**1.Bιβλιοθήκες**

* **requests:** χρησιμοποιείται για να κάνει HTTP αίτημα στο API της Alpha Vantage και να πάρει τα δεδομένα της μετοχής της google με μετοχικό σύμβολο GOOGL**.**
* **json:** χρησιμοποιείται για να διαχειριστεί αρχικά τα δεδομένα που είναι σε μορφή JSON.
* **csv:** χρησιμοποιείται για την ανάγνωση και την εγγραφή στο αρχείο csv στο οποίο βασίζεται το project (close\_prices.csv).
* **pandas:** χρησιμοποιεί DataFrames για την ανάλυση και διαχείριση δεδομένων.
* **matplotlib:** χρησιμοποιείται για την δημιουργία γραφημάτων (συνάρτηση **pyplot**) καθώς και για την πρόσβαση σε colormaps που επιτρέπουν την εφαρμογή χρωμάτων στα δεδομένα σε γραφήματα (συνάρτηση **cm**).
* **statsmodels:** παρέχει εργαλεία για στατιστική ανάλυση χρονοσειρών. Η συνάρτηση **seasonal\_decompose** χρησιμοποιείται για την αποσύνθεση της χρονοσειράς σε trend, seasonality και residual. Επίσης, χρησιμοποιείται η συνάρτηση **plot\_acf**, η οποία δημιουργεί ένα γράφημα που απεικονίζει την αυτοσυσχέτιση της χρονοσειράς για διάφορες τιμές lag.
* **scipy:** από συγκεκριμένη βιβλιοθήκη χρησιμοποιείται η συνάρτηση **gaussian\_filter1d** για την εφαρμογή Gaussian φίλτρου στα δεδομένα ώστε να τα εξομαλύνει.
* **sklearn:** από τηνσυγκεκριμένη βιβλιοθήκη χρησιμοποιούνται οι συναρτήσεις:
* **LinearRegression** είναι για την δημιουργία του μοντέλου γραμμικής παλινδρόμησης.
* **Lasso** είναι για την δημιουργία του μοντέλου που χρησιμοποιεί L1 κανονικοποίηση.
* **Ridge** είναι για την δημιουργία του μοντέλου που χρησιμοποιεί L2 κανονικοποίηση.
* **PolynomialFeatures** χρησιμοποιείται για να μετατρέψει τα χαρακτηριστικά σε πολυωνυμικά χαρακτηριστικά.
* **mean\_squared\_error, mean\_absolute\_error** είναι για τον υπολογισμό του μέσου τετραγωνικού σφάλματος και του μέσου απόλυτου σφάλματος.
* **numpy:** χρησιμοποιείται για επιστημονικούς υπολογισμούς και πιο συγκεκριμένα για τον υπολογισμό της τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος (rmse).
* **joblib:** χρησιμοποιείται για την αποθήκευση και φόρτωση μοντέλων σε δυαδική μορφή.

**2.Προεπεξεραγασία**

Αρχικά, δημιουργείται η συνάρτηση αυτοσυχέτισης (**ACF**) χρησιμοποιώντας την βιβλιοθήκη **statsmodels**, προκειμένου να αποτυπωθεί η αυτοσυσχέτιση των δεδομένων της χρονοσειράς σε διαφορετικά χρονικά διαστήματα lags. Πιο συγκεκριμένα εξετάζει πόσο σχετίζεται μια τιμή με τις τιμές πριν από t ημέρες.

Παρατηρήσεις που προκύπτουν από το γράφημα:

1. Η ACF ξεκινά κοντά στο 1 στο lag 0, υποδεικνύοντας ισχυρή αυτοσυσχέτιση στην αρχή, κάτι που είναι αναμενόμενο, καθώς μια χρονοσειρά συσχετίζεται πάντα τέλεια με τον εαυτό της. Για μικρές τιμές εξακολουθεί να υπάρχει ισχυρή συσχέτιση.
2. Καθώς αυξάνουμε τα lag , η αυτοσυσχέτιση μειώνεται σταδιακά. Αυτό υποδηλώνει ότι οι ημερήσιες τιμές κλεισίματος της μετοχής της Google συσχετίζονται σε μεγάλο βαθμό σε μικρά χρονικά διαστήματα.
3. Η ACF φθίνει σταδιακά, μέχρι να πλησιάσει σχεδόν το μηδέν. Αυτό δείχνει ότι σε μαγαλύτερα χρονικά διαστήματα η συσχέτιση μεταξύ των τιμών μειώνεται.

A graph with a blue line

Description automatically generated

Ακόμη, αποτυπώνονται στην παρακάτω γραφική παράσταση οι τιμές κλεισίματος καθώς οι εξομαλυμένες τιμές κλεισίματος που προκύπτουν από την χρήση gaussian φίλτρου με sigma=3. Η συγκεκριμένη γραφική δίνει μια συνολική εικόνα της συμπεριφοράς των τιμών κλεισίματος της μετοχής. Παρατηρείται:

1. Μέχρι το 2020 οι τιμές κλεισίματος εμφανίζουν διακυμάνσεις, αλλά και μια σταθερή ανοδική τάση.
2. Περίπου το 2021 παρατηρείται απότομη άνοδος της μετοχής.
3. Τέλος, στα μέσα του 2022 υπάρχει ξαφνική πτώση της μετοχής, η οποία στην συνέχεια σταθεροποιείται σε πολύ χαμηλές τιμές με ελάχιστες διακυμάνσεις σε σύγκριση με τις προηγούμενες περιόδους.

A graph with orange and blue lines

Description automatically generated

Επιλέχθηκε η στατιστική μέθοδος **Seasonal Decomposition** χρησιμοποιώντας το ρποσθετικό μοντέλο (additive)**,** ώστε να διασπαστεί η χρονοσειρά στα βασικά της συστατικά (observed, trend, seasonal, residual). Με αυτόν τον τρόπο θα κατανοήσουμε καλύτερα την δομή και την τάση της μετοχής.

1. Aποτυπώνονται τα αρχικά δεδομένα (Observed), τα οποία αναλύθηκαν παραπάνω.

A graph showing the growth of the stock market

Description automatically generated

1. Στην επόμενη γραφική ( Trend) αποτυπώνεται η μακροχρόνια συμπεριφορά της χρονοσειράς αφαιρώντας τι εποχικές διακυμάνσεις και τον θόρυβο. Παρατηρείται ότι μέχρι το 2021 η τάση είναι ανοδική και στην συνέχεια υπάρχει απότομη πτώση.

A graph showing the growth of the stock market

Description automatically generated

1. Στην γραφική Seasonal αποτυπώνεται η ετήσια εποχικότητα, η οποία είανι η επαναλαμβανόμενη διακύμανση των τιμών. Παρατηρείται ότι υπάρχει τακτική περιοδικότητα στις διακυμάνσεις των τιμών που επαναλαμβάνεται κάθε χρόνο.

A graph of a graph of a graph

Description automatically generated with medium confidence

1. Τέλος η γραφική Residual αποτυπώνει τις τυχαίες διακυμάνσεις μετά την αφαίρεση της τάσης και της εποχικότητας. Μεγάλη αύξηση στις αποκλίσεις συμβαίνει κατα την περόδο του 2022 που υπήρχε η μεγάλη πτώση, τα προηγούμενα χρόνια τα residuals είναι σχετικά σταθερά.

A graph of a graph of a graph

Description automatically generated with medium confidence

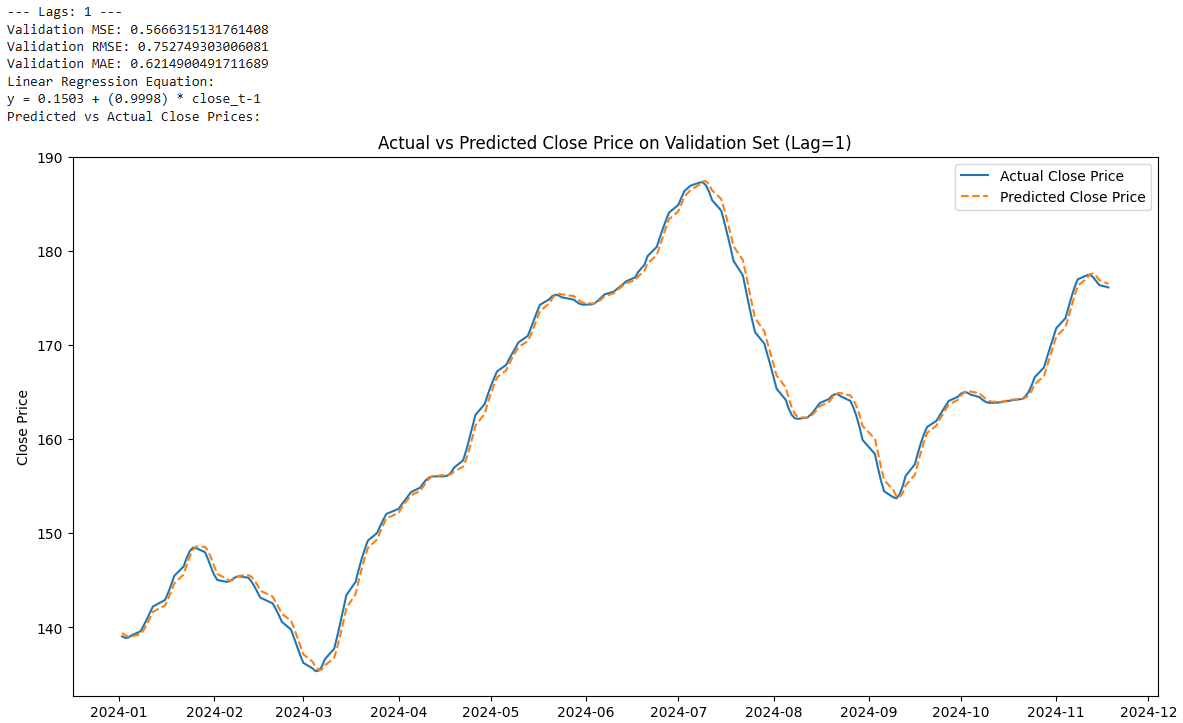
**3.Linear Regression**

**3.1 TRAIN-VALIDATION**

Το σύνολο των δεδομένων που χρησιμοποιούμε περιέχει τιμές κλεισίματος της μετοχής της GOOGL CLASS A απο το 2004 μέχρι 15-11-2024. Αρχικά, εφαρμόζεται φιλτράρισμα στα δεδομένα χρησιμοποιώντας το gaussian filter με sigma=3 και δημιουργείται στο dataframe μια στήλη closed\_smoothed που περιέχει τις τιμές που θα χρησιμοποιηθούν. Στην συνέχεια, στο dataframe δημιουργούνται και στήλες που θα περιέχουν τα lagged features δηλαδή τις παρελθοντικές τιμές. Προκειμένου να βρεθεί το σωστό πλήθος lags για την μετοχή θα δοκιμαστεί το μοντέλο για διαφορετικό πλήθος παρελθοντικών τιμών απο 1(close\_t-1) μέχρι 7 (close\_t-1…close\_t-7). Επιπλέον, διασπάτε το dataframe σε σύνολο εκπαίδευσης (train set) και σύνολο επικύρωσης (validation set), το train set θα περιέχει τιμές πριν το 2024 και το validation set θα περιέχει τις τιμές μετά το 2024 μέχρι 18-11-2024.

* Το Χ\_train, Χ\_validation είναι τα lags και το y\_train, y\_validation είναι οι στόχοι, δηλαδή οι τιμές κλεισίματος της επόμενης μέρας.

Ύστερα, εκπαιδεύεται το μοντέλο γραμμικής παλινδρόμησης στα δεδομένα εκπαίδευσης. Επίσης, πραγματοποιούνται προβλέψεις στα δεδομένα επικύρωσης και αξιολογείται το μοντέλο χρησιμοποιώντας τις μετρικές Μέσο τετραγωνικό σφάλμα (**MSE**),Τετραγωνική ρίζα του MSE (**RMSE**), Μέσο απόλυτο σφάλμα (**MAE**).

Η διαδικασία επαναλαμβάνεται για πλήθος lags από 1 μέχρι και 7 και τα αποτελέσματα για το κάθε μοντέλο αποθηκεύονται στην λίστα  **results\_list.** Για κάθε μοντέλο εκτυπώνεται η εξίσωση γραμμικής παλινδρόμησης και δημιουργείται η γραφική με τις προβλεπόμενες και τις πραγματικές τιμές για το σύνολο επικύρωσης. Παρατίθενται ενδεικτικά τα αποτελέσματα και οι γραφικές παρακάτω για lags από 1 μέχρι 4.

Εικόνα που περιέχει κείμενο, διάγραμμα, γράφημα, γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο, διάγραμμα, γραμμή, στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο, διάγραμμα, στιγμιότυπο οθόνης, γράφημα

Περιγραφή που δημιουργήθηκε αυτόματα

Τέλος ανακτώνται απο την λίστα results\_list τα σφάλματα (MSE, RMSE, MAE) για τις διάφορες τιμές των lags και δημιουργείται το παρακάτω γράφημα.

Εικόνα που περιέχει κείμενο, γραμμή, γράφημα, διάγραμμα

Περιγραφή που δημιουργήθηκε αυτόματα

Απο το γράγημα φαίνεται ότι οι μετρικές σφάλματος (MSE, RMSE και MAE) μειώνονται σημαντικά από το lag 1 έως το lag 2 και στη συνέχεια φθάνουν σε χαμηλές και σταθερές τιμές από το lag 3 και μετά. Μέχρι το lag 4, τα σφάλματα έχουν φθάσει σε  σταθερή κατάσταση, γεγονός που υποδηλώνει ότι τα πρόσθετα lag  πέραν αυτού του σημείου δεν παρέχουν ουσιαστική βελτίωση της ακρίβειας. Οπότε γι'αυτό και θα χρησιμοποιηθεί το μοντέλο που έχει δημιουργηθεί με τα 4 lag.

Μετά απο αυτήν την παρατήρηση προστέθηκαν στον κώδικα οι εντολές οι οποίες αποθηκεύουν το μοντέλο με lag = 4 χρησιμοποιώντα την βιβλιοθήκη joblib, ώστε να μπορεί να χρησιμοποιηθεί.

    # Store the model for lag 4

    if lags == 4:

        best\_model = model

        joblib.dump(best\_model, 'linear\_regression\_model.pkl')

NA ΠΩ ΓΙΑ MSE OTI EINAI MIKRO

**3.2 TEST**

Χρησιμοποιώντας το μοντέλο που αποθηκεύτηκε όπως αναφέρθηκε και παραπάνω θα προβλεφθεί με βάση τις 4 προηγούμενες τιμες (11-11-2024 μέχρι 18-11-2024) η τιμή για τις 19-11-2024 η οποία δεν βρίσκεται στο σύνολο των δεδομένων. Η τιμή που πρόβλεψε το μοντέλο είναι:



Ενώ η πραγματική τιμή της μετοχής για εκείνη την μέρα είναι+ συμπερασματα