פרוייקט מסכם נושאים מתקדמים בלמידת מכונה:

מגישים:

אורי ביטון-213868797

דניאל גולדשמיד-209328244

https://github.com/OriBiton/Final-Project-ML2קישור לגיטהאב:

**חשוב-**עבור כתיבת הקודים הטכניים נעזרנו בצ'אט גיפיטי על מנת לייעל את העבודה,את רמת הקוד ואת הזמן. את העבודה ביצענו ביחד תוך ישיבה בזומים ועבודה בשותפות מלאה.

**תקציר:**

את הפרוייקט שלנו ביצענו על דאטה של הגרלת דירות בהנחה.

בדאטה קיים מידע על ההגרלות כגון כמות נרדשמים,כמות זוכים, פרטי הפרוייקט ומבצעי הפרוייקט, זמנים של ביצוע ההגרלה וסיום הרישום, מחירי מטר ממוצע לדירות בהגרלה ועוד.

קישור לדאטה:

<https://data.gov.il/dataset/mechir-lamishtaken/resource/7c8255d0-49ef-49db-8904-4cf917586031>

את הנתונים חילצנו בעזרת האיי פי אי שלהם.

לעבודה היו שני חלקים עיקריים.

1) בניית מודל חיזוי לשאלה הבאה: "האם יישארו דירות לא מאוכלסות לאחר ביצוע של הגרלה"

2) בניית מודל קלאסטרינג לחלוקת ההגלות לקבוצות דומות שיכולות לתת לנו ולבעלי העניין תובנות חשובות בנוגע לסוגי ההגרלות השונים.

**רשימת הדברים שנעשו בפרוייקט:**

**1) הכנת הדאטה-** סידור נתונים שנכתבו בצורה שגוייה כגון "-" במקום ערך חסר או פסיקים בנתונים מספריים וכו..

מילוי ערכים חסרים

זיהוי חריגים באמצעות שיטות שנלמדו בקורס ונוסף שיטות שקיימות היום והסרתם מהדאטה.

**2) פיצ'ר אנג'ניר:**

הסרת עמודות שלא רלוונטיות להמשך הפרוייקט

הוספת עמודות חדשות שיעזרו לנו לשלבי הפרוייקט השונים כגון עמודת המטרה שלנו למודל הסופרוויזד (האם יישארו דירות ללא זוכים לאחר הגרלה)

או עמודות שיעזרו לנו להציג תובנות חשובות על הדאטה.

**3) וויזואליזציה:**

הצגת גרפים וטבלאות משמעותיות גם לבדיקה הסטטיסטית של התפלגות העמודות וגם גרפים עבור תובנות חשובות שנמצאו במהלך העבודה

**4) בניית מודל לחיזוי עמודת המטרה:**

בחלק הזה של הפרוייקט רצינו לחזות האם לאחר הגרלה יישארו דירות פנויות שלא זכו בהן.

חיזוי זה הוא חשוב כי הוא יכול לתת לבעלי העניין אינדיקציה להאם יישארו דירות לא מאוכלסות לאחר הגרלה או לא, ואיזה פיצ'רים הכי משפיעים על התוצאה. לשם כך גם נעזרנו בכלי שאפ כדי להציג את חשיבות הפיצ'רים על החיזוי.

השתמשנו ברשימת מודלים של קלאסיפיקציה ולאחר מכן השוואנו ביניהם ובחרנו במודל הטוב ביותר על סמך הדיוק שלו להמשך התהליך של בדיקת חשיבות הפיצ'רים.

**5) קלאסטרינג:**

בחלק הזה של הפרוייקט רצינו לזהות קבוצות הגרלות שדומות במאפיינים שלהם ולסווגן.

לאחר מכן בחנו אילו פיצ'רים הכי משפיעים ועליהם הרצנו כמה בדיקות וויזואליות סטטיסטיות כדי לראות את ההבדל בין הקבוצות השונות.

גם פה השוואנו בין ארבעה אלוריתמים שונים של קלאסטרינג ובחרנו את האלגוריתם המתאים ביותר לנתונים עבור המשך התהליך.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**הכנת הדאטה:**

**מילוי ערכים חסרים:**

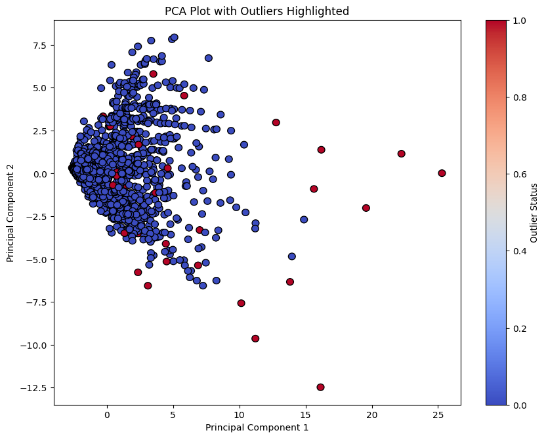
ראשית ראינו כי קיימות 2 עמודות בעלות יותר מ50% ערכים חריגים. את העמודות האלה בחרנו להוציא מהדאטה

לאחר מכן נותרנו עם 5 עמודות בעלות ערכים חסרים כאשר 3 מהן היו עמודות מספריות ו2 היו קטגוריאליות.

עבור העמודות המספריות בחרנו למלא עם החציון של העמודה משום שלא רצינו לתת חשיבות להטיית הממוצע שנגרם בגלל הערכים חריגים בעמודות הללו.

עבור העמודות הקטגוריאליות בחרנו למלא עם הערך השכיח בעמודה. שיטה זאת מאוד נפוצה ואינטואיטיבית ולכן בחרנו לבצע אותה.

**זיהוי חריגים והוצאתם מהדאטה:**

חלק זה חולק ל2:

1) זיהוי שורות חריגות בדאטה- בחלק הזה השתמשנו ב3 מודלים מוכרים:

Isolation Forest, OneClassSVM, LOF

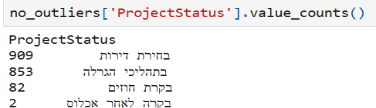
ייצגנו את הדאטה באמצעות פיסיאיי בעל 2 מימדים ואת השורות החריגות בצבע אדום כפי שניתן לראות בגרף.

השורות החריגות אלו שורות שהוסכמו על ידי שלושת המודלים השונים כחריגות (על מנת לוודא שאלו באמת שורות חריגות)  
לאחר מכן הסרנו את השורות הללו מהדאטה.

2)זיהוי של ערכים חריגים בכל עמודה ועמודה.

עבור העמודות המספריות השתמשנו באלגוריתם של איזולשן פורסט. שמנו לו בפרמטר (0.001) על מנת שיזהה רק ערכים באמת חריגים מאוד בכל אחת מהעמודות

עבור העמודות הקטגוריאליות ביצענו בדיקה וויזואלית כזאת:

וכאשר ראינו עככים מאוד חריגים כמו בדוגמא הנל החלטנו להסיר את השורות שמכילות את הערך הזה מהדאטה

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**וויזואליזציה:**

לאורך הפרוייקט היו כמה ויזואליזציות שקשורות לשלבים שונים בפרוייקט.  
נתמקד ב2 סוגים עיקריים:

**1)ויזואליזציות של התפלגות הנתונים:**

לחלק של הערכים החריגים והנתונים החסרים בצגנו ויזואליזציות על מנתלראות את התפלגות הנתונים.  
הוויזואליזציות עזרו לנו להבין איך מפולגות העמודות בעלות הערכים החסרים, ואיך מפולגות העמודות על מנת לזהות חריגים.

A graph with a bar

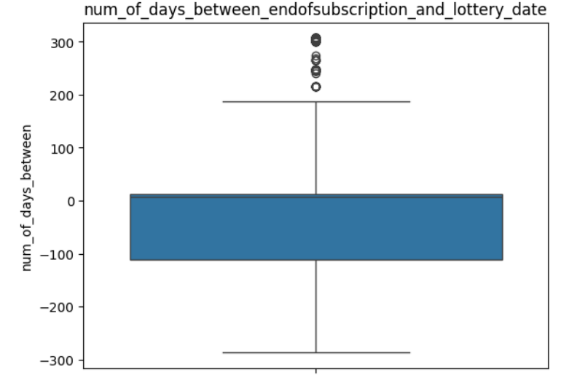
Description automatically generatedA graph of a number of winners

Description automatically generatedהנה כמה דוגמאות

A screenshot of a computer

Description automatically generated

אלו ויזואליזציות כלליות שעוזרות לנו להבין את הנתונים בצורה יותר טובה ולזהות דברים חריגים בנתונים (כמו למשל בטבלה השמאלית למטה שמנו לב שיש עמודות שאמורות להיותמספריות והגיעו כאובייקט וזה עזר לנו לזהות למה.) אלו רק כמה דוגמאות בודדות אך כמובן שבפרוייקט קיימות עוד ויזואליזציות נוספות.

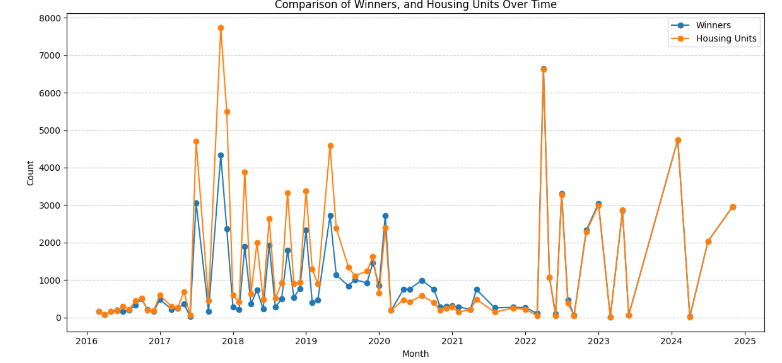
**2)וויזואליזציות מעניינות שחילצנו בחלק של הפיצ'ר אנג'ניר**

גרף זה מציג את הזמן שעבר בין סיום ההרשמה להגרלה לבין הביצוע שלה: **ניתן לראות שיותר מחצי מההגרלות בוצעו לפני שהסתיימה ההגרלה-** זאת תוצאה חריגה שיש לבדוק אותה.

A screenshot of a graph

Description automatically generated

גרפים אלו מציג את היחס בין כמות הנרשמים להגרלה לבין כמות הדירות הקיימות בהגרלה- בגרף השמאלי ניתן לראות את ההגרלות שכמות הנרשמים בהן הייתה קטנה או שווה לכמות הדירות ומימין ניתן לראות את המקרים שבהם היה "אובר הרשמה". ניתן ללמוד מכאן שברוב ההגרלות הביקוש היה גבוה בהרבה מההיצע מה שאולי יכול להסביר את העובדה שחלק מההגרלות נסגרו מוקדם מהצפוי כפי שראינו בגרף הקודם.



גרף זה מייצג לנו את ההפרש בין כמות הדירות בהגרלה לכמות הזוכים לאורך הזמן. הגרף הזה ראשית מראה לנו שבעבר היו פערים בין השניים ושברוב ההגרלות היו נותרות דירות ריקות לאחר ההגרלה לעומת היום שהפער הזה הצטמצם. הגרף הזה יכול לתת לנו אינדיקציה טובה לכך שהשנה מאוד משפיעה על האם יישארו דירות ריקות לאחר ההגרלה או לא (משתנה המטרה בחלק של החיזוי)

**פיצ'ר אנג'ניר:**

בחלק הזה הסרנו עמודות שלא רלוונטיות עבור בניית המודל כגון כמות המנצחים הכללית שבעצם מורכבת מחיבור של מנצחים מכל סוג או כמות הנרשמים שגם מורכבת מכמות הנרשמים מכל סוג  
כמו כן הסרנו עמודות כמו מספרים מזהים של ההגרלה וכו..

בחלק הזה גם יצרנו עמודות חדשות שמבטאות קומבינציה של כמה עמודות שונות בדאטה כגון ההפרש בין זמן ביצוע ההגרלה לזמן סיום ההרשמה (מה שמופיע בויזואליזציה לעיל),  
היחס בין כמות הנרשמים לכמות הדירות בהגרלה (מה שמופיע בויזואליזציה לעיל),

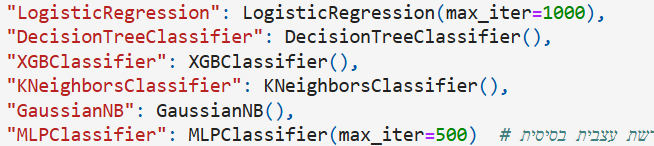
האם יישארו דירות פנויות לאחר ביצוע ההגרלה (עמודת המטרה) שנלקחה מהעמודות של כמות הבתים בהגרלה וכמות הזוכים ועוד...

**בניית מודל לחיזוי עמודת המטרה:**

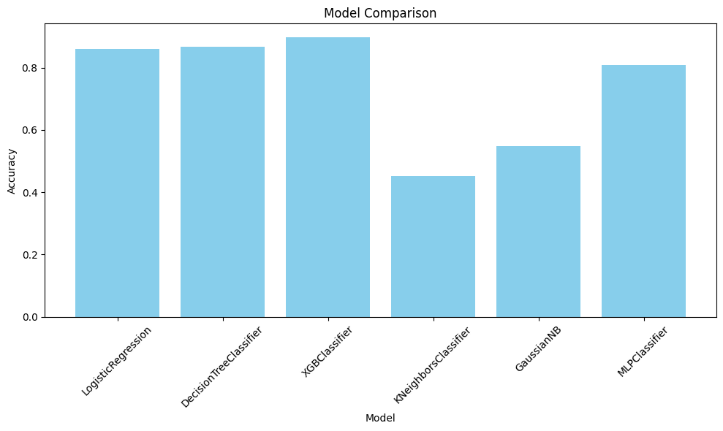
בחלק הזה של הפרוייקט רצינו לחזות האם לאחר הגרלה יישארו דירות פנויות שלא זכו בהן.

ראשית הסרנו את עמודות המנצחים מכיוון שזה לא משהו שיועיל לנו לחיזוי של הגרלות בעתיד משום שזה דבר שקורה רק לאחר ביצוע ההגרלה...

לאחר מכן קודדנו עמודות קטגוריאליות באמצעות גט דאמיז ונירמלנו את הנתונים .

רשימת המודלים שנבחנו היא:

אלו מודלים נפוצים היום בעולם של למידת מכונה וחוזים משתנים קטגורייאלים כמו עמודת המטרה.

לאחר מכן השוואנו בין המודלים על ידי בדיקת הדיוק שלהם:

XGBClassifier ניתן לראות מכאן שהמודל

הכי מתאים לנתונים שלנו ומספק את הדיוק הגבוה ביותר:

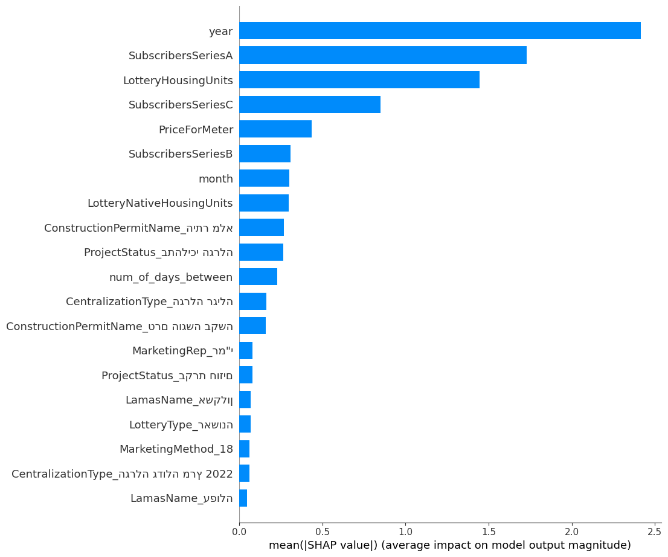


**A graph of data being displayed

Description automatically generated with medium confidenceלאחר מכן השתמשנו בכלי שאפ על מנת לראות את חשיבות הפיצ'רים על המודל:**

הגרף מציג את מידת ההשפעה של כל פיצר על ערך החיזוי ראשית כל שורה מדברת על פיצר מסויים כל נקודה מדברת על דגימה מסויימת צבע הנקודה מצביע על ערך הפיצר (לדוגמא במשתנים קטגוריאלים אדום זה 1 וכחול זה 0) המיקום על ציר האיקס מצביע על כמה זה תרם לגדילה (בכיוון ימין) או לדעיכה (בכיוון שמאל) של ערך החיזוי. לדוגמא אפשר לראות מהגרף שהשנה משפיעה ביחס הפוך על הסיכוי שישארו דירות ריקות-זאת אומרת שככל שערך הפיצר גדול יותר (אדום) כך הסיכוי שישארו דירות ריקות קטן יותר.

לדוגמא הפיצ'ר המשפיע ביותר הוא השנה-(קיבלנו רמז בגרף של הויזואליזציה) אפשר לראות שככל שהשנה יותר נמוכה (כחול) הסיכוי שישארו דירות פנויות גדל (מושך ימינה)

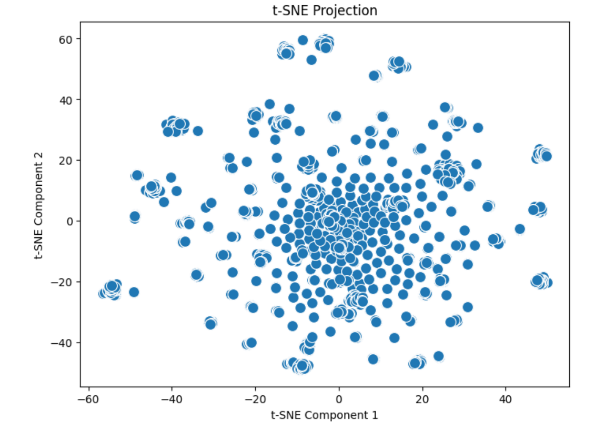


גרף שמראה את הפיצ'רים הכי משפיעים בממוצע

מכאן רואים למשל שהשנה הכי משפיעה על החיזוי .

**קלאסטרינג:**

לצורך הקלאסטרינג יצרנו עוד עמודה נוספת של עדכניות ההגרלה (תאריך ביצוע ההגרלה האחרונה פחות תאריך ביצוע ההגרלה)  
והורדנו עמודות לא רלוונטיות כמו תאריך ביצוע ההגרלה,תאריך סיום הרשמה (עמודות שהמרנו אותן לעמודה אחת של כמה זמן עבר בין סיום ההרשמה לביצוע ההגרלה בחלק של הפיצ'ר אנג'ניר), מזהה פרוייקט והגרלה,שם הקבלן וכו...

****ביצענו קידוד לעמודות הקטגוריאליות באמצעות גט דאמיז ונירמלנו את הנתונים  
  
לראות את הדאטה בצורה וויזואלית: T-SNEהצגנו גרף

אפשר לראות צפיפות של נתונים במרכז

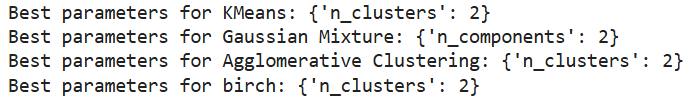
לאחר מכן עברנו לחלק של הקלאסטרינג.

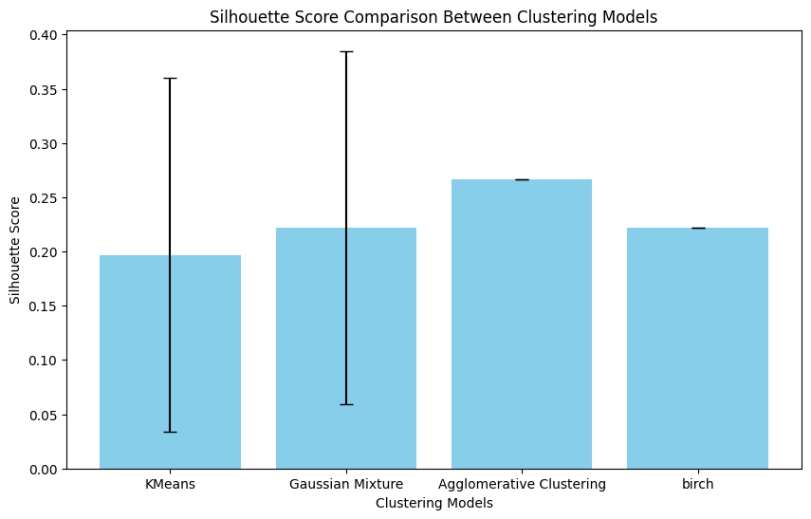
השיטות שנבדקו הן:

K-Means,GaussianMixture,AgglomerativeClustering,Birch

לכל אחת מהשיטות הרצנו מבחן שילוט על מנת למצוא את הפרמטרים האופטימליים עבור הנתונים שלנו.

להלן תוצאות ההשוואה בין המודלים:

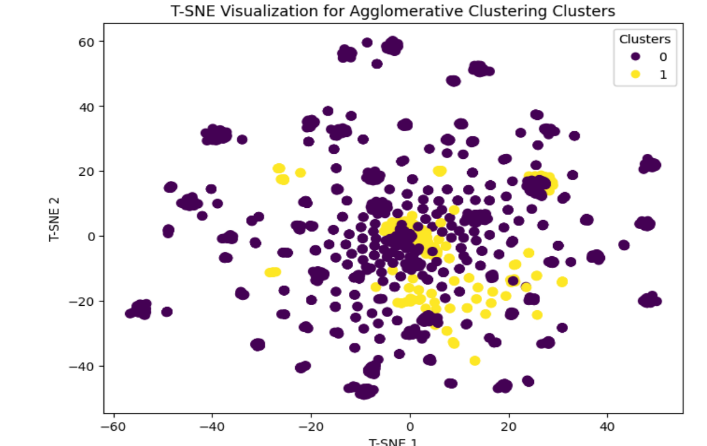




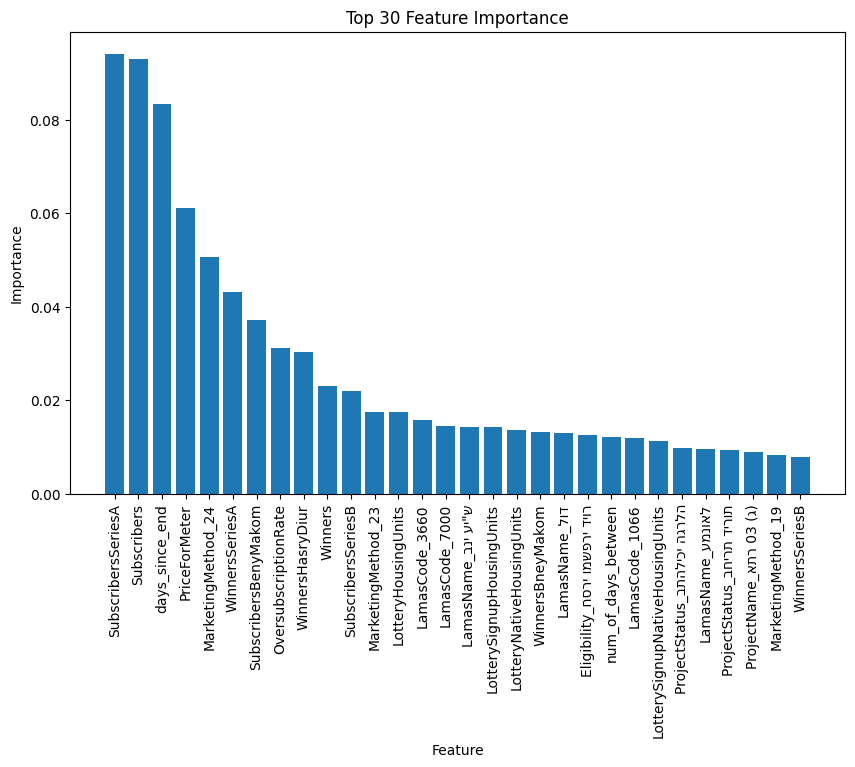
הרצנו כל אחד מהמודלים 10 פעמים על מנת למצוא את התוצאה הממוצעת שתהיה כמה שיותר מהימנת כדי לבחור את המודל הטוב ביותר לנתונים.

המודל טוב ביותר:



בנינו גרף שיראה את הקלאסטרים על הנתונים:

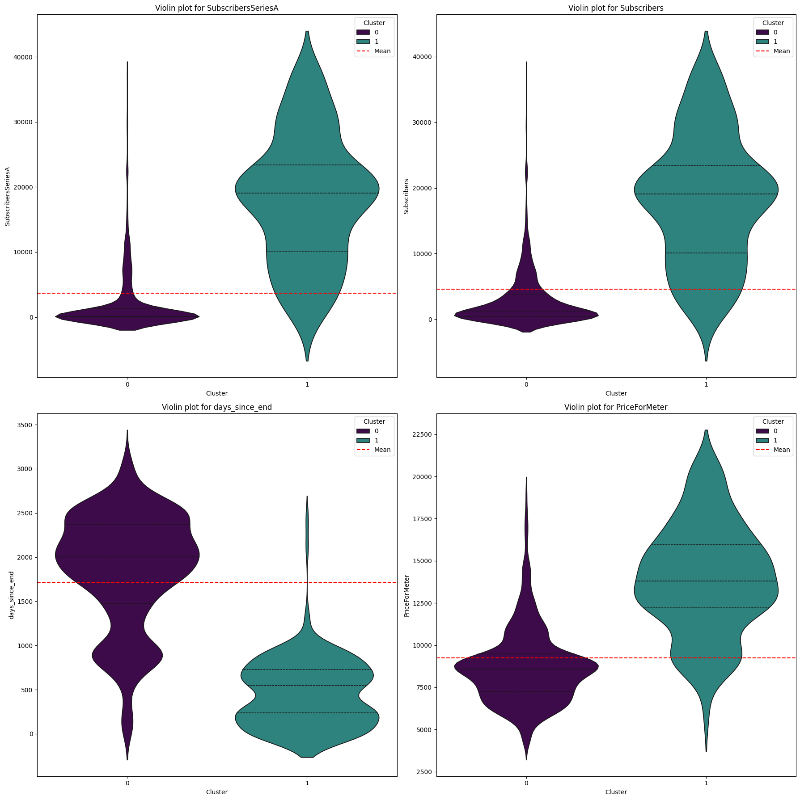
אפשר לראות שקלאסטר 0 יותר מפוזר ואילו קלאסטר 1 יותר מרוכז סביב המרכז



לאחר מכן רצינו לחפש תובנות על הקלאסטרים. לשם כך בשלב ראשון רצינו לראות מה הפיצ'רים שהכי משפיעים על הקלאסטרינג. לשם כך הרצנו מודל רנדום פורסט על הקלאסטרים כדי למצוא את החשיבות של הפיצ'רים, להלן התוצאות:

בחרנו לבצע ניתוח סטטיסטי על הקלאסטרים לפי ארבעת הפיצ'רים הכי משפיעים שהם כמות הנרשמים מסוג איי,כמות הנרשמים בסך הכל,מידת עדכניות ההגרלה והמחיר למטר

להלן התוצאות:



**להלן כמה תובנות ומסקנות:**

**קלאסטר 0: מדובר בקלאסטר של הגרלות פחות עדכניות אשר כמות הרשומים אליהן הייתה נמוכה ביחס להגרלות היותר עדכניות. הגרלות אלה מאופיינות גם במחירים זולים משמעותית ביחס להגרלות של הקלאסטר השני**

**קלאסטר 1: מדובר בהגרלות יחסית עדכניות שבוצעו אשר מחירם גבוה ביחס לקלאסטר 0 וכמות הנרשמים גבוהה משמעותית מהקלאסטר השני**

**קלאסטר 0:הגרלות ישנות וזולות שכמות הנרשמים אליה קטנה**

**קלאסטר 1: הגרלות עדכניות ויקרות שכמות הנרשמים אליהן גבוהה**

בוצע גם מבחן טי טסט על הפיצ'רים הכי משפיעים ויצא שאכן יש הבדל מובהק בממוצעים בין הקלאסטרים.