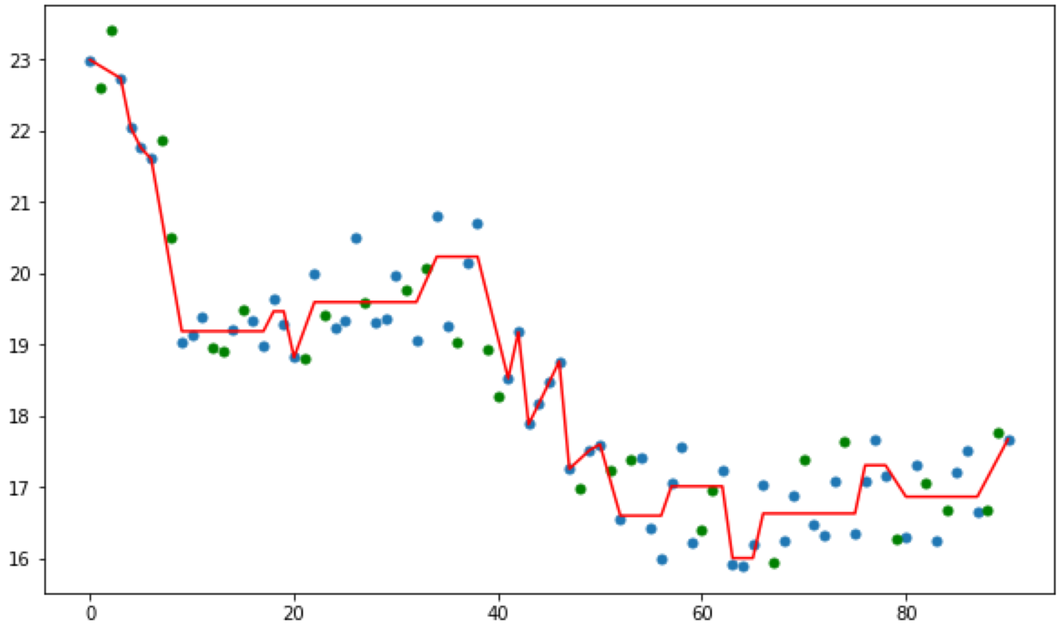
Modeling

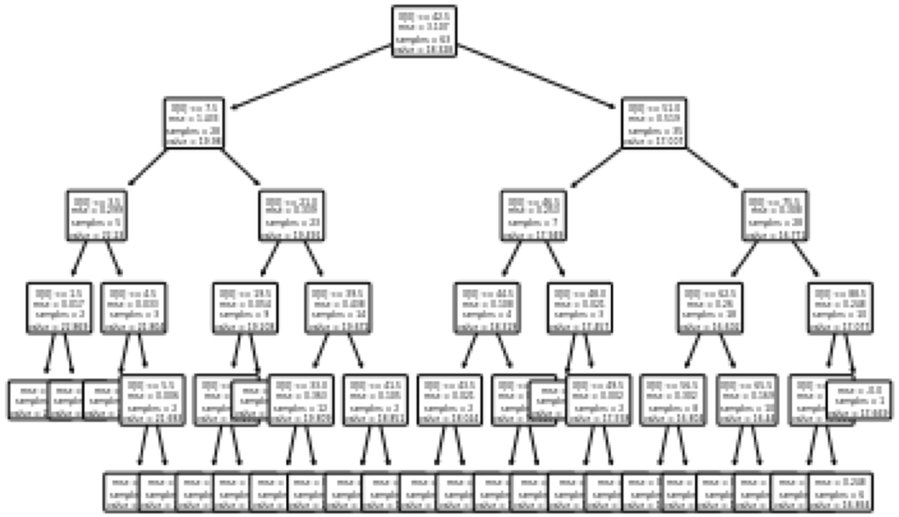




REGRESSION TREES

Modeling

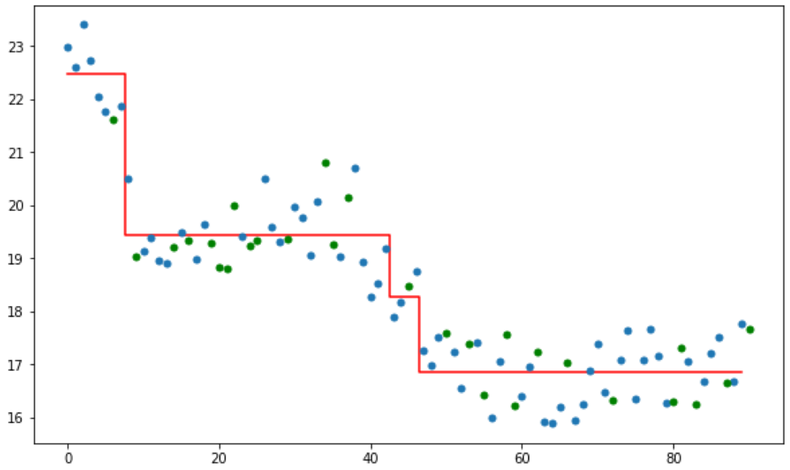




REGRESSION TREES

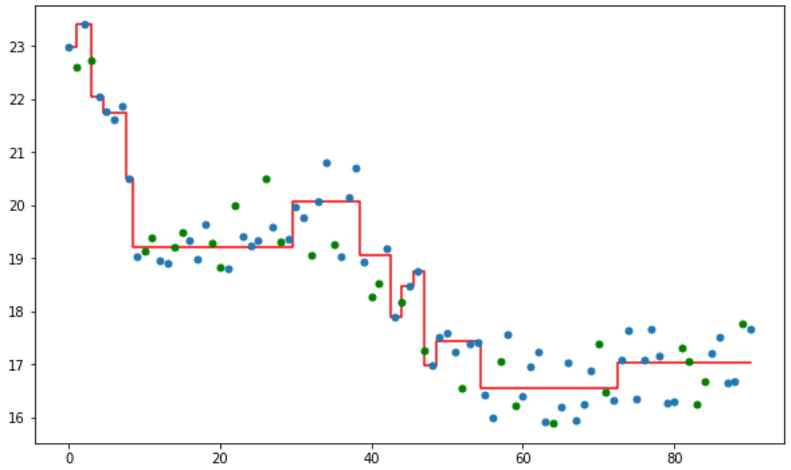
Modeling

depth=2



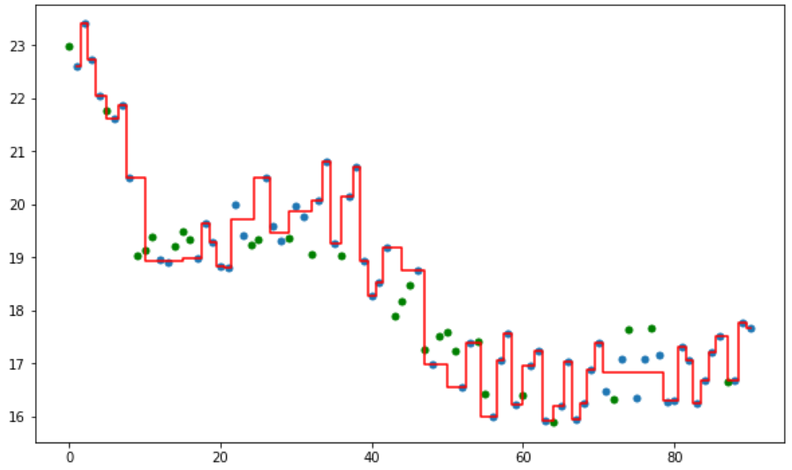
Train error: 0.557
Test error: 0.673

depth=4



Train error: 0.404
Test error: 0.655

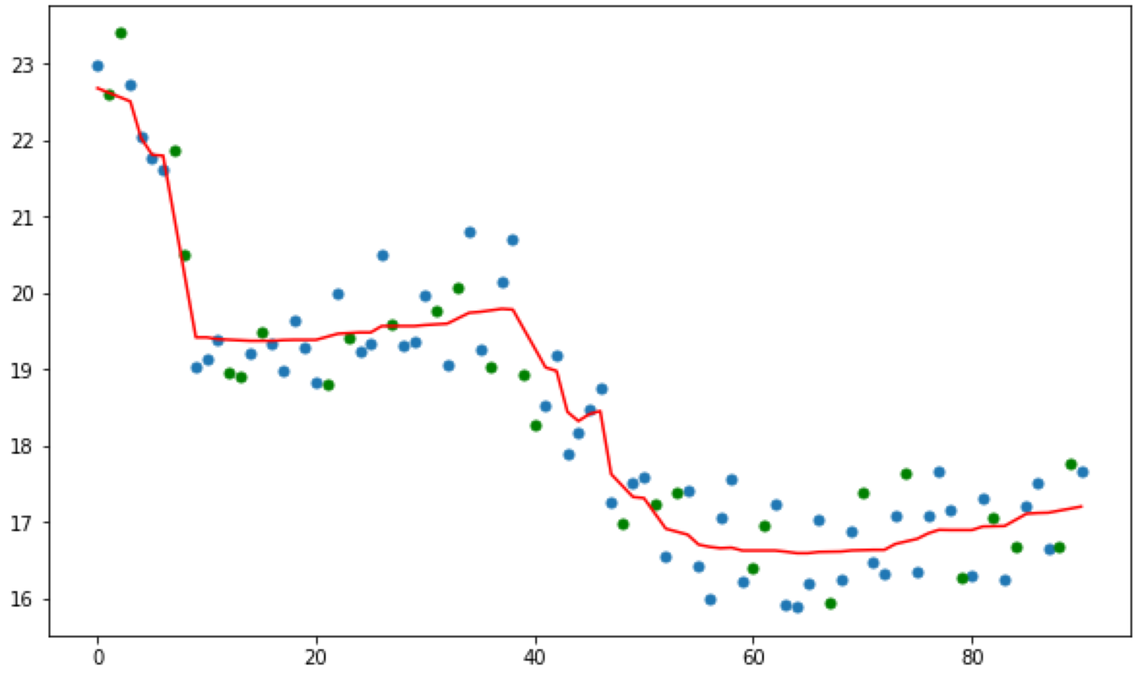
depth=8

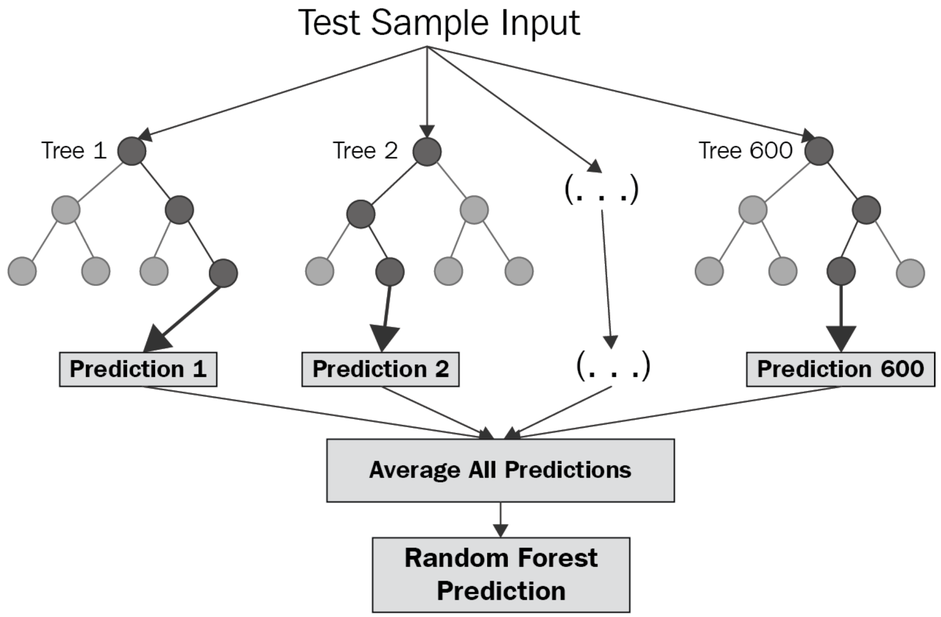


Train error: 0.141
Test error: 0.736

ENSEMBLE METHOD: RANDOM FOREST REGRESSION

Modeling

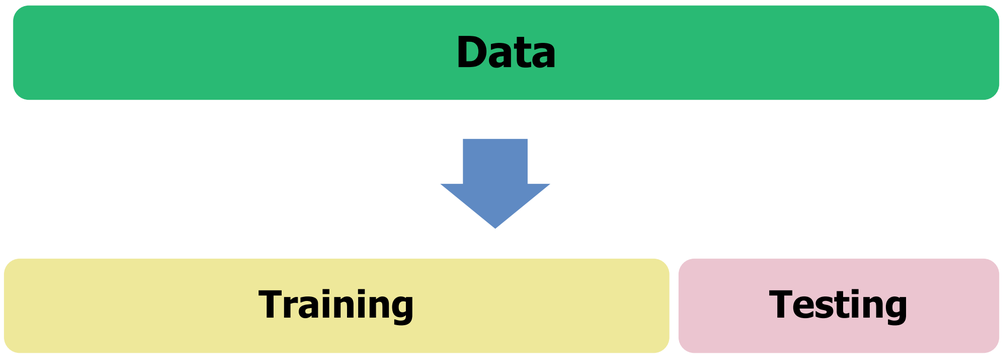


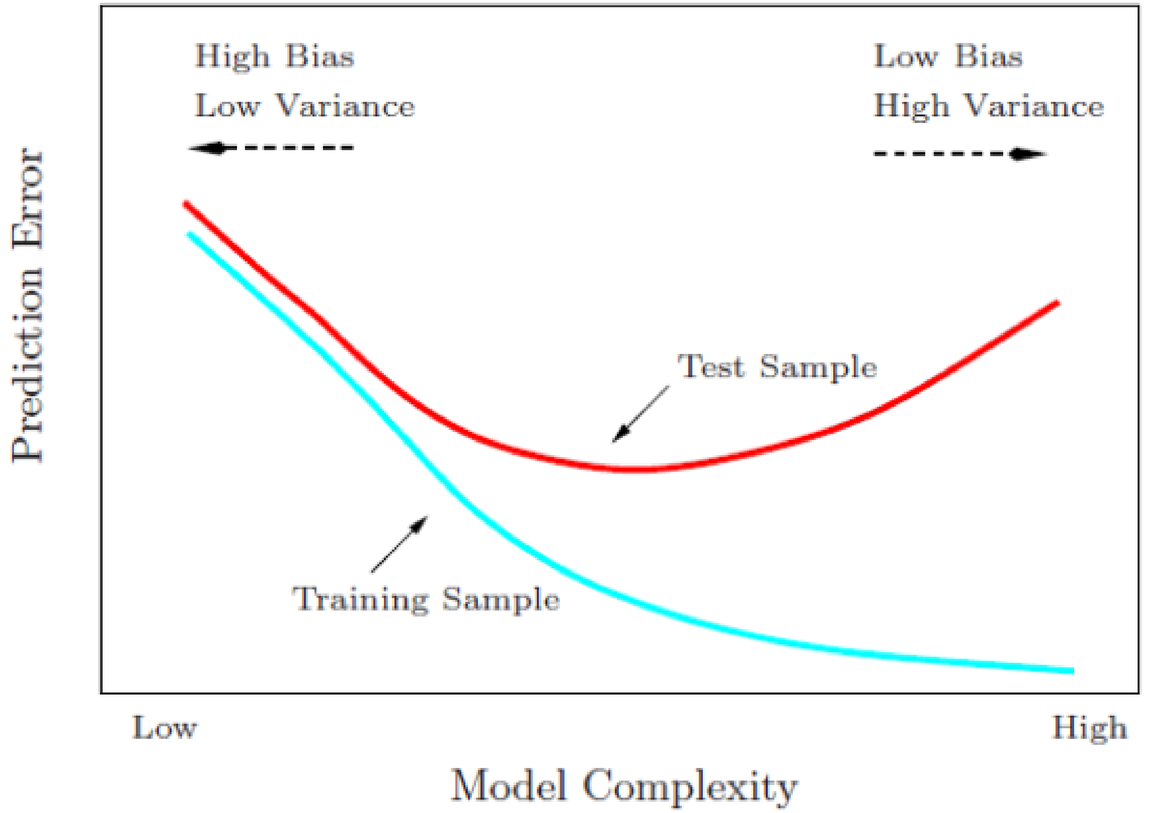


Train error: 0.4538
Test error: 0.5178

MODEL TRAINING AND TESTING

Bias-variance trade-off





NATIONAL RESEARCH
UNIVERSITY