פרויקט מסכם - מבוא ללמידת מכונה

קבוצה 15: **מתן** ****** ********* | **אסף בירן** 316354984

תקציר מנהלים

בפרויקט זה פותח מודל חיזוי המנבא האם בקשת לקוח להלוואה תאושר או תידחה. המודל יוכל לסייע לארגון בתהליך קבלת ההחלטות על אישור הלוואות, ויהווה כלי תומך לשיפור תהליכים עסקיים ולניהול סיכונים.
ראשית, בוצע ניתוח של הנתונים ועיבודם בהתאם. לאחר מכן נבחנו ארבעה מודלים: Logistic Regression, Random Forest, ANN, Decision Tree, ונמצא כי המודל בעל התוצאות הטובות ביותר הינו Random Forest.
מודל זה סיפק ציון AUC גבוה מאוד – 0.9176. לאחר מכן נבחנו הציונים ה - confusion matrix של המודל.
בשלב זה, על מנת למקסם את פוטנציאל הרווח של הארגון, הועדף מזעור ה- FPR על חשבון מדדים אחרים, זאת מכיוון שבמקרה בו מאושרת הלוואה לאדם שאינו יכול להחזיר אותה הארגון מפסיד יותר מאשר במקרה דחייה של לקוח שיחזיר את ההלוואה. לכן, לאחר ניתוח של thresholds שונים בתחזית המודל ומכיוון שמודל זה מחזיר וקטור הסתברויות ולא תחזית בינארית, נמליץ לקבוע את הרף על 0.7.
המסמך הנ"ל מספק סקירה מעמיקה על הלוגיקה מאחורי ההחלטות שבוצעו בבניית המודל ותוצאותיו.

חלקים 2 + 1 – אקספלורציה ועיבוד מקדים

מטרת חלק זה היא העמקת ההבנה של הנתונים ובניית תמונה כללית של 'התנהגות' הנתונים, וכן מימוש המסקנות על ידי ביצוע פעולות שונות בנתונים, טרם בניית המודל עצמו. כל זאת על מנת לשפר את טיב המודל ודיוקו .
בפרויקט זה נעשה שילוב בין חקר הנתונים לעיבוד המקדים. בחירה זו נעשתה על מנת לאפשר הקשר ברור וביידי יותר בין פעולות עיבוד שנעשו על מאפיין כלשהו לבין הסיבות והגורמים לפעולות אלו. לכן המאפיינים נבחנו בצורה טורית - עבור כל מאפיין נעשה ניתוח בסיסי ולאחר מכן בהתאם לממצאים נעשו פעולות עיבוד שונות או ניתוחים מתקדמים יותר.
כמו כן, הוחלט לבצע הפרדה בעיבוד הנתונים בין מודלים רגישים לערכים חרגים ונרמול כגון רגרסיה לוגיסטית ו - ANN לבין מודלים מבוססי עצי החלטה שפחות רגישים לתופעות אלו, זאת על מנת למקסם את הפוטנציאל של כל אחד מהמודלים וכדי לבחור את המודל הטוב ביותר בצורה מושכלת. על כן בחלק מהמאפיינים נעשו פעולות המיועדות למודלים רגישים ואילו פעולות אחרות המיועדות למודלים מבוססי עצי החלטה. לבסוף המודלים נבחנו על עמודות שונות, בהתאם למאפיינים התואמים להם .[1]

ראשית, נבחנו כלל העמודות הנומריות (מספריות), לאחר מכן עמודת המטרה (Loan_Approval), ולבסוף עמודות קטגוריאליות.
בכל מאפיין בחנו את התפלגות ערכיו, ואת קיומם של ערכים חריגים או חסרים.
עבור ניתוח ראשוני של משתנים נומריים נבנתה פונקציית *feature analyze*. הפונקציה מקבלת כפרמטרים את ה-DataFrame, שם המאפיין, ומספר (bins). הפונקציה מציגה היסטוגרמה של המאפיין (כאשר הפרמטר bins מאפשר לשפר את אינפורמטיביות ההיסטוגרמה בהתאם להתפלגות הפיצ'ר), נתונים סטטיסטיים של המשתנה באופן כללי (בעזרת פונקציית describe()), וכן חלוקה של נתונים אלו לפי סיווג התצפיות בעמודת המטרה. בנוסף הפונקציה מדפיסה את מספר הערכים החסרים ומספר הערכים החריגים בעמודה ואת האחוז שהם מהווים מכלל התצפיות. ערכים חריגים נקבעים על פי שיטת ה - IQR. בנוסף הפונקציה **מחזירה** משתנה מסוג tuple, אשר רכיביו הם הגבול העליון והתחתון של הטווח בו תצפיות נחשבות כחריגות.[2] ניתוח זה נועד לתת תמונה כללית של המאפיין והתפלגותו על מנת לגבש אסטרטגיה אופטימלית לעיבוד שלו.
לאחר מכן הוצגו היסטוגרמות ותרשימי קופסא של המשתנה, תוך הפרדה בין תצפיות שסווגו בעמודת המטרה כחיוביות לבין תצפיות שליליות. הפרדה זו מאפשרת הבנה עמוקה יותר של ההבדלים בין תצפיות חיוביות לתצפיות שליליות, וכן מאפשרת לזהות נטיות ומגמות הקשורות לעמודת המטרה .
לבסוף, נעשו פעולות עיבוד למאפיינים בהתאם למסקנות הניתוח, וכן נעשו ניתוחים מתקדמים יותר במידת הצורך. הניתוח הראשוני של עמודות קטגוריאליות כלל barplot המציג בציר ה - X את סוגי התיוגים השונים של המשתנה, ובציר ה- Y מוצגת כמות התצפיות שקיימת לכל תיוג, כאשר כל תיוג מפוצל לתחזיות שליליות וחיוביות. גרף זה מאפשר להבין את התפלגות

---

[1] יפורט בהמשך בחלקים הרלוונטיים

[2] זיהוי הערכים החריגים והטיפול בהם מפורט בהמשך

המשתנה הכללית, וכן את התפלגות המשתנה בקרב תצפיות חיוביות ושליליות. יחד עם הגרף הודפס גם מספר הערכים החסרים של המשתנה. לאחר מכן נעשו פעולות עיבוד למשתנים וכן ניתוח מתקדם יותר במידת הצורך.
כמו כן נבחנו קשרים בין המשתנים השונים - הן על ידי לוגיקה והבנה של המאפיינים והן על ידי מטריצת קורלציות, זאת על מנת למנוע את ההגדלה של מימדיות הבעיה לשווא. מימדיות גדולה של הבעיה יכולה לפגוע בתוצאות המודל במספר דרכים: היא יכולה לגרום ל- overfitting במודל, ובמודלים מבוססי עצי החלטה היא יכולה להפריע למודל בבחירת המאפיינים החשובים ביותר (לדוגמא: אם יש שני מאפיינים עם קורלציה מלאה, החשיבות שלהם במודל תתפצל ביניהם, על אף שהמודל יכול לקבל את "החשיבות" הזו על ידי הכנסת מאפיין אחד בלבד). נוסף על כך, מימדיות גדולה יותר מאריכה את זמן החישוב, ובכך מגבילה את האפשרות לחקור את המודל ולשפר אותו בשלבים מתקדמים יותר.

לאחר הניתוח הראשוני בוצעו פעולות עיבוד. במהלך עיבוד הנתונים בוצעו הפעולות הבאות:

<u>"ניקיון" וסידור מקדים של הנתונים:</u>

בחלק מהעמודות הקטגוריאליות היה צורך בסידור ו- 'נרמול' נתונים במהלך הניתוח הראשוני על מנת לאפשר הבנה טובה יותר של הערכים והמגמות בהמשך הניתוח ובמודלים. בין היתר בוצעו:

- נרמול עמודת Preffered_communication_Method ואיחוד הפורמט של הערכים, לדוגמא: Phone_call הומר ל- Phone.
- הומרו ערכים קטגוריאליים, לדוגמא: בכל משתנה קטגוריאלי בעל תצפיות שתויגו כ- '?' , תיוג התצפיות הומר ל- Unknown.
- המרת עמודת Gender לערכים בינאריים (Male == 0 , Female == 1) על מנת לשפר את עיבוד משתנה זה בקרב המודלים השונים .
- המשתנה הקטגוריאלי Education_Level הומר למשתנה אורדינלי לפי רמת ההשכלה .

<u>ערכים חסרים:</u>

נעשתה סקירה על מספר הערכים החסרים בעמודות על מנת לוודא שאין תצפיות בעלות ערכים חסרים רבים. מסקירה זו עלה כי אין תצפיות בעלות ערכים חסרים רבים.[3] לאחר מכן במהלך ניתוח המאפיינים זוהו מאפיינים בעלי ערכים חסרים, אשר טופלו על ידי הפעולות הבאות:

- עבור המשתנה Age - ערכים חסרים הושלמו באמצעות החציון, שמהווה תמצית חסינה לערכים חריגים, על מנת שלא לפגוע בצורה מהותית בהתפלגות המקורית של הנתונים.
- עמודה B הוסרה מכיוון שכל ערכיה היו חסרים ולכן איננה אינפורמטיבית.
- במשתנים Investment Gain / Loss מרבית הערכים הם אפס וכמות נמוכה של ערכים חסרים, על כן ערכים אלו הושלמו על ידי הערך השכיח אפס.
- במשתנים קטגוריאליים ערכים חסרים סווגו תחת התווית ' Unknown' באופן טריוויאלי.

<u>ערכים חריגים ונרמול</u>

מודלים מבוססי עצי החלטה אינם רגישים לערכים חריגים וכן אינם זקוקים לנרמול של הנתונים. פעולות אלו אף יכולות לפגוע בתוצאות מודלים מסוג זה. על כן הטיפול בערכים חריגים לא נעשה על העמודה המקורית, אלא לכל מאפיין נוצרה עמודה חדשה (לכל feature היא נקראה: {feature}_scaled) ובעמודה זו טופלו הערכים החריגים ונעשה נרמול.
באופן כללי, בעמודות נומריות הערכים החריגים זוהו בעזרת שיטת ה- IQR, מכיוון ששיטה זו אינה מניחה שהנתונים מתפלגים בקירוב נורמלית (לעומת z-score), ועל כן שיטה זו פחות מוגבלת ומתאימה למגוון רחב יותר של מאפיינים. בעת זיהוי ערכים חריגים הפעולות שנעשו הן:

- בעמודת Age נעשה "קיצוץ" (Clipping) של הערכים לפי שיטת ה – IQR, כלומר ערכים גבוהים הומרו לגבולות העליונים והתחתונים (בהתאמה) של הטווח שאינו נחשב חריג על פי שיטת ה- IQR. כך מצד אחד לא הוסר מידע לחלוטין מהנתונים, אך מצד שני רמת החריגות של התצפיות ירדה, ובכך גם השפעתן על הנרמול ועל הטיית המודלים ירדה.

---

[3] ראה נספח 1.0

- בעמודות D, C_imputed מספר הערכים החריגים היה קטן מאוד והם נעו בטווח הגיוני וסביר משאר ההתפלגות, לכן לא נעשה טיפול בערכים חריגים בעמודות אלו.
- עמודת Weekly_Work_Hours מרוכזת מאוד סביב מרכז ההתפלגות שלה, לכן למעלה מרבע מהתצפיות נחשבות כביכול כ- "ערכים חריגים" בהתפלגות זו. על כן לא ניתן לטפל בערכים חריגים בשיטת IQR עבור עמודה זו,כי פעולה כזו תפגע באינפורמטיביות העמודה. בעקבות זאת נעשה Clipping על האחוזון העליון והתחתון של העמודה. כך רק ערכים קיצוניים "קוצצו" והעמודה שומרת על האינפורמטיביות שלה.
- עמודת A הינה בעלת זנב ימני מובהק מאוד[4] ולכן נבחנו טרנספורמציות לוגריתם ושורש על העמודה. לבסוף נבחרה טרנספורמציית השורש (והוגדרה עמודה חדשה "sqrt_A" בהתאם), משום שהפכה את הנתונים לדמויי התפלגות נורמלית בצורה טובה יותר מטרנספורמציית הלוגריתם, וכן מזערה את מספר הערכים החריגים. ערכים שנשארו חריגים גם לאחר הטרנספורמציה טופלו על ידי Clipping בשיטת ה- IQR.
- במשתנים קטגוריאליים ישנם שני סוגי חריגות - תיוגים בעלי מספר קטן של תצפיות ותיוגים בעלי רוב מובהק בעמודה, אשר כוללים את רוב התצפיות. בשני המקרים נעשה איחוד של תיוגים בעלי מספר קטן של תצפיות עם תיוגים אחרים בעלי מאפיינים זהים (לדוגמא : במשתנה Job_Type התיוג Priv-house-serv בעל תצפיות מועטות ולכן אוחד עם תיוגי מיעוט נוספים בעלי אופי דומה כגון Handlers-cleaner. כמו כן, במשתנה Country_of_Residence רובן המוחלט של התצפיות מתויגות כ- USA, לכן שאר המדינות אוחדו לתיוג Other).[5]

לאחר הטיפול בערכים חריגים, נורמלו הנתונים בעזרת Z-score standartization. שיטה זו נבחרה מכיוון שהיא פחות רגישה לערכים חריגים מ- Min-Max Scaling, ולרוב מתאימה יותר למודלים מבוססי מרחק כגון רגרסיה לוגיסטית, KNN ו- SVM.
עוד גורם לבחירה בשיטה זו היא העובדה שבפרויקט זה לא נעשה שימוש ב- PCA (ב- PCA ישנה העדפה מובהקת לשימוש ב- Min-Max Scalar) .

<u>Feature Engineering</u>

על מנת לנסות ולשפר את תוצאות המודלים נעשו פעולות עיבוד נוספות על הנתונים. הפעולות שנעשו הן:

- יצירת גירסה קטגוריאלית למשתנים נומריים (Age, Weekly_Work_Hours, C_imputed, D) המחלקת את הערכים המספריים לקטגוריות מספרים על פי טווח. החלוקה נעשתה על בסיס פיזור ההתפלגות של הנתונים, שיעורי אישור הלוואה דומים, וכן משמעויות הנובעות מהבנה של הדומיין (לדוגמא חלוקה של קטגוריות גיל לפי שלבים שונים בחיים).[6] חלוקה כזו יכולה לשפר מודלים מבוססי עצי החלטה. לדוגמא: חילקנו את הגילאים לפי שלבים שונים בחיים. כך מצאנו למשל שעבור תצפיות בגילאי 18-24 שיעור אישורי ההלוואה נמוך מאוד (דבר שמסתדר עם ההבנה שלנו על הדומיין, כיוון שבגיל זה רוב האנשים עדיין לא מבוססים כלכלית). כך הוספה של משתנה זה לעץ החלטה יכולה לשפר את תוצאותיו שכן היא יכולה "לחסוך" למודל חישובים ולגרום לו לזהות באופן ישיר מגמות שהוא היה מתקשה יותר לזהות בעמודה הנומרית המקורית. במקרה של אוברפיטינג ניתן אף להחליף את העמודה הנומרית בעמודת הקטגוריות על מנת להעלות את ההכללה של המודל ולהקטין את השונות. נשים לב, שעמודות אלו נועדו למודלים מבוססי עצי החלטה - מודלים שאינם מושפעים מנרמול וערכים חריגים, ולכן החלוקה לקטגוריות נעשתה על העמודות המקוריות ולא על העמודות המנורמלות.
- יצירת גרסאות יותר "מקובצות" עבור חלק מהמשתנים הקטגוריאליים - עבור משתנים בעלי מספר רב של תיוגים, נעשה קיבוץ של התיוגים על פי התפלגות המשתנה, שיעורי אישור הלוואה דומים, וכן משמעויות הנובעות מהבנת הדומיין.
פעולה זו מאפשרת לבצע encoding בצורה טובה על המשתנה, וכן מעלה את הכללת המודל, ובכך מפחיתה את הסיכויים לאוברפיטינג .
- טיפול במשתני Investment למיניהם - בעיבוד הנתונים עלה כי משתני ה- Investment חריגים בהתפלגותם. רובן המוחלט של התצפיות שווה ל-0, ובקרב התצפיות שאינן שוות ל-0 ישנן תצפיות חריגות מאוד בערכן הגבוה. הוחלט לא להסיר תצפיות אלו, משום שמבחינת הערכים החריגים עלתה המסקנה כי ככל הנראה לא מדובר בטעות בהזנת המידע, זאת בעקבות השכיחות הגבוהה יחסית של ערכים קיצוניים והבנה של הדומיין (לדוגמא: מההסטוגרמה נובע שהרווח של למעלה מ- 120 אנשים הוא סביב ה 100000. הסיכוי שלכל כך הרבה אנשים הוקלד בטעות סכום כזה קיצוני הוא קלוש בהתחשב

---

[4] ראה נספח 1.1

[5] ראה נספח 1.3

[6] ראה נספח 1.2

בהתפלגות הנתונים, מה גם שרוב האנשים אינם משקיעים, ואנשים שמלומדים בתחום יכולים להרוויח סכומים גבוהים במיוחד). בנוסף מבחינת הנתונים עלה כי אין אף תצפית שבה גם עמודת הרווח וגם עמודת ההפסד גדולות מ-0, וברובן המוחלט של התצפיות שתי העומדות שוות ל- 0. על כן הוחלט למלא את הערכים החסרים בעמודות ב- 0, וכן ליצור משתנה חדש:

Investment_balance = Gain-Loss. משתנה זה למעשה יכול להחליף את שתי העמודות הנ"ל ובכך להקטין את מימד הבעיה ולשפר את זמן הריצה של המודלים. על משתנה זה בוצע Clipping לאחוזון העליון והתחתון וסטנדרטיזציה והתוצאה נשמרה בעמודה ייעודית. כמו כן, נוצרה עמודת *Investment_Type* שמהווה המרה של עמודת מאזן ההשקעות לעמודה קטגוריאלית (רווח, הפסד, או ניטרלי). בנוסף, מניתוח הנתונים עלה כי ישנו הבדל משמעותי בשיעור אישורי ההלוואה בין תצפיות שביצעו השקעות (כלומר הפסידו או הרוויחו) לבין אנשים שמאזנם הוא 0.[7] בצורה מעט מפתיעה אולי, גם עבור אנשים שהיו מופסדים בהשקעות, שיעור אישורי ההלוואות היה גבוה (מעל 50 אחוז, אל מול 20 אחוז בקרב הניטרליים). לכן נוצרה עמודה נוספת - *Has_Investment* - עמודה בינארית המחווה האם מאזן ההשקעות שונה מאפס. עמודות אלו נוצרו מכיוון שיש ביכולתן לשפר תוצאות של מודלים מבוססי עצים, באותו אופן כפי שפורט קודם לכן בנוגע ל- "קיבוץ" משתנים.

<u>זיהוי מגמות נוספות</u>

- בעמודות C ו־E נמצאה קורלציה לינארית חזקה.[8] שתי העמודות כללו מספר גדול יחסית של ערכים חסרים. מכיוון שמשמעות העמודות וחשיבותן למודל אינה ידועה, הוחלט ליתר ביטחון שלא למלא את הערכים החסרים באמצעים פשטניים (כגון חציון או ממוצע), שעלולים לפגוע באינפורמציה שבנתונים.
במקום זה, נעשה שימוש באלגוריתם MICE,[9] המבוסס על איטרציות של חיזוי ערכים חסרים באמצעות רגרסיות על משתנים קיימים. השיטה מאפשרת לשמר את מבנה הקשר בין C ל־E, ואף מצליחה להשלים ערכים חסרים כפולים על בסיס מידע חלקי.
לבסוף, נוצרו שתי עמודות חדשות (C_imputed, E_imputed) ללא ערכים חסרים. בעקבות הקורלציה הגבוהה הוחלט לכלול במודל רק את אחת מהן על מנת להימנע ממידע חופף ולשפר את יכולת ההכללה של המודל. העמודה שנבחרה היא C_imputed משום שמספר הערכים החסרים ב- C קטן משמעותית מ- E, ועל כן עמודה זו מכילה יותר תצפיות מקוריות ופחות "קירובים".
- ישנן עמודות קטגוריאליות לא מאוזנות, כלומר קיים תיוג שמהווה רוב מובהק במדגם. לדוגמא: שיעור הגברים במדגם הוא פי שניים בקירוב משיעור הנשים, בעמודת המדינה רוב התצפיות הם אזרחי USA וכו'. עובדה זה יכולה לפגוע בביצועי המודל ולגרום להטיה שלו. לכן בעת בחינת ביצועי המודלים השונים בחרנו להשתמש ב- StratifiedKFold, שלמעשה מבצע Kfold, אך בכל חלוקה שומר על ההתפלגות היחסית של כל משתנה זהה להתפלגותו בדאטה המקורי. כך כאשר אנו מבצעים Grid Search התוצאה שנקבל תהיה אופטימלית עבור נתונים שמתפלגים כמו הדאטה שלנו, ולא עבור נתונים שמתפלגים בצורה אקראית, וכך דיוק המודל משתפר .
- עמודת המטרה imbalanced, כלומר ישנן הרבה יותר תצפיות שליליות מחיוביות, דבר אשר יכול לפגוע בדיוק המודל וביכולת שלו לחזות תוצאה חיובית. על מנת לפתור בעיה זו בעת הפיצול לסט אימון וסט ולידציה הוגדר הפרמטר stratify וכן בחלק מהמודלים נעשה שימוש במשקולות על מנת לאזן.[10]
- המשתנה Years_of_Education לא נותח ועובד תחילה, מכיוון שעלה החשד שהוא מקביל למשתנה Education_Level. לאחר המרת המשתנה Education_Level לאורדינלי אושש החשש ונמצא שהקורלציה בין העמודות שווה ל- 1. על כן אין צורך בשתי העמודות. נבחר להשתמש ב- Education_Level מכיוון שאין בה ערכים חסרים והיא יותר אינפורמטיבית.
- ככל הנראה יש קורלציה גבוהה בין Household_Role לבין עמודות אחרות כגון Marital_Status ו- Gender. קורלציה זו נובעת מהבנה של התיוגים השונים בקטגוריות. לדוגמא גבר שמופיע כנשוי ב- Marital_Status ככל הנראה יתוייג כ- Husband בעמודת Household_Role. לכן במודלים כגון רגרסיה לוגיסטית הסרת חלק מהמאפיינים הנ"ל יכולה לשפר את מימד הבעיה מבלי לאבד שונות רבה.

---

[7] ראה נספח 1.4
[8] ראה נספחים 1.5 – 1.7
[9] Multiple Imputation by Chained Equations
[10] יפורט בהמשך

פיצול הנתונים ל - Train ו- Validation

לבסוף, פוצלו הנתונים ל- Train (80% מהתצפיות) ו- Validation (20%). בעת הפיצול הוגדר בפונקציה הפרמטר: startify=y, כאשר y הוגדר כעמודת המטרה. פעולה זו משמרת את היחס בין תצפיות חיוביות לשליליות בעמודת המטרה, כך שיחס זה בסט האימון וסט הולידציה יהיה זהה ליחס בדאטה המקורי. פעולה זו מונעת הטיה של המודל ופגיעה בתוצאותיו בעקבות חוסר האיזון של עמודת המטרה.
לאחר מכן לכל אחד מהסטים הותאמו שתי גרסאות - גרסא עבור מודלים מבוססי עצי החלטה, וגרסא עבור מודלים רגישים יותר (רגרסיה לוגיסטית ו- ANN). הסיבה לפיצול זה היא שכפי שצויין קודם לכן, מודלים כגון רגרסיה לוגיסטית ו- ANN רגישים לערכים חריגים, נתונים שאינם מנורמלים יכולים לפגוע בהם משמעותית, וכן נוטים לאוברפיטינג בקלות כאשר ישנם נתונים בעלי קורלציה גבוהה. השפעה של תופעות אילו נמוכה בהרבה במודלים מבוססי עצי החלטה, ופעולות המניעה שלהן (לדוגמא הסרת ערכים חריגים או קיצוצם) אף יכולות לפגוע ביכולת המודל לזהות תופעות ודפוסים מורכבים בדאטה, ובכך לפגוע בתוצאותיו. כך ניתן להתאים עמודות שונות לסוגי מודלים שונים על מנת למקסם את דיוק המודלים.
בשלב האחרון בוצע encoding לעמודות הקטגוריאליות בכל גרסא.

חלקים 4 + 3 - הרצת המודלים והערכתם

בסיום הכנת ועיבוד הנתונים נבחנו ארבעה מודלים, כאשר המדד להערכתם הוא AUC. עבור כל מודל נבחנו פרמטרים שונים בעזרת GridSearch על מנת לבחור בפרמטרים האופטימליים. על מנת למקסם את אופטמיליות הפרמטרים ה- GS הופעל על כלל הנתונים (סט האימון + סט הולידציה). לאחר בחירת הפרמטרים כל מודל אומן עם הפרמטרים שנבחרו על סט האימון, ולאחר מכן ביצע תחזיות על סט הולידציה ועל סט האימון. על פי תחזיות אלו חושבו ציוני AUC, אשר הודפסו על מנת לבחון את טיב המודל והאם המודל overfitted. לבסוף, על מנת לקבל תמונה כללית אף יותר של טיב המודל, בוצעה הערכה נוספת שלו באמצעות K-Fold Cross Validation עם ערך K=5.

חשוב לציין כי נוסו פרמטרים נוספים מלבד אלו שמופיעים בקוד, וכן נוסו קומבינציות שונות של עמודות שכל מודל התאמן עליהן.
לבסוף הוחלט לכלול בקוד את הפרמטרים הסופיים שנבחנו ואת העמודות שנמצאו כאופטימליות לאימון המודלים, זאת על מנת לשמור על סדר בקוד ולחסוך בזמן ריצה.
תוצאות המודלים:

Logistic Regression

בחירת פרמטרים: ל- Grid Search הוזנו ערכים שונים על מנת למצוא את הפרמטרים האופטימליים. בין היתר, נוסו מספר ערכים עבור חוזק הרגולריזציה (C)- ככל שהערך של פרמטר זה גדל הרגולריזציה חלשה יותר, וגדל הסיכוי לאוברפיטינג וכאשר הוא קטן גדל הסיכוי לאנדרפיטינג. לכן נוסו מספר ערכים שונים על מנת למצוא ערך שמאזן תופעות אלו בצורה הטובה ביותר. בנוסף נוסו שיטות שונות של penalty וסולברים שונים, על מנת למצוא את הקומבינציה שמספקת את התוצאות הטובות ביותר וממקסמת את ציון ה- AUC. מספר האיטרציות המקסימלי (max_iter) ותנאי העצירה של האלגוריתם הוגדרו כברירות המחדל, משום שהיה נראה שדי בהם על מנת שהאלגוריתם יתכנס בצורה טובה, וניסיון להגדיל את האיטרציות או להקטין את תנאי העצירה היו מעלים משמעותית את זמן הריצה ללא שיפור משמעותי בתוצאות.

הפרמטרים הנבחרים: { C': 0.1, 'max_iter': 100, 'penalty': 'l2', 'lbfgs': 'liblinear', 'tol': 0.0001 '}
תוצאות המודל: Validation AUC = 0.8945, Train AUC = 0.8954, תוצאות אלו מעידות על כך שהמודל אינו overfitted.
כמו כן ,בבדיקה הסופית(Kfold)[11] התקבל **ערך ה- AUC: 0.8938.**

[11] ראה נספח 2.0

Decision Tree

בחירת פרמטרים: על מנת למצוא פרמטרים אופטימליים נוסו מספר ערכים שונים של עומק מקסימלי (max_depth) על מנת לאפשר למודל גמישות מחד אך למנוע אוברפיטינג ולקצר את זמן הריצה. כמו כן, נבחנו שני קריטריונים לקביעת הפיצולים (criterion) על מנת לבחור בקריטריון שמספק תוצאות טובות יותר. בנוסף, נבחנו ערכים שונים למספר התצפיות המינימלי בעלה (min_samples_split) ובפיצול (min_samples_split), זאת על מנת להגיע לאיזון בין דיוק הפיצולים ותוצאותיהן לבין ההכללה של המודל והפחתת הסיכוי לאוברפיטינג. מכיוון שעמודת המטרה imbalanced, הוחלט להשתמש במודל במשקולות מאוזנות על התצפיות ('class_weight = 'balanced). פרמטר זה מוסיף לתצפיות משקולות על מנת לתקן את חוסר האיזון בעמודת המטרה, ובכך להקטין את הסיכוי להטיה במודל. לבסוף הוחלט שלא להגביל את מספר המאפיינים המקסימלי (max_features = None), על מנת לאפשר למודל לזהות מגמות מורכבות.

הפרמטרים הנבחרים: {'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 10, 'max_features': None, 'min_samples_leaf': 10, 'min_samples_split': 2}

תוצאות המודל: Validation AUC = 0.8993, Train AUC = 0.9238. ניתן לראות שישנו פער בין הציונים, אך הוא קטן (פחות מ- 0.3) מכאן נסיק שלמודל יש נטייה קלה מאוד ל- overfitting. ניתן להגדיל את ההכללה של המודל הן על ידי שינוי הפרמטרים של המודל (למשל: הקטנת max_depth, הגדרת max_features, הגדלת min_samples_split ו-min_samples_leaf וכו'), והן על ידי החלפת חלק מהמאפיינים במאפיינים כלליים יותר (לדוגמא: החלפת המאפיין Age ב- Age_Group). בנוסף, קיצוץ של התצפיות החריגות ונרמול גם היו יכולים לשפר את הכללת המודל. עם זאת, בבדיקה הסופית(Kfold)[12] התקבל **ערך ה- AUC: 0.8945.** ישנם מודלים אחרים שהגיעו לתוצאות טובות בהרבה, ועל כן לא נעשה מאמץ מיוחד על מנת לשפר את המודל.

ANN

בחירת פרמטרים: על מנת למצוא פרמטרים אופטימליים נוסו מספר גדלים שונים של שכבות חבויות (hidden_layer_sizes) וכן שתי פונקציות אקטיבציה (activation), על מנת למצוא את הקומבינציה האופטימלית שמצליחה להתחקות אחר תופעות מורכבות בדאטה מחד, אך מאידך שומרת על הכללת המודל. כמו כן ה- batch_size שנבחר הוא 100 מכיוון שזהו גודל בינוני כך שמצד אחד יהיו מספיק תצפיות ב-batch, אך מצד שני זמן הריצה יהיה סביר. מספר האיטרציות המקסימלי (max_iter) שנבחר הוא 1000, על מנת לאפשר למודל להתכנס לתוצאה טובה, אך עם זאת לשמור על זמן ריצה סביר. הסולבר (solver) שנבחר הוא sgd מכיוון שהוא היחיד שלמדנו בקורס. לבסוף, אלפא (alpha) וקצב הלמידה ההתחלתי (learning_rate_init) הוגדרו כברירת המחדל שלהם.

הפרמטרים הנבחרים: {'activation': 'relu', 'alpha': 0.0001, 'batch_size': 100, 'hidden_layer_size': (32,16),'learning_rate_init': 0.001, 'max_iter': 1000, 'solver': 'sgd'}

תוצאות המודל: Validation AUC = 0.9061, Train AUC = 0.9137. ניתן לראות שההפרש קטן מ-0.01 מה שמעיד על כך שהמודל אינו overfitted. בבדיקה הסופית (KFold)[13] התקבל **ערך ה-AUC : 0.9082.**

Random Forest

בחירת פרמטרים: ראשית, נוסו ערכים שונים עבור מספר העצים (n_estimators) על מנת לקבל תוצאות יציבות ולהגדיל את הכללת המודל אך גם לשמור על זמן ריצה סביר. בנוסף נוסו ערכים שונים של קריטריונים ועומקים מקסימליים. בניגוד לעץ ההחלטה, העומקים במודל זה היו גדולים יותר משמעותית, כי המודל מסווג על פי מספר גדול של עצים ולא על פי עץ אחד. לכן ככל הנראה ניתן להשתמש בעומקים גדולים יותר על מנת להגדיל את השונות מבלי לפגוע משמעותית בהכללה ולהגדיל את הסיכויים לאוברפיטינג. בדומה לעץ ההחלטה, נוסו ערכים שונים של min_samples_split ו- min_samples_leaf, וכן הוגדר class_weight = 'balanced'. לבסוף, בניגוד לעץ ההחלטה, במודל זה הוגדר 'max_features = 'sqrt מכיוון שהגבלה של מספר המאפיינים בכל עץ מוסיפה למודל רנדומליות, ובכך מעלה את הכללת המודל ומפחיתה את הסיכוי לאוברפיטינג.

---

[12] ראה נספח 2.1
[13] ראה נספח 2.2

פרמטרים נבחרים: {'class_weight': 'balanced', 'criterion': 'entropy', 'max_depth': 30, 'max_features': 'sqrt', 'min_samples_leaf': 3, 'min_samples_split': 10, 'n_estimators': 1700}

תוצאות המודל: Validation AUC = 0.9153, Train AUC = 0.9826. ניתן לראות כי ישנו הפרש גדול יחסית (מעל 0.06) בין הציונים. אם כן, המודל בעל נטייה יחסית גדולה ל- overfitting. כפי שנאמר בהקשר של עץ ההחלטה, היה ניתן לשפר את הכללת המודל על ידי קיצוץ תצפיות חריגות ונרמול, החלפת חלק מהמאפיינים עליהם אומן המודל במאפיינים כלליים יותר ושינוי פרמטרים (בדומה לעץ ההחלטה, וגם ישנה האפשרות להגדיל את מספר העצים.) בבדיקה הסופית (Kfold)[14] התקבל **ערך ה- AUC: 0.9176.** זהו המודל עם הערך הגבוה ביותר שהתקבל בפרויקט ועל כן הוא נבחר על מנת לספק תחזית לסט המבחן. העובדה שתוצאות הבדיקה הסופית בעזרת KFold היו קרובות לתוצאות על סט הולידציה מראה כי על אף הנטייה ל- overfitting המודל מספק תוצאות טובות מאוד, ועל כן הוחלט להשאירו בקונסטלציה הנוכחית ולא לבצע שינויים דרסטיים על מנת להוריד את ה- overfitting.

Confusion matrix: לאחר הרצת המודל נבנתה confusion matrix.[15] בשלב זה של המודל הועדף מזעור ה- FPR גם במחיר של ירידה קלה בציוני המודל האחרים, זאת מכיוון שבמקרה בו מאושרת הלוואה לאדם שלא יכול להחזירה - הארגון מפסיד יותר, אל מול מקרה דחייה של לקוח שיחזיר את ההלוואה .שיעור ה FPR במטריצה היה 0.124, כאשר מתוך 4203 תצפיות שליליות כ- 521 סווגו כחיוביות, ובאופן כללי מתוך 5536 תצפיות כ - 841 סווגו בצורה שגויה. על כן, נעשה ניתוח של thresholds שונים בתחזית המודל על מנת לשפר את התוצאות.[16] נמצא כי הסף האופטימלי הוא 0.7, עבורו התקבל FPR של 0.0428, כאשר רק 180 תצפיות שליליות סווגו כחיוביות.[17] עם זאת כ -745 תצפיות חיוביות בלבד סווגו בצורה נכונה, זאת לעומת 1013 ברף המקורי. על כן מכיוון שמודל זה מחזיר וקטור הסתברויות ולא תחזית בינארית, נמליץ לקבוע את רף התחזית על 0.7 על מנת למקסם את פוטנציאל הרווח של הארגון ולהקטנת הסיכון העסקי.

סיכום

בפרויקט זה נבנה מודל מקיף על מנת לחזות האם בקשת הלוואה תידחה או תאושר.

ראשית בוצע ניתוח מעמיק של הנתונים על מנת לזהות תופעות ומגמות בנתונים. המסקנות שנרכשו בשלב זה באו לידי ביטוי בשלב עיבוד הנתונים, ואפשרו לטפל בהם בצורה אופטימלית. כך הושלמו ערכים חסרים, בוצע נרמול ונעשו מניפולציות נוספות על עמודות בצורה המושכלת ביותר.

לאחר מכן נבחנו ארבעה מודלים שונים ולבסוף נמצא כי Random Forest סיפק את התוצאות הטובות ביותר, זאת על אף שהוא נטה מעט ל- overfitting. על כן בוצעה תחזית על סט המבחן בעזרת מודל זה, וסופקו המלצות לקביעת הרף במודל על מנת למקסם את פוטנציאל הרווח בארגון. נציין כי גם שאר המודלים הגיעו לתוצאות טובות, וקיימות מגוון דרכים לשפרן אף יותר.

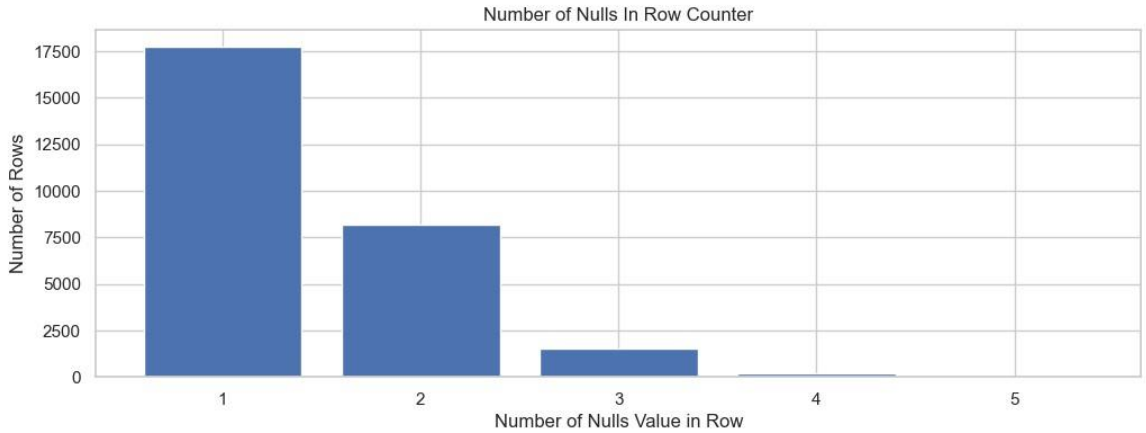
---

[14] ראה נספח 2.3
[15] ראה נספח 2.4
[16] ראה נספח 2.5
[17] ראה נספח 2.6

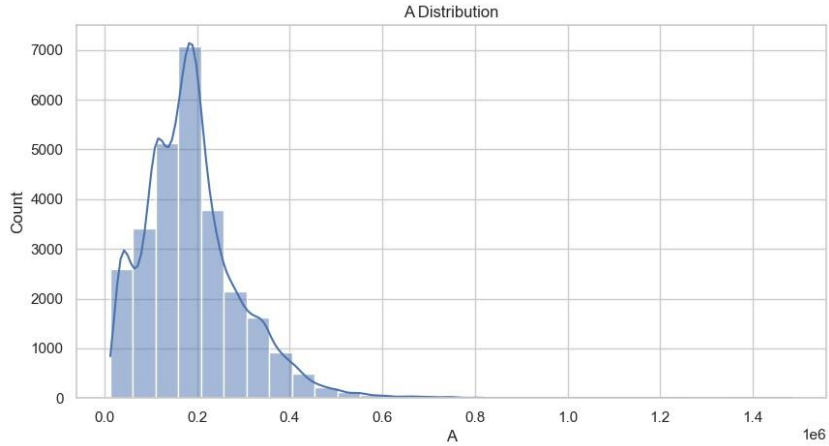
נספחים

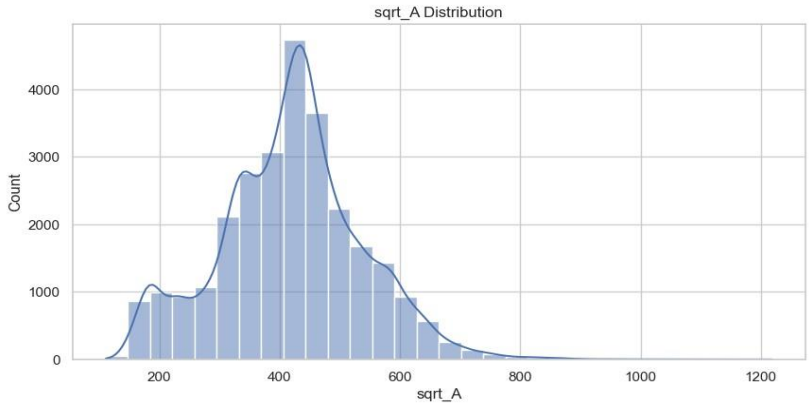
נספח 1- אקספלורציה

- **נספח 1.0 -** היסטוגרמה של מספר הערכים החסרים בכל שורה

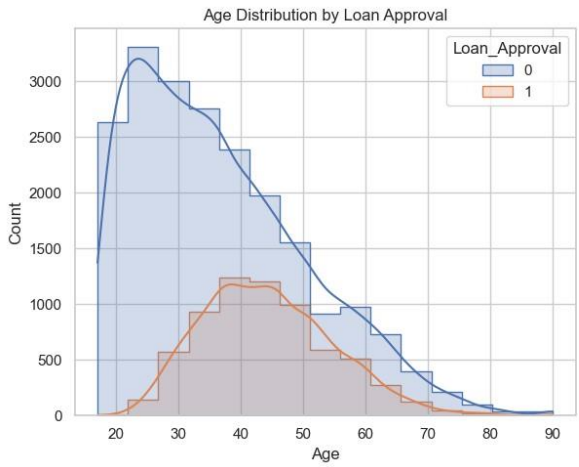


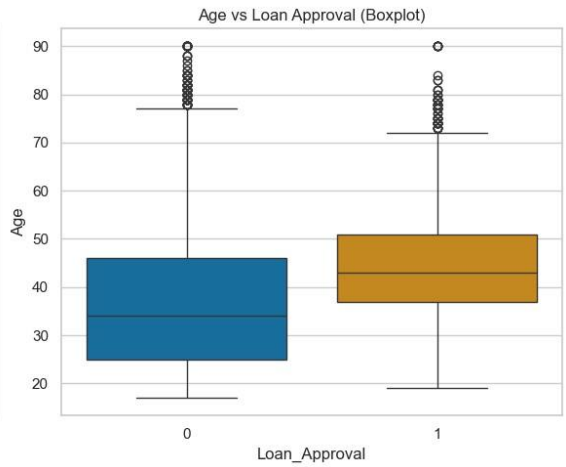
- **נספח 1.1 -** התפלגות עמודה A לפני טרנספורמציה והתפלגות עמודה sqrt_A לאחר הטרנספורמציה

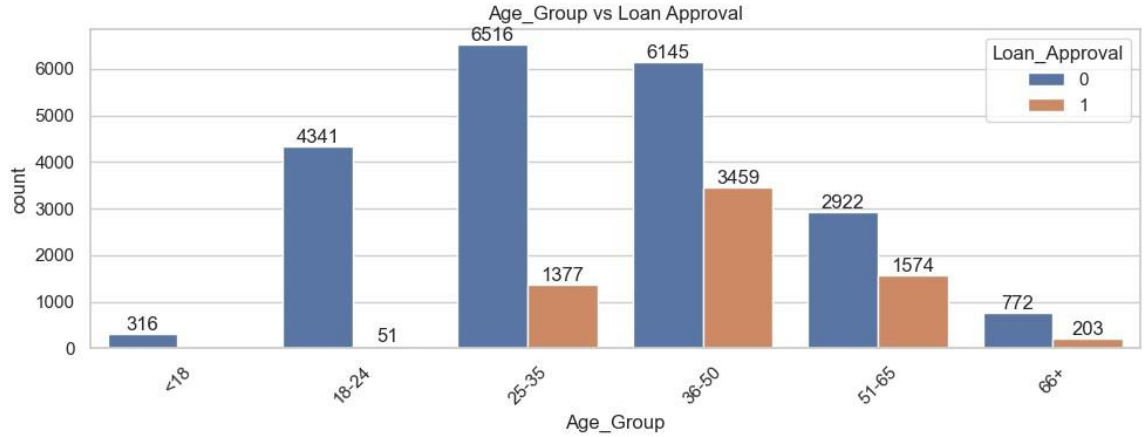




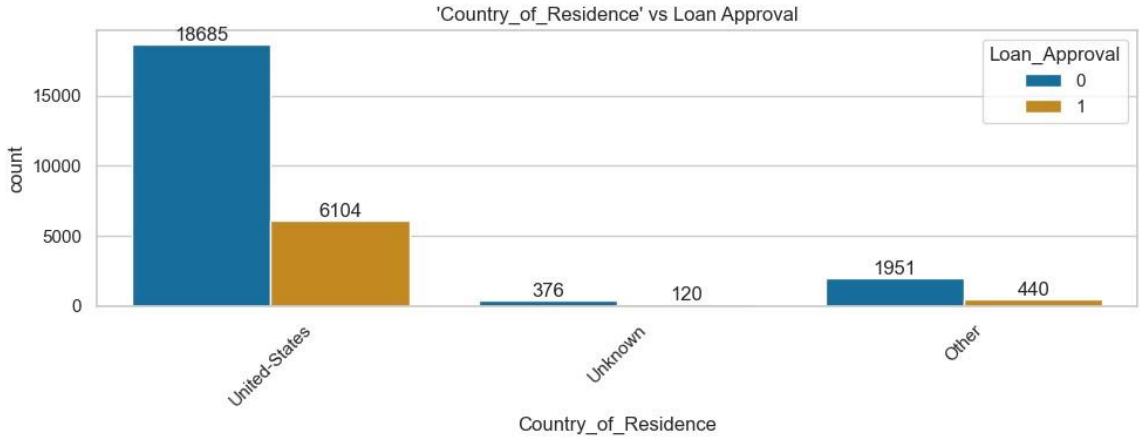
נספח 1.2 - התפלגות העמודה Age (מפוצלת לפי ערך עמודת המטרה), והתפלגות העמודה לאחר חלוקה לקבוצות קטגוריאליות (Age_Group)



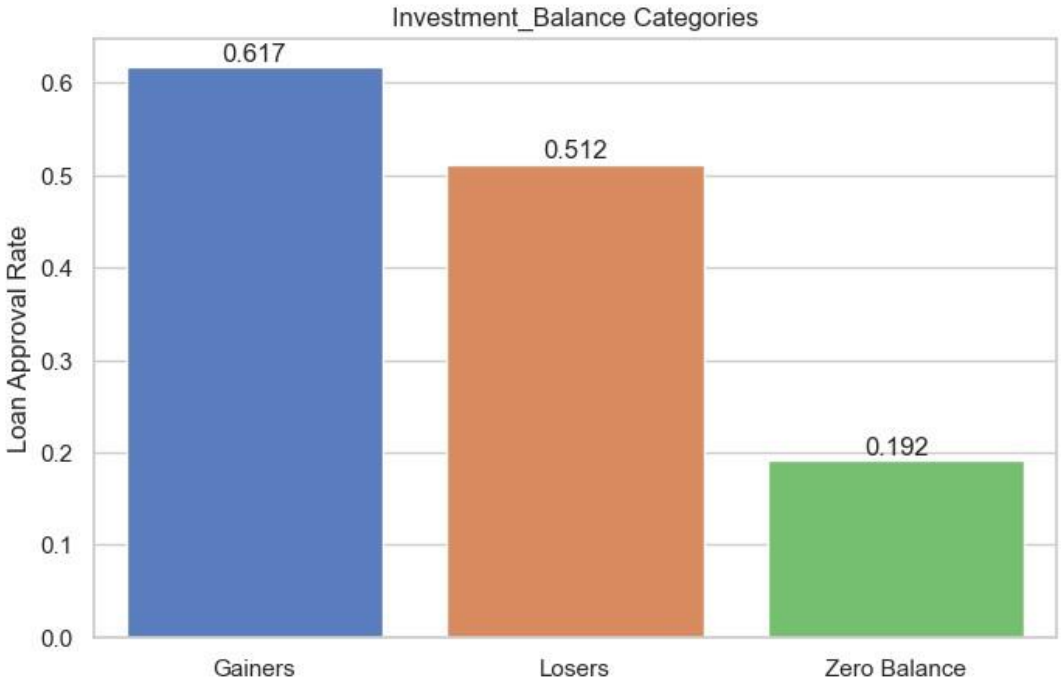




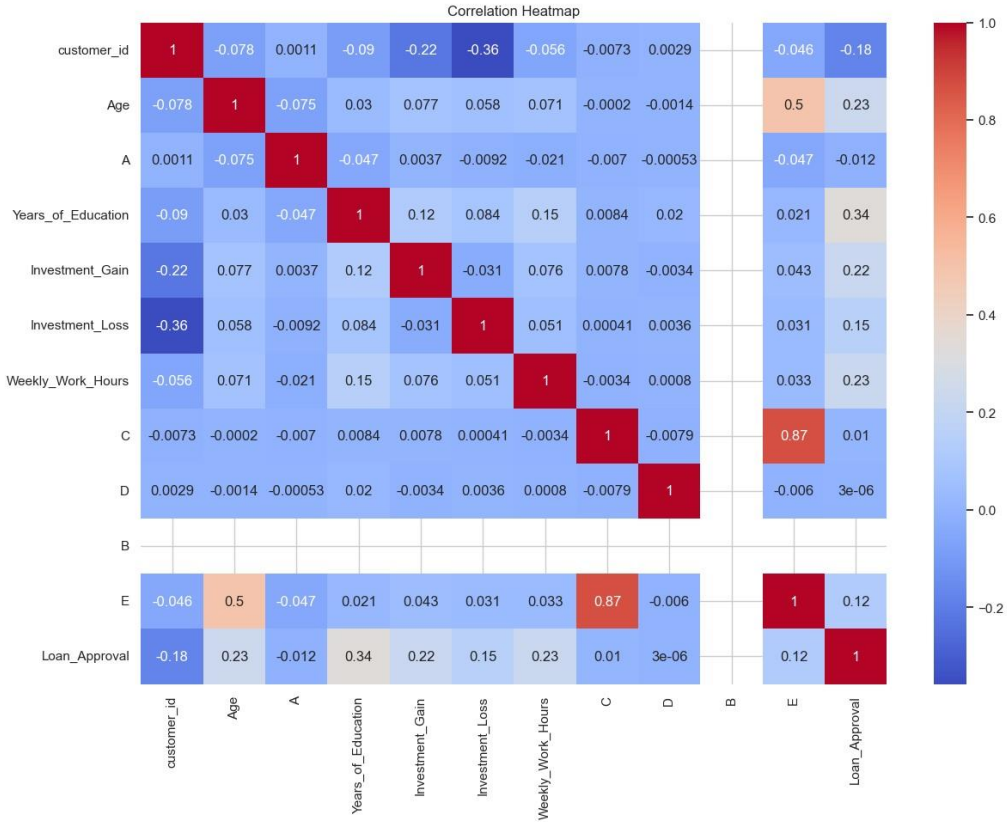
- **נספח 1.3 -** התפלגות עמודת Country_of_residence לאחר קיבוץ הערכים



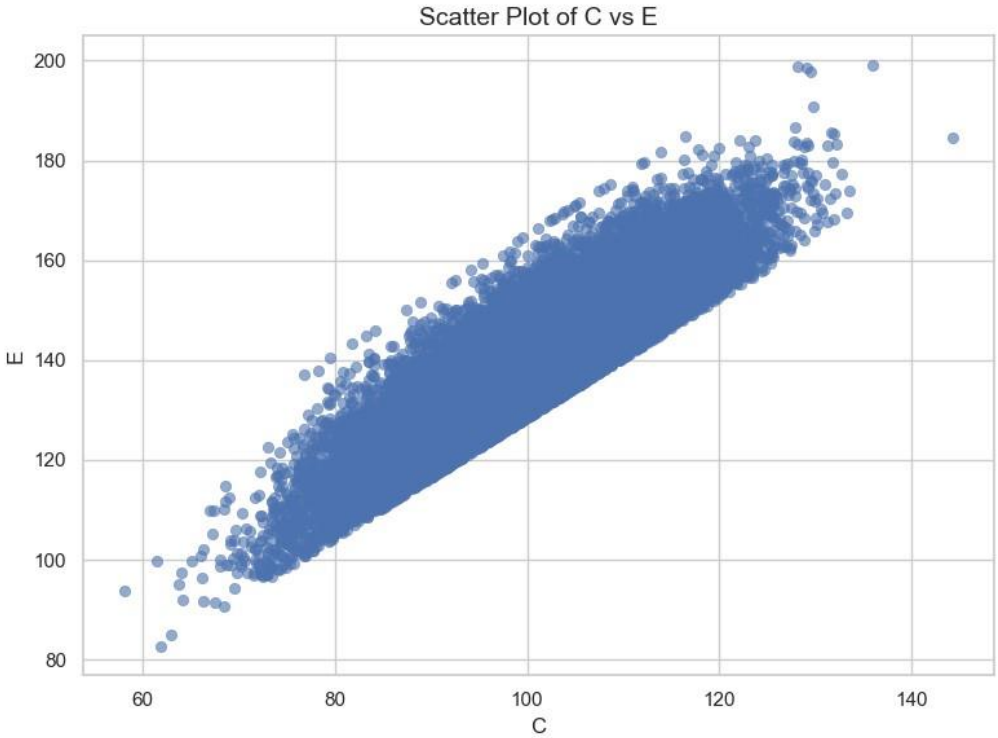
נספח 1.4 - אחוזי אישור ההלוואה בקרב אוכלוסיית המפסידים מרוויחים ונטרליים ב- Investment_Balance



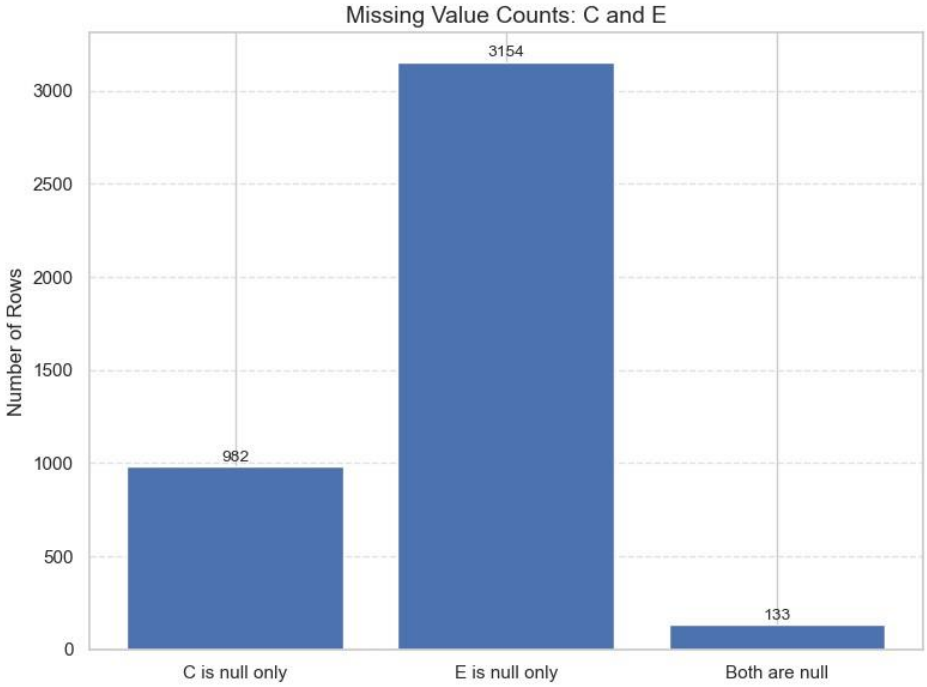
- **נספח 1.5** - מטריצת הקורלציות של המאפיינים הנומריים **לפני** עיבוד הנתונים



- **נספח 1.6** – scatter plot של E ביחס ל- C

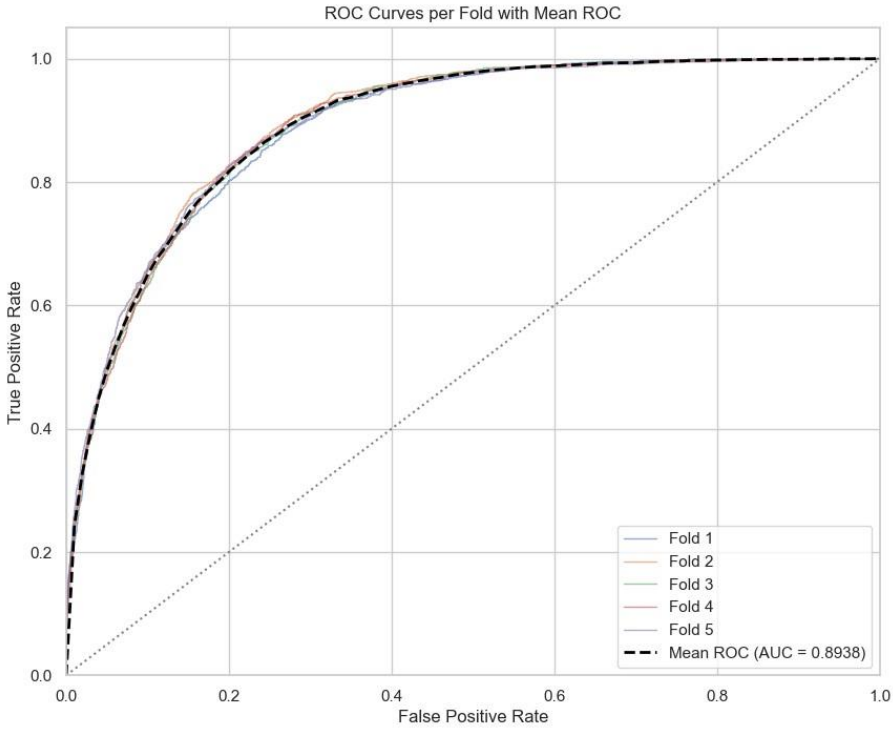


- **נספח 1.7** - מספר התצפיות עם ערכים חסרים בעמודות C,E ושתי העמודות

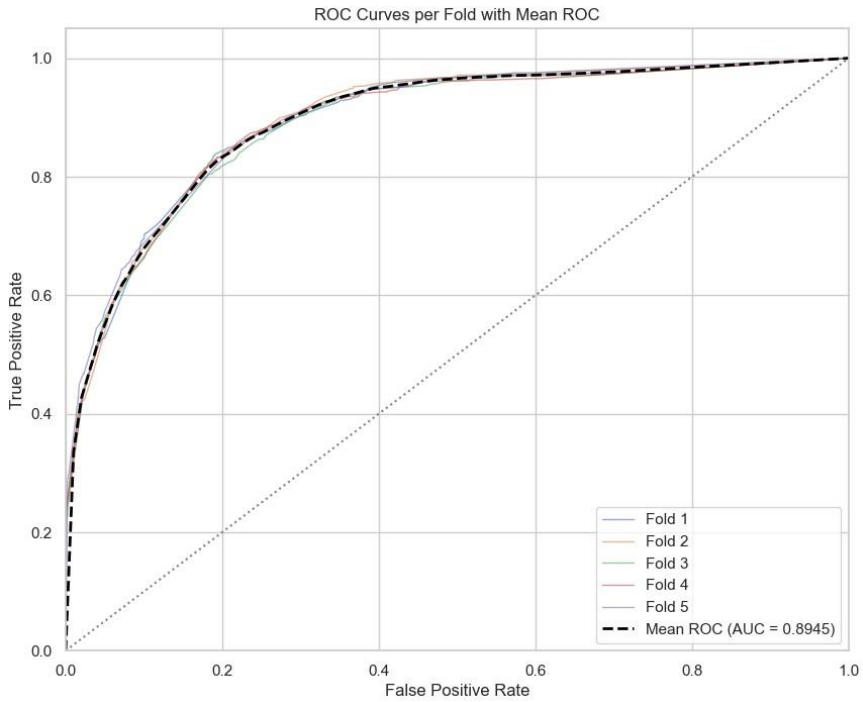


נספח 2- תוצאות המודלים

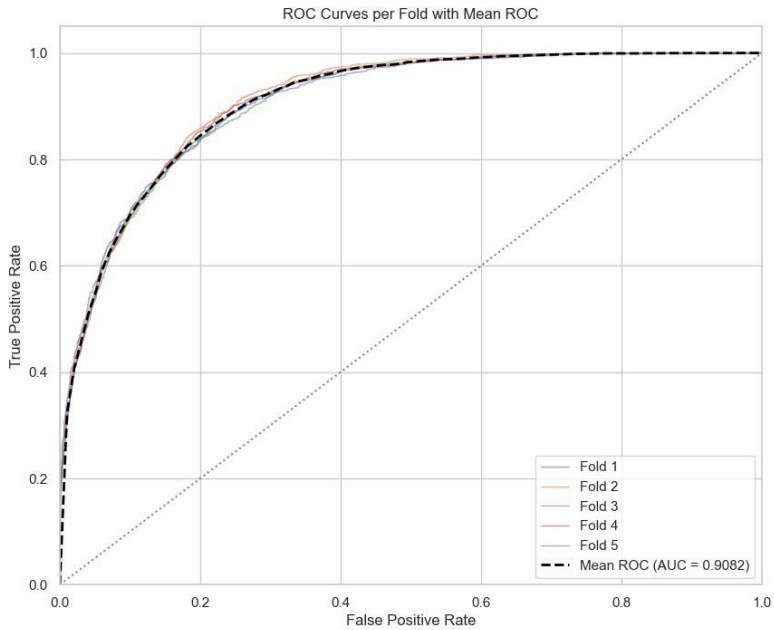
- **נספח 2.0** - פלט ROC וציון AUC ממוצע של מודל הרגרסיה הלוגיסטית



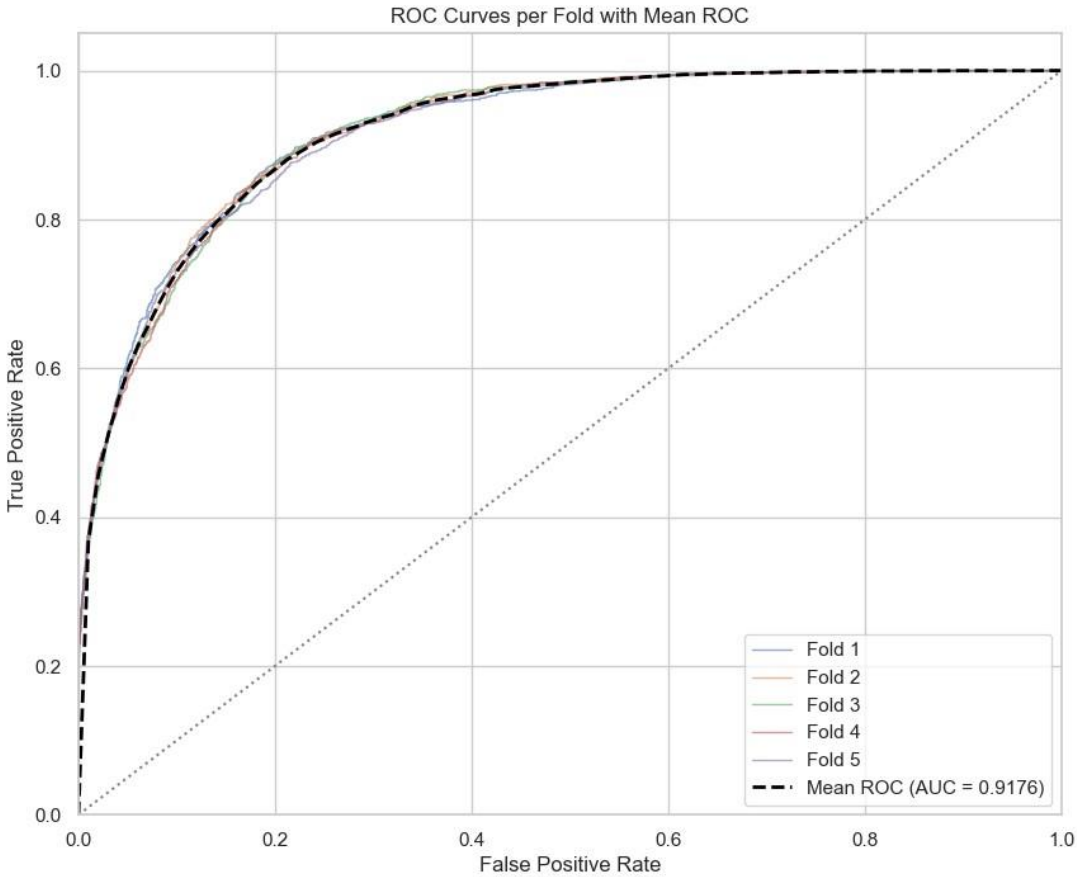
- **נספח 2.1** - פלט ROC וציון AUC ממוצע של מודל Decision Tree



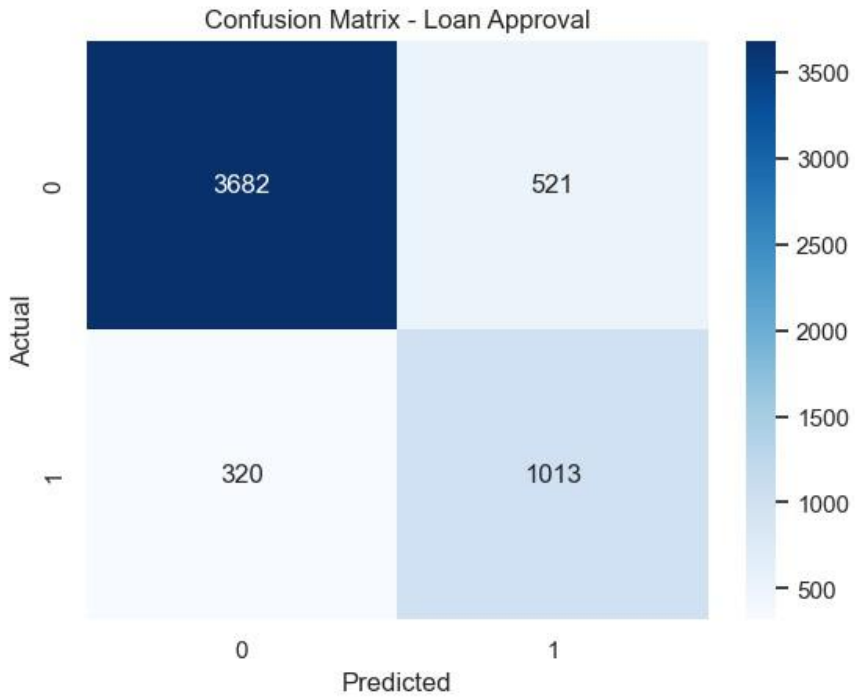
- **נספח 2.2** - פלט ROC וציון AUC ממוצע של מודל ANN



- **נספח 2.3** - פלט ROC וציון AUC ממוצע של מודל Random Forest



- **נספח 2.4 -** confusion matrix של המודל עם הרף המקורי



- **נספח 2.5 -** confusion matrix של המודל עם הרף החדש (0.7)