אוניברסיטת בן-גוריון בנגב
המחלקה להנדסת תעשייה וניהול

נושאים נבחרים בסטטיסטיקה

פרויקט הקורס | סמסטר א' תשפ"ה

חיזוי תביעות ביטוח של לקוחות

מרצה: פרופ' ישראל פרמט

מגיש: יואב אילון 209491018

תוכן עניינים

תקציר

תביעות ביטוח רכב הן נושא מרכזי עבור חברות הביטוח, שכן חיזוי נכון של הסבירות להגשת תביעה מאפשר תמחור מדויק יותר של פוליסות הביטוח, ניהול סיכונים משופר, וזיהוי מוקדם של לקוחות עם סיכון גבוה. אחת הדרכים המרכזיות לבצע חיזוי זה היא באמצעות מודלים סטטיסטיים, כאשר המטרה היא להבין אילו גורמים משפיעים על ההסתברות של לקוח להגיש תביעת ביטוח.

בפרויקט זה נעשה שימוש ברגרסיה לוגיסטית כדי לנתח את הסיכוי להגשת תביעת ביטוח בהתבסס על משתנים אישיים, מאפייני רכב, אזור מגורים, וניסיון נהיגה. רגרסיה לוגיסטית היא שיטה מתאימה במיוחד לחיזוי משתנה בינארי. השיטה משתמשת בפונקציית הלוג כדי למפות את הקשרים בין המשתנים המסבירים להסתברות להגשת תביעה. בניגוד לרגרסיה לינארית, מודל זה אינו דורש הנחות של נורמליות, ליניאריות והומוסקדסטיות בשאריות, ומתאים למצבים שבהם המשתנה המוסבר הוא קטגוריאלי.

תחילה, הסתכלתי על התפלגות המשתנים, הקשרים בין המשתנים והקשרים למשתנה המטרה. מתוך התנהגות המשתנים והקשרים האלו ניתחנו מגמות משמעותיות שישפיעו לנו על המודל ולכן יש להתייחס אליהן בדרך מסוימת - בחירת משתנים מסוימים או איחוד משתנים למאפיין חדש. לאחר מכן, בשלב בניית המודל, בחנתי אילו מאפיינים יכנסו למודל בעזרת שיטות ומבחנים סטטיסטיים, מסקנות שעלו מהכנת הנתונים ולפי היגיון שעלה ממשמעות הנתונים.

בסופו של דבר בניתי מודל רגרסיה לוגיסטית שכולל כמה משתנים אישיים על הלקוח, רכבו והתנהגותו שנמצאו משפיעים על תביעות ביטוח. חלק מיתרונות המודל הוא הפשטות שבה ניתן להשתמש בו כדי לחזות את יחס הסיכוי - פי כמה הסיכוי לתביעה משתנה (גדל/ קטן) עם שינוי של יחידה במשתנה המסביר.

מבוא

פרויקט זה עוסק בניתוח נתוני ביטוח רכב על מנת לזהות גורמים המשפיעים על הסבירות של מבוטח להגיש תביעת ביטוח בשנה הקרובה. סט הנתונים מכיל מידע על לקוחות (כולל נתונים דמוגרפיים כגון גיל, מגדר, השכלה והכנסה) לצד נתונים הקשורים להיסטוריית הנהיגה שלהם (מספר עבירות התנועה ותאונות עבר) וכן מאפיינים שונים של הרכב (סוג הרכב ובעלות עליו).

המטרה המרכזית של הפרויקט היא לבנות מודל חיזוי שיאפשר לחזות האם לקוח צפוי להגיש תביעת ביטוח, בהתבסס על הנתונים שבידינו. לחיזוי זה יש חשיבות רבה עבור חברות ביטוח, שכן הוא יכול לסייע בהערכת סיכונים, תמחור פוליסות בצורה מדויקת יותר, שיפור ניהול התביעות, ואף במניעת הונאות ביטוח. על ידי שימוש במודל חיזוי, חברות הביטוח יכולות לזהות נהגים בסיכון גבוה, להציע להם פוליסות מותאמות אישית, ולהבטיח שמדיניות התמחור שלהן הוגנת ורווחית. בנוסף, המודל מספק הבנה מעמיקה יותר של הדפוסים וההתנהגויות המובילים לתביעות ביטוח, דבר שיכול לשמש גם לשיפור תקנות הנהיגה ולפיתוח תוכניות למניעת תאונות.

מאחר והנתונים כוללים מגוון רחב של משתנים, ניתוח מעמיק יאפשר לנו לבחון אילו גורמים הם המשפיעים ביותר על הסיכוי לתביעת ביטוח. באמצעות שיטות סטטיסטיות, נוכל לאתר קשרים בין המשתנים ולבנות מודל אפקטיבי ומדויק ככל האפשר.

מטרת הפרויקט

מטרתי בפרויקט זה הינה חיזוי תביעות הביטוח של לקוחות, כתלות במאפיינים על הלקוחות, היסטוריית נהיגה ומאפייני רכב. מטרת הפרויקט היא לזהות את הגורמים המשפיעים על הסבירות של לקוח להגיש תביעת ביטוח רכב ולפתח מודל חיזוי אפקטיבי שיסייע לחברות הביטוח בהערכת סיכונים.

<u>שאלות המחקר</u>

1. אילו מאפיינים משפיעים באופן המשמעותי ביותר על חיזוי תביעת ביטוח רכב?
2. איזו שיטת מידול מספקת את החיזוי המדויק ביותר לתביעת ביטוח רכב?

<u>השערות המחקר</u>

1. מספר עבירות ותאונות גבוה יותר מגדיל את ההסתברות לתביעת ביטוח - נהגים עם היסטוריה של עבירות תנועה ותאונות נוטים להיות בעלי דפוסי נהיגה מסוכנים יותר, ולכן צפויים להגיש יותר תביעות ביטוח.
2. לקוחות בעלי נתונים פיננסיים נמוכים יהיו בעלי סבירות גבוהה יותר להגיש תביעות ביטוח רכב, בשל תלות גבוהה יותר בכיסוי - לקוחות עם הכנסה נמוכה או ניקוד אשראי נמוך עשויים להתקשות בכיסוי הוצאות בלתי צפויות, ולכן יסתמכו יותר על הביטוח במקרה של נזק לרכב.
3. חיזוי בעזרת רגרסיה לוגיסטית ייתן חיזוי מדויק יותר - רגרסיה לוגיסטית מתאימה לבעיות שבהן יש צורך לחזות משתנה בינארי, שכן היא מאפשר להעריך את הסיכוי לתביעות ביטוח תוך שקילת ההשפעה של משתנים שונים.

שיטה

נתונים

סט הנתונים בפרויקט מכיל 6,506 רשומות כאשר כל רשומה מהווה לקוח בחברת הביטוח, ולכל לקוח נאספו 17 מאפיינים. המשתנה המוסבר הינו בינארי - האם הלקוח תבע את הביטוח בשנה החולפת.

פירוט המאפיינים:

שם המשתנה	סוג המשתנה	פירוט
ID	ספירה	מספר מזהה ייחודי ללקוח
GENDER	קטגוריאלי	מגדר הלקוח (זכר/ הקבה)
EDUCATION	קטגוריאלי	רמת השכלה של הלקוח (3 רמות)
INCOME	קטגוריאלי	סיווג רמת הכנסה (4 רמות)
CREDIT_SCORE	רציף	ניקוד האשראי המייצג את היציבות הפיננסית של הלקוח. נע בין 0 ל-1.
VEHICLE_OWNERSHIP	בינארי	האם הלקוח הוא הבעלים של הרכב
VEHICLE_YEAR	בינארי	האם הרכב יוצר לפני או אחרי שנת 2015
MARRIED	בינארי	האם הלקוח נשוי
CHILDREN	בינארי	האם הלקוח בעל ילדים
POSTAL_CODE	קטגוריאלי	אזור מגורים של הלקוח (4 אזורי מגורים)
ANNUAL_MILEAGE	בדיד	כמות הקילומטרים שהלקוח נוהג בשנה
VEHICLE_TYPE	קטגוריאלי	סוג הרכב של הלקוח (2 קטגוריות)
SPEEDING_VIOLATIONS	בדיד	מספר עבירות מהירות של הלקוח
PAST_ACCIDENTS	בדיד	מספר התאונות בהן הלקוח היה מעורב בעבר
AGE	בדיד	גיל הלקוח
DRIVING_EXPERIENCE	בדיד	מספר השנים בהם הלקוח בעל רישיון נהיגה
OUTCOME	בינארי	משתנה המטרה - האם הלקוח תבע את דמי הביטוח השנה החולפת

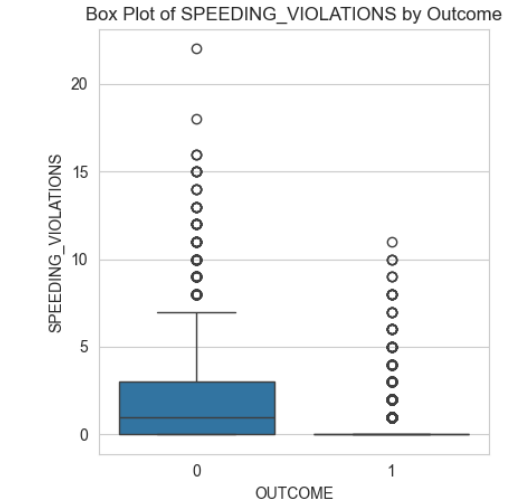
<u>טיוב הנתונים</u>

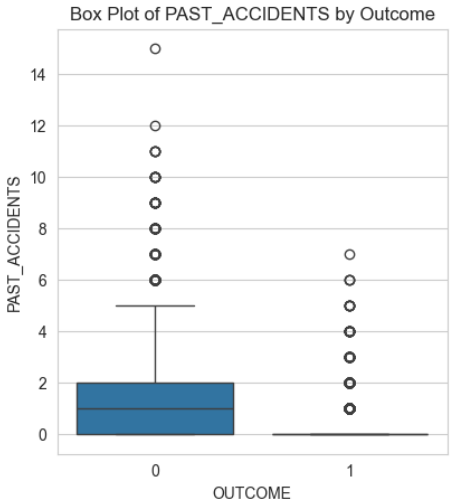
1. עדכון שנות הניסיון בנהיגה - הוספתי מאפיין נוסף לסט הנתונים - גיל הוצאת הרישיון (AGE_START_DRIVING). שמתי לב בניתוח המאפיינים שיש לקוחות שהוציאו רישיון לפני גיל 16. מאחר ונתון זה לא הגיוני, ביצעתי עדכון לשנות הניסיון של הלקוחות אשר גיל הוצאת הרישיון שלהם היה לפני גיל 16 בצורה הבאה - Age -16 , כדי שהגיל המינימלי להוצאת הרישיון יהיה לכל הפחות 16. בנספח 1 אפשר לראות את ההתפלגות לפני ואחרי השינוי
2. מיפוי משתנים לצורך ניתוח - כדי להפוך את הנתונים לקלים לניתוח סטטיסטי, בוצע מיפוי משתנים קטגוריאליים לערכים מספריים. מיפוי זה שומר על סדר היררכי (רמות השכלה והכנסה) והופך משתנים טקסטואליים לערכים מספריים מתאימים. המיפוי בוצע למשתנים הבאים:
   2.1. מגדר - הומר לערכים בינאריים, כאשר 0 מייצג זכר ו-1 מייצג נקבה - {"male": 0, "female": 1}.
   2.2. רמת השכלה - מופה באופן סדור לפי סדר ההשכלה, כך שללא השכלה קיבל את הערך הנמוך ביותר, ואוניברסיטה את הערך הגבוה ביותר - {"none": 1, "high school": 2, "university": 3}.
   2.3. רמת הכנסה - מופה בצורה מדורגת, כך שהכנסה נמוכה מקבלת ערך קטן יותר והכנסה גבוהה מקבלת ערך גבוה יותר - {"poverty": 1, "working class": 2, "middle class": 3, "upper class": 4}.
   2.4. שנת ייצור הרכב - הומר לערכים בינאריים - {"before 2015": 0, "after 2015": 1}
   2.5. אזור מגורים - קוד הדואר הומר לערכים מספריים באופן סדור - {10238: 1, 21217: 2, 32765: 3, 92101: 4}.
   2.6. סוג רכב - הומר לערכים בינאריים - {"sedan": 0, "sports car": 1}.

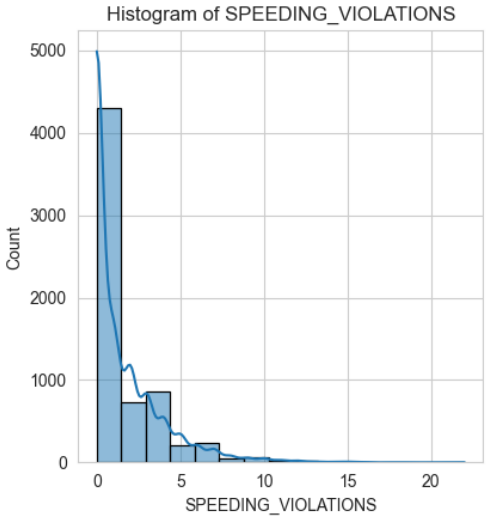
ניתוח קשרים בין המשתנים ולמשתנה המטרה

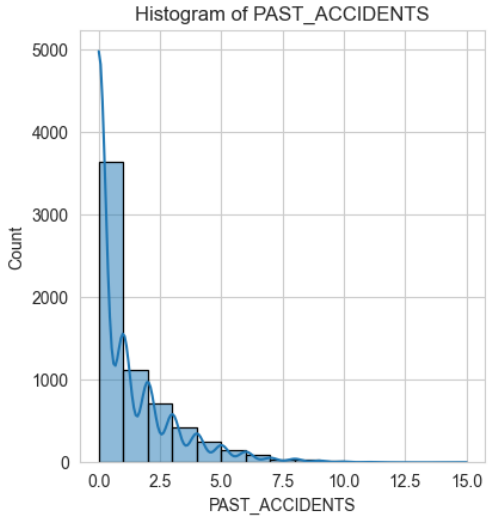
בחלק זה אתאר את התפלגות המאפיינים השונים, אראה את הקשרים בין המשתנים השונים והקשרים למשתנה המטרה. במידה ויהיה צורך לבצע שינוי/ עדכון בגלל קשר בין מאפיינים שונים, אעשה זאת לפני ביצוע המודלים כדי להביא לתוצאות מדויקות יותר במודל חיזוי. כל הגרפים מופיעים בנספח 2.

עבירות ותאונות - הנתונים מצביעים על מגמה הפוכה מהמצופה - נהגים עם יותר עבירות מהירות ותאונות עבר מגישים פחות תביעות ביטוח .ייתכן שהם נמנעים מהגשת תביעה מחשש להתייקרות הפוליסה, בעוד שנהגים זהירים יותר דווקא משתמשים בביטוח. ההיסטוגרמות מציגות התפלגות מרוכזת סביב אפס עבירות ותאונות עם זנב ימני ארוך כלומר רוב הלקוחות אינם מבצעים עבירות ותאונות כלל, ורק מיעוט מהם צובר מספר גבוה של אירועים חריגים. ממצא זה מחזק את הסברה שהגשת תביעה אינה נובעת רק מהיסטוריית נהיגה בעייתית.

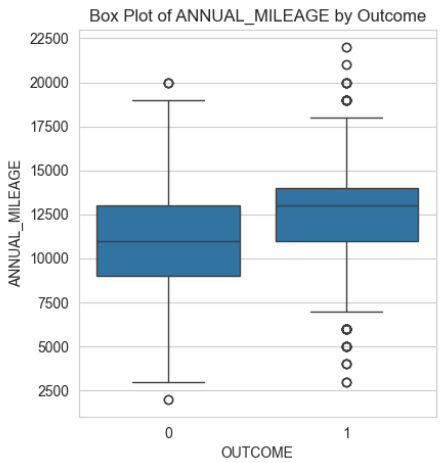




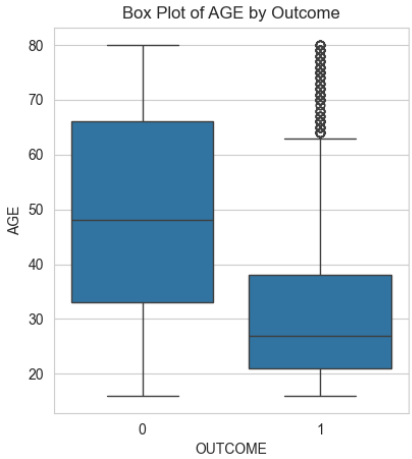




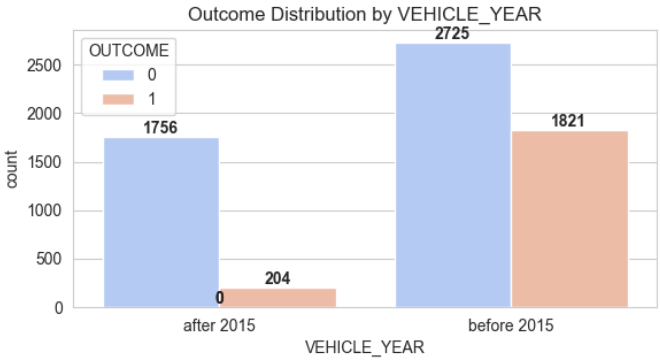
נסועה שנתית (קילומטראז') - הנתונים מראים כי לקוחות שהגישו תביעת ביטוח נוטים לנסוע מעט יותר בממוצע לעומת אלו שלא הגישו תביעה. עם זאת, ההבדלים אינם חדים, ויש חפיפה משמעותית בין הקבוצות.



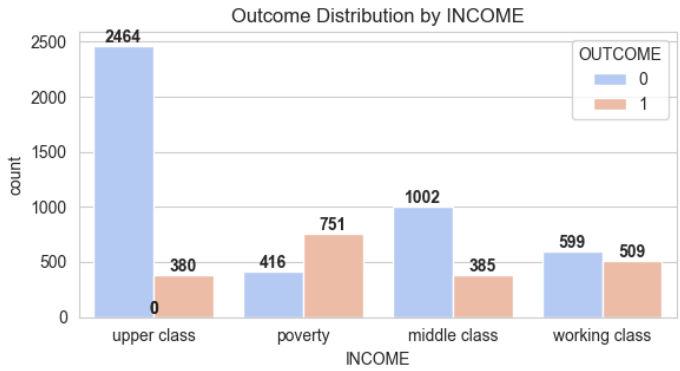
גיל - נהגים מבוגרים מגישים פחות תביעות ביטוח, בעוד שנהגים צעירים מגישים יותר. ניתן לראות שונות גדולה יותר בקרב אלו שלא הגישו תביעות, כלומר טווח הגילאים שם רחב מאוד. לעומת זאת, בקרב המגישים, הרוב המוחלט הם נהגים צעירים. בנוסף, קיימים ערכים חריגים רבים בקרב המגישים, שמייצגים נהגים מבוגרים שהגישו תביעות למרות שהמגמה הכללית הפוכה.



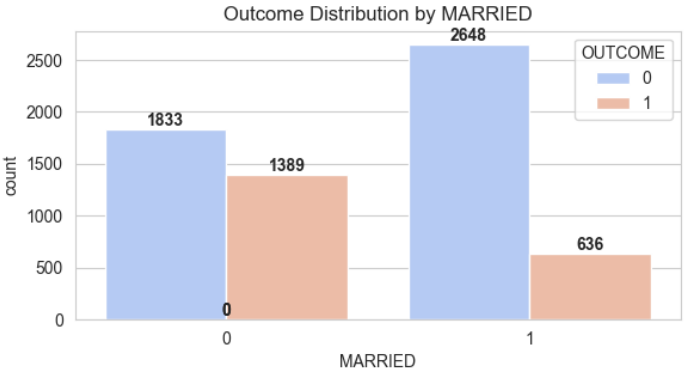
שנת ייצור הרכב - בעלי רכבים ישנים מגישים יותר תביעות ביטוח לעומת בעלי רכבים חדשים. ככל שהרכב מתיישן, כך גדל הסיכון לתקלות מכניות ולנזקים מצטברים, מה שעשוי להסביר את שיעור התביעות הגבוה בקרב רכבים שיוצרו לפני 2015. לעומת זאת, רכבים חדשים מצוידים לרוב במערכות בטיחות מתקדמות, דבר שעשוי להפחית את ההסתברות לתאונה ולצמצם את הצורך בהגשת תביעה.

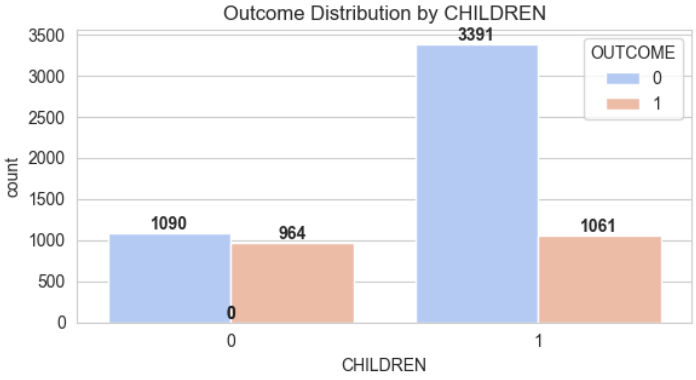


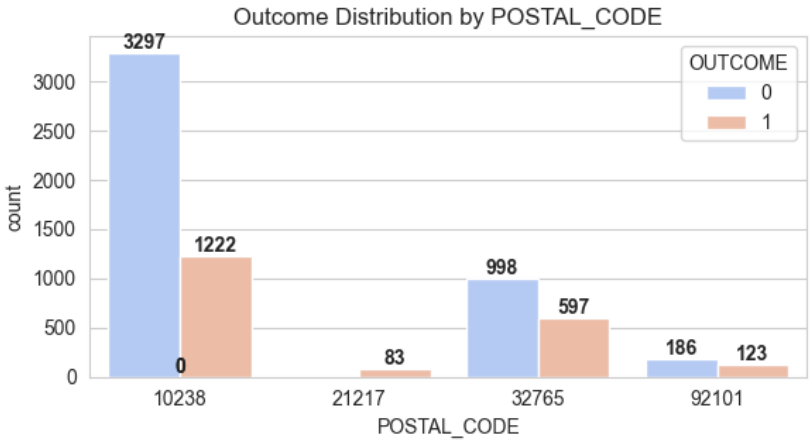
הכנסה - בעלי הכנסה גבוהה מגישים פחות תביעות ביטוח בהשוואה לבעלי הכנסה נמוכה ובינונית. ייתכן כי אנשים במעמד כלכלי גבוה מעדיפים לכסות הוצאות קטנות בעצמם כדי לשמור על עלות הפוליסה נמוכה בטווח הארוך, בעוד שבעלי הכנסה נמוכה נוטים להסתמך על הביטוח יותר בשל מגבלות פיננסיות. אפשרות נוספת היא שבעלי הכנסה גבוהה מחזיקים בפוליסות עם תנאים טובים יותר, המאפשרים להם להימנע מהגשת תביעות על נזקים קטנים.



נשואים וילדים - נהגים נשואים מגישים פחות תביעות ביטוח מאשר נהגים רווקים. הדבר עשוי לנבוע מהרגלי נהיגה זהירים יותר, נסועה נמוכה יותר או יציבות כלכלית שמאפשרת להימנע מהפעלת הביטוח עבור נזקים קטנים. לעומת זאת, נוכחות ילדים אינה משפיעה באופן משמעותי על שיעור הגשת התביעות, שכן אין פערים ניכרים בין נהגים עם וללא ילדים. נראה כי משתנה זה אינו משחק תפקיד מרכזי בהחלטה להפעיל את הביטוח.







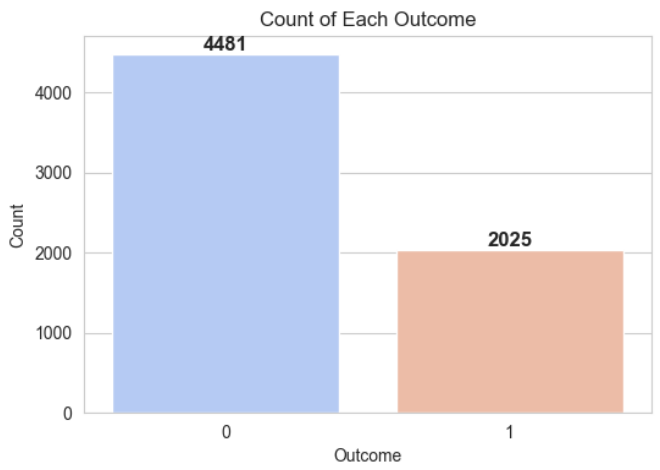
אזור מגורים - אזור המגורים נראה כגורם שמשפיע על שיעור הגשת התביעות, אך ההשפעה אינה אחידה בין האזורים. רוב המבוטחים במדגם מגיעים מאזור אחד, שבו שיעור התביעות ממוצע, בעוד שבאזורים קטנים יותר נרשמו שיעורי תביעות שונים באופן קיצוני. במיוחד בולט אזור שבו כל המבוטחים הגישו תביעה, אך גודלו הקטן מקשה על הסקת מסקנות חד-משמעיות. ייתכן שההבדלים בין האזורים נובעים מתנאים שונים, כגון מאפייני האוכלוסייה, מצב הכבישים או מדיניות ביטוח משתנה, אך ייתכן גם שמדובר בהשפעה שנובעת מגודל המדגם.

במבחנים הסטטיסטיים שנערכו נמצא קשר מובהק בין אזור המגורים לבין שיעור הגשת תביעות, וכן הבדל משמעותי בין האזורים מבחינת שיעורי התביעות. תוצאות אלו מצביעות על כך שאזור המגורים הוא משתנה בעל חשיבות, אך בשל ההבדלים בגודל המדגם בין האזורים, ייתכן שחלק מההשפעה נובעת מהטיות במדגם ולא מהבדל אמיתי בין קבוצות האוכלוסייה. נבחן את השפעת אזור המגורים בהמשך תוך שימוש במודלים השונים כדי להבין עד כמה הוא תורם לניבוי הגשת תביעות.

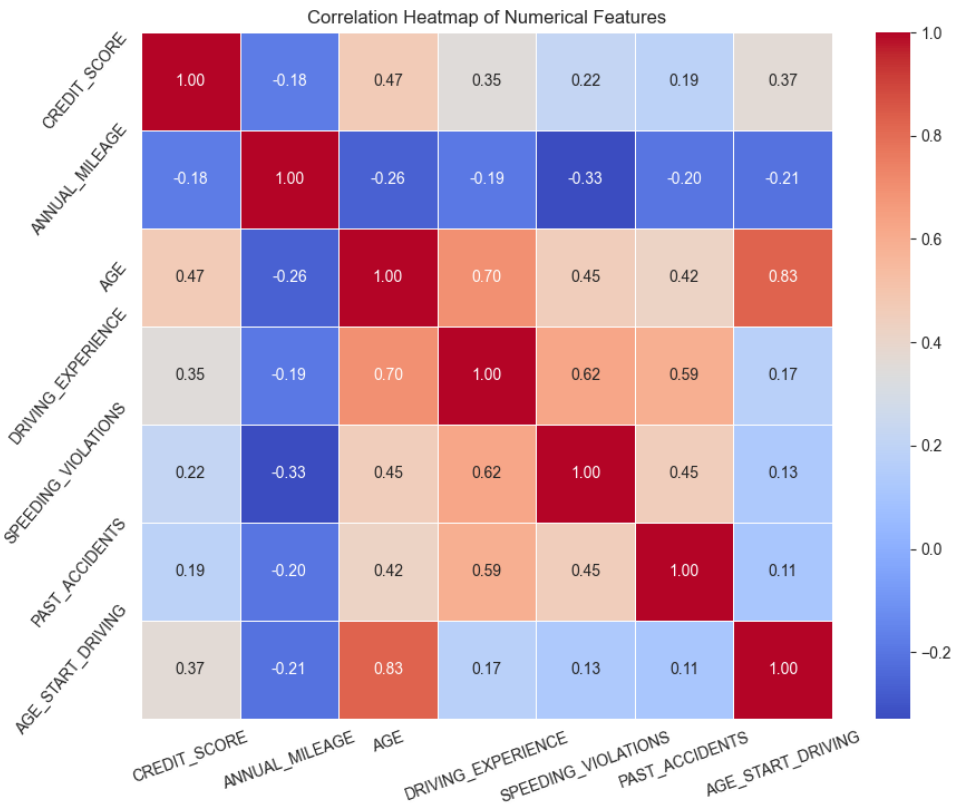
```
--- Analyzing the Influence of Postal Code on Insurance Claims ---

Chi-Square Statistic: 259.253, p-value: 0.00000
Result: There is a statistically significant relationship between postal code and claim rate.

ANOVA F-statistic: 89.949, p-value: 0.00000
Result: Significant differences in claim rates exist between postal codes.
```



הגשת תביעה (משתנה המטרה) - הגרף מציג את מספר הלקוחות שהגישו תביעת ביטוח לעומת אלו שלא. ניתן לראות כי רוב הלקוחות במדגם לא הגישו תביעה, עם 4,481 מקרים לעומת 2,025 מקרים שבהם כן הוגשה תביעה. כלומר, שיעור מגישי התביעות עומד על כ-31% מכלל המבוטחים. הפער בין הקבוצות צפוי והגיוני, שכן לא כל מבוטח נקלע למצב המחייב תביעת ביטוח. עם זאת, ניתן לראות כי מספר מגישי התביעות אינו זניח, כך שהניתוח צריך להתייחס לשתי הקבוצות בצורה מקיפה. חוסר האיזון הזה עשוי להיות רלוונטי בשלב בניית המודל, אך אינו קיצוני ולכן לא בהכרח ישפיע באופן משמעותי.

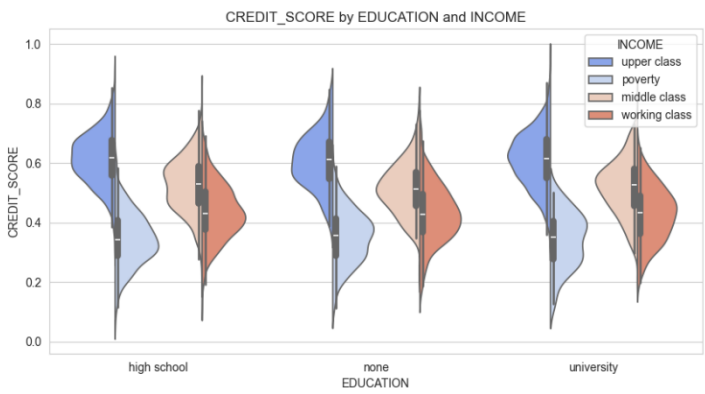


קורלציות - מפת הקורלציות מראה כי אין חשד משמעותי למולטיקולינאריות בין המאפיינים, למעט מספר קשרים בולטים. ניתן לראות קשר חזק וברור בין גיל, ניסיון נהיגה וגיל קבלת הרישיון , כאשר גיל קבלת הרישיון מחושב מתוך שני המשתנים האחרים. בשל החפיפה המשמעותית ביניהם, בעת בניית המודלים לא נשתמש בשלושתם יחד, אלא נבחר את המשתנה הרלוונטי ביותר כדי למנוע תלות מיותרת.

קשר משמעותי נוסף נמצא בין ניסיון נהיגה לבין מספר עבירות התנועה והתאונות, כאשר ככל שניסיון הנהיגה עולה, גם מספר העבירות והתאונות עולה. חשוב לציין שקשר זה לא בהכרח מעיד על סיבתיות - ייתכן שנהגים עם יותר ניסיון פשוט צוברים יותר עבירות ותאונות לאורך השנים, ולא בהכרח שניסיון נהיגה גורם ליותר עבירות ותאונות.

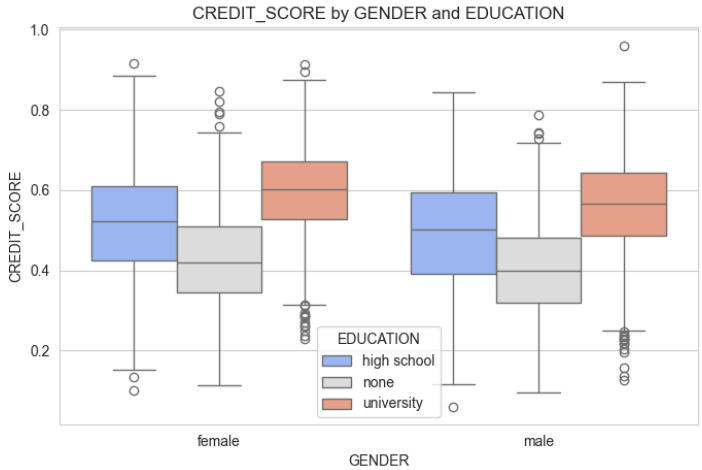
בנוסף, קיים קשר מתון בין גיל לדירוג אשראי, כך שעם העלייה בגיל, דירוג האשראי נוטה להיות גבוה יותר. קשר זה הגיוני שכן מבוגרים נוטים לצבור רקורד פיננסי יציב יותר, דבר שעשוי להשפיע על יכולתם לרכוש פוליסות ביטוח טובות יותר. קשר שלילי נוסף נמצא בין נסועה שנתית (ANNUAL_MILEAGE) לבין מספר עבירות התנועה, מה שעשוי להעיד על כך שנהגים הנוסעים מרחקים גדולים יותר, נזהרים יותר בדרכים או פשוט מפזרים את העבירות שלהם על פני קילומטראז' גדול יותר.

דירוג אשראי לפי רמת השכלה והכנסה - נבדק האם רמת ההשכלה ורמת ההכנסה משפיעות על דירוג האשראי. ניתן לראות שדירוג האשראי גבוה יותר בקרב מי שמשתייכים למעמד הכלכלי הגבוה, ללא תלות משמעותית ברמת ההשכלה. עם זאת, בקרב בעלי הכנסה נמוכה, ניתן לראות שונות גבוהה יותר, מה שעשוי להעיד על השפעה של גורמים נוספים או מעט נתונים בקטגוריה. ניתן לראות לפי טבלת השכיחויות התאמה בין רמת ההכנסה לרמת ההשכלה, לדוגמא בעלי הכנסה גבוה בעלי השכלה אקדמאית בעיקר בעוד שמעמד העובד והביניים בעיקר בעלי השכלה תיכונית.

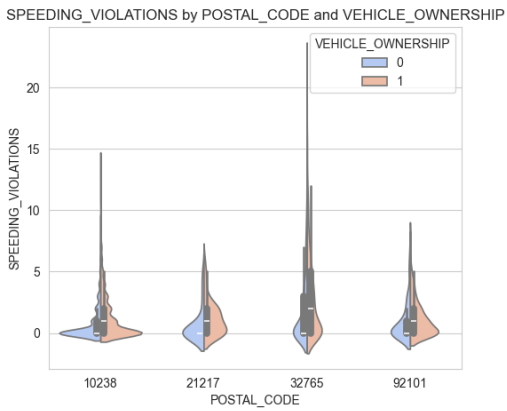


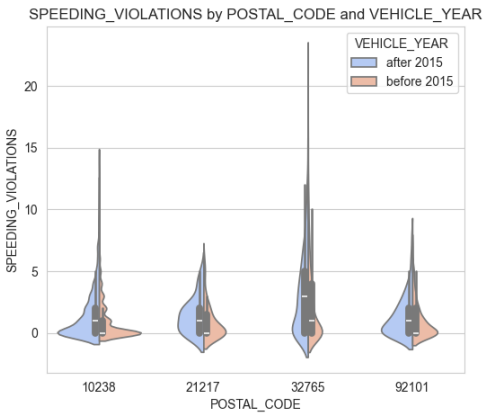
EDUCATION	none	high school	university
INCOME			
poverty	615	494	58
working class	322	571	215
middle class	184	717	486
upper class	101	909	1834

דירוג אשראי לפי מגדר ורמת השכלה - מטרת הבדיקה הייתה לבחון האם יש הבדל בין גברים לנשים מבחינת דירוג האשראי ברמות השכלה שונות. לא נראה פערים גדולים בין המגדרים, אך בקרב חסרי השכלה נראה שדירוג האשראי נמוך יותר באופן כללי, ללא קשר למגדר. זה מחזק את ההשערה כי השכלה עשויה להיות גורם מתווך בין הכנסה לדירוג אשראי (אנשים בעלי השכלה גבוהה יותר יהיו במעמד הכנסה גבוהה יותר ולכן דירוג האשראי שלהם יהיה גבוה יותר).

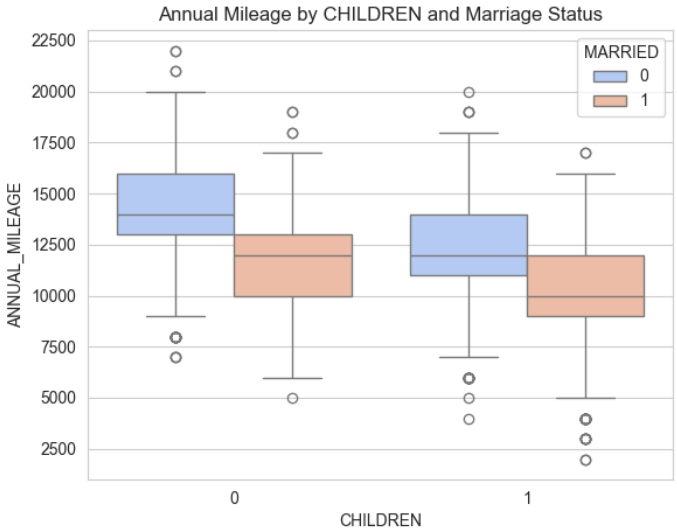


עבירות מהירות לפי אזור ובעלות על רכב או שנת ייצור - הגרפים מציגים את התפלגות מספר עבירות המהירות לפי אזור מגורים, תוך הבחנה בין בעלות על רכב ושנת ייצור הרכב. ניתן לראות שבכל האזורים, רוב הנהגים צוברים מספר נמוך של עבירות, אך קיימים חריגים עם מספר גבוה במיוחד של עבירות, בעיקר באזור 32765. כמו כן, ההבדלים בין קבוצות הבעלות על רכב ושנת הייצור אינם מובהקים, אך יש נטייה קלה לכך שלנהגים עם רכבים חדשים יותר יש שונות גבוהה יותר במספר העבירות, בעיקר באזור 21217.





נסועה שנתית לפי מצב משפחתי - הגרף מציג את התפלגות הנסועה השנתית לפי מצב משפחתי וקיום ילדים. ניתן לראות שבאופן כללי, רמת הנסועה של נשואים נמוכה יותר יחסית לרווקים, ללא תלות בקיום ילדים. כמו כן, קיום ילדים לא מוביל להבדל משמעותי בנסועה השנתית, אך יש שונות גבוהה יותר בקרב רווקים ללא ילדים.



מולטיקולינאריות

ביצעתי בדיקה לזיהוי מולטיקולינאריות בין משתנים שונים באמצעות מדד VIF, מבחני ANOVA וקורלציות מסוג פירסון. מדד VIF נועד לבחון האם יש תלות גבוהה בין משתנים רציפים, דבר שעלול לפגוע בדיוק המודל. מבחני ANOVA מאפשרים לבדוק האם משתנים קטגוריאליים יוצרים הבדלים מובהקים במשתנים רציפים, וכך להבין האם יש קשרים משמעותיים בין סוגי משתנים שונים. בנוסף, השתמשתי בקורלציות מסוג פירסון כדי לבדוק קשרים בין משתנים רציפים לבין משתנים קטגוריאליים שהומרו לערכים מספריים, שכן מדד זה מתאים במיוחד לנתונים מדורגים.

```
Variance Inflation Factor (VIF):
                Feature       VIF
1          CREDIT_SCORE  1.287110
2        ANNUAL_MILEAGE  1.174609
3                   AGE  2.241868
4    DRIVING_EXPERIENCE  3.012700
5   SPEEDING_VIOLATIONS  1.791134
6        PAST_ACCIDENTS  1.577399
```

התוצאות מראות כי אין חשד למולטיקולינאריות משמעותי בין המשתנים המספריים, שכן כל הערכים נמוכים יחסית ואינם חורגים מעבר לרף המקובל. עם זאת, ערכי ה-VIF של גיל וניסיון נהיגה גבוהים יחסית לאחרים, מה שעשוי להצביע על קשר מסוים ביניהם. מאחר וידוע שניסיון נהיגה תלוי ישירות בגיל, ייתכן שבמודלים עתידיים יהיה צורך לבחון האם להכניס את שתי המשתנים הללו, וגם את גיל הוצאת הרישיון שמחושב מהם.

```
ANOVA Test Results for Credit Score:
     Feature  F-Statistic        P-Value
0     INCOME  2675.895725   0.000000e+00
1  EDUCATION   662.103750  1.891070e-262

Spearman Correlation for Income, Education & Credit Score:
                INCOME  EDUCATION  CREDIT_SCORE   OUTCOME
INCOME        1.000000   0.553456      0.742835 -0.410487
EDUCATION     0.553456   1.000000      0.406596 -0.181677
CREDIT_SCORE  0.742835   0.406596      1.000000 -0.320337
OUTCOME      -0.410487  -0.181677     -0.320337  1.000000
```

תוצאות המבחן מראות כי הכנסה והשכלה משפיעות באופן מובהק על דירוג האשראי. ככל שהכנסה והשכלה גבוהות יותר, דירוג האשראי נוטה להיות גבוה יותר.
בנוסף, נמצא קשר שלילי בין הכנסה ודירוג אשראי לבין הגשת תביעות ביטוח, כלומר, בעלי הכנסה נמוכה ודירוג אשראי נמוך נוטים להגיש יותר תביעות. הקשר בין השכלה להגשת תביעה קיים אך חלש יותר. ממצאים אלו מצביעים על חשיבות המשתנים הכלכליים בחיזוי הסבירות לתביעת ביטוח, וכדאי לבדוק האם שילובם למדד יחיד ישפר את יכולת החיזוי.

המבחן מראה כי למצב משפחתי ולמספר הילדים יש השפעה מובהקת על כמות הנסועה השנתית. ניתן להסיק כי נשואים והורים לילדים נוהגים באופן שונה בהשוואה לאחרים.

בנוסף, המתאם מצביע על קשרים ברורים בין מספר תאונות, עבירות תנועה וניסיון נהיגה. ככל שניסיון הנהיגה עולה, כך עולה גם מספר העבירות והתאונות, מה שעשוי להעיד על כך שמשך זמן הנהיגה משפיע על הצבירה של אירועים אלו. כמו כן, נסועה שנתית נמצאת בקשר שלילי עם עבירות ותאונות, מה שעשוי לרמוז כי נהגים שנוסעים יותר הם זהירים יותר או שהרגלי הנהיגה שלהם שונים.

```
ANOVA Test Results for Annual Mileage:
    Feature  F-Statistic        P-Value
0   MARRIED  1810.215576   0.000000e+00
1  CHILDREN  1654.346538  1.877449e-322

Spearman Correlation for Mileage, Accidents, Violations & Driving Experience:
                     ANNUAL_MILEAGE  PAST_ACCIDENTS  SPEEDING_VIOLATIONS  DRIVING_EXPERIENCE   OUTCOME
ANNUAL_MILEAGE             1.000000       -0.199048            -0.307860           -0.189636  0.186677
PAST_ACCIDENTS            -0.199048        1.000000             0.575794            0.661806 -0.386990
SPEEDING_VIOLATIONS       -0.307860        0.575794             1.000000            0.721310 -0.382883
DRIVING_EXPERIENCE        -0.189636        0.661806             0.721310            1.000000 -0.508516
OUTCOME                    0.186677       -0.386990            -0.382883           -0.508516  1.000000
```

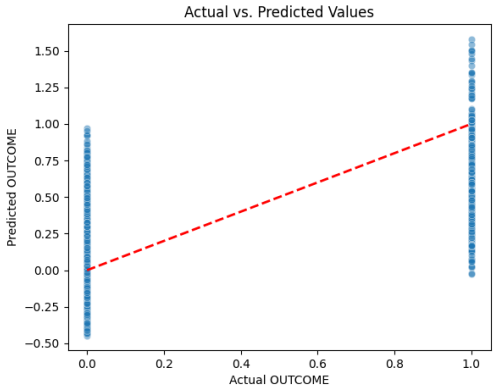
למרות שמצאנו שיש קשרים מובהקים וחזקים בין המשתנים, כרגע לא נשנה את המאפיינים. ייתכן כי בחינות מסוימות שנערכו כעת אינן משקפות את תמונת המצב האמיתית, מאחר ומשתנים נוספים שאינם נכללו בניתוח זה עשויים להשפיע על הקשרים שנצפו. בחלק של בניית המודלים, נבחן האם שילוב של חלק מהמאפיינים למדדים חדשים ומשוקללים יוכל לשפר את ביצועי המודל. שילובים אלו עשויים לסייע בהסברת שונות גבוהה יותר, להפחית השפעות לא רצויות של קשרים בין המשתנים, ולהוביל לתוצאות חיזוי טובות יותר.

בחינת מודלים

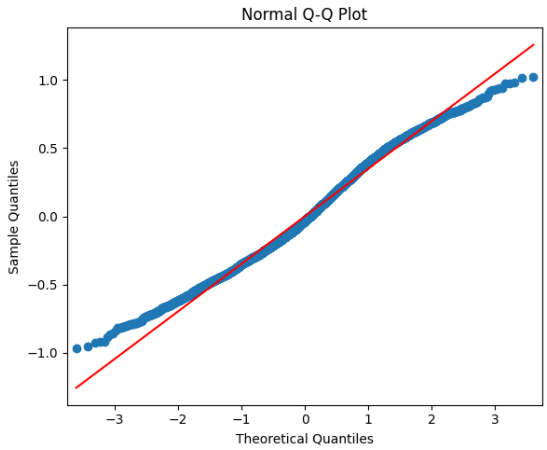
רגרסיה לינארית

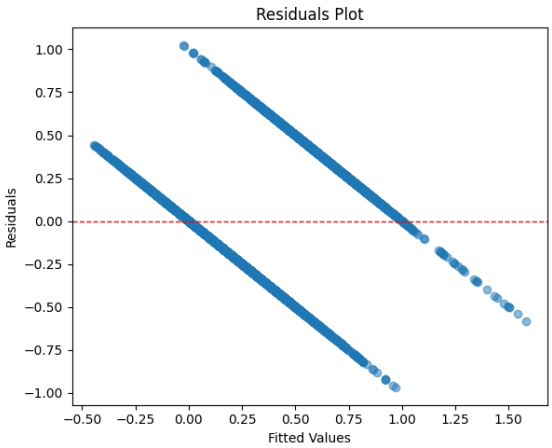
בתחילה, אפעיל רגרסיה לינארית על הנתונים. למרות שמשתנה המטרה הוא בינארי, ורגרסיה לינארית אינה אידיאלית למצבים כאלה, היא תשמש כהשוואת בסיס לבחינת השפעת המשתנים. בנוסף, מרבית המשתנים בסט הנתונים הם קטגוריאליים, מה שעשוי להשפיע על איכות ההתאמה הירודה של המודל הלינארי. נבחן את המודל באמצעות התקדמות בצעדים, כדי לזהות אילו משתנים תורמים להסברת שונות המשתנה התלוי. לאחר מכן, ננתח את תוצאות המודל ונבדוק את הנחותיו.

בנספח 3 מופיע תוצאות המודל. נוכל לראות שאיכות ההתאמה של המודל ע"י R^2_{adj} הוא יחסית נמוך ומסביר רק 43.4% מהשונות של משתנה המטרה. כמו כן בהמשך נשווה מודלים נוספים ע"י מדדי BIC/AIC. ננתח חלק ממשתני המודל - החותך אומר שהסיכוי לתבוע את הביטוח הוא 31% כאשר כל המשתנים הם אפס וקבוצות הבסיס שהיא - מגדר נקבה, לא נשואים, שנת ייצור רכב אחרי 2015 וכדומה. ניתן לראות שעלייה ביחידת גיל (שנה) מורידה את הסיכוי ב-0.43% לתביעת הביטוח ומעבר לקטגוריה של אזור מגורים 21217 מגדילה את הסיכוי ב-74% לתביעת הביטוח (באזור זה כולם תבעו את הביטוח).



ניתן לראות מהתרשימים שהמודל הלינארי לא מתאים לנתונים - התוצאה בפועל היא בינארית אבל המודל עושה חיזוי לטווח הערכים בין 0 ל-1, כלומר מה הסיכוי לדחות ולא האם לדחות. המודל נותן ערכים רציפים לעומת התוצאה הרצויה שהיא בינארית.
בנוסף, הנחות המודל הלינארי לא מתקיימות - ניתן לראות בתרשים Q-Q סטייה מהקו האלכסוני וחצייה שלו מספר פעמים ולכן השגיאות אינן מתפלגות נורמאלית בניגוד להנחות המודל. פיזור השאריות אינן מפוזרות אחיד סביב ה-0 אלא יש מבנה ברור של 2 קווים שנובעים מחוסר התאמה של משתנה מטרה בינארי לרגרסיה לינארית. בנוסף, המבחנים מאשרים שהנחות המודל לא מתקיימות.






```
Shapiro-Wilk Test: W=0.988, p-value=0.000000
Residuals are not normally distributed (Reject H0)
Breusch-Pagan Test: x²=987.798, p-value=0.000000
Heteroscedasticity detected (Reject H0)
```

רגרסיה לוגיסטית

רגרסיה לוגיסטית מתאימה לחיזוי של משתנה מטרה בינארי. בניגוד לרגרסיה לינארית, שבה נוכל לקבל כל ערך רציף, רגרסיה לוגיסטית משתמשת בפונקציית לוג כדי למפות את ערכי המשתנים המסבירים להסתברות שבאמצעותה ניתן להעריך את הסבירות לכך שמבוטח יתבע את הביטוח.

תחילה, כדי לקבל נקודת התייחסות איך ואילו משתנים מסבירים משפיעים על המשתנה המוסבר, בניתי מודל מלא שכלל את כל המאפיינים. המודל בעל ערך של מדד BIC 4,429, שהוא נמוך משמעותית מערכו של המודל הלינארי, ולכן נוכל להגיד שהמודל הלוגיסטי מתאים יותר לנתונים שלנו ולמשתנה המטרה הבינארי. בנספח 4 מוצג סיכום המודל המלא.

עם זאת, המודל מורכב וכולל מספר רב של משתנים ולכן ביצעתי שינויים מתוך הקשרים שראינו בניתוח הנתונים המקדים והתוצאות עד כה:

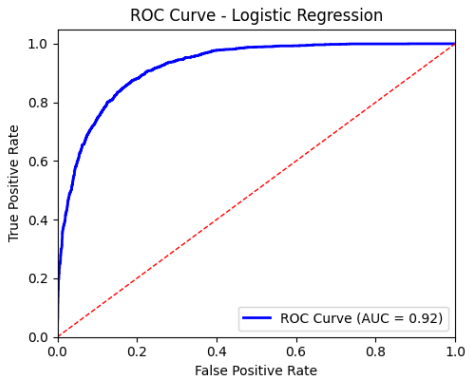
- איחוד אזורי המגורים למעט אזור 21217 בו כל המבוטחים תבעו את הביטוח.
- יצירת מדד סוציו-אקונומי המאחד את המשתנים השכלה והכנסה.
- יצירת מדד סיכון המאחד את המשתנים מספר עבירות ותאונות.
- אינטרקציה בין אזור מגורים לרמת סיכון - כדי לבחון האם הקשר ביניהם משפיע באופן שונה על ההסתברות לתביעה.
- הסרת משתנים לא מובהקים.

לאחר מכן, ביצעתי מודל עם הנתונים שנשארו וקיבלנו מודל מצומצם יותר ושמסביר את הנתונים ומשתנה המטרה בצורה טובה יותר מהמודל המלא. ערך BIC של מודל זה הוא 4,329 ולאחר מבחן LRT קיבלנו שאין הבדל מובהק סטטיסטי בין המודלים, כלומר המודל המצומצם מסביר באותה מידה כמו המודל המלא מבלי לאבד מידע חשוב. בנספח 5 מוצג סיכום המודל המצומצם.

Metric	Full Model	Reduced Model
BIC	4429.006721	4329.190434
AIC	4279.836159	4247.824672
Pseudo R^2	0.47503	0.476518
Log-Likelihood Ratio	-	-12.011486
Degrees of Freedom	-	10.000000
p-value	-	1.000000

ניתוח המודל המצומצם - כמעט כל המשתנים מובהקים, למעט אזור מגורים 21217, אשר מגדיל את הסיכוי בצורה משמעותית לתביעה (מפני שכולם תבעו את הביטוח באזור זה), ומדד הסוציו-אקונומי. למרות זאת השארתי אותם במודל מפני שהם מסבירים לדעתי ערכים חשובים במודל שתורמים להבנתו.

פירוש מקדמי המודל - גברים ביחס לנשים (קבוצת הבסיס) בהסתברות לתבוע את הביטוח פי $e^{1.1491} = 3.15$. יש קשר בין מדד הסיכון לאזור המגורים, ניתן לראות שיש השפעה חיובית, כלומר, אזור המגורים מגדיל פי 1.39 את הסיכוי להגשת תביעה.



בסופו של דבר, הוצאתי גרף שמתאר את היחס בין זיהוי נכון לזיהוי שגוי של מקרי תביעה. ככול שהעקומה רחוקה מקו האמצע ושהערך קרוב ל-1, כך המודל יותר טוב.

לכן נוכל להגיד על מודל זה שהוא מבחין היטב בין כאלו שתובעים את הביטוח לכאלו שלא.

תוצאות ומסקנות

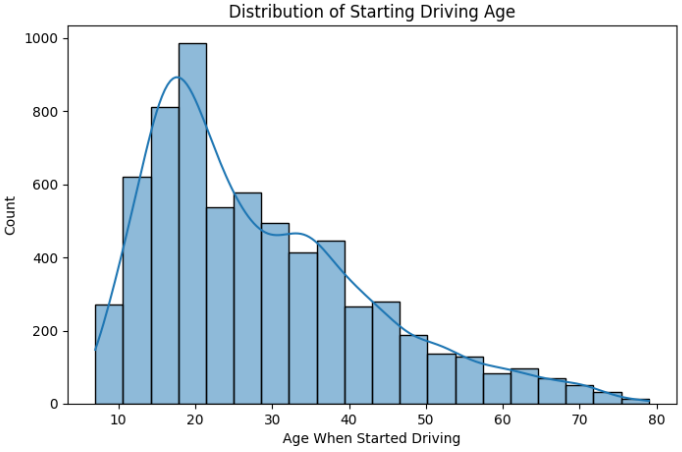
המודל הנבחר הוא המודל הלוגיסטי המצומצם. בחנו את המודלים לפי מורכבות והתאמה במדדים של R^2_{adj} ו-BIC, והוא המודל שהביא את התוצאות הטובות ביותר לעומת המודלים האחרים, משמע הוא מאוזן בין התאמה לנתונים לבין פשטות המודל. נספח 5 מציג את תוצאות המודל.

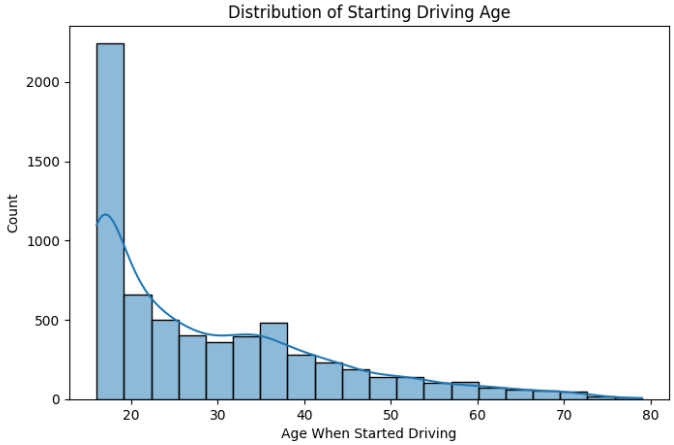
מודל זה מנסה להעריך את הסיכוי של לקוח להגיש תביעת ביטוח בהתבסס על משתנים מסבירים.
משמעות החותך במודל הוא הסיכוי לתביעת ביטוח כאשר כל הערכים הינם 0 לערכים הרציפים ובקבוצת הבסיס עבור המשתנים הבינאריים והקטגוריאליים. במקרה שלנו, קבוצת הבסיס היא נשים, לא נשואות, לא בעלות רכב, עם רכב שיוצר לאחר שנת 2015 - והסיכוי שלהן לתבוע את הביטוח הוא 19.3%.
משמעות המשתנים הוא השינוי בסיכוי לתביעת הביטוח כאשר אנו מגדילים ביח' את ערך המשתנה, לדוגמא - תוספת בשנת ניסיון של מבוטח מעלה את הסיכוי לתביעת ביטוח פי $e^{-0.1485}$, ומעבר לקבוצת הגברים מעלה את הסיכוי לתביעת ביטוח פי $e^{1.1491}$.

המודל מספק לנו כלי חיזוי אפקטיבי לאיתור לקוחות בסיכון גבוה לתביעת ביטוח. ניתן להשתמש בו לצורך תמחור פוליסות מבוסס-סיכון, שיפור מדיניות ביטוחית, והבנת גורמים מרכזיים המשפיעים על תביעות.
המשתנים השונים מאפשרים לדעת את הסיכוי הנתון לתביעת ביטוח וע"פ כך לחשב את הפרימייה ללקוח.

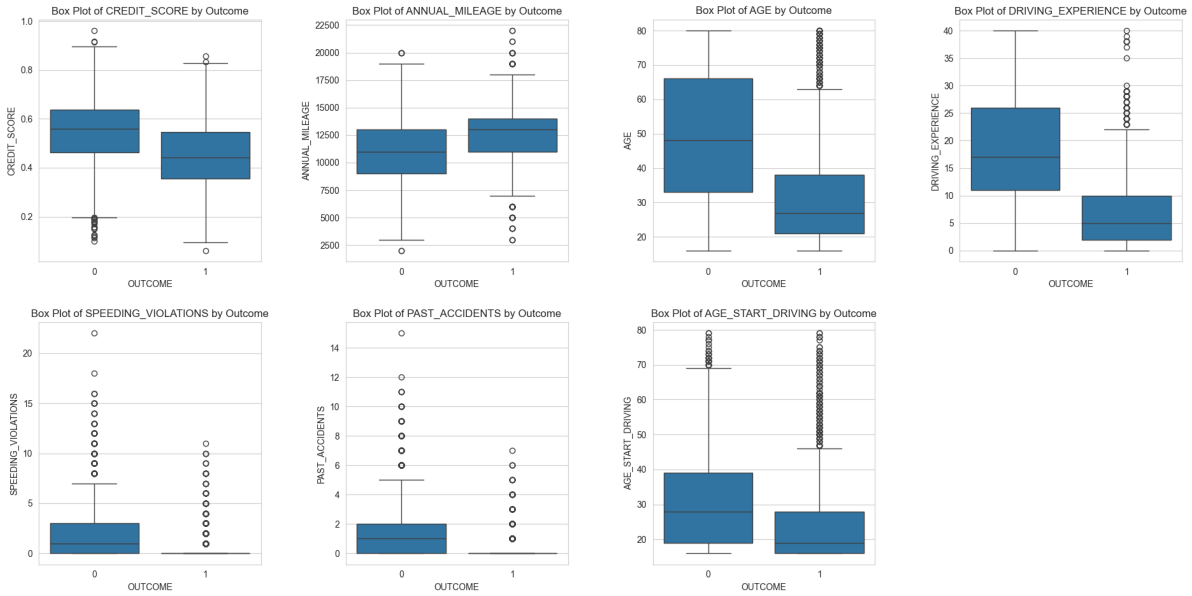
נספחים

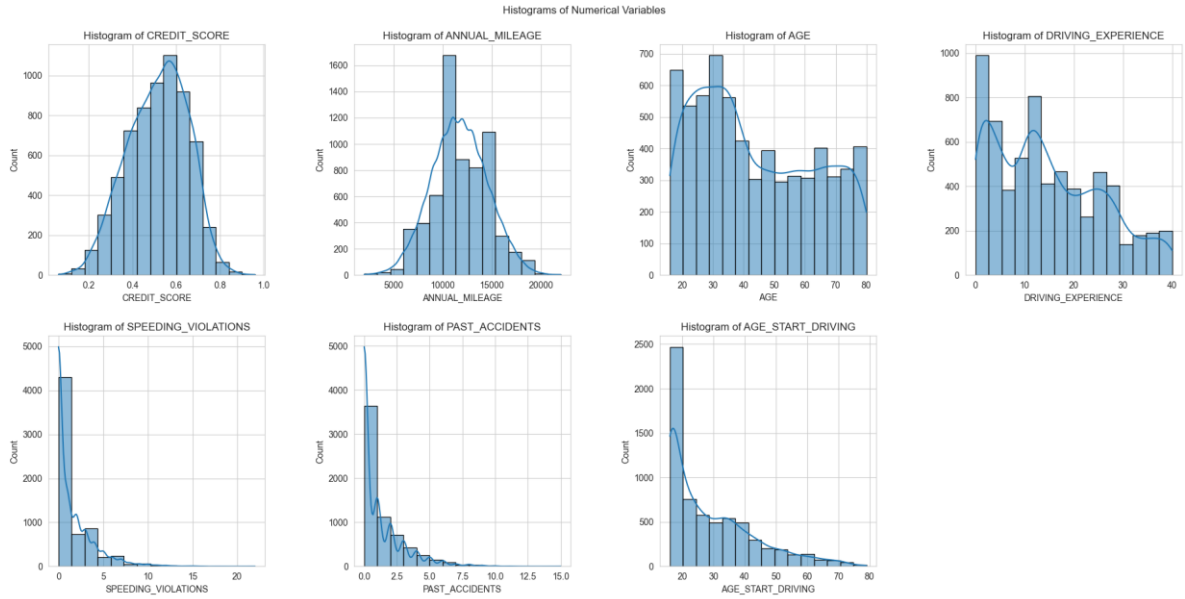
נספח 1 - גיל הוצאת רישיון

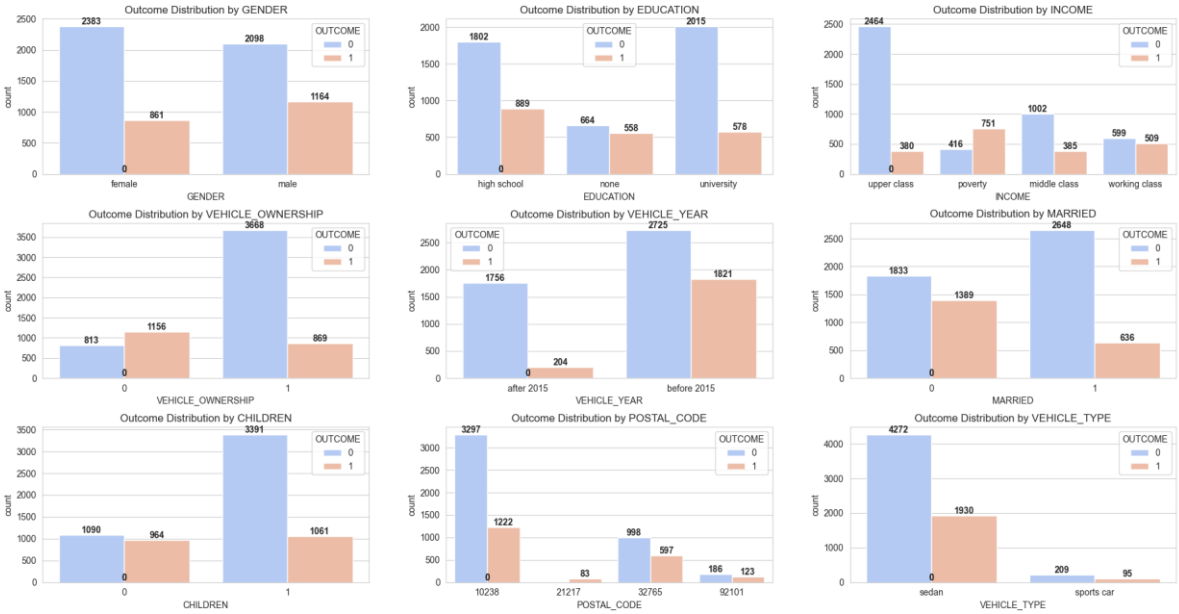




נספח 2 - גרפים על המאפיינים והקשרים




Histograms of Numerical Variables
Histogram of CREDIT_SCORE
Histogram of ANNUAL_MILEAGE
Histogram of AGE
Histogram of DRIVING_EXPERIENCE
Histogram of SPEEDING_VIOLATIONS
Histogram of PAST_ACCIDENTS
Histogram of AGE_START_DRIVING
Count
CREDIT_SCORE
ANNUAL_MILEAGE
AGE
DRIVING_EXPERIENCE
SPEEDING_VIOLATIONS
PAST_ACCIDENTS
AGE_START_DRIVING


Outcome Distribution by GENDER
2383
861
2098
1164
female
male
GENDER
Outcome Distribution by EDUCATION
1802
889
664
558
2015
578
high school
none
university
EDUCATION
Outcome Distribution by INCOME
2464
380
416
751
1002
385
599
509
upper class
poverty
middle class
working class
INCOME
Outcome Distribution by VEHICLE_OWNERSHIP
813
1156
3668
869
VEHICLE_OWNERSHIP
Outcome Distribution by VEHICLE_YEAR
1756
204
2725
1821
after 2015
before 2015
VEHICLE_YEAR
Outcome Distribution by MARRIED
1833
1389
2648
636
MARRIED
Outcome Distribution by CHILDREN
1090
964
3391
1061
CHILDREN
Outcome Distribution by POSTAL_CODE
3297
1222
83
998
597
186
123
10238
21217
32765
92101
POSTAL_CODE
Outcome Distribution by VEHICLE_TYPE
4272
1930
209
95
sedan
sports car
VEHICLE_TYPE
OUTCOME
count

נספח 3 - תוצאות המודל הלינארי

```
=== Stepwise Selection Linear Model Summary ===
                            OLS Regression Results
==============================================================================
Dep. Variable:                OUTCOME   R-squared:                       0.434
Model:                            OLS   Adj. R-squared:                  0.432
Method:                 Least Squares   F-statistic:                     331.5
Date:                Thu, 20 Feb 2025   Prob (F-statistic):               0.00
Time:                        23:16:28   Log-Likelihood:                -2371.7
No. Observations:                6506   AIC:                             4775.
Df Residuals:                    6490   BIC:                             4884.
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
DRIVING_EXPERIENCE        -0.0094      0.000    -23.944      0.000      -0.010      -0.009
const                      0.3132      0.036      8.810      0.000       0.244       0.383
VEHICLE_OWNERSHIP         -0.2548      0.010    -24.726      0.000      -0.275      -0.235
VEHICLE_YEAR_before 2015   0.1641      0.010     16.470      0.000       0.145       0.184
POSTAL_CODE_21217          0.7427      0.039     19.200      0.000       0.667       0.818
GENDER_male                0.1180      0.009     13.083      0.000       0.100       0.136
POSTAL_CODE_32765          0.1484      0.012     12.802      0.000       0.126       0.171
ANNUAL_MILEAGE          1.276e-05   2.05e-06      6.213      0.000    8.73e-06    1.68e-05
POSTAL_CODE_92101          0.1545      0.021      7.527      0.000       0.114       0.195
INCOME_poverty             0.1187      0.016      7.504      0.000       0.088       0.150
MARRIED                   -0.0307      0.010     -2.932      0.003      -0.051      -0.010
SPEEDING_VIOLATIONS       -0.0085      0.003     -3.042      0.002      -0.014      -0.003
INCOME_working class       0.0454      0.014      3.343      0.001       0.019       0.072
CHILDREN                  -0.0291      0.011     -2.642      0.008      -0.051      -0.008
AGE_START_DRIVING          0.0051      0.000     17.362      0.000       0.005       0.006
AGE                       -0.0043      0.000    -18.701      0.000      -0.005      -0.004
EDUCATION_none            -0.0261      0.012     -2.089      0.037      -0.051      -0.002
==============================================================================
```

נספח 4 - תוצאות המודל הלוגיסטי המלא

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                OUTCOME   No. Observations:                 6506
Model:                          Logit   Df Residuals:                     6484
Method:                           MLE   Df Model:                           21
Date:                Sat, 22 Feb 2025   Pseudo R-squ.:                  0.4750
Time:                        19:11:08   Log-Likelihood:                -2117.9
converged:                       False   LL-Null:                       -4034.4
Covariance Type:            nonrobust   LLR p-value:                     0.000
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     -1.8815      0.391     -4.817      0.000      -2.647      -1.116
CREDIT_SCORE               0.3374      0.429      0.787      0.431      -0.503       1.178
VEHICLE_OWNERSHIP         -1.9678      0.088    -22.248      0.000      -2.141      -1.794
MARRIED                   -0.2821      0.093     -3.043      0.002      -0.464      -0.100
CHILDREN                  -0.0980      0.092     -1.070      0.284      -0.277       0.081
ANNUAL_MILEAGE             0.0001   1.87e-05      6.792      0.000    9.04e-05       0.000
SPEEDING_VIOLATIONS       -0.0790      0.035     -2.272      0.023      -0.147      -0.011
PAST_ACCIDENTS            -0.2016      0.048     -4.181      0.000      -0.296      -0.107
AGE                       -0.0477        nan        nan        nan        nan         nan
DRIVING_EXPERIENCE        -0.0999        nan        nan        nan        nan         nan
AGE_START_DRIVING          0.0522        nan        nan        nan        nan         nan
GENDER_male                1.1541      0.085     13.652      0.000       0.988       1.320
EDUCATION_none            -0.0241      0.107     -0.226      0.821      -0.233       0.185
EDUCATION_university      -0.0102      0.099     -0.104      0.917      -0.203       0.183
INCOME_poverty             0.1385      0.150      0.924      0.356      -0.155       0.432
INCOME_upper class        -0.1100      0.130     -0.847      0.397      -0.364       0.144
INCOME_working class       0.0901      0.124      0.724      0.469      -0.154       0.334
VEHICLE_YEAR_before 2015   1.9115      0.111     17.298      0.000       1.695       2.128
VEHICLE_TYPE_sports car   -0.0978      0.181     -0.542      0.588      -0.452       0.256
POSTAL_CODE_21217         25.5245   3945.492      0.006      0.995   -7707.498    7758.547
POSTAL_CODE_32765          1.2919      0.106     12.189      0.000       1.084       1.500
POSTAL_CODE_92101          1.4089      0.177      7.970      0.000       1.062       1.755
==========================================================================================
Pseudo R² (McFadden's R²): 0.4750
AIC: 4279.8362
BIC: 4429.0067
```

נספח 5 - תוצאות המודל הלוגיסטי המצומצם

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                OUTCOME   No. Observations:                 6506
Model:                          Logit   Df Residuals:                     6494
Method:                           MLE   Df Model:                           11
Date:                Sat, 22 Feb 2025   Pseudo R-squ.:                  0.4765
Time:                        19:11:08   Log-Likelihood:                -2111.9
converged:                       False   LL-Null:                       -4034.4
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                        -1.6431      0.269     -6.100      0.000      -2.171      -1.115
VEHICLE_OWNERSHIP            -1.9933      0.086    -23.282      0.000      -2.161      -1.825
MARRIED                      -0.2767      0.090     -3.080      0.002      -0.453      -0.101
ANNUAL_MILEAGE                0.0001   1.69e-05      8.096      0.000       0.000       0.000
DRIVING_EXPERIENCE           -0.1485      0.007    -20.317      0.000      -0.163      -0.134
GENDER_male                   1.1491      0.083     13.809      0.000       0.986       1.312
VEHICLE_YEAR_before 2015      1.9183      0.108     17.791      0.000       1.707       2.130
POSTAL_CODE_21217            31.8559   8.68e+04      0.000      1.000    -1.7e+05     1.7e+05
POSTAL_CODE_OTHER             1.1194      0.105     10.688      0.000       0.914       1.325
ECONOMIC_INDEX               -0.0091      0.018     -0.505      0.614      -0.044       0.026
RISK_INDEX                   -0.4377      0.070     -6.253      0.000      -0.575      -0.300
RISK_POSTAL_INTERACTION       0.3336      0.083      4.036      0.000       0.172       0.496
==============================================================================================
Pseudo R² (McFadden's R²): 0.4765
AIC: 4247.8247
BIC: 4329.1904
```