אוניברסיטת
אריאל
בשומרון

סיווג סיכון הלוואות אשראי

GitHub : https://github.com/avivyefet/Credit_risk_classification

מגישים : חנן ג'קובס 316090877
אביב יפת 208495267

1. מבוא :

מטרת הפרויקט היא בהינתן משתנים שונים של אדם המבקש הלוואה מגוף פיננסי, לדעת לסווג את הלווה ולחזות האם הוא יוגדר כהלוואה טובה או רעה.

מבנה הפרויקט בנוי כך שראשית ביצענו חקר מעמיק אודות הפיצרים, הבנתם הכרת עולם התוכן הפיננסי וההשלכות השונות של כל פרמטר על מנת להבין טוב יותר איך כל דבר עלול להשפיע על הערך החזוי. לאחר חקר הנתונים, ביצענו את השלבים הבאים : עיבוד מקדים לנתונים על מנת להתאימם למודלים, בחינת קורולציות בין המשתנים הלא תלויים, חלוקת הנתונים לנתוני אימון ובדיקה, יישום SMOTE על סט האימון, כשלבסוף יישומנו מודלים של סיווג לקבלת התוצאות והסקת המסקנות.

2. מערך נתונים וEDA

מערך הנתונים מורכב מ- 1000 לקוחות ו-21 משתנים המתארים את דמוגרפיית הלווים ונתונים אודות אשראי. מערך הנתונים אינו מכיל ערכים חסרים, מה שמבטיח את מהימנות הניתוח של הנתונים.

מערך הנתונים מחולק ראשית למשתנה תלוי שהוא ערך המטרה שלנו-class אשר מתאר האם הלוואת הלקוח מסווג כהלוואה שאינה טובה או כהלוואה טובה. ניתוח ראשוני של ערך המטרה שלנו מגלה ש-30% מההלוואות מסווגות כגרועות, בעוד ש-70% הנותרים מסווגות כהלוואות טובות *(ראה איור מס' 1).*

המשתנים הבלתי תלויים מתחלקים למשתנים נומריים ומשתנים קטגוריאליים. המשתנים הקטגוריאליים הינם :

checking_status, credit_history, Purpose, employment, savings_status, other_parties, personal_status, property_magnitude, other_payment_plans, housing, job

המשתנים הנומריים הינם: duration, credit_amount, existing_credit, Age. הסבר אודות המשתנים ניתן לראות בנספח *(ראה טבלה מס' 1)*

האתגר הראשון בפרויקט היה לבצע EDA חקר וניתוח על מערך הנתונים במטרה להבין את המשתנים התלויים והמשתנה התלוי בצורה מעמיקה, להבין מה עומד מאחורי כל משתנה, מה המשמעות של התצפיות השונות המופיעות בנתונים ומה ניתן להסיק מאותם המשתנים.

עבור כל משתנה נומרי הצגנו את התפלגות המשתנה על ידי היסטוגרמה, גרף Boxplot ו ובעזרת פונקציית describ חישבנו ערכים סטטיסטיים – mean, std, min, max, 25%, 75%. על פי חקר הנתונים ניתן לראות כי הגיל הממוצע של מבקשי הלוואות הוא 35.5 שנים ורוב מבקשי ההלוואות הם בני 25-30 *(ראה איור מס' 2)*. 75% מההלוואות הם עד 3972 וגודל ההלוואה הממוצע הוא בסביבות 3270*(ראה איור מס' 3)*. משך ההלוואה הממוצע משתרע על פני כ-21 חודשים *(ראה איור מס' 4)* . נתונים אלה מספקים רקע משמעותי לעולם התוכן שנרצה לנתח.

כחלק מחקר הנתונים חילקנו את הגיל לעשורים והצגנו את התפלגות ההלוואות המסווגות כטובות וגרועות לפי העשורים השונים *(ראה איור מס' 5)* ניתן לראות כי מערך הנתונים מכיל לקוחות מהעשור 10-19 ועד העשור 70-79 וכי רוב הלקוחות בכל העשורים סווגו כהלוואות טובות.

מניתוח סיווג הלוואות לפי מקטעי סכום ההלוואה לפי מספר דגימאות זהה בכל מקטע נראה כי חמשת המקטעים הראשונים, בהם ההלוואות עד 5509, מציגים התנהגות דומה יחסית. עם זאת, המקטע השישי בולט עם שיעור גבוה יותר של הלוואות

המסווגות כגרועות. ממצא זה מצביע על כך שסכומי הלוואות גבוהים יותר במגזר זה עשויים להיות קשורים לסיכון מוגבר או לגורמים אחרים המשפיעים על ביצועי ההלוואה *(ראה איור מס' 6)* .

מניתוח סיווג ההלוואות לפי מקטעים של משך ההלוואה, ניתן לראות כי ככל שמשך ההלוואה גדלה כך אחוז ההלוואות המסווגות כהלוואות שאינן טובות גדל וכי בהלוואות שמשך ההלוואה בין 36 ל72 חודשים רוב ההלוואות כ52% סווגו כהלוואות שאינן טובות *(ראה איור מס' 7)*.

עבור המשתנים הקטגוריאליים הצגנו את התפלגות המשתנים ואת אחוז סיווג ההלוואה עבור כל אחד מערכי הקטגוריה.

בבחינת הדמוגרפיה של הלווים, אנו למדים שרוב מוחלט של הלווים הם עובדים זרים *(ראה איור מס' 8)*.

יתרה מכך, מעניין לציין שכ-70% מהלווים הם גברים, בעוד 30% הנותרים הם נשים *(ראה איור מס'9)*.

בנוסף לכך ראינו כי רוב הלקוחות מועסקים בין 1 ל4 שנים והם נחשבים למיונים בעבודתם וכי לרוב הלקוחות היו הלוואות בעבר שכבר שולמו.

3.הכנת מערך הנתונים ו מתודולוגיה:

3.1 הכנת מערך הנתונים:

לאחר בחינה מעמיקה והבנת מערך הנתונים ביצענו הכנה של מערך הנתונים. בניתוח EDA ראינו כי המשתנה 'personal_status' מכיל מידע אודות סטטוס הלקוח ומידע עבור מגדר הלקוח. בפרויקט החלטנו לחלק את המשתנה לשתי משתנים אשר האחד יבטא את מגדר הלקוח (male/female) והשני יבטא את סטטוס הלקוח.

בנוסף לכך, את המשתנה הנומרי 'age' הפכנו למשתנה קטגוריאלי על ידי חלוקה למקטעים, כאשר כל מקטע מתאר את עשור הגיל. לאחר חלוקה לעשורים קיבלנו 7 קטגוריות להלן ['10-19', '20-29', '30-39', '40-49', '50-59', '60-69', '70-79'].

את המשתנים הקטגוריאליים אשר מכילים שני ערכים הפכנו למשתנים בינאריים כאשר ערכי המתשנים יקבלו ערך של 1 או 0, לרבות המשתנה התלוי – סיווג ההלוואה , כאשר הלוואה המסווגת כהלוואה טובה מקבלת ערך של 1 והלוואה המסווגת כהלוואה רעה מסווגת כ0. עבור שאר המשתנים קטגוריאלים יישמנו One hot encoder. לאחר הכנת מערך הנתונים קיבלנו מערך נתונים אשר מכיל 67 עמודות ו1000 שורות.

3.2. בחינת קורלציה בין המשתנים הבלתי תלויים:

בחינת מולטיקולינאריות בין המשתנים הבלתי תלויים על ידי חישוב מתאם פירסון בין שני משתנים בלתי תלויים. בפרויקט נקבע ערך סף של 0.85 כאשר קורלציה גבוהה מערך הסף או נמוכה ממינוס ערך הסף תיחשב כקורלציה גבוהה בין המשתנים הבלתי תלויים. לאחר בחינה של הקורלציה בין המשתנים ראינו כי אין משתנים שערכם גבוהים או נמוכים מערך הסף.

3.3 חלוקת מערך הנתונים לסט אימון וסט בדיקה:

סט האימון מכיל 80% ממערך הנתונים וסט הבדיקה מכיל 20% ממערך הנתונים. לאחר מכן ביצענו סטנדרטיזציה על סט האימון וסט הבדיקה.

3.4 יישום SMOTE על סט האימון:

מניתוח הנתונים ראינו כי הדאטה אינו מאוזן, 30% מההלוואות מסווגים כהלוואה שאינה טובה. באמצעות מודלים ללמידת מכונה ננסה לנבא האם הלוואת הלקוח יסווג כהלוואה טובה או הלוואה שאינה טובה. כאשר נכניס למודל נתונים שאינם מאוזנים, המודל ילמד לסווג את ההלוואות הטובות ויתקשה לחזות את ההלוואות המסוכנות לחברה. על מנת לפתור את הבעיה השתמשנו בשיטה oversampling על ידי שימוש ב SMOTE אשר מייצר נקודות נתונים נוספות המבוססות על נקודות הנתונים המקוריים. בפרויקט בחרנו להגדיל את הנתונים של סט האימון כך שהמודל ילמד לסווג בצורה טובה יותר את הלקוחות עם הלוואה המסווגת כהלוואה שאינה טובה. עבור סט הבדיקה בחרנו לא לבצע איזון על הדאטה מכיוון שנרצה לבחון את המודלים עלפי הנתונים הגולמיים ולא לפי נקודות שהוספנו באופן מלאכותי.

3.5 יישום מודלים לסיווג והערכת המודלים

בפרויקט הושמו המודלים הבאים: Decision Tree, Random Forest, KNN Classifier, SVM, Logistic Regression, XGBoost, CatBoost. עבור כל מודל ביצענו אופטמיזציה לפרמטרים של המודלים על ידי GridSearchCV עם cv=5 במטרה לקבל את הפרמטרים האופטימליים שיביאו למודל חיזוי טוב יותר ומדוייק יותר. עבור כל מודל לסיווג הצגנו את עשרת המשתנים המשפיעים על סיווג המודל ואת הפרמטרים האופטימליים. הערכת המודלים התבצעה על ידי מדדי הערכה לסיווג: Precision, Recall, F1 וConfusion matrix.

4. תוצאות ומסקנות:

להלן המודלים האופטימיילם אשר יושמו בפרויקט:

מודל	פרמטרים אופטימליים
Decision Tree	{'criterion': 'entropy', 'max_depth': 9, 'min_samples_split': 5}
Random Forest	{'max_depth': 9, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
KNN Classifier	{'n_neighbors': 7, 'p': 1, 'weights': 'distance'}
SVM	{'C': 10, 'gamma': 0.1}
Logistic Regression	{'C': 10, 'penalty': 'l2'}
XGBoost	{'learning_rate': 0.01, 'max_depth': 11, 'n_estimators': 500}
CatBoost	{'depth': 4, 'iterations': 300, 'learning_rate': 0.1}

את תוצאות המודלים ניתן לראות בנספח מס' 6.2. על פי התוצאות ניתן לראות כי גם לאחר איזון הדאטה המודלים הצליחו בצורה טובה לנבא את סיווג ההלוואות אשר מסוגות כהלוואות טובות והתקשו לנבא את סיווג ההלוואות המסווגות כהלוואות שאינן טובות.

המודל אשר סיווג את הלקוחות בצורה הטובה ביותר הינו Random Forest, עבור מודל זה קיבלנו את מדדי הדיוק הגבוהים ביותר *(ראה איור מס' 10)*. עבור מודל זה קיבלנו F1 משוקלל של 0.79 כאשר עבור סיווג הלוואות שאינן טובות קיבלנו F1 של 0.62, ועבור הלוואות טובות קילבנו F1 של 0.86. בנוסף לכך ניתן לראות כי מתוך 60 הלוואות שאינן טובות 34 סווגו נכון ו261 סווגו הלוואות טובות למרות שאינן טובות ומתוך 140 הלוואות טובות סווגנו נכון 125 מההלוואות.

יתר על כן, בחנו את הפרמטרים אשר משפיעים על ניבאוי הסיווג, על פי המדל הנבחר הפרמטרים אשר השפיעו על הסיווג *(ראה איור מס' 17)*:

חשבון עובר ושב - מצב חשבון עובר ושב משפיע על סיווג ההלוואה, מניתוח הנתונים ראינו כי ככל שהסכום בעובר ושב גבוה יותר יותר הלוואות מסווגות כהלוואות טובות.

משך הלוואה - משך ההלוואה הינו משתנה קריטי בהחלטה האם ההלוואה תסווג כהלוואה טוב או שאינה בטוחה, ממצא זה מתיישב עם ניתוח EDA בו ראינו כי ככל שמשך ההלוואה גבוהה יותר כך הסיכוי לסווג את הלוואה כהלוואה שאינה טובה עולה.

דיור - סוג הדיור בו הלווה מתגורר. ראינו בניתוח הנתונים כי רוב הלקוחות מתגוררים בנכס שבבעלותם.

היסטוריית כרטיס האשראי - תיעוד של היסטוריית האשראי של הלווים, ניתן לראות כי כאשר היסטוריית הלווים מוגדרת כקריטית כלומר שיש ללווה הלוואות קיימות נוספות הדבר משפיע על סיווג ההלוואה.

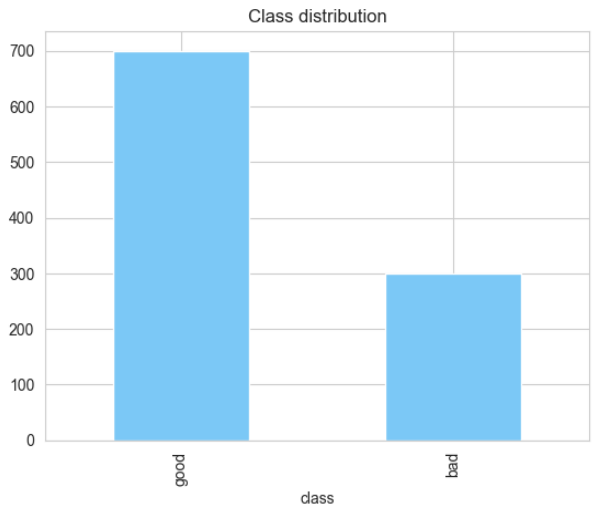
חשבון חיסכון- מתאר האם יש ללווה חשבון חיסכון ואת כמות החיסכון בחשבון, ראינו בנתיוח הנתונים כי כאשר אין ללקוח חשבון חיסכון או כאשר החיסכון של הלווה עד 500 הסיכוי שהלוקח יסווג כהלוואה שאינה טובה גבוה יותר מהלקוחות שיש להן חיסכון של 500 ומעלה.

6. נספחים :

6.1 נספחים – מערך הנתונים וEDA

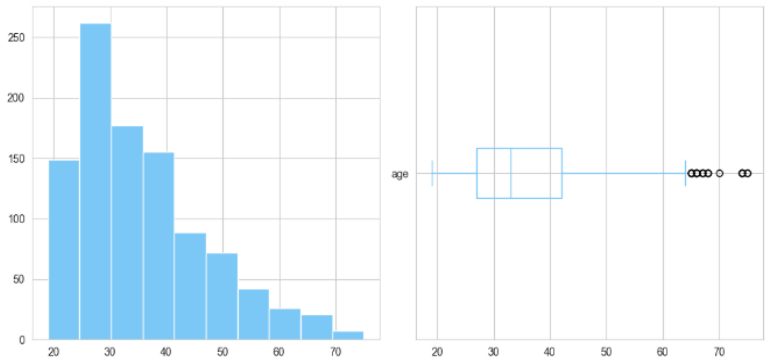
שם משתנה	הסבר	סוג משתנה	ערכי המשתנה
checking_status	הסטטוס של חשבון העובר ושב	קטגוריאלי	{ no checking,<0, 0<=X<200,>=200}
duration	משך ההלוואה בחודשים	נומרי	מספר
credit_history	היסטוריית אשראי, תיעוד של היסטוריית האשראי של הלווה	קטגוריאלי	{no credits/all paid, existing paid, delayed previously, critical/other existing credit, all paid}
Purpose	מטרת ההלוואה	קטגוריאלי	{new car, education, furniture/equipment, radio/tv, used car, business, domestic appliance, repairs, other, retraining}
credit_amount	סכום ההלוואה	קטגוריאלי	מספר
savings_status	סטטוס חשבון החיסכון של הלווה	קטגוריאלי	{>=1000, <100, 500<=X<1000, no known savings, 100<=X<500}
employment	תעסוקה, משך הזמן בו הלווה מועסק בשוק העבודה.	קטגוריאלי	{unemployed, 1<=X<4, 4<=X<7, >=7, <1}
installment_commitment	התחייבות לתשלומים, מייצג את האחוז מהכנסות הלווים המשמשות כעת להחזר הלוואות קיימות.	אורדינלי	{1, 2, 3, 4}
personal_status	סטטוס, המצב המשפחתי ומין הלווה	קטגוריאלי	{male single, female div/dep/mar, male div/sep, male mar/wid}
other_parties	האם ישנם גורמים נוספים האחראים להלוואה מלבד הלווה.	קטגוריאלי	{guarantor, none, co applicant}
residence_since	השנים שהלווה גר בבית מגוריו הנוכחי	נומרי	{1, 2, 3, 4}
property_magnitude	סוג הנכס המשמש כבטוחה להלוואה.	קטגוריאלי	{unskilled resident, high qualif/self emp/mgmt, unemp/unskilled non res}
age	גיל הלווה	נומרי	מספר
other_payment_plan	האם ללווה יש הלוואות ממקומות אחרים	קטגוריאלי	{none, bank, stores}
housing	סוג הדיור בו הלווה מתגורר	קטגוריאלי	{for free, own, rent}
existing_credits	מספר הלוואות קיימות	קטגוריאלי	{1, 2, 3, 4}
job	סוג העבודה שיש ללווה	קטגוריאלי	{skilled, unskilled resident, high qualif/self emp/mgmt, unemp/unskilled non res}
num_depen	: מספר האנשים התלויים בלווה כלכלית	קטגוריאלי	{2, 1}
own_telephone	האם ללווה יש טלפון משלו.	בוליאני	{none, yes}
foreign_worker	האם הלווה הוא עובד זר או לא.	בוליאני	{yes, no}
class	המשתנה המציין אם הלווה מסווג כטוב או רע	בוליאני	{good, bad}

איור מס' 1 – התפלגות סיווג ההלוואה



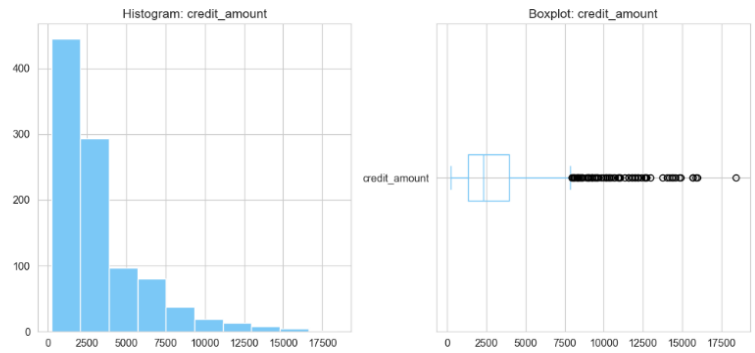
איור מס' 2 התפלגות גיל הלקוחות

count	1000.00
mean	35.55
std	11.38
min	19.00
25%	27.00
50%	33.00
75%	42.00
max	75.00



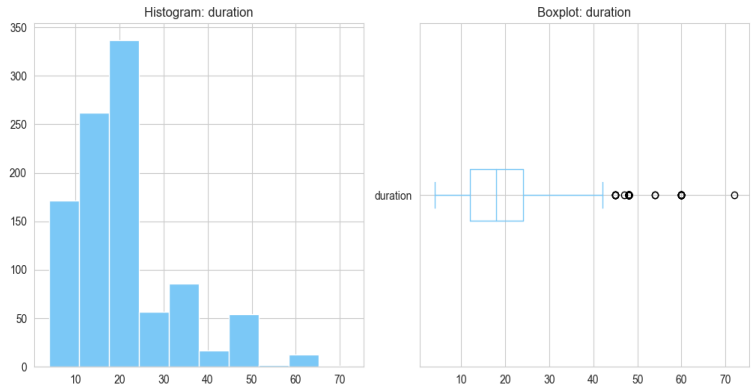
איור מס' 3 התפלגות גודל ההלוואה

count	1000.00
mean	3271.26
std	2822.74
min	250.00
25%	1365.50
50%	2319.50
75%	3972.25
max	18424.00

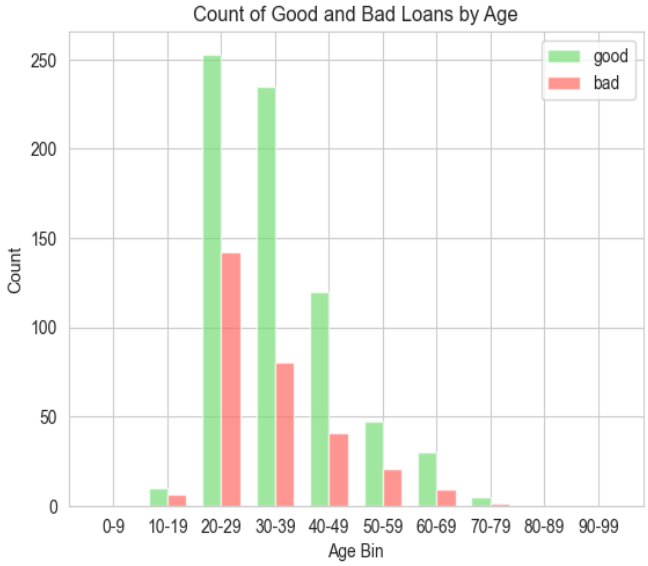


איור מס' 4 התפלגות משך ההלוואה

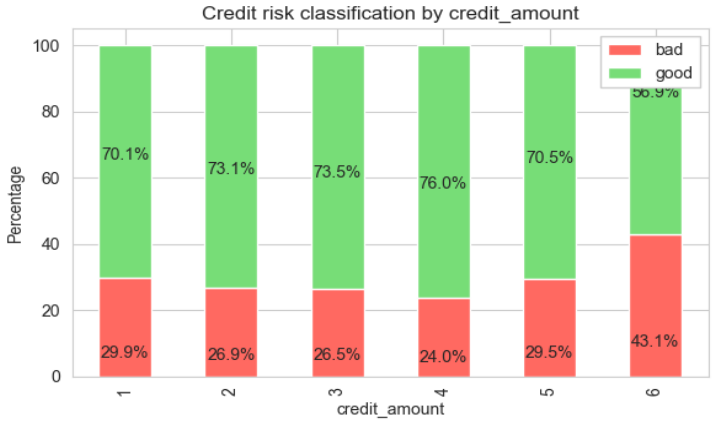
count	1000.00
mean	20.90
std	12.06
min	4.00
25%	12.00
50%	18.00
75%	24.00
max	72.00



איור מס' 5 סיווג הלוואה לפי קבוצת גיל

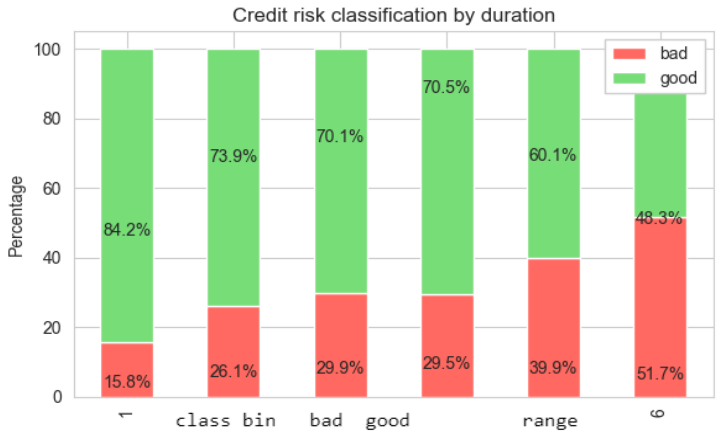


איור מס' 6 אחוז ההלוואות טובות ורעות לפי גודל הלוואה



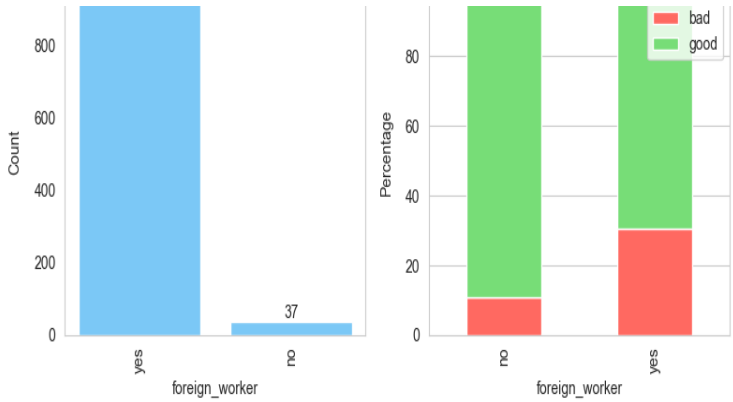
class	bin	bad	good	range
0	1	30.0	70.0	(249.999, 1203.5]
1	2	27.0	73.0	(1203.5, 1554.0]
2	3	27.0	73.0	(1554.0, 2319.5]
3	4	24.0	76.0	(2319.5, 3368.0]
4	5	30.0	70.0	(3368.0, 5509.0]
5	6	43.0	57.0	(5509.0, 18424.0]

איור מס' 7 אחוז הלוואות טובות ורעות לפי משך ההלוואה

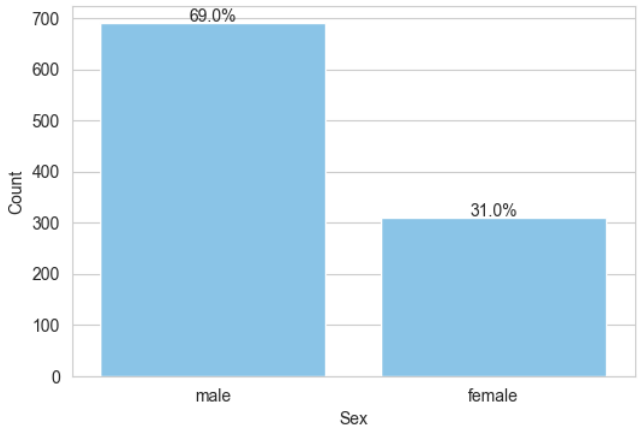


class	bin	bad	good	range
0	1	16.0	84.0	(3.999, 10.0]
1	2	26.0	74.0	(10.0, 12.0]
2	3	30.0	70.0	(12.0, 18.0]
3	4	29.0	71.0	(18.0, 24.0]
4	5	40.0	60.0	(24.0, 36.0]
5	6	52.0	48.0	(36.0, 72.0]

איור מס' 8 התפלגות עובדים זרים



איור מס' 9 התפלגות הלווים לפי מין

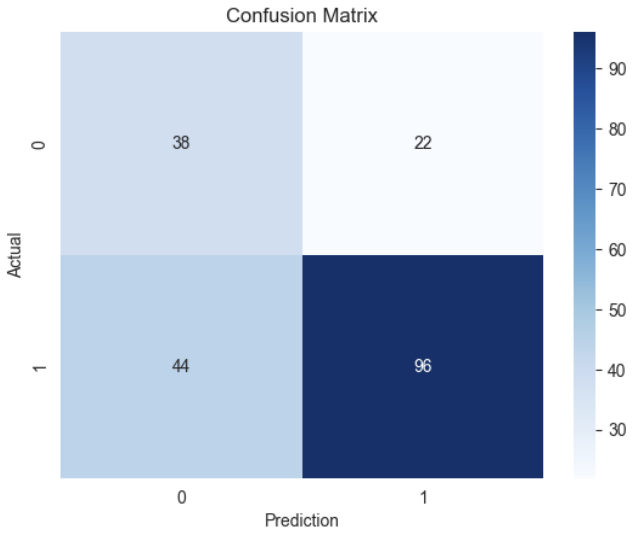


2.6 נספחים – תוצאות

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

           0       0.46      0.63      0.54        60
           1       0.81      0.69      0.74       140

    accuracy                           0.67       200
   macro avg       0.64      0.66      0.64       200
weighted avg       0.71      0.67      0.68       200
```

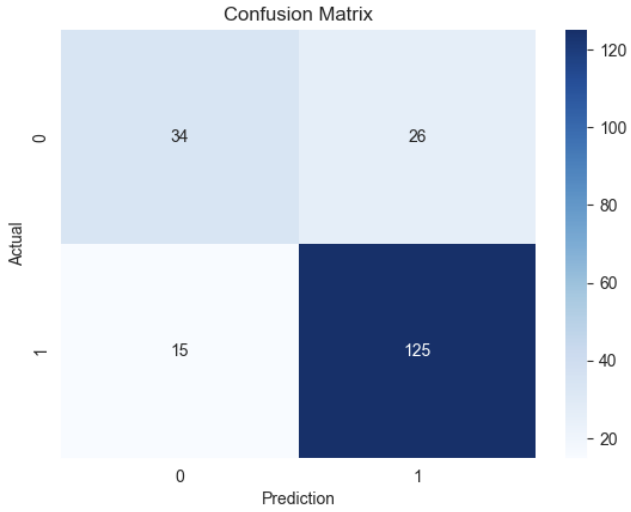


Decision Tree – 11 איור מס'

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

           0       0.69      0.57      0.62        60
           1       0.83      0.89      0.86       140

    accuracy                           0.80       200
   macro avg       0.76      0.73      0.74       200
weighted avg       0.79      0.80      0.79       200
```

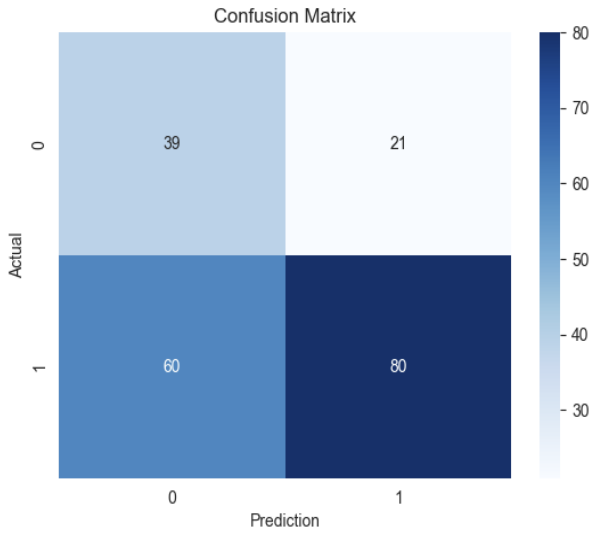


Random Forest – 10 איור מס'

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

           0       0.39      0.65      0.49        60
           1       0.79      0.57      0.66       140

    accuracy                           0.59       200
   macro avg       0.59      0.61      0.58       200
weighted avg       0.67      0.59      0.61       200
```

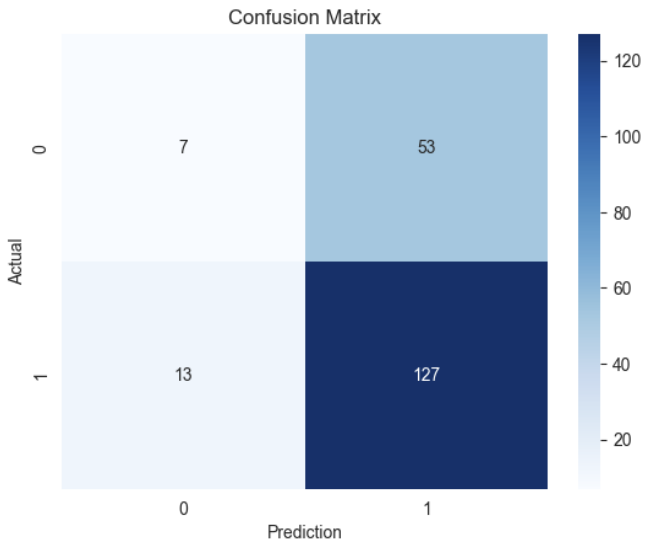


KNN - 13 איור מס'

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

           0       0.35      0.12      0.17        60
           1       0.71      0.91      0.79       140

    accuracy                           0.67       200
   macro avg       0.53      0.51      0.48       200
weighted avg       0.60      0.67      0.61       200
```

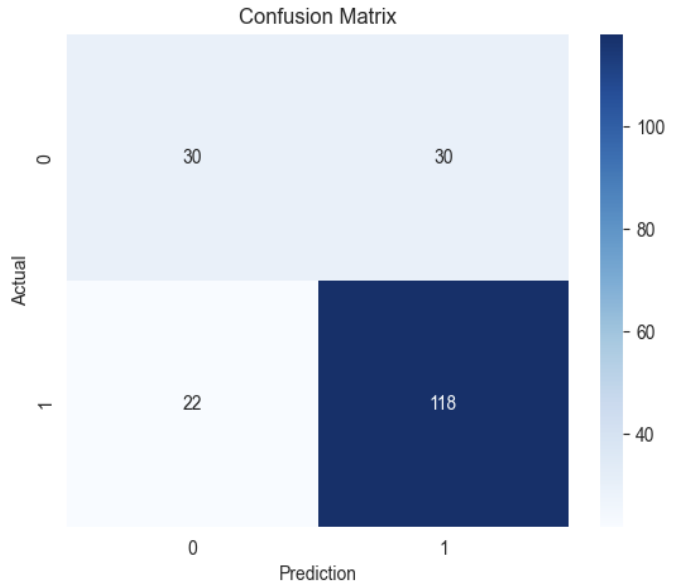


SVM - 12 איור מס'

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

           0       0.58      0.50      0.54        60
           1       0.80      0.84      0.82       140

    accuracy                           0.74       200
   macro avg       0.69      0.67      0.68       200
weighted avg       0.73      0.74      0.73       200
```

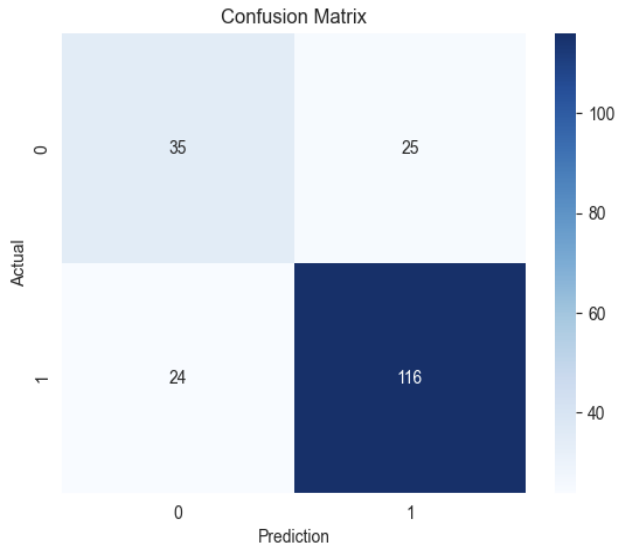


XGboost - 14 איור מס'

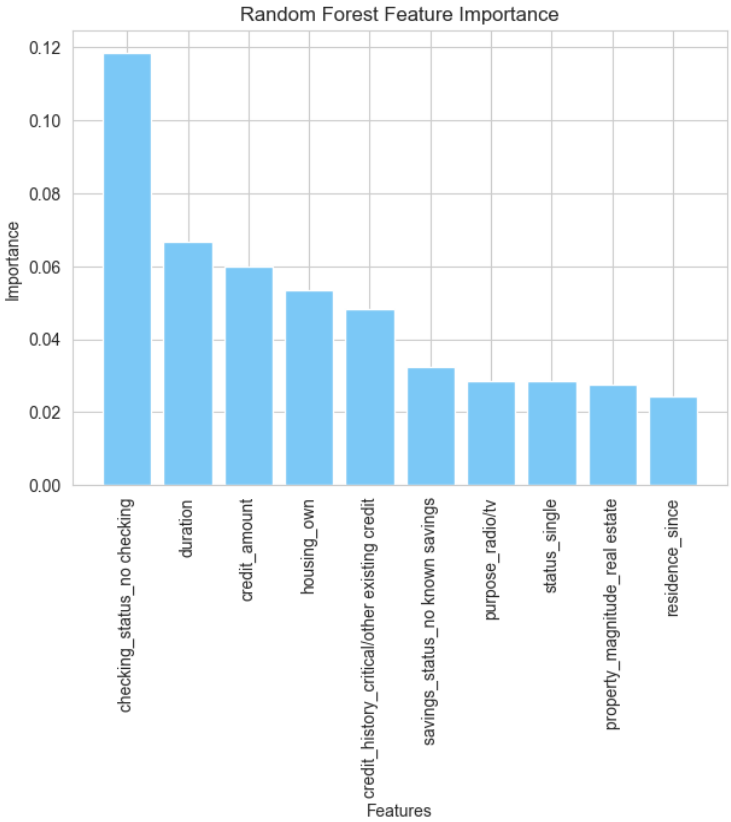
```
Classification Report:
              precision    recall  f1-score   support

           0       0.59      0.58      0.59        60
           1       0.82      0.83      0.83       140

    accuracy                           0.76       200
   macro avg       0.71      0.71      0.71       200
weighted avg       0.75      0.76      0.75       200
```



Logistic Regression - 15 איור מס'

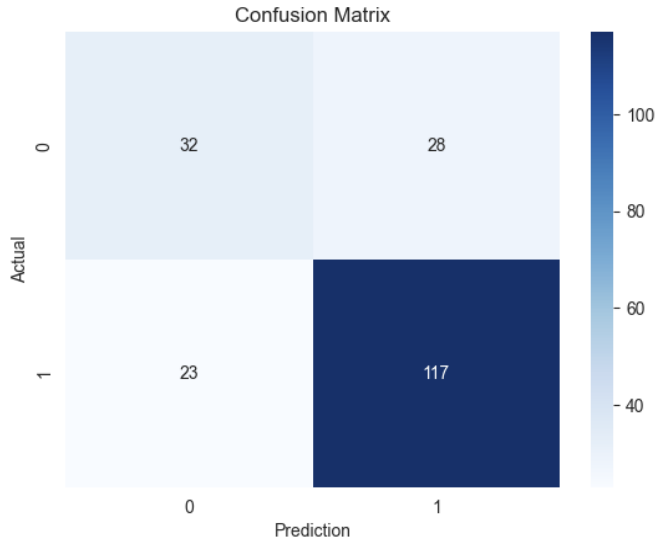


Random Forest איור מס' 17 – משתנים משמעותיים

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

           0       0.58      0.53      0.56        60
           1       0.81      0.84      0.82       140

    accuracy                           0.74       200
   macro avg       0.69      0.68      0.69       200
weighted avg       0.74      0.74      0.74       200
```



Catboost - 16 איור מס'