



Forecasting Natural Gas Spot Prices with Machine Learning

ΜΟΥΧΤΑΡΗΣ ΔΗΜΗΤΡΙΟΣ

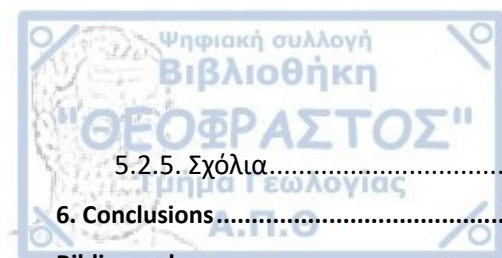
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Διατμηματικό Πρόγραμμα Μεταπτυχιακών
Σπουδών

«Έρευνα και Εκμετάλλευση Υδρογονανθράκων»



Περίληψη	1
Abstract	1
1. Εισαγωγή	2
2. Literature Review	3
3. Statistical Properties – Data	6
3.1. Assessing the Data	6
3.2. Analyzing the Dataset	9
3.2.1. Χρηματιστηριακοί Δείκτες	9
3.2.2. Exchange Rates (Συναλλαγματικές Ισοτιμίες)	10
3.2.3. West Texas Intermediate (WTI) Spot Prices	11
3.2.4. Interest Rates	11
3.2.5. Future Contracts	13
4. Methodology – Machine Learning Approaches	14
4.1. Support Vector Machines (SVM)	15
4.1.1. Μαθηματική προσέγγιση	16
4.1.2. Μέθοδοι Kernel	18
4.2. Gaussian Process Regression	19
4.3. Regression Trees	20
4.4. Ensemble of Trees	21
4.4.1. Bagging	22
4.4.2. Boosting	23
4.5. Linear Regression	24
4.6. Overfitting	25
5. Empirical Section	26
5.1. Empirical Study	26
5.1.1. Πρόβλεψη κατεύθυνσης της τιμής του Φυσικού Αερίου	26
5.1.2. Πρόβλεψη Τιμής Φυσικού Αερίου	28
5.2. Empirical Results	30
5.2.1. Χρονικός ορίζοντας t+1	30
5.2.2. Χρονικός ορίζοντας t+3	32
5.2.3. Χρονικός ορίζοντας t+5	34
5.2.4 Χρονικός ορίζοντας t+10	36



5.2.5. Σχόλια.....	38
6. Conclusions.....	39
Bibliography	41



Θα ήθελα να ευχαριστήσω τον κ. Γκόγκα Περικλή και τον κ. Σοφιανό Εμμανουήλ για την επίβλεψη, την ουσιαστική βοήθεια και την αδιάκοπη επιστημονική υποστήριξή τους καθ' όλη την διάρκεια εκπόνησης αυτής της διπλωματικής εργασίας. Η συμβολή και η καθοδήγησή τους ήταν αξιοσημείωτη για την ολοκλήρωση αυτής της ερευνητικής προσπάθειας. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου για την συνεχή υποστήριξη, υλική και μη, καθώς αποτέλεσαν καθοριστικό παράγοντα στην ολοκλήρωση των σπουδών μου.





Η δυνατότητα πρόγνωσης της τιμής του φυσικού αερίου έχει προφανή σημασία για τους παραγωγούς, καταναλωτές και επενδυτές στην σχετική αγορά του φυσικού αερίου. Στην παρούσα διπλωματική γίνεται προσπάθεια πρόγνωσης του natural gas spot price με την χρήση μεθόδων μηχανικής μάθησης και συγκεκριμένα Support Vector Machines (SVM), Regression Trees, Linear Regression, Gaussian Process Regression (GPR) και Ensemble of Trees. Τα μοντέλα εκπαιδεύτηκαν με τη χρήση ενός συνόλου 21 επεξηγηματικών μεταβλητών. Χρησιμοποιούμε την μέθοδο 5 fold-cross-validation με το 90% των δεδομένων για την εκπαίδευση των μοντέλων και το τελευταίο 10% των δεδομένων για τον έλεγχο σε άγνωστα δεδομένα. Τα αποτελέσματα δείχνουν ότι οι μέθοδοι μηχανικής μάθησης παρουσιάζουν διαφορές στην ακρίβεια πρόγνωσης των τιμών του φυσικού αερίου. Ωστόσο τα μοντέλα Bagged Trees (που ανήκει στην μέθοδο Ensemble of Trees) και Linear Regression παρουσιάζουν την καλύτερη απόδοση πρόγνωσης σε σύγκριση με τα υπόλοιπα μοντέλα.

Abstract

The ability to forecast the price of natural gas benefits stakeholders and has become a valuable tool for all market participants in competitive gas markets. In this paper an attempt is made to forecast the natural gas spot prices using machine learning methods and specifically with Support Vector Machines (SVM), Regression Trees, Linear Regression, Gaussian Process Regression (GPR) and Ensemble of Trees. These models are trained by using a set of 21 explanatory variables. We utilize the 5 fold-cross-validation method with 90% of the natural gas spot price data for model training and the remaining 10% of data for testing purposes of unknown data. The results show that these machine learning methods all have different forecasting accuracy when it comes to forecasting natural gas spot prices. However, the Bagged Trees (belonging to the Ensemble of Trees method) and Linear Regression models have an improved forecasting performance compared to the rest.

1. Εισαγωγή

Το φυσικό αέριο έχει προταθεί ως λύση για την αύξηση της ασφάλειας του ενεργειακού εφοδιασμού και την μείωση της περιβαλλοντικής ρύπανσης σε όλον τον κόσμο. Αποτελεί το δεύτερο πιο διαδεδομένο ενεργειακό εμπόρευμα, μετά το πετρέλαιο (Čeperić, Čeperić, & Žikonić, 2017). Με την αντικατάσταση του άνθρακα και την ευρεία διάδοση της χρήσης του φυσικού αερίου, η πρόγνωση της τιμής του φυσικού αερίου έχει γίνει ένα από τα πιο κρίσιμα ζητήματα σε πολλούς τομείς. Η ακριβής πρόγνωση των τιμών του φυσικού αερίου και η κατεύθυνση των αλλαγών των τιμών είναι υψίστης σημασίας, καθώς οι προβλέψεις αυτές χρησιμοποιούνται στην εμπορία προϊόντων, στον σχεδιασμό παραγωγής ηλεκτρικής ενέργειας και στην λήψη κανονιστικών αποφάσεων, καλύπτοντας τόσο την προσφορά όσο και την ζήτηση στην αγορά φυσικού αερίου.

Είναι κοινώς γνωστό ότι οι πληροφορίες είναι χρήσιμες εάν μπορούν να αξιοποιηθούν για την βελτίωση μιας διαδικασίας λήψης αποφάσεων. Στα οικονομικά, οι αποφάσεις βασίζονται συχνά σε προγνώσεις των ανοδικών ή καθοδικών κινήσεων της μεταβλητής στόχου. Έτσι το directional forecasting μπορεί να παρέχει ένα χρήσιμο πλαίσιο για την εκτίμηση της αξίας που παρουσιάζει η οικονομική πρόγνωση (Blaskowitz & Herwartz, 2011). Μια αναμενόμενη άνοδος της τιμής ενός εμπορεύματος θα αυξήσει την ζήτησή ενώ μία αναμενόμενη μείωση θα οδηγήσει σε πώληση.

Στην παρούσα διπλωματική εστιάζουμε στις βραχυπρόθεσμες προγνώσεις του Natural Gas Spot Price, προβλέποντας 1, 3, 5 και 10 μέρες μπροστά. Στόχος της διπλωματικής αυτής είναι η δημιουργία ενός αξιόπιστου μοντέλου πρόγνωσης της τιμής του natural gas spot prices με την χρήση μεθόδων μηχανικής μάθησης, καθώς και να εξετάσει τα πλεονεκτήματα και την αποτελεσματικότητα της μηχανικής μάθησης στην πρόγνωση τιμών του φυσικού αερίου. Το σύνολο των επεξηγηματικών μεταβλητών επιλέχθηκε με βάση την σχετική βιβλιογραφία και την ικανότητα που παρουσιάζουν στην ενίσχυση της προγνωστικής ικανότητας της τιμής του φυσικού αερίου και γενικά σχετίζονται με αυτό. Οι επιλεγμένες μεταβλητές τροφοδοτήθηκαν στη συνέχεια σε έναν σημαντικό αριθμό μοντέλων πρόγνωσης, μέσω μιας διαδικασίας μάθησης training-testing με αποτέλεσμα να προκύψουν τα πιο αποτελεσματικά και με το μικρότερο σφάλμα μοντέλα για την πρόγνωση του φυσικού αερίου.

Λόγω των σημαντικών οικονομικών αποτελεσμάτων που προκύπτουν από την πρόγνωση τιμών, έχουν διερευνηθεί και μελετηθεί πολλές τεχνικές (Ying , et al., 2010) (Aksoy & Selbas, 2019) ειδικά στις χρονοσειρές που σχετίζονται με την ηλεκτρική ενέργεια. Μερικές από αυτές είναι: τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks) (De Felice & Xin Yao, 2011), (Liu, et al., 1996), η προσέγγιση Fuzzy Logic (Abdel-Aal, 2004) και πολλές άλλες συμπεριλαμβανομένων των κλασσικών στατιστικών προσεγγίσεων όπως multiple linear regression και ARMA (Autoregressive Moving Average) (Park, El-Sharkawi, Marks, Atlas, & Damborg, 1991). Ωστόσο, δημοσιεύσεις στον χώρο της πρόγνωσης τιμών φυσικού αερίου και της αγοράς του είναι σχετικά σπάνιες.

Μία από τις λίγες μελέτες που επιδιώκουν να προβλέψουν τις κινήσεις των τιμών του φυσικού αερίου για την αγορά των ΗΠΑ, είναι αυτή των Buchanan et al. (Buchanan, Hodges, & Theis, 2001), όπου αναλύονται οι εμπορικές θέσεις (trader positions) που δημοσιεύονται σε εβδομαδιαία βάση. Οι Nguyen et al. (2010) κάνουν πρόγνωση στις τιμές του φυσικού αερίου, αλλά βασίζονται σε μηνιαία προθεσμιακά προϊόντα και συμβόλαια μελλοντικής εκπλήρωσης (monthly forward products – futures) αντί να επικεντρώνονται στις τρέχουσες τιμές. Χρησιμοποιώντας ένα multi-layer perceptron (MLP), μια ειδική μορφή ANN, παρατήρησαν πως τα μοντέλα με την χρήση GARCH υπερέχουν έναντι των MLP.

Μία άλλη μελέτη που αναλύει τις τιμές του φυσικού αερίου είναι αυτή των Salehnia et al. (2013). Δοκιμάζουν διάφορα μη γραμμικά μοντέλα: το Local Linear Regression (LLR), το Dynamic Local Linear Regression (DLLR) και τα Artificial Neural Networks (ANN). Χρησιμοποιούν ημερήσιες, εβδομαδιαίες και μηνιαίες spot τιμές του Henry Hub από το 1997 έως το 2012. Καταλήγουν στο συμπέρασμα ότι το MAPE (Mean Absolute Percentage Error) για το μοντέλο DLLR είναι χαμηλότερο σε σχέση με τα μοντέλα του ANN, αλλά η πρόγνωση τιμής αυτού του μοντέλου δεν είναι ιδιαίτερα εντυπωσιακή και αρκετά noisy¹. Οι Serletis & Shahmoradi (2006) αναλύουν τα χαρακτηριστικά μεταβλητότητας των μελλοντικών συμβολαίων (future contracts) φυσικού αερίου Henry Hub χρησιμοποιώντας μοντέλα GARCH. Πιστεύουν πως οι μεγάλες κεφαλαιακές απαιτήσεις και οι σημαντικοί χρόνοι παράδοσης που σχετίζονται με την παραγωγή και την

¹ Όταν συλλέγονται δεδομένα, υπάρχει πάντα κάποιο περιθώριο λάθους και ανακρίβειας, οπότε στα δεδομένα που συλλέγονται παρουσιάζεται κάποιο σφάλμα. Το σφάλμα αυτό αποκαλείται noise και δημιουργεί προβλήματα για τους αλγόριθμους μηχανικής μάθησης.

παράδοση ενέργειας, καθιστούν αυτές τις αγορές πολύ ευαίσθητες στις ανισορροπίες μεταξύ της ζήτησης και της προσφοράς, γεγονός που οδηγεί στην υψηλή αστάθεια των τιμών. Επιπλέον, θεωρούν πως η υψηλή μεταβλητότητα επηρεάζεται από τις καιρικές συνθήκες και τους περιορισμούς χωρητικότητας που μπορούν να παρουσιαστούν.

Σύμφωνα με τους Mishra & Smyth (2016), διαπιστώθηκε ότι οι μελλοντικές τιμές φυσικού αερίου (futures) δεν υπερτερούν στην πρόγνωση, έναντι ενός random walk μοντέλου. Διαπιστώνουν ότι οι τιμές spot και futures του φυσικού αερίου είναι προβλέψιμες μέσω της εφαρμογής unit root tests, τα οποία επιτρέπουν ταυτόχρονα την ύπαρξη ετεροσκεδαστικότητας και πολλαπλά structural breaks. Από την άλλη, τα ευρήματα στην προαναφερθείσα βιβλιογραφία παρέχουν μικτά αποτελέσματα κατά την εφαρμογή unit root tests για τον έλεγχο ύπαρξης της υπόθεσης αποτελεσματικής αγοράς. Για παράδειγμα η υπάρχουσα βιβλιογραφία ισχυρίζεται ότι οι τιμές του φυσικού αερίου είναι non-stationary. Ορισμένοι συγγραφείς υποστηρίζουν πως αυτό είναι συνέπεια της αναποτελεσματικής αντιμετώπισης της ετεροσκεδαστικότητας σε δεδομένα υψηλής συχνότητας. Κατά την τελευταία δεκαετία, οι προσεγγίσεις ANN συνδυαζόμενες με άλλες μεθόδους (συνήθως fuzzy ή εξελικτικές μεθόδους) χρησιμοποιούνται συχνά στις προβλέψεις τιμών και όγκων στις αγορές ενέργειας. Σαν μειονέκτημα ωστόσο, εμφανίζονται πολλά προβλήματα στον σχεδιασμό ενός συστήματος πρόγνωσης που να βασίζεται στα ANN για πρακτικούς σκοπούς. Εμφανίζονται δύο προβλήματα που σχετίζονται με τα ANN: 1) το “overfitting” (Το βασικό πρόβλημα στην εφαρμογή συστημάτων μηχανικής μάθησης είναι ότι ενώ από καθαρά πρακτική άποψη το να βρεθούν μοτίβα δεν είναι ιδιαίτερα δύσκολο, το να βρεθούν ωστόσο αυτά που λειτουργούν αξιόπιστα σε καταστάσεις της πραγματικής ζωής εμπεριέχουν αρκετά εμπόδια) – το πρόβλημα αυτό δεν παρουσιάζεται μόνο στα ANN αλλά και σε άλλες περιστάσεις και 2) η «κατάρρα των διαστάσεων» (προκαλείται από την αύξηση της πολυπλοκότητας που σχετίζεται με την προσθήκη επιπλέον διαστάσεων στο νευρωνικό δίκτυο). Υπό αυτές τις συνθήκες, παρόλο που ταιριάζουν (fitting) καλά τα προηγούμενα δεδομένα, οι προβλέψεις μπορεί να είναι μη ικανοποιητικές.

Τα SVR (Support Vector Regression machines) προτάθηκαν από τους Drucker et al. (1996) και τους Vapnik et al. (1996), ενώ χρησιμοποιήθηκαν από τους Chen et al. (2004) και Mohandes (2002) στην πρόγνωση για την αντιμετώπιση των παραπάνω προβλημάτων.

Η μηχανική μάθηση θεωρείται ένας τομέας με μια πληθώρα δυνατοτήτων και προοπτικών στην ακαδημαϊκή κοινότητα, ενώ ταυτόχρονα παρουσιάζει πολλές χρήσεις και στην πραγματική

επιχειρηματική εφαρμογή. Οι τεχνολογικοί γίγαντες όπως η Microsoft, η Google και το Facebook εξακολουθούν να θεωρούν πως ακόμη και τα πιο πολύτιμα και εξελιγμένα συστήματα μηχανικής μάθησης, αν και είναι χρήσιμα σε πολλές περιπτώσεις, παραμένουν ανελαστικά και δαπανηρά για τις εταιρείες και τους πελάτες τους (Čerperić, Čerperić, & Žikonić, 2017). Μερικά από τα προβλήματα που παρουσιάζονται με την χρήση της μηχανικής μάθησης είναι ότι αν επιτραπεί στον αλγόριθμο να αναζητήσει δεδομένα χωρίς ξεκάθαρους φραγμούς και όρια μπορεί να εντοπίσει μοτίβα δίχως νόημα που μπορεί να φαίνονται ισχυρά στο σετ εκπαίδευσης αλλά θα αποτύχει σε άγνωστα δεδομένα και συνεπώς στην πραγματική εφαρμογή. Επιπλέον όταν οι ερευνητές υπέρ-περιπλέκουν τα μοντέλα τους, συνήθως προσθέτοντας πολλές παραμέτρους προκειμένου να πάρουν τα αποτελέσματα που αναζητούν, ένα πρόβλημα υπέρ προσαρμογής προκύπτει (overfitting) οδηγώντας εκ νέου σε μια πολύ επιτυχημένη απόδοση στα δεδομένα εκπαίδευσης αλλά με απογοητευτικά αποτελέσματα σε άγνωστα δεδομένα (Čerperić, Čerperić, & Žikonić, 2017). Σύμφωνα με την ανασκόπηση της βιβλιογραφίας παρατηρείται ότι οι μεθοδολογίες της μηχανικής μάθησης παράγουν μεγαλύτερη ακρίβεια πρόγνωσης σε σύγκριση με τις τυπικές οικονομετρικές μεθόδους και τυπικά υπερτερούν έναντι του RW μοντέλο, ενώ οι οικονομετρικές προσεγγίσεις συχνά αποτυγχάνουν σε αυτό. Στην εργασία αυτή λοιπόν, προσπαθούμε να ανακαλύψουμε την πιθανή σχέση μεταξύ των τιμών του φυσικού αερίου και άλλων οικονομικών μεταβλητών χρησιμοποιώντας μεθοδολογίες μηχανικής μάθησης με ημερήσια δεδομένα, καθώς και να εντοπίσουμε μοντέλα που έχουν την δυνατότητα να προβλέπουν με επιτυχία τις τιμές του φυσικού αερίου. Τα μοντέλα που εκπαιδεύτηκαν ήταν τα Linear Regression, Interactions Linear, Robust Linear, Fine Tree, Medium Tree, Coarse Tree, Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM, Boosted Trees, Bagged Trees, Squared Exponential, Matern 5/2 GPR, Exponential GPR και Rational Quadratic.

3. Statistical Properties – Data

3.1. Assessing the Data

Για την εκπαίδευση και τον έλεγχο των μοντέλων συγκεντρώθηκε ένα dataset που αποτελείται από 2423 ημερήσιες τιμές του Natural Gas Spot Price από την βάση δεδομένων του Energy Information Administration και 21 σχετικές οικονομικές μεταβλητές από την βάση δεδομένων της Federal Reserve Bank of Saint Louis και του Yahoo Finance για την περίοδο 3 Δεκεμβρίου 2010 έως 18 Σεπτεμβρίου 2020. Επιπλέον υπολογίστηκε και προστέθηκε στο dataset το momentum των τελευταίων 5 και 10 ημερών, καθώς και το moving average 5 και 10 ημερών. Με εξαίρεση τα interest rates, όλες οι μεταβλητές μετατράπηκαν σε φυσικούς λογάριθμους. Στον Πίνακα 1 συγκεντρώνονται κάποιες βασικές στατιστικές ιδιότητες των μεταβλητών αυτών.

Πίνακας 1: Ιδιότητες Δευτερευόντων Μεταβλητών (Regressors).

#	Name	Mean	Standard Deviation	Skewness	Kurtosis	Variance
Panel A: <u>Stock Indices</u>						
1	NASDAQ Composite Index	5289.44	2132.48	0.61	-0.4	4549055
2	S&P 500 Index	2112.97	607.77	0.18	-0.92	369500
3	Dow Jones Industrial Average Index	18797.36	5195.66	0.3	-1.04	27003372
Panel B: <u>Exchange Rates</u>						

4	USD/EUR	0.19	0.09	0.3	-1.29	0.008
5	JPY/USD	4.62	0.14	-0.82	-0.61	0.019
6	USD/GBP	0.37	0.1	0.19	-1.5	0.010
Panel C: <u>WTI Spot Price</u>						
7	Cushing, OK WTI Spot Price FOB	4.17	0.37	-0.57	0.34	0.1401
Panel D: <u>Interest Rates</u>						
8	Effective Federal Funds Rate	0.638	0.77	1.17	-0.15	0.5974
9	5-Year Breakeven Inflation Rate	1.7	0.32	-0.84	1.57	0.107
10	10-Year Breakeven Inflation Rate	1.94	0.33	-0.44	0.36	0.1142
11	1-Year Treasury Constant Maturity Rate	0.75	0.82	1.1	-0.23	0.6757
12	10-Year Treasury Constant	2.22	0.61	-0.4	0.51	0.3736

	Maturity Rate					
13	Bank Prime Loan Rate	3.76	0.75	1.18	-0.12	0.5732
Panel E: <u>Future Contracts</u>						
14	Natural Gas Futures Contract 1	1.1	0.26	-0.171	-0.51	0.0686
15	Natural Gas Futures Contract 2	1.11	0.25	-0.174	-0.62	0.0626
16	Natural Gas Futures Contract 3	1.13	0.24	-0.163	-0.69	0.0573
17	Natural Gas Futures Contract 4	1.15	0.23	-0.096	-0.75	0.0519
18	OK Crude Oil Future Contract 1	4.181	0.371	-0.53	0.15	0.138
19	OK Crude Oil Future Contract 2	4.189	0.358	-0.37	-0.45	0.1283
20	OK Crude Oil Future Contract 3	4.195	0.348	-0.27	-0.81	0.1213

21	OK Crude Oil Future Contract 4	4.199	0.341	-0.22	-0.95	0.1165
----	--------------------------------	-------	-------	-------	-------	--------

3.2. Analyzing the Dataset

3.2.1. Χρηματιστηριακοί Δείκτες

Χρησιμοποιήθηκαν τρεις χρηματιστηριακοί δείκτες (Panel A) για να εξετάσουμε την πιθανή σχέση μεταξύ της χρηματιστηριακής αγοράς και αυτής του φυσικού αερίου. Οι δείκτες αυτοί είναι οι Dow Jones Industrial Average Index, S&P 500 Index και NASDAQ Composite Index.

3.2.1.1. Dow Jones Industrial Average Index

Ο χρηματιστηριακός δείκτης Dow Jones Industrial Average (DJIA), γνωστός επίσης και ως Dow 30, παρακολουθεί 30 μεγάλες δημόσιες blue chip εταιρείες με αξιοσημείωτα σταθερά κέρδη, που δραστηριοποιούνται στο Χρηματιστήριο της Νέας Υόρκης (NYSE). Ορισμένες από αυτές τις εταιρείες περιλαμβάνουν την Walt Disney Company, την Exxon Mobil Corporation και την Microsoft Corporation. Ο DJIA είναι ο δεύτερος παλαιότερος δείκτης αγοράς των ΗΠΑ και ένας από τους πιο σημαντικούς δείκτες του κόσμου και χρησιμοποιείται ως βάση αναφοράς της συνολικής αγοράς, ενώ παλαιότερος θεωρείται ο Dow Jones Transportation Average. Ο DJIA σχεδιάστηκε προκειμένου να χρησιμεύσει ως πληρεξούσιο για την υγεία και σταθερότητα της ευρύτερης οικονομίας των ΗΠΑ.

3.2.1.2. S&P 500 Index

Ο S&P 500 είναι ένας χρηματιστηριακός δείκτης που παρακολουθεί τις μετοχές 500 μεγαλοκεφαλαιούχων (large-cap) εταιρειών των ΗΠΑ. Αντιπροσωπεύει την απόδοση του χρηματιστηρίου αναφέροντας τα ρίσκα και τις αποδόσεις των μεγαλύτερων εταιριών. Οι επενδυτές τον χρησιμοποιούν ως σημείο αναφοράς της συνολικής αγοράς, με την οποία συγκρίνονται οι υπόλοιπες επενδύσεις. Ο S&P έχει περισσότερες large-cap μετοχές από τον Dow Jones Industrial Average. Επιπλέον έχει λιγότερες μετοχές που σχετίζονται με την τεχνολογία σε σύγκριση με την μετοχή της Nasdaq. Συγκεκριμένα τον Απρίλιο του 2020, το 57% των πιστώσεων (allocations) Nasdaq αφορούσαν την τεχνολογία πληροφοριών σε σύγκριση με το 23% για το S&P 500 εκείνη την χρονική περίοδο. Ο δείκτης αυτός σταθμίζεται από ένα ανώτατο float-adjusted όριο αγοράς. Μετρά δηλαδή μόνο τις μετοχές που είναι διαθέσιμες στο δημόσιο. Δεν ασχολείται

με μετοχές που κατέχονται από ομάδες ελέγχου, άλλες εταιρείες ή κυβερνητικές υπηρεσίες (Amadeo, 2020).

3.2.1.3. Nasdaq Composite Index

Ο Nasdaq Composite Index είναι ένας μεγάλος χρηματιστηριακός δείκτης σταθμισμένου κεφαλαίου αγοράς με περισσότερες από 2500 μετοχές, αμερικανικές αποθήκες καταθέσεων (depository receipts, ADRs) και επενδυτικά καταπιστεύματα ακινήτων (real estate investment trusts REITs), μεταξύ άλλων. Η σύνθεση του δείκτη αποτελείται σχεδόν κατά 50% από τεχνολογικές μετοχές, καταναλωτικές υπηρεσίες, υγειονομική περίθαλψη και χρηματοδοτήσεις των ανερχόμενων σημαντικών βιομηχανιών. Ο δείκτης αυτός δεν περιορίζεται μόνο σε εταιρείες που έχουν έδρα στις ΗΠΑ – κάτι που τον ξεχωρίζει από τους περισσότερους δείκτες. Περιλαμβάνει τέλος όλες τις μετοχές της Nasdaq που δεν είναι παράγωγα, προνομιούχες μετοχές (preferred shares), κεφάλαια, χρηματιστηριακά κεφάλαια (exchange-traded funds ETFs) ή χρεόγραφα (debenture securities). Αξίζει να αναφερθεί πως παρά τις διαφορές που παρουσιάζουν οι 3 αυτοί δείκτες, τείνουν να κινούνται μαζί (Chen J. , Nasdaq Composite Index, 2019).

3.2.2. Exchange Rates (Συναλλαγματικές Ισοτιμίες)

Δεδομένου ότι το πετρέλαιο εμπορεύεται σε δολάρια ΗΠΑ (USD) και του γεγονότος ότι η συναλλαγματική ισοτιμία EUR/USD χρησιμοποιείται για το 28% του παγκόσμιου όγκου ημερήσιας συναλλαγματικής ισοτιμίας (Dimitriadou, Gogas, Papadimitriou, & Plakandaras, 2018), επιλέξαμε επίσης αυτήν την συναλλαγματική ισοτιμία μαζί με τις ισοτιμίες JPY/USD (Japanese Yen/United States Dollars Foreign Exchange Rate) και USD/GBP (United States Dollars/Great Britain Pound Foreign Exchange Rate), καθώς οι δύο αυτές είναι οι αμέσως πιο διαδεδομένες συναλλαγματικές ισοτιμίες μετά από αυτή του EUR/USD (Panel B).

Ως συναλλαγματική ισοτιμία ορίζουμε την αξία του νομίσματος ενός έθνους έναντι του νομίσματος ενός άλλου έθνους ή οικονομικής ζώνης. Οι περισσότερες συναλλαγματικές ισοτιμίες είναι free-floating και θα αυξηθούν ή θα μειωθούν με βάση την προσφορά και την ζήτηση της αγοράς. Ορισμένα νομίσματα δεν είναι free-floating και έχουν περιορισμούς. Μια free-floating συναλλαγματική ισοτιμία αυξάνεται και πέφτει λόγω αλλαγών στην αγορά ξένου συναλλάγματος (Chen J. , Exchange Rate Definition, 2020). Οι συναλλαγματικές ισοτιμίες που χρησιμοποιήθηκαν είναι οι EUR/USD, GBP/USD και USD/JPY.

3.2.3. West Texas Intermediate (WTI) Spot Prices

Επιπλέον, ακολουθώντας τη βιβλιογραφία που αποδίδει τις διακυμάνσεις των τιμών του φυσικού αερίου σε ακραίες καταστάσεις προσφοράς και ζήτησης, συνυπολογίσαμε και τις spot prices του Cushing, OK WTI Spot Price FOB (Dollars per Barrel) (Panel C). Το West Texas Intermediate είναι ένα αργό πετρέλαιο που χρησιμεύει ως ένα από τα κύρια παγκόσμια σημεία αναφοράς του πετρελαίου. Προέρχεται κυρίως από την ενδοχώρα του Τέξας και είναι ένα από τα πετρέλαια υψηλότερης ποιότητας στον κόσμο, το οποίο είναι ιδιαίτερα εύκολο να διωλιστεί. Το WTI συχνά συγκρίνεται με το αργό πετρέλαιο Brent, το οποίο αποτελεί και αυτό σημείο αναφοράς για τα δύο τρίτα των παγκόσμιων συμβάσεων πετρέλαιο που βασίζονται στο πετρέλαιο που εξάγεται στην Βόρεια Θάλασσα. Το WTI Crude Oil Spot Price είναι η τιμή για την άμεση παράδοση του πετρελαίου West Texas Intermediate, επίσης γνωστό και ως Texas Light sweet. Το WTI ειδικότερα είναι χρήσιμο για την τιμολόγηση οποιουδήποτε προϊόντος πετρελαίου στην Αμερική (Chen J. , West Texas Intermediate (WTI) , 2020).

3.2.4. Interest Rates

Συλλέχτηκαν επίσης τα κύρια interest rates που χρησιμοποιήθηκαν ως σημεία αναφοράς για τις ΗΠΑ και την Ευρωπαϊκή οικονομία συναλλαγών, συγκεκριμένα τα Effective Federal Funds Rate, 5-Year Breakeven Inflation Rate, 10-Year Breakeven Inflation Rate, 1-Year Treasury Constant Maturity Rate, 10-Year Treasury Constant Maturity Rate και Bank Prime Loan Rate (Panel D).

3.2.4.1. Effective Federal Funds Rate

Το ενεργό επιτόκιο ομοσπονδιακών κεφαλαίων (Effective Federal Funds Rate), είναι το επιτόκιο που οι τράπεζες χρεώνουν η μία την άλλη για δάνεια μίας ημέρας για να καλύψουν τις απαιτήσεις του αποθεματικού τους. Καθορίζεται από την Ομοσπονδιακή Επιτροπή Ανοικτής Αγοράς (Federal Open Market Committee FOMC) και αποτελεί το πιο ισχυρό και με την μεγαλύτερη επιρροή επιτόκιο στην οικονομία ενός έθνους, καθώς επηρεάζει την απασχόληση, την ανάπτυξη και τον πληθωρισμό. Η FOMC ορίζει ένα επιτόκιο ομοσπονδιακών κεφαλαίων οκτώ φορές τον χρόνο, με βάση τις επικρατούσες οικονομικές συνθήκες. Το επιτόκιο αυτό μπορεί να επηρεάσει τα

βραχυπρόθεσμα επιτόκια στα καταναλωτικά δάνεια και στις πιστωτικές κάρτες, καθώς και να επηρεάσει το χρηματιστήριο (Chen J. , Federal Funds Rate, 2020).

3.2.4.2. 5-Year Breakeven Inflation Rate

Ο συγκεκριμένος ρυθμός πληθωριστικού συντελεστή (Breakeven Inflation Rate) αντιπροσωπεύει ένα μέτρο του αναμενόμενου πληθωρισμού που προέρχεται από 5ετες Χρεόγραφα σταθερής διάρκειας του δημοσίου (5-Year Treasury Constant Maturity Securities) και από 5ετές χρεόγραφα σταθερής διάρκειας με δείκτη πληθωρισμού (5-Year Treasury Inflation-Indexed Constant Maturity Securities). Η τελευταία μεταβλητή υποδηλώνει το πόσο αναμένουν, οι συμμετέχοντες στην αγορά, να είναι ο πληθωρισμός την επόμενη 5ετία, κατά μέσο όρο (Interest Rate Spreads, 2021).

3.2.4.3. 10-Year Breakeven Inflation Rate

Για αυτόν τον ρυθμό πληθωριστικού συντελεστή ισχύουν τα ίδια που ισχύουν και στον 5ετή, απλώς με προσαρμοσμένα δεδομένα για περίοδο 10 ετών.

3.2.4.4. 1-Year Treasury Constant Maturity Rate

Ο δείκτης αυτός δημοσιεύεται από το Federal Reserve Board με βάση την μέση απόδοση μιας σειράς χρεόγραφων (Treasury securities) του Δημοσίου, έχοντας ταυτόχρονα όλα προσαρμοσμένα στο ισοδύναμο της διάρκειας ενός έτους. Το 1-year constant maturity Treasury (CMT) αντιπροσωπεύει την απόδοση των πιο πρόσφατα δημοπρατημένων χρεόγραφων του Δημοσίου, ενός χρόνου. Η μηνιαία τιμή CMT ενός χρόνου είναι ένας δημοφιλής δείκτης υποθηκών με τον οποίο συνδέονται πολλές υποθήκες ρυθμιζόμενου επιτοκίου (adjustable-rate mortgages ARMs). Καθώς αλλάζουν οι οικονομικές συνθήκες, οι δανειστές χρησιμοποιούν αυτόν τον δείκτη (ο οποίος ποικίλλει) για να προσαρμόσουν τα επιτόκια προσθέτοντας έναν συγκεκριμένο αριθμό εκατοστιαίων μονάδων που ονομάζεται margin (το οποίο δεν ποικίλλει) στον δείκτη για τον καθορισμό του επιτοκίου που πρέπει να πληρώσει ένας δανειολήπτης. Όταν ο δείκτης αυτός ανεβαίνει, τότε τα επιτόκια των δανείων που συνδέονται με αυτό αυξάνονται επίσης (Chen J. , One-Year Constant Maturity Treasury, 2020).

3.2.4.5 10-Year Treasury Constant Maturity Rate

Για τον 10-ετή αυτόν δείκτη ισχύουν τα ίδια που ισχύουν και για το 1-Year Treasury Constant Maturity Rate, απλώς με προσαρμοσμένα όλα τα δεδομένα για περίοδο 10 ετών.

Το Bank Prime Loan Rate (επιτόκιο τραπεζικού δανεισμού) είναι το επιτόκιο που χρεώνουν οι εμπορικές τράπεζες τους πιο αξιόπιστους πελάτες του, κυρίως μεγάλες εταιρείες. Το επιτόκιο τραπεζικού δανεισμού καθορίζεται σε μεγάλο βαθμό από το επιτόκιο ομοσπονδιακών κεφαλαίων, το οποίο είναι το επιτόκιο μιας ημέρας που οι τράπεζες χρησιμοποιούν για να δανείζουν η μια στην άλλη. Αποτελεί επιπλέον την βάση ή το σημείο εκκίνησης για τα περισσότερα επιτόκια, συμπεριλαμβανομένων των επιτοκίων για υποθήκες, για δάνεια μικρών επιχειρήσεων ή για προσωπικά δάνεια (παρόλο που το επιτόκιο αυτό ενδέχεται να μην αναφέρεται συγκεκριμένα ως συστατικό του επιτοκίου που χρεώνεται τελικά). Ο κίνδυνος μη εκπλήρωσης υποχρεώσεων είναι ο κύριος καθοριστής του επιτοκίου που χρεώνει μια τράπεζα σε έναν δανειολήπτη. Επειδή οι καλύτεροι πελάτες μιας τράπεζας έχουν ελάχιστες πιθανότητες αθέτησης υποχρεώσεων, η τράπεζα μπορεί να τους χρεώσει ένα ποσοστό χαμηλότερο από το επιτόκιο που χρεώνει σε έναν πελάτη που έχει μεγαλύτερη πιθανότητα μη εκπλήρωσης υποχρεώσεων (Chen J. , Prime Rate, 2020).

3.2.5. Future Contracts

Τέλος συλλέχθηκαν Future Premiums (Μελλοντικά Ασφάλιστρα) προκειμένου να εξετάσουμε την ικανότητά τους να προβλέπουν spot prices μέσω της αξιολόγησης 8 Future Contracts (Panel E).

3.2.5.1 Natural Gas Futures Contracts

Ένα συμβόλαιο μελλοντικής εκπλήρωσης φυσικού αερίου, όπως όλα τα εμπορεύματα, είναι μια σύμβαση που υποχρεώνει τον αγοραστή να αγοράσει μια συγκεκριμένη ποσότητα φυσικού αερίου σε προκαθορισμένη τιμή σε μια μελλοντική ημερομηνία. Οι ημερομηνίες παράδοσης ορίζονται γύρω στην 15^η ημέρα του επόμενου μήνα. Τα συμβόλαια αυτά επιτρέπουν στους συμμετέχοντες στην αγορά σημαντική δραστηριότητα αντιστάθμισης προκειμένου να γίνει διαχειρίσιμο το ρίσκο που παρουσιάζεται στην εξαιρετικά ευμετάβλητη τιμή του φυσικού αερίου, η οποία κατευθύνεται από την ζήτηση που σχετίζεται με τον καιρό (weather related demand). Συνοψίζοντας, τα μελλοντικά συμβόλαια φυσικού αερίου αποτελούν το τρίτο μεγαλύτερο συμβόλαιο μελλοντικής εκπλήρωσης φυσικών εμπορευμάτων στον κόσμο κατ' όγκο. Χρησιμοποιείται ευρέως ως εθνική τιμή αναφοράς για το φυσικό αέριο, το οποίο συνεχίζει να αυξάνεται ως παγκόσμια πηγή

ενέργειας. Τέλος συνιστούν ένα ανεξάρτητο και αυτόνομο εμπόρευμα (NYMEX: Your Home for Henry Hub Natural Gas, n.d.).

3.2.5.2. Cushing, OK Crude Oil Future Contracts

Το Cushing, Oklahoma είναι η τοποθεσία παράδοσης για το συμβόλαιο μελλοντικής εκπλήρωσης της NYMEX Light Sweet Crude Oil. Πιο συγκεκριμένα τα μελλοντικά συμβόλαια αργού πετρελαίου είναι συμβόλαια μελλοντικής εκπλήρωσης στα οποία οι αγοραστές και οι πωλητές πετρελαίου συντονίζονται και συμφωνούν να παραδώσουν συγκεκριμένες ποσότητες αργού πετρελαίου σε μια δεδομένη και προκαθορισμένη ημερομηνία και τιμή στο μέλλον. Τα μελλοντικά συμβόλαια που αποτελούν σημείο αναφοράς για το αργό πετρέλαιο στις ΗΠΑ περιλαμβάνει το West Texas Intermediate (WTI), ένα πετρέλαιο συγκεκριμένου βαθμού που έχει αρκετά χαμηλή πυκνότητα και περιεκτικότητα σε θείο που το καθιστά σχετικά εύκολο να διυλιστεί. Ιστορικά, έχει εμπορευτεί στο New York Mercantile Exchange και ως εκ τούτου πολλοί έμποροι αναφέρονται στα συμβόλαια αυτά ως συμβόλαια μελλοντικής εκπλήρωσης αργού πετρελαίου NYMEX WTI. Οι προδιαγραφές για τα συμβόλαια αυτά καθορίζονται κατά τέτοιο τρόπο ώστε να επιτρέπεται στους συμμετέχοντες στην αγορά να τα εμπορεύονται ομοιόμορφα. Κάθε συμβόλαιο καλύπτει 1000 βαρέλια και οι ημερομηνίες παράδοσης είναι διαθέσιμες έως και εννέα χρόνια στο μέλλον (Caplinger, 2016).

4. Methodology – Machine Learning Approaches

Η μηχανική μάθηση είναι μια μέθοδος ανάλυσης δεδομένων που αυτοματοποιεί την δημιουργία αναλυτικών μοντέλων. Αποτελεί έναν από τους πιο ταχεία αναπτυσσόμενους τεχνικούς τομείς σήμερα, ο οποίος βρίσκεται στην διασταύρωση της επιστήμης των υπολογιστών (computer science) και της στατιστικής, καθώς και στον πυρήνα της τεχνητής νοημοσύνης και της επιστήμης δεδομένων (data science). Η πρόσφατη πρόοδος και εξέλιξη στην μηχανική μάθηση οφείλεται στην ανάπτυξη νέων αλγόριθμων μάθησης και θεωρίας και φυσικά στην συνεχιζόμενη «έκρηξη» της διαθεσιμότητας μεγάλου όγκου διαδικτυακών δεδομένων αλλά και στην υπολογιστική ισχύ χαμηλού κόστους που προσφέρεται. Ουσιαστικά, μέσω της μηχανικής μάθησης ο υπολογιστής αναλύει μεγάλο όγκο δεδομένων, εξάγει αυτόματα πληροφορίες και τις χρησιμοποιεί για να κάνει προγνώσεις, να αποκρυπτογραφήσει εάν η πρόγνωση ήταν ορθή και, αν ήταν λανθασμένη, να μάθει από αυτήν για να πραγματοποιήσει στο μέλλον μια πιο σωστή πρόγνωση. Η υιοθέτηση μεθόδων μηχανικής μάθησης με την χρήση δεδομένων μπορούν να εντοπιστούν σε όλη την

επιστήμη, την τεχνολογία και το εμπόριο, οδηγώντας έτσι σε μια πιο τεκμηριωμένη λήψη αποφάσεων σε πολλούς τομείς της ζωής, συμπεριλαμβανομένων την εκπαίδευση, την χρηματοοικονομική, την μοντελοποίηση και το μάρκετινγκ (Jordan & Mitchell, 2015).

Η επίδραση της μηχανικής μάθησης έχει επίσης γίνει έντονα αισθητή στην επιστήμη υπολογιστών (computer science) και σε ένα μεγάλο εύρος κλάδων που ασχολούνται με data intensive ζητήματα, όπως στις καταναλωτικές υπηρεσίες, στην διάγνωση βλαβών σε σύνθετα συστήματα και στον έλεγχο εφοδιαστικών αλυσίδων.

Στην παρούσα διπλωματική, εφαρμόζονται 5 μέθοδοι πρόγνωσης: Linear Regression, Regression Trees, Support Vector Machines, Gaussian Process Regression Models και Ensemble of Trees.

4.1. Support Vector Machines (SVM)

Τα Support Vector Machines αποτελούν μια μεθοδολογία επιβλεπόμενης² μηχανικής μάθησης που χρησιμοποιείται στην ταξινόμηση δεδομένων. Η βασική ιδέα ενός SVM είναι η επιλογή ενός μικρού αριθμού δεδομένων (data points) από ένα ευρύτερο σύνολο δεδομένων που ονομάζονται Support Vectors (SV) και ορίζουν τα όρια των 2 διαφορετικών κλάσεων, ενώ στο μέσο των ορίων αυτών βρίσκεται ο βέλτιστος διαχωριστής των δεδομένων (ο οποίος απέχει την ίδια απόσταση από τις δύο κλάσεις) – Αναλυτικά αυτό περιγράφεται στο Σχήμα 1. Η μεθοδολογία μπορεί να γενικευτεί και να εφαρμοστεί σε περιπτώσεις που περιλαμβάνουν περισσότερες από 2 κλάσεις. Παρόλα αυτά, στην μελέτη αυτή, εφαρμόζουμε directional forecasting, επομένως η δυαδική έκδοση του μοντέλου είναι επαρκής. Τα SVM μπορούν να επιλύσουν και προβλήματα regression, τα οποία ονομάζεται Support Vector Regression (SVR). Τα SVR επιλύουν προβλήματα regression με βάση τα support vector machines (Plakandaras, Rangan, Gogas, & Papadimitriou, 2015). Τα SVR επιπλέον έχουν ισχυρή ικανότητα μάθησης, κάτι που τα καθιστά κατάλληλα για ένα μικρό σύνολο δειγμάτων και έχουν μεγάλη επιτυχία πρόγνωσης, ενώ ταυτόχρονα αποδίδουν πολύ καλά αποτελέσματα σε άγνωστα δεδομένα. Το θετικό αυτό στοιχείο προκύπτει διότι μπορούν να επεξεργαστούν διάφορες εργασίες καθώς προσαρμόζονται και ταυτόχρονα αποκτούν πρόσβαση σε ικανοποιητικές απαντήσεις. Συγκεκριμένα, μια μη γραμμική σχέση μπορεί να μετατραπεί σε

²Στην επιβλεπόμενη μηχανική μάθηση, ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένες εισόδους σε γνωστές και επιθυμητές εξόδους (σύνολο εκπαίδευσης), έχοντας ως απώτερο στόχο την γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο (σύνολο ελέγχου).

γραμμική με την χρήση των SVR και έτσι η εργασία να επιλυθεί εύκολα (Su, Zhang, Zhu, Zha, & Wen, 2019).

4.1.1. Μαθηματική προσέγγιση

Θεωρούμε ένα dataset (διάνυσμα-vectors) $\mathbf{x}_i \in \mathbb{R}^n$ ($i=1, 2, \dots, N$) το οποίο διαχωρίζεται σε δύο κατηγορίες (output targets) $y_i \in \{-1, +1\}$. Εάν οι δύο αυτές κατηγορίες είναι γραμμικά διαχωρισμένες, τότε ορίζουμε ένα όριο διαχωρισμού ως:

$$f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i - b = 0$$

Συνυπολογίζοντας ότι:

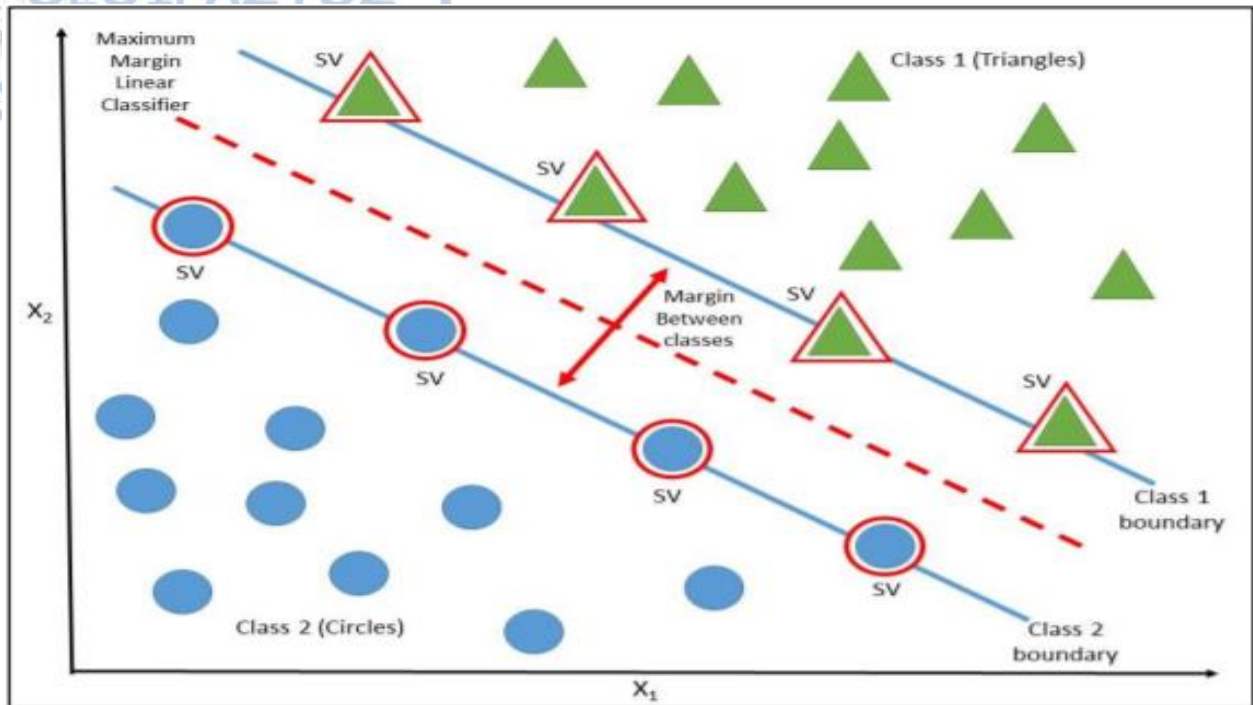
$$\mathbf{w}^T \mathbf{x}_i - b > 0 \text{ for } y_i \in +1$$

Και

$$\mathbf{w}^T \mathbf{x}_i - b < 0 \text{ for } y_i \in -1$$

Ωστε $y_i f(\mathbf{x}_i) > 0, \forall i$. Όπου $\mathbf{w}$ είναι το διάνυσμα των παραμέτρων και b είναι το bias.

Ο βέλτιστος διαχωριστής (hyperplane) αναγνωρίζεται ως το όριο καθορισμού που ταξινομεί κάθε διάνυσμα δεδομένων στην αντίστοιχη κατηγορία και έχει την μέγιστη απόσταση (συνχά καλείται margin) και από τις δύο κατηγορίες. Τα οριακά σημεία δεδομένων (marginal data points) που καθορίζουν την θέση του hyperplane ονομάζονται Support Vectors (SVs). Στο Σχήμα 1, τα Support Vectors παριστάνονται με το έντονο περίγραμμα, οι margin lines (γραμμές περιθωρίου) (που καθορίζουν την απόσταση του hyperplane από κάθε κατηγορία) παριστάνονται με τις μη διακοπτόμενες γραμμές και το hyperplane παριστάνεται με την διακεκομμένη γραμμή (Athanasiou, Gogas, & Papadimitriou, 2020).



Σχήμα 1. Επιλογή Hyperplane και support vectors. Οι δύο κατηγορίες αντιπροσωπεύονται από κύκλους και τρίγωνα. Τα support vectors υποδεικνύονται από τους έντονους κοίλους κύκλους και τρίγωνα (κόκκινου χρώματος), οι margin lines (γραμμές περιθωρίου) αντιπροσωπεύονται από τις συνεχείς γραμμές και το hyperplane (γραμμικός διαχωριστής) αντιπροσωπεύεται από την διακεκομμένη γραμμή (Dimitriadou, Gogas, Papadimitriou, & Plakandaras, 2018).

Η λύση στο πρόβλημα εντοπισμού του βέλτιστου hyperplane μπορεί να δοθεί από την διαδικασία Lagrange που χρησιμοποιεί την ακόλουθη εξίσωση:

$$\min_{\mathbf{w}, b, \xi} \max_{\alpha, \mu} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i - \sum_{j=1}^N \alpha_j [y_j (\mathbf{w}^T \mathbf{x}_j - b) - 1 + \xi_j] - \sum_{k=1}^N \mu_k \xi_k \right\}$$

Όπου ξ_i μετρά την απόσταση του διανύσματος $\mathbf{x}_i$ από το hyperplane όταν υπάρξει λανθασμένη ταξινόμηση. Τα $\alpha_1, \alpha_2, \dots, \alpha_n$ είναι οι μη αρνητικοί πολλαπλασιαστές Lagrange.

Το hyperplane ορίζεται επομένως ως:

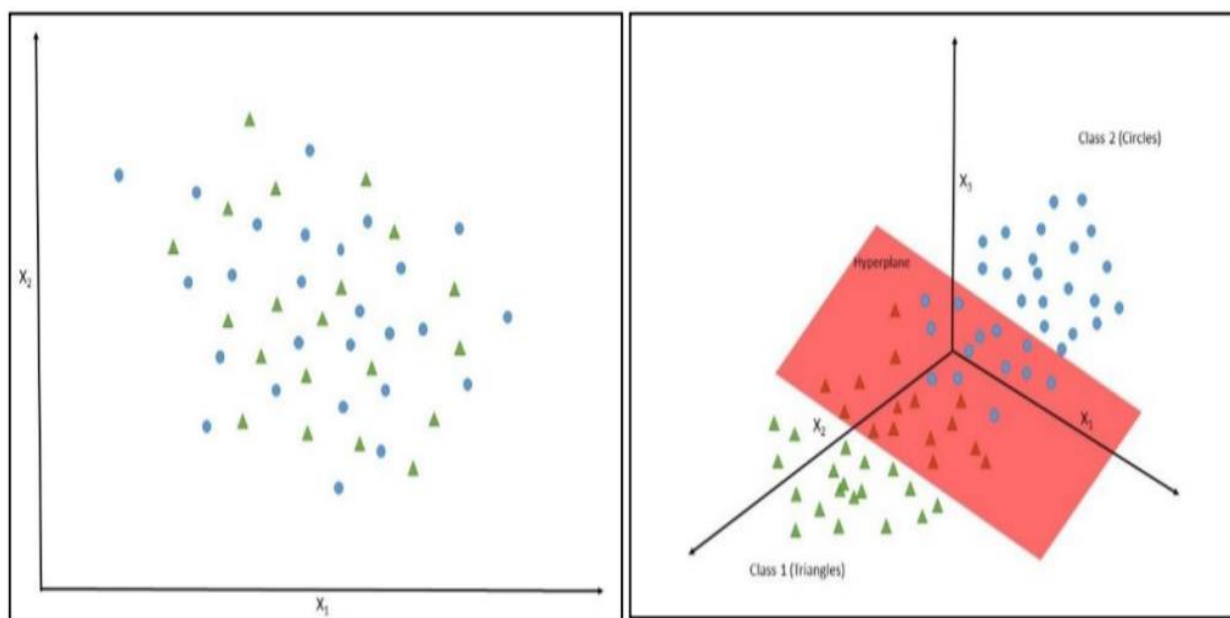
$$\hat{\mathbf{w}} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

$$\hat{b} = \hat{\mathbf{w}}^T \mathbf{x}_i - y_i, i \in V,$$

Όπου $V = \{i : 0 < y_i < C\}$, είναι το σύνολο των support vector δεικτών.

4.1.2. Μέθοδοι Kernel

Σε αρκετές περιπτώσεις η ανάλυση μη γραμμικών δεδομένων παρουσιάζει αρκετά προβλήματα, καθώς σε έναν μονοδιάστατο χώρο δεν μπορούν να διαχωριστούν σε κλάσεις τα δεδομένα αυτά. Για τον λόγο αυτό, τα μοντέλα SVM χρησιμοποιούν Kernel functions (πυρήνες) για την αποτελεσματική αντιμετώπιση αυτού του προβλήματος. Οι λειτουργίες Kernel προβάλλουν τα σημεία του χώρου δεδομένων σε έναν χώρο υψηλότερων διαστάσεων, που καλείται feature space. Ο χώρος διαστάσεων αυξάνεται έως ότου τα δεδομένα στο feature space να είναι γραμμικώς διαχωρίσιμα (ή τουλάχιστον όσο το δυνατόν πλησιέστερα στον γραμμικό διαχωρισμό). Όταν η λειτουργία Kernel είναι μη γραμμική, τότε το παραγόμενο μοντέλο SVM είναι επίσης μη γραμμικό. Στο Σχήμα 2, απεικονίζεται ένα σύνολο δεδομένων (dataset) που απαρτίζεται από δύο κατηγορίες που δεν είναι γραμμικά διαχωρίσιμες στον αρχικό δισδιάστατο χώρο (αριστερό γράφημα). Μετά την προβολή σε έναν χώρο υψηλότερων διαστάσεων (δεξί γράφημα), ο γραμμικός διαχωρισμός καθίσταται εφικτός.



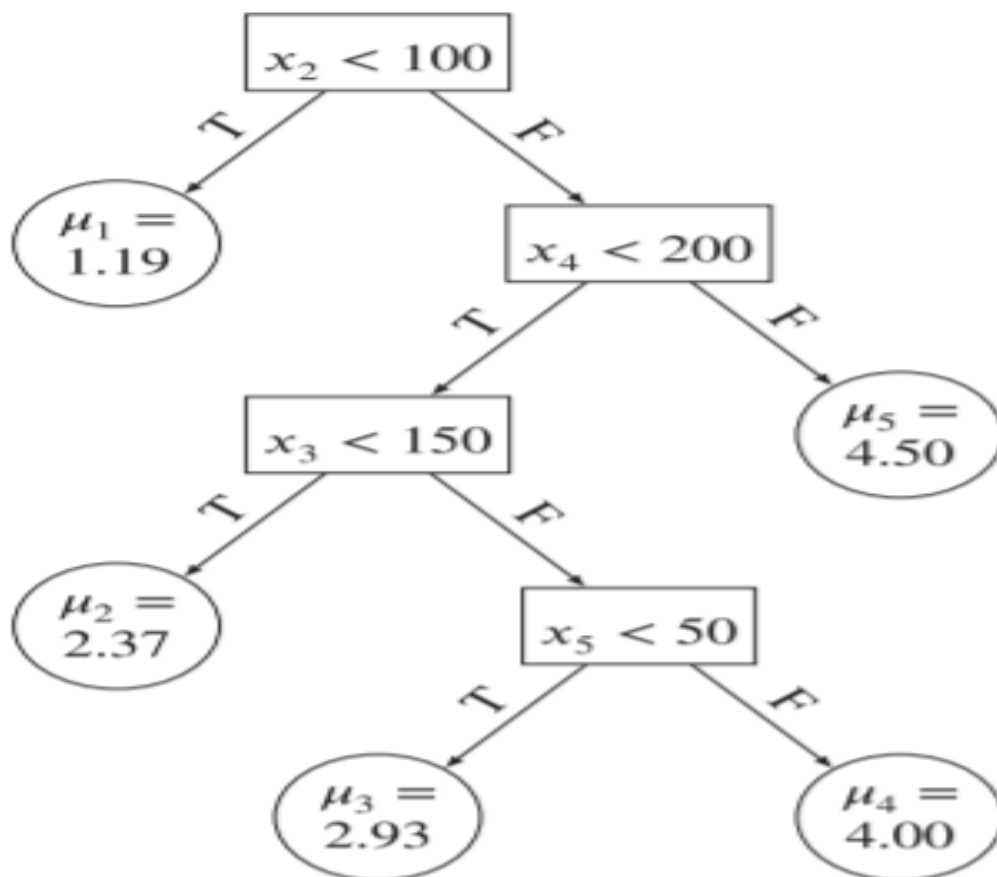
Σχήμα 2. Το σενάριο μη διαχωρίσιμων δεδομένων σε 2 κλάσεις στις 2 διαστάσεις(Αριστερά) και η περίπτωση στην οποία παρατηρείται διαχωρισμός στο feature space μετά την προβολή σε χώρο ανώτερων διαστάσεων(Δεξιά) (Dimitriadou, Gogas, Paradimitriou, & Plakandaras, 2018).

4.2. Gaussian Process Regression

Το Gaussian Processes είναι μια ευέλικτη κατηγορία μη παραμετρικών μοντέλων μηχανικής μάθησης που χρησιμοποιούνται κατά κύριο λόγο για την μοντελοποίηση χωρικών (spatial) δεδομένων και δεδομένων που αφορούν χρονοσειρές. Τα Gaussian μοντέλα χρησιμοποιούνται συνήθως για την επίλυση δύσκολων προβλημάτων μηχανικής μάθησης. Είναι ιδιαίτερα χρήσιμα και ελκυστικά λόγω της ευέλικτης μη παραμετρικής φύσης τους και της υπολογιστικής απλότητας που παρουσιάζουν. Μια κοινή εφαρμογή των Gaussian Processes είναι το regression (παλινδρόμηση). Για παράδειγμα, εάν έχουμε ελλιπή γεωγραφικά δεδομένα για τον καιρό (π.χ. θερμοκρασία ή υγρασία) πως μπορεί κάποιος να αποκτήσει τιμές για μη γνωστές ή μη εποπτευόμενες περιοχές; Εάν λοιπόν υπάρχουν βάσιμοι λόγοι για την κανονική κατανομή των δεδομένων της περιοχής, τότε η χρήση του GP μοντέλου θα μπορούσε να αποτελέσει μια συνετή επιλογή (Powell, 2021). Το Gaussian Process Regression βασίζεται στον προσδιορισμό μιας κατάλληλης λειτουργίας kernel (kernel function) ή ενός μέτρου ομοιότητας μεταξύ σημειακών δεδομένων (data points) των οποίων οι θέσεις είναι γνωστές. Σε σύγκριση με άλλες μεθόδους μηχανικής μάθησης, τα πλεονεκτήματα του GPR έγκεινται στο ότι είναι σε θέση να ενσωματώνει ανεμπόδιστα πολλαπλές εργασίες μηχανικής μάθησης, όπως η εκτίμηση παραμέτρων. Επιπλέον διαθέτει εξαιρετική απόδοση με ένα σχετικά μικρό σύνολο δεδομένων εκπαίδευσης (training dataset) για την πραγματοποίηση της πρόγνωσης. Ωστόσο ένα γνωστό πρόβλημα που παρουσιάζεται είναι ότι λόγω της πολυπλοκότητας υπολογισμού κατά την πρόγνωση, καθίσταται ανέφικτη η αποτελεσματικότητα του GPR για μεγάλα σύνολα δεδομένων σύμφωνα με τους Su et al. (2019).

4.3. Regression Trees

Τα Decision Trees είναι μία από τις προγνωστικές προσεγγίσεις μοντελοποίησης που χρησιμοποιούνται στην στατιστική, στο data mining, και στην μηχανική μάθηση. Χρησιμοποιεί ένα decision tree (ως μοντέλο πρόγνωσης) προκειμένου να μεταβεί από τις παρατηρήσεις για ένα αντικείμενο (το οποίο αντιπροσωπεύεται στα κλαδιά) σε συμπεράσματα σχετικά με την τιμή-στόχο του αντικειμένου (που απεικονίζεται στα φύλλα). Τα Decision trees όπου η μεταβλητή στόχος μπορεί να πάρει συνεχείς τιμές (συνήθως πραγματικούς αριθμούς) ονομάζονται regression trees. Ας υποθέσουμε ότι έχουμε συμεταβλητές (covariates) $x = (x_1, \dots, x_5)$ και εξαγόμενο y . Μια αναπαράσταση ενός υποθετικού, απλού regression tree φαίνεται στο Σχήμα 3. Κάθε σημείο στο οποίο υπάρχει ένας δυαδικός διαχωρισμός απόφασης (binary decision split) καλείται κόμβος (node). Για παράδειγμα, σύμφωνα με το Σχήμα 3 στον αρχικό κόμβο (root node) υπάρχει μια συνθήκη $x_2 < 100$. Εάν το $x_2 < 100$ είναι αλήθεια, τότε ακολουθούμε την διαδρομή προς τα αριστερά. Διαφορετικά ακολουθούμε το μονοπάτι προς τα δεξιά. Υποθέτοντας ότι το $x_2 < 100$ είναι αλήθεια, βλέπουμε ότι φτάνουμε σε έναν κόμβο που δεν διαχωρίζεται. Το σημείο αυτό καλείται τερματικός κόμβος (terminal node) και η παράμετρος $\mu_1 = 1.19$ είναι η ανατεθειμένη τιμή του $E[y|x]$ για οποιοδήποτε x όπου $x_2 < 100$. Ας υποθέσουμε ότι το $x_2 < 100$ δεν είναι αλήθεια. Έτσι, μετακινώντας κατά μήκος της δεξιάς πλευράς, εμφανίζεται ένας άλλος εσωτερικός κόμβος με την συνθήκη $x_4 < 200$. Αυτή η συνθήκη θα ελεγχόταν και εάν η συνθήκη είναι αληθινή, ακολουθούμε την διαδρομή προς τα αριστερά. Αυτή η διαδικασία συνεχίζεται μέχρι να φτάσουμε σε έναν τερματικό κόμβο και η παράμετρος μ_i σε αυτόν τον τερματικό κόμβο ορίζεται ως η τιμή του $E[y|x]$, όπου μ_i είναι η μέση παράμετρος του i κόμβου για το regression tree (Tan & Roy, 2019). Έτσι για παράδειγμα, σε ένα ζήτημα με $x_{k2} = 120$, $x_{k3} = 115$, $x_{k4} = 191$ και $x_{k5} = 56$ θα ανατεθεί μια τιμή $\mu_2 = 2,37$ για $E[y|x]$. Η τιμή θα ήταν ακριβώς η ίδια για ένα άλλο ζήτημα k' που αντίθετα με το k έχει συμεταβλητές $x_{k'1} = 130$, $x_{k'2} = 135$, $x_{k'3} = 92$, $x_{k'4} = 183$ και $x_{k'5} = 10$.



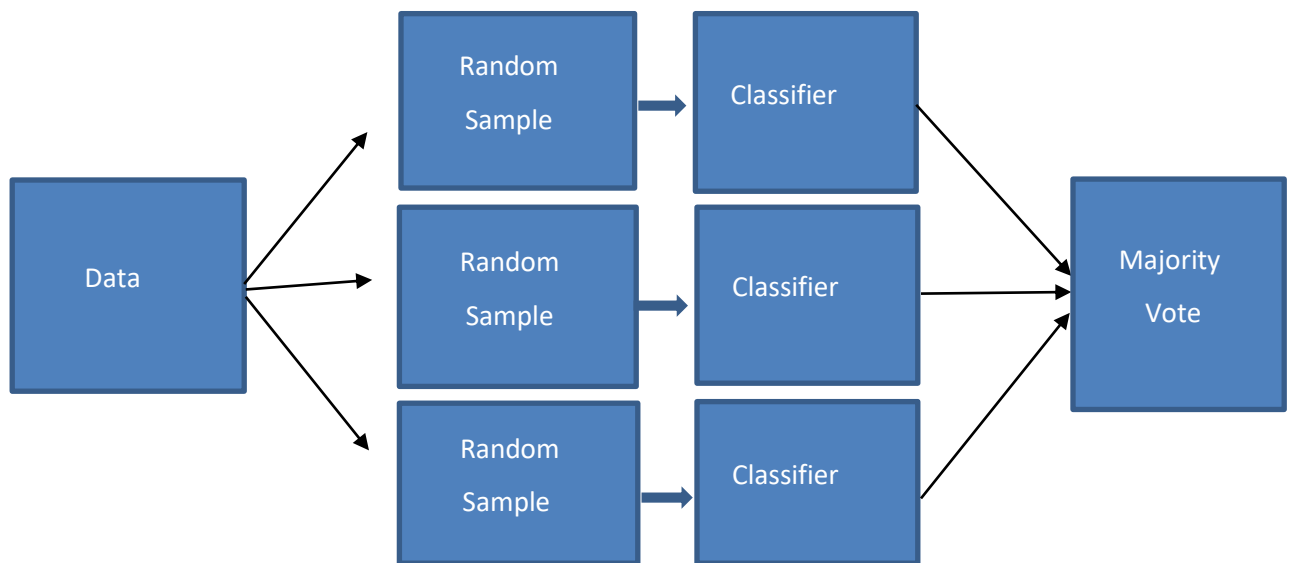
Σχήμα 3. Παράδειγμα ενός Regression Tree $g(x; T, M)$ όπου το μ_i είναι η μέση παράμετρος του i κόμβου για το Regression Tree (Tan & Roy, 2019).

4.4. Ensemble of Trees

Ένα Ensemble of Trees σχηματίζεται από πολλά μεμονωμένα δέντρα που αθροίζονται διαδοχικά. Τα Decision Trees είναι ένας από τους πιο αποτελεσματικούς και ερμηνεύσιμους αλγόριθμους ταξινόμησης, ενώ οι τεχνικές ensemble έχουν αποδειχθεί ότι επιλύουν προβλήματα που προέρχονται από το overfitting και την διακύμανση. Οι μέθοδοι ensemble προτιμώνται επειδή συνδυάζουν πολλά decision trees για να παράγουν καλύτερη απόδοση πρόγνωσης, σε αντίθεση με την χρήση ενός μόνο decision tree. Η βασική αρχή που διέπει το μοντέλο του ensemble είναι ότι μια ομάδα αδύναμων learners συγκεντρώνονται και σχηματίζουν έναν ισχυρό learner. Οι βασικές τεχνικές για την εκτέλεση ensemble decision trees είναι οι Bagging και Boosting (Zimmermann, 2008).

4.4.1. Bagging

Το Bagging (Bootstrap Aggregation) χρησιμοποιείται όταν στόχος μας είναι να μειώσουμε την διακύμανση ενός decision tree. Κατά την διαδικασία αυτή, η βασική ιδέα είναι δημιουργηθούν πολλά υποσύνολα δεδομένων από το training sample επιλεγμένα τυχαία με αντικατάσταση. Κάθε συλλογή δεδομένων υποσυνόλου χρησιμοποιείται για την εκπαίδευση των αντίστοιχων decision trees. Έτσι ως αποτέλεσμα, καταλήγουμε σε ένα σύνολο διαφορετικών μοντέλων. Τέλος χρησιμοποιείται ο μέσος όρος όλων των προβλέψεων που προέκυψαν από διαφορετικά δέντρα, ο οποίος είναι πιο ισχυρός και ακριβής από ένα μοναδικό decision tree.

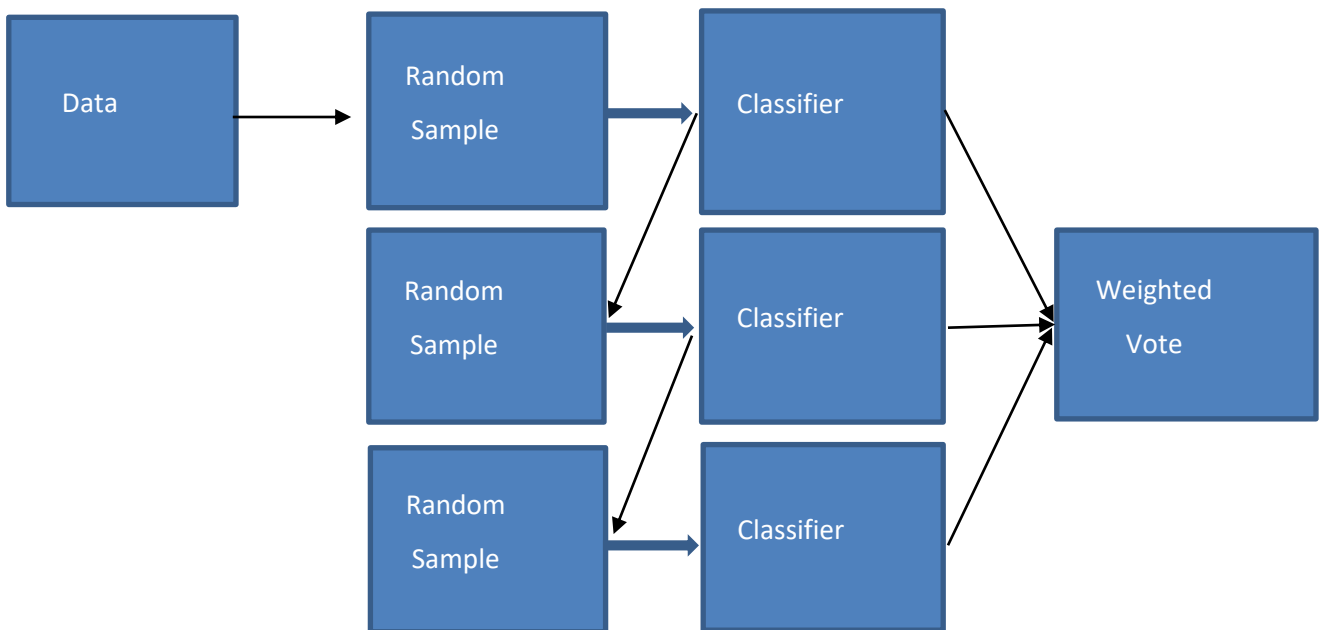


Σχήμα 4. Σχηματική επεξήγηση του Bagging (Peterson, 2018).

Τα Random Forest είναι μια επέκταση του bagging. Προχωράει ένα επιπλέον βήμα, κατά το οποίο εκτός από την λήψη του τυχαίου υποσυνόλου δεδομένων, λαμβάνει επιπλέον την τυχαία επιλογή χαρακτηριστικών σε κάθε node αντί να χρησιμοποιεί όλα τα χαρακτηριστικά για την ανάπτυξη δέντρων. Το πλεονέκτημα της χρήσης της τεχνικής Random Forest είναι ότι διαχειρίζεται υψηλό αριθμό δεδομένων με μεγάλη αποτελεσματικότητα, ενώ το μειονέκτημα είναι ότι η τελική πρόγνωση βασίζεται στις μέσες προβλέψεις των subset trees (δέντρα υποσυνόλων), επομένως δεν θα αποδώσει ακριβείς τιμές για το μοντέλο regression.

4.4.2. Boosting

Το Boosting είναι μια άλλη τεχνική του ensemble που αποσκοπεί στην δημιουργία ενός συνόλου προβλέψεων. Η τεχνική αυτή ξεκινά προσαρμόζοντας ένα αρχικό μοντέλο (π.χ. ένα δέντρο ή ένα linear regression) στα δεδομένα. Στη συνέχεια, κατασκευάζεται ένα δεύτερο μοντέλο που εστιάζει στην ακριβή πρόβλεψη των περιπτώσεων όπου το πρώτο μοντέλο δεν έχει καλή απόδοση. Ο συνδυασμός των δύο αυτών μοντέλων είναι καλύτερος από το κάθε μεμονωμένο μοντέλο. Έπειτα επαναλαμβάνεται η διαδικασία του boosting πολλές φορές. Κάθε διαδοχικό μοντέλο επιχειρεί να διορθώσει τις αδυναμίες και τα σφάλματα του συνδυασμένου ενισχυμένου συνόλου όλων των προηγούμενων μοντέλων.. Συνδυάζοντας ολόκληρο το σετ στο τέλος μετατρέπει τους αδύναμους learners σε ένα μοντέλο με καλύτερη απόδοση.



Σχήμα 5. Σχηματική επεξήγηση του Boosting (Peterson, 2018).

Το Gradient Boosting είναι μια επέκταση της μεθόδου boosting.

Gradient Boosting = Gradient Descent + boosting method

Συγκεκριμένα, χρησιμοποιεί αλγόριθμο gradient descent ³(καθόδου κλίσης) που έχει την ικανότητα να βελτιστοποιήσει οποιαδήποτε διαφοροποιημένη λειτουργία απώλειας (διαφορά μεταξύ πραγματικών και προβλεπόμενων τιμών). Ένα σύνολο δέντρων (ensemble of trees)

³ Το Gradient Descent είναι ένας αλγόριθμος βελτιστοποίησης που χρησιμοποιείται κατά την εκπαίδευση ενός μοντέλου μηχανικής μάθησης. Χρησιμοποιείται για την ενημέρωση παραμέτρων σε ένα μοντέλο.

χτίζεται ένα προς ένα και τα μεμονωμένα δέντρα αθροίζονται διαδοχικά. Το επόμενο δέντρο επιδιώκει να ανακτήσει την απώλεια. Το πλεονέκτημα της τεχνικής Gradient Boosting είναι ότι αποδίδει πολύ αποτελεσματικά όταν χρησιμοποιούνται unbalanced δεδομένα (π.χ. στην αξιολόγηση κινδύνου σε πραγματικό χρόνο). Τα βασικά του μειονεκτήματα είναι το γεγονός ότι είναι επιρρεπείς σε overfitting και ότι απαιτείται προσεκτικός συντονισμός διαφορετικών υπερ – παραμέτρων (hyper – parameters).

4.5. Linear Regression

Στην στατιστική, το linear regression είναι μια γραμμική προσέγγιση στην μοντελοποίηση της σχέσης μεταξύ μιας βαθμιδωτής απόκρισης και μιας ή περισσότερων επεξηγηματικών μεταβλητών (επίσης γνωστές και ως εξαρτώμενες και ανεξάρτητες μεταβλητές). Συνεπώς, η μέθοδος linear regression (γραμμική παλινδρόμηση) είναι ένας βασικός και κοινώς χρησιμοποιούμενος τύπος προγνωστικής ανάλυσης.

Η γενική ιδέα του regression βασίζεται στην εξέταση δύο βασικών πραγμάτων:

1. Ένα σύνολο μεταβλητών πρόγνωσης είναι αποτελεσματικό στην πρόγνωση μιας εξαρτώμενης μεταβλητής.
2. Ποιες μεταβλητές συγκεκριμένα αποτελούν σημαντικούς παράγοντες πρόγνωσης της εξαρτώμενης και με ποιο τρόπο επηρεάζουν την εξαρτώμενη μεταβλητή.

Οι εκτιμήσεις αυτές του regression χρησιμοποιούνται για να εξηγήσουν την σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών (Statistics Solutions, 2013). Η απλούστερη μορφή της εξίσωσης του regression με μια εξαρτώμενη και μια ανεξάρτητη μεταβλητή καθορίζεται από τον τύπο:

$$y = c + b * x$$

όπου y = εκτιμώμενη τιμή εξαρτώμενης μεταβλητής

c = σταθερά

b = συντελεστής παλινδρόμησης (regression coefficient)

x = τιμή ανεξάρτητης μεταβλητής.

Οι τρεις κύριες χρήσεις που προκύπτουν από την ανάλυση του regression είναι:

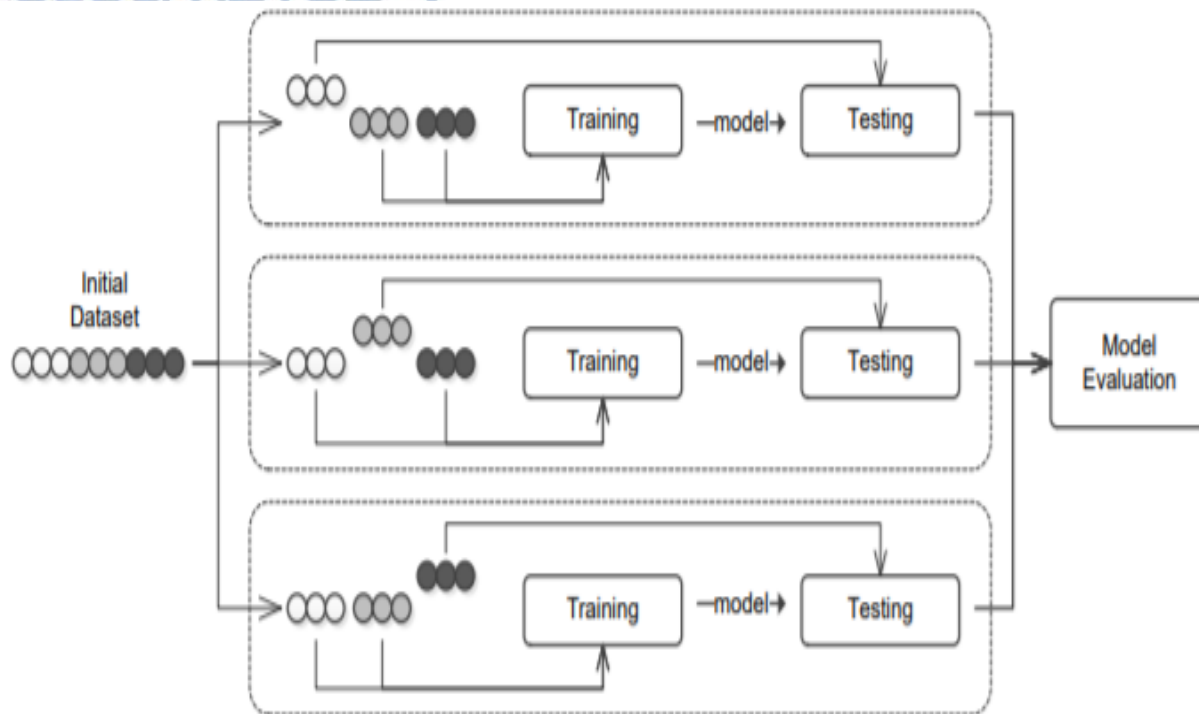
- a) Το regression μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της επίδρασης που ασκούν οι ανεξάρτητες μεταβλητές σε μια εξαρτημένη μεταβλητή.
- b) Δεύτερον, μπορεί να χρησιμοποιηθεί για την πρόγνωση των επιδράσεων ή των επιπτώσεων των αλλαγών. Δηλαδή η ανάλυση του regression μας βοηθά να

κατανοήσουμε πόσο μεταβάλλεται η εξαρτημένη μεταβλητή με μια αλλαγή σε μία ή περισσότερες ανεξάρτητες μεταβλητές.

c) Τρίτον, η ανάλυση του regression προβλέπει τάσεις και μελλοντικές τιμές.

4.6. Overfitting

Ένα πιθανό πρόβλημα που μπορεί να προκύψει κατά την εκπαίδευση μοντέλων μηχανικής μάθησης είναι το πρόβλημα του overfitting. Αυτή είναι μια κατάσταση όπου το επιλεγμένο μοντέλο είναι πολύ καλά προσαρμοσμένο στα δεδομένα εκπαίδευσης αλλά αποτυγχάνει στην περιγραφή της πραγματικής διαδικασίας δημιουργίας δεδομένων και συγκεκριμένα στα δεδομένα ελέγχου (out-of-sample). Συγκεκριμένα, το overfitting παρατηρείται κυρίως όταν ένα μοντέλο μαθαίνει τις λεπτομέρειες και τον θόρυβο των δεδομένων εκπαίδευσης (training data) σε τέτοιο βαθμό ώστε να επηρεάζει αρνητικά την απόδοση του μοντέλου κατά την χρήση νέων δεδομένων. Αυτό σημαίνει ότι ο θόρυβος ή οι τυχαίες διακυμάνσεις στα δεδομένα εκπαίδευσης συλλέγονται και μαθαίνονται ως έννοιες από το μοντέλο. Το βασικό πρόβλημα σε αυτό είναι ότι αυτές οι έννοιες δεν ισχύουν για νέα δεδομένα και έτσι επηρεάζουν αρνητικά την ικανότητα γενίκευσης των μοντέλων. Προκειμένου λοιπόν να αποφευχθεί το over-fitting (προσαρμογή του μοντέλου στα δεδομένα και όχι στο φαινόμενο), χρησιμοποιείται η μέθοδος k-fold cross validation στο κομμάτι του training. Σύμφωνα λοιπόν με την μεθοδολογία cross-validation το training sample (δείγμα εκπαίδευσης) χωρίζεται σε n μέρη. Μετά την επιλογή μιας αρχικής διαμόρφωσης (initial configuration), το μοντέλο εκπαιδεύεται n φορές στα μέρη $n-1$, διατηρώντας κάθε φορά ένα μέρος του συνόλου δεδομένων για σκοπούς δοκιμής (testing). Η ακρίβεια του in-sample με την επιλεγμένη διαμόρφωση είναι απλά η μέση τιμή των προβλέψεων σε όλα τα τμήματα n . Μετά την αλλαγή των παραμέτρων του μοντέλου, πραγματοποιείται εκ νέου η επαναληπτική μέθοδος εκπαίδευσης (iterative training scheme) έως ότου βρεθεί το σετ των παραμέτρων με το οποίο επιτυγχάνεται το ελάχιστο σφάλμα πρόγνωσης. Το εκπαιδευτικό αυτό σχήμα καλείται « n -fold cross validation» και μια επισκόπηση αυτού απεικονίζεται στο Σχήμα 6.



Σχήμα 6. Επισκόπηση ενός 3-fold Cross Validation training scheme. Κάθε υποσύνολο χρησιμοποιείται ως δείγμα δοκιμής (testing), ενώ τα υπόλοιπα χρησιμοποιούνται για την εκπαίδευση του μοντέλου για κάθε συνδυασμό τιμών των παραμέτρων (Dimitriadou, Gogas, Papadimitriou, & Plakandaras, 2018).

5. Empirical Section

5.1. Empirical Study

5.1.1. Πρόβλεψη κατεύθυνσης της τιμής του Φυσικού Αερίου

Προκειμένου να ελεγχθεί η ικανότητα γενίκευσης του επιλεγμένου μοντέλου, το dataset χωρίζεται σε δύο μέρη: το πρώτο 90% αποτελεί τα δεδομένα εκπαίδευσης (in-sample) και το εναπομείναν 10% των πιο πρόσφατων παρατηρήσεων αποτελεί τα δεδομένα ελέγχου (out-of-sample).

Επιλέχθηκε επιπλέον 5-fold cross validation για την μελέτη αυτή προκειμένου να αποφευχθεί το over fitting των δεδομένων. Εκπαιδεύτηκαν αυτοπαλίνδρομα (Autoregressive-AR) μοντέλα προσθέτοντας κάθε φορά ένα-ένα τα lags. Δοκιμάστηκαν συνολικά lags από 1 έως και 15. Έπειτα από την εκπαίδευση των SVM AR μοντέλων για την εύρεση του πιο αντιπροσωπευτικού αριθμού Autoregressive Lags, δεν προκύπτει κάποιο ικανοποιητικό μοντέλο. Έτσι λοιπόν συμπεραίνεται πως το series δεν έχει memory, δηλαδή οι παρελθοντικές τιμές δεν επηρεάζουν τις μελλοντικές. Για τον λόγο αυτό επιλέχτηκαν επομένως όλα τα Lags, με όλες τις μεταβλητές έχοντας πάντα την κατεύθυνση του natural gas spot prices ως target (κύρια μεταβλητή) για την εφαρμογή των



μοντέλων του classification learner. Τα αποτελέσματα της μελέτης παρουσιάζονται στον Πίνακα 2:

Πίνακας 2. Τα ποσοστά ακρίβειας πρόγνωσης των μοντέλων του classification learner για τα δεδομένα εκπαίδευσης (in-sample) την χρονική στιγμή $t+1$.

Fine Tree : 53,5 %	Coarse Gaussian SVM : 58,5 %
Medium Tree : 56,9 %	Fine KNN : 52,4 %
Coarse Tree : 56,6 %	Medium KNN : 56,6%
Linear Discriminant : 57,7 %	Coarse KNN : 56,4 %
Quadratic Discriminant : 50,9 %	Cubic KNN : 56,8%
Logistic Regression : 56,9 %	Weighted KNN : 54,9 %
Linear SVM : 58,5 %	Boosted Trees : 57,5 %
Quadratic SVM : 55,6 %	Bagged Trees : 55,2 %
Cubic SVM : 53,4 %	Subspace Discriminant : 58,5 %
Fine Gaussian SVM : 57,2 %	Subspace KNN : 52,7 %
Medium Gaussian SVM : 58,3 %	RUS Boosted Trees : 53,5 %

Τα ποσοστά ακρίβειας πρόγνωσης που προκύπτουν από όλα τα μοντέλα δεν είναι ικανοποιητικά, καθώς κανένα μοντέλο δεν ξεπερνάει το 60%, όσον αφορά την επιτυχία πρόβλεψης. Τα παραπάνω λοιπόν αποτελούν ένδειξη πως τα δεδομένα μας ανταποκρίνονται σε αποτελεσματική αγορά. Η υπόθεση της αποτελεσματικής αγοράς δηλώνει ότι: 1) οι τρέχουσες τιμές περιλαμβάνουν όλες τις διαθέσιμες πληροφορίες και προσδοκίες, 2) οι τρέχουσες τιμές είναι η καλύτερη προσέγγιση της εσωτερικής αξίας, 3) οι αλλαγές των τιμών οφείλονται σε απρόβλεπτα γεγονότα και 4) λανθασμένες εκτιμήσεις τιμών συμβαίνουν αλλά όχι σε προβλέψιμα μοτίβα που μπορούν να οδηγήσουν σε consistent outperformance (Fama, 1970). Μια αγορά δηλαδή καλείται αποτελεσματική όταν δεν καθίσταται δυνατή για πρόγνωση με την χρήση παρελθοντικών πληροφοριών και δεδομένων. Σημαίνει δηλαδή πως όλες οι διαθέσιμες πληροφορίες έχουν ήδη ενσωματωθεί στην τιμή του asset και άρα δεν μπορεί να χρησιμοποιηθεί καμία επιπλέον πληροφορία για την περαιτέρω πρόγνωση (semi-strong form of efficiency).

5.1.2. Πρόβλεψη Τιμής Φυσικού Αερίου

Ξεκινήσαμε την εκπαίδευση των AR μοντέλων προσθέτοντας κάθε φορά ένα επιπλέον lag, ενώ εκπαιδεύτηκαν μοντέλα με έως και 15 lag προκειμένου να βρούμε τον βέλτιστο αριθμό lag. Χωρίσαμε το δείγμα μας σε δύο μέρη. Το πρώτο από τις 19 Νοεμβρίου 2010 έως τις 19 Σεπτεμβρίου του 2019, από το οποίο χρησιμοποιήθηκαν 2180 τιμές για την εκπαίδευση των μοντέλων (in-sample) και το δεύτερο μέρος που αποτελείται από τις 20 Σεπτεμβρίου του 2019 έως τις 18 Σεπτεμβρίου του 2020, από το οποίο χρησιμοποιήθηκαν οι υπόλοιπες 243 τιμές για την αξιολόγηση της ακρίβειας πρόγνωσης που παρουσιάζει το out-of-sample τμήμα των μεταβλητών. Χρησιμοποιήσαμε και εδώ το 5-fold cross validation για την αποφυγή του overfitting. Η βέλτιστη δομή υστέρησης (δηλαδή ο βέλτιστος αριθμός lag που θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου) επιλέγεται με βάση το ελάχιστο RMSE που παρουσιάζεται κατά το training των in-sample δεδομένων.

$$RMSE = \sqrt{\frac{\sum_{t=1}^T (\hat{y}_t - y_t)^2}{T}},$$

όπου $\hat{y}$ = forecasted value, y = actual value, T = observations

Η ακρίβεια πρόγνωσης του κάθε μοντέλου για το out-of-sample dataset υπολογίζεται από τον τύπο του Root Mean Square Error (RMSE) και έτσι όσο μικρότερο το RMSE, δηλαδή το σφάλμα, τόσο καλύτερη και πιο ακριβής είναι η πρόγνωση του κάθε μοντέλου. Εκπαιδεύοντας λοιπόν το Linear SVM μοντέλο με εναλλαγές στον αριθμό των lag που χρησιμοποιούμε κάθε φορά, καταλήγουμε ότι με την χρήση των 14 πρώτων lag έχουμε το μικρότερο $RMSE = 0,04196$. Μέσω της παραπάνω διαδικασίας παίρνουμε τα RMSE για το in-sample dataset.

Γίνεται σύγκριση ουσιαστικά των πραγματικών τιμών με των παραγόμενων από το μοντέλο τιμών. Η παραπάνω διαδικασία πραγματοποιείται για κάθε έναν από τους χρονικούς ορίζοντες $t+1$, $t+3$, $t+5$ και $t+10$. Επιπλέον προσθέτουμε ένα μοντέλο Random Walk για κάθε χρονικό ορίζοντα ($t+1$, $t+3$, $t+5$, $t+10$) με τα αντίστοιχα RMSE για το in-sample και out-of-sample του καθενός. Το RW αυτό δημιουργήθηκε προκειμένου να γίνει σύγκριση με τα υπόλοιπα μοντέλα και να βγει συμπέρασμα για το πόσο αντιπροσωπευτικό και αποτελεσματικό είναι το κάθε σύνθετο μοντέλο. Έπειτα λοιπόν από την παραπάνω διαδικασία προκύπτουν τα αποτελέσματα των RMSE για in-sample και out-of-sample για τον χρονικό ορίζοντα $t+1$, όπως φαίνονται στον πίνακα 3.

Πίνακας 3. Τα RMSE για *in-sample* και *out-of-sample* data για τον χρονικό ορίζοντα $t+1$.

Models	In sample RMSE	OOS RMSE
RW	0.042643	0.057435
Linear Regression	0.038421	0.062872
Interactions Linear	0.1009	1.560244
Robust Linear	0.039067	0.05736
Fine Tree	0.04992	0.071181
Medium Tree	0.045707	0.083954
Coarse Tree	0.047189	0.083388
Linear SVM	0.038581	0.056694
Quadratic SVM	0.044703	0.161214
Cubic SVM	0.058968	0.456275
Fine Gaussian SVM	0.098278	0.461634
Medium Gaussian SVM	0.042352	0.243223
Coarse Gaussian SVM	0.046224	0.079625
Boosted Trees	0.065151	0.06169
Bagged Trees	0.041597	0.061089
Squared Exponential	0.039915	0.216353
Matern 5/2 GPR	0.039915	0.164098
Exponential GPR	0.039989	0.098513
Rational Quadratic	0.040069	0.120337

Στα αποτελέσματά που προκύπτουν στον πίνακα 3 παρατηρούμε πως υπάρχουν αρκετά μοντέλα που παρουσιάζουν overfitting, επομένως επιλέγουμε να επικεντρωθούμε και να ασχοληθούμε μόνο με τα μοντέλα που δεν παρουσιάζουν overfitting. Τα μοντέλα λοιπόν που παραλείπουμε είναι αυτά που εμφανίζονται με το χαρακτηριστικό κίτρινο φόντο, και συγκεκριμένα τα Interactions Linear, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM, Squared Exponential, Matern 5/2 GPR, Exponential GPR και Rational Quadratic.

5.2. Empirical Results

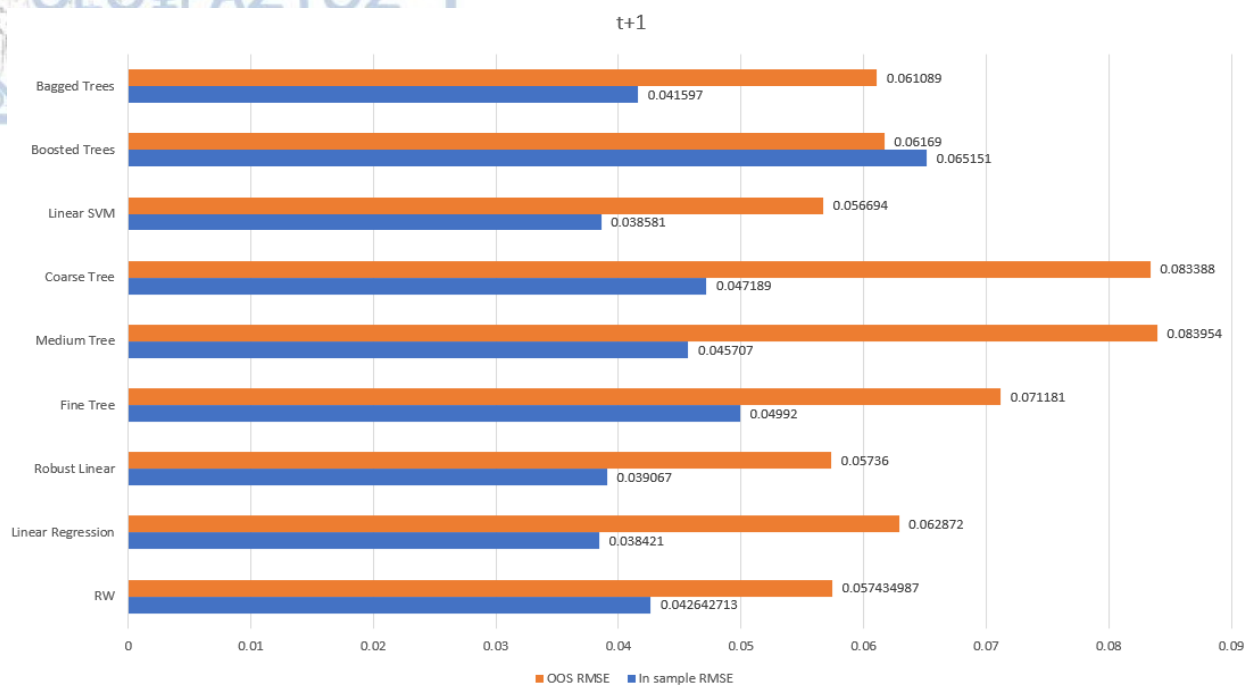
Με τα παραπάνω δεδομένα λοιπόν γίνεται η μελέτη και ο υπολογισμός των RMSE για τους χρονικούς ορίζοντες $t+1$, $t+3$, $t+5$ και $t+10$, τόσο για IS όσο και για OOS.

5.2.1. Χρονικός ορίζοντας $t+1$

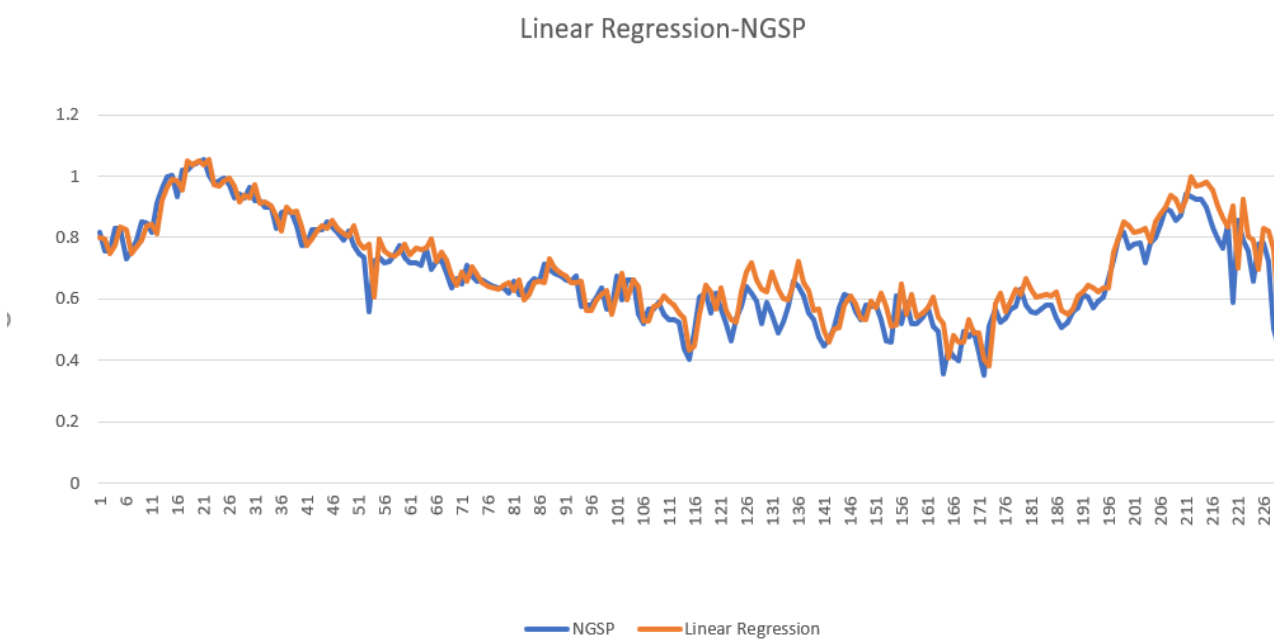
Σύμφωνα με το Γράφημα 1 και τον Πίνακα 4 παρατηρούμε πώς για τον χρονικό ορίζοντα $t+1$ το βέλτιστο μοντέλο είναι το Linear Regression με το μικρότερο in-sample RMSE = 0,03842, ενώ το μικρότερο RMSE από τα out-of-sample data το έχει το Linear SVM με RMSE = 0,056694. Με βάση τις παρατηρήσεις αυτές λοιπόν καταλαβαίνουμε πως το Linear SVM αποτελεί το δεύτερο καλύτερο μοντέλο, μετά το Linear Regression, αφού έχει το χαμηλότερο RMSE στα OOS data και το δεύτερο χαμηλότερο RMSE στα training data (IS). Επιπλέον και το Robust Linear παρουσιάζει πολύ καλά αποτελέσματα, καθώς έχει το δεύτερο χαμηλότερο RMSE στα testing data και το τρίτο χαμηλότερο στα training data. Το Random Walk που δημιουργήσαμε, ανταποκρίνεται πολύ καλά στα OOS data. Μπορούμε λοιπόν να συμπεράνουμε γενικά πως τα γραμμικά μοντέλα είναι σε θέση να προβλέψουν με υψηλή ακρίβεια τις τιμές του natural gas για την επόμενη ημέρα. Όπως φαίνεται στον Πίνακα 4 το βέλτιστο μοντέλο, που είναι το Linear Regression, μπορεί να προβλέψει με υψηλή ακρίβεια άγνωστα δεδομένα.

Πίνακας 4. Τα RMSE για IS και OOS για τον χρονικό ορίζοντα $t+1$.

$t+1$	In sample	OOS
Models	RMSE	RMSE
RW	0.042643	0.057435
Linear Regression	0.038421	0.062872
Robust Linear	0.039067	0.05736
Fine Tree	0.04992	0.071181
Medium Tree	0.045707	0.083954
Coarse Tree	0.047189	0.083388
Linear SVM	0.038581	0.056694
Boosted Trees	0.065151	0.06169
Bagged Trees	0.041597	0.061089



Γράφημα 1. Η σύγκριση των RMSE ανάμεσα σε IS και OOS για τον χρονικό ορίζοντα t+1.



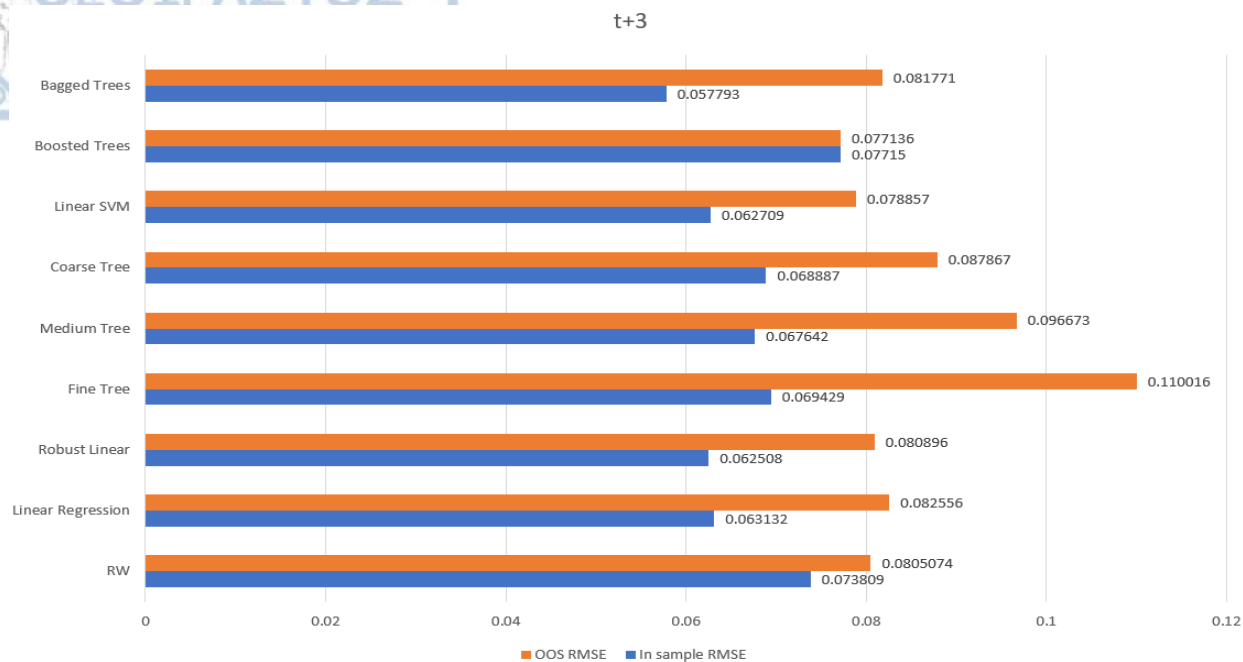
Γράφημα 2. Σύγκριση των actual τιμών του στόχου (natural gas spot prices) με τις fitted τιμές του βέλτιστου μοντέλου για t+1.

5.2.2. Χρονικός ορίζοντας t+3

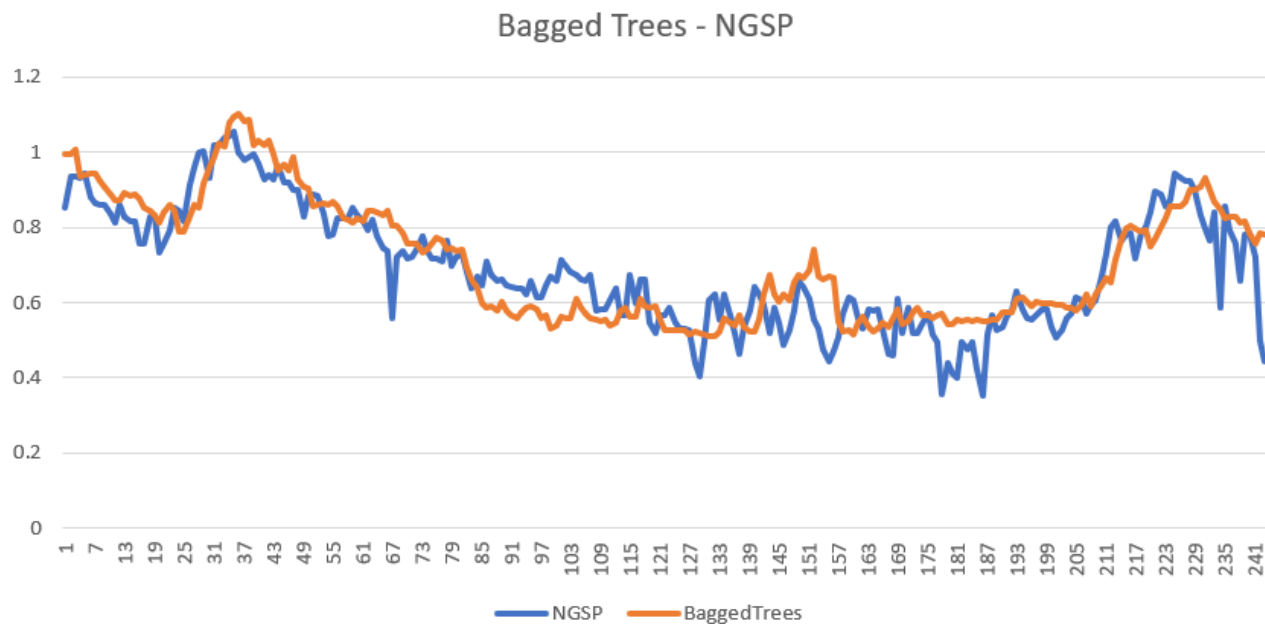
Εξετάζοντας τα δεδομένα του Πίνακα 5 παρατηρούμε πως το βέλτιστο μοντέλο για τον χρονικό ορίζοντα t+3, με βάση τα in-sample δεδομένα, είναι το Bagged Trees με $RMSE = 0,057793$ στα training data. Ταυτόχρονα, το μικρότερο RMSE στα out-of-sample δεδομένα το παρουσιάζει το Boosted Trees με $RMSE = 0,077136$. Σύμφωνα με τις παρατηρήσεις αυτές λοιπόν είναι ξεκάθαρο πως το καλύτερο μοντέλο τον χρονικό ορίζοντα t+3 αποτελεί το Bagged Trees, ενώ δεύτερο καλύτερο θεωρείται το Robust Linear με IS $RMSE = 0,062508$. Μετά το Boosted trees, ακολουθεί το Linear SVM με $RMSE = 0,078857$ στα testing data και ταυτόχρονα αποτελεί το τρίτο καλύτερο του IS με $RMSE = 0,062709$.

Πίνακας 5. Τα RMSE για IS και OOS για τον χρονικό ορίζοντα t+3.

t+3	In sample	OOS
Models	RMSE	RMSE
RW	0.073809	0.080507
Linear Regression	0.063132	0.082556
Robust Linear	0.062508	0.080896
Fine Tree	0.069429	0.110016
Medium Tree	0.067642	0.096673
Coarse Tree	0.068887	0.087867
Linear SVM	0.062709	0.078857
Boosted Trees	0.07715	0.077136
Bagged Trees	0.057793	0.081771



Γράφημα 3. Η σύγκριση των RMSE ανάμεσα σε IS και OOS για τον χρονικό ορίζοντα t+3.



Γράφημα 4. Σύγκριση των actual τιμών του στόχου (natural gas spot prices) με τους fitted τιμές του βέλτιστου μοντέλου για t+3.

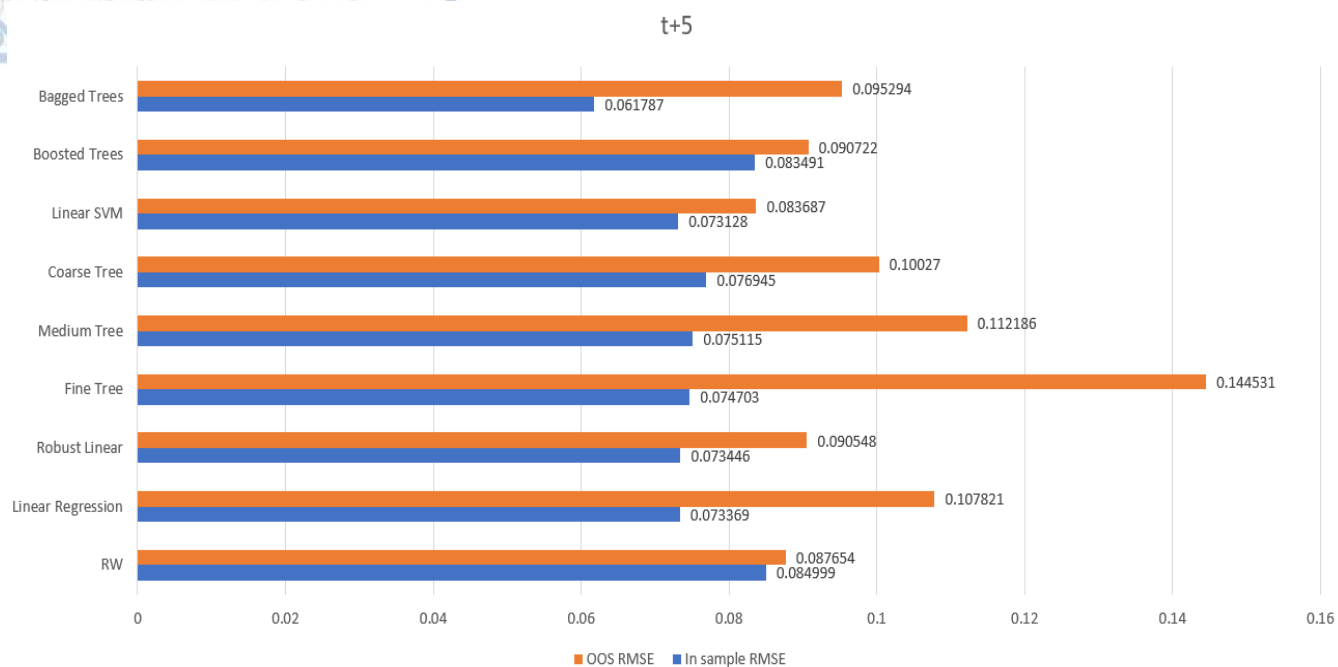
Αξίζει να σημειωθεί, ότι στις δυο αυτές χρονικές περιστάσεις ($t+1$ και $t+3$), το RMSE των Boosted Trees στο OOS ήταν το χαμηλότερο από το αντίστοιχο του in-sample. Ωστόσο το φαινόμενο αυτό παρατηρείται επειδή υπάρχει εξάρτηση από το δείγμα, καθώς τα «άγνωστα» δεδομένα που χρησιμοποιήθηκαν για την out-of-sample πρόγνωση δεν χρησιμοποιήθηκαν κατά την διάρκεια της φάσης εκπαίδευσης των μοντέλων.

5.2.3. Χρονικός ορίζοντας $t+5$

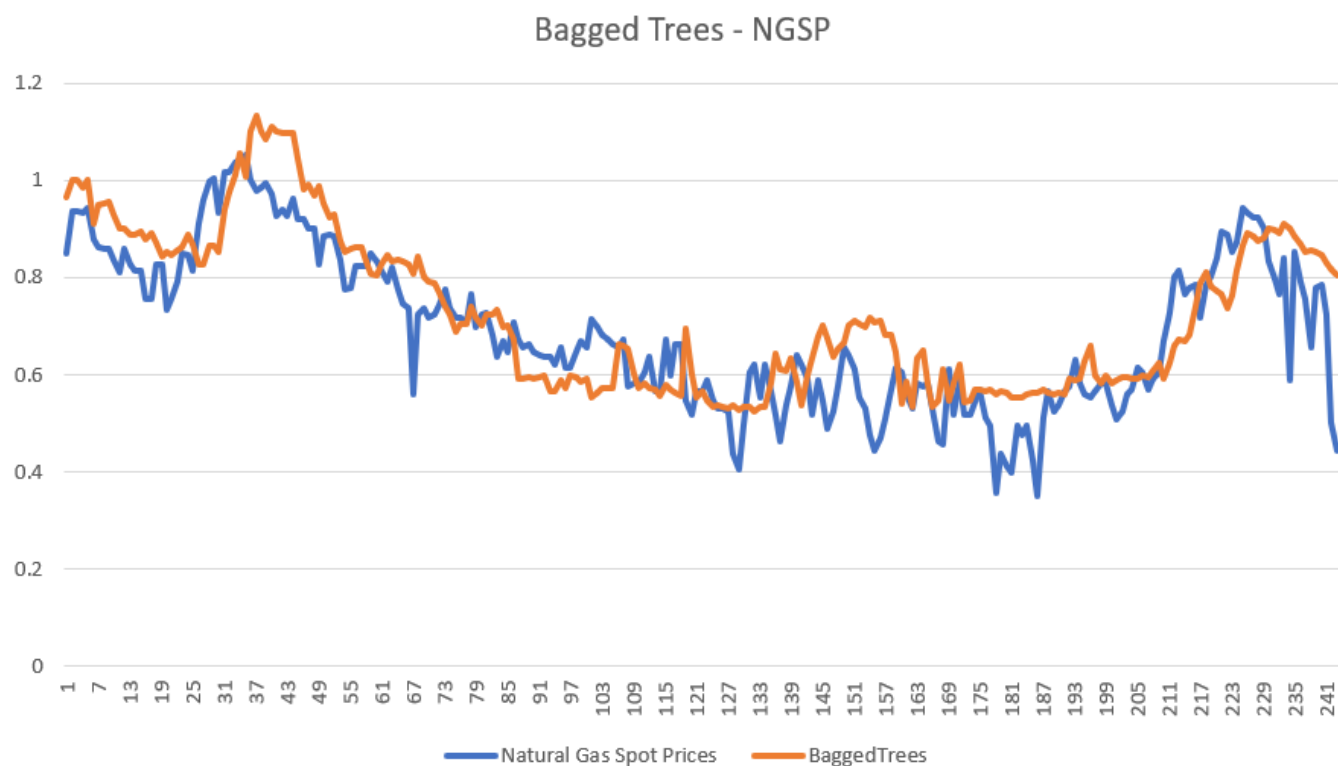
Από την ανάλυση των δεδομένων του Πίνακα 6, προκύπτει ότι το μοντέλο με το μικρότερο δυνατό σφάλμα στα in sample data είναι το Bagged Trees για τον χρονικό ορίζοντα $t+5$. Επιπλέον το μοντέλο του Linear SVM παρουσιάζει πολύ καλή προγνωστική ικανότητα καθώς έχει το καλύτερο $RMSE = 0,083687$ στο out-of-sample data και ταυτόχρονα το δεύτερο καλύτερο στο in-sample data. Καταλήγουμε ότι το βέλτιστο μοντέλο αποτελεί το Bagged Trees με in-sample $RMSE = 0,061787$, ενώ δεύτερο καλύτερο είναι το Linear SVM σύμφωνα με τα training data. Αξίζει επιπλέον να σημειωθεί πως το Robust Linear παρουσιάζει καλά αποτελέσματα, καθώς κατέχει το τρίτο χαμηλότερο out-of-sample $RMSE = 0,090548$.

Πίνακας 6. Τα RMSE για IS και OOS για τον χρονικό ορίζοντα $t+5$.

$t+5$	In sample	OOS
Models	RMSE	RMSE
RW	0.084999	0.087654
Linear Regression	0.073369	0.107821
Robust Linear	0.073446	0.090548
Fine Tree	0.074703	0.144531
Medium Tree	0.075115	0.112186
Coarse Tree	0.076945	0.10027
Linear SVM	0.073128	0.083687
Boosted Trees	0.083491	0.090722
Bagged Trees	0.061787	0.095294



Γράφημα 5. Η σύγκριση των RMSE ανάμεσα σε IS και OOS για τον χρονικό ορίζοντα t+5.

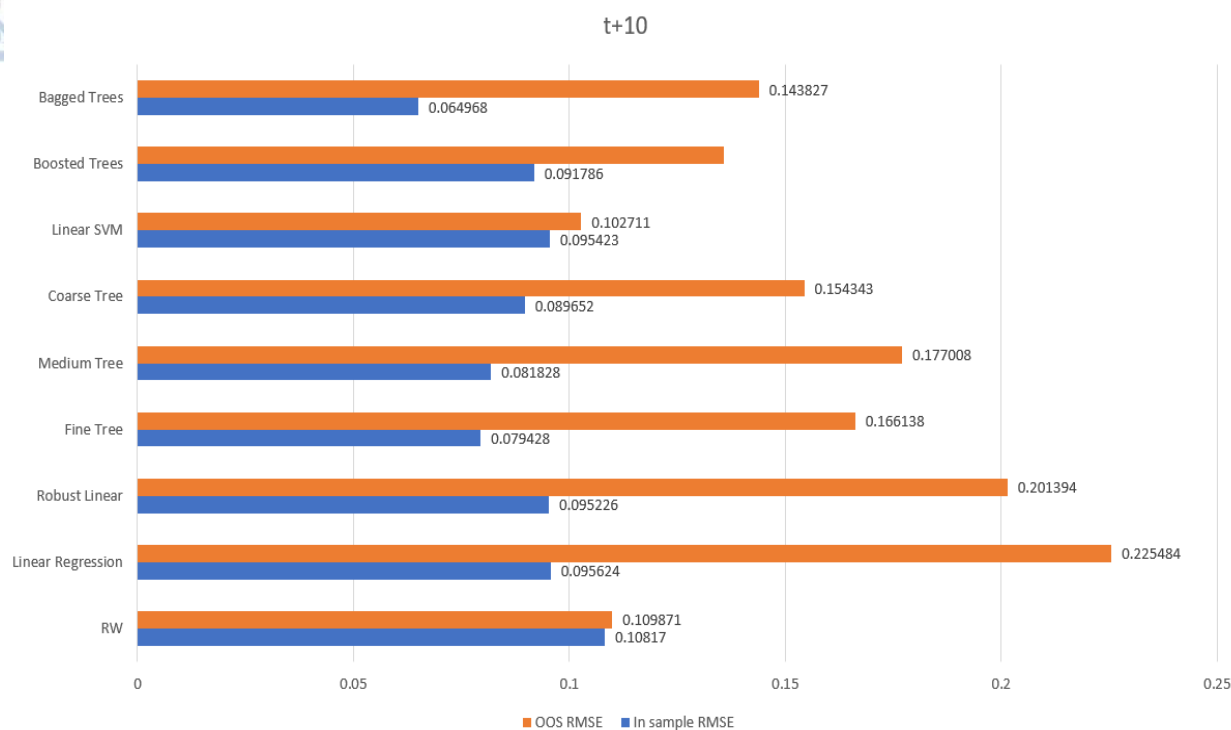


Γράφημα 6. Σύγκριση των actual τιμών του στόχου (natural gas spot prices) με τις fitted τιμές του βέλτιστου μοντέλου για t+5.

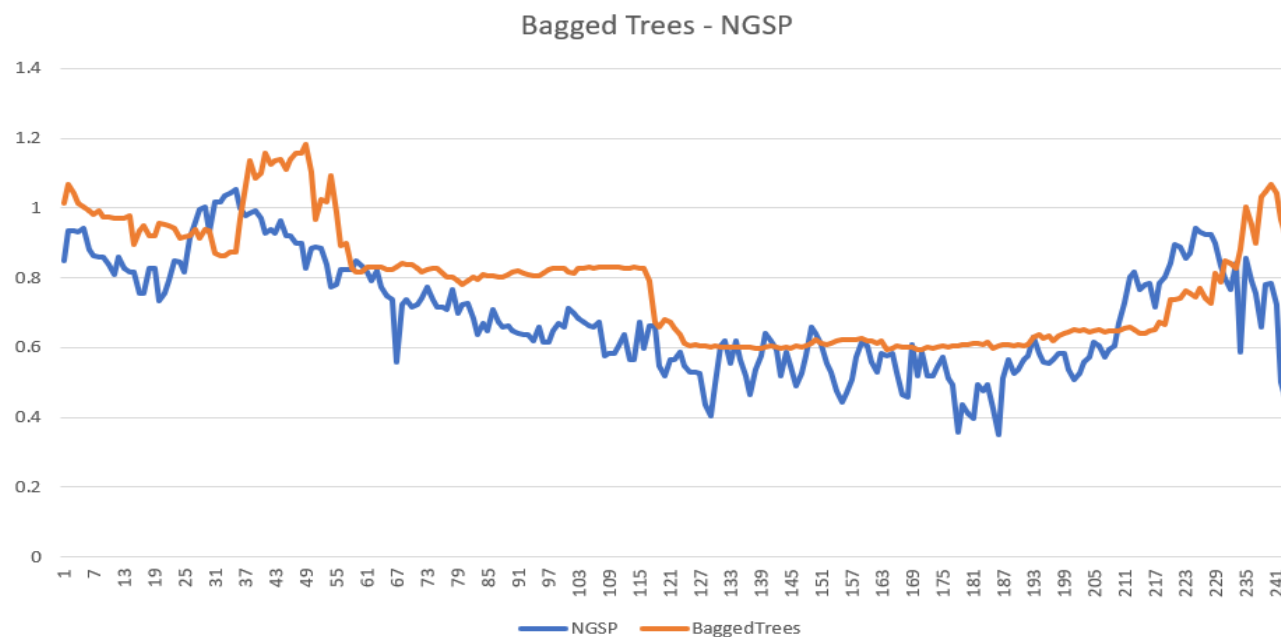
Εξετάζοντας τα δεδομένα του Πίνακα 7 για τον χρονικό ορίζοντα t+10 μπορούν να εξαχθούν κάποια βασικά συμπεράσματα. Πρώτον παρατηρείται πως καλύτερο μοντέλο αποτελεί αυτό του Bagged Trees με $RMSE = 0,064968$, σύμφωνα με τα training data (in-sample). Επιπλέον διακρίνεται πως δεύτερο αποτελεσματικότερο μοντέλο είναι το Fine Tree με in-sample $RMSE = 0,079428$. Αξίζει να σημειωθεί τέλος πως το μοντέλο με το μικρότερο $RMSE$ στα testing data αποτελεί το Linear SVM με $RMSE = 0,102711$.

Πίνακας 7. Τα $RMSE$ για IS και OOS για τον χρονικό ορίζοντα t+10.

Models	In sample	OOS
	RMSE	RMSE
RW	0.10817	0.109871
Linear Regression	0.095624	0.225484
Robust Linear	0.095226	0.201394
Fine Tree	0.079428	0.166138
Medium Tree	0.081828	0.177008
Coarse Tree	0.089652	0.154343
Linear SVM	0.095423	0.102711
Boosted Trees	0.091786	0.135707
Bagged Trees	0.064968	0.143827



Γράφημα 7. Η σύγκριση των RMSE ανάμεσα σε IS και OOS για τον χρονικό ορίζοντα t+10.



Γράφημα 8. Σύγκριση των actual τιμών του στόχου (natural gas spot prices) με τις fitted τιμές του βέλτιστου μοντέλου για t+10.



Ενδιαφέρον παρουσιάζει το γεγονός ότι το μοντέλο Linear SVM ξεπερνάει σε όλες τις χρονικές περιστάσεις το μοντέλο του Random Walk στο out-of-sample dataset, από άποψη σφάλματος κατά την πρόγνωση. Παρόλα αυτά το Random Walk που δημιουργήσαμε παρουσιάζει πολύ καλά αποτελέσματα σε όλες τις χρονικές στιγμές, ενώ ταυτόχρονα συμπεραίνουμε πως όλα τα γραμμικά μοντέλα έχουν την δυνατότητα να προβλέψουν με υψηλή ακρίβεια τις τιμές του natural gas και παρουσιάζουν πολύ καλή απόδοση με μικρές τιμές RMSE. Πολύ καλή προγνωστική ικανότητα παρουσιάζουν και τα Bagged Trees, έχοντας το μικρότερο training σφάλμα σε όλες τις χρονικές περιστάσεις εκτός από την $t+1$. Συνοψίζοντας μπορεί να εξαχθεί το βασικό συμπέρασμα πως όσο απομακρυνόμαστε χρονικά τόσο αυξάνεται το RMSE των μοντέλων και επομένως τόσο χειροτερεύει η ακρίβεια πρόγνωσης.

Η αποτελεσματική πρόγνωση της μελλοντικής μεταβολής της τιμής του φυσικού αερίου αποκτά ολοένα και μεγαλύτερη σημασία, καθώς μειώνει σημαντικά τα εμπόδια για την προμήθεια φυσικού αερίου σε πελάτες σε εθνικό επίπεδο, ενισχύοντας τον ανταγωνισμό. Επιτρέπει, επιπλέον, στους επενδυτές να διαχειρίζονται με βέλτιστο τρόπο την στρατηγική τους.

Στη παρούσα διπλωματική μελετήθηκε η αποτελεσματικότητα των αλγορίθμων μηχανικής μάθησης στην πρόγνωση του natural gas spot prices, με την χρήση 21 ερμηνευτικών μεταβλητών (regressors) που επιλέχθηκαν με βάση την οικονομική θεωρία, καθώς σύμφωνα με την σχετική βιβλιογραφία επηρεάζουν την διακύμανση της τιμής του Natural Gas Spot Price. Ο ορίζοντας πρόβλεψης ήταν μία, τρεις, πέντε και δέκα μέρες μετά ($t+1$, $t+3$, $t+5$ και $t+10$). Συγκεκριμένα χρησιμοποιήθηκαν μακροοικονομικοί και χρηματιστηριακοί δείκτες, συναλλαγματικές ισοτιμίες, επιτόκια, οι spot τιμές και τα future contracts του Oklahoma West Texas Intermediate Crude Oil, καθώς και τα αντίστοιχα future contracts του natural gas.

Για την εκπαίδευση των μοντέλων συγκεντρώθηκε ένα dataset 2423 ημερήσιων τιμών για την χρονική περίοδο 3 Δεκεμβρίου 2010 έως 18 Σεπτεμβρίου 2020. Το dataset αυτό χωρίστηκε σε δύο υποσύνολα, το πρώτο από τις 19 Νοεμβρίου 2010 έως τις 19 Σεπτεμβρίου του 2019, ήτοι 2180 τιμές για την εκπαίδευση των μοντέλων και το δεύτερο μέρος που αποτελείται από τις 20 Σεπτεμβρίου του 2019 έως τις 18 Σεπτεμβρίου του 2020, ήτοι οι υπόλοιπες 243 τιμές για τον έλεγχο της ικανότητας γενίκευσης των μοντέλων σε άγνωστα δεδομένα που δεν χρησιμοποιήθηκαν στην διαδικασία της εκπαίδευσης των μοντέλων πρόβλεψης. Οι χρονικές υστερήσεις του natural gas spot price χρησιμοποιήθηκαν για να αξιοποιήσουμε στο μοντέλο εκπαίδευσης πιθανή μακρά μνήμη (long range dependence) στον μηχανισμό που δημιουργεί τις τιμές του natural gas spot price. Προκειμένου να αποφύγουμε την περίπτωση της υπερεκπαίδευσης (overfitting), χρησιμοποιήθηκε η μέθοδος του 5-fold cross validation. Η διαδικασία αυτή επιτρέπει την επιλογή μοντέλου όπου υπάρχει ισορροπία στην ακρίβεια πρόβλεψης in-sample και out-of-sample. Έτσι επιτυγχάνεται μια ισορροπημένη σχέση bias που αφορά το σφάλμα πρόβλεψης στα in-sample δεδομένα και variance που αφορά το σφάλμα πρόβλεψης στα out-of-sample δεδομένα.

Για την εύρεση του βέλτιστου AR (αυτοπαλίνδρομου μοντέλου) χρησιμοποιήθηκε το Linear SVM με την προσθήκη ενός επιπλέον lag σε κάθε βήμα. Έτσι, καταλήξαμε ότι ο άριστος αριθμός

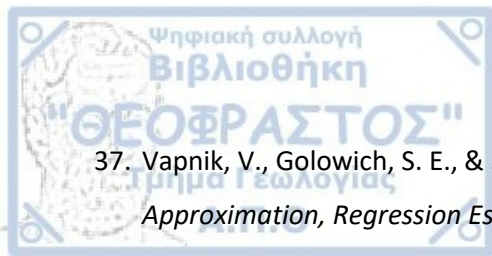
υστερήσεων είναι 14. Στο βέλτιστο αυτό AR μοντέλο λοιπόν προστέθηκαν οι επεξηγηματικές μεταβλητές και πραγματοποιήθηκε η εκπαίδευση των μοντέλων μηχανικής μάθησης. Αξίζει να σημειωθεί πως τα μοντέλα Interactions Linear, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM, Squared Exponential, Matern 5/2 GPR, Exponential GPR και Rational Quadratic παρουσίασαν overfit και για τον λόγο αυτό απορρίφθηκαν. Σύμφωνα με τα αποτελέσματα, από όλα τα μοντέλα που επιλέχθηκαν για την πρόγνωση, αυτά που δεν παρουσίασαν overfitting και είχαν την καλύτερη απόδοση πρόγνωσης από άποψη μικρότερου σφάλματος (RMSE) ήταν τα Bagged Trees και Linear Regression στους αντίστοιχους ορίζοντες πρόβλεψης ($t+1$, $t+3$, $t+5$, $t+10$) που επιλέξαμε να μελετηθούν. Μια απλή γραμμική παλινδρόμηση στην ουσία παρουσιάζει καλύτερα αποτελέσματα συγκριτικά με τις υπόλοιπα μοντέλα που εκπαιδεύσαμε, ακόμα και από τα μοντέλα του machine learning. Από την έρευνά μας λοιπόν καταλήγουμε ότι για τον βραχυπρόθεσμο χρονικό ορίζοντα $t+1$ το μοντέλο που πραγματοποιεί πρόγνωση με μεγαλύτερη επιτυχία στις τιμές του φυσικού αερίου είναι το Linear Regression, ενώ για τους πιο απομακρυσμένους χρονικούς ορίζοντες $t+3$, $t+5$ και $t+10$ το πιο αποτελεσματικό μοντέλο είναι το Bagged Trees.

1. Abdel-Aal, R. E. (2004). Short-Term Hourly Load Forecasting Using Abductive Networks. *IEEE TRANSACTIONS ON POWER SYSTEMS*, 19(1), 164-173. doi: 10.1109/TPWRS.2003.820695
2. Aksoy, B., & Selbas, R. (2019, June 14). Estimation of Wind Turbine Energy Production Value by Using Machine Learning Algorithms and Development of Implementation Program. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 1-13.
doi:<https://doi.org/10.1080/15567036.2019.1631410>
3. Amadeo, K. (2020, September 17). *The S&P 500 and How It Works*. Retrieved from The Balance: <https://www.thebalance.com/what-is-the-sandp-500-3305888>
4. Athanasiou, A. F., Gogas, P., & Papadimitriou, T. (2020, December). Forecasting S&P 500 spikes: an SVM approach. *Digital Finance*. doi:10.1007/s42521-020-00024-0
5. Blaskowitz, O., & Herwartz, H. (2011). On economic evaluation of directional forecasts. *International Journal of Forecasting*, 27(4), 1058-1065. doi:10.1016/j.ijforecast.2010.07.002
6. Buchanan, W. K., Hodges, P., & Theis, J. (2001, May). Which way the natural gas price: an attempt to predict the direction of natural gas spot price movements using trader positions. *Energy Economics*, 23(3), 279-293. doi:10.1016/S0140-9883(00)00074-8
7. Caplinger, D. (2016, July 12). *What Are Crude Oil Futures and How Do They Work?* Retrieved from TheMotleyFool: <https://www.fool.com/investing/2016/07/12/what-are-crude-oil-futures-and-how-do-they-work.aspx#:~:text=Crude%20oil%20futures%20are%20futures,given%20date%20in%20the%20future.>
8. Čeperić, E., Čeperić, V., & Žiković, S. (2017, September 8). Short-term forecasting of natural gas prices using machine learning and feature selection algorithms. (H. Lund, Ed.) *Energy*, 140, 893-900.
9. Chen, B. J., Chang, M. W., & Lin, C. J. (2004, November). Load forecasting using support vector Machines: a study on EUNITE competition 2001. *IEEE Transactions on Power Systems*, 19(4), 1821-1830. doi:10.1109/TPWRS.2004.835679
10. Chen, J. (2019, June 25). *Nasdaq Composite Index*. Retrieved from Investopedia: <https://www.investopedia.com/terms/n/nasdaqcompositeindex.asp>



11. Chen, J. (2020, January 31). *Exchange Rate Definition*. Retrieved from Investopedia:
<https://www.investopedia.com/terms/e/exchangerate.asp>
12. Chen, J. (2020, September 22). *Federal Funds Rate*. Retrieved from Investopedia :
<https://www.investopedia.com/terms/f/federalfundsrate.asp>
13. Chen, J. (2020, February 21). *One-Year Constant Maturity Treasury*. Retrieved from Investopedia: <https://www.investopedia.com/terms/c/cmtindex.asp>
14. Chen, J. (2020, June 30). *Prime Rate*. Retrieved from Investopedia:
<https://www.investopedia.com/terms/p/primerate.asp>
15. Chen, J. (2020, December 31). *West Texas Intermediate (WTI)* . Retrieved from Investopedia:
<https://www.investopedia.com/terms/w/wti.asp>
16. De Felice, M., & Xin Yao. (2011, August). Short-Term Load Forecasting with Neural Network Ensembles: A Comparative Study [Application Notes]. *IEEE Computational Intelligence Magazine*, 6(3), 47-56. doi:10.1109/MCI.2011.941590
17. Dimitriadou, A., Gogas, P., Papadimitriou, T., & Plakandaras, V. (2018, October 13). Oil Market Efficiency under a Machine Learning Perspective. *Forecasting*, 1(1), 157-168.
doi:10.3390/forecast1010011
18. (n.d.). *Dow Jones Industrial Average (DJIA)*. Corporate Finance Institute. Retrieved January 18, 2021, from Corporate Finance Institute:
<https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/dow-jones-industrial-average-djia/>
19. Drucker, H., Burges, J. C., Kaufman, L., Smola, A., & Vapnik, V. (1996). *Support Vector Regression Machines*.
20. Fama, E. F. (1970, May). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, 25(2), 383-417.
21. *Interest Rate Spreads*. (2021, January 15). Retrieved from FRED:
[https://fred.stlouisfed.org/series/T5YIE#:~:text=The%20breakeven%20inflation%20rate%20represents,Constant%20Maturity%20Securities%20\(TC_5YEAR\).](https://fred.stlouisfed.org/series/T5YIE#:~:text=The%20breakeven%20inflation%20rate%20represents,Constant%20Maturity%20Securities%20(TC_5YEAR).)
22. Jordan, M. I., & Mitchell, T. M. (2015, July 17). Science. *Machine learning: Trends, perspectives, and prospects*, 349(6245), 255-260.
23. Liu, K., Subbarayan, S., Shoults, S. S., Manry, M. T., Kwan, C., Lewis, F. I., & Naccarino, J. (1996, May). Comparison of very short-term load forecasting techniques. *IEEE Transactions on Power Systems*, 11(2), 877-882. doi:10.1109/59.496169

24. Mishra, V., & Smyth, R. (2016, April). Are natural gas spot and futures prices predictable? *Economic Modelling*, 54, 178-186. doi:10.1016/j.econmod.2015.12.034
25. Mohandes, M. (2002). Support Vector Machines for short-term electrical load forecasting. *Energy Research*, 26(4), 335-345. doi:doi.org/10.1002/er.787
26. Nguyen, H. T., & Nabney, I. T. (2010, September). Short-term electricity demand and gas price forecasts using wavelet transforms and adaptive models. *Energy*, 35(9), 3674-3685. doi:10.1016/j.energy.2010.05.013
27. NYMEX: Your Home for Henry Hub Natural Gas. (n.d.). Retrieved from cmegroup: <https://www.cmegroup.com/trading/energy/nymex-natural-gas-futures.html>
28. Park, D. C., El-Sharkawi, M. A., Marks, R. J., Atlas, L. E., & Damborg, M. J. (1991, May). Electric load forecasting using an artificial neural network. *IEEE Transactions on Power Systems*, 6(2), 442-449. doi:10.1109/59.76685
29. Peterson, K. (2018, September). Resting Heart Rate Variability Can Predict Track and Field Sprint Performance. *OA Journal - Sports*, 1.
30. Plakandaras, V., Rangan, G., Gogas, P., & Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic Modelling*, 45, 259-267.
31. Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. In *Advanced Lectures on Machine Learning* (pp. 63-71).
32. Salehnia, N., Falahi, M. A., Seifi, A., & Mahdavi Adeli, M. H. (2013, September). Forecasting natural gas spot prices with nonlinear modeling using Gamma test analysis. *Journal of Natural Gas Science and Engineering*, 14, 238-249. doi:10.1016/j.jngse.2013.07.002
33. Seeger, M. (2004). Gaussian Processes for Machine Learning. *International Journal of Neural Systems*, 24(2), 69-106. doi:10.1142/S0129065704001899
- Serletis, A., & Shahmoradi, A. (2006, September). Returns and volatility in the NYMEX Henry Hub natural gas futures market. *OPEC Energy Review*, 30(3). doi:10.1111/j.1468-0076.2006.00167.x
34. Statistics Solutions. (2013). Retrieved from <https://www.statisticssolutions.com/what-is-linear-regression/>
35. Su, M., Zhang, Z., Zhu, Y., Zha, D., & Wen, W. (2019, May 3). Data Driven Natural Gas Spot Price Prediction Models Using Machine Learning Methods. *Energies*, 12(9). doi:org/10.3390/en12091680
36. Tan, Y. V., & Roy, J. (2019). Bayesian additive regression trees and the General BART model. In *Statistic in Medicine* (pp. 5048-5069).



37. Vapnik, V., Golowich, S. E., & Smola, A. (1996). *Support Vector Method for Function Approximation, Regression Estimation and Signal Processing*.
38. Ying , C., Luh, P., Yige, Z., Michel , L., Coolbeth, M., Guan, C., . . . Rourke, S. J. (2010, February 322-330). Short-Term Load Forecasting: Similar Day-Based Wavelet Neural Networks. *IEEE Transactions on Power Systems*, 25(1).
39. Zimmermann, A. (2008). Ensemble-Trees: Leveraging Ensemble Power Inside Decision Trees. In *Lecture Notes in Computer Science* (Vol. 5255, pp. 76-87). Springer, Berlin, Heidelberg.
doi:https://doi.org/10.1007/978-3-540-88411-8_10