HW3_Wet

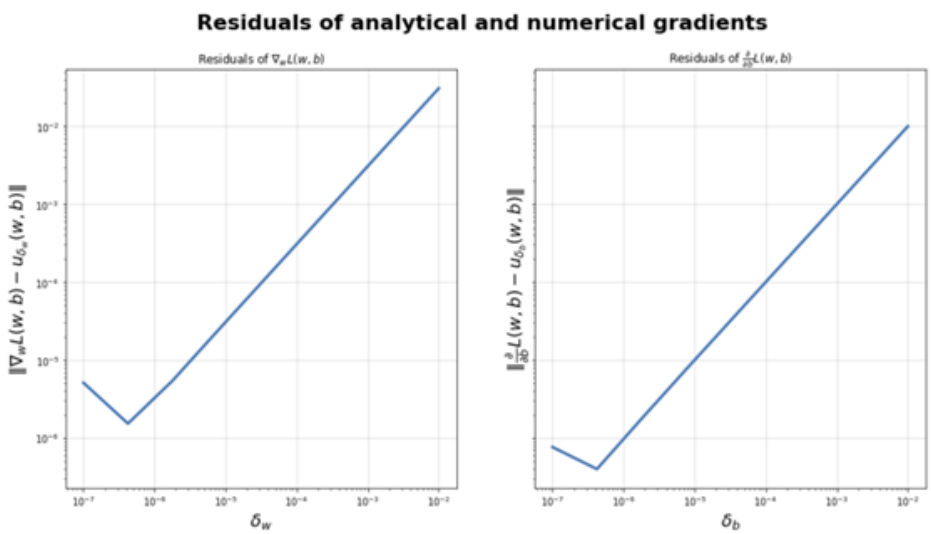
עומר טאוב – 316497122

אורי זהר – 205960750

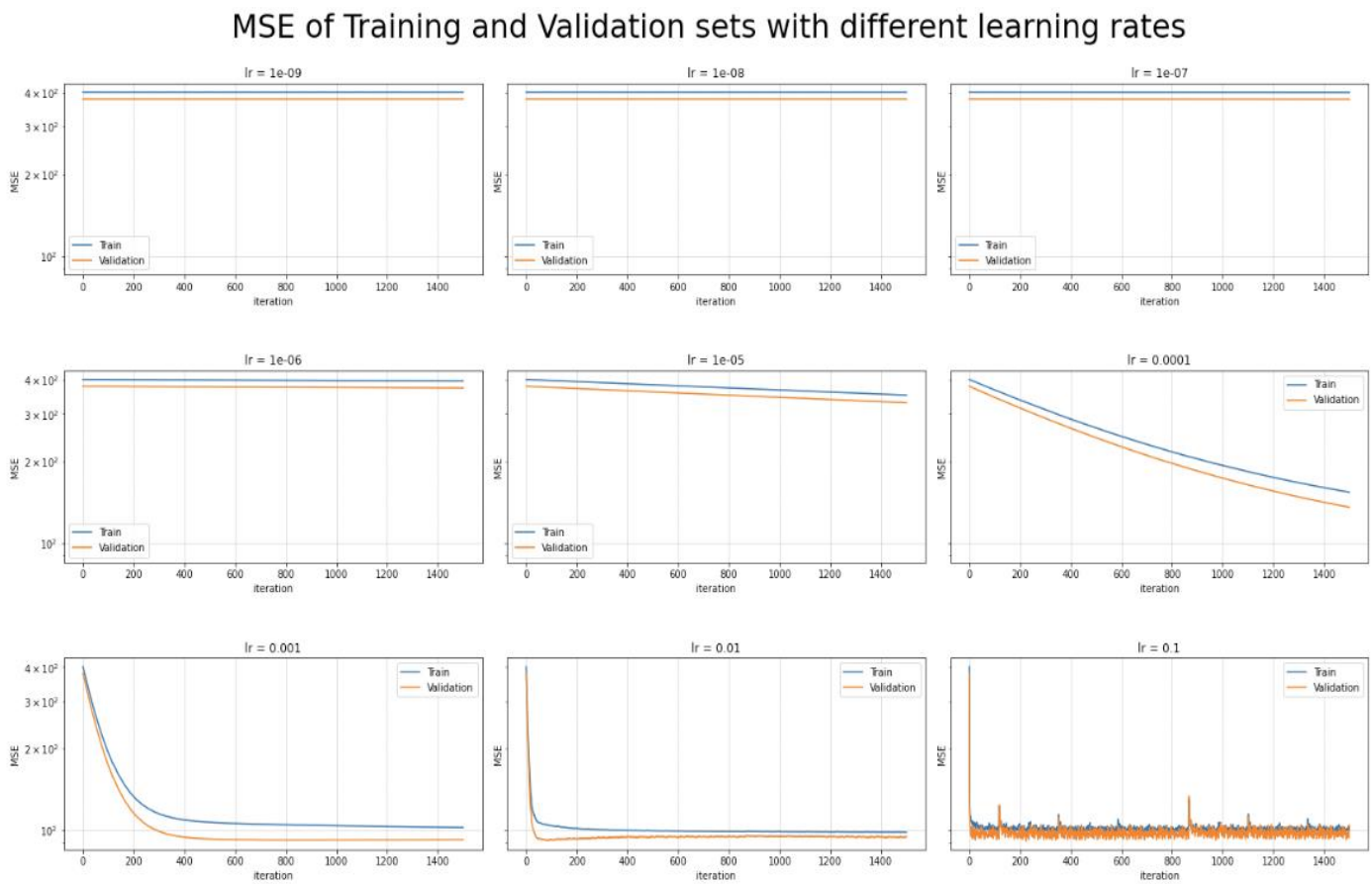
Q(1)

$$\frac{\partial}{\partial b}L(\underline{w},b)=\frac{\partial}{\partial b}\frac{1}{m}\sum_{i=1}^{m}\left(\underline{w}^{T}\underline{x}_i+b-y_i\right)^2=\frac{1}{m}\sum_{i=1}^{m}2\left(\underline{w}^{T}\underline{x}_i+b-y_i\right)$$

Q(2)



Q(3)



נשים לב למספר מסקנות שמתקבלות מהגרפים על קצב ההתכנסות (גודל צעד) :

1. אם נבחר את קצב ההתכנסות להיות קטן מידי נגיע אל המינימום רק לאחר מספר גדול מאוד של איטרציות, ע"פ הגרפים ניתן לראות שקצב התכנסות קטן מידי הוא לכל הערכים שקטנים או שווים ל- 0.0001.

2.אם נבחר קצב התכנסות גדול מידי לא נצליח להתכנס אל המינימום אלא רק נתקרב אליו (במקרים מסויימים אפילו נפספס אותו), ע"פ הגרפים ניתן לראות שקצב התכנסות גדול מידי הוא לערכים גדולים או שווים ל-0.1.

3.ניתן לראות שבין 0.01 ל-0.001 נקבל תוצאות טובות, אנחנו מקבלים את השגיאה המינימאלית גם בסט האימון וגם בסט הוולידציה במספר יחסית נמוך של איטרציות.

הבחירה שלנו לערך של קצב ההתכנסות האופטימלי הוא 0.001, הוא מתכנס בצורה יציבה אחרי מספר נמוך של איטרציות לערך המינימום.

עבור קצב ההתכנסות האופטימאלי נראה שכבר אחרי 800 איטרציות אין שינוי עבור ערכי שגיאות האימון והוולידציה ולכן לא נקבל תוצאות טובות עבור הגדלת מספר האיטרציות.

Section 2

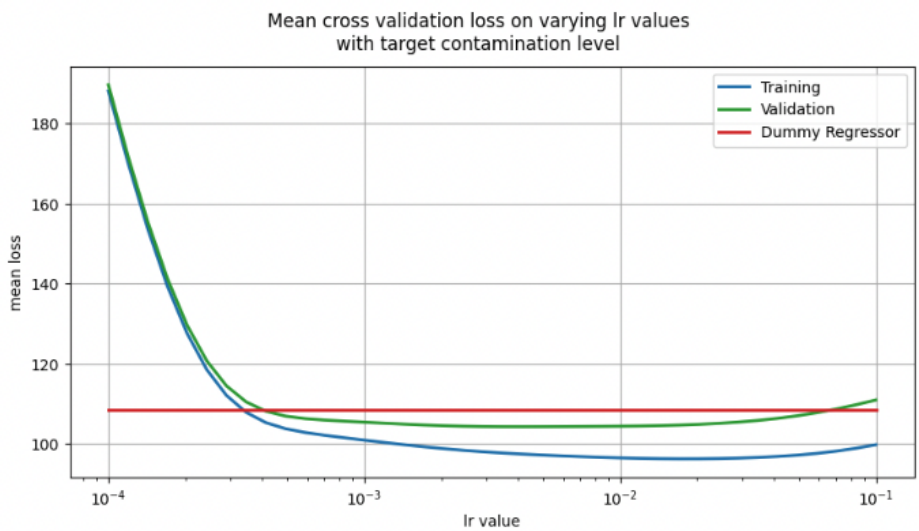
Q(4)

Model	Section	Train MSE	Valid MSE
		Cross Validated	
Dummy	**2**	108.10	108.42

Q(5)

קצב הלמידה האופטימלי (עבורו קבלנו את שגיאת הוולידציה הקטנה ביותר) הוא

Optimal_lr = 0.004124626382901352



Model	Section	Train MSE	Valid MSE
Dummy	2	108.10	108.42
Linear	2	97.39	104.22

Q(6)

עבור שני המודלים הנ״ל נירמול הפיצ'רים של סט האימון לא היה משנה את התוצאות.

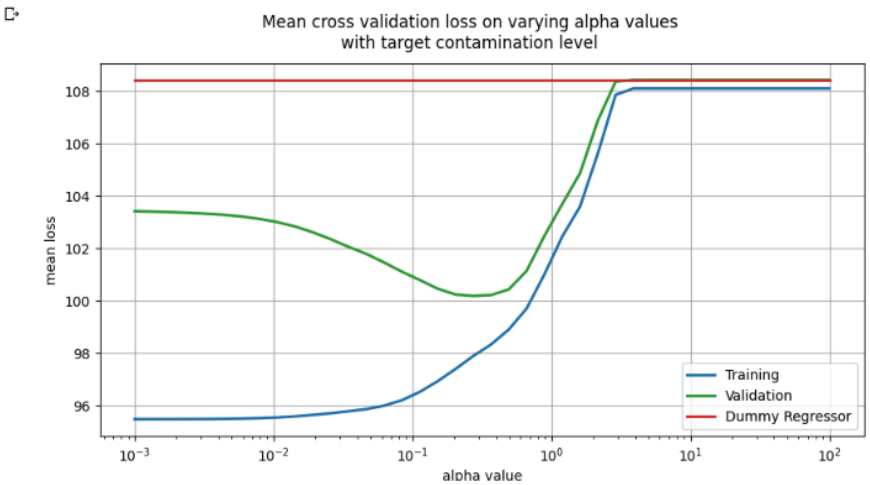
1. Linear Regression : זהו מודל לינארי בו כל כניסה בווקטורים w ו-b לומדת את הדאטה המתאים לכניסה המתאימה בווקטור x בהתאמה. כלומר הלמידה מתבצעת לכל פיצ'ר (כניסה בווקטור x) בצורה נפרדת. כתוצאה מכך נקבל שנרמול פיצ'ר לא ישפיע על הביצועים אלא רק על סדרי גודל.
2. Dummy : במודל זה התחזית תלויה בממוצע של כל הלייבלים בסט האימון ואינה תלויה בכלל בפיצ'רים לכן עבור מודל זה פעולת הנרמול לא תשנה את הביצועים.

Section 3

Q(7)

ערך אלפא האופטימלי (עבורו קבלנו את שגיאת הוולידציה הקטנה ביותר) הוא

Optimal alpha=0.2728333376486767



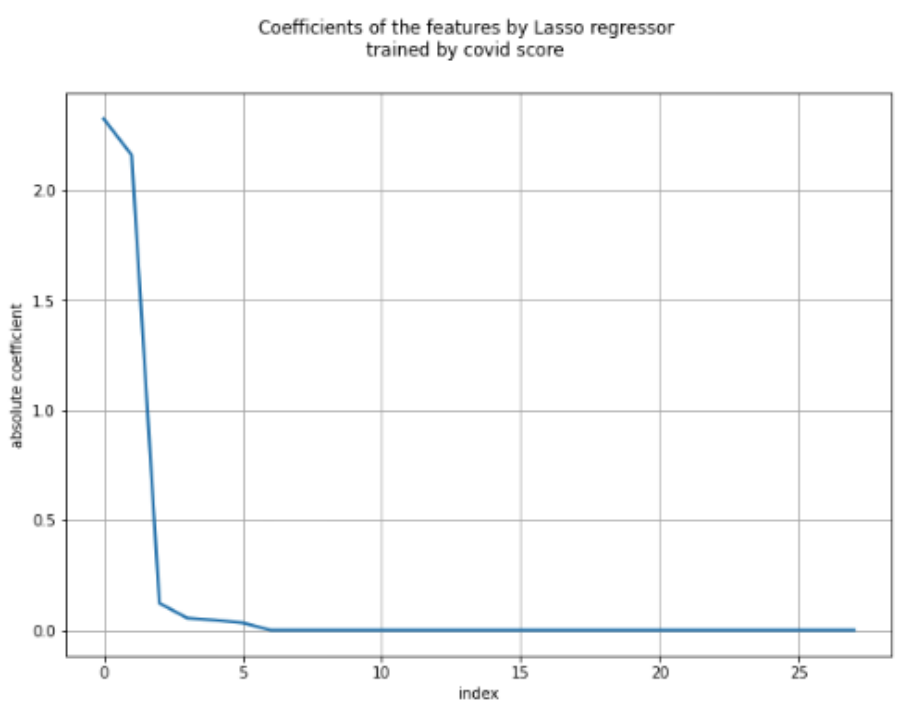
Q(8)

Model	Section	Train MSE	Valid MSE
Dummy	2	108.10	108.42
Linear	2	97.39	104.22
Lasso Linear	3	97.88	100.17

Q(9)

sugar_levels	2.3256087584934084
PCR_01	2.1611692695609457
low_appetite	0.12283656981561519
PCR_10	0.05479358693257271
current_location_x	0.045453952014649766

Q(10)



Q(11)

גודל המקדם של כל פיצ'ר הוא המשקל שהמודל נתן לו כדי לחזות את הערך מטרה, לכן ככל שהמקדם גדול יותר הפיצ'ר הזה חשוב יותר לחיזוי.

ניתן לראות כי המקדמים עבור רוב הפי'צרים קרובים מאוד ל-0 לכן הכפלה של פי'צר (דגימה חדשה) במשקל זה לא ישפיע רבות(אם בכלל) על ניבוי הערך. ולעומת זאת לערכים מאוד גדולים השפעה גדולה על ערך התוצאה.

Q(12)

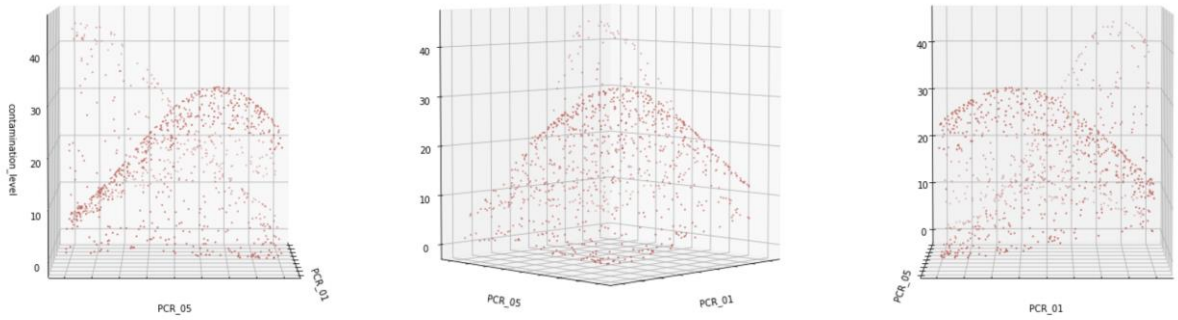
במודל הלאסו בשונה מהמודל הלינארי יש גם את האיבר של הרגולריזציה, איבר זה כן מושפע מהבדלים בסדרי הגודל בין פיצרים שונים. כפי שהסברנו בשאלה מספר 6, הנירמול משנה את סדרי הגדול של הכניסות בווקטור w ובמקרה בו לא היה רגולריזציה הביצועים על מודל האימון לא היו משתנים אבל בתוספת רגולריזציה יש גם חשיבות לסדרי הגודל ואלה משפיעים על החשיבות של הפיצ'ר בפרדיקציה.

בגלל שאנחנו רוצים לתת לכל פיצ'ר משקל שווה בקבלת ההחלטות לביצוע פרדיקציה, הביצועים במודל הלאסו לאחר הנירמול יהיה טובים יותר.

Section 4: Polynomial fitting (visualization)

Q(13)

Three dimensional scatter plot of PCR_01, PCR_05, and contamination_level



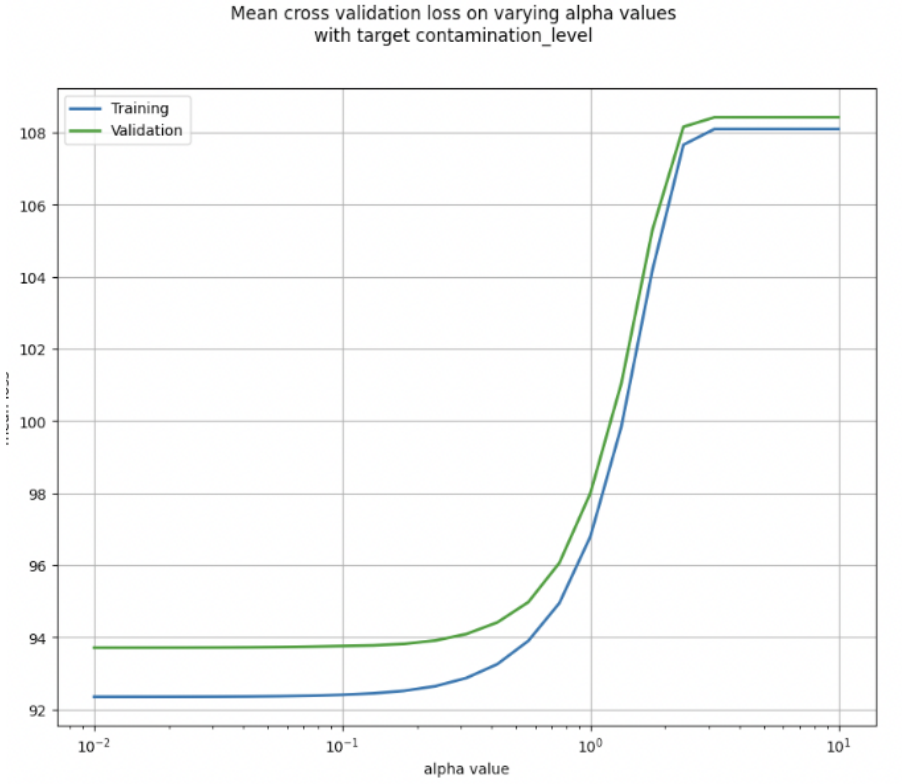
ניתן לראות שפיזור הנקודות מזכיר פרבולה תלת מימדית, מפיזור הנקודות ניתן להבין שהדאטה אינו לינארי ובחירת מודל לא ליניארי יניב תוצאות טובות יותר.

בעזרת מידע זה נשקול לבחור מודל רגרסיה פולינומי עם פולינום בדרגה גבוהה. הסיבה לכך היא שמודלים של רגרסיה פולינומית יכולים להתאים לקשרים לא ליניאריים, ופרבולה תלת מימדית היא פונקציה לא ליניארית.

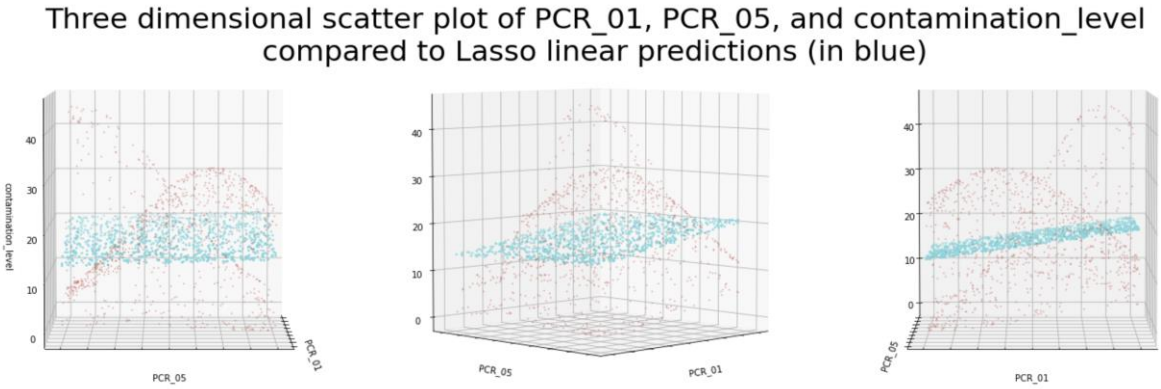
Q(14)

ערך אלפא האופטימלי (עבורו קבלנו את שגיאת הוולידציה הקטנה ביותר) הוא 0.001

(השגיאה עבור סט הוולידציה היא 105.47 והשגיאה עבור סט האימון היא 104.62 (עם האלפא האופטימלי



Q(15)



Q(16)

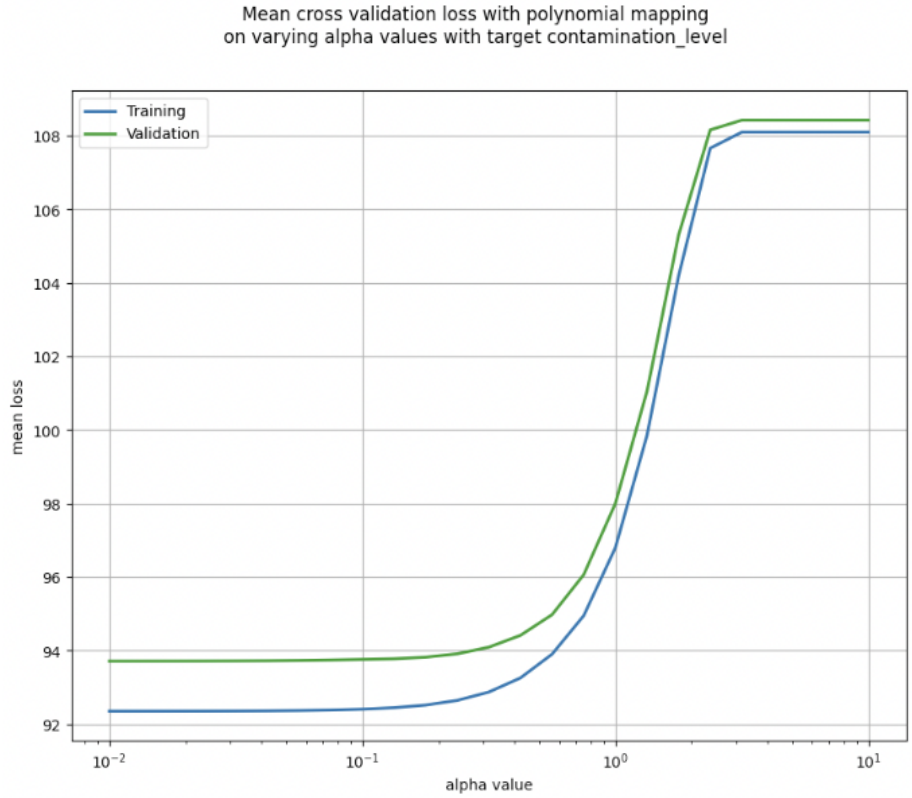
חשוב לבצע נרמול אחרי ביצוע מיפוי פולינומיאלי בגלל שלאחר המיפוי נוצרים לנו פיצרים חדשים שלא היו וייתכן ויחרגו מטווחי הערכים המנורמלים שהיו לנו לפני .

כפי שתיארנו לפני מודל הלאסו מושפע מסדרי גודל שונים בין הפיצ'רים שלו ולכן על מנת לא שלא נפגע בביצועים עלינו לנרמל ולוודא שכל הפיצרים יהיו בעלי אותם סדרי גודל.

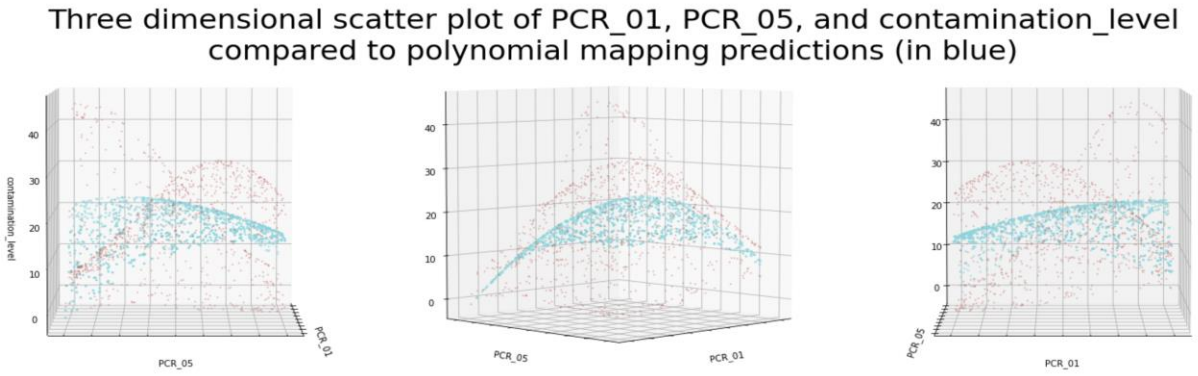
Q(17)

ערך אלפא האופטימלי (עבורו קבלנו את שגיאת הוולידציה הקטנה ביותר) הוא 0.01

(השגיאה עבור סט הוולידציה היא 93.72 והשגיאה עבור סט האימון היא 92.36 (עם האלפא האופטימלי



Q(18)



Q(19)

מההצגה ויזואלית הראשונה ראינו כי הדאטה מתפזר בצורה של פולינומיאלית ורצינו להבין האם בחירה של מודל פולינומיאלי ישפר את הביצועים.

כפי שראינו, הצלחנו להקטין את השגיאה וכמסקנה מכך נבין שעלינו לבחור מודל שאינו ליניארי על מנת לנבא טוב יותר את הדאטה.

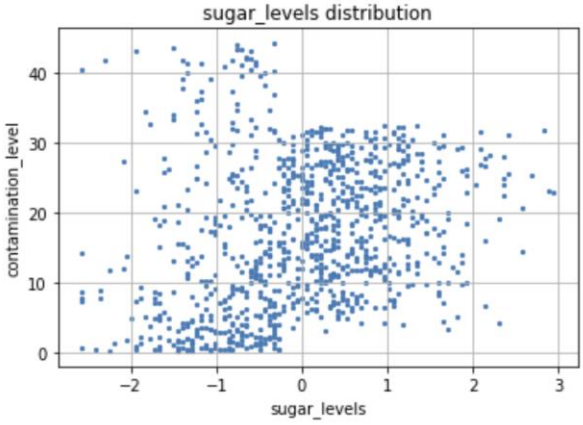
עדיין ניתן לראות שקבלנו שגיאה יחסית גבוהה ושימוש במיפוי פולינומיאלי אינו מספיק על מנת לקבל פרדיקציה אמינה- נחוץ מרחב היפוטזות גדול יותר.

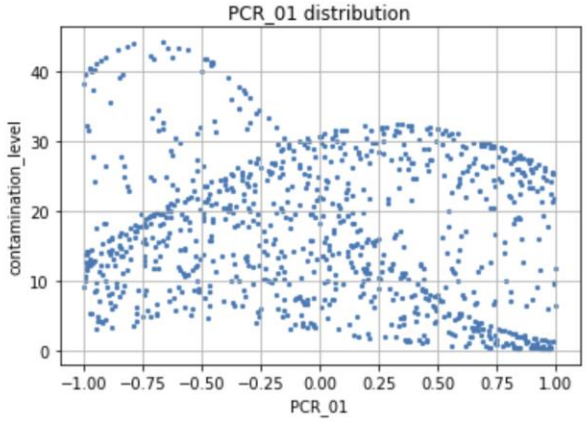
Section 5 :RandomForest fitting of the CovidScore

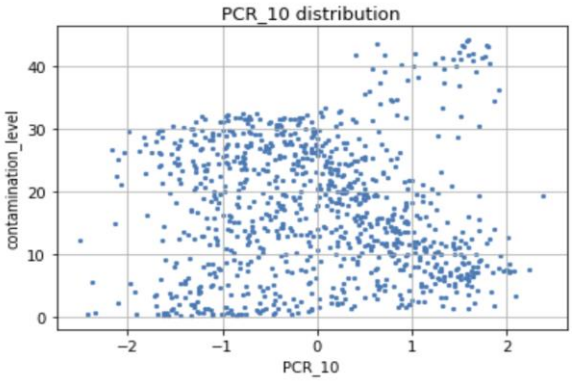
Q(20)

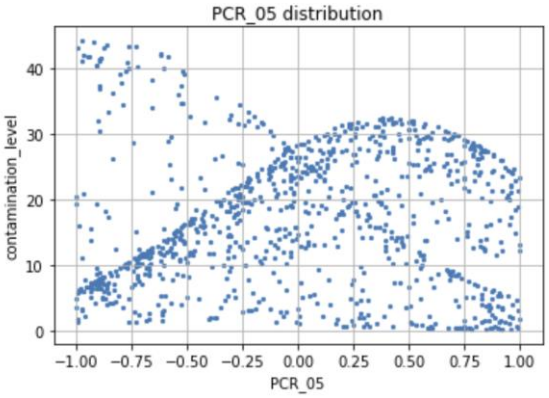
חיפשנו מועמדים מתאימים לבדוק האם מיפוי פולינומי/גאוסיאני יכולים לשפר את היכולות שלהם לניבוי המודל. על מנת שמיפוי כזה יוביל לשיפור בביצועים נרצה לוודא שהקשר בין הפיצ'ר לערך המטרה אינו ליניארי וגם שלפיצר חשיבות גבוהה בניבוי המודל. בסעיפים קודמים ראינו שהפיצ'רים PCR_01, sugar_levels, PCR_10, low_apetite, current_location_x הם 5 הפיצרים בעלי הcoeffiecient הגבוהה ביותר במודל הלאסו ועל כן ניתן להסיק שיש להם חשיבות גבוהה בניבוי המודל ולכן נסמן אותם כמועמדים ובנוסף גם PCR_01 ו-PCR_05 נסמן כמועמדים כי ראינו שהדאטה שלהם מתפלג בצורה לא לינארית אך בצורה שניתן לנבות את משתנה המטרה.

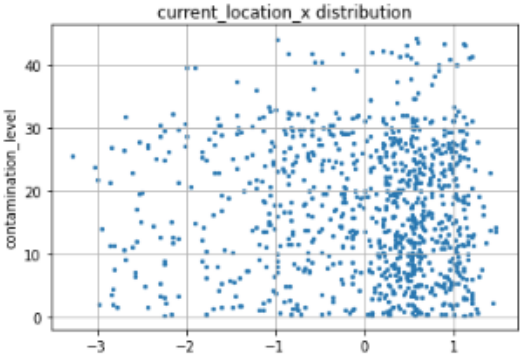
ננסה לראות שאכן הקשרים של המועמדים אינו ליניארי ולשם כך נציג את הגרפים ה-2 מימדיים שלהם כנגד ערך המטרה :

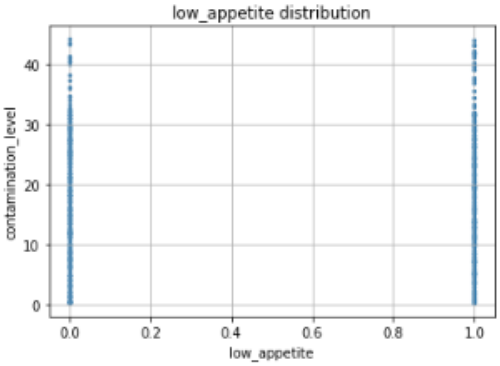

sugar_levels distribution
contamination_level
sugar_levels


PCR_01 distribution
contamination_level
PCR_01


PCR_10 distribution
contamination_level
PCR_10


PCR_05 distribution
contamination_level
PCR_05


current_location_x distribution
contamination_level


low_appetite distribution
contamination_level
low_appetite

כפי שניתן לראות הפיצ'ר low_appetite הוא פיצ'ר בינארי לכן החלטנו לא לבצע עליו מיפוי נוסף ולהשאיר אותו כמו שהוא.

בדיקה נוספת שבצענו בניסיון לפסול מועמדים הוא בדיקת הקורלציה שלהם לערך המטרה - קורלציה גבוהה מידי תראה על קשר ליניארי שלא נרצה להרוס אותו - אבל הערכים לא היו מספיק גבוהים על מנת שנפסול מועמדים ללא בדיקה.

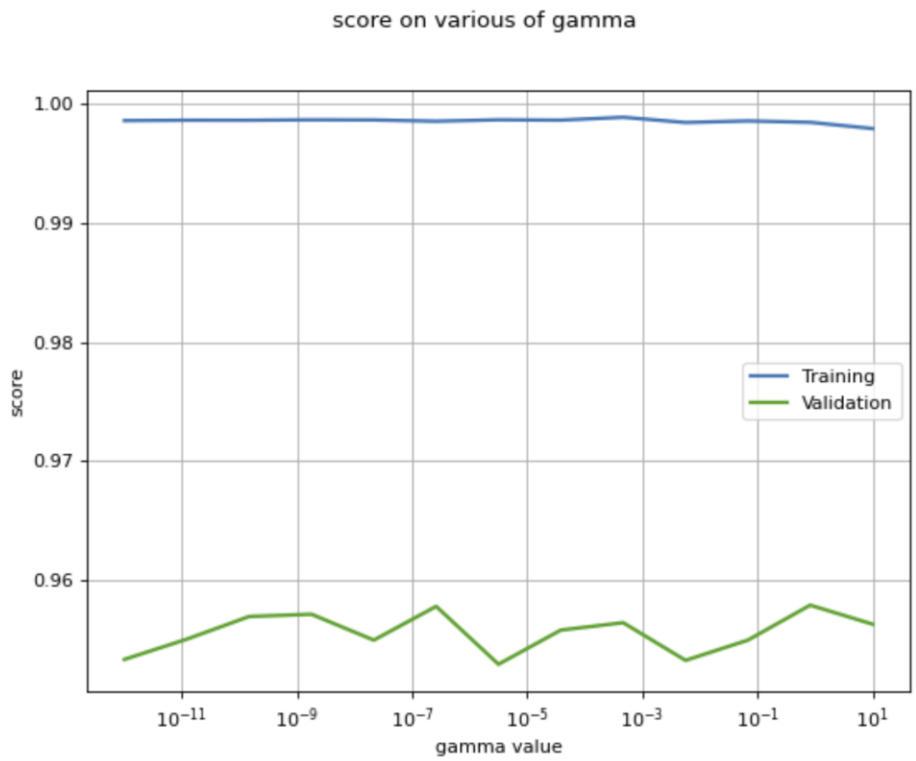
Contamination_level	
current_location_x	0.027698
sugar_levels	0.254438
PCR_01	0.174791
0.032456	PCR_05
PCR_10	0.041070

כעת נרצה לבדוק עבור אילה מהפיצ'רים(מתוך ארבעת הפיצ'רים) נרצה למפות בעזרת RBF mapping ואיזה בעזרת Polynomial mapping. בשביל לדעת לאיזה פיצ'ר מתאים איזה מיפוי הרצנו partitions ל-3 רשימות (rbf, poly no_map) מתוך רשימת הפיצ'רים הרלוונטים למיפוי. עבור כל חלוקה כזאת נבצע cross validation על סט האימון ועבור החלוקה הטובה ביותר נבחר אותה בתור החלוקה שאיתה נמשיך כ- rbf_features, polynomial_features.

לכן הפיצ'רים שהחלטנו לעשות מיפוי RBF הם - [PCR_01].

הפיצ'רים שהחלטנו לעשות מיפוי פולינומיאלי הם - [PCR_01, PCR_5].

עבור פיצ'רים אלו החלטנו למצוא את הgamma המתאים:



הגאמה שאותו למדנו הוא : 10^{-6}

Q(21)

נסביר למה ואיך שגיאות האימון והולידציה ישתנו כאשר נשתמש במיפוי RBF על הדאטה במודל .RandomForest

Training_error- שגיאת האימון צפויה לקטון כאשר נשתמש במיפוי RBF מכיוון שכאשר אנחנו משתמשים במיפוי זה על פי'צר אנחנו מגדילים את מרחב הפי'צר, וכן הופכים את המודל למורכב יותר. במודל זה פיצ'רים טובים(לפי entropy לדוגמא) צפויים שיהיו בהרבה עצי החלטה חלשים מכיוון שהם עוזרים יותר ועוזרים לנבא את ה target variable. לכן ביצוע RBF על פיצ'רים יהפוך את אותם פיצ'רים לאקספרסיבים יותר ובעזרת הטרנספורמציה יוכלו להפריד בצורה טובה יותר. כך יוכל המודל ללמוד את סט האימון בצורה מדויקת יותר ובכך שגיאת האימון צפויה לקטון.

Validation_error - שגיאת הולידציה צפויה לקטון או לגדול זאת בהתאם לדאטה ולמספר הפי'צרים. כאשר משתמשים ב RBF המודל שלנו מורכב יותר, וכתוצאה מכך נפגעת היכולת ההכללה שלו על מידע שלא אומן ועלול להתקיים Ovefitting על סט האימון.

כאשר יש מספיק samples והפי'צר לאחר המיפוי נותן יכולת ביטוי טובה לtarget שגיאת הולידציה צפויה אז לקטון.

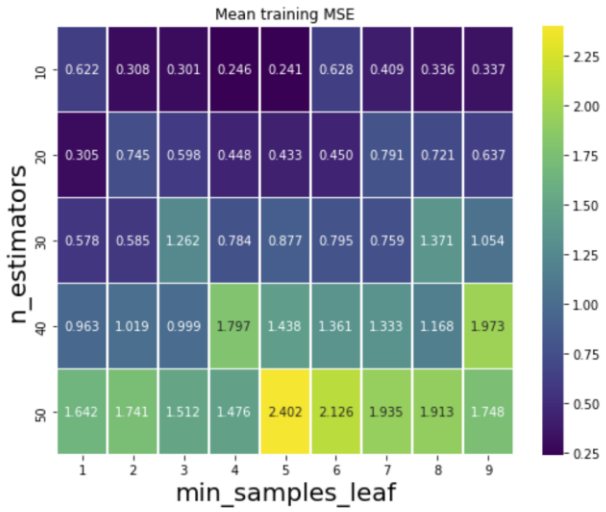
- חשוב לציין כי במודלים מסוג ensemble קשה יותר להגיע למצב OverFitting מכיוון הסתמכות המודלים החלשים שכל אחד מהם הוא underfitting לדאטה - כפי שלמדנו בהרצאה.

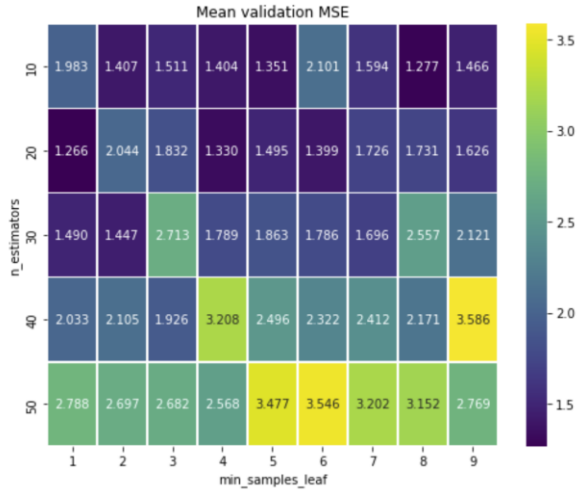
שאלה 22

ההבדל בין המודל RandomForest לשאר מודלי הרגירסיה הפולינומים שעשינו עד עכשיו הוא שמודל RandomForest הוא מסוג Ensemle - מורכב מהרבה עצי החלטה חלשים לעומת מודל אחד לינארי בשאר המודלים. בנוסף, למודל זה יש את היכולת להשתמש במספר קטן יותר מהפיצ'רים שנמצאים בdata שלנו מכיוון שכל עץ חלש בוחר פי'צרים לפי ההערכה שלו כמה הפיצ'ר יעזור לנבא. על כן פיצ'רים אשר לא עוזרים כלל לעצים החלשים לא ישתתפו בניבוי המודל ensemble של כולם. הבדל נוסף במודל זה לעומת שאר המודלים הלינאריים הוא למודל זה יש נטייה פחותה יותר ל overfitting מאשר שאר המודלים מכיוון שכפי שהזכרנו, מורכב מהרבה עצים חלשים שלכל אחד יש נטייה קטנה מאוד לoverfitting בעצמו לכן יקטין את הסיכוי של המודל החזק להיות ב overfitting בעצמו.

שאלה 23

מיקדנו את החיפוש שלנו בטווח הנ"ל לאחר מספר הרצות על טווחים גדולים יותר וקפיצות בין פרמטרים גדולים יותר.





כפי שניתן לראות ערך ה - MSE הטוב ביותר עבור ה - train הינו : 0.241

וערך ה - MSE הטוב ביותר עבור ה - test הינו:1.266

הפרמטרים הטובים ביותר עבור ה validation set הם

min_sample leafs =1

n_estimators = 20

שאלה 24

Model	Section	Train MSE	Valid MSE
		cross validated	
Dummy	2	108.10	108.42
Linear	2	97.39	104.22
Lasso	3	97.88	100.17
RF Regressor	5	0.241	1.266

שאלה 25

Model	Section	Train MSE	Valid MSE	Test MSE
	25	cross validated		Retrained
Dummy	2	108.10	108.42	108.126
Linear	2	97.39	104.22	98.324
Lasso	3	97.88	100.17	98.293
RF Regressor	5	0.241	1.266	4.762

המודל אשר נתן את התוצאה הטובה ביותר הוא הRandomForest Regressor שכן שגיאת הMSE שלו קטנה פי בערך 25 מכל מודל אחר שתרגיל על סט המבחן.

Dummy model - כפי שניתן לראות מודל זה בעל הביצועים הגרועים ביותר והיה ניתן להסיק זאת מעצם פעולתו, חישוב ממוצע של ה-target ללא התחשבות במידע על הפיצ'רים. ניתן לראות כי הוא סובל מ-underfitting שכן שגיאת האימון שלו גבוהה באופן יחסי .

Linear model - זו המודל שני הגרוע ביותר מבחינת הביצועים, בשונה ממודל ה-dummy הוא כן מתייחס אל הפיצ'רים אך אינו מצליח לסווג אותם כראוי. מודל זה סובל מ-overfitting כי כפי שניתן לראות קיים הבדל יחסית גדול בין השגיאה של סט האימון לשגיאה של סט הוולידציה.

Lasso model - ניתן לראות כי המודל הנ"ל מתנהג בצורה דומה למודל הליניארי אך ההבדל בניהם הוא הרגולריזציה שמונעת את ה-overfitting שנוצר ואכן כפי שניתן לראות יש שיפור בביצועים על סט הוולידציה.

RandomForestRegressor Model - זהו המסווג הטוב ביותר ביותר מבין כל המסווגים, ניתן לראות שמאופן פעולתו בו הוא בוחר את הפיצ'רים האינפורמטיבים ביותר (ובנוסף בגלל שהפיצ'רים הנ"ל אינם ליניארים אז מיפוי

באמצעות RBF אפילו ישפר את הביצועים) ונמנע משימוש בפיצ'רים שאינם תורמים ללמידת הדאטה (בכך גם נמנע overfitting) מתקבלות תוצאות גבוהות משמעותית משאר המסווגים.