

תקציר מנהלים

במהלך פרויקט זה עמדה בפנינו המטרה לבנות מודל שיחזה עבורנו ויסווג את סט הנתונים שלנו לפי Label 0 ו-1. זאת באמצעות סט נתונים המורכב מפיצ'רים וסיווגים קיימים. מימשנו מודלים שונים תוך התנסות וחקירה ושאפנו למצוא את המודל המתאים ביותר על מנת להצליח לחזות סט נתונים ללא תוצאה ידועה. השתמשנו בשיטות שונות והשוונו מודלים שונים במטרה ברורה: למצוא את המודל המתאים ביותר לסט הנתונים כך שיחזה באופן הטוב ביותר את סט המבחן שלנו.

לאחר מכן השתמשנו הכלים אשר נלמדו בכיתה על מנת לנקות את הנתונים החריגים (בחנו 2 שיטות) והפעלנו אלגוריתמים כגון PCA, פיצ'ר סלקשן לפי קורלציה על מנת לצמצם את מספר הפיצ'רים בנתונים שלנו. החלטנו לחקור 2 כיוונים טיפול במשתנים הקטגוריאליים. אשר הניבו תוצאות מעניינות ומגוונות.

חקרנו לעומק כיווני Clustering באמצעות פונקציות שונות חיפשנו סוג מסוים אשר יספק לנו מידע חדש על טיב ואודות הנתונים (לא ניתן לנו מידע על מקור הנתונים). לאחר מבט ויזואלי (כשהיה ניתן) ושימוש בחישוב קורלציה הוספנו עמודת קלוסטרנינג בתקווה לשיפור תוצאות המודל.

חקרנו לעומק את בעיית ה-imbalanced dataset כיוון שהסט נתונים שלנו סבל ממנה בצורה חמורה.

לאחר התלבטות החלטנו על כיוון טיפול בבעיה זאת (SMOT oversampling) ועיצבנו את סט הנתונים לפיה תוך ניהול סיכוני overfitting וחלוקה לסט ולידציה.

לאחר מכן, השתמשנו בשיטות שונות ובאלגוריתמי למידת מכונה אשר למדנו בקורס ובעזרת 2 המדדים איתם הונחינו לקבוע, בחרנו להשתמש במודל רשת נוירונים עם קידוד בשיטת OneHot עם מדד AUC של 0.805

ומדד חדש של 0.668 על סט הולידאציה.

לבסוף יצרנו 2 קבצי CSV קובץ אחד כמו שנדרשנו אשר מכיל את ההסיתברויות ואחד נוסף המכיל את הקביעה (0 או 1).

בדוח זה נתאר את הפעולות וההחלטות שבצענו עד כדי קביעת המודל הטוב ביותר.

חלק א' אקספלורציה

בשלב האקספלורציה נרצה לקבל מידע על הנתונים. כלומר נרצה להבין מה סוג המידע בכל אחד מהפיצ'רים(העמודות) וכן נרצה להבין באיזה אופן המידע מתפלג ומה הם הטווחים בהם המידע מתפלג. ראשית הצגנו מדגם של 5 רשומות ראשונות עבור כל העמודות.(נספח 1) לאחר חקירה עמודה של ההתפלגויות והערכים בשימוש בגרפים היסטוגרמות ופירוטים אשר מאפשרת חבילת pandas. הסקנו את הנתונים הבאים:

Feature	Data options		Distribution
0	-0.490607	9.092011	נורמלית
1	2.437300	248.877854	נורמלית אסימטרי
2	0	81.200000	נורמלית אסימטרי
3	1	100.000000	נורמלית
4	3	100.000000	נורמלית אסימטרי
5	D-P		התפלגות בדידה(ערכים קטגוריאליים)
6	a0-a30		התפלגות בדידה(ערכים קטגוריאליים)
7	0.45	3.435	נורמלית
8	-0.492857	2.128	נורמלית
9	1	12.00	בדידה
10	-0.195661	0.954	2 התפלגויות שנראות נורמליות(אחת קטנה ואחת גדולה)

11	984.50	1038.9	נורמלית
12	983.70	1040.9	נורמלית
13	0 1 unknown		התפלגות בדידה(ערכים קטגוריאליים)
14	0.0 -62.62 mm		התפלגות בדידה(ערכים קטגוריאליים)
15	0	143.0	לא ידוע
16	3.9	46.1	נורמלית
17	-0.7	36.9	נורמלית
18	A -P		התפלגות בדידה(ערכים קטגוריאליים)
19	A-P		התפלגות בדידה(ערכים קטגוריאליים)
20	11	135.0	לא ידוע
21	0	69.0	ערכים בדידים, נורמלית
22	2010	2012	בדידה (שנים)

23	0	670	ערכים בדידים, נורמלית
24	0.001355	1426.45	נורמלית אסימטרי
label	0	1	בינארית

על מנת למצוא קורלציה בין הנתונים בדקנו עמודות שערכיהם הם נומריים בלבד מכיוון שעל פיטצרים קטגוריאליים לא ניתנים לבדיקה על סמך מה השיטות שנלמדו בכיתה. עמודות בעלות קורלציה גבוהה(על פי הגרף קורלציה נספח ד):

0.91	8,17	0.89	0,1
0.96	11,12	0.89	1,2
0.86	16,17	0.98	7,16
		0.89	7,17

חלק ב' עיבוד מקדים

בשלב העיבוד המקדים ביצענו מניפולציות ושינויים בסט הנתונים על בסיס הנלמד בכיתה ומדריכים בKaggle.

הסרת חריגים

ראשית, הזנו נתונים חסרים באמצעות ממוצע העמודה עבור כל אחת מהעמודות המספריות(נומריים).

לאחר מכן ניקינו נתונים חריגים מן העמודות הנומריים לפי Z scoren כפי שנלמד בתרגול.מצאנו את ציון z עבור כל אחת מנקודות הנתונים במערך הנתונים, ואם ציון z גדול מ - 3, נוכל לסווג נקודה זו כOutliner וכנקודה חריגה(כל נקודה מחוץ ל -3 סטיות תקן). לדוגמה עמודה 0 עם ערכים רחוקים מעל הערך 4.63 . ביצענו הסרת חריגים 3 סטיות תקן הורידו מעל 8.3% מן השורות והוסרו 1859 רשומות. **לאחר מכן** בחנו אפשרות להחסיר פחות שורות באמצעות שברון עבור ההתפלגויות הנורמליות(quantile)ע"י קיצוץ של 0.01 אחוז מהקצוות.(חזרנו לשלב זה על מנת לבחון כיצד ניקוי שונה של חריגים משפיע על המודל).

בחירת פיצ'רים והסרת עמודות בעלות קורלציה גבוהה

בחלק האקספלורציה ציינו כי חלק מן פיצרים הם בעלי קורלציה גבוהה מאוד. וכעת בצענו הסרה של עמודות(Feature selection) אשר הראו קורלציה גבוהה מאוד. הסרנו את העמודות הבאות: ['1','11','17','16','2']. בשלב זה הצגנו את הנתונים בשנית בעזרת פונקציה ווידאו כעת לאחר הסרת חריגים הנתונים מתפלגים באופן צפוף ולמראית עין הנתונים נראו פחות מפוזרים.

אסטרטגיות דגימה מחדש לטיפול בחוסר איזון(Imbalanced dataset)

השלב הבא(אשר נלמד שלאחר מעשה) היה לסדר את החוסר איזון אשר קיים מבחינת מספר שורות אשר מסווגות ל'1' ול'0' (5240 לעומת 16921). כדי להתמודד עם החוסר איזון השתמשנו בטכניקה שאומצה באופן נרחב להתמודדות עם Datasets מאוד לא מאוזנים נקראת דגימה מחדש(Resampling). שיטה זאת מורכבת מהסרת דגימות ממעמד הרוב (תת-דגימה) ו / או הוספת דוגמאות נוספות ממעמד המיעוט (דגימת יתר). היישום של דגימת יתר(over-sampling) הוא שכפול רשומות אקראיות מהסיווג הפחות נפוץ, מה שעלול לגרום לoverfitting. ב-under-sampling, הטכניקה כוללת הסרת רשומות אקראיות ממעמד הרוב, מה שעלול לגרום לאובדן מידע. לאחר שבנינו פונקציות עבור 2 השיטות, החלטנו להשתמש בפונקציית over-sampling וטיפלנו בחשש מאוברפיטינג ע"י שימוש בחבילת Smoth המייצרת שורות סנתטיות עבור הלייבל הפחות מיוצג. בנוסף דאגנו תמיד לבדוק עם סט ולידציה שלא יוצרו ממנו שורות סנתטיות.לאחר סידור חוסר האיזון בשיטה שנבחרה(over-sampling) מספר הרשומות עלה ל 29,346.

טיפול בעמודות והמשתנים הקטגוריאליים

השתמשנו ב-2 דרכים על מנת לטפל במשתנים הקטגוריאליים:

- One Hot Encoding - זה בעצם הייצוג של משתנים קטגוריים כקטורים בינאריים. ערכים קטגוריים אלה ממופים לראשונה לערכים שלמים. לאחר מכן כל ערך שלם מיוצג כקטור בינארי שכולו 0 (למעט האינדקס של המספר השלם המסומן כ-1).
- Target Encoding - קידוד מבוסס יעד הוא מספור משתנים קטגוריים באמצעות עמודת Label. בשיטה זו אנו מחליפים את המשתנה הקטגורי במשתנה מספרי חדש אחד בלבד ומחליפים כל קטגוריה של המשתנה הקטגורי בהסתברות המתאימה למטרה (אם קטגורית) או ממוצע היעד (אם מספרי) לפי מספר המופעים.

ממדיות הבעיה: למה ממדיות גדולה עלולה ליצור בעיה?

נראה כי אכן ממדיות הבעיה גדולה מידי. שכן על פי כלל האצבע עבור N פיצ'רים יש צורך ב-N² רשומות.

שיטה	מספר פיצ'רים	כמות דאטה (לפי כלל אצבע)
One Hot Encoding	100	9409
Target Encoding	24	400

סט הנתונים שלנו מורכב מ-29,346 רשומות, יש בידינו כמות גדולה מן הנדרש. הבעיה בממדיות גדולה היא שככל שיש יותר מאפיינים נדרש יותר מידע ויש חשש ל-overfitting כפי שגם תארנו בהסבר על הקידוד מכון מטרה.

נרמול

הבא, נבצע נרמול לנתונים. בבעיה שלנו יש צורך בנרמול הנתונים. יש חשיבות לנרמול מכיוון שקיימים ערכים שהטווחים שלהן שונים מאוד ופעם זה בעצם נותן השפעה גדולה מאוד לפיצ'רים עם הטווחים הגדולים, לכן נבצע נרמול. בנוסף על מנת לבצע PCA נבצע כנדרש נרמול ראשוני.

ביצוע PCA

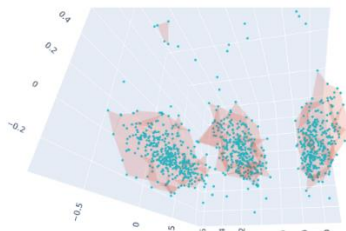
כפי שהזכרנו בהרצאה, על מנת לבצע PCA כנדרש הפרדנו את המשתנים הנומריים ואת המשתנים הבינאריים ובצענו PCA בנפרד עבור כל דאטה פריים לאחר מכן חיברנו אותם בחזרה.

שיטה	מספר פיצ'רים מקורי	לאחר PCA
One Hot Encoding	100	89
Target Encoding	24	18

ביצוע Clustering

על מנת לבצע Clustering השתמשנו בפונקציית Visualizer K-Elbow המיישמת את השיטה לבחירת המספר האופטימלי של האשכולות (Clusters) לאשכול K-אמצעי. כלומר ע"י שימוש בפונקציה זאת קיבלנו אומדן למספר הקלוסטרס בדאטאסט. לאחר חיפוש ארוך בעזרת פונקציה מצאנו שילוב של 3 עמודות אשר היה בעל קורלציה של 0.33 אחוז לעמודת הלייבל. נציין כי הקורלציה עלולה הייתה להיות בגלל החיפוש הנרחב (כפי שראינו בכיתה כאשר הוצג לפנינו גרף התאבדויות לצד גרף פיתוח האינטרנט) ולכן ניקח את השימוש בקלוסטר זה בעירבון מוגבל.

3d point clustering



חלק ג'-מימוש המודלים

בחלק הראשון, תחת המודלים הבסיסיים נבחרו המודלים הבאים:
 1. מודל Logistic regression, במודל זה השתמשו השתמשנו באותם ההיפר פרמטרים ל-2 קבצי הנתונים (Target encoding, One-hot encoding) כאשר הפרמטרים היו:

הסבר	ערך Gridsearch	משתנה
שיטת הקנס בו ישתמש המודל בעת פתרון בעיית האופטימיזציה.	l1	penalty
פרמטר רגולציה משמש בשביל למנוע אוברפיטינג ככל שערכו של הפרמטר גדול יותר קיימת פחות רגולציה.	1	C
האלגוריתם בו יעשה שימוש המודל בשביל לפתור את בעיית האופטימיזציה.	'liblinear'	solver
קריטריון עצירה שבנוי "סובלנות".	100	max_iter
חסם עליון למספר האיטרציות שהמודל יכול לבצע עד הפסקת ריצות המודל	0.0001	tol

רמת הדיוק אליה הגענו בתצורות השונות היו:

- עבור Target encoding הדיוק שהושג היה שווה ל- 0.872
- עבור Hot encoding הדיוק שהושג היה שווה ל- 0.890

עבור מודל זה בצענו גם grid search ועבורו:

הסבר	ערך Gridsearch עבור Target decoding	ערך Gridsearch עבור One-Hot decoding	משתנה
שיטת הקנס בו ישתמש המודל בעת פתרון בעיית האופטימיזציה.	l1	l1	penalty
פרמטר רגולציה משמש בשביל למנוע אוברפיטינג ככל שערכו של הפרמטר גדול יותר קיימת פחות רגולציה.	0.61584	0.615848	C
האלגוריתם בו יעשה שימוש המודל בשביל לפתור את בעיית האופטימיזציה.	'liblinear'	liblinear	solver
קריטריון עצירה שבנוי "סובלנות".	1000	1600	max_iter
חסם עליון למספר האיטרציות שהמודל יכול לבצע עד הפסקת ריצות המודל	0.0001	0.0001	tol

רמת הדיוק אליה הגענו בתצורות השונות היו:

סוג/ מדד	Avg AUC	מדד חדש	validation
Target encoding	0.87	0.538	0.779
Hot encoding	0.89	0.598	0.805

2. המודל השני בו השתמשנו היה מודל K nearest neighbor, ההיפר פרמטר שנבחר עבור מודל זה היה:

הסבר	ערך Gridsearch עבור One-Hot Hot decoding	ערך Gridsearch עבור Target decoding	משתנה
מספר השכנים שאותם על המודל לבדוק על מנת לתת פרדיקציה לערך הדגימה	30	28	N_neighbiors
פונקציית משקל המשמשת בחיזוי: <u>uniform</u> : משקולות אחידות. משקל כל הנקודות בכל שכונה.	distance	weights	weights

			distance: נקודות משקל לפי היפוך המרחק שלהם. במקרה זה, לשכנים קרובים יותר של נקודות שאילתה תהיה השפעה רבה יותר משכנים שנמצאים רחוק יותר.
--	--	--	---

רמת הדיוק הטובות ביותר אליה הגענו בתצורות השונות:

- עבור Target encoding הדיוק (AUC) שהושג היה שווה ל- 0.870
- עבור Hot encoding הדיוק (AUC) שהושג היה שווה ל- 0.799

מודלים מתקדמים

בחלק השני של בחירת המודלים נבחרו 2 מודלים מתקדמים כאשר גם בחלק זה, עבור כל אחד מהמודלים האלו נבחרו 2 תצורות של הנתונים, כלומר בסכ"ה 2 סוגי מודלים כאשר לכל סוג יש 2 קבצי נתונים שונים.

3. מודל Random Forest. במודל זה בשביל למקסם את מידת הדיוק של המודל, אל מול זמן חישוב סביר. הרצנו את המודל על מספר שונה של עצים בכל "יער" (הדבר בוצע מספר פעמים בשביל לתת אמינות סטטיסטית לגודל היער הנבחר). לאחר מכן בצענו חיפוש Gridsearch כדי לוודא את שאר הפרמטרים לאחר בדיקה זו נבחרו ההיפר פרמטרים הבאים למודל זה:

הסבר	ערך Gridsearch עבור One-Hot Hot decoding	ערך Gridsearch Target decoding	משתנה
מספר השכנים שאותם על המודל לבדוק על מנת לתת פרדיקציה לערך הדגימה	300	300	n_estimators
מספר התכונות שיש לקחת בחשבון כשמחפשים את הפיצול הטוב ביותר.	sqrt	auto	max_features
הפונקציה למדידת איכות פיצול. הקריטריונים הנתמכים הם: Gini Impurity - "gini" "entropy" - על פי רווח המידע.	entropy	entropy	criterion
העומק המרבי של העץ. אם אין, אז הצמתים מורחבים עד שכל העלים טהורים או עד שכל העלים מכילים פחות מדגימות המינימום דגימות.	10	10	max_depth
המספר המינימלי של הדגימות הנדרש להיות בצומת עלים.	8	1	min_samples_leaf
המספר המינימלי של הדגימות הדרוש לפיצול צומת פנימי.	5	2	min_samples_split

רמת הדיוק אליה הגענו בתצורות השונות היו:

סוג / מדד	Avg AUC	מדד חדש	validation
Target encoding	0.91	0.635	0.802
Hot encoding	0.92	0.639	0.806

4. מודל ANN. עבור מודל זה נבחרו ההיפר פרמטרים הבאים:

הסבר	ערך Gridsearch עבור One-Hot Hot decoding	ערך Gridsearch Target decoding	משתנה
פונקציית האקטיבציה לשכבות הנסתרות	relu	tanh	activation
אלפא הוא פרמטר למונח רגולציה, המכונה מונח penalty, הנלחם בהתאמת יתר על ידי הגבלת גודל המשקולות.	1e-08	0.1	alpha

hidden_layer_sizes	10	10	פרמטר המציין את מספר שכבות הנסתרות ומספר הניורונים בכל שכבה מהשכבות השונות.
max_iter	500	500	חסם עליון למספר האיטרציות שהמודל יבצע.
random_state	3	2	קובע יצירת מספרים אקראיים למשקולות ואתחול ה-Biases.
solver	lbfgs	lbfgs	השיטה בה נבצע את אופטימיזצית המיסקול בכל ריצה

רמת הדיוק אליה הגענו בתצורות השונות היו:

validation	מדד חדש	Avg AUC	סוג/ מדד
0.805	0.599	0.89	Target encoding
0.803	0.629	0.91	Hot encoding

חלק ד'- הערכת אחד המודלים שבהם השתמשנו:

המודל אשר אותו בחרנו לנתח בעזרת ה Confusion matrix הוא מודל ANNN עבור ריצת המודל על חלק validation מקובץ הנתונים עבור OneHot incoding. כאשר $n=2216$ מתקבלת המטריצה הבאה:

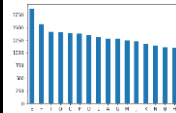
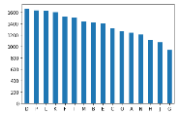

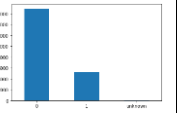
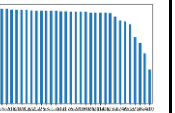
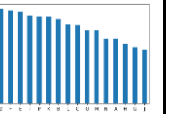
	Actual "Yes"	Actual "No"	Sum rows(the system predictions)
predicted "Yes"	TP=289	FP=100	389
predicted "No"	FN=284	TN=1358	1642
Sum columns(the actual labels of the data)	573	1458	

מקרא תאי המטריצה:

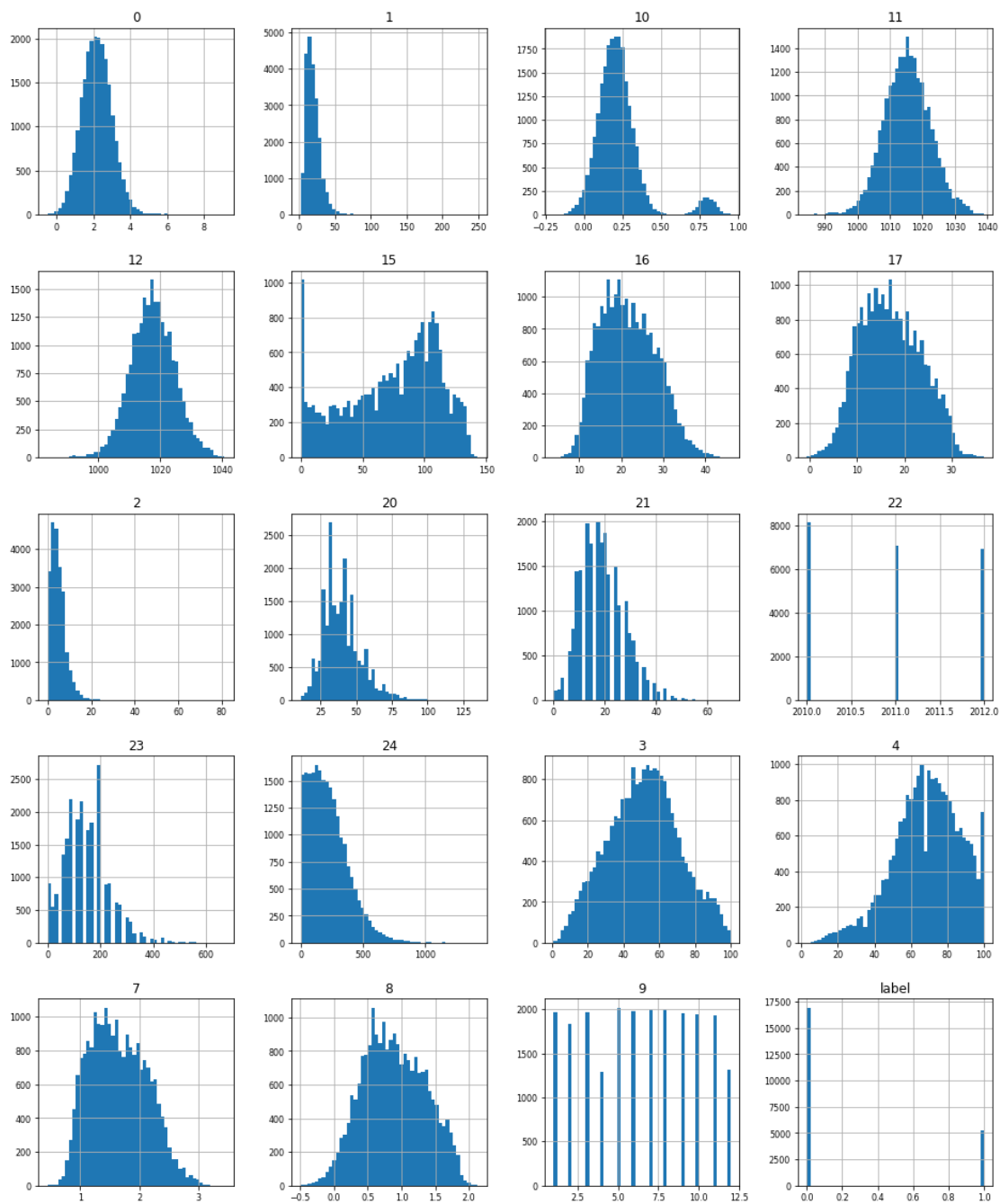
- TP = המודל ניבא ערך של 1 לתצפית וערך התצפית הוא אכן 1
- TN = המודל ניבא ערך של 0 לתצפית וערך התצפית הוא אכן 0
- FP = המודל ניבא ערך של 1 לתצפית אך ערך התצפית הוא 0
- FN = המודל ניבא ערך של 0 לתצפית אך ערך התצפית הוא 1
- רמת הדיוק שמתקבלת מהמודל לפי המטריצה (מדד ה-Accuracy) שווה ל-0.784 כלומר 78.4%
- ובמדד החדש על סט הולידשן היא 0.530.

נספחים

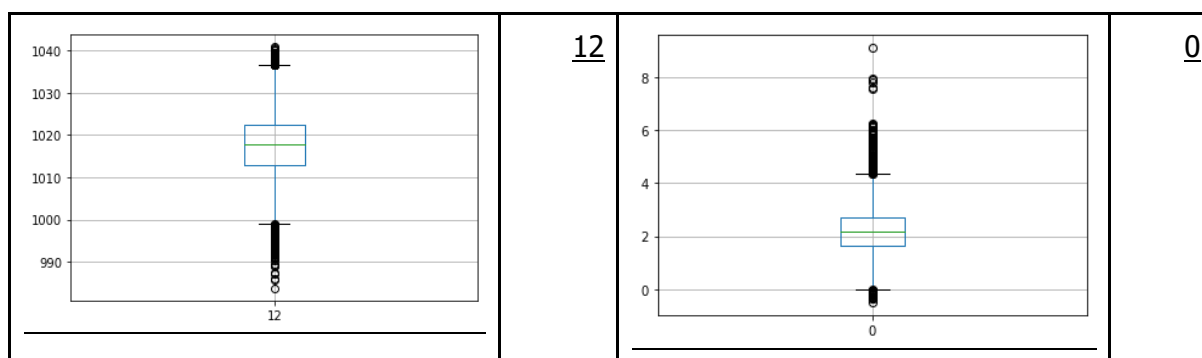
נספח א: Barplot עבור העמודות הלא נומריות:

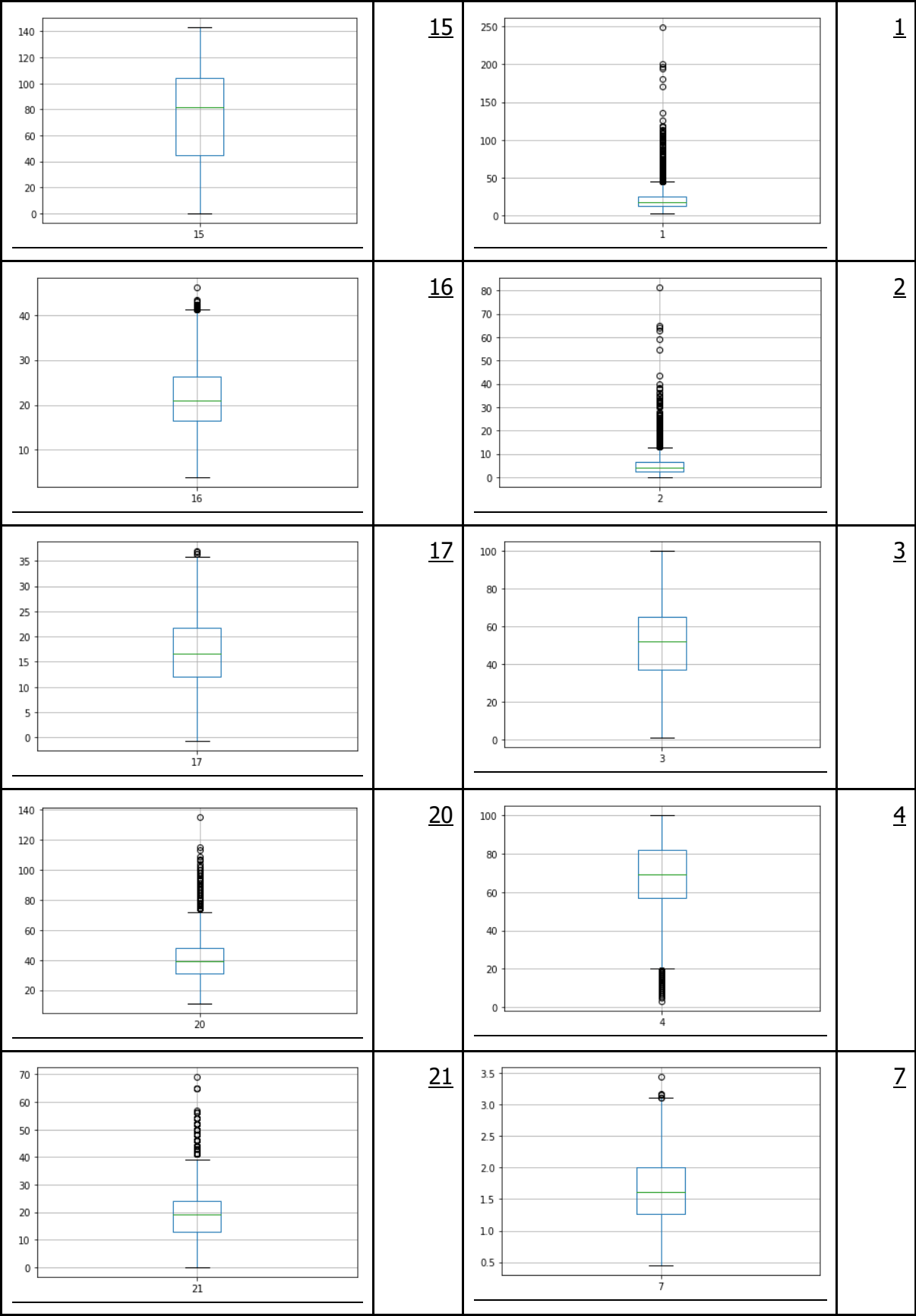
19	18	14	13	6	5
					

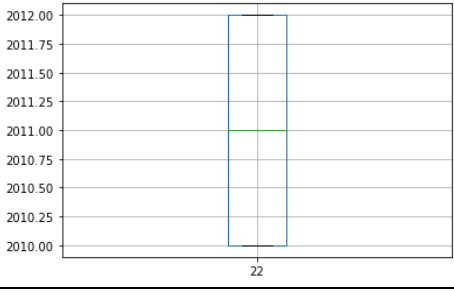
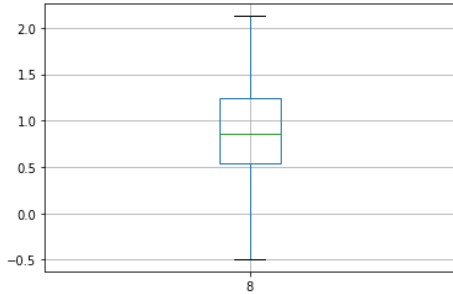
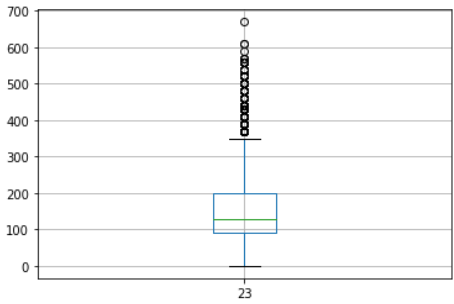
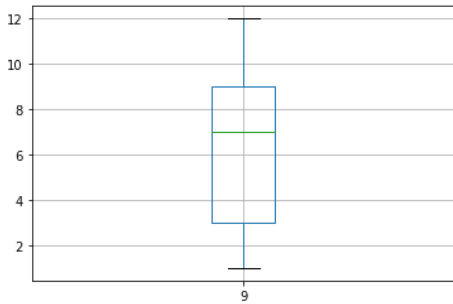
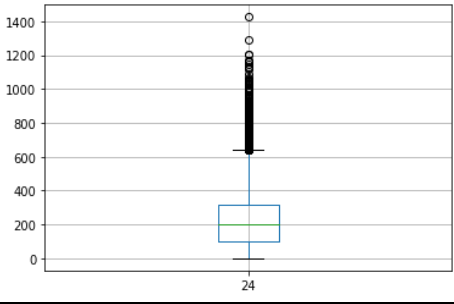
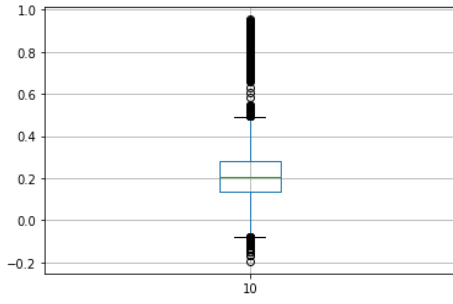
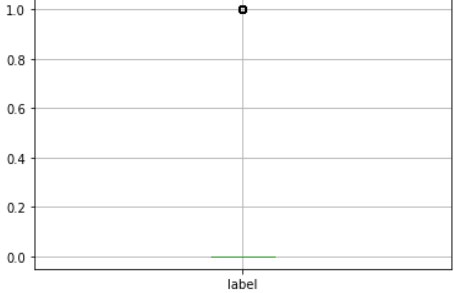
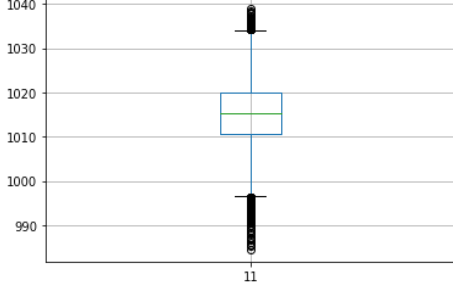
נספח ב: היסטוגרמה עבור העמודות הנומריות:



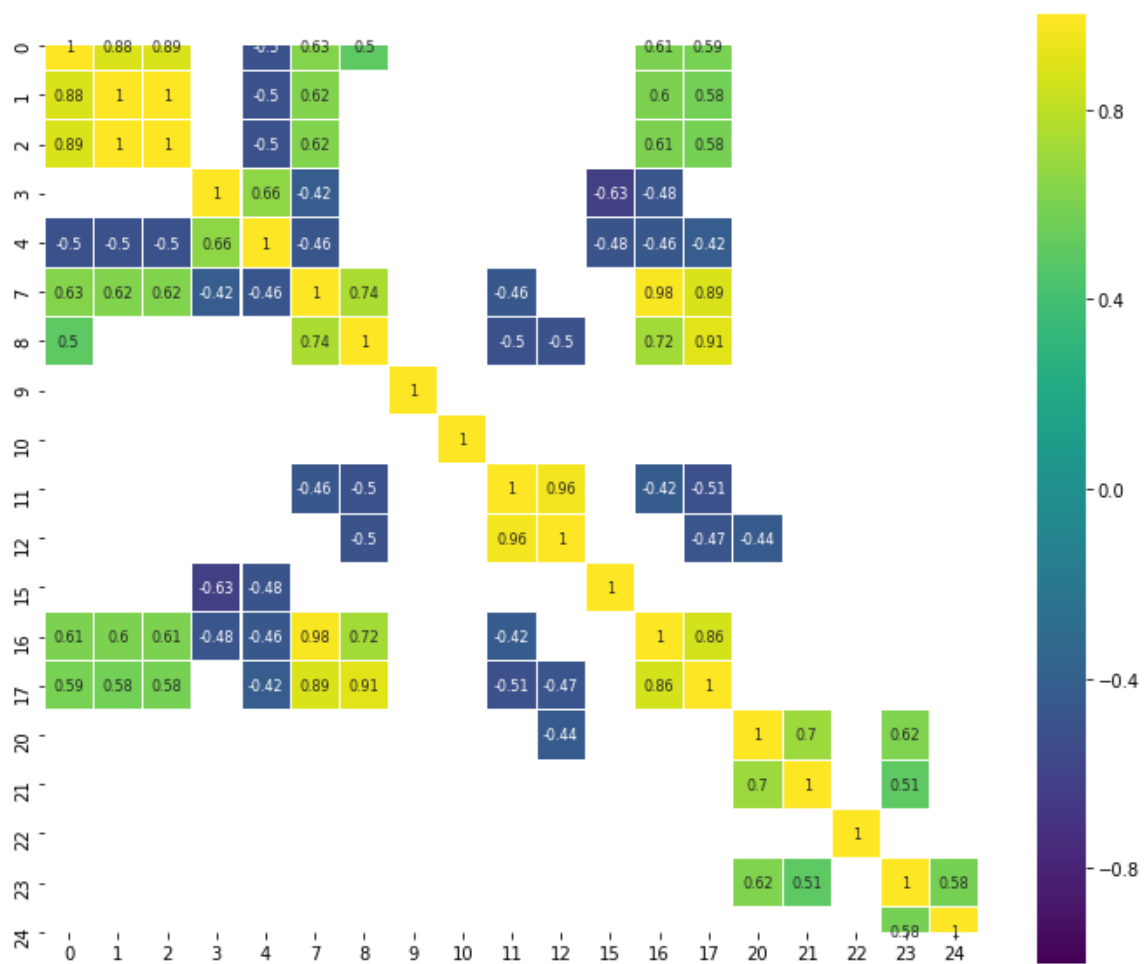
נספח ג: גרפי BoxPlot עבור העמודות הנומריות



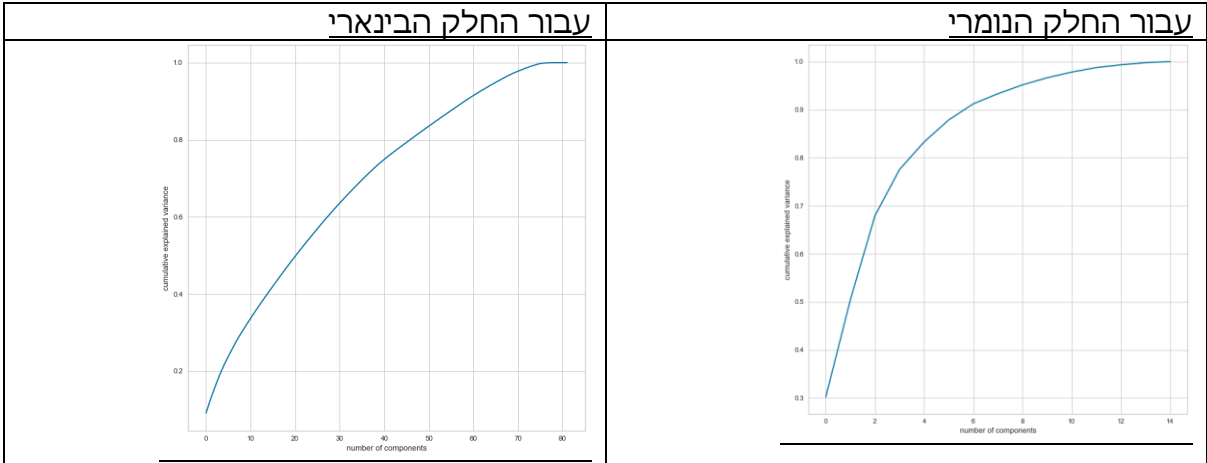


	<u>22</u>		<u>8</u>
	<u>23</u>		<u>9</u>
	<u>24</u>		<u>10</u>
	Label		<u>11</u>

נספח ד : גרף קורולציה בין כל פיטצור נומרי

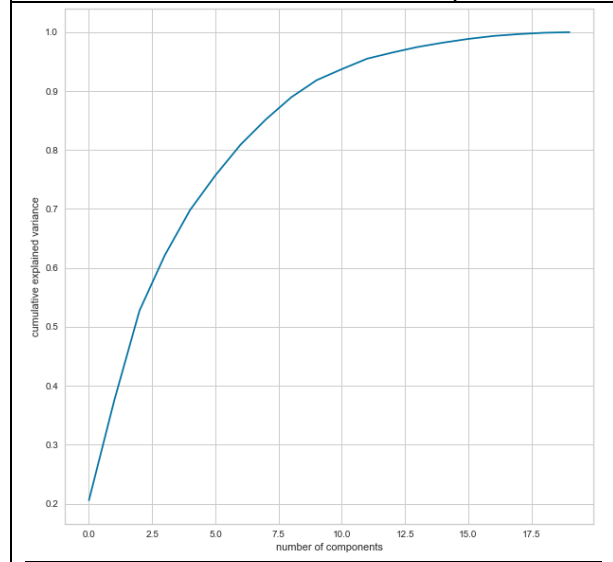


נספח ה - PCA עבור One-Hot



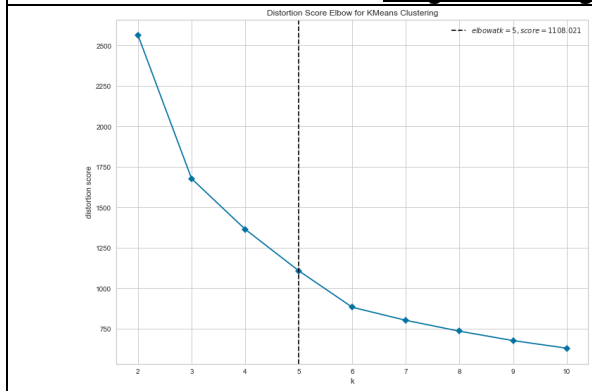
נספח ה – PCA עבור target

עבור החלק הנומרי

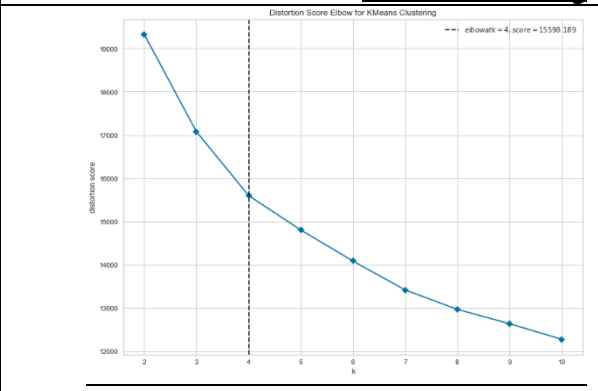


נספח ו – חיפוש מספר הקלוסטרינג לפי Elbow

Target decoding

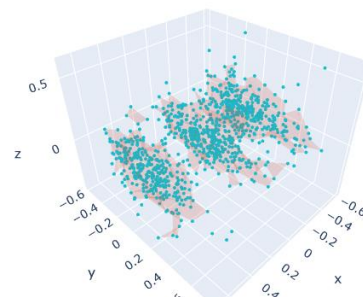


OneHot decoding

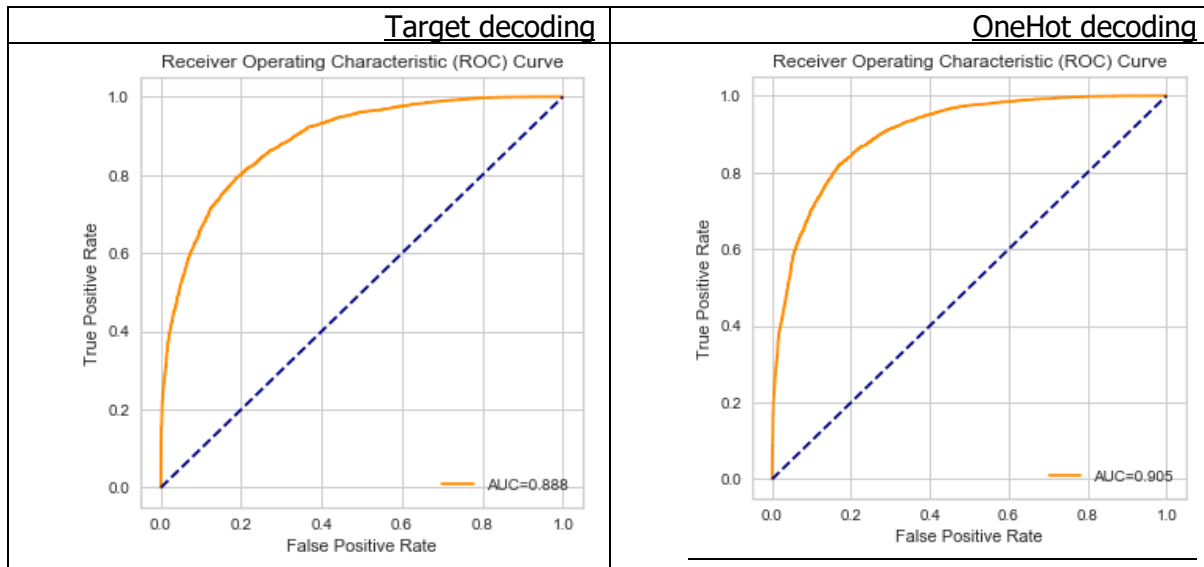


נספח ז – ויזואליזציה וחיפוש עמודת קלוסטרינג בעלת זיקה לליבל

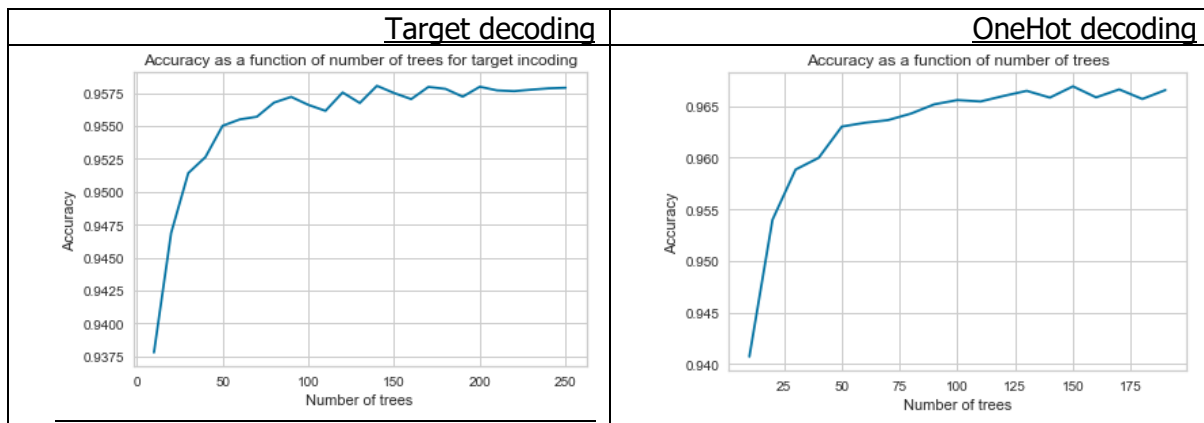
3d point clustering

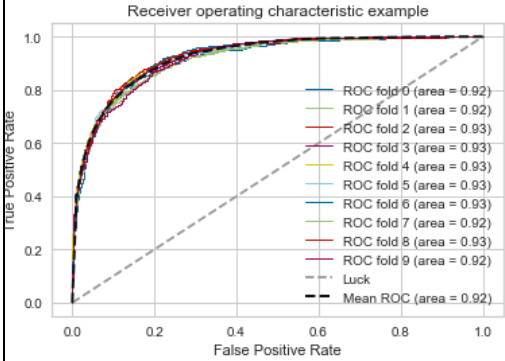
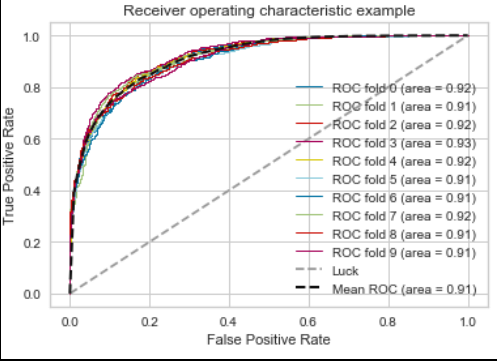
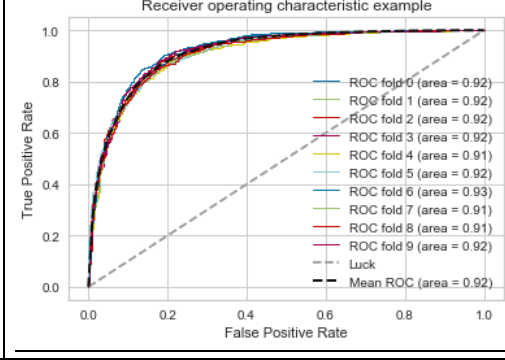
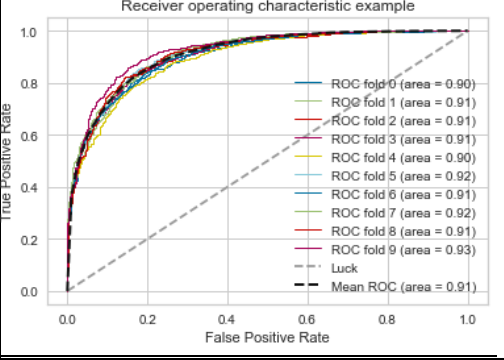
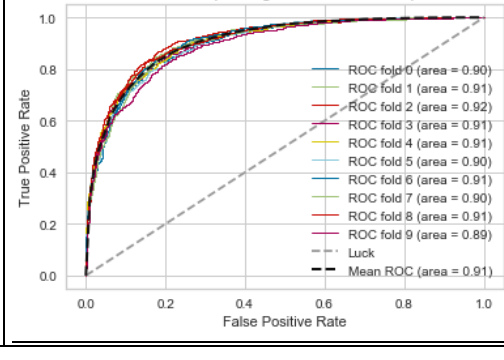
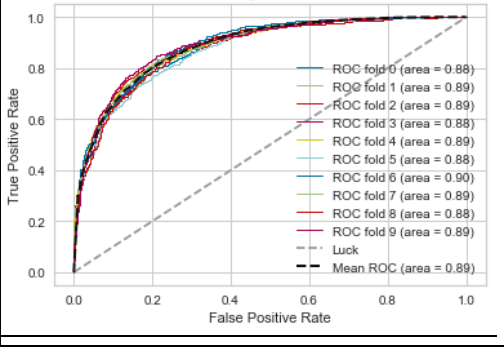


נספח ח – תוצאות ראשוניות עבור רגרסייה לוגיסטית לפני חיפוש גרד



נספח ט – חיפוש מספר עצים ראשוני לפי accuracy



מודל/קידו	OneHot decoding	Target decoding
1		
Random Forest		
ANN		
Logistic regression		

נספח כ – מטריצת הקונפיון לפי סט הולידציה

