**פרויקט רשתות נוירונים**

**TRADING BOT**

**דוד אילוז 209959220**

**דור אגבבה 208133116**

**תקציר:**

בפרויקט זה פיתחנו מודל לחיזוי ערכי מניות באמצעות רשת נוירונים מסוג LSTM.  
מטרת הפרויקט הייתה ליצור מודל שמסוגל לחזות ערכים עתידיים של מניות בהתבסס על נתונים היסטוריים.

התחלנו את העבודה בקריאת הנתונים , הבנו שהמידע אינו נקי ועלינו לבצע ניקוי למידע. לשם כך ביצענו מחיקה של ערכים ריקים ועמודות שכל המידע בהן 0. לאחר מכן ביצענו מחקר קצר על הפרמטרים(הפיצ'רים) שבידנו בדאטה וסיננו את זה מספר מועט יותר שנותנים לנו מספיק מידע ולא גורעים מאיתנו דבר כדי לעבוד. סידרנו את הדאטה כך שיהיה ברמת ה-5 דק, ולא ברמת הדקה (נפרט בהמשך מדוע). ערבבנו אותו ברמה יומית ,כך שמידע של יום אחד לא יתערבב עם יום אחר.  
בנוסף גילינו כי הדאטה צריך להתקבל בצורה של מערך 3 מימדים, למעשה מערכים דו מימדיים שמייצגים את המידע הקודם לערך שברצוננו לשערך, לכן סידרנו אותו כך.

לבסוף יצרנו סקריפט שהוא בעצם ה baseline שלנו עם פונקציה גנרית שתקבל אפשריות לבדוק hyperparameters אחרים .   
הבעיה שהתעסקנו איתה היא בעצם בעיית regression ולא בעיית קלסיפיקציה כמו שלרוב אנחנו מתעסקים, ולכן צפה בעיה כיצד נמדוד דיוק ,לשם כך חקרנו ומצאנו מטריקה mean absolute error percentage . אז הדפסנו כמדדים מספר גרפים , גם את השגיאה וגם את המדד שתיארנו.

לבסוף, השקענו מאמצים משמעותיים במציאת hyperparameters , בצורה שתמנע overfitting ותביא לדיוק מקסימלי עבור ה validation set .

לבסוף, שיפרנו משמעותית את התוצאות מהbaseline ע"י שינוי והתאמת ה hyperparameters השונים , הפחתנו את הloss כך שישאף ל0.

**מקור הנתונים:**

<https://www.kaggle.com/datasets/debashis74017/stock-market-data-nifty-50-stocks-1-min-data/discussion/409605>

מחברת לדוגמא שהמודל שלה שימש כbaseline :

<https://www.kaggle.com/code/alirezanematolahy/tehran-stock-prediction-using-lstm-model#The-Model>

**תיאור הבעיה:**

הבעיה המרכזית בפרויקט היא חיזוי ערך המנייה העתידי בהתבסס על נתוני העבר ונתונים כלכליים שונים . באופן כללי חיזוי מדויק של ערכי מניות הוא משימה מורכבת שמושפעת ממספר רב של גורמים שלא בהכרח ידועים, כולל מגמות בשוק, נתונים כלכליים, חדשות, ודינמיקה פסיכולוגית של המשקיעים.   
מטרת הפרויקט היא לפתח רשת נוירונים המסוגלת לנתח את הנתונים הזמינים ולהסיק תחזיות מדויקות לעתיד, אשר יכולות לסייע בקבלת החלטות השקעה מושכלות שתשמש משקיעים.

משקיעים רבים מחפשים כלים שניתן לסמוך עליהם ולהשתמש בהם כעזר לקבלת החלטות או אף כמקבל ההחלטות ללא התערבות שלהם כלל. לכן חיזוי ערך המנייה יכול להוות עבורם ניתוח מושלם וגם לתת להם כיוון אם עלה ובכמה , או אם ירד ובכמה. ובכך ניתן להסיק אם כדאי להתחיל עסקה או לא. הוספנו כאחד המדדים בעבודה , כמה הגרף המדומה דומה לגרף המקורי בצורתו (עלייה וירידה).

עבור הבעיה שלנו **האתגר הראשון** היה למצוא מודל מתאים מבין המודלים שלמדנו עבור הבעיה שלנו, ומצאנו שמודל LSTM יודע להתמודד טוב עם בעיות מהסוג שלנו , שעליו לקבל מידע של סדרת זמן. למעשה המודל נדרש לקבל sequence של מידע (מספר נקודות זמן לפני כן) ולהחליט מה תהיה התוצאה של נקודת הזמן הבאה. ניתן למצוא מספר רב של פתרון בעיות דומות לשלנו בחיזוי ערכי מנייה במודל LSTM ,שבו המודל LSTM נבחר על פני רבים אחרים שלא הצליחו לתת את אותן התוצאות שבהשוואה לLSTM .

מקורות מידע שהסתמכנו עליהם בבחירת מודל LSTM:

<https://neptune.ai/blog/predicting-stock-prices-using-machine-learning>

<https://www.nature.com/articles/s41599-024-02807-x>

<https://www.sciencedirect.com/science/article/abs/pii/S0957417420304334>

**אתגר נוסף** היה להחליט מה הוא גודל הזמנים בין דגימה לדגימה, משמע מה מרווחי הזמן מקנייה לקנייה מבחינת המשתמש הסופי, והמצד שלנו ראינו מה בעצם תהיה רמת הרעש שתכנס למודל . ככל שיחידת המידה תהיה קטנה יותר כך הגרף יהיה רועש יותר וגם החיזוי יהיה פחות נקי אך פוטנציאל הרווח יגדל (יותר עסקאות). הבחירה הייתה נתונה בידנו כיוון שה dataset שקיבלנו הינו ברמת דקה . לכן יכולנו לבחור מספר אפשריות . האפשריות שהעלנו בפנינו שהיו נכונות לדעתנו לדון עליהן :

|  |  |  |  |
| --- | --- | --- | --- |
| מרווח דגימה | רווח פוטנציאלי | רמת רעש פוטנציאלית | רמת דיוק |
| דקה אחת | גבוהה מאוד | גבוהה מאוד | נמוך |
| חמש דקות | גבוהה | גבוהה | בינוני |
| יום אחד | נמוך | נמוך | גבוהה |

הבנו כי יש יחס ישר בין רווח פוטנציאלי ורמת הרעש, ויחס הפוך עם רמת הדיוק. היינו יכולים לבחור גם יחידת מידה אחרת בין 5 דק ליום, אך רצינו לשמור על רווח פוטנציאלי גבוהה . רוב המודלים שפגשנו ברשת, בנו מודל שרמתו היא יום אחד , בגלל שתי סיבות עיקריות , הראשונה היא חוסר בdataset ברמה קטנה מיום, והשנייה היא רמת הרעש שמקשה על המודל.  
אנחנו בחרנו לבסוף לקחת מודל עם מרווח דגימה של 5 דק .

**אתגר שלישי** שהתמודדנו איתו היה לסנן מתוך כל הנתונים שקיבלנו את הנתונים המתאימים לבעיה שלנו וסינון מידע לא רלוונטי. המידע הגיע עשיר בפרמטרים לכל רשומה (60 פרמטרים) . רוב הפרמטרים היו בעצם אינדיקטורים פיננסיים , שחלקם חוזרים על עצמם וחלקם היו ריקים. בבחירת המדדים לחיזוי ערך המנייה, הדגשנו את הערך העכשווי של המנייה כגורם החשוב ביותר, בעוד שאר הפרמטרים משמשים כהשלמות שמסייעות למודל לחזות בצורה מדויקת יותר את ערך המנייה העתידי. לאחר בחינה מעמיקה של מחקרים ושימוש נפוץ בקרב משקיעים, החלטנו לכלול את המדדים הבאים, אשר נחשבים לפופולריים ביותר לחיזוי מהיר ומדויק של ערך מניה:

- EMA5 (Exponential Moving Average) ממוצע נע אקספודנציאלי מחושב על פני 5 ימי מסחר. זהו אינדיקטור שמבליט שינויים קצרים-טווח בערך המניה ומגיב במהירות לשינויים בשוק. EMA5 משמש משקיעים לזיהוי מגמות שוק בזמן אמת ולספק אינדיקציה מוקדמת לשינויי מגמה.

- EMA20 ממוצע נע אקספודנציאלי מחושב על פני 20 ימי מסחר. בניגוד ל-EMA5, EMA20 מעניק מבט רחב יותר על המגמות, תוך שמירה על רגישות מסוימת לשינויים קצרים. הוא מאפשר למשקיעים לזהות מגמות בינוניות-טווח ומסייע באישור המגמות שזוהו על ידי EMA5.

MACD (Moving Average Convergence Divergence) - אינדיקטור זה מבוסס על ההבדל בין שני ממוצעים נעים אקספודנציאלים– EMA12 ו-EMA26. MACD מספק תמונה על כיוון המגמה, עוצמתה ושינויי המגמה הקרבים. מדד זה פופולרי במיוחד בקרב משקיעים בשל היכולת שלו לזהות נקודות כניסה ויציאה אידיאליות מהשוק.

הבחירה במדדים אלה התבססה על השימוש הנפוץ שלהם בקרב משקיעים, המאפשר לחזות ערך מניה בקצב מהיר ובדיוק גבוה, תוך זיהוי מגמות בשוק בזמן אמת לזמן הקרוב. אם היה ניתן להשתמש גם במדד ה volume היינו משתמשים , אך הוא היה ריק ..

מקורות מידע שהתבססנו עליהם:

<https://www.tokenmetrics.com/blog/best-indicators-for-crypto-trading-and-analysis#1-moving-averages>

<https://groww.in/p/exponential-moving-average>

<https://realtrading.com/trading-blog/short-term-trading-indicators/>

למידה מחוברות של אחרים :

<https://www.kaggle.com/code/carlmcbrideellis/lstm-time-series-stock-price-prediction-fail>

<https://www.kaggle.com/code/alirezanematolahy/tehran-stock-prediction-using-lstm-model#The-Model>

<https://www.youtube.com/watch?v=Vfx1L2jh2Ng&t=562s&ab_channel=LazyProgrammer>

הצלחנו ללמוד ממחברות אחרות, שאכן המודל מגיע להישגים טובים, שכדאי לנו להשתמש בdropout מה שהתברר לא כהכי טוב בסופו של דבר.   
איזה מודל בסיסי כדאי להתחיל ממנו, ולאיזה תוצאות אנשים הגיעו.   
למדנו על פרופורציית חלוקת הנתונים , ולמדנו כיצד למדוד את המודל בסופו של דבר , ומה הטעיות שרוב האנשים עושים.

**ניתוח נתונים ראשוני :**

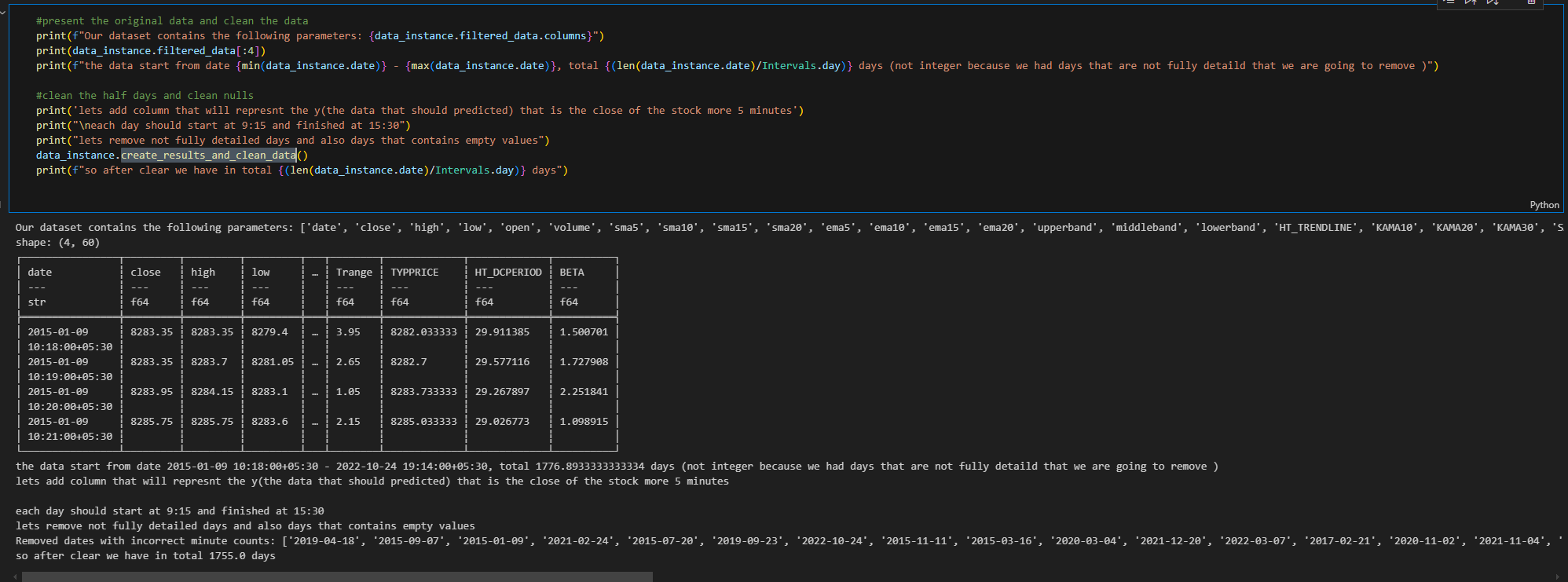
מקור נתונים :

<https://www.kaggle.com/datasets/debashis74017/stock-market-data-nifty-50-stocks-1-min-data/discussion/409605>

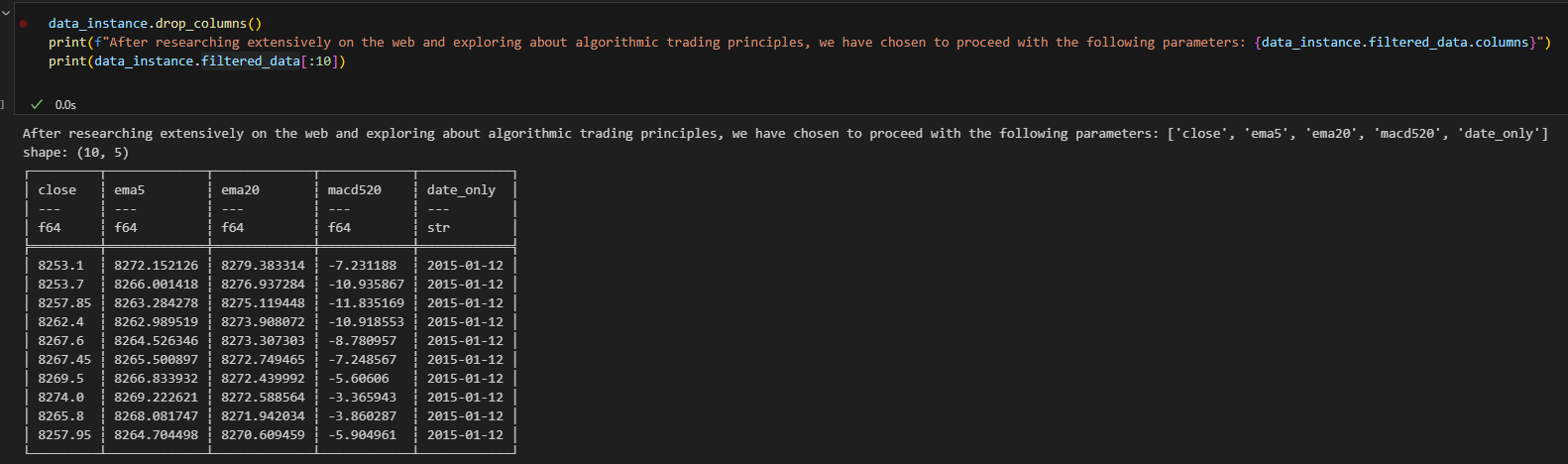
שם הקובץ : NIFTY 50\_minute\_data\_with\_indicators.csv

מקור הנתונים שלקחנו הוא של מדד ולא מנייה בודדת, המדד מורכב מ 50 מניות. אנחנו בחרנו להתמקד במדד עצמו, שגם אותו ניתן לרכוש, ובכך אנחנו זוכים למידע יותר שקט משל מנייה רגילה, ואפקטים/תנודות שיש למנייה רגילה.   
קראנו את המידע וגילינו שיש לנו מספר עמודות ריקות , אז התחלנו לסנן את המידע , מחקנו עמודות שיש להן מידע ריק.  
חקרנו מעט ברשת והבנו שיום מסחר רגיל מתחיל בשעה 9:15 ונגמר ב15:30 .  
בדקנו ממתי עד מתי יש לנו מידע . והתחלנו גם בסינון ימים שהם לא שלמים או שמכילים חצאי ימים למיניהם . נשארנו עם 1755 ימים ,כמות לא מבוטלת של ימים (קרוב ל7 שנים של מסחר, 5 ימי מסחר בשבוע).

קצת מניתוח ראשוני של הדאטה.



כפי שתיארנו קודם לכן , בחרנו את העמודות הנ"ל (EMA5,EMA20 ,MACD520 ,CLOSE) מתוך כל העמודות הנ"ל.



ובנוסף הוספנו עמודה שבעצם תתאר את y ,את המטרה. לכל רשומה הy שלה זה הclose של המנייה בעוד 5 דק.