מעבדה במדעי הנתונים

גלי קולני 208156463

יהל חותם 213508096

נושא הפרויקט:

אנליזה על הלוואות בנקאיות.

לינק ל-GitHub : https://github.com/yaahle/Final-Project-Lab/tree/main

בחירת הנושא ומוטיבציה:

- אוסף הנתונים שבחרנו עוסק בהלוואות בנקאיות.
- כל רשומה היא הלוואה אחת, וכוללת מידע על הלווה (פירוט בהמשך).
- כל רשומה כוללת גם את סטטוס ההלוואה, כלומר האם הלווה הצליח להחזיר את ההלוואה בזמן המוגדר או לא.
- **אוספי הנתונים נלקחו מהאתר Kaggle – בחרנו שני קבצי נתונים שונים :**

1. https://www.kaggle.com/datasets/mirzahasnine/loan-data-set?select=loan_train.csv
2. https://www.kaggle.com/datasets/mirzahasnine/loan-data-set?select=loan_test.csv

- כולל שני קבצי נתונים השונים אשר לשני הקבצים יש את אותם הפיצ'רים מלבד הפיצ'ר על סטטוס ההלוואה אשר קיים רק באחד מקבצי הנתונים.

רקע לנושא:

מהי הלוואה ?

- הלוואה היא עסקה בה אדם או גוף משפטי נותן סכום כסף לאדם או גוף אחר, לתקופה מוגבלת בזמן, לפי תנאים שנקבעים מראש. הגוף הנותן את הכסף נקרא "מלווה", והגוף שמקבל את הכסף נקרא "לווה". סכום הכסף עצמו שניתן בהלוואה נקרא "קרן."
- סיבות אפשריות להלוואה :
אירועים רפואיים חריגים, רכישת דירה / רכב, הקמת וקידום עסקים עצמאיים, לימודים מקצועיים ורכישת השכלה, אירועים ועוד.
- אנשים רבים צריכים לקחת במהלך חייהם הלוואה מסיבות שונות.
- **למה לנו לדעת זאת ?-** חשוב לבנקים לדעת נתונים אלו, על מנת לדעת למי לתת הלוואה ולמי לא.
-

תיאור המידע:

- מבנה הנתונים הינו שני קבצי csv.
- גודל המדגם: 980 רשומות.
קובץ ראשון - בערך 610 רשומות.
קובץ שני - בערך 370 רשומות.
- הסטטיסטיקה התיאורית תוצג עבור איחוד שני הקבצים.

* כמות תכונות/פיצ'רים:
קובץ ראשון- 12 פיצ'רים(כולל תוצאה סופית)
קובץ שני- 11 פיצ'רים(לא כולל תוצאה סופית)

פיצרים:

- מגדר - גבר/אישה
- סטטוס משפחתי - נשוי/לא נשוי
- מספר מעורבים -אפס/ אחד/שניים/שלושה+
- רמת השכלה - בעל תואר/ללא תואר
- סטטוס העסקה - שכיר/עצמאי
- משכורת הלווה השנתית במיליונים
- משכורת הלווה המשני במיליונים(אם יש)
- גודל ההלוואה בעשרות מיליונים
- משך ההלוואה בשנים
- סטטוס ציון credit - טוב / לא טוב
- מקום מגורים - עירוני/חצי עירוני/כפרי
- סטטוס החזרת ההלוואה - רק בקובץ הראשון -החזיר / לא החזיר

שאלות המחקר:

- **שאלת המחקר הראשית**

האם ניתן לחזות לפי התכונות השונות, האם הלווה יחזיר את ההלוואה בזמן?

CORRELATION / COLLINEARITY

- תחילה המרנו את כל הערכים הטקסטואליים לערכים מספרים .
- **המידע אחרי ההמרה :**

	Gender	Married	Dependents	Education	Self_Employed	Applicant_Income	Coapplicant_Income	Loan_Amount	Term	Credit_History	Area	Status
0	0.0	0.0	0.0	1	0.0	584900	0.0	15000000	360.0	1.0	0	1.0
1	0.0	1.0	1.0	1	0.0	458300	150800.0	12800000	360.0	1.0	2	0.0
2	0.0	1.0	0.0	1	1.0	300000	0.0	6600000	360.0	1.0	0	1.0
3	0.0	1.0	0.0	0	0.0	258300	235800.0	12000000	360.0	1.0	0	1.0
4	0.0	0.0	0.0	1	0.0	600000	0.0	14100000	360.0	1.0	0	1.0
5	0.0	1.0	2.0	1	1.0	541700	419600.0	26700000	360.0	1.0	0	1.0
6	0.0	1.0	0.0	0	0.0	233300	151600.0	9500000	360.0	1.0	0	1.0
7	0.0	1.0	3.0	1	0.0	303600	250400.0	15800000	360.0	0.0	1	0.0
8	0.0	1.0	2.0	1	0.0	400600	152600.0	16800000	360.0	1.0	0	1.0
9	0.0	1.0	1.0	1	0.0	1284100	1096800.0	34900000	360.0	1.0	1	0.0
10	0.0	1.0	2.0	1	0.0	320000	70000.0	7000000	360.0	1.0	0	1.0
11	0.0	1.0	2.0	1	NaN	250000	184000.0	10900000	360.0	1.0	0	1.0
12	0.0	1.0	2.0	1	0.0	307300	810600.0	20000000	360.0	1.0	0	1.0
13	0.0	0.0	0.0	1	0.0	185300	284000.0	11400000	360.0	1.0	2	0.0
14	0.0	1.0	2.0	1	0.0	129900	108600.0	1700000	120.0	1.0	0	1.0
15	0.0	0.0	0.0	1	0.0	495000	0.0	12500000	360.0	1.0	0	1.0
16	0.0	0.0	1.0	0	0.0	359600	0.0	10000000	240.0	NaN	0	1.0
17	1.0	0.0	0.0	1	0.0	351000	0.0	7600000	360.0	0.0	0	0.0
18	0.0	1.0	0.0	0	0.0	488700	0.0	13300000	360.0	1.0	2	0.0
19	0.0	1.0	0.0	1	NaN	260000	350000.0	11500000	NaN	1.0	0	1.0

- **מטריצת קורולציה(מעל 0.3):**

	Gender	Married	Dependents	Education	Self_Employed	Applicant_Income	Coapplicant_Income	Loan_Amount	Term	Credit_History
Gender	1.000000	-0.337228	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Married	-0.337228	1.000000	0.351099	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Dependents	NaN	0.351099	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Education	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN
Self_Employed	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN
Applicant_Income	NaN	NaN	NaN	NaN	NaN	1.000000	NaN	0.519334	NaN	NaN
Coapplicant_Income	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN
Loan_Amount	NaN	NaN	NaN	NaN	NaN	0.519334	NaN	1.000000	NaN	NaN
Term	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN
Credit_History	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000
Area	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

העמודות עם קורולציה גבוהה:

- Gender - Married (binary - binary)
- Dependents - Married (ordinal - binary)
- Loan Amount – Applicant Income (continuous - continuous)

- מבין העמודות הבעייתיות, נרצה להוריד features בצורה חכמה, ולכן מכל זוג נוריד את העמודה בעלת הקורלציה הכי נמוכה עם העמודה הסופית(Status) :
Gender, Dependents, Application income.
- העמודה coapplicant income כוללת ערכי הכנסה בדומה לעמודה Applicant_Income (שאותה מחקנו), ויש תלות גבוהה בין הערכים שלהם. היא כוללת את הערכים האלה רק במקרים שבהם בכלל יש co-applicant, ואחרת כוללת 0 . לכן, התלות הזו לא הופיעה במטריצת הקורלציה. כדי לטפל בבעיה זו, נמיר את העמודה הזאת לעמודה בינארית:אם אין co-applicant, יהיה 0 , אם יש, יהיה 1. כך נטפל בבעיית התלות ונשמור על המידע של קיום ה-co-applicant.

Handling Missing Values: Imputation:

- נשתמש ב KNN, כאשר K=5 על מנת למלא את הנתונים החסרים בדטה שלנו.
- בנוסף, על העמודות הקטגוריאליות עשינו את הפונקציה round על מנת שזה יצא ערכים הגיוניים, כלומר אחד או אפס.
- נשים לב שהערכים לא ייצאו מגדר הגבולות האפשריים גם לאחר הפעלת round, בגלל ש KNN לוקח רק ממוצע על ערכים שכבר קיימים ולכן לא ייתן ערך גדול מהמקסימום או קטן מהמינימום.

המידע המקורי

	Married	Education	Self_Employed	Loan_Amount	Coapplicant	Term	Credit_History	Area	Status
0	0.0	1	0.0	15000000	0.0	360.0	1.0	0	1.0
1	1.0	1	0.0	12800000	1.0	360.0	1.0	2	0.0
2	1.0	1	1.0	6600000	0.0	360.0	1.0	0	1.0
3	1.0	0	0.0	12000000	1.0	360.0	1.0	0	1.0
4	0.0	1	0.0	14100000	0.0	360.0	1.0	0	1.0
5	1.0	1	1.0	26700000	1.0	360.0	1.0	0	1.0
6	1.0	0	0.0	9500000	1.0	360.0	1.0	0	1.0
7	1.0	1	0.0	15800000	1.0	360.0	0.0	1	0.0
8	1.0	1	0.0	16800000	1.0	360.0	1.0	0	1.0
9	1.0	1	0.0	34900000	1.0	360.0	1.0	1	0.0
10	1.0	1	0.0	7000000	1.0	360.0	1.0	0	1.0
11	1.0	1	NaN	10900000	1.0	360.0	1.0	0	1.0
12	1.0	1	0.0	20000000	1.0	360.0	1.0	0	1.0
13	0.0	1	0.0	11400000	1.0	360.0	1.0	2	0.0
14	1.0	1	0.0	1700000	1.0	120.0	1.0	0	1.0
15	0.0	1	0.0	12500000	0.0	360.0	1.0	0	1.0
16	0.0	0	0.0	10000000	0.0	240.0	NaN	0	1.0
17	0.0	1	0.0	7600000	0.0	360.0	0.0	0	0.0
18	1.0	0	0.0	13300000	0.0	360.0	1.0	2	0.0
19	1.0	1	NaN	11500000	1.0	NaN	1.0	0	1.0

לאחר מילוי הערכים

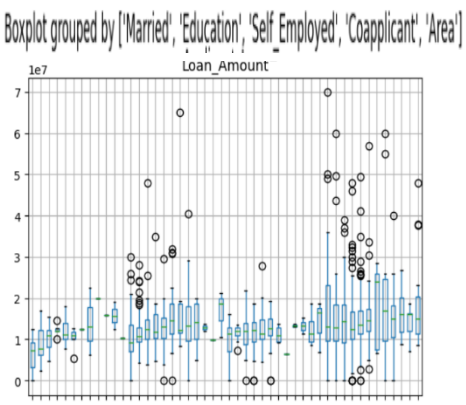
	Married	Education	Self_Employed	Loan_Amount	Coapplicant	Term	Credit_History	Area	Status
0	0.0	1.0	0.0	15000000.0	0.0	360.0	1.0	0.0	1.0
1	1.0	1.0	0.0	12800000.0	1.0	360.0	1.0	2.0	0.0
2	1.0	1.0	1.0	6600000.0	0.0	360.0	1.0	0.0	1.0
3	1.0	0.0	0.0	12000000.0	1.0	360.0	1.0	0.0	1.0
4	0.0	1.0	0.0	14100000.0	0.0	360.0	1.0	0.0	1.0
5	1.0	1.0	1.0	26700000.0	1.0	360.0	1.0	0.0	1.0
6	1.0	0.0	0.0	9500000.0	1.0	360.0	1.0	0.0	1.0
7	1.0	1.0	0.0	15800000.0	1.0	360.0	0.0	1.0	0.0
8	1.0	1.0	0.0	16800000.0	1.0	360.0	1.0	0.0	1.0
9	1.0	1.0	0.0	34900000.0	1.0	360.0	1.0	1.0	0.0
10	1.0	1.0	0.0	7000000.0	1.0	360.0	1.0	0.0	1.0
11	1.0	1.0	0.2	10900000.0	1.0	360.0	1.0	0.0	1.0
12	1.0	1.0	0.0	20000000.0	1.0	360.0	1.0	0.0	1.0
13	0.0	1.0	0.0	11400000.0	1.0	360.0	1.0	2.0	0.0
14	1.0	1.0	0.0	1700000.0	1.0	120.0	1.0	0.0	1.0
15	0.0	1.0	0.0	12500000.0	0.0	360.0	1.0	0.0	1.0
16	0.0	0.0	0.0	10000000.0	0.0	240.0	0.8	0.0	1.0
17	0.0	1.0	0.0	7600000.0	0.0	360.0	0.0	0.0	0.0
18	1.0	0.0	0.0	13300000.0	0.0	360.0	1.0	2.0	0.0
19	1.0	1.0	0.0	11500000.0	1.0	360.0	1.0	0.0	1.0

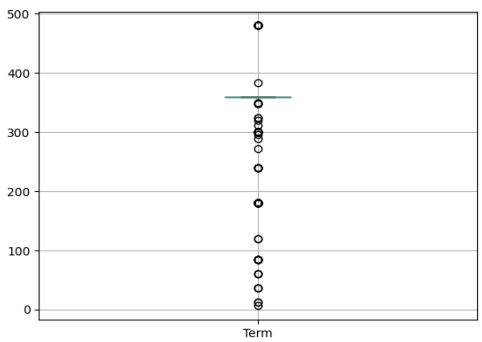
round

	Married	Education	Self_Employed	Loan_Amount	Coapplicant	Term	Credit_History	Area	Status
0	0.0	1.0	0.0	15000000.0	0.0	360.0	1.0	0.0	1.0
1	1.0	1.0	0.0	12800000.0	1.0	360.0	1.0	2.0	0.0
2	1.0	1.0	1.0	6600000.0	0.0	360.0	1.0	0.0	1.0
3	1.0	0.0	0.0	12000000.0	1.0	360.0	1.0	0.0	1.0
4	0.0	1.0	0.0	14100000.0	0.0	360.0	1.0	0.0	1.0
5	1.0	1.0	1.0	26700000.0	1.0	360.0	1.0	0.0	1.0
6	1.0	0.0	0.0	9500000.0	1.0	360.0	1.0	0.0	1.0
7	1.0	1.0	0.0	15800000.0	1.0	360.0	0.0	1.0	0.0
8	1.0	1.0	0.0	16800000.0	1.0	360.0	1.0	0.0	1.0
9	1.0	1.0	0.0	34900000.0	1.0	360.0	1.0	1.0	0.0
10	1.0	1.0	0.0	7000000.0	1.0	360.0	1.0	0.0	1.0
11	1.0	1.0	0.0	10900000.0	1.0	360.0	1.0	0.0	1.0
12	1.0	1.0	0.0	20000000.0	1.0	360.0	1.0	0.0	1.0
13	0.0	1.0	0.0	11400000.0	1.0	360.0	1.0	2.0	0.0
14	1.0	1.0	0.0	1700000.0	1.0	120.0	1.0	0.0	1.0
15	0.0	1.0	0.0	12500000.0	0.0	360.0	1.0	0.0	1.0
16	0.0	0.0	0.0	10000000.0	0.0	240.0	1.0	0.0	1.0
17	0.0	1.0	0.0	7600000.0	0.0	360.0	0.0	0.0	0.0
18	1.0	0.0	0.0	13300000.0	0.0	360.0	1.0	2.0	0.0
19	1.0	1.0	0.0	11500000.0	1.0	360.0	1.0	0.0	1.0

Outlier Detection:

- ישנם **שני features נומריים**: Loan Amount ו-Term .
- נרצה למצוא את נקודות הקצה בהתאם לשכבות האוכלוסייה המתקבלות מה features הקטגוריאליים ולהסיר אותן.
- נשתמש בboxplot על השכבות השונות של האוכלוסייה עבור הפיצ'רים הנומרים כדי למצוא את נקודות הקצה בכל שיכבה באוכלוסיה ולא פשוט למצוא נק' קצה באופן כללי אלא בצורה חכמה.



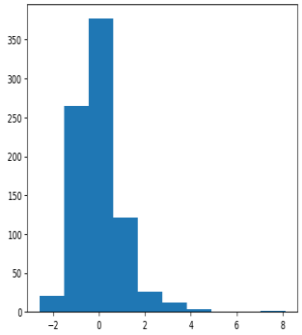


- ניתן לראות מה data set-שתקופת ההלוואה השכיחה ביותר היא 360 חודשים (30 שנה), 835 נקודות מתוך 980, כלומר 145 נקודות הנחשבות לחריגות.
- מההבנה שלנו של המידע, החלטנו לקחת את כל הנקודות בין 10 שנים (120) ל-40 שנים (480), שזוהי התקופה שבדרך כלל לוקחים הלוואה (למדנו זאת ממקורות שונים).לכן, לא נצטרך לבדוק את הboxplot על כל שכבות האוכלוסייה. כל שאר הנקודות , הגדרנו כ outliers-והסרנו אותן.
- סך הכל מחקנו 16 נקודות.
- נמצאו 138 נקודות קצה
- נמחק אותן מהמידע
- נשארו 843 רשומות

Standardization / Normalization:

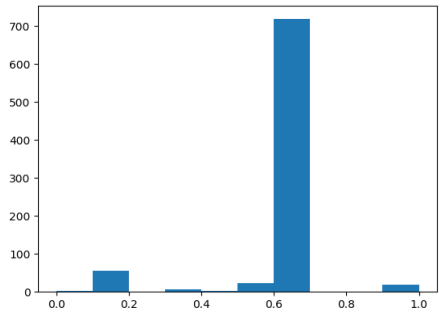
- ישנן שתי עמודות עם ערכים נומריים במידע:העמודה Loan_Amount והעמודה Term. נשים לב שפעולת סטנדרטיזציה עדיפה אם ההתפלגות דומה להתפלגות נורמלית, אחרת פעולת נורמליזציה תהיה עדיפה.

Loan Amount



- הנתונים דומים להתפלגות נורמלית. נחשב ממוצע מדגם:
  - 13376196.319
- נחשב שונות מדגמית:
  - 29517592893974.18
- נפעיל את הטרנספורמציה:
  - נחסר את הממוצע מכל נקודה
  - נחלק בסטיית תקן

Term



- הנתונים אינם דומים להתפלגות נורמלית ,נחשב מקסימום
  - 480.0
- נחשב מינימום
  - 120.0
- נפעיל את הטרנספורמציה:
  - נחסר את המינימום מכל נקודה
  - נחלק במרחק בין המקסימום למינימום

הנתונים אחרי הטרנספורמציה:

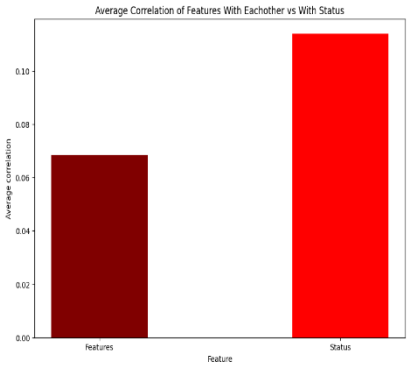
	Married	Education	Self_Employed	Loan_Amount	Coapplicant	Term	Credit_History	Area	Status
0	0.0	1.0	0.0	0.326679	0.0	0.666667	1.0	0.0	1.0
1	1.0	1.0	0.0	-0.101226	1.0	0.666667	1.0	2.0	0.0
3	1.0	0.0	0.0	-0.256828	1.0	0.666667	1.0	0.0	1.0
4	0.0	1.0	0.0	0.151627	0.0	0.666667	1.0	0.0	1.0
6	1.0	0.0	0.0	-0.743083	1.0	0.666667	1.0	0.0	1.0
7	1.0	1.0	0.0	0.482281	1.0	0.666667	0.0	1.0	0.0
8	1.0	1.0	0.0	0.676783	1.0	0.666667	1.0	0.0	1.0
10	1.0	1.0	0.0	-1.229339	1.0	0.666667	1.0	0.0	1.0
11	1.0	1.0	0.0	-0.470780	1.0	0.666667	1.0	0.0	1.0
12	1.0	1.0	0.0	1.299190	1.0	0.666667	1.0	0.0	1.0
13	0.0	1.0	0.0	-0.373529	1.0	0.666667	1.0	2.0	0.0
15	0.0	1.0	0.0	-0.159577	0.0	0.666667	1.0	0.0	1.0
16	0.0	0.0	0.0	-0.645832	0.0	0.333333	1.0	0.0	1.0
17	0.0	1.0	0.0	-1.112638	0.0	0.666667	0.0	0.0	0.0
18	1.0	0.0	0.0	-0.003975	0.0	0.666667	1.0	2.0	0.0
19	1.0	1.0	0.0	-0.354079	1.0	0.666667	1.0	0.0	1.0
20	1.0	0.0	0.0	-0.568031	0.0	0.666667	0.0	0.0	0.0
22	1.0	0.0	0.0	-0.334629	1.0	0.666667	0.0	1.0	0.0
23	1.0	0.0	0.0	-0.412429	1.0	0.666667	0.0	2.0	0.0
24	1.0	1.0	0.0	0.346129	1.0	0.666667	1.0	1.0	0.0

היתכנות:

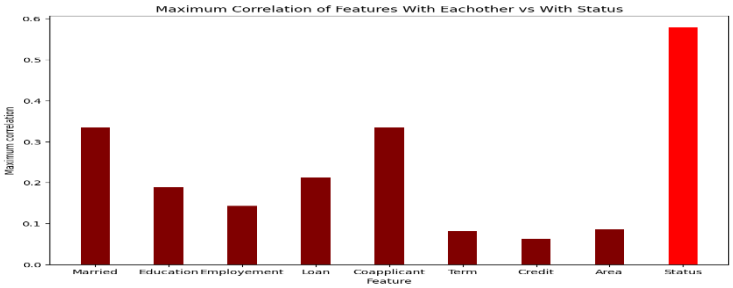
- כדי להראות היתכנות, נשתמש בקורלציה של המשתנים השונים. נשתמש בערך המוחלט של הקורלציה. אם יש היתכנות לפרויקט, נצפה שהקורלציה של המשתנים המסבירים אחד עם השני תהיה נמוכה מאשר הקורלציה שלהם עם המשתנה המוסבר.
- נראה זאת באמצעות מטריצות שונות:

- קורלציה ממוצעת וקורלציה מקסימלית.

- נשווה בין ממוצע הקורלציה של המשתנים המסבירים זה אל זה, לעומת ממוצע הקורלציה שלהם עם המשתנה המוסבר:
- ניתן לראות שהקורלציה הממוצעת גבוהה יותר עבור המשתנה המוסבר.



- נשווה בין הקורלציה המקסימלית עבור כל משתנה מסביר לבין הקורלציה המקסימלית של המשתנה המוסבר:
- ניתן לראות שהקורלציה המקסימלית גבוהה יותר עבור המשתנה המוסבר.



הטיות בנתונים :

- חשוב לשים לב, שרוב האנשים שלוקחים הלוואה בנקאית , מראש חושבים שיאשרו להם את ההלוואה.
- ניתן להסתכל על זה בתור סוג של selection bias . האנשים שנמצאים במחקר/בנתונים שלנו יהיו שונים מאשר כלל האנשים באוכלוסייה. כלומר, האנשים המגיעים למחקר מראש שונים מאדם אקראי באוכלוסייה.
- אבל, בסופו של דבר מטרת המודל היא לעזור לבנק להחליט החלטות בצורה נכונה, והבנק גם ככה רואה רק אנשים שבאים לבקש את ההלוואה, ולכן ההטיה לא מפריעה.
- ההטיה תהיה קיימת גם בתוך הבנק ולכן נכון לאמן את המודל עם הטיה זו אך חשוב לשים לב אליה.

מודלים:

- בפרויקט שילבנו 2 data שונים עם אותם פיצ'רים מלבד ה-label של התוצאה, כלומר החזיר / לא החזיר. הדטה בייס הראשון כלל פיצ'ר זה והשני לא. לכן, בשלב זה, אנו חייבים לפצל את המידע מכיוון שעל מנת לבנות מודל לפרויקט שלנו נוכל להשתמש רק בדטה המתויג.
- נשארנו לנו 503 תצפיות אחרי ההפרדה. נפריד את המידע ל-train , 80% ו- test 20% .

- יש לנו 8 פיצ'רים ומשתנה מוסבר, ה-target .
- מודל התחלתי : בחרנו להשתמש במודל של **רגרסיה לוגיסטית** עם כל הפיצ'רים.
- **מדדים :** sencitivity specitivity ו-accuracy.

1. **מטריצת בלבול :**

	Positive(predict)	Negative(predict)
Positive(real)	**65 TP**	**1 FN**
Negative(real)	**19 FP**	**16 TN**

Sensitivity = 65 / (65+1) = 98.4%

Specificity = 16 /(19+16) = 45.7%

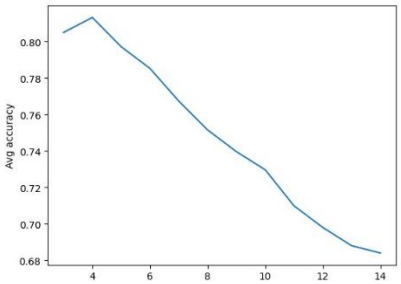
2. המדד השני שהשתמשנו בו הוא accuracy :

קיבלנו accuracy = 80.2% .

- **מסקנות :**

מהמדד הראשון נראה שהמודל יזהה בצורה טובה את האנשים שהחזירו את ההלוואה אבל בצורה פחות טובה את אלה שלא החזירו. לפי המדד השני, קיבלנו רמת דיוק יחסית גבוה בשביל מודל התחלתי.

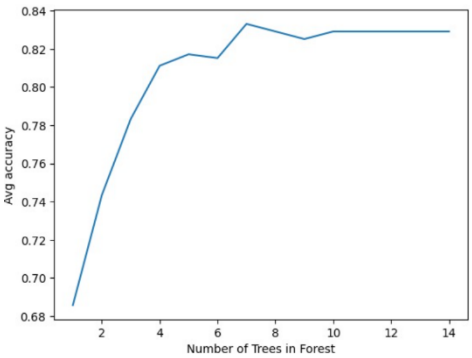
- מודל נוסף שבחרנו להשתמש בו הוא **עץ החלטה**:
-תחילה, ניסנו עץ החלטה עם עומק לא מוגבל על נתוני האימון וקיבלנו accuracy =72.27%.
-הסיבה שהדיוק ירד הוא שהעמוק הלא מוגבל גרם ל-overfitting.
-בדקנו איזה עומק הכי טוב לעץ , השתמשנו ב-k-fold-validation כאשר k=12 . בדקנו את כל העומקים מ-3 עד 15 והעומק הכי טוב שמצאנו הוא 4.



- אחרי שמצאנו שהעומק הכי טוב לעץ שלנו הוא 4, אימנו שוב את המודל על ה-train ובדקנו על ה-test וקיבלנו accuracy=80.2% .

מודל סופי : Random Forest –

- אנו כבר יודעים שהעומק האופטימלי עבור כל עץ הינו 4, אבל אנו לא יודעים כמה עצים ירכיבו את היער. נבדוק באמצעות k-fold-validation עבור k=12 מעץ אחד עד 15 עצים. ניתן לראות שכמות העצים האופטימלית היא 7 עצים.



- נאמן את המודל על כל ה-training ונבדוק אותו על ה-test על מנת לקבל תוצאות סופיות.
- קיבלנו accuracy=81.18% .
- **מטריצת בלבול :**

	Positive(predict)	Negative(predict)
Positive(real)	**65 TP**	**1 FN**
Negative(real)	**18 FP**	**17 TN**

נקבל :

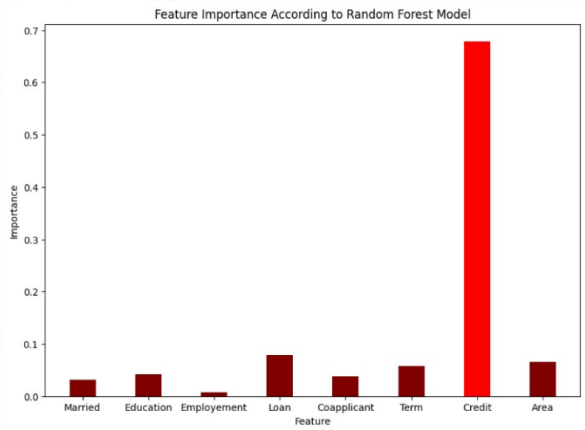
Sensitivity = 65 / (65+1) = 98.4%

Specificity = 17/(18+17) = 48.57%

מסקנות:

- ניתן לראות שמודל זה הוא הטוב ביותר מכל המודלים שבדקנו.
- ניתן לראות שמדד ה- accuracyהוא הגבוה ביותר וכך גם מדדי ה-sensitivity ו-specificity, כלומר מודל זה מזהה בצורה טובה יותר גם אנשים אשר לא החזירו את ההלוואה.
- בחרנו במודלים אלה בפרויקט שלנו מכיוון שאנו רוצים לנבא משתנה בינארי. תחילה, ניסנו מודל רגרסיה לוגיסטית למקרה שההשפעה של הפיצ'רים היא ליניארית ולאחר מכן ניסנו מודל של עץ החלטה למקרה שההשפעה של הפיצ'רים מתנהגת בצורה אחרת וגילינו שהיא לא מתנהגת בצורה אחרת מכיוון שקיבלנו accuracy זהה .
- השתמשנו בשיטת ה-ensemble יער אקראי כדי לשפר את ביצועי העצים.

Feature importance:



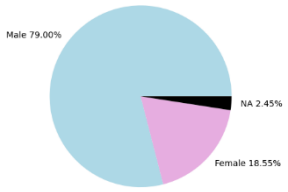
- לפי המודל הסופי שלנו, Random Forest , פלטנו את חשיבות הפיצ'רים:
- ניתן לראות באמצעות גרף זה , שמבחינת information gain הפיצ'ר הכי חשוב הוא ה- credit score ולאחריו ה-loan וה-term .
ניתן לראות שהפיצ'רים שהכי פחות חשובים הם ה-Married ו- Employment .
- חשוב לציין שגם בתוצאות אלה נכלל ה-bias שדיברנו עלי למעלה. כלומר, יכול להיות במציאות הפיצ'רים של Married ו- Employment הם הרבה יותר חשובים אבל האנשים שנמצאים בנתונים הם אנשים אשר ניגשו לקחת את ההלוואה ולכן זה מבטל חשיבות של הפיצ'רים המסוימים האלה/נראים פחות חשובים מאשר שהם יהיו במציאות.

רעיונות לשיפור:

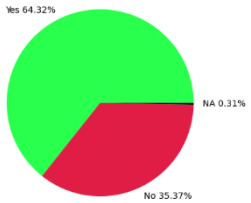
- בפרויקט זה אחת מהבעיות המרכזיות שנתקלנו בהן היא שהייתה לנו כמות יחסית קטנה של מידע. אם נוכל לאסוף מידע נוסף , ביצועי המודל היו יכולים להשתפר.
- יכול להיות שמודל הרבה יותר מסובך כמו רשת נוירונים בהינתן כמות גבוהה מאוד של מידע היה מוציא ביצועים הרבה יותר טובים. אך, מכיוון שלא הייתה לנו כמות גדולה של מידע נמנענו מלנסות מודל זה מכיוון שזה היה גורם ל-overfitting .

נספחים:
תכונות/פיצ'רים :

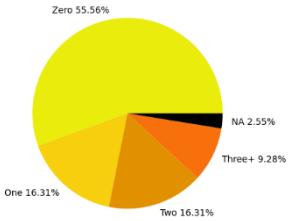
1. מגדר: גבר \ אישה



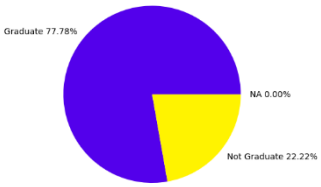
2. סטטוס משפחתי: נשוי \ לא נשוי



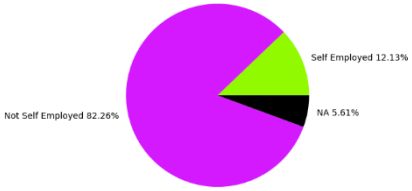
3. מספר מעורבים (ילדים): אחד \ שניים \ שלושה+



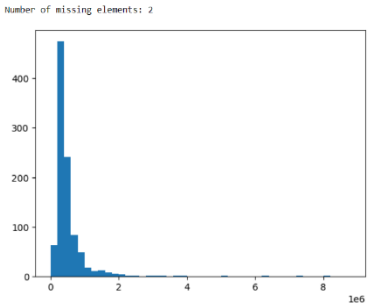
4. רמת השכלה: בעל תואר \ ללא תואר



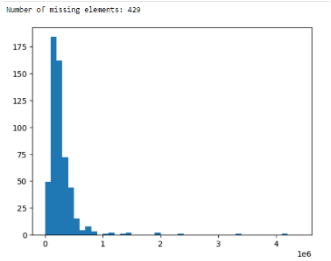
5. סטטוס העסקה: שכיר \ עצמאי



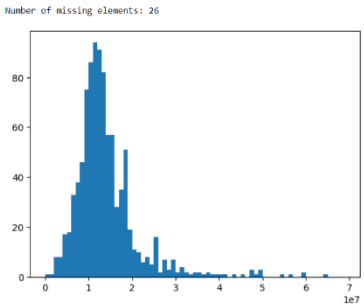
6. משכורת הלווה (שנתית) במיליונים:



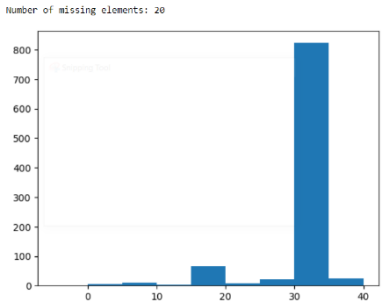
7. משכורת הלווה המשני במיליונים (אם יש, שנתית): לא לכולן יש לווה משני,לכן חסרים הרבה ערכים



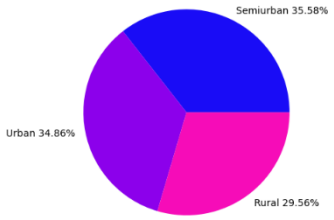
8. גודל ההלוואה בעשרות מיליונים:



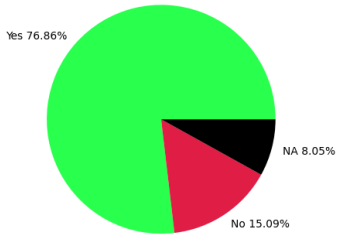
9. משך ההלוואה (בשנים, במקור חודשים):
חצי עירוני / כפרי



11 . מקום מגורים: עירוני/



10. סטטוס ציון credit: טוב \ לא טוב



12. סטטוס החזרת הלוואה:

החזיר/לא החזיר

(רק קובץ ראשון)