מבוא למערכות לומדות - דו"ח, תרגיל בית 3

מגישים: עומר שמחי - 316572593, חן פרץ - 204219638

30 ביוני 2021

1. חקר והכנת נתונים

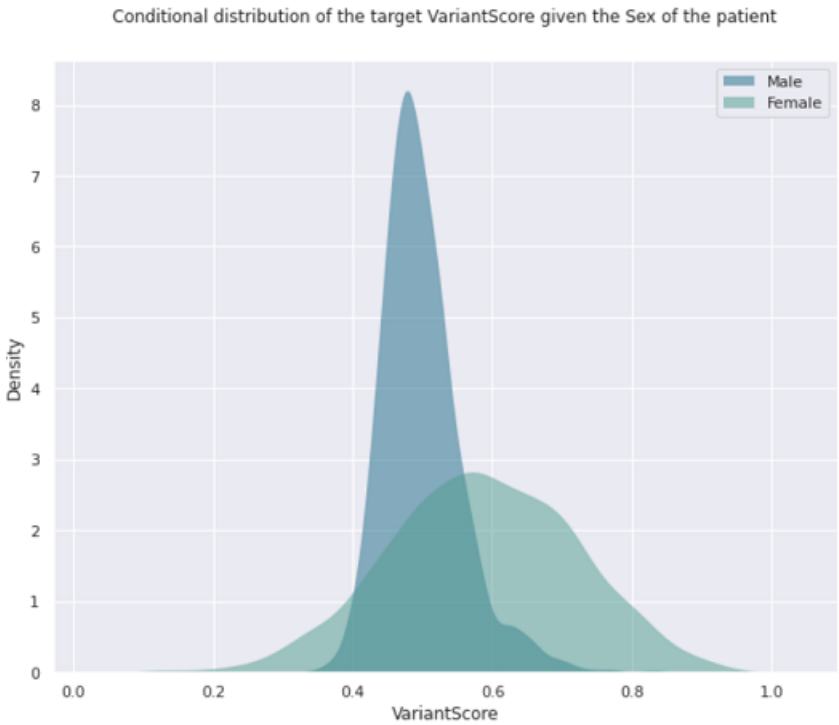
פיצול הנתונים

1. בקוד בלבד, סכום הספרות האחרונות של תעודות הזהות שלנו הוא $8+3=11$ אשר יהיה בשימוש בהמשך.

לפני העיבוד המקדים

2. להלן גרף של צפיפות ההסתברות של $VariantScore$ בהינתן ה-Sex של הנבדק:

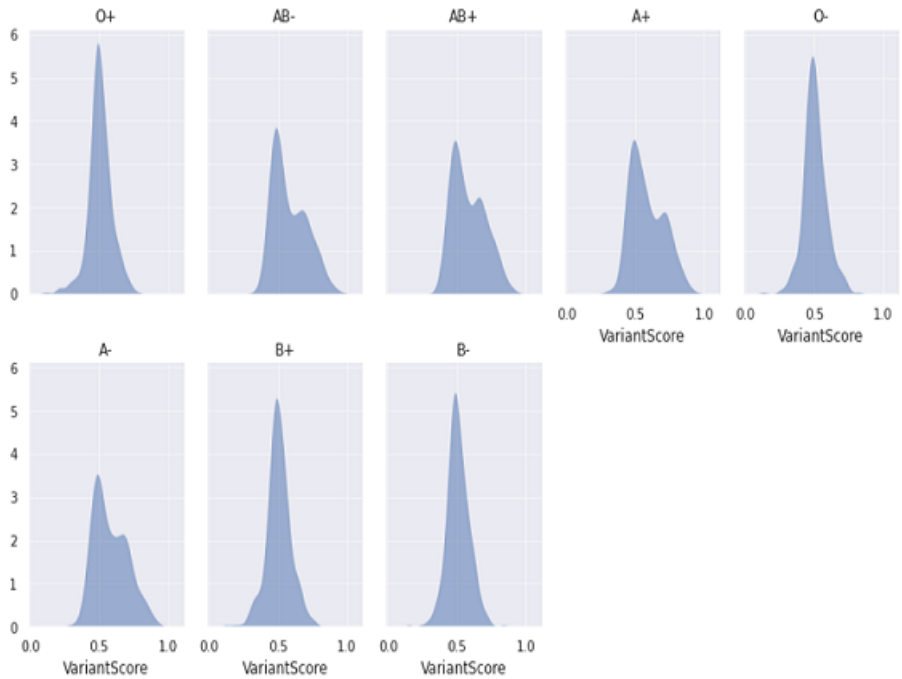
איור 1: צפיפות ההסתברות של $VariantScore$ בהינתן Sex



ניתן לזהות דפוס דומה בהתפלגות לפי שני המגדרים - שניהם תואמים באופן יחסית טוב את הקרנל הגאוסיאני (כאן ב-$KDEplot$ נתנו לו את הקרנל הגאוסיאני הדיפולטי) - "הפעמון" חלק יחסית. עוד אבחנות לגבי הגרף המתקבל - **השונות** בהתפלגות גבוהה יותר מאשר אצל הנשים, דבר שמתבטא בכך שהתיוגים של הדטא של הזכר מתפרסים בתחום צר יותר של ה-$VariantScore$ (ב-$[0.4, 0.6]$ רוב הערכים) ואילו אצל הנשים בתחום רחב יותר (ב-$[0.3, 0.85]$ רוב הערכים) בניגוד לכך, **התוחלת** של הגברים והנשים דומה (ציר הסמטרייה של "הפעמון") - אצל הגברים בערך 0.5 ואצל הנשים בערך 0.55.

3. להלן גרף של צפיפות ההסתברות של $VariantScore$ בהינתן ה-$BloodType$ של הנבדק:

איור 2: צפיפות ההסתברות של $VariantScore$ בהינתן $BloodType$



ניתן לזהות דפוס לפי קבוצות סוגי הדם - עבור $AB-, A-, A+, AB+$ קיבלנו צפיפות כמעט זהה (מעין "דבשת כפולה" רחבה יחסית - שונות גדולה יחסית) ואילו לקבוצות הדם $O+, O-$ ו־$B+, B-$ ניתן לזהות "פעמון" גאוסיאני חלק יחסית וצר (שונות גבוהה יותר באופן ניכר). כמו כן, עבור **כל הגרפים** ניתן לראות שהתוחלת היא בערך $VariantScore \approx 0.5$.

עיבוד מקדים

4. את ייצוג הפיצ'ר Sex הפכנו מייצוג **בינארי** (ובפרט נומרי) מייצוג **קטגורי.** כלומר, נתנו את הספרה 1 עבור נשים ו־0 עבור גברים. בחרנו בייצוג הנ"ל היות שרואים שוני יחסית משמעותי (כפי שתואר בשאלה 2) בין פונקציית צפיפות ההסתברות של שני המגדרים השונים. כלומר, הם **שונים מאופן מהותי** מבחינת התיוגים שניתנו להם בגלל מגדרם (ולא רק מעצם השוני במגדר). לכן, החלטנו שיהיה מתאים לתת ייצוג שמשמר את השוני ביניהם (הרי M ו־F שקול ל־0 ו־1 מבחינת ייצוג השוני בין הפיצ'רים).

5. את ייצוג הפיצ'ר של $bloodType$ החלטנו לשנות בהתאם לפונקציות הצפיפות ביחס למשתנה המטרה $VariantScore$. בשאלה 3 שנענתה קודם לכן, שמנו לב לחלוקה של למעשה שתי קבוצות של סוגי דם המתנהגים באופן כמעט זהה ביחס ל־$VariantScore$ ($AB-, A-, A+, AB+$ בקבוצה אחת ו־$O+, O, B+, B-$ השניה) כאשר מדברים על צפיפות יחסית. על כן, נתנו ייצוג **בינארי** (ובפרט נומרי) אשר יחליף את סוגי הדף בקבוצה הראשונה לספרה 0 ואילו את סוגי הדם בקבוצה השני לספרה 1. כמו קודם,

היות שהעניין שלנו בסופו של דבר הוא ההשפעה על התיוג ביחס למשתנה המטרה, החלטנו על ייצוג שמחדד את השוני בין הקבוצות ולא בשוני בין סוגי הדם.

6. כאשר קיבלנו לידינו את הדטא בתרגיל הנוכחי, העברנו אותו בכמה תחנות על מנת להכינו לאימון ופרדיקציה בסעיפים הבאים. שלבי ההכנה היו שלבים שנקטנו בשתי העבודות הקודמות ויפורטו להלן:

- השלמנו את כל הערכים החסרים (ערכי NaN) עבור כל אחד מהפיצ'רים בטבלה - ההשלמה של הערכים החסרים התבצעה ע"י השלמה בהתאם להתפלגות של הערכים **הקיימים.** זאת כמובן, רק אם חסרים ערכים.
- לאחר מכן, ביצענו את שינוי הייצוג של $Sex, bloodType$ כפי שפורט הסעיפים הקודמים.
- לאחר שכל הערכים הושלמו והמשתנים הקטגוריים שינו את ייצוגם לנומרי, הצגנו $plot$ המציג טבלה של הקורלציה בין כל שניים מהפיצ'רים. עבור זוג פיצ'רים שהראו קורלציה של $\approx 95\%$ ומעלה, הורדנו את אחד מהם בצורה שרירותית (כיוון שהוא לא יתרום מידע נוסף ולכן אין צורך בשניהם).
- לאחר מכן, על מנת להוריד את השפעת ה־$outliers$, נרמלנו את כל הדטא לפי נרמול $zscore$. הנרמול מצמצם את ההשפעה של ה־$outliers$ בכל שמצמצם את המרחק בין נקודות המייצגות את הדטא. נעיר כאן שלא ניתן **להוריד** את ה־$outliers$ כי אנו מנסים לדמות סביבת מבחן אמיתית ולכן לא נמחק דוגמא בטסט אבל כן נוכל לנרמל את הדטא ובכך להוריד את ההשפעה של ה־$outliers$ כאמור.

2. הערכה

7. להלן טבלה המציגה את ערכי ה־MSE עבור $train\,set, validation\,set$ כאשר מבצעים פרדיקציה עם $DummyRegressor$ ומעריכים את ה־$preformence$ עם $CV\_evaluation$:

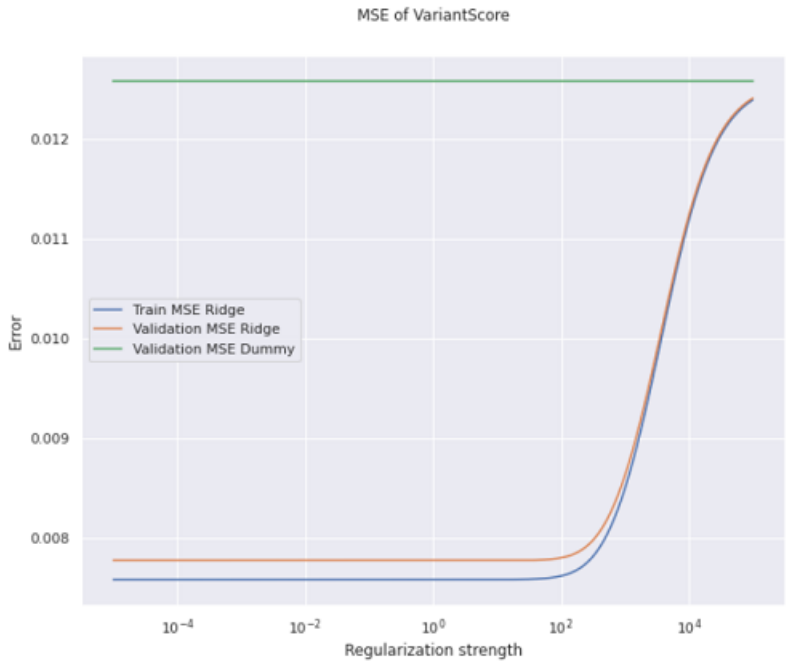
איור 3: הערכת התוצאות של $CV\_evaluation$ עם פרדיקציה לפי ה־$DummyRegressor$

Model	Section	Train MSE	Valid MSE
Dummy	2	0.0125773	0.0125841

3. רגרסור ליניארי בסיסי

8. להלן הגרף המציג את ה־MSE של ה־$VariantScore$ כפונקציה של ה־$regularizatios\_strength$:

איור 4: ה־MSE של ה־$VariantScore$ כפונקציה של $regularizatios\_strength$ עבור $Ridge\,regressor$



9. להלן טבלת השגיאות לפי $train\_set, validation\_set$ של ה־$Ridge\,regressor$ עבור ההייפר־פרמטרים הטובים ביותר:

איור 5: MSE של ה־$train\_set, validation\_set$ לפי ההייפר־פרמטרים הטובים ביותר

Model	Section	Train MSE	Valid MSE	Best hyperparameter
Dummy	2	0.0125773	0.0125841	-
Basic linear	3	0.00758314	0.0077756	15.167168884709241

10. לאחר שאימנו על ה־$train\_set$ כולו, להלן 5 הפיצ'רים בעלי המקדמים הגדולים ביותר (בערך מוחלט, לפי הסדר) עבור הרגרסור המתקבל:

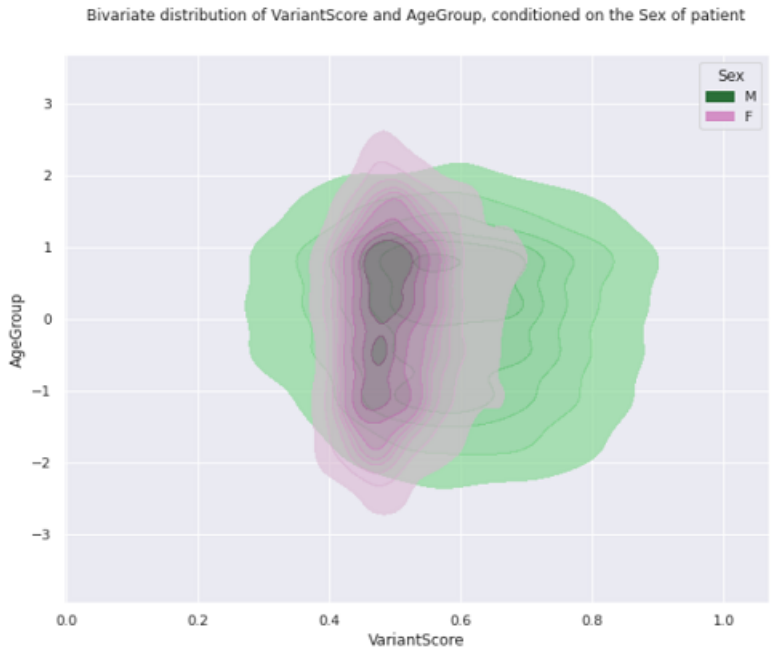
איור 6: 5 הפיצ'רים בעלי המקדמים הגדולים ביותר

	feature	coefficients
22	Sex	0.082456
14	PCR_72	0.037177
3	BloodType	0.032583
9	PCR_19	0.015843
13	PCR_7	0.015122

4. רגרסור ליניארי היררכי

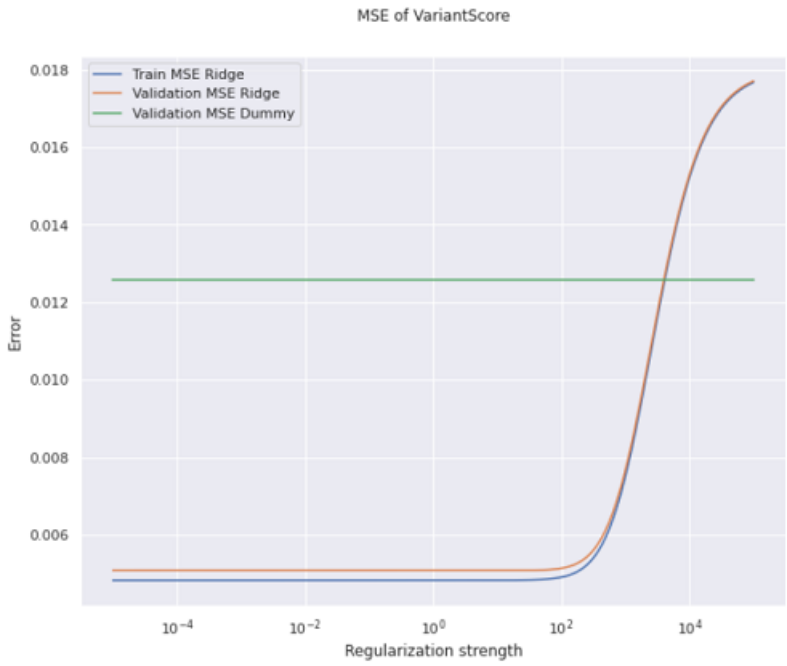
11. להלן גרף המציג את הצפיפות של ההסתברות של $VariantScore, AgeGroup$ כתלות בפיצ'ר Sex לפי שני המגדרים שלו (כזכור - 0 = זכר, 1 = נקבה):

איור 7: $bivariate\,distribution\,of\,VariantScore, AgeGroup\,conditioned\,on\,Sex$

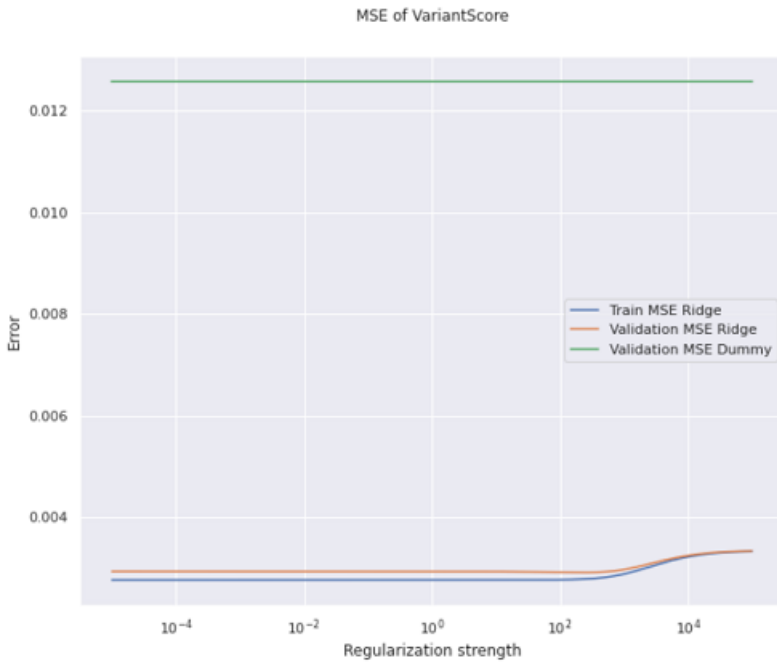


12. נבצע $tuning$ ל־$regularizatios\_strength$ של ה־$Ridge\,regressor$ על שני הסטים של Sex (זכר ונקבה בנפרד). להלן התוצאות:

איור 8: $Male$


MSE of VariantScore
Train MSE Ridge
Validation MSE Ridge
Validation MSE Dummy
0.018
0.016
0.014
0.012
0.010
0.008
0.006
Error
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
Regularization strength

איור 9: $Female$



איור 10: MSE של ה־$train\_set, validation\_set$ לפי ההייפר־פרמטרים הטובים ביותר עבור ה־$Ridge\,regressor$

Subset	Optimal Strength	Optimal Validation MSE
Male	15.1672	0.00508763
Female	217.112	0.00290677

13. נחזור על אותו תהליך עבור ה־$multilevel\,regressor$. להלן התוצאות:

איור 11: MSE של ה־$train\_set, validation\_set$ לפי ההייפר־פרמטרים הטובים ביותר עבור ה־$multilevel\,regressor$

Model	Section	Train MSE	Valid MSE	Best hyperparameter
Dummy	2	0.0125773	0.0125841	-
Basic linear	3	0.00758314	0.0077756	15.167168884709241
Multilevel linear	4	0.00380682	0.00399969	Male: 15.167168884709241, Female: 217.11179456945052

14. קיבלנו שגיאה נמוכה יותר הן עבור $train\_set$ והן עבור $validation\_set$ עבור $multilevel\,regressor$ ביחס ל־$basic\,regressor$ מהחלק הקודם. בשאלה 2, הבחנו בכך שיש **שונות** שונה עבור גברים ונשים. לכן, בניגוד למודל $basic\,regressor$, מודל ה־$multilevel\,regressor$ לומד מודל נפרד עבור גברים ומודל עבור הנשים ואז יודע להתאים בין המודלים כדי לספק פרדיקציה טובה יותר ע"ס הדטא. כלומר היכולת של ה־$multilevel\,regressor$ לעבוד עם שני המגדרים בנפרד (בניגוד ל־$basic\,regressor$ שמתייחס ל־Sex ללא הבדל במגדר) מאפשר לו לבצע פרדיקציה שמתכללת את שני המגדרים בצורה מדוייקת יותר.

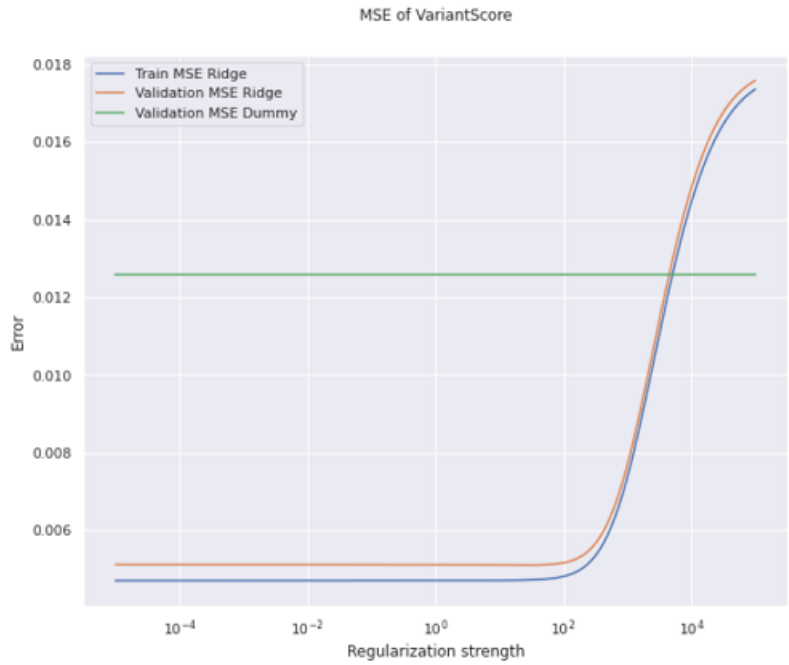
5. אימון לפי רגרסור פולינומי

15. כאשר משתמשים ב־$feature\,mapping$ כפי שתואר בשאלה נצפה לשגיאת אימון **קטנה יותר**. בשאלה 11 ראינו כי לא ניתן להסביר את בין הפיצ'רים למשתנה המטרה ע"י מודל ליניארי ולכן ייתכן כי מודל שאיננו ליניארי, לדוגמא מודל פולינומי יידע להסביר טוב יותר. בנוסף וחשוב מכך, ניתן להסתכל כל מודל ליניארי כאל מקרה פרטי של מודל פולינומי (פשוט ניתן לפיצ'רים "הריבועיים" משקל 0). על כן, נצפה שככזה, נשיג לכל הפחות שגיאה כמו שהשגנו במקרה הליניארי ואכן, התוצאות מראות שהמודל הפולינומי (במקרה שלנו ריבועי) משיג שגיאה קטנה יותר (בהמשך).

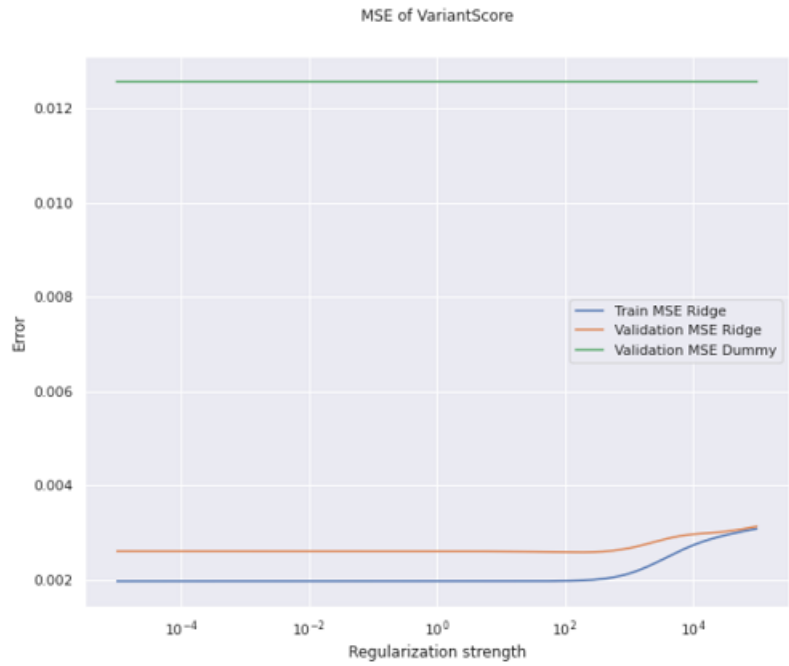
16. כמו קודם, נצפה גם עבור ה־$validation\_set$ לשגיאת אימון **קטנה יותר,** בדומה להסבר שנתנו בסעיף הקודם. בשונה מה־$train\_set$, כאן השיפור בשגיאה יכול להיות פחות משמעותי - היות שהשיפור ב־$train$ עשוי להיות משמעותי (שוב, ראו בהמשך תוצאות שמאמתות זאת) בעקבות $overfitting$, נראה את השגיאה ב־$validation\_set$ עולה (כלומר השגיאה נהיית פחות טובה, באופן יחסי, על שאף שצפוי שיפור ביחס למודל הקודם בכל זאת).

17. נבצע $tuning$ ל־$regularizatios\_strength$ של ה־$Ridge\,regressor$ על שני הסטים של Sex (זכר ונקבה בנפרד) כאשר נוספו עמודות עבור "הדטא בחזקת 2". להלן התוצאות:

איור 12: $Male$


MSE of VariantScore
Train MSE Ridge
Validation MSE Ridge
Validation MSE Dummy
0.018
0.016
0.014
0.012
0.010
0.008
0.006
Error
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
Regularization strength

איור 13: $Female$



איור 14: MSE של ה־$train\_set, validation\_set$ לפי ההייפר־פרמטרים הטובים ביותר עבור ה־$Ridge\,regressor$

Subset	Optimal Strength	Optimal Validation MSE
Male	24.094	0.00510881
Female	153.437	0.00258015

18. להלן טבלה של ה־MSE של ה־$train\_set, validation\_set$ של הרגסורים **מהחלק הנוכחי והקודם** לאחר שביצענו להם $tuning$ עבור כל אחד מבין שני המגדרים:

איור 15: טבלה של ה־MSE של ה־$train\_set, validation\_set$ של הרגסורים לפי הסטים השונים ולפי מגדר

Multilevel Model	Section	Sex	Train MSE	Valid MSE
Linear	4	M	0.00483835	0.00508763
Polynomial	5	M	0.00472393	0.00510881
Linear	4	F	0.00277479	0.00290677
Polynomial	5	F	0.00197811	0.00258015

19. ניתן לראות מהטבלה שהמיפוי הפולינומי שביצענו שיפר את התוצאות, כלומר הקטין את ה־MSE גם ב־$train\_set$ וגם ב־$validation\_set$ עבור **שני המגדרים (למרות שאצל הנשים היה שיפור משמעותי בהרבה ביחס לגברים).** עבור שני המגדרים ככל הנראה המודל הליניארי עשה $underfit$ לדטא, כלומר, חזה פחות טוב את הלייבל $VariantScore$ בהשוואה למודל הפולינומי. הסיבה היא ככל הנראה שהמודל הליניארי פחות **מורכב** מהמודל הפולנומי, בדומה להסבר בשאלה 15, המודל הפולינומי יכול לתאר גם קשרים לא ליניאריים בין הפיצ'רים השונים ולכן לאמן ולחזות יותר טוב את הלייבל ע"ס הדטא. בהסתכלות על הצד השני של המטבע, כפי שתיארנו בשאלה 15, ניתן לחשוב על הגדלת מורכבות הסיווג (העלאה בחזקה של הדטא) כעל $overfitting$ שיגרום לשגיאה ב־$validation\_set$ לקטון פחות באופן יחסי מאשר הקטנת השגיאה ב־$train\_set$ (כמובן שהחיפוש שלנו הוא אחרי ה־$sweet\ spot$ שתביא לתוצאות מדוייקות).

20. להלן טבלה המציגה את ה־MSE של ה־$train\_set$ ו־$validation\_set$ לפי הרגרסורים השונים בתוספת ה־$Multilevel\ polynomial$ של החלק הנוכחי:

איור 16: טבלה של ה־MSE של ה־$train\_set, validation\_set$ של הרגסורים השונים

Model	Section	Train MSE	Valid MSE	Best hyperparameter
Dummy	2	0.0125773	0.0125841	-
Basic linear	3	0.00758314	0.0077756	15.167168884709241
Multilevel linear	4	0.00380682	0.00399969	Male: 15.167168884709241, Female: 217.11179456945052
Multilevel polynomial	5	0.00335197	0.00384026	Male: 24.09403560239527, Female: 153.43684089300132

6. בחינת המודלים על ה־$Test\_set$

21. להלן טבלה המציגה את ה־MSE של ה־$train\_set$ ו־$validation\_set$ לפי הרגרסורים השונים בתוספת **התוצאות החדשות** של כל אחד מהרגרסורים על ה־$test\_set$:

איור 17: טבלה של ה־MSE של ה־$train\_set, validation\_set, train\_set$ של הרגסורים השונים

Model	Section	Train MSE	Valid MSE	Test MSE
Dummy	2	0.0125773	0.0125841	0.0115347
Basic linear	3	0.00758314	0.0077756	0.00747632
Multilevel linear	4	0.00380682	0.00399969	0.003723
Multilevel polynomial	5	0.00335197	0.00384026	0.00333152

22. כפי שניתן לראות מהטבלה בסעיף הקודם, המודל הטוב ביותר על ה־$test\_set$ (כלומר, זה שמשיג את ה־MSE הנמוך ביותר) הוא המודל של $Multilevel\,poly$. ניתן לראות כי ה־MSE על ה־$test\_set$ יחסית קרוב ל־MSE שמושג תוך שימוש ב־$CV\_evaluation$, וכמו כן, התוצאות על ה־$test\_set$ הן מקבילות לתוצאות בשימוש ה־$CV\_evaluation$ מבחינת טיב השגיאה שהמודלים משיגים (כלומר, מודל שהשיג שגיאה יותר נמוכה ביחס למודל אחר ב־$CV\_evaluation$, ישיג שגיאה נמוכה יותר מאותו מודל גם על ה־$test\_set$). מכך ניתן להסיק שהמודלים שלנו עובדים טוב באופן כללי, ולא מתקיים $overfitting/underfitting$ על ה־$train\_set$.

23. **המודל הנבחר הוא** $Multilevel\,poly$ שכבר הוצג בסעיפים הקודמים. הוא המודל שהשיג את ה־MSE הנמוך ביותר עבור ה־$Test\_set$ מבין כל אלו שניסינו. אעיר כי ניסינו לא מעט מודלים נוספים כדי לנסות לשפר את השגיאה (הכוונה במודלים **שלא** ניסינו כחלק מהעבודה והסעיפים הקודמים), ביניהם:

$$ElasticNet, Lasso, Bayesian\,Regression$$

ועוד כ־10 אחרים ולמרות זאת אף אחד מהם לא השיג שגיאה טובה יותר.