4. Modelling

מטרת שלב זה:

לבחור את המודלים בהם תשתמשו, להכיר אותם, לתכנן את הניסוי, לבנות פייפליין מסודר בקוד להרצת המודלים תוך קריאות אל מסד הנתונים של הפרויקט, בנוסף ביצוע הרצות ראשוניות של המודל/מודלים לקבלת תוצאות.

בדוח זה עליכם לפרט על מה שביצעתם (מה שלא רלוונטי ניתן לא לכלול):

- **בחירת שיטות מודל רלוונטיות -**
- Random Forest: מודל המבוסס על "יער" של עצי החלטה הפועלים במקביל לחיזוי תוצאה משותפת. בחרנו בו בשל יכולתו להתמודד עם נתונים פיננסיים תנודתיים מבלי לבצע "התאמת יתר" (Overfitting) ובזכות היציבות שלו בזיהוי קשרים מורכבים בין משתני מאקרו למחיר המניה.

- XGBoost: אלגוריתם מתקדם בשיטת ה-Boosting, הלומד מטעויות קודמות ומשפר את עצמו בכל איטרציה. בחרנו בו כיוון שהוא נחשב למודל החזק ביותר כיום לחיזוי נתונים טבלאיים, המאפשר להגיע לדיוק גבוה מאוד תוך ניצול יעיל של כוח עיבוד.

- Logistic Regression: מודל סטטיסטי קלאסי המשמש כנקודת ייחוס (Baseline) לבדיקת קשרים ליניאריים פשוטים. השתמשנו בו כדי להוכיח כי השוק מתנהג בצורה מורכבת שאינה ניתנת לחיזוי באמצעות קו ישר, מה שחיזק את הצורך במודלים המורכבים יותר שהרצנו.

- **תכנון ניסוי – (Test Design) -**

בשלב הראשון, חילקנו את הנתונים לפי שיטת החלוקה Train/Validation/Test.

בשלב השני, הגדרנו קריטריונים להצלחת המודל לפי המדדים הבאים המתבססים על מטריצת הבלבול:

- **Accuracy (דיוק כללי):** זהו המדד הראשוני ליציבות המודל. יעד ההצלחה הוגדר כ-60%, והמודל הנבחר (Random Forest) עקף אותו משמעותית עם 86% דיוק.

- **Precision (דיוק בחיזוי):** זהו המדד החשוב ביותר למשקיע כדי למנוע כניסה לעסקאות מפסידות (False Positive). המודל השיג 0.90 (90%) בקטגוריה זו, מה שמעיד על רמת אמינות גבוהה מאוד של סיגנל הקנייה.

- **Recall (זיהוי הזדמנויות):** המודל השיג 84% מה שמראה שהוא יודע "לתפוס" את רוב הזדמנויות הרווח בשוק ולא להישאר בחוץ בימים ירוקים (יום בו מחיר המניה נסגר במחיר גבוהה יותר ממחיר הפתיחה שלו או ממחיר הסגירה של יום הקודם).
- **F1-Score (המדד המאזן):** מדד זה (שעמד על 0.87) מוכיח שהמודל מאוזן - הוא לא רק "מנחש" עליות כל הזמן, אלא מבצע הפרדה איכותית בין ימי רווח לימי הפסד.

חלוקת נתונים - השתמשנו בשיטת Train/Validation/Test ולשם כך בחרנו לחלק את הנתונים לפי סדר כרונולוגי ולא באקראי מהסיבה שבשוק ההון לא ניתן לערבב זמנים, המודל חייב ללמוד מהעבר כדי לחזות את העתיד.

- לקחנו את 70% התאריכים הראשונים ביותר (4,155 שורות) כדי שהמודל יתאמן ויבין איך השוק מתנהג.
- המשכנו עם 15% נוספים (891 שורות) כקבוצת בדיקת ביניים, שדרכה אנחנו מוודאים שהמודל לא סתם מנחש ומבצעים התאמות.
- בסוף, שמרנו את ה-15% הכי עדכניים (891 שורות) למבחן הסופי והאמיתי - לראות איך המודל מתמודד עם נתונים חדשים לגמרי שהוא מעולם לא ראה.

החלוקה הזו היא קריטית כי היא מונעת זליגת מידע מהעתיד ומדמה מצב של מסחר אמיתי בזמן אמת.

- במידת הצורך יש להפעיל אלגוריתם **לבחירת מאפיינים (Feature selection)** – הפעלנו אלגוריתם בקוד להצגת ה5 מאפיינים שהכי השפיעו על התוצאות של המודל. הוא מדרג את המשתנים לפי תרומתם לדיוק הפיצולים בעצי ההחלטה, ככל שמשתנה מסוים עזר להפריד טוב יותר בין ימי עליות לימי ירידות, כך דירוגו במודל עלה.

<u>5 המאפיינים המובילים במודל Random Forest:</u>

```
daily_return_high    0.340273
daily_return_low     0.316228
VIX                  0.057080
volume               0.053226
SP500                0.046468
dtype: float64
```

<u>5 המאפיינים המובילים במודל Random Forest:</u>

```
daily_return_low       0.467775
daily_return_high      0.196680
adj_close              0.047603
CONSUMER_CONFIDENCE    0.041343
SP500                  0.039188
dtype: float32
<Figure size 800x600 with 0 Axes>
```

תיעוד הנחות והגדרות -

- הנחות לגבי נתונים:

1. הנחנו שהעבר משפיע על העתיד ולכן כשחילקנו את הנתונים ל-Train ו-Test, השתמשנו ב-shuffle=False. הנחנו שסדר הימים הוא קריטי.
2. הנחנו ששורות עם ערכים חסרים הן "רעש" שעלול להטעות את המודל ולכן הסרנו אותן.
3. הנחנו שהדבר הכי חשוב למשקיע הוא העלייה/ירידה של המניה ולא המחיר המדויק באגורות, ולכן הפכנו את הבעיה לחיזוי של 0 או 1.

- התאמות המודל:

1. המרנו את סוגי הנתונים שהגיעו מטבלאות הSQL כ"טקסט" או אובייקט. ביצענו התאמה והמרנו אותם למספרים כי המודל שלנו יודע לבצע חישובים רק על מספרים.
2. טיפלנו בערכים חסרים על ידי ניקוי הטבלאות משורות ריקות כדי למנוע שגיאות אימון.
3. יצרנו עמודה חדשה של 0 ו1 לצורך דיוק המודל והחלטה עסקית.
4. סידרנו את כל ה-DataFrame לפי תאריך לפני החלוקה, כדי להבטיח שהמודל לומד על תקופה מוקדמת ובודק את עצמו על תקופה מאוחרת יותר.

- **הציגו תוצאות ראשוניות של הרצת מודל**

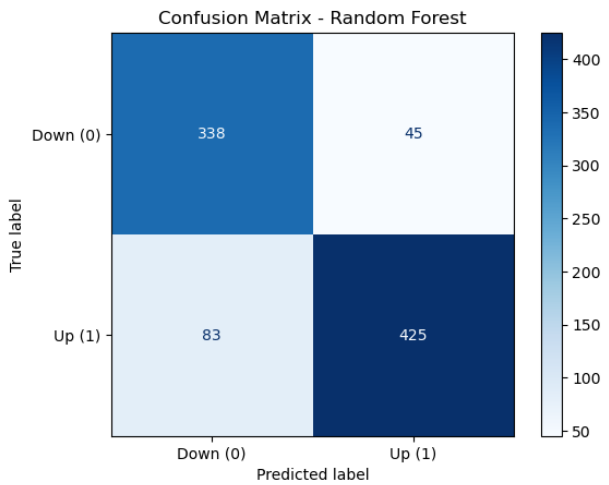
Random Forest

```
Classification Report - Random Forest:
              precision    recall  f1-score   support

           0       0.80      0.88      0.84       383
           1       0.90      0.84      0.87       508

    accuracy                           0.86       891
   macro avg       0.85      0.86      0.85       891
weighted avg       0.86      0.86      0.86       891
```

<u>Confusion Matrix</u>



הפרמטרים שהשתמשנו במודל:

```
--- Features Importance: Random Forest ---
daily_return_high       0.340273
daily_return_low        0.316228
VIX                     0.057080
volume                  0.053226
SP500                   0.046468
open_price              0.044214
adj_close               0.043690
US10Y                   0.041024
CONSUMER_CONFIDENCE     0.022106
INFLATION_CPI           0.019378
FED_RATE                0.016313
dtype: float64
```

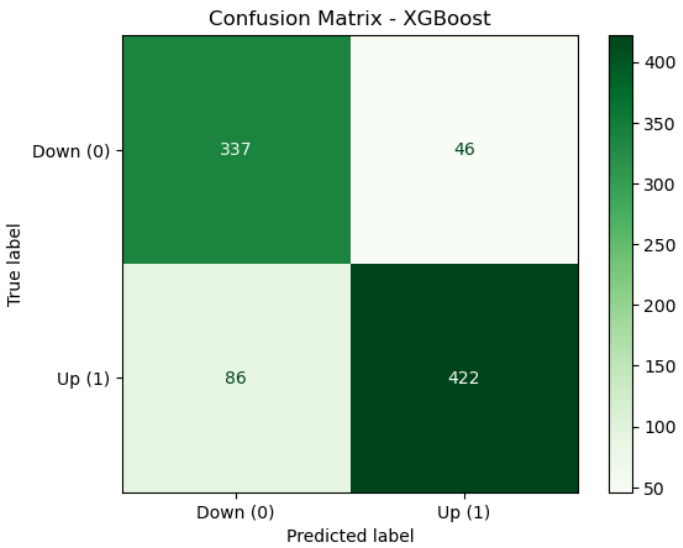
XGBoost

```
Classification Report - XGBoost:
              precision    recall  f1-score   support

           0       0.80      0.88      0.84       383
           1       0.90      0.83      0.86       508

    accuracy                           0.85       891
   macro avg       0.85      0.86      0.85       891
weighted avg       0.86      0.85      0.85       891
```

Confusion Matrix



הפרמטרים שהשתמשנו במודל:

```
--- Features Importance: XGBoost ---
daily_return_low        0.467775
daily_return_high       0.196680
adj_close               0.047603
CONSUMER_CONFIDENCE     0.041343
SP500                   0.039188
VIX                     0.036697
FED_RATE                0.036138
open_price              0.034511
INFLATION_CPI           0.034249
US10Y                   0.033923
volume                  0.031893
dtype: float32
```

מודל ה Random Forest - נבחר כמודל המועדף מאחר שהשיג דיוק כללי מעט גבוה יותר (כ-86%) בהשוואה למודל XGBoost שהגיע לכ-85%. בנוסף, ערכי ה Precision, Recall-ו F1-Score-במודל זה מצביעים על איזון טוב בין זיהוי נכון של ימי עלייה לבין הימנעות מסיווגים שגויים. ערך F1-Score הגבוה יחסית מעיד כי המודל מצליח לשלב בצורה יעילה בין דיוק החיזוי לבין כיסוי נכון של המקרים החיוביים. רמת ה Precision-הגבוהה מצביעה על כך שרוב התחזיות החיוביות אכן תואמות את התוצאה בפועל, דבר שמפחית את הסיכון לקבל החלטות שגויות. לכן, בשל השילוב בין דיוק מעט גבוה יותר ויציבות במדדי הביצוע, נבחר מודל Random Forest להמשך העבודה והניתוח בפרויקט.

האקדמית עמק יזרעאל
ע"ש מקס שטרן

דו"ח התקדמות להגשה לשלב Modelling

שמות הסטודנטים : נועם פטיש וניקול בלישר

מס' ת"ז : 212205082, 328654546

שם הפרויקט : חיזוי מניות

שם המנחה : ג'ניה גוטפריד

בשלב זה חיברנו את המסד נתונים שנמצא בSQL לקוד בפייתון להרצת המודל.
שלפנו נתונים, ניקינו ערכים חסרים, המרנו טקסטים ואובייקטים למספרים לצורך חישובים.
לאחר מכן, ביצענו חלוקת נתונים לפי שיטת Train/Validation/Test , בחרנו כמה מודלים לצורך השוואה ובדיקה של המודל.
ביצענו הרצה ראשונית ומצאנו כי באלגוריתם של Random Forest כי רמת הדיוק יותר גבוהה ולכן אנו רואות לנכון לבחור בשיטה זו בהמשך הפרויקט שלנו.

חתימת הסטודנט : נועם פטיש

חתימת הסטודנט : ניקול בלישר

חתימת המנחה _____

תאריך 13/04/2026