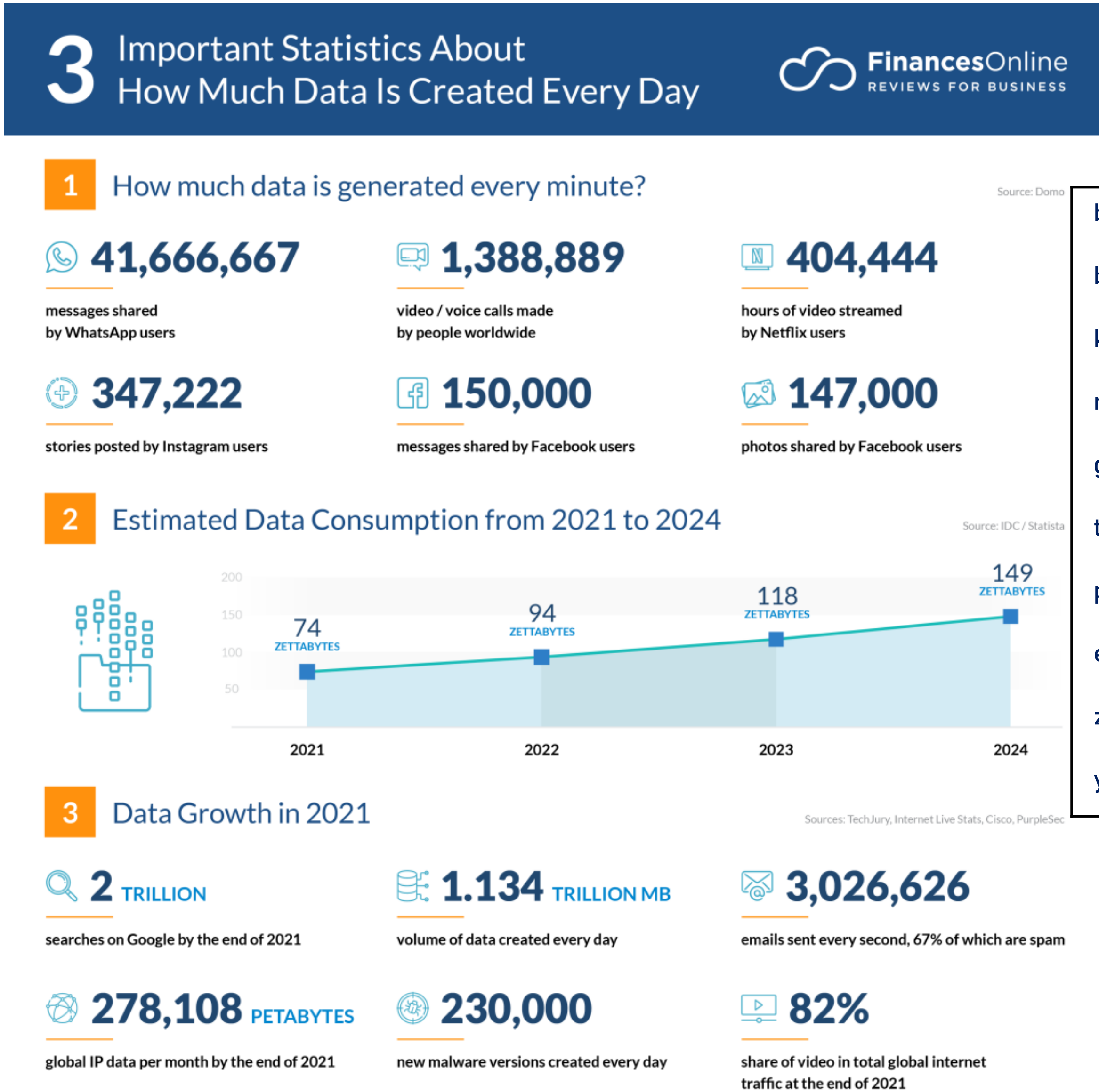
Machine Learning per la Finanza

Docenti: Zelda Marino e Paolo Zanetti

Email: {zelda.marino;paolo.zanetti}@uniparthenope.it

How Much Data Is Created Every Day



bit	b	1
byte	B	8 bit
kilobyte	KB	10^3 bytes
megabyte	MB	10^6 bytes
gigabyte	GB	10^9 bytes
terabyte	TB	10^{12} bytes
petabyte	PB	10^{15} bytes
exabyte	EB	10^{18} bytes
zettabyte	ZB	10^{21} bytes
yottabyte	YB	10^{24} bytes

https://financesonline.com/how-much-data-is-created-every-day/#:~:text=Every%20day%20Big%20Data%20statistics,(2021%2C%20February%209).

What is Machine Learning

- Machine learning is a branch of AI
- The idea underlying machine learning is that we give a computer program access to lots of data and let it learn about relationships between variables and make predictions
- Some of the techniques of machine learning date back to the 1950s but improvements in computer speeds and data storage costs have now made machine learning a practical tool

Software

- There a several alternatives such as Python, R, MatLab, Spark, and Julia
- Need ability to handle very large data sets and availability of packages that implement the algorithms.
- Python seems to be winning at the moment
- Libraries such as Numpy, Pandas, Scikit-Learn (Sklearn), and Tensorflow make it easy to handle large data sets and implement machine learning algorithms in Python

Machine Learning vs. Automation

- Computers have been used to automate many business decisions (payroll, sending out invoices, summarizing sales by region, etc)
- This is digitization: the third industrial revolution
- Machine learning is central to the fourth industrial revolution where computers are used to create intelligence

Example: Loan Applications (digitization vs. ML)

- ❖ If loan officers applied certain known rules we could digitize their activities
- ❖ If we did not know the rules used, we could use ML to determine them
- ❖ But we could go one step further and use ML to improve upon the rules for accepting or rejecting loans

Traditional statistics

- Means, SDs
- Probability distributions
- Significance tests
- Confidence intervals
- Linear regression
- etc

The new world of statistics

- Huge data sets
- Fantastic improvements in computer processing speeds and data storage costs
- Machine learning tools are now feasible
- Can now develop non-linear prediction models, find patterns in data in ways that were not possible before, and develop multi-stage decision strategies
- New terminology: features, labels, activation functions, target, bias, supervised/unsupervised learning……

Traditional Statistics vs Machine Learning

Statistics

Develop Hypothesis

↓

Collect Data

↓

Test Hypothesis

Machine Learning

Collect Data

↓

Try different models

↓

Find Patterns or Develop a Predictive Tool

Types of Machine Learning

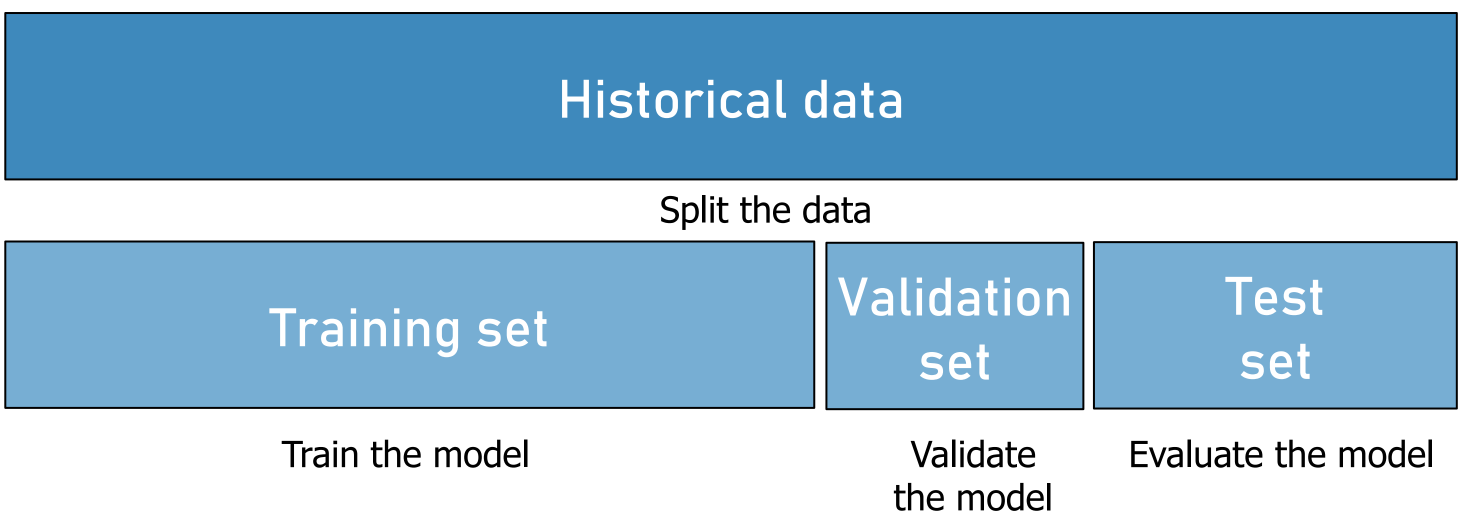
- Unsupervised learning (find patterns)
- Supervised learning (predict numerical value or classification)
- Semi-supervised learning (only part of data has values for, or classification of, target)
- Reinforcement learning (multi-stage decision making)

Applications of ML

- Credit decisions
- Classifying and understanding customers better
- Portfolio management
- Private equity
- Anti-money laundering
- Identifying fraudulent transactions
- Language translation
- Voice recognition
- Biometrics
- etc

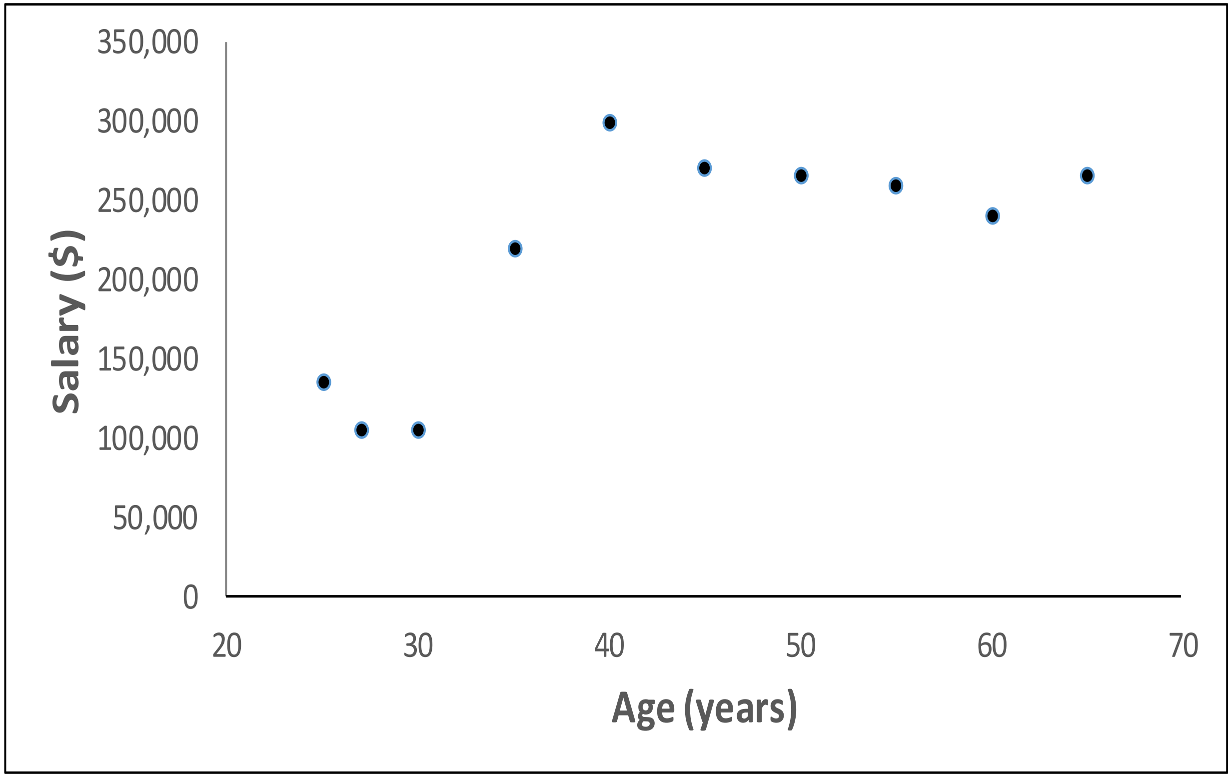
Model evaluation

- Divide data into three sets
  - Training set
  - Validation set
  - Test set
- Develop different models using the training set and examine how well they generalize to new data using the validation set
- Rule of thumb: increase model complexity until model no longer generalizes well to the validation set
- The test set is used to provide a final out-of-sample indication of how well the chosen model works



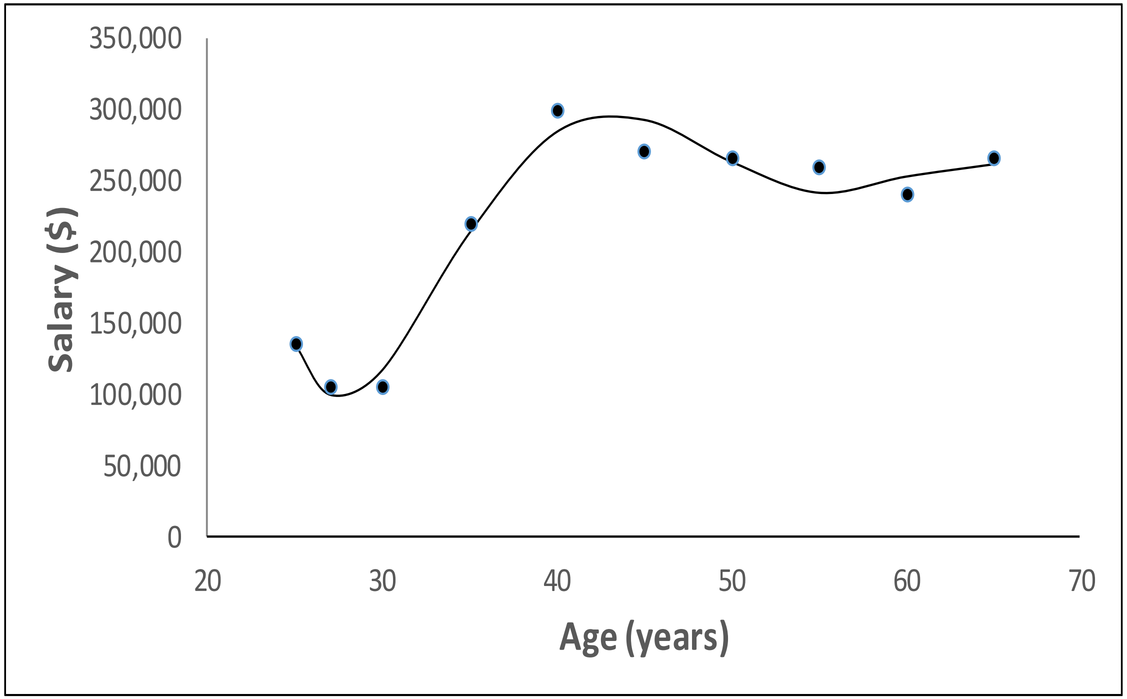
A Baby Data Training Set
(Salary as a function of age for a certain profession in a certain area)

Age (years)	Salary ($)
25	135,000
55	260,000
27	105,000
35	220,000
60	240,000
65	265,000
45	270,000
40	300,000
50	265,000
30	105,000



A Good Fit (Y = Salary, X = Age)

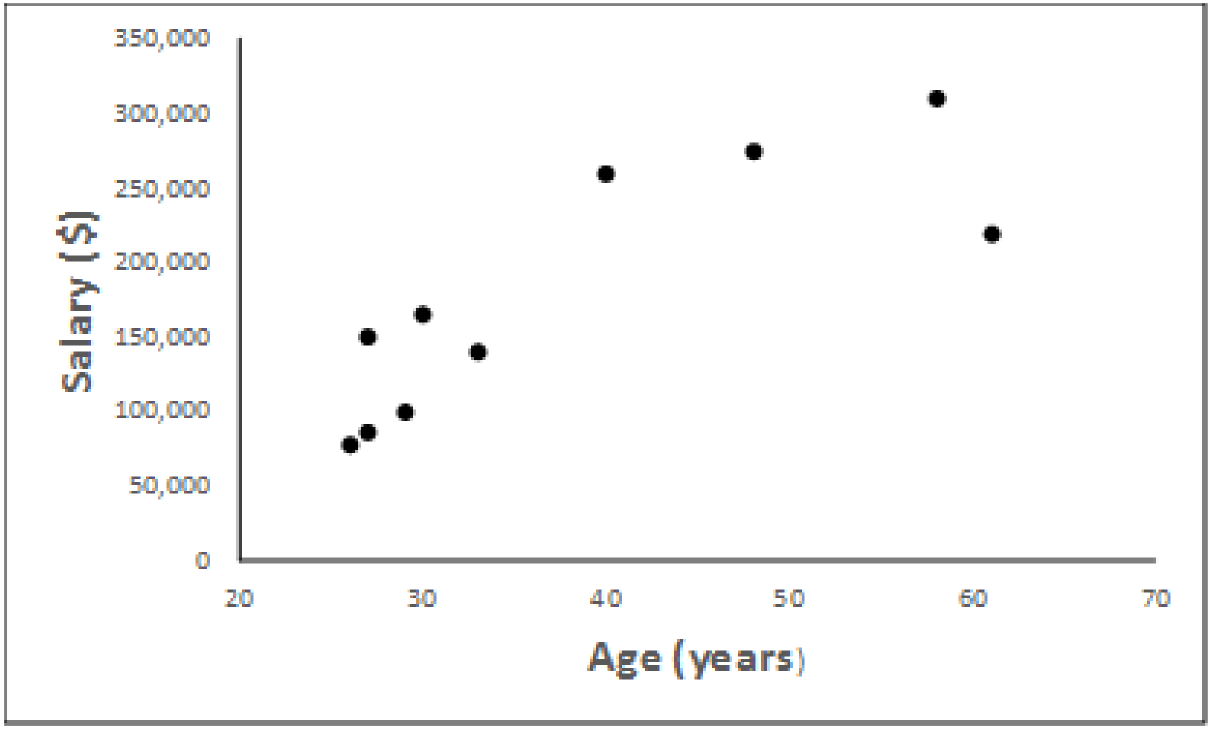
$$Y = a + b_1X + b_2X^2 + b_3X^3 + b_4X^4 + b_5X^5$$



Standard deviation 12.902$ (RMSE root mean square error)

An Out-of-Sample Validation Set

Age (years)	Salary ($)
30	166,000
26	78,000
58	310,000
29	100,000
40	260,000
27	150,000
33	140,000
61	220,000
27	86,000
48	276,000

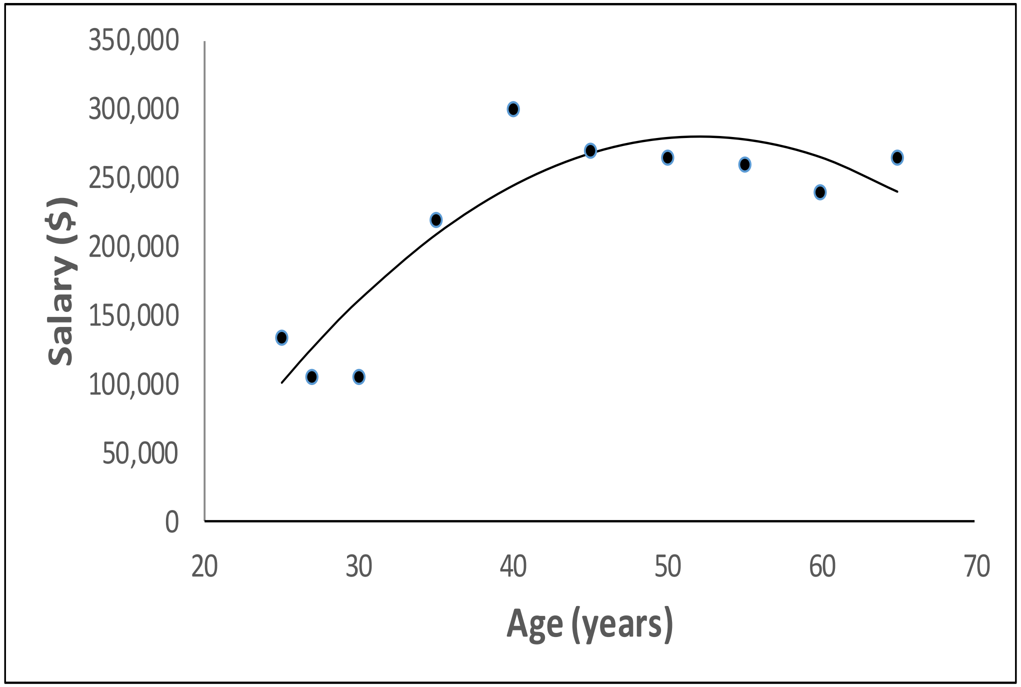


The Fifth Order Polynomial Model Does Not Generalize Well

- The root mean squared error (rmse) for the training data set is $12,902
- The rmse for the validation data set is $38,794
- We conclude that the model overfits the data

Quadratic Model for Baby Data Set

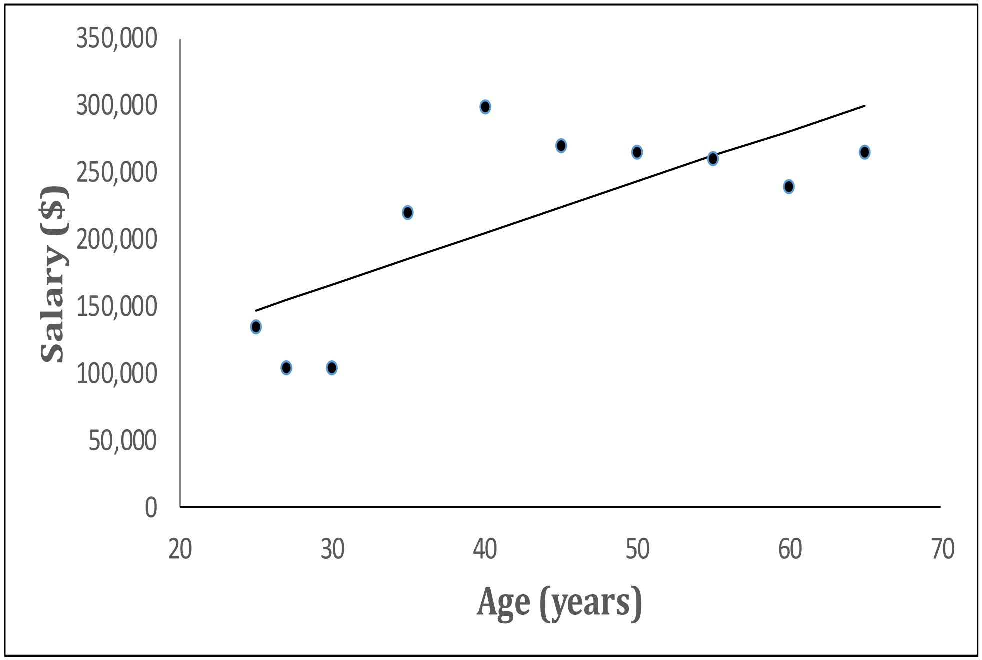
❖ $Y = a + b_1 X + b_2 X^2$



Standard deviation 32.932$ (RMSE root mean square error)
Standard devian on validation set 38.794$

Linear Model for Baby Data Set

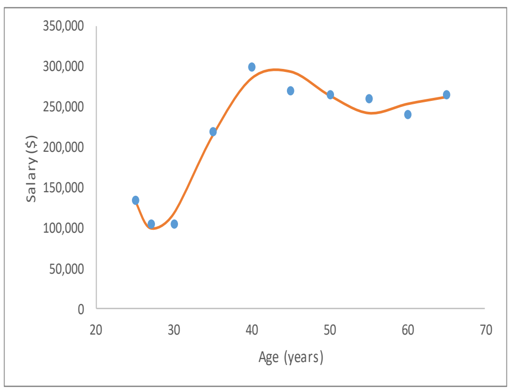
$$Y = a + b_1 X$$

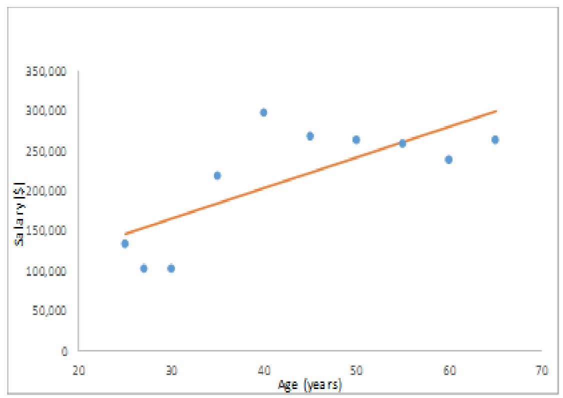


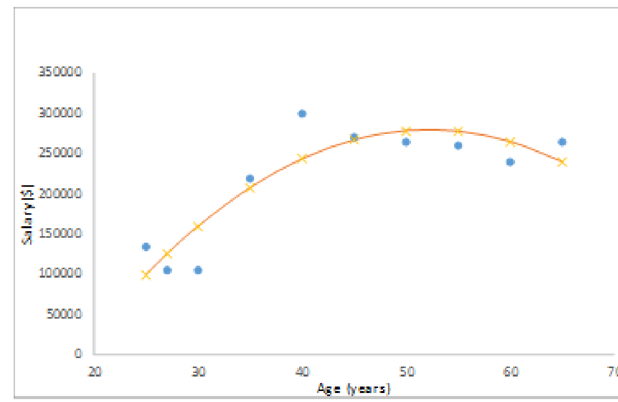
The linear model under-fits while the 5^{th} degree polynomial

	Polynomial of degree 5	Quadratic model	Linear model
Training set	12, 902	32,932	49,731
Validation set	38,794	33,554	49,990

Overfitting/Underfitting;
predicting salaries for people in a certain profession in a certain area







Overfitting

Underfitting

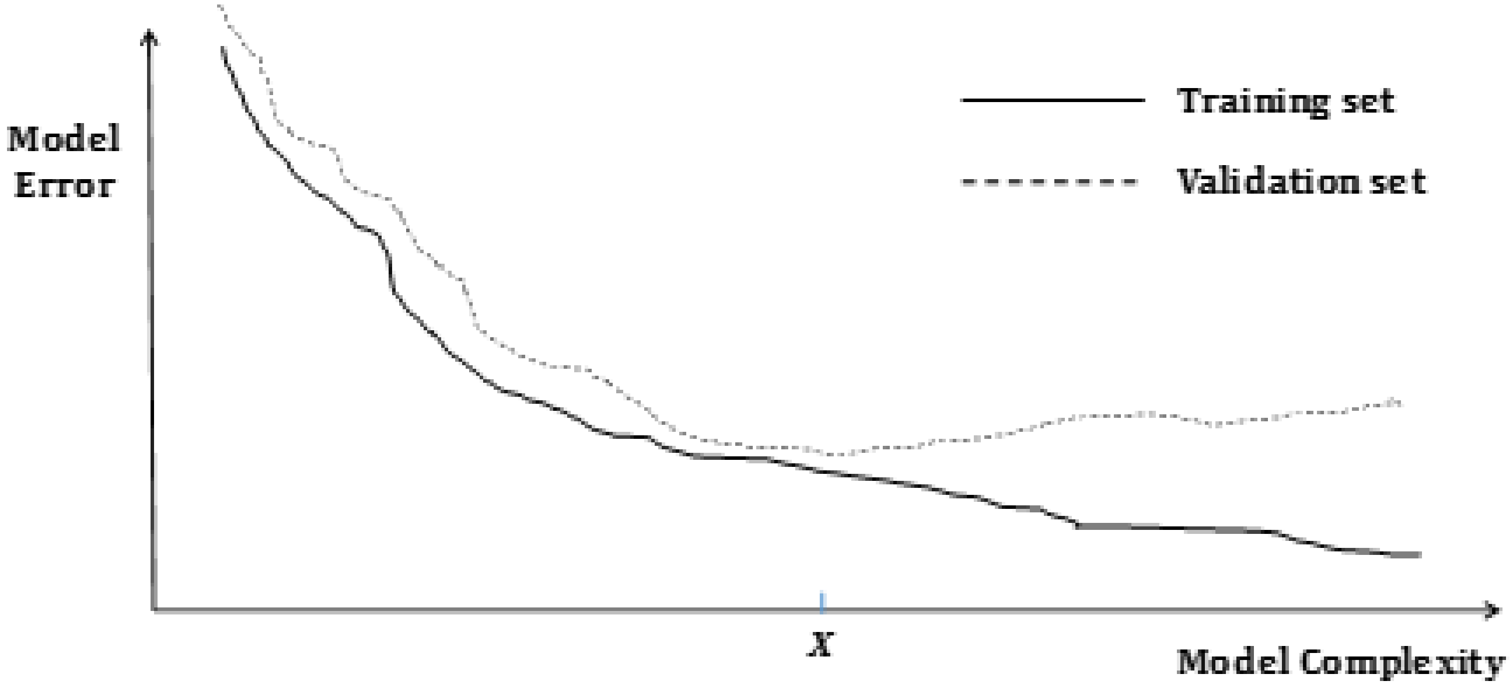
Best model?

Test Set Results for Quadratic Model

Age (years)	Salary ($)	Predicted salary ($)	Error ($)
26	110,000	113,172	−3,172
52	278,000	279,589	−1,589
38	314,000	232,852	+83,148
60	302,000	264,620	+37,380
64	261,000	245,457	+15,543
41	227,000	249325	−22,325
34	200,000	199,411	+589
46	233,000	270,380	−37,380
57	311,000	273,883	−37,117
55	298,000	277,625	+20,375

SD of error is $34,273

Typical Pattern of Errors for Training Set and Validation Set

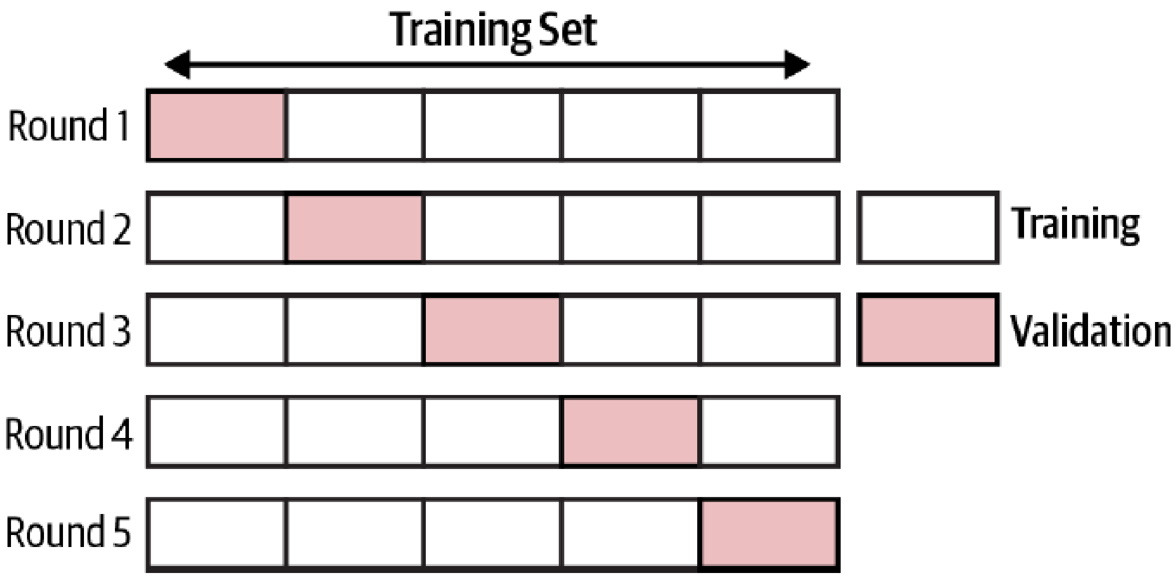


Bias-variance trade-off

❖ Bias refers to error caused by underfitting

❖ Variance refers to errors caused by overfitting

Cross Validation



Data Cleaning

- ❖ Dealing with inconsistent recording
- ❖ Removing unwanted observations
- ❖ Removing duplicates
- ❖ Investigating outliers
- ❖ Dealing with missing items

Tidy Data

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

Name	Age	Size	Country
Sara	"26"	1.78	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Tidy Data

Name	Age	Size	Country
Sara	"26"	1.78	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

Tidy Data

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

Homogeneity

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Homogeneity

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

Data types

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien		1.80	"FR"

Data types

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien	28	1.80	"FR"

Missing values

Reasons:
- Data entry
- Error
- Valid missing value

Solutions:
- impute
- drop
- keep

Bayes Theorem

$$P(Y|X) = \frac{P(X\&Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y|X) = \frac{P(X \text{ and } Y)}{P(X)} \qquad P(X|Y) = \frac{P(X \text{ and } Y)}{P(Y)}$$

$$P(X \text{ and } Y) = P(Y|X)P(X) \qquad P(X \text{ and } Y) = P(X|Y)P(Y)$$

Example

We observe that 90% of fraudulent transactions are for large amounts late in the day. Also 3% of transactions are for large amounts late in the day and 1% of transactions are fraudulent

$$P(\text{fraud}|\text{large\&late}) = \frac{P(\text{large\&late}|\text{fraud})P(\text{fraud})}{P(\text{large\&late})} = \frac{0.9 \times 0.01}{0.03} = 0.3$$

Bayes can be counterintuitive

- One person in ten thousand has a certain disease
- A test is 99% accurate (i.e., if person has the disease the test gets this right 99% of the time; similarly when the person does not have the disease the test is right 99% of the time)
- You test positive
- What is the chance that you have the disease?
- X=test positive, Y=has disease, $\bar{Y}$= does not have disease
- $P(X|Y) = 0.99;\ P(Y) = 0.0001$
- $P(X) = P(X|Y)P(Y) + P(X|\bar{Y})P(\bar{Y}) = 0.99 \times 0.0001 + 0.01 \times 0.9999 = 0.0101$
- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{0.99 \times 0.0001}{0.0101} = 0.0098$