בס"ד

שם הפרויקט –DL &CHARGEBACKD

שם הקורס : למידה עמוקה

פרטי המגיש :

שם - יאיר עמר

תז - 314715111

**תקציר הפרויקט :**

המטרה הראשית של הפרויקט הינה להפחית מקרים של הכחשת עסקה

על ידי חיזוי ההסתברות שהלקוח יבצע הכחשת עסקה בעזרת מודל.

הנתונים התקבלו מחברת AVATRADE

גודל הדאטה המקורי הוא כ24,400 שורות ו כ24 עמודות

ואחרי עיבוד והפיכת הדאטה למאוזן מספר השורות יורד

ל285 ומספר העמודות עולה למעל 1000.

השתמשתי בארכיטקטורת רשת נוירונים מסוג Sequential

התוצאות המתקבלות מהמודל הינם גבוהות במיוחד במספר המטריקות שבהם השתמשתי , וזאת למרות שבגרף הקורלציות לא נמצא שום קשר חיובי חזק בין משתנה המטרה למשתני הקלט.

# מבוא

AVATRADE הינה חברה מובילה בישראל למסחר עצמאי, באמצעות חוזה הפרשים (CFD)

בחברה קיים מגוון רחב של אפיקי השקעה.

לדוגמא: מדדים, מט"ח, מניות,אג"ח, מטבעות וסחורות.

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.

החברה מנסה לאזן בין הרצון לנהל את הסיכון להונאה, לבין הנוחות וחווית המשתמש של לקוחותיה בעת ביצוע המסחר שבסופו של דבר מעודדת לקוחות להתחבר ולבצע פעולות בחשבון.

לדוגמה אימות דו-שלבי מקשה על ההתחברות ולכן מפחית מספר כניסות ופעולות.

כיום אחוז הכחשות עסקה המדווחות עומדות על **כחצי אחוז** כאשר עד לאחוז אחד חברות האשראי מאפשרות המשך התקשרות ללא התערבות וחסימה.

**מטרת הפרוייקט**

להפחית את כמות מקרי ההונאה, על מנת להמנע מסנקציות של חברות האשראי,

שזה הסיכון העיקרי של החברה.

בעזרת מודל שחוזה בהסתברות גבוהה לדרג את הפוטנציאל לביצוע chargeback אצל הלקוחות החברה תוכל לנקוט באמצעי ביטחון מיוחדים עד כדי חסימת משתמש

# נתונים :

הנתונים התקבלו מחברת AVATRADE בקובץ אקסל , הנתונים במבנה אחיד טבלאי

משיחה עם האיראי בחברה עולה שגם הערכים החריגים הגיוניים והם לא טעות.

משתני הקלט הינם : 'user', 'deposit\_attempts', 'sucess\_deposit\_attempts', 'num\_cards',

'num\_cards\_success', 'first\_deposit', 'last\_deposit',

'deposit\_sucess\_ratio', 'deposit\_time\_delta', 'signup\_time\_delta',

'calls', 'docs\_uploaded', 'docs\_approved', 'approval\_rate',

'sales\_view', 'conversion\_owner', 'sales\_agent',

'account\_create\_on\_date', 'country\_name', 'tag', 'channel',

'ava\_business', 'profile\_state', 'num\_closed\_trade'

משתנה המטרה הינו: 'IsFraud'

סטטיסטקות בסיסיות על חלק מהעמודות :



הנתונים התקבלו בשני קבצים אחד של כל הדאטה ואחד רק של מי שהכחיש עסקה אז הוספתי את משתנה המטרה לקובץ הראשי,הצגתי היסטוגרמות או פלוט סכימה לזהות ערכים חריגים

הדאטה היה מאוד לא מאוזן ברמה של רק כחצי אחוז היה למקרים של הכחשת עסקה לכן ביצעתי undersample

מבחינת ערכים חסרים בגלל שלא היה מספר משמעותי פשוט זרקתי את השורות עם הערכים חסרים.

הצגתי גרף של קורלציות לזיהוי קשרים חזקים בין משתני הקלט למטרה ולא נמצא משהו משמעותי

בחלוקה למבחן השתמשתי ב20 אחוז ובאימון ב80 אחוז שמתוכם בשלב הפיט הקצאתי 15 אחוז לואלידציה.

גודל הדאטה המקורי הוא כ24,400 שורות ו כ24 עמודות ואחרי עיבוד והפיכת הדאטה למאוזן מספר השורות יורד ל285 ומספר העמודות עולה למעל 1000

# מתודולוגיה

בדוגמה זו השתמשנו בארכיטקטורת רשת נוירונים מסוג Sequential מהספרייה TensorFlow שמתאימה לבצע סיווג בינארי בעזרת שכבות מדויקות.

הבחירה ב־Sequential נעשתה משום שהיא פשוטה לשימוש ומתאימה למקרים שבהם התקבלו נתונים בעלי מבנה אחיד,כלומר שכל המאפיינים בקבוצת הנתונים מתאימים להיות חלק מהקלט של הרשת הנוירונית.

הארכיטקטורה כוללת חמישה שכבות Dense, כאשר כל שכבה מכילה מספר מסוים של יחידות (units). השכבות הביניים מכילות 32, 16, ו־8 יחידות בהתאמה, כל אחת עם פעולת הפעלה ReLU. שכבת הפלט, שבה יש יחידה אחת, משתמשת בפונקציית הפעולה Sigmoid לסיווג דו־קטגורי .

הכניסה לרשת היא וקטור עם מספר מאפיינים על פי מספר המאפיינים במערך X\_processed\_train.

נעשה שימוש בפונקציית k\_cross\_validation\_with\_nn לצורכי אימון ואימות ובשלב האימון מתבצעת חלוקה נוספת ל85 אימון 15 מבחן בעזרת **validation\_split=0.15**

הקריטריון המרכזי שנעשה שימוש בו לאימון ובחירת היפר פרמטרים הוא מדד **F1-score**,

בנוסף נעשה שימוש במדדים נוספים כגון **Recall AUC Precisionו Accuracy**

לאחר קביעת ארכיטקטורת הרשת, אימון המודל נעשה באמצעות חיפוש אופטימלי של הפרמטרים הרלוונטיים (batch size, epochs, optimizer) על ידי RandomizedSearchCV של ספריית scikit-learn.

החיפוש האופטימלי מתבצע על מנת למצוא את ההיפרפרמטרים שמביאים את הביצועים האופטימלים של המודל ביחס לנתונים האימון.

# תוצאות

משתנה המטרה שאנחנו רוצים לחזות הוא בינארי,

בגלל שהדאטה היה לא מאוזן השיטה שבא השתמשתי זרקה הרבה מהדאטה וכנראה גם דוגמאות חשובות לכן הרצתי את כל הקוד כמה פעמים לצרכי השוואה כאשר בכל פעם מתבצע k\_cross\_validation ואלה המדדים הרלוונטים בממוצע לאחר כל התהליך:

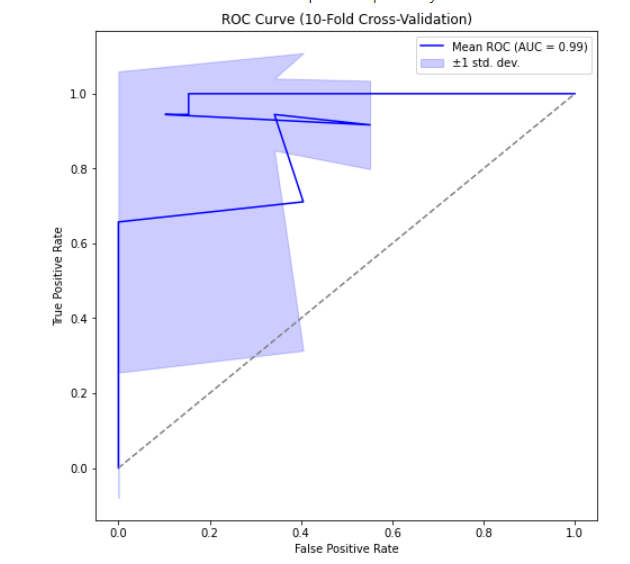
**תמונה שמכילה טקסט, צילום מסך, גופן, מספר

התיאור נוצר באופן אוטומטי**תמונה שמכילה טקסט, צילום מסך, גופן, מספר

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, צילום מסך, גופן, מספר

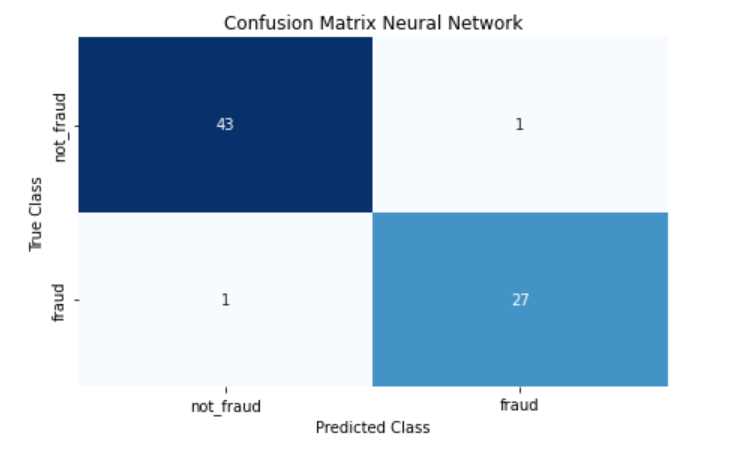
התיאור נוצר באופן אוטומטי

הצגה של השטח בדוגמא אחת אחרי k-fold:

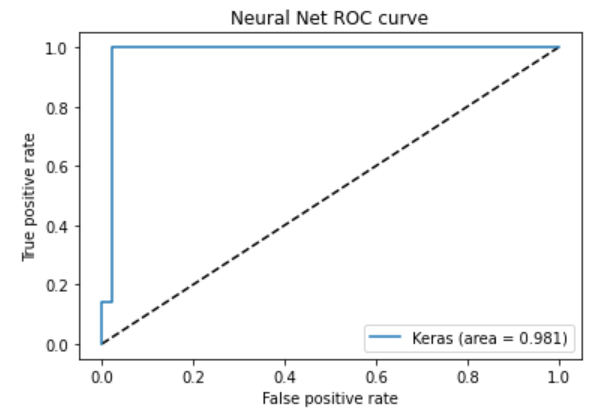


אפשר לראות שממוצע השטח מאוד גבוה ברמה שאפילו מחשידה.

הצגה של התוצאות במטריצה:



הצגה של השטח בהרצה אחת פשוטה :



# סיכום ומסקנות:

לסיכום המודל מחזיר תוצאות מדהימות ברמה שאפילו מעלות חשד לזליגת

דאטה ואחרי ווידוא של התוצאות אפשר להשתמש בהם לצורכי החברה .

פעולות אפשריות להמשך :

* לבחון מספר מודלים נוספים ולשם כך לעשות בחירת פיט’צרים ובסופו של דבר

להשוות בין כול המודלים .

* לבחון דרכים נוספות לטיפול בדאטה לא מאוזן .
* להוריד ערכים חריגים .
* לטפל במקרים מיוחדים כמו עמודת המדינה כדי להפחית את כמות העמודות

אחרי שלב הקידוד חם.

# קישור לפרויקט

<https://github.com/yairamar097/DL-CHARGEBACKD.git>