ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Ενότητα 3 – Εισαγωγή στη Μηχανική Μάθηση

Μέθοδοι Μηχανικής Μάθησης στα Χρηματοοικονομικά

Αθανάσιος Σάκκας, Επ. Καθηγητής, ΟΠΑ

Τι είναι η Μηχανική Μάθηση

- Η μηχανική μάθηση είναι κλάδος της τεχνητής νοημοσύνης.
- Η ιδέα της μηχανικής μάθησης είναι ότι δίνουμε σε ένα πρόγραμμα υπολογιστή πρόσβαση σε πολλά δεδομένα και του αφήνουμε να μάθει για τις σχέσεις μεταξύ μεταβλητών και να κάνει προβλέψεις.
- Μερικές από τις τεχνικές της μηχανικής μάθησης χρονολογούνται από τη δεκαετία του 1950, αλλά οι βελτιώσεις στις ταχύτητες υπολογιστών και το κόστος αποθήκευσης δεδομένων έχουν πλέον κάνει τη μηχανική μάθηση ένα πρακτικό εργαλείο.

Software

- Υπάρχουν πολλές εναλλακτικές όπως Python, R, MatLab, Spark και Julia.

- Απαιτείται ικανότητα χειρισμού πολύ μεγάλων συνόλων δεδομένων και διαθεσιμότητα πακέτων που υλοποιούν τους αλγόριθμους.

- Η Python φαίνεται να κερδίζει αυτή τη στιγμή.

- Βιβλιοθήκες όπως οι Numpy, Pandas, Scikit-Learn (Sklearn) και Tensorflow διευκολύνουν το χειρισμό μεγάλων συνόλων δεδομένων και την εφαρμογή αλγορίθμων μηχανικής μάθησης στην Python.

Machine Learning (ML) vs. Automation

- Οι υπολογιστές έχουν χρησιμοποιηθεί για την αυτοματοποίηση πολλών επιχειρηματικών αποφάσεων (μισθοδοσία, αποστολή τιμολογίων, σύνοψη πωλήσεων ανά περιοχή, κ.λπ.).

- Αυτή είναι η ψηφιοποίηση: η τρίτη βιομηχανική επανάσταση.

- Η μηχανική μάθηση είναι κεντρική στην τέταρτη βιομηχανική επανάσταση όπου οι υπολογιστές χρησιμοποιούνται για τη δημιουργία (τεχνητής) νοημοσύνης.

Παράδειγμα: Αιτήσεις δανείου (Ψηφιοποίηση vs. ML)

- Εάν οι υπάλληλοι δανείων εφάρμοζαν ορισμένους γνωστούς κανόνες, θα μπορούσαμε να ψηφιοποιήσουμε τις δραστηριότητές τους.

- Εάν δεν γνωρίζαμε τους κανόνες που χρησιμοποιήθηκαν, θα μπορούσαμε να χρησιμοποιήσουμε το ML για να τους προσδιορίσουμε.

- Θα μπορούσαμε όμως να πάμε ένα βήμα παραπέρα και να χρησιμοποιήσουμε το ML για να βελτιώσουμε τους κανόνες αποδοχής ή απόρριψης δανείων.

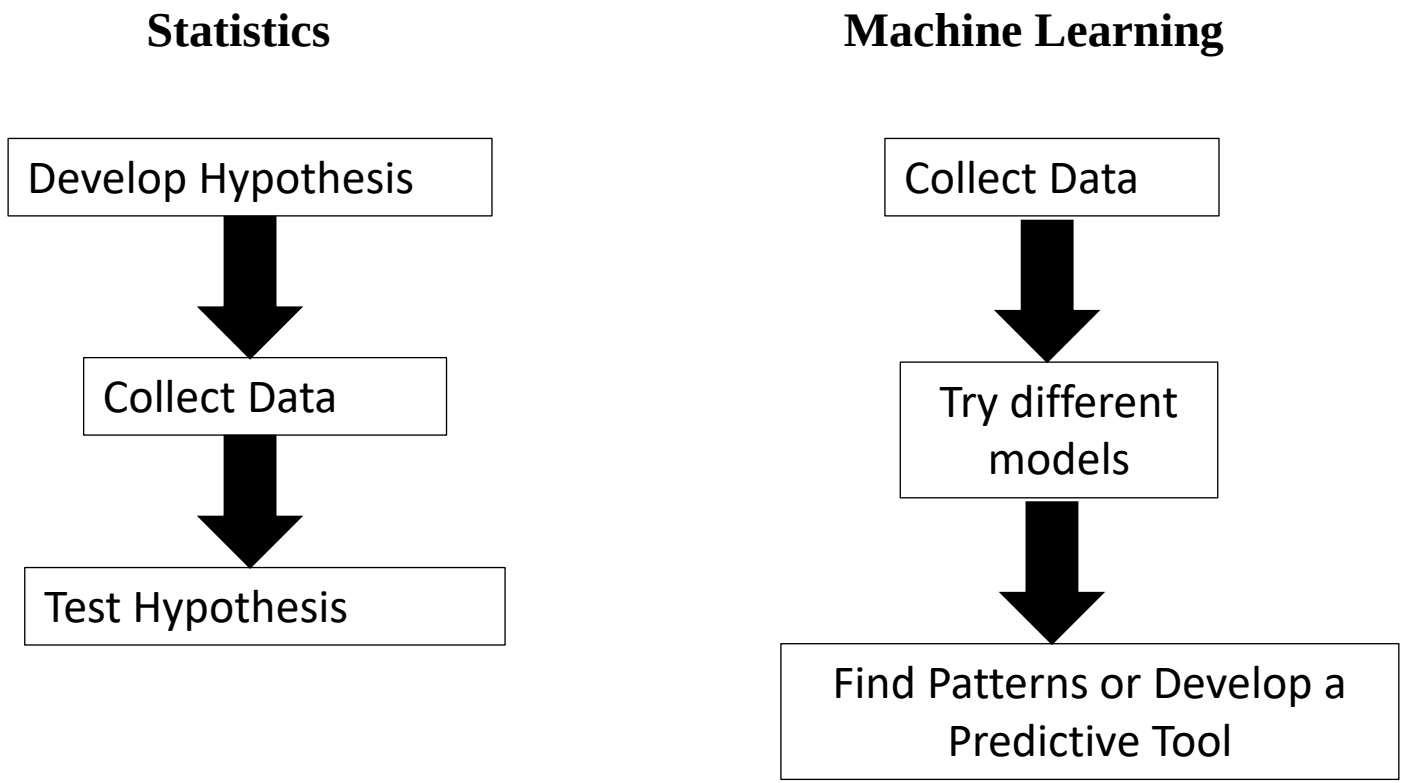
Traditional statistics

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

Ο νέος κόσμος των statistics

- Τεράστια σύνολα δεδομένων.
- Φανταστικές βελτιώσεις στις ταχύτητες επεξεργασίας υπολογιστή και στο κόστος αποθήκευσης δεδομένων.
- Τα εργαλεία μηχανικής μάθησης είναι πλέον εφικτά.
- Μπορεί τώρα να αναπτυχθούν μη γραμμικά μοντέλα πρόβλεψης, να βρεθούν μοτίβα σε δεδομένα με τρόπους που δεν ήταν δυνατό πριν και να αναπτυχθούν στρατηγικές λήψης αποφάσεων σε πολλά στάδια.
- Νέα ορολογία: features, labels, activation functions, target, bias, supervised/unsupervised learning……

Traditional Statistics vs Machine Learning



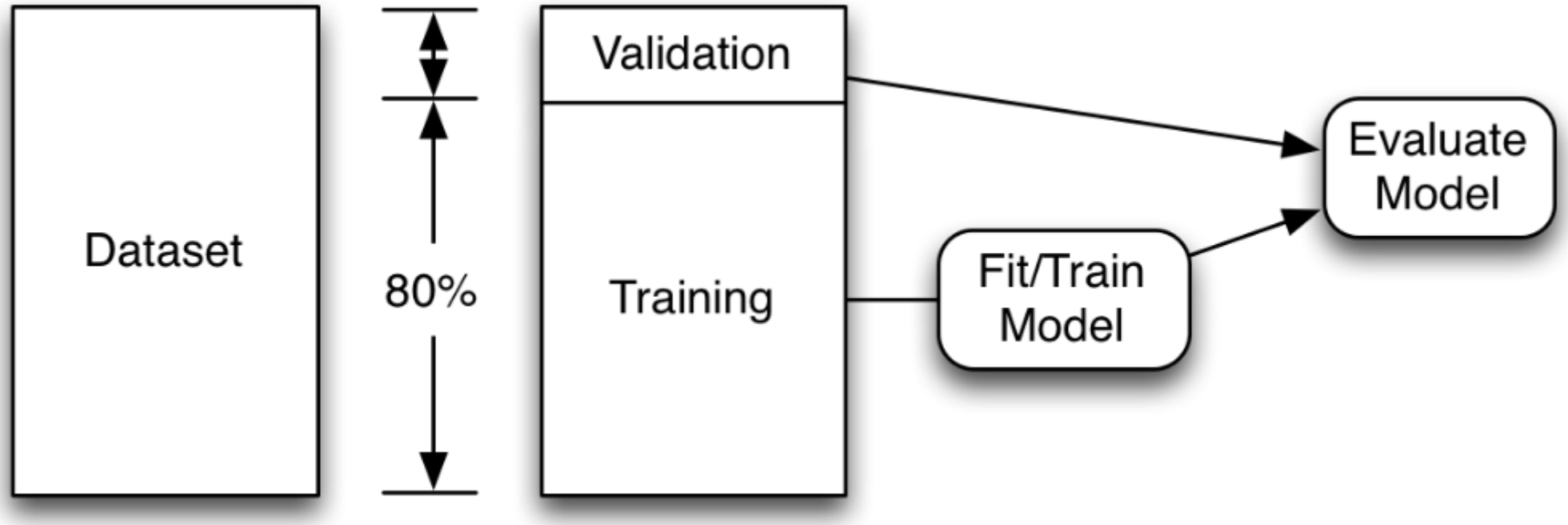
Τύποι Μηχανικής Μάθησης

- **Unsupervised learning** (find patterns) - Εκμάθηση χωρίς επίβλεψη (εύρεση προτύπων).
- **Supervised learning** (predict numerical value or classification)- Εποπτευόμενη μάθηση (πρόβλεψη αριθμητικής τιμής ή ταξινόμηση).
- **Semi-supervised learning (**only part of data has values for, or classification of, target) - Ημι-εποπτευόμενη μάθηση (μόνο μέρος των δεδομένων έχει τιμές για, ή ταξινόμηση του, στόχου).
- **Reinforcement learning** (multi-stage decision making) - Ενισχυτική μάθηση (η λήψη αποφάσεων σε πολλά στάδια).

ML Good Practice (1)

- Χωρίστε τα δεδομένα σε τρία sets
  - Training set
  - Validation set
  - Test set
- Αναπτύξτε διαφορετικά μοντέλα χρησιμοποιώντας το training set και εξετάστε πόσο καλά γενικεύονται σε νέα δεδομένα χρησιμοποιώντας το validation set.
- Rule of thumb: αυξήστε την πολυπλοκότητα του μοντέλου έως ότου το μοντέλο δεν γενικεύεται πλέον καλά στο validation set.
- Το test set χρησιμοποιείται για να παρέχει μια τελική ένδειξη εκτός δείγματος (out-of-sample) για το πόσο καλά λειτουργεί το επιλεγμένο μοντέλο.

ML Good Practice (2)



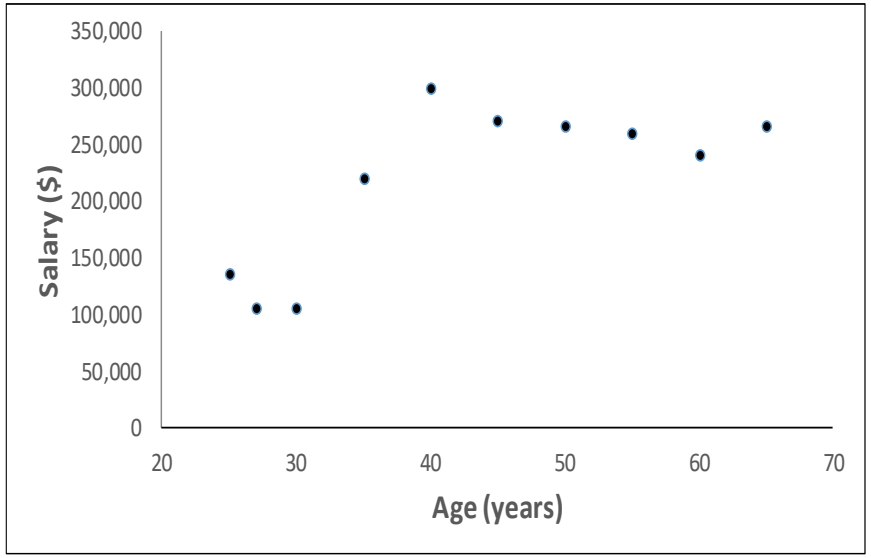
Παράδειγμα

Μισθός σε συνάρτηση με την ηλικία για ένα συγκεκριμένο επάγγελμα σε μια συγκεκριμένη περιοχή.
Αρχείο *Salary vs. Age Example.xlsx*

Training set

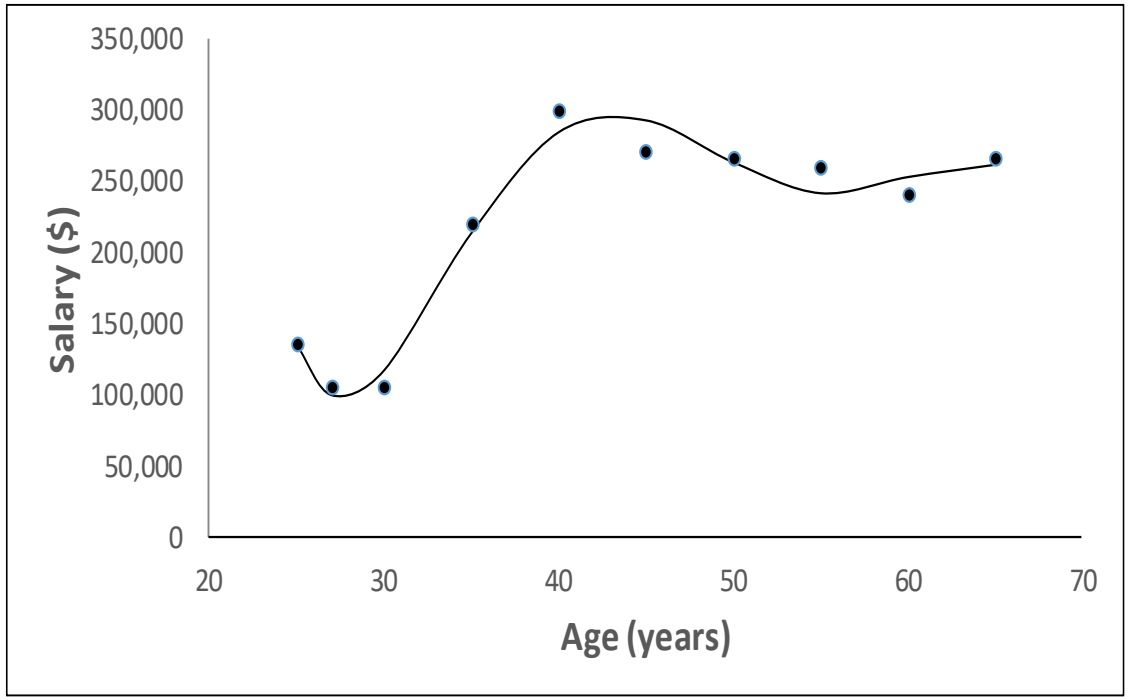
Age (years)	Salary ($)
25	135,000
55	260,000
27	105,000
35	220,000
60	240,000
65	265,000
45	270,000
40	300,000
50	265,000
30	105,000

Scatter plot για το training set



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

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

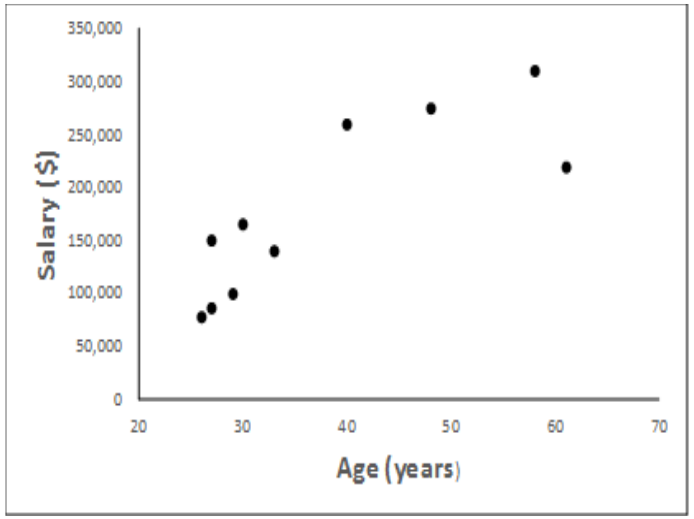


- Το root mean squared error (rmse) (η τυπική απόκλιση της διαφοράς μεταξύ της εκτίμησης/πρόβλεψης του μοντέλου και της πραγματικής τιμής) για το training dataset είναι $12,902.

Ένα 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

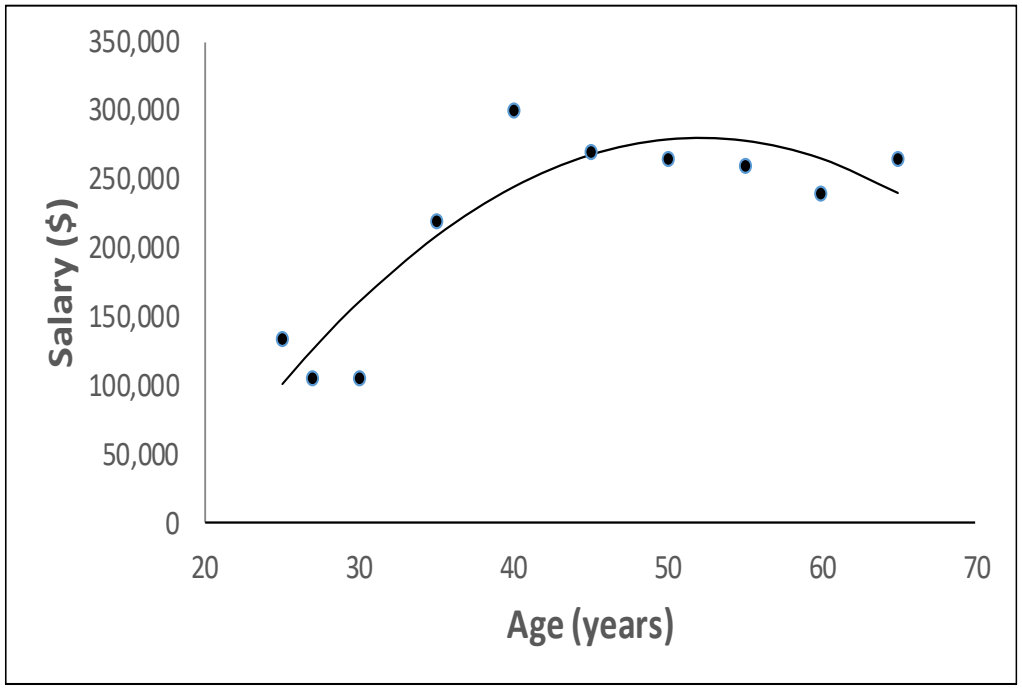
Scatter Plot για το Validation Set



- Το root mean squared error (rmse) για το training dataset είναι $12,902.
- Το rmse για το validation dataset is $38,794, πολύ υψηλότερο από το rmse για το training dataset ($12,902).
- Συμπεραίνουμε ότι το μοντέλο overfits τα δεδομένα, δε γενικεύεται καλά στα νέα δεδομένα.
- Το πολυωνυμικό μοντέλο πέμπτης τάξης δεν γενικεύεται καλά.

Quadratic Model

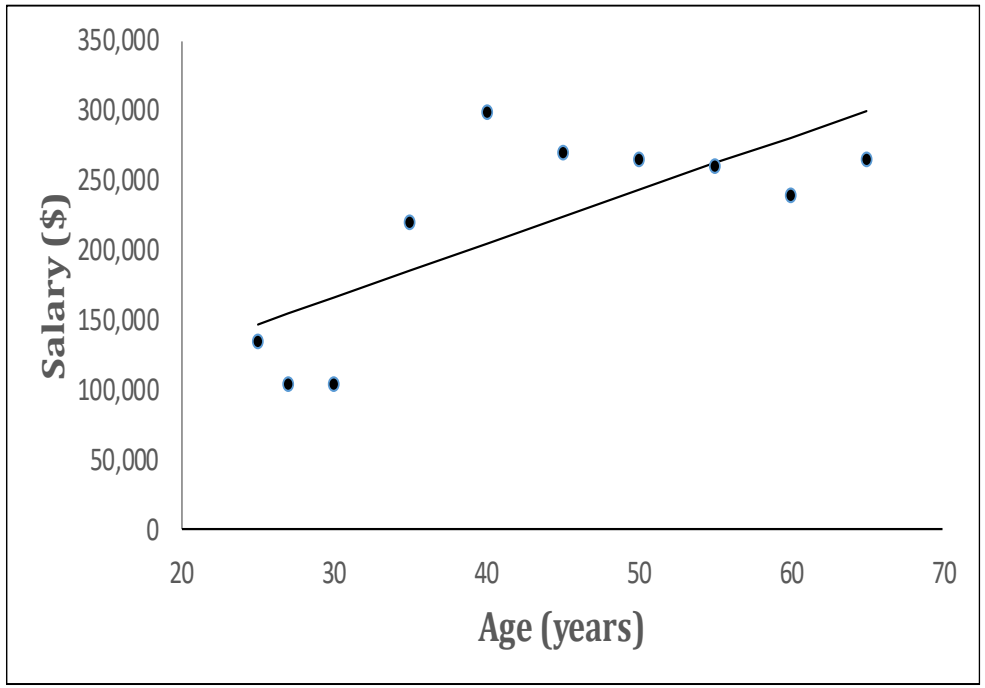
$$Y = a + b_1X + b_2X^2$$



- Το rmse για το training dataset είναι $32,932
- Το rmse για το validation dataset είναι $33,554

Linear Model

$$Y = a + b_1X$$



- Το rmse για το training dataset είναι $49,731
- Το rmse για το validation dataset είναι $49,990

Σύνοψη Αποτελεσμάτων: RMSE

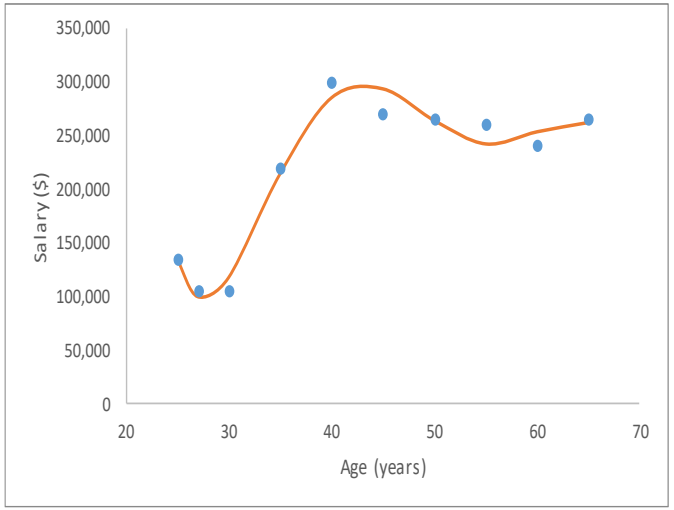
	Polynomial of degree 5	Quadratic model	Linear model
Training set	12, 902	32,932	49,731
Validation set	38,794	33,554	49,990

- Το γραμμικό μοντέλο (linear model) *underfits*: το rmse του training set είναι όσο μεγάλο είναι και το rmse του validation set.

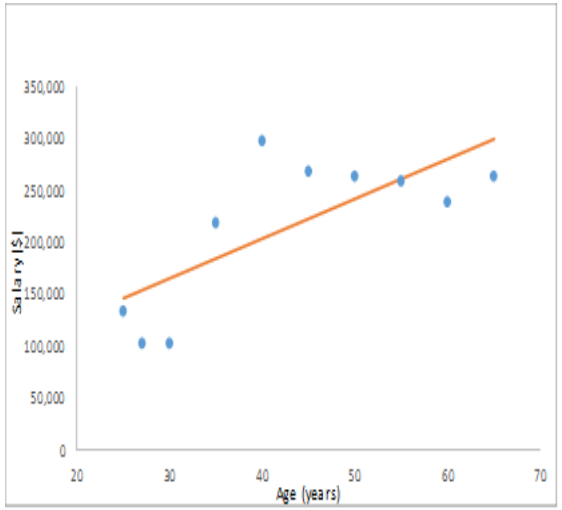
- Το πολυώνυμο 5ης τάξης (5th degree polynomial) *overfits*: το rmse του training set είναι πολύ μικρό ενώ, το rmse του validation set είναι πολύ μεγαλύτερο.

- Το quadratic μοντέλο είναι το πιο ακριβές, έχει το μικρότερο rmse στο rmse από τα υπόλοιπα μοντέλα και δεν έχει μεγάλη διαφορά από το rmse του training set. Αλλά πόσο ακριβές είναι το quadratic μοντέλο; Η απάντηση στη διαφάνεια 18.

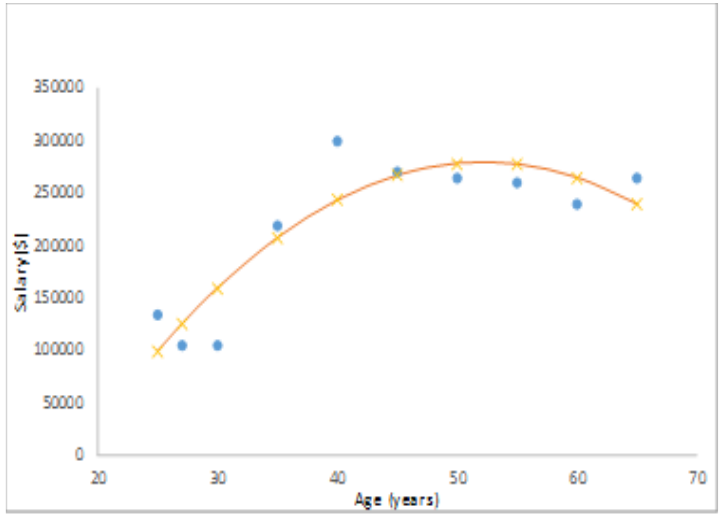
Overfitting/Underfitting



Overfitting



Underfitting



Best model?

Πόσο ακριβές είναι το quadratic μοντέλο;

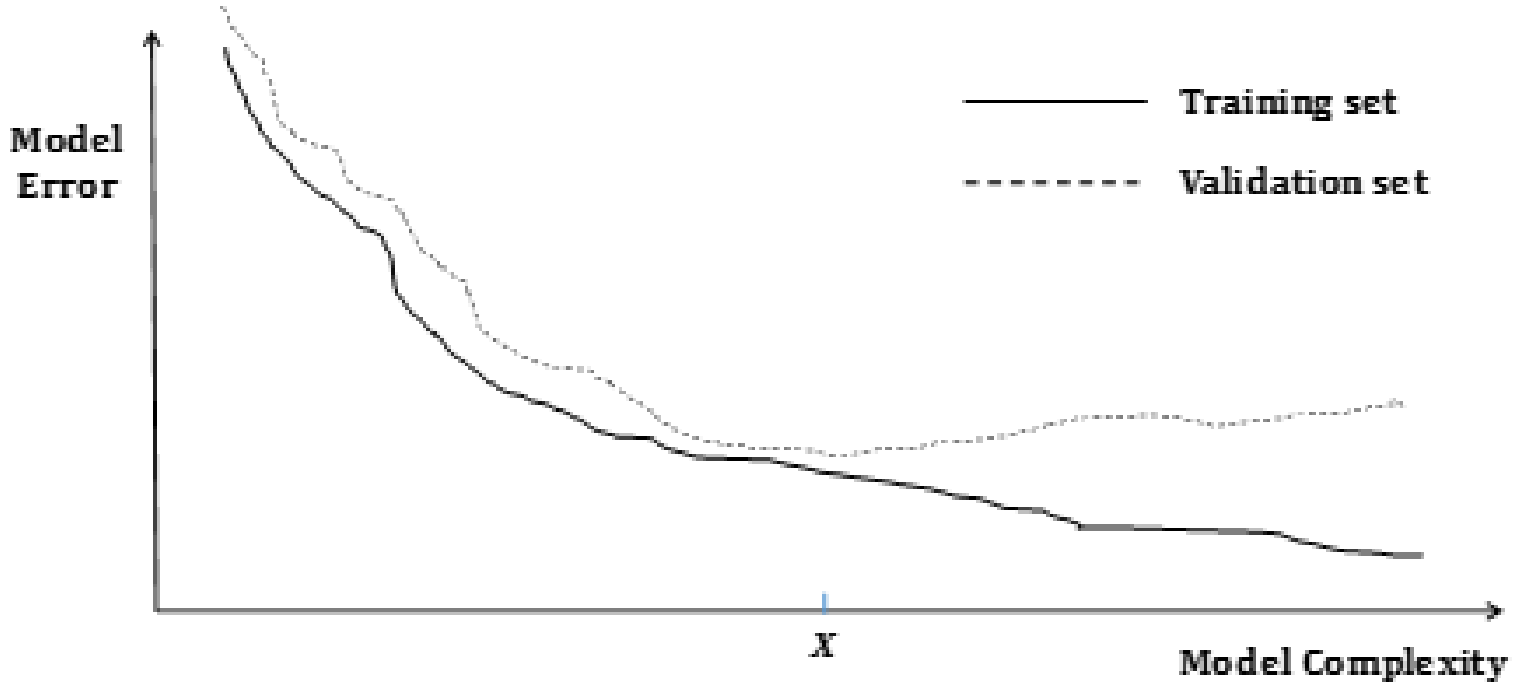
Πρέπει να υπολογίσουμε το rmse του test dataset.

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

Το SD of error (rmse) είναι $34,273

Rule of thumb: Η πολυπλοκότητα του μοντέλου πρέπει να αυξάνεται έως ότου το μοντέλο δεν γενικεύεται πλέον καλά στο out-of-sample set.

Typical Pattern of Errors for Training Set and Validation Set



Bias-variance trade-off

- Το bias αναφέρεται στο σφάλμα (error) που προκαλείται από underfitting.
- Το variance αναφέρεται στο σφάλμα (error) που προκαλείται από overfitting (υπάρχει random noise στο training set).

Data Cleaning

- Αντιμετώπιση inconsistent recording
- Αφαίρεση ανεπιθύμητων παρατηρήσεων
- Αφαίρεση διπλότυπων (duplicates)
- Διερεύνηση outliers
- Αντιμετώπιση στοιχείων που λείπουν (missing items)