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

2/12/2021

פרויקט חלק א'

קורס לימוד מכונה

מרצה: ד"ר ניר ניסים

עוזר הוראה: דור זזון

מגישות:

נוי כשר 314963810

ניצן אגם 313360489

קבוצה: 3

תוכן עניינים

Data Collection and Sensing

Data Collection

ה-Data collection הוא השלב הראשון בתהליך לימוד המכונה והוא מייצג את **העולם האמיתי** עליו רוצים ללמוד. זהו ייצוג ראשוני של המידע – איסוף המידע וביצוע **labeling** למידע שנאסף; כל ישות שעובר את תוכנית ההכשרה ב-data collection מייצג **דגימה** אחת. איסוף המידע בצורה איכותית מהווה בסיס חשוב למשימת לימוד המכונה, ולכן על הישויות להיות מייצגות וללא חזרתיות, ממצות, אמינות ושלמות (התמודדות עם ערכים חסרים).

Sensing

שלב החישה הוא תהליך בו מודדים כל אחת מהדגימות בסט הנתונים בצורה עקבית והגיונית. שלב זה מקשר בין **העולם האמיתי** לבין **הנתונים הגולמיים**. פעולת החישה מתבצעת באותו האופן על כל דגימה ובנפרד.

בקורס למדנו על שני סוגי חישה, נסווג כל אחד מהמשתנים לפי סוג החישה המתאים לו:

חישה סטטית	חישה דינמית
החישה תייצר את אותם ערכים ל-sample ללא קשר לזמן בו היא מתבצעת. • enrollee_id • city • city_development_index • gender • major_discipline • training_hours * יצאנו מנקודת הנחה שאינדקס התפתחות העיר אינו משתנה	הערך המתקבל תלוי בזמן בו ביצעו את הדגימה. • relevent_experience • education_level • experience • enrolled_university • company_size • company_type • last_new_job

ניתן לראות ששני סוגי החישה בוצעו ב-Data שלנו, שכן אם נבצע דגימה נוספת על הישותים בתאריכים שונים חלק מהמאפיינים יישארו זהים וחלקם ישתנו.

Sensing שלא בוצע על הדאטה

חישה סטטית - מספר הצעות עבודה בסיום התוכנית. חישה זו תסייע במשימת הלימוד משום שאנו מעריכים כי קיים קשר ישר בין מדד זה לבין ההסתברות לשינוי מקום העבודה בסיום התוכנית.

חישה דינמית - שביעות רצון הישות בתחילת התוכנית, בחציון ובסיומה. חישה זו תסייע במשימת הלימוד משום שאנו מעריכים כי קיים קשר בין מדד זה ושינויו לאורך הזמן לשינוי במקום העבודה בסיום התוכנית.

קטגוריית וסוג משימת הלמידה

קטגוריות משימת הלמידה היא **למידה מונחית** (Supervised Learning), כיוון שערך היעד (ערך ה-Y) ידוע עבור כל דגימה ב-Data. סוג משימת הלמידה היא **משימת סיווג בינארית** (Binary Classification, i.e. Concept Learning), שכן משתנה היעד (target) נע בין הערכים 0 ו-1. משימת למידה נוספת שניתן לבצע על סט נתונים זה היא **משימת חיזוי** (Prediction/Regression Task). במשימה זו נבצע חיזוי למשתנה יעד רציף, המנבא את השכר ההתחלתי של אותו ישות בחברה לאחר סיום הכשרתו בתוכנית.

Dataset Creation

Exploratory Data Analysis

בחלק זה נבצע בחינה של המשתנים שקיבלנו ללא שינויים ב-Data ונסיק על משימת הלימוד בהתאם. קישור לנספח Describe

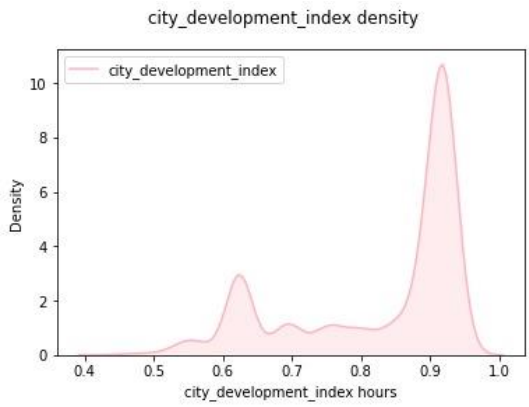
Enrollee_ID

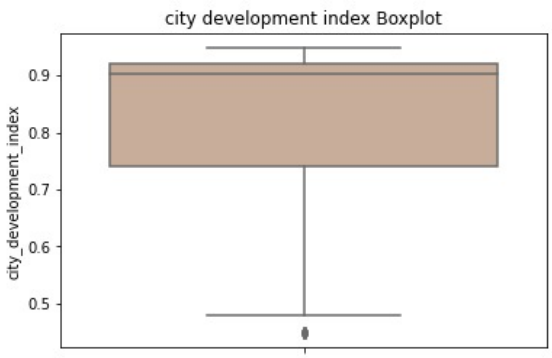
משתנה מסביר המציין מספר זהות ייחודי עבור כל ישות והוא המזהה של כל sample, קיימות 15,326 ישויות ב-Data. ביצענו בדיקה שאין ערכים כפולים בפיצ'ר זה. ערך משתנה זה אינו משפיע על משימת הלימוד.

משתנים רציפים:

City_Development_Index

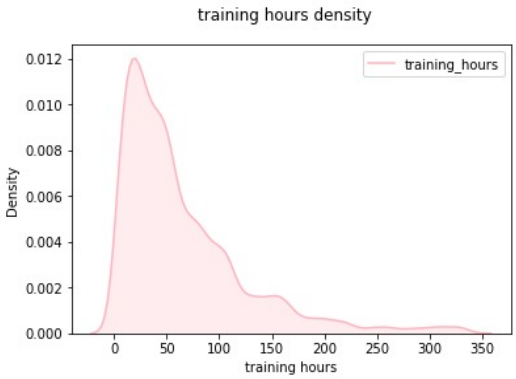
משתנה מסביר המציין את רמת הפיתוח של העיר בה הישות גרה. טווח הערכים נע בין 0 ל-1, ובפועל הערך המינימלי הוא 0.448, הערך המקסימלי הוא 0.949, הערך הממוצע הוא 0.829063 וסטיית התקן היא 0.123161. ניתן לראות כי רוב הישויות מגיעות מערים מפותחות יחסית. נשער כי ישויות המגיעות מערים מפותחות יחסית לא יחפשו שינוי, מכיוון שחייהן נמצאים במקום יציב יחסית.

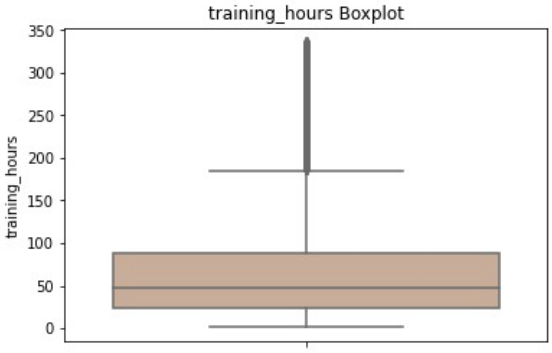




training_hours

משתנה מסביר המציין את מספר שעות הכשרת הישות. הערך המינימלי היא שעת הכשרה, הערך המקסימלי הוא 366 שעות הכשרה, החציון עומד על 47 שעות הכשרה, הערך הממוצע 65.33 וסטיית התקן היא 60.007145. ניתן לראות בתרשים הצפיפות כי התפלגות המשתנה היא קירוב של לוג נורמלית אסימטרית בעלת זנב ימני. נעריך כי ככל שמספר שעות ההכשרה יעלו, כך יעלה רצון הישות להישאר במקום העבודה.



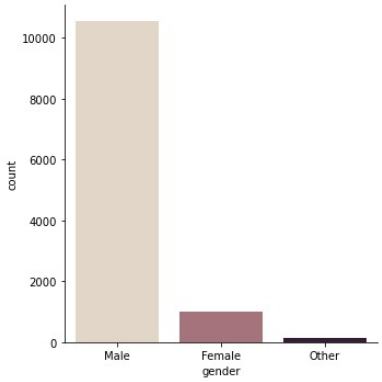


משתנים קטגוריאליים:

City

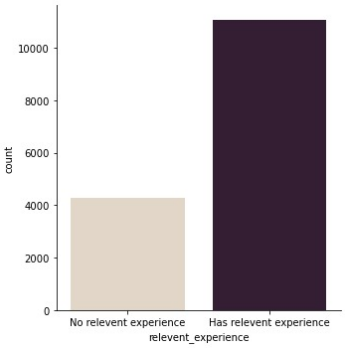
משתנה מסביר המציין את קוד העיר שהישות גרה בה. קיימות 122 ערים שונות ב-Data הנתון. רוב הישויות מגיעות מעיר מספר 106.

Gender

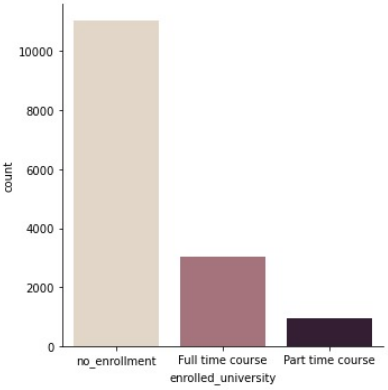


משתנה מסביר המציין את מגדריות הישות. קיימים שלושה ערכים שונים למשתנה זה: Male, Female, Other. הערך השכיח ביותר הוא Male. קיימים samples מועטים עבור Other & Female, וניתן להעריך כי למשימת הלימוד יהיה קשה להכריע האם ישויות אלו ירצו לעזוב את עבודתן.

Relevent_Experience



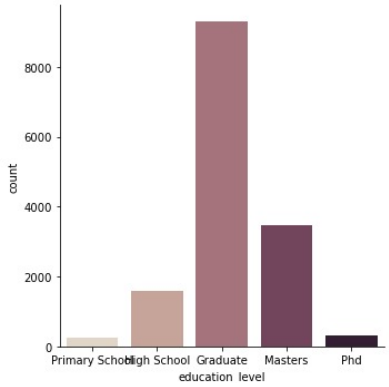
משתנה מסביר המציין האם יש ליישות ניסיון תעסוקתי רלוונטי. קיימים שני ערכים שונים למשתנה זה: Has/No relevant experience. לרוב הישויות קיים ניסיון תעסוקתי.



Enrolled_University

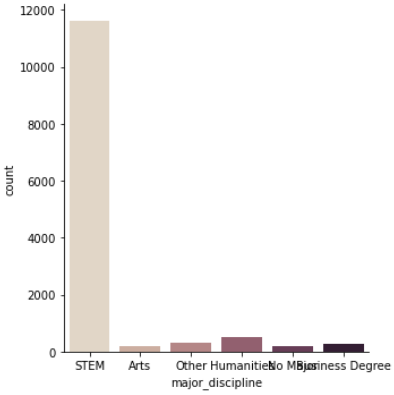
משתנה מסביר המציין כמה הישות מעורבת בלימודים אקדמיים אם בכלל. למשתנה שלושה ערכים:

no_enrollment, Full time course, Part time course. רוב הישויות לא לוקחות קורסים בזמן ההכשרה באוניברסיטה. נעריך כי ישויות הנמצאות במסגרת אקדמאית לא יחפשו שינוי במקום העבודה כיוון שהן יחפשו יציבות כלכלית בשלב זה בחייהם.



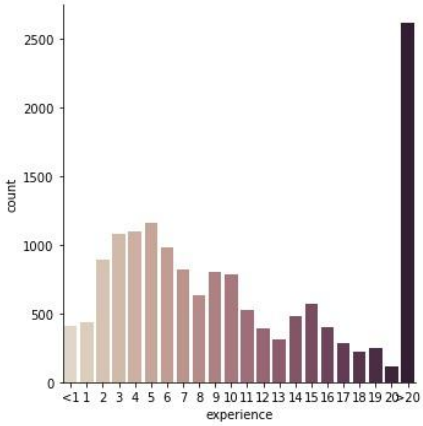
Education_Level

משתנה מסביר המציין את הרמה האקדמית של הישות, בעל חמישה ערכים שונים. רוב הישויות בוגרי תואר ראשון (Graduate), ונראה שקיים קשר בין משתנה זה לבין Enrolled_university, כיוון שרוב בוגרי התואר הראשון סיימו את ההתחייבות האקדמית שלהם לעת עתה. ניתן לראות כי ההתפלגות יחסית סימטרית.



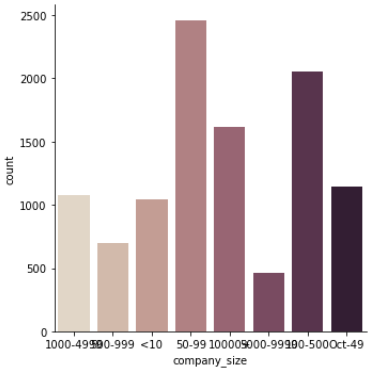
Major_Discipline

משתנה מסביר המציין את תחום ההתמחות האקדמי של הישות, בעל שישה ערכים שונים. כ-75% מהישויות בעלי התמחות ב- STEM (תחום המדע והטכנולוגיה), וייתכן שתחום ההתמחות לא יעזור במשימת הלמידה (זאת נבדוק בהמשך).



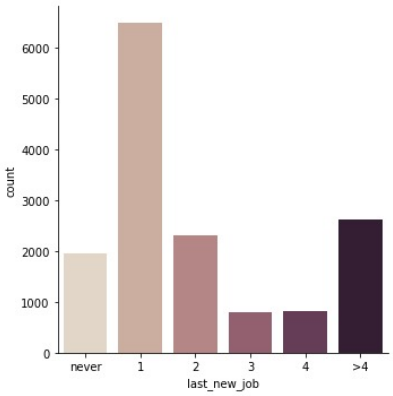
Experience

משתנה מסביר המציין את הניסיון התעסוקתי של הישות, בעל 22 ערכים. המשתנה מייצג שנים עגולות, והערכים בדידים - נעים בין 1 למספר שנים הגדול מ-20. רוב הישויות בעלות יותר מ-20 שנות ניסיון, ומנגד קיימות לא מעט ישויות עם שנות ניסיון נמוכות יחסית (מתיישב יחד עם המסקנה הקודמת, שרובם בוגרי תואר ראשון).



Company_Size

משתנה מסביר המציין את גודל החברה הנוכחית של הישות, בעל שמונה ערכים. הנפוץ מביניהם הוא גודל חברה של 50-99 עובדים. נבחין כי קיימות טעויות במשתנה זה הנובע מפונקציות נתונים ב-Excel (למשל: גודל חברה Oct-49 במקום 10-49), אותן עלינו לתקן בשלב ה- Pre processing.



Company_Type

משתנה מסביר המציין את סוג הארגון בחברה הנוכחית של הישות, בעל שישה ערכים. הערך הנפוץ הוא חברה פרטית בע"מ (Pvt Ltd), ונעריך שהתנאים בחברות פרטיות טובים באופן יחסי ולכן תהיה נטייה להישאר במקום עבודה זה. בנוסף, נעריך כי קיים קשר בין חברות יציבות מהסקטור הפרטי והציבורי לבין רמת התפתחות גבוהה בעיר.

Last_New_Job

משתנה מסביר המציין את ההפרש בשנים בין מקום העבודה הקודם לבין מקום העבודה הנוכחי, כלומר מתי התרחש שינוי העבודה האחרון. למשתנה קיימים שישה ערכים שונים, הערך הנפוץ להפרש הוא שנה. ערך זה עשוי ללמד אותנו שעובדים אלו יעדיפו להישאר במקום העבודה ולצבור ניסיון. מההיסטוגרמה ניתן לראות כי מדובר בהתפלגות אסימטרית עם זנב ימני.

Target - משתנה מוסבר, משתנה היעד

משתנה בינארי המציין האם הישות חיפשה שינוי תעסוקתי בתום ההכשרה (=1), או לא (=0). שאר המשתנים מסייעים להסביר את משתנה זה, ובאמצעותם נוכל להכריע עבור Data חדש האם ישות מסוימת תחפש שינוי תעסוקתי. כ-75% מהישויות לא חיפשו שינוי תעסוקתי בסוף ההכשרה ואילו הנותרות כן.

איזון ב-Data תלוי במשתנה היעד, ומשמעות שכיחות הערך 0 (75%) היא **שה-Data אינו מאוזן.**

קורלציה בין המשתנים – קישור לנספח מפת חום

מתאם הינו מדד סטטיסטי המבטא את המידה שבה שני משתנים קשורים זה לזה בקשר לינארי. נשים לב כי סט הנתונים כולל משתנים רציפים לצד משתנים קטגוריאליים. על כן, נחשב את הקורלציה ואת עוצמת הקשר בין משתנים אלו באמצעות הכלים הבאים:

- מתאם פירסון - עבור מקרים של רציף-רציף.
- יחס קורלציה - עבור מקרים של קטגוריאלי-רציף.
- Cramer's V or Theil's U - עבור מקרים של קטגוריאלי-קטגוריאלי.

דוגמאות לקשרים מעניינים בין משתנים מסבירים:

- (city VS city_development_index (corr = 1 - קורלציה גבוהה מעידה על תופעת המולטי-קולינריות שעלולה לפגוע באמינות אומדני המודל, ולכן בשלב ה-Feature Selection נשקול להסיר את אחד מהמשתנים.
- company_size VS company_type (corr = 0.38)
- relevant_experience VS enrolled_university (corr = 0.39)

בעזרת קשרים אלו ננסה ליצור בשלב Feature extraction משתנים חדשים שיחזו בצורה מיטבית האם העובד יעזוב את מקום עבודתו בתום ההכשרה.

pre-processing

בשלב זה נבצע מניפולציות על ה-Row Data כדי שנוכל להשתמש בו בצורה מיטבית למשימת הלמידה.

Clean noise

ראשית, נתקן את הערך 'Oct-49' בעמודת *Company_Size* לערך '10-49'.

כמו כן, בדקנו שישויות בעלות השכלה תיכונית/יסודית אינן בעלות תחום התמחות אקדמי כיוון שהן עדיין לא עוסקות בפעילות אקדמית כלשהי.

Redundancy in data

ניתן לבדוק בקלות את החזרתיות ב-Data לפי המזהה enrollee_id, ולאחר בדיקה זו ניתן להגיד שאין כפילות בנתונים שעלולה לפגוע באמינותם.

Missing values - קישור לנספח ערכים חסרים

כחלק ממשימת הלימוד עלינו לדעת כיצד להתמודד עם ערכים חסרים ב-Data. במקרה שלנו חסרים 16,659 ערכים ב-7 מהפיצ'רים, ובשלב זה נבחן מהי הדרך הנכונה להתמודד עם הבעיה על מנת להכין את ה-Data למשימת הלימוד.

- **מחיקת משתנה gender** - בעמודה זו קיימים כ-3000 נתונים חסרים (כחמישית מה-Data). בנוסף, שמנו לב שמשתנה זה אינו חיוני להכרעת המשתנה המוסבר כלומר הקורלציה ביניהם זניחה, ועל כן החלטנו להסיר משתנה זה.
- **מחיקת דגימות** - קיימות 10 דגימות ב-Data בהן חסר ערך בכל שבעת הפיצ'רים העלולים להכיל ערך חסר. בחרנו להסיר את דגימות אלו, כיוון שקשה יהיה להיעזר בכלים שנלמדו על מנת להשלים את עמודות אלו מבלי לייצר רעש. בנוסף, הוחלט כי נסיר כל דגימה שחסרים בה ערכים לפחות ב-4 מהפיצ'רים.
- **השלמת ערך No Major** - נצא מנקודת הנחה שישויות בעלות השכלה תיכונית/יסודית אינן בעלות רקע אקדמי (עדיין), לכן דגימות אלו קיבלו את הערך 'No major' במשתנה major_discipline.

- **השלמת ערך STEM** - קיימת קורלציה חזקה בין education_level ו-major_discipline, וקיימות דגימות בהן הישות בעלת השכלה אקדמית כלשהי (Graduate, Masters, Phd) אך ללא תחום התמחות. מכיוון שמרבית הדגימות בעלות הערך STEM, שכן ישויות אלו עוברות הכשרה ל-DS, נבחר להשלים את ערך זה לכל אותן דגימות.
- **השלמת ערך שכיח** - קיימת קורלציה חזקה בין company_size ו-company_type, וקיימות דגימות רבות בהן קיים ערך במשתנה אחד וחסר בשני. נבחר להשלים ערכים אלו בעזרת הערך הנפוץ ביותר, לדוגמא: עבור חברה המונה פחות מעשרה עובדים, נמצא מהו סוג החברה הנפוץ ביותר ונשלים זאת בכל אחד מהערכים החסרים.
- **יצירת קטגוריית None** - כאשר חסר ערך ב-company_size וגם ב-company_type, נשלים את הערך ע"י הוספת הקטגוריה 'None'. בהמשך ניצור פיצ'ר חדש שמתבסס על יכולת ההבחנה בין ישות בתוכנית בעלת מקום עבודה (אשר לה קיים ערך לפחות באחד משני המשתנים) או לא.
- **שימוש באלגוריתם KNN** - את שאר הערכים החסרים השלמנו באמצעות אלגוריתם K-nearest neighbors, השומר על השונות של ה-Dataset, וכן הוא מדויק ויעיל יותר משיטות אחרות שלמדנו בקורס (לדוגמא ערך ממוצע שאינו מתאים לערכים קטגוריאליים). זהו אלגוריתם מבוסס מרחק בין דגימות שונות, המשלים את הערכים החסרים בעזרת הדגימות הדומות ביותר. הגדרנו את הפרמטר של האלגוריתם להיות k=1 על מנת להימנע מערכים לא שלמים. על מנת להשתמש באלגוריתם זה ביצענו דיסקרטיזציה ונרמול של ה-data, עליהם נפרט בשלב ה-feature representation.

<u>Data type conversions</u>

בחרנו להמיר את המשתנה experience מקטגוריאלי לרציף. ראשית, כיוון שטבעי יותר למדל יחידות מידה של שנים כמשתנה רציף, ושנית כי קיים ריבוי קטגוריות למשתנה זה ועלינו לשאוף להוריד את מימדיות המודל.

<u>Proportions in the data</u>

כפי שכבר ציינו, ה-Data הנתון **אינו מאוזן**, שכן מרבית העובדים יבחרו שלא לשנות את מקום עבודתם. אנו סבורות כי הבעיה משקפת את המציאות כפי שהיא, כיוון שבמציאות רוב שוק העבודה אינו מחפש שינוי בהעסקתם ועל כן לא נבצע שינויים בשלב זה.

segmentation

בשלב זה נרצה לחדד את המידע בכל פיצ'ר ולהשאיר ב-Data את המידע הרלוונטי ביותר. תחילה שקלנו לערוך את המשתנה city ולמחוק את תת המחרוזת '_city' מכלל הדגימות, כך שישאר מספר העיר בלבד. לאחר בחינה נוספת של ה-Data ובמיוחד לאור הקורלציה

המושלמת בין פיצ'ר זה ל-city_development_index, החלטנו להוריד את עמודת city מהמודל ולכן לא ביצענו את שלב זה.

feature extraction

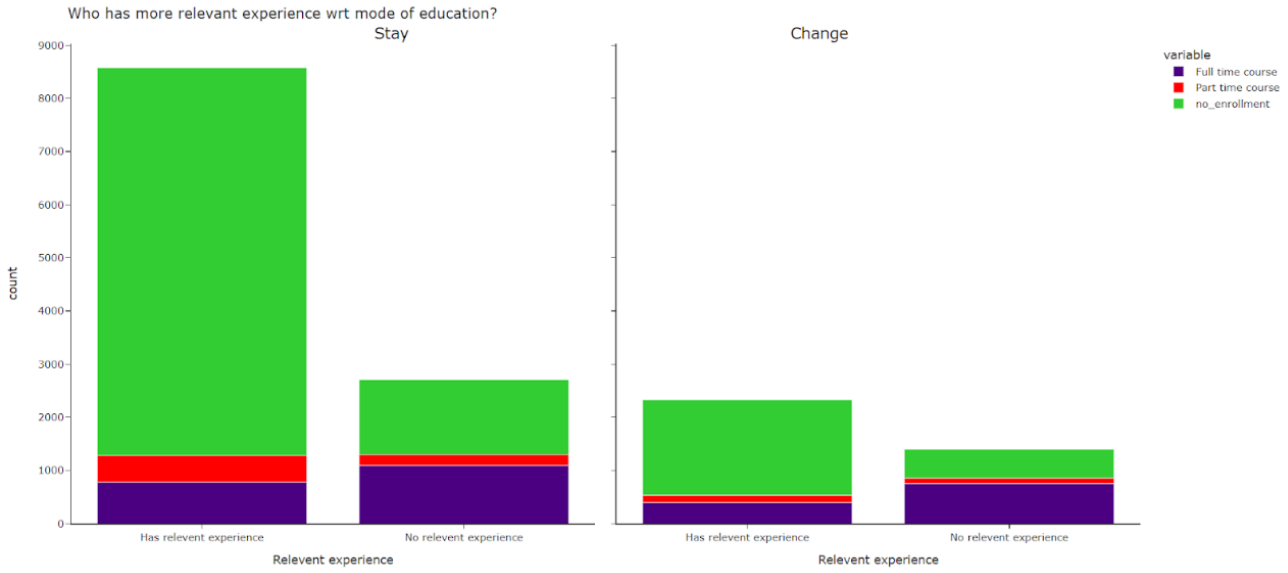
בשלב זה ננסה למצוא קשרים מעניינים בין משתנים וליצור משתנים חדשים על מנת לשפר את יכולת משימת הלמידה. מכיוון שמשימת הלימוד שלנו הינה לבצע קלסיפיקציה לעובדים בתכנית ההכשרה, מטרתנו היא ליצור פיצ'רים שיסייעו באבחנה בין הרצון להישאר בעבודה או לחפש שינוי.

כמות הפיצ'רים אינה משתנה בין כל דגימה, ועל כן נשתמש במתודולוגיית *Fixed number of Features* בשיטת *Knowledge Based*. הידע שלנו בנושא מתבסס על תחום התעסוקה בעולמות התוכן של DS, מכיוון שיש לנו ידע מספק בתחום אין צורך להיעזר במומחה.

- **pop_enrollee** - משתנה בינארי המקבל ערך 1 אם המתמודד מגיע עם נתונים פופולריים ב-Data, כלומר ה-major_discipline שלו הוא STEM **וגם** ה-company_type הוא Pvt Ltd, ו-0 אחרת.

חילצנו את משתנה זה כיוון ששני ערכים אלו הינם הפופולריים ביותר בקטגוריות שלהם, והשילוב בין שניהם עשוי לבטא קשר חזק יותר עם משתנה המטרה.

- **is_working** - משתנה בינארי המקבל 1 אם לישות קיים ערך לפחות באחד משני המשתנים company_type **או** company_size ו-0 אחרת. עבור חלק נכבד מהישויות בתוכנית לא קיים מידע באף אחד מהמשתנים הנ"ל וההיגיון מוביל למסקנה שהם לא עובדים. כחלק משיפור יכולת משימת הלימוד נרצה לבחון את הקשר בין קיום מקום עבודה של ישות בתוכנית ההכשרה להכרעה להישאר בחברה.
- **university+relevent_exp** - כפי שניתן לראות בגרף, מצאנו קשר מעניין בין המשתנה enrolled_university ל-relevent_experience :



- הסיכויים לחיפוש שינוי תעסוקתי גבוהים יותר עבור ישויות ללא ניסיון שהם סטודנטים "במשרה מלאה".
- ישויות בעלות ניסיון רלוונטי ללא קורסים באוניברסיטה הן הקבוצה הגדולה ביותר בקרב העוזבים את מקום עבודתם.
- רוב הסטודנטים "במשרה חלקית" לא יחפשו שינוי תעסוקתי ללא תלות בניסיון הרלוונטי, לכן ישות במשרה חלקית תקבל ערך אחיד במשתנה החדש ללא תלות בניסיון.

כתוצאה מהתובנות הנ"ל, יצרנו חמישה ערכים שונים לקטגוריה עבור כל הקומבינציות האפשריות פחות אחד. **דוגמא ל-Dataset לאחר חילוץ הפיצ'רים:**

Index	velopmen	relevent_experience	enrolled_university	education_level	major_discipline	xperience	ompany_siz	company_type	st_new_jo	iining_hou	pop_enrollee	is_working	university+relevent_exp	target
0	0.91	No relevent experience	no_enrollment	High School	No Major	2	None	None	never	16	0	0	0	1
1	0.897	Has relevent experience	no_enrollment	Masters	STEM	6	1000-4999	Pvt Ltd	1	262	1	1	2	0
3	0.91	No relevent experience	no_enrollment	High School	No Major	3	None	None	1	35	0	0	0	0
4	0.555	Has relevent experience	Full time course	Graduate	STEM	4	None	None	1	43	0	0	3	1
5	0.897	Has relevent experience	no_enrollment	Masters	STEM	>20	500-999	Pvt Ltd	2	18	1	1	2	0

feature representation

לאחר שחילצנו את המשתנים, עלינו להחליט כיצד נרצה להציג את הישויות לפיהם. השינויים שביצענו בשלב זה נובעים מהרצון להכין את ה-Data למשימות הלימוד בצורה המיטבית ביותר.

- **קידוד משתנים קטגוריאלים אורדינליים** - משתנים בעלי משמעות לסדר בין הערכים הוחלפו לייצוג מספרי (enrolled_university, education_level, last_new_job, company_size).
- **קידוד משתנים קטגוריאלים נומינליים** - משתנים ללא משמעות לסדר בין הערכים צריכים להיות מיוצגים בצורה אחרת, שכן אין משמעות למרחקים בניהם. בחרנו לבצע את הקידוד בשיטת **Base-N** עם בסיס 5, הידוע כ-Quinary system. שיטה זו מתמודדת עם ריבוי מימדיות בצורה יעילה ומפחיתה את מספר העמודות הנדרשות כדי לייצג ביעילות את הנתונים לעומת משתני דמה.
  את הקידוד ביצענו עבור major_discipline, company_type והמשתנה החדש מהשלב הקודם university+relevent_exp.
- **נרמול ערכים** - על מנת למנוע הטיה בתהליך הלמידה ולהשיג קנה מידה אחיד, ביצענו נרמול של המשתנים הרציפים experience, training_hours על פי מדד **min_max**:

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- **הצגה בוליאנית** - relevent_experience הוא בינארי מטבעו והומר לייצוג מספרי של 1 ו-0.

לאחר כלל השינויים ה-Dataset מיוצג באופן הבא:

Index	velopmen	ent_exper	lled_unive	ucation_le	or_discipli	or_discipli	experience	mpany_si	npany_typ	npany_typ	st_new_jo	training_hours	op_enrolle	s_working	y+relever	y+relever	target
0	0.91	0	0	1	0	1	0.0952381	8	0	1	0	0.0447761	0	0	0	1	1
1	0.897	1	0	3	0	2	0.285714	5	0	2	1	0.779104	1	1	0	2	0
3	0.91	0	0	1	0	1	0.142857	8	0	1	1	0.101493	0	0	0	1	0
4	0.555	1	2	2	0	2	0.190476	8	0	1	1	0.125373	0	0	0	3	1
5	0.897	1	0	3	0	2	1	4	0	2	2	0.0507463	1	1	0	2	0

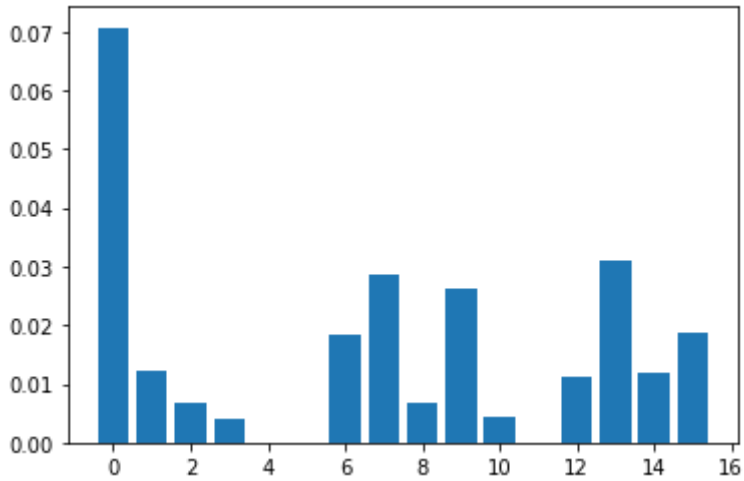
feature selection

כחלק מה-Dataset קיימים משתנים שאינם אינפורמטיביים מספיק על מנת לתרום למשימת הלימוד, וחלקם אף עלולים להטעות את מודל הלמידה. כתוצאה מכך, נרצה לספק למודל הלמידה סט פיצ'רים קטן אך מייצג ככל הניתן. ראשית, נמחק מה-data את המשתנה city שאינו מניב מידע חדש כלל, והוא בעל קורלציה מושלמת למשתנה city_development_index. בנוסף, נמחק את משתנה enrollee_id שהינו חד ערכי ואינו תורם למשימת הלימוד.

כמו כן, עבור שאר הפיצ'רים השתמשנו במתודולוגיית **Quantitative evaluation**, שבבסיסה מודדים ומחשבים את חשיבות ותרומת הפיצ'רים למשימת הלמידה.

בחרנו ב-**Filter method** ובפרט ב-**information gain**: מדד זה מחושב בין כל שני משתנים, ובהינתן משתנה אחד, הוא מודד את הפחתת אי-הוודאות/אי-הסדר של המשתנה השני.

שיטה זו מיוצגת ע"י הגרף:



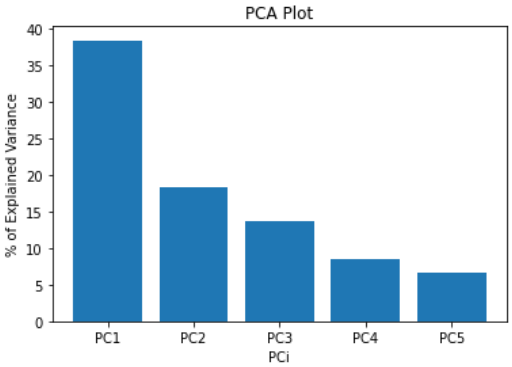

```
0.070702410228324 91
0.012061764255869356
0.006701866317534444
0.0039616922825049805
9.657214661218383e-05
5.956372443982616e-05
0.018471107667686892
0.02868938181455638
0.00693069725388229
0.026179878688914693
0.004287822918620288
0.0
0.011260274043649598
0.030945573399667303
0.011926022598020669
0.018792062543934174
```

המשתנים אותם בחרנו להשאיר במודל בסדר יורד לפי מדד ה-information gain:

city_development_index, is_working, company_size, company_type_1, university+relevent_exp_1 ,experience, relevent_experience, university+relevent_exp_0, pop_enrollee.

dimensionality reduction

זהו השלב האחרון בתהליך ה-Dataset Creation, אותו נבצע על מנת להוריד את מספר המימדים הנדרשים להסביר את ה-Dataset ולשפר את דיוק המודל. נבצע אותו ע"י שימוש בשיטת **Principal Component Analysis**. אנו מודעות לטרייד-אוף בין אחוז השונות המוסברת לבין מספר המשתנים ולכן ניסינו למצוא את נקודת האיזון ביניהם. לאחר שלב ה-feature selection נותרנו עם שמונה משתנים, ולאחר ביצוע ה-PCA קיבלנו חמישה משתנים חדשים המהווים קומבינציה לינארית של הקודמים להם, השונות המוסברת שלהם באחוזים היא: [38.4, 18.4, 13.7, 8.5, 6.6].



נוותר על משתנה PC5, שהינו בעל השונות המוסברת הנמוכה ביותר, ובעזרת ארבעת המשתנים הנותרים המודל שלנו יצליח להסביר כ-79% מהשונות הכוללת.

model training

כדי להעריך נכונה את השגיאה נבחר להשתמש בשיטת **K Fold Cross Validation.**

בחרנו בשיטה זו על מנת להגדיל את ה-Exploration שלנו על ה-Data. בעזרת חזרתיות בחינת המודל על חלקים שונים של ה-Data, נוכל להעריך את שגיאת המודל בצורה אמינה יותר, וכך ניצור מודלים אשר יסייעו לנו לבצע את משימת הלימוד בצורה טובה יותר על סט נתונים שלא פגשנו - ובפועל להגיע לתוצאות טובות יותר בעולם האמיתי.

על מנת להתקרב לנקודת ה-Optimize Fitting **נבחר ב-K=10**, כמקובל בספרות עבור Dataset גדול יחסית. בשיטה זו נחלק את הData שלנו ל-10 חלקים, ובכל איטרציה נאמן את המודל על 9 חלקים ונשאיר חלק אחד בחוץ שיהווה כ-Validation set. בנוסף, נחשב בכל איטרציה את מדד השגיאה/דיוק של המודל ולבסוף נחשב את ממוצע השגיאות של כלל האיטרציות.

למרות שזמן הריצה של שיטת holdout קטן משל K fold, בחרנו להימנע משיטה זו כיוון שה-Data שלנו אינו מאוזן, ובשיטה זו אנחנו עלולים לפספס דגימות חשובות בעת תהליך הלמידה.

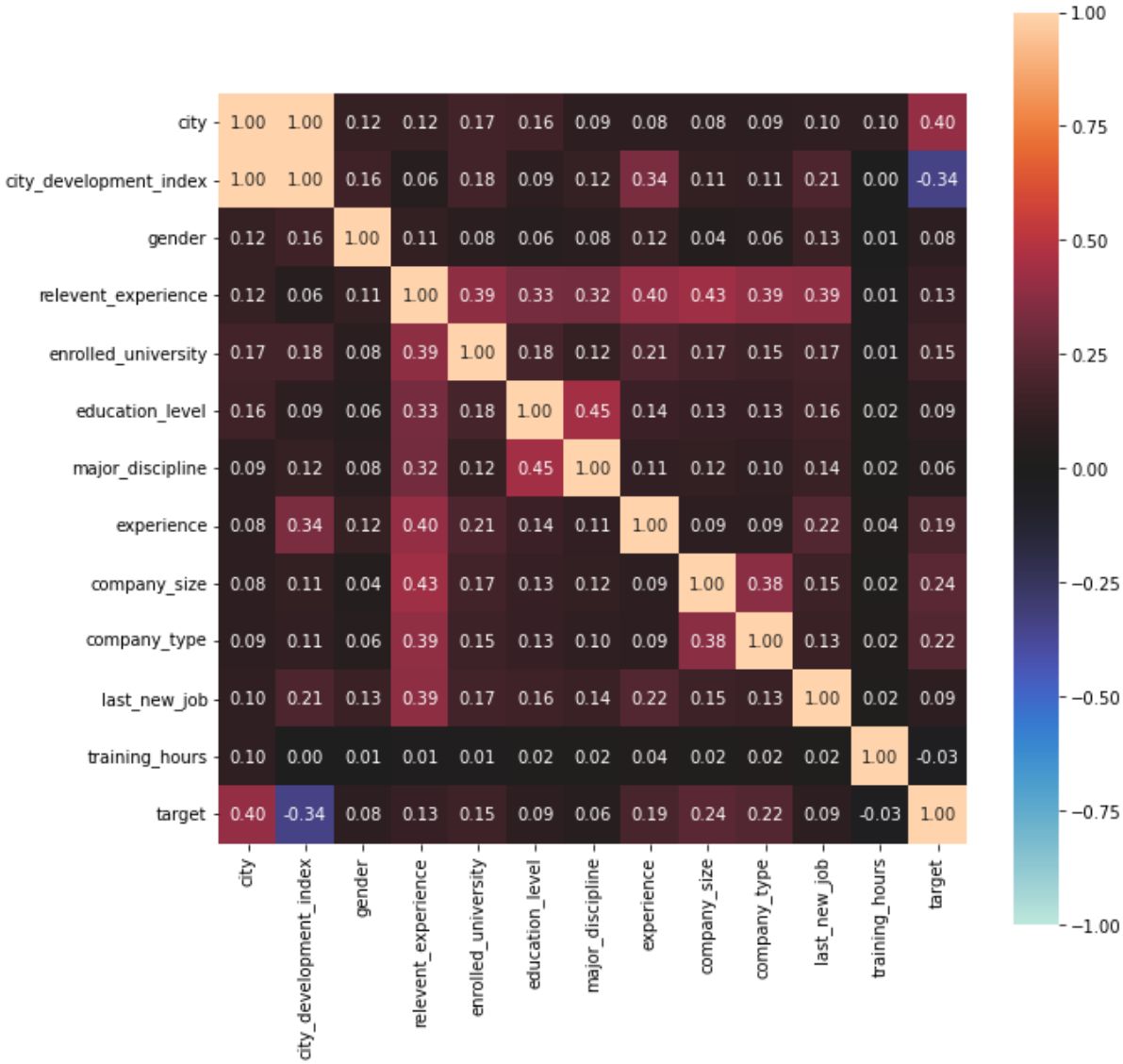
בשיטת Leave one out יכולנו להגיע לרמת דיוק גבוהה באופן יחסי, אך ה-Dataset שלנו גדול וזמן הריצה בשיטה זו יהיה ארוך מאוד. בנוסף, שיטה זו עלולה להוביל אותנו ל-overfitting.

נספחים

נספח פלט Describe

```
Out[9]:
       city_development_index  training_hours        target
count            15326.000000    15326.000000  15326.000000
mean                 0.829063       65.337466      0.249576
std                  0.123161       60.007145      0.432782
min                  0.448000        1.000000      0.000000
25%                  0.740000       23.000000      0.000000
50%                  0.903000       47.000000      0.000000
75%                  0.920000       88.000000      0.000000
max                  0.949000      336.000000      1.000000
```

נספח מפת חום עבור קורלציה בין משתנים מסבירים



נספח ערכים חסרים

```
city                        0
city_development_index      0
gender                   3639
relevent_experience         0
enrolled_university       311
education_level           370
major_discipline         2237
experience                 48
company_size             4779
company_type             4943
last_new_job              332
training_hours              0
target                      0
dtype: int64
```