דוח מסכם קבוצה 36-  
פרויקט למידת מכונה

**קורס : למידת מכונה**

**מגישים: אלעד שניאור - 318299120 , איתי אלפסה -208677617**

**תאריך 11/08/2024**

**תוכן עניינים**

[חלק ראשון – אקספלורציה](#_heading=h.tn78iy2zrb8e) 3-4

חלק שני- עיבוד מקדים 5

חלק שלישי-הרצת מודלים 6-7

חלק רביעי-הערכת מודלים 7

חלק חמישי-ביצוע פרדיקציה 7

חלק שישי - שימוש בכלים שלא נלמדו בקורס 7

חלוקת אחריות 8

נספחים 9-19

# 

**חלק ראשון - אקספלורציה :**השלב הראשון שעשינו היה להסתכל על כל הדאטה ולהבין עם איזה פיטצ’רים אנחנו הולכים להתמודד. ניסינו להבין מה הערכים האפשריים של כל פיטצ’ר, מה הסוג שלו, ואיך הוא מתפלג.

התפלגות הפיטצ’רים קטגוריאליים :

ע"פ גרפי ההתפלגות של הפיטצ’רים קטגוריאליים (ראה איור מספר 14) רצינו לראות בעיקר האם קיימים ערכים מסוימים מהם יש לנו כמות נתונים קטנה מאוד דבר העלול להשפיע על טיב האמידה שלנו של הערכים האלו באותם פיטצ'רים, כמות הקבוצות להם מתחלקות ואת הפרופורציה בחלוקה של הנתונים. בשלב זה גם בדקנו באינטרנט את ההתפלגות של פיטצ'רים מסוימים בעולם כדי להבין האם הדאטה שלנו אמינה ונותנת מדגם תואם לאוכלוסייה. לדוגמא אחוז מוגבלי עבודה( 16% בעולם), אחוז בעלי הפרעות אישיותיות באוכלוסיה וכו'... ,  
עשינו זאת כדי לקחת את זה בחשבון ולדעת כיצד נתמודד בהמשך עם מילויי הערכים החסרים .

שמנו לב כי הפיטצ'רים: עבד בעבר, קבוצת גיל, מוגבלות, האם מפתח ובעל בעיות מנטליות הינם פיטצ'רים המתחלקים לשתי קבוצות בלבד. ולכן רצינו לראות את היחס ביניהם כדי להיעזר בזה במילוי הערכים החסרים.   
לעומת זאת שמנו לב כי הפיטצ'רים: מדינות וניסיון תעסוקתי הינם פיטצ'רים אשר מחולקים לקבוצות רבות וכך גם הפרופורציה שלהם, ולכן הבנו כי נדרש להתמודד איתם בהמשך בצורה פרטנית.   
עבור ניסיון לימודים ,ומין ראינו כי מדובר בקבוצות ברורות אשר קיים בפרופורציה ויחס בין קבוצות אלו ולכן עלינו לשמר את זה על מנת לא לפגוע בדאטה. וכך עשינו בהמשך עם ההתמודדות איתם.  
פיטצ'ר C שאין לנו כותרת עבורו ולכן איננו יודעים להסביר מה טיבו. עבורו ראינו כי גם במקרה שלו קיימים יחסים בין הקבוצות אותם הוא מחלק ולכן נתמודד איתו בדרך זהה כי נצטרך לשמר את יחס זה בהמשך.   
ערכים חסרים:

לאחר ההתבוננות בפיטצ’רים הקטגוריאליים ניגשנו להסתכל על כמות הערכים החסרים בדאטה ,ושמנו לב לכמות החוסרים המשמעותי אשר קיים עבור פיטצ'ר ניסיון תעסוקתי (26% ערכים ריקים).

התפלגות עמודת המטרה:

בחרנו לבדוק כיצד מתפלגת עמודת המטרה הלייבל (ראה איור 13). בחרנו לבדוק את זה על מנת שנוכל להבין אם המדגם שקיבלנו אמין ומתאר בצורה טוב את מה שהולכים לחזות. לדוגמא אם היינו רואים שבסט הנתונים יש לנו רק 20% שאינם התקבלו לעבודה (או להפך) הינו מבינים שאין לנו מדגם מספיק מייצג עבור המודלים שנשתמש בהמשך. במקרה הזה הדטה מחולקת בצורה אמינה וטובה היות וקיים פיזור מאוזן בין הדוגמאות. בנוסף ראינו שאם היינו לוקחים מודל "טיפש" שאומר לא עבור כל מועמד שהיה מגיע היינו מקבלים auc : 0.536 .

התפלגות פיטצ'רים נומרים :   
 כדי לאמוד את התפלגות הפיטצ’רים הנומרים הצגנו גרפי התפלגות ובנוסף הוצאנו טבלת נתונים סטטיסטים אשר מציגה לנו מספרים סטטיסטים של כל פיטצ’ר .כעת נתאר את התפלגות כל פיטצ'ר נומרי

* פיטצ'ר שנות ניסיון – לפי גרף התפלגות של הפיטצ'ר (ראה איור 1) אנחנו רואים שהתפלגות הנתון אינה נורמלית. קיימת הטיה גדולה של הנתון לערכים קטנים מהממוצע ובנוסף קיים זנב ארוך לערכים הגבוהים. ע"פ הגרף התחתון ניתן לראות כי הפיטצ'ר מתפלג בצורה זהה עבור אלו שהתקבלו ולא התתקבלו לעבודה. באמצעות גרף ה – box plot (ראה איור 11) יכולנו לראות שיחסית קיימים מספר ערכים מועט שהינם מעל האחוזון העליון ולכן כמות החריגים הצפויה נמוכה.
* פיטצ'ר A - לפי גרף התפלגות של הפיטצ'ר (ראה איור 2) אנחנו רואים שהתפלגות הנתון הינה נורמלית מובהקת ההתפלגות הינה די סימטרית וללא הטיה. באמצעות גרף ה – box plot (ראה איור 8) אנו מסיקים כי קיימת בנתון כמות ערכים גבוהה של חריגים בערכים הנמוכים והגבוהים.
* פיטצ'ר B - לפי גרף התפלגות של הפיטצ'ר (ראה איור 3) אנחנו רואים שהתפלגות הנתון אינה נורמלית. קיימת הטיה גדולה של הנתון לערכים קטנים מהממוצע ובנוסף קיים זנב ארוך לערכים הגבוהים. באמצעות גרף ה – box plot (ראה איור 9) יכולנו לראות שיחסית קיימים מספר ערכים מעטים שהינם מעל האחוזון העליון ולכן כמות החריגים הצפויה נמוכה.
* פיטצ'ר D- לפי גרף התפלגות של הפיטצ'ר (ראה איור 4) אנחנו רואים שהתפלגות הנתון הינה נורמלית מובהקת ההתפלגות הינה סימטרית וללא הטיה. באמצעות גרף ה – box plot (ראה איור 10) אנו מסיקים כי קיימת בנתון ערכים חריגים בכמות לא גבוהה אשר גם היא סימטרית בערכים הגבוהים והנמוכים בהתפלגות.
* פיטצ'ר משכורת קודמת- לפי גרף התפלגות של הפיטצ'ר (ראה איור 5) אנחנו רואים שהתפלגות הנתון אינה נורמלית. קיימת הטיה גדולה של הנתון לערכים קטנים מהממוצע ובנוסף קיים זנב ארוך לערכים הגבוהים. ע"פ הגרף התחתון ניתן לראות כי הפיטצ'ר מתפלג בצורה זהה עבור אלו שהתקבלו ולא התתקבלו לעבודה. באמצעות גרף ה – box plot (ראה איור 7) יכולנו לראות שקיים ריכוז ערכים חריגים מעל האחוזון העליון.

שימוש בגרפים אלו עזר לנו להבין בעיקר את דרך ההתמודדות הטובה ביותר עבור שלוש פעולות מרכזיות : מילוי ערכים חסרים, שיטת נרמול וטיפול בערכים חריגים.

לאחר התבוננות בכלל הפיטצ’רים, והתמודדות איתם כפי שיתואר בהמשך בחלק השני - עיבוד מקדים . הצגנו טבלת קורלציה אשר מציגה את ההתנהגות הקורלטיבית בין הפיטצ’רים.

טבלת קורלציה :

מטבלת הקורלציה בשילוב עם ויזואליזציה של heatmap של כלל הפיטצ'רים (ראה טבלה 6) הצלחנו לקבל הערכה טובה של קורלטיביות בין הפיטצ'רים השונים. התבדנו לראות שבין מספר פיטצ'רים שהערכנו שתהיה קורלטיביות מובהקת לא היו כגון : פיטצ'ר שכר קודם אל מול שנות ניסיון (0.37) ופיטצ'ר רמת השכלה אל מול שכר קודם (0.01). מול זאת מצאנו שתי קורלציות די חזקות שבאמצעותן התחלנו בתהליך הורדת המימדיות. מצאנו שבין פיטצ'ר 'B' ופיטצ'ר שנות ניסיון קיימת קורלציה גבוהה של 0.92 . ובנוסף ל html ו לcss יש התאמה לינארית מלאה.  
לכן בחלק הבא של הפרויקט נוריד את הפיטצ'רים האלו כחלק מפעולות אקטיביות לצמצום מימדים.

התפלגות וקורלציה לעמודת הלייבל :

מעמודת הלייבל אנחנו יכולים לאמוד עד כמה קובץ הנתונים שניתן לנו מאוזן, האם יש לנו איזשהם פרמטרים עם קשר משמעותי לחיזוי הלייבל או להפך. מהתפלגות הפרמטרים ביחס לעמודת הלייבל (ראה איור 12) אנו רואים כי הפרמטר בעל הקורלציה הגבוהה ביותר לעמוד הלייבל היא עמודה 'D' עם ערך של 0.41. ועמדוה 'node' עמודה שאנחנו יצרנו על ידי החלוקה ורואים קשר של 0.35 .איננו מזהים עמודה עם קורלציה ישירה לערך הlabel שנרצה להתייחס אליה באופן מיוחד. אנו רואים כ-16 עמודות עם ערך קורלציה ללייבל שנמוך מ-0.1 . נכון לכרגע לא נרצה לבטל עמודות אלו אך נוכל לקחת את זה בחשבון בחשיבותנו לעמודות אלו.

מסקנות עיקריות מחלק זה :

* פיטצ’רים נומרים שראינו מתפלגים בצורה קרובה לנורמלית- מילאנו ערכים חסרים בממוצע, והורדת ערכים חריגים בעזר סטיות תקן .
* פיטצ’רים נומרים שראינו שלא מתפלגים בצורה נורמלית - מילאנו ערכים חסרים בחציון, והורדת ערכים חריגים בהתאם לmin -max -scaling .
* פיטצ’רים קטגורילים - מילאנו הערך השכיח ביותר על מנת לשמור על המאזן והפרופורציה שראינו והסקנו בחלק זה ,היה מספר ניסיונות למלא ערכים חסרים לפי הפרופורציה של המדגם אך ראינו שזה לא שיפר לנו את הדיוק והשארנו את מילוי החסרים בשכיח.
* בעזרת טבלה הקורלציה ראינו שיש שני פיטצ’רים בעלי קורלציה גבוהה שיהיה ניתן להשאיר פיטצ’ר אחד על מנת להקטין את הממדיות ולכן אנו עושים זאת בהמשך.
* לפיטצ’ר מדינות וניסיון תעסוקתי נדרש התמודדות אחרת, ולכן התחלנו לחשוב על דרכים יצירתיות כיצד להתמודד איתם, יפורט בהמשך.

חלק שני- עיבוד מקדים :

חלק זה הוא חלק שלאורך כל הפרויקט היה תחת שינויים ועבודה מתמדת בניסיון לשפר את התוצאות. בחלק זה בעיקר גרם לנו לנסות לחשוב בדרכים יצירתיות תוך כדי שימוש בכלים מגוונים ויישומם בכל מיני דרכים. בחלק זה התחלנו בלשנות את הפיטצ’רים הקטגוריאליים לערכים מספריים. הבנו שיש לנו כמה סוגים ולכל אחד התייחסנו בצורה שונה. למדנו על שתי דרכים חדשות לטיפול במידע קטגוריאלי והמרתו לערכים מספריים ויישמנו את שיטת target encoding עליה הרחבנו בחלק 6. ראינו שהשימוש בה הגיוני ונכון לסוג דאטה שלנו.  
לאחר מכן עברנו לטפל בערכים חסרים. זיהינו הרבה ערכים חסרים ובשלב זה היה חשיבה רבה על הערכים אותם נמלא מתוך ההבנה כי זה ישפיע על המדגם ובסופו של דבר על המודל. לערכים הקטגוריאליים בחרנו למלא על פי הרוב ולערכים הנומריים מילאנו על פי ההתפלגות (ממוצע עבור נורמלי וחציון עבור לא נורמלי). את הפיטצ’ר ניסיון תעסוקתי ( אשר כולל יכולות טכנולוגיות ) אשר היה לו 14000 ערכים חסרים, בחרנו לנכון למלא רק בערכים אשר יש בפיטצ’ר האם מפתח -ערך אמת ובכל עמודה של היכולות החדשות שמנו את הערך הממוצע על מנת לא להשפיע על המדגם בצורה קיצונית ולנסות להפעיל חשיבה הגיונית שאם יש לנו ידע שהמועמד הינו מפתח אזי כנראה שיש לו ידע ממוצע .   
לאחר מכן עברנו לטפל בערכים חריגים. מתוך ההבנה כי אין לנו הרבה דאטה , בחרנו למלא בשדות אשר היה ערך חריג, את הערך הממוצע או החציון בהתאם להתפלגות, במקום למחוק/להתעלם מכל הדוגמה הקיימת .   
בשלב האחרון החלטנו לנרמל את חלק מהפיטצ'רים. ההחלטה הזאת נובעת משתי סיבות עיקריות, הראשונה על מנת להשתמש במודלים לינארים וזה רלוונטי לבצע תהליך נרמול עבורם. סיבה נוספת היא שקיבלנו יותר דיוק בהרצת שאר המודלים ולכן בחרנו להשתמש בסט הנתונים המנורמל. כמו כן רצינו לנרמל את הפיטצ'רים שהוספנו כ- 11 פיטצ’רים לפי הקבוצות המתאימות בקטגוריה (ראה טבלה 2). אך לאחר שראינו כי זה עדין לא מביא לנו דיוק מספק וכי יתכן שפעולה זו לא נותנת משמעות לתכונות שנמצאות בתוך כל קבוצה, השתמשנו בשיטת tfidf (מפורט בקוד).  
בשלב הבא בפרויקט היה ניסיון לצמצם את המימדיות- לשם כך השלב הראשון היה לבדוק את הקורלציה בין הפיטצ’רים, וכך בחרנו לעשות צעד אקטיבי ולהוריד 2 פיטצ’רים שהיו בעלי ערך קורלציה ביניהם של מעל 0.9.  
טכניקה נוספת להורדת מימדים בה השתמשנו הינה PCA. רצינו להשתמש בה כדי לנסות להשתמש במודל שמסוגל לשמור על פיזור הdata במרחב בצורה טובה ובכל זאת להצליח להוריד מימדים. נוסיף לכך שגם מבחינה פרקטית היות ולא התבקשנו במסגרת המשימה להסביר את דרך הפעולה של המודל לא היה לנו אכפת לקבל פיטצ'רים חדשים שאינם בעלי הסבר הגיוני. ראינו כי דבר זה אינו משפר את המודל הנבחר ומוריד את הדיוק בצורה קיצונית ולכן זה בהערה בקוד ובחרנו לא להשתמש בזה בשלב האימון וגם בשלב הבחינה של המודל.   
שלב זה היה שלב משמעותי בפרויקט, לאורך הדרך היו ניסיונות רבים לטפל בנתונים כדי לנסות לקבל תוצאות מיטביות.   
\* מודלים ותהליכים שניסינו אבל לא נכנסו לקוד כי ראינו שלא תרמו להצלחת המודל :   
-נסינו להשתמש בשיטת Naïve Base על ה -11 פיטצ’רים שהוספנו מתוך עמודת ניסיון תעסוקתי ולנסות להמיר אותם להסתברות אך זה לא תרם למודל.  
-ניסינו להשתמש בשיטת ICA לנסות לצמצם את הממדיות בנוסף לשיטת PCA וגם זה לא תרם.  
-ניסינו להשתמש בשיטת בחירת פיטצ'רים בבחירה אחורנית וקיבלנו כי זמן ההרצה מאוד גבוהה ולכן בחרנו לוותר עליו ולא להמשיך איתו בשיטת הורדת המימדים.

לבסוף הפרמטרים, והמודלים אשר נמצאים בפרויקט הם המדויקים ביותר שהצלחנו ללמד את המודל על מנת להוציא דיוק מרבי.

חלק 3- הרצת המודלים:

הסבר משמעות ההיפר פרמטרים עבור המודל הנבחר :

ראשית נציין, כפי שהראנו באמצעות גרף במחברת הפרויקט, כי יישום המודלים ללא אף כיוון והגדרה של ההיפר פרמטרים הביאו תוצאות סבירות עבור כלל המודלים (ערכי auc בין 0.7-0.8). השוואה זו ביצענו כדי להדגיש את משמעות ההיפר פרמטרים עבור כלל המודלים ואת חשיבות הדיוק שלהם. השתמשנו בשיטת Gridisearch שמאפשרת לבצע בדיקה מעמיקה לכל מודל על סט היפר פרמטרים שאנו מגדרים, ומציאת סט הפרמטרים הטוב ביותר תחת בדיקה של k-fold-cv עם ערך של 5 פולדים. נשים לב כי בפרויקט השארנו רק את הערכים הכי טובים שיצאו לנו כדי לצמצם את זמן הריצה.

במודל מסוג RFC אנחנו צריכים להתבונן מה ההשפעה של שונות והטיה על עצי החלטה, היות ומודל זה מבוסס על עצי החלטה. נשים לב שלעומת עצי החלטה יש במודל RFC חסינות הרבה יותר גבוהה להגעה להתאמה יתר לסט האימון ונוכל לקחת את זה בחשבון. הרעיון באופן כללי שחוזר ברוב ההיפר - פרמטרים היא השפעת ההיפר-פרמטר על עומק העצים. העמקת העץ מובילה בצורה ישירה להשפעה על שונות והטיה של המודל. נסביר כיצד זה מתרחש עבור כל היפר-פרמטר.

1. בחירת בשיטת בדיקת ביצועים של כל חלוקה של העץ ע"י entropy מובילה לבדיקה יותר מאוזנת ולכן משפיעה על ביצועים מאוזנים של העץ מבחינת שונות והטיה.

2. עומק העץ - פרמטר עומק העץ בעץ החלטה בודד הוא ערך משמעותי בהיבט שונות והטיה של מודל. ככל העץ יותר עמוק זה מוביל להתאמה יותר גדולה של המודל לסט הנתונים ובכך הגדלת השונות של המודל. אם נבחר להגביל את עומק העץ נגיע למצב שבו הטית המודל תגדל כיוון שהוא לא יצליח לנתח את המידע ואת התפלגות הנתונים. לעומת זאת בשיטה מבוססת עצי החלטה עם ריבוי עצים ככל שנגדיל את עומק העץ עקרונית נמקד את המודל יותר על סט הנתונים ונקבל תוצאות טובות יותר. כדי להימנע מהגעה להתאמת יתר לסט הנתונים אנחנו נמנענו מלתת עומק של יותר מ-50.

3. מספר דוגמאות מינימלי בעלה - נתון זה משפיע גם על השונות וההטיה של המודל. באם נבחר כמות עלים גדולה אנחנו נמנע מהעץ מלהעמיק ולנתח את המידע לעומק ובכך נגדיל את ההטיה אך אם נתן ערכים מאוד קטנים בכל עלה נגדיל בהכרח את עומק העץ כך שימשיך לבצע עוד ועוד פיצולים ובכך נגדיל את השונות שלו והמרוכבות.

4. מספר דוגמאות מינימליות לפיצול - כמו בהיפר פרמטר הקודם גם פרמטר זה ישפיע על כמות הפיצולים שכל עץ מבצע. ככל שנגדיל את המספר ,נמנע מהעץ להעמיק ולבצע חלוקות עבור מספר דוגמאות קטן ובכך הגדלת הערך תמנע את העמקתו ובכך תגדיל את הטיית המודל בעוד הקטנת המספר לערכים קטנים תגרום להעמקה של העץ והתאמה יתרה לסט הנתונים ובכך הגדלת השונות של המודל.

5. מספר עצי החלטה במודל - היפר פרמטר זה משפיע בצורה חיובית בהקטנת השונות של המודל ככל שכמות העצים תגדל. ולכן הגדלת כמות העצים תועיל לנו בהצלחת חיזוי יותר ויותר טובה בעוד השונות לא גדלה ובנקודה מסוימת כנראה תתקבע בערך מסוים.

חשיבות הפיטצ’רים על פי ניתוח של המודל הנבחר :

כפי שניתן לראות בגרף המתאר את חשיבות כל פיטצ'ר להצלחת המודל RFC (ראה איור 15) אנחנו יכולים לראות כי הפיטצ'רים המשמעותיים ביותר להצלחת המודל היו :

* 'D' ,node, 'typerscript','A' , שכר קודם, שנות ניסיון, 'server'

מבחינה הגיונית אנו יכולים להבין מדוע שכר עבודה קודם ושנות ניסיון רלוונטים לקבלה לעבודה הנוכחית. שנות ניסיון מצביעות על ניסיון תעסוקתי שיש למועמד ולכן כמובן משפיע על רלוונטיות למשרה. השכר הקודם מקבל מקום גבוהה ממנו בצורה פחות ברורה אך כנראה מצביע על הקשר לרמת ההתאמה של המועמד למשרה.   
ישנם שני פיטצ'רים במקום הראשון והרביעי שאנחנו לא יודעים את טיבם בדיוק אך אנו יודעים שהם מתפלגים נורמלית ויכול להיות שהתפלגותם משפיעה על הרלוונטיות שלהם.  
בנוסף שלושת הפיטצ'רים שיצרנו מתוך עמודת הניסיון התעסוקתי תרמו רבות למודל העץ ולכן זה מצביע שככל הנראה המשרה מתאפיינת ביכולות פיתוח ברשת ( שרת, פיתוח אתרים, ועיצוב אתרים) וזה עוזר לנו להצביע על אופי המשרה שלפיה אנחנו מסווגים.

חלק רביעי - הערכת המודלים:

בחלק זה היינו צריכים להעריך את המודל בכמה דרכים. הדרך הראשונה היית Matrix Confusion. הערכה זאת עזרה לנו להבין איפה אנחנו טועים יותר ומהן הטעויות שלנו במודל ,האם אנחנו טועים יותר ואומרים שאדם התקבל לעבודה והוא לא התקבל או הפוך. מפני שבמקרה שלנו אין משמעות עסקית( לדוגמא שהבוס אומר שיותר חשוב לנו לא לקבל מישהו שלא מתאים מאשר לא לקבל מישהו מתאים) אז לא היה ניסיון להטות את אחד הצדדים ,והיה נסייון רק לשפר שנעשה כמה שפחות טעויות באופן כללי . לאחר מכן הצגנו גרפים של כל מודל ובדקנו את ROC שלהם והצגנו בגרף באדום את המודל הטוב ביותר בעזרת שיטת - k-fold-cv שבעזרתה בחרנו את המודל.   
ולבסוף בדקנו האם המודל שלנו הוא Overfitted פירוט על ההחלטה מופיע במחברת הקוד מתחת לחלק זה.

חלק חמישי-ביצוע פרדיקציה:

לאחר העיבוד המקדים, בחירת המודלים והערכת המודלים בחרנו במודל**RFC . בחרנו בו כי** הוא נתן לנו דיוק מרבי, שגיאה מודל נמוכה, ופרדיקציה מדויקת. נוסף לכך רצינו להשתמש במודל זה היות ויכולנו ממנו לקבל את חשיבות הפרמטרים בצורה יותר מדויקת ולהבין את דרך העבודה שלו מאחורי הקלעים בצורה טובה יותר.

בשלב האחרון באמצעות שימוש במתודת ה- pipeline בחרנו רק בפעולות הרלוונטיות שביצענו גם על סט האימון לעבור גם על test שלנו ולבסוף לתת הערכת הסתברותית לכל אדם להתקבל לעבודה .

חלק 6 – שימוש בכלים שלא נלמדו בקורס:  
בשלב עיבוד סט המידע התמודדנו עם בעיה של התייחסות לפיטצ'רים קטגוריאליים שאינם בינאריים. תחילה ביצענו המרה של הערכים הקטגוריאליים שהופיעו כמחרוזות לערכים נומרים בין 0 למספר הערכים האפשריים ללא התחשבות בערך עצמו שמוחלף. לדוגמא, פיטצ'ר רמת ההשכלה בו הופיעו 5 אופציות שונות של רמת השכלה אפשרית המרנו כל אופציה בערך נומרי (תואר ראשון הוחלף ב- 0 , דוקטורט הוחלף ב- 1 וכו'...).  
פעולה זו הביאה לנו תוצאות auc נמוכות בשימוש במודל מסוג רגרסיה לוגיסטית ולכן התחלנו לחפש דרכים נוספות לביצוע encoding לעמודות אלו. לבסוף בחרנו להשתמש בשיטת Target Encoding שעליה נתאר בחלק זה.   
**Target Encoding** -זוהי שיטת המרת מידע קטגוריאלי למידע נומרי תוך התחשבות בהסתברות הופעת כל ערך אפשרי ביחס לעמודת המטרה (לייבל). השיטה עובדת בצורה הפשוטה ביותר בכך שכל ערך קטגוריאלי יוחלף בהסתברות המותנית – מה ההסתברות לקבל ערך לייבל -1 (משמעותו להתקבל לעבודה) בתנאי שמופיע הערך הקטגוריאלי.   
באמצעות שיטה זו כל ערך קטגוריאלי יוחלף בערך נומרי בין 0 ל-1 (מנורמל) שערכו בעל משמעות (שלא כמו בשיטה הקודמת שביצענו) ובהתאם לחשיבותו להתאמה לייבל. שיטה זו דומה ברעיון המימוש שלה לשיטות של bayes ולכן שיטה זו גם נקראת Bayesian mean method. מעבר לכך ניתן לבצע, כפי שיישמנו, הוספת היפרפרמטר שמהווה "משקולת" באמצעותה משלבים לתוך חישוב ההסתברות את ההסתברות של ערך הלייבל להיות 1 בכלל הדתה. תוספת זו נקראת smooth. מטרתה העיקרית של תוספת זו היא להימנע ממצבים בו ערך קטגוריאלי שמופיע בצורה יחסית נמוכה בקובץ המידע שלנו יקבל ערך הסתבורתי נמוך מאוד. ככל שכמות המידע שלנו נמוכה יותר נרצה להתאים ערך משקולת גדול יותר.   
בחרנו בשיטה זו ממספר סיבות. הראשונה: למנוע הגעה למימדיות יתר בשימוש בשיטה הנפוצה של one hot endcoding (פתיחת פיטצ'ר חדש לכל ערך קטגוריאלי והפיכתו לבינארי).והשנייה: אפשרנו למודלים מבוססי משקולות לקבל ערך שמתאים את עצמו להתאמתו ללייבל ובכך מאפשר למודל להתאים משקולת שערך התא הקטגוריאלי מוביל אותה לערך יותר גדול/ קטן בהתאם להסתברות.

נספחים

1. אחריות בפרויקט :

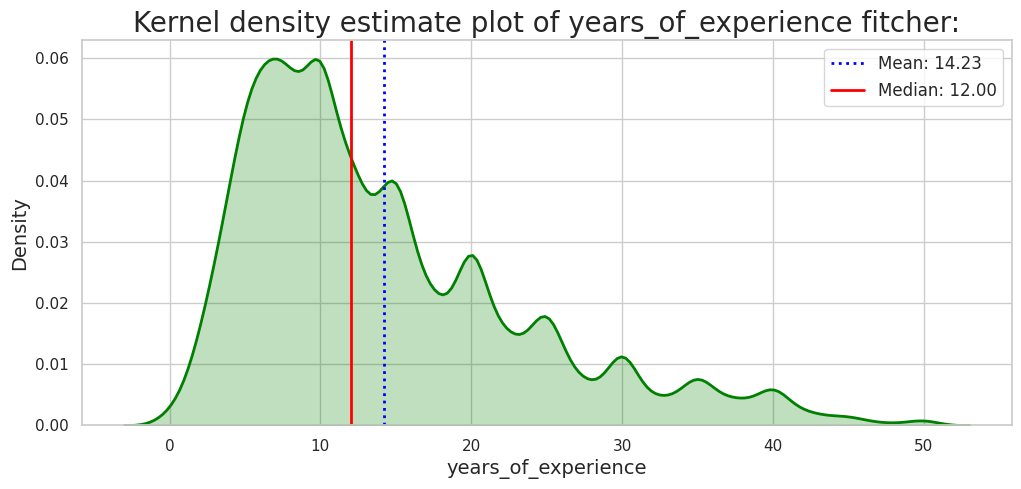
במהלך הפרויקט השתדלנו כמה שיותר לשמור על עבודה משותפת כדי לאפשר גם חשיבה משותפת ופתירת רעיונות בצורה יצירתית עם שני מוחות במקום אחד. ובנוסף כך גם לאפשר למידה של שנינו מהעבודה על הפרויקט.

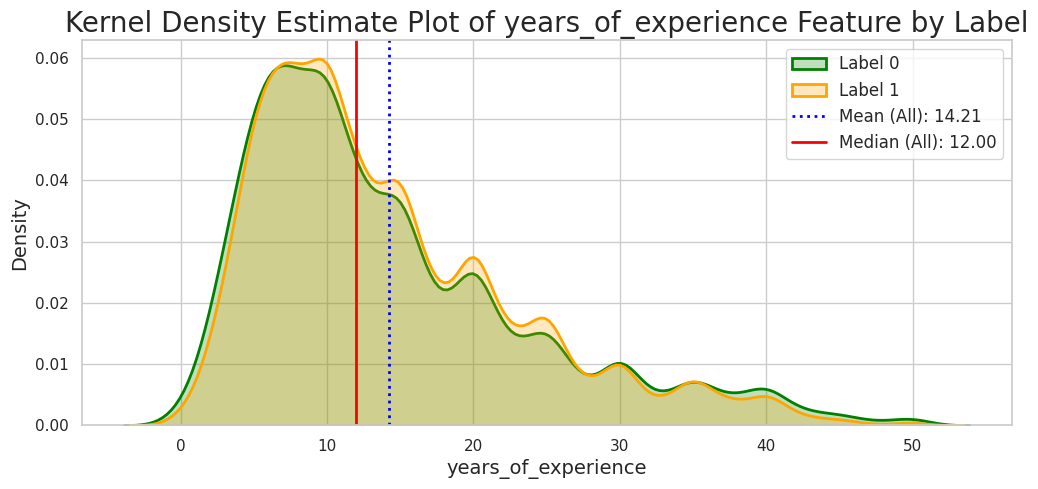
אלעד – עבדתי תחילה על חלק מספר 1 ונתתי הסבר לחלק הנומרי , חלק מספר 2 עבדנו בצורה משותפת לכל הדרך ועשינו אותו יחד, אך כחלק מההתמודדות עם הפיטצ'רים הקטגוריאליים אני התמודדתי עם הפיטצ'ר של המדינות ולאחר מכן על התוספת של tfidf בנוסף עם נרמול הדאטה. בחלק מספר 3 פיצלנו כך שאני עבדתי על יישום מודל KNN וRFC ולאחר מכן בשלב האחרון אני עיצבתי את חלק זה יחד עם התייחסות לשאלות שנשאלו . לבסוף אני ביצעתי את חלק מספר 6 ויישום שלו.

איתי – עבדתי בחלק 1 על ההסבר החלק הקטגוריאלי , חלק מספר 2 כמו שאלעד רשם לאורך הדרך עבדנו עליו יחד, אך כחלק מההתמודדות עם הפיטצ'ארים הקטגוריאליים אני התמודדתי עם הפיטצ'ר של ניסיון תעסוקתי ,מילוי ערכים חסרים, ערכים חריגים, בדיקת הקטנות מימדיות .בחלק מספר 3 אני עבדתי על יישום LGR ועל MLP, ולבסוף אני ביצעתי את חלק מספר 5 בדיקת המודלים שבחרנו והערכת המודלים .

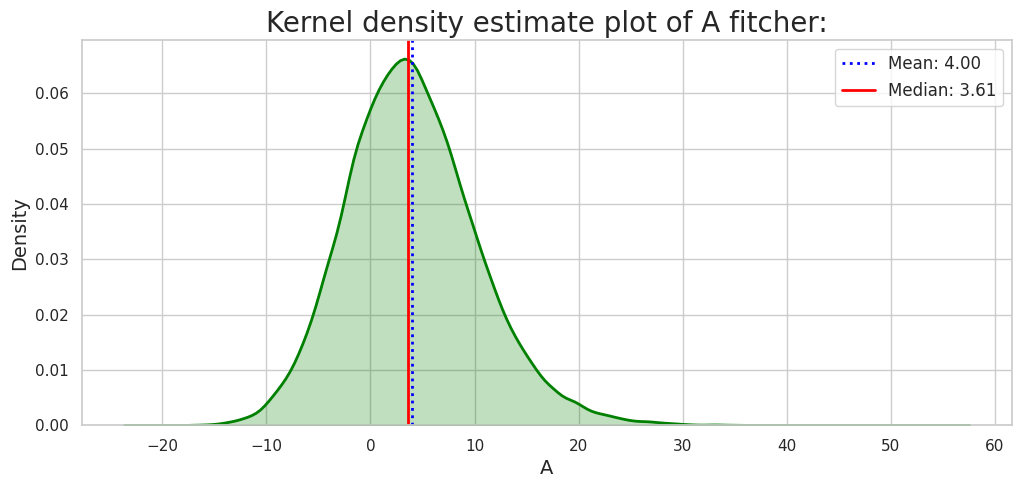
ברמה האישית מעבר לכך נהנינו לעבוד בצוות על הפרויקט, הרגשו שכל החומר התיאורטי שלמדנו לאורך הקורס משתלב בפרויקט . התמודדות עם הדאטה וכיווני המחשבה הגבירו בנו את הרצון לשפר את המודל ככל שניתן על מנת להשיג AUC גבוהה דבר אשר גרם לנו לחקור לעומק כל מודל ומודל , שיטות נוספות להתמודדות עם דאטה , ולהיכנס יותר לעומק הדברים בתעשייה ובתחום למידת מכונה .

1. גרפים וטבלאות :



**

*1 - גרפי התפלגות פיטצ'ר שנות ניסיון*



*2 גרף התפלגות פיטצ'ר 'A'*

*תמונה שמכילה טקסט, עלילה, תרשים, קו

התיאור נוצר באופן אוטומטי3 גרף התפלגות פיטצ'ר 'B'*

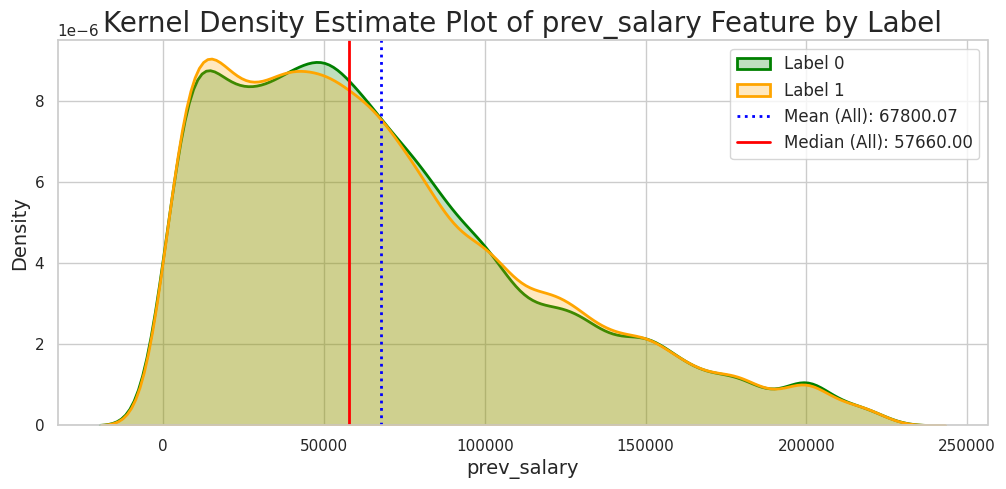
תמונה שמכילה טקסט, עלילה, תרשים, קו

התיאור נוצר באופן אוטומטי

*4גרף התפלגות פיטצ'ר 'D'*

תמונה שמכילה טקסט, תרשים, עלילה, קו

התיאור נוצר באופן אוטומטי

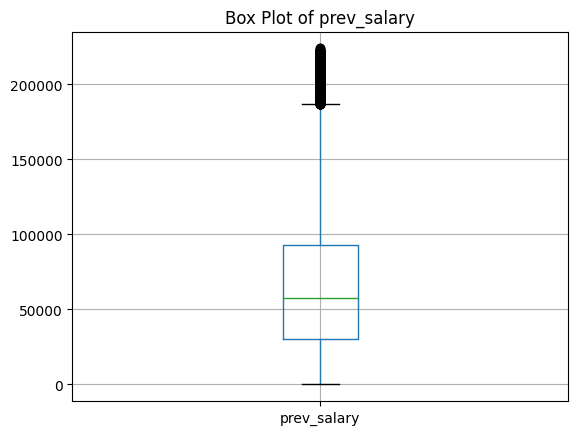
**

*5 - גרפי התפלגות פיטצ'ר שכר קודם*

תמונה שמכילה צילום מסך, אדום, טקסט, דפוס

התיאור נוצר באופן אוטומטי

*6 טבלת heatmap עם ערכי קורלציה של הפיט'צרים*



*7 box plot של פיטצ'ר שכר קודם*

תמונה שמכילה טקסט, צילום מסך, קו, תרשים

התיאור נוצר באופן אוטומטי

*8 box plot של פיטצ'ר 'A'*

תמונה שמכילה טקסט, צילום מסך, קו, עלילה

התיאור נוצר באופן אוטומטי

*9 box plot של פיטצ'ר 'B'*

תמונה שמכילה טקסט, תרשים, צילום מסך, קו

התיאור נוצר באופן אוטומטי

*10 box plot של פיטצ'ר 'D'*

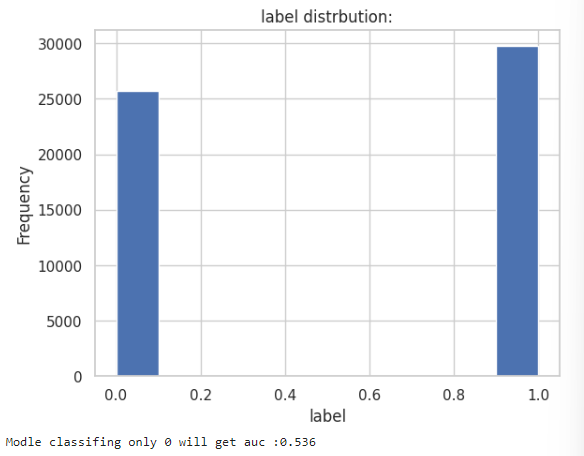


*11 box plot של פיטצ'ר שנות ניסיון*

תמונה שמכילה טקסט, תפריט, צילום מסך

התיאור נוצר באופן אוטומטי

*12 קורלציית פיטצ'רים ביחס לעמודת label*



*13 התפלגות פיטצ'ר label*

תמונה שמכילה טקסט, עיגול, צילום מסך, תרשים

התיאור נוצר באופן אוטומטי

*14 התפלגות פיטצ'רים קטגוריאלים*



*15 חשיבות פיטצ'רים במודל RFC שבחרנו*



*טבלה 1 - נתונים סטטיסטים על פיטצ'רים נומרים*

|  |  |
| --- | --- |
| **Category** | **Technology** |
| **Web Development** |  |
| **Front-End** | HTML/CSS, JavaScript, TypeScript, React.js, Vue.js, Angular, Angular.js, jQuery, Next.js, Gatsby, Svelte |
| **Back-End** | Node.js, Express, PHP, ASP.NET Core, ASP.NET, Django, Flask, Ruby on Rails, Spring, Laravel, FastAPI, Symfony, Phoenix, Play Framework |
| **Programming Languages** | Python, Java, C#, C++, C, Go, Rust, Ruby, Kotlin, Swift, Dart, R, Perl, Groovy, Scala, Objective-C, Delphi, Elixir, Haskell, Lua, Julia, F#, Deno, LISP, Erlang, Crystal, OCaml, Fortran, COBOL, APL |
| **Databases** |  |
| **SQL** | SQL, PostgreSQL, MySQL, Microsoft SQL Server, SQLite, MariaDB, Oracle, IBM DB2 |
| **NoSQL** | MongoDB, Redis, DynamoDB, Cassandra, Couchbase, CouchDB, Neo4j, Elasticsearch, Firebase Realtime Database, Cloud Firestore |
| **Cloud & DevOps** |  |
| **Platforms** | AWS, Microsoft Azure, Google Cloud Platform, Heroku, DigitalOcean, IBM Cloud or Watson, Oracle Cloud Infrastructure, VMware, Linode, OVH |
| **Tools & Technologies** | Docker, Kubernetes, Terraform, Ansible, Puppet, Chef, Pulumi, OpenStack |
| **Version Control & Package Managers** | Git, npm, Yarn, Homebrew |
| **Mobile Development** | Xamarin, Unity 3D, Unreal Engine |
| **Scripting & Automation** | Bash/Shell, PowerShell, VBA, MATLAB, SAS |
| **Frameworks & Libraries** | Spring, Ruby on Rails, Django, Flask, Laravel, Symfony, FastAPI, Phoenix, Play Framework, ASP.NET, ASP.NET Core, Blazor, Nuxt.js |
| **Specialized Languages & Tools** | Solidity (Blockchain), Rust (Systems Programming), Crystal (Fast Ruby-like language), Fortran (Scientific Computing), COBOL (Business, finance, administrative), APL (Array processing) |
| **Game Development** | Unity 3D, Unreal Engine |
| **Other Tools & Service** | Firebase, Managed Hosting, Colocation |

*טבלה 2 - חלוקת עמודת ניסיון תעסוקתי לקבוצות*