פרויקט גמר בנושאים נבחרים בסטטיסטיקה

14/3/2024

מאי חכים ת.ז. 209043017
גלעד ארז ת.ז. 314898453
ליאל פרטוש ת.ז. 312196561

תוכן עניינים

תקציר

בפרויקט זה בחנו את החשיבות של בחירת משתנים בשני סוגים של בעיות: רגרסיה וסיווג. בבעיית הרגרסיה, בחנו את ההשפעה של שינוי הקריטריון באלגוריתם Stepwise Regression (הקנס) על בחירת המשתנים למודל ועל יכולת החיזוי שלו. השערת המחקר שלנו הייתה שקנס נמוך יוביל לריבוי משתנים מסבירים ולהתאמת יתר, וקנס גבוה יוביל למיעוט משתנים ומודל לא יעיל, ועל כן הקנס האידיאלי יהיה בנקודה מסוימת באמצע, זאת ללא קשר לכמות המשתנים במודל ההתחלתי (ריק, מלא או "בינוני"). בבעיית הסיווג, בחנו את השיטות הבאות לבחירת משתנים: חשיבות משתנים בעזרת Random Forest ו-XGBoost, דירוג המשתנים בעזרת Fisher Score ובחירת משתנים באמצעות Chi Squared test לאי תלות. את יעילות השיטות השונות בחנו באמצעות ארבעה מודלי סיווג: רגרסיה לוגיסטית, Random Forest, XGBoost ורשת נוירונים. השערת המחקר שלנו הייתה כי שיטה לבחירת משתנים תתאים למודלים שמנגנון הסיווג שלהם דומה למנגנון הדירוג שלה. הנתונים בהם השתמשנו לבדיקת ההשערות שלנו הם רשימת השירים הכי מושמעים ב-Spotify לשנת 2023, כאשר משתנה המטרה הוא כמות ההשמעות של שיר. במודל הרגרסיה רצינו לחזות את מספר ההשמעות ובמודלי הסיווג חילקנו את השירים לפופולריים או לא פופולריים עפ"י כמות ההשמעות שלהם (אם מספר ההשמעות גדול מערך החציון- השיר פופולרי). בניסוי הרגרסיה, ראינו כי מספר המשתנים הנבחרים השתנה עבור ערכים שונים של גודל הקנס ומודלים התחלתיים שונים. כאשר הקנס קטן יחסית, האלגוריתם הכניס יותר משתנים למודל. ריבוי משתנים אמנם שיפר את התוצאות על סט האימון אך לא נצפתה מגמה ברורה בסט הבחינה ועל כן לא מצאנו מודל התחלתי אופטימלי ל- Stepwise Regression. בניסוי הסיווג, לא מצאנו שיטה אידיאלית עבור אף מודל ברמת ביטחון של 90%.

מבוא

בתחום הסטטיסטיקה ולימוד מכונה, בחירת המשתנים היא בעלת חשיבות מרכזית בחידוד מודלי חיזוי. עיסוק זה רלוונטי גם בבעיות רגרסיה וגם בבעיות סיווג, כאשר היעילות של מודלים תלויה במשתנים שנבחרו בתבונה. בפרויקט זה נבחן את ההשפעה של שיטות שונות לבחירת משתנים על הביצועים של מודל רגרסיה ושל מודלי סיווג.

ניתוח רגרסיה: ברגרסיה, אנחנו מתמקדים בבדיקת אלגוריתם Stepwise-Regression. באופן ספציפי, אנו שואפים לפענח את ההשפעה של שינוי הקריטריון של האלגוריתם (רכיב הקנס על הוספת משתנים), לצד בחירת נקודת ההתחלה - בין אם זה מודל ריק, מלא או "בינוני".

ניתוח סיווג: בתחום הסיווג אנו מתמקדים בזיהוי שיטות מתאימות לביצוע Feature Selection על פני ארכיטקטורות מודלים שונות.

לשם כך, אנו משתמשים במערך נתונים שמקורו באתר Kaggle, המאגד את השירים המושמעים ביותר ב-Spotify בשנת 2023 ומספק תכונות הקשורות לכל שיר.

החלק המרכזי בניתוח שלנו הוא חיזוי הפופולריות של השירים, הבא לידי ביטוי בחיזוי המשתנה streams - כמות ההשמעות ב-Spotify בשנת 2023.

בניתוח הרגרסיה אנו בוחנים את הביצועים של מודלים שאומנו עם קומבינציות שונות של גודל הקנס ושל כמות המסבירים ההתחלתית.

בניתוח מודלי הסיווג אנו בוחנים את ההשפעה של ארבע שיטות לבחירת משתנים: Feature Importance של Random Forest ושל XGBoost, דירוג עפ"י Fisher Score ו-Chi Squared test לאי תלות. בעזרת השיטות השונות אנו מאמנים ארבעה מודלי סיווג

שונים- רגרסיה לוגיסטית, Random Forest, XGBoost ורשת נוירונים במטרה לבדוק האם יש שיטת בחירת משתנים מסוימת שמתאימה לכל מודל סיווג באופן מובהק.

מטרות המחקר

בפרויקט השווינו בין דרכים שונות לבחירת משתנים, בשני מקרים שונים: בעיית רגרסיה ובעיית סיווג.

- עבור רגרסיה בחנו את אלגוריתם Stepwise-Regression.
  שאלה: כיצד משפיעים שינויים בקריטריון של האלגוריתם (או ליתר דיוק, בקנס על הוספת משתנים) ונקודת ההתחלה (מודל ריק, מלא, או בינוני), על ביצועי המודל על סט האימון וסט הבחינה?
  השערה: קנס נמוך מדי יביא לריבוי משתנים, ולהתאמת יתר. קנס גבוה מדי יביא למיעוט משתנים ולמודל לא מוצלח. הקנס האידיאלי יהיה קנס "בינוני" בלי קשר למודל ההתחלתי.
- עבור סיווג בדקנו שיטות שונות לבחירת משתנים.
  שאלה: אילו שיטות לחישוב Feature Selection יתאימו לאילו סוגי מודלים?
  השערה: שיטה לבחירת משתנים תתאים למודלים שמנגנון הסיווג שלהם דומה למנגנון הדירוג שלה. למשל: דירוג לפי XGBoost מתאים למודלים מבוססי עצים.

שיטה

הנתונים:

מאגר הנתונים שלנו נלקח מאתר Kaggle ומציג את השירים הכי מושמעים ב-Spotify בשנת 2023. האתר לא מציין כיצד הנתונים נאספו.

עבור כל שיר תועדו: שם השיר, שם/ות האמן/ים, מספר האמנים (שתרמו לשיר), תאריך הפרסום (שנה, חודש ויום בחודש), בכמה רשימות השמעה של ספוטיפיי השיר מופיע (נאסף גם עבור Apple ו-Deezer, פלטפורמות האזנה אחרות), האם השיר מופיע בדירוג של ספוטיפיי ואם כן באיזה מקום מדורג (נאסף גם עבור Apple, Deezer ו-Shazam), מדד bpm- beats per minute, מפתח, סוג סולם (מינור/מג'ור), danceability% (מדד המציין עד כמה ניתן לרקוד לצלילי השיר), valence% (עד כמה התוכן של השיר חיובי), energy% (עד כמה השיר נתפס כאנרגטי), acousticness% (כמות הצלילים האקוסטים בשיר), instrumentalness% (כמות התוכן האינסטרומנטלי בשיר), liveness% (כמות המוזיקה החיה בשיר), speechiness% (כמות המילים בשיר) ו-streams (כמות ההשמעות ב-Spotify בשנת 2023). המשתנה אשר בחרנו לחזות הוא streams - כמות ההשמעות ב-Spotify בשנת 2023.

מתוך 974 הרשומות הקצנו 754 רשומות לסט האימון ואת ה-200 הנותרות לסט הבחינה.

הזנת הנתונים:

עקב טעות בשלב ה-data acquisition בערך של המשתנה Streams ברשומה הבאה, הסרנו אותה.

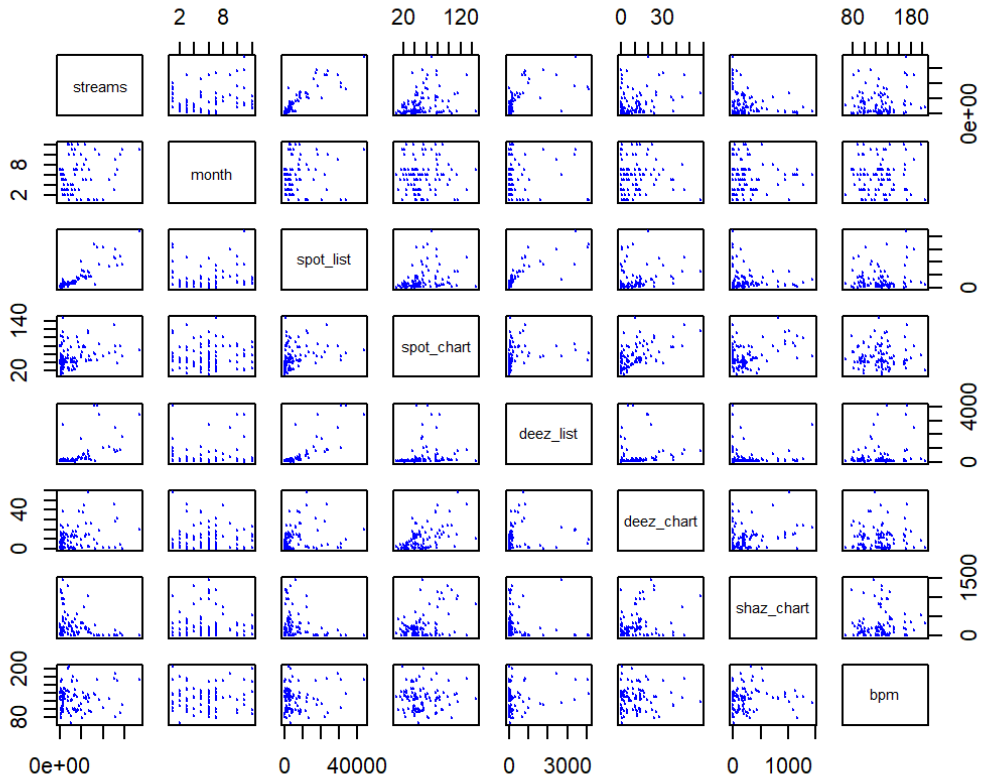
track_name	artist(s)_name	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists	in_apple_charts	in_deezer_playlists	in_deezer_charts	in_shazam_charts
Love Grows (	Edison Li	1	1970	1	1	2877	0	BPM110KeyAMo	16	0	54	0	0

ערכים חסרים:

המשתנים בהם היו רשומות עם ערכים חסרים הם: key ו-in_shazam_charts (האם השיר בדירוג של Shazam). השלמנו את הערכים החסרים באופן הבא: כשליש מהערכים בעמודת in_shazam_charts היו 0 ולכן כך גם השלמנו את הערכים החסרים. את הערכים החסרים במשתנה key השלמנו בערך "Unknown" כיוון שיש שירים ללא מפתח.

קורלציות בין משתנים:

בתרשים הבא חיפשנו קורלציות בין חלק מהמשתנים הכמותיים. ניתן לראות קורלציה חיובית בין משתנה המטרה למשתנים הקשורים לפלטפורמות האזנה שונות: in_spotify_playlists, in_spotify_charts, in_deezer_playlists, in_deezer_charts, in_shazam_charts. הקורלציות האלה (ובעיקר הקורלציה עם in_spotify_playlists שנראית ברורה יותר מהאחרות) מסתדרות עם ההיגיון לפיו שירים המופיעים ברשימות השמעה/דירוג שירים מובילים כנראה פופולריים ויושמעו פעמים רבות. בייחוד רשימות השמעה ב-Spotify, שכן זאת פלטפורמת ההאזנה של משתנה המטרה. המשתנים הללו גם בקורלציה בינם לבין עצמם, ולכן נצפה לראות את חלקם בתור משתנים מסבירים אך לא בהכרח את כולם, כיוון שזה עלול ליצור מולטיקולינאריות. בנספח 1 מצורף החלק השני של התרשים, נציין כי הקורלציות בו פחות ברורות. נספח 2 מתאר את הקורלציות בין המשתנים בצורה של מפת חום ותומך בקורלציות שתוארו לעיל.



ניסוי 1: רגרסיה

המודל: רגרסיה לינארית מרובה, עם משתני דמה ואינטראקציה.

הנחות המודל:

- לינאריות - הקשר בין המסבירים למוסבר הוא קשר לינארי.
- שוויון שונויות השגיאות - שונות השגיאה קבועה ולא תלויה בערכי המסבירים או המוסבר.
- נורמליות השגיאות - השגיאה מתפלגת נורמלית עם שונות קבועה ותוחלת 0.
- אי תלות השגיאות - שתי שגיאות שונות אינן תלויות זו בזו.

מהלך הניסוי:

כדי להבין כיצד משפיעים **הקנס-פר-פרמטר** ו**המודל ההתחלתי** באלגוריתם Stepwise-Regression, הרצנו את האלגוריתם עם קומבינציות שונות של גודל הקנס ונקודת ההתחלה, ובכל שילוב בדקנו את ביצועי המודל על סט האימון שלו - ועל סט בחינה בלתי מעורב. הביצועים נמדדו על פי מדד MSE, שהוא גודל השגיאה הריבועית הממוצע של המודל על סט הנתונים.

נזכיר שמדד BIC (Bayes Information Criterion) מחושב באופן הבא:

$$BIC = -2\ln(L) + k*\ln(n)$$

כש-k הוא מספר הפרמטרים במודל (בלי החותך), n מספר התצפיות, ו-L היא נראות המודל והנתונים.

אנחנו הוספנו פרמטר (נקרא לו s) שישמש כקבוע לשליטה בגודל הקנס. כעת המדד הוא:

$$metric = -2\ln(L) + s*k*\ln(n)$$

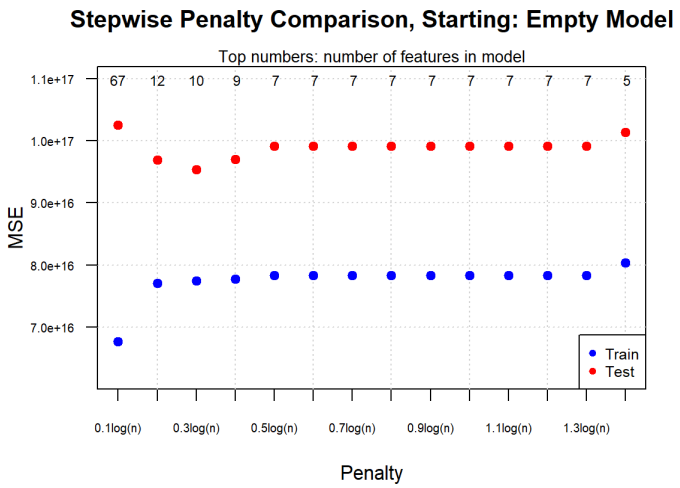
כאשר נתנו לאלגוריתם לרוץ עם ערכי s שנעים מ0.1 עד 1.4 בקפיצות של 0.1.

את הבדיקה הזו עשינו כשנקודת ההתחלה של האלגוריתם היא אחת משלוש:

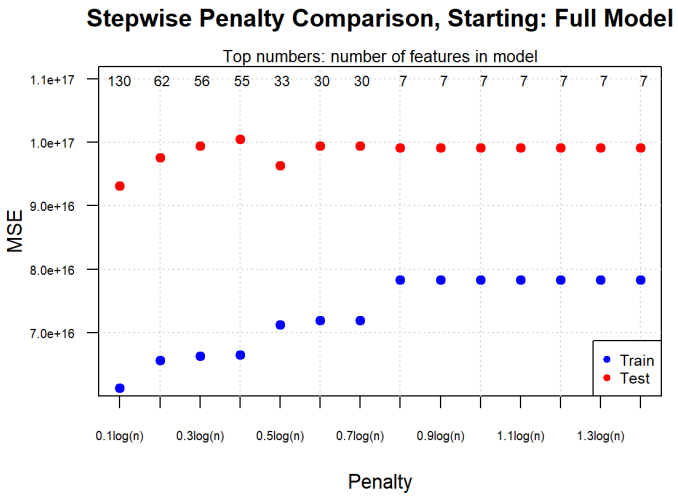
1. מודל ריק, ללא מסבירים
2. מודל מלא, עם מסבירים רציפים, משתני דמה ואינטרקציות של הקטגוריאליים
3. מודל "בינוני" בו רק המסבירים הרציפים

תוצאות:

ראשית, הרצנו את הניסוי כשהמודל ההתחלתי הוא מודל ריק, והוצאנו את התרשים הבא:



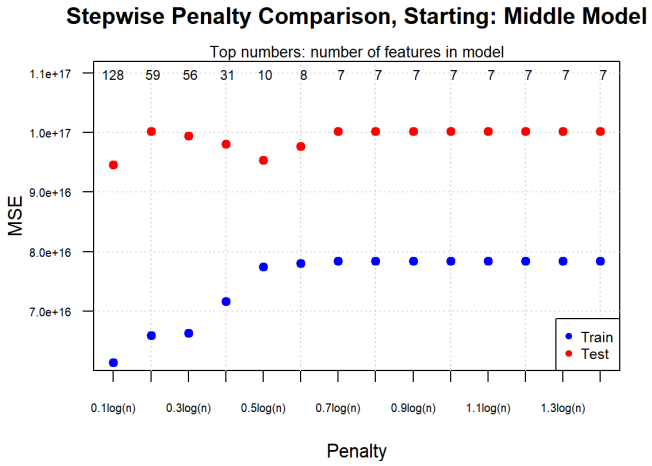
בתרשים ניתן לראות את מספר הפרמטרים בכל מודל (המספרים בחלק העליון של הגרף), את ה-MSE של סט האימון בכחול ושל סט הבחינה באדום. ראשית, כצפוי, ה-MSE של האימון נמוך משל הבחינה בכל גודל קנס. אנחנו לא מופתעים לראות את זה כי סביר שהמודל מתפקד טוב יותר על הנתונים אותם למד.
לא נראית מגמה ברורה בהפרשים בין הבחינה לאימון, מלבד בערך הקנס הנמוך ביותר שהביאו לריבוי משתנים (67) - ולמודל שבבירור נכנס למצב של התאמת יתר, כך שביצועיו על סט האימון השתפרו על חשבון הבחינה.
הרצנו את אותה הבדיקה, כשהפעם נקודת ההתחלה הייתה מודל מלא, וקיבלנו את התוצאה הבאה:



במקרה הזה ניתן לראות מספר הבדלים לעומת התרשים הקודם:

1. האלגוריתם מתכנס למספר מסבירים גדול יותר כשהקנס קטן (132,62,56), אך מגיע לאותו מספר מסבירים (7) כשהקנס גדול, בלי תלות בנקודת ההתחלה. משום שהאלגוריתם חמדני ולא אופטימלי הוא עלול להתכנס לנקודות מקסימום מקומי שקרובה לנקודת ההתחלה שלו.
2. ככל שהקנס גדל ומופחתת כמות המסבירים - נפגעים ביצועי המודל על סט האימון (MSE גדול). כשהמסבירים מתמעטים נראית יציאה ממצב של התאמה גבוהה לסט האימון. לעומת זאת, סט הבחינה כמעט ולא מושפע מהשינוי במספר המסבירים.

לבסוף, בחנו זאת שוב כשהמודל ההתחלתי הוא מודל בו יש רק מסבירים רציפים, כנקודת "אמצע" בין המודל המלא לריק.



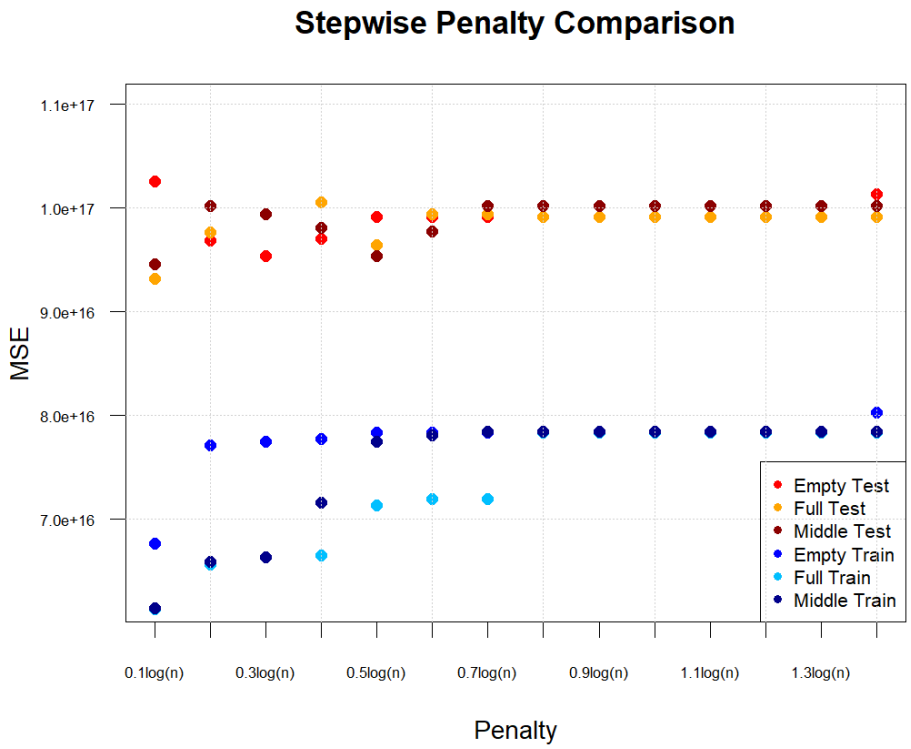
נוסף על התנהגות דומה של ה-MSE ביחס לגודל הקנס, ראינו שגם כאן כשהקנס גדול האלגוריתם מתכנס לאותם שבעה משתנים מסבירים בעלי התרומה הגדולה ביותר למודל.
החלטנו לבדוק מה הם אותם מסבירים:

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -5.609e+09  1.944e+09  -2.884 0.004021 **
artist_count          -3.616e+07  1.102e+07  -3.281 0.001078 **
released_year          2.941e+06  9.648e+05   3.048 0.002372 **
in_spotify_playlists   3.613e+04  1.888e+03  19.132  < 2e-16 ***
in_spotify_charts      5.023e+06  6.575e+05   7.640  5.9e-14 ***
in_apple_playlists     2.615e+06  1.617e+05  16.169  < 2e-16 ***
in_shazam_charts      -6.434e+05  7.645e+04  -8.416  < 2e-16 ***
energy_.              -2.283e+06  6.022e+05  -3.791 0.000161 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 286600000 on 843 degrees of freedom
Multiple R-squared:  0.7423,    Adjusted R-squared:  0.7401
F-statistic: 346.8 on 7 and 843 DF,  p-value: < 2.2e-16
```

בהמשך, נראה שהמשתנים האלה נבחרו גם על פי שיטות נוספות לבחירת משתנים, לצורך משימת סיווג לשתי מחלקות (שנוצרו ממשתנה המטרה הרציף streams).

על מנת להשוות בין נקודות ההתחלה האפשריות, הוצאנו את הגרף שמציג את ביצועי כל המודלים שיצרנו על סטי האימון והבחינה:



בתרשים ניתן לראות איך עבור קנסות קטנים, כשהאלגוריתמים מעדיפים מודלים מרובי משתנים, המודלים הנבחרים הם בעלי ביצועים מגוונים - על פי נקודת ההתחלה. ככל שהקנס גדל, ופחות משתנים יכולים להיכנס למודל - כך נקודת ההתחלה נהיית פחות רלוונטית ושונות הביצועים קטנה.
אפשר לראות גם את הירידה הברורה בביצועי המודלים על סט האימון כאשר הקנסות גדולים, לעומת ביצועים די דומים על סט הבחינה.

מסקנות ודיון:

1. אלגוריתם Stepwise-Regression לבחירת משתנים מתכנס לנקודת מקסימום מקומית, בה מספר המסבירים, כצפוי, מושפע מגודל הקנס ומנקודת ההתחלה.
2. בבעיות עם משתנים רבים, נקודת ההתחלה משמעותית בעיקר כשהקנס קטן (ומאפשר הכנסה של מספר מסבירים גדול). כשמספר המקומות מוגבל יותר - סביר שהאלגוריתם יבחר בסט משתנים אחיד ומובהק, בלי תלות בנקודת ההתחלה.
3. ככל שהמודל עשיר יותר במסבירים, ביצועיו על סט האימון ישתפרו. ככל שיש יותר מסבירים המודל מקבל יותר חופש (המימד של מרחב החיזוי גדל), והשונות המוסברת גדלה או לכל הפחות לא קטנה. עם זאת, על סט הבחינה קשה להבחין במגמה ברורה. יתכן שהסיבה היא שריבוי המשתנים מוביל להתאמת יתר של המודל לסט האימון, ופוגע ביכולת ההכללה של המודל.
4. הביצועים על סט האימון טובים מהביצועים על סט הבחינה בכל המודלים שנבחנו.
5. נקודת ההתחלה וגודל הקנס של האלגוריתם לא השפיעו באופן ברור על יכולות המודל הנבחר על סט הבחינה, לכן אין לנו המלצה חד משמעית לגבי דרך מועדפת להריץ Stepwise-Regression.

ניסוי 2: סיווג

המודלים:

ביצענו השוואה בין ארבע דרכים לבחירת משתנים: Random Forest, XGBoost, מבחן חי בריבוע לאי תלות ומדד Fisher-Information, ובדקנו את השפעתן על יכולות הסיווג של ארבעה מודלים: רגרסיה לוגיסטית, Random Forest, XGBoost ו-Neural Network.

הנחות המודלים:

- הנחת לינאריות עבור ה-Log(odds) **ברגרסיה לוגיסטית**
- התצפיות לא תלויות אחת בשניה

מהלך הניסוי:

המרת משתנה המטרה לקטגוריאלי:

כיוון שהמודלים הנבחרים משמשים לטובת פתרון בעיית סיווג, החלטנו להמיר את משתנה המטרה לקטגוריאלי: אם מספר ההשמעות גדול מערך החציון, ערכו יהיה 1, אחרת 0. שיטה זו הפכה את הבעיה שלנו למאוזנת.

בחירת המשתנים:

בחרנו להשוות בין ארבע שיטות של Feature Selection:

- **Random Forest-** בשלב הראשון אימנו מודל Random Forest בעזרת סט האימון עם כל המשתנים, ולאחר מכן ביצענו חישוב של Feature Importance על בסיסו. Feature Importance מחשב את חשיבות המשתנה המסביר על ידי מדידת הירידה ב-impurity של הצומת (בעזרת מדדים כמו ג'יני או אנטרופיה) שכל משתנה גורם כאשר הוא משמש לפיצול. משתנים שמביאים לירידה הגדולה ביותר ב-impurity נחשבים חשובים יותר. לבסוף, קיבלנו סט משתנים מדורג (נספח 3), ומתוכו בחרנו את 15 המשתנים בעלי הדירוג הגבוה ביותר (נספח 7).

- **XGBoost-** בדומה ל-Random Forest, בשלב הראשון אימנו את המודל, ולאחר מכן ביצענו חישוב של Feature Importance ל-XGBoost. החשיבות של כל משתנה מחושבת על סמך התדירות שבה נעשה שימוש בכל משתנה לפיצול הנתונים על פני כל עצי ההחלטה. משתנים המשמשים בחלק העליון של העצים (ליד השורש) לקבלת החלטות נחשבים חשובים יותר. לבסוף, קיבלנו סט משתנים מדורג (נספח 4), ומתוכו בחרנו את 15 המשתנים הטובים ביותר (נספח 8).
- **Fisher Score-** בשיטה זו מתבצע Feature Importance באמצעות מציאת צירופים לינאריים של משתנים המפרידים בצורה הטובה ביותר בין המחלקות השונות בנתונים. לכל אחד מהמשתנים ניתן ציון פישר כך שככל שהציון גבוה יותר כך המשתנה טוב יותר (נספח 5). גם כאן בחרנו את 15 המשתנים הטובים ביותר (נספח 9).
- **Chi-test-** שיטה סטטיסטית המסייעת לזהות את המשתנים הרלוונטיים ביותר במערך הנתונים על ידי מבחן אי תלות בין כל משתנה מסביר למשתנה היעד (נספח 6). עבור שיטה זו בחרנו את 15 המשתנים המובהקים ביותר (נספח 10).

עבור כל אחת מהשיטות קיבלנו סט משתנים שונה, כאשר היו משתנים משמעותיים שהופיעו בכלל הסטים. למשל, in_spotify_playlists הגיע למקום הראשון/שני בכל המדדים, באופן לא מפתיע- שכן כמות המופעים של השיר בפלייליסטים עשויה לגרום להשמעות מרובות של השיר (בהתאם לקורלציות שמצאנו בחלק הקודם של הדו"ח).

<u>אימון המודלים</u>

לכל סט משתנים ולכל סוג מודל ביצענו Hyperparameter tuning בעזרת grid search של ספריית scikit-learn עם cross validation של 5 חזרות. בסוף נשארנו עם 16 קומבינציות של סוג המודל (עם ההיפר-פרמטרים הטובים ביותר שנבחרו עבורו ב-grid search), ותת סט המשתנים שנבחר בשיטה מסוימת.

<u>Bootstrap</u>

כדי לייצג את יכולות הסיווג של המודלים שלנו באופן יותר אינפורמטיבי מאומד נקודתי (accuracy), השתמשנו בשיטת bootstraping ליצירת רווחי סמך. מדובר בשיטה א-פרמטרית לשערוך התפלגות של סטטיסטי, בה עושים שימוש בדגימה אקראית חוזרת ונשנית עם החזרות, מתוך המדגם הבודד שבידינו. מכל דגימה מחושב הסטטיסטי, ולפי צפיפות הסטטיסטיים כולם ניתן לבנות רווח סמך עם רמת ביטחון. אנחנו לבצע 1000 דגימות bootstrap, ובחרנו ברווח עם רמת ביטחון של 90%.

תוצאות:

תוצאות ה-accuracy שהתקבלו לכל אחד מהמודלים עבור כל שיטת Feature Selection:

- תוצאות ה-Feature Selection עבור רגרסיה לוגיסטית:

	Random Forest	XGBoost	Fisher Score	Chi-Squared
Accuracy	0.835	0.829	**0.846**	0.842
רווחי סמך של Accuracy	0.79 ,0.875	0.785 ,0.87	0.805, 0.885	0.8 ,0.885

- תוצאות ה-Feature Selection עבור XGBoost:

	Random Forest	XGBoost	Fisher Score	Chi-Squared
Accuracy	0.815	**0.834**	0.83	0.811
רווחי סמך של Accuracy	0.77 ,0.86	0.79 ,0.875	0.785 ,0.875	0.765 ,0.855

- תוצאות ה-Feature Selection עבור Random forest:

	Random Forest	XGBoost	Fisher Score	Chi-Squared
Accuracy	0.830355	0.83424	**0.840195**	0.834445
רווחי סמך של Accuracy	0.785,0.875	0.79,0.875	0.795,0.88	0.79,0.88

- תוצאות ה-Feature Selection עבור Neural network:

	Random Forest	XGBoost	Fisher Score	Chi-Squared
Accuracy	**0.846**	0.845	0.805	0.81
רווחי סמך של Accuracy	0.8 ,0.88	0.805 ,0.885	0.755 ,0.85	0.76 ,0.855

מסקנות ודיון

ניתן לראות כי עבור כל מודל, רווחי הסמך של תתי הסט של המשתנים חופפים. כלומר, ברמת ביטחון של 90% לא ניתן לומר שיש הבדל מובהק בין שיטות בחירת המשתנים עבור מודלים שונים.

אם נסתכל על ערכי ה-accuracy הממוצעים, ניתן לראות כי התוצאה הטובה ביותר עבור רגרסיה לוגיסטית התקבלה עבור המשתנים שנבחרו עם מדד פישר, במודל ה-XGBoost עבור המשתנים שנבחרו עם XGBoost, במודל ה-Random Forest עבור המשתנים שנבחרו עם מדד פישר וברשת הנוירונים עבור המשתנים שנבחרו עם Random Forest. התוצאה עבור מודל ה-XGBoost אמנם תואמת את השערת המחקר שלנו, אך ברמת ביטחון של 90% לא ניתן להגיד שלשיטה זו יש עדיפות על פני האחרות.

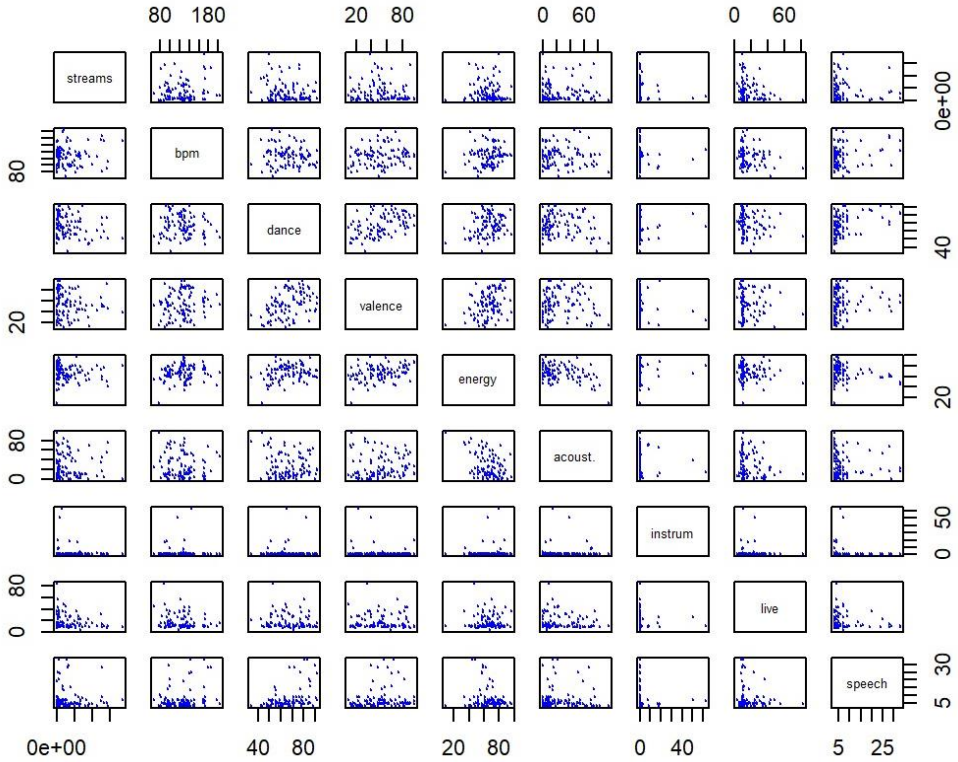
כפי שציינו לאורך הדו"ח, היו מספר משתנים בעלי קורלציה גבוהה יחסית עם משתנה המטרה שגם דורגו במקומות הראשונים בשיטות בחירת המשתנים השונות. יתכן שהם מאפשרים הפרדה טובה בין המחלקות בבעיית הסיווג שבחרנו ועל כן לא ראינו הבדלים משמעותיים בביצועי המודלים שנבחנו.

המסקנה שלנו היא שאין "כלל אצבע" לבחירת משתנים, ועל כן יש לבדוק כל מקרה לגופו בהתאם לסט הנתונים וסוג הבעיה אשר מעוניינים לפתור.

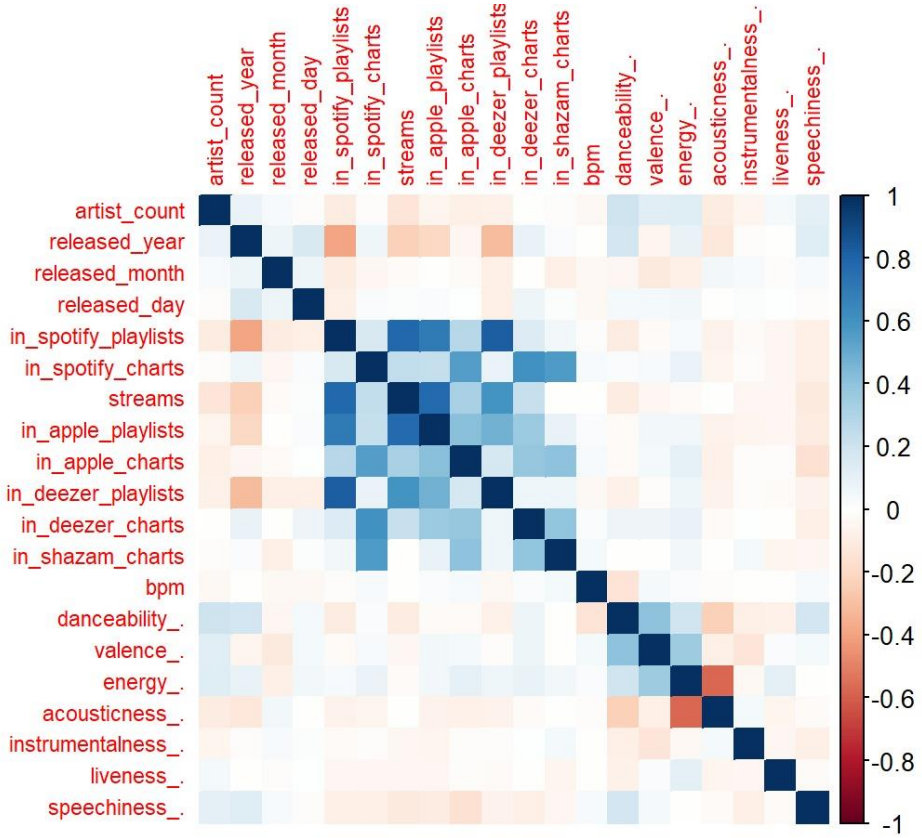
למחקר המשך, היינו מציעים לבדוק ספים שונים של כמות המשתנים הנבחרים בכל שיטה. מפאת חוסר בזמן בדקנו רק את ה-15 הטובים ביותר, אך ייתכן שבחירה של מספר משתנים שונה הייתה יוצרת הפרדה משמעותית יותר, שהייתה עוזרת להכריע בין יעילות השיטות השונות.

נספחים:

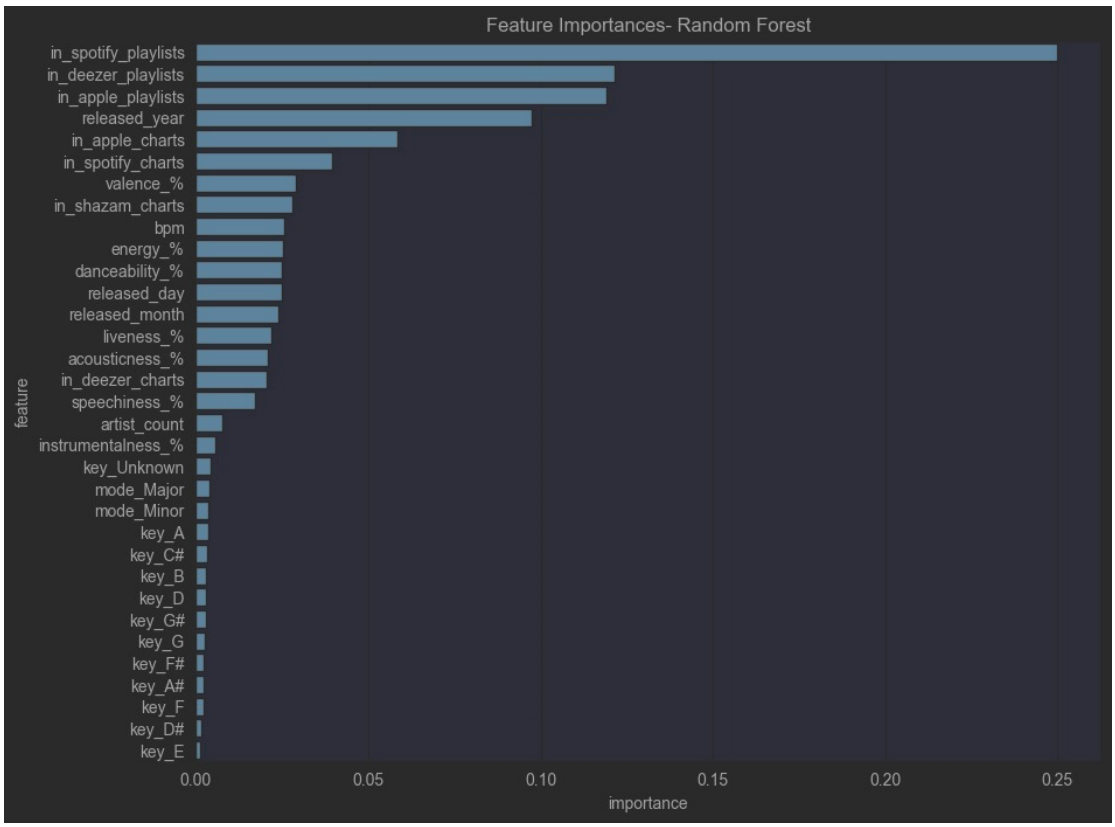
נספח 1: Scatter Plots בין המסבירים הכמותיים



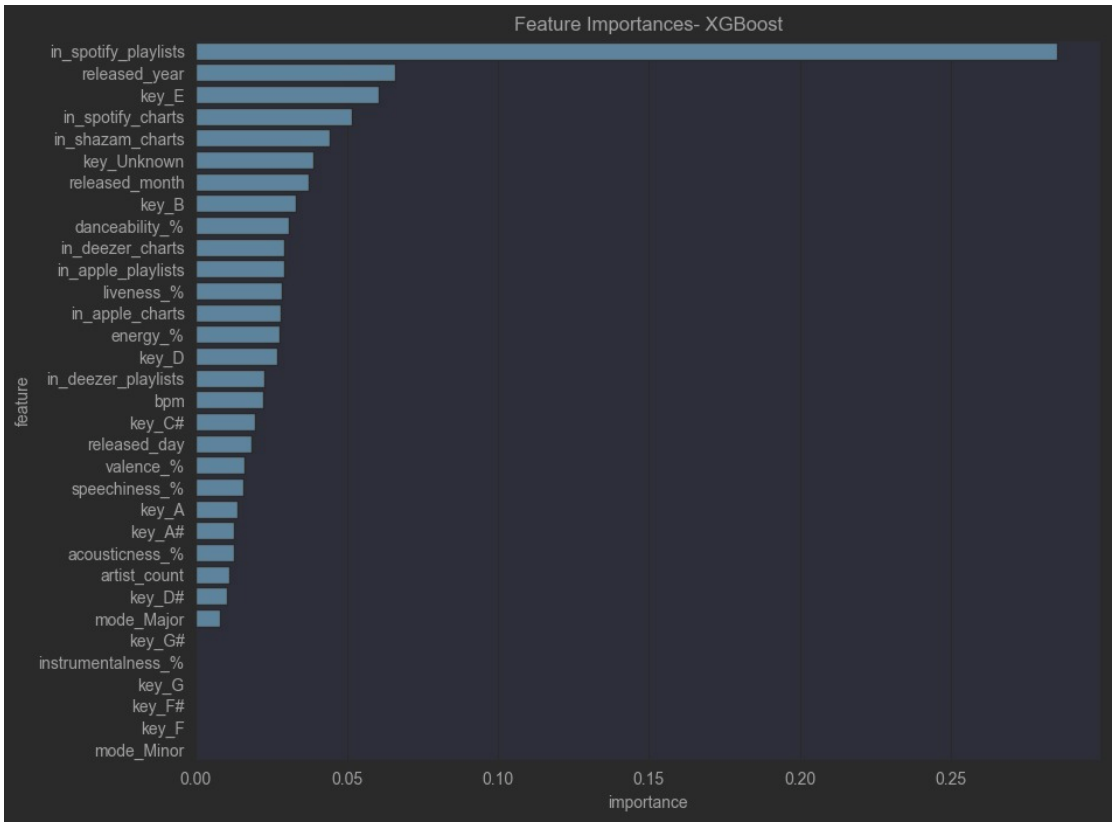
נספח 2: קורלציות בין המסבירים הכמותיים



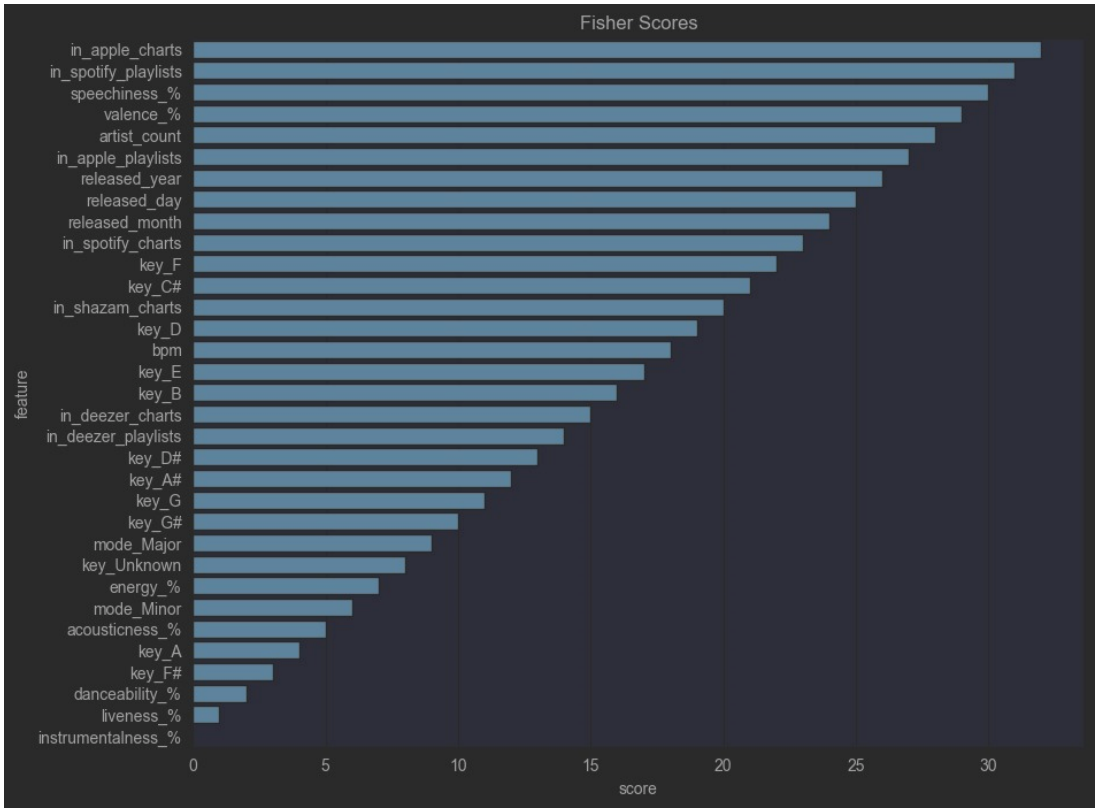
נספח 3: דירוג חשיבות הפיצ'רים עפ"י Random Forest



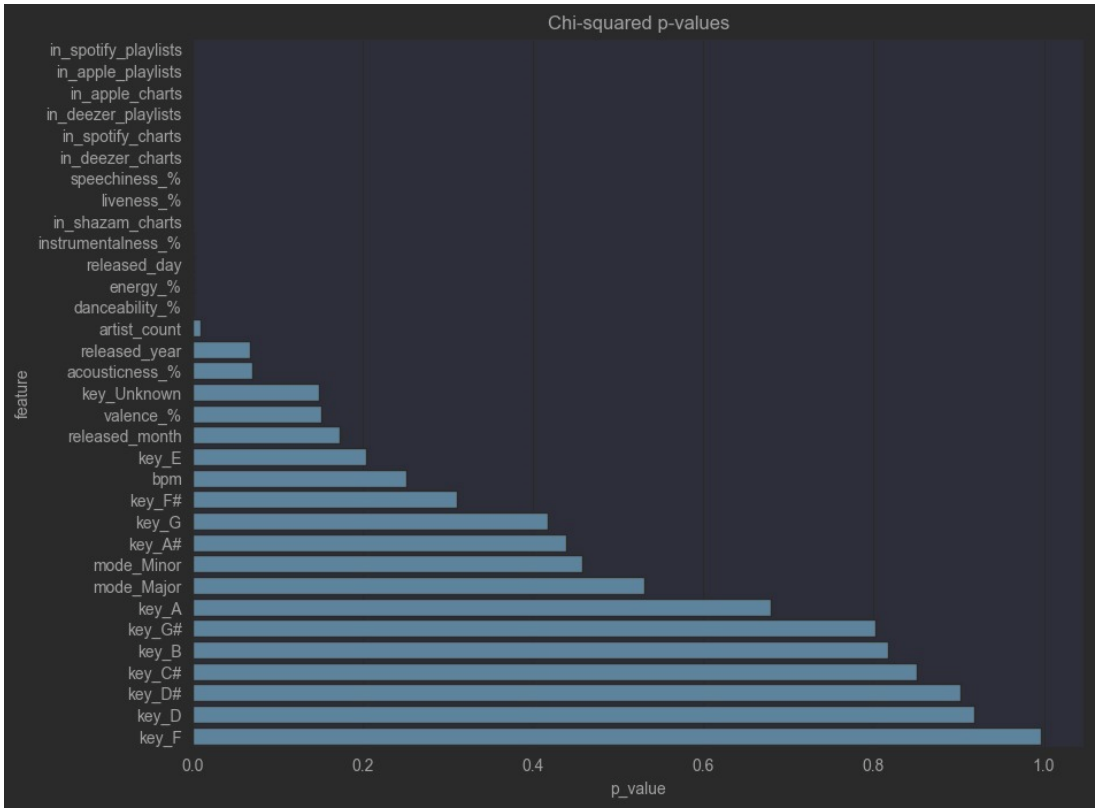
נספח 4: דירוג חשיבות הפיצ'רים עפ"י XGBoost



<u>נספח 5:</u> דירוג חשיבות הפיצ'רים עפ"י Fisher Score



<u>נספח 6:</u> דירוג חשיבות הפיצ'רים עפ"י Chi-Squared



נספח 7: 15 המשתנים שנבחרו ע"פ Random Forest

```
Top 15 features for Random Forest:
4     in_spotify_playlists
8      in_deezer_playlists
6       in_apple_playlists
1            released_year
7          in_apple_charts
5        in_spotify_charts
13               valence_%
10        in_shazam_charts
11                     bpm
14                energy_%
12          danceability_%
3             released_day
2           released_month
17              liveness_%
15          acousticness_%
```

נספח 8: 15 המשתנים שנבחרו ע"פ XGBoost

```
Top 15 features for XGBoost:
4     in_spotify_playlists
1            released_year
25                    key_E
5        in_spotify_charts
10        in_shazam_charts
30             key_Unknown
2           released_month
21                    key_B
12          danceability_%
9         in_deezer_charts
6       in_apple_playlists
17              liveness_%
7          in_apple_charts
14                energy_%
23                    key_D
```

<u>נספח 9:</u> 15 המשתנים שנבחרו ע"פ Fisher Score

```
Top 15 features for Fisher Score:
7          in_apple_charts
4     in_spotify_playlists
18           speechiness_%
13               valence_%
0             artist_count
6       in_apple_playlists
1            released_year
3             released_day
2           released_month
5        in_spotify_charts
26                   key_F
22                  key_C#
10        in_shazam_charts
23                   key_D
11                     bpm
```

<u>נספח 10:</u> 15 המשתנים שנבחרו ע"פ Chi-Squared

```
P-values under 0.05 in Chi-Squared test:
4     in_spotify_playlists
6       in_apple_playlists
7          in_apple_charts
8      in_deezer_playlists
5        in_spotify_charts
9         in_deezer_charts
18           speechiness_%
17              liveness_%
10        in_shazam_charts
16      instrumentalness_%
3             released_day
14               energy_%
12          danceability_%
0             artist_count
```