בס"ד

דו"ח סופי – פרויקט גמר

נושא הפרויקט – חיזוי הכחשות עסקה בחברת השקעות בעזרת מודלים של למידת מכונה.

מנחים אקדמאיים

ד"ר רון הירשפרונג וד"ר אילן לאופר

מרכז הקורס

ד"ר אילן לאופר

מנחה תעשייתי

שאדי ג'אבר

מגישים

רוני קרוניץ

יאיר עמר

מוריה מרגליות

תוכן עניינים

1. תקציר מנהלים

תיאור הארגון:

AvaTrade היא חברה מובילה למסחר עצמאי באמצעות פלטפורמות מסחר מקוונות, המאפשרות לאנשים לסחור בשווקים פיננסיים מכל רחבי העולם. החברה נוסדה בשנת 2006 וצמחה לחברה בולטת בתעשייה, בזכות מחויבותה לחדשנות, אמינות ושירות לקוחות זמין 6 ימים בשבוע במגוון אמצעי תקשורת כמו טלפון, מייל, וווטסאפ ופייסבוק. החברה מספקת שירותי מסחר עצמאי בבורסה ללקוחות ביותר מ-150 מדינות, עם משרדים ברחבי העולם ומשרתת מעל 400 אלף לקוחות.

המודל הרווחי של החברה:

החברה רווחית בזכות ניצול מרווחים (spreads), עמלות לילה ועמלות אי פעילות. AvaTrade מתחרה עם מספר זירות מסחר כמו eToro, Plus500 ו-Trading212.

הגדרת הבעיה:

החברה מתמודדת עם שני סוגי הונאות עיקריים: הכחשות עסקה לא כנות (Friendly Chargeback) והונאות פליליות (Criminal Chargeback). החברה מנסה לאזן בין ניהול סיכון ההונאה לבין חווית המשתמש, כדי לעודד לקוחות להתחבר ולבצע פעולות בחשבון.

מטרות הפרויקט:

מטרת הפרויקט המרכזית היא הפחתת כמות מקרי הכחשות העסקה, על ידי בניית מודל לחיזוי ההסתברות שהלקוח יכחיש עסקה. התוצאות הנגזרות מהמטרה המרכזית כוללות הפחתת נזק כלכלי, צמצום בעיות רגולטוריות והפחתת כמות שיחות הטלפון.

אילוצי הפרויקט:

במהלך רוב הסמסטר הראשון, אחת מחברות הצוות התמודדה עם קשיים אישיים שהשפיעו על המשך השתתפותה בפרויקט. בנוסף, אחד מחברי הצוות היה במילואים במשך יותר מ-150 ימים במהלך המלחמה, ולאחר מכן תרם כליה שגררה אשפוז והתאוששות, מה שצמצם את זמינותו לפרויקט. בנוסף, בוצע ניסיון לבדוק אפשרות לקבל הקלטות של שיחות בין הלקוחות לאנשי השירות של החברה, במטרה להשתמש במודלי עיבוד טקסט (NLP) לניתוח השיחות. עם זאת, נמסר לנו מהחברה שפרויקט כזה ידרוש כוח מחשוב אדיר שאינו ריאלי עם המחשבים שברשותנו, ודורש הגעה פיזית למשרדי החברה. בנוסף, ביקשנו מהחברה מדדים ומשתנים רלוונטיים כגון TP, FP, TN, FN, כדי לחשב מדדים רלוונטיים נוספים כמו AUC, Precision, Recall F1 score ו-Accuracy. אולם, נענינו כי בשלב זה אין ברשותם את הנתונים הדרושים. בנוסף, החברה נמצאת כעת בתהליך מעבר לשימוש בשירותי ענן, דבר שמקטין את הזמינות וההיענות לצרכינו במסגרת הפרויקט.

<u>שיטות וטכניקות:</u>

דרכי טיפול בדאטה:

- ללא טיפול מיוחד
- SMOTE (Synthetic Minority Over-sampling Technique)
- Under Sampling

מודלים לחיזוי:

- AdaBoost
- Random Forest
- Gradient Boosting (בעקבות זמן ריצה ארוך הוחלט להשתמש רק בHGB)
- Decision Tree
- Logistic Regression
- Hist Gradient Boosting
- ANN
- Voting Classifier

<u>תהליך אימון, ובדיקת המודלים והשיטות:</u>

על מנת להעריך את ביצועי המודלים תחת התנאים השונים, השתמשנו בשיטת קיפול צולב מאוזן חוזר (Repeated Stratified K-Fold) עם 10 חלוקות (k=10). שיטה זו מחלקת את הנתונים לעשרה חלקים באופן ששומר על הפרופורציות של המחלקות בכל חלק. חזרה על התהליך מספר פעמים מאפשרת קבלת הערכות יציבות ואמינות יותר על ידי הפחתת ההשפעה של חלוקה מקרית של הנתונים.

<u>תוצאות והערכת ביצועים:</u>

לאחר אימון המודל, נבחנו ביצועיו על סט המבחן, בעזרת מגוון מדדים שנבחרו לשם הערכת דיוקו ויעילותו בזיהוי מקרי הכחשת עסקה.

מדדים סטטיסטיים:

כל אחד מהמודלים הוערך לפי המדדים: דיוק (Accuracy), זיכרון (Recall), ספציפיות (Specificity), דיוק משוקלל (Precision), F1-Score ו-ROC AUC.

מדדים כמותיים:

מדד המודד את כמות הכחשות העסקה, באחוזים, עבור כל חודש. מדד זה יספק לחברת AvaTrade תמונת מצב כיום בנוגע לכמות הכחשות העסקה ובנוגע לעונתיות של הכחשות העסקה. כיום, התאריכים המסופקים על ידי החברה בנתונים שהתקבלו לצורך הפרויקט, מספקים מידע על תאריך הצטרפות הלקוח ועל האם ביצע הכחשה או לא. לחברה אין מידע בנוגע לזמן ההכחשה.

מדד כמותי נוסף הוא אחוז הכחשות העסקה ביחס לסך נתוני העסקאות שיש ברשות החברה.

מדד איכותני:

מדד זה הינו שאלון שהועבר בין העובדים, על מנת שנוכל להעריך את תרומת המודל שלנו לחברה. ההערכה תבוצע על ידי השוואה בין תגובות המשיבים על הסקר לפני הכנסת המודל שלנו לארגון, לבין תגובות המשיבים על הסקר לאחר הכנסת המודל שלנו לארגון.

לאחר ניתוח תוצאות המבחן והערכת ביצועי המודל, נמצא כי המודל HGB הציג ביצועים טובים ואמינים במיוחד. ההסתמכות על מדד ה-Recall הייתה מרכזית, מאחר שהעדיפות ניתנה להפחתת שגיאות מסוג False Negatives כדי להבטיח את זיהוי כל מקרי ההונאה הפוטנציאליים.

ההשוואה בין שיטות הטיפול השונות בנתונים הראתה שהשיטה הנבחרת, Under-sampling, הציגה את הביצועים המאוזנים והמדויקים ביותר, תוך הימנעות מבעיית Overfitting שעלולה להופיע בשיטות אחרות.

מימוש הפתרון והערכתו:

הפתרון בפרויקט התבסס על חלוקת הנתונים לסטים של אימון (80% מהנתונים) ושל מבחן (20% מהנתונים), תוך שימוש בשיטת Repeated Stratified KFold לוולידציה מדויקת. המודל שנבחר הוא HistGradientBoostingClassifier, שעבר אופטימיזציה באמצעות חיפוש רשת (Grid Search). תהליך העיבוד המקדים כלל סקלינג של נתונים נומריים, קידוד תכונות קטגוריאליות, ו-Undersampling לאיזון הדאטה.

לאחר אימון המודל, ביצועיו נבחנו על ידי מדדים שונים. המודל הציג תוצאות מצוינות, במיוחד במדד ה-Recall, שבו זוהו מרבית מקרי ההונאה עם מינימום שגיאות מסוגFalse Negatives. מבין שיטות הטיפול השונות, Under-sampling נמצאה כאפקטיבית ביותר, והפתרון הכללי עמד ביעדי הפרויקט, עם תחזיות מדויקות ואמינות.

בסופו של דבר, הפתרון הנבחר עמד ביעדים שהוגדרו בתחילת הפרויקט והוכיח את יכולתו לספק תחזיות מדויקות ואמינות, המסייעות לחברה בהתמודדות עם סיכונים כלכליים ורגולטוריים הנובעים ממקרי הכחשת עסקה.

לסיכום, הפרויקט התמקד בהפחתת מקרי הונאה, על ידי פיתוח מודלים לחיזוי הכחשות עסקה, תוך שימוש במגוון שיטות טיפול בדאטה ומודלים לחיזוי.

מצורף קישור לגיטהאב, בו ניתן לצפות במאמרים ובקוד:
https://github.com/yairamar097/Final_Project.git

2. תיאור הארגון

AvaTrade הינה חברה מובילה למסחר עצמאי בעזרת פלטפורמות מסחר מקוונות, המאפשרות לאנשים לסחור ממקומות שונים בעולם, בשווקים פיננסיים שונים מרחבי העולם. החברה נוסדה בשנת 2006 ומאז צמחה לחברה בולטת בתעשייה. היא צברה מוניטין גבוה בזכות מחויבותה לחדשנות, אמינות ושירות לקוחות זמין 6 ימים בשבוע. שירות הלקוחות זמין באמצעות מגוון אמצעי תקשורת כגון טלפון, מייל, ווטסאפ ופייסבוק.

AvaTrade מספקת ללקוחות שירותי מסחר עצמאי בבורסה, ביותר מ-150 מדינות ברחבי העולם, עם משרדים בישראל, איטליה, פולין, אירלנד, מקסיקו, סין, יפן, אוסטרליה ועוד. החברה משרתת מעל 400 אלף לקוחות המבצעים למעלה מ-3 מיליון פעולות מסחר בהיקף של למעלה מ-60 מיליון דולר בחודש.

<u>מכשירים פיננסיים:</u>

AvaTrade מציעה מגוון רחב של מכשירים פיננסיים, כולל מניות, סחורות, מדדים, מט"ח, מטבעות קריפטוגרפים ועוד. מבחר רחב זה מאפשר לסוחרים לגוון את השקעותיהם ולהתאים את אסטרטגיות המסחר שלהם לתנאי שוק שונים.

מכשיר פיננסי נוסף הוא חוזה הפרשים (CFD, Contract for Difference). חוזה הפרשים הוא מכשיר פיננסי המאפשר למשקיעים לנצל את השינויים במחיר של נכס מסוים מבלי להחזיק בו בפועל. בחוזה מסוג זה, שני צדדים, הקונה והמוכר, מסכימים להחליף את ההפרש במחיר של הנכס מהתאריך בו החוזה נכנס לתוקף ועד לתאריך הסגירה שלו. ההפרש יכול להיות חיובי או שלילי, תלוי בכיוון שבו הנכס התנהל מאז הכניסה לחוזה. המשקיעים בחוזי הפרשים יכולים לנסות להרוויח הן מעליות והן מירידות במחירים של מגוון נכסים, כולל מניות, מט"ח, מדדים, סחורות ועוד. אחת התכונות המרכזיות של חוזי הפרשים היא היכולת לסחור במינוף, מה שאומר שהמשקיעים יכולים לפתוח עסקאות גדולות יותר מההון שהם בפועל מחזיקים, תמורת הפקדה (מרג'ין) שהיא רק חלק קטן מערך העסקה. אך, יש לזכור כי המינוף מגביר גם את הסיכון ויכול להוביל להפסדים גדולים ולא רק לרווחים.

<u>פלטפורמות מסחר:</u>

החברה מספקת לסוחרים פלטפורמות מסחר ידידותיות למשתמש ועשירות בתכונות. AvaTrade תומכת בפלטפורמות פופולריות כגון MetaTrader 4, MetaTrader 5, והפלטפורמה העצמאית שלה, AvaTradeGo. פלטפורמות אלו מציעות כלי תרשימים מתקדמים, כלים לניתוח טכני, אינדיקטורים הניתנים להתאמה אישית, נתוני שוק בזמן אמת ויכולות ביצוע הזמנות. בנוסף, החברה מספקת מערך הדרכה נרחב הכולל מידע לסוחר המתחיל, מאמרים, מילון מונחים וסרטוני הדגמה והדרכה.

סוגי חשבונות:

AvaTrade פונה לסוחרים ברמות ניסיון שונות על ידי הצעת סוגים שונים של חשבונות, כולל חשבונות סטנדרטיים, חשבונות אסלאמיים (תואמים לחוקי השריעה), וחשבונות הדגמה לתרגול מסחר. לכל סוג חשבון יש תכונות והטבות משלו המותאמות לדרישות הסוחר הספציפיות.

אבטחת המידע של הסוחרים ורגולציה:

AvaTrade נותנת עדיפות לאבטחה והגנה על הכספים של לקוחותיה. כדי להבטיח סביבת מסחר בטוחה, הארגון מקפיד על מסגרות רגולטוריות חזקות.

להלן רשימת גופים רגולטורים בולטים המפקחים על הפעילות של AvaTrade:

- הבנק המרכזי של אירלנד: AvaTrade מורשה על ידי הבנק המרכזי של אירלנד ומפוקחת על ידו. פיקוח רגולטורי זה נועד להבטיח שהארגון עומד בסטנדרטים פיננסיים מחמירים ושומר על שקיפות בפעילתו.
- הוועדה לשירותים פיננסים (FSC) - איי הבתולה הבריטיים: AvaTrade מוסדרת גם על ידי הנציבות לשירותים פיננסיים באיי הבתולה הבריטיים. רשות רגולטורית זו מפקחת ומפוקחת על פעילות הארגון לקידום נהלי מסחר הוגנים ואתיים.
- רשות ניירות ערך והשקעות אוסטרלית (ASIC): עבור פעילותה באוסטרליה, AvaTrade מוסדרת על ידי ASIC. גוף רגולטורי זה שומר על האינטרסים של סוחרים אוסטרלים ומבטיחים כי AvaTrade עומדת בדרישות הרגולטוריות הנדרשות במדינה.

כמו כן, AvaTrade פועלת בכפוף לחוקים ולמדיניות בכל מדינה ומדינה בה החברה פעילה, וביניהן גם הרשות לניירות ערך במדינת ישראל.

המודל הרווחי של החברה:

- **spreads** (מרווח): spreads, מרווח, הוא ההפרש בין מחיר הקנייה למחיר המכירה של מכשיר פיננסי או כלי מסחר אחר. זהו הרווח שהחברה מרוויחה מכל עסקה, והוא עשוי להשתנות ממכשיר למכשיר.
- **עמלת לילה**: עמלת לילה היא עמלה שנלקחת על פוזיציה פתוחה בשעה 00:00 בלילה. העמלה עשויה להשתנות ממכשיר למכשיר.
- **עמלת אי פעילות**: החברה גובה עמלת אי פעילות רבעונית ושנתית בחשבון במקרה של חוסר פעילות במשך 3 חודשים או שנה.

לחברה מספר זירות מסחר מתחרות, ביניהן:

- ETORO
- 500PLUS
- TRADING212

תהליך התחברות וביצוע עסקאות:

1. לקוח מגיע לאתר החברה (לוחץ על קישור אל האתר/האפליקציה דרך חיפוש בגוגל, קליק על גבי פרסומת וכדומה.)
2. הלקוח מחליט אם להירשם באופן מידי או לקבל מידע לגביי החברה.
3. הלקוח מחליט אם לפתוח חשבון אמיתי למסחר או חשבון דמה.
4. לאחר שלקוח פתח חשבון הוא יקבל שיחת טלפון משירות\שימור לקוחות.
5. אם הלקוח פתח חשבון אמיתי הוא ימלא את המסמכים הנדרשים לרישום .
6. במצב ואכן אישרו ללקוח את כל המסמכים הלקוח יוכל להתחיל להפקיד כסף ולסחור.

3. הגדרת הבעיה

כיום החברה מתמודדת בעיקר עם שני סוגי הונאות:

1. Friendly Chargeback – כשהלקוח מכחיש עסקה בשל הפסד.

2. Criminal Chargeback – לקוח שגנב את כרטיסי אשראי והפקיד לחשבון מסחר שלו.

מטרת החברה היא לאזן בין הרצון להפחית את הסיכון להונאה, לבין הנוחות וחווית המשתמש של לקוחותיה, בעת ביצוע. לדוגמה, אימות דו-שלבי מקשה על ההתחברות, פוגע בחוויית המשתמש, וכתוצאה מכך עלול לפגוע ברווחי החברה.

כיום, אחוז הכחשות העסקה המדווחות עומד על כחצי אחוז, כאשר עד לאחוז אחד חברות האשראי מאפשרות המשך התקשרות ללא התערבות וחסימה. הסיכון העיקרי של החברה הוא התערבות של חברות האשראי, ולכן שואפת לצמצם את אחוז הכחשות העסקה ככל הניתן, על מנת להשיג מרווח ביטחון גדול יותר.

4. מטרות הפרויקט

מטרה מרכזית

מטרת הפרויקט המרכזית היא הפחתת כמות מקרי הכחשות העסקה, על מנת להימנע מסנקציות של חברות האשראי, על ידי בניית מודל שיריצו כל תקופה והמודל יחזיר הסתברות שהלקוח יכחיש עסקה.

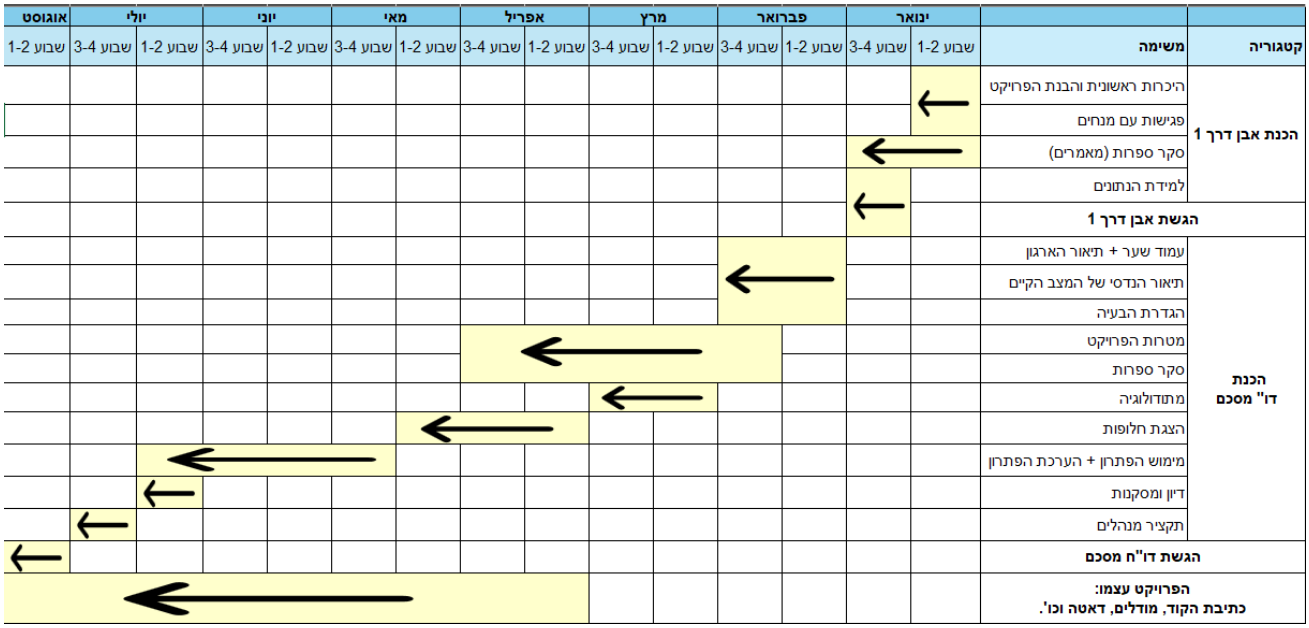
מטרות ותוצאות נגזרות

- החברה תהיה חשופה פחות להחזרים כספיים ללקוחות ולקנסות מצד חברת האשראי – פחות נזק כלכלי.
- החברה נדרשת להוכיח שהיא עושה מספיק תהליכים על מנת למנוע מקרי ביצוע עסקאות ע"י גנבים, או, שההכחשה הזו היא אשמת הלקוח. דבר זה מורכב וקשה וע"י זיהוי כללי של מקרי הכחשת עסקה – נוכל לצמצם גם את הבעיה הזו.
- הפחתת החשיפה לבעיות רגולטוריות.
- הפחתה בכמות שיחות טלפון, שנועדו לבדוק את האדם הסוחר, על מנת למנוע הונאות.

5. תרשים גאנט

תרשים הגאנט המוצג בפרק זה משקף את תהליך העבודה בפרויקט, כולל כל שלבי התכנון, הביצוע, וההערכה. התרשים מספק סקירה של לוח הזמנים לכל משימה, החל מהבנת צרכי החברה ובחינת האפשרויות השונות, דרך יישום הפתרון הנבחר, ועד להגשת הדו"ח הסופי. הגאנט ממחיש את סדר הפעולות, את ההתקדמות בכל שלב ואת האינטגרציה בין הפעילויות השונות לאורך חודשי הפרויקט, תוך הדגשת המועדים הקריטיים והעמידה ביעדים שנקבעו.

להלן תרשים הגאנט:



6. סקר ספרות

תקציר מאמר Artificial Intelligence and Fraud Detection:

המאמר "Artificial Intelligence and Fraud Detection" מאת יאנג באו, ג'ילס הילארי ו-בין קה, מציג את האתגרים וההזדמנויות בשימוש בטכנולוגיות בינה מלאכותית (AI) ולמידת מכונה (ML) לזיהוי ומניעת הונאות. המאמר מדגיש כי למרות ההתקדמות הטכנולוגית, רק 13% מהארגונים משתמשים בטכנולוגיות אלו לזיהוי הונאות. אחת הסיבות לכך היא שהונאות מסוגים שונים דורשות גישות שונות, והונאות אשר קל יחסית לזהות, כמו הונאות בכרטיסי אשראי, מותאמות יותר לשימוש ב-ML. [1]

אתגרים בזיהוי הונאות

המאמר מפרט את האתגרים העיקריים בזיהוי הונאות באמצעות ML:

1. **חוסר איזון בנתונים**: הונאות הן נדירות, ולכן יש מעט נתונים ללמד את האלגוריתמים.
2. **התנגדות פעילה מצד מבצעי הונאות**: אנשים המבצעים הונאה, מנסים לשנות את דפוסי הפעילות כדי להקשות על הזיהוי.
3. **זיהוי מאוחר של הונאות**: חלק מההונאות מתגלות רק לאחר זמן רב, דבר המקשה על אימון האלגוריתמים.
4. **מגבלות בדאטה**: איכות הנתונים, שלמותם ואמינותם משפיעות רבות על יכולת המודלים לזהות הונאות.

תרומת המאמר לפרויקט

המאמר מספק סקירה נרחבת על האתגרים וההזדמנויות בשימוש בטכנולוגיות בינה מלאכותית לזיהוי הונאות, דבר שיכול לסייע בבניית מודל לחיזוי הסתברות הכחשת עסקה בפלטפורמת מסחר. השימוש בדאטה איכותי ומגוון הוא קריטי להצלחת המודל, וכן הבנת דפוסי ההתנהגות של הרמאים והתאמת המודל לשינויים בדפוסים אלו.

תקציר מאמר Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review:

המאמר "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review" מציג סקירה שיטתית של הספרות הקיימת בנוגע לזיהוי הונאות פיננסיות באמצעות למידת מכונה (ML). ההונאות הפיננסיות הפכו לאיום נפוץ על חברות וארגונים, וטכניקות קונבנציונליות, כמו אימותים ידניים, אינן מדויקות, יקרות וגוזלות זמן. לכן, גישות מבוססות ML יכולות לסייע בזיהוי הונאות פיננסיות על ידי ניתוח כמויות גדולות של נתונים. [2]

טכניקות למידת מכונה נפוצות

המאמר מסכם את טכניקות ה-ML הנפוצות ביותר לזיהוי הונאות:

- **תמיכה במכונות ווקטוריות (SVM)**
- **רשתות נוירונים מלאכותיות (ANN)**
- **מודלים סמויים של מרקוב (HMM)**
- **רגרסיה לוגיסטית**
- **עצי החלטה**
- **כללי למידת בייס (Naive Bayes)**
- **שיטות אנליזה מקובצות (Clustering)**

הסקירה מצביעה על כך שהונאות בכרטיסי אשראי הן הסוג הנפוץ ביותר של הונאה שנחקרה באמצעות טכניקות אלו.

תרומת המאמר לפרויקט

המאמר מספק הבנה רחבה של השיטות הנפוצות והאתגרים בזיהוי הונאות פיננסיות באמצעות ML. התובנות מהמאמר יכולות לסייע בבניית מודל לחיזוי הסתברות הכחשת עסקה בפלטפורמת מסחר על ידי שימוש בטכניקות ML מתקדמות ויישום מדדים להערכת ביצועים כמו דיוק, רגישות, זיהוי שגוי ועוד. הבנת הדפוסים השונים של ההונאות והטכניקות המתאימות לזיהוי שלהם תאפשר יצירת מודל חיזוי מדויק ואמין.

תקציר מאמר MetaFraud: A Meta-Learning Framework for Detecting Financial Fraud:

מבוא ומוטיבציה

המאמר "MetaFraud: A Meta-Learning Framework for Detecting Financial Fraud" מאת אחמד עבאסי, קונאן אלברכט, אנתוני ואנס וג'יימס הנסן מציג את MetaFraud, מסגרת למידת מטא (Meta-Learning) חדשנית לזיהוי הונאות פיננסיות. זיהוי הונאות פיננסיות הוא אתגר מרכזי בטכנולוגיות אינטליגנציה עסקית (BI), והמאמר מדגיש את הצורך בשיטות זיהוי חזקות יותר על מנת להתמודד עם ההשפעות השליליות של הונאות על חברות, משקיעים וכלכלות. [3]

מתודולוגיה

המאמר משתמש בגישת עיצוב מדעי (Design Science) לפיתוח מסגרת MetaFraud. המסגרת כוללת ארבעה מרכיבים עיקריים:

1. **בנייה של תכונות מקיפות**: שימוש בנתונים כספיים ציבוריים ובניית מדדים חדשים המשלבים מידע ארגוני ותעשייתי.
2. **למידה חצי-מובחנת ואקטיבית**: שיפור ביצועי המודל על ידי התאמה ושיפור מתמידים של אלגוריתמים.
3. **גנרליזציה מרובדת (Stacked Generalization)**: שילוב של מספר מודלים בסיסיים לכדי מודל עליון שמנצל את החוזקות של כל מודל בסיסי.
4. **מדדי רמת ביטחון**: המערכת מייצרת מדדי ביטחון לכל תחזית, מה שמאפשר זיהוי אמין יותר של הונאות פיננסיות.

תוצאות ותרומה לפרויקט

המאמר מציג תוצאות ניסויים המעידים על כך שכל אחד ממרכיבי המסגרת תורם לשיפור הביצועים הכוללים של המסגרת. הניסויים מראים שהמסגרת MetaFraud עולה על שיטות זיהוי הונאות קיימות ומספקת תחזיות אמינות עם מדדי ביטחון גבוהים.

תרומת המאמר לפרויקט

המאמר מספק תובנות חיוניות לבניית מודל לחיזוי ההסתברות שלקוח של פלטפורמת מסחר יכחיש עסקה. הבנה מעמיקה של שיטות מתקדמות בזיהוי הונאות פיננסיות ושילוב של מספר מודלים בסיסיים לכדי מודל עליון יכולות לשפר את דיוק המודל ולספק תחזיות אמינות יותר. בנוסף, השימוש במדדי ביטחון יכול לסייע בזיהוי מקרים עם סבירות גבוהה להונאה.

תקציר מאמר Fraud Detection in Online Payments using Machine Learning Techniques:

מבוא ומוטיבציה

המאמר "Fraud Detection in Online Payments using Machine Learning Techniques" מאת ו' סידאיה, פ' אנג'ניולו, מ' ראמש ו-י' האריטה דן באתגרים וההזדמנויות בזיהוי הונאות בעסקאות מקוונות באמצעות טכניקות למידת מכונה. בעידן שבו עסקאות מקוונות נעשות נפוצות יותר ויותר, הסיכון להונאות גם הוא גדל. המחקר מציע מודל מבוסס למידת מכונה לזיהוי הונאות בעסקאות, תוך שימוש בנתונים ציבוריים ושיטות מתקדמות כמו XGBoost להשגת תוצאות מדויקות ומהירות. [4]

מתודולוגיה

המחקר משתמש במערך נתונים מ Kaggle-הכולל 636,263 רשומות עם 10 תכונות שונות. הנתונים עוברים תהליך קדם עיבוד (Data Preprocessing) כולל המרת ערכי מחרוזת לערכים מספריים, הסרת ערכים ריקים ותקנון הנתונים לטווח קבוע. לאחר מכן, מוחלים אלגוריתמים של למידת מכונה כגון יער אקראי (Random Forest) ו XGBoost-על הנתונים המסווגים כדי לזהות הונאות.

תוצאות

המחקר מצביע על כך ש XGBoost-מציג את הביצועים הטובים ביותר בזיהוי הונאות בעסקאות מקוונות עם דיוק של 97.91%, בעוד ש Random Forest-הציג דיוק של 86.68%. הניסויים כללו שימוש במטריצת בלבול (Confusion Matrix) להערכת ביצועי המודלים.

תרומת המאמר לפרויקט

המאמר מספק הבנה מעמיקה של טכניקות מתקדמות בזיהוי הונאות מקוונות, אשר יכולה לסייע בבניית מודל לחיזוי ההסתברות שלקוח של פלטפורמת מסחר יכחיש עסקה. שימוש בשיטות כמו XGBoost מאפשר דיוק גבוה וזיהוי מהיר של דפוסי הונאה, מה שיכול לשפר את אמינות המודל בפרויקט. התובנות מהממצאים יוכלו לשמש לפיתוח מערכת זיהוי הונאות המותאמת לצרכים הספציפיים של פלטפורמת מסחר.

תקציר מאמר Detecting Chargebacks in Transaction Data with Artificial Neural Networks:

מבוא ומוטיבציה

המאמר "Detecting Chargebacks in Transaction Data with Artificial Neural Networks" מאת תיאודור גונתר ואוטו פאג'לס-פייקלס מציג מחקר הבוחן את השימוש ברשתות נוירונים מלאכותיות (ANN) לחיזוי החזרות עסקאות (Chargebacks) בעסקאות מקוונות. תהליך החזר העסקה הוא תהליך יקר עבור הסוחרים, וכולל אובדן הכנסות ודמי טיפול גבוהים לבנקים. המחקר מתמקד בפיתוח מודל לחיזוי עתידי של החזרות עסקאות במטרה לצמצם עלויות כספיות ולשפר את רווחיות הסוחרים. [5]

מתודולוגיה

המחקר השתמש במודל של רשת נוירונים מלאכותית שסווגה כמודל למידה מונחית. הנתונים נאספו מחברת NinjaCall, חברה טכנולוגית המתמקדת בשירותי תקשורת בינלאומית. המודל אומן עם נתוני עסקאות מהחברה, אשר כללו תכונות כגון כמות העסקה, מדינת התשלום, כתובת ה-IP ועוד. הנתונים חולקו לערכות אימון, ולידציה ובדיקה.

תוצאות

המודל הצליח לסווג כ-60% מהחזרות העסקאות הנכונות, אך סבל מהטיית יתר בכך שסיווג כמות גבוהה של עסקאות כהחזרים, עם דיוק של 13% בלבד בעסקאות שסווגו כהחזרים. על אף הדיוק הנמוך, המודל הראה פוטנציאל כשימוש ככלי משלים למערכת הסיווג הקיימת של NinjaCall, שיכולה לסייע באיתור עסקאות חשודות ולדרוש אימות נוסף לפני אישורן.

תרומת המאמר לפרויקט

המאמר מספק תובנות חשובות לבניית מודל לחיזוי הסתברות הכחשת עסקה בפלטפורמת מסחר. השימוש ברשתות נוירונים מלאכותיות מאפשר זיהוי של דפוסים מורכבים בעסקאות, דבר שיכול לשפר את יכולת החיזוי של המודל בפרויקט. בנוסף, היישום של המודל ככלי משלים יכול לעזור בצמצום עלויות הקשורות להחזרים ולהגדיל את שביעות הרצון של הלקוחות.

ככלל, המאמרים מציגים מספר מגבלות שחוזרות על עצמן במסגרת המחקרים בנוגע לחיזוי הסתברויות שונות הקשורות בחיזוי הונאה. מגבלות אלה כוללות:

1. **חוסר איזון בנתונים**: אחוז נמוך של החזרות בעסקאות עלול להוביל להטיות במודל ולדיוק נמוך בזיהוי ההחזרים.
2. **השתנות דפוסי ההחזרים**: דפוסי ההחזרים יכולים להשתנות עם הזמן, מה שעלול להשפיע על הדיוק של המודל אם לא מתבצע עדכון תדיר של הנתונים.
3. **איכות הנתונים**: איכות הנתונים והשלמות שלהם משפיעות רבות על ביצועי המודל. למשל, הנתונים ששימשו לאימון המודל במאמר האחרון כבר עברו סינון על ידי מערכת הסיווג הקיימת של NinjaCall, מה שיכול להוביל להטיות נוספות.

מגבלות אלו מצביעות על הצורך בשיפור ואיזון הנתונים, וכן בעדכון מתמיד של המודל כדי להתמודד עם השינויים בדפוסי ההחזרים.

<u>סקר ספרות אינטגרטיבי</u>

<u>מבוא לזיהוי הונאות הכחשות עסקה</u>

הונאות הכחשות עסקה (Chargeback Fraud) הן אתגר משמעותי עבור סוחרים במערכת התשלומים המקוונת. הונאות אלו מתרחשות כאשר לקוח מבצע עסקה חוקית אך לאחר מכן מכחיש אותה בטענה שלא ביצע אותה, מה שמוביל להחזר כספי על חשבון הסוחר. תופעה זו גורמת להפסדים כלכליים משמעותיים לסוחרים, מאחר והם לא רק מאבדים את ההכנסה מהעסקה אלא גם משלמים עמלות נוספות על ההחזר. [4]

החשיבות הכלכלית והבעיות הרגולטוריות

הונאות הכחשות עסקה מהוות בעיה כלכלית חמורה עבור סוחרים, שכן הן משפיעות ישירות על הרווחיות ועל עלויות התפעול. החזרים כספיים מצריכים תשלום עמלות נוספות לספקי השירותים הפיננסיים, מה שמגדיל את ההפסדים הכלכליים. בנוסף, סוחרים נדרשים להשקיע משאבים נוספים במעקב אחר עסקאות ובחקירת מקרים של הונאה, דבר שמגביר את העלויות הכוללות של העסק. [4] בנוסף להיבט הכלכלי, ישנם גם קשיים רגולטוריים הנוגעים לזיהוי וטיפול בהונאות הכחשות עסקה. התקנות והחוקים המשתנים בין מדינות שונות מציבים אתגרים נוספים לסוחרים בניהול עסקאות בינלאומיות ובהתמודדות עם מקרים של הונאה. [2]

הקשיים בזיהוי הונאות הכחשות עסקה

זיהוי הונאות הכחשות עסקה הוא אתגר מורכב, מכיוון שמקרים רבים נראים כלגיטימיים במבט ראשון. לקוחות עשויים להגיש תלונות על עסקאות שהם אכן ביצעו, תוך ניצול הפרצות במערכות הפיננסיות. בנוסף, הקושי להבדיל בין טעויות אמיתיות של לקוחות לבין כוונות זדון מקשה על זיהוי מדויק של מקרים אלו. [3]

סוגי הרמאיות בהכחשות עסקה

ישנם מספר סוגים של הונאות הכחשות עסקה, כולל הונאות עסקה לא כנות (Friendly Fraud), שבהן לקוחות מכחישים עסקאות שהם אכן ביצעו, והונאות צד שלישי, שבהן צד שלישי מבצע את העסקה ללא ידיעת הלקוח האמיתי. סוגי הרמאיות השונים מצריכים שימוש בשיטות זיהוי שונות ובניית מודלים מותאמים לכל סוג של הונאה. [4]

טכניקות לזיהוי הונאות הכחשות עסקה

זיהוי הונאות הכחשות עסקה מצריך שימוש בטכניקות מתקדמות ללמידת מכונה ובינה מלאכותית. אחד הכלים המרכזיים הוא שימוש במודלים של רשתות נוירונים מלאכותיות (Artificial Neural Networks). מודלים אלו יכולים לנתח כמויות גדולות של נתונים ולזהות דפוסים חריגים בהתנהגות הלקוחות, דבר המאפשר גילוי מוקדם של הונאות פוטנציאליות. מודלים של רשתות נוירונים מלאכותיות (ANNs) הם מהכלים החזקים ביותר בתחום זיהוי הונאות הכחשות עסקה. רשתות נוירונים מסוגלות לזהות דפוסים חריגים בהתנהגות הלקוחות ולספק זיהוי מוקדם של הונאות הכחשות עסקה. [3] שימוש ברשתות נוירונים מלאכותיות מאפשר זיהוי מוקדם ומדויק של הונאות הכחשות עסקה על ידי למידה והבנה של דפוסים מורכבים בתוך נתוני העסקאות. [4]

שיטות נוספות כוללות שימוש במודלים של למידת מכונה מבוססי עץ כמו Random Forest ו-XGBoost. מודלים אלו מפיקים החלטות מתוך מספר רב של עצי החלטה, ומספקים דיוק גבוה בזיהוי

דפוסים חריגים בעסקאות. הם מאפשרים ניתוח רב-שכבתי של הנתונים וזיהוי תבניות שאינן גלויות באמצעים אחרים. [3]

שילוב אלגוריתמים ומודלים היברידיים

השילוב של מספר אלגוריתמים במודל אחד יוצר מודל היברידי, המנצל את היתרונות של כל אחד מהאלגוריתמים המשולבים. לדוגמה, שילוב של ANN עם Random Forest יכול לשפר את היכולת לזהות הונאות על ידי שילוב הכוח של ניתוח דפוסים מורכבים עם היכולת להתמודד עם מגוון רחב של נתונים. [2] מודלים היברידיים אלו מאפשרים זיהוי מדויק יותר של הונאות ומפחיתים את כמות הניבויים השגויים.

מטריקות הערכה של המודלים לזיהוי הונאות הכחשות עסקה

הערכת ביצועי המודלים לזיהוי הונאות הכחשות עסקה מצריכה שימוש במטריקות הערכה מדויקות ומגוונות. מטריקות אלו מאפשרות לבחון את דיוק המודלים, הרגישות שלהם, ויכולת הזיהוי של דפוסים חריגים. שימוש במטריקות אלו מבטיח שהמודלים מסוגלים לזהות הונאות בצורה יעילה ואמינה. דיוק (Accuracy) היא אחת המטריקות המרכזיות בהערכת ביצועי המודלים. דיוק המודל מודד את אחוז התחזיות הנכונות מכלל התחזיות. עם זאת, במקרים של דאטה לא מאוזן, כמו בזיהוי הונאות הכחשות עסקה, דיוק גבוה עלול להטעות אם המודל מנבא בעיקר את הקטגוריה הדומיננטית. לכן, חשוב להשתמש במטריקות נוספות. רגישות (Recall) היא מטריקה נוספת שמודדת את יכולת המודל לזהות את כל המקרים האמיתיים של הונאה מתוך כלל המקרים האמיתיים. מטריקה זו חשובה במיוחד בזיהוי הונאות, כיוון שהיא מאפשרת להעריך את יכולת המודל לזהות מקרים אמיתיים של הונאה ולהפחית את מספר המקרים המוחמצים. [3] דיוק תחזיתי (Precision) מודד את אחוז התחזיות הנכונות מתוך כלל התחזיות החיוביות של המודל. מטריקה זו חשובה להערכת המודל בזיהוי הונאות, שכן היא מאפשרת להבין כמה מתוך המקרים שזוהו כהונאות הם באמת הונאות. [4] מטריקה נוספת היא ציון (F1 Score). ציון הוא מטריקה משולבת המביאה בחשבון הן את הדיוק התחזיתי והן את הרגישות. ציון זה מאפשר להעריך את ביצועי המודל בצורה מאוזנת כאשר יש צורך להתחשב גם ברגישות וגם בדיוק התחזיתי. [2] בנוסף, Area Under the Curve (AUC) של עקומת ה- Receiver Operating Characteristic (ROC) הוא כלי חשוב להערכת ביצועי המודל. ה-AUC מודד את יכולת המודל להבחין בין מקרים של הונאה למקרים שאינם הונאה על פני טווח ערכים של סף החלטה. AUC גבוה מעיד על יכולת גבוהה של המודל לזהות נכון הונאות ולהפחית את מספר התחזיות השגויות. [3] מטריצת בלבול (Confusion Matrix) היא כלי נוסף המאפשר להעריך את ביצועי המודל על ידי הצגת התפלגות התחזיות הנכונות והשגויות של המודל. מטריצה זו מאפשרת לראות את מספר התחזיות הנכונות של הונאות ולא הונאות, וכן את מספר התחזיות השגויות של כל אחת מהקטגוריות. [3] מדד נוסף הוא Normalized Discounted Cumulative Gain, המשמש להערכת איכות הסדר של התחזיות. מדד זה בוחן כמה טוב המודל מצליח להניח את התחזיות הנכונות במקומות הראשונים. [4]

התמודדות עם Overfitting ו Underfitting

בבניית מודלים לזיהוי הונאות, חשוב להתמודד עם בעיות של Overfitting ושל Underfitting. Overfitting מתרחש כאשר המודל מתאים יותר מדי לנתוני האימון ולכן לא מצליח להכליל טוב על נתונים חדשים. אחת הטכניקות להתמודדות עם Overfitting היא שימוש ב-Dropout, דבר המפחית את מורכבות המודל על ידי הסרה אקראית של יחידות נוירון במהלך האימון. [4] Underfitting מתרחש כאשר המודל אינו מתאים מספיק לנתוני האימון ולכן לא מצליח לתפוס את הדפוסים המהותיים. שיפור המודל על ידי הוספת שכבות נוספות או הגדלת מספר הנוירונים בכל שכבה עשוי לעזור להתמודד עם Underfitting. [3] שימוש בטכניקות אלו מאפשר לשפר את ביצועי המודלים ולהבטיח שהם יכולים לזהות הונאות הכחשות עסקה בצורה אמינה ויעילה.

בהתבסס על סקר הספרות, הבנת הטכניקות והמטריקות המתקדמות לזיהוי הונאות הכחשות עסקה, בפועל יבוצעו את השלבים הבאים:

בתחום הטיפול בדאטה, המיקוד יהיה בניקוי הנתונים והסרת ערכים חסרים או לא תקינים, ולאחר מכן אשתמש בטכניקות נרמול סקלת הנתונים על מנת להבטיח שכל המאפיינים יהיו בטווח ערכים דומה. כמו כן, אשתמש בטכניקות איזון כמו SMOTE להגדלת מספר הדוגמאות מהמיעוט על מנת לשפר את ביצועי המודלים ולמנוע בעיות הנגרמות מדאטה לא מאוזן. בשלב הבא, נבחן מספר מודלים לזיהוי הונאות הכחשות עסקה, כולל רשתות נוירונים מלאכותיות (ANNs), מודלים של למידת מכונה מבוססי עץ כמו Random Forest ו-XGBoost, ומודלים היברידיים שמשלבים אלגוריתמים שונים על מנת לנצל את היתרונות של כל אחד מהם. כדי להעריך את ביצועי המודלים, נשתמש במטריקות הערכה מגוונות כמו דיוק (Accuracy), רגישות (Recall), דיוק תחזיתי (Precision), ציון F1, Area Under the Curve (AUC) של עקומת ה-Receiver Operating Characteristic (ROC) ומטריצת בלבול (Confusion Matrix). כלים אלו יספקו תמונה מקיפה של ביצועי המודלים ויכולת הזיהוי של הונאות הכחשות עסקה. בנוסף לכך, נשים דגש על התמודדות עם Overfitting ו Underfitting על ידי שימוש בטכניקות כמו Dropout להפחתת מורכבות המודל והוספת שכבות נוספות או הגדלת מספר הנוירונים בכל שכבה במידת הצורך. באמצעות שלבים אלו, נוכל לבנות מודלים יעילים ומדויקים לזיהוי הונאות הכחשות עסקה, שיסייעו לסוחרים להתמודד עם תופעה זו בצורה אפקטיבית.

6.1. ביבליוגרפיה

[1] Y. Bao, G. Hilary, and B. Ke, "Artificial intelligence and fraud detection," *Innovative Technology at the Interface of Finance and Operations: Volume I*, vol. 1, pp. 223-247, 2022. Available: Springer, https://link.springer.com/chapter/10.1007/978-3-030-75729-8_8 [Accessed: August 14, 2024].

[2] A. Abbasi, C. Albrecht, A. Vance, and J. Hansen, "Metafraud: A meta-learning framework for detecting financial fraud," *MIS Quarterly*, vol. 36, no. 4, pp. 1293-1327, 2012. Available: JSTOR, https://www.jstor.org/stable/41703508 [Accessed: August 14, 2024].

[3] U. Siddaiah, P. Anjaneyulu, Y. Haritha, and M. Ramesh, "Fraud Detection in Online Payments using Machine Learning Techniques," *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2023, pp. 268-273. Available: IEEE Xplore, https://ieeexplore.ieee.org/document/10142404 [Accessed: August 14, 2024].

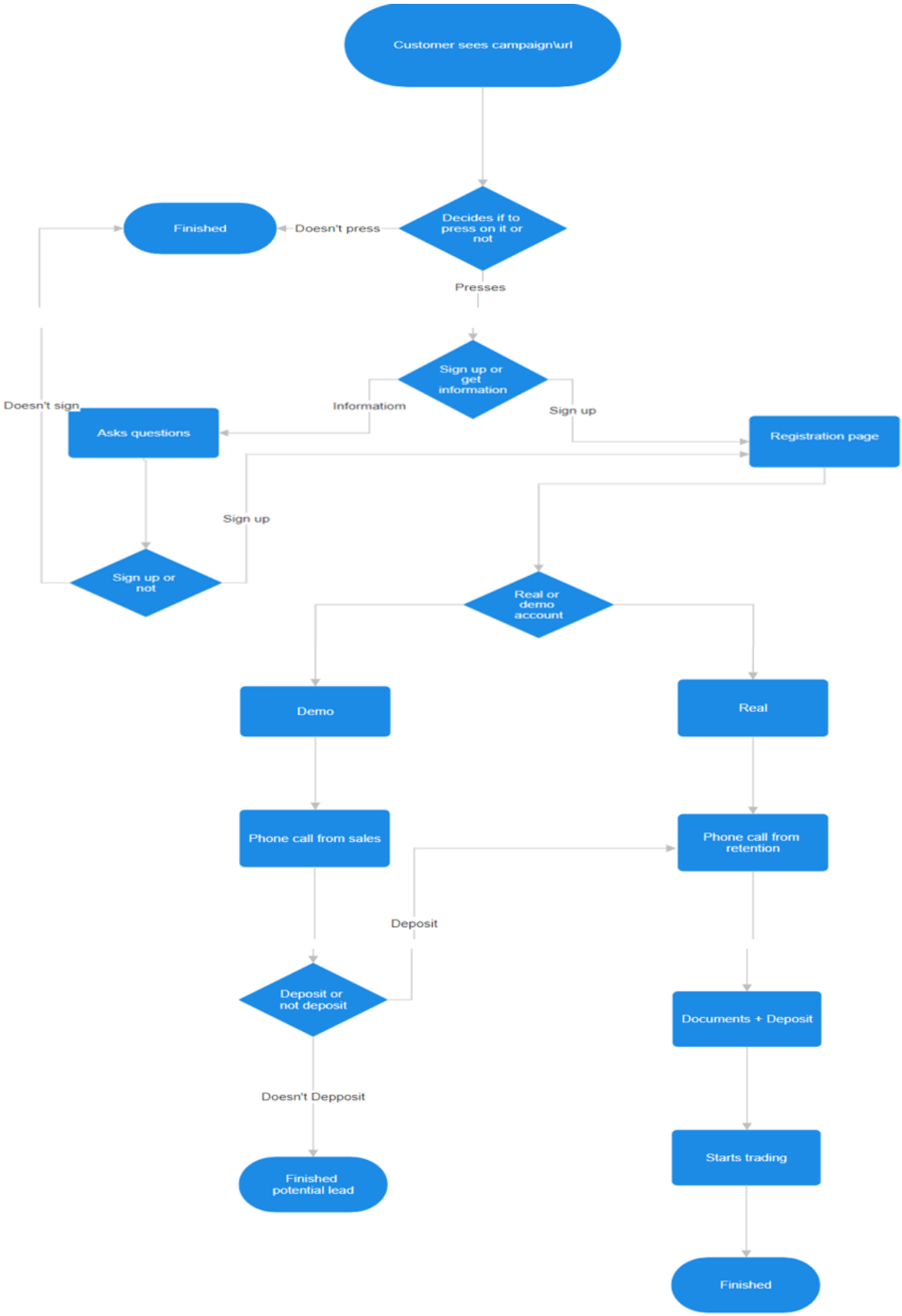
[4] T. Günther and O. Pagels-Fick, "Detecting Chargebacks in Transaction Data with Artificial Neural Networks," 2022. Available: DiVA, https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1706610&dswid=-7822 [Accessed: August 14, 2024].

[5] A. Ali, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, M. A. Nasser, and A. Saif, "Financial fraud detection based on machine learning: a systematic literature review," *Applied Sciences*, vol. 12, no. 19, p. 9637, 2022. Available: MDPI, https://www.mdpi.com/2076-3417/12/19/9637 [Accessed: August 14, 2024].

7. תיאור הנדסי של המצב הקיים

בפרק זה יוצג התיאור ההנדסי של המצב הקיים. התרשים המובא מטה מתאר את התהליך שעובר הלקוח החל מהרגע בו נחשף לקמפיין, ועד לתחילת פעילות המסחר שלו בפלטפורמת AvaTrade. התהליך מתואר בשלבים מפורטים החל מהחלטת הלקוח אם להירשם או לא, דרך סוג החשבון הנבחר (חשבון אמיתי או חשבון הדגמה), ועד לפעולות שנעשות על מנת להבטיח את הפעלת החשבון והתממשקות הלקוח עם נציגי השירות. התיאור מציג את נקודות המפתח בהחלטות שהלקוח צריך לקחת ואת הפעולות המבוצעות על ידי החברה כדי להמיר לקוח מתעניין ללקוח פעיל.

גרף 1: תרשים התיאור ההנדסי



<u>להלן המדדים בעזרתם נבחן את המודל:</u>

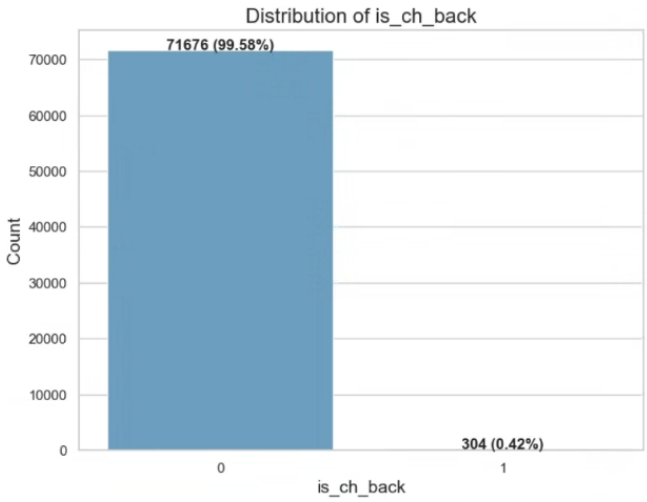
מטריקות הערכת מודלים:

מאחר והמודל עוסק בבעיית קליספיקציה, בחינת טיב המודל תתבצע באמצעות המטריקות הבאות: דיוק (Accuracy), זיכרון (Recall), ספציפיות (Specificity), דיוק משוקלל F1-Score (Precision), ו-ROC AUC.

מדדים כמותיים:

- מדד המודד את כמות הכחשות העסקה, באחוזים, עבור כל חודש. מדד זה יספק לחברת AvaTrade תמונת מצב כיום בנוגע לכמות הכחשות העסקה ובנוגע לעונתיות של הכחשות העסקה. בעתיד, לאחר שהמודל יכנס לשימוש, החברה תוכל למדוד זאת שוב ולהשוות חודש לחודש/תקופה לתקופה. יש לשים לב שההשוואה תבוצע עבור חודש/תקופה כרגע, ועבור אותו החודש/אותה התקופה בעתיד. כיום, התאריכים שיש בידי החברה, מספקים מידע על תאריך הצטרפות הלקוח ועל האם ביצע הכחשה או לא. לחברה אין מידע בנוגע לזמן ההכחשה ולזמן העסקה. בנוסף לכך, כשיהיו לחברה את תאריכי כל עסקה, החברה תוכל לקבל נתונים נוספים כגון כמות עסקאות ממוצעת לפני ביצוע הכחשת עסקה, דבר שיכול לעזור לחברה לדעת האם, לפי הסטטיסטיקות, ההכחשה היא הונאה או הכחשה בשל גניבת פרטים.
- כיום, אחוז הכחשות העסקה מתוך סך הנתונים שהתקבלו מהחברה, עומד על 0.42%. לאחר זמן מה מהכנסת המודל לשימוש בחברה, החברה תוכל למדוד זאת שוב ולהשוות.

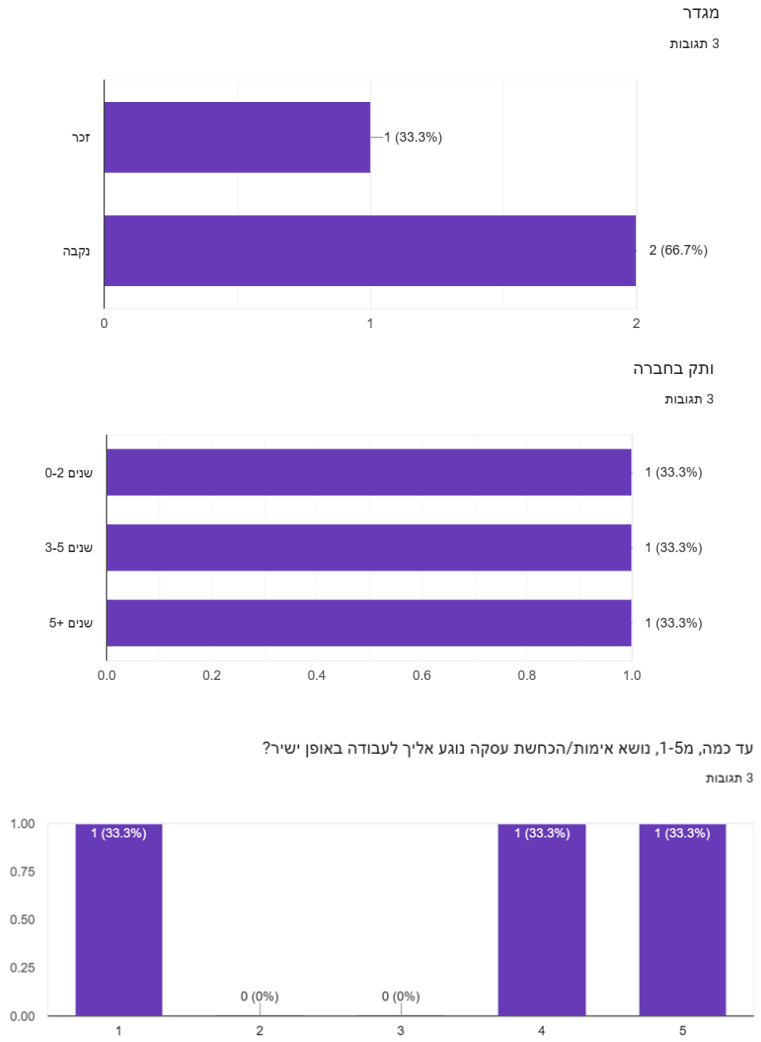
גרף 2: אחוז הכחשות העסקה מסך העסקאות



מדד איכותני:

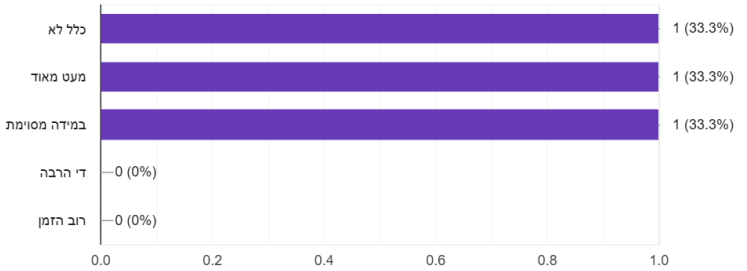
הועבר שאלון בין העובדים, על מנת שנוכל להעריך את תרומת המודל שלנו לחברה. ההערכה תבוצע על ידי השוואה בין תגובות המשיבים על הסקר לפני הכנסת המודל שלנו לארגון, לבין תגובות המשיבים על הסקר לאחר הכנסת המודל שלנו לארגון.

גרף 3: תוצאות סקר איכותני



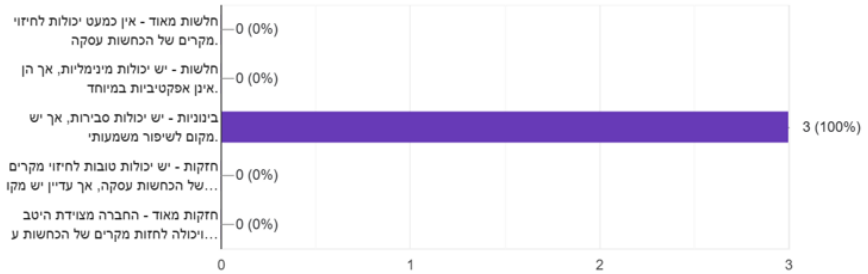
לדעתך, מ1-5, כמה מהזמן שלך בעבודה מוקדש לטיפול באימות/הכחשת עסקה?

3 תגובות



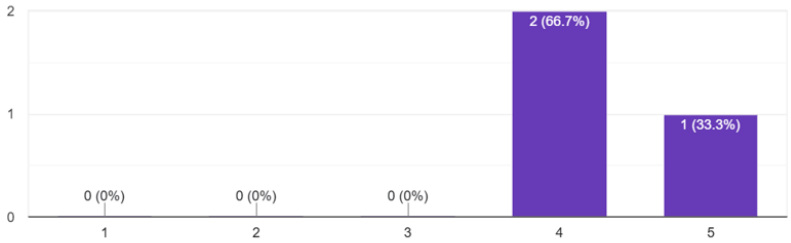
איך היית מדרג את היכולות הנמצאות כיום בחברה, לצורך חיזוי מקרים של הכחשות עסקה?

3 תגובות



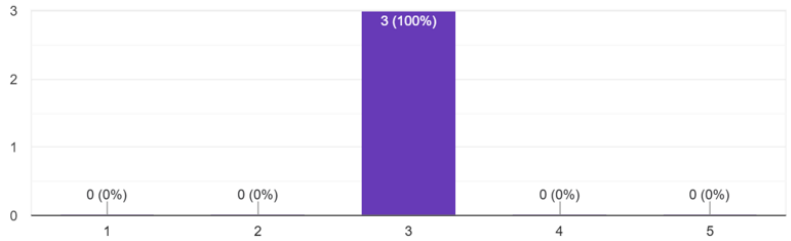
כיצד היית מדרג את הצורך בשיפור יכולות החברה, לטובת מניעת הכחשות עסקה?

3 תגובות



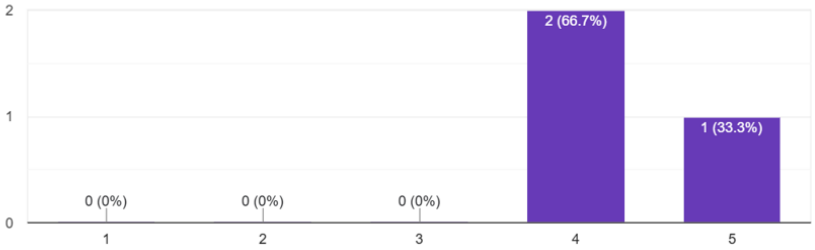
כיום, איך היית מדרג את טיב השימוש במודלים של למידת מכונה, לצורך מניעת הכחשות עסקה?

3 תגובות



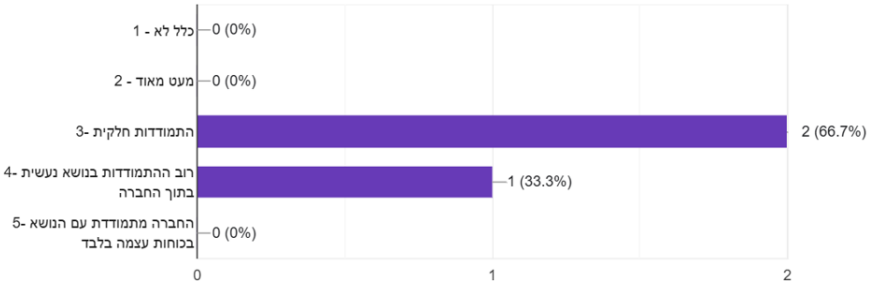
עד כמה, מ1-5, אתה מאמין ששימוש במודלים של למידת מכונה יכול ליצור שינוי מהותי בנוגע להכחשות עסקה?

3 תגובות



כמה, לדעתך, החברה מתמודדת עם הכחשות עסקה בעצמה?

3 תגובות



נשמח לשמוע מכם המלצות, הערות, הארות... :)

תודה רבה!

תגובה אחת

נושא הכחשת עסקה הוא נושא קירטי לחברה, חברות הארשאי הגדולות חכולות להטיל סנקציות על חברות שאחוז הכחשת העסקה עולה על 1% מסך העסקאות וזה נושא שמעסיק את החברה במידת מה וינם תהליכים שנעשים ברקע להתמודד עם הנושא הזה אך לא מספיק, בגלל שיש נושאים נוספים אחרים, ושימוש בלמידת מכונה יכול לסייע המון, בעיקר בגלל שניתן לזהות אלמנטים ייחודיים אצל הלקוחות שנוטים להכחיש עסקה.

8. מתודולוגיה

פרק זה נועד לתיאור השלבים שבוצעו לאורך הפרויקט. תחילה, יוצגו השלבים באופן כללי, ולאחר מכן יוצג פירוט עבור כל אחד מהשלבים. לבסוף, יוצג פירוט על אופן מימוש השלבים בקוד.

להלן הצגת השלבים השונים:

חקר מצב קיים והגדרת הבעיה: חקר המצב הקיים בעזרת: למידת הארגון, דיוק הבעיה וההשלכות של הכחשות עסקה, הצבת יעדים ומטרות ברורים.

למידת הנתונים: בחינת ולמידת הנתונים בעזרת: הצגת גרפים של התפלגויות הנתונים הצגת השוואה בין קבוצת מי שביצע הכחשת עסקה לבין מי שלא, הצגת טבלאות עם סיכומים מתמטיים.

טיפול בנתונים – טיפול והתמודדות עם: ערכים חסרים, ערכים חריגים ודאטה לא מאוזן על ידי מספר שיטות.

הרצת המודלים – בחינת מספר מודלים המתאימים לבעיית קליספיקציה:
ANN, Logistioc Regression, Decision Tree, HGB, Random Forest, AdaBoost, ו- Voting Classifier.

הערכת המודל והשיטה - שימוש במטריקות הערכה בעזרת טבלה מסכמת שכוללת : TP, TN, FP, FN , Precision, Recall, Specificity ,F1, Accuracy, Auc
• שימוש בגרפים של ROC Curve ו Feature Importances

הטמעה בארגון – העברת קוד מסודר שירוץ כל תקופה מסוימת וידרג את הסיכוי של כל לקוח להכחשת עסקה

לצורך בחירת שיטת הטיפול בנתונים הטובה ביותר עבור הנתונים הקיימים, נוצרו שלושה קבצים. בכל אחד משלושת הקבצים, נעשה טיפול מקדים בנתונים ונבחנו כל אחד מהמודלים. שלושת הקבצים כמעט זהים, כאשר ההבדל המרכזי הוא שיטת הטיפול. הבדל נוסף בוצע לאחר בחירת שיטת הטיפול והמודל שהציגו את הביצועים הטובים ביותר, תוך התחשבות בצרכי הפרויקט, נוספו גרפים נוספים על מנת להציג את הערכת המודל בצורה מעמיקה יותר.

עבור בחינת המודלים ובחירת המודל הטוב ביותר לחברה, כמו גם עבור בחירת שיטת הטיפול. השלבים בוצעו תחילה על נתוני האימון, לאחר מכן, על נתוני המבחן, כאשר נתוני האימון מהווים 80% מהנתונים הכוללים, ונתוני המבחן מהווים 20% מהנתונים הכוללים. בעזרת נתוני האימון, בוצעה

וולידציה בשיטת Repeated Stratified KFold. לבסוף, המודל הנבחר בשיטה הנבחרת, נבדק על נתוני המבחן.

<u>תיאור ועיבוד מקדים של הנתונים:</u>

נבצע הסתכלות על הנתונים, נציג גרפים להמחשתם, קורלציות בין העמודות והצגת ניתוחים סטטיסטיים (חציון, ס.תקן, ממוצע וכדומה.).

גודל הנתונים: 38 עמודות ו-71991 שורות, כאשר כל שורה מתארת לקוח.

<u>להלן רשימת העמודות מהנתונים, בהן בוצע שימוש:</u>

- **user:** מזהה ייחודי לכל לקוח.
- **deposit_attempts:** מספר הניסיונות להפקדה שבוצעו על ידי הלקוח.
- **success_deposit_attempts:** מספר ההפקדות שהצליחו מתוך כל הניסיונות להפקדה.
- **num_crads:** מספר הכרטיסים (כרטיסי אשראי) שהוזנו על ידי הלקוח.
- **num_crads_success:** מספר הכרטיסים שההפקדות בעזרתם הצליחו.
- **deposit_success_ratio**: היחס בין מספר ההפקדות שהצליחו לבין סך כל ניסיונות ההפקדה.
- **deposit_time_delta:** הפרש הזמן (בימים) בין ההפקדה הראשונה לאחרונה.
- **signup_time_delta:** הפרש הזמן (בימים) בין תאריך הרישום של הלקוח לבין תאריך ההפקדה הראשונה.
- **calls:** מספר השיחות שהלקוח קיים עם נציגי השירות.
- **docs_uploaded:** מספר המסמכים שהלקוח העלה.
- **docs_approved:** מספר המסמכים שאושרו.
- **approval_rate:** שיעור האישורים מתוך המסמכים שהועלו.
- **country_name:** שם המדינה של הלקוח.
- **channel:** הערוץ דרכו הגיע הלקוח (כגון פרסום, המלצה, וכו').
- **ava_business:** יחידת העסק של הלקוח (כגון Ava Financial, Ava Capital).
- **profile_state:** מצב הפרופיל של הלקוח (כגון אמיתי, דמו וכדומה).
- **num_closed_trades:** מספר העסקאות שסגר הלקוח.
- **business_group_name:** שם הקבוצה העסקית שהלקוח משתייך אליה.
- **is_ch_back**: האם הלקוח ביצע הכחשת עסקה (chargeback).

<u>להלן עמודות שחושבו עבור המודל:</u>

- **days_between_deposits:** עמודה זו מציגה את מספר הימים בין ההפקדה הראשונה לבין תאריך ההפקדה האחרונה של הלקוח. מטרת העמודה היא אפשור הערכת תדירות ההפקדות של הלקוח ולזהות מגמות או שינויים בהתנהגות ההפקדות לאורך זמן.

- **days_to_first_deposit:** עמודה זו מציגה את מספר הימים בין תאריך יצירת החשבון לבין תאריך ההפקדה הראשונה של הלקוח. מטרת עמודה זו היא לספק מידע על הזמן שלקח ללקוח לבצע הפקדה ראשונה לאחר רישום, מה שיכול להעיד על מידת ההתחייבות או הנכונות של הלקוח להשקיע.
- **days_between_trades:** עמודה זו מציגה את מספר הימים בין העסקה הסגורה הראשונה לבין העסקה הסגורה האחרונה של הלקוח. מטרת עמודה זו היא להציג את תדירות העסקאות של הלקוח ומסייע להבין את דפוסי המסחר שלו.
- **success_ratio_crads:** עמודה זו מציגה את העסקאות שבוצעו באמצעות כרטיסי אשראי ואושרו, ביחס לכלל העסקאות בכרטיסי אשראי. מטרת עמודה זו היא אפשור הערכת היעילות והאמינות של כרטיסי האשראי שהלקוח משתמש בהם.
- **approval_ratio_docs:** עמודה זו מציגה את המסמכים שאושרו, ביחס לכלל המסמכים שהועלו. עמודה זו משמשת כמדד לאיכות המסמכים שהלקוח מעלה והמהימנות שלו בתהליך האימות.
- **avg_days_between_deposits:** עמודה זו מציגה את הזמן הממוצע, בימים, בין ההפקדות של כל לקוח. מטרת עמודה זו היא אפשור הבנת דפוסי ההפקדות של הלקוח ולזהות אם יש תקופות של פעילות מוגברת או מואטת.
- **avg_days_between_trades:** עמודה זו מציגה את הזמן הממוצע, בימים, בין עסקאות סגורות של כל לקוח. עמודה זו משמשת כמדד לפעילות המסחר של הלקוח, ומאפשרת להעריך את קצב ואת תחלופת העסקאות.
- **first_deposit_month:** עמודה זו מציגה את החודש בו בוצעה ההפקדה הראשונה. מטרת עמודה זו היא סיפוק מידע עונתי או תקופתי על פעילות הלקוח, שיכול לשמש בניתוח מגמות עונתיות.
- **last_deposit_month:** עמודה זו מציגה את החודש בו בוצעה ההפקדה האחרונה של הלקוח. עמודה זו משמשת כמדד לפעילות העדכנית של הלקוח, ומאפשרת זיהוי שינויים בהתנהגות ההפקדות.
- **first_deposit_weekday:** עמודה זו מציגה את היום בשבוע בו בוצעה ההפקדה הראשונה. מטרת עמודה זו היא אפשור זיהוי דפוסים יומיים בפעילות הלקוח, למשל אם הלקוח נוטה להפקיד בימים מסוימים בשבוע.

<u>השלמת ערכים חסרים וטיפול בערכים חריגים</u>:

כחלק מהכנת הנתונים לשימוש במודל, בוצעו מספר פעולות על מנת לטפל בערכים חסרים ובערכים חריגים, על מנת להימנע מהטיות המודל ומחלוקה באפס.

כחלק מהטיפול בערכים חריגים בוצע, בין היתר, סינון רשומות. השלמת הערכים החסרים נעשתה בהתאם לנתוני כל עמודה. למשל, ביצוע ממוצע עבור העמודהavg_days_between_deposit , ומילוי בערך אפס עבור ערכים חסרים בעמודה success_ratio_crad, מתוך הנחה שאם חסר ערך- זה כי לא בוצעו עסקאות בכרטיסי אשראי על ידי הלקוח ברשומה זו.

לאחר מכן, בוצע חילוק של הנתונים לאימון (Train) ולמבחן (Test), כאשר חלק האימון מהווה 80% מהנתונים, וחלק המבחן מהווה 20% מהנתונים.

<u>עיבוד מאפיינים:</u>

הופעלה שיטת OneHotEncoder, שנועדה להמיר את המאפיינים הקטגוריאלים לייצוג בינארי. בנוסף, הושמטה העמודה הראשונה מכל קטגוריה, כדי למנוע מולטיקולינאריות בין העמודות. הקידוד הוגדר כך שיתעלם מערכים לא מוכרים בנתונים החדשים.

נוסף על כך, בוצע נרמול בשיטת StandardScaler, כדי לבצע אחידות למאפיינים הנומריים. פעולה זו חשובה על מנת להבטיח שכל תכונה תתרום באופן שווה למודל.

<u>פונקציות שנבנו לאימון ולהערכת המודל:</u>

פונקציית evaluate_model

פונקציה זו מחשבת את ביצועי המודל בעזרת המדדים הבאים:

Precision, Recall, Specificity, F1 Score, Accuracy ו-AUC. בנוסף, הפונקציה מחזירה את ה-FPR וה-TPR.

פונקציית draw_roc_curve_k_fold

פונקציה זו מציירת את עקומת ה-ROC לאחר ביצוע קיפול (Cross-Validation) של k קפלים. הפרמטרים שפונקציה זו מקבלת, הם רשימת TPR ו-FPR עבור כל קיפול. בנוסף, הפונקציה מקבלת AUC וthreshold, שהוגדר בברירת מחדל של 0.5, מאחר שהערכים החזויים הם בינאריים (לקוח יבצע הכחשה, או לא.).

פונקציית plot_feature_importance

פונקציה זו מציגה גרף של המאפיינים החשובים של המודל. הפרמטרים שפונקציה זו מקבלת, הם חשיבות המאפיינים הממוצעת, המודל, מספר התכונות אותן נרצה להציג ובאיזו דרך חושבו הנתונים.

פונקציית create_nn_model

פונקציה זו יוצרת ומקמפלת מודל רשת נוירונים (Neural Network). פונקציה זו מקבלת את הפרמטרים הבאים: האלגוריתם בו ישתמש המודל לעדכון משקולות, ברירת מחדל היא 'adam' ואת שיעור ההשמטה (Dropout Rate) להקטנת ה-Overfitting, כאשר ברירת מחדל היא 0.2.

פונקציית create_model

פונקציה זו בונה את המודל הנבחר, בעזרת הפרמטרים הבאים: סוג המודל, משקולות קטגוריאליים (במידה ונדרשים) ופרמטרים אופטימליים.

פונקציית search_hyperparameters

פונקציה זו מבצעת חיפוש אחר ההיפר פרמטרים האופטימליים למודלים שונים, באמצעות שיטת קיפול צולב, ונקיטת מספר צעדים חשובים לעיבוד ולהערכת הביצועים של המודל.

הפונקציה מגדירה את מדד ה-Recall כפונקציית המידוד שישמש ב-Grid Search. כלומר, אנו מחפשים פרמטרים שימקסמו את מדד ה-Recall.

GridSearchCV הוא כלי חשוב המאפשר לנו לבדוק אוטומטית שילובים שונים של פרמטרים על ידי חיפוש ברשת הפרמטרים שהוגדרה. הוא מבצע אימונים רבים ובודק כל שילוב של פרמטרים כדי למצוא את השילוב שמניב את הביצועים הטובים ביותר לפי מדד ה-Recall. השילוב האופטימלי נבחר מתוך תוצאות ה-Grid Search.

הפונקציה מגדירה רשת פרמטרים. עבור כל סוג מודל, מוגדרת רשת פרמטרים שממנה ייבחרו הפרמטרים האופטימליים באמצעות Grid Search. רשת זו כוללת שילובים שונים של פרמטרים אפשריים.

להלן רשימת היפר-הפרמטרים עבור כל מודל:

- RandomForest: מספר העצים (n_estimators), עומק מקסימלי של העצים (max_depth), ומספר מינימלי של דוגמאות לפיצול (min_samples_split).
- AdaBoost: מספר האסטימטורים (n_estimators) וקצב הלמידה (learning_rate).
- GradientBoost: מספר האסטימטורים (n_estimators), קצב הלמידה (learning_rate), ועומק מקסימלי של העצים (max_depth).
- DecisionTree: עומק מקסימלי של העצים (max_depth) ומספר מינימלי של דוגמאות לפיצול (min_samples_split).
- LogisticRegression: פרמטר הקנס (C) וסוג העונש (penalty).
- HistGradientBoost: קצב הלמידה (learning_rate), מספר האיטרציות (max_iter), ועומק מקסימלי של העצים (max_depth).
- NeuralNetwork: גודל המיני-חבילה (batch_size), מספר האפוקים (epochs), ואופטימייזר (optimizer).

הפונקציה משתמשת ב RepeatedStratifiedKFold בשלב הולדיציה כדי לוודא שהחלוקה של משתנה המטרה בכל קיפול תהיה מאוזנת כלומר בכל קיפול יהיה כמעט את אותו יחס של מספר המקרים שבהם בוצעה הכחשת עסקה בחלק שמוקצה לטסט וזה חשוב במיוחד בדאטה לא מאוזן וככה מאפשר הערכת ביצועים יציבה ומונע הטיה של הנתונים

פונקציית k_cross_validation

פונקציה זו הינה הפונקציה הראשית, אשר מבצעת קריאה לפונקציות שהוזכרו לעיל. פונקציה זו מבצעת קיפול מוצלב עם חיפוש אחר היפר-פרמטרים ואימון מחדש על כל סט האימון, היא מחזירה טבלה מסכמת עם ביצועי המודל ,וגרפים להמחשת התוצאות .

היא משמשת להערכת ביצועי מודל ולמציאת הפרמטרים האופטימליים עבורו.

<u>פרמטרים:</u>

X: תכונות הקלט (DataFrame).

y: תוויות היעד (Series).

model_type: סוג המודל (כגון RandomForest, AdaBoost, NeuralNetwork).

k: מספר הקיפולים בקיפול מוצלב (ברירת מחדל: 10).

class_weights: משקלים למחלקות, נועד לאזן את הקטגוריות בנתונים (ברירת מחדל: {0: 1540, 1: 389}).

pred_threshold: סף החיזוי לקביעת התוויות (ברירת מחדל: 0.5).

dropout_rate: שיעור ה-Dropout עבור רשתות נוירונים (ברירת מחדל: 0.2).

<u>שלבי ריצת הפונקציה:</u>

חיפוש היפר-פרמטרים

הפונקציה מתחילה בחיפוש אחר הפרמטרים האופטימליים למודל שנבחר באמצעות הפונקציה search_hyperparameters. תהליך זה כולל חיפוש ברשת הפרמטרים ובחירת הפרמטרים הטובים ביותר על פי מדד הביצוע Recall.

המדד העיקרי שנלקח בחשבון בעת ניתוח ביצועי המודל היה Recall, מאחר ש"מחיר הטעות" של החברה נמוך על זיהוי שווא כחיובי, לעומת "מחיר הטעות" שהחברה עלולה לשלם על זיהוי שווא כשלילי. במקרה של זיהוי שווא כחיובי, ה"מחיר" יהיה שיחת טלפון ללקוח החשוד, על ידי נציג אנושי, בעוד שה"מחיר" על זיהוי שווא כשלילי, עלול לגרור קנסות ומחירים רגולטוריים.

תוצאות הקיפול המוצלב

לאחר חיפוש ההיפר-פרמטרים, הפונקציה שומרת את תוצאות הקיפול המוצלב עבור כל קיפול:

ערכי TP (True Positives), TN (True Negatives), FP (False Positives), FN (False Negatives), מדדי Recall, Specificity, Precision, F1 Score, Accuracy ו-AUC ושיעורי TPR (True Positive Rate) ו-FPR (False Positive Rate).

התוצאות נשמרות בטבלה וממנה מחושבים ממוצע וסטיית התקן הכללים של מטריקות הערכה.

הדפסת עקומת ROC

הפונקציה מציירת את עקומת ה-ROC הממוצעת מהקפלים המוצלבים באמצעות הפונקציה draw_roc_curve_k_fold.

הדפסת feature_importance

הפונקציה מחזירה גרף של העמודות בעלי החשיבות המירבית למודל באמצעות הפונקציה plot_feature_importance

אימון המודל שנבחר עם ההיפר פרמטרים האופטימליים שנבחרו מחדש ובחינה פעם אחת נוספת על תוצאות החלק של הטסט מתוך החלק של האימון והחזרת תוצאות אמת ראשוניות

הפונקציה מחזירה טבלה מסכמת עם ביצועי המודל ,ההיפר פרמטרים האופטימלים,וגרפים להמחשת התוצאות .

<u>אימון המודל</u>

לאחר ניתוח מעמיק של הנתונים ובדיקת מודלים שונים, נבחר המודל HistGradientBoostingClassifier בשילוב עם Undersampling, מאחר שהתברר כי הוא מספק את התוצאות הטובות ביותר עבור המטרה שלנו. בתהליך זה התמודדנו עם בעיית חוסר האיזון בנתונים על ידי שימוש ב- RandomUnderSampler, שאפשר איזון בין הדוגמאות החיוביות והשליליות במהלך האימון.

לאחר שנבחר המודל המוביל, בוצע אימון על כל נתוני האימון בעזרת הפרמטרים האופטימליים שנמצאו. בתהליך האימון נעשה שימוש בנתונים שעברו את כל שלבי העיבוד המקדים.

<u>בחינת התוצאות</u>

לאחר האימון, בוצע ניבוי תוויות עבור סט הבדיקה והערכה מעמיקה של ביצועי המודל. ניתוח מטריצת הבלבול (Confusion Matrix) הדגיש את רמת הדיוק של המודל, והתוצאות הוצגו בצורה גרפית כדי להקל על הפרשנות. בנוסף, נוצר גרף ROC שבחן את היחס בין שיעור הנכונות החיוביות לבין שיעור התוצאות השגויות החיוביות, מה שסיפק אינדיקציה נוספת לגבי רמת הדיוק של המודל.

בהמשך, בוצעה חקירה של ההסתברויות שחושבו לכל לקוח במטרה לזהות חריגים, ויזואליזציות נוספות, כמו BoxPlot וScatterPlot, סייעו להבין את החריגות הללו. לבסוף, נעשה שימוש בטכניקת Permutation Importance כדי לקבוע את חשיבות העמודות עבור המודל, והצגת 15 העמודות החשובות ביותר בגרף מפורט. כל אלה תרמו להבנה מעמיקה של ביצועי המודל והגורמים המשפיעים על תוצאותיו, ואפשרו להציג תובנות משמעותיות לגבי המודל שנבחר.

<u>תהליך בניית המודל והפייפליין לשימוש בנתונים חדשים</u>

<u>הכנת הנתונים לצורך תהליך האימון:</u>

כתבנו פונקציה בשם prepare_data, אשר מבצעת את שלב העיבוד המקדים על עמודות הנתונים כמו המרת תאריכים, חישוב משתנים חדשים, מילוי ערכים חסרים, וטיפול בערכים אינסופיים.

לאחר הכנת הנתונים, ביצענו חלוקה של הנתונים לנתוני אימון ונתוני בדיקה באמצעות StratifiedShuffleSplit, על מנת לשמור על היחס בין הקבוצות השונות.

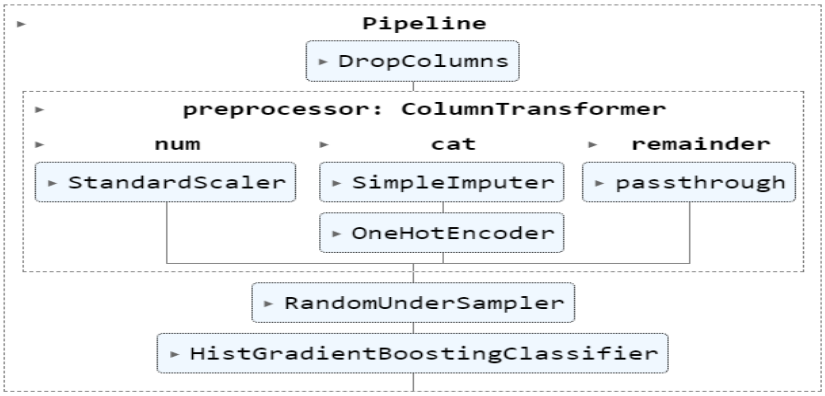
בניית הפייפליין לאימון:

שימוש ב-ColumnTransformer כדי לבנות פייפליין לעיבוד מקדים הכולל סקלינג למשתנים מספריים ו-OneHotEncoding למשתנים קטגוריים.

יצירת פייפליין מלא הכולל שלב של הסרת עמודות לא רלוונטיות, עיבוד מקדים של הנתונים, תהליך של Undersampling, ולבסוף את המודל HistGradientBoostingClassifier עם ההיפרפרמטרים הכי טובים שקיבלנו בשלב האימון.

את הפייפליין הזה אימנו על הנתונים בעזרת הפונקציה fit.

גרף 4: תרשים הפייפליין לאימון



הערכת ביצועי המודל:

לאחר האימון, השתמשנו במודל כדי לבצע תחזיות על סט הנתונים של הבדיקה.

תהליך השימוש בנתונים חדשים:

לאחר שהיינו מרוצים מביצועי המודל, הבנו שאין צורך לבצע את תהליך ה-Undersampling על נתונים חדשים, ולכן יצרנו פייפליין חדש עבור תחזיות על נתונים חדשים:

בניית פייפליין לתחזיות:

יצרנו פייפליין חדש, הנקרא inference_pipeline, המכיל את שלבי ההסרה של עמודות לא רלוונטיות, את העיבוד המקדים (שהשתמשנו בו גם באימון) ואת המודל המאומן.

הפייפליין הזה משמש אותנו כעת לתחזיות על נתונים חדשים, מבלי לבצע את ה-Undersampling.

שימוש בפייפליין לתחזיות:

בנינו פונקציה בשם preprocess_and_predict, שנועדה לטפל בנתונים חדשים ולבצע תחזיות על סמך מודל קיים. הפונקציה מקבלת שני קלטים: הנתונים החדשים שברצוננו להעריך, והפייפליין שנקרא inference_pipeline, המכיל את המודל שאחראי על החיזויים.

השלב הראשון שהפונקציה מבצעת הוא הכנת הנתונים החדשים באמצעות קריאה לפונקציה בשם prepare_data. שלב זה חיוני מכיוון שלפני שהנתונים יכולים לשמש כקלט למודל, עליהם לעבור תהליך של הכנה. תהליך זה יכול לכלול מספר פעולות כמו ניקוי נתונים, המרת משתנים לייצוגים מתאימים. מטרת ההכנה היא לוודא שהנתונים יותאמו בצורה הטובה ביותר למודל שהוכשר.

לאחר שהנתונים מוכנים, הפונקציה משתמשת בפייפליין החיזוי (inference_pipeline) כדי לחשב את ההסתברות שהמקרים שבנתונים החדשים הם מקרים של הונאה. הפייפליין הזה מכיל את המודל שלמד והותאם למטרה זו, ובשלב זה הוא מקבל את הנתונים שהוכנו ומחזיר ערכים המייצגים את הסיכוי לכך שכל מקרה יזוהה כהונאה.

השלב הבא בפונקציה הוא להוסיף את ערכי ההסתברות הללו כעמודה חדשה לנתונים המקוריים. עמודה זו נקראת Fraud_Probability, והיא מאפשרת לראות את ההערכה של המודל לכל מקרה בנפרד.

מעבר לכך, מתווספת עמודה נוספת בשם Fraud_Label, שמציינת האם המקרה נחשב כהונאה על פי ההערכה של המודל. הערכה זו מתבצעת על בסיס סף קבוע, ובמקרה הזה הסף הוא 0.5. כלומר, אם ההסתברות המחושבת עבור מקרה מסוים עולה על 50%, המקרה יסומן כהונאה.

כדי לשפר את ההבנה ולהציג את הנתונים בצורה אינטואיטיבית יותר, ההסתברויות להונאה מועברות ממצב של מספר עשרוני לאחוזים. כך, למשל, הסתברות של 0.75 תהפוך ל-75%, מה שמקל על התפיסה של רמת הסיכון לכל מקרה.

בסיום, הנתונים מסודרים מחדש לפי עמודת ההסתברות להונאה, מהגבוה לנמוך. כך, השורות שבהן ההסתברות להונאה היא הגבוהה ביותר יופיעו בראש הטבלה. לבסוף, הטבלה הסופית נשמרת כקובץ CSV בשם fraud_predictions.csv, שבו מופיעות רק העמודות החשובות שנבחרו להצגה.

הפונקציה מחזירה את הטבלה הסופית, שמכילה את כל המידע הדרוש על המקרים החדשים, כולל חיזוי הסיכויים להונאה וזיהוי המקרים שדורשים תשומת לב מיוחדת.

סיכום

על ידי הפרדה בין תהליך האימון ותהליך התחזיות, הצלחנו לבנות מודל מדויק ויעיל שמתמודד בצורה מיטבית עם בעיית חוסר האיזון בנתונים. הפייפליין הסופי שלנו מותאם באופן ייחודי לתחזיות על נתונים חדשים, ומאפשר שימוש יעיל במודל שנבנה על בסיס נתוני העבר.

9. הצגת חלופות

אנו עוסקים בקוד, ולפיכך החלופות האפשריות שנציג הן:

1. דרכי טיפול בדאטה
2. מודלים לחיזוי

כל חלופה היא שילוב של כל אחת מדרכי הטיפול בדאטה, עבור כל אחד מהמודלים.

<u>דרכי טיפול בדאטה – הכנת הנתונים:</u>

1. **הכנסת הדאטה למודל, ללא טיפול מיוחד:**

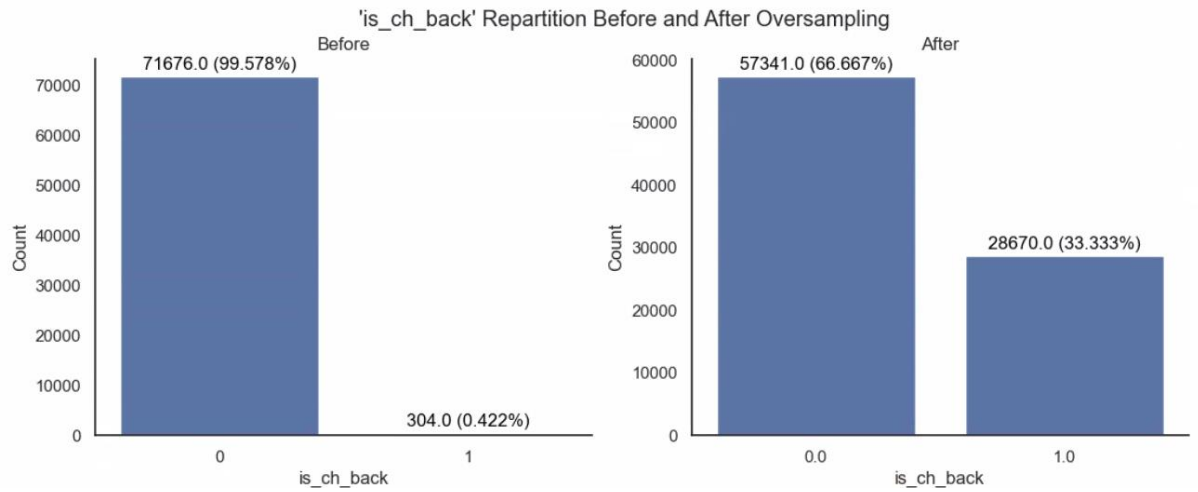
בהתאם לגישה זו, הנתונים יישארו כפי שהם, ללא איזון או טיפול מיוחד. מטרת גישה זו היא להוות קו בסיס להשוואת הגישות הבאות. על ידי שימוש בטכניקה זו, ניתן לראות את השפעת חוסר האיזון על ביצועי המודל, ולהבין את הצורך בטכניקות הבאות שיוצגו.

2. **SMOTE (Synthetic Minority Over-sampling Technique):**

טכניקה זו נועדה לאזן את הנתונים על ידי יצירת דגימות סינתטיות של מחלקת המיעוט. בטכניקה זו, עבור כל דגימה ממחלקת המיעוט נבחרים K השכנים הקרובים ביותר באמצעות מרחק אוקלידי, ובאמצעותם נוצרות דוגמאות חדשות על ידי חישוב ממוצע משוקלל בין הדוגמה המקורית לשכניה.

טכניקה זו מסייעת להגדיל את מספר הדגימות ממחלקת המיעוט מבלי לשכפל דגימות קיימות, ובכך לשפר את האיזון בין המחלקות.

גרף 5: הדגמת המצב לפני ואחרי ה-SMOTE

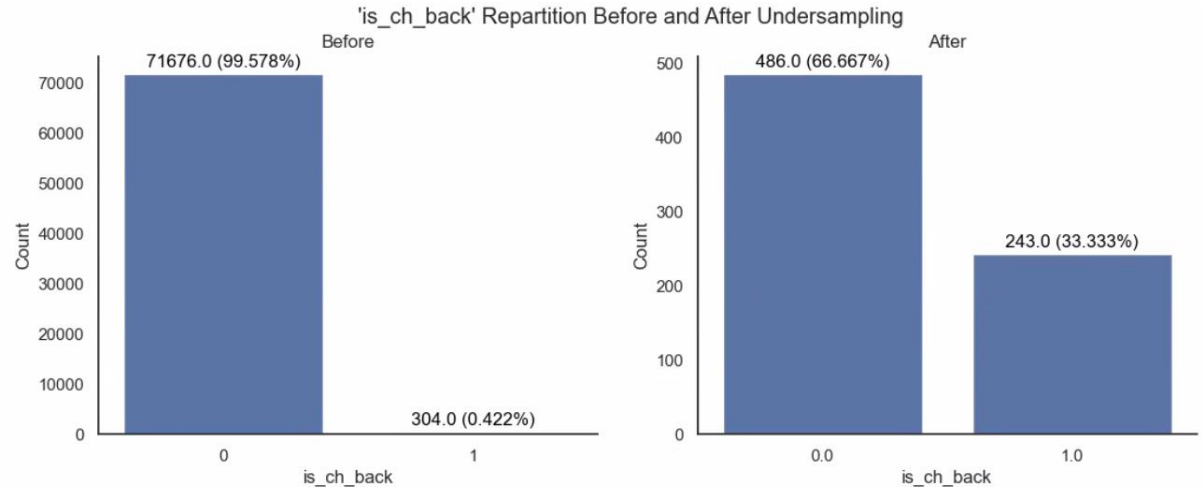


בהתאם לגרף, ניתן לראות שכדרכה של טכניקה זו, <u>הוגדלו</u> מספר הדגימות <u>ממחלקת המיעוט</u>, על ידי יצירת דוגמאות חדשות המבוססות על דוגמאות קיימות, דבר <u>שהגדיל</u> באופן יחסי לדאטה את מחלקת <u>המיעוט</u>.

3. **Under Sampling:**

בגישה זו מצמצמים את מספר הדגימות ממחלקת הרוב, על מנת לאזן את הנתונים. איזון הנתונים נעשה על ידי בחירת תת-קבוצה ממחלקת הרוב, כך שגודל מחלקת הרוב יהיה דומה לגודל מחלקת המיעוט. בעזרת הקטנת גודל מחלקת הרוב, טכניקה זו מאפשרת למודל לתת חשיבות שווה לכל מחלקה במהלך האימון. יחד עם זאת, טכניקה זו עלולה להוביל לאיבוד מידע חשוב ממחלקת הרוב, ולכן יש להשתמש בה בזהירות.

גרף 6: הדגמת המצב לפני ואחרי ה-Under Sampling



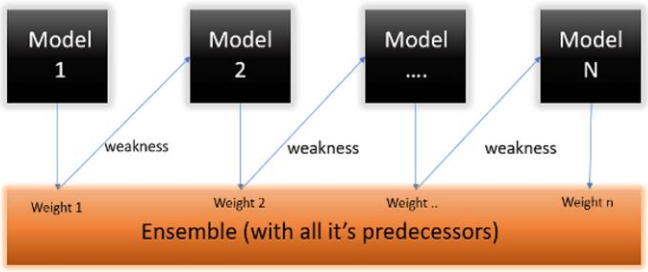
בגרף זה ניתן לראות, שכדרכה של שיטה זו, נזרקו רשומות ממחלקת הרוב (קבוצה 0- ללא הונאות), דבר שהגדיל באופן יחסי לדאטה את מחלקת המיעוט (קבוצה 1- עם הונאות), על מנת ליצור איזון יחסי בין הקבוצות.

נבחן את ששת המודלים הבאים:

1. **AdaBoost**

AdaBoost הוא מודל המשלב מספר מודלים פשוטים (Weak Learners), על מנת ליצור מודל חזק יותר. המודל פועל כך שכל מודל חדש מתקן שגיאות של המודלים הקודמים, עם דגש על דוגמאות שכשלו בעבר. השיטה עמידה בפני התאמת יתר ומשפרת את ביצועי המודלים בצורה מדורגת.

גרף 7: אופן הפעולה של מודל AdaBoost

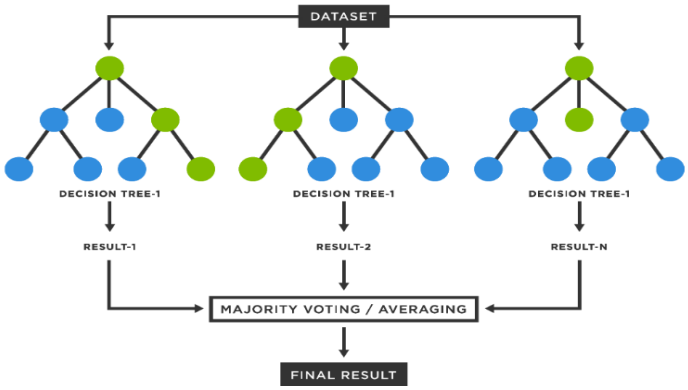


הגרף מציג את עקרון הפעולה של אלגוריתם AdaBoost. התמונה מתארת סדרה של מודלים , ... (Model 1, Model 2, Model (Model N, כאשר כל מודל מתקן את החולשות של המודל הקודם לו. כל מודל מקבל משקל (Weight) בהתאם לביצועיו, והמודלים מחוברים יחד ליצירת מודל משולב (Ensemble) המתחשב בכל המודלים הקודמים. אלגוריתם AdaBoost מחזק את הדגמים שמבצעים היטב ומשקלל את תוצאותיהם כדי לשפר את הדיוק הכולל של התחזית. גרף זה לקוח מהאתר MDPI וניתן לגשת אליו דרך הקישור: https://www.mdpi.com/2227-9717/11/3/761

2. Random Forest

Random Forest הוא מודל Ensemble המורכב ממספר רב של עצי החלטה בלתי תלויים. המודל פועל כך שכל עץ מתבסס על דגימה שונה של הנתונים ושל הפיצ'רים, והחלטת הסיווג נקבעת על פי "הצבעת רוב" של העצים. המודל מתאים למגוון רחב של בעיות סיווג ורגרסיה, ומקטין את בעיית התאמת היתר.

גרף 8: אופן הפעולה של מודל Random Forest

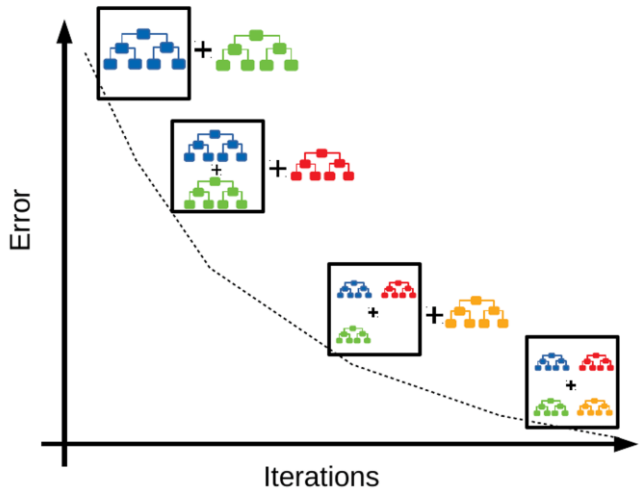


הגרף מציג את מבנה המודל Random Forest, שהוא מודל Ensemble המורכב ממספר רב של עצי החלטה בלתי תלויים. המודל עובד כך שכל עץ מתבסס על דגימה שונה של הנתונים ושל הפיצ'רים, והחלטת הסיווג או הרגרסיה מתקבלת על פי "הצבעת רוב (Majority Voting) "של כל העצים. בדרך זו, המודל מתאים למגוון רחב של בעיות סיווג ורגרסיה ומקטין את בעיית התאמת היתר שנמצאת לעיתים קרובות בשימוש בעץ החלטה יחיד. גרף זה לקוח מהאתר Medium וניתן לגשת אליו דרך הקישור: https://medium.com/@denizgunay/random-forest-af5bde5d7e1e

3. Gradient Boosting

Gradient Boosting הוא מודל Boosting שמשפר את ביצועי המודל על ידי בניית מודלים חדשים שמתקנים שגיאות של המודלים הקודמים. המודל פועל כך שהוא מתחיל ממודל בסיסי ונמשך באופן מדורג, כאשר כל מודל מנסה לשפר את התחזית של המודלים הקודמים על ידי חיזוי השגיאות (השאריות) שהם השאירו. השיטה גמישה ובעלת ביצועים גבוהים.

גרף 9: אופן הפעולה של מודל Gradient Boosting

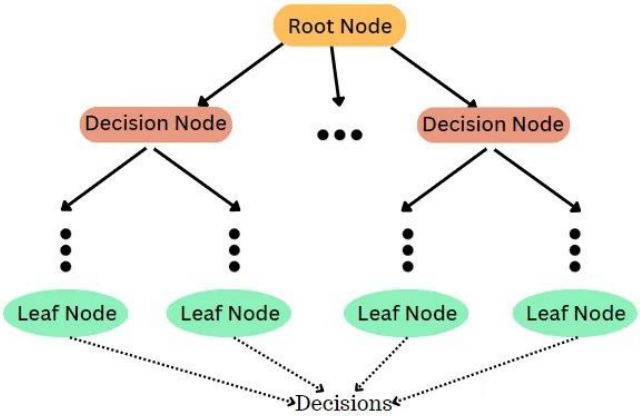


הגרף מציג את תהליך ה-Gradient Boosting בצורה גרפית, שבו כל מודל עוקב מתווסף במטרה להפחית את השגיאה שנשארה מהממודלים הקודמים. הגרף מציג את הירידה בשגיאה (Error) לאורך מספר האיטרציות (Iterations), כאשר כל תוספת של מודל חדש (מיוצגת על ידי עצים בצבעים שונים) תורמת לשיפור הדיוק הכללי של התחזיות. גרף זה לקוח מהאתר AlmaBetter וניתן לגשת אליו דרך הקישור:
https://www.almabetter.com/bytes/tutorials/data-science/gradient-boosting

4. Decision Tree

Decision Tree הוא מודל סיווג, שכל צומת בו מייצגת בדיקת תכונה, וכל ענף מייצג תוצאת בדיקה אפשרית. כל עלה בעץ מייצג תוצאה סופית של סיווג. המודל פשוט להבנה ולפרשנות ומתאים לבעיות סיווג עם יחסי גומלין לא ליניאריים בין הפיצ'רים. ניתן לשלוט בעומק העץ ובמספר הצמתים כדי למנוע התאמת יתר.

גרף 10: אופן הפעולה של מודל Decision Tree

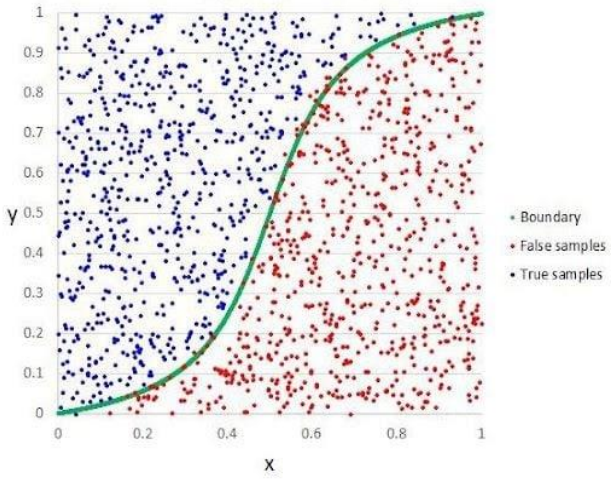


הגרף מציג מבנה בסיסי של עץ החלטה. בעץ ההחלטה ישנו צומת שורש (Root Node) שממנו מתחיל תהליך קבלת ההחלטות. מהצומת הראשי יוצאים מספר צמתי החלטה (Decision Nodes), שמפצלים את הנתונים על פי קריטריונים שונים. הפיצולים נמשכים עד להגעה לצמתי עלים (Leaf Nodes), שמייצגים את ההחלטות הסופיות. כל צומת עלה מכיל את תוצאת הסיווג או התחזית הסופית, לאחר שעבר דרך הצמתים הקודמים. גרף זה לקוח מהאתר Medium וניתן לגשת אליו דרך הקישור:
https://medium.com/@nidhigh/decision-trees-a-powerful-tool-in-machine-learning-dd0724dad4b6

5. Logistic Regression

Logistic Regression הוא מודל ליניארי הממיר קלט להסתברות של סיווג בינארי, באמצעות פונקציית הסיגמואיד. המודל מתאים במיוחד לבעיות סיווג בינארי ויכול להתמודד עם נתונים ליניאריים ולא ליניאריים. הוא מהיר, פשוט, קל להבנה, ומספק תוצאות טובות כאשר המחלקות ניתנות להפרדה ליניארית.

גרף 11: המחשת תוצאות על פי מודל Logistic Regression

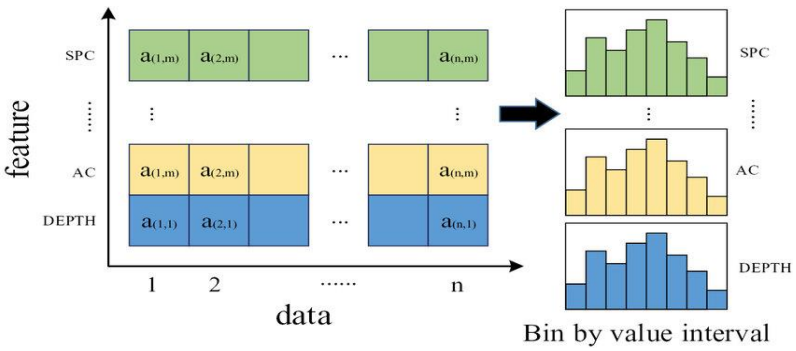


הגרף ממחיש את פעולת מודל Logistic Regression בסיווג נתונים, תוך הצגת הגבול בין המעמד החיובי (אדום) לשלילי (כחול). הקו הירוק מציין את גבול ההחלטה של המודל, כאשר נתונים משמאלו מסווגים כשליליים ומשמאלו כמתארים חיוביים. הגרף מדגיש את אתגרי המודל במקרים של חפיפה בין המעמדות, המובילים לשגיאות סיווג. גרף זה לקוח מ-LinkedIn וניתן לגשת אליו דרך הקישור: https://www.linkedin.com/pulse/logistic-regression-anil-mahanty/

6. Hist Gradient Boosting

Hist Gradient Boosting מאיץ את תהליך הבניה של המודלים החלשים על ידי סיכום הנתונים בהיסטוגרמות, מה שמפחית את זמן החישוב ואת דרישות הזיכרון. המודל מתאים במיוחד למערכי נתונים גדולים מאוד, תוך שמירה על דיוק גבוה ושיפור ביצועים. המודל מתמודד עם נתונים מסובכים ומחלקות לא מאוזנות, ומספק גמישות ואופטימיזציה גבוהה בתהליך האימון.

גרף 12: אופן הפעולה של מודל Hist Gradient Boosting



הגרף מציג את תהליך הבניה של היסטוגרמות עבור כל תכונה (feature) במערך נתונים. בשלב הראשון, הנתונים המקוריים לכל

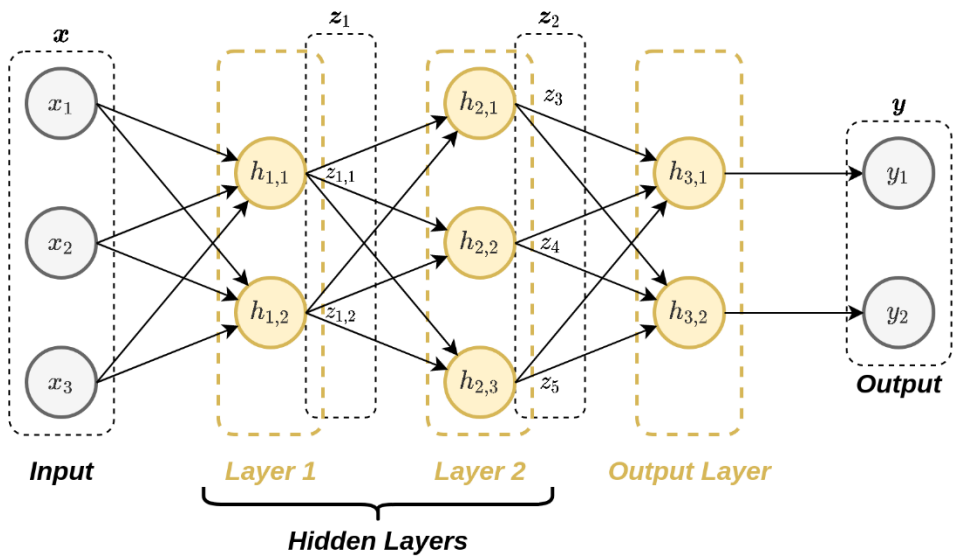
תכונה מחולקים לערכים שונים ומאורגנים בטבלאות. לאחר מכן, לכל תכונה נבנות היסטוגרמות שמחלקות את הערכים לאינטרוולים (bins). תהליך זה עוזר להפחית את זמן החישוב ואת דרישות הזיכרון על ידי סיכום הנתונים בהיסטוגרמות, והוא משמש במודלים כמו Hist Gradient Boosting כדי לשפר ביצועים ולשמור על דיוק גבוה. גרף זה לקוח מהאתר ResearchGate וניתן לגשת אליו דרך הקישור:

https://www.researchgate.net/figure/Principle-of-histogram-based-decision-tree-algorithm_fig4_336227458

7. ANN

רשתות נוירונים מלאכותיות הן טכניקת עיבוד מידע המחקה את התנהגות הרשתות הנוירוניות הביולוגיות. הן מורכבות מצמתים (נוירונים מלאכותיים) המחוברים ביניהם בקשרים משוקללים, אשר יכולים לעבד מידע ולהעבירו הלאה בשכבות הרשת. תהליך הלמידה ברשתות אלה מתבצע באמצעות עדכון משקלי הקשרים על סמך הנתונים המוזנים לרשת, בתהליך הנקרא אימון. רשתות נוירונים משמשות למגוון רחב של תחומים כמו זיהוי תמונה וקול, עיבוד שפה טבעית, חיזוי נתונים, ואוטומציה של תהליכים שונים. מודלים מתקדמים יותר כמו רשתות נוירונים המותאמות לעלויות שגיאה שונות וללמידת חיזוק עמוקה, מאפשרים שיפורים משמעותיים בביצועים, ומפחיתים שגיאות במגוון רחב של יישומים.

גרף 13: אופן הפעולה של ANN



הגרף מציג רשת עצבית מלאכותית (Artificial Neural Network - ANN) עם שכבת קלט, שתי שכבות חבויות ושכבת פלט. רשת זו מקבלת שלושה קלטים (x_1, x_2, x_3), מעבירה אותם דרך שתי שכבות חבויות $h_{1,1}, h_{1,2}, h_{2,1}, h_{2,2}, h_{2,3}$, ומפיקה שני פלטים (y_1, y_2). כל נוירון בשכבה מחובר לכל הנוירונים בשכבה הבאה, כך שכל קלט משפיע על כל הנוירונים בשכבות הבאות עד פלט הסופי.

8. Voting Classifier

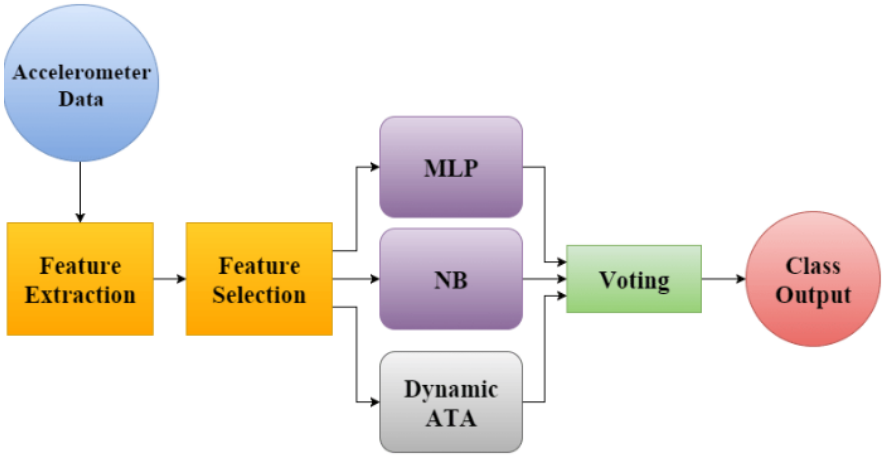
Voting Classifier הוא מודל היברידי. מודל זה משתמש בגישה המשלבת מספר אלגוריתמים של למידת מכונה, על מנת לשפר את ביצועיו הכוללים של המודל.

בפרויקט זה המודל ישלב את המודלים הבאים:

a. Random Forest Classifier
b. HistGradientBoostingClassifier
c. Logistic Regression

המודל ההיברידי משתמש בכוחם המשותף של שלושת המודלים שצוינו לעיל, על מנת לשפר את דיוק התחזיות, להקטין שגיאות, ולהגביר את יציבות התחזיות. היתרון המרכזי של מודל זה הוא היכולת לקחת את היתרונות היחסיים של כל מודל משולב, על מנת להשיג תוצאות טובות יותר בהשוואה לשימוש במודל יחיד.

גרף 14: אופן הפעולה של המודל ההיברידי Voting Classifier



הגרף מתאר את אופן הפעולה של מודל Voting Classifier היברידי, המשתמש בנתוני האקסלרומטר. הנתונים עוברים שלבי חילוץ ובחירת מאפיינים, שלאחריהם נעשה שימוש בשלושה מודלים (MLP, NB, Dynamic ATA) לקבלת תחזיות. לבסוף מתקיים תהליך הצבעה המשלב את התחזיות של המודלים ליצירת פלט הסיווג הסופי. גרף זה לקוח מהאתר ResearchGate וניתן לגשת אליו דרך הקישור: https://www.researchgate.net/figure/Hybrid-Learning-Model-with-Two-Off-Line-Machine-Learning-Models-Purple-and-One-On-Line_fig3_307968015

10. מימוש הפתרון

שלבי הפתרון בוצעו תחילה על נתוני האימון. לאחר מכן, על נתוני המבחן, כאשר נתוני האימון מהווים 80% מהנתונים הכוללים, ונתוני המבחן מהווים 20% מהנתונים הכוללים. בעזרת נתוני האימון, בוצעה וולידציה בשיטת Repeated Stratified KFold. לבסוף, המודל הנבחר בשיטה הנבחרת, נבחן על נתוני המבחן.

המודל הנבחר מקבל מקרים חדשים ומבצע עליהם תחזיות לסיווג מקרים כהכחשת עסקה או לא, תוך התמודדות עם חוסר איזון בנתונים.

נבנו שלושה קבצים נפרדים, כאשר בכל קובץ נעשה שימוש בשיטת טיפול שונה להתמודדות עם הנתונים הלא מאוזנים, ובה בחנו כל אחד מהמודלים. הפרדה זו נעשתה על מנת לבדוק איזו שיטה תוביל לתוצאות הטובות ביותר.

עבור כל אחד מהקבצים, נעשו השלבים הבאים:

<u>עיבוד מקדים של הנתונים:</u>

תחילה, בוצע ניתוח מקיף של הנתונים, כולל הצגת סטטיסטיקות, גרפים, וקורלציות בין העמודות השונות, כדי להבין את מבנה הנתונים ולזהות דפוסים חשובים. לאחר מכן, חושבו ונוספו עמודות חדשות. עמודות אלו בעלות חשיבות לניתוח והבנה מעמיקה יותר של התנהגות הלקוחות.

בנוסף, טופלו ערכים חסרים וערכים חריגים. כחלק מטיפול זו, סוננו רשומות חריגות, על מנת למנוע מרשומות אלה לגרום להטיית המודל.

<u>חלוקת הנתונים לאימון ולמבחן:</u>

הנתונים חולקו לסטים של אימון ושל מבחן באמצעות שיטת Stratified Shuffle Split, השומרת על יחס מאוזן בין הקבוצות השונות ומבטיחה שמודל האימון יבוסס על נתונים מגוונים ורלוונטיים.

<u>בניית פייפליין לעיבוד ואימון המודל:</u>

נעשה שימוש בכלי ColumnTransformer, לצורך בניית פייפליין לעיבוד מקדים. פייפליין זה כלל ביצוע שינוי קנה המידה של הנתונים הנומריים, והבאתם לטווח מסויים (Data Scaling). בנוסף, הפייפליין כולל קידוד של תכונות קטגוריאליות באמצעות One Hot Encoder. כמו כן, הפייפליין כלל תהליך של Under Sampling לשם איזון הדוגמאות החיוביות והשליליות, דבר התורם לשיפור הביצועים של המודל. לסיום, שולב בפייפליין המודל Hist Gradient Boosting Classifier, שנבחר לאחר בדיקות מקיפות ונמצא כמתאים ביותר למטרות הפרויקט. המודל עבר אופטימיזציה לפרמטרים החשובים ביותר בעזרת חיפוש רשת (Grid Search), דבר המבטיח שהפרמטרים שהתקבלו הם המתאימים ביותר לאימון המודל.

<u>הערכת ביצועי המודל:</u>

לאחר האימון, המודל נבחן על סט הנתונים של הבדיקה, וחושבו מגוון מדדי ביצועים: Precision, Recall, Specificity, F1-score, Accuracy ו-AUC. מדדים אלו נועדו לוודא שהמודל מצליח לזהות בצורה טובה, יחסית, מקרים של הכחשת עסקה. כמו כן, נוצרו גרפים וטבלאות להצגה ויזואלית של הנתונים.

במודל נעשה שימוש בסף חיזוי (Threshold) של 0.5, כדי להחליט אם מקרה מסוים יסווג כהכחשת עסקה. המשמעות היא, שאם ההסתברות המחושבת על ידי המודל עבור מקרה מסוים הייתה מעל 50%, המקרה סומן כהכחשת עסקה. לפיכך, בבחינת החריגים כהכחשה או לא, נעשתה הסתכלות על הסתברות מתחת לחצי ומעל לחצי, בהתאמה.

<u>בניית פונקציות לביצוע וולידציה ולהערכת תוצאות המודל:</u>

- draw_roc_curve_k_fold: פונקציה זו מציירת את עקומת ה-ROC הממוצעת לאחר ביצוע קיפול צולב (Cross-Validation) של k קפלים. הפונקציה מאפשרת להציג את ביצועי המודל בצורה גרפית, כולל סטיית התקן של הביצועים.
- plot_feature_importance: פונקציה זו מציגה גרף של המאפיינים החשובים (Feature Importance) למודל, ומאפשרת להבין אילו תכונות הכי השפיעו על תחזיות המודל.
- evaluate_model: פונקציה זו מעריכה את ביצועי המודל באמצעות חישוב מדדים שונים, ומחזירה אותם יחד עם שיעורי חיוביים אמיתיים ושקריים לכל קיפול.
- search_hyperparameters: פונקציה זו מבצעת חיפוש אחר הפרמטרים האופטימליים למודלים שונים, תוך שימוש בשיטת קיפול צולב (Cross-Validation), ומבטיחה שהמודל יעבור אופטימיזציה מלאה.

<u>בניית פייפליין לתחזיות על נתונים חדשים:</u>

לאחר שהוצגו תוצאות משביעות רצון מהמודל, נבנה פייפליין חדש, inference_pipeline, שכלל את שלבי העיבוד המקדים ואת המודל המאומן, אך ללא שלב ה-Undersampling, שאינו נחוץ לנתונים חדשים.

נבנתה פונקציה בשם preprocess_and_predict, שמכינה את הנתונים החדשים, מחשבת את ההסתברויות להכחשת עסקה, מוסיפה עמודות המצביעות על ההסתברות להונאה ועל הזיהוי שלה, ומסדרת את הנתונים לפי סדר יורד של ההסתברות להכחשה. פונקציה זו מאפשרת לבצע תחזיות בצורה יעילה ומהירה, ומסייעת בזיהוי מקרים בעלי פוטנציאל להכחשת עסקה בצורה ברורה ומדויקת.

תוצאות חיזוי והשפעת שיטות הטיפול השונות על ביצועי המודלים:

טבלה 1: דוגמה להצגת הנתונים שנתקבלו במסגרת החיזוי

	user	conversion_owner	tag	country_name	business_group_name	account_create_on_date	Fraud_Probability	Fraud_Label
0	14669433	AvaFX Customer Service	185134	Colombia	Ava Financial	2023-03-09	0.000	False
1	15071430	AD Unassigned	avafx	Brazil	Ava Financial	2023-06-26	99.984	True
2	15946975	Giuseppe Moscato	avafx	Italy	Ava Capital	2024-01-04	0.000	False
3	16530149	AvaFX Customer Service	fusionpartners_google_za_search_nonbrand	South Africa	Ava Financial	2024-07-23	0.000	False
4	14617898	AvaFX Customer Service	126295	Mexico	Ava Financial	2023-02-15	0.001	False
...	...	...	...	...	...	...	...	...
14391	14618811	Daniel Krastev	2195	South Africa	Ava Financial	2023-02-15	0.000	False
14392	16046291	S Adler	188233	Austria	Ava Capital	2024-02-04	0.000	False
14393	16338307	Tsvetan Metodiev Petrov	avafx	Bangladesh	Ava Financial	2024-05-14	0.000	False
14394	15786228	Teddy Orenstein	62206	Germany	Ava Capital	2023-11-13	0.322	False
14395	14574828	Eran Atia	avafx	Israel	ATrade	2023-01-26	0.000	False

14396 rows × 8 columns

בטבלה זו ניתן לראות את תוצאות החיזוי של המודל. עמודת Fraud_Probability מציגה את ההסתברות המשוערת להכחשת עסקה, בעוד שעמודת Fraud_Label מציינת את ההחלטה הסופית של המודל בהתבסס על הסתברות זו.

טבלה 2: התוצאות של המודלים השונים לאחר הטיפול בנתונים באמצעות Under-sampling

Metric	AdaBoost	DecisionTree	HistGradientBoost	LogisticRegression	RandomForest	Sequential_NN	Voting_CR	Mean_SD_Summary
TN	(96.0, 1.225)	(94.8, 1.095)	(96.0, 1.225)	(95.4, 1.342)	(96.2, 0.837)	(94.8333, 2.324)	(239.5, 2.121)	(116.105, 0.552)
TP	(47.2, 1.095)	(47.0, 1.0)	(47.2, 1.095)	(47.4, 0.894)	(47.6, 1.14)	(45.7778, 5.495)	(118.0, 1.414)	(57.168, 1.666)
FN	(1.2, 1.643)	(2.4, 1.342)	(1.2, 1.643)	(1.8, 1.789)	(1.0, 0.707)	(2.3667, 2.382)	(3.5, 2.121)	(1.924, 0.542)
FP	(1.4, 1.14)	(1.6, 0.548)	(1.4, 1.14)	(1.2, 0.447)	(1.0, 1.0)	(2.8222, 5.476)	(3.5, 2.121)	(1.846, 1.754)
AUC	(0.9812, 0.017)	(0.9712, 0.01)	(0.987, 0.011)	(0.9954, 0.004)	(0.9958, 0.005)	(0.9918, 0.005)	(0.9957, 0.003)	(0.988, 0.005)
Accuracy	(0.9822, 0.015)	(0.9726, 0.011)	(0.9822, 0.015)	(0.9794, 0.015)	(0.9863, 0.011)	(0.9644, 0.038)	(0.9808, 0.012)	(0.978, 0.01)
F1-score	(0.9733, 0.022)	(0.9592, 0.016)	(0.9733, 0.022)	(0.9694, 0.022)	(0.9794, 0.016)	(0.9409, 0.105)	(0.9712, 0.017)	(0.967, 0.033)
Precision	(0.9758, 0.033)	(0.9517, 0.027)	(0.9758, 0.033)	(0.964, 0.035)	(0.9794, 0.014)	(0.9425, 0.108)	(0.9712, 0.017)	(0.966, 0.032)
Recall	(0.9713, 0.023)	(0.967, 0.011)	(0.9713, 0.023)	(0.9753, 0.009)	(0.9794, 0.021)	(0.9419, 0.112)	(0.9712, 0.017)	(0.968, 0.036)
Specificity	(0.9877, 0.017)	(0.9753, 0.014)	(0.9877, 0.017)	(0.9815, 0.018)	(0.9897, 0.007)	(0.9757, 0.024)	(0.9856, 0.009)	(0.983, 0.006)

מטבלה זו ניתן להסיק כי שיטת הטיפול Under-sampling, על פי מדד הRecall, נותנת תוצאות טובות מאוד.
ניתן לשים לב כי לפי שיטה זו, אין הבדל רב בין המודלים השונים.

טבלה 3: תוצאות המודלים לאחר הטיפול בנתונים באמצעות SMOTE

Metric	AdaBoost	DecisionTree	HistGradientBoost	LogisticRegression	RandomForest	Sequential_NN	Voting_CR	Mean_SD_Summary
TN	(11404.8, 7.918)	(11420.4, 9.236)	(11445.2, 5.718)	(11397.4, 7.127)	(11445.6, 4.879)	(11421.0333, 17.34)	(28571.0, 5.657)	(13872.205, 4.271)
TP	(5723.4, 4.827)	(5719.8, 12.296)	(5730.2, 1.304)	(5720.6, 6.309)	(5729.8, 2.387)	(5732.0889, 3.179)	(14331.5, 0.707)	(6955.341, 3.978)
FN	(63.4, 7.635)	(47.8, 8.871)	(23.0, 5.385)	(70.8, 6.797)	(22.6, 4.615)	(47.1667, 17.26)	(3.5, 0.707)	(39.752, 5.105)
FP	(10.6, 4.827)	(14.2, 12.296)	(3.8, 1.304)	(13.4, 6.309)	(4.2, 2.387)	(1.9111, 3.179)	(99.5, 4.95)	(21.087, 3.623)
AUC	(0.9996, 0.0)	(0.9973, 0.0)	(1.0, 0.0)	(0.9985, 0.0)	(1.0, 0.0)	(0.9988, 0.001)	(1.0, 0.0)	(0.999, 0.0)
Accuracy	(0.9957, 0.001)	(0.9964, 0.0)	(0.9984, 0.0)	(0.9951, 0.001)	(0.9984, 0.0)	(0.9971, 0.001)	(0.9976, 0.0)	(0.997, 0.001)
F1-score	(0.9936, 0.001)	(0.9946, 0.001)	(0.9977, 0.001)	(0.9927, 0.001)	(0.9977, 0.0)	(0.9957, 0.002)	(0.9964, 0.0)	(0.995, 0.001)
Precision	(0.989, 0.001)	(0.9917, 0.002)	(0.996, 0.001)	(0.9878, 0.001)	(0.9961, 0.001)	(0.9918, 0.003)	(0.9931, 0.0)	(0.992, 0.001)
Recall	(0.9982, 0.001)	(0.9975, 0.002)	(0.9993, 0.0)	(0.9977, 0.001)	(0.9993, 0.0)	(0.9997, 0.001)	(0.9998, 0.0)	(0.999, 0.001)
Specificity	(0.9945, 0.001)	(0.9958, 0.001)	(0.998, 0.0)	(0.9938, 0.001)	(0.998, 0.0)	(0.9959, 0.002)	(0.9965, 0.0)	(0.996, 0.001)

מטבלה זו ניתן להסיק כי שיטת הטיפול בנתונים SMOTE, על פי מדד הRecall, מספקת תוצאות טובות "מידי", המבטאות Overfitting על הנתונים. אנו רוצים להימנע מOverfitting, מאחר שמודל שנותן תוצאות טובות "מידי", עלול לא לייצג נאמנה את המציאות. מסיבה זו, גם כאן נסתכל על מודל הHGB, שנותן את הממוצע הנמוך ביותר, אך גם הוא עדיין גבוה מידי.

טבלה 4: תוצאות המודלים ללא טיפול מיוחד בנתונים

Metric	AdaBoost	DecisionTree	HistGradientBoost	LogisticRegression	RandomForest	Sequential_NN	Voting_CR	Mean_SD_Summary
TN	(11453.4, 6.58)	(11445.2, 4.919)	(11443.6, 12.681)	(11453.6, 5.595)	(11465.0, 1.871)	(11466.4556, 4.44)	(28644.5, 16.263)	(13910.251, 5.096)
TP	(25.4, 5.225)	(26.0, 3.808)	(30.6, 2.881)	(22.2, 3.271)	(14.8, 2.28)	(2.6667, 5.746)	(64.5, 20.506)	(26.595, 6.409)
FN	(14.8, 6.34)	(23.0, 4.743)	(24.6, 12.361)	(14.6, 5.177)	(3.2, 1.483)	(1.7444, 4.349)	(57.0, 19.799)	(19.849, 6.257)
FP	(23.2, 5.215)	(22.6, 4.278)	(18.0, 3.24)	(26.4, 3.209)	(33.8, 2.683)	(45.9333, 5.776)	(26.0, 15.556)	(27.99, 4.486)
AUC	(0.9937, 0.003)	(0.7668, 0.042)	(0.9208, 0.052)	(0.9904, 0.008)	(0.9942, 0.006)	(0.9844, 0.052)	(0.9943, 0.001)	(0.949, 0.024)
Accuracy	(0.9967, 0.0)	(0.996, 0.001)	(0.9963, 0.001)	(0.9964, 0.0)	(0.9968, 0.0)	(0.9959, 0.0)	(0.9971, 0.0)	(0.996, 0.0)
F1-score	(0.5682, 0.066)	(0.532, 0.065)	(0.5946, 0.062)	(0.518, 0.032)	(0.4439, 0.064)	(0.0754, 0.148)	(0.6007, 0.089)	(0.476, 0.036)
Precision	(0.6421, 0.074)	(0.5319, 0.06)	(0.5782, 0.126)	(0.6157, 0.074)	(0.8203, 0.093)	(0.1899, 0.321)	(0.7251, 0.062)	(0.586, 0.093)
Recall	(0.5226, 0.108)	(0.5356, 0.084)	(0.63, 0.064)	(0.4567, 0.067)	(0.3048, 0.049)	(0.0549, 0.118)	(0.5304, 0.166)	(0.434, 0.04)
Specificity	(0.9987, 0.001)	(0.998, 0.0)	(0.9979, 0.001)	(0.9987, 0.0)	(0.9997, 0.0)	(0.9998, 0.0)	(0.9991, 0.001)	(0.999, 0.001)

מטבלה זו ניתן להסיק כי ללא טיפול בנתונים, הRecall נמוך מאוד ולא נותן תוצאות מספקות.
בנוסף, ניתן לראות שאפילו בשיטת החיזוי הזו, מודל הHGB מספק את התוצאות הטובות ביותר.

לסיכום, לאחר ביצוע תהליך מקיף של עיבוד נתונים מקדים, אימון ובחינת המודלים השונים, ניתן להסיק כי שיטת הטיפול בנתונים משפיעה באופן משמעותי על ביצועי המודלים ועל היכולת שלהם לזהות מקרים של הכחשת עסקה. שיטת ה-Under-sampling נמצאה כאפקטיבית במיוחד, כשהיא מספקת תוצאות טובות ומאוזנות ללא נטייה לOverfitting. לעומת זאת, שיטת ה-SMOTE הראתה ביצועים טובים מדי, מה שמעיד על סיכון לOverfitting, ואילו הנתונים ללא טיפול מיוחד הובילו לתוצאות נמוכות מדי. בבחינת שלוש השיטות המוזכרות, מודל ה-HistGradientBoosting התגלה כיציב ומדויק, דבר המקנה לו יתרון משמעותי בבחירת המודל הסופי לפרויקט זה.

בסופו של דבר, שילוב מודל HGB עם שיטת ה-Under-sampling, נבחר כפתרון האופטימלי.

11. הערכת הפתרון

נבחרו שיטת הטיפול Under-sampling על הנתונים, והמודל Hist Gradient Boosting. לצורך בחינת התוצאות של חלק המבחן, יבוצע שימוש במספר מדדים סטטיסטיים, אשר לפיהם הוערך כל אחד מהמודלים, עבור כל אחת מהשיטות. לאחר מכן, תבוצע השוואה באמצעים וויזואלים הבאים: ROC ו-Confusion Matrix. לבסוף, יוצג הסבר מפורט על תוצאות הערכת הפתרון עבור כל אחד מהמודלים וכל אחת מהשיטות, אשר יבסס את המסקנות שלנו.

<u>מדדים סטטיסטיים:</u>

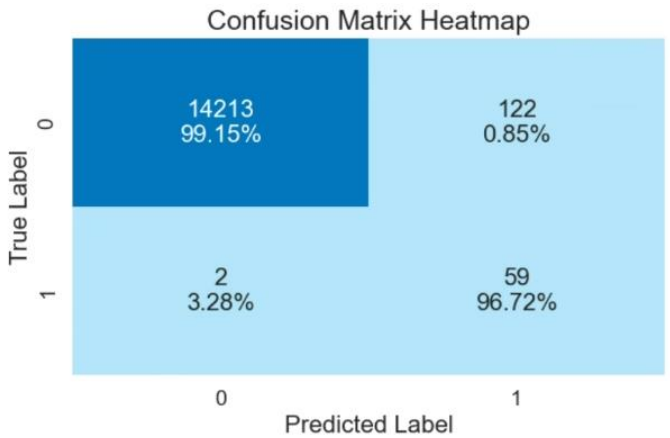
i. Accuracy (דיוק)
ii. Recall
iii. Specificity
iv. Precision
v. F1-Score
vi. AUC
vii. ROC

<u>אמצעים וויזואליים:</u>

1. **Confusion Matrix**

מטריצת הבלבול מציגה את מספר המקרים בכל קטגוריה: חיוביים אמיתיים (TP), חיוביים שקריים (FP), שליליים אמיתיים (TN) ושליליים שקריים (FN). מטריצה זו מאפשרת לנו להבין כמה תחזיות נכונות ולא נכונות המודל שלנו מבצע. ככל שמספר ה-TP וה-TN גבוהים יותר, כך המודל שלנו טוב יותר בזיהוי הונאות ואי הונאות.

גרף 15: מטריצת הבלבול



הגרף מציג את מטריצת הבלבול (Confusion Matrix) עבור המודל הנבחר.

מטריצה זו מציגה את התוצאות באופן הבא:

- True Label: הציר האנכי מייצג את התוויות האמיתיות של הדוגמאות.
- Predicted Label: הציר האופקי מייצג את התוויות שחזה המודל.

<u>תיאור הנתונים:</u>

- (TN) True Negative: מספר המקרים שהיו בפועל שליליים וגם נחזו כשליליים (0,0) - במקרה הזה 14,213 – 99.15%.
- (FP) False Positive: מספר המקרים שהיו בפועל שליליים אך נחזו כחיוביים (1,0) - במקרה הזה 122 – 0.85%.
- (FN) False Negative: מספר המקרים שהיו בפועל חיוביים אך נחזו כשליליים (0,1) - במקרה הזה 2 – 3.28%.
- (TP) True Positive: מספר המקרים שהיו בפועל חיוביים וגם נחזו כחיוביים (1,1) - במקרה הזה 59 – 96.72%.

2. **עקומת ROC**

עקומת ה-ROC מציגה את שיעור החיוביים האמיתיים (TPR) לעומת שיעור החיוביים השקריים (FPR) עבור ספי החלטה שונים. ה-AUC (שטח מתחת לעקומה) מציין את היכולת הכללית של המודל להבדיל בין מחלקות. ערך AUC קרוב ל-1 מעיד על מודל טוב.

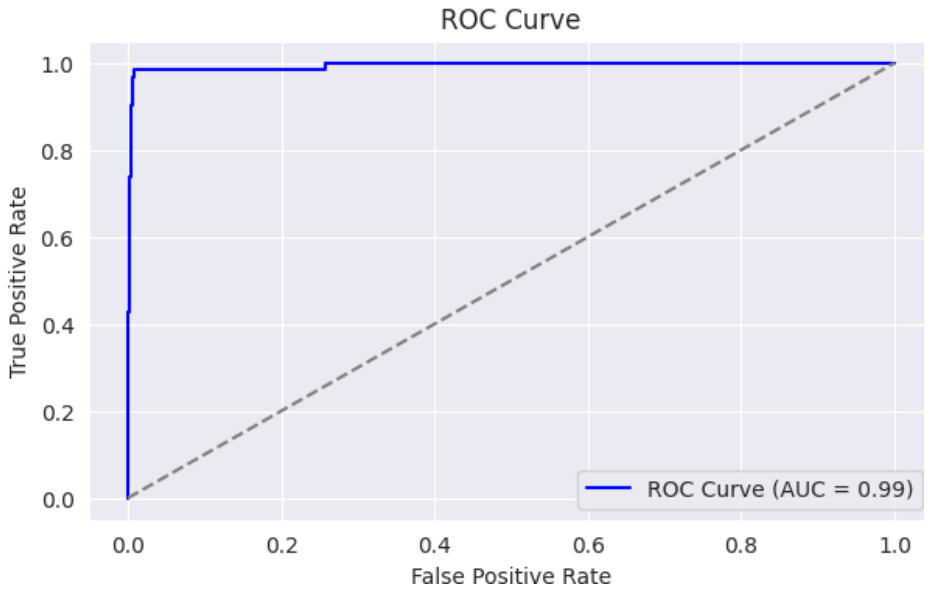
עקומה שמתקרבת לפינה העליונה השמאלית מעידה על מודל טוב יותר בזיהוי הונאות:

- False Positive Rate נמוך מצביע על כך שהמודל מצליח לשמור על שיעור נמוך של חיוביים שגויים, דבר המצביע על כך שהמודל מזהה בצורה מדויקת את רוב המקרים השליליים.
- True Positive Rate גבוה מצביע על כך שהמודל מזהה בצורה מוצלחת את רוב המקרים החיוביים, כלומר הוא מצליח לזהות רוב המקרים בהם קיימת הכחשה אמיתית של העסקה.

<u>תיאור הגרף:</u>

- הקו הכחול: מייצג את ביצועי המודל עבור ערכי סף שונים.
- הקו האפור: מייצג את קו הבסיס בו המודל הוא מקרי לחלוטין (מודל שלא מזהה כלל). כל מודל שמעל הקו הזה נחשב לטוב יותר ממודל אקראי.
- שטח תחת העקומה (AUC - Area Under the Curve): ה-AUC הוא ערך בין 0 ל-1, כאשר ערך קרוב ל-1 מעיד על ביצועים טובים מאוד של המודל.

גרף 16: עקומת ROC



גרף זה מציג את ביצועי המודל לחיזוי הכחשות עסקה באמצעות עקומת ROC, באופן הבא:

True Positive Rate (TPR): ידוע גם כרגישות (Recall), והוא היחס בין המקרים החיוביים שנחזו נכון (True Positives), לבין סך המקרים החיוביים בפועל. $\frac{TP}{TP+FN} = \text{TPR}$

False Positive Rate (FPR): היחס בין המקרים השליליים שנחזו בטעות כחיוביים (False Positives), לבין סך המקרים השליליים בפועל. $\frac{FP}{FP+TN} = \text{FPR}$

<u>ניתוח העקומה:</u>

- הקו הכחול: ניתן לראות כי במקרה זה הקו מתקרב מאוד לפינה השמאלית העליונה, דבר המעיד על ביצועים טובים של המודל.
- הקו האפור: ניתן לראות כי במקרה זה עקומת המודל נמצא הרבה מעל הקו האקראי.
- שטח תחת העקומה (AUC - Area Under the Curve): במקרה זה, ה-AUC הוא 0.99, דבר המצביע על כך שהמודל מבצע חיזוי בצורה מצוינת:.

3. גרף Boxplot המוצג יחד עם גרף Scatter Plot

גרף Boxplot מציג את התפלגות הנתונים באמצעות רבעונים, ומאפשר לזהות את הערכים החריגים.

- הקווים העליונים והתחתונים של התיבה מייצגים את הרבעון הראשון (Q1) ואת הרבעון השלישי (Q3)
- הקו בתוך התיבה מייצג את החציון, שהוא הערך האמצעי בנתונים.
- הקווים היוצאים אנכית לתיבה מייצגים את גבולות הנתונים שאינם חריגים, הנקבעים על פי 1.5 פעמים ההפרש הבין-רבעוני (IQR, Interquartile Range). חריגים מוגדרים

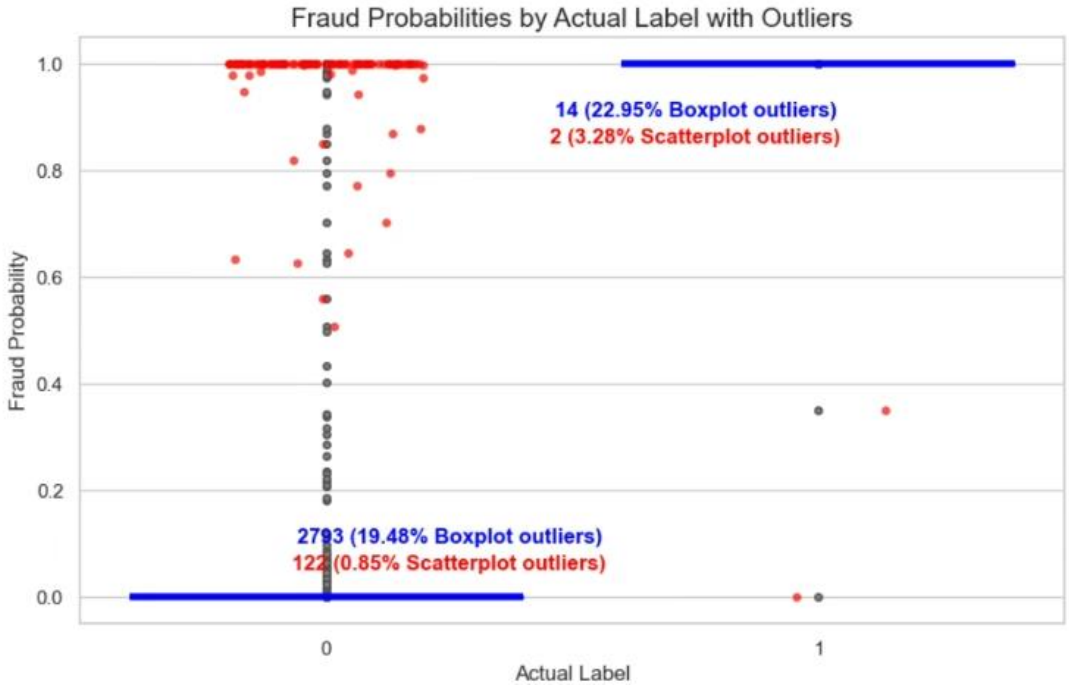
כערכים שנמצאים מחוץ לטווח של 1.5 פעמים ההפרש הבין-רבעוני (IQR) מכל צד של התיבה.

- הנקודות מחוץ לקווים מייצגות את הערכים החריגים.

גרף Scatter Plot מציג את הקשר בין שני משתנים, באמצעות נקודות המפוזרות בשטח מערכת הצירים.

- ציר ה-X מציג את המשתנה הראשון.
- ציר ה-Y מציג את המשתנה השני.

גרף 17: סבירויות להונאה שהמודל טעה בחיזוייו



הגרף מציג את ההסתברויות להכחשת עסקה שחזה המודל עבור כל לקוח, כאשר 0 מצביע על אי-הכחשה, ו-1 מצביע על הכחשה. ציר X: מציג האם המודל מסווג את הלקוח כמכחיש או לא. ציר Y: מציג את ההסתברות להכחשת עסקה שהמודל מצא עבור כל לקוח. נקודות בגרף: כל נקודה מייצגת לקוח אחד.

<u>ניתוח הגרף:</u>

הנקודות האפורות מייצגות חריגים בהתאם לגרף הBox Plot הסטנדרטי, בהם המודל לא חזה באופן אבסולוטי את מקרי הכחשת העסקה.

בהתאם לthreshold שהוגדר בפרק מתודולוגיה, מטרת המודל היא לשייך מקרים כהכחשת עסקה פוטנציאלית, או לא, בהתאם לקרבה היחסית של הסתברותם ל1 או ל0. לפיכך, בנוסף לגרף הBox Plot, מוצג גרף Scatter Plot. גרף זה מתאר מקרים חריגים על פי הגדרת הthreshold, על מנת לקבל תמונת מצב מדויקת יותר בנוגע לטעויות המודל בפועל.

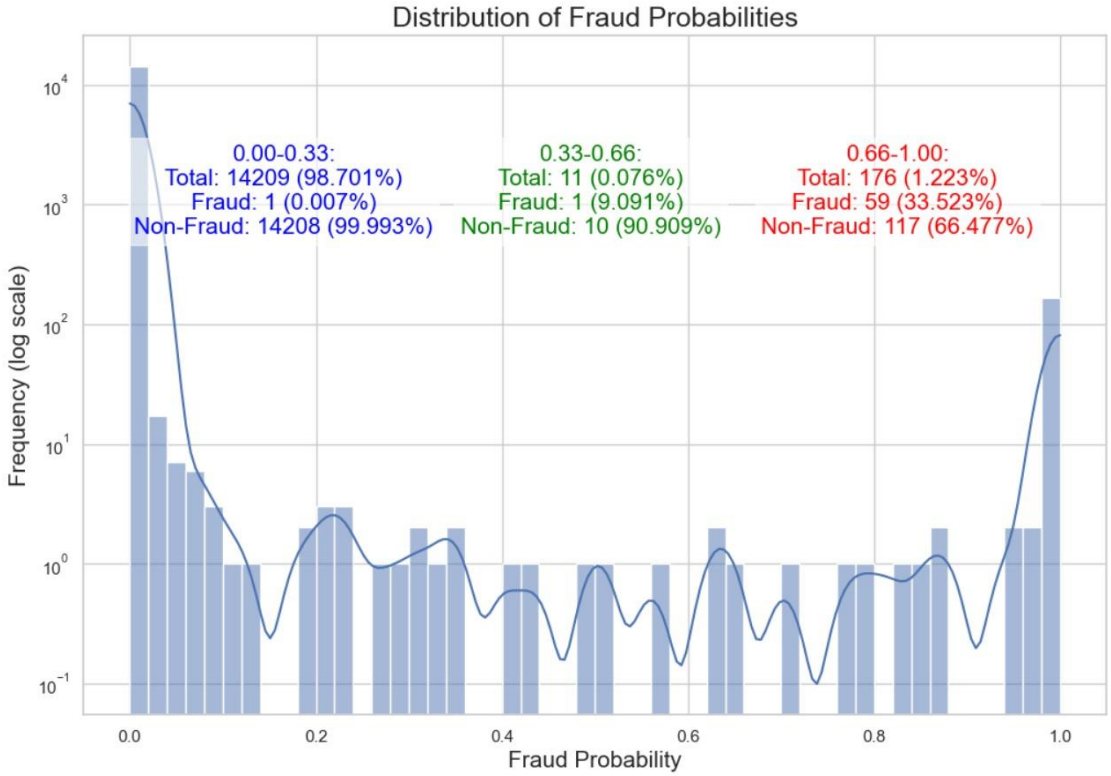
ניתן לראות ששיעור החריגים בפועל, קטן מזה המוצג בגרף הBox Plot. עבור לקוחות שסווגו כ-0 (אי הכחשה), על פי גרף הBox Plot 19.48% מהמקרים הוגדרו חריגים, בעוד שעל פי גרף

הScatter Plot רק 0.85% מהמקרים הוגדרו כחריגים. עבור לקוחות שסווגו כ-1 (הכחשה), 22.95% מהמקרים הוגדרו כחריגים על פי הBox Plot, בעוד שרק 3.28% מהמקרים הוגדרו כחריגים על פי הScatter Plot. פערים אלה מדגישים את חשיבות נקודת המבט הנוספת שמספק לנו גרף הScatter Plot, על מנת להבין את תוצאות המודל בצורה מדוייקת יותר.

4. היסטוגרמת החיזוי להונאה

היסטוגרמה היא כלי גרפי להצגת התפלגות של קבוצות נתונים על ידי חלוקתם למספר רווחים (bins). ההיסטוגרמה מראה כמה ערכים קיימים בכל רווח מסוים, והיא שימושית במיוחד כאשר רוצים לראות את התפלגות הנתונים ואת הצורה הכללית של התפלגות זו.

גרף 18: היסטוגרמת החיזוי להונאה



גרף זה מציג את התפלגות ההסתברויות להונאה שחזה המודל עבור כל לקוח. ההיסטוגרמה מחולקת לשלוש קבוצות הסתברויות ומוצגת בסקלה לוגריתמית בציר ה-Y.

<u>ניתוח הגרף:</u>

- צירים:
  - ציר X – הסתברות, לפי חיזוי המודל, שהלקוח יבצע הכחשת עסקה.
  - ציר Y – סך המקרים בהסתברות זו (מציר X).
- הגרף מחולק לשלוש קבוצות שונות:

1. **כחול**: שליש תחתון של הנתונים (0.00-0.33)

- סך כל המקרים בשליש זה: 14,209 (98.701% מכלל המקרים בגרף)

- הכחשה: 1 מקרה (0.007% משליש זה)
- ללא הכחשה: 14,208 מקרים (99.993% משליש זה)

רוב המקרים שהמודל חזה כמקרים ללא הכחשת עסקה- נמצאים בטווח הקרוב ל0 בהסתברות להכחשה. עובדה זו מצביעה על דיוק גבוה של המודל בזיהוי מקרים ללא הכחשת עסקה.

2. **ירוק**: שליש אמצעי של הנתונים (0.33-0.66)
- סך כל המקרים בשליש זה: 11 (0.076% מכלל המקרים בגרף)
- הכחשה: 1 מקרה (9.091% משליש זה)
- ללא הכחשה: 10 מקרים (90.909% משליש זה)

השליש האמצעי הוא טווח בו המודל מתקשה להחליט האם מדובר בלקוח שיבצע הכחשה או לא. ולכן, העובדה שמספר המקרים בטווח זה הוא קטן מאוד, היא דבר חיובי.

3. **אדום**: שליש עליון של הנתונים (0.66-1.00)
- סך כל המקרים בשליש זה: 176 (1.223% מכלל המקרים בגרף)
- הכחשה: 59 מקרים (33.523% משליש זה)
- ללא הכחשה: 117 מקרים (66.477% משליש זה)

רוב המקרים שהמודל חזה כהכחשת עסקה נמצאים בטווח זה, דבר המעיד על חיזוי טוב של המודל.

4. טבלת סיכום למדדים

הטבלה מציגה את תוצאות הביצועים של המודל באמצעות מדדים סטטיסטיים שונים. מדדים אלו משמשים להערכת איכות המודל ויכולתו לזהות בצורה נכונה מקרים של הונאה, לעומת מקרים שאינם הונאה.

טבלה 5: סיכום המדדים של המודל שנבחר

Metric	TN	TP	FN	FP	Precision	Recall	Specificity	F1-score	Accuracy	AUC
Score	14213.0	59.0	2.0	122.0	0.325967	0.967213	0.991489	0.487603	0.991386	0.993595

נראה כי המודל הראה יכולת טובה במיוחד בזיהוי מקרים חיוביים אמיתיים (Recall) עם ערך של 0.967.

<u>ניתוח הביצועים – מדדים:</u>

- Precision: אחוז המקרים שהמודל חזה כחיוביים והיו נכונים מתוך סך כל המקרים שהמודל חזה כחיוביים.
- Recall: אחוז המקרים החיוביים שהמודל הצליח לזהות נכון מתוך סך המקרים החיוביים בפועל.

- Accuracy: האחוז הכולל של התחזיות הנכונות (TP + TN) מתוך סך כל הדוגמאות.
- Specifity: אחוז המקרים השליליים שנתגלו נכון מתוך כלל המקרים השליליים בפועל.
- F1-score: ממוצע משוקלל בין Precision לבין Recall.
- Accuracy: אחוז החיזויים הנכונים מתוך כלל המקרים.
- AUC: שטח מתחת לעקומת הROC, כפי שהוסבר לעיל (סעיף 2- עקומת ROC).

<u>חישוב המדדים:</u>

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{213,14 + 59}{2 + 122 + 213,14 + 59} \approx 0.9914$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{59}{2 + 59} \approx 0.9672$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{59}{122 + 59} \approx 0.326$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{213,14}{122 + 213,14} \approx 0.9915$$

$$\text{F1 Score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2 = 2 \times \frac{0.326 \times 0.9672}{0.326 + 0.9672} \approx 0.488$$

<u>תוצאות המדדים:</u>

Accuracy: 99.14%

Recall: 96.72%

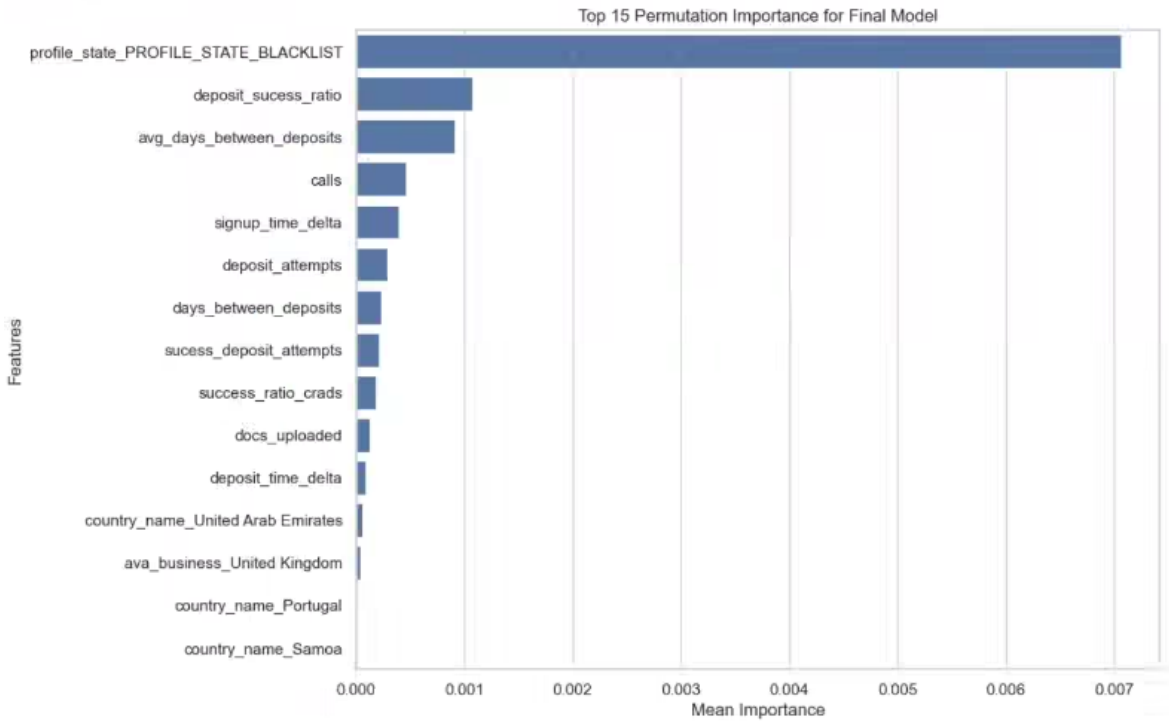
Precision: 32.6%

Specificity: 99.15%

F1 Score: 48.8%

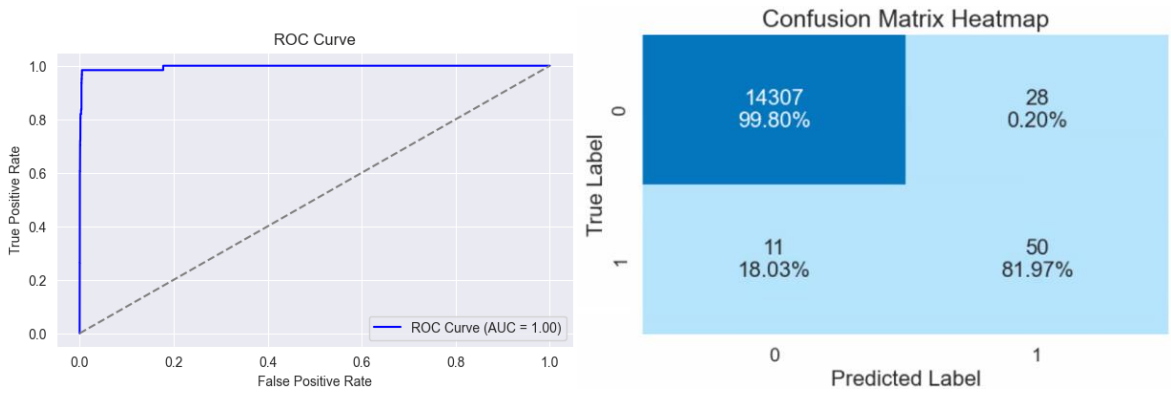
לאחר הערכת המודל, המודל מציג את 15 העמודות החשובות ביותר להצלחת חיזויו.

גרף 19: 15 העמודות החשובות



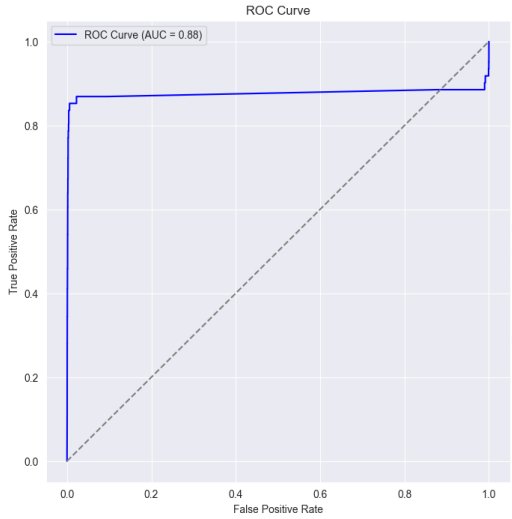
להלן השוואת המדדים הסטטיסטיים במודל HGB לפי שתי השיטות הנוספות לטיפול בדאטה, שלא נבחרו:

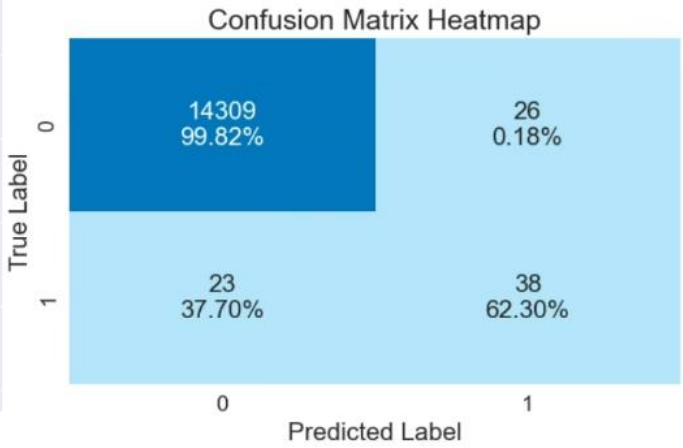
גרף 20: שיטת הטיפול SMOTE



Metric	TN	TP	FN	FP	Precision	Recall	Specificity	F1-score	Accuracy	AUC
Score	14307.0	50.0	11.0	28.0	0.641026	0.819672	0.998047	0.719424	0.997291	0.99606

גרף 21: ללא טיפול מיוחד בנתונים





Metric	TN	TP	FN	FP	Precision	Recall	Specificity	F1-score	Accuracy	AUC
Score	14309.0	38.0	23.0	26.0	0.59375	0.622951	0.998186	0.608	0.996596	0.829538

12. דיון ומסקנות

במהלך פרויקט זה, התמקדנו בניתוח הנתונים, בבניית מודלים ובהערכתם, כאשר המטרה הייתה לפתח מודל שיעזור בזיהוי ובהפחתת מקרי הכחשות עסקה עבור חברת AvaTrade. מטרה זו חשובה לחברה, על מנת להפחית את הסיכונים הכלכליים ואת הסיכונים הרגולטוריים אליהם היא חשופה בעת ביצוע הכחשות עסקה על ידי לקוחות. בניית והערכת המודל נעשתה תוך התחשבות במורכבות ובחוסר האיזון של הנתונים.

<u>ניתוח ביצועי המודל:</u>

במסגרת הפרויקט נבחנו מספר מודלים ושיטות לטיפול בנתונים. מההשוואות שבוצעו בין המודלים ובין הטכניקות השונות, נבחר המודל Hist Gradient Boosting בשיטת הטיפול Under-sampling. המדדים הסטטיסטיים מצביעים על כך שהמודל מצליח לזהות בצורה טובה את מרבית המקרים החיוביים (הכחשות עסקה), תוך שמירה על שיעור נמוך של טעויות מסוג False Positives. המדד העיקרי שנלקח בחשבון בעת ניתוח ביצועי המודל היה Recall, מאחר ש"מחיר הטעות" של החברה נמוך על זיהוי שווא כחיובי, לעומת "מחיר הטעות" שהחברה עלולה לשלם על זיהוי שווא כשלילי.
המודל הנבחר בשיטה הנבחרת, הראה ביצועים טובים במיוחד במדד הקריטי Recall, תוך שמירה על איזון בין זיהוי המקרים החיוביים (הכחשות עסקה), לבין שמירה על כמות מינימלית של אזעקות שווא.

הביצועים שהוצגו בשיטות הוויזואליות בפרק הערכת הפתרון, מציגים גם הם יכולות טובות של המודל. בפסקאות הבאות יוצגו המסקנות משיטות אלה.

Confusion Matrix:

המודל מזהה בצורה טובה מאוד את המקרים השליליים (TN גבוה)- לקוחות שלא ביצעו הכחשת עסקה וזוהו ככאלה, ומצליח לזהות גם את המקרים החיוביים האמיתיים בצורה מספקת (TP)- לקוחות מכחישים שזוהו ככאלה. בנוסף לכך, יש לו כמות מסוימת של זיהוי חיובי שקרי (FP)- לקוחות שלא ביצעו הכחשת עסקה, שזוהו בטעות כמבצעים הכחשת עסקה, וכמות מעטה מאוד של זיהוי שלילי שקרי (FN)- לקוחות מכחישים שזוהו בטעות כלא מכחישים. כלומר, המודל מצליח לזהות את רוב המקרים החיוביים, אך עדיין יש לו מקום לשיפור בזיהוי נכון של מקרים שליליים שזוהו בטעות כחיוביים (FP). **מאחר ולפי מטרת הפרויקט מחיר הטעות על FP הוא נמוך, ניתן להסיק לבסוף שהמודל מספק תוצאות טובות מאוד לצרכים הנדרשים.**

עקומת ROC:

העקומה המוצגת וה-AUC הגבוה מצביעים על כך שהמודל מבצע בצורה מצוינת בזיהוי הכחשות עסקאות. הוא מצליח להבחין היטב בין מקרים של הכחשות עסקה לבין מקרים שאינם הכחשות, דבר המקנה לו ביצועים גבוהים הן בזיהוי החיובים הנכונים והן בהימנעות מזיהוי שגוי של חיובים.

Box Plot & Scatter Plot:

הגרף מציג כי המודל מצליח לסווג בצורה נכונה את רוב המקרים בהתאם לסף (threshold) שנקבע. ישנם מקרים בודדים שסווגו כחריגים על ידי גרף הScatter Plot, אך העובדה שמקרים אלה מהווים אחוז קטן בלבד מכלל המקרים, מצביעה על כך שלרוב המודל מצליח לסווג בצורה טובה את המקרים.

היסטוגרמה:

- דיוק בזיהוי הכחשת עסקה בהסתברות גבוהה:
המודל מצליח לזהות באופן טוב מקרים של הכחשת עסקה בטווח ההסתברויות הגבוהות (0.66-1.00).
- חוסר דיוק בטווח הביניים:
בטווח ההסתברויות הבינוניות (0.33-0.66) המודל מתקשה להבחין בין הכחשת עסקה לבין אי-הכחשה. ככל שהמודל מסווג ערכים בטווח שקרוב יותר ל0 או ל1- כך הוא קרוב יותר לדיוק. באזור זה יש אחוז נמוך מאוד של מקרים (0.076% מכלל המקרים בגרף), דבר המצביע על יכולות טובות סיווג של המודל, אך עדיין משאיר מקום קטן לשיפורו.
- דיוק בזיהוי אי-הכחשת עסקה:
המודל מציג ביצועים טובים מאוד בזיהוי מקרים אלה, במיוחד בטווח ההסתברויות הנמוכות (0.00-0.33).

בהתבסס על התוצאות שהתקבלו, ניתן להסיק כי המודל Hist Gradient Boosting בשיטת הטיפול Under-sampling, הוא המודל המתאים ביותר לצרכי חברת AvaTrade. מודל זה יבטיח את צמצום הסיכונים הכלכליים והרגולטוריים עבור החברה.

<u>המלצות לעתיד:</u>

- **שימוש במודל תוחלת הרווח שנוסח עבור החברה:**

$$P * (\text{money}_{\text{loss}} + \text{regulation}_{\text{loss}}) + (1 - P) * \text{money}_{\text{profit}}$$

P היא ההסתברות להכחשת עסקה. הסתברות זו מוכפלת בהפסדים הכלכליים והרגולטוריים האפשריים, בשל הכחשה אפשרית. בנוסף, קיימת ההסתברות שלא תהיה הכחשת עסקה, שהיא המשלים להסתברות להכחשת עסקה. הסתברות זו מוכפלת ברווח הכלכלי האפשרי.

מודל זה מאפשר לחברה לשקלל את הרווח האפשרי, למול ההפסד האפשרי, ולקבל החלטות מושכלות בנוגע לאישור או לדחיית עסקאות, על בסיס ניתוח סיכונים מתמטי. למודל זה יש יתרונות רבים, כגון איזון בין רווח לבין הפסד, דבר המאפשר הסתכלות רחבה ומאוזנת. בנוסף, מודל זה מאפשר יכולת להגדיר את תוחלת הרווח המינימלית הנדרשת מלקוח שעלול להוות סיכון לחברה, וכך לשלוט ברמת הסיכון וברמת ההפסד האפשרי שהחברה מוכנה לספוג.
עם זאת, על החברה להגדיר טווח סיכון רגולטורי ולקחת בחשבון שהגדרה זו ככל הנראה תצריך נרמול לערך הכספי, על מנת להשתמש במודל בצורה נכונה מבלי להטות אותו עם מספרים גדולים ולהתאימו לטווח. כמו כן, על החברה להחליט מהי תוחלת הרווח המינימלית, כפי שצוין לעיל.
אנו ממליצים לחברה להמשיך ולפתח את המודל, בהתאם לשינויים שיהיו בעתיד, באם יהיו.

- **איסוף נתונים נוספים והרצת המודל מחדש:**

איסוף נתונים והרצת המודל הנבחר עליהם, יאפשרו את דיוק התחזיות ואת התאמת המודל למציאות המשתנה. איסוף הנתונים כולל הן נתונים על לקוחות נוספים, הן את הנתונים הקיימים לאורך זמן, והן נתונים נוספים כגון תאריכי הכחשת עסקה (כיום המידע היחיד בנוגע להכחשת לקוח הוא האם בוצעה הכחשה או לא).

- **הצגת מדד כמותי נוסף:**

כפי שתואר במדדי הפרויקט, איסוף תאריכי הכחשת עסקה יוכל להוות מדד נוסף ומשמעותי להצלחת הפרויקט.

- **שיפור המודל ובחינתו:**

כפי שהורחב לעיל, יש למודל מקום קטן לשיפור בנושא חיזוי שליליים כחיוביים (False Positives), ובנושא סיווג לקוחות שהסתברותם להכחשת עסקה נעה סביב החצי. יתרה מזאת, עם איסוף נתונים חדשים והזמן החולף, יש לבחון את המודלים הנוספים שהבאנו. יש לקחת בחשבון שהמודל המועדף והשיטה הנבחרת עשויים להשתנות לאורך הזמן. בנוסף לכך, מומלץ לבחון באופן תדיר שיטות חדשות ומודלים חדשים, ולהתעדכן בהתפתחויות בתחום הMachine Learning, ובחידושים שהוא מביא עימו.

- **בחינת הThreshold ועדכונו לפי צורך:**

מומלץ לבחון את הThreashold שהוגדר בפרויקט זה, ולעדכנו בהתאם לרמת הסיכון שהחברה רוצה ויכולה לקחת. יש לוודא כי הThreashold נשאר רלוונטי ומותאם לצרכי החברה גם במשך הזמן, בהתחשב בשינויים הרגולטוריים והכלכליים.

<u>תרומות:</u>

כל אחד מחברי הצוות תרם בצורה משמעותית להצלחת הפרויקט.

- **קודים:** רוני ויאיר הובילו את העבודה על פיתוח הקודים.
- **הצגת אבן דרך**: רוני ומוריה הציגו את אבן דרך אחת.

- **חיפוש מאמרים:** שלושתנו סייענו בחיפוש המאמרים.
- **סקר ספרות:** יאיר ביצע את הסקירה האינטגרטיבית, בעוד שמוריה עסקה בסקירה המבוססת על מאמרים, בסיוע סיכומי המאמרים של רוני.
- **כתיבת הדוח:** כל חברי הצוות היו מעורבים בכתיבת הדוח. יאיר היה אחראי על החלקים הראשוניים כמו סקר הספרות, בעוד שמוריה התמקדה בפרקים האחרונים, כגון תקציר מנהלים, הצגת חלופות והערכת הפתרון. רוני התמקדה בפרקי המתודולוגיה ומימוש הפתרון.

המשימות חולקו בצורה מאוזנת כדי להבטיח שכל היבטי הפרויקט יקבלו תשומת לב ויושלמו ברמה הגבוהה ביותר.