אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

# מערכות לומדות וכריית נתונים
# 364-2-1651

## פרוייקט – חיזוי שכר אצל עובדי IBM


***מגישים***

גלית פליקשטיין 204382949

יובל לוי 305730897


***תאריך***

17/07/21

***מרצה***

בעז לרנר

## תקציר

ע"י שימוש בכלי לימוד המכונה הנלמדים בקורס, ניתוחים גרפיים כמותיים ואיכותיים נוכל להבין טוב יותר את ההשפעה של גורמים שונים על השכר של עובד בחברת IBM . (זאת בהנחה שיש דמיון בדרישות הגיוס ובשכר שלהם בשאר השווקים בעולם).

פתרון הבעיה עשוי לעזור הן בצד של המעסיק והן בצד של העובד, ומכאן הערך המוסף שלו.

מבחינת המעסיק,לחיזוי עלות ההעסקה השפעה ישירה על הדוחות הכספיים של החברה, וכך השיקולים האם לגייס עובד חדש/ לקדם עובד קיים מגובי נתונים המסייעים בבחינת הכדאיות הכלכלית.

מבחינת העובד, לחיזוי השכר הצפוי השפעות בטווח הקצר של התאמת מקום עבודה זה או אחר לשיקוליו, אך גם השפעות בטווח הרחוק יותר של בחירת כיווני קריירה חדשים/ נוספים.

פרוייקט זה נבנה על בסיס שימוש במטולוגיית CRISP-DM לכל שלביו.

תחילה הובן הצורך בפיתוח כלי הניבוי מתוך הבנת הצורך של העולם העסקי. לאחר מכן, הוכנו הנתונים עד כדי מיצוי טוב ומייצג של סט הנתונים את הבעיה שנרצה לחקור. בשלב המידול הראשון בוצעה בדיקה האם עלינו לנתח בעיה זו כבעיית סיווג. כדי למצוא את מספר המחלקות לסווג אליהן, הורץ אלגוריתם ה KMEANS ונמצאה כמות האשכולות האופטימלית לפי מדדים שונים. בהערכה מחודשת של הבעיה הובן שהפתרון אינו אינפורמטיבי דיו, והבעיה הומרה להיות בעיית פרדיקציה.

לטובת בעיה זו הורצו ארבעה מודלים: רגרסיה ליניארית, יער אקראי, רשת ניורונים וxgboost, וחיזויהם הושוו ע"י מדדים סטטיסטים שונים. שם נצפה כי רשת הניורונים והרגרסיה הליניארית הניבו את אותם הניבויים וכי ליער האקראי היו הביצועים הטובים ביותר.

הערה: השוני בין הביצועים שהוצגו בכיתה לבין אלה המוצגים בעבודה נובעים כיוון שהעבודה הוצגה טרם ידיעתינו על הצורך בקביעת seed (כפי שהתקבל בריג'קטים על ההגשה הראשונה).

את תוצאות ביצועי כל המודלים, היפרהפרמטרים הטובים ביותר שנמצאו באימון, ובדיקות נוספות שנעשו יהיה ניתן למצא בקובץ המצורף הכולל את הקוד.

לנוחיותך, ניתן לגשת אליו גם כאן.

https://colab.research.google.com/drive/1QtxUGC7EaZ-oOR2xRkVKkEmppVtoir-i?authuser=1#scrollTo=pZgT-l626Zog

## תוכן עניינים



## מקרא טבלאות ואיורים

## Business understanding

במסגרת פרוייקט הקורס, בחרנו לבחון את הגורמים המשפיעים עם גובה השכר באחת החברות הגדולות במשק -IBM.

IBM מהווה את אחד התאגידים הגדולים והוותיקים בעולם המחשוב ,ובמשך שנים רבות היה הגדול בתחום זה.

התאגיד עוסק במגוון תחומי חומרה ,תוכנה ושירותים :מעבדים ,מערכות מחשב שלמות בגדלים שונים ,ציוד היקפי ,תוכנה , תוכנות בסיסי נתונים ,שרתי יישומים, כלי פיתוח, ייעוץ ועוד .הוא המעסיק הגדול בעולם של עובדים בתחום טכנולוגיית המידע . בשנת 2012 העסיק IBM כ-450,000 עובדים ברחבי תבל ובהם גם ב"יבמ ישראל."

כעת, רוב המועסקים הפוטנציאלים מחפשים את השכר הצפוי להם במנועי החיפוש ומחשבוני השכר, ברשתות החברתיות או בהשוואה עם חברים במשרות דומות במקומות שונים. בדבר זה קיימת בעיה מכיוון שישנה שונות מאוד גדולה בין משכורות כתלות במיקומן בארץ, בהכנסות של החברה ובגורמים רבים נוספים.

מטרתנו היא לתת לעובד הפוטנציאלי אינדיקציה על השכר שאותו יוכל לבקש בבואו להצעת שכר, או להעלאה במקום עבודתו. ומאידך, לתת למעסיק את המידע על על עלות ההעסקה של עובדים קיימים ופוטנציאליים. הכלי יתן לשני הצדדים את היכולת לבצע החלטה מושכלת בבואם לחתום על החוזה.

## Data understanding

### איסוף הנתונים

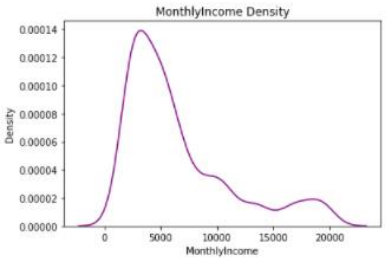
סט הנתונים נלקח מאתר קאגל, בכתובת :

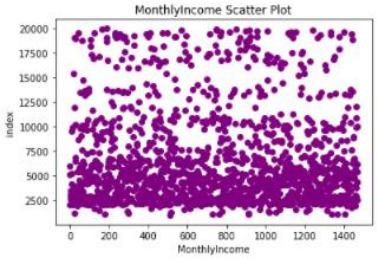
https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv

הסט כלל תחילה 1470 רשומות, משתנה מטרה אחד – רציף, 34 משתנים מסבירים כאשר מתוכם 12 רציפים, 13 בדידים ו9 קטגוריאלים (מתוכם שלושה בוליאנים).

### תיאור הנתונים

#### *משתנה מוסבר- שכר חודשי*




משתנה המטרה: גובה השכר החודשי ($). זהו משתנה רציף אשר תיאורתית יכול לקבל ערכים מ0 (מתנדב) ועד לתקרת השכר הקיימת בעולם.

בפועל ניתן לראות כי בסט הנתונים שלנו נראה כי השכר הכי נמוך שנצפה הוא 1,009 בעוד התקרה שנדגמה הייתה 19,999.

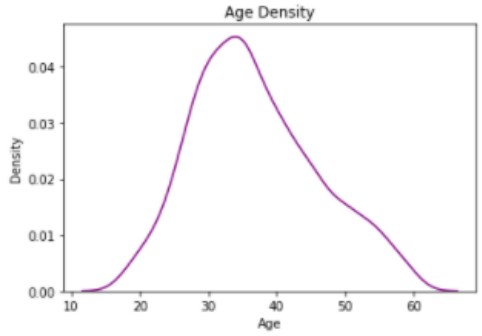
הממוצע בסט נתונים זה היה 6,503 עם סטיית תקן של 4707.95 והחציון 4813.5.

ניתן לראות בגרף הצפיפות שההתפלגות נראת יחסית נורמאלית בחלקה השמאלי בעוד יש זנב ימני עבה, דבר העולה בקנה אחד עם השכר הגבוה בדרגים הגבוהים בהנהלת החברות.

גם בתרשים הפיזור ניתן להבחין שמרבית הצפיפות היא בשלישי התחתון שלו המסמל משכורות עד 7500, והצפיפות מדיללת עבור שכר גבוה.

*איור 1 צפיפות ופיזור משתנה המטרה - השכר החודשי*

## משתנה מסביר - גיל



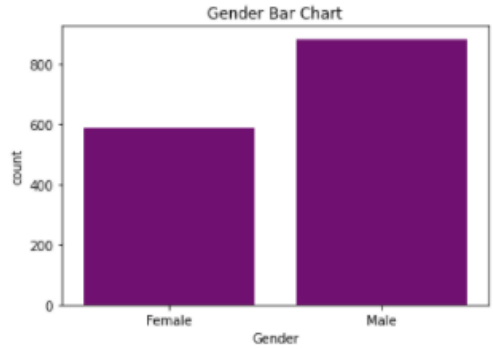

*איור 2 - צפיפות משתנה מסביר – גיל*

ניתן לראות את צפיפות הגילאים אשר מתפלגת די נורמלית, כאשר הגיל המינימאלי במדגם הוא 18 והמקסימאלי הוא 60.

ממוצע הגילאים הוא 36.92 עם סטיית תקן של 9.13, והחציון 36. שוב, דבר העולה בקנה אחד עם הגילאים הצעירים בענף ההייטק כפי שנצפו גם בכתבות על מצב השוק כיום.

## משתנה מסביר - מגדר

```
Male      882
Female    588
Name: Gender, dtype: int64
```

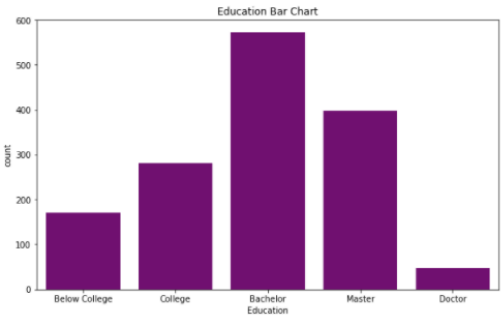



*איור 3 - השוואה בין כמות הגברים והנשים במדגם*

בסט הנתונים שלנו נדגמו 882 גברים לעומת 588 נשים, פער שאינו בטל בשישים.

שוב דבר היכול להיות אידיקציה לכך שהסט מייצג את העולם אמיתי כיום. לא ניכנס כאן לסיבות שנשים פחות מתברגות בשוק ההייטק, אך כן ניתן לקרא על כך במאמר של דינה פיין-קושניר שצורף בבילגיורפיה.

## משתנה מסביר – גובה השכלה




*איור 4 - השוואה בין כמות עובדים בעלי רמות השכלה שונות*

באופן לא מפתיע מרבית העובדים (כ600) הם בעלי תואר 'בוגר' (משמע תואר ראשון), ואחריהם בוגרי תואר שני. מה שכן היה מפתיע מעט לראות הוא שישנם מעט מאוד בוגרי דוקטורט, מאידך הרבה עובדים אשר ללא השכלה כלל או בוגרי מכללות. בארץ ניתן לחשוב על יוצאי יחידות טכנולוגיות אשר מגיעים עם ניסיון מהצבא, אך בארה"ב ניתן להניח שרוב המקצועות המאויישים ע"י אותם העובדים הם אינן משרות פיתוח, אלא משרות כמו שיווק ומכירות שניתן ללמוד אותן בהכשרות ספציפיות ע"י המעסיק.

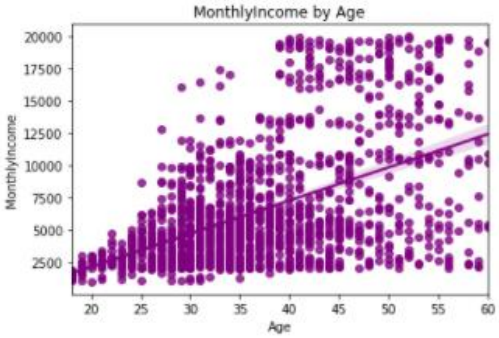
## חקר הנתונים

בשלב זה נבדקו קשרים בין משתנים שונים בסט הנתונים שלנו למשתנה המטרה.

### הקשר בין השכר החודשי לגיל העובד

באופן שאינו מפתיע ניתן לראות בגרף זה את הקשר הליניארי שקיים בין גיל העובד לעליה בשכרו.

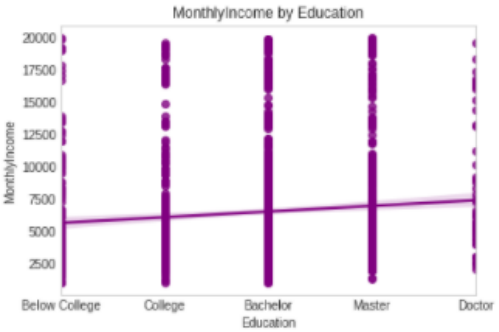
כמובן שהגיל עצמו אינו מנבא את השכר, אלא זה כי קיימים משתנים מתווכים[1], כי ככל שגיל העובד עולה כך גם עולות שנות הניסיון שלו, השנים שלו עם אותו המנהל והותק שלו בחברה, וכל אלה דברים הידועים כמשפיעים על גובה השכר.




*איור 5 גובה השכר כתלות בגיל העובד*

### הקשר בין השכר החודשי לגובה שכר העובד

בבדיקת הקשר בין גובה התואר לבין השכר ציפינו לראות קשר יותר ליניארי ומובהק שייתן עדיפות עבור תארים גבוהים יותר. כן ניתן לראות מגמת עליה קלה בין הממוצי השכר של התצפיות של העובדים ללא התואר לבין ממוצע השכר של העובדים בעלי הדוקטורט, אך לא עלה בקנה אחד עם מה שציפינו לראות.




*איור 6 גובה השכר כתלות בגובה התואר האקדמי של העובד*

### איכות הנתונים

בחלק זה ניסינו להבין גם עד כמה הנתונים שקיבלנו בסט זה על העובדים בשנת 2017, מאפיינים טוב מספיק את העולם האמיתי שלנו כיום, כמה זה מייצג את המגמות בעולם ההייטק המערבי והאם ניתן להכליל זאת על מה שמתרחש בחברות גדולות אחרות גם היום. לשם כך קראנו מחקרים ונתונים של הלמ"ס על עובדי ההיטק בישראל כיום והיה ניכר כי ישנן מגמות דומות בין סט הנתונים לבין מה שקורה בפועל בארץ כיום. לדוגמא דברים שעולים בקנה אחד עם סט הנתונים הוא שרב העובדים בהייטק הם גברים, שככל שהגיל עולה מספר המועסקים יורד, ושלא נמצא הבדל מובהק בין העסקה של בוגרי אוניברסיטיאות ומכללות אקדמיות.

## Data preparation

בשלב זה נעשה מניפולציות על הדאטה הגולמי (Sensed raw data) שלנו בכדי שנוכל להשתמש בו ביעילות רבה יותר בשלבים הבאים. מטרתו לעשות הדגשה לאלמנטים שנרצה. מעין פילטר לניקוי רעשים בדאטה ומחיקת לכלוך למודל הלומד העתידי, שיקבל דאטה איכותי ככל הניתן ונקי מנתונים שעשויים להטות את הניבוי בגלל כשלים לוגיים.

### ניקוי סט הנתונים

בסט הנתונים שנבחר לפרוייקט לא היו ערכים חסרים להשלמה, אז במקרה שלנו מה שבוצע בשלב זה הוא בדיקה האם יש כפילויות של שורות בדאטה, ונצפה כי בסט הנתונים שלנו כל רשומה מתוך 1470 הרשומות מופיעה פעם אחת בלבד.

1 הסתייגות: זה נאמר תחת ההנחה שחוקרים את עולם ההיטק. במשלחי יד אחרים יכול להיות שגם עם עליית שנות הניסיון הוותק לא יעלו את השכר (למשל שחקני תאטרון שישתכרו בהתאם להצגות והמוניטין, או לחלופין מוכרות בגדים ששכרן לא עולה בהתאם לוותק)

כבר בשלב זה הורדו משתנים לא רלוונטים:

- משתנים אשר עבור כל הרשומות יש בהם את אותו הערך(Over18, StandardHours, EmployeeCount)
משתנים אשר להם אותו הערך עבור כל הרשומות אותו הערך מהווים יתירות ולא מוספים לנו מידע חדש. (גם אם תגיע בסט הבחינה תצפית עם ערך שונה, המודלים לא יואמנו לערכים שונים, למשל בעץ היא לא תהיה קריטריון לפיצול לעולם).
- משנים יחודיים לכל רשומה (EmployeeNumber)
מכיוון שהוא יכול לגרום ל- overfitting עם משתנה המטרה. נרצה שהמודל יבין אם העובד משתכר כפי שמשתכר לא לפי שמו אלה לפי המאפיינים שלו כמו מגדר, השכלה, מצבו המשפחתי וכו'. נתון זה הוא מעין מפתח מזהה של הרשומה.
- משתנים אשר מכילים חלק ממרכיבי שכר העובד (DailyRate, HourlyRate, MonthlyRate)
לוגית, לא נכון יהיה לנבא את שכר העובד לפי מרכיבי שכרו משתנים אלו ככל הנראה לא יתקבלו בסט הבחינה עבור משתנה אשר לא עובד בחברה וכך אנחנו מתמודדים עם יתירות (Handling Redundancy)
- משתנה המייצג באופן חד-חד ערכי משתנה אחר (JobRole~Department) – עבור כל תפקיד שאינו 'מנהל' ידוע באופן חח"ע מאיזו מחלקה הוא מגיע, לכן ההתייחסו למשתנה זה נכללה עבור הJobRole בעוד המחלקה ירדה.

## שינוי פורמט הנתונים

בוצעו המרות Data Types: עבור כל המשתנים הקטגוריאלים נוספו משתני דמה נומריים מכיוון שסוג המשימה היא חיזוי ובחרנו לאמנה גם ברשת ניורונים, לאלו ללא חשיבות לסדר (כמו מגדר וסטטוס משפחתי, הוכנסו ערכים מ0 ומעלה כתלות בכמות הקטגוריות שהיו, ולאלו עם חשיבות בסדר בוצעה חלוקה לפי תחומי הערכים) ל5 קטגוריות.

- המרת משתנים קטגוריאלים לנומיליים(ללא סדר)
- המרות של משתנה קטגוריאלי לאורדינאלי
- דיסקריטיזציה של משתנים רציפים
- המרת הסקאלה של כל המשתים בסט לסקאלה של 0-1 לטובת אימון הרשת.

מכיוון שרצינו שסט זה יאומן גם על רשת ניורונלית מכיוון שנדרש להמנע משקל רב יותר לפיצ'רים שהם בסקאלה גבוהה יותר, כל הפיצ'רים ינורמלו להיות באותה הסקאלה (בעבודה השימוש היה בMinMaxScaler).

## יצירת משתנים חדשים

מכיוון שבבחינת הבעיה שלנו רצינו לפתור שתי בעיות, האחת היא ניבוי של שכר עובד בבואו לשיחת שכר (משמע עובד שכבר עובד בארגון) והשנייה לנבא את פוטנציאל ההשתכרות של עובד חדש המגיע לראיון עבודה ולהצעת שכר ראשונית. וכך היה עלינו לעשות את ההבחנה בין עובד "חדש" ל"ותיק". אך יש כשל לוגי בניבוי של שכר עובד חדש על בסיס נתונים של עובדים אשר עובדים בחברה כבר תקופה. למשל, לעובד חדש שיגיע לראיון לא תהיה משמעות למשתנים כמו כמה הוא מסופק מהחברה, או כמה שנים עברו מאז הקידום האחרון שלו בחברה.

אי לכך, יצרנו הבדלה בין עובדים חדשים לותיקים באופן מלאכותי, כאשר יצרנו אידיקטור והתייסנו לכל עובד אשר עובד פחות משנתיים בחברה כעובד חדש וכל עובד אשר עובר יותר משנתיים כעובד ותיק.

ורק עבור העובדים הותיקים יצרנו שני משתנים חדשים, אשר אחד נותן ציון לסיפוק הכללי שלו עם כל המשתנים אשר קשורים לסיפוק, ופיצ'ר נוסף אשר ממשקל את המשתנים אשר קשורים לקידום הכללית שלו בחברה.

בחרנו את המשתנים אשר נראו לנו כרלוונטים להשתייך לכל אחת מהקבוצות ומשקלנו אותם לפי החשיבות שלהם[2].

(ע"ע בנספחים)

## בחירת הפיצ'רים

2 החשיבות שלהם כפי שנראתה לנו נכונה מהיכרות עם עובדים בעבודות שלנו ומקורסים כמו התנהגות ארגונית.

זהו השלב המרכזי שבו נרצה להוציא את כל הפיצ'רים כדי לקבל סט פיצ'רים מצומצם אופטימלי ואינפורמטיבי ללא פיצ'רים מיותרים (קורלטיבים אחד לשני), ככל הניתן על מנת לשפר את תהליך החיזוי שלנו ולא להריץ מודלים על סט עם יתירות. לכן, על מנת לבחור את הפיצ'רים הרלוונטים לאימון בדקנו את מפת החום, ובדקנו את המשתנים אשר ביניהם קיים מתאם גבוה מ0.4.

בשלב זה הוסר האינדיקטור והמשתנים הקורלטיביים ביותר, והמשתנים אשר מאפיינים עובד אשר כבר עובד בארגון.

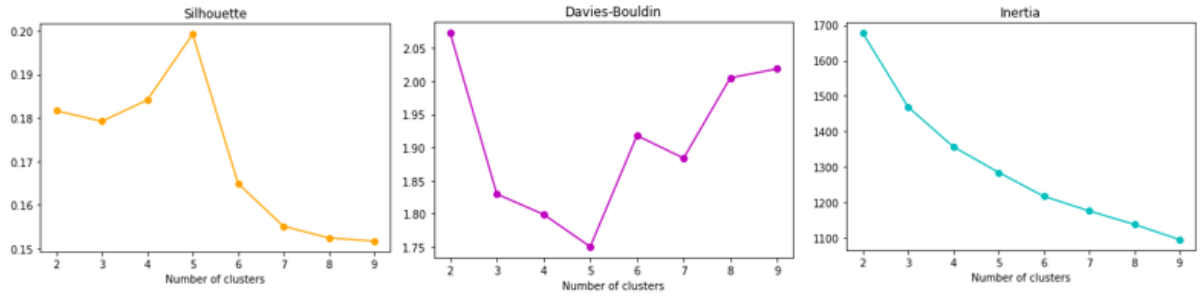
<u>הערה:</u> ההנחה היא שבמעמד ראיון העבודה, הפרטים הבאים יהיו שקופים הן למועמד והן למעסיק: זכאות למניות בחברה, באיזו תדירות הוא צפוי לטוס לפגישות עבודה, אם מעוגן בחוזה שעליו לעשות שעות נוספות, לאיזו דרגה של משרה הוא צפוי להיכנס(Job Level) ולאיזה תפקיד. התיחסותנו למשתנה 'אחוז ההעלאה בשכר' עבור עובדים חדשים אשר מתראיינים למשרה היא אחוז השיפור בשכר לעומת המשרה הקודמת.

## Modeling

### Kmeans

בשלב המידול ראשית נלקח סט הנתונים ללא התגיות והורץ האלגוריתם לאשכול – Kmeans.

הכוונה בניתוח בגישה הלא-מונחית הזו הייתה לבדוק האם קיימת חלוקה לאשכולות, שתבדיל בצורה טובה בין פערי השכר של העובדים. עפ"י המדדים השונים שנלקחו בחשבון למדידת טיב האשכול, 5 אשכולות היוו את החלוקה הטובה ביותר.




*איור 7- כמות אשכולות לפי מדדי טיב האשכול השונים*

ואולם, בהתאם למטודולוגיית CRISP-DM, לאורה נכתבה העבודה, לאחר שלב המידול והערכה, חזרנו חזרה לשלב הראשון של הבנת בעיה ועלתה התהייה אם חלוקה לחמש מחלקות זה הדבר הנכון לעשות בהינתן הבעיה שרצינו לפתור.

נמצא כי אין הגיון רב בחלוקה שכזו לאשכולות כיוון שעוד בשלב הגיוס, עובד מיועד לתפקיד מסויים במחלקה/ אגף מסויימים (עם רמות שכר מתאימות). כך למשל, סביר שמפתח בעל ניסיון יקבל שכר גבוה יותר מעובדת במשאבי האנוש, וכן הד"ר יותר מבוגר האוניברסיטה הטרי.

בנוסף, חלוקה זו אינה אינפורמטיבית דיה: רמות השכר שנמדדו נעו כאמור בין 1k ל-20k, כך שחלוקה ל-5 אשכולות מספקת טווח רב מידיי לחיזוי שאינו מספק ערך מוסף רב לשני הצדדים.

פתרון של בעיית סיווג למחלקות לא באמת מתאים לבעיה הנחקרת.

מעבר ממשימת סיווג למשימת חיזוי. כמה פרטים טכניים על המימוש: כל המודלים הורצו על אותו סט הנתונים, אשר תחילה פוצל ל train and test ביחס של 0.8 ו0.2 בהתאמה. כל מודל הורץ תחילה עם היפרהפרמטרים הדיפולטיים שלו, ולאחר מכן נעשה tuning להיפרפרמטרים בכל מודל על מנת לשפר את ביצועיהם. באימון המודלים עבור ההיפרפרמטרים השונים בוצע grid search עם CV=10.

## linear Regression

משהבנו שדרוש חיזוי מדויק יותר לשכר הצפוי של הפרט, הגדרנו את הבעיה בצורתה הנוכחית: "חיזוי שכר אצל עובדי IBM".

**Basic LinearRegression**
explained_variance_score: 0.9214829772032165
RMSE: 1299.5770500573753
R2: 0.9214063055451537
MAE: 1001.4840134460121

זוהי בעיית רגרסיה קלאסית, ואכן השיטה הראשונה למידול היא הגישה הנאיבית של רגרסיה ליניארית פשוטה.

בקובץ המצורף למסמך זה, ניתן למצוא את השלב המקדים של בדיקת הנחות המודל, וכן השוואה למודלים נוספים כמו Lasso ו-Ridge, אך מאחר וזה אינו נושא הקורס, לא יורחב עליהן, אלא רק ידווחו התוצאות:

המודל הניב ביצועים יפים בדמות אחוז שונות מוסבר של 92% ו-$R^2$ בשיעור דומה.

עם זאת, נמצאו גם טעויות ניבוי של מעל 1k $, אשר וודאי לעובד הישראלי אינן בטלות בשישים, וכך פנינו לשיטות חדשות יותר

## Random forest

האלגוריתם זה מרחיב ומשפר את הניבוי של עץ ההחלטה המסורתי בעזרת יצירת יער. בשיטה זו נייצר B מדגמי bootstrap כשכל אחד משמש לבניית עץ החלטה ע"י m משתנים מתוך p הקיימים בקובץ המאוחד. כל אחד יניב את ניבויו ובסוף התהליך נמצע את כלל הניבויים לכדי תוצאה אחת – התוצאה סופית.

**Improved RandomForestRegressor**
explained_variance_score: 0.9479438798507033
RMSE: 1060.4770260412415
R2: 0.9476657304989752
MAE: 823.1476959323938

כמו בעץ ההחלטה הרגיל, גם כאן מספר רב של היפר-פרמטרים שאפשר לכוונן.
אנו בחרנו להתמקד ב-4:

Bootstrap – כל עץ מאומן על מדגם Bootstrap של הנתונים.

N Estimators – זהו מספר העצים ביער.

Max Depth – זהו העומק המקסימלי אליו כל עץ יוכל להגיע. ככל שערך זה גבוה יותר, כך הסיכון ל-overfitting גדל.

Min Sample Split – מספר התצפיות המינימלי הדרוש לפיצול צומת פנימי (שאינו עלה). ככל שערך זה נמוך יותר, כך הסיכון ל-overfitting גדל.

## XGboost

זוהי שיטת Ensemble נוספת שנלמדה, רק שהפעם גידול היער נעשה בצורה איטרטיבית, כשכל עץ אינו בלתי-תלוי בקודמו. נתחיל מניחוש ראשוני ונחשב שארית, עבורה נתאים עץ החלטה במטרה למזער אותה. האלגוריתם חוזר על התהליך בצורה כזו שכל עץ מנסה לתקן את השגיאות של קודמיו. סוף התהליך יגיע כשהתפלגות השאריות תהיה נורמלית (הטעות תהא ללא דפוס מסוים).

כמו ביער, גם כאן מספר היפר-פרמטרים שאפשר לכוונן. נבחר להתמקד ב-:

Max Depth – זהו העומק המקסימלי אליו כל עץ יוכל להגיע. ככל שערך זה גבוה יותר, כך הסיכון ל-overfitting גדל.

Min Child Weight – זהו פרמטר הגזימה של העץ, נעצור את הפיצול כשמשקל כל צומת עובר רף מסוים. מן הסתם ככל שהרף נמוך יותר, הסיכון ל-overfitting גדל.

Eta – זהו קצב הלמידה. כמו בכל תהליך איטרטיבי גם כאן, הניבוי החדש מוכפל בקצב למידה שקובע את גודל הצעד הנלקח בשיפור.

Subsample – מאחר ותהליך האימון איטי מאוד, נבחר בתת-מדגם אקראי עליו נריץ את המודל.

Colsample Bytree – מאחר ותהליך האימון איטי, נרצה לבחור גם אחוז מסוים מהפיצ'רים (נבחרים אקראית) לכל עץ.

שני ההיפר-פרמטרים האחרונים מסייעים בהתכנסות מהירה יותר של האלגוריתם וגם עוזרים למנוע התאמת -יתר.

**Improved XGBRegressor**
explained_variance_score: 0.9454720024062774
RMSE: 1085.633537181501
R2: 0.9451533457985457
MAE: 849.922156509088

## MLP

המודל האחרון שנבחן הוא רשת הנוירונים הקלאסית, עם שכבה חבויה אחת. כאמור, לצורך האימון היה צורך בעיבוד מקדים של הנתונים כיוון שהקלט לרשת צריך להיות נומרי בלבד, וכזכור מקורס הקדם "לימוד מכונה", המשקלים ברשת מתעדכנים בהתאם לפונקציית ההפסד (MSE) בתהליך ה-BackPropagation. כאן ההיפר-פרמטרים שכוונו היו:

Activation – היחידה. מאחר ומדובר בבעיית רגרסיה קלאסית ואין צורך להטיל את החיזויים לסקאל אחרת.

Learning Rate - קצב הלמידה. הניבוי החדש מוכפל בקצב למידה שקובע את גודל הצעד הנלקח בתהליך ה- Gradient Descent לשיפור. ככל שהקצב נמוך יותר יש צורך ביותר Epochs (הם כמות הפעמים האלגוריתם הלמידה יעבוד על סט האימון).

Learning Rate Init – זהו הגודל ההתחלתי לקצב הלמידה. הרעיון להתאים קצב למידה משתנה, הולך וקטן עם הזמן. נתחיל מצעדי שיפור גדולים בכיוון הגרדיאנט ונסיים ב-Fine Tuning בלבד.

Alpha – פרמטר רגולריזציה שמוסיפים לפונקציית ה-cost ובכך מקטינים את הסיכון להתאמת-יתר.

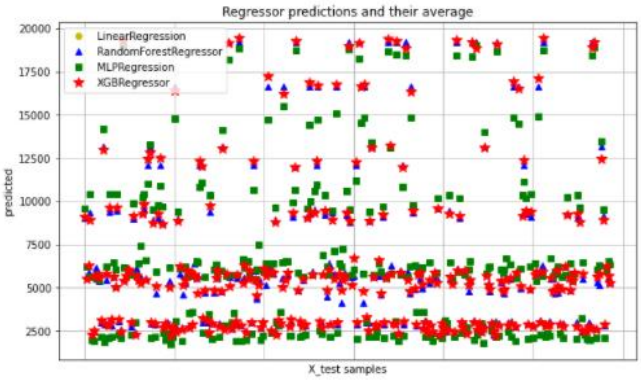
Hidden Layer Sizes – מספר הנוירונים בשכבה החבויה.

**Improved MLPRegressor**
explained_variance_score: 0.921483146173903
RMSE: 1299.5749538058842
R2: 0.9214065590922847
MAE: 1001.4800204849021

## Evaluation

בשלב זה בוצעה השוואת תוצאות התחזיות של המודלים השונים אשר אימנו על סט הבחינה שלנו. התוצאות הן החיזויים של המודלים שאומנו על סט האימון שהושאר בתחילת העבודה. נבדק למי מהם היו הביצועים הטובים ביותר על בסיס ארבעה מדדים שמודדים את טיב החיזוי – $R^2$ , RMSE, MAE, $\eta^2$. את הנוסחאות והסברים עליהם ניתן למצא בטבלה 1 ובנספחים בהתאמה.

.




*איור 8- השוואת חיזוי המודלים השונים: רגרסיה ליניארית, FR, MLP, וXGBOOST*

חדי ההבחנה ישימו לב כי ישנם ארבעה מודלים ורק שלושה ייצוגים בגרף הנ"ל. הדבר נובע מכיוון שהרשת הנוירונלית והרגרסיה הליניארית מניבים את אותם הניבויים. ניתן לראות זאת באיור 9 בנספחים את ההשוואה בזוגות של המודלים.

*טבלה 1 - השוות מדדי הביצוע של המודלים השונים*

| Formula | Metric | Random Forest | XGboost | MLP | Linear Regression |
|---|---|---|---|---|---|
| $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}$ | RMSE | **1060.477** | 1083.949 | 1299.574 | 1299.577 |
| $1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2}$ | $R^2$ | **0.947** | 0.9451 | 0.9214 | 0.9214 |
| $1-\frac{var(y-\hat{y})}{var(y)}$ | η2 (explained_variance_score ) | **0.9479** | 0.9454 | 0.9214 | 0.9214 |
| $\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\hat{y}_i\rvert$ | MAE | **823.147** | 849.922 | 1001.480 | 1001.484 |

## סיכום, דיון ומסקנות

מכיוון שנצפה באופן ויזואלי באיור 8 שרוב החזיויים שהתקבלו עבור המודלים היו קרובים דיים וחזו טוב יחסית את סט הבחינה, מירב ההתלבטות הייתה האם מתבקש לשפר את זמן ריצת המודל או את יכולות הניבוי שלו. מכיוון שהמערכת לא עובדת על מידע שנדרש להיות מנותח בזמן אמת נבחר לשפר את ביצועיו. כאשר המדדים היו טובים דיים מבחינת הביצועים, הדבר הנכון מבחינתנו היה לשקלל את ניבויי המודלים השונים לכדי ניבוי אחד, על סמך הממוצע שלהם .

מדדי הביצוע עבור החיזוי האחרון התקבלו:

**New Pred**
RMSE: 1073.1749618928807
R2: 0.9464049474795355
explained_variance_scoremlp: 0.9466086590634372
MAE: 823.3834623882859

כפי שנאמר בכיתה, בהתאם ל-*CRISP-DM* יש להסתכל על התוצאות בחשיבה ביקורתית.

- השכר תלוי במרכיבים נוספים אחרים שלא היו קיימים בסט הנתונים כגון: כמה מועמדים היו על המשרה הספציפית, האם החברה בהליך גדילה, כמה מועמדים חסרים בשוק במשק וכו' שהיו חסרים בסט הנתונים שלנו.
- שכר ברוטו זה לא הכל, יש מעטפת של תנאים נוספים שגם על פיהם העובד מחליט האם כדאי לו לעבוד בחברה מסויימת או לא.
- נדגמה חברה אחת בלבד, וגם הנתונים שנקלחו על שכר העובדים עבור סניף שלא בארץ, אנו מודעים לבעייתיות בכך, בתחום "חם" כזה יש לקחת בחשבון גם את המתחרים.

## ביבליוגרפיה

1. הקשר בין הגעה לעבודה לבין מגדר + קשר בין הגעה והשכלה :

https://employment.molsa.gov.il/Research/Documents/X12828.pdf

2. שחיקה וגובה שכר

/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5972736

3. נשים והשכלה

https://www.tandfonline.com/doi/full/10.1080/13803611.2016.1256222

4. מגדר והכנסה

https://www.payscale.com/data/gender-pay-gap

5. מדדי ביצוע למשימת רגרסיה:

https://vijay-choubey.medium.com/how-to-evaluate-the-performance-of-a-machine-learning-model-d12ce920c365

6. רמת שביעות הרצון של נשים בתעשיית ההיטק בישראל

http://aranne5.bgu.ac.il/others/Fine-kushnirDina.pdf

## נספחים

*1. הפיצ'רים החדשים שנוספו:*

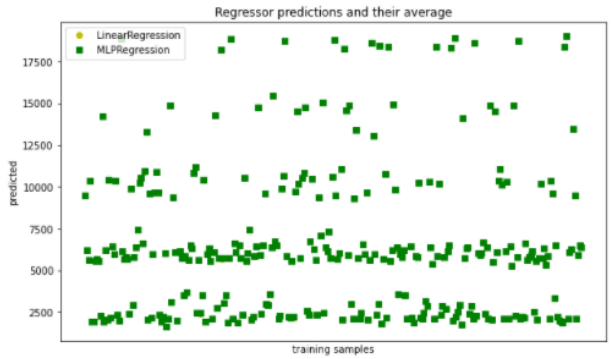
```
df_dataset['worker'] =0.2 * df_dataset['JobInvolvement'] +0.2 * df_dataset['PercentSalaryHike'] +0.2 * df_dataset['TrainingTimesLastYear']  +0.4 * df_dataset['catYearsSinceLastPromotion']

df_dataset['Satisfaction'] =  0.3 * df_dataset['JobSatisfaction'] + 0.2 * df_dataset['EnvironmentSatisfaction'] + 0.3 * df_dataset['catYearsWithCurrManager']+ 0.1 * df_dataset['catYearsSinceLastPromotion']+ 0.1 * df_dataset['RelationshipSatisfaction']
```

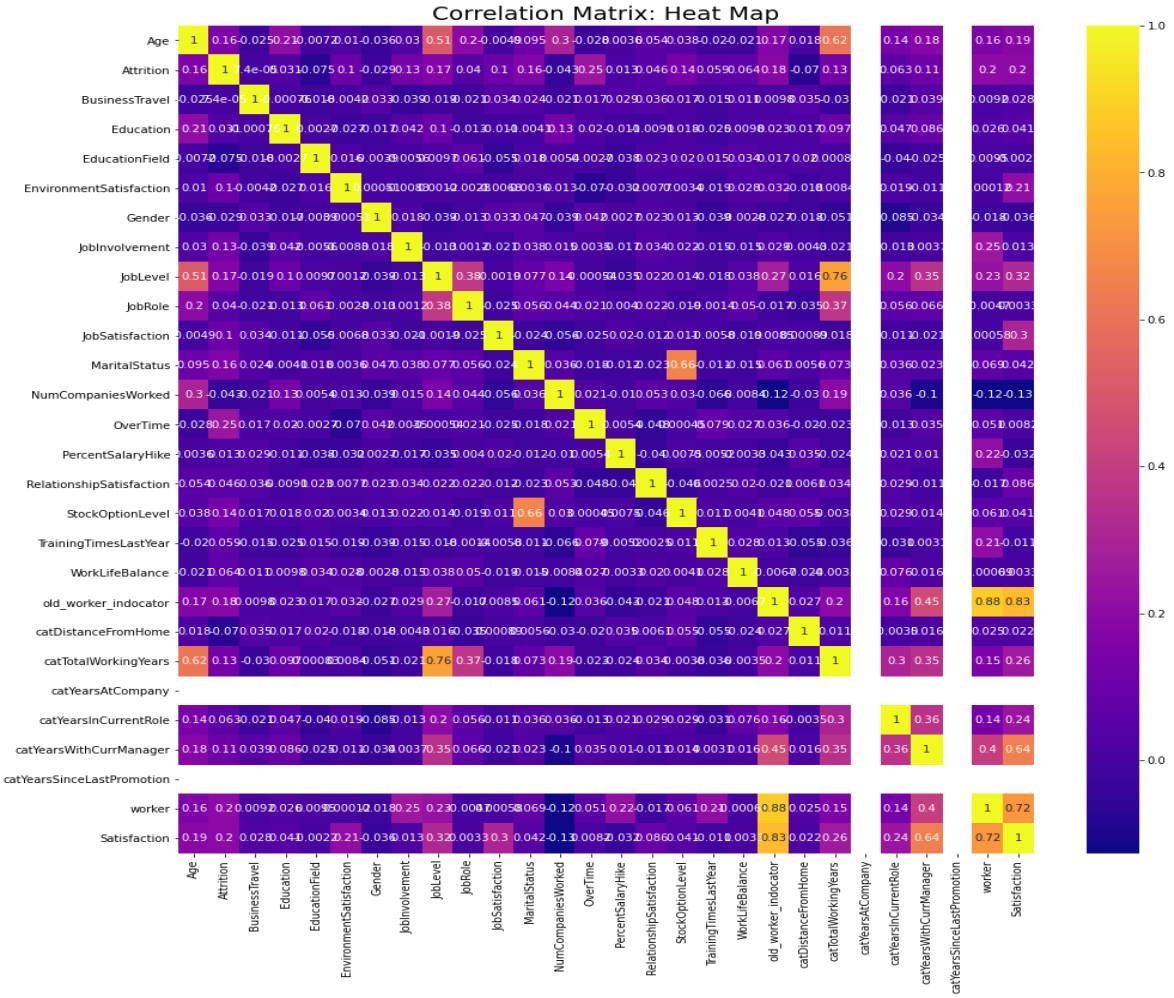
*2. טבלה 2 - הפיצ'רים הסופיים*

| מספר סידורי | שם משתנה | שמות המשתנים כפי שמופיעים בסט הנתונים |
|---|---|---|
| | הכנסה חודשית | MonthlyIncome |
| .1 | מגדר | Gender |
| .2 | שחיקה | Attrition |
| .3 | רמת השכלה | Education |
| .4 | ציון משוקלל לעובד קיים | worker |
| .5 | מרחק מהבית | DistanceFromHome |
| .6 | תדירות נסיעות לחו"ל | BusinessTravel |
| .7 | תחום לימודים | EducationField |
| .8 | שעות נוספות | OverTime |
| .9 | תפקיד | JobRole |
| .11 | אחוז העלאה בשכר | PercentSalaryHike |
| .12 | מניות | StockOptionLevel |
| .13 | דירוג תפקיד | JobLevel |
| .14 | איזון עבודה-פנאי | WorkLifeBalance |

*3. איור 9 - חיזויים של רגרסיה ליניארית ורשת נוירונלית*



*4. איור 10- מפת חום*




*5. איור 10- משתנים בעלי מתאם גבוה מ0.4*

| | Pair | Corr |
|---|---|---|
| 45 | (old_worker_indocator, worker) | 0.878802 |
| 46 | (old_worker_indocator, Satisfaction) | 0.834832 |
| 18 | (JobLevel, catTotalWorkingYears) | 0.755081 |
| 62 | (worker, Satisfaction) | 0.724570 |
| 25 | (MaritalStatus, StockOptionLevel) | 0.662577 |
| 59 | (catYearsWithCurrManager, Satisfaction) | 0.644746 |
| 1 | (Age, catTotalWorkingYears) | 0.616547 |
| 0 | (Age, JobLevel) | 0.509604 |
| 43 | (old_worker_indocator, catYearsWithCurrManager) | 0.451130 |
| 58 | (catYearsWithCurrManager, worker) | 0.403307 |