Final Project

Stock Market Prediction in the U.S. - Analysis and Forecasting of S&P 500 Trends

מגישים – שינה תחאוחו 213381569 ודין תחאוכו 318291705

מנחה – ג'ניה גוטפריד

מנחה בוחן - ד"ר לואי עבדאללה

תאריך בחינה – 28/8/2025

.

תוכן עניינים

The Max Stern Yezreel Valley College
Dept. of Information Systems

המכללה האקדמית עמק יזרעאל
ע"ש מקס שטרן
החוג למערכות מידע

06/10/2022

תבנית בחירת נושא לפרויקט גמר

מגישים: שינה תחאוחו ודין תחאוחו

התמחות: מדע הנתונים

שם הפרויקט:
מערכת מתקדמת לזיהוי מגמות בבורסה.

תיאור תמציתי:

שוק ההון מהווה את אחד המרכיבים המרכזיים במערכת הכלכלית העולמית, ומאופיין בדינמיות גבוהה ובתנודתיות רבה. משקיעים וחברות בכל רחבי העולם מתמודדים עם הצורך להבין ולחזות את מגמות השוק על מנת לקבל החלטות מושכלות ואסטרטגיות. פרויקט זה מתמקד בפיתוח מערכת מבוססת נתונים שמטרתה לחזות עליות וירידות בבורסה.

המערכת תשלב אלגוריתמים מתקדמים של למידת מכונה לניתוח נתונים היסטוריים ותבניות מסחר, ותשתמש בדאטה רחב היקף הכולל נתוני מניות, מחזורי מסחר, חדשות כלכליות ועוד. המטרה היא לספק כלי ניתוח מתקדם שמסייע למשקיעים בזיהוי מגמות עתידיות ובצמצום הסיכונים. פרויקט זה משלב בין תחומי מדעי המידע, כלכלה ולמידת מכונה, תוך שאיפה ליצירת תובנות משמעותיות ובעלות ערך בעולם המסחר.

פרויקט גמר 2024

Business Understanding Report

<u>חיזוי הבורסה באר"הב</u>

מגישים : שינה תחאוחו 213381569 , דין תחאוכו 318291705

תאריך הגשה : 15/12/2024

מוגש ל : ג'ניה גוטפריד

משימה 1 - רקע

1.1 קביעת יעדים עסקיים

במסגרת פרויקט הגמר שלנו לשנת 2024 אנו נעסוק בחיזוי מגמות בשוק המניות בישראל.
אנו רוצים להציע כלי תומך החלטות למשקיעים ולספק תובנות שיעזרו למקסם רווחים ולהפחית סיכונים.

תחום שוק ההון מאופיין בתנודות גבוהות ובצורך מתמיד לקבלת החלטות מדויקות ומבוססות מידע לכן המטרה המרכזית שלנו היא לפתח מודל חיזוי לשוק ההון הישראלי שיספק תחזיות מדויקות על מגמות שוק המניות.

1.2 רקע עסקי

מטרות -

1. **בניית מודל חיזוי:** פיתוח מודל למידת מכונה שמנבא תנועות מניות ברמת דיוק גבוהה בהתבסס על נתונים היסטוריים ועדכניים.

2. **זיהוי מאפיינים מרכזיים:** איתור הגורמים המשפיעים ביותר על תנועות מחירי מניות, כמו נפח מסחר, סנטימנט שוק או אינדיקטורים מקרו-כלכליים.

3. **שיפור קבלת החלטות:** מתן תובנות פעולה למשקיעים כדי לקבל החלטות מושכלות לגבי רכישה ומכירה של מניות.

4. **הערכת ביצועי המודל:** הערכת ביצועי המודל באמצעות מדדים מתאימים כמו דיוק, זיהוי ואמינות כמו F1 score כדי להבטיח את אמינותו.
5. **יכולת הרחבה:** הבטחת יכולת להתאים את הפתרון לכל השווקים בעולם, ובכך להעניק יישום רחב יותר.

<u>משאבים זמינים</u>

1. **נתונים היסטוריים**

- נתוני שערי מניות , מחזורים יומיים , נפח מסחר ותנועות בשוק לאורך 10 שנים אחרונים
- נתונים חיצוניים כגון שיעורי ריבית, דוחות כלכליים וכו

2. **כלים טכנולוגיים**

- שימוש בשפות תכנות כמו Python ובספריות מתקדמות ללמידת מכונה כגון Scikit -learn , TensorFlow
- גישה למערכות ענן לניתוח נתונים בהיקפים גדולים (Big Data)

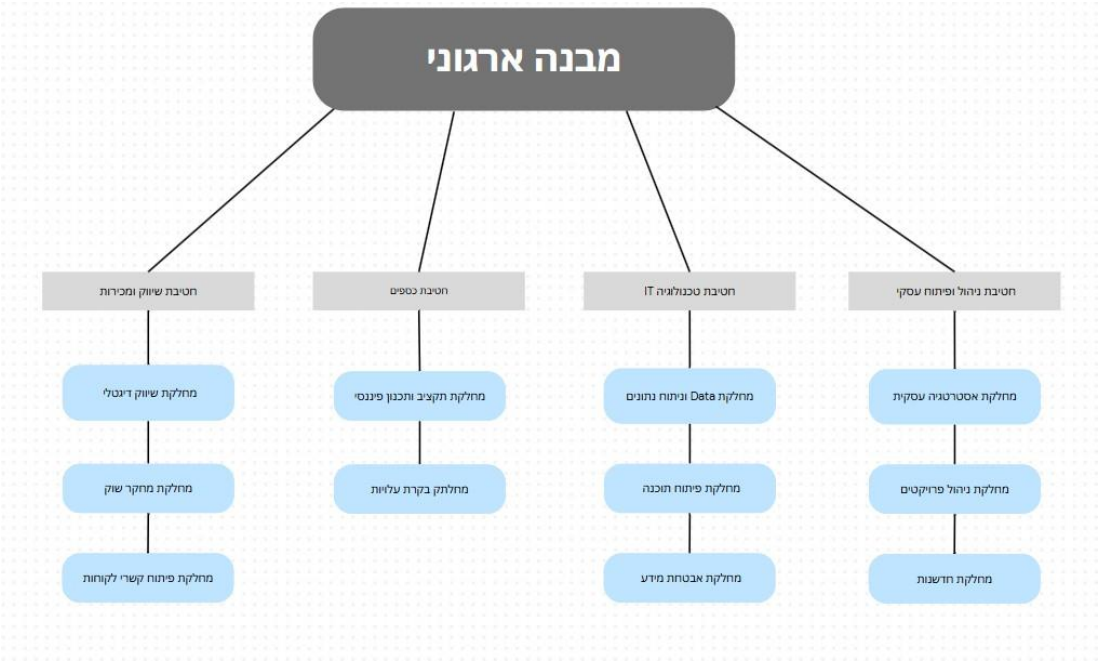
3. **צוות מקצועי**

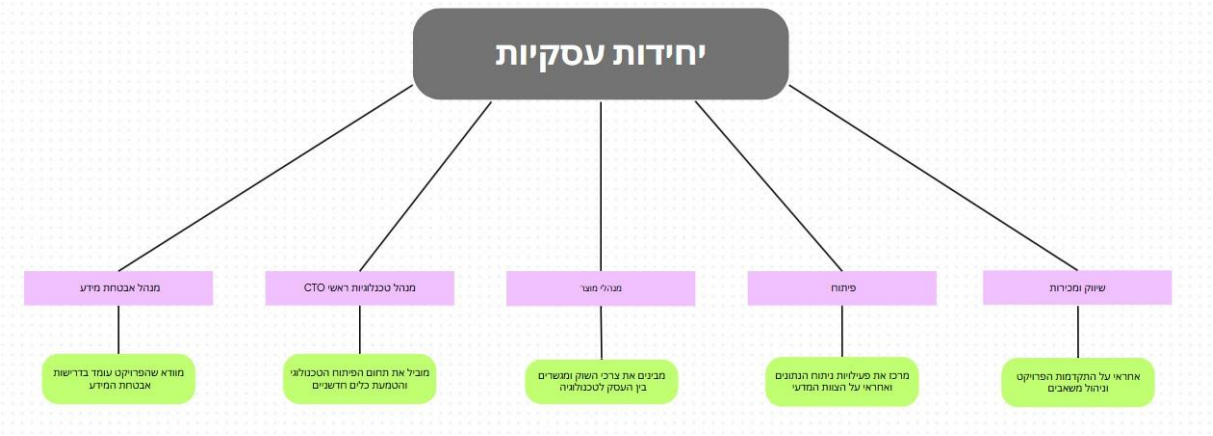
- מומחים בתחום הפיננסים, מנתחי נתונים, ומפתחים בעלי ניסיון הפרויקטי למידת מכונה

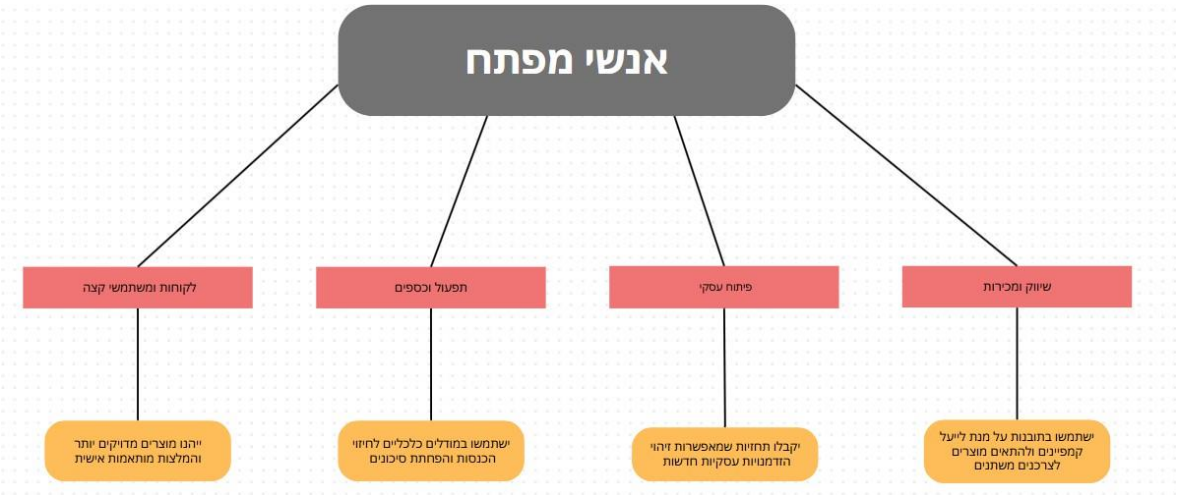
<u>בעיות</u>

- היעדר כלים מדויקים בזמן אמת לחיזוי מחירי מניות מותאמים לשוק הישראלי.
- הצורך בניתוחים מתקדמים לתמיכה בהחלטות פיננסיות בעלות משמעות.

קביעת מבנה ארגוני:







תיאור תחום הבעיה

תחום הבעיה:

התחום המרכזי של המיקוד הוא אנליטיקה פיננסית בשוק הון, המשתלב עם מתודולוגיות מדעי הנתונים. הפרוייקט ממוקם בצומת שבין אנליטיקה חיזויית לבין קבלת החלטות פיננסיות.

תיאור כללי של הבעיה:

תנועות מחירי המניות הן בעלות תנודתיות גבוהה ומושפעות ממגוון גורמים כמו אירועים גיאופוליטיים, מדדים כלכליים ומגמות בענף. משקיעים מתמודדים עם אתגרים בקבלת החלטות מושכלות בשל חוסר הוודאות הטבוע.

דרישות מקדימות של הפרוייקט:

- **מוטיבציה:** המוטיבציה נובעת מהביקוש הגובר לאסטרטגיות השקעה מבוססות נתונים ומהצורך במודלים חיזויים אמינים בשוק הישראלי.

- **שימוש נוכחי במדעי הנתונים:** בעוד שמוסדות פיננסיים גדולים משתמשים באלגוריתמים מתקדמים, למשקיעים קטנים חסרים כלים נגישים לניתוח וחיזוי.

תיאור הפתרון הנוכחי

פתרונות קיימים:

- כיום, משקיעים רבים מסתמכים על כלים ציבוריים זמינים, כמו גרפים של מניות ואינדיקטורים בסיסיים לניתוח טכני.

- פתרונות מתקדמים כמו מערכות מסחר אלגוריתמיות, זמינים אך לעיתים קרובות אינם נגישים למשקיעים קטנים בשל עלויות גבוהות או מורכבות טכנית.

<u>יתרונות וחסרונות של הפתרונות הנוכחיים:</u>

יתרונות:

- כלים ציבוריים הם פשוטים וקלים לשימוש.
- מערכות מתקדמות מספקות יכולות חזקות למשקיעים גדולים.

חסרונות:

- כלים ציבוריים חסרים דיוק חיזוי והתאמה לגורמים ייחודים לשוק.
- מערכות מתקדמות אינן זמינות למשקיעים פרטיים או קטנים עקב מגבלות במשאבים.

רמת הקבלה:

כלים ציבוריים מתקבלים באופן רחב על ידי משקיעים פרטיים אך לעיתים קרובות זוכים לביקורת על תועלתם המוגבלות בשווקים משתנים במהירות. פתרונות מתקדמים נהנים מקבלה בקרב מוסדות פיננסיים אך אינם דמוקרטיים

2. יעדים עסקיים וקריטריונים להצלחה

2.1 מטרות עסקיות

- **תיאור הבעיה**

הפרויקט נועד לטפל בחיזוי תנודות שוק המניות בישראל אשר מושפעות מגורמים דינמיים רבים כמו אירועים גיאופוליטיים וכו. כדי לאפשר לחברות ואנשים פרטיים לבצע החלטות השקעה מבוססות יותר. הבעיה המרכזית היא חוסר היכולת של המשקיעים לזהות מגמות בזמן אמת , מה שגורם להפסדים או הזדמנויות שלא מנוצלות.

- **שאלות עסקיות מדויקות**

1. איך ניתן לחזות שינויים במדדי המניות בטווח קצר (יומי / שבועי) ?
2. אילו גורמים חיצוניים (כגון חדשות כלכליות או נתוני מאקרו) משפיעים בצורה המשמעותית ביותר על הבורסה הישראלית?
3. כיצד ניתן לשפר את דיוק התחזיות לעומת מודלים קיימים ?
4. מהן המגמות המתפתחות בענפי שוק שונים כמו הייטק, פיננסים, ותעשייה מסורתית ?
5. אילו מניות צפויות לעלות בערך שלהן במהלך פרק זמן נתון ?

- **דרישות עסקיות**

1. שמירה על דרישות לקוחות קיימות כלומר אספקת מידע עדכני ומהיר ללא עיכובים , הבטחת פרטיות, ואבטחת המידע של הלקוחות המשתמשים במערכת.
2. הגדלת הזדמנויות מכירה צולבת - פיתוח דוחות ותובנות המציעים ללקוחות מוצרים פיננסיים מתאימים על בסיס פרופיל הסיכון שלהם
3. שיפור חווית משתמש - פלטפורמה אינטואיטיבית וקלה לשימוש שתספק נתונים זמן אמת שתאפשר ללקוחות גישה לתחזיות בצורה ברורה
4. דיוק - המודל חייב להשיג רמת דיוק גבוהה (למשל מעל 75%) כדי להבטיח שימוש מעשי עבור המשקיעים

- **יתרונות צפויים במונחים עסקיים**

1. **שיפור בקבלת החלטות -** העצמת משקיעים לקבלת החלטות מושכלות יותר אשר עשויות לשפר את ביצועי התיק שלהם

2. **נגישות -** מתן כלים למשקיעים פרטיים הדומים לאלו המשמשים סוחרים מוסדיים וגישור הפער באנליטיקה שוקית

3. **הבנת השוק -** מתן תובנות על הגורמים המרכזיים המניעים תנועות מניות בשוק , התרומות להבנת השוק

4. **השפעה כמותית -** לדוגמא המודל יכול לשאוף לשפר את התשואות על תיקי השקעות ב 5% לפחות במהלך שישה חודשים עבור משתמשים המאמצים את התובנות

2.2 קריטריוני הצלחה

ההצלחה של הפרויקט שלנו תימדד על פי יכולתו לחזות במדויק תנועות מחירי מניות (עליה או ירידה) בהקשר של שוק ההון, קריטריוני ההצלחה מחולקים לקטגוריות אובייקטיביות וסובייקטיביות כדי לספק מסגרת הערכה רחבה ומאוזנת.

קריטריונים אובייקטיביים

1. **דיוק החיזוי:**
   - המודל צריך להשיג דיוק של לפחות 75% על נתוני בדיקה שלא נראו בעבר. מדדים כמו Recall, Precision, ו- F1 Score ישמשו להערכת היעילות של המודל.

2. **זיהוי גורמים מרכזיים:**
   - המודל צריך לזהות ולדרג בהצלחה את הגורמים המשפיעים ביותר על תנועות מחירי מניות בשוק ההון. דוגמאות כוללות נפח מסחר, אינדקטורים מקרו-כלכליים וסנטימנט מחדשות.

3. **עקביות בין מגזרים**
   - המודל חייב להפגין ביצועים עקביים בין מגזרים שונים בשוק ההון.

4. **יכולת הרחבה וגמישות:**
   - המערכת צריכה להיות מסוגלת לשלב מקורות נתונים חדשים או להסתגל לשווקים פיננסיים אחרים מכל העולם במידת הצורך.

5. **השפעה עסקית**
   - הפגנת שיפורים מדידים בתוצאות קבלת החלטות, כמו שיפור תשואות בתיקים ב5% לפחות עבור משתמשים המיישמים את תחזיות המודל במשך שישה חודשים.

קריטריונים סובייקטיביים

- **תובנות מתוך ניתוח הנתונים:** הצלחה סובייקטיבית תימדד ביכולת שלנו לזהות מגמות או דפוסים חשובים בשוק ההון מתוך הנתונים, כמו התנהגות מניה מסוימת בתגובה לאירועים כלכליים או נפחי מסחר.

- **פרשנות תוצאות המודל:** נותני החסות לפרויקט (למשל, מנחים או קהל היעד של ההגשה) צריכים להבין בקלות את תוצאות המודל ואת התובנות העסקיות שהפקנו ממנו. הצלחה תוגדר כאשר המצגת או הדוח יזכו למשוב חיובי על בהירותם.

- **יכולת יישומית:** המודל צריך להיות גמיש מספיק לשימוש על מניות אחרות או תקופות זמן שונות, דבר שיתבטא במשוב מהמנחה.

רשימת משימות למימוש קריטריוני הצלחה

1. **יישור מטרות עסקיות וקריטריונים להצלחה**
   - הבטחה שלכל מטרה עסקית (כמו השגת דיוק חיזוי, יצירת השפעה עסקית) יש קריטריון הצלחה מקבילה.

2. **הגדרת בעלי עניין להערכות סובייקטיביות**
   - זיהוי ויישור עם בעלי עניין מרכזיים, כגון יועצים אקדמיים ואנליסטים פיננסיים, לקביעת ציפיות למדידת הצלחה סובייקטיבית.

3. **מעקב ואימות מדדים**
   - הערכת התקדמות הפרויקט באופן שוטף מול הקריטריונים האובייקטיביים והסובייקטיביים להבטחת התאמה למטרות שהוגדרו

תובנות הניתנות לפעולה:

- התחזיות וההסברים של המודל צריכים להיות ברורים הניתנים ליישום עבור משקיעים פרטיים ואנליסטים פיננסיים.

אימות מומחים:

- משוב ממומחים בתעשייה או יועצים אקדמיים צריך לאשר את הרלוונטיות והיישומיות המעשית של תובנות המודל.

התאמה למגמות שוק:

- התוצרים של המודל צריכים להתאים למגמות השוק במציאות, כפי שייבחנו על ידי מומחים המכירים את שוק ההון הישראלי.

תרומה אקדמית:

- הפרויקט צריך לתרום לתחום מדעי הנתונים על ידי תיעוד שיטות ותוצאות שיכולות לשמש לצרכים חינוכיים או מחקריים עתידיים.

3.הערכת מצב

בהערכת המצב הראשונית לפרויקט, התמקדנו בזיהוי המשאבים והנתונים הזמינים לצורך חיזוי שוק ההון. הנתונים שבידינו כוללים סדרות זמן יומיות של מחירי מניות (פתיחה, סגירה, גבוה, נמוך, מחיר מותאם ומחזורי מסחר), כאשר הם מבוססים כרגע על חברת Amazon אך ניתנים להתאמה לכל חברה אחרת לפי הצורך. המטרה היא להבין את יכולות החיזוי תוך התמודדות עם אתגרים כמו חוסר נתונים, מגבלות חישוביות ודיוק החיזוי. לצד זאת, אני בוחנת את הכלים הטכנולוגיים שברשותי, הכוללים חומרה מתאימה וכלי ניתוח מבוססי Python, וכן משאבים נוספים כמו מקורות נתונים משלימים. הערכה זו מאפשרת לי למפות את התהליך ולבנות תכנית פעולה אפקטיבית לפרויקט.

- **נתונים פיננסיים**
  - **Open, High, Low, Close - מחירים יומיים**
  - **Adjusted Close - מחיר מותאם**
  - **Volume - מחזור מסחר יומי**
  - **תאריכים מדוייקים**
  - **שם החברה (דינמי) - הנתונים ניתנים להתאמה לכל חברה שהיא בין אם זה Google, Amazon , Apple וכו באמצעות שינוי בקוד לקבלת נתונים של חברות שונו**

כוח אדם דרוש להשלמת הפרויקט

- **צוות מדעי הנתונים:**

אשר אחראים על ניקוי ועיבוד הנתונים, בניית מודלים לחיזוי סדרות זמן והערכת ביצועי המודלים כדי לשפר את הדיוק שלהם. מומחים אלה צריכים לשלוט בכלי ניתוח נתונים כמו Python ובספריות רלוונטיות כגון Pandas ו-TensorFlow, וכן להיות בעלי ניסיון בעבודה עם סדרות זמן וסטטיסטיקה.

- **מומחים פיננסיים**

תפקידם לספק הבנה מעמיקה של שוק ההון ולזהות את הגורמים המשפיעים על מחירי המניות. הם מתרגמים את תוצאות המודלים לתובנות עסקיות שיכולות לסייע בקבלת החלטות, ומציעים רעיונות לשיפור המודלים על ידי שילוב משתנים רלוונטיים כמו נתוני מקרו-כלכלה או אירועים פוליטיים.

- **אנשי IT ואבטחת מידע**

מטפלים בגישה לנתונים ממאגרים חיצוניים, חיבור לממשקי API, ושימוש בשרתי ענן במידת הצורך. הם אחראים על ניהול התשתיות הטכנולוגיות שמאפשרות את החישובים ואת גישה לנתונים בזמן אמת.

- **יועצים עסקיים**

יספקו פרספקטיבה נוספת על תוצאות הניתוח, ובמנהל פרויקט שיבטיח שהפרויקט מתקדם לפי התכנון ויתמודד עם בעיות לוגיסטיות

מהם גורמי הסיכון הגדולים ביותר המעורבים?

- **סיכון טכנולוגי:** אי התאמה בין המודלים המדעיים לבין יכולות המחשוב הקיימות. חוסר יציבות במערכות יכולות להוביל לעיכובים בתהליך.

- **סיכון נתוני:** איכות הנתונים היא קריטית. יש לקחת בחשבון שהנתונים עשויים להיות לא מדויקים או חסרים, דבר שיכול לפגוע ביכולת המודלים לבצע תחזיות מדויקות.

- **סיכון רגולטורי:** קיימת האפשרות שחוקים או רגולציות ישתנו במהלך הפרויקט, דבר שיכול להשפיע על השימוש בנתונים או על המודלים שנבנים.

- **סיכון בשוק:** שינוי במגמות שוק בזמן הפרויקט עשוי להקטין את היכולת לחזות במדויק את התנודות בשוק.

תוכנית מגירה

- **תוכנית גיבוי טכנולוגי:** יצירת מערכות גיבוי ויכולת אוטומטית להפעיל מודלים חלופיים במקרה של כשל טכנולוגי.

- **תוכנית לניהול נתונים:** לפתח תוכניות לשיפור איכות הנתונים, כולל תיקון נתונים חסרים ומבצעי ניקוי נתונים.

- **תוכנית רגולציה:** להכיר את כל החוקים ורגולציות שמסבכים את השימוש בנתונים ולהתעדכן בהן בזמן אמת.

- **תוכנית לניהול סיכונים בשוק:** לשלב מודלים שמעדכנים תחזיות בהתאם למגמות השוק בזמן אמת, ולהפעיל מערכות לניהול סיכון לאור שינויים חדים בשוק.

4. מלאי משאבים

4.1 משאבי חומרה

לחברה יש תמיכה בשני שרתי פיתוח חזקים, שכל אחד מהם יהיה מצויד ב 32 ליבות , 64GB של זיכרון RAM ואחסון SSD בנפח 2TB. בנוסף אנו משתמשים בשירותי AWS, שמאפשרים גמישות בהרחבת משאבי המחשוב והאחסון בהתאם לצרכים. הרשת המחברת בין השרתים פועלת ברוחב פס של 1Gbps, ויש גיבוי של 100Mbps למקרים של עומסים חריגים או תקלות. צוות ה IT שלנו אחראי באופן קבוע על התחזוקה , עדכון ותיקון תקלות כך שהמערכת תמיד מוכנה לעבודה שוטפת.

4.2 זיהוי מקורות נתונים ומאגרי מידע

בנוגע למקורות נתונים, כרגע אנו עובדים עם נתוני סדרות זמן הכוללים מחירים יומיים של מניות ומחזורי מסחר בפורמט CSV. עם זאת, במידת הצורך, נרחיב את בסיס הנתונים על ידי שימוש במקורות חיצוניים כמו Yahoo Finance API כדי לקבל נתונים עדכניים נוספים. הנתונים מאוחסנים כרגע באופן מקומי, אך נשקול לעבור לאחסון מבוסס ענן כמו Google Sheets או SQL לטובת עבודה גמישה ונוחה יותר. כרגע אין לנו צורך ברכישת נתונים חיצוניים, אך נבחן את האפשרות לשלב נתוני מקרו-כלכלה בעתיד. כמו כן, לא זיהינו בעיות אבטחה אשר מונעות גישה לנתונים הדרושים לנו.

4.3 זיהוי משאבי כוח אדם

בפרויקט עובדים מומחים מתחומים שונים כדי להבטיח מענה מקצועי בכל שלב. בתחום מדעי הנתונים ישנם שני אנליסטים בעלי ניסיון רב ב python וב -SQL המובילים את הפיתוח והניתוח. בנוסף מנהל השיווק ומנהל המכירות משתפים פעולה ומספקין הבנה עמוקה של הצרכים העסקיים צוות מסדי הנתונים כולל מנהל מסד נתונים בכיר האחראי על ניהול ואחזקת המידע, בנוסף לכך הצוות ה IT שלנו כולל שני מנהל רשת ומומחה סייבר. צוות זה מוודא שהמערכת מאובטחת וזמינה לאורך כל הדרך, תוך מניעה של תקלות או גישה לא מורשית

משימה 5 - דרישות הנחות ואילוצים

5.1 קביעת דרישות

הדרישה המרכזית בפרויקט היא להגיע למודל חיזוי מדויק שיוכל לשמש ככלי תומך בקבלת החלטות בשוק ההון. מעבר לכך, בחנו דרישות נוספות:

- אין מגבלות אבטחה או חוק משמעותיות על הנתונים שבהם אנו משתמשים כרגע, מכיוון שמדובר בנתוני סדרות זמן פומביים הנגישים ממקורות חיצוניים כמו Yahoo Finance.
- תזמון הפרויקט מוגדר מראש, ושנינו מיושרים היטב לעמידה בלוח הזמנים שנקבע, תוך התחשבות בשלבים השונים (ניקוי נתונים, בניית המודל, וניתוח התוצאות).
- תוצאות הפרויקט אינן מיועדות לפריסה מסחרית כרגע, אלא להגשה כחלק מפרויקט גמר, ולכן אין צורך בממשקים מתקדמים כמו פרסום באינטרנט או אינטגרציה למסדי נתונים.

5.2 הבהרת הנחות

במסגרת הבהרת ההנחות, לקחנו בחשבון את ההשפעות הכלכליות שעלולות להשפיע על תהליך הפרויקט ותוצאותיו. אחת ההנחות המרכזיות היא שאין גורמים כלכליים ישירים, כמו דמי ייעוץ או מוצרים מתחרים, שיכבידו על הפרויקט. עם זאת, אנו מודעים לכך שריבית היא גורם כלכלי משמעותי בשוק ההון, ולכן יש לה פוטנציאל להשפיע על הנתונים ועל דיוק החיזוי. ריבית עשויה להשפיע על מחירי המניות באופן עקיף, למשל דרך שינויים בשווי החברות, בתזרימי המזומנים שלהן או במחזורי המסחר. בשלב זה, אנו מניחים שהשפעות הריבית משתקפות במחירי המניות ההיסטוריים שעליהם מבוסס המודל שלנו.

בנוסף, ישנה הנחה כי הנתונים שבידינו אמינים, אך אנו מודעים לכך שייתכן שיהיו אי-שלמות או חוסרים בנתונים, ובמקרה כזה נדרש להשלים ממקורות חיצוניים. לבסוף, נותני החסות שלנו מצפים לקבל תוצאות מעשיות וברורות לחיזוי, תוך הבנה בסיסית של המודלים שבהם נעשה שימוש, אך ללא דרישה להיכנס לעומק הטכני של המודלים עצמם.

5.3 אימות אילוצים

במסגרת אימות האילוצים של הפרויקט, ביצענו בדיקות כדי להבטיח שהעבודה מתבצעת בצורה חלקה וללא מגבלות שעלולות לעכב את ההתקדמות. ראשית, יש לנו גישה מלאה לכל הנתונים הדרושים, ואין צורך בסיסמאות או באישורים מיוחדים כדי לעבוד עם מקורות המידע שלנו, מאחר שהם מבוססים על נתונים ציבוריים ונגישים כמו Yahoo Finance.

שנית, אימתנו את כל האילוצים החוקיים הקשורים לשימוש בנתונים, ואנו בטוחים שאין מגבלות רגולטוריות או חוקיות על הנתונים שבידינו, מאחר שהם זמינים לשימוש פומבי למטרות ניתוח ולמידה.

לבסוף, מבחינת אילוצים כספיים, הפרויקט עומד בתקציב שהוגדר מראש. אנו משתמשים בכלים חינמיים כמו Python, Google Colab ומאגרים ציבוריים, כך שאין הוצאות משמעותיות מעבר להשקעה בזמן ובמשאבים שכבר יש ברשותנו. עם זאת, במידה ויידרש שילוב נתונים נוספים או חישובים מתקדמים, נבחן את האפשרות להרחיב את התקציב בהתאם לצורך.

באופן כללי, כל האילוצים המרכזיים טופלו ואומתו, ואנו בטוחים שהפרויקט מתקדם בכיוון הנכון ללא חסמים משמעותיי

6. סיכונים ותוכניות חירום

זיהוי והתמודדות עם סיכונים פוטנציאליים הם חיוניים להצלחת פרויקט חיזוי המניות. חלק זה מזהה סיכונים מרכזיים הקשורים ללוחות זמנים, מגבלות תקציב, איכות נתונים ותוצאות, יחד עם תוכניות חירום להתמודדות עם כל סיכון.

הערכת סיכונים ותוכניות חירום

סיכוני לוחות זמנים

- **סיכון:** ייתכן שהפרויקט ייקח יותר זמן מהמתוכנן עקב מורכבויות לא צפויות בעיבוד נתונים, פיתוח מודלים או שלבי אימות.

תוכנית חירום:
- יצירת לוח זמנים מפורט עם אבני דרך קטנות לניטור התקדמות באופן שוטף.
- הקצאת זמן נוסף לשלבים קריטיים, כמו ניקוי נתונים וכיוונון מודלים.
- מתן עדיפות למשימות והתמקדות במטרות המרכזיות במקרה של אילוצי זמן.

סיכונים פיננסיים

- **סיכון:** מגבלות תקציב או עלויות בלתי צפויות, כמו רכישת נתונים חיצוניים או משאבים חישוביים נוספים, עלולות להכביד על משאבי הפרויקט.

תוכנית חירום:
- שימוש במערכי נתונים בקוד פתוח וכלים חינמיים ככל הניתן.
- ניצול משאבים אקדמיים, כגון קרדיטים לשירותי ענן או תשתית חישובית המסופקים על ידי המכללה.
- תכנון הדרגתי של הפרויקט להימנעות מהוצאות מיותרות בשלב מוקדם.

סיכוני נתונים

- **סיכון:** הנתונים עשויים להיות באיכות נמוכה, לא שלמים, או לא מייצגים את המורכבות של שוק המניות.

תוכנית חירום:
- יישום תהליכי ניקוי ועיבוד יסודיים לטיפול בנתונים חסרים או רעשים.
- השלמת פערי נתונים על ידי מקורות חלופיים, כמו APIs פיננסיים או יצירת נתונים סינתטיים.
- הערכת איכות הנתונים באופן שוטף והתאמת המודל בהתאם.

סיכוני תוצאות

- **סיכון:** ייתכן שתוצאות הראשוניות לא יעמדו בציפיות, עם דיוק נמוך של המודל או היעדר תובנות ישימות.

תוכנית חירום:

- בחינה מחודשת של בחירת מאפיינים ועיבודם כדי להבטיח שהמודל כולל משתנים רלוונטיים.
- ניסוי עם אלגוריתמים חלופיים או מודלים היברידיים לשיפור הביצועים.
- קבלת משוב ממנחים אקדמיים ומומחים פיננסיים כדי לחדד את הגישה.

7.טרמינולוגיה

כדי להבטיח תקשורת חלקה בין בעלי עניין טכניים ולא טכניים המעורבים בפרויקט חיזוי המניות, חיוני לקבוע הבנה משותפת של מונחים מרכזיים.

מילון מונחים

מונחים עסקיים

1. **מניה:** נייר ערך המייצג בעלות בחברה ומעניק לבעלים תביעה על חלק מנכסי החברה ורווחיה.
2. **בורסה:** מרכז בה נסחרים מניות, אג"ח ונגזרים.
3. **תנודתיות שוק:** רמת השונות במחירי מניות לאורך זמן, המושפעת מגורמים כמו שינויים כלכליים, אירועים גיאופוליטיים או סנטימנט המשקיעים.
4. **נפח מסחר:** המספר הכולל של מניות שנסחרו במהלך תקופת זמן מסוימת.
5. **דוח רווח:** דוח פיננסי שמפרסמת חברה המתאר את רווחיותה וביצועיה הפיננסיים בתקופה מסוימת.
6. נטישה Churn - תהליך שבו לקוחות עוזבים את השירות או המוצר.

מונחי מדעי הנתונים

1. **למידת מכונה:** תחום של בינה מלאכותית שבו מודלים לומדים דפוסים מנתונים כדי לבצע תחזיות או קבלת החלטות ללא תכנות מפורש.
2. **דיוק חיזוי:** אחוז התוצאות החזויות נכון בהשוואה לתוצאות בפועל.
3. **מאפיינים (features):** משתנים או תכונות המשמשים כקלט במודל למידת מכונה, כגון נפח מסחר או אינדיקטורים כלכליים.
4. **התאמת יתר (Overfitting):** מצב שבו מודל למידת מכונה מציג ביצועים טובים על נתוני אימון אך ביצועים גרועים על נתונים שלא נראו בשל מורכבות יתרה.
5. **ולידציה צולבת (Cross-Validation):** טכניקה להערכת ביצועי מודל למידת מכונה על ידי חלוקת הנתונים לתתי-קבוצות לאימון ובדיקה.
6. **ניתוח סנטימנט:** שימוש בעיבוד שפה טבעית להערכת הטון (חיובי, שלילי או ניטרלי) של נתוני טקסט, כגון חדשות פיננסיות או פוסטים במדיה חברתי

מונחים ייחודיים לפרויקט

1. **תנועת מניה:** שינוי במחיר מניה, עלייה או ירידה, לאורך תקופת זמן מסוימת.
2. **גורמים משפיעים מרכזיים:** גורמים המשפיעים משמעותית על תנועות מחירי מניות, כגון סנטימנט שוק או תנאים מקרו-כלכליים.
3. **נתונים סינתטיים:** נתונים שנוצרו באופן מלאכותי למילוי פערים או לשיפור מערך הנתונים לאימון מודלים של למידת מכונה.
4. **לוח מחוונים:** ממשק חזותי המציג מדדים ותובנות מרכזיות שנוצרו על ידי מודל החיזוי.

8. עלויות ותועלות

חלק זה מעריך את העלויות המשוערות הקשורות לפרויקט חיזוי המניות ומשווה אותן עם התועלות הפוטנציאליות. על ידי שקילת ההוצאות מול הערך של התובנות והתוצרים, ניתוח זה מדגיש את הכדאיות וההשפעה הכוללת של הפרויקט.

עלויות משוערות

1. **איסוף נתונים ונתונים חיצוניים:**
   - נכון לעכשיו, אנו משתמשים בנתונים ציבוריים הזמינים בחינם ממקורות כמו Yahoo Finance. במידה ונדרש להרחיב את הנתונים באמצעות מקורות חיצוניים בתשלום, העלות תעמוד על כ-100-200 דולר לשירותי API מתקדם (למשל TradingView או Alpha Vantage).

2. **פריסת תוצאות:**
   - מכיוון שמטרת הפרויקט היא אקדמית, אין צורך בפריסה מסחרית או מערכת מורכבת להצגת התוצאות. ההצגה תבוצע בצורה מקומית או באמצעות כלים חינמיים כמו Jupyter Notebook או Google Colab, ולכן אין עלויות מיוחדות בשלב זה.

3. **עלויות תפעול:**
   - שימוש בפלטפורמות ענן כמו Google Colab הוא חינמי, אלא אם נדרשת הרחבת משאבים (GPU, אחסון נוסף), שעלולה לעלות כ-10-30 דולר לחודש במקרה של שימוש מתקדם.

היתרונות הפוטנציאליים

1. **המטרה העיקרית המושגת:**
   - המטרה המרכזית של הפרויקט היא לבנות מודל לחיזוי סדרות זמן שיוכל לשמש כבסיס לתובנות בשוק ההון. הצלחה בפרויקט תספק כלי חיזוי בעל ערך שיכול להיות מורחב או מותאם לשימוש מעשי.

2. **תובנות נוספות מחקירת נתונים:**
   - מעבר למודל החיזוי, ניתוח הנתונים עשוי לחשוף דפוסים, מגמות והתנהגויות מעניינות בשוק המניות, אשר יכולים לשמש להבנה עמוקה יותר של גורמים כלכליים המשפיעים על מחירי מניות.

3. **יתרונות מהבנת נתונים טובה יותר:**
   - התהליך כולו יספק לנו, כצוות, הבנה מעמיקה של עבודה עם סדרות זמן, בניית מודלים וניתוח פיננסי. ידע זה הוא בעל ערך להמשך דרכנו, בין אם במישור האקדמי או המקצועי.

<u>סיכום השורה התחתונה</u>

הפרויקט מצריך השקעה כספית נמוכה מאוד, בעיקר בשל שימוש בכלים ונתונים חינמיים. מנגד, היתרונות של הפרויקט הם משמעותיים מאוד, הן ברמה המעשית (כלי חיזוי מדויק) והן ברמה האקדמית והמקצועית (ידע וניסיון מעמיק בניתוח נתונים). לכן, השורה התחתונה היא שהפרויקט עומד ביחס עלות-תועלת חיובי, ויש לו פוטנציאל לספק ערך רב.

משימה 9 מטרות מדעי הנתונים וקריטריונים להצלחה

9.1 מטרות מדעי הנתונים

הגדרת הבעיה במדעי הנתונים:
המטרה העיקרית היא לבנות מודל סיווג שיזהה לקוחות בסיכון לנטישה על בסיס נתוני התנהגות, דמוגרפיה והיסטוריית רכישות. בנוסף, נשתמש ב-רגרסיה כדי לחזות את הערך הפוטנציאלי של לקוחות ולשפר את מאמצי השיווק.

מטרות טכניות:

1. בניית מודל חיזוי מבוסס **רגרסיה לינארית ועצי החלטה**, לדוגמה שימוש ב- Decision Trees, XGBoost
2. יצירת תחזיות מחירים עם תוקף של שלושה חודשים קדימה, כאשר דיוק התחזית ייבחן מול נתוני האמת (Actuals).
3. חישוב ביצועי המודל באמצעות מדדים כמותיים כמו Mean Absolute Error (MAE) ו-Root Mean Squared Error (RMSE), והגעה לערכים נמוכים ככל הניתן.

תוצאות מספריות רצויות:

- דיוק של לפחות 85% במודל הסיווג (Accuracy).
- שיעור זיהוי חיובי אמיתי (True Positive Rate) של לפחות 80%.
- תחזיות שיביאו להפחתת נטישת לקוחות ב-10%.

9.2 קריטריוני הצלחה במדעי הנתונים

שיטות להערכת מודל:

- דיוק (Accuracy): מדד אחוז התחזיות הנכונות מכלל התחזיות.
- F1-Score: איזון בין דיוק לזיהוי, חשוב במיוחד אם עלות נטישה גבוהה.

אמות מידה להצלחה:

- הצלחה תוגדר כמודל שמגיע לדיוק של לפחות 85% עם F1-Score מעל 0.8.
- רמת ביצועים עקבית על סט נתונים חדש שלא נכלל באימון (Validation Set).

מדידות סובייקטיביות:

- שביעות רצון של מנהלי צוותים עסקיים מהמסקנות והתובנות.
- תהליך העבודה צריך להיות מובן ונגיש לצוותים טכניים ולא טכניים כאחד.

פריסת מודל כתנאי להצלחה:

- תוצאות המודל יפורסמו בלוח מחוונים אינטראקטיבי שמשולב בתהליכים קיימים.
- הצלחה תכלול הפחתת זמן תגובה למקרים של לקוחות בסיכון, עד 48 שעות מקבלת התראה.

תכנון ראשוני לפריסה:

1. הגדרת תשתית לאחסון וטעינת מודל בזמן אמת.
2. יצירת ממשקי API לחיבור בין המודל לבין מערכות CRM. 3.
פיילוט קטן עם 10% מנתוני הלקוחות לפני פריסה מלאה.

משימה 10 תוכנית הפרויקט

תוכנית פרויקט למדעי הנתונים: מסמך אב לניהול מוצלח

לפני תחילת הפרויקט, נערכו פגישות עם כל בעלי העניין, כולל צוותי מדעי הנתונים, אנליסטים עסקיים, מנהלי פרויקטים ונציגי מחלקות רלוונטיות. בפגישות אלו הוצגה תוכנית הפרויקט, מטרותיה, שלביה והמשאבים הנדרשים. המשתתפים סיפקו משוב, והוכנסו התאמות בהתאם להערות שצצו, כדי להבטיח שכל המעורבים מודעים ומסכימים על התוכנית.

פעילויות הפרויקט:

1. **איסוף והכנת נתונים:**
   - זיהוי מקורות נתונים פנימיים וחיצוניים (2-3 ימים).
   - ניקוי ועיבוד נתונים, כולל טיפול בחוסרים ובפורמטים שונים (2-3 ימים).

2. **חקר נתונים ואנליזה ראשונית:**
   - זיהוי תבניות ודפוסים ראשוניים בעזרת כלי ויזואליזציה (שבוע).

3. **פיתוח מודלים:**
   - יצירת מודלים ראשוניים של סיווג ורגרסיה (שבוע).
   - חזרה על שלבי הדוגמנות בהתאם לתוצאות הבדיקות (2-3 שבועות).

4. **הערכת ביצועי מודלים:**
   - בדיקת דיוק, F1-Score, ו-ROC-AUC באמצעות סט בדיקה (2-3 ימים).
   - שיפור המודלים על בסיס הערכות הביצועים (2-3 ימים).

5. **פריסת התוצאות:**
   - פיתוח לוח מחוונים אינטראקטיבי ופריסת תשתית API (שבועות 2-3).
   - בדיקות משתמשים ופיילוט על 10% מנתוני הלקוחות (2-3 שבועות).

6. **מעקב ותחזוקה:**
   - מעקב אחר ביצועי המודלים בשטח (חודש ראשון לאחר הפריסה).
   - שיפור רציף של המודלים על בסיס נתונים חדשים.

משאבים ומאמץ

- משאבי תוכנה: Python, Google Colab, ספריות ניתוח נתונים (Pandas, NumPy), וספריות ויזואליזציה.
- משאבי חומרה: מחשב עם מעבד i5 ומעלה, 8GB RAM.
- מאמץ: כ-20-25 שעות עבודה משותפת לשני חברי הצוות.

נקודות החלטה ובקשות עיון

- לאחר שלב איסוף הנתונים: אישור כי הנתונים תקינים ומוכנים לעבודה.
- לאחר ניתוח הנתונים (EDA): בחירה במשתנים המרכזיים לחיזוי.
- לאחר הערכת ביצועי המודל: החלטה אם לבצע חזרתיות על שלב הדוגמנות.

משימה 11 הערכה ראשונית של כלים וטכניקות

בחירת כלים

לאור הצרכים העסקיים והפרוייקטים, נבחרו הכלים הבאים כמתאימים ביותר:

1. **Python:**
   - *יתרונות:*
     - שפה גמישה ופופולרית מאוד במדעי הנתונים.
     - מציעה ספריות עשירות כמו NumPy, pandas, scikit-learn ו-TensorFlow לניתוח נתונים, למידת מכונה ועיבוד שפה טבעית.
     - מתאימה לכל שלבי הפרויקט, החל מאיסוף נתונים ועד לפריסה.
2. **R:**
   - *יתרונות:*
     - חזקה במיוחד לניתוח סטטיסטי וויזואליזציה.
     - מתאימה לפרויקטים עם דגש על מודלים סטטיסטיים ותובנות ויזואליות.
     - מציעה ספריות כמו ggplot2 ו-dplyr.
3. **SQL:**
   - *יתרונות:*
     - הכרחית לעבודה עם מסדי נתונים ולביצוע שאילתות מורכבות.
     - מאפשרת גישה וניתוח מהירים לנתונים ממקורות מבוזרים.
4. **Power BI/:**
   - *יתרונות:*
     - כלים עוצמתיים ליצירת דוחות ודשבורדים אינטראקטיביים.
     - משמשים לניתוח תובנות ולהצגתן לבעלי העניין בצורה ידידותית.

בחירת טכניקות

בהתאם לסוג הבעיה העסקית והנתונים הקיימים, נבחרו הטכניקות הבאות:

1. **סיווג (Classification):**
   - מתאים כאשר המטרה היא לחזות קטגוריות מוגדרות מראש (למשל, האם לקוח יעזוב או לא).
   - *שימוש* במודלים כמו Logistic Regression, Decision Trees ו-Random Forests.

2. **אשכולות (Clustering):**
   - מתאים לזיהוי קבוצות הומוגניות בתוך הנתונים (למשל, פילוח לקוחות).
   - שימוש באלגוריתמים כמו K-Means ו-Hierarchical Clustering.

3. **רגרסיה (Regression):**
   - משמשת לחיזוי משתנה רציף (למשל, תחזית מכירות).
   - שימוש במודלים כמו Linear Regression ו-Gradient Boosting.

4. **עיבוד שפה טבעית (NLP):**
   - מתאים לניתוח טקסטים ומסמכים (למשל, ניתוח סנטימנט או מיון מיילים).
   - שימוש בספריות כמו NLTK ו-spaCy.

5. **למידת מכונה (Machine Learning):**
   - כולל טכניקות מתקדמות לחיזוי ולמידה אוטומטית של תבניות.
   - שימוש במודלים כמו Support Vector Machines ו-Neural Networks.

Stock Market Prediction in the U.S. - Analysis and Forecasting of S&P 500 Trends

data understanding report

המכללה האקדמית
עמק יזרעאל
ע"ש מקס שטרן

מגישים - שינה תחאוחו 213381569 , דין תחאוכו 318291705

מנחה - ג'ניה גוטפריד

מקורות נתונים

נתוני מניות S&P 500

הנתונים שלנו נלקחו באמצעות ספריית YFinance ב Python , המאפשרת גישה למידע פיננסי היסטורי ועדכני של מניות, מדדים וכו..
בחרנו להשתמש בספריה זו משום שהיא חינמית, נגישה ומהימן לנתוני שוק ההון.
בנוסף Yfinance מספקת גישה למידע רב כולל מחירי פתיחה, סגירה , שיאים יומיים ונפחי מסחר שישמשו אותנו כבסיס לאנליזה.

הנתונים שלנו כוללים

- נתוני מחיר מניה יומי (פתיחה, סגירה, שיא, שפל)
- נפח מסחר יומי
- נתונים מתואמים (Adjusted close) המתחשבים בדיבידנדים ובשינויים נוספים
- טווח זמן - הנתונים כוללים מחירים היסטוריים בטווחים משתנים

נתוני ריביות

נתוני הריביות של ארה"ב נלקחו ממאגר FRED federal reserve economic data. מאגר זה מספק נתוני ריביות כלליים כמו ריבית הפדרל ריזרב, ריביות אג"ח ממשלתיות וריביות נוספות.

ריבית היא אחד הגורמים המרכזיים המשפיעים על שוק המניות -

- ריביות גבוהות מייקרות את עלויות ההלוואות וההשקעות, דבר שעלול להקטין את רווחי החברות ולהוריד את מחירי המניות
- ריביות נמוכות - מעודדות השקעות, מגדילות את הביקוש למניות ויכולות להעלות את מחירי השוק

לא נרכשו נתונים נוספים, כל הנתונים נלקחו ממקורות פתוחים ואמינים, כך שאין צורך להתחשב בעלויות נתונים

בדיקת נתונים ראשונית

נתוני מניות + נתוני רביות

- תאריך (Date)
- מחיר פתיחה (Open)
- מחיר סגירה (Close)
- המחיר הגבוה ביותר לאותו יום (High)
- המחיר הנמוך ביותר לאותו יום (Low)
- מחיר מתואם (Adj Close)
- נפח המסחר (Volume)
- ריבית

המאפיינים החשובים

1. תאריך (Date)

תאריך הוא מאפיין קריטי בניתוח סדרות זמן, שכן הוא מאפשר מעקב אחר שינויים לאורך זמן. הוא משמש כנקודת ייחוס לשילוב נתוני המניות ונתוני הריביות שייבאנו. בעזרת התאריך אנו יכולים לבדוק איך אירועים כלכליים או פוליטיים מסוימים משפיעים על מחירי מניות.

<u>שימוש אפשרי</u>

- בניית גרפים המציגים מגמות לאורך זמן
- זיהוי דפוסים עונתיים, האם יש עלייה במחירים בתקופה מסוימת של השנה

2. מחירי פתיחה (Open)

מחיר פתיחה מייצג את המחיר שבו בוצעה העסקה הראשונה ביום המסחר. הוא משקף את מצב השוק לאחר אירועים שקרו מחוץ לשעות המסחר (כגון הודעות כלכליות או חדשות חשובות)
למשל אם במהלך הלילה פורסם דוח חיובי על החברה, מחיר הפתיחה עשוי להיות גבוה יותר מסגירת היום הקודם.

<u>השפעה של נתוני סוף יום</u>

- מחיר הפתיחה מושפע ממחיר הסגירה של היום הקודם ומציפיות המשקיעים, ולכן הוא מספק תובנה על איך שוק המניות מגיב לאירועים שקרו לאחר הסגירה

<u>ניתוח מסחר יומי</u>

- סוחרים יומיים מתמקדים בפערים בין מחירי פתיחה למחירי המסחר במהלך היום לכן זה קריטי עבורם

3. מחיר סגירה (Close)

מחיר הסגירה נחשב למחיר החשוב ביותר ביום המסחר, מכיוון שהוא משקף את הערכת השוק את שווי המניה בסיום יום המסחר. הוא משמש כבסיס לניתוחים יומיים ומספק נקודת ייחוס ליום הבא

<u>חישוב תשואות</u>

תשואות יומיות מחושבות על בסיס השינוי בין מחירי הסגירה של ימים עוקבים. זהו מדד מרכזי בניתוחים פיננסיים והשקעות

תשואה יומית = מחיר סגירה (יום קודם) / (מחיר סגירה (יום נוכחי) - מחיר סגירה(יום קודם))

<u>החלטות משקיעים</u>

משקיעים מוסדיים ומנהלי קרנות משתמשים במחיר הסגירה כדי להעריך את ביצועי המניות בתיקי ההשקעות שלהם

4. מחיר מתואם (Adj Close)

זהו המדד המדויק ביותר לערך המניה, מכיוון שהוא מתחשב באירועים טכניים כמו חלוקת דיבידנדים או פיצולי מניות. הוא מאפשר להשוות בין מניות לאורך זמן בצורה הוגנת יותר

<u>שימוש אפשרי</u>

- חישוב תשואות יומיות, שבועיות או חודשיות
- זיהוי מגמות בשוק או בביצועי מניות ספציפיות
- הוא יכול להוות בסיס למודלים חיזוי ולמידת מכונה.

5. ריבית (interest Rate)

ריבית היא גורם מאקרו כלכלי המשפיע ישירות על השוק. כאשר הריבית עולה, עלות גיוס הכספים עולה, מה שמוביל פעמים רבות לירידה במחירי המניות
ריבית יכולה לשמש להבנת הקשרים בין שווקים שונים, כמו שוק המניות ושוק האג"ח

6. נפח המסחר (Volume)

נפח המסחר מעיד על מידת העניין והפעילות של המניה. נפחים גבוהים במיוחד עשויים להצביע על אירועים חריגים כמו דיווחי רווחים או חדשות כלכליות.

נפח מסחר בשילוב עם מחירים יכול לסייע לזהות מגמות בשוק - האם המחיר עולה יחד עם עלייה בנפח המסחר? זה עשוי לאותת על מגמה חיובית.

מאפיינים לא רלוונטים

- **מחירי שיא ושפל (High, Low)**

<u>לא משקפים את מצב השוק ביום יום המסחר</u>

בעוד שמחירי הפתיחה והסגירה מספקים תובנות ברורות על ציפיות המשקיעים (פתיחה) ותוצאות המסחר (סגירה), המחירים הגבוה והנמוך מציינים רק נקודות קצה זמניות שייתכן ולא מייצגות את הביצועים בפועל.

<u>תנודתיות רגעית</u>

מחירי ה-High וה-Low יכולים להיות תוצאה של עסקאות בודדות (מקרים חריגים) שאינן משקפות את המגמה האמיתית או את תחושת השוק הכללית.
לדוגמה: אם מישהו ביצע עסקה חריגה שגרמה למחיר להגיע לשיא רגעי, המחיר הגבוה לא בהכרח מייצג את ערך המניה עבור רוב המשקיעים.

<u>מידע פחות שימושי למודלים לחיזוי</u>

מודלים לחיזוי ביצועי מניות מתמקדים לרוב במחירי הפתיחה, הסגירה והתשואות. מחירי ה-High וה-Low לא תמיד מוסיפים מידע ייחודי שמסייע בתחזיות.
במקרים רבים, הנתונים האלו מתואמים עם מחירי הפתיחה והסגירה

כמות הנתונים הקיימת

הנתונים שיש לנו מכסים תקופה של 15 שנים (ינואר 2010 עד ינואר 2025)
יש לנו 3775 רשומות (שורות) , וכל שורה מייצגת יום מסחר ב S&P 500

15 שנה היא תקופה ארוכה שמכסה מגוון רחב של מצבים בשוק ההון, כולל תקופות גאות ושפל כלכלי (למשל משבר הקורונה ב 2020, התאוששות הכלכלה, שינויי ריבית משמעותיים וכו)
כמות הנתונים מספקת כדי לזהות מגמות כלליות, כמו מגמת עלייה ארוכת טווח במדד ה s&p 500 ותנודתיות לאורך זמן.

דיוק התחזיות תלוי בשיטת המידול שאנו נבחר לעשות. אם אנו נבחר להשתמש במודלים פשוטים כמו רגרסיה לינארית, הכמות שיש לנו בהחלט לא מספיקה כדי לבצע תחזיות בסיסיות.
אם נבחר להשתמש במודלים מתקדמים יותר, הכמות תספיק ברוב המקרים, אבל איכות התחזיות תושפע גם מאיכות הנתונים, מהשפעות חיצוניות (כמו אירועים כלכליים) ומהיכולת של המודל ללכוד דפוסים מורכבים.

בחירת Power BI ככלי לניתוח ושילוב נתונים

Power BI הוא כלי חזק לניתוח נתונים ויצירת ויזואליזציות מתקדמות. שילוב הנתונים שבחרת באמצעות Power BI הוא בחירה מצוינת, שכן הוא מאפשר:

- <u>שילוב מקורות שונים -</u>
  שילבנו נתוני מניות מ-yfinance יחד עם נתוני ריביות שנאספו מ-FRED.

- <u>ויזואליזציה ברורה -</u>
  יצירת גרפים ודוחות דינמיים להבנת דפוסים והשפעות.

- <u>נוחות בשילוב נתונים מורכבים-</u>
  Power BI מציע כלים פשוטים למיזוג נתונים על בסיס עמודות משותפות (כמו תאריך).

זיהוי נתונים חסרים בערכי הריבית

במהלך שילוב הנתונים מ-FRED (נתוני ריבית) ו-yfinance (נתוני מניות), זוהו ערכים חסרים בעמודת הריבית. הדבר התרחש בעיקר בימים שבהם אין נתוני מסחר במניות (כגון סופי שבוע או חגים) או ימים שבהם מאגר ה-FRED לא מספק נתונים מלאים.
ערכים חסרים מהווים בעיה משמעותית בניתוח, במיוחד כאשר מדובר בפרמטר כמו הריבית שמשפיע ישירות על שוק המניות.

<u>גורמים אפשריים לנתונים חסרים</u>

- חוסר התאמה בלוחות זמנים -
  נתוני מניות מופיעים עבור ימי מסחר בלבד, בעוד שנתוני ריבית עשויים להיכלל גם בימים בהם אין מסחר.

- היעדר נתוני ריבית ספציפיים -
  ייתכן שחלק מהתאריכים ב-FRED חסרים נתוני ריבית עקב עדכונים שטרם הושלמו במערכת.

<u>השפעת הנתונים החסרים על הניתוח</u>

- השפעה על המתאם -
  חוסר ברציפות בנתוני הריבית עלול להטות את החישובים הסטטיסטיים, כמו מתאם בין ריבית לתשואת מניות.

- השפעה על המודלים -
  מודלים לחיזוי, כמו רגרסיה, עלולים להיכשל או להיות לא מדויקים כאשר יש חוסרים משמעותיים בנתוני קלט.

טיפול בערכים חסרים

בפרויקט נתקלנו ב כ 12 שורות בעמודת הריבית שהכילו ערכים חסרים. מכיוון שמדובר בנתוני סדרות זמן שבהם יש תלות בין ערכים סמוכים, בחרנו למלא את הערכים החסרים על בסיס בערך של היום הקודם.

תהליך המילוי

1. <u>זיהוי ערכים חסרים</u>
   זיהינו את השורות בעמודת הריבית עם ערכים חסרים באמצעות בדיקת ()isnull

2. <u>שיטת המילוי</u>
   לכל ערך חסר מילאנו את הערך מהשורה הקודמת (שיטת Forward Fill). שיטה זו מתאימה במיוחד לנתוני סדרות זמן שבהם קיימת המשכיות טבעית בערכים לאורך הזמן

```
merged_df['EFFR'] = merged_df['EFFR'].ffill()
```

נתוני הריביות מתארים מגמה לאורך זמן, ושיטת ה Forward Fill מניחה שהערך של היום הקודם הוא הניחוש הסביר ביותר לערך ביום החסר.
שיטה זו קלה ליישום, מהירה להבנה, ומספקת מענה טוב כשמספר הערכים החסרים קטן יחסית בהשוואה לכמות הנתונים שיש

תיאור הנתונים

כמות הנתונים

כמות הנתונים היא מדד קריאי שמכתיב את האיכות והדיוק של הניתוחים שלנו. נתונים רבים יותר יכולים להביא לתובנות מדויקות יותר וליכולת לבצע תחזיות עם רמת ביטחון גבוהה יותר.
עם זאת גם כמות גדולה מדי של נתונים יכולה להכביד על הערכת ועל הניתוחים, לכן יש צורך באיזון בין הכמות לבין הרלוונטיות של הנתונים.

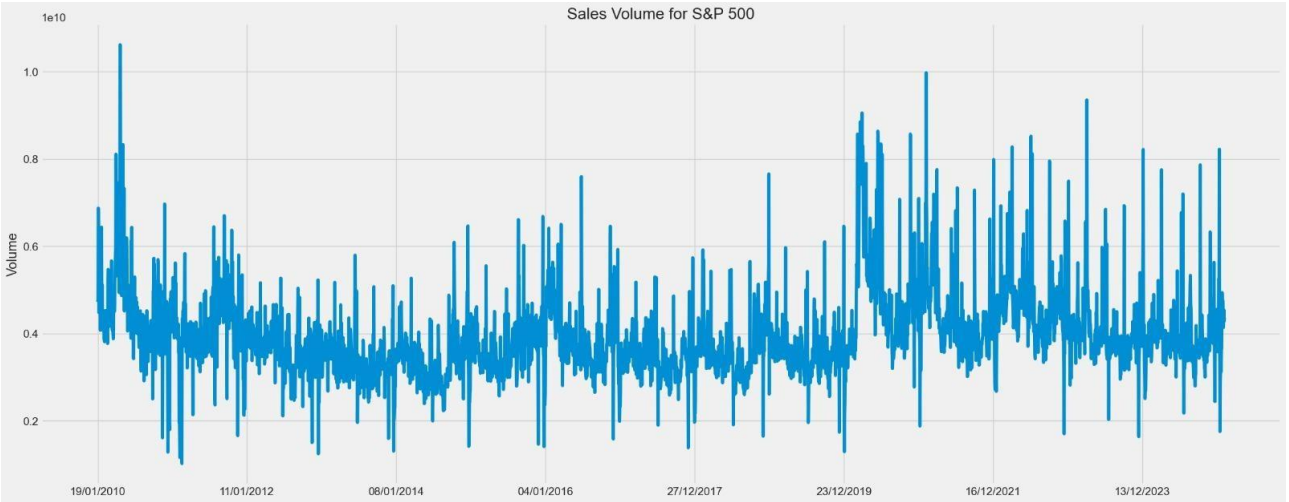
הנתונים שיש לנו מכסים תקופה של 15 שנים (ינואר 2010 עד ינואר 2025)

- מספר תצפיות (שורות) - במקרה שלנו יש 3775 ימים של נתוני s&p500
- מספר המשתנים (עמודות) - 8 מאפיינים שיש בכל תצפית (מחיר פתיחה, מחיר סגירה, נפח מסחר, ריבית, גבוהה, נמוך, סגירה מותאם)

סוגי ערכים

interest Rate	Volume	Low	High	Adj Close	Close	open	Date
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Date

חקר נתונים



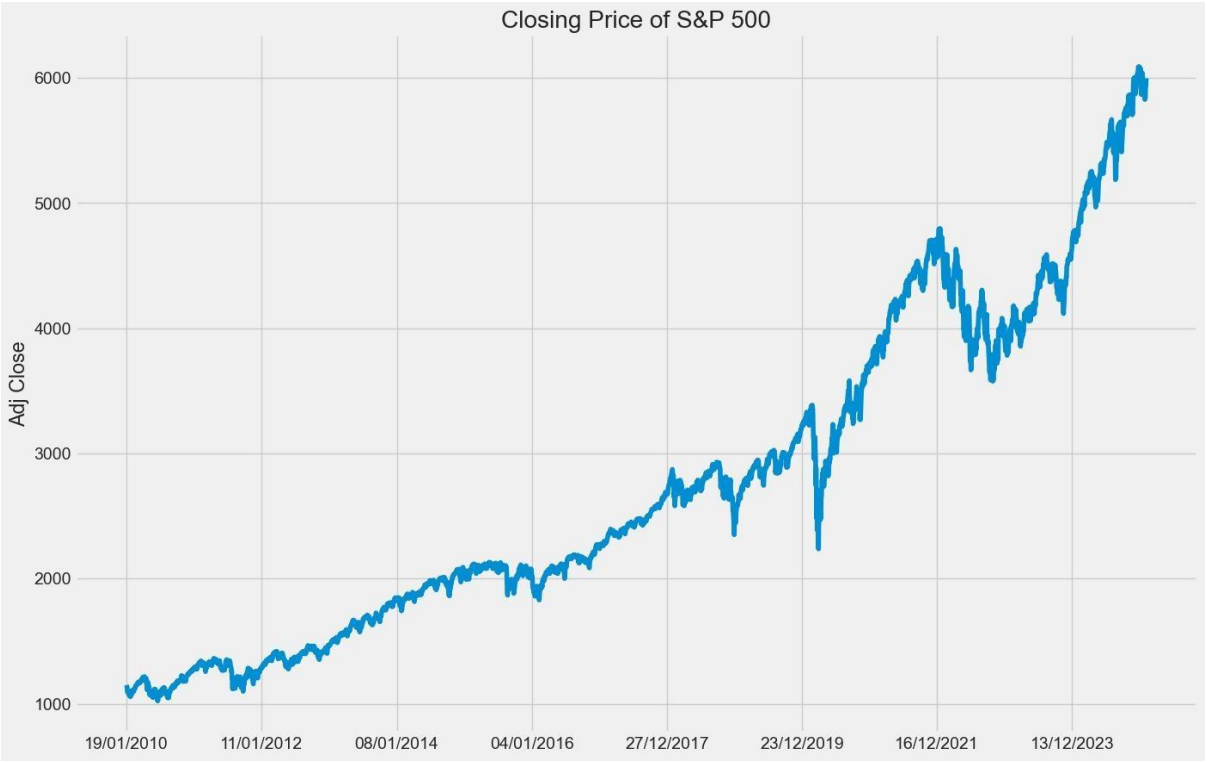
הגרף מתאר את נפח המסחר (volume) במדד ה s&p 500 לאורך זמן. ניתן להבחין בשונות משמעותית בנפח המסחר, עם תקופות של עלייה חדה וירידה

ניתוח הגרף

- נפח מסחר גבוה משקף פעילות מוגברת בשוק, שיכולה להיות תוצאה של אירועים כלכליים או פוליטיים משמעותיים
- אירועים כמו משברים כלכליים, החלטות של הבנק הפדרלי על הריבית או פרסום דוחות רבעוניים עשויים להסביר את התנודתיות
- בניתוח נפח מסחר ניתן לשלב אינדקטוריים טכניים כמו OBV(On-Balance-Volume)h או שימוש במודלים סטטיסטיים כדי לזהות קשרים בינו לבין שינויים במחיר המניה
- ניתן לראות שינויים חדים בנפח המסחר בזמנים מסוימים. למשל בזמן משברים כלכליים (כמו משבר הקורונה בשנת 2020) יש עלייה חדשה בנפח, שיכולה להצביע על פאניקה בשוק

ניתוחים סטטיסטיים אפשריים

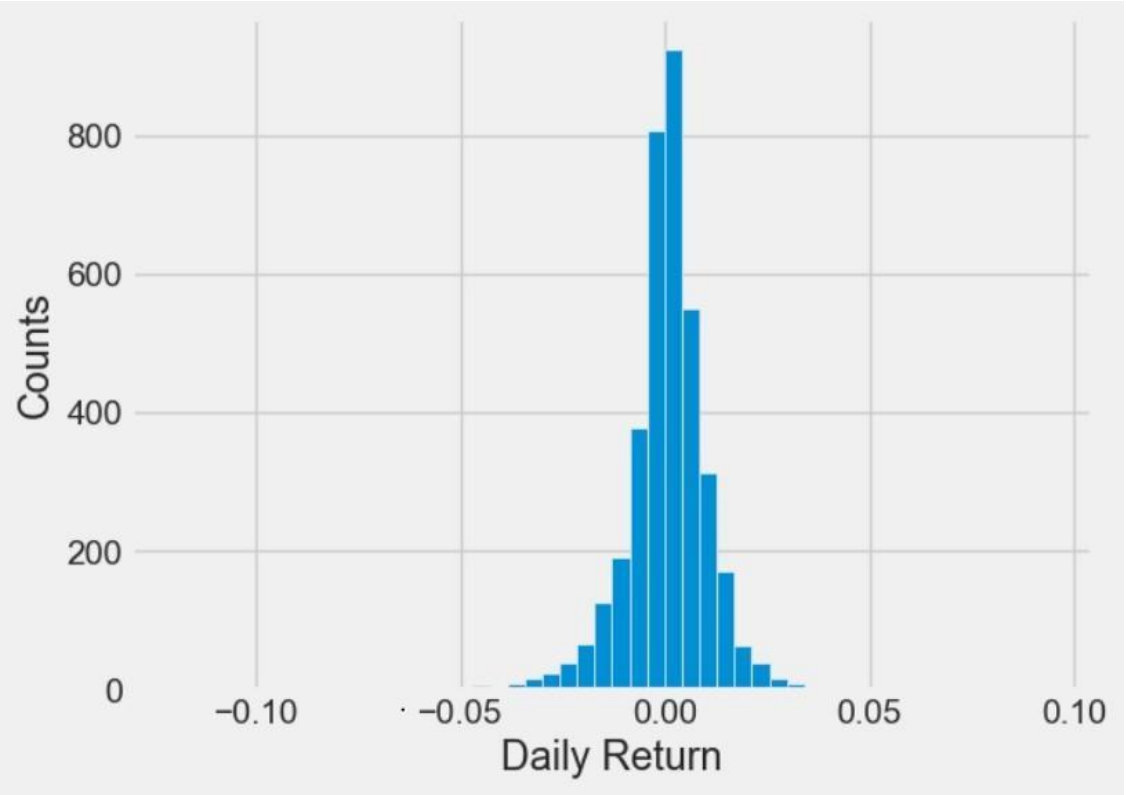
- ניתן לחשב את ממוצע נפח המסחר לאורך השנים ולבחון האם קיים גידול בנפח לאורך הזמן
-
- ניתן לבדוק סטיית תקן שמספקת תובנה על רמת התנודתיות



הגרף מתאר את המחיר הסגור המותאם של המדד שלנו לאורך זמן, עם מגמת עלייה מובהקת.

ניתוח

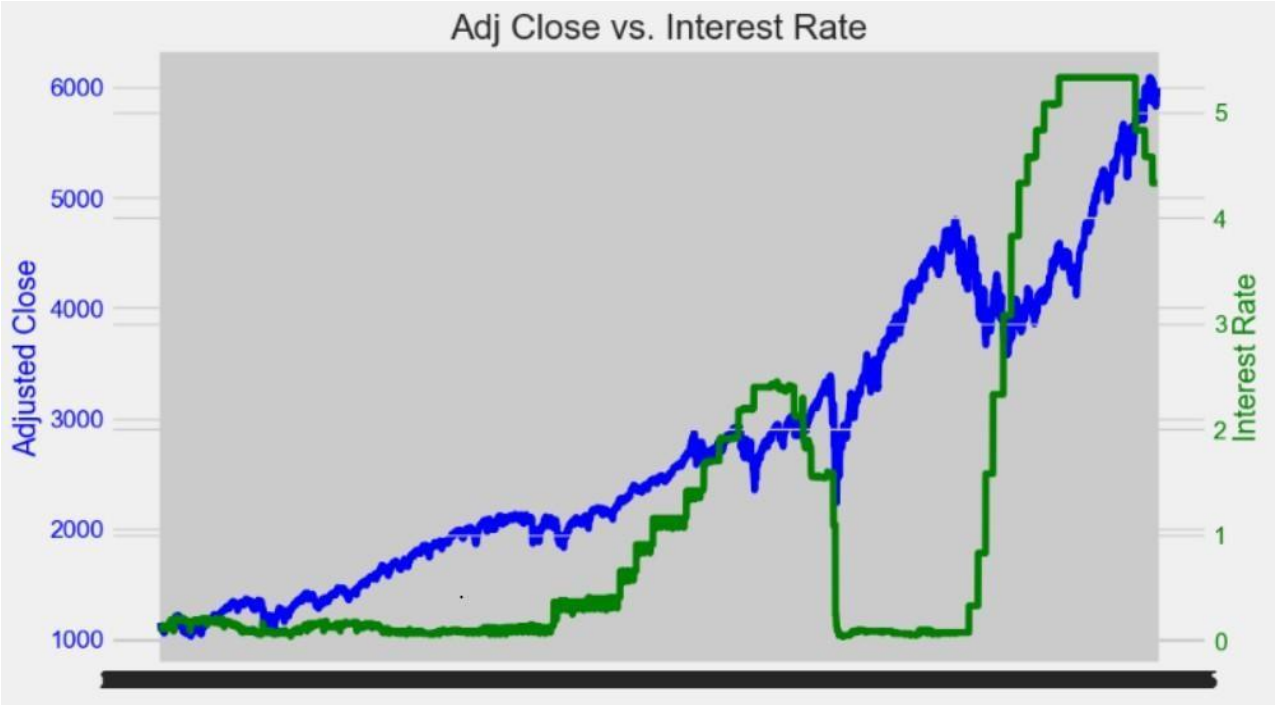
- העלייה מציינת את מגמת הצמיחה בשוק ההון האמריקאי לאורך השנים
- תנודות חזקות (כמו בשנת 2020) עשויים לשקף משברים גלובליים (למשל הקורונה)
- ניתן להוסיף ממוצעים נעים (MA) כדי לזהות נקודות תמיכה והתנגדות או לזהות מומנטום בשוק
- ייתכן שניתן לזהות מחזורים (Cycles) בעליות וירידות במדד. מחזוריות זו יכולה להיות קשורה לאירועים כלכליים כמו משברי ריבית



היסטוגרמה של התשואות היומית ב S&P 500

ניתוח

- רוב התשואות מרוכזות סביב 0, מה שמראה שהרוב המוחלט של הימים הם בעלי שינויים מינוריים
- זנבות ארוכים משני צידי ההיסטוגרמה עשויים להצביע על ימים עם תשואות קיצוניות שיכולים להיות תוצאה של אירועים פתאומיים
- התפלגות נורמלית - ההיסטוגרמה מזכירה התפלגות נורמלית, עם ריכוז גבוהה של תשואות סביב 0. כדאי לבדוק עד כמה ההתפלגות קרובה לנורמלית
- הזנבות חשובים במיוחד. הם משקפים ימים עם שינויים משמעותיים בשוק שיכולים להיות קריטיים להבנת סיכונים



גרף המשלב את מחיר הסגירה המתואם עם שיעורי הריביות

ניתוח

- ניתן להבחין בקשר מסוים בין עלייה בריבית לירידה בטווח הקצר בשוק ההון, אך בטווח הארוך ישנה מגמת עלייה
- ריבית גבוהה יכולה להפחית את הנזילות בשוק, מה שמוביל לירידה זמנית בערך המניות
- קשר זה יכול לשמש מודלים של חיזוי לתמחור מניות וזיהוי מניות לפי שינויים צפויים בריבית
- כל שהריבית עולה, יש נטייה לירידה בשוק ההון, מכיוון שהמשקיעים מעדיפים להשקיע באג"ח שמניבות רווח בטוח יותר. עם זאת, השוק מתאושש בטווח הארוך.
- לעיתים, שוק ההון מגיב לשינויים בריבית בעיכוב. ניתן לנתח את העיכוב הזה באמצעות מודלים סטטיסטיים (למשל, cross-correlation).

מענה על רשימת המשימות

2. אילו סוגי השערות גיבשנו לגבי הנתונים?

- **השערה ראשונית על מחירי סגירה:** המחירים צפויים לעלות בטווח הארוך (מגמת שוק שורי), עם תנודתיות מוגברת בזמן משברים כלכליים או שינויים בריבית.

- **השערה על הקשר בין נפח מסחר למחיר:** נפח מסחר גבוה עשוי להתרחש במקביל לתנודות מחירים חדות (עליות או ירידות).

- **השפעת הריבית:** שינויים בריבית משפיעים באופן שלילי על השוק בטווח הקצר, אך השוק מתאושש בטווח הארוך.

- **מחירי פתיחה מול סגירה:** בימים תנודתיים, הפער בין מחיר הפתיחה למחיר הסגירה יהיה גדול יותר.

2. כיצד שינו חקירות אלה את ההשערה הראשונית שלנו?

לאחר בדיקת הנתונים:

- **מתאם בין ריבית למחירי סגירה:** גילינו שייתכן עיכוב בהשפעת הריבית על מחירי המדד. כלומר, עליית ריבית לא תמיד מתבטאת בירידות מיידיות.

- **התנהגות נפח המסחר:** נפח מסחר גבוה נצפה בעיקר בתקופות של משבר כלכלי או הכרזות משמעותיות, אך לא תמיד מלווה בעליות מחירים.

- **התפלגות מחירים:** במהלך ניתוח התפלגות התשואות היומיות, התגלה שהתפלגות היא כמעט נורמלית, אך עם זנבות שמצביעים על ימים קיצוניים.

3. האם החקירות חשפו מאפיינים חדשים?

- **מאפיינים משמעותיים שהתגלו:**
  - קיימת עונתיות מסוימת במחירים ובנפח המסחר, כמו רבעון רביעי חזק בדרך כלל.
  - נפח מסחר קיצוני מתרחש לעיתים בסמוך לאירועים כלכליים משמעותיים (כמו הכרזות ריבית או משברים גלובליים).
  - המחירים מגיבים בצורה חזקה יותר לשינויים מהירים בריבית בהשוואה לשינויים מתונים.

Stock Market Prediction in the U.S. - Analysis and Forecasting of S&P 500 Trends

Data Preparation Report

המכללה האקדמית
עמק יזרעאל
ע"ש מקס שטרן

מגישים – דין תחאוכו 318291705, שינה תחאוחו 213381569 מנחה

– ג'ניה גוטפריד

1. Selecting Data

מקור

השתמשנו בספריה של **Yahoo Finance** בפייתון לכריית נתונים, הספריה מספקת מידע היסטוריה על מניות הכוללת, תאריך, מחיר סגירה משוכלל, מחיר סגירה, מחיר הכי גבוה, מחיר הכי נמוך, מחיר פתיחה, ונפח מסחר. בנוסף הורדנו נתונים על ריבית ממאגר **FRED federal reserve economic data**.

בחירת מאפיינים

על סמך המטרה שלנו לחזות מחירי סגירה של המניה S&P 500, בחרנו את התכונות הבאות כרלוונטיות מה – Dataset שלנו:

- **תאריך:** הכרחי לניתוח סדרות זמן.
- **מחיר סגירה ומחיר סגירה משוכלל:** התכונות האלו מייצגות את המטרה שלנו, בייחוד מחיר סגירה משוכלל, מכיוון שהיא לוקחת בחשבון דיבדנדים והנפקת מניות חדשות.
- **מחיר גבוה, מחיר נמוך ומחיר פתיחה:** בדרך כלל תכונות אלו יהיו אינדקטורים למחיר הסגירה.
- **נפח:** מייצג את פעילות השוק ואת הנדילות של המנייה, תכונה פוטנציאלית להשפיע על תנועות המחיר של המניה.
- **ריבית:** יכול להשפיע על פעילות השוק וביצועי השוק.

לא וויתרנו על שום מאפיין מכיוון שכל מאפיין ב dataset רלוונטי לחיזוי המטרה שלנו.

2. Cleaning Data

נתונים חסרים

- היו חורים בנתונים של שיעורי הריבית שהורדנו. עשינו backward-fill לתאריכים בהם היו חסרים שיעורי ריבית, מכיוון ששיעורי ריבית לא קופצים מיום ליום.

שגיאות בנתונים

- ווידאנו שאכן יש מגמה כלשהי בעזרת גרפים של תאריך ומחיר סגירה.

עקביות בקוד

- המרנו את התאריך לפורמט של **YYYY-MM-DD**, הורדנו את ה Ticker של המניה מכיוון שאנחנו עושים על מניה אחת בלבד.

לא היה צורך לוותר על תכונה או מקרה משמעותי בנתונים, מלבד ערכים ספציפיים עם חוסר עקביות.

3. Constructing New Data

כדי לשפר את החיזוי של המודל שלנו, ייצרנו תכונות חדשות כמו:

- **טווח מחיר (גבוה-נמוך):** נגזרת המייצגת את התנודתיות היומית של המניה.
- **תשואה יומית:** מחושב ממחירי סגירה כשינוי באחוזים, חשוב לניתוח ומידול של סדרות זמן.
- **תנודה ממוצעת 20 יום, 50 יום:** לראות מגמות מחירי המניה לטווח הקצר והבינוני.
- **שינויים בשיעורי הריבית:** משקף את המומנטום הכלכלי.

4. Integrating Data

שילוב הנתונים על המניה של Yahoo Finance ונתונים על שיעורי הריבית של FRED נעשו עם שימוש בתאריך כמזהה ייחודי, תוך מיזוג אינדקטורים כלכליים עם נתוני המניה. התקבל מערך נתונים מקיף המשלב ביצועי שוק המניות והקשר כלכלי, מה שמבטיח התאמה כרונולוגית.

5. Formatting Data

עקב מספר אפשרויות גישות מודל (Linear Regression, Random Forest, Neural Netwroks) ביצענו את הדברים הבאים:

- נרמול נתונים (Min-Max scaling) להבטיח ערך שווה בין תכונות ואימון מודל יציב.
- פיצול הנתונים ל training set ו testing set.

6. Exploratory Data Analysis (EDA)

	Adj Close	Close	High	Low	Open	Volume	Interest Rate
count	3775.000000	3775.000000	3775.000000	3775.000000	3775.000000	3.775000e+03	3775.000000
mean	2741.750058	2741.750058	2755.786281	2725.619154	2741.238127	3.939924e+09	1.240177
std	1273.743103	1273.743103	1280.114582	1266.902882	1273.740396	9.486365e+08	1.720700
min	1022.580017	1022.580017	1032.949951	1010.909973	1027.650024	1.025000e+09	0.040000
25%	1729.335022	1729.335022	1731.654968	1717.159973	1724.889954	3.374750e+09	0.090000
50%	2438.209961	2438.209961	2449.120117	2428.199951	2436.500000	3.782490e+09	0.190000
75%	3828.680054	3828.680054	3859.574951	3801.350097	3834.905029	4.302215e+09	1.900000
max	6090.270020	6090.270020	6099.970215	6079.979980	6089.029785	1.061781e+10	5.330000

Adj Close

- **ממוצע (2652~):** מצביע על ממוצע מחיר סגירה משוכלל לאורך תקופת הזמן הנתונה, מציע את גובה מחיר המניה הממוצע.
- **סטיית תקן (1274~):** מדגיש את התנודתיות הגבוה במחיר יומי של המניה.
- **מינימום (1022~) ומקסימום (6090~):** מציג את הקצוות של השוק, טווח המחיר של המניה.

High and Low

- **ממוצע (נע בטווח של Adj Close):** נע בסביבת הטווח של מחיר סגירה משוכלל, טיפוסי בשוקים יציבים.
- **סטיית תקן (1267~ ו 1280~):** מצביע על תנודתיות יומית משמעותית, עקבית עם תנועות שוק אקטיביות.
- הקרבה בין גבוה לנמוך משקף יציבות בהתנהגות השוק בלי קפיצות לא מוסברות.

Open Prices

- דומה ל **High and Low**, מראה יציבות במחיר הפתיחה ביחס לתנודתיות יומית.
- מצביע על מחירי פתיחת שוק יציבות, עם תנודות מתואמות עם השיא והשפל היומי.

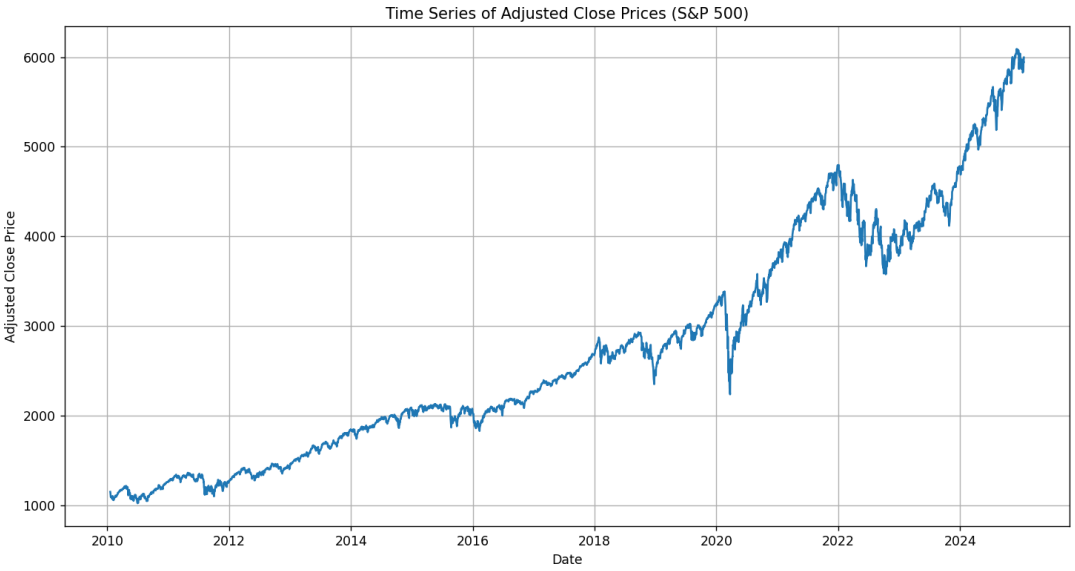
Volume

- **ממוצע (3.94 מליארד):** משקף את פעילות המסחר היומית הממוצעת, משקף את נזילות השוק.
- **סטיית תקן (948 מיליון):** מדגיש שונות משמעותית, מה שמצביע על תנודות ברמות פעילות המשקיעים, במיוחד סביב חדשות כלכליות גדולות או אירועים המשפיעים על השוק.
- **מינימום ומקסימום:** ממחיש קיצוניות של פעילות בשוק ונזילות, יכול להיות עקב אירועי שוק או חגים גדולים.

Interest Rate

- **ממוצע (1.24%):** שיעורי ריבית ממוצעים נמוך, ככל הנראה מעיד על סביבת הריבית הנמוכה הממושכת בעקבות המשבר ב-2008 ומגפת הקורונה.
- **סטיית תקן (1.72):** תקופות של שינוי מדיניות כלכלית וציפיות השוק.
- **טווח (0.04% - 5.33%):** מצביע על שינויים קיצוניים במדיניות שיעורי הריבית מנמוך מאוד לגבוה, מה שמשפיע על החלטות השקעה ודינמיות השוק.

הנתונים הסטטיסטיים האלה מדגישים את האופי המורכב של נתוני שוק המניות ומסייעים להבהיר את גישת המודלים, תוך דגש על הצורך בנרמול הנתונים ועיבוד נתונים אפשרית בעתיד כדי להבטיח מודל חיזוי יעיל.



מגמה:

- בכללי, מחיר סגירה משוכלל של מנית S&P 500 נוטה למגמת עליה לטווח הארוך, עקב צמיחת השוק וצמיחה כלכלית כללית.
- התנודתיות לטווח הקצר יותר מייצג תגובות שוק לאירועים כמו דוחות כלכליים, אירועים גיאופוליטיים, שינויי מדיניות וסנטימנט המשקיעים.

דפוסים תקופתיים:

למרות ששוק המניות בדרך כלל מגמות תקופתיות אחידות בהשוואה לקמעונאות, ניתן להבחין גם כאן בדפוסים חוזרים:

- תקופות מסוימות מציגות עלייה בנפח המסחר, כמו בתקופת הדוחות הכספיים (דוחות רבעוניות), הצהרות כלכליות, או אי ודאות גיאופוליטית (מלחמות).
- ניתן לזהות דפוסים, כמו תיקוני שוק מחזוריים.

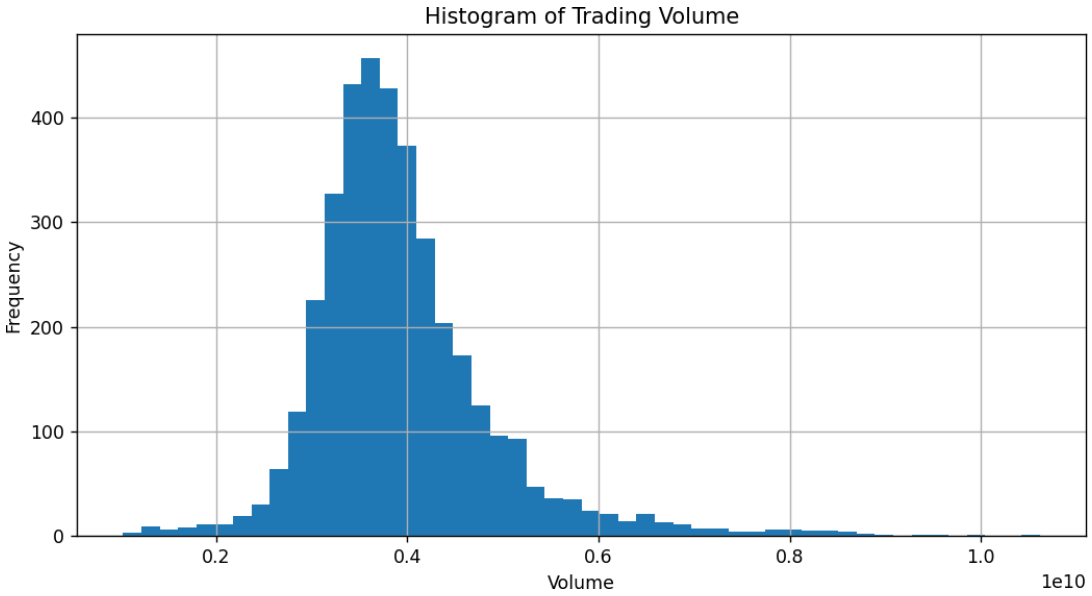
תנודתיות וחריגות:

קפיצות חדות או נפילות חדות מייצגות תקופות של תנודתיות גבוהה, בדרך כלל כתגובה לאירוע כלכלי משמעותי, לדוגמא, תקופת הקורונה, המשבר הכלכלי ב2008, יום שני השחור 1987 וכו'.

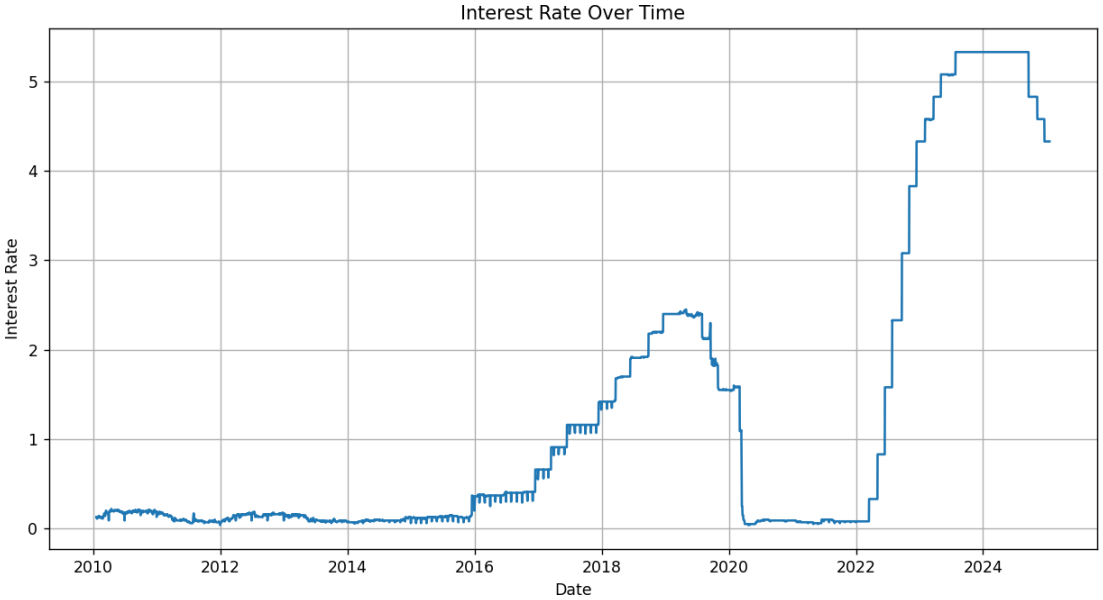
- זיהוי אירועים כאלה הכרחי להבנת הדינמיות של שוק המסחר ולשיפור דיוק החיזוי של המודל.

ניתוח מגמה:

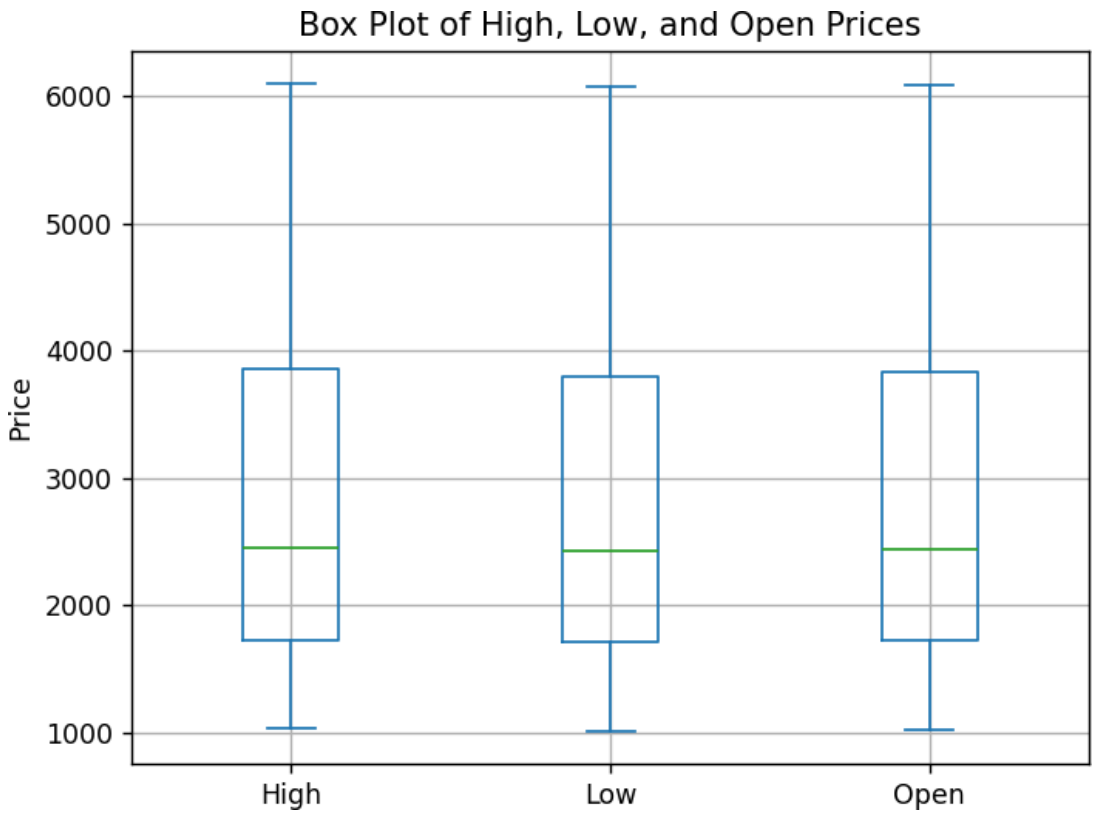
- מגמת העלייה הממושכת לטווח הרחוק מצביע את הצמיחה הכלכלית.
- ממוצע נע 20/50/100 יכולות להדגיש את המגמות הבסיסיות ולהפחית רעשים, מה שיסייע לחיזוי תנועות מחירים עתידיות.



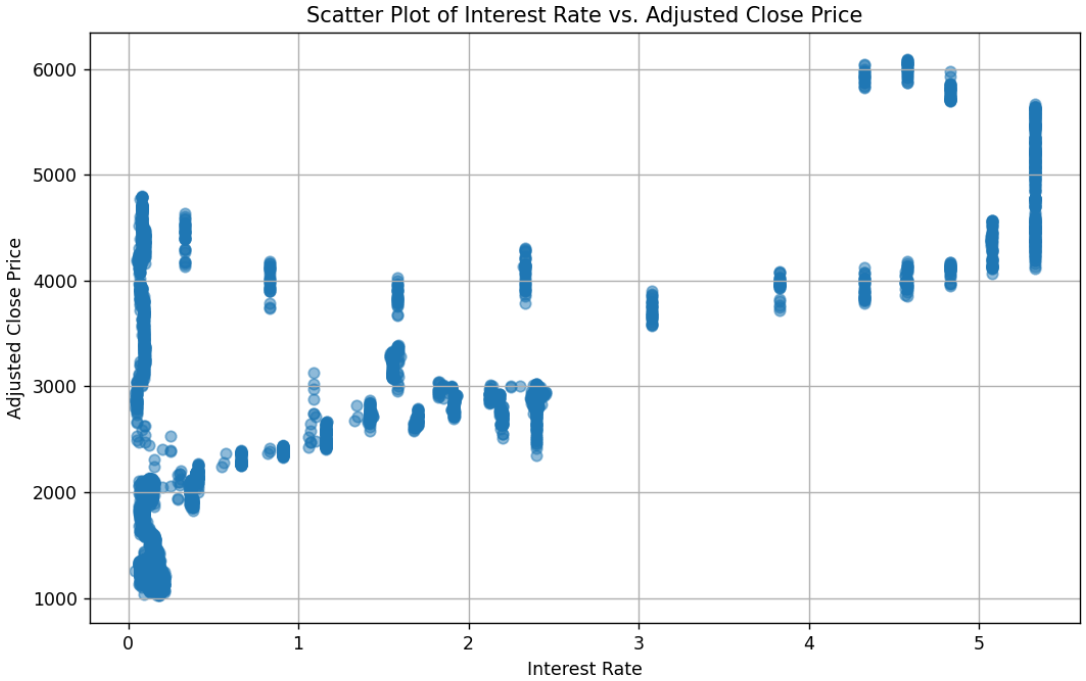
- **מטרה:** מציג את התפלגות התדירות של נפח המסחר היומי של S&P 500.
- **תובנות:**
  - ההיסטוגרמה ממחיש את ריכוז מחזורי המסחר, ומציינת את רמות הפעילות האופייניות של השוק.
  - רוב נפח המסחר מתרחש בטווח מסויים, מה שמדגיש את נדילות המניה.
  - המלבנים הקטנים יותר בטווח הנפחים הגדולים מרמזים על עליות מדדמנות, רוב הסיכויים אירועי שוק משמעותיים.



- **מטרה:** מתאר מגמות בשיעורי הריבית לאורך מסגרת הזמן הנתון.
- **תובנות:**
  - מדגיש תקופות של עלייה או ירידה בשיעורי הריבית, שעלולים להיות מתואמים עם מחזורים כלכליים.
  - שימושי להבנת ההקשר המאקרו-כלכלי המשפיע על מחירי המניות ופעילות השוק.
  - יכול להראות חפיפה עם אירועים קריטיים בשוק ההון או תקופות יציבות המשתקפות בתנודתיות במחירי המניה.



- **מטרה:** מסכם את ההתפלגות, החציון וחריגות במחירים הגבוהים, הנמוכים והפתיחה.
- **תובנות:**
  - השונות הכוללת של השוק במסחר יומי.
  - מדגיש חריגות המצביעים על אירועי שוק או ימי מסחר תנודתיים מאוד.
  - חציון וטווחים בין-רבעוניים עוזרים להבנת טווח המסחר האופייני של מחירי המניות.



- **מטרה:** חקירה של הקשר הישיר בין אינדיקטורים כלכליים כמו שיעורי ריבית ומחירי סגירה משכוללים של המניה.
- **תובנות:**
  - זיהוי אם יש תיאום ויזואלי בין שינויים בשיעורי הריבית לביצועי השוק.
  - תיאום יכול להצביע על כך ששיעורי הריבית כגורם משפיע על מחירי המניות.
  - פיזור מצביע על שונות במחירי המניות ביחס לשינויים בריבית.

הגרפים והתובנות מספקות קשר בסיסי לבחירת מודל חידוי מתאים, תוך הדגשת הרלוונטיות של התכונות והמאפיינים של הנתונים.

Stock Market Prediction in the U.S. - Analysis and Forecasting of S&P 500 Trends

Modelling Report

המכללה האקדמית עמק יזרעאל

ע"ש מקס שטרן

מגישים - שינה תחאוחו 213381569, דין תחאוחו 318291705

תאריך הגשה - 10/5/2025

מוגש למנחה - גניה גוטפריד

בחירת טכניקות מידול

לצורך חיזוי מחיר הסגירה של מניית S&P 500, נבחרו ארבעה מודלים מייצגים ממשפחות שונות - רגרסיה ליניארית, Random Forest, XGBoost ו־LightGBM.
הבחירה במודלים אלו נועדה להשוות בין גישות פשוטות (ליניארית) לבין גישות מתקדמות המבוססות על עצים ומנגנוני Boosting.
המודלים נבחרו על רקע הבנה שהתנהגות מניה יכולה להיות מושפעת הן מגורמים ישירים (כמו מחיר אתמול) והן מקשרים מורכבים (אינדיקטורים, מגמות)

כל מודל נבחן על היכולת שלו לנבא את מחירי הסגירה בצורה מדויקת, רציפה, ויציבה, תוך התאמה לסוג הנתונים – שהם במקרה זה: מניה בעלת תנודתיות נמוכה יחסית.

Linear Regression

רגרסיה ליניארית היא שיטה סטטיסטית לחיזוי ערך מספרי על בסיס משתנים אחרים.
המודל מניח שקיים קשר קווי (ליניארי) בין המשתנה התלוי (y) למשתנים הבלתי תלויים.(X)

מבנה מתמטי

המשוואה הכללית – $y^ = \beta0 + \beta1x1 + \beta2x2 +\cdots+ \beta nxn$

- Y^ - תחדית הערך
- X1,X2,...,Xn - הפיצרים (המשתנים הבלתי תלויים
- B0 – מקדם חיתוך עם ציר Y
- Bi - מקדמים שמתארים את השפעת כל פיצר

איך מתבצע האימון?

באימון, המודל מחפש את ערכי β שמצמצמים את פונקציית העלות -

$$Loss = \sum(y^i-yi)^2$$

זהו סכום ריבועי השגיאות - (MSE) מודל טוב הוא כזה שבו השגיאה הממוצעת קטנה.

רלוונטיות לחיזוי מניות -

- מתאימה במיוחד במקרים שבהם אין שינויים חדים במחיר -כלומר, במניות עם תנודתיות נמוכה.
- המודל יכול ללמוד מגמות כלליות של עלייה/ירידה.
- אידיאלי כשמחיר הסגירה מושפע חזק ממחיר האתמול – שזה נפוץ במדדים כמו S&P 500.

מה נותן המודל -

- תחדית יציבה למחר בהתבסס על אתמול
- פשוט להסבר ולפרשנות
- בסיס טוב להשוואה מול מודלים מתקדמים

Random Forest

Random Forest הוא מודל Ensemble כלומר הוא מאגד מספר רב של מודלים קטנים (במקרה הזה עצי החלטה) כדי ליצור חיזוי חזק ומדויק יותר.

איך זה עובד?

- בונה מספר רב של עצי החלטה שונים.
- כל עץ מאומן על מדגם אקראי שונה מהדאטה (Bootstrap Sampling)
- בכל עץ בוחרים תת-קבוצה אקראית של פיצ'רים.
- התחזית הסופית (ברגרסיה) היא ממוצע של כל תחזיות העצים

$$y^ = 1/T \sum y^$$

רכיבים עיקריים:

- n_estimators: כמה עצים יהיו ביער
- max_depth: העומק המקסימלי של כל עץ
- min_samples_split: המינימום של דגימות לפיצול נוסף

מאפיינים חשובים

- לא רגיש לרעש נקודתי
- לא מניח קשר ליניארי
- מתאים גם לנתונים מורכבים

רלוונטיות לחיזוי מניות:

- יכול לזהות קשרים לא ליניאריים בין משתנים – כמו השפעה מורכבת של נזילות, תנודתיות, או אינדיקטורים טכניים.
- מתאים למצבים שבהם התנהגות השוק לא אחידה (למשל: כשמחיר מושפע משילוב של כמה גורמים בלתי צפויים)
- לא מניח קשר ישר – מה שיכול לעזור כשלמניה יש קפיצות חדות

מה נותן המודל

- גמישות בלמידת דפוסים מורכבים
- יציבות מול רעש נקודתי (שכיח בשוק)
- אבל פחות טוב בזיהוי מגמות מתמשכות (trend), כי הוא לא "חוזה קדימה" – אלא מחלק קבוצות על בסיס ערכים קודמים.

XGBOost Regressor

מודל Boosting חכם שמבצע סדרה של תיקוני שגיאות – כל עץ חדש מנסה לשפר את התחזית של הקודמים. כל התחזית הסופית היא סכום של התחזיות הקודמות.

איך זה עובד?

1. מתחילים עם ניחוש ראשון פשוט (למשל ממוצע)
2. בונים עץ שמנבא את השאריות (ההבדל בין התחזית האמיתית לישנה)
3. מוסיפים את החידוי של העץ הזה לניחוש הקודם.
4. חוזרים שוב על התהליך עם העץ הבא.

פרמטרים חשובים:

- learning_rate: עד כמה כל עץ משפיע על המודל הסופי
- n_estimators: מספר העצים
- max_depth: עומק העצים
- early_stopping_rounds: עצירה מוקדמת אם אין שיפור

מה מיוחד בXGBoost-

- ביצועים גבוהים
- תמיכה ב Regularization (שליטה על מורכבות)
- מהיר מאוד

רלוונטיות לחיזוי מניות:

- מתאים לנתוני שוק עם התנהגות לא ליניארית ולא יציבה -כמו מניות עם תנודתיות גבוהה.
- יודע להתמודד עם תבניות מורכבות, כמו כשמחיר מושפע במקביל ממספר פקטורים
- מאפשר שליטה בפרמטרים כמו learning_rate כדי להימנע מ - overfitting-תכונה חשובה בשוק תנודתי.

מה נותן המודל:

- יכולת למידה עמוקה של דפוסים
- מתאים לתחזיות מבוססות אינדיקטורים
- אך דורש תכונות זמן כדי להבין מגמה – אחרת הוא חוזה ממוצע

LightGBM Regressor

LightGBM הוא Gradient Boosting בדיוק כמו XGBoost, אבל משתמש בטכניקות מתקדמות שגורמות לו להיות מהיר וקל יותר מבחינת משאבים.

איך זה עובד?

- בונה עצים לפי leaf-wise growth – כלומר קודם מפתח את העלה שבו הטעות הכי גדולה (ולא לפי עומק)
- מחלק את הערכים באמצעות Histogram-based decision - מתאים במיוחד לנתונים גדולים.

בכל שלב:

- נבחר פיצ'ר שיחלק את הדאטה בצורה הכי טובה
- הפיצול ייעשה בנקודת עלה (Leaf) עם השגיאה הכי גדולה
- הערכים בדאטה ממופים לבאקטים (Bins) כדי להאיץ את המיון

פרמטרים חשובים:

- n_estimators: מספר האיטרציות (עצים)
- max_depth: עומק מקסימלי של עץ
- learning_rate: *משקל של כל עץ*
- num_leaves: מספר העלים המקסימלי

יתרונות:

- מהיר מאוד
- מתאים לדאטה גדול
- תמיכה במקביליות(Parallel Learning)

רלוונטיות לחיזוי מניות:

- מצוין כשעובדים עם כמות גדולה של מניות או נתונים היסטוריים רבים -למשל תחזית יומית ל־1000 מניות.
- טוב לאופטימיזציה של תיקים - אפשר לאמן אותו מהר על הרבה מניות במקביל.
- מתאים לנתונים מורכבים או לא אחידים – אך, כמו XGBoost, מתקשה ללא תכונות זמן.

מה נותן המודל:

- ביצועים גבוהים במערכות מסחר בזמן אמת
- תומך בהפעלה מהירה על נתונים גדולים
- צריך הכנה נכונה של פיצ'רים מבוססי זמן כדי להיות אפקטיבי

Model Assumptions and Data Preparation

שלב ראשון

ייבאנו את הנתונים מקובץ CSV המכיל נתונים פיננסיים יומיים.
המראת עמודת התאריך לפורמט זמן נדרשת כי אנחנו עובדים עם סדרת זמן – בהמשך נשתמש ברצף הכרונולוגי כדי לחזות את העתיד על סמך העבר.
זה חיוני לכל מודל שמתבסס על סדר –ולא מודלים שמערבבים נתונים אקראית.

```
# Load Data

df = pd.read_csv('gpt.csv')
df['Date'] = pd.to_datetime(df['Date'])
```

שלב שני

יצרנו תכונות חדשות (פיצ'רים) שמייצגות את ערכי האתמול של כל משתנה.
למשל Close_lag, הוא מחיר הסגירה של אתמול.

למה זה חשוב?
המודלים לא יכולים להשתמש בנתונים של העתיד כדי לחזות את העתיד — הם חייבים לקבל רק מידע מהעבר.
זו התאמה חשובה לכל המודלים ובעיקר למודלים של סדרות זמן.
בלי זה, המודל היה "רואה" את המטרה לפני הזמן - דליפת מידע.

```
df['Close_lag'] = df['Close'].shift(1)
df['Open_lag'] = df['Open'].shift(1)
df['High_lag'] = df['High'].shift(1)
df['Low_lag'] = df['Low'].shift(1)
df['Volume_lag'] = df['Volume'].shift(1)
df['Vix_lag'] = df['Vix'].shift(1)
df['InterestRate_lag'] = df['Interest Rate'].shift(1)
df['GDP_lag'] = df['GDP'].shift(1)
df['target'] = df['Close'].shift(-1)
```

שלב שלישי

הגדרנו את מה שאנחנו רוצים לחזות – מחיר הסגירה של יום מחר.
עשינו זאת על ידי הדחה של עמודת Close שורה אחת למעלה ,((shift(-1) כך שהמודל ילמד את הקשר בין היום לבין מחר.

שלב רביעי

הסרנו את כל העמודות שמכילות את הערכים של היום (t), והשארנו רק את ערכי האתמול.(t-1)

למה זה חשוב?
אם נשאיר את הנתונים של היום בזמן שאנחנו מנסים לחזות את מחר — המודלים ילמדו "לרמות."
התוצאה תהיה תחזית מדויקת מדי אבל לא אמינה -זה נקרא דליפת מידע.
המודלים חייבים ללמוד רק מתוך מידע שזמין ברגע האמיתי כלומר, רק מה שהיה עד אתמול.

```
df = df.drop(['Close', 'Open', 'High', 'Low', 'Volume', 'Vix', 'Interest Rate', 'GDP'], axis = 1)
```

שלב חמישי

יצרנו את מערך הקלטים (X) ואת ערך המטרה (Y)
זהו הפורמט שמודלי למידת מכונה צריכיX –

- X תכונות שנכנסות למודל
- Y הערך שהוא מנסה לחזות

```
features = ['Close_lag', 'Open_lag', 'High_lag', 'Low_lag', 'Volume_lag','Vix_lag', 'InterestRate_lag', 'GDP_lag']
target = 'target'
X = df[features]
y = df[target]
```

סיכום

על מנת להכין את הנתונים לבניית מודלים, בוצעו מספר שלבים בקוד לצורך שמירה על מבנה סדרת הזמן ומניעת דליפת מידע.

תחילה נטענו הנתונים מקובץ CSV והומרה עמודת התאריך לפורמט מתאים. לאחר מכן נוצרו תכונות מסוג - lagערכים של היום הקודם עבור כל משתנה - מתוך מטרה להשתמש רק במידע היסטורי לצורך התחזית.

עמודת המטרה הוגדרה כערך מחיר הסגירה של היום הבא (target = Close.shift(-1)
על מנת למנוע מצב בו המודל יראה מידע עתידי, הוסרו העמודות המקוריות של היום (Close, Open וכו'), ונשמרו רק ערכי האתמול.
שורות עם ערכים חסרים הוסרו, והנתונים חולקו ל־Train/Testלפי סדר כרונולוגי (80/20), מתוך שמירה על המשכיות הזמן.

שלב ההכנה בוצע באופן זהה עבור כל ארבעת המודלים, תוך התאמה לדרישות של כל אחד מהם

Test Design

לאחר עיבוד הנתונים והגדרת משתני הקלט (X) ועמודת המטרה (Y), בוצעה חלוקה של הנתונים לשני סטים

- סט אימון (Train) – 80% הראשונים של הנתונים – משמש לאימון המודלים \
- סט בדיקה (Test) - 20% האחרונים של הנתונים - משמש לבדיקה בלבד

החלוקה התבצעה לפי סדר כרונולוגי באמצעות החיתוך הבא בקוד

```
split_index = int(len(X) * 0.8)

X_train, X_test = X[:split_index], X[split_index:]
y_train, y_test = y[:split_index], y[split_index:]
```

לא נעשה שימוש בפונקציית train_test_split() משום שמדובר בבעיה של סדרת זמן, בה השמירה על רצף כרונולוגי היא חיונית.
ערבוב של שורות עלול להוביל לדליפת מידע עתידי ולפגוע באמינות התחזית.

מדדי ההערכה שנבחרו

למדידת הביצועים של כל מדד השתמשנו בשני מדדים מרכזיים

1. **MSE – Mean Squared Error**
מודד את ממוצע ריבועי השגיאות בין התחזית לבין הערך האמיתי

MSE = 1/n ∑ (y^i−yi)^2

   - ככל שהערך נמוך יותר – התחזית מדויק יותר
יחידות המדד הן בריבוע של יחידות המטרה

2. **R^2 – מקדם ההסברה**
מדד סטטיסטי שמראה איזה אחוז מהשונות בנתונים מוסברת על ידי המודל

   - ערך של 1 = חיזוי מושלם
   - ערך של 0 - המודל לא חזה טוב יותר מהממוצע

למה בחרנו דווקא בהם?

- MSEנפוץ מאוד במודלים של רגרסיה – נותן אינדיקציה ישירה לגודל השגיאה
- R^2קל לפרש – הוא אומר "עד כמה המודל הסביר את הדאטה"
- השילוב ביניהם נותן תמונה גם על גודל השגיאה וגם על איכות ההתאמה הכללית

הרצת איטרציות

כל מודל הורץ שלוש פעמים (שלוש איטרציות) כאשר בכל הרצה כוונו פרמטרים שונים לצורך טיוב ביצועים.
לדוגמה -

- Random Forest: שינוי n_estimators ו־max_depth
- XGBoost־ LightGBM שינוי גם של learning_rate

בכל הרצה שמרנו את התוצאות (mse, r2) והשווינו אותן בהמשך. תהליך זה מאפשר לזהות האם שינוי פרמטרים שיפר את התחזית או לא.

Model Description

בשלב זה נבנו ארבעה מודלים מסוג רגרסיה. כל אחד מהם רץ בשלוש איטרציות שונות, עם טיוב פרמטרים.
להלן תיאור כל מודל, הפרמטרים שנבחרו, והאופן בו הם הופעלו בקוד -

איטרציה ראשונה

```
# Models at iteration 1

linearRegression = LinearRegression()
randomForest = RandomForestRegressor(random_state = 42)
xGBoost = xgb.XGBRegressor(objective = 'reg:squarederror', random_state = 42)
lightGBM = lgb.LGBMRegressor(random_state = 42)
```

מודל	איטרציה	2^R	MSE	הסבר
Linear Regression	1 - ברירת מחדל	0.9889	4831.50	המודל התאים קו ליניארי מדויק לנתונים. הצליח לזהות את הקשר הישיר בין ערכי האתמול למחיר של מחר. התחזית עקבה כמעט בדיוק אחר המחיר האמיתי.
Random Forest	1 - ברירת מחדל	0.4715	230690.89	המודל הריץ מספר עצים בעומק ברירת מחדל. חזה תחזיות שטוחות, לא הצליח לקלוט מגמה. ביצועים נמוכים יחסית.
XGBoost	1 - ברירת מחדל	0.40890	257988.60	המודל רץ עם עומק קטן ומספר עצים מועט. לא טוייב. לא הצליח ללמוד תבניות חזקות. תחזית ממוצעת וחלשה.
LIghtGBM	1 - ברירת מחדל	0.4690	231775.11	ביצועים מעט טובים אך נמוכים, תחזית קבועה יחסית, מגיב מהר אך לא מדויק

איטרציה שנייה

```
# Models at iteration 2

linearRegression = LinearRegression()
randomForest = RandomForestRegressor(random_state = 42, n_estimators = 200, max_depth = 10)
xGBoost = xgb.XGBRegressor(objective = 'reg:squarederror', random_state = 42, n_estimators = 200, max_depth = 4, learning_rate = 0.1)
lightGBM = lgb.LGBMRegressor(random_state = 42, n_estimators = 200, max_depth = 5, learning_rate = 0.1)
```

מודל	איטרציה	2^R	MSE	הסבר
Linear Regression	2 - ברירת מחדל	0.9889	4831.50	לא שונה מהאיטרציה הראשונה. המודל עדיין מספק תחזית מדויקת מאוד ועוקב אחרי מגמת הנתונים בצורה מצוינת.
Random Forest	2	0.4671	232,594.74	הוגדל מספר העצים והוגבל עומק כדי לנסות למנוע תחזיות שטוחות. התוצאה דומה מאוד לאיטרציה 1 – לא נרשמה כמעט כל השפעה.
XGBoost	2	0	252,243.09	נוספו עצים, עומק רדוד יחסית, ולמידה זהירה יותר. חל שיפור קטן ב־ R^2, אך עדיין תחזית ממוצעת ושטוחה.
LIghtGBM	2	0.4609	235,299.69	נעשה טיוב זהה כמעט ל־ XGBoost. התוצאה נשארה דומה לאיטרציה 1 – מעט פחות טובה. המודל מהיר, אך לא מדויק.

איטרציה שלישית

```
Models at iteration 3

nearRegression = LinearRegression()
ndomForest = RandomForestRegressor(random_state = 42, n_estimators = 300, max_depth = 20, min_samples_split = 5)
Boost = xgb.XGBRegressor(objective = 'reg:squarederror', random_state = 42, n_estimators = 300, max_depth = 6, learning_rate = 0.05, subsample = 0.8
ghtGBM = lgb.LGBMRegressor(random_state = 42, n_estimators = 300, max_depth = 8, learning_rate = 0.05, num_leaves = 31, subsample = 0.8)
```

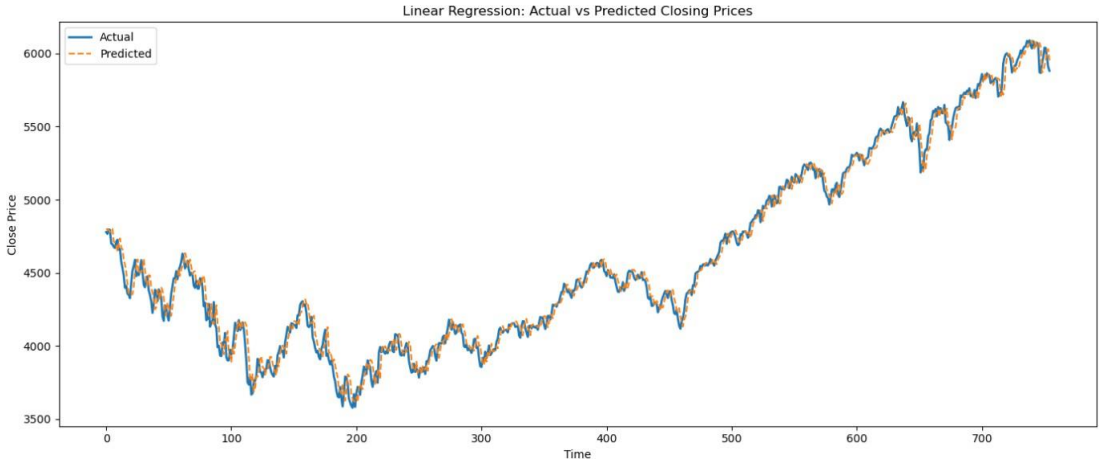
מודל	איטרציה	2^R	MSE	הסבר
Linear Regression	3 ברירת מחדל	0.9889	4831.50	כמו בשתי האיטרציות הקודמות – תחזית מדויקת, עקבית לחלוטין, מציינת התאמה חזקה של הדאטה למודל ליניארי.
Random Forest	3	0.4673	232,514.36	תוצאה כמעט זהה לאיטרציה 2. מראה שהמודל יציב אך לא מדויק – ממשיך לחזות ערכים שטוחים.
XGBoost	3	0.4357	246,291.36	שיפור קל נוסף לעומת R^2 ב־ איטרציה 2, אך עדיין תחזית כללית ולא מדויקת מספיק. מתחיל להתייצב, אך רחוק מהמודל הליניארי.
LIghtGBM	3	0.4678	232,267.86	תוצאה כמעט זהה לאיטרציה 2. מראה יציבות, אך דיוק לא משתפר. לא מצליח לנצל את הפיצ'רים כדי להבין מגמה.

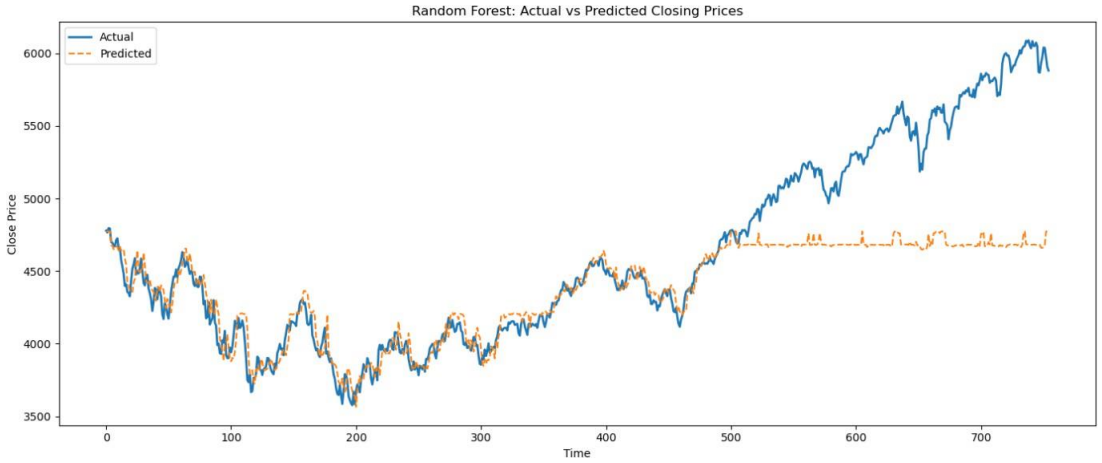
מכיוון שאנו רוצים לחזות מחיר סגירה עתידי, ולקחנו את הפיצ'רים כנתונים מהיום הקודם, חילוק הנתונים שלנו לקבוצת למידה וקבוצה בדיקה, לא נעשתה באופן רנדומלי. תחילה מיינו את הנתונים לפי תאריך, ולאחר מכן פיצלנו את הנתונים ל80% קבוצת לימוד, 20% קבוצת בדיקה. כמו כן, בגרפים שנציג בהמשך, ההתחלה של הגרף היא ה80% והסוף היא 20% כלומר הסוף של הגרף היא החיזוי על הקבוצת בדיקה.

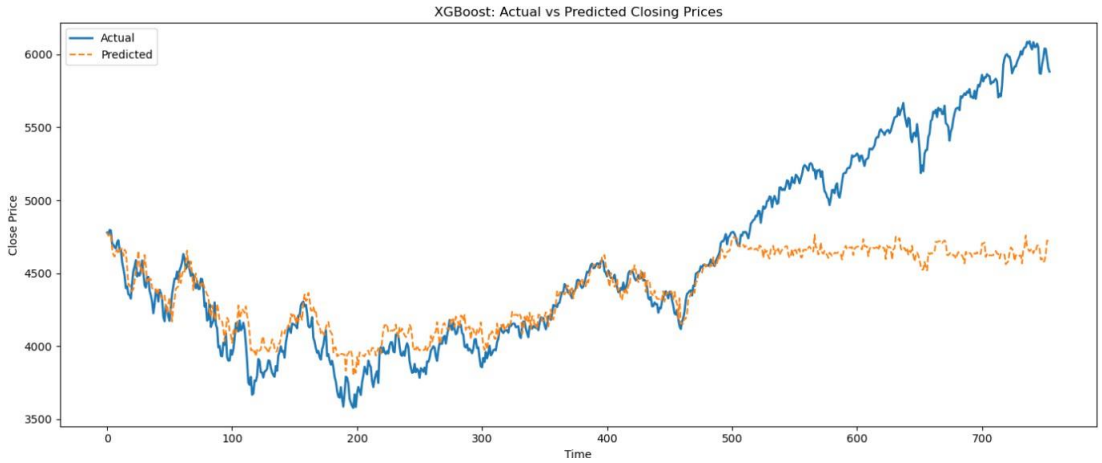
לאחר הרצת המודלים ב-4 איטרציות סך הכל (ראשונה עם פרמטרים כברירת מחדל), ושרטטנו גרף של כל מודל המציג את החיזוי של המודל לצד המחיר האמיתי. אנו מבחינים שבכל איטרציה, הרגרסיה לינארית היא המודל שהכי דייק בחיזוי, המודל הצליח לקלוט מגמות של המניה, ולחזות את מחיר הסגירה העתידי בצורה מרשימה. לעומת זאת, מודלי העץ לא כל כך דייקו בחיזוי מהתחלה, ובחיזויים על קבוצת הבדיקה, המודל משתטח, כלומר לא מצליח לקלוט מגמות.

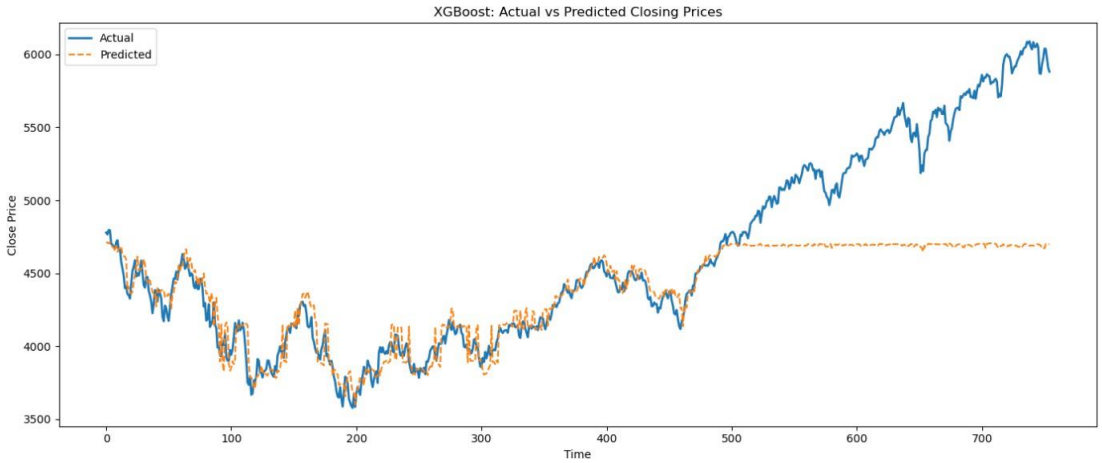
התוצאות בהחלט הגיונית מכיוון שבדרך כלל, מחיר הסגירה העתידי (של מחר) מתואם עם מחיר הסגירה של היום, כלומר אין שינויים קיצוניים בין מחירי סגירה בטווח של יום. לכן הרגרסיה לינארית מצליחה לתפוס את המגמה של המניה בצורה מרשימה, היא בעצם מניחה, אם מחיר הסגירה היום היא 4500 אז מחיר הסגירה מחר תהיה $4500 + \varepsilon$. רגרסיה לינארית יודעת לעקוב אחרי מגמה, במיוחד לטווח הקצר.

לעומת זאת, מודלי עצים דורשים אינדיקציה של זמן, אך לעומת שהוספנו פיצ'רים מהונדסים, התוצאות לא היו מעודדות. בנוסף, הם מתאימים את עצמם יותר מדי לקבוצת הלמידה, כך שאם מחירי הסגירה בקבוצת הבדיקה מחוץ לטווח של קבוצות הלמידה, המודלים ישתטחו בדרך כלל בחיזוי על קבוצת הבדיקה.






XGBoost: Actual vs Predicted Closing Prices
Actual
Predicted
Close Price
Time
6000
5500
5000
4500
4000
3500
0
100
200
300
400
500
600
700


XGBoost: Actual vs Predicted Closing Prices
Actual
Predicted
Close Price
Time
6000
5500
5000
4500
4000
3500
0
100
200
300
400
500
600
700

לסיכום, לאחר הרצת המודלים מספר פעמים עם פרמטרים שונים, הגענו למסקנה שהמודל הטוב ביותר לחיזוי מחירי מניה עתידיים לטווח הקצר היא רגרסיה לינארית, גם במדדי ההערכה וגם לפי הגרפים שצירפנו. מודל הרגרסיה הלינארית הצליחה לחזות בצורה מרשימה מחירי סגירה עתידיים, עם סטיות לא גדולות. שאר המודלים כפי שציינו השתטחו בקבוצת הבדיקה, למרות ניסיונות שיפור. לכן החלטנו להמשיך את הפרוייקט עם רגרסיה לינארית.

Evaluation Report

Stock Market Prediction in the U.S. - Analysis and Forecasting of S&P 500 Trends

המכללה האקדמית עמק יזרעאל

ע"ש מקס שטרן

מגישים – שינה תחאוחו 213381569, דין תחאוכו 318291705

מוגש ל – גניה גוטפריד

תאריך הגשה – 30/5/2025

מטרה עסקית

מטרת הפרויקט היא לחזות את ערך הסגירה של מדד S&P 500 לטווח של עד שבוע (7 ימים) קדימה ,בהתבסס על נתונים היסטוריים וכלכליים הכוללים: מחירי עבר, נפחי מסחר, ריבית, תוצר מקומי ,(GDP) ומדד התנודתיות.(VIX)

לצורך כך, נבנים 7מודלים שונים ,שכל אחד מהם מיועד לחיזוי של יום אחר בעתיד:

- מודל 1 לחיזוי יום אחד קדימה
- מודל 2 לחיזוי יומיים קדימה
- וכן הלאה עד יום 7

גישה זו מאפשרת תחזיות מפורטות לטווח קצר, תומכת בקבלת החלטות פיננסיות לטווח של שבוע, ומתאימה לאופי השוק בו השינויים היומיים יחסית מדודים אך משמעותיים.

בהירות הצגת התוצאות

לכל אחד מהמודלים בוצעו שלוש הרצות (איטרציות) תוך שמירה על עקביות במבנה הנתונים, והצגת התוצאות ב־ MSEו־R^2.
נכון לשלב זה, נבחן המודל לחיזוי יום 1 קדימה (target_1d) בלבד, בשלבים מתקדמים יורחבו המודלים ל־7 ימים.

ממצאים ייחודיים

מודל Linear Regression לחיזוי יום אחד קדימה הפיק תוצאה גבוהה ויציבה:

- $R^2 = 0.9889$, MSE = 4831.50 בכל ההרצות
- התוצאה עקבית, ברורה, ומבוססת על קשרים ליניאריים חזקים בין פיצ'רים ליעד

למרות שמדדים כאלה עלולים להיראות חריגים, במקרה זה הם סבירים: התחזית היא לטווח קצר מאוד (יום אחד), ולכן השוק צפוי יחסית.

התאמה למטרות העסקיות

המודל לחיזוי יום אחד קדימה מספק תוצאה איכותית ביותר - ברמת דיוק שמתאימה לחיזוי עסקי.
בעתיד, כל אחד מהמודלים לחיזוי ימים נוספים (2-7) ייבחן בנפרד, אך צפוי ש־רמת הדיוק תלך ותרד ככל שמתרחקים מהיום הנוכחי -בשל עלייה באי-הוודאות וחולשת הקשרים הליניאריים.

דירוג מודלים לפי התאמה

דירוג	מודל	2^R	MSE	הערות
1	Linear Regression	0.9889	4831.50	מדויק, קל לפרשנות
2	LightGBM	0.467	232,000	יציב יחסית
3	Random forest	0.467	232,000	מעט פחות יציב
4	XGBboost	0.42	250,000	הביצועים הנמוכים ביותר

שאלות עסקיות חדשות שעלו

- כיצד לשפר את רמת הדיוק ככל שמתרחקים מהיום הנוכחי?
- האם נוכל לזהות מגמות במקום תחזיות נקודתיות?
- האם נוכל לשלב סנטימנט חדשותי או מידע חיצוני בתור פיצ'רים?

סקירת תהליך העבודה

הצלחות

- בנייה נכונה של פיצרים מושהים
- אימון מודלים מרובים והשוואה עקבית
- ניתוח תוצאה גבוה שהובילה להבנה של דליפת מידע ותיקון

אתגרים

- תוצאה גבוהה מדי בתחילה (2^R = 0.9999) העידה על דליפת מידע - טופלה בהסרת Close, Adj Close
- ביצועים חלשים למודלים מורכבים לעומת פשוטים - מצריך חשיבה על התאמת מודל לבעיה

תובנות

- ככל שמתרחקים מיום התחזית - הקשרים הסטטיסטיים נחלשים
- חשוב למדוד כל מודל באופן עצמאי, ולא להסתמך על תחזיות קודמות
- מודל פשוט יכול להיות מדויק יותר ממודלים מורכבים כשיש מעט רעש בנתונים

Deployment Report

המכללה האקדמית
עמק יזרעאל

תכנית פריסה(Deployment Plan)

בפרויקט זה נפרס מודל לחיזוי מחיר הסגירה של מדד **S&P 500** המודל משולב באתר דשבורד ייעודי: המשתמש מזין תאריך, והמערכת מחזירה **חיזוי מספרי** ומציגה **גרף השוואה** בין הערכים בפועל לבין התחזית עבור חלון הזמן הרלוונטי.

ארכיטקטורת פריסה -

- **Frontend (Dashboard)** -

דף אינטרנט (HTML/CSS/JS) עם שדה להזנת תאריך, כפתור "חשב תחזית", אזור לתצוגת התוצאה המספרית, ורכיב גרפי המציג סדרות "Actual" ו-"Predicted".

- **Backend -**

שירות יישומי (python/ flask) שמטעין את המודל המאומן, מקבל בקשת חיזוי עם תאריך, מכין את הפיצ'רים לתאריך היעד, מפיק תחזית ומחזיר גם נתוני סדרה לציור הגרף.

- **מקורות נתונים –**

נתוני עבר של S&P 500 מועשרים במדדים מאקרו-כלכליים (כגון VIX, ריבית, GDP).הנתונים מתעדכנים מעת לעת לתחזיות עדכניות.

ממשק המשתמש -

1. המשתמש מזין תאריך 2.
מקבל מחיר סגירה חזוי
3. מוצג גרף של Actual vs predicted

- **הרחבות עתידיות :**הוספת דשבורד אינטראקטיבי (פילטרים, טווחי תאריכים), ודוחות תובנות (טבלאות ביצועים, פירוק תרומת פיצ'רים)

תכנון ניטור ותחזוקה

כדי להבטיח שהמודל ישמור על אמינות ורלוונטיות לאורך זמן, נקבעו מספר מנגנונים לניטור ותחזוקה. ראשית, יש לעקוב באופן רציף אחרי מחירי הסגירה בפועל של מדד S&P 500 ולהשוות אותם מול התחזיות שהמודל מפיק. בנוסף, יש לעקוב אחרי נתוני ריבית ונתוני תוצר מקומי גולמי (GDP), מאחר שאלו גורמים מאקרו־כלכליים מרכזיים שיכולים להשפיע על תוצאות המודל.

הערכת הביצועים של המודל תתבצע באמצעות מדדי דיוק מקובלים -

- R^2 - מקדם התאמה
- MSE - שגיאת ריבוע ממוצעת
- RMSE

מדדים אלה יחושבו באופן קבוע, והם ישמשו לבחינת יציבות ואמינות התחזיות לאורך זמן. בנוסף, תיבדק התאמה חזותית באמצעות גרפים המשווים בין ערכים נצפים לבין תחזיות המודל.

נקבעו קריטריונים ברורים לזיהוי מצב שבו המודל "פג תוקף". בין היתר, אם ערך ה־ R^2 ירד באופן עקבי מתחת ל־0.4, או אם ערכי ה־ MAEאו RMSE יעלו מעבר לספים שהוגדרו מראש, תישקל החלפה או עדכון של המודל. גם הופעת דפוסי נתונים חדשים או אירועים כלכליים חריגים שאינם מתואמים עם נתוני האימון עלולים להוביל לצורך בעדכון.

במסגרת התחזוקה השוטפת, המודל יאומן מחדש בפרקי זמן קבועים עם נתונים עדכניים, על מנת לשמור על רמת דיוק גבוהה. בנוסף, יישמר תיעוד של כל גרסה והקוד ינוהל באמצעות מערכת בקרת גרסאות (GitHub), כדי לאפשר שקיפות ומעקב אחרי השינויים. במקרה שבו יתגלה שמודל חדש מציג תוצאות ירודות בהשוואה לקודמו, תתאפשר חזרה מהירה לגרסה הקודמת.