

# U-CAR

When Data & Digital  
marketing meets



## Final Project- Ucar

## למידת מכונה בתחום

## מימון הרכבים

## תוכן עניינים

3	.....:תקציר
4	.....:מבוא
6	.....:הכנת הנתונים
8	.....:תכנון מודל סיווג
11	.....:סטטיסטיקות
16	.....:(Evaluation) :תוצאות הרצות האלגוריתמים והערכת המודל
18	.....:מסקנות ודיון
19	.....:מקורות

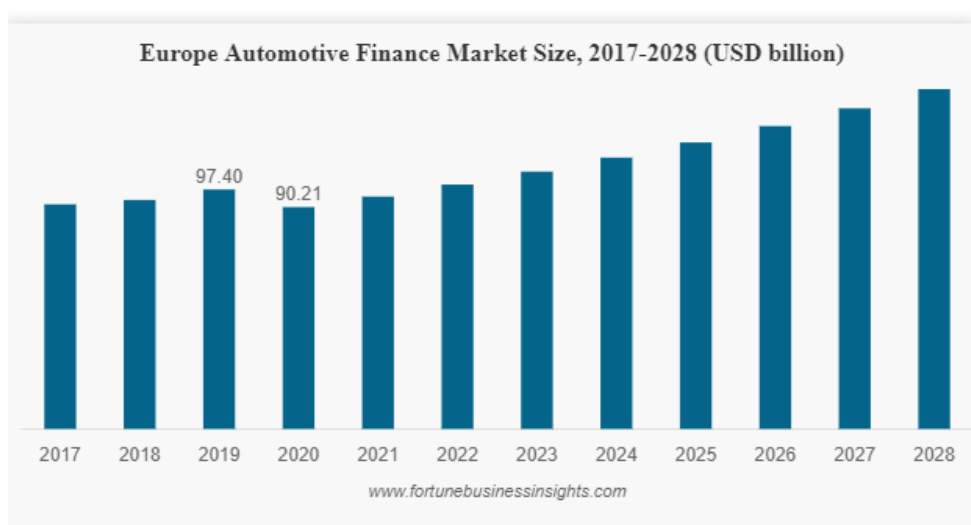
## תקציר:

כיום מתחוללת מהפכה בעולם המודרני ומי שלא רוצה להישאר מאחור בוחר לשלב את תחום למידת המכונה (ML) בעסקים השונים. במאמר זה נדגים כיצד תחום ה-ML עשוי לסייע לעולם מימון הרכבים בחיסכון בזמן הן ללקוח והן לחברה, וכמובן בשיפור רווחיות החברה. במציאות של היום, קניית רכב חדש עלולה להיות תהליך ארוך, יקר ומייגע המצריך משאבים רבים מהאדם הממוצע. לעיתים קרובות הוא אינו מקבל את התוצאה הרצויה. לכן, האלטרנטיבה של קניית רכב באמצעות סוכנויות למימון רכבים נראית יותר אופטימלית לאנשים רבים. כמו כן, גם בתהליך בחירת רכב דרך סוכנות למימון רכבים ההצעה ללקוח לא תמיד עולה בקנה אחד עם רצונו ועם משאביו. אחת המטרות ששמנו לפנינו היא ליצור מודל הסתברותי לחיזוי הסכמה או אי-הסכמה של הלקוח להצעה כלשהי שתינתן לו. המודל בין שלל יתרונותיו הרבים עשוי לסייע במציאת העסקה האידיאלית בעבור אותו לקוח, בנוסף לתרום להגדלת רווחיות החברה עצמה, והחשוב מכל לחסוך זמן יקר לשני הצדדים. מערך הנתונים שלנו הינו מערך נתונים אנונימי שהגיע לידנו מאחת סוכנויות הרכבים בארץ הכולל בתוכו הצעות שניתנו ללקוחות. לאחר בדיקה יסודית של מספר מודלים, בחרנו במודל שנתן את התוצאות הטובות ביותר- Gradient Boosting עם accuracy של 0.71, F1-i העומד על 0.71 .


## מבוא:

בכבישי ישראל נעים 3,840,104 כלי רכב מנועיים, כך מפרסמת הלשכה המרכזית לסטטיסטיקה (נכון לאפריל 2021) אשר מסכמת, כמדי שנה, את מספר כלי הרכב בכבישי ישראל. מדובר בעלייה של 4.1% בהשוואה לנתוני 2020. מתוך כלי רכב אלה 3,312,200 היו כלי רכב פרטיים, והיתר היו משאיות, אופנועים, מוניות ורכבים מסחריים [1]. כל הנתונים שהצגנו לעיל ממחישים בצורה מיטבית את תרבות הצריכה של הקניין הישראלי וכמה הנוחות חשובה לו. בנוסף, לפי נתוני משרד הרישוי, על 726 אלף כלי רכב רשום שעבוד בנקאי. 176 אלף מכוניות חדשות - 52% מסך כלי הרכב שעלו ב-2016 - נרכשו בהלוואות בנקאיות. כלומר, אפשר להסיק כי רבים לא מוכנים להתפשר על הנוחות שלהם ולמרות שאין בבעלותם את הסכום הדרוש לקניית רכב, הם פונים לבנקים במטרה לקבל הלוואות שונות [2].

יתרה מכך, שוק מימון הרכב העולמי צפוי לגדול מ-245.62 מיליארד דולר ב-2021 ל-385.42 מיליארד דולר ב-2028 [3].



כאן Ucar, שמבינה לליבו של האזרח הפשוט ולקושי העומד בפניו לרכוש רכב חדש (בטח ובטח במחירי השוק כיום), נכנסת לתמונה ולכן מציעה ללקוחות שלה שיטה חדשנית שפותחה בשילוב למידת מכונה. על מנת שסוכנות הרכב Ucar תדע להתאים עבורו את ההצעה האידיאלית והמשתלמת ביותר היא מתבססת על דאטה מהימן ופיצ'רים שונים. אחת הבעיות בתחום מימון הרכבים שעשויה להיפתר היא חיסכון אדיר בכספי לקוחות רבים שהולכים לאיבוד – בהתבסס על הדאטה של הפרויקט מדובר על כ-967,104,000 שקלים מהצעות שלקוחות סירבו להן. ייתכן כי באמצעות התחשבות בצורכי הלקוח, מתן מימון מתאים יותר ופריסת תשלומים רחבה יותר, ככל הנראה המספר היה קטן בהרבה ומוביל לכך שהחברה תרוויח אולי עוד כמה מיליונים. כיוון שהמודל שבנינו נותן איזושהי הערכה על סיכוי ההסכמה של הלקוח, אותו סוכן שיושב מול הלקוח יכול להריץ את הנתונים במחשב ולשפר את ההצעה עד שיגיע לאחוז הסכמה שיתאים עבורם.



U-Car

←

Age

25

Gender

Male

Price

45000

Funding percent

75%

Funding amount

33750.00

Monthly payment

595.16

Payment duration

56.71

Condition

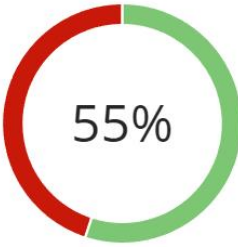
Used

Amount of seats

5

☒ Work Car

☐ SUV



תמונה 1: סימולציה לשימוש במודל : [UCAR APP](#)

### הצלחה מול כישלון:

עוד בתחילת העבודה על הפרויקט בהתחשב בדאטה שקיבלנו ובמידע אודות הלקוח, הבנו ש- 70% במדד ה-Accuracy ייחשב מבחינתנו כהצלחה בפרויקט.

## הכנת הנתונים:

✕ איסוף הדאטה: התקבל הדאטה הגולמי (11949 רשומות) עם הפיצ'רים הבאים: מספר עסקה,

מותג הרכב, דגם הרכב, האם חדש או ישן?, גיל הלקוח, סכום המימון לרכב שיינתן ללקוח,

מחיר הרכב, החזר חודשי, מגדר הלקוח, ופונקציית המטרה- האם התקבלה ההצעה.

○ לכל סוג רכב שהיה קיים בסט הנתונים הוספנו את המאפיינים הבאים: מספר

מושבים, האם זה רכב עבודה/פרטי/4X4.

✕ עיבוד הדאטה: טיפול בדאטה, סינון הדאטה

○ הסרת Null :

▪ מחיקת 150 רשומות ללא דגם ומותג כלל

▪ מחיקת 45 רשומות ללא מחיר רכב ומחיר למימון

○ מחיקת ערכי זבל: 30

▪ 30 רשומות הדגם/המותג לא אמינות

○ מחיקת ערכי-קיצון (ראה [תמונה 1](#)):

▪ מחיקת רשומה שבה סכום המימון גבוה מסכום הרכב

▪ מחיקת 45 רשומות שבהן הגיל היה מעל 90 או מתחת ל18

▪ מחיקת 168 רשומות שבהן הוצע ללקוח מימון למשך כ-20 שנה ומעלה.

(לאחר מחקר והתעמקות בנושא החלטנו על המספר 20)

○ השלמת ערכים חסרים:

▪ 1402 ערכי גיל הלקוח שהושלמו באמצעות KNN (כאשר K=5)

○ המרת הדאטה מעברית לאנגלית בצורה עקבית

○ שינוי סוג פיצ'ר מקטגוריאלי לבינארי מ-Gender --> isMale ,

isNew <-- new\_old

○ יצירת פיצ'רים חדשים באמצעות חישובים על פיצ'רים קיימים כגון:

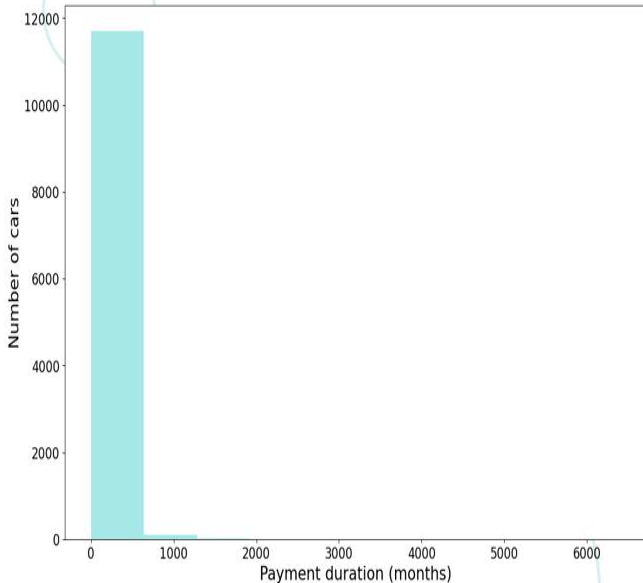
▪ הוספת אחוז המימון -

$$\text{FundingPercent} = (\text{funding\_amount} / \text{car\_price}) * 100$$

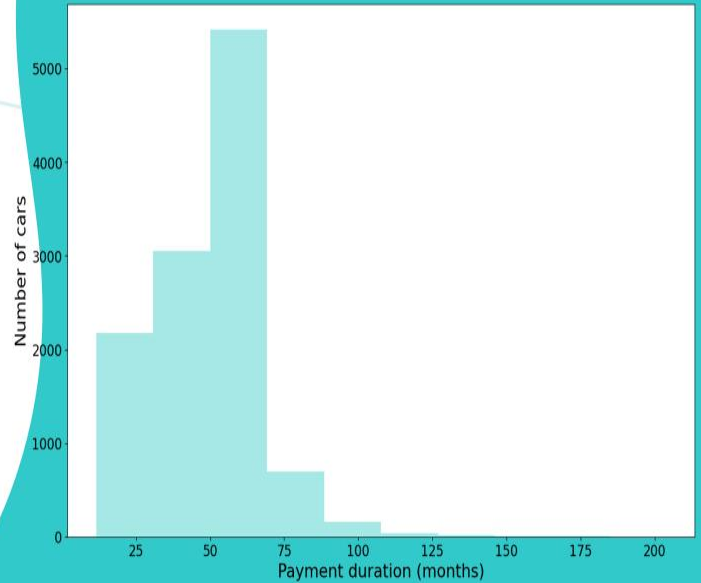
▪ חישוב כמה חודשים להחזר הלוואה-

$$\text{payment\_duration} = \text{funding\_amount} / \text{month\_payment}$$

## Payment Duration Histogram Before Remove Outliers:



## Payment Duration Histogram After Remove Outliers:

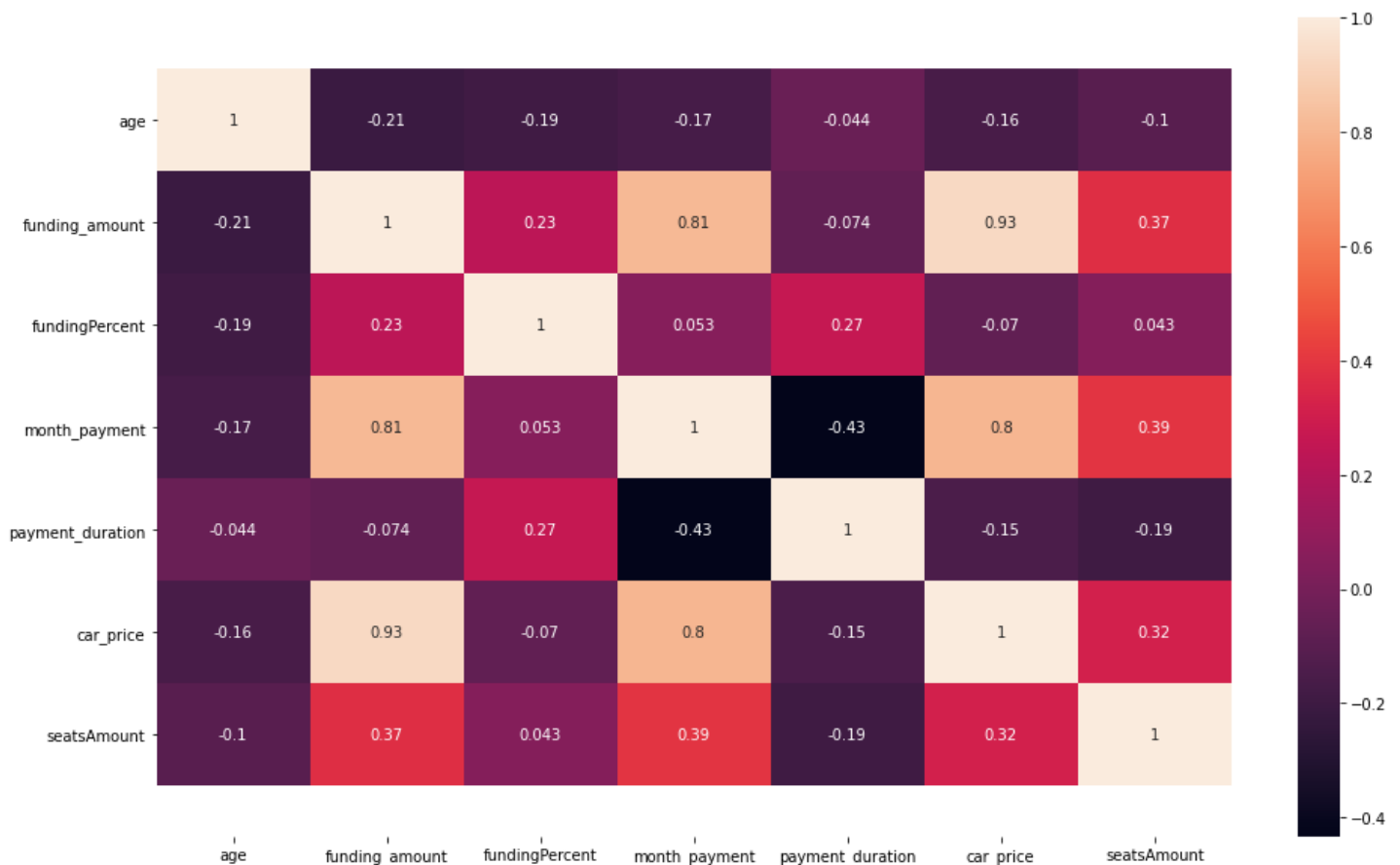


תמונה 2: התפלגות תקופת ההלוואה בחודשים לפני ואחרי מחיקת ערכי קיצון

### ○ הפיצ'רים:

- brand – מותג הרכב (Mercedes, Fiat, Hyundai, etc.)
- model – דגם הרכב
- Age – גיל הלקוח שמקבל את ההצעה
- funding amount – סכום ההצעה שמוגשת
- fundingPercent – אחוז המימון
- month payment – החזר חודשי
- payment duration – תקופת ההלוואה בחודשים
- car price – מחיר הרכב
- is male – מגדר הלקוח האם גבר או אישה?
- isNew – האם מדובר ברכב חדש או ישן?
- isWorkCar – האם מדובר ברכב עבודה?
- isSUV – האם מדובר ברכב שטח?
- seatsAmount – מספר המושבים ברכב
- ind sold – פונקציית המטרה – הסכים או לא?

הערה חשובה: במאגר הנתונים לא קיים מידע מעמיק על הלקוח כמו: גובה הכנסה חודשית, קיים בית בבעלותו, כמה ילדים יש לו, האם הוא נשוי/ רווק, האם פשט בעבר את הרגל, מקום מגורים וכדומה... חשוב לנו לציין זאת כיוון שבמידה וכן היה בידינו מידע שכזה יכולנו לשפר את דיוק המודל ואת אמינותו.



תמונה 3: מפת חום המראה את רמת המתאם בין הפיצ'רים השונים

בחרנו להשאיר את כל הפיצ'רים הקיימים כיוון שאין יותר מידי, וכל אחד מהם תורם להבנת הדאטה בצורה עמוקה יותר.

### תכנון מודל סיווג:

בשלב תכנון המודל ניצב בפנינו האתגר, מהו האלגוריתם הטוב ביותר שיענה בצורה המדויקת על שאלת המחקר שלנו. לשם כך, ניסינו מגוון רחב של אלגוריתמי חיזוי מסוג קלסיפיקציה כגון: רגרסיה לוגיסטית, רשתות נוירונים, עצי החלטה, יערות רנדומיים, SVM, Gradient Boosting, Xgboost. המשותף לכל אלה היא העובדה שכולם מספקים תוצאה הסתברותית שמתאימה לאופי שאלת המחקר- מהי ההסתברות שלקוח יסכים או לא להצעה כלשהי שתינתן לו?

○ Gradient Boosting- האלגוריתם שנתן את החיזוי הטוב ביותר מבין כולם. הגברת הגרדיאנט היא טכניקת למידת מכונה נפוצה בעיקר לסיווג ורגרסיה, אשר משלבת מספר מסווגים חלשים למודל חיזוי חזק. בנוסף, פחות נוטה ל-Over-fitting (במערך נתונים בעל מימדים נמוכים כמו במקרה שלנו).

○ Neural Network- רשת נוירונים שואפת לזהות קשרים בסיסיים במערך נתונים באמצעות תהליך המחקר את האופן שבו פועל המוח האנושי.

○ Logistic Regression- רגרסיה לוגיסטית היא מודל סטטיסטי המתאר קשר אפשרי בין משתנה איכותי/קטגורי, המכונה "המשתנה המוסבר", ובין משתנים אחרים המכונים "משתנים מסבירים". המודל מאפשר לאמוד את מידת ההשפעה של שינוי בערכו כל אחד מהמשתנים המסבירים על ערכו של המשתנה המוסבר. במילים אחרות, המודל מאפשר לאמוד מתאמים בין המשתנים המסבירים למשתנה המוסבר. (ראה [Feature Importance](#))



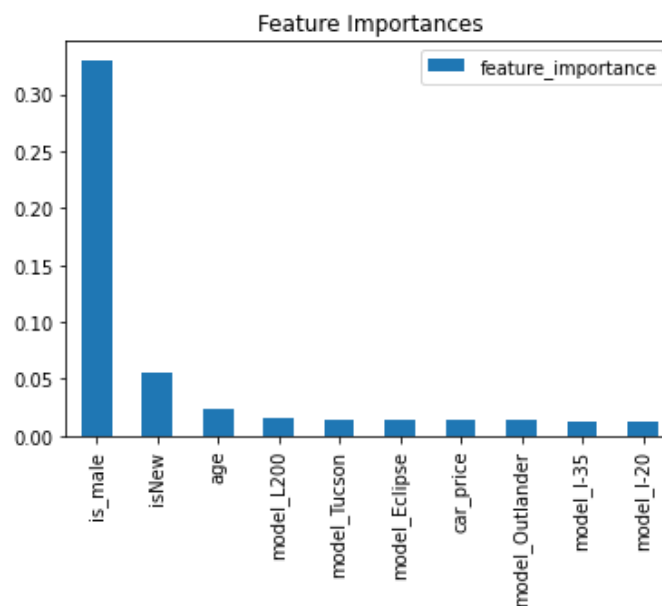
### ○ בחירת הפיצ'רים

יותר מידי פיצ'רים לא תורם ומקשה על הניתוח ← בזבוז זמן, כסף ומשאבים.  
כאשר מדברים על דאטה קיים רעש רקע ולכן לאחר בחירת הפיצ'רים החשובים שתורמים לניתוח נצמצם את הרעש.

### מה הכוונה בבחירת פיצ'ר?

כאשר קיים סט אימון נתון שמיוצג על ידי  $d$  פיצ'רים. נרצה:  
א. להקטין את מספר הפיצ'רים ל- $k$  ( $k \ll d$ )  
ב. לוודא שלאחר צמצום הבעיה, היא ניתנת לפתרון באמצעות למידת מכונה (כלומר שיעור הדיוק של האלגוריתם).

### דירוג הפיצ'רים



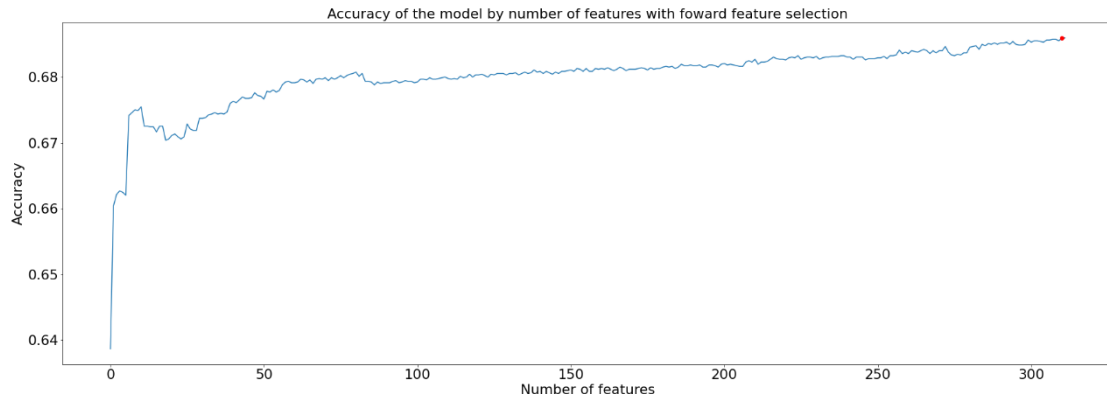
תמונה 4: שיטת ניקוד הפיצ'רים על סמך Xgboost

## שיטות בחירה לאחר הדירוג

$X'$  - סט פיצ'רים ;  $x_i$  - מס' הפיצ'ר

### Forward Algorithm ○

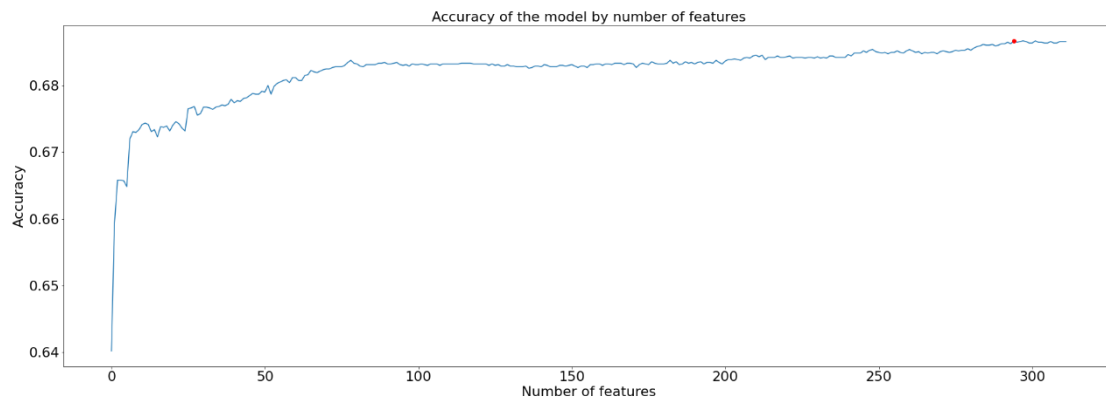
התחלנו עם קבוצה ריקה של פיצ'רים  $X' = \emptyset$ , בכל שלב הוספנו את הפיצ'ר הכי טוב בהתאם לניקוד,  $X' = X \cup x_i$  וביצענו הערכה למודל. עצרנו כאשר מדד ה-Accuracy לא השתנה.



תמונה 5: Forward Algorithm - feature selection

### Backward Algorithm ○

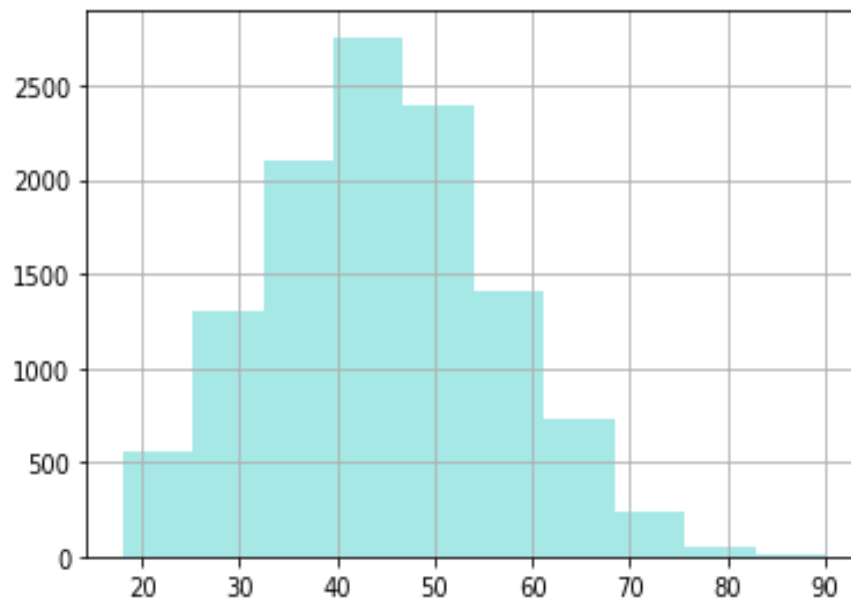
התחלנו עם סט של פיצ'רים  $X' = X$ , בכל שלב הורדנו את הפיצ'ר בעל הניקוד הגרוע ביותר,  $X' = X - x_i$  וביצענו הערכה למודל. עצרנו כאשר מדד ה-Accuracy לא השתנה.



תמונה 6: Backward Algorithm - feature selection

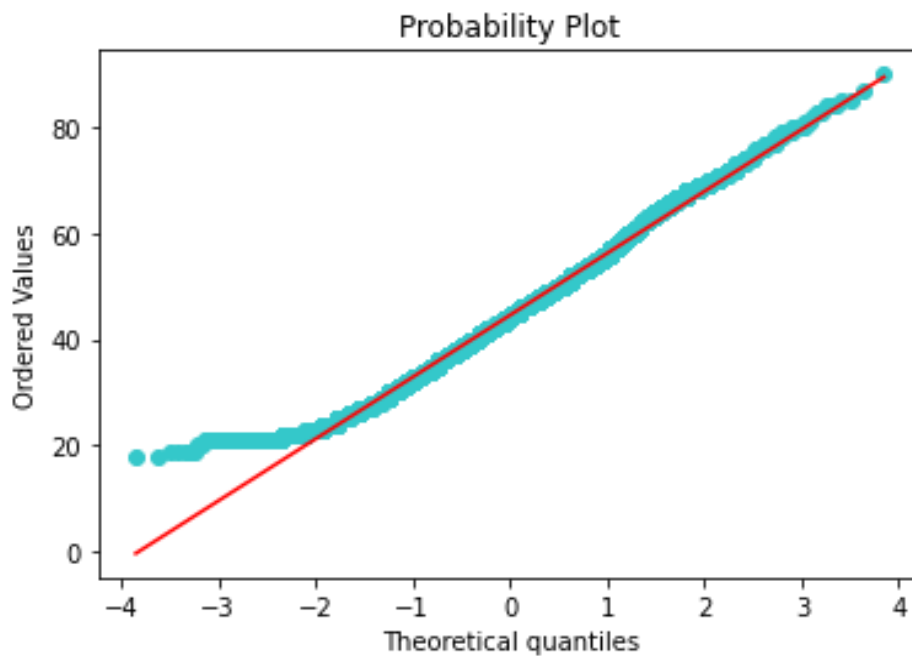
## סטטיסטיקות:

ראשית, לפני ביצוע המבחנים הסטטיסטיים היה עלינו להבין באילו מבחנים מדובר- מבחנים פרמטריים/א-פרמטריים ועל כן השתמשנו במספר דרכים כדי להיות בטוחים.



תמונה 7: *Histogram normality test(age)*

נראה כי ההיסטוגרמה לא בצורת פעמון ולכן השתמשנו ב- qqplot כשיטה נוספת לבדיקת נורמליות.



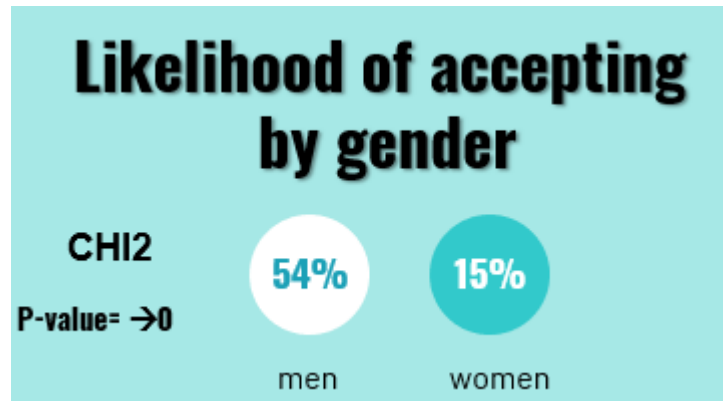
תמונה 8: *qqplot normality test(age)*

בנוסף, כדי להיות בטוחים ב-100% ביצענו את מבחן שפירו (המבחן המדויק ביותר) שנתן את החותמת הסופית כי אכן הדאטה לא מתפלג נורמלית.

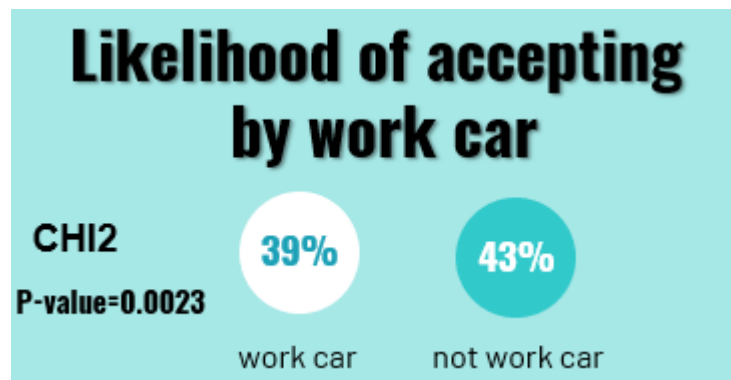
מבחנים סטטיסטיים שביצענו:

מבחני חי בריבוע:

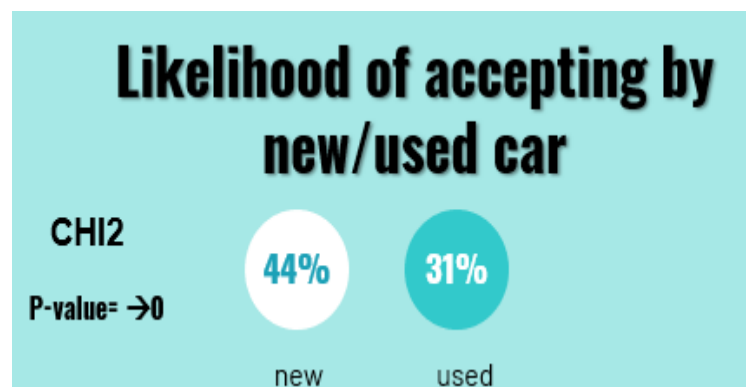
✓ האם יש הבדל בין אחוז ההסכמה של הגברים והנשים לגבי ההצעות שהם מקבלים?



✓ האם יש הבדל בין אחוז ההסכמה לגבי רכב עבודה או רכב אחר?

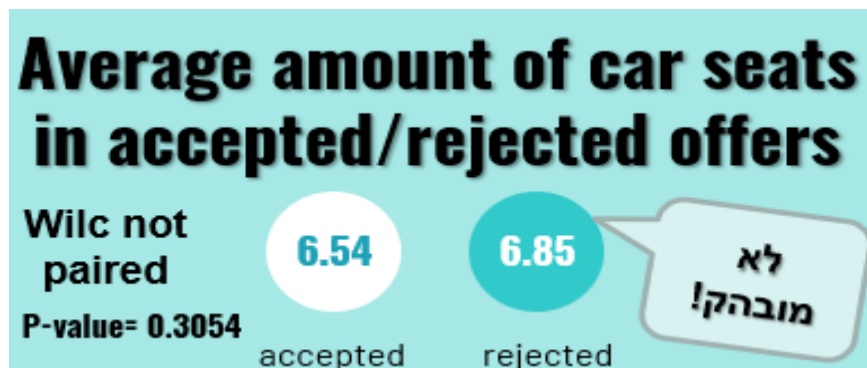


✓ האם יש הבדל באחוז ההסכמה ברכב חדש למשומש?

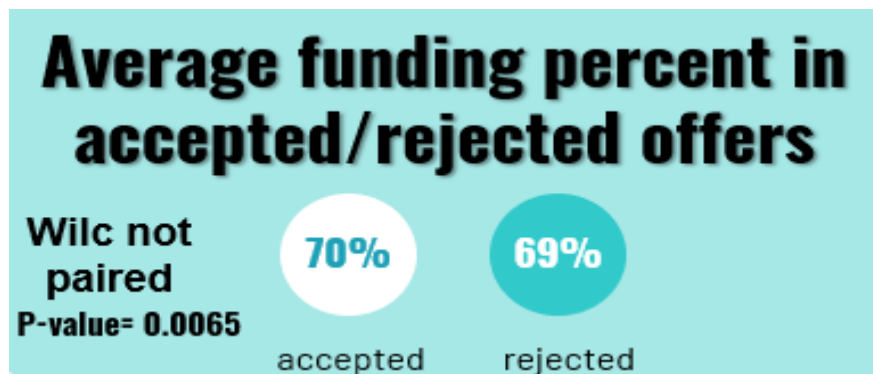


מבחני וילקוקסון לא מזוג:

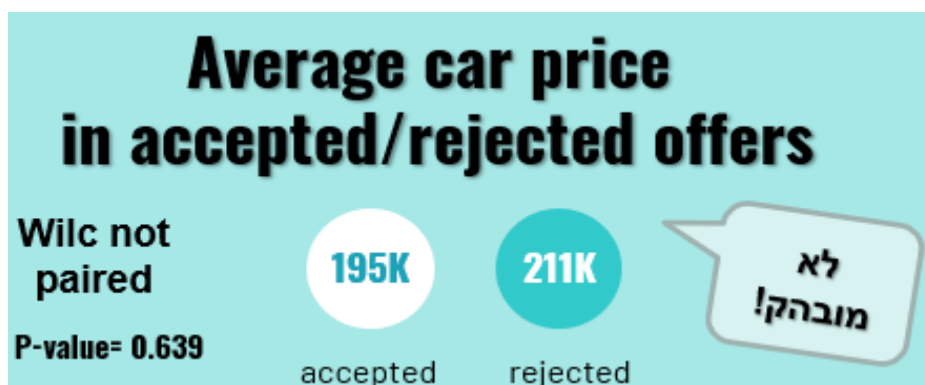
✓ האם יש הבדל בכמות הממוצעת של המושבים ברכב בין הצעות שהתקבלו לבין הצעות שלא התקבלו?



✓ האם יש הבדל בין אחוז המימון הממוצע בהצעות שהתקבלו לבין הצעות שלא התקבלו?



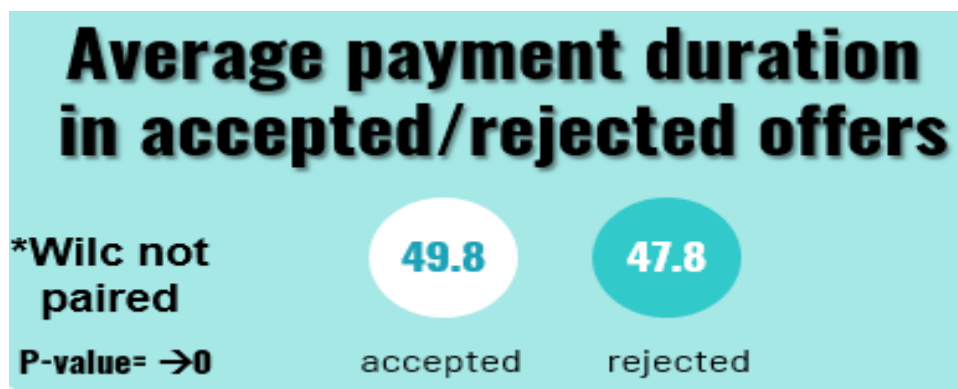
✓ האם יש הבדל בין מחיר הרכב הממוצע בהצעות שהתקבלו לבין הצעות שלא התקבלו?



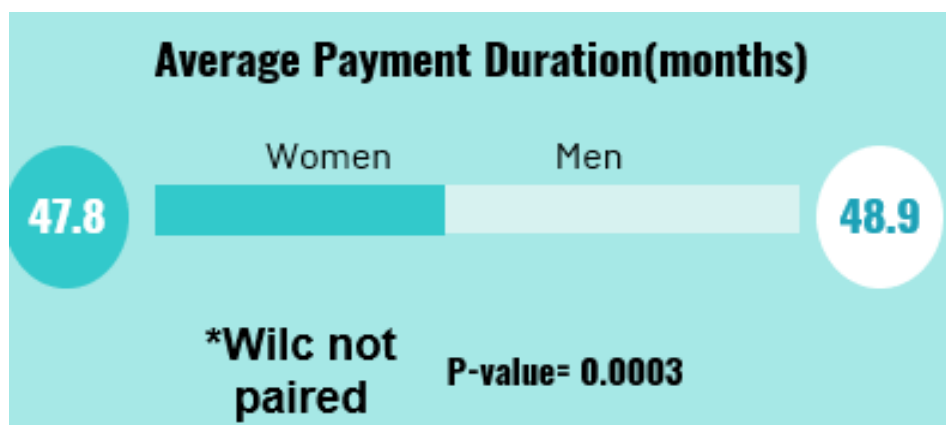
✓ האם יש הבדל במחיר הרכב הממוצע בהצעות שהתקבלו בין נשים וגברים?



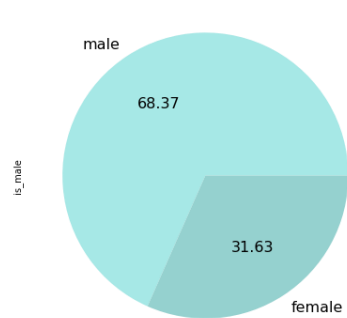
✓ האם יש הבדל בתקופת ההלוואה הממוצעת בהצעות שהתקבלו לבין הצעות שלא התקבלו?



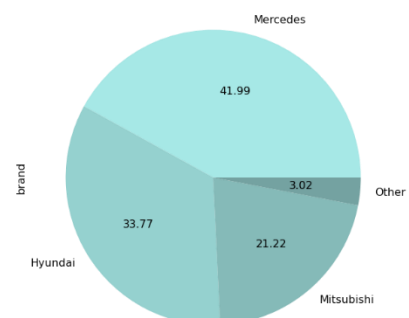
✓ האם יש הבדל בתקופת ההלוואה הממוצעת שניתנת בין נשים וגברים?



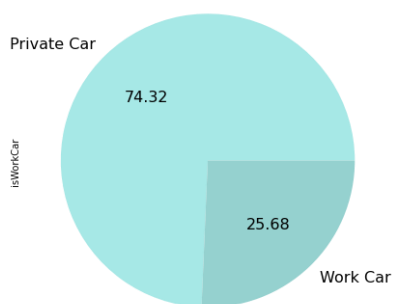
## סטטיסטיקה תיאורית:



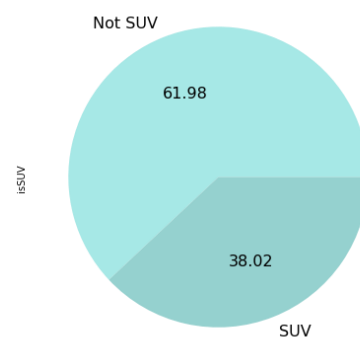
תמונה 10: התפלגות המגדר הלקוח במערך



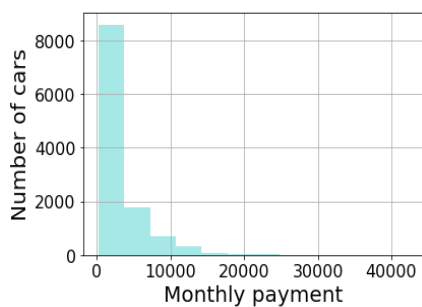
תמונה 9: התפלגות סוגי המותגים במערך הנתונים הנתונים



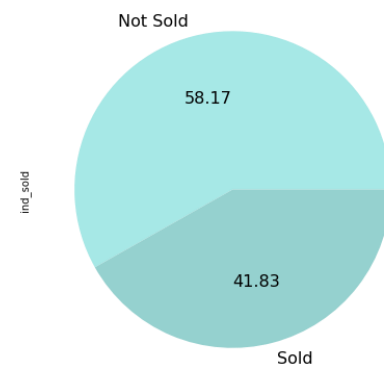
תמונה 12: התפלגות רכבי עבודה ופרטיים



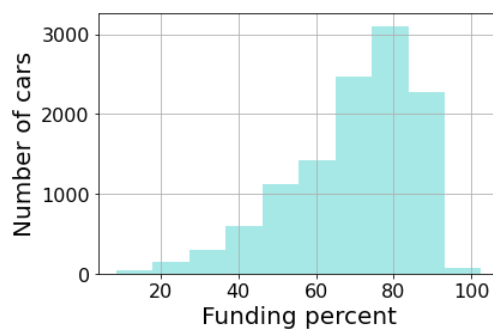
תמונה 11: התפלגות רכבי שטח



תמונה 14: התפלגות מס' הרכבים והמימון החודשי שהם מקבלים



תמונה 13: התפלגות ההצעות שהתקבלו / לא



תמונה 15: התפלגות הרכבים ואחוז המימון שניתן להם

## תוצאות הרצות האלגוריתמים והערכת המודל (Evaluation):

בניסיון להגיע לתוצאות טובות אף יותר מהמודלים הבסיסיים שעובדים עם כל המערך הנתונים, בנינו מגוון של מודלים כדלהלן:

- מודל שהופיע בו דאטה רק על נשים: נעשה כיוון שאחוזי ההסכמה אצל גברים ונשים היו שונים מאוד (ראה מבחן סטטיסטי תואם, ראה תמונה 16) ורצינו לראות איך האלגוריתם יגיב לכל מחלקה בנפרד.  
בתוך מחלקת הנשים היה חוסר איזון בין ההצעות שנשים הסכימו להצעות שסירבו להן, ולכן ביצענו איזון בין המחלקות כך ש-50% היו הצעות שהסכימו ו-50% שלא הסכימו. תוצאות המודל לא היו מספקות בשל כך המשכנו לנסות מודלים אחרים.
- מודל שהופיע בו דאטה רק על גברים: כמו במודל הנשים, גם כאן פיצלנו למחלקה נפרדת כדי לראות איך האלגוריתם יגיב. במקרה של הגברים לא היה צורך באיזון המחלקות, מאחר והן היו מאוזנות כבר. תוצאות המודל היו דומות לזה של הנשים, ולא היו מספקות.
- מודל של רכבי עבודה
- מודל של רכבים חדשים
- מודל של רכבים משומשים
- מודל של הרכבים הנמכרים
- מודל ללא הפיצ'רים דגם ומותג הרכב: מבין כל המודלים המודל הנ"ל סיפק את ה-Accuracy הגבוה ביותר עם חיזוי של 0.71

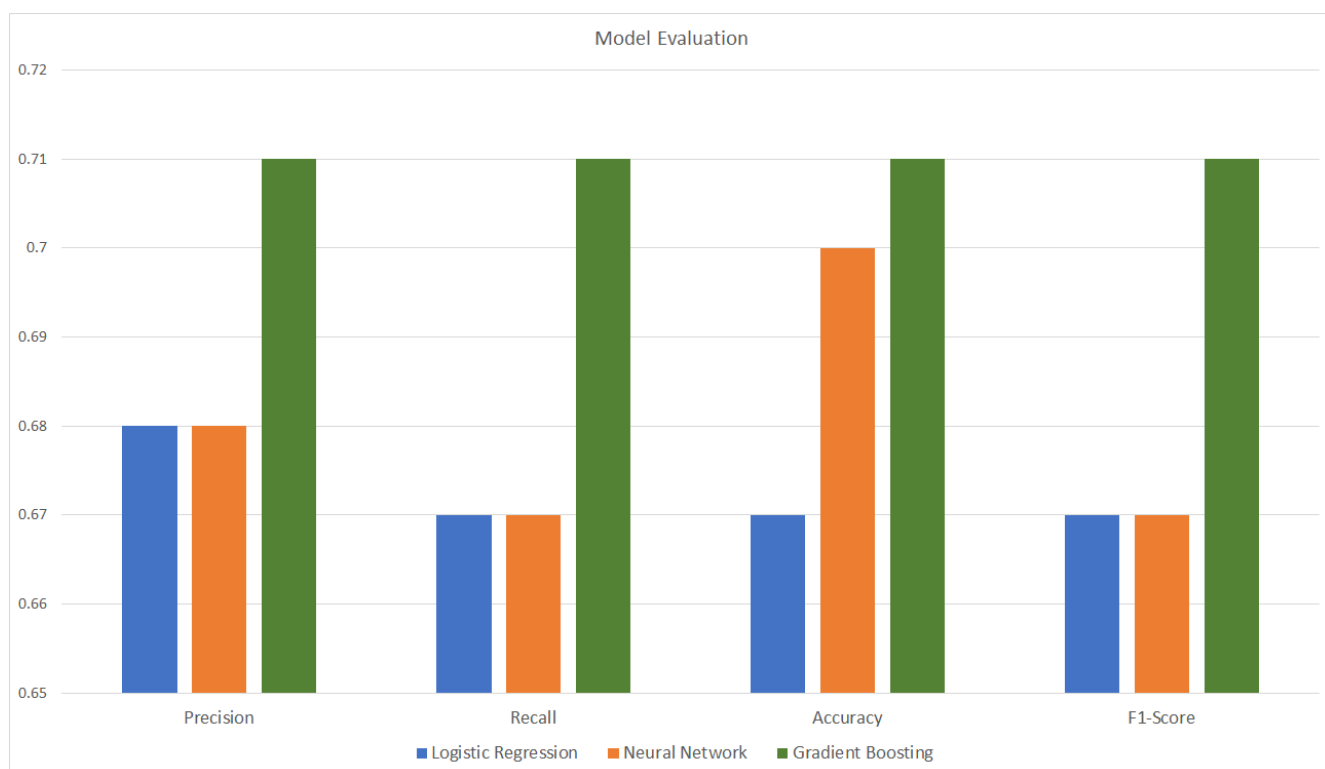
שימוש ב-Tableau ליצירת ויזואליזציה לייצוג דאטה (קורס שנלמד בסמסטר א'):

Age Group	Gender	
	Female	Male
18-29	17.1%	53.8%
30-39	26.4%	64.7%
40-49	2.8%	53.8%
50-59	14.5%	46.2%
60-69	23.2%	42.2%
70+	12.3%	46.9%

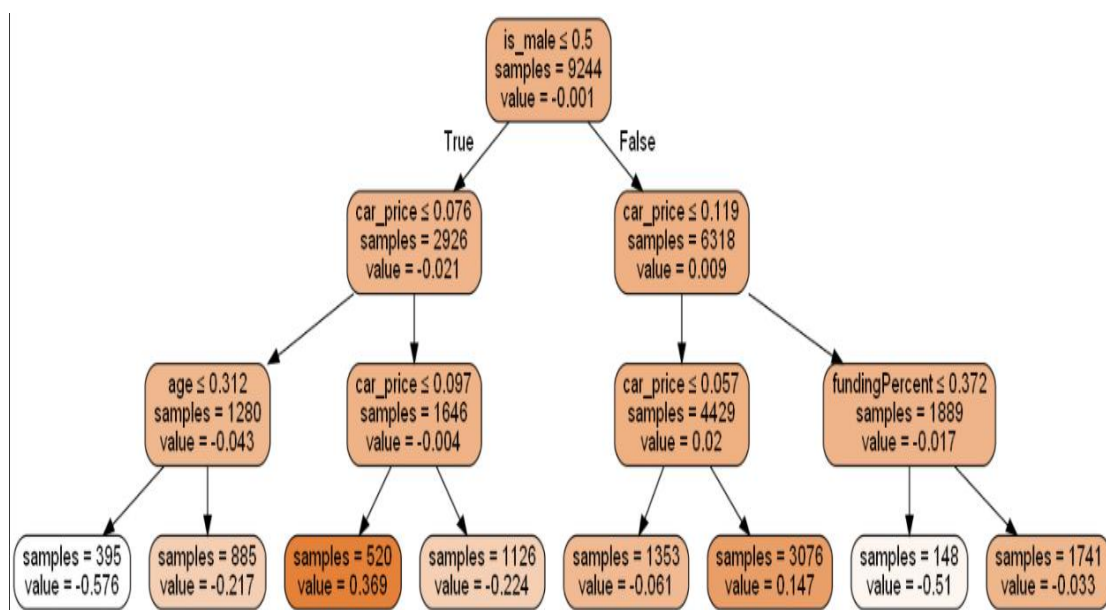
תמונה 16: טבלת התפלגות אחוזי ההסכמה בין נשים וגברים בשכבות הגילאים השונות

ניתן לראות בגרף מטריקות הערכה (תמונה 17) כי האלגוריתם של Gradient Boosting מספק את התוצאות הטובות ביותר עם 0.71 בכל המדדים של Precision, Recall, Accuracy, F1. כפי שהגדרנו בשלב ה-discovery הצלחה מבחינתנו הייתה אחוז דיוק של 70% במדד ה-Accuracy, ולכן ניתן לומר שהמודל עמד בהצלחה ביעד שהצבנו ואף הצליח מעבר למצופה גם בשאר המדדים.

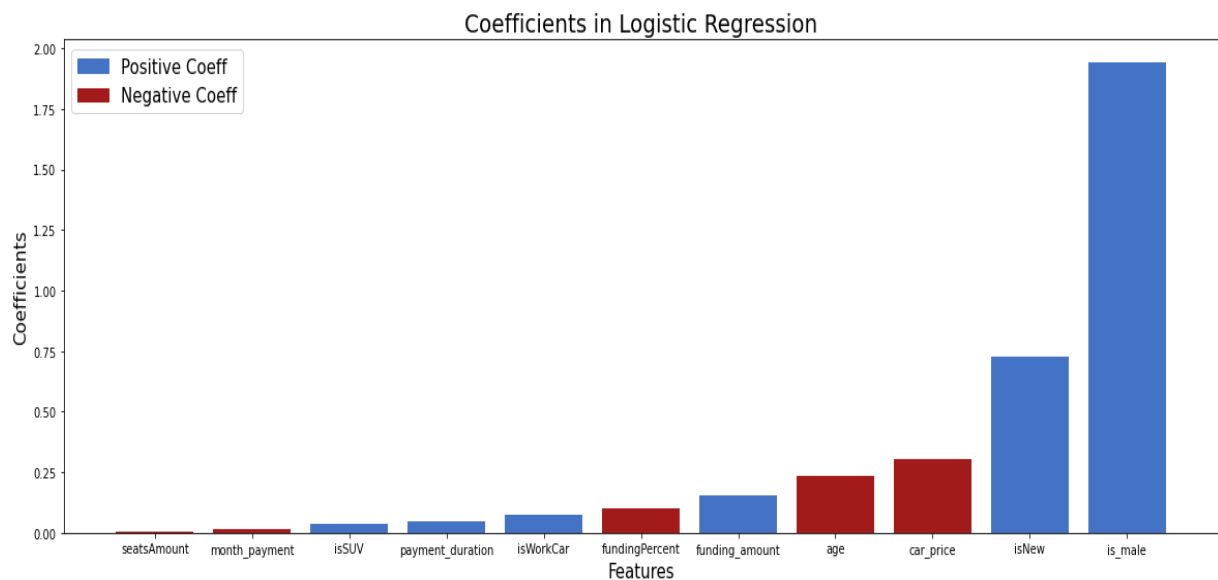




תמונה 17: גרף מטריקות הערכה של האלגוריתמים השונים



תמונה 18: אחד מעצי ההחלטה לפירוש ה- Gradient Boosting



תמונה 19: שיטת ניקוד הפיצ'רים על סמך Logistic Regression ( $\theta$ )

### מסקנות ודיון:

במציאות של היום, כשהערך של השקל מאפשר לקנות פחות מכפי שהיה ניתן בעבר, מציאות בה אנשים חיים ממשכורת של שכר מינימום, ומתקשים לסיים את החודש בכבוד מבלי להיכנס למינוס כל שני וחמישי. מרבית האוכלוסייה לא מוכנה להתפשר על הנוחות שלה וכן תעדיף לרכוש אוטו. השאלה היא מה המחירים שהיא משלמת בעבור הנוחות שלה? ככל הנראה, רוב האנשים יעדיפו לקנות אוטו חדש למרות שאין באמצעותם את המשאבים המתאימים לכך. ייתכן כי מניעים כמו מה אנשים חושבים עליהם יכתיבו בעבורם את המחיר שהם ישלמו עבור אוטו שכזה.

Ucar יכולה להיות בעבור רבים מהם פתרון הולם וחדשני שמתאים את צורכי הלקוח באמצעות בינה מלאכותית ולהוות X-Factor בכל התחום של מימון רכבים. כפי שהצגנו בשלב ה-discovery, המטרות העיקריות שלנו הן לדאוג לנוחות ולרווחיות הן של הלקוח והן של החברה כדי ליצור מצב שכל הצדדים מרוצים.

מובן לנו, כי ניתן לשפר את המודל ולהביאו לגבהים חדשים בהסתמך על דאטה מהימן יותר, ועם פיצ'רים חדשים אודות הלקוח (כמו גובה הכנסה חודשית, קיים בית בבעלותו, כמה ילדים יש לו, האם הוא נשוי/ רווק, האם פשט בעבר את הרגל, מקום מגורים וכדומה) שלא סופקו לנו בתחילת התהליך. בעולם מתוקן, ככל הנראה סוכנות רכבים שמספקת מימון ללקוחות תרצה לדעת למי היא נותנת מימון ואילו במודל שלנו פרט למאפיינים שוליים כמו גיל ומגדר לא ידוע כלל על מצבו הכלכלי של הלקוח. כלומר, כדי שלקוח מסוים יקבל הלוואה, החברה צריכה לדעת שאכן יש באפשרותו להחזיר את ההלוואה במלואה + הריבית עליה.

לסיכום, זכינו לחוות איך מתנהל Data-Analytics-Life-Cycle מתחילתו ועד סופו, ליישם כלים תיאורטיים שנלמדו בקורסים קודמים באופן מעשי כשרק נפתח בפנינו עולם ה-Big Data.

## מקורות:

הפתעה: מספר המכוניות לנפש בישראל נמוך בהשוואה למערב – כלכליסט אתר חדשות מוביל בישראל [1]

חשיפה: כמה מכוניות בישראל נרכשו בהלוואה בנקאית? – דה מרקר הוא עיתון יומי כלכלי ישראלי ואתר חדשות כלכלי [2]

שוק מימון הרכב העולמי צפוי לגדול – חברה המתמחה בחקר שווקים ומספקת שירותים מותאמים אישית עם התמקדות נלהבת על דיוק הנתונים [3]