



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Προγραμματιστικά Εργαλεία και Τεχνολογίες για Επιστήμη Δεδομένων
ακαδημαϊκό έτος 2020-21

Title : Time Travel Project

Μπάκος Παναγιώτης
pmpakos@hotmail.com
A.M.: 03002978
PhD Student @ CSLab

1 Εισαγωγή

Στα πλαίσια της συγκεκριμένης εργασίας θα εργαστούμε πάνω στο dataset με τις ιστορικές τιμές των μετοχών του Χρηματιστηρίου της Νέας Υόρκης από το 1962 έως το 2017. Συγκεκριμένα, τα δεδομένα αποτελούνται από logs, ξεχωριστό για καθεμία μετοχή που θα μελετήσουμε, στο οποίο για κάθε ημερομηνία αποθηκεύονται :

- οι τιμές ανοίγματος, κλεισίματος της μετοχής τη συγκεκριμένη μέρα (*Open, Close*)
- η ελάχιστη και η μέγιστη τιμή που έφτασε η μετοχή τη συγκεκριμένη μέρα (*Low, High*)
- ο όγκος αγοραπωλησιών της μετοχής τη συγκεκριμένη μέρα (*Volume*)

Με βάση αυτά τα δεδομένα, καλούμαστε να παράξουμε δύο ακολουθίες, **Small** και **Large**, στις οποίες θα δίνεται η κατάλληλη σειρά κινήσεων (αγοραπωλησίες μετοχών), ώστε να επιτύχουμε μεγάλο κέρδος. Τα όρια για τις δύο ακολουθίες είναι 1K και 1M κινήσεις αντίστοιχα.

Για κάθε ακολουθία ακολουθήθηκε διαφορετική προσέγγιση, γι αυτό και δημιουργήθηκαν δύο ξεχωριστά Jupyter Notebooks. Για τη **Small**, οι μόνες κινήσεις που εξετάσαμε είναι οι "*Buy-Low*" και "*Sell-High*", ενώ για τη **Large** κάνουμε αποκλειστικά Intra-day trading, οπότε οι δύο πιθανοί συνδυασμοί κινήσεων είναι "*Buy-Open -> Sell-High*" και "*Buy-Low -> Sell-Close*". Παρακάτω θα παρουσιαστεί η μέθοδος παραγωγής κάθε ακολουθίας.

2 Προεπεξεργασία dataset

Η μορφή των παρεχόμενων δεδομένων είναι ξεχωριστό txt αρχείο για κάθε μετοχή. Διαβάζουμε επομένως κάθε αρχείο και το περνάμε στο pandas dataframe `stock_data`. Μετατρέπουμε τη στήλη της ημερομηνίας στην κατάλληλη μορφή, ταξινομούμε τις εγγραφές του dataframe με βάση την ημερομηνία, καθώς με τον τρόπο αυτό θα "διασχίσουμε" το dataset, και καθαρίζουμε το dataset από τιμές μετοχών που ισούνται με το 0 σε οποιαδήποτε από τις στήλες *Open, Close, Low, High*. Η μείωση του μεγέθους του dataset μετά από αυτήν την αλλαγή ήταν ελάχιστη, έως μηδαμινή

Επιπροσθέτως, για την ακολουθία **Large** μόνο, λόγω του Intra-day trading που θα ακολουθήσουμε σαν τακτική, δημιουργούμε δύο έξτρα στήλες στο dataframe, μία για κάθε συνδυασμό κινήσεων ("*Buy-Open -> Sell-High*" και "*Buy-Low -> Sell-Close*"), καθώς και δύο στήλες "*MaxIntra*" και "*MaxProfit*", που θα μας υποδεικνύουν ποιες μετοχές να εξετάσουμε κάθε μέρα. Αυτές θα αναλυθούν στην ενότητα που αφορά τη **Large** ακολουθία.

3 Ακολουθία Small

Θέλουμε αρχικά να δημιουργήσουμε μια ακολουθία κινήσεων, με όριο να είναι 1000. Επειδή το εύρος των ημερομηνιών είναι αρκετά μεγάλο, από το 1962 έως το 2017, μπορούμε να ορίσουμε ότι θα σπάσουμε αυτό το εύρος σε χρονικά παράθυρα. Το μέγεθος του παραθύρου ορίζεται ως το floor της διαίρεσης του πλήθους των ημερών του dataset με το πλήθος κινήσεων που έχουμε στη διάθεσή μας (1000), διαιρεμένο δια 2, καθώς εντός κάθε παραθύρου θα εκτελούμε 2 κινήσεις.

Η μία κίνηση θα είναι *Buy-Low*, την ημέρα που η μετοχή εμφανίζει τη χαμηλότερη τιμή εντός του παραθύρου, και η άλλη θα είναι *Sell-High*, την ημέρα που η μετοχή θα εμφανίζει τη μέγιστη τιμή εντός του παραθύρου. Φυσικά, θα πρέπει η ημερομηνία πώλησης να είναι μεταγενέστερη της ημερομηνίας αγοράς (και σίγουρα όχι η ίδια ημέρα, καθώς δεν είναι δυνατόν να γνωρίζουμε ποια από τις δύο τιμές της μετοχής προηγήθηκε μέσα στη μέρα).

Αρχίζουμε να διατρέχουμε το dataset με αυτά τα χρονικά παράθυρα. Κρατάμε υπό εξέταση τις μετοχές που εμφανίζονται εντός του παραθύρου¹. Στη συνέχεια, προσπαθούμε να βρούμε ποια μετοχή θα είναι αυτή που θα μας δώσει το μεγαλύτερο κέρδος, με βάση το balance (υπόλοιπο στον τραπεζικό μας λογαριασμό) που έχουμε διαθέσιμο πριν την έναρξη του συγκεκριμένου παραθύρου.

¹Για την ακολουθία **Small**, επιλέχθηκε να εξεταστούν όλες οι μετοχές, και να μην γίνει κάποια επιλογή στο preprocessing, καθώς υπήρχε η υπολογιστική ικανότητα για να εξεταστούν όλες.

Διατρέχοντας τις εγγραφές της εξεταζόμενης μετοχής, εντοπίζουμε την ημέρα που η μετοχή αυτή εμφανίζει την ελάχιστη τιμή της, καθώς και το μέγιστο κέρδος που μπορεί να προκύψει από μια τέτοια αγοραπωλησία εντός του παραθύρου. Εάν ισχύουν τα εξής:

- η ημερομηνία αγοράς (Low) προηγείται της ημερομηνίας πώλησης (High)
- η αγοραπωλησία μας προσφέρει το μέγιστο κέρδος
- το balance επαρκεί για να αγοράσουμε έστω μία μετοχή της εξεταζόμενης εταιρίας

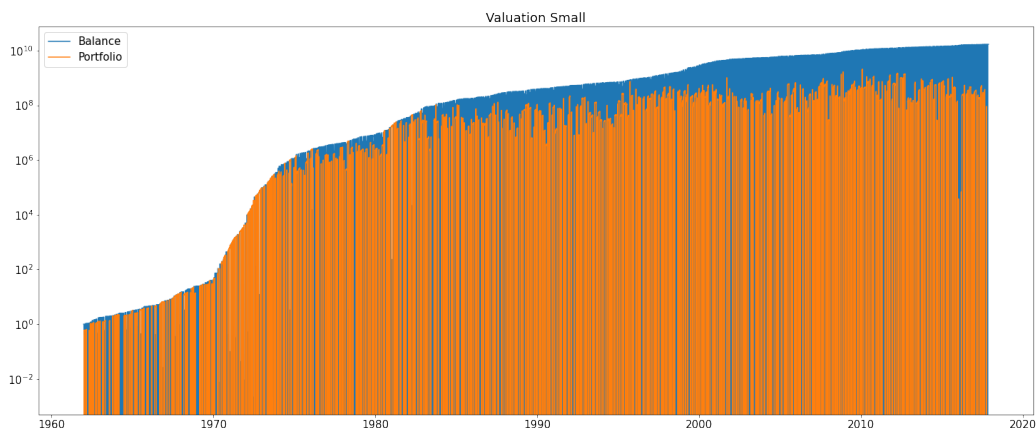
τότε αποφασίζουμε ότι αυτή η μετοχή θα είναι η επιλογή για το συγκεκριμένο παράθυρο, και υπολογίζουμε το πλήθος των μετοχών που θα αγοράσουμε και θα πουλήσουμε. Εδώ μπαίνει το constraint ότι μπορούμε να αγοράσουμε ή να πουλήσουμε έως το 10% του όγκου των μετοχών τις ημέρες αγοράς ή πώλησης αντίστοιχα, το οποίο και λαμβάνουμε υπόψιν. Λαμβάνουμε ακόμη υπόψιν και το ποσοστό προμήθειας που επιβάλλεται σε κάθε αγοραπωλησία μετοχών (1%).

Να σημειωθεί εδώ ότι δεν υπολογίζουμε με τον βέλτιστο τρόπο το ποια μετοχή προσφέρει το μεγαλύτερο κέρδος, καθώς κάνουμε μονό πέρασμα στις εγγραφές κάθε μετοχής. Με ένα διπλό loop θα μπορούσαμε να εντοπίσουμε με καλύτερο τρόπο τη βέλτιστη μετοχή, αλλά το γεγονός ότι έχουμε όλες τις μετοχές υπό εξέταση, θα οδηγούσε σε μεγάλο χρόνο εκτέλεσης και αποφεύχθηκε, μιας και με αυτόν τον τρόπο πετύχαμε ούτως ή άλλως ένα σχετικά καλό αποτέλεσμα.

Μετά από όλα αυτά, το balance που προέκυψε όταν τελείωσε η χρονική διάτρεξη του dataset είναι **16909153723.90599 \$** μετά από **972 κινήσεις**. Ο λόγος που οι κινήσεις είναι λιγότερες από 1000, είναι γιατί σε κάποια παράθυρα το μονό πέρασμα που περιγράφηκε προηγουμένως δεν βρήκε κάποια κίνηση να εκτελέσει.

Πέρα από το κέρδος που παρουσιάσαμε παραπάνω, θέλουμε να έχουμε και εικόνα της αποτίμησης για κάθε μέρα του dataset. Ως αποτίμηση ορίζουμε την αξία των μετοχών που βρίσκονται στην κατοχή μας στο κλείσιμο (*Close*) κάθε μέρας. Επειδή σε κάθε χρονικό παράθυρο έχουμε μόνο μία μετοχή στην κατοχή μας, η δημιουργία της δομής Portfolio είναι ιδιαίτερα απλή, και γίνεται κατά τη διάρκεια του loop που περιγράφηκε προηγουμένως.

Επομένως, για κάθε μέρα, κρατάμε στα dictionaries Balance και Portfolio το **κέρδος** και την **αποτίμηση** ανά ημέρα, τα οποία και θα απεικονίσουμε παρακάτω.



4 Ακολουθία Large

Στο μέρος αυτό της άσκησης επικεντρωνόμαστε στο Intra-day trading. Σύμφωνα με αυτό, αγοράζουμε και πουλάμε την ίδια μέρα κάποια μετοχή, εκμεταλλευόμενοι τη διακύμανση της τιμής της μετοχής εντός της ημέρας. Συγκεκριμένα, οι δύο πιθανοί συνδυασμοί κινήσεων που θα εξετάσουμε είναι "*Buy-Open -> Sell-High*" και "*Buy-Low -> Sell-Close*".

Αρχικά, κατά το διάβασμα του dataset, δημιουργούμε δύο νέες στήλες, τις "**O->H**" και "**S->H**", οι οποίες περιέχουν τη διαφορά που προκύπτει αν εκτελέσουμε κάποιον εκ των δύο συνδυασμών, λαμβάνοντας υπόψιν και την προμήθεια που καλούμαστε να πληρώσουμε για κάθε transaction.

Στη συνέχεια, αφαιρούμε τις γραμμές του dataframe στις οποίες και οι δύο αυτές στήλες είναι αρνητικές ή μηδενικές, καθώς πρόκειται για κινήσεις που ούτως ή άλλως ο αλγόριθμος δεν θα επιλέξει, καθώς θα ασχοληθούμε μόνο με Intra-day trading. Με τον τρόπο αυτό πετυχαίνουμε και αρκετά μεγάλη μείωση του μεγέθους του dataframe που θα εξετάσουμε, σε ποσοστό 65%.

Τέλος, δημιουργούμε δύο επιπλέον στήλες που θα βοηθήσουν στην επιλογή των κατάλληλων μετοχών για κάθε μέρα, τις *MaxIntra* και *MaxProfit*. Η πρώτη περιέχει το max μεταξύ των προηγούμενων δύο πιθανών συνδυασμών κινήσεων, ενώ στη δεύτερη υπολογίζουμε το μέγιστο κέρδος που μπορεί να προκύψει για τη μετοχή αυτή κατά την ημέρα αυτή. Συγκεκριμένα, αυτό υπολογίζεται ως το γινόμενο του *MaxIntra* πολλαπλασιασμένο με το 10% του όγκου της μετοχής, που είναι και το μέγιστο πλήθος που μπορούμε να αγοράσουμε για τη συγκεκριμένη μετοχή.

Στη διάθεσή μας έχουμε 1 εκατομμύριο κινήσεις. Καταλαβαίνουμε ότι μπορούμε να κάνουμε παραπάνω από μία κινήσεις κάθε μέρα του dataset. Συγκεκριμένα, αν διαιρούσαμε το εύρος ημερομηνιών με 1M, θα προέκυπτε ότι μας επιτρέπονται 70 κινήσεις τη μέρα. Όμως, καλύτερα θα είναι να μπορούμε να προσαρμόσουμε το όριο των επιτρεπόμενων κινήσεων ανά ημέρα, μιας και οι πρώτες δεκαετίες του dataset περιέχουν λίγες μετοχές, ενώ μετά τη δεκαετία του '90 υπάρχει μία έκρηξη από διαφορετικές μετοχές. Επομένως, μετά από κάθε ημέρα του dataset, προσαρμόζω το όριο κινήσεων ανά ημέρα, με βάση το πόσες κινήσεις και πόσες μέρες μας απομένουν. Με τον τρόπο αυτό θα μπορούσαμε να εκμεταλλευτούμε την ύπαρξη πολλών διαθέσιμων μετοχών στα τελευταία χρόνια του dataset. Για κάθε μετοχή θα χρειαστεί να "σπαταλάμε" δύο κινήσεις, όπως δείξαμε και προηγουμένως.

Στη συνέχεια, διατρέχουμε το dataset ανά ημέρα (παράθυρο μίας ημέρας πλέον). Κρατάμε τις γραμμές του dataframe της συγκεκριμένης ημέρας, και ταξινομούμε με βάση την στήλη *MaxProfit*, επιλέγοντας τις top επιτρεπόμενες-ανά-ημέρα μετοχές και κρατάμε σε μια μεταβλητή *temp_balance* το υπόλοιπο του λογαριασμού πριν αρχίσει η μέρα, το οποίο και είναι η αγοραστική μας δύναμη για τη μέρα αυτή. Στη συνέχεια, ελέγχουμε για κάθε μία από αυτές τις μετοχές εάν έχουμε τη δυνατότητα να την αγοράσουμε για τη μέγιστο πιθανό Profit που μας παρέχει (με βάση τη *MaxIntra* τιμή δηλαδή). Αν έχουμε, την αγοράζουμε, και περνάμε το κατάλληλο transaction, "*O->H*" ή "*L->C*", στο log κινήσεων. Στη συνέχεια, αλλάζουμε κατάλληλα το *temp_balance*, ώστε στη συνέχεια, όταν εξετάζουμε την επόμενη μετοχή, η αγοραστική μας δυνατότητα να έχει επηρεαστεί κατάλληλα.

Πλέον, λόγω της εφαρμογής μόνο Intra-day trading, καταλήγουμε στο τέλος της ημέρας να μην έχουμε καμία μετοχή στην κατοχή μας. Επομένως, το Portfolio δεν υφίσταται πλέον σαν έννοια. Κρατάμε επομένως ένα dictionary μόνο για το Balance και πώς εξελίσσεται αυτό μέρα με τη μέρα. Καταλήγουμε η ακολουθία κινήσεων που προκύπτει να δίνει κέρδος **295659440855.4203 \$** μετά από ακριβώς **1000000** κινήσεις. Παρακάτω φαίνεται και η εξέλιξη του Balance μας μέσα στα χρόνια.

