Πρόβλεψη δημοτικότητας ειδήσεων στα κοινωνικά δίκτυά μέσω μοντέλων μηχανικής μάθησης

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Μανδηλαρά Ιωάννα

Επιβλέπων Καθηγητής: Κουσουρής Κωνσταντίνος

Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος του 2019

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών

Πρόβλεψη δημοτικότητας ειδήσεων στα κοινωνικά δίκτυά μέσω μοντέλων μηχανικής μάθησης

Διπλωματική Εργασία

της

Μανδηλαρά Ιωάννας

Επιβλέπων: Κουσουρής Κωνσταντίνος
Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Σεπτεμβρίου 2019.

(*Υπογραφή)*	(*Υπογραφή)*	(*Υπογραφή)*
……	……	……
Κουσουρής Κωνσταντίνος	Τζαμαριουδάκη Αικατερίνη	Τσιπολίτης Γεώργιος
Επίκουρος Καθηγητής	Ερευνήτρια Α	Καθηγητής ΣΕΜΦΕ
Ε.Μ.Π	ΕΚΕΦΕ Δημόκριτος	Ε.Μ.Π.

Αθήνα, Σεπτέμβριος του 2019

(*Υπογραφή)*

……

ΜΑΝΔΗΛΑΡΑ ΙΩΑΝΝΑ

Διπλωματούχος Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Τομέας Μαθηματικών

“Predicting the future isn’t magic, it’s artificial intelligence.”

~Dave Waters

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως στόχο την ανάδειξη της Μηχανικής Μάθησης, καθώς πρόκειται για έναν κλάδο που αναπτύσσεται διαρκώς και αποτελεί ένα νέο πεδίο ενδιαφέροντος σε διάφορους τομείς της επιστήμης. Η αφθονία περιεχομένου που δίνουν τα τελευταία χρόνια τα κοινωνικά δίκτυα μπορεί να χρησιμοποιηθεί κατάλληλα στον τομέα της Μηχανικής Μάθησης.

Στην συγκεκριμένη πτυχιακή εργασία παρουσιάζονται μοντέλα παλινδρόμησης με την βοήθεια των αλγορίθμων μηχανικής μάθησης. Κατασκευάζονται γραμμικά και μη γραμμικά μοντέλα με μεθόδους ενδυνάμωσης με την βοήθεια δέντρων απόφασης και νευρωνικά δίκτυα.

Το σύνολο δεδομένων προς επεξεργασία αποτελείται από ειδήσεις για 4 διαφορετικά θέματα και την δημοτικότητα τους στα κοινωνικά δίκτυα Facebook, LinkedIn και Google+. Για την επεξεργασία των δεδομένων χρησιμοποιήθηκε η γλώσσα προγραμματισμού ‘Python’ και οι αντίστοιχες βιβλιοθήκες της για την κατασκευή κάθε μοντέλου.

Πραγματοποιείται επεξεργασία στο σύνολο δεδομένων πριν την δημιουργία μοντέλων μηχανικής μάθησης ώστε να είναι κατάλληλα για χρήση. Μελετώντας τους αλγορίθμους μηχανικής μάθησης, αναλύεται η διαδικασία επιλογής των παραμέτρων κάθε μοντέλου με σκοπό την δημιουργία ενός αποδοτικού μοντέλου. Ακολουθεί η σύγκριση των μοντέλων μηχανικής μάθησης ώστε εξαχθούν τα ανάλογα συμπεράσματα. Βασιζόμενοι στα συμπεράσματα που προκύπτουν γίνεται έλεγχος γενίκευσης του μοντέλου σε ειδήσεις με διαφορετικό θέμα.

Λέξεις κλειδιά

Μηχανική Μάθηση, Νευρωνικά Δίκτυα, Μέθοδοι ενδυνάμωσης, Κοινωνικά Δίκτυα, Παλινδρόμηση , Δέντρα απόφασης, Ενδυναμωμένα δέντρα απόφασης

ABSTRACT

This diploma thesis aims to promote machine learning, as it is a field that is constantly developing and is a new field of interest in various fields of science. The abundance of content of social networks can be used appropriately in the field of machine learning.

In this thesis, regression models are presented with the help of machine learning algorithms. Linear and nonlinear models are created with boosting methods, where as weak learners used decision trees and neural networks.

The dataset to be processed consists of news about 4 different topics and their feedback on social platforms Facebook, LinkedIn and Google +. The programming language 'Python' and its respective libraries for the construction of each model were used to process the data.

The dataset is processed before the creation of machine learning models to be suitable for use. Studying the algorithms of machine learning, the process of selecting the parameters of each model is analyzed in order to create an efficient model. Following is a comparison of the machine learning models to draw the appropriate conclusions. Based on the resulting conclusions, the model's generalization is checked in news with a different topic.

Keywords

Machine Learning, Neural Networks, Boosting methods, Social Media, Regression, Decision Trees, Boosted Decision Trees

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω μια σειρά από ανθρώπους που με βοήθησαν για να υλοποιηθεί η συγκεκριμένη πτυχιακή εργασία.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Κουσουρή για την ανάθεση της συγκεκριμένης πτυχιακής εργασίας σε εμένα, καθώς επίσης και για την βοήθεια του για να μάθω περισσότερα πράγματα στο συγκεκριμένο αντικείμενο και να εξελίξω τις γνώσεις μου.

Επίσης, θα ήθελα να ευχαριστήσω όλους τους υπόλοιπους καθηγητές του τμήματος Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών για τις γνώσεις που αποκόμισα πάνω στην επιστήμη των Εφαρμοσμένων Μαθηματικών.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου που με στηρίζει τόσα χρόνια στην ακαδημαϊκή μου πορεία και μου έδωσε την δυνατότητα, το ήθος και τα εφόδια να φτάσω ως εδώ και τους φίλους μου που με στηρίζουν κάθε μέρα και είναι πάντα δίπλα μου.

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή

Αντικείμενο πτυχιακής εργασίας

Η επίλυση υπολογιστικών προβλημάτων μέσω αλγοριθμικών διαδικασιών αποτελεί βασικό κομμάτι της έρευνας, των επιστημών αλλά και της καθημερινότητας. Πιο συγκεκριμένα, οι τεχνικές μηχανικής μάθησης σε συνδυασμό με την στατιστική έχουν ως στόχο την εξαγωγή συμπερασμάτων από την επεξεργασία δεδομένων. Η χρήση των τεχνικών μηχανικής μάθησης στοχεύει στην αποτελεσματική ανάλυση μεγάλου όγκου δεδομένων και στην δημιουργία μη γραμμικών μοντέλων παλινδρόμησης.

Στις μέρες μας, η αφθονία περιεχομένου που προσφέρουν οι πλατφόρμες των κοινωνικών δικτύων έχει επιφέρει άνοδο σε τομείς της έρευνας, όπως εξόρυξη δεδομένων, μηχανική μάθηση κλπ. Για αυτόν τον λόγο, η παρούσα πτυχιακή εργασία έχει ως τελικό στόχο την ανάλυση δεδομένων από τα κοινωνικά δίκτυα Facebook, Google+ και LinkedIn με απώτερο σκοπό αφενός την πρόβλεψη της δημοτικότητας ειδήσεων στις συγκεκριμένες πλατφόρμες και αφετέρου την σύγκριση των μοντέλων πρόβλεψης που χρησιμοποιούνται για την εξαγωγή προβλέψεων.

Η επεξεργασία των δεδομένων πραγματοποιήθηκε με την βοήθεια της γλώσσας '*Python'*. Πιο συγκεκριμένα, χρησιμοποιήθηκαν αρκετές βιβλιοθήκες της συγκεκριμένης γλώσσας για την υλοποίηση των αλγορίθμων.

Δομή Πτυχιακής Εργασίας

Στο παρόν σημείο, κρίνεται σκόπιμο να δοθεί μια σύντομη περιγραφή του κάθε κεφαλαίου που περιέχεται στην εργασία. Το πρώτο κεφάλαιο αφορά την περιγραφή του προβλήματος, το σύνολο δεδομένων και την πρώτη επεξεργασία του με στόχο την εξαγωγή συμπερασμάτων μεταξύ των μεταβλητών. Στην συνέχεια, ακολουθεί το δεύτερο κεφάλαιο, όπου περιγράφεται η έννοια της μηχανικής μάθησης και ορισμένες σημαντικές πληροφορίες σχετικά με την θεωρία που θα χρησιμοποιηθούν στα επόμενα κεφάλαια. Το τρίτο κεφάλαιο αποτελεί το θεωρητικό μέρος της συγκεκριμένης πτυχιακής εργασίας και περιγράφονται οι αλγόριθμοι που θα χρησιμοποιηθούν και κάποια σημαντικά μέτρα απόδοσης ενός μοντέλου. Τέλος, στο τελευταίο κεφάλαιο, παρουσιάζονται τα αποτελέσματα, η υλοποίηση των αλγορίθμων και τα συμπεράσματα που προκύπτουν ύστερα από την επεξεργασία του συνόλου δεδομένων.

Κεφάλαιο 1 : Περιγραφή του προβλήματος

Το συγκεκριμένο πρόβλημα που θα μελετηθεί αφορά την πρόβλεψη της δημοτικότητας ενός άρθρου στο Facebook, σχετικά με την οικονομία. Στην συνέχεια, μελετάται η γενίκευση του συγκεκριμένου προβλήματος σε άρθρα με διαφορετικό θέμα. Το σύνολο δεδομένων προς επεξεργασία είναι από το UCI Machine Learning Repository[1].

1.1 Περιγραφή συνόλου δεδομένων

Το σύνολο δεδομένων προς επεξεργασία παρέχει πληροφορία σχετικά με άρθρα από ιστοσελίδες συγκέντρωσης δεδομένων όπως Google News και Yahoo! News και το ανάλογη ανατροφοδότηση (‘feedback’) τους στις πλατφόρμες του Facebook, Google+ και LinkedIn. Τα δεδομένα που συλλέχθηκαν αφορούν μια περίοδο 8 μηνών, μεταξύ του Νοεμβρίου του 2015 και του Ιουλίου του 2016 και χωρίζονται σε 4 θέματα : Οικονομία, Microsoft, Obama και Παλαιστίνη ,δηλαδή αφορούν ένα τομέα, μια εταιρεία, ένα πρόσωπο και μια χώρα. Πιο συγκεκριμένα, το συγκεκριμένο σύνολο δεδομένων παρέχει γνώση για ειδήσεις σχετικά με καθημερινά σημαντικά θέματα. Είναι χωρισμένα σε δύο υποσύνολα δεδομένων, τα οποία είναι:

- *1ο αρχείο δεδομένων* : Το 1ο αρχείο δεδομένων αποτελείται από 11 μεταβλητές, οι οποίες δίνουν πληροφορίες σχετικά με τα άρθρα. Οι μεταβλητές αυτές είναι:

IDLink	: Μοναδική ταυτότητα συνδέσμου του άρθρου
Title	: Τίτλος του άρθρου σύμφωνα με την πηγή κοινωνικού δικτύου
Headline	: Επικεφαλίδα του άρθρου σύμφωνα με την πηγή κοινωνικού δικτύου
Source	: Πηγή σύμφωνα με την οποία δημοσιεύτηκε το άρθρο
Topic	: Θέμα του άρθρου
Publish Date	: Ημερομηνία και ώρα δημοσίευσης της είδησης
Sentiment Title	: Βαθμολογία της ανάλυσης συναισθήματος του τίτλου της είδησης
Sentiment Headline	: Βαθμολογία της ανάλυσης συναισθήματος του ‘headline’ της είδησης
Facebook	: Δημοτικότητα του άρθρου στο Facebook μετά από 2 ημέρες
Google+	: Δημοτικότητα του άρθρου στο Google+ μετά από 2 ημέρες
LinkedIn	: Δημοτικότητα του άρθρου στο LinkedIn μετά από 2 ημέρες

- *2ο αρχείο δεδομένων* : Το 2ο σύνολο δεδομένων είναι χωρισμένο σε 12 αρχεία με 145 μεταβλητές το καθένα, δηλαδή όλους του συνδυασμούς των 4 θεμάτων και του

[1] https://archive.ics.uci.edu/ml/datasets/News+Popularity+in+Multiple+Social+Media+Platforms

'feedback' των άρθρων στις πλατφόρμες του Facebook, Google+ και LinkedIn. Οι 145 μεταβλητές παρέχουν πληροφορία σχετικά με την εξέλιξη της δημοτικότητας κάθε είδησης σε μια περίοδο 2 ημερών. Πιο συγκεκριμένα, κάθε περίπτωση περιγράφεται από 145 μεταβλητές : το μοναδικό IDLink της είδησης και οι 144 μετρήσεις της δημοτικότητας του άρθρου σε διαστήματα 20 λεπτών, TS_i για i = 1,…, 144.

1.1.1 Περιγραφή μεταβλητών TS και Sentiment Score

Η δημοτικότητα κάθε άρθρου αντιπροσωπεύει ένα διαφορετικό είδος πληροφορίας, ανάλογα με την πλατφόρμα.

- Η δημοτικότητα ενός άρθρου στο Facebook και στο LinkedIn δείχνει τον αριθμό των δημοσιεύσεων που απέκτησε, σύμφωνα με το μοναδικό URL το οποίο χρησιμοποιείται σαν μέτρο δημοτικότητας.
- Η δημοτικότητα ενός άρθρου στο Google+ δείχνει τον αριθμό των φορών που ένας χρήστης δήλωσε ότι του αρέσει. Ο λόγος που δεν δηλώνει τον αριθμό των δημοσιεύσεων είναι λόγω τεχνικών περιορισμών από το Google+.

Σε ορισμένες περιπτώσεις, η μέτρηση της δημοτικότητας μιας είδησης σε ένα συγκεκριμένο χρονικό διάστημα δεν ήταν εφικτή. Αυτές οι περιπτώσεις έχουν τιμή στην δημοτικότητα του άρθρου -1. Μια τέτοια περίπτωση μπορεί να συμβεί όταν μια είδηση προτείνεται στα κοινωνικά δίκτυα ύστερα από 2 ημέρες της δημοσίευσης της στο Google News και Yahoo! News.

Οι μεταβλητές Sentiment Title και Sentiment Headline δείχνουν την βαθμολογία ανάλυσης συναισθήματος που έχει ο τίτλος και η επικεφαλίδα της είδησης. Η ανάλυση συναισθήματος είναι ένα πεδίο της εξόρυξης κειμένων ('text mining'), το οποίο αναλύει τις γνώμες και τα συναισθήματα των ανθρώπων ως προς προϊόντα, υπηρεσίες , θέματα κλπ. Το διάστημα τιμών των συγκεκριμένων μεταβλητών είναι το [-1,1]. Θετικές λέξεις ή προτάσεις έχουν θετική βαθμολογία ανάλυσης συναισθήματος , αρνητικές λέξεις έχουν αρνητική βαθμολογία ενώ ουδέτερες έχουν μηδενική βαθμολογία. (1)

1.1.2 Προ επεξεργασία συνόλου δεδομένων

Πριν την ανάλυση του προβλήματος στο συγκεκριμένο σύνολο δεδομένων, ακολουθήθηκε μια επεξεργασία των δεδομένων έτσι ώστε να είναι στην σωστή μορφή για την μελέτη του προβλήματος. Αρχικά, τα 12 αρχεία που περιείχαν την εξέλιξη της δημοτικότητας των άρθρων δεν περιείχαν τον ίδιο αριθμό ειδήσεων για κάθε τομέα με το τελικό αρχείο που αποτελείται από τις 11 μεταβλητές. Για παράδειγμα, για τα άρθρα με θέμα 'Microsoft':

Αριθμός άρθρων στο Facebook:	18531
Αριθμός άρθρων στο Google+:	20702
Αριθμός άρθρων στο LinkedIn:	20702
Αριθμός άρθρων στο συνολικό αρχείο:	21858

Οπότε έπρεπε να βρεθούν τα κοινά IDLink μεταξύ των 4 αρχείων, τα οποία είναι για κάθε τομέα:

Θέμα άρθρου	**Αριθμός κοινών άρθρων**
Microsoft	17957
Ομπάμα	26936
Οικονομία	28842
Παλαιστίνη	7678

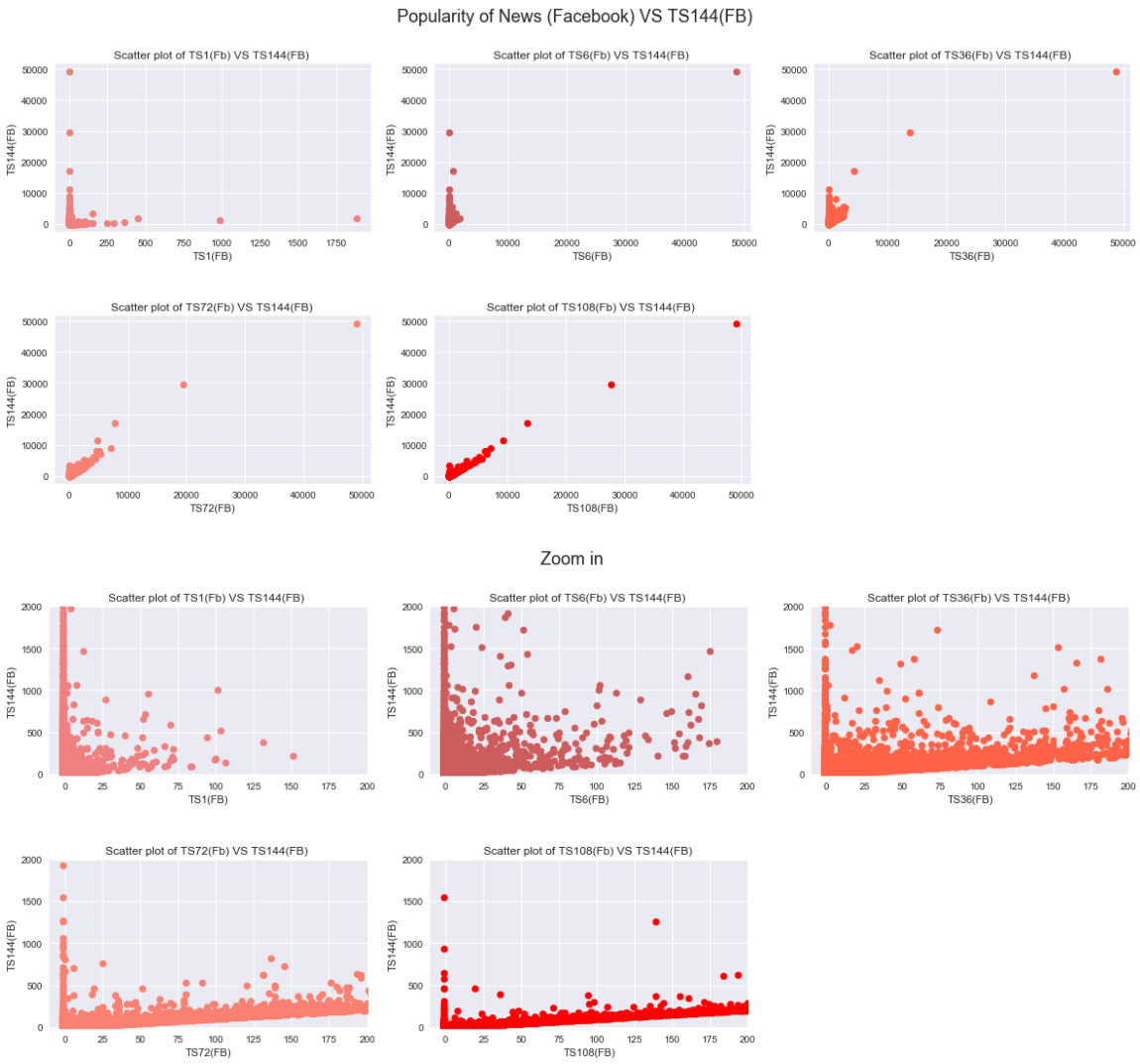
Παρατηρείται ότι τα άρθρα με θέμα την οικονομία έχουν το μεγαλύτερο μέγεθος και για αυτό η μελέτη θα γίνει πάνω σε αυτά και στην συνέχεια το μοντέλο που θα προκύψει θα εφαρμοσθεί στα σύνολα δεδομένων των υπόλοιπων τομέων.

Στόχος του προβλήματος είναι να γίνει η πρόβλεψη της δημοτικότητας των άρθρων της οικονομίας στην πλατφόρμα του Facebook μετά από 2 ημέρες μέσω ενός μοντέλου με τις παρακάτω μεταβλητές:

- ***$TS_1(FB)$, $TS_6(FB)$, $TS_{36}(FB)$, $TS_{72}(FB)$, $TS_{108}(FB)$*** : Δημοτικότητα των άρθρων στο Facebook τα πρώτα 20', τις πρώτες 2 ώρες, τις πρώτες 12 ώρες, την 1η ημέρα και τις πρώτες 36 ώρες αντίστοιχα.
- ***$TS_1(G+)$, $TS_6(G+)$, $TS_{36}(G+)$, $TS_{72}(G+)$, $TS_{108}(G+)$, $TS_{144}(G+)$*** : Δημοτικότητα των άρθρων στο Google+ τα πρώτα 20', τις πρώτες 2 ώρες, τις πρώτες 12 ώρες, την 1η ημέρα και τις πρώτες 36 ώρες αντίστοιχα.
- ***$TS_1(LD)$, $TS_6(LD)$, $TS_{36}(LD)$, $TS_{72}(LD)$, $TS_{108}(LD)$, $TS_{144}(LD)$*** : Δημοτικότητα των άρθρων στο LinkedIn τα πρώτα 20', τις πρώτες 2 ώρες, τις πρώτες 12 ώρες, την 1η ημέρα και τις πρώτες 36 ώρες αντίστοιχα.
- ***SentimentHeadline, SentimentTitle***
- ***Is Monday, Is Tuesday, Is Wednesday, Is Thursday, Is Saturday, Is Sunday*** : Δυαδικές μεταβλητές που λαμβάνουν την τιμή 1 αν το άρθρο δημοσιεύτηκε την συγκεκριμένη ημέρα αλλιώς 0.
- ***Is Morning, Is Afternoon, Is Evening, Is Night*** : Δυαδικές μεταβλητές που λαμβάνουν την τιμή 1 αν το άρθρο δημοσιεύτηκε το συγκεκριμένο μέρος της ημέρας, αλλιώς 0. Κάθε διάστημα της ημέρας αποτελείται από 6 ώρες.

1.1.3 Συσχετίσεις μεταβλητών με την μεταβλητή απόκρισης

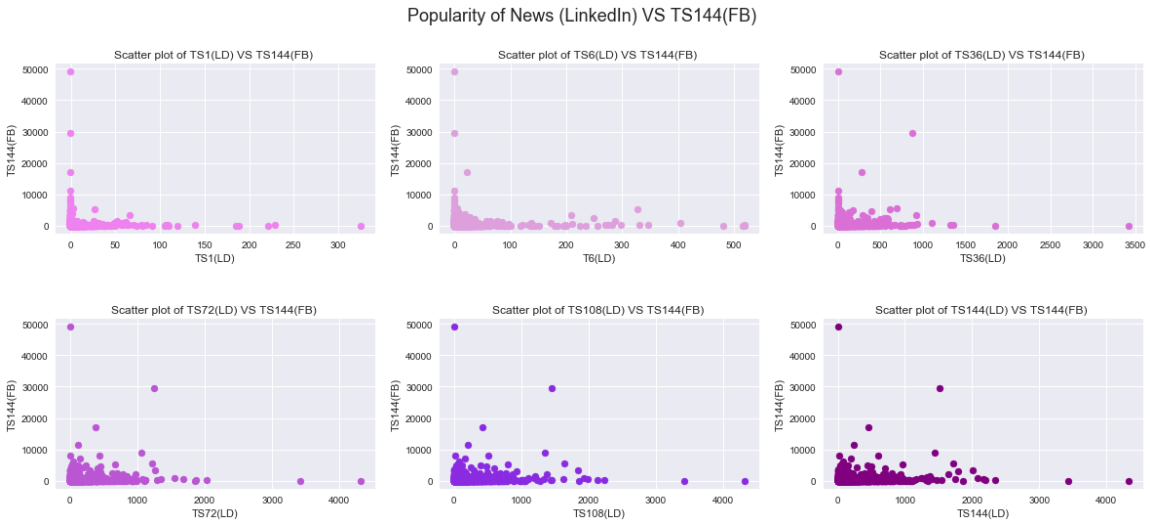
Παρακάτω παρουσιάζονται τα διαγράμματα κάθε μεταβλητής TS συναρτήσει της μεταβλητής απόκρισης, δηλαδή εκείνης που θα προβλεφθεί η τιμή, για τα δεδομένα της οικονομίας.



Διάγραμμα 1.1: Συσχετίσεις μεταξύ των TS(FB) συναρτήσει του TS144(FB)

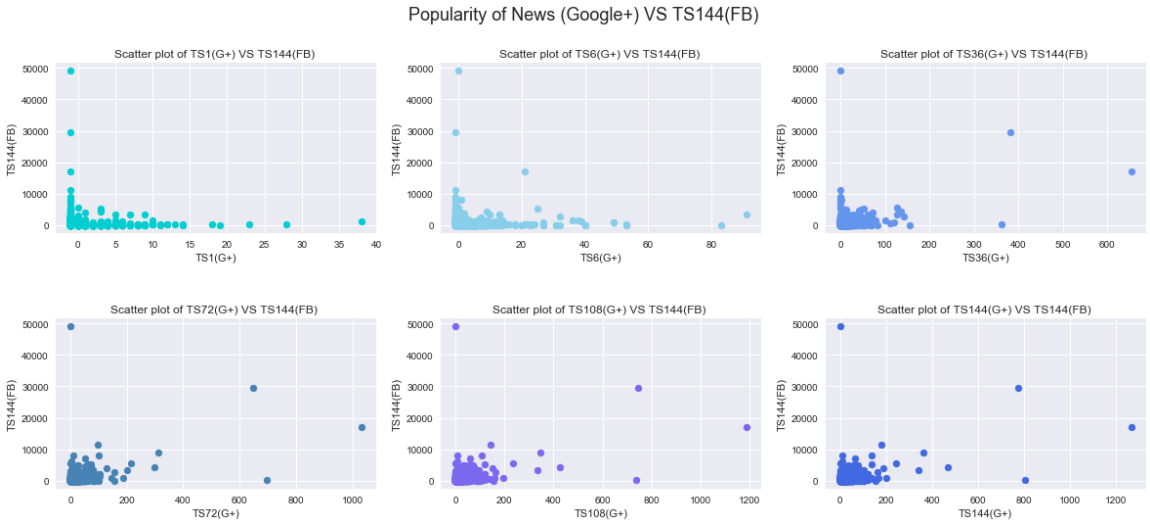
Παρατηρείται ότι με την πάροδο του χρόνου, η μεταβλητή απόκρισης έχει μια γραμμική εξάρτηση από τις μεταβλητές TS36(FB), TS72(FB) και TS108(FB), οι οποίες δείχνουν την δημοτικότητα των ειδήσεων στην πλατφόρμα του Facebook μετά από 12, 24 και 36 ώρες από την δημοσίευση του άρθρου αντίστοιχα. Επιπλέον, οι ειδήσεις με λιγότερες κοινοποιήσεις στο Facebook τα πρώτα 20 λεπτά και τις πρώτες 2 ώρες αποκτούν πολύ περισσότερες κοινοποιήσεις ύστερα από 2 ημέρες(TS144(FB).

Στην συνέχεια, παρουσιάζονται τα διαγράμματα συσχετισμού των μεταβλητών που αφορούν το LinkedIn και Google+.



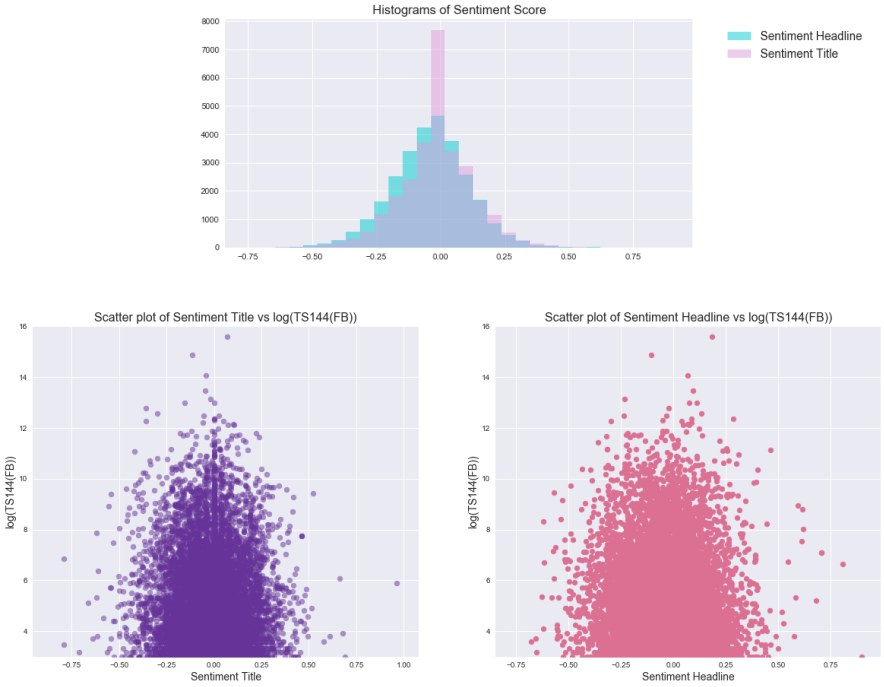
Διάγραμμα 1.2: Συσχετίσεις μεταξύ των TS(LinkedIn) συναρτήσει του TS144(FB)

Σύμφωνα με τα διαγράμματα 1.2, φαίνεται ότι οι ειδήσεις με λιγότερες κοινοποιήσεις στην πλατφόρμα του LinkedIn έχουν μεγαλύτερη δημοτικότητα στην πλατφόρμα του Facebook.



Διάγραμμα 1.3: Συσχετίσεις μεταξύ των TS(Google+) συναρτήσει του TS144(FB)

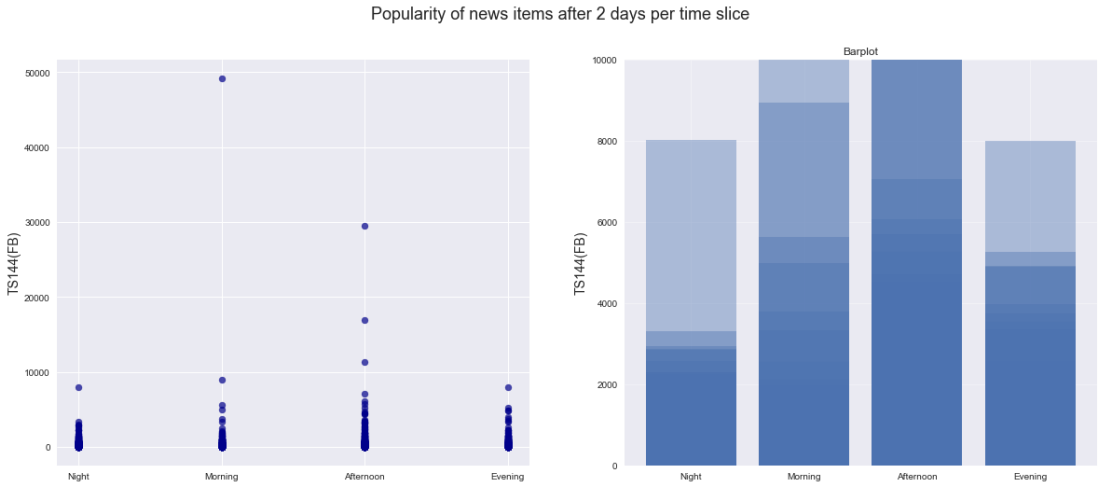
Παρόμοια συμπεράσματα προκύπτουν και για την δημοτικότητα των ειδήσεων στο Google+, δηλαδή ειδήσεις που δεν είναι τόσο διάσημες στο Google+ είναι πιο διάσημες στην πλατφόρμα του Facebook.

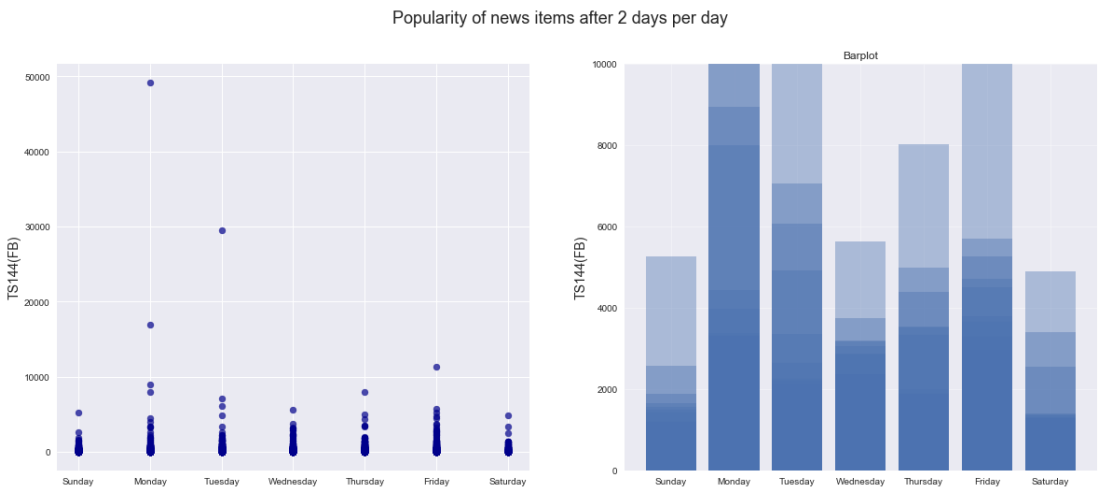


Διάγραμμα 1.4: Συσχετίσεις μεταξύ της βαθμολογίας ανάλυσης συναισθήματος συναρτήσει του TS144(FB)

Όπως φαίνεται παραπάνω, ειδήσεις με ουδέτερη βαθμολογία ανάλυσης συναισθήματος για τον τίτλο και την επικεφαλίδα της είδησης είναι πιο διάσημα στην πλατφόρμα του Facebook ύστερα από 2 ημέρες.

Η ανάλυση που έγινε για τις κατηγορικές μεταβλητές παρουσιάζεται παρακάτω:





Διάγραμμα 1.5: TS144(FB) συναρτήσει του διαστήματος της ημέρας & της ημέρας δημοσίευσης

Στα παραπάνω διαγράμματα, παρατηρείται ότι το μεγαλύτερο ποσοστό των ειδήσεων που έχουν δημοσιευθεί απόγευμα και Δευτέρα ή Παρασκευή έχει το μεγαλύτερο πλήθος κοινοποιήσεων στο Facebook ύστερα από τις 2 ημέρες της δημοσίευσης τους.

Συνοψίζοντας, κάθε μεταβλητή έχει ένα διαφορετικό τρόπο συσχέτισης με την μεταβλητή απόκρισης και για αυτό χρησιμοποιείται στο μοντέλο παλινδρόμησης που θα κατασκευαστεί.

1.1.4 Αφαίρεση τιμών της μεταβλητής απόκρισης

Το τελευταίο βήμα της προ επεξεργασίας των δεδομένων της οικονομίας αποτελεί η αφαίρεση εκείνων των δειγμάτων , όπου η μεταβλητή απόκρισης λαμβάνει μηδενικές τιμές ή -1. Ο λόγος που αφαιρέθηκαν οι αρνητικές τιμές, είναι διότι δεν δηλώνουν την δημοτικότητα του νέου[2] και αντίστοιχα η αφαίρεση των μηδενικών τιμών έγινε διότι χρειάστηκε στην συνέχεια, στον έλεγχο απόδοσης του μοντέλου. Η ίδια διαδικασία ακολουθήθηκε στα δεδομένα της Παλαιστίνης, Microsoft και Ομπάμα ώστε να γίνει ο έλεγχος γενίκευσης του μοντέλου που θα προκύψει από τα δεδομένα οικονομίας.

[2] Αναλυτικότερα στο 1.1.1

Κεφάλαιο 2 : Εισαγωγή στην Μηχανική Μάθηση

Σε διάφορους τομείς της επιστήμης χρειάζεται να επιλυθεί ένα πρόβλημα βασισμένο σε ένα μεγάλο σύνολο δεδομένων. Για την επίλυση ενός τέτοιου προβλήματος στον υπολογιστή χρειάζεται ένας αλγόριθμος. Ένα αλγόριθμος είναι μια ακολουθία οδηγιών που μετασχηματίζουν τις τιμές εισόδου σε τιμές εξόδου. Μέσω διάφορων αλγορίθμων στόχος είναι η κατασκευή ενός αποδοτικού μοντέλου, το οποίο μπορεί να ανιχνεύσει μοτίβα στα δεδομένα. Τέτοιου είδους μοτίβα μπορούν να βοηθήσουν έτσι ώστε να γίνουν προβλέψεις. Υποθέτοντας ότι το κοντινό μέλλον δεν είναι πολύ διαφορετικό από το παρελθόν όταν γίνεται η συλλογή δεδομένων, οι προβλέψεις που θα γίνουν μπορούν να θεωρηθούν σωστές. Αυτό αποτελεί την αρχή της μηχανικής μάθησης.

Η μηχανική μάθηση αποτελεί ένα πεδίο της επιστήμης που δίνει την δυνατότητα στους υπολογιστές να 'μάθουν' χωρίς να είναι αυστηρά προγραμματισμένοι. Βασίζεται σε αλγόριθμους που κατασκευάζουν μοντέλα που μπορούν να εκπαιδευθούν από σύνολα δεδομένων χωρίς όμως να βασίζονται σε προγραμματισμό βασισμένο σε κανόνες.

Ορίζεται ένα μοντέλο με κάποιες παραμέτρους και η εκμάθηση είναι εκτέλεση ενός αλγορίθμου, ο οποίος βελτιστοποιεί αυτές τις παραμέτρους βασιζόμενος στο σύνολο δεδομένων εκπαίδευσης, δηλαδή στην 'παλιότερη γνώση'. Η μηχανική μάθηση χρησιμοποιεί την θεωρία της στατιστικής για την δημιουργία αυτών των μοντέλων, γιατί ο κύριος στόχος της είναι να εξάγει συμπεράσματα από το σύνολο δεδομένων. (2)

2.1 Είδη μηχανικής μάθησης

Στην μηχανική μάθηση, οι αλγόριθμοι που χρησιμοποιούνται χωρίζονται σε κατηγορίες, ανάλογα με τον σκοπό τους. Παρακάτω θα εξηγηθούν τα δύο πιο δημοφιλή είδη μηχανικής μάθησης.

Επιτηρούμενη μάθηση

Η επιτηρούμενη μάθηση είναι η εκπαίδευση ενός μοντέλου, γνωρίζοντας τις μεταβλητές εισόδου και εξόδου. Πιο αναλυτικά, χρησιμοποιείται όταν διαθέτουμε μεταβλητές εισόδου x και μεταβλητές εξόδου y και γίνεται υλοποίηση ενός αλγορίθμου, ο οποίος 'μαθαίνει' μια συνάρτηση να αντιστοιχίζει τις μεταβλητές εισόδου σε μεταβλητές εξόδου , $y = f(x)$. Στόχος είναι η προσέγγιση των τιμών εξόδου μέσω της συνάρτησης f σε τέτοιο βαθμό ώστε όταν τροφοδοτηθεί το μοντέλο με νέες τιμές εισόδου οι τιμές εξόδου που θα προκύψουν για τις νέες τιμές εισόδου να έχουν όσο το δυνατόν μικρότερη απόκλιση από τις πραγματικές τιμές. Η εκπαίδευση του μοντέλου σταματάει όταν ο αλγόριθμος επιτυγχάνει ένα ικανοποιητικό επίπεδο απόδοσης του μοντέλου.

Τα προβλήματα επιτηρούμενης μάθησης μπορούν να χωριστούν σε 2 κατηγορίες προβλημάτων, οι οποίες είναι:

- *Ταξινόμηση (Classification)*

  Ταξινόμηση είναι μια διαδικασία κατηγοριοποίησης των δεδομένων στον επιθυμητό αριθμό κλάσεων, όπου για κάθε κλάση ορίζεται μια ετικέτα.

  Υπάρχουν 2 είδη ταξινόμησης, τα οποία είναι τα παρακάτω:

  1. Δυαδική ταξινόμηση: Ταξινόμηση σε 2 κλάσεις δεδομένων

  2. Ταξινόμηση πολλαπλών κλάσεων: Ταξινόμηση σε περισσότερες από 2 κλάσεις.

- *Παλινδρόμηση (Regression)*

  Παλινδρόμηση είναι μια διαδικασία αναγνώρισης προτύπων και πρόβλεψης μιας πραγματικής τιμής μιας μεταβλητής βασισμένης σε ένα μοντέλο με επεξηγηματικές μεταβλητές[3].

Οι πιο γνωστοί αλγόριθμου επιτηρούμενης μηχανικής μάθησης παρουσιάζονται παρακάτω:

- ✓ Πλησιέστερος γείτονας (Nearest Neighbor)
- ✓ Support Vector Machine
- ✓ Δέντρα απόφασης (Decision Tree)
- ✓ Τυχαίο δάσος δέντρων απόφασης (Random Forest)
- ✓ Λογιστική Παλινδρόμηση (Logistic Regression)
- ✓ Νευρωνικά Δίκτυα (Neural Networks)

Μη επιτηρούμενη μάθηση

Η μη επιτηρούμενη μάθηση είναι η εκπαίδευση ενός μοντέλου, γνωρίζοντας μόνο τις μεταβλητές εισόδου. Στόχος αυτής της μάθησης είναι η μοντελοποίηση της δομής των δεδομένων ή της κατανομής τους ώστε να ‘μάθει’ περισσότερα από αυτά. Ονομάζεται μη επιτηρούμενη, διότι σε σχέση με την επιτηρούμενη, δεν υπάρχουν κλάσεις δεδομένων για να κατατάξει τα δεδομένα ή σωστές τιμές για να κάνει προβλέψεις.

[3] Θα αναλυθεί στην συνέχεια στο κεφάλαιο 3

Μια σημαντική κατηγορία της μη επιτηρούμενης μηχανικής μάθησης είναι η παρακάτω:

- *Ομαδοποίηση (Clustering)*

  Ομαδοποίηση είναι μια διαδικασία διαχωρισμού των δεδομένων σε ομάδες με βάση τις ομοιότητες τους. Αποτελεί μία από τις κατηγορίες της μη επιτηρούμενης μάθησης, διότι δεν υπάρχει προγενέστερη γνώση για το κριτήριο, βάσει του οποίου γίνεται ο διαχωρισμός σε ομάδες. Ένας γνωστός αλγόριθμος ομαδοποίησης δεδομένων είναι ο k-mean Clustering.

Πίνακας 2.1: Σύγκριση Επιτηρούμενης και Μη Επιτηρούμενης Μηχανικής Μάθησης

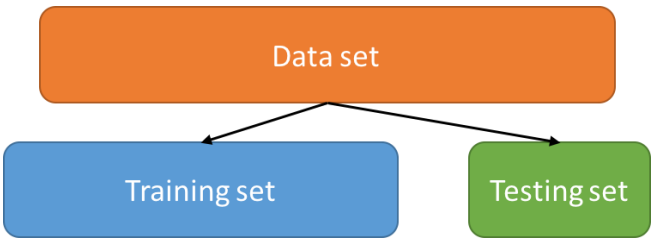
	Επιτηρούμενη	**Μη επιτηρούμενη**
Μεταβλητές εισόδου	Κωδικοποιημένα δεδομένα	Μη κωδικοποιημένα δεδομένα
Υπολογιστική πολυπλοκότητα	Απλούστερο πρόβλημα	Πολύπλοκο πρόβλημα
Ακρίβεια	Υψηλή ακρίβεια	Λιγότερη υψηλή ακρίβεια

2.2 Διαχωρισμός δεδομένων

Στην μηχανική μάθηση, το πρώτο βήμα που γίνεται για την δημιουργία μοντέλου πρόβλεψης είναι η αξιοποίηση των παρατηρήσεων και η μετατροπή τους σε δεδομένα κατάλληλα προς επεξεργασία. Επόμενο βήμα είναι ο διαχωρισμός του σύνολού δεδομένων σε δύο υποσύνολα , τα οποία είναι τα εξής :

- *Σύνολο εκπαίδευσης* : Σύνολο δεδομένων, πάνω στο οποίο βασίζεται η κατασκευή των μοντέλων πρόβλεψης.

- *Σύνολο αξιολόγησης* : Σύνολο δεδομένων , πάνω στο οποίο εξετάζεται η απόδοση των μοντέλων πρόβλεψης.



Εικόνα 2.1: Διαχωρισμός συνόλου δεδομένων

Συνήθως, το σύνολο εκπαίδευσης (training set) αποτελείται από το 70% των παρατηρήσεων και το σύνολο αξιολόγησης από το υπόλοιπο 30%. Ο διαχωρισμός του συνόλου δεδομένων σε δύο υποσύνολα γίνεται με τυχαίο τρόπο.

Μετά την δημιουργία μοντέλου πρόβλεψης , βασισμένου στο σύνολο εκπαίδευσης, γίνεται η αξιολόγηση του μοντέλου με την βοήθεια προβλέψεων στο σύνολο αξιολόγησης. Με αυτόν τον τρόπο, εξετάζεται η απόδοση του μοντέλου που δημιουργήθηκε, διότι το σύνολο αξιολόγησης περιέχει 'γνωστές' τιμές για την μεταβλητή που θα προβλεφθεί . Συνεπώς, είναι εύκολο να ελεγχθεί αν οι προβλέψεις που πραγματοποιήθηκαν στο μοντέλο αξιολόγησης είναι σωστές.

2.3 Υπερεκπαίδευση

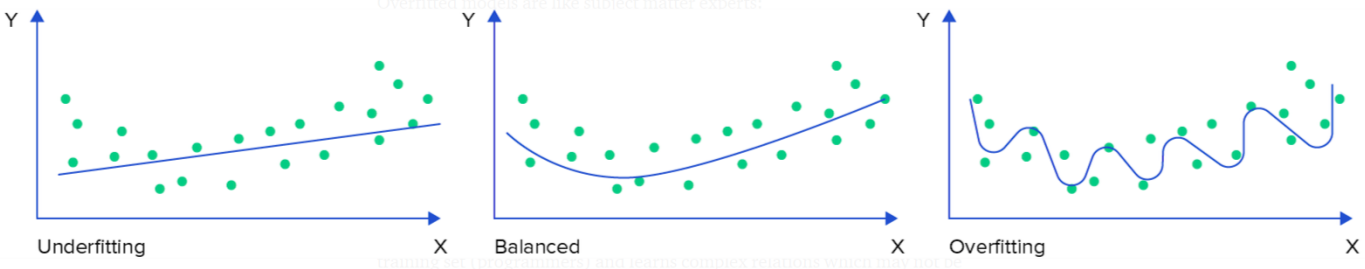
Το μεγαλύτερο πρόβλημα που συναντάται στην εφαρμογή αλγορίθμων επιτηρούμενης μηχανικής μάθησης είναι η υπερεκπαίδευση και η υποεκπαίδευση.

Υπερεκπαίδευση (overfitting)

Η υπερεκπαίδευση αναφέρεται σε ένα μοντέλο που έχει προσαρμοστεί στο σύνολο εκπαίδευσης σχεδόν τέλεια. Συμβαίνει όταν ένα μοντέλο μαθαίνει κάθε λεπτομέρεια του συνόλου εκπαίδευσης, σε βαθμό που επηρεάζει την απόδοση του μοντέλου στα νέα δεδομένα. Αυτό έχει σαν αποτέλεσμα να έχει αρνητικό αντίκτυπο στα νέα δεδομένα με αποτέλεσμα το μοντέλο να μην είναι ικανό να γενικευθεί. (3)

Υποεκπαίδευση (underfitting)

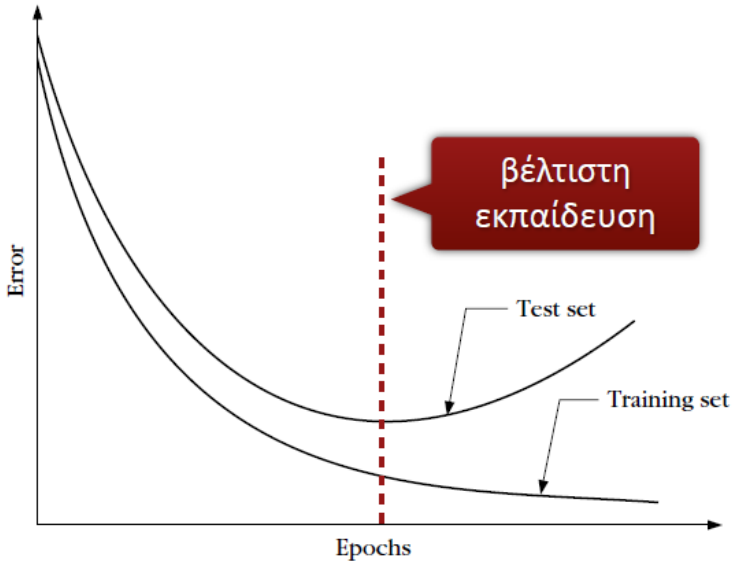
Η υποεκπαίδευση αναφέρεται σε ένα μοντέλο που δεν μπορεί ούτε να προσαρμοστεί στο σύνολο εκπαίδευσης ούτε να γενικευθεί σε νέα δεδομένα. Συμβαίνει όταν ένα μοντέλο είναι πολύ απλό, δηλαδή βασίζεται σε πολύ λίγες μεταβλητές ή δεν υπάρχει συσχέτιση μεταξύ αυτών και της μεταβλητής εξόδου, με αποτέλεσμα να μην έχει την ικανότητα να εκπαιδευθεί. (3)



Εικόνα 2.2: Υποεκπαίδευση - Υπερεκπαίδευση

Στο παραπάνω διάγραμμα, παρουσιάζονται η περιπτώσεις της υπερεκπαίδευσης και της υποεκπαίδευσης. Στο πρώτο διάγραμμα φαίνεται ότι η ευθεία προσαρμογής στα δεδομένα δεν ανταποκρίνεται στην πραγματικότητα, στο 2ο παρουσιάζεται η περίπτωση ενός αποδοτικού μοντέλου και στο τελευταίο η περίπτωση υπερεκπαίδευσης, όπου η ευθεία έχει προσαρμοστεί τέλεια στα δεδομένα.

Ακολούθως παρουσιάζεται η βέλτιστη εκπαίδευση έτσι ώστε να μην συμβεί υπερεκπαίδευση.



Εικόνα 2.3: Βέλτιστη Εκπαίδευση

Κεφάλαιο 3 : Θεωρία

3.1 Προ - Επεξεργασία δεδομένων

Συχνά, τα σύνολα δεδομένων περιέχουν χαρακτηριστικά με διαφορετικό εύρος τιμών, διαφορετικές κλίμακες μέτρησης ή διαφορετικές μονάδες μέτρησης. Για αυτό καθίσταται αναγκαία η επεξεργασία των δεδομένων, καθώς οι περισσότεροι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν την Ευκλείδεια απόσταση, η οποία επηρεάζεται από τα εύρος τιμών των μεταβλητών. Συνεπώς, αν οι τιμές ενός χαρακτηριστικού είναι μεγαλύτερης τάξης από τις τιμές ενός άλλου χαρακτηριστικού, τότε το συγκεκριμένο χαρακτηριστικό θα αποκτήσει μεγαλύτερο βάρος στην κατασκευή του μοντέλου, κάτι που δεν είναι σωστό.

Οι μέθοδοι που χρησιμοποιούνται για να γίνουν οι τιμές των χαρακτηριστικών συγκρίσιμες είναι οι παρακάτω:

- Κανονικοποίηση

  Η κανονικοποίηση αντικαθιστά τις τιμές των μεταβλητών με τον παρακάτω τύπο:

$$x' = \frac{x - \mu}{\sigma} \quad (3.1)$$

  Όπου *x* : η τιμή της μεταβλητής
  x':η τιμή της νέας μεταβλητής
  μ : μέση τιμή της μεταβλητής
  σ : τυπική απόκλιση της μεταβλητής

  Με τον συγκεκριμένο τρόπο τα χαρακτηριστικά αποκτούν τιμές με μέση τιμή ≈ 0 και τυπική απόκλιση ≈ 1 , δηλαδή αποκτούν μια κοινή κλίμακα τιμών.

- Αλλαγή κλίμακας(‘*scaling*’)

  Με την αλλαγή κλίμακας γίνεται αντικατάσταση των τιμών των μεταβλητών με τον παρακάτω τύπο:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \left[(\max - min) + min\right] \quad (3.2)$$

  Όπου *x* : η τιμή της μεταβλητής
  x': η τιμή της νέας μεταβλητής
  min(x) : η ελάχιστη τιμή της μεταβλητής
  max(x): η μέγιστη τιμή της μεταβλητής
  max, min: μέγιστη και ελάχιστη τιμή του διαστήματος

  Οι τιμές των μεταβλητών αποκτούν εύρος τιμών στο διάστημα (min,max).Συνήθως μετασχηματίζονται στο [0,1] ή [-1,1] πριν την κατασκευή νευρωνικών δικτύων.

3.2 Παλινδρόμηση

Η παλινδρόμηση είναι μια μέθοδος πρόβλεψης μιας μεταβλητής, μεταβλητή απόκρισης Y, όταν γνωρίζουμε τις τιμές κάποιας ή κάποιων άλλων μεταβλητών X, επεξηγηματικές μεταβλητές. Αυτή η τεχνική χρησιμοποιείται για την μοντελοποίηση και την ανάλυση αριθμητικών δεδομένων και πιο συγκεκριμένα για την κατασκευή ενός μοντέλου, έστω Y= g(X), έτσι ώστε στο μέλλον να γίνει ο προσδιορισμός της τιμής του Y με βάση την τιμή του X.

3.2.1 Γραμμική Παλινδρόμηση

Η γραμμική παλινδρόμηση είναι μια τεχνική πρόβλεψης μιας γραμμικής σχέσης μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών.

Ένα *απλό γραμμικό μοντέλο* είναι ένα μοντέλο με την παρακάτω μορφή :

$$y = w_0 + w_1 x + \varepsilon \quad (3.3)$$

όπου *y*: μεταβλητή απόκρισης(εξαρτημένη μεταβλητή)

x : επεξηγηματική μεταβλητή(ανεξάρτητη μεταβλητή)

ε : σφάλμα πρόβλεψης

Η συγκεκριμένη τεχνική μπορεί να γενικευθεί σε περιπτώσεις, όπου το σύνολο δεδομένων έχει περισσότερες από μία μεταβλητές και τότε πρέπει να κατασκευαστεί ένα *πολλαπλό γραμμικό μοντέλο*. Ένα πολλαπλό γραμμικό μοντέλο είναι της παρακάτω μορφής :

$$y = w_0 + w_i x_i \, , i = 1, \dots , N \quad (3.4)$$

όπου w_i : συντελεστές ή βάρη του μοντέλου

w_0 : σταθερά του μοντέλου

N : αριθμός των μεταβλητών

Πιο συγκεκριμένα η γραμμική παλινδρόμηση υπολογίζει τα βάρη w_i για το υπερεπίπεδο που αντιστοιχεί στην εξίσωση $y - w_0 + w_i x_i \geq 0$. Στον δυσδιάστατο χώρο αντιστοιχεί στην ευθεία γραμμή της μορφής $y - w_0 + w_i x_i = 0$.

Πραγματοποιείται η εκπαίδευση ενός μοντέλου γραμμικής παλινδρόμησης χρησιμοποιώντας το σύνολο εκπαίδευσης και στην συνέχεια με βάση την εξίσωση (3.4) πραγματοποιούνται οι προβλέψεις στο σύνολο αξιολόγησης. Ο όρος ‘εκπαίδευση’ αναφέρεται στην εκτίμηση των συντελεστών w_i του μοντέλου που ελαχιστοποιούν την απόκλιση από την πραγματική τιμή της Y, δηλαδή το σφάλμα. Για την εύρεση των συντελεστών, χρησιμοποιείται η *μέθοδος των Ελαχίστων Τετραγώνων*.

Μέθοδος Ελαχίστων Τετραγώνων

Με την μέθοδο Ελαχίστων Τετραγώνων, επιλέγεται την ευθεία εκείνη που προσαρμόζεται καλύτερα στα δεδομένα που διαθέτουμε. Πιο συγκεκριμένα, επιλέγεται την ευθεία που ελαχιστοποιεί το σφάλμα, δηλαδή την απόκλιση της προβλεπόμενης τιμής από την πραγματική.

Έστω ένα σύνολο εκπαίδευσης $X = \{ x_1,\ldots, x_N\}$ με $xi = (x_{i1},\ldots,x_{id})$ και $\hat{y}_i = y(x_i)$, την προβλεπόμενη τιμή.

Η ποσότητα που εκφράζει το συνολικό σφάλμα ονομάζεται συνάρτηση κόστους και δίνεται από τον παρακάτω τύπο :

$$L(w) = \sum_{i=1}^{N} (y_i - w_0 - w_i x_i)^2 = \sum_{\iota=1}^{N} e_i{}^2 \quad (3.5)$$

Ο στόχος λοιπόν είναι η ελαχιστοποίηση της συνάρτησης κόστους ή αλλιώς του αθροίσματος τετραγωνικών σφαλμάτων e_i με τον προσδιορισμό των κατάλληλων συντελεστών.

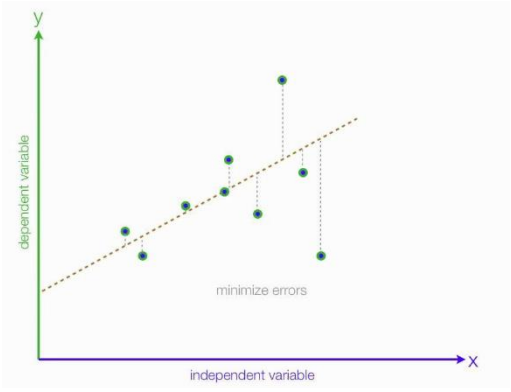
Στην περίπτωση της απλής γραμμικής παλινδρόμησης, ο προσδιορισμός των συντελεστών που ελαχιστοποιούν την παραπάνω σχέση πραγματοποιείται με τη επίλυση δύο απλών εξισώσεων:

$$\hat{w}_o = \bar{y} - \hat{w}_1 \quad (3.6)$$

$$, i = 1, .. , d$$

$$\hat{w}_1 = \frac{\sum_{i=1}^{d}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{d}(x_i - \bar{x})^2} \quad (3.7)$$

όπου $\bar{x}, \bar{y}$ οι μέσες τιμές της επεξηγηματικής μεταβλητής και της μεταβλητής απόκρισης αντίστοιχα.

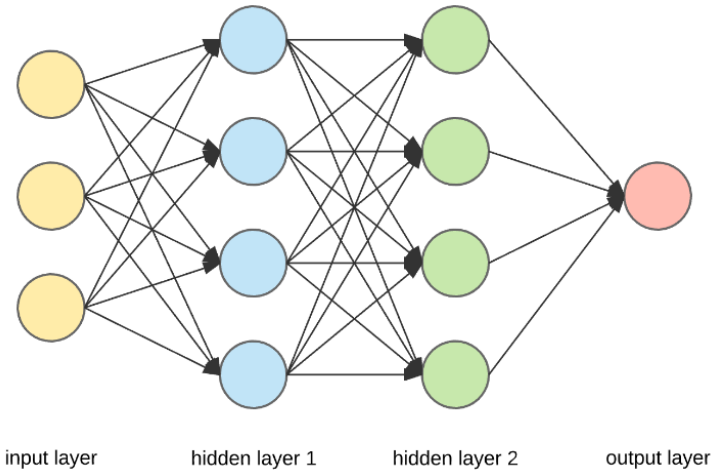


Εικόνα 3.1: Εκτιμώμενη ευθεία παλινδρόμησης με την εφαρμογή της μεθόδου Ελαχίστων Τετραγώνων

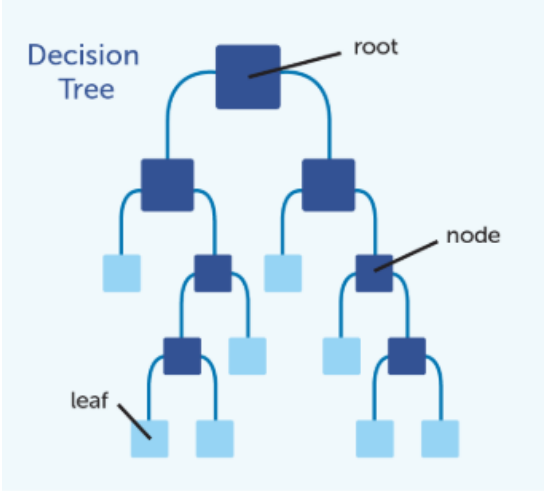
3.2.2 Μη γραμμική Παλινδρόμηση

Η μέθοδος της γραμμικής παλινδρόμησης εφαρμόζεται αρκετά καλά σε σύνολα, όπου η μεταβλητή απόκρισης έχει γραμμική σχέση με τις επεξηγηματικές μεταβλητές. Όμως, τις περισσότερες φορές, η καμπύλη πρέπει να προσαρμοστεί κατάλληλα πάνω στα δεδομένα, έτσι ώστε να δημιουργηθεί ένα μοντέλο πρόβλεψης με υψηλή απόδοση. Με τους αλγόριθμους μηχανικής μάθησης, κατασκευάζονται μη γραμμικά μοντέλα, τα οποία αποδίδουν καλύτερα από ένα κλασσικό γραμμικό μοντέλο. Οι αλγόριθμοι που χρησιμοποιήθηκαν για την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας είναι οι παρακάτω:

- Δέντρα αποφάσεων (‘Decision Trees’)
- Νευρωνικά Δίκτυα (‘Neural Networks’)



Εικόνα 3.2: Αναπαράσταση Νευρωνικού Δικτύου

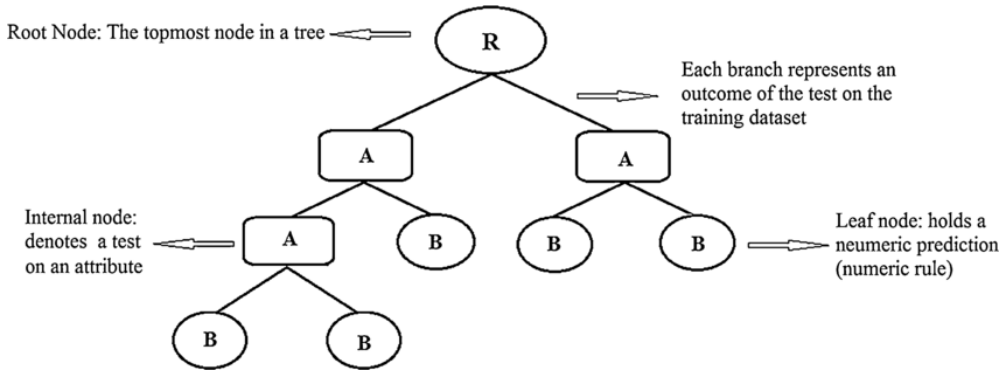


Εικόνα 3.3 : Αναπαράσταση Δέντρου απόφασης

3.3 Δέντρα απόφασης

Τα δέντρα απόφασης αποτελούν μία από τις πιο απλές και χρήσιμες δομές στον τομέα της Μηχανικής Μάθησης. Τα δέντρα απόφασης είναι μοντέλα της επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιούνται στην παλινδρόμηση (4). Στόχος τους είναι η πρόβλεψη της τιμής της μεταβλητής απόκρισης μέσω κανόνων εκμάθησης από τις επεξηγηματικές μεταβλητές. (5)

Ένα δέντρο απόφασης είναι μια δομή, όπου κάθε εσωτερικός κόμβος('decision node') αντιπροσωπεύει μια δοκιμή σε ένα χαρακτηριστικό , κάθε κόμβος φύλλου ή τερματικός κόμβος('leaf/terminal node') αντιπροσωπεύει το τελικό αποτέλεσμα, δηλαδή την προβλεπόμενη τιμή και τα 'κλαδιά'('branches') αντιπροσωπεύουν τις συζεύξεις χαρακτηριστικών που οδηγούν στους τελικούς κόμβους (4). Η δομή ενός δέντρου απόφασης παρουσιάζεται παρακάτω:



Εικόνα 3.4 : Δομή Δέντρου απόφασης

Ένα δέντρο απόφασης ακολουθεί την παρακάτω διαδικασία:

- Υποθέτουμε ότι διαθέτουμε ένα σύνολο εκπαίδευσης $X = \{x_1,\ldots, x_N\}$ με $x_ι = (x_{i1},\ldots,x_{id})$ και $\hat{y}_i = y(x_i)$, την προβλεπόμενη τιμή.

- Για κάθε μεταβλητή $x_1,\ldots, x_N$ και για κάθε τιμή $v \in R$ ακολουθείται η εξής διαδικασία:
  1. Διαχωρίζεται το σύνολο εκπαίδευσης σε δύο υποσύνολα :

$$I_< = \{i : x_{ij} < v\} \text{ και } I_> = \{i : x_{ij} \geq v\}$$

  2. Υπολογισμός της προβλεπόμενης τιμής κάθε δείγματος εκπαίδευσης

$$\beta_< = \frac{\sum_{ι \in I_<} y_i}{|I_<|} \text{ και } \beta_> = \frac{\sum_{ι \in I_>} y_i}{|I_>|}$$

3. Υπολογισμός μέσου τετραγωνικού σφάλματος μέσω του τύπου:

$$MSE = \sum_{\iota \in I_<} (y_i - \beta_<)^2 + \sum_{\iota \in I_>} (y_i - \beta_>)^2$$

- Επιλογή διαχωρισμού δείγματος εκπαίδευσης με την μικρότερη τιμή μέσου τετραγωνικού σφάλματος , δηλαδή επιλογή της καλύτερης μεταβλητής.

Η παραπάνω διαδικασία εφαρμόζεται σε κάθε κόμβο του δέντρου απόφασης. Στην αρχή της διαδικασίας, γίνεται ο πρώτος διαχωρισμός δεδομένων σε ολόκληρο το σύνολο εκπαίδευσης στον 1^o πρώτο κόμβο και ακολουθείται η ίδια διαδικασία στους επόμενους εσωτερικούς κόμβους με αρχικό μέγεθος δείγματος εκπαίδευσης αυτό που προκύπτει μετά από τον διαχωρισμό.

Παρατηρήσεις

- Κατά την διάρκεια της διαδικασίας μπορεί από έναν κόμβο να μην προκύπτει άλλος διαχωρισμός διότι υπάρχει μόνο ένα δείγμα αξιολόγησης, τότε αυτός ο κόμβος ονομάζεται κόμβος φύλλου(‘leaf node’).
- Πολλές φορές, επιλέγεται το μέγιστο βάθος του δέντρου απόφασης να είναι μικρό έτσι ώστε να αποφευχθεί η υπερεκπαίδευση. Ένα δέντρο απόφασης με μεγάλο βάθος έχει μεγάλο ρίσκο να υπερεκπαιδευθεί και για αυτόν τον λόγο ‘επεμβαίνουμε’ σε ορισμένες παραμέτρους ενός δέντρου απόφασης.

Πλεονεκτήματα και Μειονεκτήματα Δέντρων απόφασης

✓ Εύκολα στην κατανόηση, ερμηνεία των αποτελεσμάτων και στην οπτικοποίηση
✓ Λιγότερη προ-επεξεργασία των δεδομένων, διότι δεν επηρεάζεται τόσο από ακραίες τιμές(‘outliers’) και ελλιπείς τιμές(‘missing values’)
✓ Εφαρμογή σε αριθμητικές και κατηγορικές μεταβλητές
✓ Μη παραμετρική μέθοδος, δηλαδή δεν χρησιμοποιεί υποθέσεις για τον χώρο κατανομής των μεταβλητών
✓ Μπορεί να εντοπίσει μη γραμμικές σχέσεις μεταξύ των μεταβλητών

× Υπερεκπαίδευση των δεδομένων(‘overfitting’)
× Έλλειψη πληροφορίας όταν κατηγοριοποιεί μεταβλητές σε διαφορετικές κατηγορίες.
× Αστάθεια, δηλαδή μικρές διακυμάνσεις στο σύνολο δεδομένων μπορεί να προκαλέσουν μεγάλες αλλαγές στην δομή του δέντρου απόφασης.

3.3.1 'Ensemble' μέθοδοι

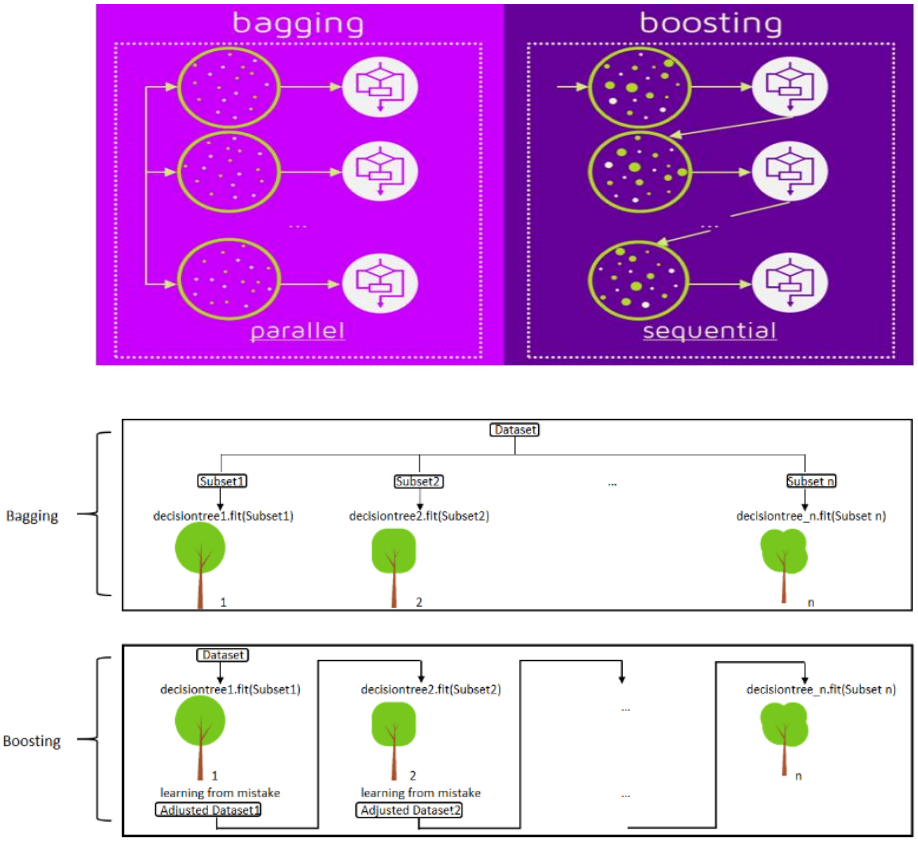
Κατά την δημιουργία δέντρων απόφασης υπάρχουν διάφοροι παράγοντες που επηρεάζουν τα αποτελέσματα που προκύπτουν, όπως η σημαντικότητα μιας μεταβλητής που θα χρησιμοποιηθεί για τον διαχωρισμό του δείγματος σε έναν κόμβο. Για αυτό τον λόγο, ο συνδυασμός πολλών δέντρων απόφασης μπορεί να λειτουργήσει καλύτερα από ένα δέντρο. Αυτή είναι η κύρια ιδέα των 'ensemble' μεθόδων.

Η ιδέα αυτών των τεχνικών είναι ότι συνδυάζουν 'αδύναμα' μοντέλα('weak learners') έτσι ώστε να δημιουργήσουν ένα 'ισχυρό' μοντέλο. Στην συγκεκριμένη περίπτωση, ως αδύναμα μοντέλα μπορούν να χαρακτηριστούν τα δέντρα απόφασης και ως ισχυρά ο συνδυασμός αυτών, δηλαδή ένας δάσος δέντρων απόφασης.

Οι δυο τεχνικές που χρησιμοποιούνται είναι οι παρακάτω:

- 'Bagging': *Στόχος αυτής της τεχνικής είναι η μείωση της διασποράς ενός δέντρου απόφασης. Η ιδέα αυτής της τεχνικής είναι η δημιουργία διαφόρων υποσυνόλων από το σύνολο εκπαίδευσης , τα οποία επιλέγονται με τυχαίο τρόπο και η εκπαίδευση ενός δέντρου απόφασης από αυτά. Το τελικό αποτέλεσμα είναι ο μέσος όρος όλων των προβλέψεων που προκύπτουν από τα διαφορετικά δέντρα.* (6)

- 'Boosting': *Στόχος αυτής της τεχνικής είναι να δημιουργήσει μια συλλογή από διάφορα δέντρα απόφασης. Η ιδέα είναι να 'μάθει' κάθε δέντρο απόφασης από το προηγούμενο δέντρο απόφασης. Πιο συγκεκριμένα, το σύνολο εκπαίδευσης που χρησιμοποιείται για κάθε μέλος αυτής της συλλογής επιλέγεται με βάση τη απόδοση του προηγούμενου δέντρου απόφασης. Οι παρατηρήσεις που έχουν μεγάλη απόκλιση ανάμεσα στην προβλεπόμενη και την πραγματική τιμή επιλέγονται πιο συχνά από αυτές που έχουν μικρότερη απόκλιση.*
  Το τελικό αποτέλεσμα είναι το βεβαρυμμένο μέσο όρο ('weighted average') όλων των δέντρων απόφασης. Με άλλα λόγια, προσαρμόζονται διαδοχικά δέντρα απόφασης και στόχος είναι να μειωθεί το σφάλμα βάσει του προηγούμενου. (7)



Εικόνα 3.5 : Bagging VS Boosting

Οι δύο παραπάνω τεχνικές δίνουν μοντέλα με καλύτερη απόδοση από αυτά που προκύπτουν με ένα δέντρο απόφασης και για αυτό είναι πολύ γνώστες. Στην συνέχεια, για την υλοποίηση της συγκεκριμένης διπλωματικής εργασίας θα χρησιμοποιηθεί η μέθοδος ενδυνάμωσης(‘boosting’).

3.3.2 Adaboost Μέθοδος

Input:

- Έστω (x_1, y_1) , … , (x_N, y_N) η ακολουθία των δειγμάτων του συνόλου εκπαίδευσης, όπου y∈R
- Κατασκευή 'weak learner'
- Αριθμός επαναλήψεων: ακέραιος αριθμός T

Αρχικοποίηση:

- Αριθμός επανάληψης t = 1
- Αρχικοποίηση κατανομής βαρών $w_t(i) = 1 / n$, για ∀i =1,… N
- Μέσος όρος της συνάρτησης κόστους $\overline{L_t} = 0$

Επαναλήψεις όσο $\overline{L_t} < 0.5$ ή $t \leq T$

- Κάθε παρατήρηση του συνόλου εκπαίδευσης του 'weak learner' αντιστοιχίζεται με το αντίστοιχο βάρος
- Δημιουργία μοντέλου παλινδρόμησης : $f_t(x) \rightarrow y$
- Υπολογισμός τιμής του σφάλματος για κάθε παρατήρηση του συνόλου εκπαίδευσης : $l_t(i) = | f_t(x_i) - y_i |$
- Υπολογισμός της τιμής της συνάρτησης κόστους $L_t(i)$ για κάθε παρατήρηση του συνόλου εκπαίδευσης, χρησιμοποιώντας μία από τις παρακάτω μορφές:
  - ☑ Γραμμική : $L_t(i) = \frac{l_t(i)}{\max_{i=1,\dots,N} l\_t\ (i)}$
  - ☑ Εκθετική : $L_t(i) = 1 - \exp(-\frac{l_t(i)}{\max_{i=1,\dots,N} l_t(i)})$ σ
  - ☑ Τετραγωνική : $L_t(i) = (\frac{l_t(i)}{\max_{i=1,\dots,N} l\_t\ (i)})^2$

- Υπολογισμός του μέσου όρου του σφάλματος : $\overline{L_t} = \sum_{i=1}^{N} L_t(i) D_t(i)$
- Ορισμός $\beta_t = \frac{\overline{L_t}}{1 - \overline{L_t}}$
- Αλλαγή τιμής βάρους : $w_{t+1} = \frac{D_t \beta_t^{(1-L_t(i))}}{Z_t}$,όπου Z_t ένας παράγοντας κανονικοποίησης έτσι ώστε D_t να είναι κατανομή.
- Ορισμός t =t +1

Output: Τελική υπόθεση

- $f_{fin}(x) = \inf(\ y \in Y : \ \sum_{t:\ f_t(x) \leq y} log \frac{1}{\beta_t} \geq \frac{1}{2} \sum_t log \frac{1}{\beta_t})$

Μία από τις πιο γνωστούς 'boosting' αλγορίθμους για προβλήματα παλινδρόμησης είναι ο Adaboost. Παρακάτω παρατίθεται ο κώδικας (8), ο οποίος θα εξηγηθεί στην συνέχεια:

Input:

- Έστω (x_1, y_1) , … , (x_N, y_N) η ακολουθία των δειγμάτων του συνόλου εκπαίδευσης, όπου $y \in R$
- Επιλέγεται ο αριθμός δέντρων απόφασης T

Αρχικοποίηση:

- t = 1
- Το βάρος κάθε παρατήρησης του συνόλου εκπαίδευσης έχει ομοιόμορφη κατανομή, ώστε κάθε παρατήρηση να έχει την ίδια ευκαιρία να 'επιλεχθεί' στο πρώτο σύνολο εκπαίδευσης, δηλαδή στο πρώτο δέντρο απόφασης.
- Ο μέσος όρος της συνάρτησης κόστους(επιλέγεται μία από τις παραπάνω 3) λαμβάνει την τιμή 0.

Ο λόγος που χρησιμοποιείται ο μέσος όρος $\bar{L_t}$ είναι διότι επηρεάζει την παράμετρο β_t, η οποία ανανεώνει τα βάρη κάθε παρατήρησης με καινούργιες τιμές. Όσο πιο χαμηλή είναι η τιμή του $\bar{L_t}$, τόσο πιο χαμηλή είναι η τιμή του β_t, με αποτέλεσμα η τιμή του $\log(1/\beta_t)$(βάρος κάθε δέντρου απόφασης) να είναι μεγαλύτερη και κατά συνέπεια στο τελευταίο βήμα να έχει μεγαλύτερη επιρροή στην τελική απόφαση.

Επαναλήψεις όσο $\bar{L_t} < 0.5$ ή $t \leq T$

- Εκπαίδευση του δέντρου απόφασης
- Υπολογισμός του σφάλματος κάθε δείγματος & της τιμής της συνάρτησης κόστους και στην συνέχεια του μέσου όρου της.
- Υπολογισμός της τιμής του β_t και ανανέωση των βαρών κάθε δείγματος με την καινούργια κατανομή. Το βάρος κάθε δείγματος πολλαπλασιάζεται με την τιμή του β_t και στην συνέχεια όλα τα βάρη κανονικοποιούνται για να αποτελούν κατανομή.

Output: Τελική Υπόθεση:

- Η τιμή της εξόδου είναι η επιβαρυμένη διάμεσός (9)

3.3.3 Gradient Boosting Μέθοδος

Η μέθοδος Gradient Boosting είναι μια επέκταση μιας μεθόδου ενδυνάμωσης. Πιο συγκεκριμένα, χρησιμοποιεί τον 'gradient descent' αλγόριθμο.

Gradient Boosting= Gradient Descent + Boosting.

Η διαδικασία εκμάθησης στην συγκεκριμένη μέθοδο προσαρμόζεται διαδοχικά στα μοντέλα έτσι ώστε να παρέχει μια ακριβέστερη εκτίμηση της τιμής της μεταβλητής απόκρισης. Η κύρια διαφορά ανάμεσα στους αλγόριθμους ενδυνάμωσης, Adaboost και Gradient Descent, είναι ο τρόπος με τον οποίο εντοπίζουν τις 'αδυναμίες' των 'weak learners', δηλαδή ο αλγόριθμος Adaboost επιβαρύνει τα δείγματα με υψηλό βάρος, ενώ ο αλγόριθμος Gradient Descent χρησιμοποιεί τις κλίσεις(gradient) συναρτήσεων κόστους. (6)

Ο αλγόριθμος Gradient Boosting περιλαμβάνει 3 στοιχεία :

1. Βελτιστοποίηση συνάρτησης κόστους : *Η επιλογή συνάρτησης κόστους εξαρτάται από τύπο προβλήματος που πρέπει να επιλυθεί.*

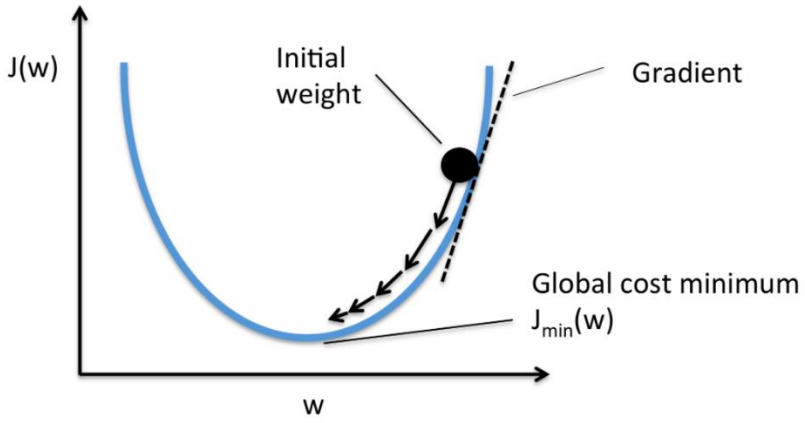
2. Έναν 'weak learner' που πρέπει να κάνει προβλέψεις : *Τα δέντρα απόφασης χρησιμοποιούνται, όπως και στην Adaboost , ως 'weak learners'.*

3. Ένα πρόσθετο μοντέλο που προσθέτει τους 'weak learners' έτσι ώστε να ελαχιστοποιηθεί η τιμή της συνάρτησης κόστους : *Η τεχνική της Gradient Descent είναι να ελαχιστοποιεί την τιμή της συνάρτησης κόστους, με την προσθήκη δέντρων*. (10)

Όσον αφορά τον αλγόριθμο Gradient Descent, παρακάτω παρουσιάζεται συνοπτικά η διαδικασία που ακολουθεί ο συγκεκριμένος αλγόριθμος:

Στόχος του είναι η ελαχιστοποίηση της συνάρτησης κόστους J(w), η οποία είναι παραμετροποιημένη από ένα μοντέλο παραμέτρων w. (11)

I. Αρχικοποίηση των βαρών με τυχαίο τρόπο
II. Υπολογισμός της κλίσης G της συνάρτησης κόστους , δηλαδή $G = \partial J(w)/\partial w$. Η τιμή της εξαρτάται από τις τιμές των παραμέτρων του μοντέλου και την συνάρτηση κόστους.
III. Ανανέωση τιμών των βαρών σύμφωνα με τον τύπο: $w = w - \eta G$, όπου η καλείται ρυθμός εκμάθησης('learning rate') και καθορίζει το αριθμό των βημάτων που χρειάζονται για να βρεθεί το ελάχιστο της συνάρτησης κόστους.
IV. Επανάληψη μέχρι να σταματήσει να μειώνεται η τιμή της συνάρτησης κόστους ή να τερματιστεί κάποιο κριτήριο.



Εικόνα 3.6: Οπτικοποίηση Αλγορίθμου Gradient Descent

Παρακάτω παρατίθεται ο κώδικας της μεθόδου Gradient Boosting (12), ο οποίος θα αναλυθεί στην συνέχεια:

Input

- Έστω (x_1, y_1) , … , (x_N, y_N) η ακολουθία των δειγμάτων του συνόλου εκπαίδευσης, όπου $y \in R$
- Έστω συνάρτηση κόστους, η οποία είναι παραγωγίσιμη, $L(y_i, F(x))$
- Αριθμός επαναλήψεων, ακέραιος αριθμός M

Αρχικοποίηση

- Αρχικοποίηση μοντέλου με μια σταθερή τιμή: $F_o(x) = arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$
- Αριθμός επανάληψης m=0

Επαναλήψεις όσο $m \leq M$

- Υπολογισμός $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \dots, N$
- Προσαρμογή δέντρου απόφασης στις τιμές r_{im} και δημιουργία τερματικών κόμβων $\{R_{jm}\}_1^{J_M}$
- Για j=1,…, J_M : υπολογισμός $\gamma_{jm} = arg\min_{\gamma} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i + \gamma))$
- Ανανέωση τιμών : $F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_m I(x \in R_{jm})$

Output:

$F_M(x)$

Input:

- Έστω (x_1, y_1) , … , (x_N, y_N) η ακολουθία των δειγμάτων του συνόλου εκπαίδευσης, όπου y∈R
- Έστω συνάρτηση κόστους, η οποία είναι παραγωγίσιμη, $L(y_i,F(x))$ και αξιολογεί την απόδοση του μοντέλου, δηλαδή πόσο καλά μπορούμε να προβλέψουμε την τιμή απόκρισης. Συνήθως, χρησιμοποιείται η συνάρτηση Ελαχίστων Τετραγώνων (LS Function) , η οποία δίνεται από τον τύπο (3.8) και είναι παραγωγίσιμη :

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \text{ , όπου } \hat{y}_i : \text{προβλεπόμενη τιμή} \ \& \ i = 1, \dots N \quad (3.8)$$

Αρχικοποίηση

- Αρχικοποίηση μοντέλου με μια σταθερή τιμή: $F_o(x) = arg\min_{\gamma}\sum_{i=1}^{N} L(y_i, \gamma)$, όπου η μεταβλητή γ δηλώνει την προβλεπόμενη τιμή. Χρειάζεται να βρεθεί η τιμή του γ, δηλαδή η προβλεπόμενη τιμή, που ελαχιστοποιεί το άθροισμα της συνάρτησης κόστους κάθε δείγματος εκπαίδευσης. Με άλλα λόγια, δημιουργείται ένας κόμβος φύλλου('leaf') , όπου προβλέπει ότι όλα τα δείγματα εκπαίδευσης έχουν βάρος F_0.
- Αρχικοποίηση τιμής m =0.

Επαναλήψεις όσο m ≤ M : Δημιουργία 'for' βρόγχου στον αριθμό των δέντρων

- Υπολογισμός $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, i = 1, \dots, N$.Αν χρησιμοποιηθεί η συνάρτηση Ελαχίστων Τετραγώνων, τότε:
$$r_{im} = -2[-((y_i - \hat{y}_i)^2)] = 2(y_i - \hat{y}_i)^2 \text{ , } \iota = 1, \dots, N$$
το οποίο ονομάζεται ψευδό- σφάλμα(pseudo- residual), διότι είναι πολλαπλασιασμένο με το 2.
- Προσαρμογή δέντρου απόφασης στις τιμές r_{im} και δημιουργία τερματικών κόμβων $\{R_{jm}\}_1^{J_M}$
- Για j=1,…, J_M : υπολογίζεται $\gamma_{jm} = arg\min_{\gamma}\sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i + \gamma))$, δηλαδή χρειάζεται να βρεθεί η τιμή του γ που ελαχιστοποιεί το παραπάνω άθροισμα. Η διαφορά με την εύρεση του F_0 είναι ότι λαμβάνεται υπόψη η προηγουμένη προβλεπόμενη τιμή F_{m-1} και ότι αφορά τα δείγματα σε κάθε τερματικό κόμβο έτσι όπως έχουν διαχωριστεί.
- Ανανεώνονται οι τιμές των προβλέψεων $F_m(x) = F_{m-1}(x) + v\sum_{j=1}^{J_m}\gamma_m I(x \in R_{jm})$, όπου η παράμετρος ν καλείται 'Learning Rate'. Όσο πιο χαμηλή τιμή έχει η παράμετρος ν, τόσο πιο πολύ μειώνεται η επίδραση που έχει κάθε δέντρο στην τελική πρόβλεψη , με αποτέλεσμα να βελτιώνεται η ακρίβεια κατά την διάρκεια του αλγορίθμου.
- Output: Στο τέλος του αλγορίθμου, το τελικό αποτέλεσμα που προκύπτει είναι η προβλεπόμενη τιμή $F_M(x)$.

3.4 Νευρωνικά Δίκτυα

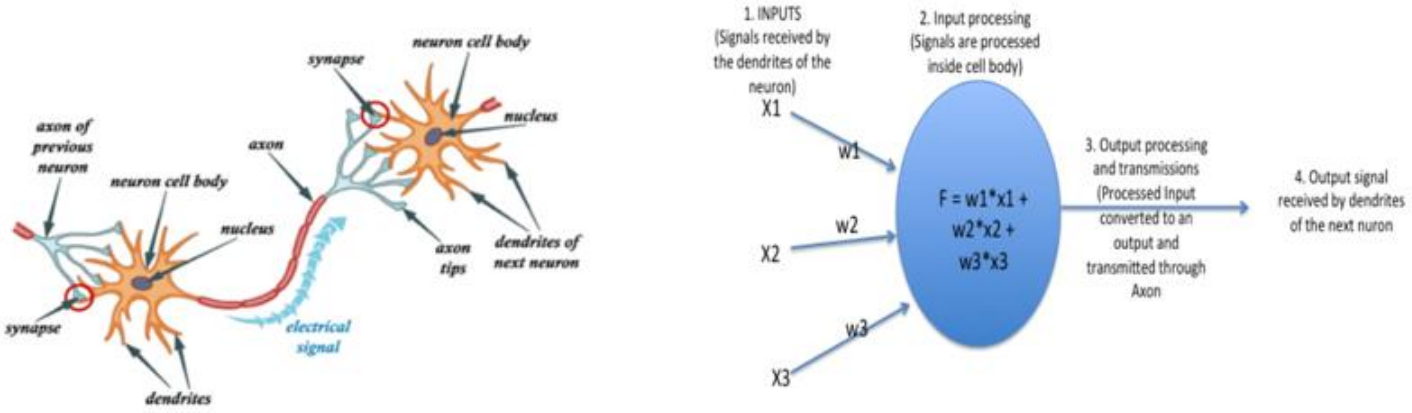
Τα νευρωνικά δίκτυα αποτελούν έναν από τους σημαντικότερους τομείς της μηχανικής μάθησης. Ο λόγος που είναι ευρέως γνωστά στον τομέα της Ανάλυσης Δεδομένων είναι διότι μπορούν να χρησιμοποιηθούν για την επίλυση πολύπλοκων καθημερινών προβλημάτων. Μπορούν να 'μάθουν' και να μοντελοποιήσουν τις σχέσεις μεταξύ των εισόδων και των εξόδων ενός μοντέλου , οι οποίες δεν συνδέονται μεταξύ τους με γραμμικό τρόπο. Αποτελούν μια εναλλακτική λύση σε κλασσικά μη γραμμικά προβλήματα.

Ορισμός Νευρωνικού Δικτύου

Ένα νευρωνικό δίκτυο είναι ένα δίκτυο που αποτελείται από ένα πλήθος απλών, αλλά με μεγάλο βαθμό διασύνδεσης, στοιχείων ή κόμβων που ονομάζονται *νευρώνες* και είναι διατεταγμένοι σε στρώματα που επεξεργάζονται πληροφορίες. Η ικανότητα επεξεργασίας του δικτύου 'αποθηκεύεται' στα βάρη , η οποία αποκτάται από μια διαδικασία αναγνώρισης προτύπων. (13)

Ονομάστηκαν νευρωνικά δίκτυα λόγω του ότι χρησιμοποιούν την διαδικασία του ανθρώπινου εγκέφαλου για να αναπτύξουν αλγόριθμους που μπορούν να λύσουν προβλήματα πρόβλεψης και να μοντελοποιήσουν περίπλοκα μοτίβα.

Ο ανθρώπινος εγκέφαλος αποτελείται από 100 δισεκατομμύρια κύτταρα ή νευρώνες, τα οποία επεξεργάζονται την πληροφορία με την μορφή ηλεκτρικών σημάτων. Εξωτερικές πληροφορίες / ερεθίσματα λαμβάνονται από τους δενδρίτες του νευρώνα, επεξεργάζονται στο σώμα των κυττάρων/νευρώνων, μετατρέπονται σε μια έξοδο και περνούν στον επόμενο νευρώνα. Ο επόμενος νευρώνας μπορεί να επιλέξει είτε να τις αποδεχθεί είτε να τις απορρίψει ανάλογα με τη δύναμη του σήματος (14). Παρακάτω παρουσιάζεται η δομή ενός Νευρωνικού δικτύου σε σχέση με την δομή του ανθρωπίνου εγκεφάλου:



Εικόνα 3.7 : Σύγκριση δομής ανθρώπινου εγκεφάλου με νευρωνικό δίκτυο

3.4.1 Perceptron

Ορισμός Perceptron

Ένας perceptron δέχεται κάποιες δυαδικές τιμές εισόδου $x_1, x_2, \ldots$, και παράγει μια τιμή εξόδου. Ο τρόπος υπολογισμού της τιμής εξόδου γίνεται με την βοήθεια των βαρών $w_1, w_2, w_3, ..$, τα οποία είναι πραγματικές τιμές που δείχνουν τον βαθμό 'σημαντικότητας' των τιμών εισόδου στην τιμή εξόδου. Η τιμή εξόδου είναι της μορφής 0 ή 1 και καθορίζεται από το αν το επιβαρυμένο άθροισμα $\sum_j w_j x_j$ είναι μεγαλύτερο ή ίσο από μια συγκεκριμένη τιμή, η οποία καλείται τιμή του κατωφλιού(threshold value). Όπως τα βάρη αποτελούν παράμετρο του μοντέλου, έτσι και το κατώφλι είναι ένας πραγματικός αριθμός που αποτελεί παράμετρο του νευρώνα.

$$\text{Εξοδος} = \begin{cases} 0, & \text{αν } \sum_j w_j x_j \leq b \\ 1, & \text{αν } \sum_j w_j x_j > b \end{cases}, \text{όπου } b: \text{ κατώφλι}$$

Ένα νευρωνικό δίκτυο αποτελείται από πολλούς Perceptron , οι οποίοι σχηματίζουν τα στρώματα ('layers'). Η διαδικασία που ακολουθεί είναι η εξής : Το πρώτο στρώμα εξάγει κάποιες αποφάσεις σύμφωνα με τις τιμές εισόδου, στην συνέχεια το δεύτερο στρώμα εξάγει τις αποφάσεις βασισμένο στα αποτελέσματα που προέκυψαν από το πρώτο στρώμα και συνεχίζεται η ίδια διαδικασία στα υπόλοιπα στρώματα μέχρι να προκύψει η τιμή εξόδου.

Όμως, μια μικρή αλλαγή στα βάρη μπορεί να προκαλέσει αλλαγή και στην τιμή εξόδου και κατά συνέπεια να δημιουργηθεί ένα μη αποδοτικό μοντέλο . Για αυτόν τον λόγο είναι αναγκαία η ύπαρξη μιας συνάρτησης ενεργοποίησης. (15)

3.4.2 Συνάρτηση ενεργοποίησης

Ορισμός συνάρτησης ενεργοποίησης:

Η συνάρτηση ενεργοποίησης είναι μια μαθηματική εξίσωση που καθορίζει την τιμή εξόδου ενός νευρώνα.

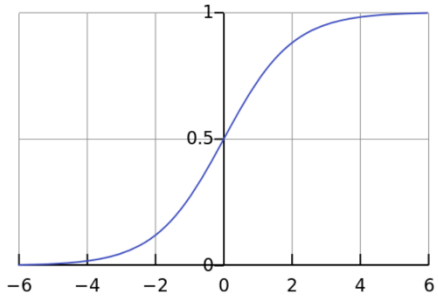
Το σύνολο τιμών που μπορεί να προκύψει από τις μεταβλητές εξόδου των νευρώνων μπορεί να είναι (-inf,+inf) . Για αυτόν τον λόγο, χρησιμοποιούνται οι συναρτήσεις ενεργοποίησης, οι οποίες αποφασίζουν αν ο νευρώνας πρέπει να 'ενεργοποιηθεί' ή όχι. Μια συνάρτηση ενεργοποίησης κανονικοποιεί την τιμή εξόδου κάθε νευρώνα στο διάστημα (0,1) ή (-1,1).

Οι συναρτήσεις ενεργοποίησης που θα χρησιμοποιηθούν στην συνεχεία είναι οι παρακάτω:

1. *Σιγμοειδής*

   Η σιγμοειδής συνάρτηση ενεργοποίησης είναι από τις πιο γνώστες στον τομέα της μηχανικής μάθησης. Ο τύπος της δίνεται ακολούθως:

$$f(x) = \frac{1}{1+e^{-x}} \quad (3.9)$$



Εικόνα 3.8 : Σιγμοειδής Συνάρτηση

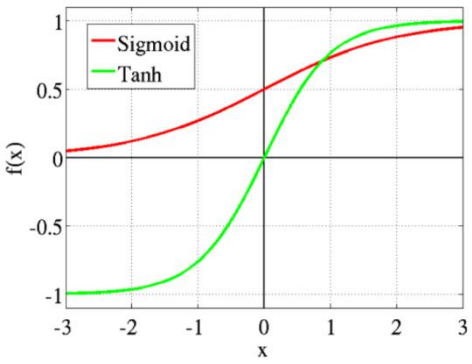
- ✓ Μη γραμμική και το σύνολο τιμών είναι το [0,1]. Γνωρίζουμε ότι η πιθανότητα ενός ενδεχομένου λαμβάνει τιμές στο διάστημα [0,1], οπότε η συγκεκριμένη συνάρτηση είναι ιδιαίτερα χρήσιμη για προβλήματα όπου χρειάζεται να γίνει η πρόβλεψη μιας πιθανότητας ως τιμή εξόδου.
- ✓ Η συγκεκριμένη συνάρτηση είναι παραγωγίσιμη , οπότε είναι δυνατή η εύρεση της κλίσης μιας σιγμοειδής καμπύλης σε δύο σημεία.
- ✓ Στο διάστημα τιμών [-2,2] οι τιμές του y μειώνονται απότομα. Μικρές αλλαγές στις τιμές του x σε αυτό το διάστημα μπορεί να επιφέρουν σημαντικές αλλαγές στις τιμές του y , με αποτέλεσμα να δημιουργούν σημαντικές αλλαγές στις προβλέψεις.

- ✗ Για πολύ υψηλές ή χαμηλές τιμές του x δεν παρατηρείται σημαντική αλλαγή στην πρόβλεψη με αποτέλεσμα να δημιουργηθεί το ‘vanishing gradient πρόβλημα’, δηλαδή η κλίση να έχει χαμηλή τιμή ή να έχει σχεδόν εξαφανιστεί. Το αποτέλεσμα που μπορεί να προκληθεί είναι να ‘αρνηθεί’ το νευρωνικό δίκτυο να εκπαιδευθεί περισσότερο ή να είναι αρκετά αργό μέχρι να φθάσει σε μια ακριβή πρόβλεψη, διότι επηρεάζεται η ανανέωση των βαρών.

2. *Υπερβολική εφαπτομένη ή Tanh*

Ο τύπος της συγκεκριμένης συνάρτησης δίνεται παρακάτω:

$$f(x) = \frac{2}{1+e^{-2x}} - 1 = 2sigmoid(2x) - 1 \quad (3.10)$$



Εικόνα 3.9 : Tanh VS Sigmoid

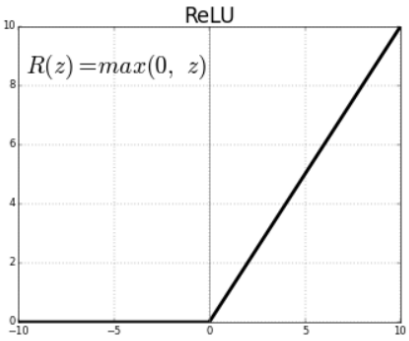
- ✓ Μη γραμμική με σύνολο τιμών στο [-1,1].
- ✓ Οι τιμές εξόδου y είναι κεντραρισμένες στο 0, με αποτέλεσμα να είναι ευκολότερο να μοντελοποιήθουν τιμές εισόδου x που είναι 'ισχυρά' αρνητικές, ουδέτερες και 'ισχυρά' θετικές.

✗ Ίδια μειονεκτήματα με την σιγμοειδή συνάρτηση

3. *Relu*

Ο τύπος της συγκεκριμένης συνάρτησης δίνεται παρακάτω:

$$f(x) = \max(0, x) \quad (3.11)$$



Εικόνα 3.10: Relu συνάρτηση

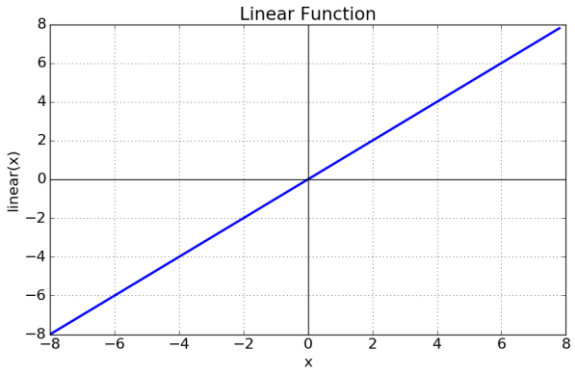
- ✓ Μη γραμμική με σύνολο τιμών [0,+inf)
- ✓ Γρήγορη σύγκλιση νευρωνικού δικτύου

× Μηδενική ή αρνητική τιμή στην τιμή εισόδου επιφέρει μηδενική τιμή στην κλίση('gradient') της συνάρτησης με αποτέλεσμα το νευρωνικό δίκτυο να μην μπορεί να εκτελέσει την 'Back propagation' μέθοδο (θα εξηγηθεί στην συνέχεια) και να μην μπορεί να μάθει.

4. *Γραμμική*

Συνήθως σε προβλήματα παλινδρόμησης στο τελευταίο στρώμα χρησιμοποιείται η γραμμική συνάρτηση ενεργοποίησης. Ο τύπος της συγκεκριμένης συνάρτησης είναι :

$$f(x) = x \quad (3.12)$$

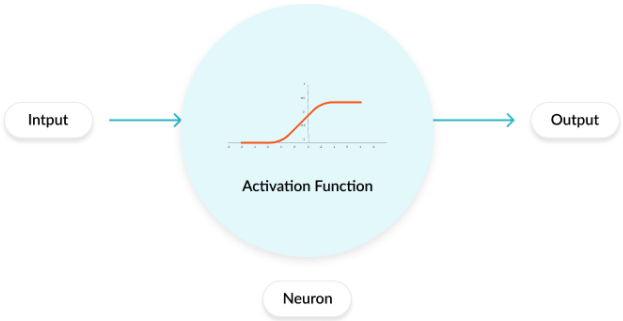


Εικόνα 3.11: Γραμμική συνάρτηση ενεργοποίησης

✓ Η γραμμική συνάρτηση αποτελεί έναν γραμμικό συνδυασμό των τιμών εισόδου του πρώτου στρώματος ανεξάρτητα από τον αριθμό των στρωμάτων.
× Σύνολο τιμών (-inf,+inf)
× Η παράγωγος της είναι 1, δηλαδή δεν εξαρτάται από την τιμή εισόδου με αποτέλεσμα στην μέθοδο 'Back propagation' τα βάρη να μην ανανεώνονται σωστά έτσι ώστε να δοθεί μια καλύτερη πρόβλεψη.

Στην συνέχεια, γνωρίζοντας πλέον την χρησιμότητα την συνάρτησης ενεργοποίησης συνοψίζεται παρακάτω η διαδικασία που ακολουθείται μεταξύ των νευρώνων :

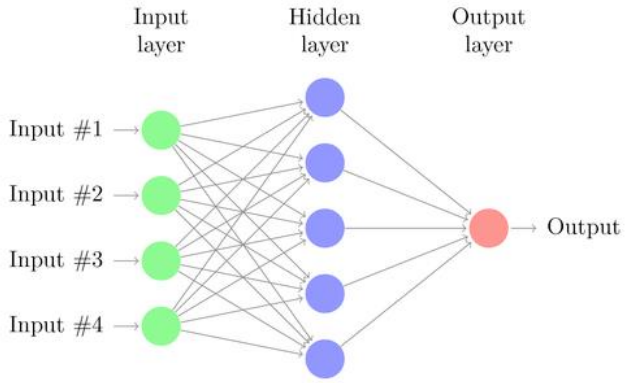
Οι τιμές εισόδου τροφοδοτούνται στον νευρώνα, έπειτα κάθε νευρώνας πολλαπλασιάζει αυτές τις τιμές με το αντίστοιχο βάρος του και προκύπτει η τιμή εξόδου του νευρώνα, η οποία μεταφέρεται στον επόμενο νευρώνα. Ο ρόλος της συνάρτησης ενεργοποίησης είναι ότι αντιστοιχίζει τις τιμές εισόδου σε τιμές εξόδου που χρειάζονται, έτσι ώστε το νευρωνικό δίκτυο να λειτουργεί σωστά.



Εικόνα 3.12 : Διαδικασία νευρωνικού δικτύου μεταξύ των νευρώνων

Αρχιτεκτονική Νευρωνικού Δικτύου

Η απλούστερη αρχιτεκτονική ενός νευρωνικού δικτύου αποτελείται από ένα στρώμα εισόδου (input layer), ένα κρυφό στρώμα(hidden layer) και ένα στρώμα εξόδου(Output layer) [4] και παρουσιάζεται παρακάτω:



Εικόνα 3.13 : Νευρωνικό Δίκτυο με ένα κρυφό στρώμα

Ορισμός συνάρτησης κόστους

Η συνάρτηση κόστους είναι μια μαθηματική συνάρτηση, η οποία υπολογίζει την απόκλισή της προβλεπόμενη τιμής, τιμή εξόδου του νευρωνικού δικτύου, με την πραγματική τιμή.

Σε προβλήματα παλινδρόμησης χρησιμοποιείται η συνάρτηση των μέσων τετραγωνικών σφαλμάτων σαν συνάρτηση κόστους, της οποίας ο τύπος δίνεται ακολούθως:

[4] Τα στρώματα των νευρωνικών δικτύων αποτελούνται από τους νευρώνες

$$L = MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 , i = 1, \dots, N \quad (3.13)$$

όπου N : αριθμός δειγμάτων

y_i: πραγματική τιμή

$\hat{y}_i$: προβλεπόμενη τιμή

3.4.3 Feed Forward Μέθοδος

Στόχος της συγκεκριμένης μεθόδου είναι ο υπολογισμός των ενεργοποιήσεων κάθε νευρώνα σε κάθε διαδοχικό κρυφό στρώμα μέχρι να καταλήξει στην τιμή εξόδου. (16)

Η διαδικασία θα αναλυθεί σε ένα νευρωνικό δίκτυο που αποτελείται από 1 κρυφό στρώμα (Εικόνα 3.13 : Νευρωνικό Δίκτυο με ένα κρυφό στρώμα).

Αρχικά, κατασκευάζονται Μ γραμμικοί συνδυασμοί των τιμών εισόδου $x_1,\dots,x_D$ με την ακόλουθη μορφή:

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} , \text{όπου } j = 1, \dots, M \ (3.14)$$

όπου $w_{ji}^{(1)}$: $\beta\acute{\alpha}\rho\eta$

$w_{j0}^{(1)}$: $\mu\epsilon\rho o\lambda\eta\psi\acute{\iota}\alpha('bias')$

a_j : ενεργοποίηση

Στην παραπάνω εξίσωση ο εκθέτης (1) δηλώνει ότι η συγκεκριμένη παράμετρος ανήκει στο 1ο στρώμα του νευρωνικού δικτύου.

Στην συνέχεια, κάθε μία από τις ενεργοποιήσεις μετατρέπεται σε τιμή εξόδου κάθε νευρώνα μέσω μιας παραγωγίσιμης, μη γραμμικής συνάρτησης ενεργοποίησης , σύμφωνα με τον ακόλουθο τύπο:

$$z_j = h(a_j) , \text{όπου } h: \text{συνάρτηση ενεργοποίησης } (3.15)$$

Η ίδια διαδικασία συνεχίζεται μέχρι το τελευταίο στρώμα, δηλαδή π.χ. για το 2ο στρώμα, το κρυφό στρώμα, οι παραπάνω τιμές z_j συνδυάζονται γραμμικά για να δώσουν τις τιμές εξόδου κάθε νευρώνα του 2ου στρώματος , σύμφωνα με τον παρακάτω τύπο :

$$a_k = \sum_{i=1}^{M} w_{kj}^{(2)} x_j + w_{k0}^{(2)} \text{ , όπου } k = 1, \dots, M \quad (3.16)$$

Στο τέλος, προκύπτει η τιμή εξόδου του νευρωνικού δικτύου, η οποία δίνεται από τον ακόλουθο τύπο :

$$y_{NN}(x, w) = f\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right) \quad (3.17)$$

όπου *f*: συνάρτηση ενεργοποίησης που χρησιμοποιείται για να δώσει τις τιμές εξόδου του νευρωνικού δικτύου. (17)

Στην συνέχεια, χρησιμοποιείται η μέθοδος 'Back propagation' για ελαχιστοποιηθεί η συνάρτηση κόστους και κατά συνέπεια να εκπαιδευθεί το νευρωνικό δίκτυο.

3.4.4 Back propagation Μέθοδος

Στόχος ενός νευρωνικού δικτύου είναι να εκπαιδευθεί , δηλαδή να ελαχιστοποιηθεί η συνάρτηση κόστους με αποτέλεσμα να ανανεωθούν τα βάρη.

Στην μέθοδο 'Back propagation' χρειάζεται να ελαχιστοποιηθεί η συνάρτηση κόστους. Ο τρόπος για να γίνει αυτή η ελαχιστοποίηση είναι να ανανεωθούν οι τιμές των βαρών, οι οποίες επηρεάζουν τις προβλεπόμενες τιμές[5]. Κάθε βάρος ανανεώνεται σύμφωνα με τον παρακάτω τύπο:

$$w_t = w_{t-1} - v_t \text{ , όπου } t : \alpha\rho\iota\theta\mu\acute{o}\varsigma\ \epsilon\pi\alpha\nu\acute{\alpha}\lambda\eta\psi\eta\varsigma \quad (3.18)$$

Ο τρόπος με τον οποίον υπολογίζεται η αλλαγή των τιμών στα βάρη εξαρτάται από τον αλγόριθμο εκμάθησης που χρησιμοποιείται. Η κλασσική μέθοδος 'Back propagation' που χρησιμοποιείται γίνεται με τον αλγόριθμο Gradient Descent[6], ο οποίος υπολογίζει την κλίση της συνάρτησης κόστους για κάθε βάρος :

$$v_t = \eta \nabla_{w_{t-1}} L(w_{t-1}) \text{ , όπου } \eta : \rho\upsilon\theta\mu\acute{o}\varsigma\ \epsilon\kappa\mu\acute{\alpha}\theta\eta\sigma\eta\varsigma \quad (3.19)$$

[5] Αναλυτικότερα στο βιβλίο Pattern Recognition and Machine Learning / Christopher M. Bishop
[6] Εξήγηση στο υποκεφάλαιο 3.3.3

Ο ρυθμός εκμάθησης παίζει σημαντικό ρόλο στην συγκεκριμένη μέθοδο. Η εύρεση κατάλληλου ρυθμού εκμάθησης μπορεί να είναι δύσκολη ορισμένες φόρες.

- Πολύ χαμηλή τιμή ρυθμού εκμάθησης μπορεί να επιφέρει μια καλή λύση αλλά μια αργή διαδικασία.
- Πολύ υψηλή τιμή ρυθμού εκμάθησης μπορεί να επιφέρει μια τελείως λάθος λύση, παρόλο που η διαδικασία είναι πιο γρήγορη.

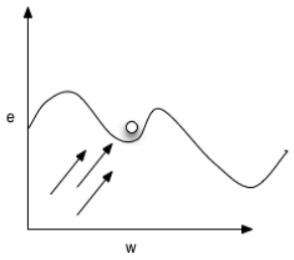
Momentum Propagation

Η ορμή (momentum) εισάγει έναν επιπλέον όρο στην εξίσωση (3.19), όπως φαίνεται παρακάτω:

$$v_t = \eta \nabla_{w_{t-1}} L(w_{t-1}) + \lambda v_{t-1}\,, \text{όπου } \lambda : \text{ορμή} \quad (3.20)$$

Η μέθοδος 'Momentum Propagation' αποτελείται από δύο παράμετρούς εκπαίδευσης: τον ρυθμό εκμάθησης και την ορμή. Η προσθήκη του επιπλέον όρου της ορμής είναι χρήσιμη γιατί εισάγεται σε τρέχων βάρος η προηγούμενη ποσότητα αλλαγής βάρους.

Με αυτή την τεχνική, μπορεί να αποφευχθεί να βρεθεί κάποιο βάρος σε ένα τοπικό ελάχιστο, διότι η ορμή δίνει επιπλέον 'δύναμη' προς μια κατεύθυνση ώστε να μετακινηθεί το βάρος από ένα τοπικό ελάχιστο.



Εικόνα 3.14: Χρησιμότητα της παραμέτρου της ορμής

Αλγόριθμος μάθησης Adam για ανανέωση τιμών των βαρών

Ο αλγόριθμος εκμάθησης Adam είναι μια μέθοδος ανανέωσης των βαρών του νευρωνικού δικτύου, η οποία υπολογίζει τους ρυθμούς εκμάθησης για κάθε παράμετρο. Πιο συγκεκριμένα, ο συγκεκριμένος αλγόριθμος εκτιμά την 1^{η} και την 2^{η} ροπή αδράνειας[7] της κλίσης(gradient) της συνάρτησης κόστους , δηλαδή τον μέσο και την διασπορά, για να καθορίσει τις διορθώσεις στα βάρη.

[7] n – οστή ροπή αδράνειας δίνεται από τον τύπο $m_n = E[x^n]$

Ο ψευδοκώδικας του συγκεκριμένου αλγόριθμου (18) δίνεται παρακάτω[8]:

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
$\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
$\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

Συνοψίζοντας, η διαδικασία που ακολουθεί το νευρωνικό δίκτυο είναι η εξής:

Αρχικά, ορίζεται η αρχιτεκτονική του νευρωνικού δικτύου, δηλαδή επιλογή αριθμού κρυφών στρωμάτων και επιλογή σε καθένα από αυτά των αριθμών των νευρώνων από τους οποίους αποτελούνται. Στην συνέχεια, γίνεται η τροφοδότηση του νευρωνικού δικτύου με τα δείγματα εκπαίδευσης. Με την διαδικασία 'Feed forward' προκύπτει η τιμή εξόδου του νευρωνικού δικτύου και έπειτα με την διαδικασία 'Back propagation' γίνεται η ανανέωση των βαρών που ελαχιστοποιούν την συνάρτηση κόστους με τον αλγόριθμο εκμάθησης που επιλέχθηκε. Αυτό αποτελεί την εποχή του νευρωνικού δικτύου. Πιο αναλυτικά, μια εποχή ορίζεται ως μια υπερπαράμετρος του μοντέλου που καθορίζει τον αριθμό των φορών που θα εφαρμοσθεί ο αλγόριθμος εκμάθησης που επιλέχθηκε σε ολόκληρο το σύνολο δεδομένων. Με άλλα λόγια, μια εποχή αποτελείται από ένα 'πέρασμα' της 'feed forward' και της 'Back propagation' σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης.

Πολλές φορές , είναι προτιμότερο να τροφοδοτείται το νευρωνικό δίκτυο με μικρότερα υποσύνολα δεδομένων του συνόλου εκπαίδευσης. Επιλέγεται ένας αριθμός δειγμάτων που χρησιμοποιείται σε ένα πέρασμα της 'Feed forward' και της 'Back propagation' και αυτό καλείται 'batch size'. Τα πλεονεκτήματα αυτής της τεχνικής είναι:

- Το 'Batch size' ελέγχει την ακρίβεια της εκτίμησης του σφάλματος της κλίσης('gradient') της συνάρτησης κόστους όταν το νευρωνικό δίκτυο εκπαιδεύεται.
- Επηρεάζει την ταχύτητα και την ευστάθεια του αλγορίθμου μάθησης.

[8] Αναλυτικότερα στην παραπομπή (14)

Ο τελευταίος ορισμός που πρέπει να δοθεί είναι αυτός των επαναλήψεων. Αριθμός επαναλήψεων είναι ο αριθμός των ‘περασμάτων’ , όπου κάθε ‘πέρασμα’ χρησιμοποιεί το ‘batch size’ και πιο συγκεκριμένα είναι η ‘feed forward’ και η ‘back propagation’.

Θα δοθεί ένα παράδειγμα για να γίνουν πιο κατανοητές οι παραπάνω έννοιες. Έστω ότι το σύνολο δεδομένων αποτελείται από 1000 δείγματα και το ‘batch size’ είναι 500, τότε θα γίνουν 2 επαναλήψεις για να ολοκληρωθεί μια εποχή.

3.5 Μετρικές αξιολόγησης μοντέλου

Η κατασκευή ενός μοντέλου παλινδρόμησης μέσω αλγορίθμων μηχανικής μάθησης αποτελεί το βασικό βήμα που πρέπει να γίνει όταν μελετάται ένα σύνολο δεδομένων. Στην συνέχεια, είναι πολύ σημαντικό να μελετηθεί η απόδοση του μοντέλου και για αυτό τον λόγο χρησιμοποιούνται διάφορες μετρικές ανάλογα με το είδος του προβλήματος (ταξινόμηση, παλινδρόμηση κλπ.). Πιο συγκεκριμένα, κατασκευάζεται ένα μοντέλο και έπειτα βάσει της γνώσης που αποκτούμε από τις μετρικές απόδοσης του μοντέλου συνεχίζουμε μέχρι να επιτύχουμε την επιθυμητή ακρίβεια.

3.5.1 Μέσο τετραγωνικό σφάλμα

Σε προβλήματα παλινδρόμησης, η πιο συνηθισμένη μετρική που χρησιμοποιείται για την αξιολόγηση του μοντέλου είναι το μέσο τετραγωνικό σφάλμα. Η συγκεκριμένη μετρική χρησιμοποιείται στην γραμμική παλινδρόμηση αλλά αποτελεί και μία από τις συνηθισμένες συναρτήσεις κόστους στα δέντρα αποφάσεων και στα νευρωνικά δίκτυα.

Έστω (x_1, y_1) , … , (x_N, y_N) η ακολουθία των δειγμάτων του συνόλου εκπαίδευσης, όπου $y \in R$, τότε το μέσο τετραγωνικό σφάλμα δίνεται από τον παρακάτω τύπο:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 , \qquad \text{όπου } \hat{y}_i : \text{προβλεπόμενη τιμή } \& \; i = 1, \dots N \quad (3.21)$$

Το μέσο τετραγωνικό σφάλμα υπολογίζει τον μέσο όρο των τετραγωνικών σφαλμάτων των προβλέψεων, δηλαδή για κάθε δείγμα αξιολόγησης υπολογίζει την τετραγωνική διαφορά μεταξύ των προβλεπόμενης τιμής και της πραγματικής τιμής και στο τέλος υπολογίζει τον μέσο όρο όλων αυτών των τιμών.

Μεγαλύτερη τιμή → *Λιγότερο αποδοτικό μοντέλο*

Μηδενική τιμή → *‘Τέλειο’ μοντέλο*

Ένα σημαντικό μειονέκτημα που παρουσιάζει η παραπάνω μετρική είναι ότι επηρεάζεται πολύ από 'noisy' δεδομένα. Μια κακή πρόβλεψη, δηλαδή με μεγάλη απόκλιση από την πραγματική τιμή, μπορεί να επηρεάσει την τιμή του τετραγωνικού σφάλματος, διότι υπολογίζει την τετραγωνική διάφορα με αποτέλεσμα να είναι δύσκολο να αποφασίσουμε πόσο καλά αποδίδει το μοντέλο που δημιουργήθηκε. Συχνά, αντί του μέσου τετραγωνικού σφάλματος χρησιμοποιείται η τετραγωνική ρίζα του, η οποία είναι πιο 'ευαίσθητη' σε 'outliers' , δηλαδή απομακρυσμένες τιμές από την πραγματική τιμή.

3.5.2 Συντελεστής προσδιορισμού R^2

Σε ένα μοντέλο παλινδρόμησης, η παρακάτω ποσότητα καλείται συντελεστής προσδιορισμού, λαμβάνει τιμές στο διάστημα [0,1] και εκφράζει το ποσοστό της διασποράς της μεταβλητής απόκρισης που εξηγείται με βάση το μοντέλο παλινδρόμησης.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \qquad \text{όπου } \hat{y}_i : \text{προβλεπόμενη τιμή } \& \; i = 1, \dots N \quad (3.22)$$

Ο αριθμητής στην παραπάνω σχέση είναι το μέσο τετραγωνικό σφάλμα και ο παρονομαστής η διασπορά της μεταβλητής απόκρισης. Όσο πιο μεγάλη τιμή έχει το MSE, τόσο μικρότερη τιμή έχει το R^2. Με την προσθήκη καινούργιων μεταβλητών στο μοντέλο η τιμή του συντελεστή προσδιορισμού βελτιώνεται και για αυτό χρησιμοποιείται πολλές φορές ο διορθωμένος συντελεστής προσδιορισμού. Ο τύπος του δίνεται παρακάτω:

$$\bar{R}^2 = R^2 - \frac{k-1}{N-k}(1 - R^2), \qquad \text{όπου } \kappa: \text{αριθμός των μεταβλητών} \quad (3.23)$$

Ο διορθωμένος συντελεστής προσδιορισμού αυξάνεται μόνο αν η μεταβλητή που προστίθεται είναι σημαντική.

3.5.3 Διασπορά και μέση τιμή των σφαλμάτων

Ένα άλλο μέτρο αξιολόγησης του μοντέλου αποτελεί η διασπορά των αποκλίσεων του μοντέλου. Στην κλασσική στατιστική, η διασπορά ενός δείγματος δίνεται από τον τύπο:

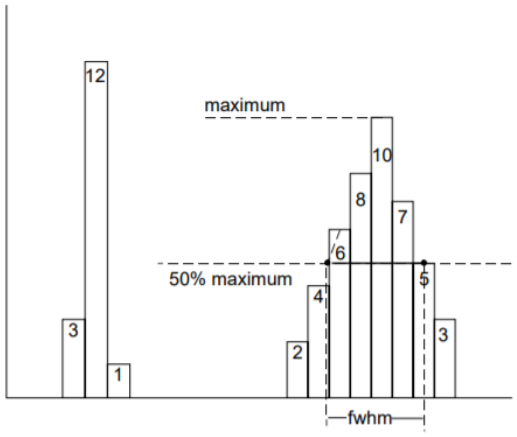
$$\sigma^2 = \frac{1}{N}\sum_{\iota=1}^{N}(x_i - \bar{x})^2 \,, (3.24)$$

Μετά την κατασκευή του μοντέλου προκύπτουν οι προβλεπόμενες τιμές της μεταβλητής απόκρισης, με βάση αυτές υπολογίζονται οι αποκλίσεις τους από τις πραγματικές τιμές. Ένας τρόπος αξιολόγησης της απόδοσης του μοντέλου είναι ο υπολογισμός της διασποράς των αποκλίσεων. Όμως, η διασπορά των αποκλίσεων είναι 'ευάλωτη' σε απομακρυσμένες τιμές ('outliers') και για αυτόν τον λόγο πολλές φορές χρησιμοποιείται η 'αποκομμένη' διασπορά. Πιο συγκεκριμένα, υπολογίζεται η διασπορά των αποκλίσεων με ένα συγκεκριμένο εύρος τιμών, έτσι ώστε να μην επηρεάζεται από ακραίες τιμές και κατά συνέπεια να αποτελεί μια αξιόπιστη μετρική για το μοντέλο.

Η ίδια ακριβώς τεχνική χρησιμοποιείται στον υπολογισμό των μέσης τιμής των σφαλμάτων, ώστε να μπορούμε να διεξάγουμε σωστά συμπεράσματα για την απόδοση του μοντέλου.

3.5.4 Full Width Half Maximum (*FWHM*)

Η τελευταία μετρική αξιολόγησης του μοντέλου που κατασκευάστηκε αποτελεί το full width half maximum. Ύστερα από την κατασκευή του μοντέλου, κατασκευάζονται τα ιστογράμματα των αποκλίσεων, τα οποία δείχνουν την κατανομή των αποκλίσεων. Παρακάτω φαίνεται η διαδικασία εύρεση του FWHM.



Εικόνα 3.15 : Εύρεση FWHM με την βοήθεια ιστογράμματος

Συνεπώς, το FWHM αποτελεί ένα μέτρο διασποράς των αποκλίσεων χωρίς να λαμβάνει υπόψιν τις ακραίες τιμές.

Κεφάλαιο 4 : Επεξεργασία δεδομένων

4.1 Γραμμικό μοντέλο με μέθοδο Ελαχίστων Τετραγώνων

Με την βοήθεια της προγραμματιστική γλώσσας '*Python'* κατασκευάστηκε ένα μοντέλο γραμμικής παλινδρόμησης με την μέθοδο Ελαχίστων Τετραγώνων.

Πριν την κατασκευή του μοντέλου γραμμικής παλινδρόμησης, το σύνολο δεδομένων διαχωρίστηκε στα σύνολα αξιολόγησης και εκπαίδευσης αντίστοιχα. Το σύνολο αξιολόγησης αποτελείται από το 30% των παρατηρήσεων και το σύνολο εκπαίδευσης από το υπόλοιπο 70%.

Οι συντελεστές του μοντέλου γραμμικού παλινδρόμησης που προέκυψαν παρουσιάζονται παρακάτω:

Πίνακας 4.1: Συντελεστές γραμμικού μοντέλου για κάθε μεταβλητή

Σταθερά	-0.184

Αριθμητικές Μεταβλητές	Συντελεστής
Sentiment Title	-2.025
Sentiment Headline	-0.139
TS1(FB)	-0.001
TS6(FB)	-0.014
TS36(FB)	-0.130
TS72(FB)	-0.003
TS108(FB)	1.148
TS1(G+)	-0.879
TS6(G+)	1.352
TS36(G+)	-1.747
TS72(G+)	1.288
TS108(G+)	-5.835
TS144(G+)	-5.098
TS1(LD)	0.206
TS6(LD)	-0.085
TS36(LD)	0.059
TS72(LD)	0.002
TS108(LD)	-0.837
TS144(LD)	0.785

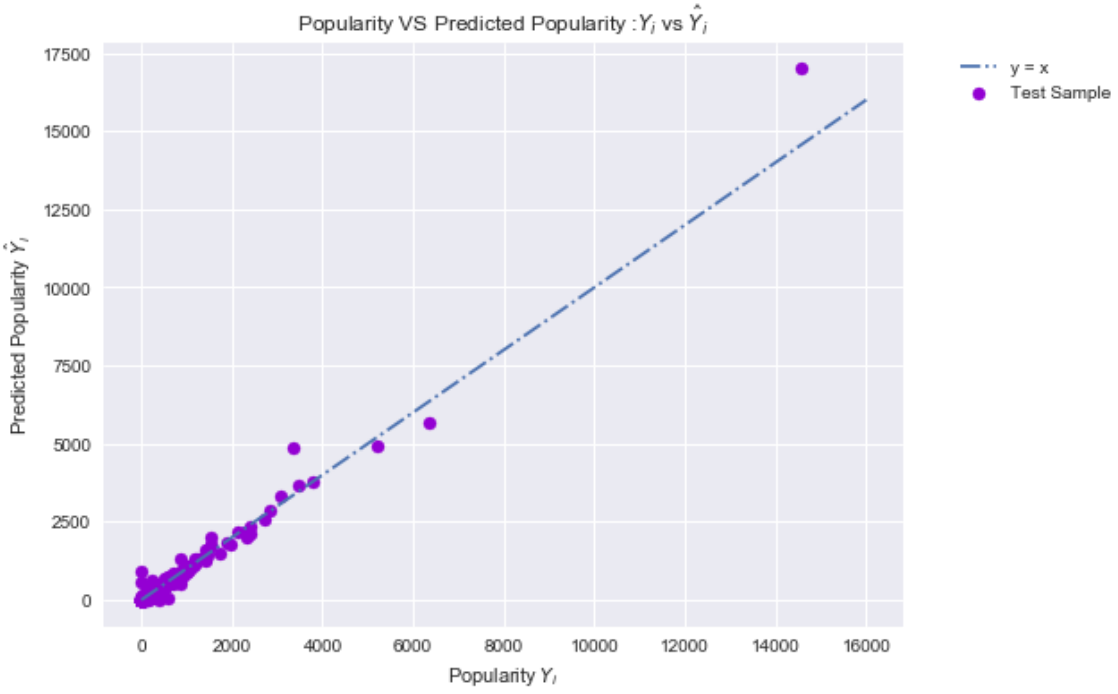
Κατηγορικές Μεταβλητές	Συντελεστής
Is Friday	-0.274
Is Monday	0.036
Is Saturday	3.415
Is Sunday	-0.565
Is Thursday	-0.214

Is Tuesday	-1.368
Is Wednesday	-1.030
Is Afternoon	0.097
Is Evening	-0.576
Is Morning	-0.109
Is Night	0.588

Η ευθεία γραμμικής παλινδρόμησης που κατασκευάστηκε είναι της παρακάτω μορφής :

$$y = -0.184 - 2{,}025x_{ST} - 0.139x_{SH} - 0.001x_{TS1(FB)} + \cdots - 0.109x_{Is\ Morning}$$

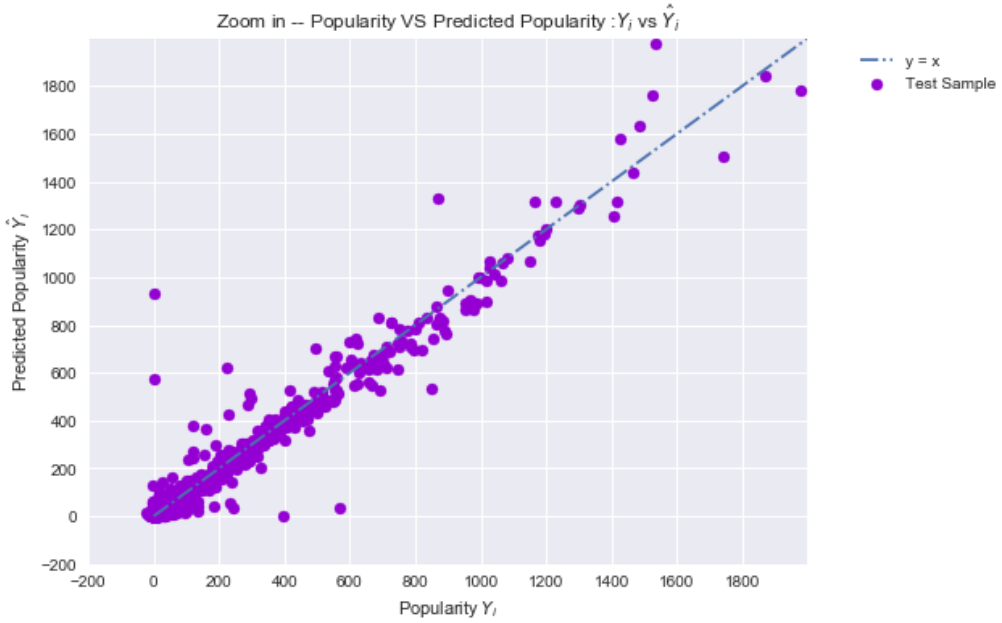
Στο παρακάτω διάγραμμα δίνεται η αναπαράσταση της εκτιμώμενης ευθείας με την μέθοδο Ελαχίστων Τετραγώνων. Στο κατακόρυφο άξονα απεικονίζονται οι προβλεπόμενες τιμές της μεταβλητής απόκρισης του δείγματος αξιολόγησης, δηλαδή του TS144(FB), και στον οριζόντιο άξονα απεικονίζονται οι πραγματικές τιμές της. Το παρακάτω διάγραμμα δείχνει την προσαρμογή του μοντέλου στο σύνολο αξιολόγησης, δηλαδή το μέγεθος των αποκλίσεων μεταξύ των πραγματικών τιμών και των προβλεπόμενων τιμών.



Διάγραμμα 4.1: Αναπαράσταση εκτιμώμενης ευθείας γραμμικής παλινδρόμησης

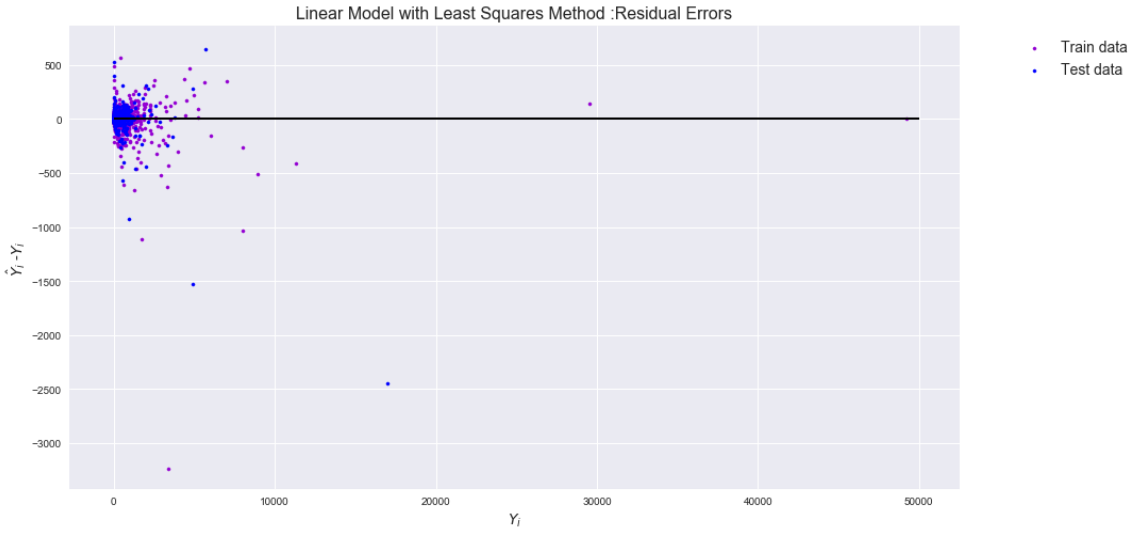
Όμως, η μεταβλητή απόκρισης λαμβάνει ένα μεγάλο εύρος τιμών διότι μετράει των αριθμό των κοινοποιήσεων μιας είδησης με θέμα την οικονομία στο Facebook και για αυτό τον λόγο δεν είναι ευδιάκριτο στο παραπάνω διάγραμμα το μέγεθος των αποκλίσεων από την

ευθεία y = x. Για αυτό, στο ακόλουθο διάγραμμα οι άξονες λαμβάνουν τιμές στο διάστημα [-200,2.000], ώστε να είναι ευδιάκριτο το μέγεθος των αποκλίσεων των παρατηρήσεων από την ευθεία y = x.



Διάγραμμα 4.2: Zoom in διαγράμματος 4.1

Σύμφωνα με το διάγραμμα 4.2, η προσαρμογή της ευθείας στο σύνολο αξιολόγησης φαίνεται να είναι ικανοποιητική, όμως για την εξαγωγή των τελικών συμπερασμάτων θα χρησιμοποιηθούν τα μέτρα απόδοσης που αναφέρθηκαν στο κεφάλαιο 3. Επιπλέον, χρειάζεται να εξεταστεί αν συνέβη υπερεκπαίδευση και για αυτόν τον λόγο είναι χρήσιμο το ακόλουθο διάγραμμα.



Διάγραμμα 4.3: Εκτιμώμενα σφάλματα

Στο παραπάνω διάγραμμα, απεικονίζονται στον y άξονα οι αποκλίσεις των προβλεπόμενων τιμών και των πραγματικών τιμών της μεταβλητής απόκρισης και στον x άξονα οι πραγματικές τιμές. Η ευθεία με το μαύρο χρώμα είναι οι ευθεία $\hat{y}_i - y_i = 0$. Το ιδανικό σενάριο για ένα μοντέλο είναι μην παρουσιάζουν αποκλίσεις από την πραγματική τιμή οι προβλέψεις που γίνονται, όμως αυτό είναι ανέφικτο όταν κατασκευάζονται μοντέλα καθημερινών προβλημάτων, οπότε στόχος είναι μικρότερη δυνατή τιμή που μπορεί να αποκτήσει αυτή η απόκλιση.

Παρατήρηση

Στο διάγραμμα 4.3, τα σύνολα εκπαίδευσης και αξιολόγησης δεν προσαρμόζονται τέλεια, οπότε η σχέση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών δεν είναι ξεκάθαρα γραμμική.

Ο συντελεστής R^2 είναι 0,98, δηλαδή αρκετά κοντά στο 1. Όμως, δεν μπορούμε να βασιστούμε στην συγκεκριμένη τιμή [9]. Ακόμη η τιμή του MSE $_{\text{Test Sample}}$ είναι 2062,8 και η τιμή του MSE $_{\text{Train Sample}}$ είναι 1526,5. Παρατηρείται ότι οι τιμές του MSE για το δείγμα αξιολόγησης και το δείγμα εκπαίδευσης δεν είναι χαμηλές και έχουν μεγάλη απόκλιση μεταξύ τους. Η υψηλή τιμή των MSE φαίνεται από το διάγραμμα 4.3, στο οποίο υπάρχουν τιμές πολύ απομακρυσμένες από την ευθεία $\hat{y}_i - y_i = 0$, με αποτέλεσμα να επιβαρύνουν κατά έναν μεγάλο βαθμό την συγκεκριμένη μετρική.

Λόγω των 'αδυναμιών' των παραπάνω μέτρων απόδοσης , θα χρησιμοποιηθούν σαν μέτρα απόδοσης του μοντέλου η μέση τιμή και η τυπική απόκλιση των ποσοστιαίων σφαλμάτων. Το ποσοστιαίο σφάλμα ορίζεται ως εξής:

$$Error\ \% = \frac{\hat{y}_i - y_i}{y_i}\ , i = 1, \dots, N$$

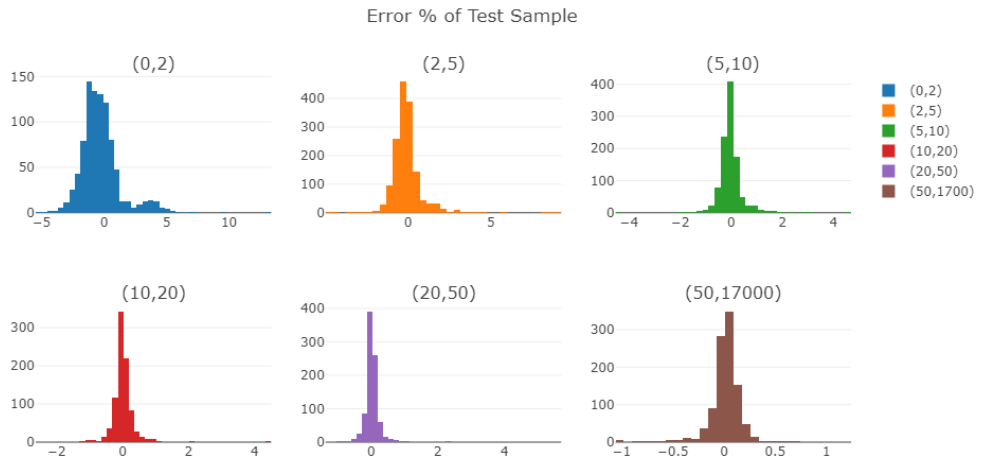
Ο λόγος που χρησιμοποιούνται τα συγκεκριμένα μέτρα απόδοσης θα εξηγηθεί παρακάτω.

Στην συνέχεια, κατασκευάστηκαν τα ιστογράμματα των ποσοστιαίων σφαλμάτων ανάλογα με τις τιμές της προβλεπόμενης τιμής. Πιο συγκεκριμένα, δημιουργήθηκαν τα ιστογράμματα των ποσοστιαίων σφαλμάτων που αντιστοιχούν στα διαστήματα τιμών της πραγματικής τιμής y_i [0,2], [2,5], [5,10], [10,20], [20,50] και [50,17000]. Επιλέχθηκαν τα συγκεκριμένα διαστήματα τιμών έτσι ώστε ο αριθμός παρατηρήσεων που περιέχουν να είναι περίπου ίδιος.

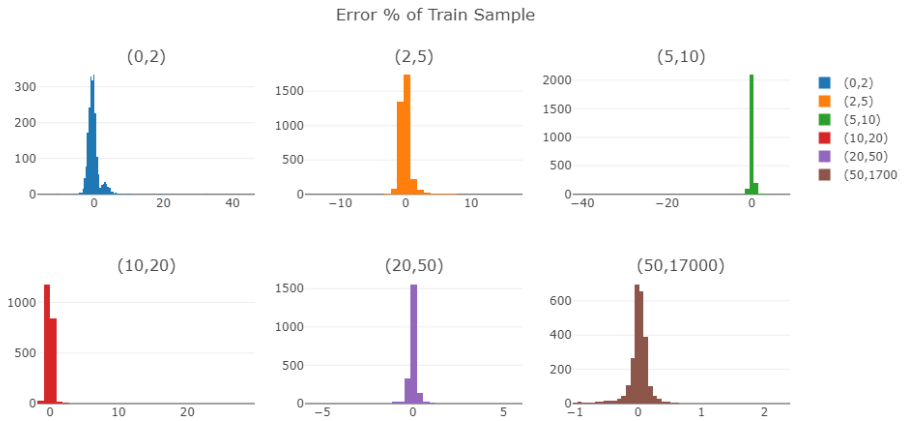
[9] Περισσότερες λεπτομέρειες στο κεφάλαιο 3.5.2

		[0,2]	[2,5]	[5,10]	[10,20]	[20,50]	[50,50.000]
Length of each slice	Test Sample	923	1543	1057	903	884	1041
	Train Sample	2143	3518	2432	2075	2129	2520

Παρατηρείται ότι το πλήθος τιμών της μεταβλητής απόκρισης είναι μεγαλύτερο για χαμηλές τιμές και για αυτό τον λόγο έγινε η επιλογή των συγκεκριμένων διαστημάτων, στα οποία θα γίνει η μελέτη.



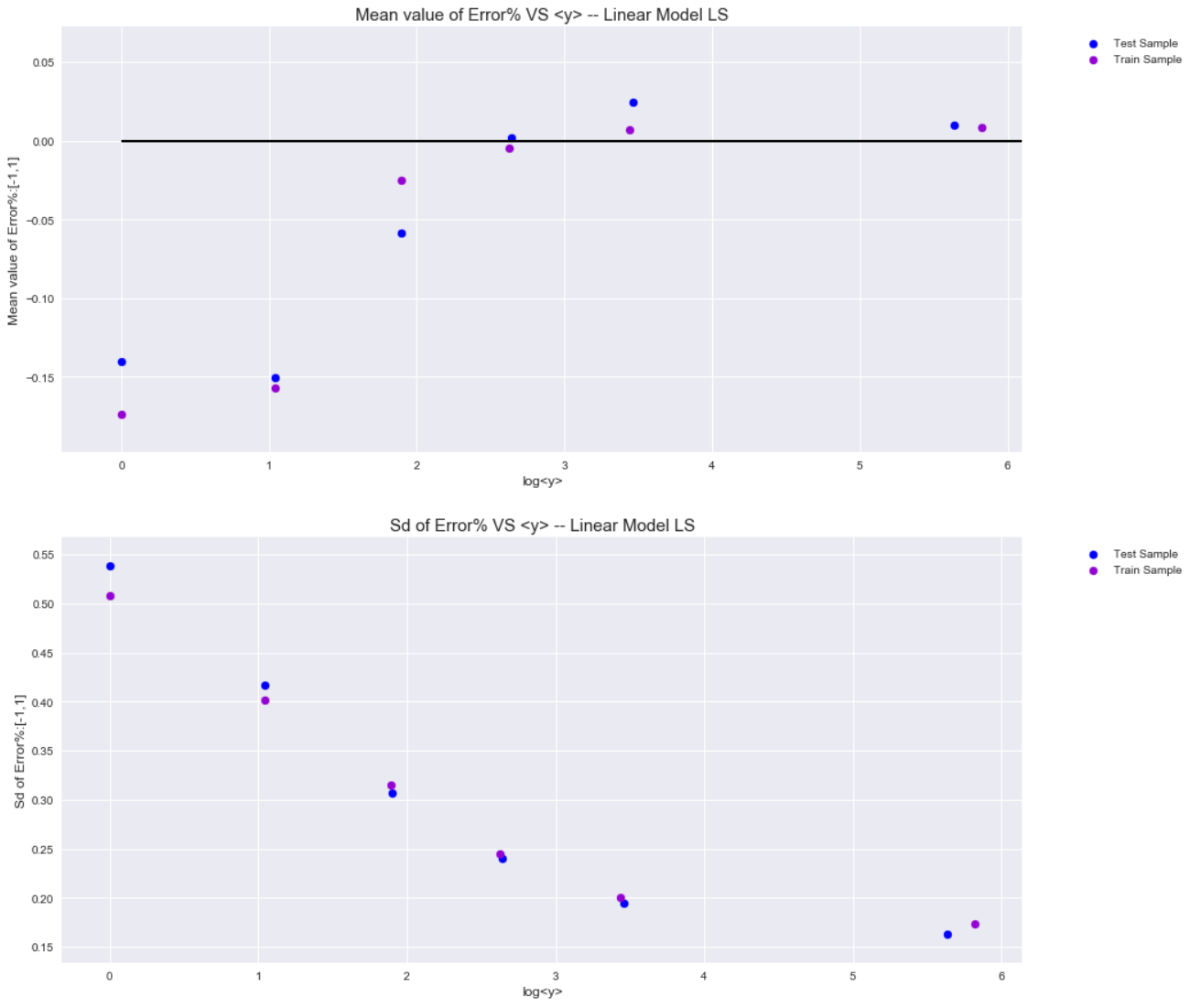
Διάγραμμα 4.4: Ιστογράμματα ποσοστιαίων σφαλμάτων του δείγματος αξιολόγησης



Διάγραμμα 4.5: Ιστογράμματα ποσοστιαίων σφαλμάτων του δείγματος εκπαίδευσης

Σύμφωνα με τα παραπάνω διαγράμματα, παρατηρείται ότι υπάρχουν ακραίες τιμές με μεγάλη απόκλιση από την πραγματική τιμή. Συνεπώς, ένας τρόπος αξιολόγησης απόδοσης του μοντέλου είναι ο περιορισμός των τιμών που λαμβάνουν τα ποσοστιαία σφάλματα και η μελέτη της συμπεριφοράς της μέσω των μέτρων απόδοσης.

Παρακάτω παρουσιάζονται τα διαγράμματα της μέσης τιμής και τυπικής απόκλισης των ποσοστιαίων σφαλμάτων με εύρος τιμών [-1,1] . Στον οριζόντιο άξονα απεικονίζεται ο λογάριθμος της μέσης τιμής της μεταβλητής απόκρισης για κάθε διάστημα και στον κατακόρυφο άξονα η μέση τιμή και η τυπική απόκλιση των ποσοστιαίων σφαλμάτων των δειγμάτων αξιολόγησης και εκπαίδευσης αντίστοιχα.



Διάγραμμα 4.6: Μέση & Τυπική απόκλιση των ποσοστιαίων σφαλμάτων περιορισμένων στο[-1,1]

Σημείωση

Η μέση τιμή των ποσοστιαίων σφαλμάτων χρησιμοποιείται σαν μέτρο απόδοσης του μοντέλου και δείχνει τον μέσο όρο των ποσοστιαίων σφαλμάτων, δηλαδή τον μέσο όρο του σφάλματος που προκύπτει από το μοντέλο για κάθε διάστημα τιμών της μεταβλητής απόκρισης. Δείχνει την ανικανότητα του μοντέλου να αναγνωρίσει την πραγματική σχέση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών. Μπορεί να αποτελέσει ένα μέτρο μεροληψίας του μοντέλου. Μοντέλα με υψηλή μεροληψία 'δίνουν προσοχή' στο σύνολο εκπαίδευσης και απλοποιούν το μοντέλο

Η τυπική απόκλιση των ποσοστιαίων σφαλμάτων δείχνει την απόκλιση των δειγμάτων από την πραγματική τιμή. Μοντέλα με μεγάλη διασπορά 'δίνουν πολύ μεγάλη προσοχή' στο σύνολο εκπαίδευσης, με αποτέλεσμα να μην μπορούν να γενικευθούν για νέα σύνολα και κατά συνέπεια να υπάρχουν υψηλά ποσοστά σφαλμάτων στα σύνολο αξιολόγησης.

Σύμφωνα με τα παραπάνω, η βέλτιστη απόδοση επιτυγχάνεται με χαμηλή μεροληψία και χαμηλή διασπορά και κατά συνέπεια τυπική απόκλιση

Παρατηρήσεις

1. Από το διάγραμμα της μέσης τιμής των ποσοστιαίων σφαλμάτων, γίνεται αντιληπτό ότι στα πρώτα διαστήματα τιμών της y η μέση τιμή των ποσοστιαίων σφαλμάτων έχει απόκλιση από την ευθεία x=0.
2. Από το διάγραμμα της τυπικής απόκλισης των ποσοστιαίων σφαλμάτων, η τυπική απόκλιση του δείγματος αξιολόγησης με την αντίστοιχη του δείγματος εκπαίδευσης δεν έχουν μεγάλη απόκλιση. Στα πρώτα διαστήματα τιμών της y οι τιμές της τυπικής απόκλισης είναι υψηλές. Για παράδειγμα, στο πρώτο διάστημα τιμών η τυπική απόκλιση είναι περίπου 0.5, ενώ όσο αυξάνονται οι τιμές της y η τυπική απόκλιση ελαττώνεται.

Συνεπώς, είναι ένα αρκετά αποδοτικό μοντέλο που έχει περιθώρια βελτίωσης.

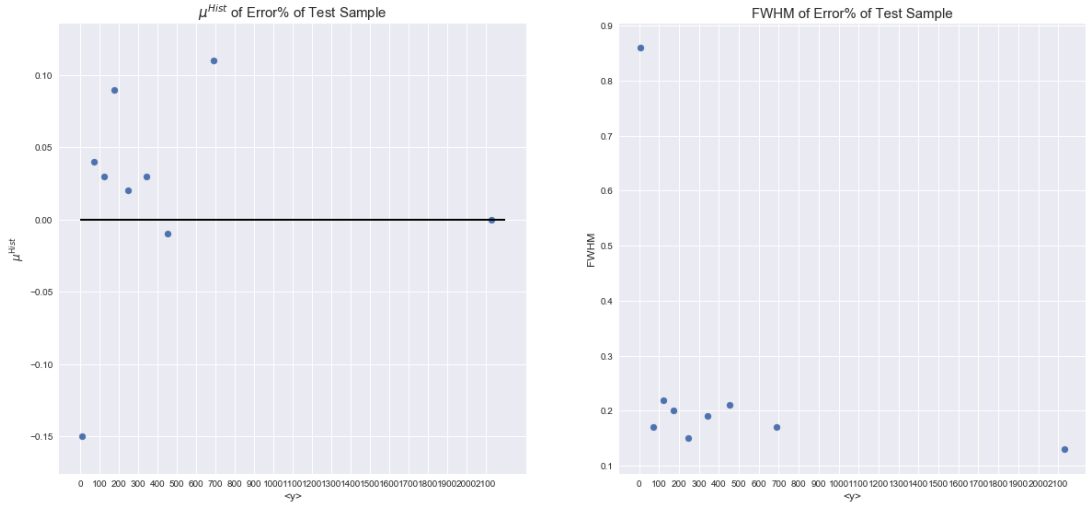
Πρώτη προσέγγιση για την απόδοση του μοντέλου

Παραπάνω παρουσιάστηκε ο τελικός τρόπος προσέγγισης του μοντέλου για την απόδοση του. Η πρώτη προσέγγιση που πραγματοποιήθηκε παρουσιάζεται παρακάτω:

1. Διαχωρισμός δεδομένων σε σύνολο εκπαίδευσης και αξιολόγησης με 50% των παρατηρήσεων το καθένα
2. Κατασκευή γραμμικού μοντέλου με μέθοδο Ελαχίστων Τετραγώνων
3. Εύρεση τιμών ποσοστιαίων σφαλμάτων για κάθε διάστημα τιμών της πραγματικής τιμής της μεταβλητής απόκρισης. Τα διαστήματα που επιλέχθηκαν είναι : [0,50], [50,100], [100,150], [150,200], [200,300], [300,400], [400,500], [500,1000], [1000,12000]. Όμως στην συνέχεια ,διερευνώντας το πρόβλημα, παρατηρήθηκε ότι η μεταβλητή απόκρισης λαμβάνει μικρότερες τιμές και κατά συνέπεια με τον διαχωρισμό των διαστημάτων που επιλέχθηκαν, τα πρώτα διαστήματα περιλαμβάνουν μεγαλύτερο πλήθος τιμών. Για αυτό τον λόγο στην τελική προσέγγιση με την μέθοδο Ελαχίστων Τετραγώνων επιλέχθηκαν διαφορετικά διαστήματα.
4. Αφαίρεση δειγμάτων με μηδενικές τιμές της μεταβλητής απόκρισης
5. Εύρεση FWHM (μέτρο διασποράς) και της μέσης τιμής του ιστογράμματος(μέτρο μέσης τιμής) των ποσοστιαίων σφαλμάτων για κάθε διάστημα.

Ακολούθως παρουσιάζονται τα ιστογράμματα του συνόλου αξιολόγησης για κάθε διάστημα τιμών της πραγματικής τιμής της μεταβλητής απόκρισης και η μέθοδος εύρεσης FWHM σε καθένα από αυτά.

Διάγραμμα 4.7: Full Width Half Maximum



Διάγραμμα 4.8: Μέση τιμή ιστογράμματος & FWHM του δείγματος αξιολόγησης

Παρατηρήσεις

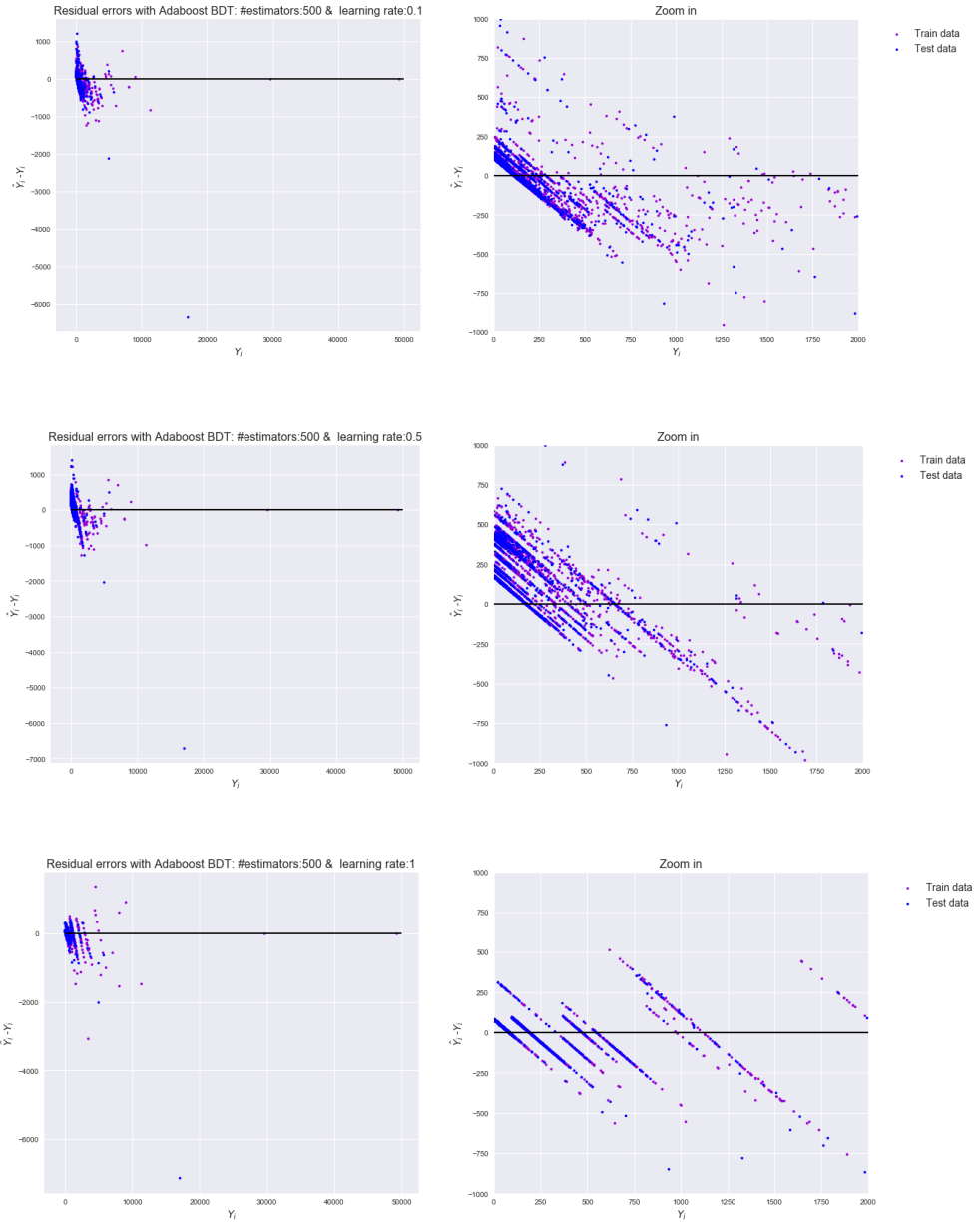
I. Στο διάγραμμα της μέσης τιμής των ποσοστιαίων σφαλμάτων συναρτήσει της πραγματικής τιμής της μεταβλητής απόκρισης, η μέση τιμή κάθε διαστήματος είναι κοντά στο 0, όμως υπάρχουν κάποιες ακραίες τιμές και κατά συνέπεια δεν επιλέχθηκε σαν μέτρο θέσης, διότι επηρεάζεται από αυτές.

II. Το 'Full Width Half Maximum' των ποσοστιαίων σφαλμάτων παρουσιάζει διακυμάνσεις ανάλογα με το διάστημα τιμών της μεταβλητής απόκρισης και κατά συνέπεια δεν επιλέχθηκε σαν μέτρο απόδοσης της διασποράς.

Σύμφωνα με τις παραπάνω παρατηρήσεις, η $2^{η}$ προσέγγιση του προβλήματος είχε ικανοποιητικά αποτελέσματα και για αυτό επιλέχθηκε.

4.2 Μη γραμμικό μοντέλο με Adaboost Boosted Decision Trees

Με την μέθοδο Adaboost, έγινε η κατασκευή ενός μη γραμμικού μοντέλου για διαφορετικές τιμές του ρυθμού εκμάθησης και αριθμό εκτιμητών = 500. Τα αποτελέσματα των εκτιμώμενων σφαλμάτων παρουσιάζονται παρακάτω:



Διάγραμμα 4.9: Μη γραμμικό μοντέλο με Adaboost για διαφορετικές τιμές του ρυθμού εκμάθησης

Παρατηρείται ότι εμφανίζονται περίεργες δομές με την κατασκευή μη γραμμικών μοντέλων με Adaboost. Συνεπώς, τα δεδομένα δεν προσαρμόζονται σωστά με την συγκεκριμένη μέθοδο. Έγινε δοκιμή διαφορετικών τιμών του ρυθμού εκμάθησης και του αριθμού εκτιμητών αλλά η συμπεριφορά των δεδομένων δεν άλλαξε. Συνεπώς, η κατασκευή ενός μη γραμμικού μοντέλου με την βοήθεια ενός δάσους δέντρων απόφασης έγινε με την μέθοδο Gradient Boosting, η οποία όπως θα φανεί στην συνέχεια έδωσε πολύ καλά αποτελέσματα.

4.3 Μη γραμμικό μοντέλο με Gradient Boosted Decision Trees

Σε αυτό το κεφάλαιο, εξετάζεται η απόδοση ενός μη γραμμικού μοντέλου για το συγκεκριμένο σύνολο δεδομένων με την βοήθεια ενός δάσους δέντρων απόφασης με την μέθοδο Gradient Boosting. Η κατασκευή του μη γραμμικού μοντέλου έγινε με την βοήθεια της προγραμματιστικής γλώσσας '*Python*'.

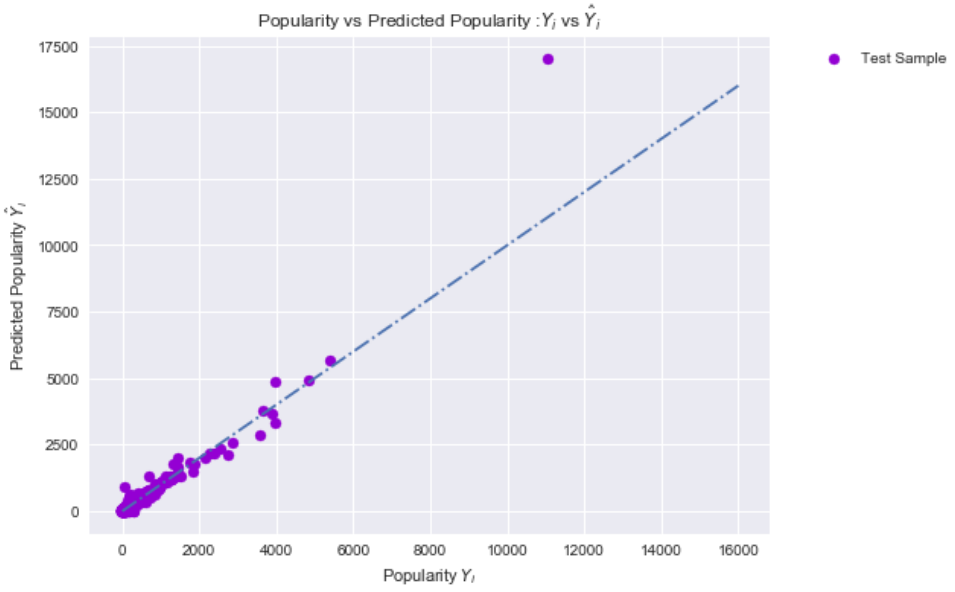
4.3.1 1η προσέγγιση

Αρχικά, οι παράμετροι που επιλέχθηκαν για την κατασκευή του μοντέλου ήταν οι παρακάτω:

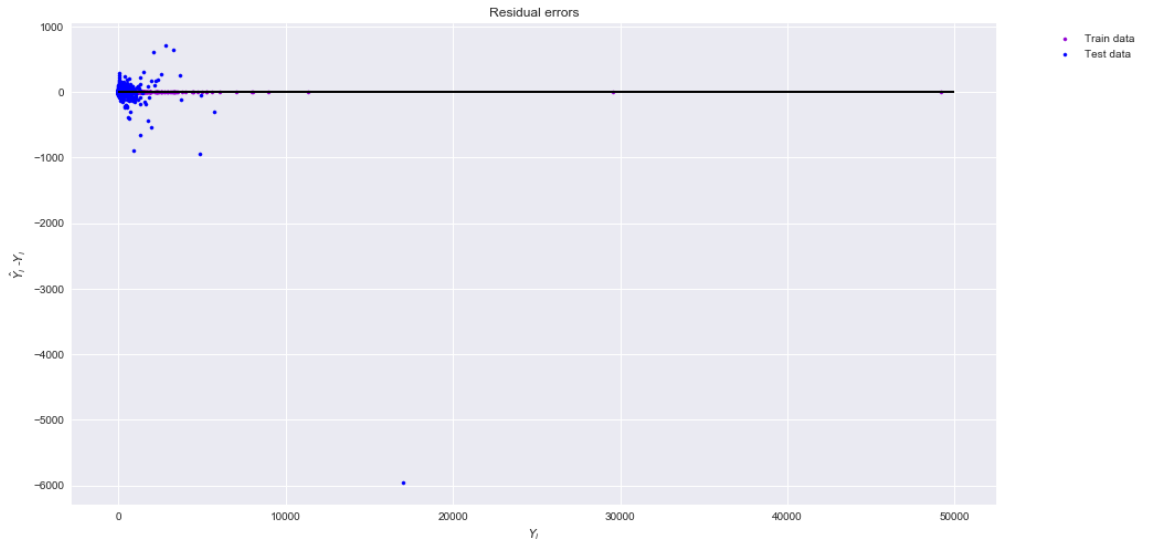
Πίνακας 4.2: Παράμετροι για την δημιουργία μοντέλου με Gradient - Πρώτη προσέγγιση

Βάθος δέντρου απόφασης	3
Ρυθμός εκμάθησης	0.5
Αριθμός δέντρων απόφασης	500
Συνάρτηση κόστους	Συνάρτηση Ελαχίστων Τετραγώνων

Κατασκευάστηκαν τα 2 παρακάτω διαγράμματα, μέσω των οποίων μπορεί να εξεταστεί η προσαρμογή των δειγμάτων αξιολόγησης για το μοντέλο που δημιουργήθηκε.



Διάγραμμα 4.10: Προβλεπόμενη τιμή VS Πραγματική τιμή του συνόλου αξιολόγησης



Διάγραμμα 4.11: Εκτιμώμενα Σφάλματα με Gradient Boosting

Από τα παραπάνω διαγράμματα γίνεται αντιληπτό ότι έχει συμβεί υπερεκπαίδευση. Παρατηρείται ότι για τα δείγματα εκπαίδευσης (*μωβ χρώμα*) η προβλεπόμενη τιμή της μεταβλητής απόκρισης που προέκυψε έχει πολύ μικρή απόκλιση από την πραγματική τιμή, ενώ για τα δείγματα αξιολόγησης (*μπλε χρώμα*) οι αποκλίσεις είναι πολύ μεγαλύτερες. Συνεπώς, οι τιμές των παραμέτρων που επιλέχθηκαν πρέπει να τροποποιηθούν και να βρεθεί ένας συνδυασμός αυτών των παραμέτρων που να δίνει ένα αποδοτικό μοντέλο, χωρίς παράλληλα να συμβαίνει υπερεκπαίδευση.

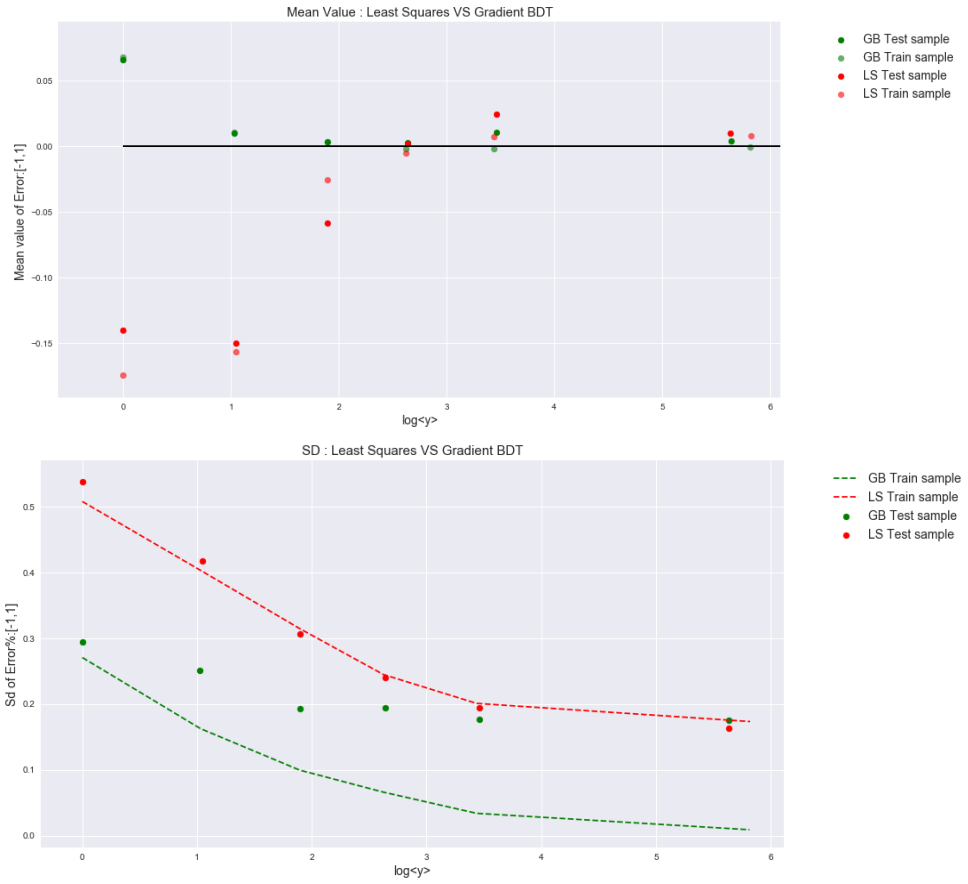
Ένας τρόπος επαλήθευσης του παραπάνω συμπεράσματος είναι η σύγκριση των τιμών του μέσου τετραγωνικού σφάλματος για τα δείγματα αξιολόγησης και εκπαίδευσης. Είναι γνωστό ότι το συγκεκριμένο μέτρο απόδοσης επηρεάζεται από ακραίες τιμές, όμως μπορεί να δώσει ικανοποιητικά αποτελέσματα όταν συμβαίνει υπερεκπαίδευση. Οι τιμές που προέκυψαν είναι :

$$MSE_{Train} = 0.46 \qquad MSE_{Test} = 6668.31$$

Η απόκλιση μεταξύ των δύο τιμών είναι πολύ μεγάλη, οπότε επαληθεύεται το γεγονός ότι συνέβη υπερεκπαίδευση. Πιθανός λόγος υπερεκπαίδευσης είναι ο υψηλός αριθμός δέντρων απόφασης σε συνδυασμό με την σχετικά ‘μέτρια’ τιμή του ρυθμού εκμάθησης. Πιο συγκεκριμένα, μεγάλη τιμή στον ρυθμό εκμάθησης έχει ως αποτέλεσμα ένα δέντρο απόφασης να ‘μαθαίνει’ γρήγορα από το προηγούμενο δέντρο. Ο ρυθμός εκμάθησης δείχνει την ποσότητα ‘αντίληψης’ του λάθους στα επόμενα διαδοχικά δέντρα. Συνεπώς, μια σχετικά μεγάλη τιμή του ρυθμού εκμάθησης σε συνδυασμό με μια μεγάλη τιμή

αριθμού δέντρων απόφασης έχει σαν αποτέλεσμα το μοντέλο να ‘εκπαιδεύεται’ σχετικά γρήγορα οπότε να μην χρειάζεται επιπλέον δέντρα απόφασης.

Τέλος, κατασκευάστηκαν τα ακόλουθα διαγράμματα για να συγκριθεί η απόδοση του γραμμικού με την αντίστοιχη του μη γραμμικού μοντέλου.



Διάγραμμα 4.12: Μέση τιμή & τυπική απόκλιση ποσοστιαίων σφαλμάτων με Gradient (1η προσέγγιση)

Όπως και στο γραμμικό μοντέλο, υπολογίστηκαν για τα ίδια διαστήματα τιμών της πραγματικής τιμής της μεταβλητής απόκρισης οι μέσες τιμές και οι τυπικές αποκλίσεις των ποσοστιαίων σφαλμάτων των δειγμάτων αξιολόγησης και εκπαίδευσης με εύρος τιμών [-1,1]. Στα παραπάνω διαγράμματα, με κόκκινο χρώμα απεικονίζεται το μοντέλο γραμμικής παλινδρόμησης με την μέθοδο Ελαχίστων Τετραγώνων (*LS*) και με πράσινο χρώμα το μοντέλο μη γραμμικής παλινδρόμησης με την μέθοδο Gradient Boosting (*GB*).

Όπως φαίνεται στο 2ο διάγραμμα, η τυπική απόκλιση των δειγμάτων αξιολόγησης παρουσιάζει μεγάλη απόκλιση από την αντίστοιχη των δειγμάτων εκπαίδευσης στο μη γραμμικό μοντέλο, ενώ δεν συμβαίνει το ίδιο στο γραμμικό μοντέλο. Άρα, έχει συμβεί υπερεκπαίδευση και κατά συνέπεια χρειάζεται να γίνουν αλλαγές στις τιμές των παραμέτρων.

4.3.2 2η προσέγγιση : Υπερεκπαίδευση

Σύμφωνα με τα συμπεράσματα που προέκυψαν από την 1η προσέγγιση του προβλήματος με την μέθοδο Gradient Boosting, στόχος είναι η εύρεση των 'βέλτιστων' τιμών των 3 βασικών παραμέτρων που επηρεάζουν την απόδοση του μοντέλου. Οι συγκεκριμένες παράμετροι, δηλαδή το βάθος κάθε δέντρου απόφασης, ο ρυθμός εκμάθησης και ο αριθμός των εκτιμητών, δηλαδή των δέντρων απόφασης, επηρεάζονται μεταξύ τους. Για παράδειγμα, μεγάλο βάθος μπορεί να έχει ως αποτέλεσμα το μοντέλο που θα κατασκευαστεί να προσαρμόζεται πολύ καλά στα δεδομένα, οπότε να χρειάζεται μεγάλος αριθμός ρυθμού εκμάθησης και μικρός αριθμός εκτιμητών έτσι ώστε να αποφευχθεί η υπερεκπαίδευση. Γενικώς, προτιμάται μικρός βάθος για κάθε δέντρο ώστε ο συνδυασμός αυτών των 'αδύναμων' δέντρων με κατάλληλες τιμές ρυθμού εκμάθησης και αριθμού εκτιμητών να επιφέρει ένα αποδοτικό μοντέλο.

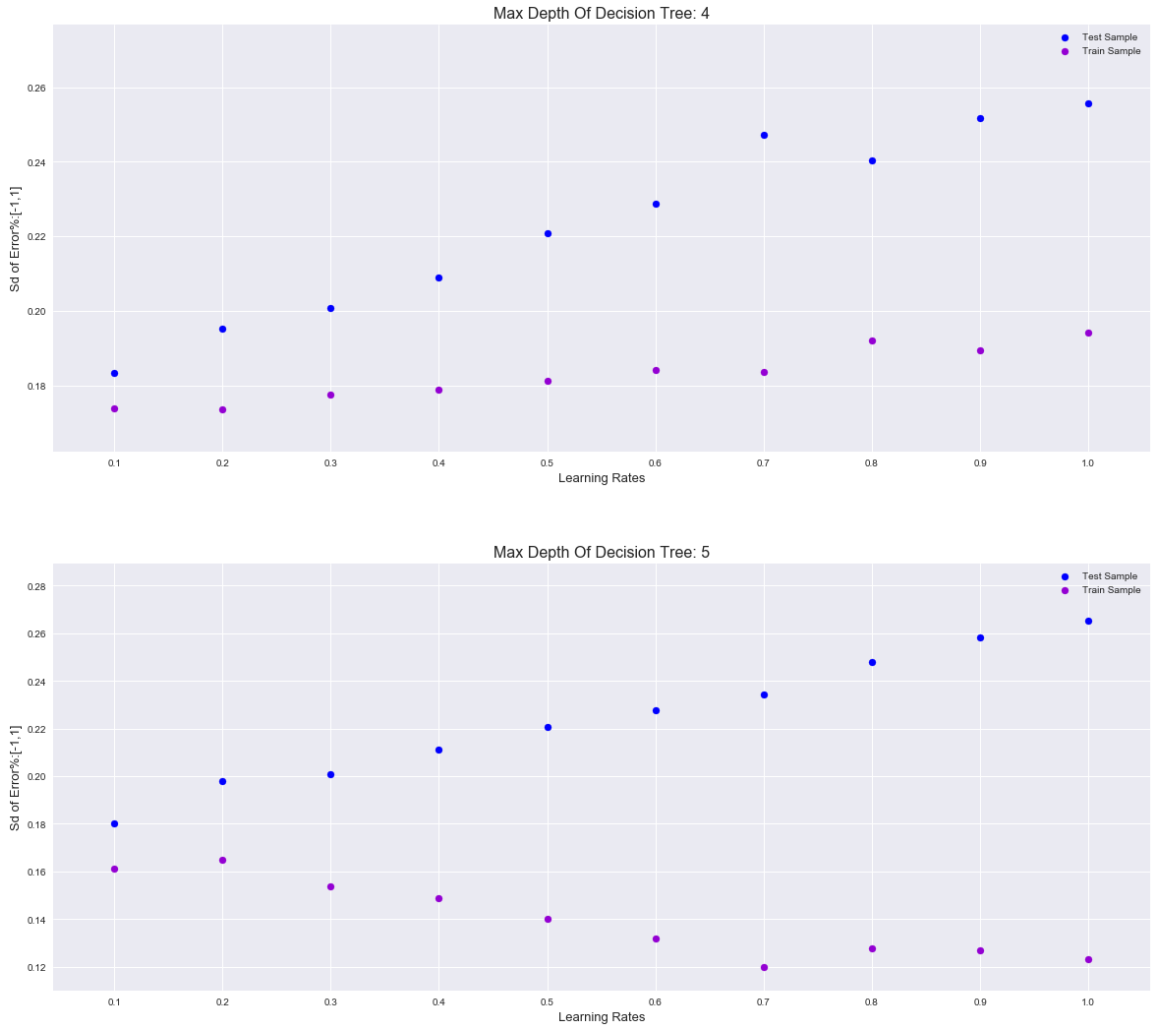
Σχέση εξάρτησης του ρυθμού εκμάθησης με το βάθος δέντρου απόφασης

Σταθεροποιώντας τον αριθμό των εκτιμητών, δηλαδή 500, στόχος είναι η εύρεση του κατάλληλου συνδυασμού των τιμών του βάθους και του ρυθμού εκμάθησης που δίνει ικανοποιητικά αποτελέσματα. Οι τιμές για κάθε παράμετρο παρουσιάζονται ακολούθως:

Πίνακας 4.3:Τιμές του ρυθμού εκμάθησης και του βάθους δέντρων για αριθμό εκτιμητών : 500

Ρυθμός εκμάθησης	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
Βάθος δέντρου απόφασης	[3, 4, 5]

Στα παρακάτω διαγράμματα απεικονίζονται οι τυπικές αποκλίσεις των ποσοστιαίων σφαλμάτων με εύρος τιμών [-1,1] των δειγμάτων εκπαίδευσης και αξιολόγησης για τις διαφορετικές τιμές του ρυθμού εκμάθησης.



Διάγραμμα 4.13: Τυπική απόκλιση ποσοστιαίων σφαλμάτων συναρτήσει του ρυθμού εκμάθησης για βάθη δέντρων 4,5,6

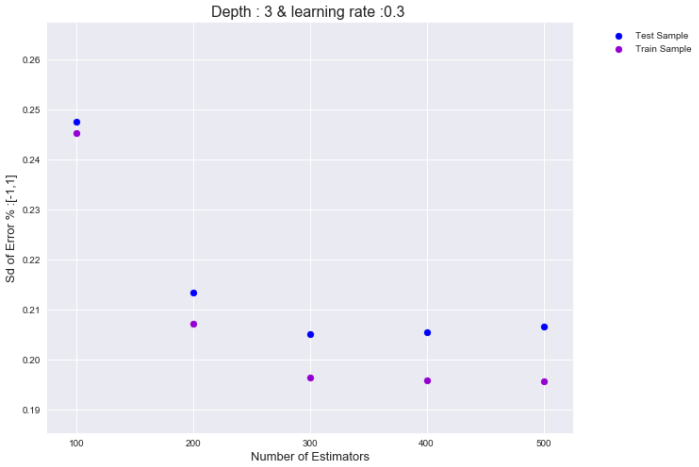
Παρατηρείται ότι όσο αυξάνεται η τιμή του ρυθμού εκμάθησης τόσο μεγαλώνει η απόκλιση της τυπικής απόκλισης του δείγματος αξιολόγησης με το δείγμα εκπαίδευσης. Συνεπώς, αυτό αποτελεί ένδειξη υπερεκπαίδευσης, διότι η τυπική απόκλιση του δείγματος αξιολόγησης αυξάνεται ενώ η αντίστοιχη του δείγματος εκπαίδευσης μειώνεται.

Ο συνδυασμός τιμών που επιλέχθηκαν για το βάθος και τον ρυθμό εκμάθησης είναι:

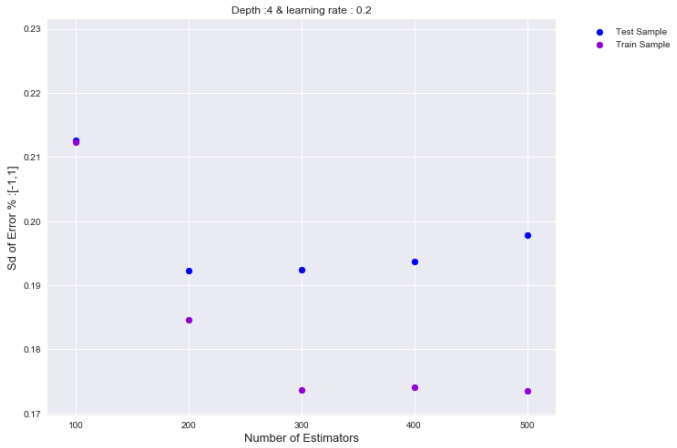
Πίνακας 4.4: Επιλογή παραμέτρων για 500 εκτιμητές (GB)

Βάθος	**Ρυθμός εκμάθησης**
3	0.3
4	0.2
5	0.2

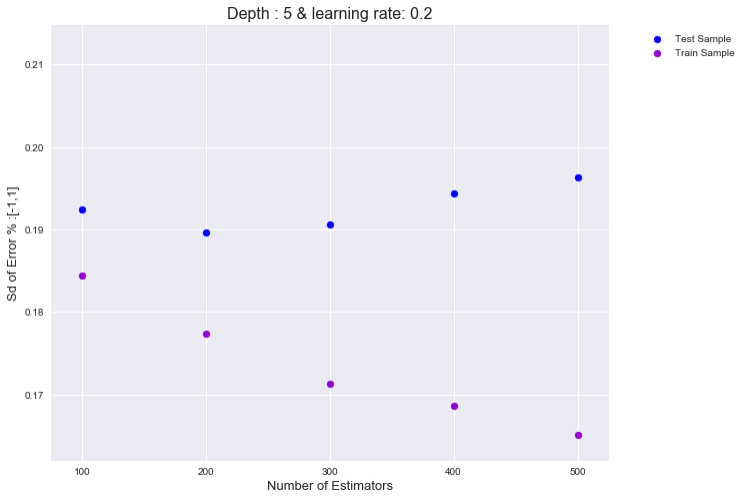
Στην συνέχεια, για τον παραπάνω συνδυασμό τιμών των δύο παραμέτρων μεταβάλλεται ο αριθμός εκτιμητών στο διάστημα [100,500] με βήμα 100.



Διάγραμμα 4.14: Τυπική απόκλιση ποσοστιαίων σφαλμάτων συναρτήσει του αριθμού εκτιμητών για βάθος 3



Διάγραμμα 4.15: Τυπική απόκλιση ποσοστιαίων σφαλμάτων συναρτήσει του αριθμού εκτιμητών για βάθος 4



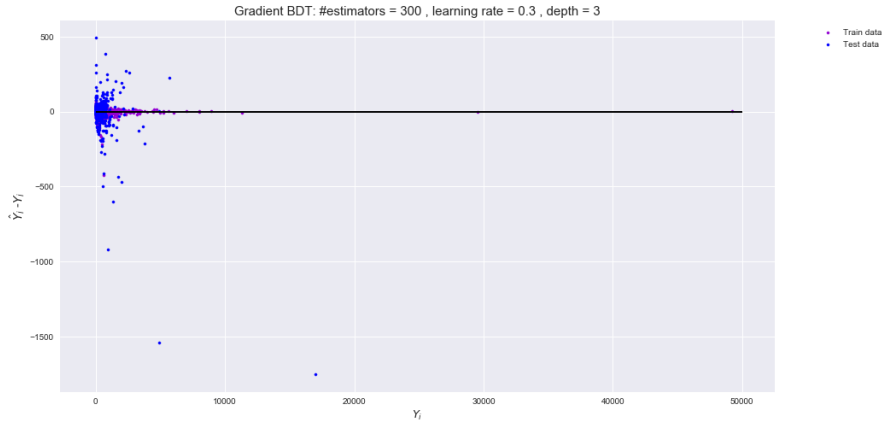
Διάγραμμα 4.16: Τυπική απόκλιση ποσοστιαίων σφαλμάτων συναρτήσει του αριθμού εκτιμητών για βάθος 5

Η τελική επιλογή της 2ης προσέγγισης για τις τιμές των παραμέτρων είναι:

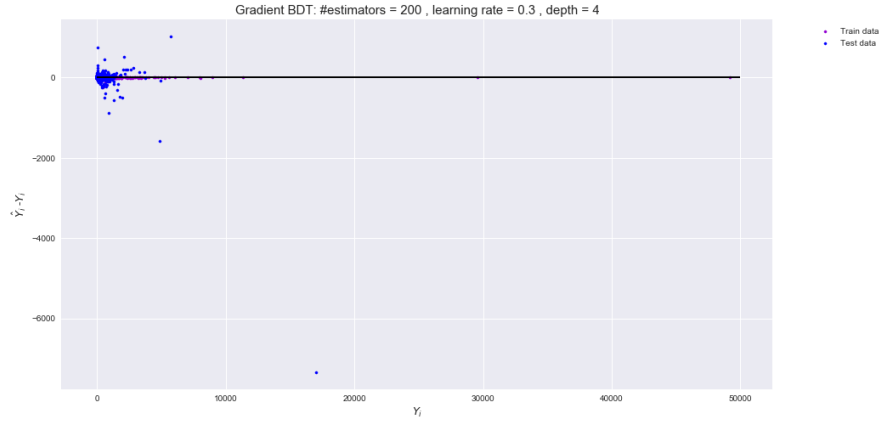
Πίνακας 4.5: Τελική επιλογή παραμέτρων 2ης προσέγγισης

Βάθος	**Ρυθμός εκμάθησης**	**Αριθμός εκτιμητών**
3	0.3	300
4	0.2	200
5	0.2	200

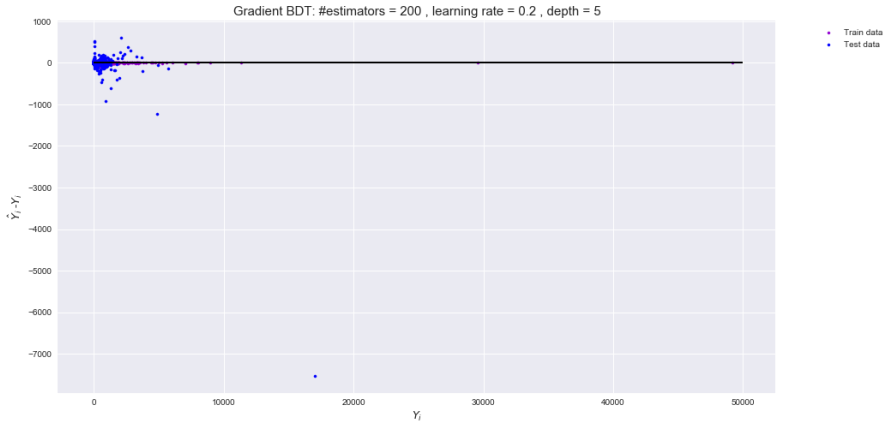
Για τις παραπάνω τιμές των παραμέτρων κατασκευάστηκαν τα αντίστοιχα μοντέλα. Το 1ο διάγραμμα σε κάθε συνδυασμό παραμέτρων απεικονίζει την απόκλιση των δειγμάτων αξιολόγησης και εκπαίδευσης συναρτήσει της πραγματικής τιμής της μεταβλητής απόκρισης και το 2ο απεικονίζει την συνάρτηση κόστους για το σύνολο εκπαίδευσης και αξιολόγησης συναρτήσει των επαναλήψεων, δηλαδή τον αριθμό των εκτιμητών.



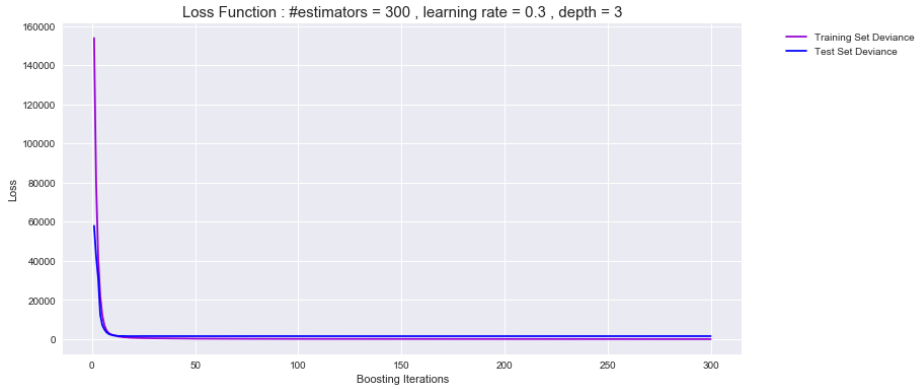
Διάγραμμα 4.17: Εκτιμώμενα σφάλματα για βάθος δέντρων : 3 (2η προσέγγιση)



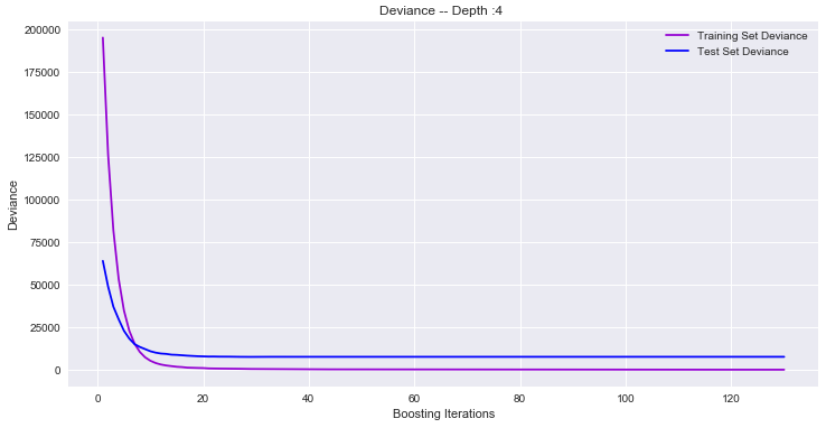
Διάγραμμα 4.18:Εκτιμώμενα σφάλματα για βάθος δέντρων : 4 (2η προσέγγιση)



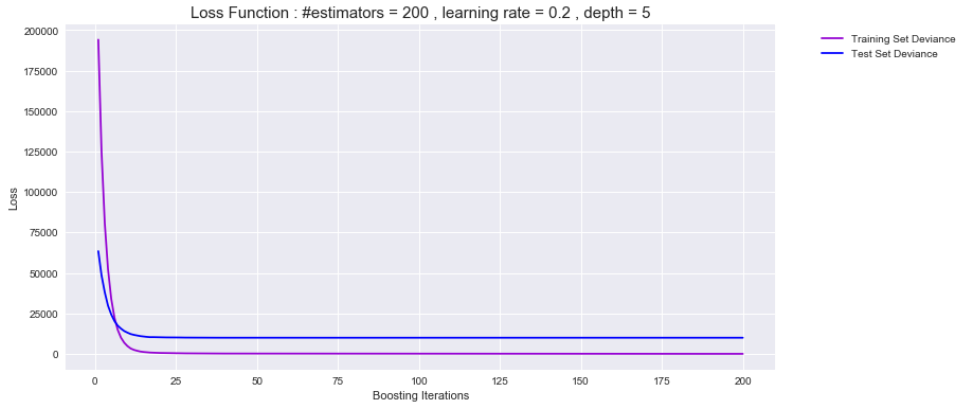
Διάγραμμα 4.19: Εκτιμώμενα σφάλματα για βάθος δέντρων : 5 (2η προσέγγιση)



Διάγραμμα 4.20: Συνάρτηση κόστους συναρτήσει του αριθμού των εκτιμητών για βάθος δέντρων 3 ($2^η$ προσέγγιση)



Διάγραμμα 4.21: Συνάρτηση κόστους συναρτήσει του αριθμού των εκτιμητών για βάθος δέντρων 4 ($2^η$ προσέγγιση)



Διάγραμμα 4.22: Συνάρτηση κόστους συναρτήσει του αριθμού των εκτιμητών για βάθος δέντρων 5 ($2^η$ προσέγγιση)

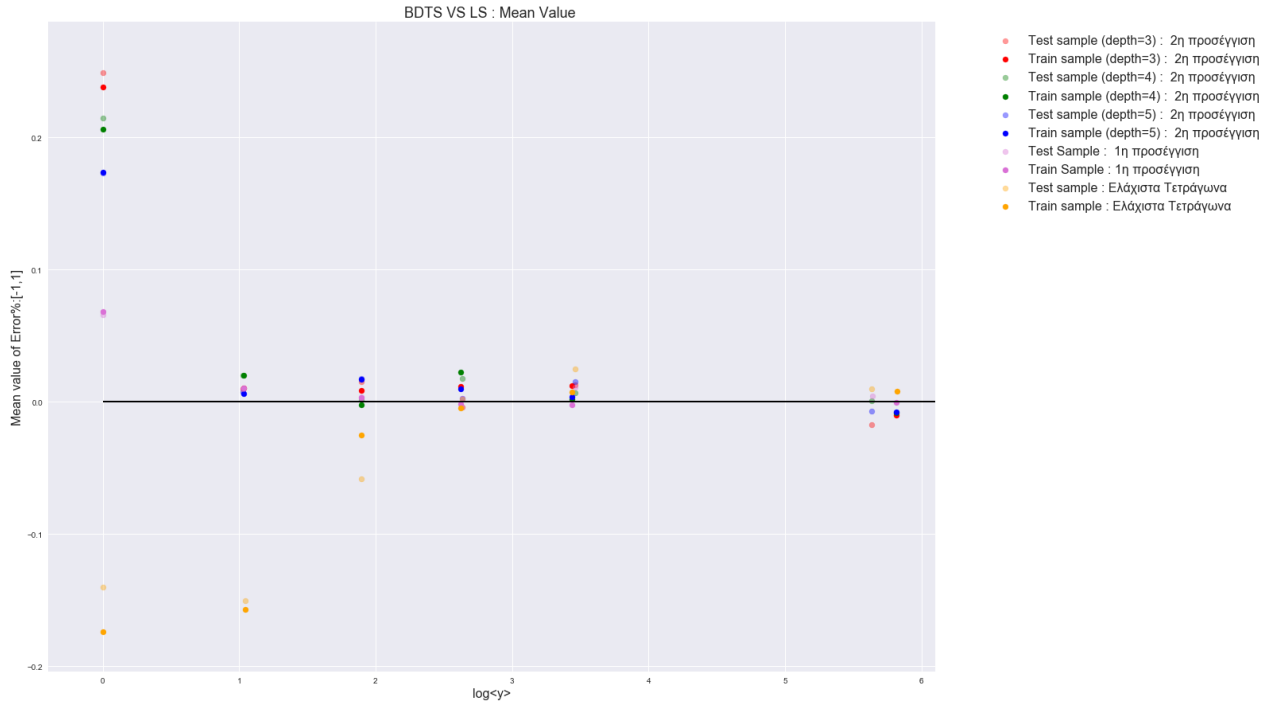
Παρατηρήσεις:

I. Σύμφωνα με τα $1^α$ διαγράμματα για κάθε συνδυασμό τιμών, τα δείγματα αξιολόγησης τείνουν να έχουν μεγαλύτερες αποκλίσεις από αυτές των δειγμάτων εκπαίδευσης. Συνεπώς, ο συνδυασμός τιμών που επιλέχθηκε δείχνει ότι συνέβη υπερεκπαίδευση.

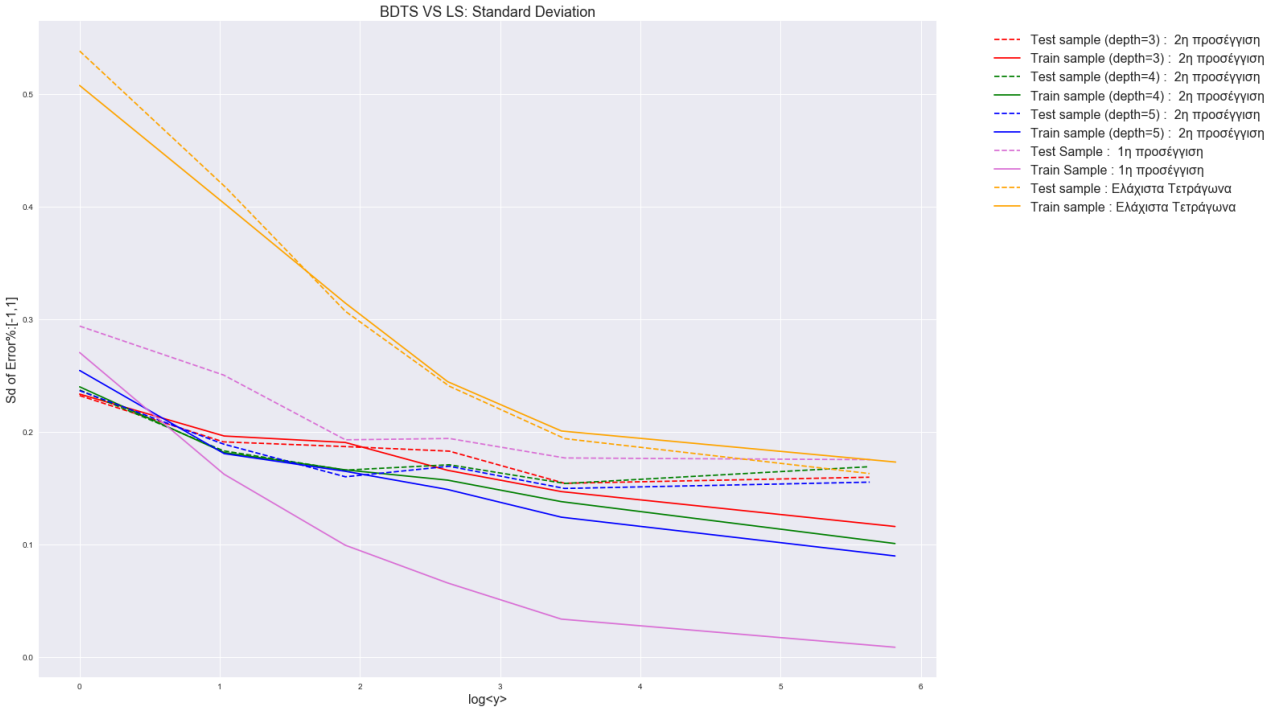
II. Σύμφωνα με τα $2^α$ διαγράμματα για κάθε συνδυασμό τιμών, έπειτα από κάποιο αριθμό επαναλήψεων, δηλαδή αριθμό δέντρων, οι τιμές των συναρτήσεων κόστους των συνόλων εκπαίδευσης και αξιολόγησης διαφέρουν κατά μια σταθερή τιμή.

Συνδυάζοντας τις παρατηρήσεις I και II, συμπεραίνεται ότι χρειάζεται μικρότερος αριθμός δέντρων για επιλεχθούν οι τιμές για τις άλλες δύο παραμέτρους.

Σύγκριση μοντέλων 2ης προσέγγισης με το μοντέλο γραμμική παλινδρόμησης



Διάγραμμα 4.23: Σύγκριση μοντέλων 2ης προσέγγισης με το γραμμικό μοντέλο (Μέση τιμή)



Διάγραμμα 4.24: Σύγκριση μοντέλων 2ης προσέγγισης με το γραμμικό μοντέλο (Τυπική απόκλιση)

Τα παραπάνω δύο διαγράμματα απεικονίζουν την μέση τιμή και την τυπική απόκλιση των ποσοστιαίων σφαλμάτων στο [-1,1] για τα σύνολα αξιολόγησης και εκπαίδευσης αντίστοιχα.

Παρατηρήσεις:

I. Η τυπική απόκλιση του γραμμικού μοντέλου για τα δείγματα αξιολόγησης και εκπαίδευσης λαμβάνει μεγαλύτερες τιμές από την αντίστοιχη των μη γραμμικών μοντέλων. Όμως, τα μη γραμμικά μοντέλα παρουσιάζουν μεγαλύτερη απόκλιση μεταξύ της τυπικής απόκλισης των δειγμάτων αξιολόγησης και εκπαίδευσης (*επιβεβαιώνεται η ένδειξη υπερεκπαίδευσης*).

II. Η μεγαλύτερη απόκλιση στην τιμή της τυπικής απόκλισης του δείγματος αξιολόγησης με την αντίστοιχη του δείγματος εκπαίδευσης παρουσιάζεται στο μη γραμμικό μοντέλο $1^{ης}$ προσέγγισης.

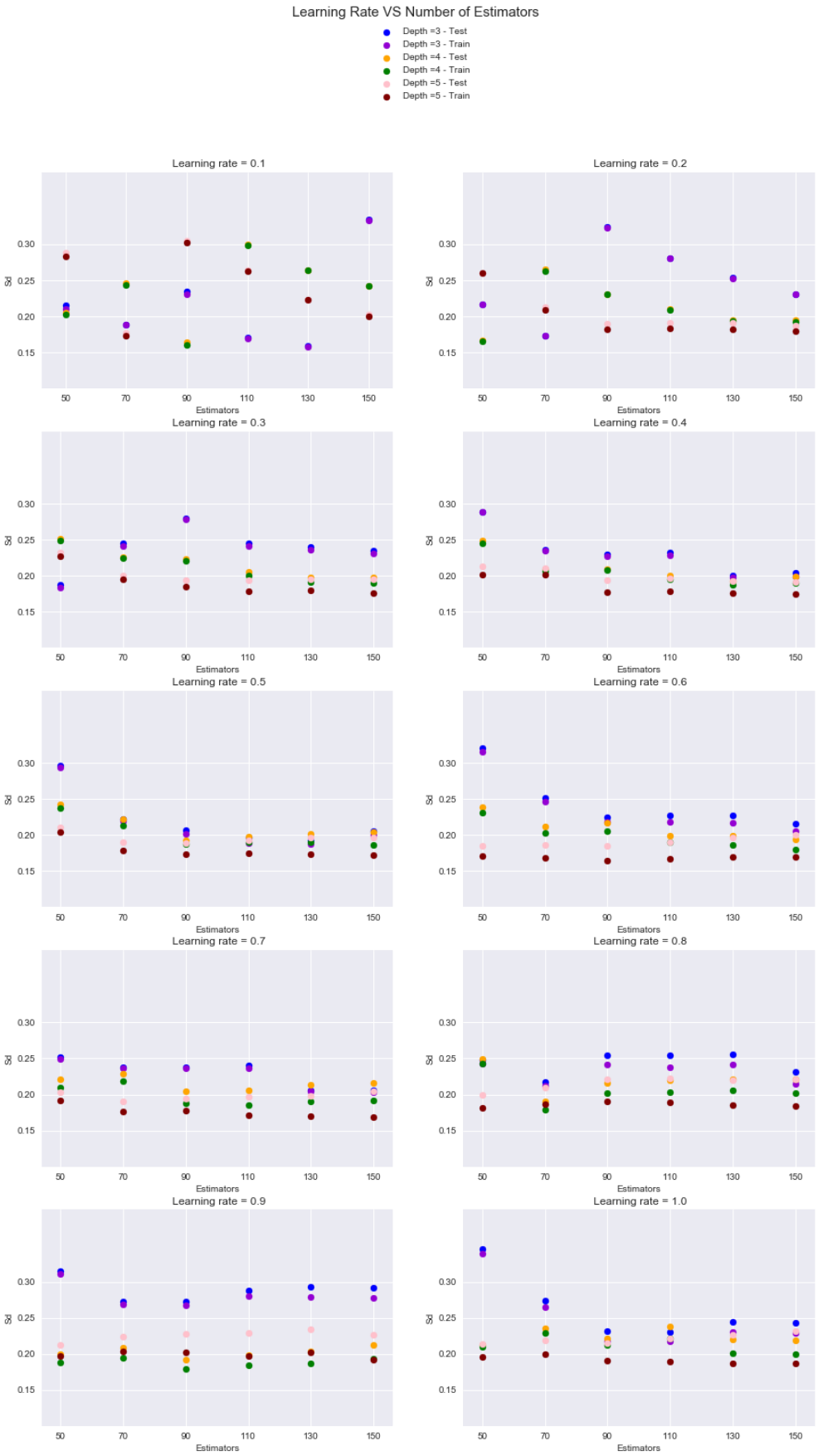
4.3.3 $3^{η}$ προσέγγιση: Τελική

Αρχικά στην τελική προσέγγιση του προβλήματος, μεταβλήθηκαν ταυτόχρονα οι τιμές των 3 βασικών παραμέτρων. Παρακάτω φαίνεται το εύρος τιμών της καθεμίας.

Πίνακας 4.6: Τιμές παραμέτρων βάθους, αριθμού εκμάθησης και αριθμού εκτιμητών

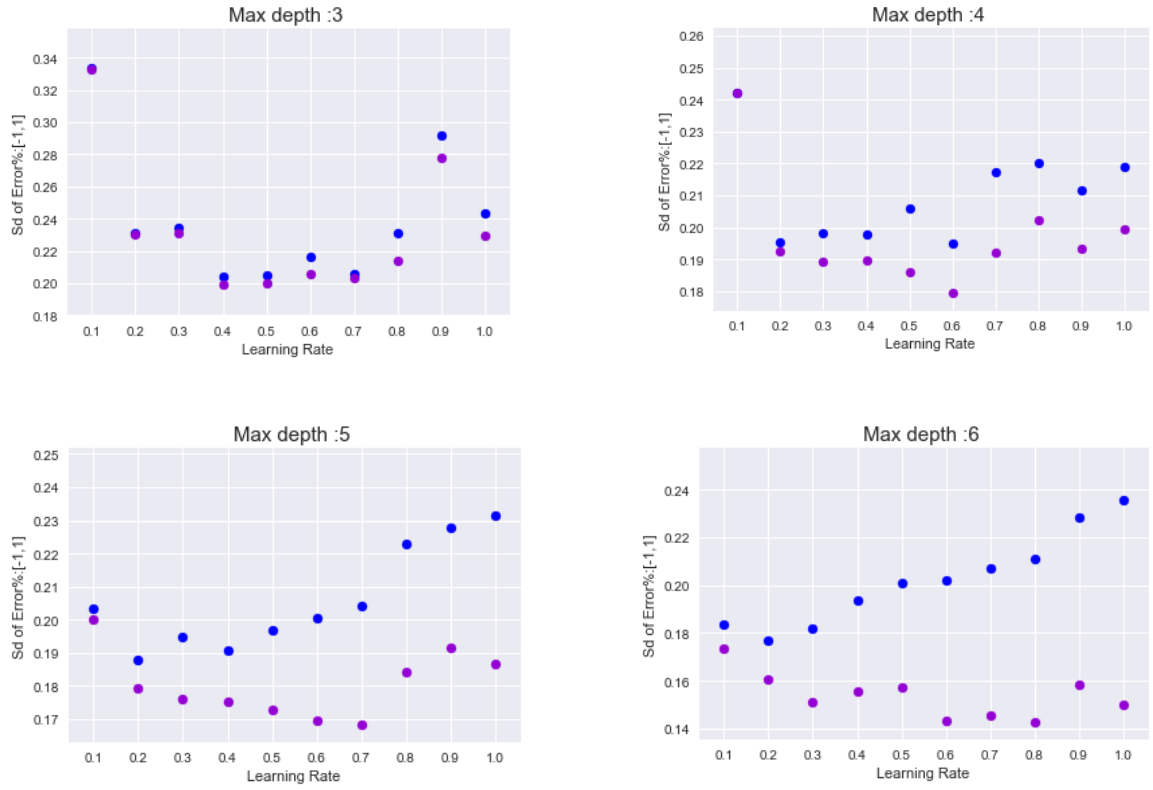
Ρυθμός εκμάθησης	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
Βάθος δέντρου απόφασης	[3, 4, 5]
Αριθμός εκτιμητών	[50, 70, 90, 110, 130, 150]

Στα παρακάτω διαγράμματα αναπαρίσταται η τυπική απόκλιση των ποσοστιαίων σφαλμάτων με εύρος τιμών [-1.1] συναρτήσει του αριθμού των εκτιμητών για κάθε τιμή του ρυθμού εκμάθησης. Σε καθένα από αυτά, απεικονίζεται η τυπική απόκλιση του δείγματος αξιολόγησης και εκπαίδευσης για κάθε βάθος.


Learning Rate VS Number of Estimators
Depth =3 - Test
Depth =3 - Train
Depth =4 - Test
Depth =4 - Train
Depth =5 - Test
Depth =5 - Train
Learning rate = 0.1
Learning rate = 0.2
Learning rate = 0.3
Learning rate = 0.4
Learning rate = 0.5
Learning rate = 0.6
Learning rate = 0.7
Learning rate = 0.8
Learning rate = 0.9
Learning rate = 1.0
Sd
Estimators
0.15
0.20
0.25
0.30
50
70
90
110
130
150

Όμως, η εύρεση βέλτιστων τιμών για τις 3 παραμέτρους, όπως φαίνεται στα παραπάνω διαγράμματα, είναι δύσκολο να πραγματοποιηθεί ταυτόχρονα. Συνεπώς, σταθεροποιώντας μία τιμή από αυτές (*2η προσέγγιση*) και μεταβάλλοντας τις άλλες δύο προκύπτουν αποτελέσματα ευκολότερα προς κατανόησή.

Σταθεροποιώντας τον αριθμό εκτιμητών ίσο με 150 και μεταβάλλοντας τον ρυθμό εκμάθησης και το βάθος δέντρων προέκυψαν τα παρακάτω διαγράμματα:



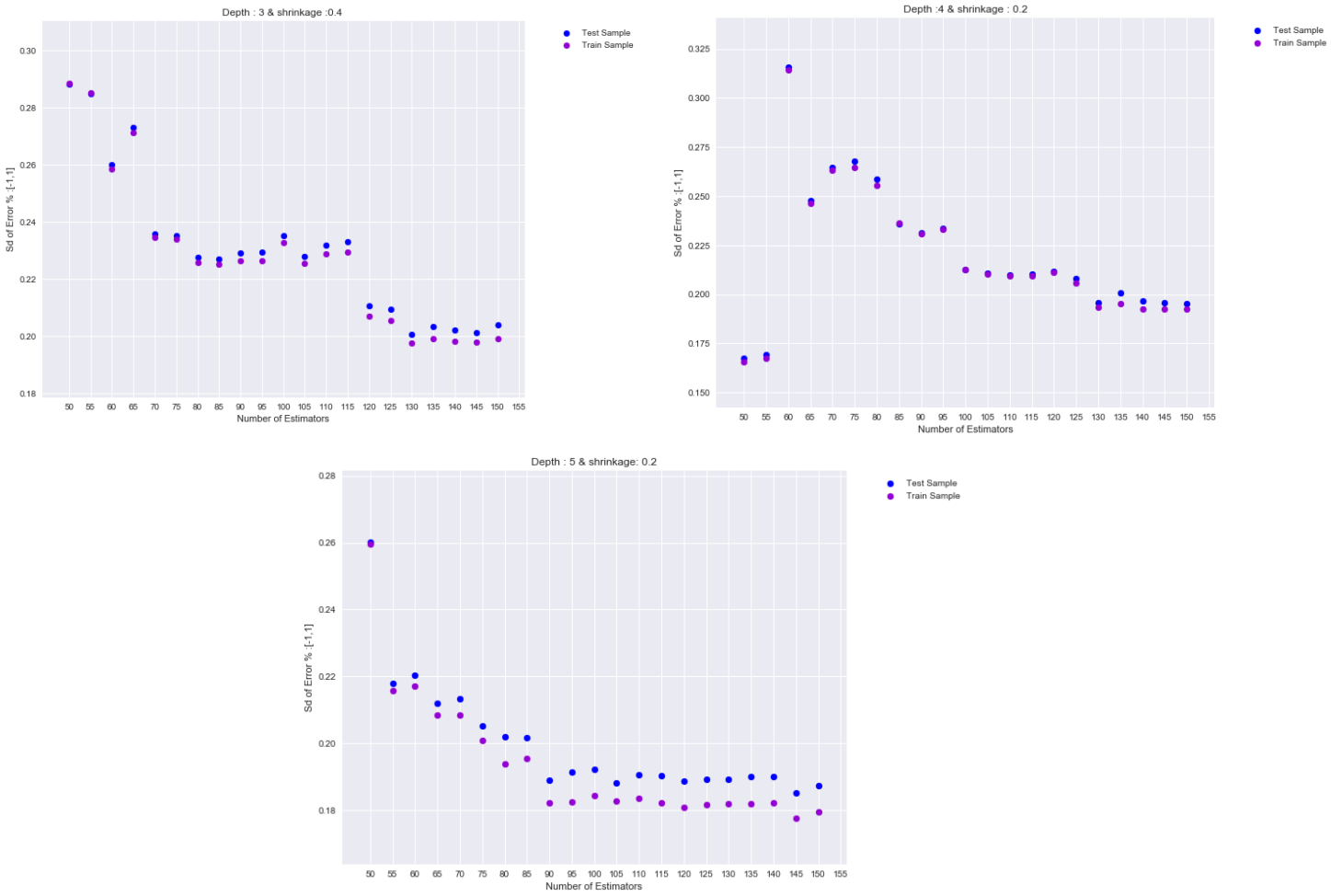
Διάγραμμα 4.25: Τυπική απόκλιση συναρτήσει του ρυθμού εκμάθησης για κάθε βάθος (3η προσέγγιση)

Επιλέχθηκαν εκείνες οι τιμές του ρυθμού εκμάθησης για τιμές βάθους =3,4,5 που παρουσιάζουν μικρή απόκλιση μεταξύ του δείγματος αξιολόγησης και εκπαίδευσης. Δεν επιλέχθηκε τιμή του ρυθμού εκμάθησης για βάθος = 6 , διότι παρατηρείται ότι υπάρχουν μεγάλες αποκλίσεις στην τυπική απόκλιση.

Πίνακας 4.7: Επιλογή παραμέτρων για 150 εκτιμητές(3η προσέγγιση)

Βάθος	**Ρυθμός εκμάθησης**
3	0.4
4	0.2
5	0.2

Στην συνέχεια, μεταβλήθηκε ο αριθμός εκτιμητών για τις παραπάνω τιμές.



Διάγραμμα 4.26: Επιλογή τελικών τιμών των παραμέτρων 3ης προσέγγισης

Οι τελικές τιμές παραμέτρων που επιλέχθηκαν είναι οι εξής:

Πίνακας 4.8: Τελικές παράμετροι 3ης προσέγγισης

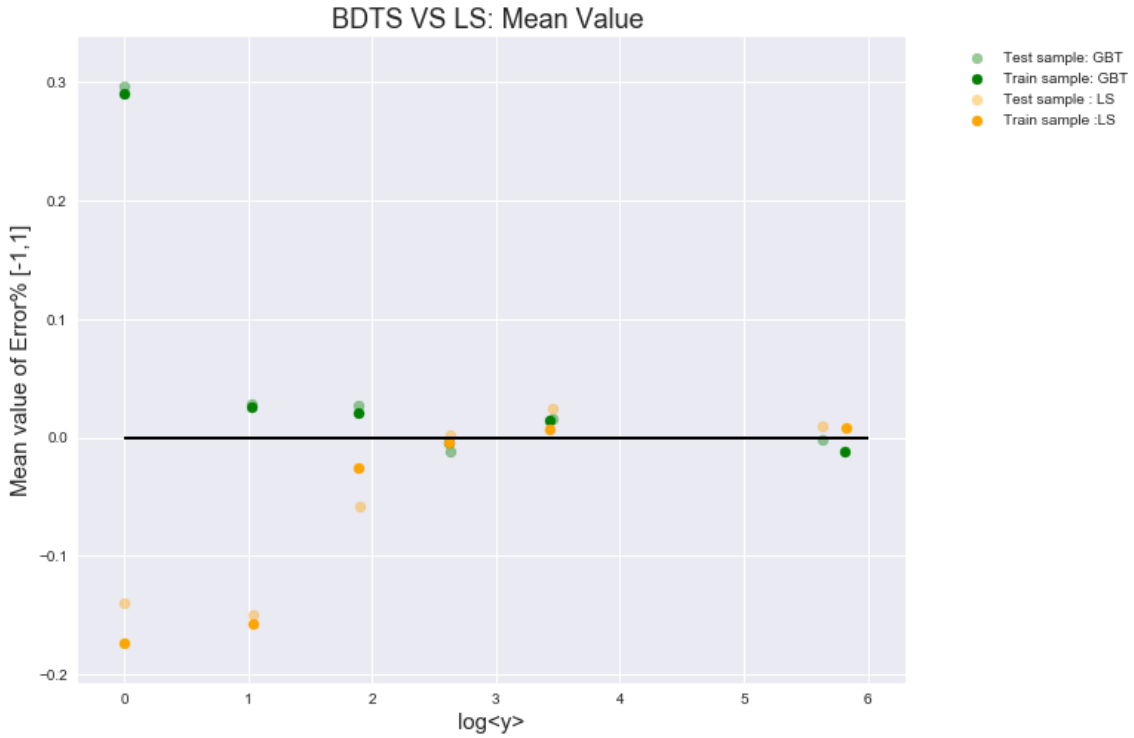
Βάθος	**Ρυθμός εκμάθησης**	**Αριθμός εκτιμητών**
3	0.4	80
4	0.2	130
5	0.2	55

Για κάθε συνδυασμό των παραπάνω παραμέτρων κατασκευάστηκαν τα αντίστοιχα μοντέλα. Το μοντέλο με τα καλύτερα αποτελέσματα είναι το παρακάτω:

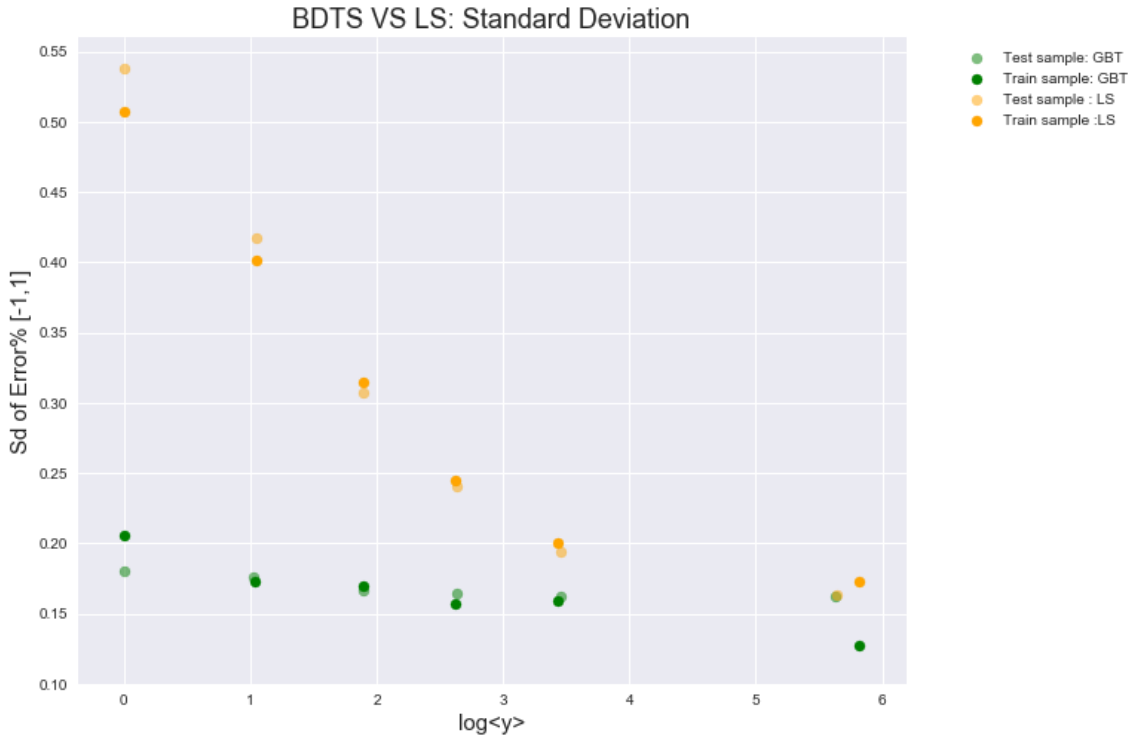
Βάθος	**Ρυθμός εκμάθησης**	**Αριθμός εκτιμητών**
4	0.2	130

Διάγραμμα 4.27: Εκτιμώμενα σφάλματα τελικού μοντέλου με Gradient

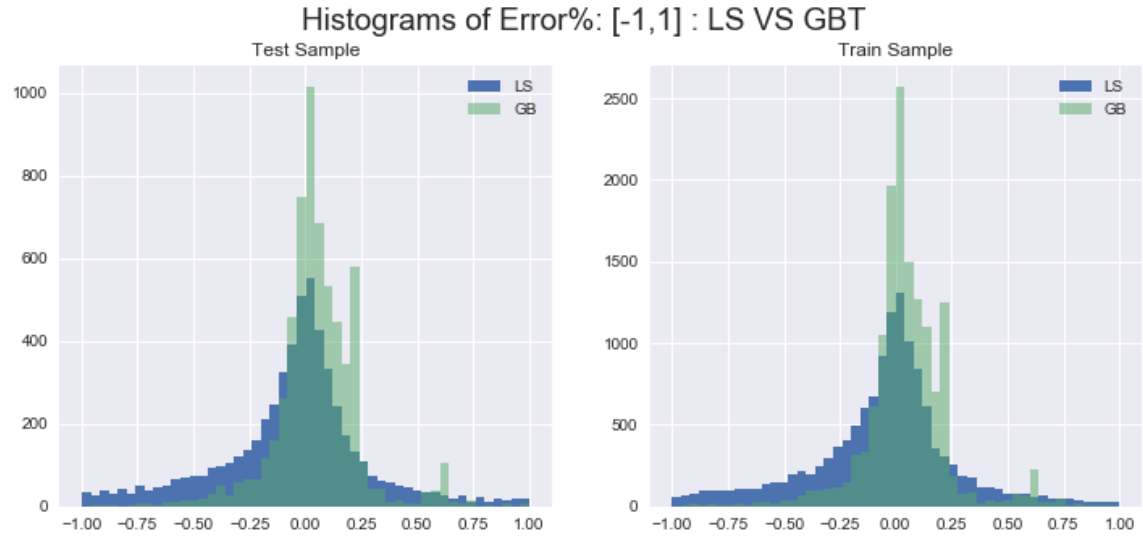
4.3.4 Σύγκριση μη γραμμικού μοντέλου με Gradient και γραμμικού μοντέλου



Διάγραμμα 4.28: Gradient VS Least Squares Method : Μέση τιμή ποσοστιαίων σφαλμάτων συναρτήσει της λογαρίθμου της μέσης τιμής της πραγματικής τιμής της μεταβλητής απόκρισης



Διάγραμμα 4.29: Gradient VS Least Squares Method : Τυπική απόκλιση ποσοστιαίων σφαλμάτων συναρτήσει της λογαρίθμου της μέσης τιμής της πραγματικής τιμής της μεταβλητής απόκρισης



Διάγραμμα 4.30: Ιστογράμματα : Gradient BDT VS Least Square Method

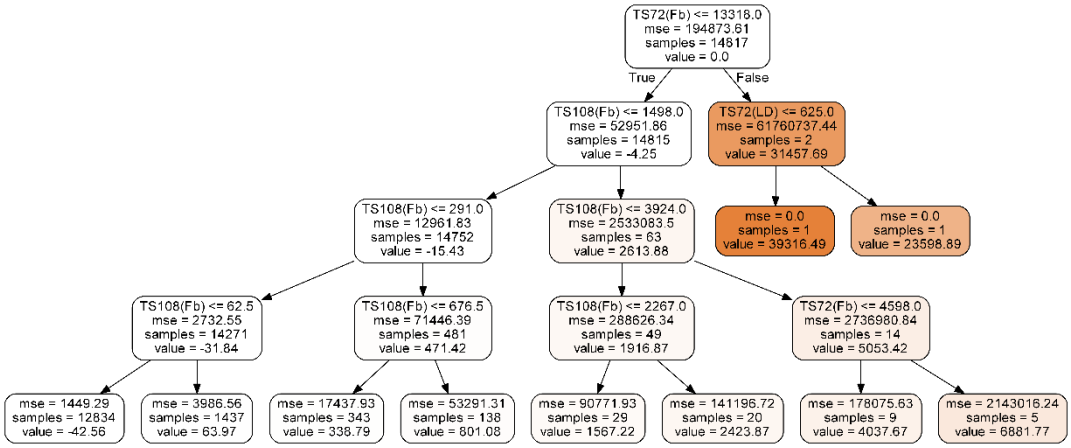
Συμπεράσματα:

I. Σύμφωνα με το διάγραμμα 4.29, οι τιμές της τυπικής απόκλισης των ποσοστιαίων σφαλμάτων για το μη γραμμικό μοντέλο είναι χαμηλότερες από τις αντίστοιχες του γραμμικού μοντέλου. Επιπλέον, έχουν μικρή απόκλιση οι τιμές της τυπικής απόκλισης του δείγματος αξιολόγησης με τις αντίστοιχες του δείγματος εκπαίδευσης για το μη γραμμικό μοντέλο.

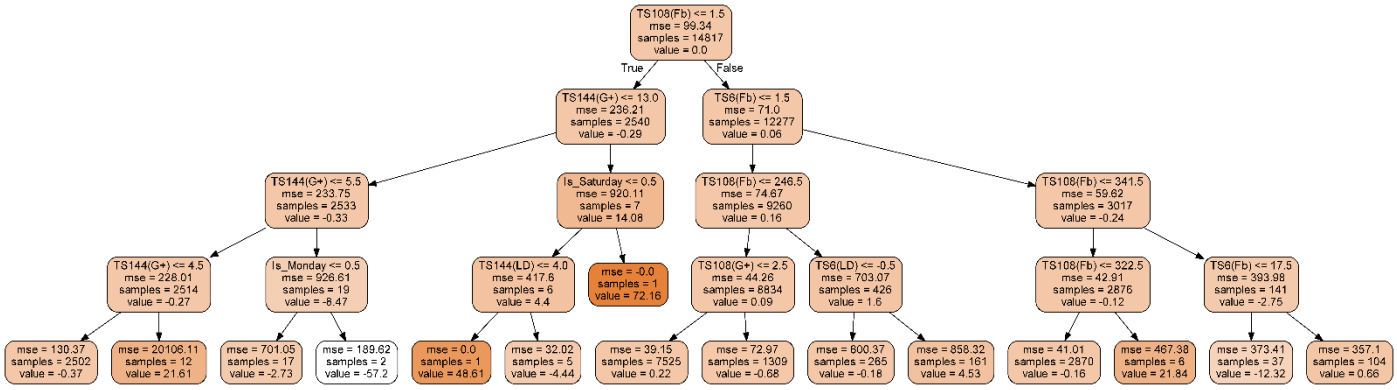
II. Η μέση τιμή των ποσοστιαίων σφαλμάτων για τα δείγματα αξιολόγησης και εκπαίδευσης είναι κοντά στο 0.

III. Στα ιστογράμματα, στο μη γραμμικό μοντέλο οι μηδενικές τιμές ποσοστιαίων σφαλμάτων έχουν μεγαλύτερη συχνότητα από τις αντίστοιχες του γραμμικού μοντέλου, δηλαδή το μη γραμμικό μοντέλο είναι πιο αποδοτικό αφού έχει μεγαλύτερο πλήθος μηδενικών αποκλίσεων.

Παρακάτω παρουσιάζεται η οπτικοποίηση του 1ου και του 130ου δέντρου απόφασης:



Διάγραμμα 4.31: Οπτικοποίηση πρώτου δέντρου απόφασης



Διάγραμμα 4.32: Οπτικοποίηση τελευταίου δέντρου απόφασης

4.4 Νευρωνικά Δίκτυα

Η κατασκευή μοντέλων με την χρήση Νευρωνικών Δικτύων έγινε με την βοήθεια της προγραμματιστικής γλώσσας '*Python*', πιο συγκεκριμένα με την βοήθεια της *Keras*[10].

Για την κατασκευή ενός Νευρωνικού Δικτύου χρειάζεται να βρεθούν οι κατάλληλες τιμές ορισμένων παραμέτρων, όπως και στα δάση δέντρων απόφασης. Αρχικά, επιλέγεται η αρχιτεκτονική δικτύου, έπειτα επιλέγεται η συνάρτηση ενεργοποίησης σε κάθε στρώμα και ο αριθμός των νευρώνων από τους οποίους αποτελείται. Στην συνέχεια, πρέπει να γίνει η επιλογή του αλγορίθμου μάθησης ('optimizer') για την μέθοδο 'Back propagation' και η τιμή του ρυθμού εκμάθησης. Τέλος, επιλέγεται η συνάρτηση κόστους και ο αριθμός των εποχών και του 'batch size'.

Πραγματοποιήθηκαν διαφορετικές προσεγγίσεις για την κατασκευή του νευρωνικού δικτύου με τις κατάλληλες παραμέτρους, διότι η καθεμιά από τις παραμέτρους επηρεάζει την άλλη και κατά συνέπεια επηρεάζει την απόδοση του μοντέλου.

4.4.1 1η προσέγγιση: Αλλαγή κλίμακας δεδομένων

Πριν την κατασκευή του νευρωνικού δικτύου, έγινε αλλαγή κλίμακας στα δεδομένα. Ο λόγος που έγινε αλλαγή κλίμακας είναι διότι το εύρος τιμών των μεταβλητών είναι πολύ μεγάλο με αποτέλεσμα το πρόβλημα να είναι δύσκολο να μοντελοποιηθεί.

Συνάρτηση Ενεργοποίησης : Σιγμοειδής

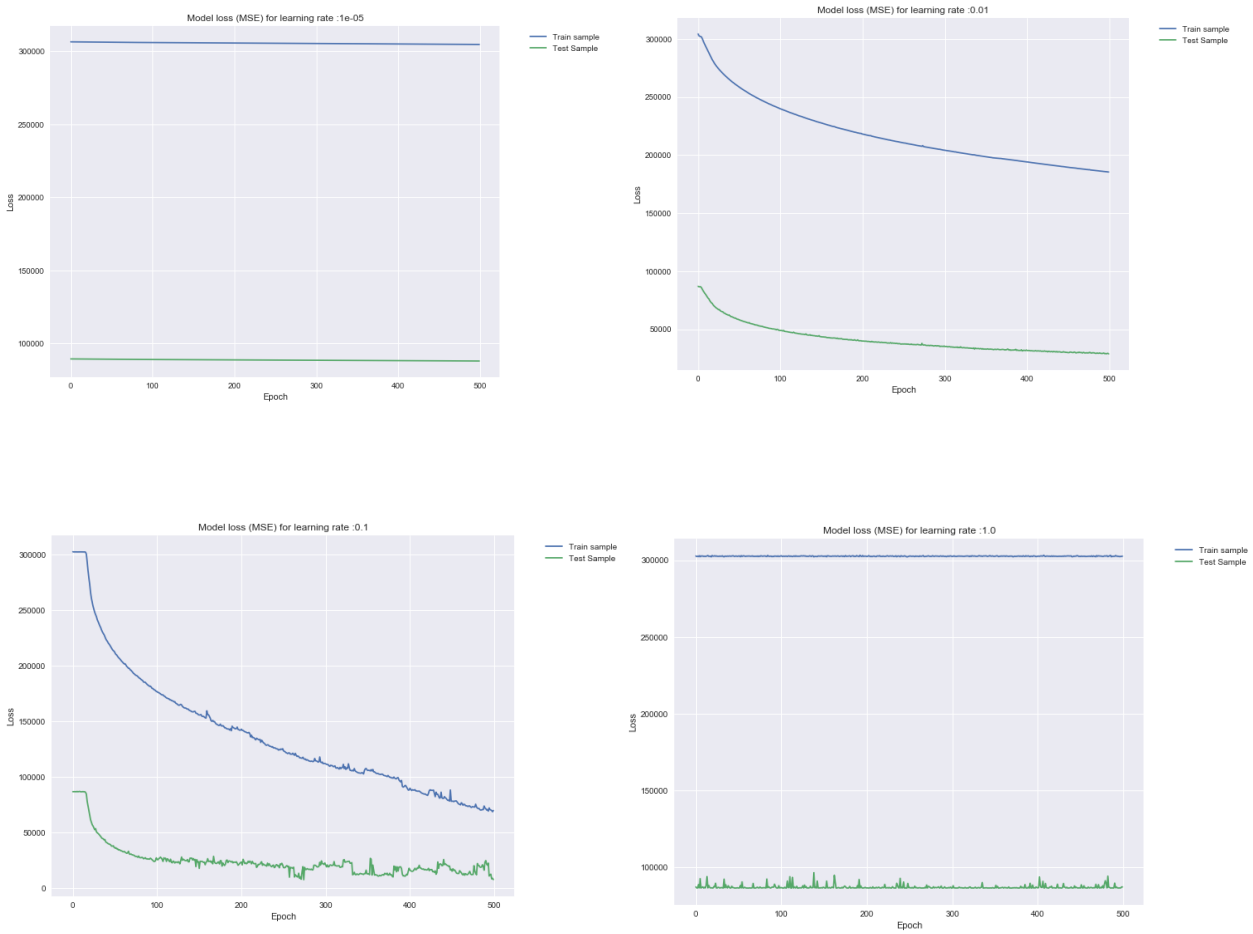
Το νευρωνικό δίκτυο που κατασκευάστηκε είναι το εξής:

Πίνακας 4.9: Αρχιτεκτονική Νευρωνικού Δικτύου για την εύρεση ρυθμού εκμάθησης με συνάρτηση ενεργοποίησης : Σιγμοειδής (1η προσέγγιση)

Στρώμα Εισόδου	30	**Συνάρτηση ενεργοποίησης**	Σιγμοειδής
Κρυφό στρώμα	30	**Συνάρτηση ενεργοποίησης**	Σιγμοειδής
Στρώμα εξόδου	1	**Συνάρτηση ενεργοποίησης**	Γραμμική
'Optimizer'		ADAM	
Συνάρτηση κόστους		MSE	
Αριθμός εποχών		500	
'Batch size'		100	

[10] Αναλυτικότερα στο Παράρτημα

Για την συγκεκριμένη αρχιτεκτονική νευρωνικού δικτύου χρειάζεται να βρεθεί η τιμή του ρυθμού εκμάθησης που δεν θα επιφέρει υπερεκπαίδευση ή υποεκπαίδευση στο σύνολο δεδομένων. Ακολούθως παρουσιάζονται τα διαγράμματα της συνάρτησης κόστους συναρτήσει του αριθμού των εποχών για κάθε νευρωνικό δίκτυο που κατασκευάστηκε για τις διαφορετικές τιμές ρυθμού εκμάθησης [10^{-5} , 10^{-2} , 10^{-1} , 1] .



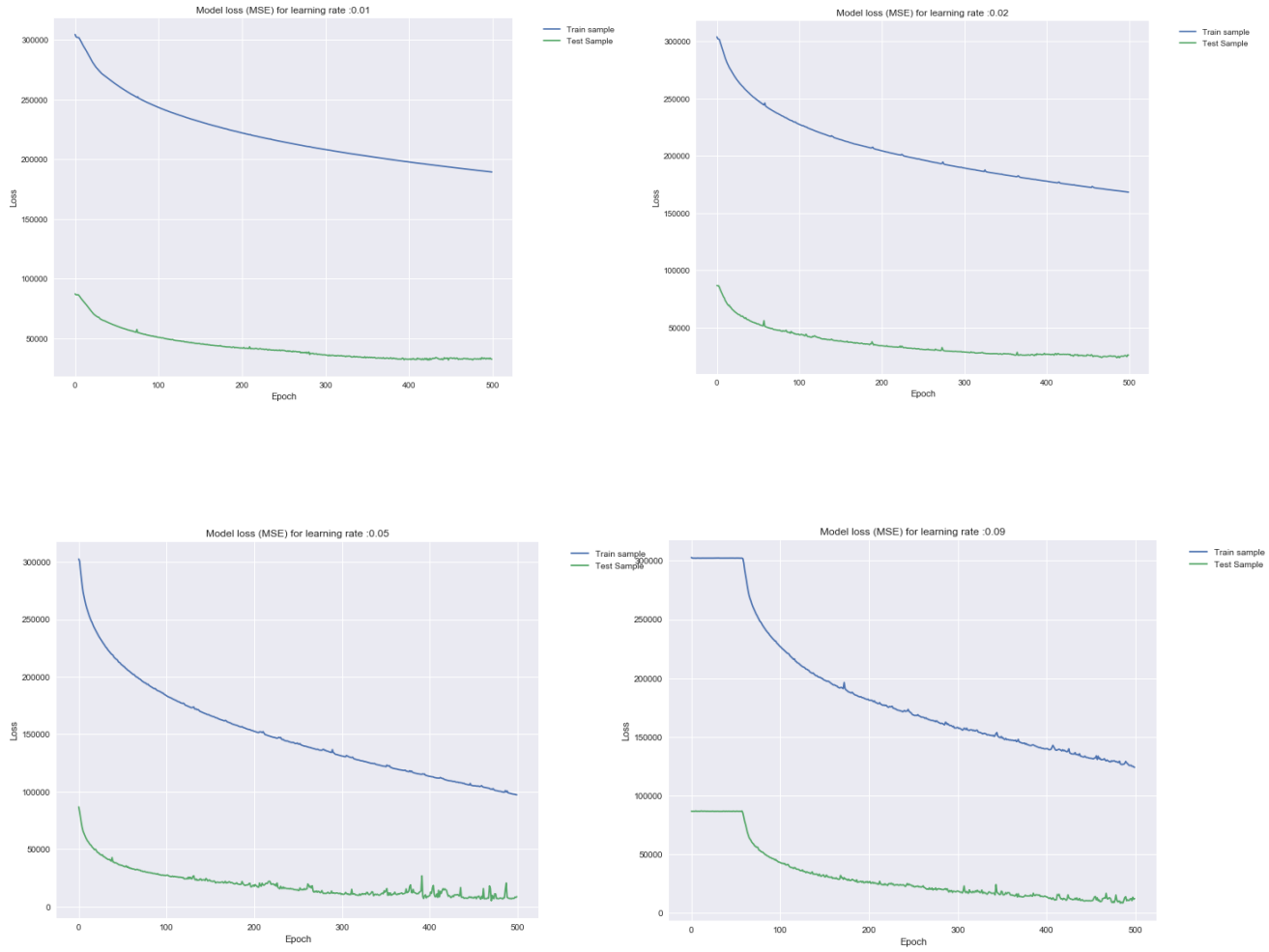
Διάγραμμα 4.33: Συνάρτηση κόστους συναρτήσει του αριθμού εποχών για διαφορετικές τιμές ρυθμού εκμάθησης με συνάρτηση ενεργοποίησης: σιγμοειδής ($1^η$ προσέγγιση)

Παρατηρήσεις

I. Σύμφωνα με τα παραπάνω διαγράμματα, ακραίες τιμές του ρυθμού εκμάθησης (10^{-5} και 1)έχουν σαν αποτέλεσμα να μην εκπαιδεύεται το μοντέλο, διότι οι τιμές των συναρτήσεων κόστους παραμένουν σταθερές και δεν μειώνονται καθώς αυξάνονται οι εποχές.

II. Η τιμή του ρυθμού εκμάθησης = 0.1 προκαλεί διακυμάνσεις στην συνάρτηση κόστους.

Ακολουθείται η ίδια διαδικασία με τιμές του ρυθμού εκμάθησης : [0.01, 0.02, 0.05, 0.09].



Διάγραμμα 4.34: Συνάρτηση κόστους συναρτήσει του αριθμού εποχών για τιμές ρυθμού εκμάθησης στο διάστημα [0.01,1] με συνάρτηση ενεργοποίησης: σιγμοειδής (1η προσέγγιση)

Σύμφωνα με τα παραπάνω διαγράμματα, επιλέχθηκε η τιμή για τον ρυθμό εκμάθησης να είναι 0.02. Όμως, ένα σημαντικό πρόβλημα που ήδη φαίνεται να υπάρχει είναι ότι οι τιμές της συνάρτησης κόστους των δειγμάτων αξιολόγησης και εκπαίδευσης δεν μηδενίζουν ύστερα από κάποιο αριθμό εποχών, ούτε εμφανίζεται σημείο ελαχίστου.

Στην συνέχεια πραγματοποιήθηκε η ίδια διαδικασία για την συνάρτηση ενεργοποίησης *Relu*.

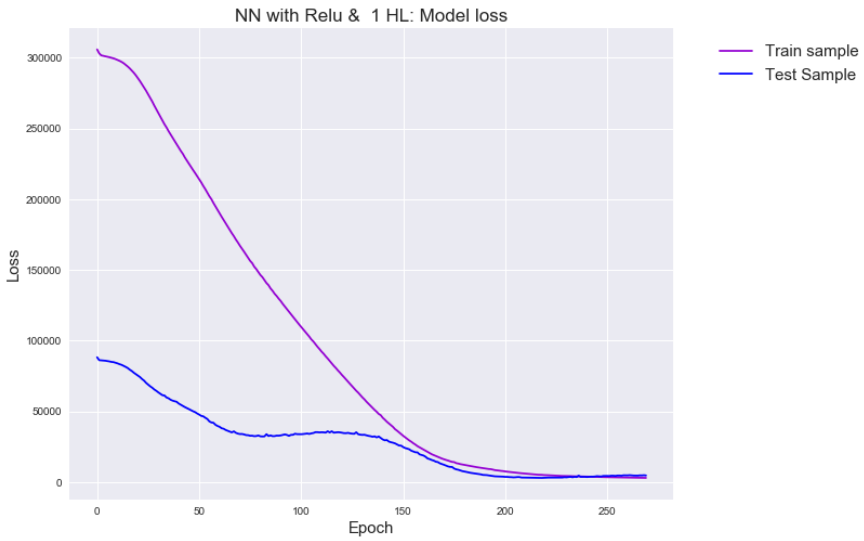
Συνάρτηση Ενεργοποίησης : Relu

Το νευρωνικό δίκτυο που κατασκευάστηκε είναι το εξής:

Πίνακας 4.10: Αρχιτεκτονική Νευρωνικού Δικτύου για την εύρεση ρυθμού εκμάθησης με συνάρτηση ενεργοποίησης : Relu (1η προσέγγιση)

Στρώμα Εισόδου	30	**Συνάρτηση ενεργοποίησης**	Relu
Κρυφό στρώμα	30	**Συνάρτηση ενεργοποίησης**	Relu
Στρώμα εξόδου	1	**Συνάρτηση ενεργοποίησης**	Γραμμική
'Optimizer'		ADAM	
Συνάρτηση κόστους		MSE	
Αριθμός εποχών		270	
'Batch size'		100	

Η τιμή του ρυθμού εκμάθησης που επιλέχθηκε είναι 0.001. Παρακάτω παρουσιάζεται το διάγραμμα της συνάρτησης κόστους συναρτήσει του αριθμού εποχών για την συγκεκριμένη τιμή του ρυθμού εκμάθησης.



Διάγραμμα 4.35: Συνάρτηση κόστους συναρτήσει του αριθμού εποχών για την συνάρτηση ενεργοποίησης: Relu και ρυθμό εκμάθησης : 0.001(1η προσέγγιση)

Στην συνέχεια, κατασκευάστηκαν νευρωνικά δίκτυα με 1,2 και 3 κρυφά στρώματα για κάθε συνάρτηση ενεργοποίησης και την αντίστοιχη τιμή του ρυθμού εκμάθησης που επιλέχθηκε.

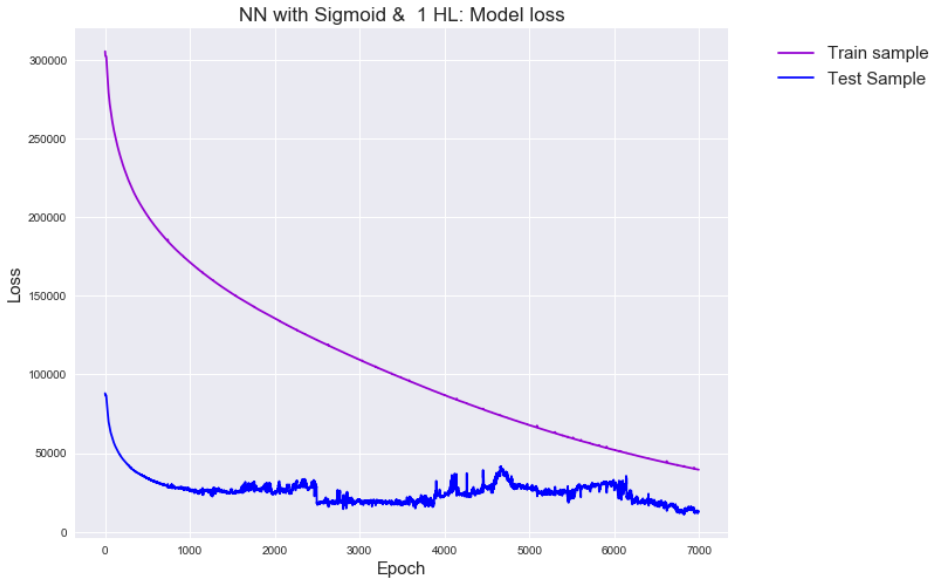
Συνάρτηση Ενεργοποίησης : Σιγμοειδής

Για την συγκεκριμένη συνάρτηση ενεργοποίησης, έγινε δοκιμή μεγαλύτερου αριθμού εποχών, διότι η συνάρτηση κόστους δεν μηδενιζόταν μετά από κάποιο αριθμό εποχών. Επιπλέον, αλλάχθηκε ο αριθμός του 'batch size', λόγω υπολογιστικής πολυπλοκότητας.

Πιο συγκεκριμένα, κατασκευάστηκε το νευρωνικό δίκτυο με την παρακάτω αρχιτεκτονική:

Πίνακας 4.10: Νευρωνικό Δίκτυο με 7000 εποχές και συνάρτηση ενεργοποίησης: Σιγμοειδής (1η προσέγγιση)

Στρώμα Εισόδου	30	**Συνάρτηση ενεργοποίησης**	Σιγμοειδής
Κρυφό στρώμα	30	**Συνάρτηση ενεργοποίησης**	Σιγμοειδής
Στρώμα εξόδου	1	**Συνάρτηση ενεργοποίησης**	Γραμμική
'Optimizer'		ADAM	
Συνάρτηση κόστους		MSE	
Αριθμός εποχών		7000	
'Batch size'		512	

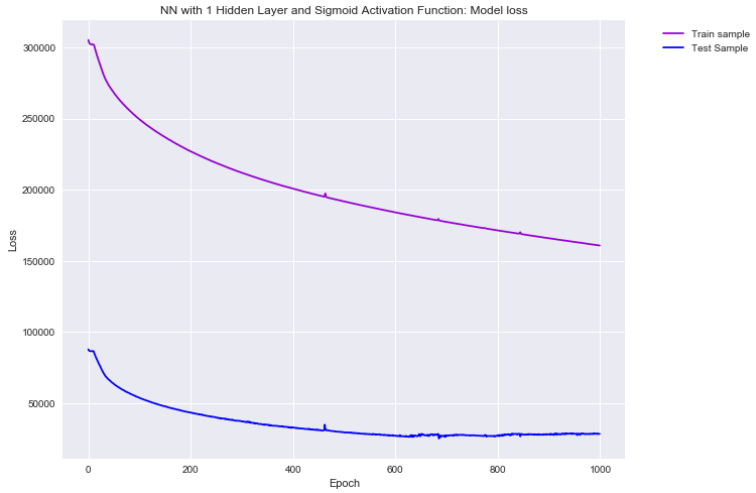


Διάγραμμα 4.36: Συνάρτηση κόστους συναρτήσει αριθμού εποχών (7000) (1η προσέγγιση Σιγμοειδής)

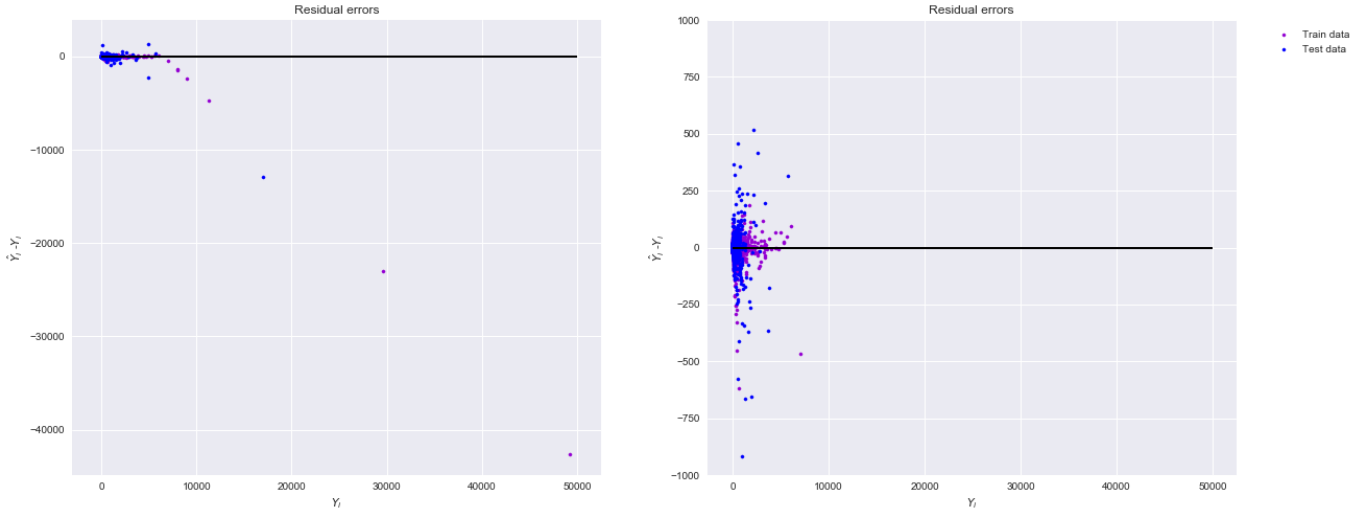
Σύμφωνα με το παραπάνω διάγραμμα, παρουσιάζονται διακυμάνσεις της συνάρτησης κόστους του δείγματος αξιολόγησης μετά από 1000 εποχές, ενώ η συνάρτηση κόστους του δείγματος εκπαίδευσης μειώνεται ομοιόμορφα.

Ακολούθως παρουσιάζονται τα αποτελέσματα που προέκυψαν από την κατασκευή νευρωνικών δικτύων με 1,2 και 3 κρυφά στρώματα με σιγμοειδή συνάρτηση ενεργοποίησης και 1000 εποχές.

- *1 κρυφό στρώμα*

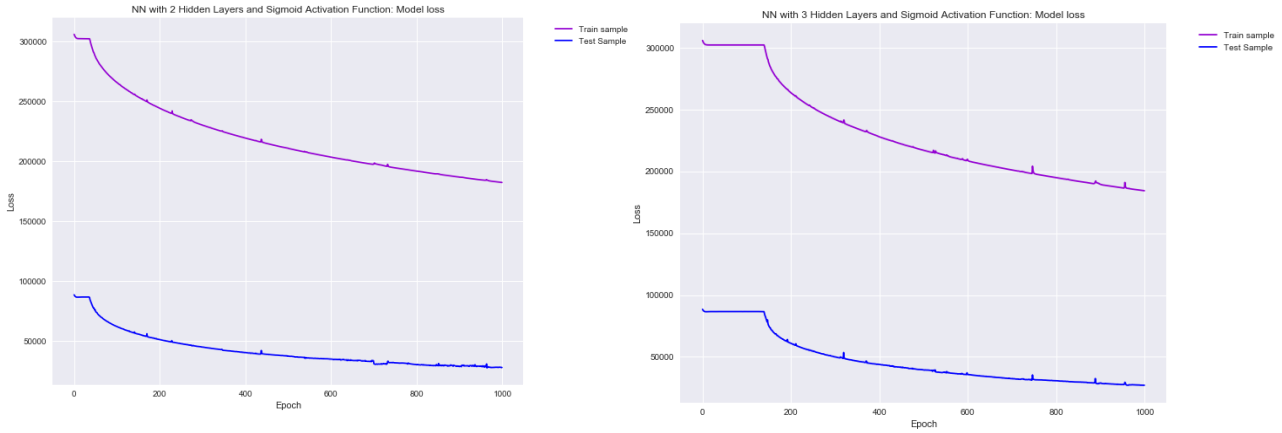


Διάγραμμα 4.37: Νευρωνικό Δίκτυο με 1 κρυφό στρώμα, σιγμοειδή : Συνάρτηση κόστους (1η προσέγγιση)



Διάγραμμα 4.38: Νευρωνικό Δίκτυο με 1 κρυφό στρώμα, σιγμοειδή : Εκτιμώμενα σφάλματα (1η προσέγγιση)

- *2 και 3 κρυφά στρώματα*



Διάγραμμα 4.39: Συναρτήσεις κόστους για νευρωνικά δίκτυα με 2 και 3 κρυφά στρώματα και σιγμοειδή(1η προσέγγιση)

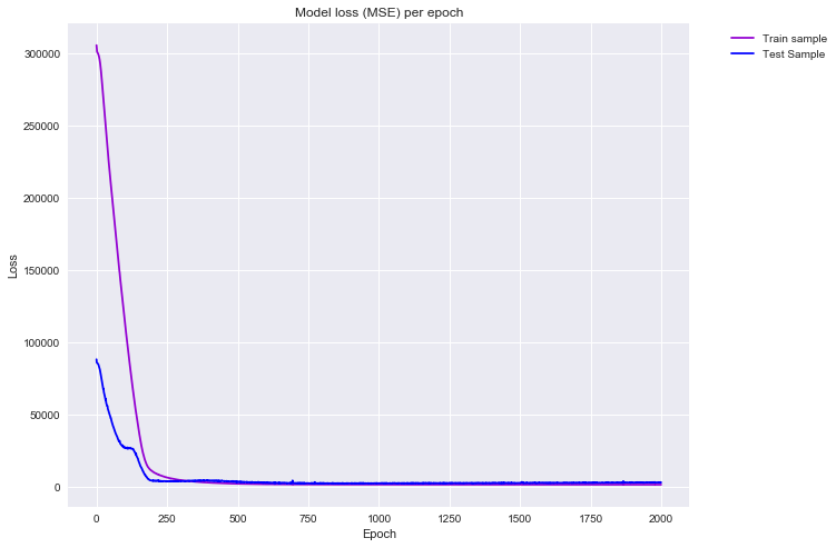
Παρατηρείται ότι με την προσθήκη στρωμάτων δεν βελτιώνεται η απόδοση του μοντέλου, δηλαδή η συνάρτηση κόστους συνεχίζει να μην μειώνεται και να μην λαμβάνει μηδενικές τιμές μετά από κάποιο αριθμό εποχών.

Συνάρτηση Ενεργοποίησης : Relu

Για την συγκεκριμένη συνάρτηση ενεργοποίησης επιλέχθηκαν 2000 εποχές.

Πίνακας 4.11: Νευρωνικό Δίκτυο με 2000 εποχές και συνάρτηση ενεργοποίησης: Relu (1η προσέγγιση)

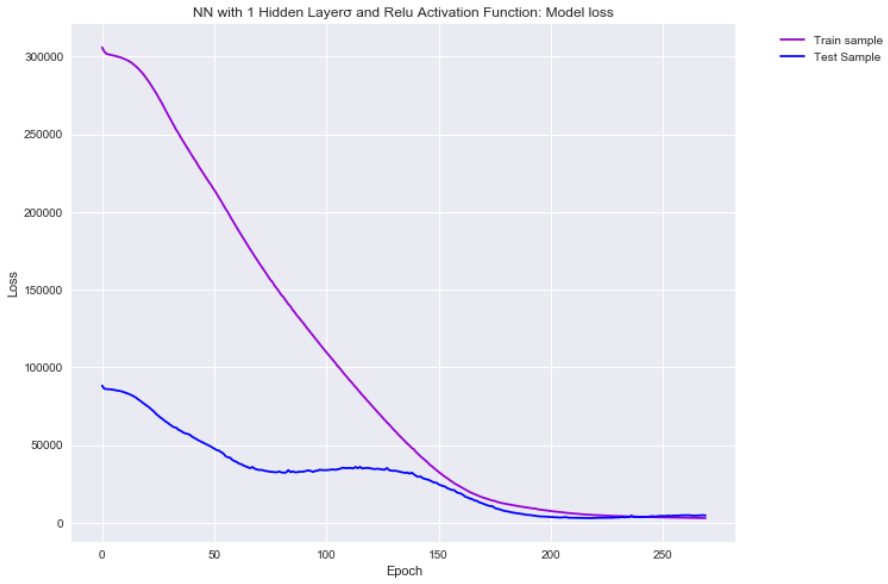
Στρώμα Εισόδου	30	**Συνάρτηση ενεργοποίησης**	Relu
Κρυφό στρώμα	30	**Συνάρτηση ενεργοποίησης**	Relu
Στρώμα εξόδου	1	**Συνάρτηση ενεργοποίησης**	Γραμμική
‘Optimizer’		ADAM	
Συνάρτηση κόστους		MSE	
Αριθμός εποχών		2000	
‘Batch size’		100	



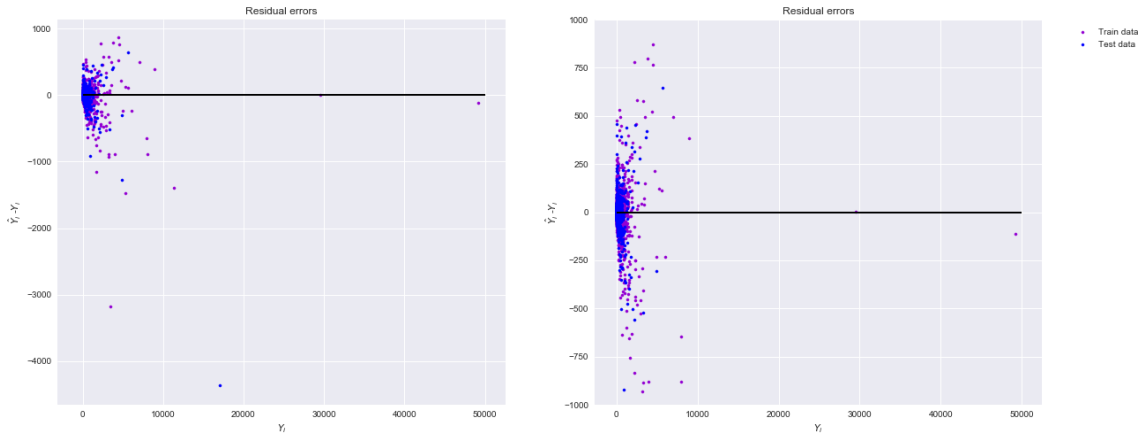
Διάγραμμα 4.40: : Συνάρτηση κόστους συναρτήσει αριθμού εποχών (2000) (1η προσέγγιση Relu)

Στο παραπάνω διάγραμμα, γίνεται αντιληπτό ότι δεν χρειάζεται μεγάλος αριθμός εποχών, διότι η συνάρτηση κόστους μηδενίζει σε μικρότερο αριθμό εποχών. Οπότε κατασκευάζεται νευρωνικό δίκτυο με 270 εποχές και 1,2 και 3 κρυφά στρώματα.

- *1 κρυφό στρώμα*

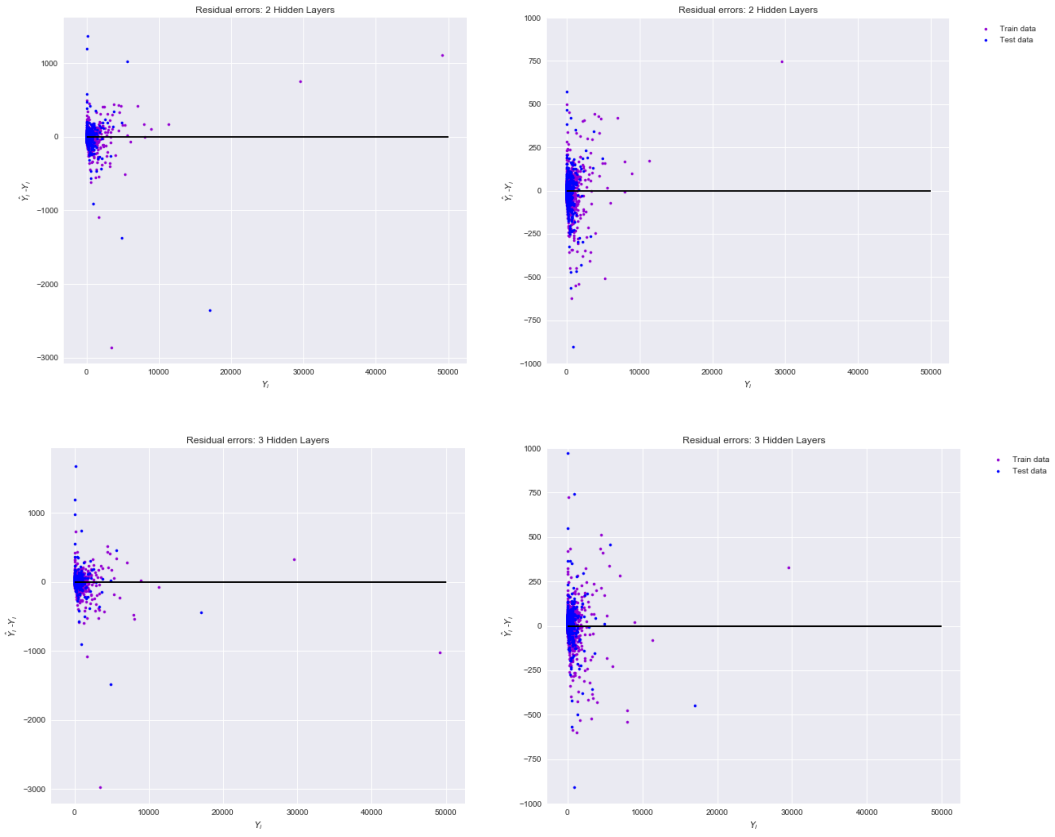


Διάγραμμα 4.41: Νευρωνικό Δίκτυο με 1 κρυφό στρώμα, Relu : Συνάρτηση κόστους (1η προσέγγιση)



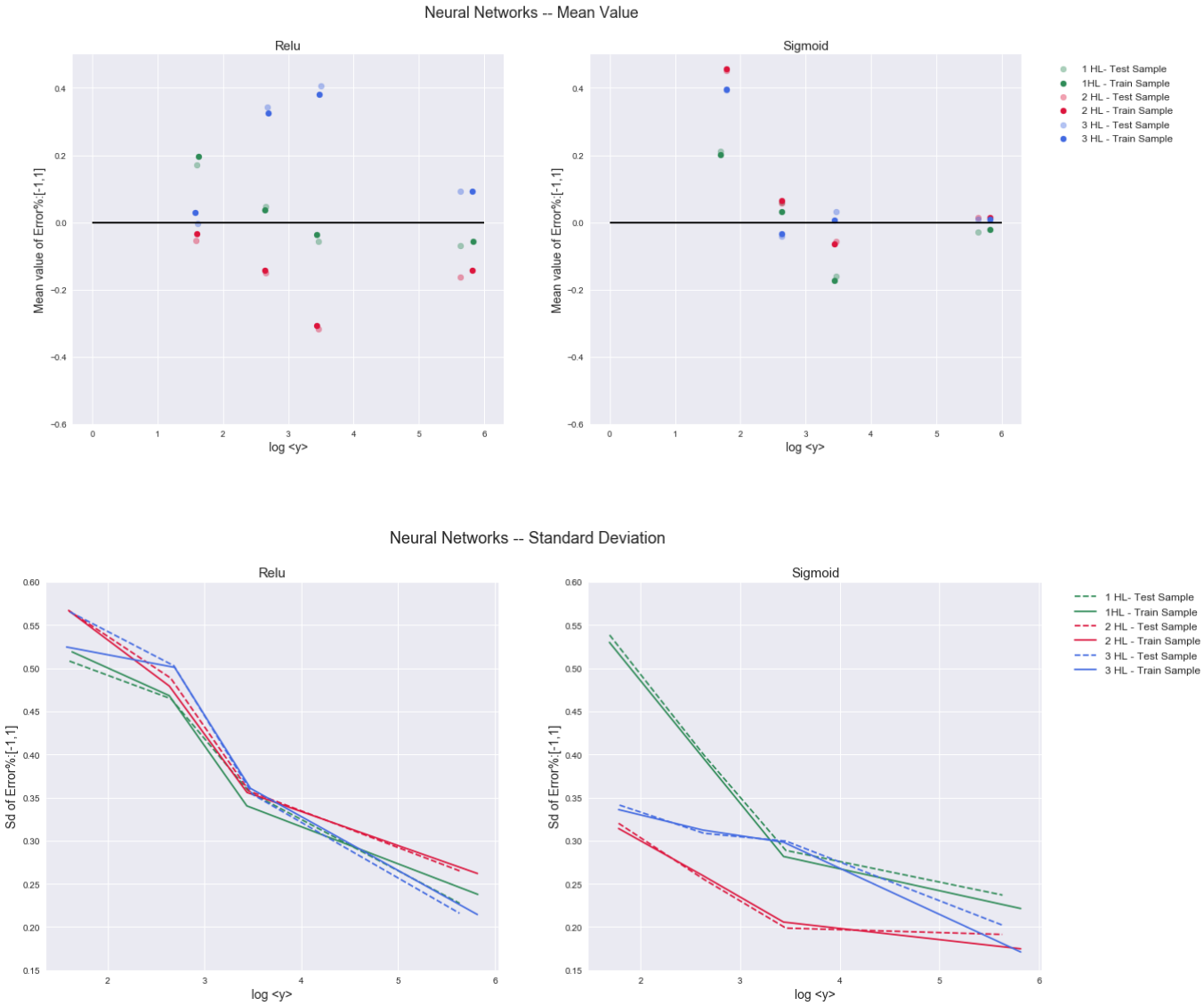
Διάγραμμα 4.42: Εκτιμώμενα σφάλματα 1ης προσέγγισης NN με Relu και 1 κρυφό στρώμα

Σύμφωνα με τα παραπάνω διαγράμματα, παρατηρείται ότι παρόλο που η συνάρτηση κόστους έχει μια σχετικά καλή συμπεριφορά, οι αποκλίσεις των δειγμάτων αξιολόγησης και εκπαίδευσης είναι σχετικά μεγάλες. Με την προσθήκη 2 και 3 στρωμάτων, αυτές οι αποκλίσεις δεν μειώνονται, όπως φαίνεται ακολούθως:



Διάγραμμα 4.43 Εκτιμώμενα σφάλματα 1ης προσέγγισης NN με Relu για 2 και 3 κρυφά στρώματα αντίστοιχα

Σύγκριση αποτελεσμάτων



Διάγραμμα 4.44: Σύγκριση αποτελεσμάτων 1ης προσέγγισης NN

Συμπεράσματα

I. Η μέση τιμή των ποσοστιαίων σφαλμάτων των δειγμάτων εκπαίδευσης και αξιολόγησης έχει σχετικά μικρές τιμές και γύρω από το 0, ενώ η τυπική απόκλιση δεν μειώνεται ομοιόμορφα.

II. Χρειάζεται μετασχηματισμός στα δεδομένα, γιατί δεν προκύπτουν αποδοτικά μοντέλα. Ο λόγος που συμβαίνει αυτό είναι διότι η αλλαγή κλίμακας είχε σαν αποτέλεσμα να χάνεται η πληροφορία που δίνουν οι υψηλές τιμές, διότι όλες ύστερα από τον μετασχηματισμό έλαβαν την τιμή 1.

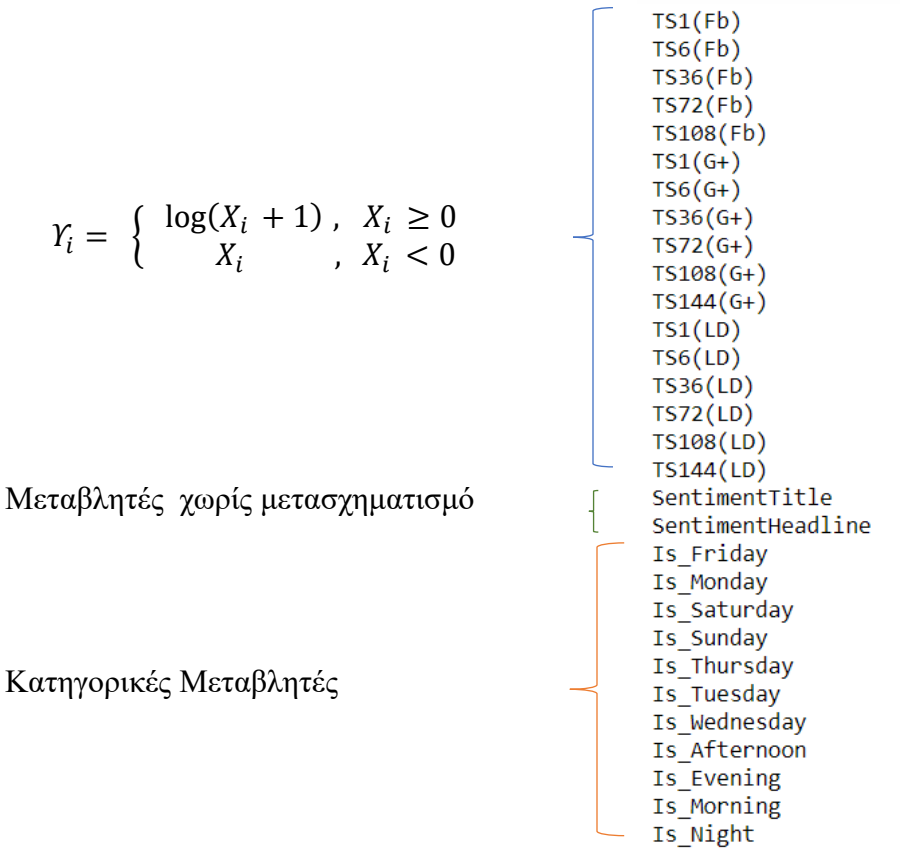
4.4.2 2η προσέγγιση : Μετασχηματισμός δεδομένων

Τα αποτελέσματα που προέκυψαν από την 1η προσέγγιση του προβλήματος με νευρωνικά δίκτυα δείχνουν ότι το νευρωνικό δίκτυο δεν μπορεί να εκπαιδευθεί, παρόλο που έχει γίνει αλλαγή κλίμακας στα δεδομένα. Με την 2η προσέγγιση του προβλήματος, δηλαδή τον μετασχηματισμό δεδομένων, στόχος είναι να εκλεχθεί αν μπορεί να εκπαιδευθεί το νευρωνικό δίκτυο και να δημιουργήσει ένα αποδοτικό μοντέλο.

1ο Βήμα

Ο μετασχηματισμός που έγινε στις επεξηγηματικές μεταβλητές παρουσιάζεται παρακάτω:

$$Y_i = \begin{cases} \log(X_i + 1), & X_i \geq 0 \\ X_i, & X_i < 0 \end{cases}$$



Πιο αναλυτικά, οι μεταβλητές που δεν μετασχηματίστηκαν είναι το Sentiment Title και Sentiment Headline, διότι λαμβάνουν τιμές στο διάστημα [-1,1]. Επιπλέον οι κατηγορικές μεταβλητές, όπως φαίνεται παραπάνω, δεν μετασχηματιστήκαν. Τέλος, εφαρμόστηκε ο μετασχηματισμός (1) στις μεταβλητές που δείχνουν την δημοτικότητα του άρθρου σε κάποιο χρονικό διάστημα, διότι έχουν μεγάλο εύρος τιμών. Επειδή αυτές οι μεταβλητές λαμβάνουν μηδενικές τιμές προστέθηκε στον λογάριθμο ο όρος 1 ώστε να αντιστοιχίζονται σε μηδενικές τιμές, μετά τον μετασχηματισμό.

2ο Βήμα

Αλλαγή κλίμακας στο διάστημα [-1,1] των αριθμητικών μεταβλητών, δηλαδή Sentiment Title, Sentiment Headline και TS_i.

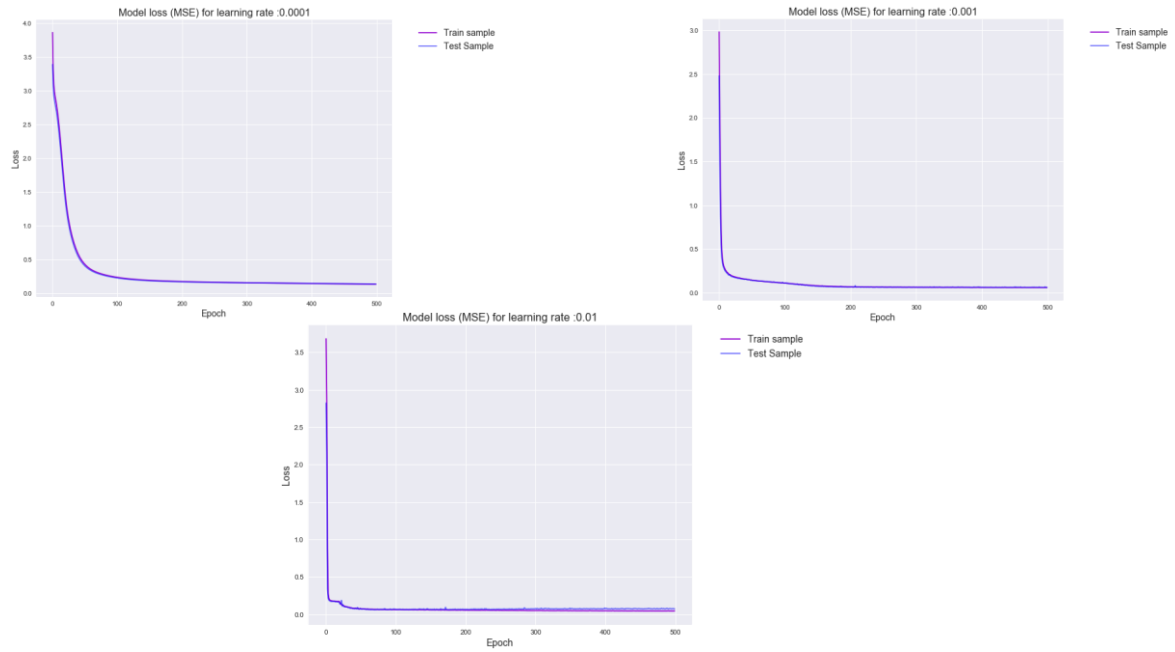
Στην συνέχεια, έγινε η επιλογή του ρυθμού εκμάθησης για κάθε συνάρτηση ενεργοποίησης για νευρωνικό δίκτυο με την παρακάτω αρχιτεκτονική και τις παρακάτω παραμέτρους:

Πίνακας 4.12: Νευρωνικό Δίκτυο 2ης προσέγγισης για κάθε τιμή του ρυθμού εκμάθησης

Στρώμα Εισόδου	30	**Συνάρτηση ενεργοποίησης**	Σιγμοειδής/ Tanh/ Relu
Κρυφό στρώμα	30	**Συνάρτηση ενεργοποίησης**	Σιγμοειδής/ Tanh/ Relu
Στρώμα εξόδου	1	**Συνάρτηση ενεργοποίησης**	Γραμμική
'Optimizer'	ADAM		
Συνάρτηση κόστους	MSE		
Αριθμός εποχών	500		
'Batch size'	100		
Ρυθμός εκμάθησης	[0.0001, 0.001, 0.01]		

Τα αποτελέσματα που προέκυψαν για κάθε συνάρτηση ενεργοποίησης είναι τα ακόλουθα:

Συνάρτηση Ενεργοποίησης : Σιγμοειδής



Διάγραμμα 4.45: Συνάρτηση κόστους με σιγμοειδή (2η προσέγγιση)

Πίνακας 4.13: Αποτελέσματα για σιγμοειδή (2η προσέγγιση)

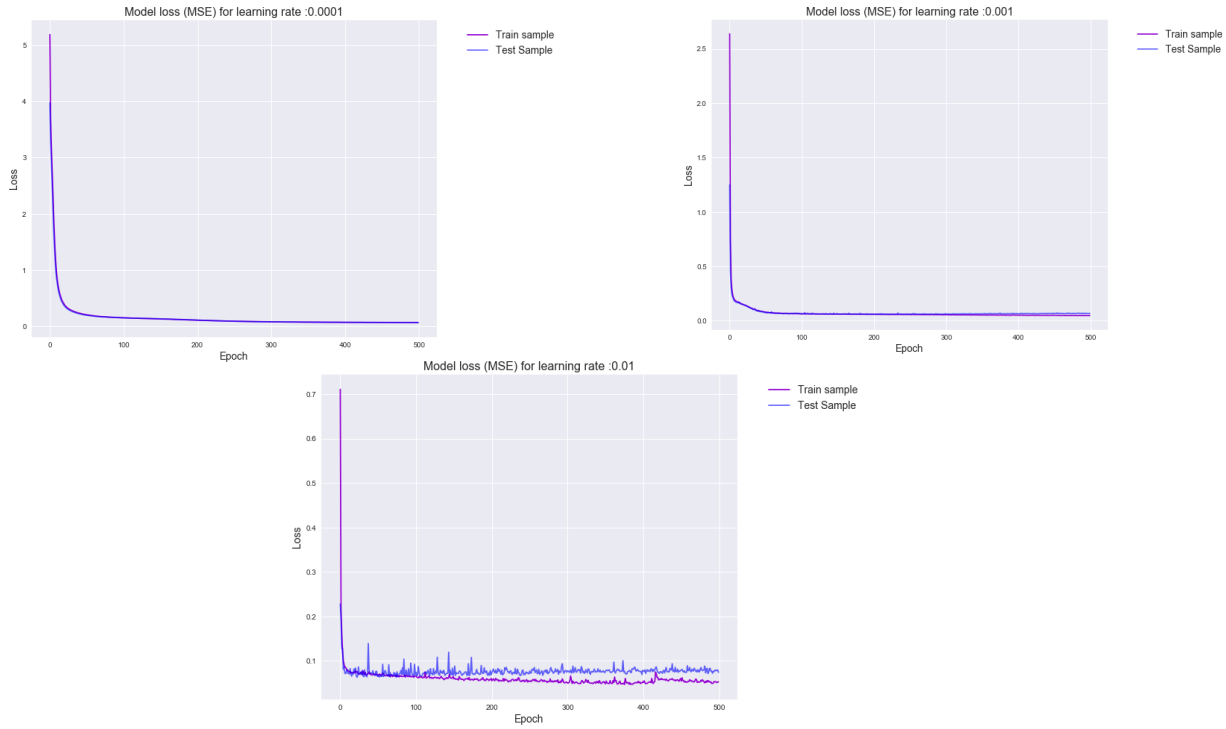
Σιγμοειδής	***Δείγμα αξιολόγησης***		***Δείγμα εκπαίδευσης***	
Ρυθμός εκμάθησης	*Μέση τιμή*	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*
0.0001	*0.033*	*0.241*	*0.031*	*0.242*
0.001	*-0.001*	*0.145*	*-0.001*	*0.142*
0.01	*0.013*	*0.156*	*0.129*	*0.143*

Σύμφωνα με τα διαγράμματα της συνάρτησης κόστους, πιο απότομη κλίση έχει για ρυθμό εκπαίδευσης 0.001. Επιπλέον, χαμηλότερες τιμές για την μέση τιμή και την τυπική απόκλιση των ποσοστιαίων σφαλμάτων στο [-1,1] για το δείγμα αξιολόγησης και εκπαίδευσης παρουσιάζονται για ρυθμό εκμάθησης = 0.001.

Συνάρτηση Ενεργοποίησης : Tanh

Στην 2η προσέγγιση δεν χρησιμοποιήθηκε η συνάρτηση Tanh, διότι έχει παρόμοια χαρακτηριστικά με την σιγμοειδή. Στην τελική προσέγγιση όμως θα χρησιμοποιηθεί, διότι μπορεί να επιφέρει καλύτερα αποτελέσματα.

Τα αποτελέσματα που προέκυψαν για την συγκεκριμένη συνάρτηση ενεργοποίησης είναι:



Διάγραμμα 4.46: Συνάρτηση κόστους με Tanh(2η προσέγγιση)

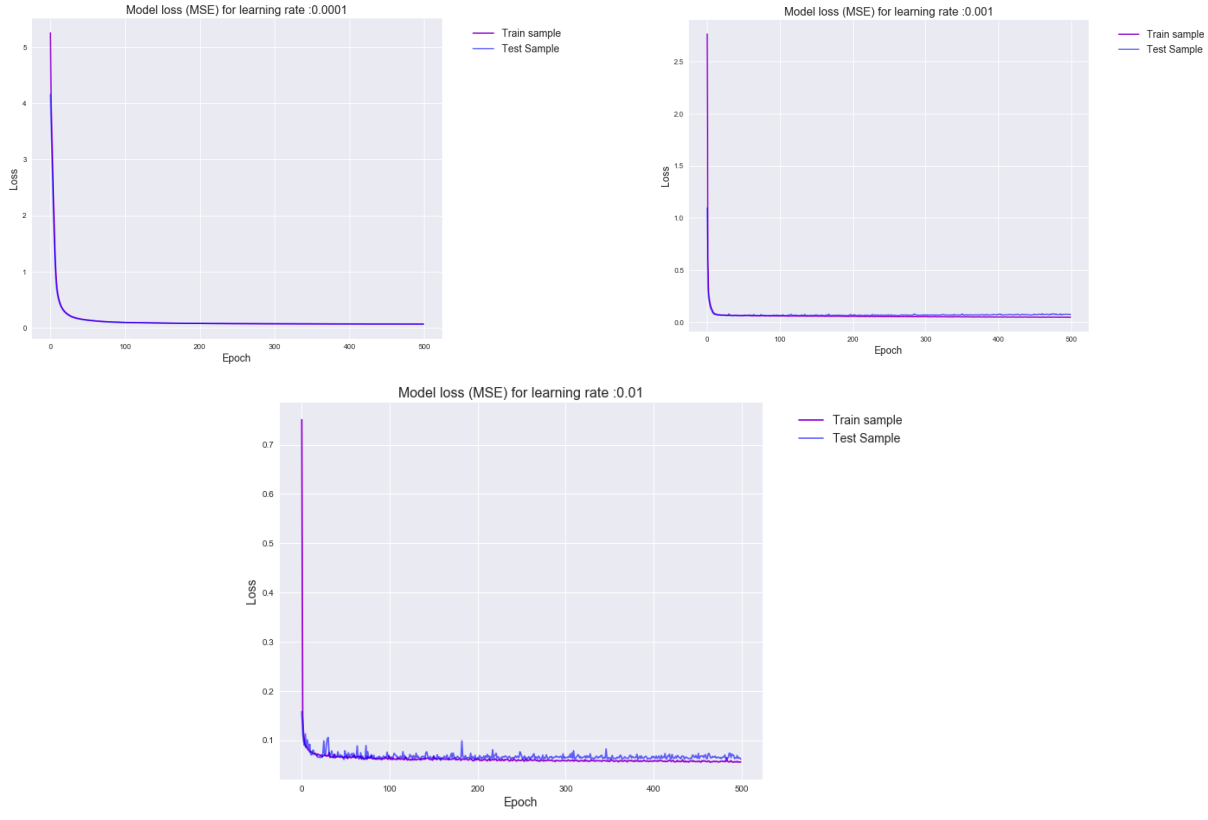
Πίνακας 4.14: Αποτελέσματα για tanh (2η προσέγγιση)

Tanh	***Δείγμα αξιολόγησης***		***Δείγμα εκπαίδευσης***	
Ρυθμός εκμάθησης	*Μέση τιμή*	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*
0.0001	*0.006*	*0.171*	*0.005*	*0.170*
0.001	*0.032*	*0.153*	*0.033*	*0.147*
0.01	*-0.010*	*0.151*	*-0.008*	*0.142*

Με τον ίδιο τρόπο, όπως στην σιγμοειδή συνάρτηση ενεργοποίησης, επιλέχθηκε ρυθμός εκμάθησης = 0.001.

Συνάρτηση Ενεργοποίησης : Relu

Τα αποτελέσματα που προέκυψαν για την συνάρτηση ενεργοποίηση Relu είναι τα ακόλουθα:



Διάγραμμα 4.47: Συνάρτηση κόστους με Relu (2η προσέγγιση)

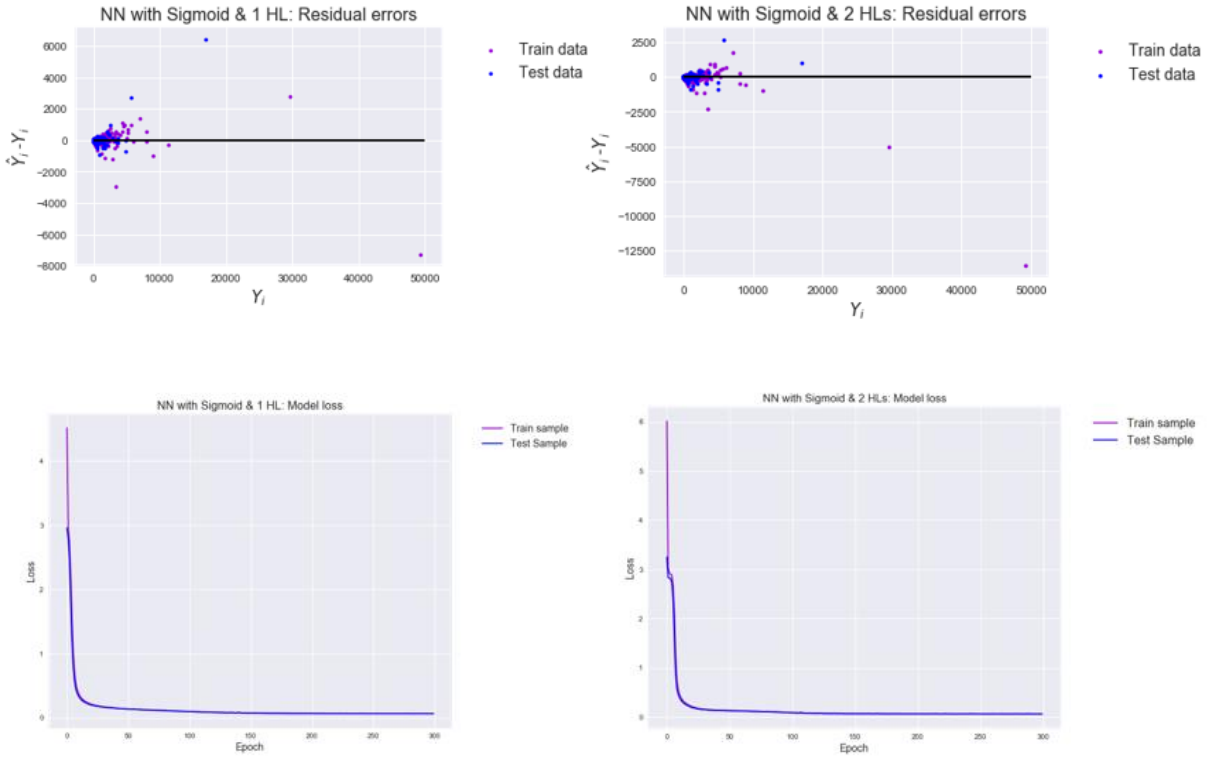
Πίνακας 4.15: Αποτελέσματα για relu (2η προσέγγιση)

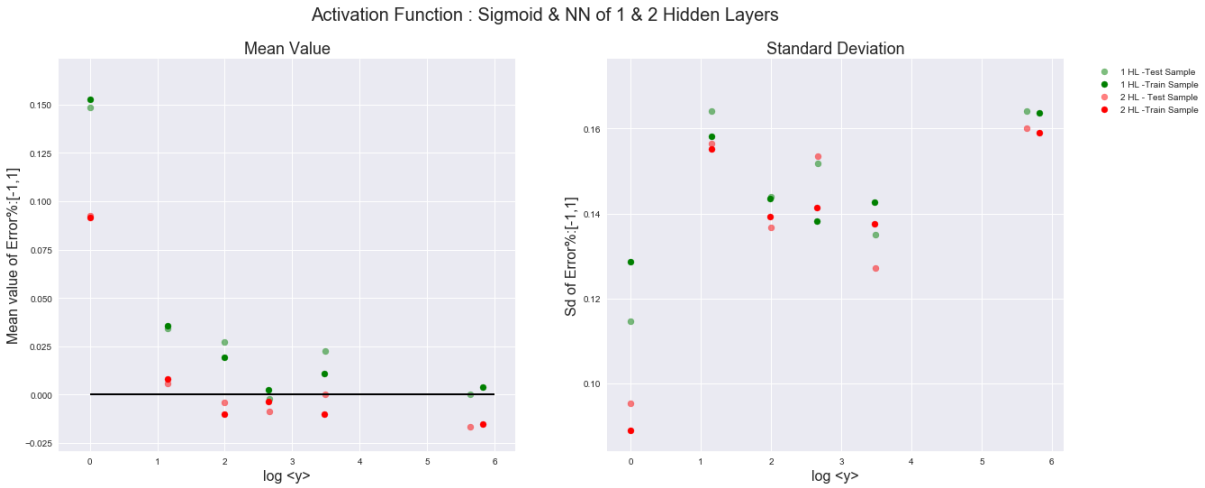
Tanh	***Δείγμα αξιολόγησης***		***Δείγμα εκπαίδευσης***	
Ρυθμός εκμάθησης	*Μέση τιμή*	*Τυπική απόκλιση*	*Μέση τιμή*	*Τυπική απόκλιση*
0.0001	*0.001*	*0.153*	*0.009*	*0.152*
0.001	*-0.048*	*0.142*	*-0.047*	*0.134*
0.01	*0.005*	*0.144*	*0.006*	*0.141*

Για την συνάρτηση ενεργοποίησης Relu, επιλέχθηκε ρυθμός εκμάθησης = 0.0001.

Τέλος, για κάθε συνάρτηση ενεργοποίησης δημιουργήθηκαν νευρωνικά δίκτυα με 1 και 2 κρυφά στρώματα 15 νευρώνων και με ρυθμό εκμάθησης εκείνον που επιλέχθηκε.

1. *Σιγμοειδής συνάρτηση ενεργοποίησης*

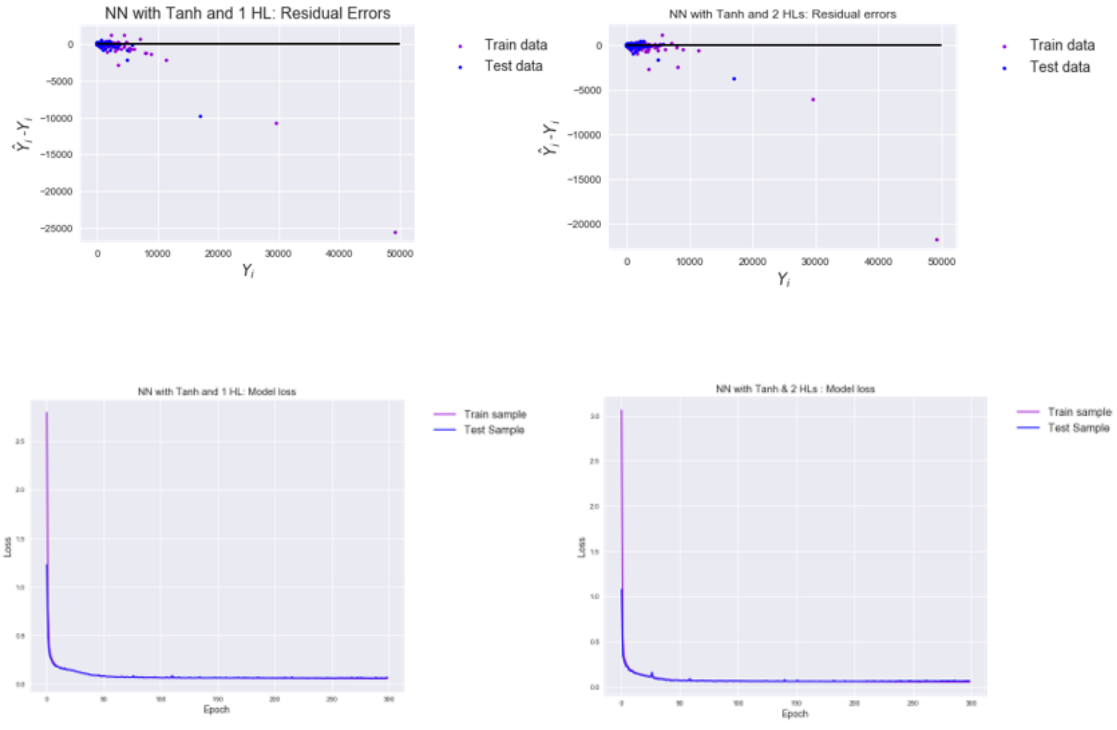


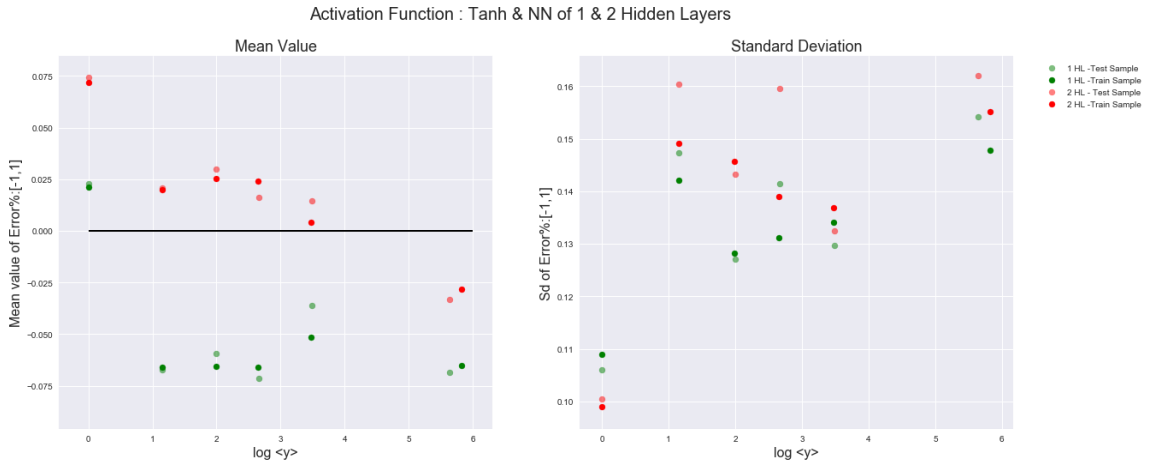


Διάγραμμα 4.48: Αποτελέσματα για σιγμοειδή (2η προσέγγιση)

Παρατηρείται ότι με την προσθήκη στρωμάτων, η απόδοση του μοντέλου βελτιώνεται ελάχιστα, για αυτό δεν συνεχίστηκε η προσθήκη επιπλέον στρώματος.

2. *Tanh συνάρτηση ενεργοποίησης*

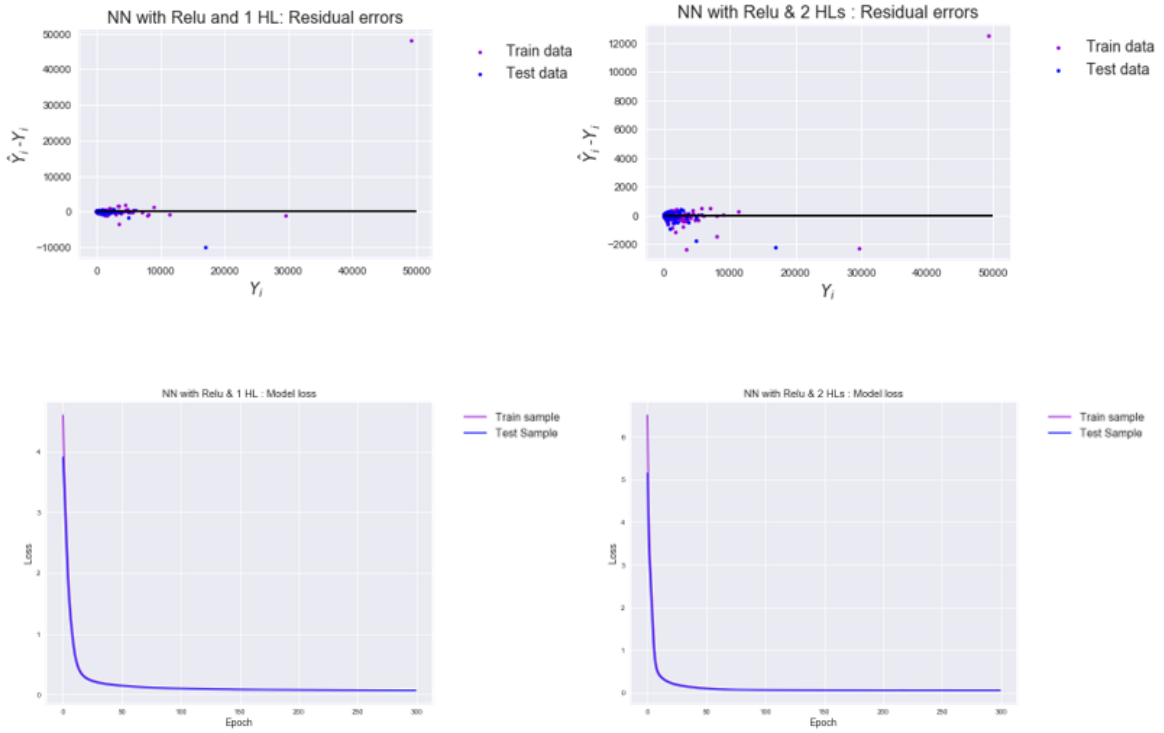


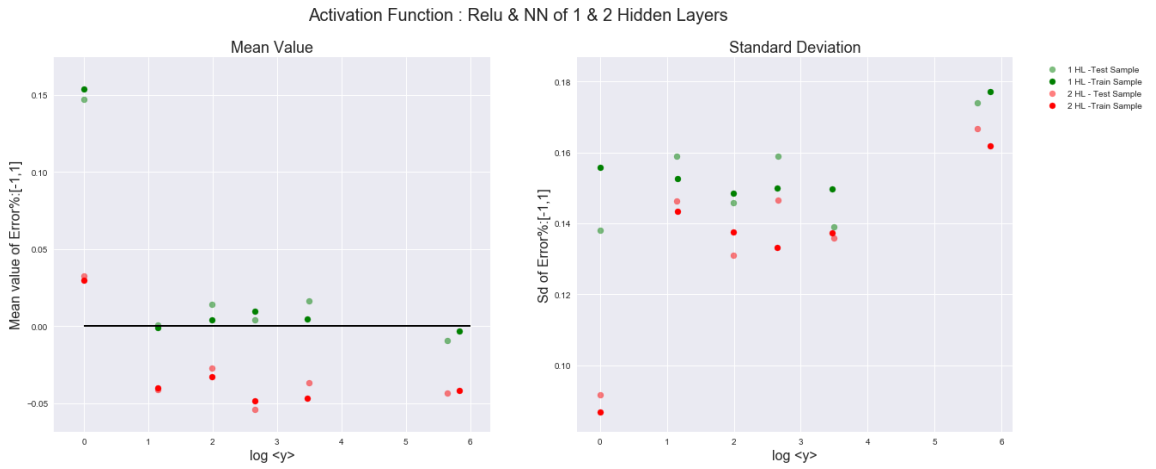


Διάγραμμα 4.49: Αποτελέσματα για την Tanh (2η προσέγγιση)

Τα ίδια συμπεράσματα προκύπτουν και για την tanh συνάρτηση ενεργοποίησης, κάτι το οποίο αναμέναμε διότι οι σιγμοειδής με την υπερβολική εφαπτομένη έχουν παρόμοια χαρακτηριστικά. Στην τελική σύγκριση, θα φανεί αν προκύπτουν καλύτερα αποτελέσματα με την υπερβολική εφαπτομένη.

3. *Relu συνάρτηση ενεργοποίησης*



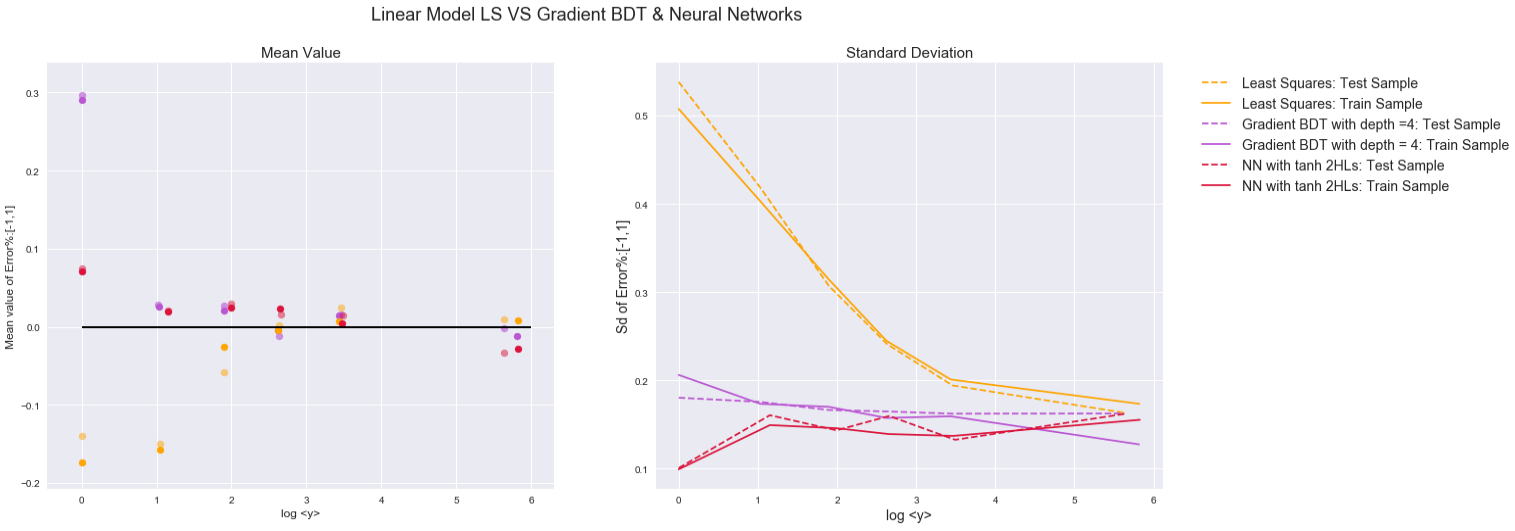


Διάγραμμα 4.50: Αποτελέσματα για την Tanh (2η προσέγγιση)

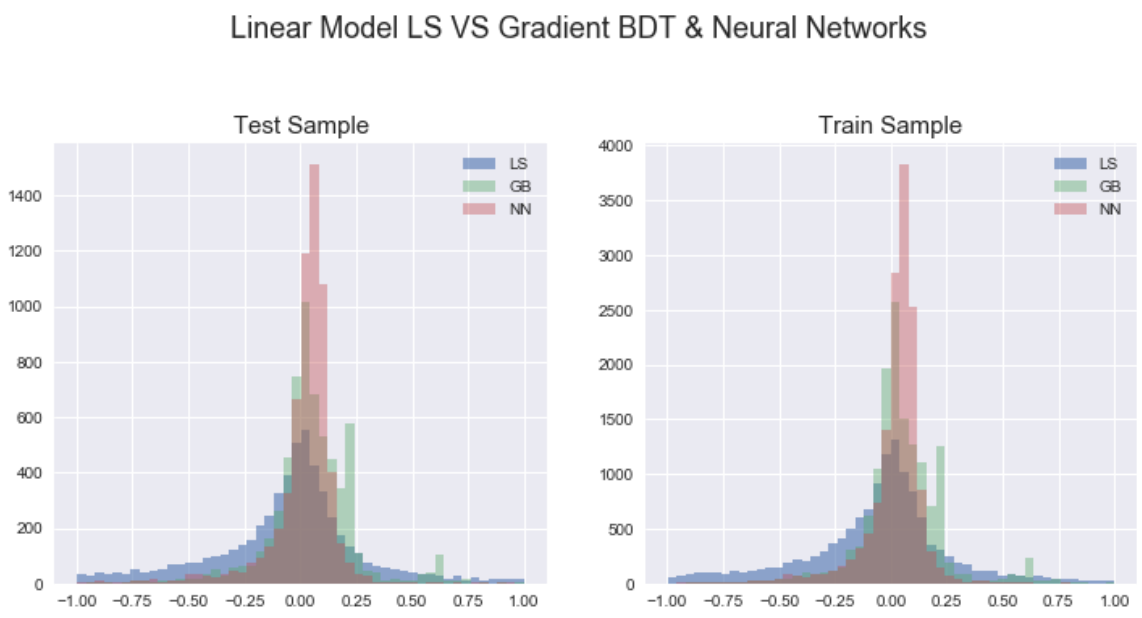
Από τα παραπάνω διαγράμματα, παρατηρείται ότι για την συνάρτηση ενεργοποίησης Relu προκύπτουν λίγο καλύτερα αποτελέσματα με την προσθήκη επιπλέον στρωμάτων.

4.4.3 Σύγκριση μη γραμμικού μοντέλου με NN , μη γραμμικού μοντέλου με Gradient και γραμμικού μοντέλου

Από τα παραπάνω αποτελέσματα, επιλέχθηκε το νευρωνικό δίκτυο με 2 κρυφά στρώματα και συνάρτηση ενεργοποίησης την υπερβολική εφαπτομένη, διότι είχε καλύτερη απόδοση.



Διάγραμμα 4.51: Σύγκριση μοντέλου με Gradient, γραμμικού μοντέλου και μοντέλου με NN



Διάγραμμα 4.52: Σύγκριση μοντέλου με Gradient, γραμμικού μοντέλου και μοντέλου με NN(Ιστογράμματα)

Παρατηρήσεις

I. Η μέση τιμή των ποσοστιαίων σφαλμάτων στο [-1,1] για κάθε διάστημα τιμών της πραγματικής τιμής της TS144(FB) έχει χαμηλότερη τιμή για το μοντέλο των Νευρωνικών Δικτύων. Το ίδιο συμβαίνει για την τυπική απόκλιση.

II. Σύμφωνα με τα ιστογράμματα, τα ποσοστιαία σφάλματα του μοντέλου των νευρωνικών δικτύων στο [-1,1] παρουσιάζουν την μεγαλύτερη συχνότητα μηδενικών τιμών.

Συμπέρασμα

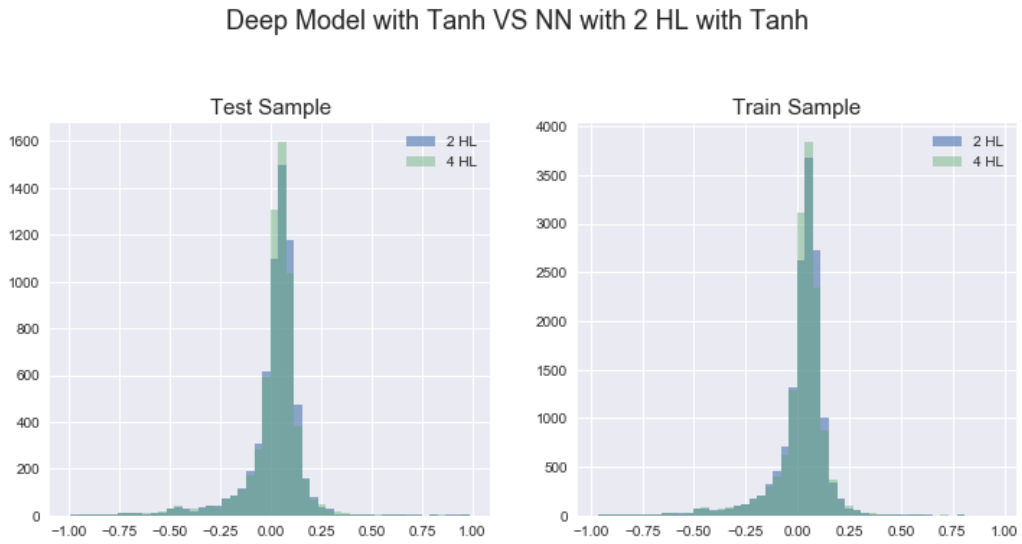
Το μοντέλο με την καλύτερη απόδοση είναι αυτό των Νευρωνικών Δικτύων, αμέσως καλύτερο το μοντέλο του δάσους δέντρων απόφασης και τέλος το γραμμικό μοντέλο.

4.4.4 Δημιουργία ‘deep’ μοντέλου

Τέλος, για το σύνολο των δεδομένων οικονομίας, κατασκευάστηκε ένα ‘deep’ μοντέλο, δηλαδή με αρκετά κρυφά στρώματα και πολλούς νευρώνες(πιο πολύπλοκη αρχιτεκτονική), για να εξεταστεί αν βελτιώνεται η απόδοση του μοντέλου. Το νευρωνικό δίκτυο που κατασκευάστηκε παρουσιάζεται παρακάτω:

Πίνακας 4.16: 'Deep' Μοντέλο

Στρώμα Εισόδου	30	**Συνάρτηση ενεργοποίησης**	Tanh
1ο Κρυφό στρώμα	60	**Συνάρτηση ενεργοποίησης**	Tanh
2ο Κρυφό στρώμα	50	**Συνάρτηση ενεργοποίησης**	Tanh
3ο Κρυφό στρώμα	40	**Συνάρτηση ενεργοποίησης**	Tanh
4ο Κρυφό στρώμα	30	**Συνάρτηση ενεργοποίησης**	Tanh
Στρώμα εξόδου	1	**Συνάρτηση ενεργοποίησης**	Γραμμική
‘Optimizer’	ADAM		
Συνάρτηση κόστους	MSE		
Αριθμός εποχών	300		
‘Batch size’	100		
Ρυθμός εκμάθησης	0.0001		



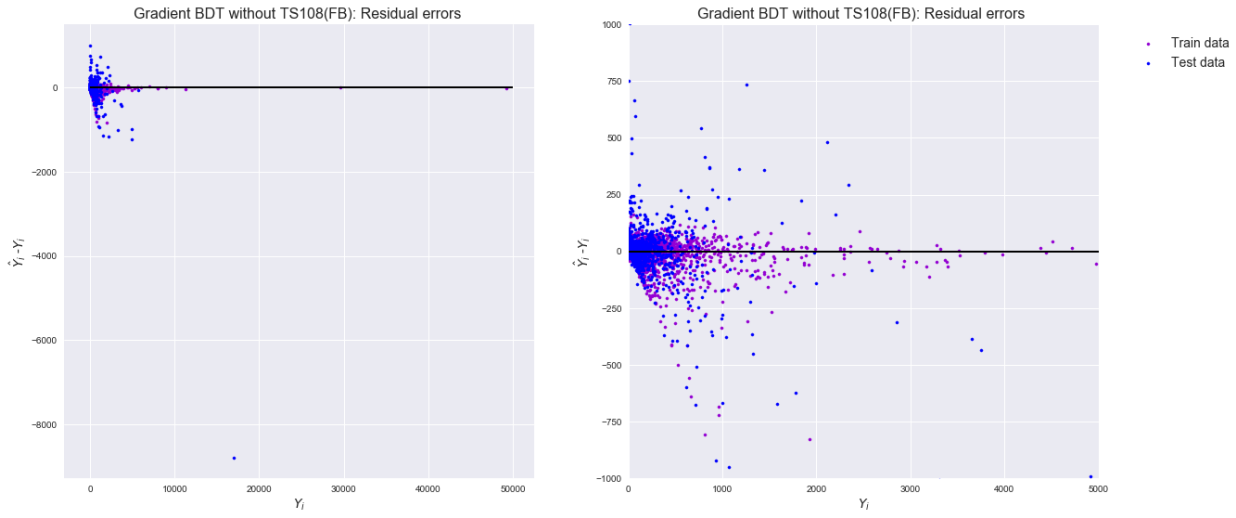
Διάγραμμα 4.53: Σύγκριση ‘deep’ μοντέλου με μοντέλο με Tanh(2ης προσέγγισης)

Παραπάνω, παρουσιάζονται τα ιστογράμματα των ποσοστιαίων σφαλμάτων στο [-1,1] για το νευρωνικό δίκτυο με την υπερβολική εφαπτομένη και τα 2 κρυφά στρώματα και για το νευρωνικό δίκτυο με την πολύπλοκη αρχιτεκτονική. Παρατηρείται ότι δεν βελτιώνεται κατά έναν μεγάλο βαθμό η απόδοση του μοντέλου με την πιο πολύπλοκη αρχιτεκτονική, διότι οι μηδενικές τιμές των ποσοστιαίων σφαλμάτων έχουν μεγάλη συχνότητα και για τα δύο νευρωνικά δίκτυα.

4.5 Αφαίρεση μεταβλητής TS108(FB)

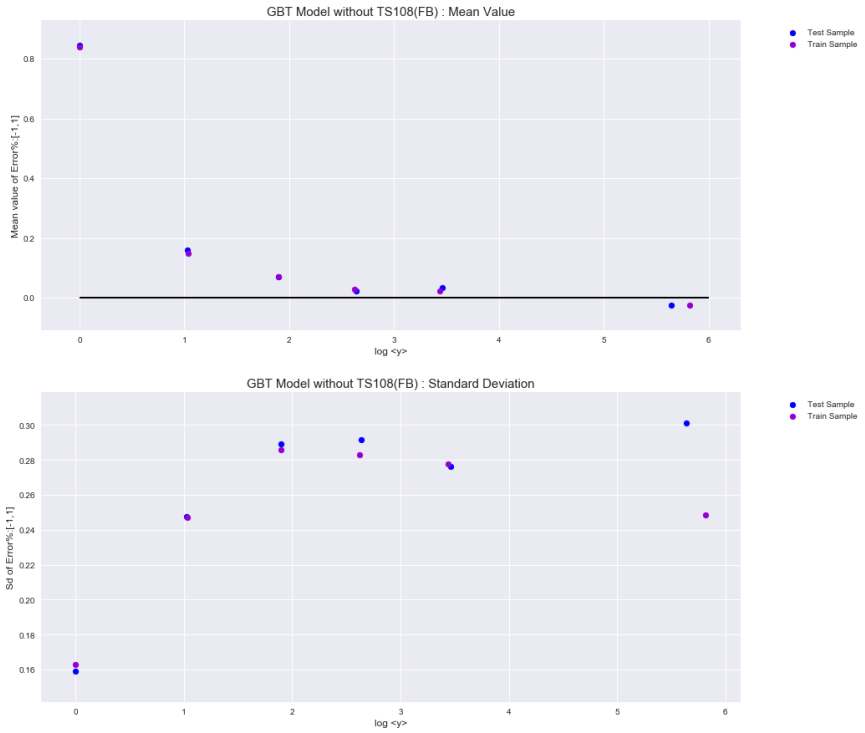
Η μεταβλητή TS108(FB) δείχνει την δημοτικότητα ενός άρθρου οικονομίας μετά από 36 ώρες, δηλαδή μετά από μιάμιση ημέρα ή διαφορετικά δείχνει την δημοτικότητα ενός άρθρου οικονομίας 12 ώρες πριν ολοκληρωθούν 2 ημέρες, δηλαδή η τιμή της μεταβλητής απόκρισης. Για αυτόν τον λόγο, η συγκεκριμένη τιμή επηρεάζει πολύ την τιμή της μεταβλητής απόκρισης μιας και αποτελεί μια από τις επεξηγηματικές μεταβλητές. Με την αφαίρεση της ελέγχεται αν η απόδοση του μοντέλου ελαττώνεται.

Κατασκευάστηκε ένα μοντέλο χωρίς αυτή την επεξηγηματική μεταβλητή, με την μέθοδο Gradient Boosting και τιμές παραμέτρων ίδιες με αυτές του καλύτερου μοντέλου του 4.2.3.(βάθος δέντρων: 4 , αριθμός εκτιμητών: 130 , ρυθμός εκμάθησης: 0.2).



Διάγραμμα 4.54: Εκτιμώμενα σφάλματα με την αφαίρεση μεταβλητής

RMSE $_{\text{Train Sample}}$ = 26.1
RMSE $_{\text{Test Sample}}$ = 124.9



Διάγραμμα 4.55: Μεση τιμή και Τυπική απόκλιση μοντέλου με την αφαίρεση μεταβλητής

Παρατηρήσεις

I. Η απόδοση του μοντέλου μειώνεται, διότι η απόκλιση μεταξύ της πραγματικής τιμής και της προβλεπόμενης είναι μεγαλύτερη για το δείγμα αξιολόγησης, σύμφωνα με το διάγραμμα 4.54.

II. Η ρίζα του μέσου τετραγωνικού σφάλματος του δείγματος εκπαίδευσης έχει σχετικά μεγάλη απόκλιση από την αντίστοιχη του δείγματος αξιολόγησης.

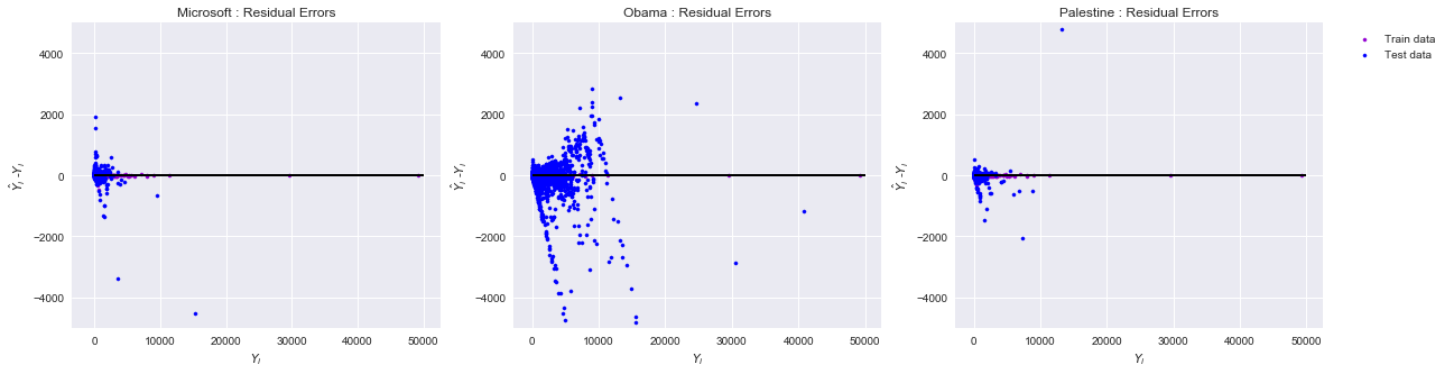
III. Η τυπική απόκλιση των ποσοστιαίων σφαλμάτων στο [-1,1] αυξάνεται όσο μεγαλώνει η τιμή της μέσης τιμής της πραγματικής τιμής.

Συνεπώς, η απόδοση του μοντέλου μειώνεται και συμπεραίνεται ότι η συγκεκριμένη μεταβλητή είναι σημαντική για το μοντέλο μας.

4.6 Επεξεργασία δεδομένων των τομέων Παλαιστίνης, Ομπάμα και Microsoft

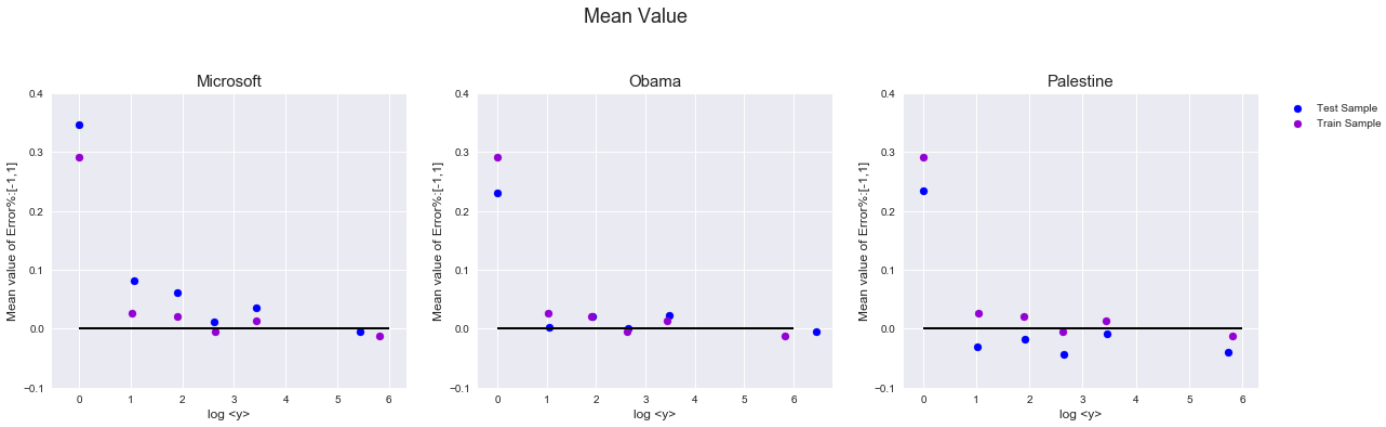
Για τους υπόλοιπους τομείς, Παλαιστίνη, Ομπάμα και Microsoft, εφαρμόστηκε το μοντέλο που εκπαιδεύθηκε στο σύνολο εκπαίδευσης του συνόλου δεδομένων της οικονομίας. Πιο συγκεκριμένα, εφαρμόστηκε το 'βέλτιστο' μοντέλο με την μέθοδο Gradient Boosting.

Τα παρακάτω διαγράμματα δείχνουν την απόκλιση από την πραγματική τιμή της μεταβλητής απόκρισης στο y άξονα συναρτήσει της πραγματικής τιμής στον x άξονα:



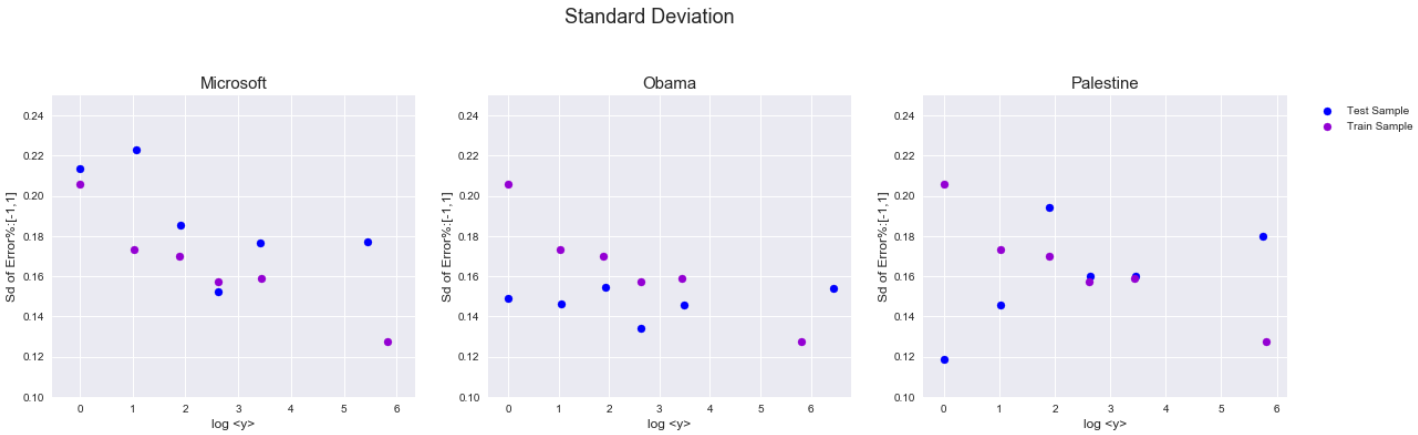
Διάγραμμα 4.56: Εκτιμώμενα σφάλματα για τους υπόλοιπους τομείς

Παρατηρείται ότι το μοντέλο που κατασκευάστηκε για τα δεδομένα της οικονομίας έχει σχετικά καλή απόδοση για τα δεδομένα της Παλαιστίνης και τα αντίστοιχα του Microsoft, ενώ για τα δεδομένα του Obama προκύπτουν πολύ μεγάλες αποκλίσεις για το σύνολο αξιολόγησης.



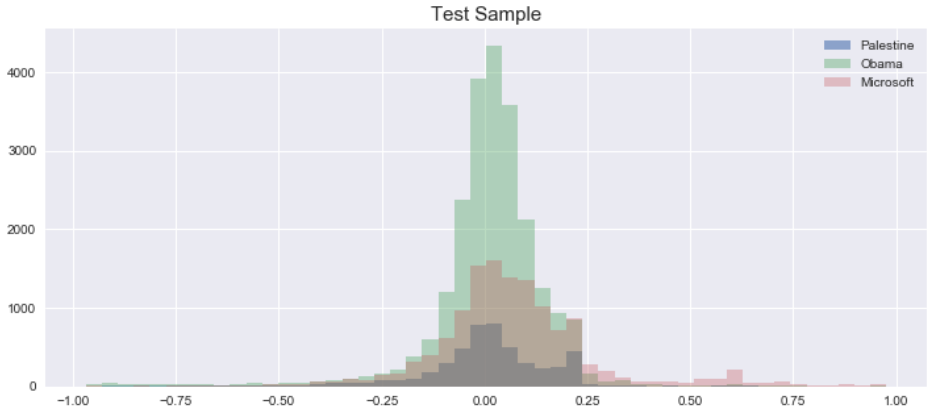
Διάγραμμα 4.57: Μέση τιμή για τους υπόλοιπους τομείς

Στην συνέχεια, κατασκευάστηκαν τα διαγράμματα της τυπικής απόκλισης και της μέσης τιμής των ποσοστιαίων σφαλμάτων στο διάστημα τιμών [-1,1].



Διάγραμμα 4.58: Τυπική απόκλιση για τους υπόλοιπους τομείς

Σύμφωνα με τα παραπάνω διαγράμματα, γίνεται αντιληπτό ότι η μέση τιμή των ποσοστιαίων σφαλμάτων στους υπόλοιπους 3 τομείς έχει παρόμοια αποτελέσματα. Ενώ, για τα διαγράμματα της τυπικής απόκλισης, δεν προκύπτουν τα ίδια συμπεράσματα. Για τα δεδομένα με θέμα τον Ομπάμα, η τυπική απόκλιση για το σύνολο εκπαίδευσης έχει μεγάλη απόκλιση με την αντίστοιχη του συνόλου αξιολόγησης και για τα άλλα 2 σύνολα δεδομένων, η τυπική απόκλιση έχει παρόμοια συμπεριφορά με σχετικά χαμηλές τιμές αλλά μεγάλες αποκλίσεις σε ορισμένα διαστήματα της πραγματικής τιμής y. Τέλος, κατασκευάστηκαν τα ιστογράμματα των ποσοστιαίων σφαλμάτων στο [-1,1].



Διάγραμμα 4.59: Ιστογράμματα για τους υπόλοιπους τομείς

Παρατηρείται ότι τα ποσοστιαία σφάλματα των παρατηρήσεων με θέμα τον Ομπάμα έχουν περισσότερες μηδενικές τιμές από τα αντίστοιχα των άλλων 2 θεμάτων. Όμως, στα δεδομένα με θέμα τον Ομπάμα υπάρχουν πολλές ακραίες τιμές, σύμφωνα με το διάγραμμα 4.56. Συνεπώς, περιορίζοντας τα ποσοστιαία σφάλματα στο διάστημα [-1,1], αυτές οι τιμές

δεν λαμβάνονται υπόψιν και κατά συνέπεια τα ιστογράμματα δεν μπορούν να βοηθήσουν στον έλεγχο απόδοσης των μοντέλων..

Συνοψίζοντας, το μοντέλο που κατασκευάστηκε με βάση τα δεδομένα της οικονομίας έχει μια σχετικά καλή απόδοση στα δεδομένα με θέμα την Παλαιστίνη και το Microsoft αλλά καθόλου καλή απόδοση για τα δεδομένα με θέμα τον Ομπάμα. Συνεπώς, το συγκεκριμένο μοντέλο δεν μπορεί να γενικευθεί για ειδήσεις με διαφορετικό θέμα.

Βιβλιογραφία

1. Torgo, Nuno Moniz and Luis. Multi-Source Social Feedback of Online News Feeds. 8 Μαρτιος 2018.

2. Aplaydin, Ethem. *Introduction to Machine Learning.* Λονδίνο : s.n., 2010.

3. Brownlee, Jason. *Overfitting and Underfitting With Machine Learning Algorithms.* 21 Μαρτιος 2016.

4. Yadav, Prince. *Decision Tree in Machine Learning.* 14 Νοεμβριος 2018.

5. Li, Lorraine. *Classification and Regression Analysis with Decision Trees.*

6. Nagpal, Anuja. *Decision Tree Ensembles- Bagging and Boosting.* 17 Οκτώβρης 2017.

7. Opitz, Richard Maclin and David. An Empirical Evaluation of Bagging and Boosting. 1997, σ. 6.

8. Solomatine, D. L. Shrestha & D. P. Experiments with AdaBoost.RT, an Improved Boosting. 2005.

9. *AdaBoost.RT: a Boosting Algorithm for Regression Problems.* Shrestha, D. P. Solomatine & D. L.

10. Brownlee, Jason. *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning.* 9 Σεπτέμβρης 2016.

11. Dhandhania, Keshav. *How to understand Gradient Descent, the most popular ML algorithm.* 18 Ιούνιος 2018.

12. *Stochastic gradient boosting.* Friedman, Jerome H. 2002.

13. Gurney, Kevin. *An introduction to neural networks.* Λονδινο : s.n., 2003/2004.

14. Mahanta, Jahnavi. *Introduction to Neural Networks, Advantages and Applications.* 10 Ιούλιος 2017.

15. Nielsen, Michael. *Neural Networks and Deep Learning.* 1995.

16. Yiu, Tony. *Understanding Neural Networks.*

17. Bishop, Christopher M. *Pattern Recognition and Machine Learning.* s.l. : Springer, 2006.

18. Kingma, Diederik P. ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.

Παράρτημα

Για την υλοποίηση των παραπάνω μοντέλων χρησιμοποιήθηκε η γλώσσα *'Python'*.

- Χρησιμοποιήθηκε η βιβλιοθήκη *Scikit-learn* για την κατασκευή των γραμμικών και μη γραμμικών μοντέλων.

Γραμμικό μοντέλο:	Linear Regression
Μοντέλο με Adaboost:	Adaboost Regressor
Μοντέλο με Gradient:	Gradient Regressor

- Χρησιμοποιήθηκε η βιβλιοθήκη *Keras* για την κατασκευή νευρωνικών δικτύων και πιο συγκεκριμένα η *Keras Regressor,* η οποία υλοποιεί την διασύνδεση με την *Scikit-learn.*

- Για την κατασκευή των διαγραμμάτων χρησιμοποιήθηκε η *Matplotlib, Plotly* και η *Seaborn.*

- Σημαντικές βιβλιοθήκες για την επεξεργασία δεδομένων είναι η *NumPy*, SciPy και *Pandas.*

- Οι βιβλιοθήκες Graphviz και Pydotplus χρησιμοποιήθηκαν για την οπτικοποίηση των δέντρων.