****

**ML Project**

**Part A**

**Data collection and Sensing**

**מהו הData collection ? איזה סוג Sensing בוצע על הדאטה (סטטי \ דינמי), הסבירו ?**

ה Data Collection במקרה שלנו הוא אוסף מודעות דרושים, מודעות אמתיות ומזויפות שנאספו ממקורות שונים ואודותיהן תועדו ב DB.על הדאטה בוצע תהליך Sensing סטטי שהרי הנתונים נאספו ותויגו בנקודת זמן מסוימת, וסביר כי המידע בבסיס הנתונים לא ישתנה לאורך זמן.

**הציעו סוגSensing שלא בוצע על הדאטה, והסבירו איך הוא יכול לעזור למשימת הלימוד (אין צורך להוסיף אותו פיזית, אלא רק ברמה הקונספטואלית)**

הוספת שדה שמתעד את מספר התגובות / הכניסות שמתקבלות עבור כל מודעת דרושים.

שדה זה יתעדכן לאורך זמן ויאפשר למודל לזהות דפוסים או מגמות חריגות שעלולים להצביע על כך שהמודעה היא הונאה. אם המודל יזהה לדוגמה מודעת דרושים שמקבלת המון תגובות בפרק זמן קצר מאוד , זו עלולה להיות אינדיקציה לכך שזו מודעה מזויפת. שילוב חישה דינמית יאפשר למודל מעקב רציף אחרי הפרמטר של מספר תגובות / כניסות למודעות הדרושים מה שיעזור לו לסווג את המודעה כמהימנה או כהונאה וישפר את הדיוק שלו.

**רעיון נוסף -** מעקב אחרי שדה של משוב משתמשים על מודעות הדרושים.עבור כל מודעה יהיה תיעוד של משוב שנתנו לה המשתמשים המשתמשים יוכלו לציין עבור מודעה האם הם מצאו אותה כמהימנה או שלדעתם מדובר בהונאה. תגובות המשתמשים ישתנו לאורך זמן ולכן מדובר בחישה דינמית.הוספה של שדה כזה יכולה לספק משוב בעל ערך רב למודל שישפר את יכולות הסיווג שלו.הוספת מעקב אחר משוב מהמשתמשים יכולה לשפר את הדיוק של המודל וללמד אותו לזהות דפוסים ומגמות שיקלו על זיהוי מודעות מזויפות בעתיד.

**מהי קטגוריית וסוג משימת הלמידה (נלמד במצגת 3)? הסבירו את תשובתם. האם ניתן להשתמש בנתונים כדי לבצע עוד סוג של משימת למידה ?**

קטגוריית משימת הלמידה היא למידה מפוקחת Supervised Learning וסוג משימת הלמידה הוא סיווג בינארי.

למידה מפוקחת מתייחסת ללמידת מכונה בה המודל מאומן על נתונים מתויגים כלומר הקלט של המודל כבר מתויג בהתאם למציאות. במקרה שלנו הנתונים עליהם יאומן המודל הם בעלי תווית שמציינת האם המודעה היא הונאה או לא ולכן נוכל לומר שמדובר בלמידה מפוקחת. סיווג בינארי הוא סוג של למידה מפוקחת שבה המודל חייב לסווג את הקלטים שהוא מקבל לאחת משתי קטגוריות במקרה שלנו מדובר בסיווג בינארי בו הקטגוריות הן "מודעה מזויפת" ו"מודעה אמתית".

**Exploratory data analysis**

**סקירה כללית של המשתנים ב - data set מצורפת בטבלה הבאה :**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **שם המשתנה** | **מסביר/מוסבר** | **סוג המשתנה** | **Null values amount** | **טווח ערכים** | **תיאור המשתנה** | **הקשר של המשתנה למשימת הלמידה** |
| job\_id | אינדקס | רציף | 0 |  | מזהה ייחודי לכל מודעה | מזהה ייחודי לכל רשומה בדאטה ייתכן ולא יעשה בו שימוש למשימת הלמידה |
| title | מסביר | טקסטואלי | 0 |  | כותרת המודעה | כותרת המודעה – אורכה והמילים הנפוצות בה עלולות לסייע בזיהוי מודעות מזויפות |
| location | מסביר | קטגוריאלי | 281 | 2749 unique locations | מיקום המפרסם | המיקום יוכל לסייע באיתור המקומות בהם תופעת המודעות המזויפות רווחת |
| department | מסביר | קטגוריאלי | 9226 | 1152 unique Departments | המחלקה בה נמצא התפקיד המפורסם במודעה | כמות הערכים החסרים בשדה זה גדולה וכך גם השוני בין הערכים הקיימים |
| salary\_range | מסביר | קטגוריאלי | 11991 | 782 unique  Ranges -  Notice that not all ranges completely numeric | טווח השכר המוצע במודעה | כמות הערכים החסרים בשדה זה גדולה וכך גם השוני בין הערכים הקיימים |
| company\_profile | מסביר | טקסטואלי | 2684 |  | תיאור קצר אודות החברה המפרסמת | קיום תיאור של החברה המפרסמת, זיהוי אורך ממוצע של התיאור וזיהוי מילים נפוצות בתיאור במודעות שקריות יעזרו לנו במשימת הסיווג |
| description | מסביר | טקסטואלי | 1 |  | תיאור המשרה המוצעת במודעה | זיהוי אורך ממוצע של תיאורי משרה אמתיים וזיהוי מילים נפוצות במודעות שקריות יכולים להועיל למשימת הסיווג. |
| Requirements | מסביר | טקסטואלי | 2181 |  | תיאור רשימת הדרישות עבור המשרה | רשימת דרישות דלה או ארוכה באופן חריג תוכל לעזור בזיהוי מודעות מזויפות |
| benefits | מסביר | טקסטואלי | 5744 |  | רשימת הטבות המוצעת על ידי המעסיק | רשימת הטבות דלה או ארוכה באופן חריג תוכל לרמוז על רמת האמינות של המודעה |
| telecommuting | מסביר | בינארי | 0 | 0/1 | 1 במידה והמשרה מציעה עבודה מקוונת | באמצעות בחינת כמות המשרות המקוונות בדאטה והבחנה בין התפלגות המודעות המקוונות המזויפות והאמיתיות נוכל להבין אם שדה זה מהווה אינדיקציה שיכולה לשמש את המערכת הלומדת שלנו. |
| has\_company\_logo | מסביר | בינארי | 0 | 0/1 | 1 במידה ולוגו החברה מופיע במודעה | בחינת המצאות לוגו החברה או חסרונו במודעות אמיתיות ומזויפות יוכלו לעזור לנו בסיווג המודעה |
| has\_questions | מסביר | בינארי | 0 | 0/1 | 1 במידה ובמודעה מופיעות שאלות סינון | בחינת המצאות שאלות סינון במודעות אמיתיות ומזויפות יוכלו לעזור לנו בסיווג המודעה |
| employment\_type | מסביר | קטגוריאלי | 2768 | 5 unique values (Full-time, contract, Part-time, Temporary, Other) | סוג המשרה (מלאה / חלקית / התמחות וכ"ו ...) | בחינת סוגי המשרות הנפוצים במודעות מזויפות ואמיתיות יוכלו לעזור לנו במשימת הסיווג. |
| required\_experience | מסביר | קטגוריאלי | 5622 | 7 unique values (Executive, Mid-senior level,  Entry level, Associate…) | ניסיון דרוש למשרה | חוסר מתאם בין דרישות הניסיון למאפיינים אחרים במודעה יוכלו לעזור בזיהוי של משרות מזויפות. |
| required\_education | מסביר | קטגוריאלי | 6490 | 13 unique values (Bachelor's Degree, High School,  Unspecified…) | רמת השכלה דרושה למשרה | חוסר מתאם בין רמת ההשכלה הדרושה למאפיינים אחרים במודעה יוכלו לעזור בזיהוי משרות מזויפות. |
| industry | מסביר | קטגוריאלי | 3929 | 129 different Industry values | התעשייה אליה מסווגת המשרה | נבחן האם קיימות תעשיות בהן תופעת המודעות המזויפות רווחת יותר. |
| function | מסביר | קטגוריאלי | 5164 | 38 unique values of function names | אופי המשרה | נבחן עבור אילו סוגי משרות תופעת הזיוף יותר שכיחה. |
| fraudulent | מוסבר | בינארי | 0 | 0/1 | 1 במידה והמודעה היא זיוף | באמצעות משתנה המטרה נבחן את איזון הדאטה |

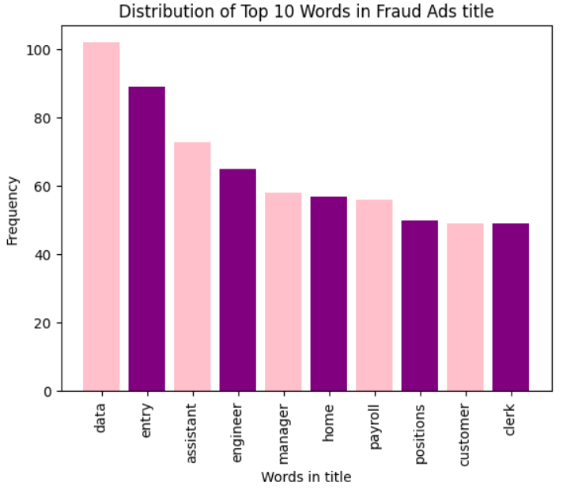
**ניתוח מעמיק של המשתנים -** תחילה בחנו את התפלגות משתנה המטרה בדאטה

מצאנו כי 4.9% (705) מהרשומות הן רשומות המייצגות מודעות מזויפות ו95.1% (13599) מהרשומות הן רשומות שמייצגות מודעות אמיתיות

הבחנו כי הדאטה לא מאוזן באופן משמעותי ונתייחס לכך בהמשך.

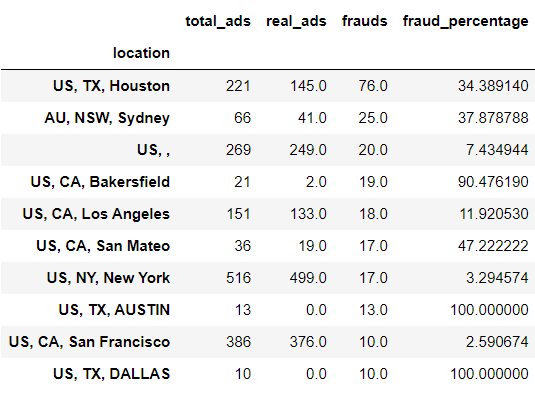
בשלב הבא בחנו משתנים שונים בדאטה על פי פרמטרים שונים בניסיון להעמיק ולהבין את הדאטה שלרשותנו יותר טוב.



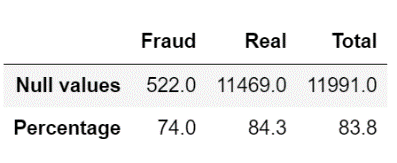
**Job title analysis -** עבור כותרת המודעה בחנו את כמות המילים הממוצעת עבור מודעות מזויפות ועבור מודעות אמיתיות , מצאנו כי ההבדל בין הממוצעים לא משמעותי במידה רבה

בנוסף חילצנו את עשרת המילים הנפוצות ביותר בכותרת של מודעות מזוייפות ובחנו את השכיחות שלהן בדאטה באמצעות הגרף המצורף.

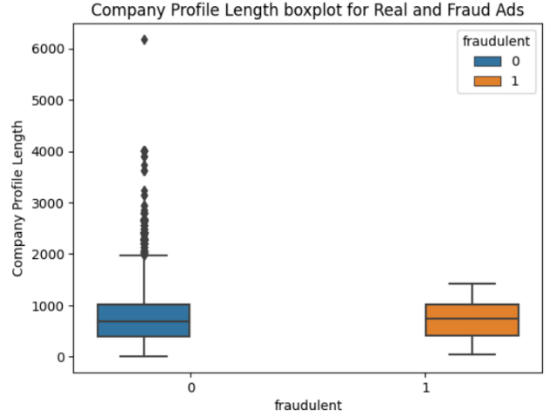
נוסף על כך חילצנו את כל המילים שמופיעות רק במודעות מזויפות ולא הופיעו מעולם במודעות אמיתיות.

**Location Analysis -** עבור שדה המיקום מצאנו כי ישנם 2749 מיקומים ייחודייםוכי קיימות 281 מודעות ללא מיקום, מתוכן 15 מודעות בלבד שהן מודעות מזוייפות.מצאנו את עשרת המיקומים בהם כמות המודעות המזויפות היא הגדולה ביותר והצגנו את הנתונים בטבלה הסיכומית המצורפת.

**Department Analysis -** עבור שדה זה זיהינו כמות ערכים חסרים משמעותית לצד חוסר אחידות. בשדה מחלקה יכולה להיות כמספר המחלקה או כשם המחלקה והמשמעות עלולה להיות שונה עבור ארגונים שונים , נוסף על כך קיימים 1152 ערכים ייחודיים המייצגים מחלקות שונות לכן כמות הרשומות בכל קטגוריה לא גדולה ומשמעותית מספיק. ניכר כי היכולת של המשתנה הזה לתרום למודל שלנו תהיה מוגבלת לכן סביר שנסיר אותו בהמשך.



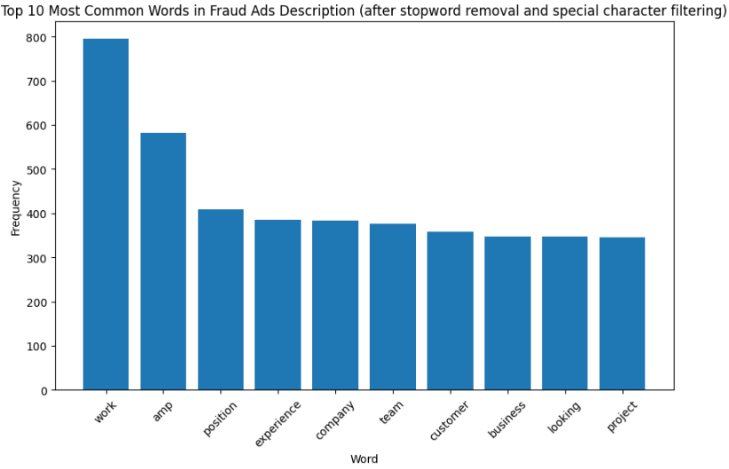
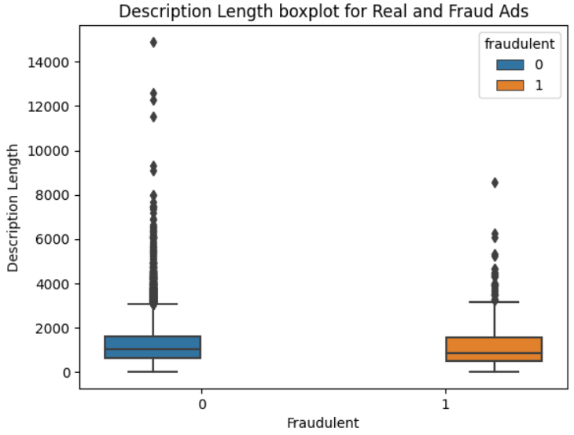
**Salary range Analysis –** גם בשדה זה זיהינו כמות ערכים חסרים גדולה מאוד לצד כמות קטגוריות גדולה ושונות רחבה בערכים. ניכר כי היכולת של המשתנה הזה לתרום למודל שלנו תהיה מוגבלת לכן סביר שנסיר אותו בהמשך. התפלגות הערכים החסרים מוצגת בטבלה המצורפת.

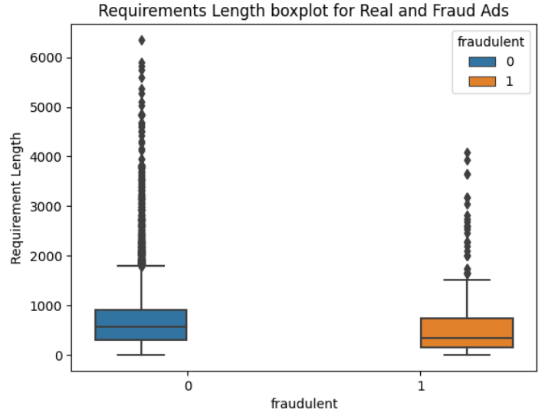


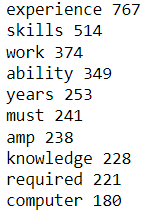
**Company profile Analysis -** עבור פרופיל החברה בחנו את התפלגות האורך של השדה עבור מודעות אמיתיות ומזויפות עשינו זאת באמצעות ה- Box plot ובאמצעות בחינת הממוצעים בקטגוריות השונות.

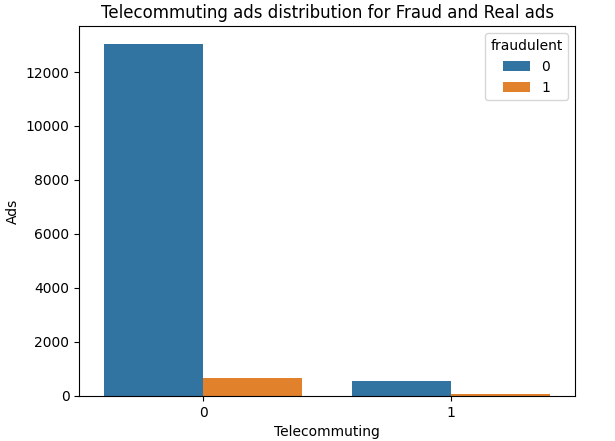
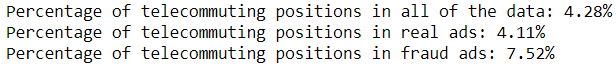
התפלגויות האורכים דומות אך ניכר פיזור גדול יותר במודעות האמיתיות בנוסף האורך הממוצע של פרופיל החברה (764) במודעות אמיתיות גדול יותר מבמודעות מזויפות (717).

**Description Analysis -** עבור שדה תיאור המשרה בחנו את עשרת המילים הנפוצות ביותר במודעות מזוייפות ואת שכיחות הופעתן.בנוסף בחנו את התפלגות האורך של השדה עבור מודעות אמיתיות ומזויפות באמצעות ה- Box plot המצורף. מסקנותינו עבור שדה זה הן כי התפלגויות אורך התיאור דומות במודעות מזויפות ואמיתיות אך קיים הבדל בפיזור הנתונים – עבור מודעות אמיתיות הפיזור גדול יותר. באופן שתואם את הגרף מצאנו הבדל בין ממוצע האורכים כאשר האורך הממוצע של תיאור משרה אמיתית ארוך יותר מתיאור של משרה מזויפת.

****

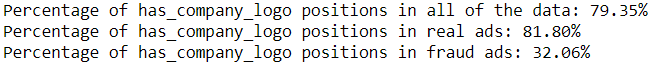
**Requirements Analysis -** עבור שדה הדרישות ביצענו בדיקה דומה לזו שביצענו עבור פרופיל החברה – בדיקה הבוחנת את התפלגות אורך השדה באמצעות תרשים Box Plot וגם כאן ההתפלגות בקרב המודעות המזויפות והאמיתיות היו דומות.

בנוסף בחנו את רשימת 10 המילים הנפוצות ביותר במודעות מזויפות

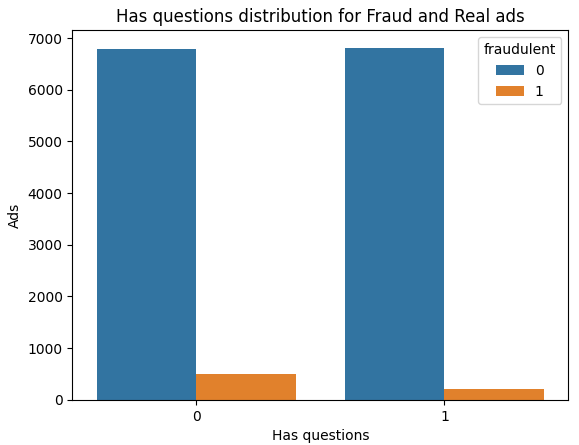
******Telecommuting Analysis -** עבור המשתנה הבינארי של עבודה מהבית בדקנו את אחוזי המשרות שמציעות עבודה מהבית בדאטה כולו במשרות מזויפות בלבד ובאמיתיות בלבד

התוצאות מוצגות בטבלה ובגרף המצורפים.

משניהם רואים באופן ברור את העובדה שרוב המשרות (אמיתיות ומזויפות) הן לא עבודה מהבית.

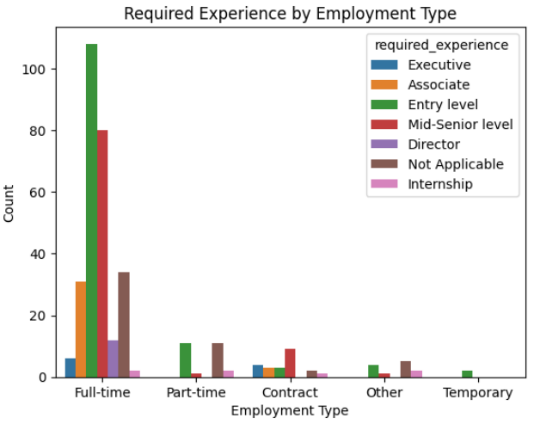
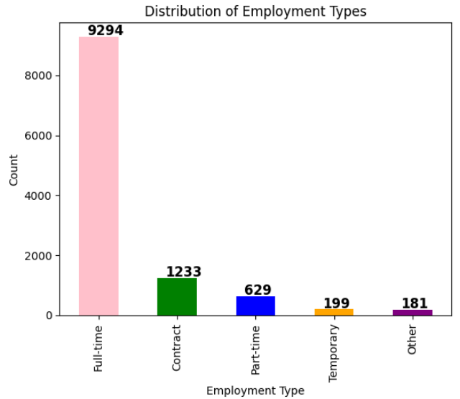
**Has Company logo analysis –**

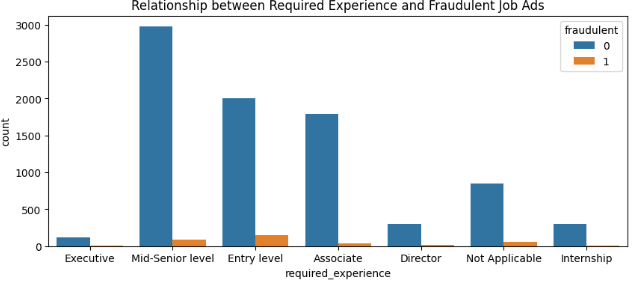
****עבור המשתנה של המצאות לוגו החברה בדקנו את אותם הדברים, מצאנו שבמשרות המזויפות אחוזי ההופעה של הלוגו הם קטנים יותר כלומר ברוב המשרות המזויפות לא יופיע לוגו החברה בעוד שברוב המשרות האמיתיות כן יופיע לוגו החברה.

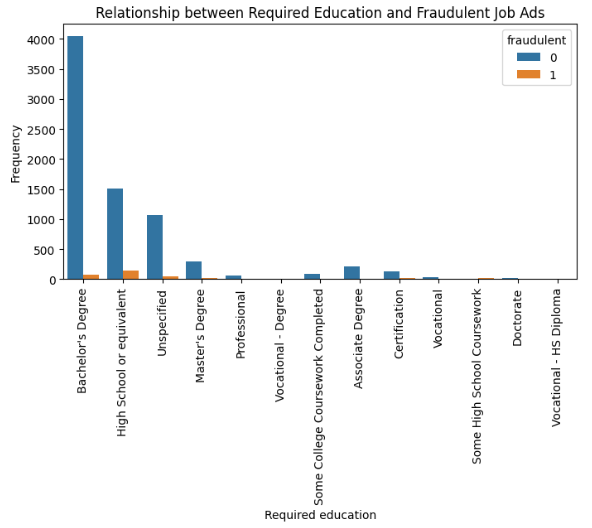


**Has questions analysis -** את משתנה זה בחנו באופן דומה לשני המשתנים הבינאריים האחרים. כאן מצאנו שבמשרות אמיתיות הסיכוי שיהיו שאלות סינון במודעה הוא בערך 50% בעוד שהסיכוי שיופיעו שאלות סינון במודעות מזויפות הוא 28% בלבד. גם כאן בחרנו להציג את המידע באמצעות גרף עמודות :

**Employment type analysis –** עבור המשתנה הקטגוריאלי בחנו את שכיחות ההופעה של הקטגוריות באמצעות גרף העמודות המצורף משמאל.

****בחנו את הקטגוריות גם תוך התייחסות למודעות מזויפות ואמיתיות ולא ניכר קשר מיוחד. בנוסף בחנו את הקשר בין הניסיון הדרוש לבין סוגי העבודה עבור מודעות מזויפות בלבד וזיהינו כי עבור רוב הקטגוריות לא דרוש ניסיון או דרוש ניסיון דל (גרף ימני).

**Required Experience -** בחנו את שכיחות הקטגוריות השונות בדאטה במודעות מזויפות ואמיתיות וראינו שהדרישה הנפוצה ביותר במודעות אמיתיות היא mid-senior בעוד שבמודעות מזויפות זו דווקא Entry-level באופן הגיוני במודעות מזויפות הניסיון הדרוש הוא דל יותר כפי שראינו גם בגרף הקודם.

**Required Education Analysis -** בחנו את שכיחות הקטגוריות במודעות המזויפות והשוונו באמצעות גרף לשכיחות הקטגוריות במודעות האמיתיות מצאנו שהדרישה הנפוצה ביותר במודעות המזויפות היא High school ובאמיתיות היא Bachelor’s Degree מה שכמובן מתיישב עם ההיגיון שבמודעות מזויפות רמת ההשכלה הדרושה תהיה נמוכה יותר כדי לנסות ללכוד כמה שיותר קורבנות.

**קשרים בין משתנים -** בחנו את הקשרים בין המשתנים הטקסטואלים על ידי חיפוש המצאות מילות מפתח משדה אחד בשדה האחר. בחנו האם עבור מודעות מזויפות תהיה חוסר התאמה בין השדות הטקסטואליים.

**Function words in The Title -** בחנו את כמות הכותרות בהן מופיעה מילה שתואמת לשדה function מצאנו שגם במודעות אמיתיות וגם במזויפות האחוזים די דומים.

**Required Experience in the description -** בחנו את כמות התיאורים בהם מופיעה מילה שמציינת את הניסיון הדרוש גם כאן האחוזים היו דומים בין מודעות מזויפות (7%) למודעות אמיתיות (10%).

**Industry in the description -** בחנו את כמות הרשומות בהן בשדה התיאור מופיע איזכור לתעשייה, כמות המודעות בהן יש התאמה בין התיאור לתעשייה עבור מודעות מזויפות היא 25.6% ועבור מודעות אמיתיות היא 31.4% אחוז מעט גבוה יותר.

**Department in the description -** באופן דומה בחנו את כמות הרשומות בהן המחלקה מופיעה בשדה התיאור וכאן מצאנו שהכמות עבור מודעות מזויפות ואמיתיות כמעט זהה בערך 17% . באופן שאינו דומה את הציפיה שלנו מידת ההתאמה במודעות המזויפות הייתה מעט גבוהה יותר.

**Pre-processing**

**בדיקת יתירות בדאטה – Redundancy check in the data -** כדי לבחון האם בדאטה קיימות כפילויות או יתירות מידע החלטנו להגדיר את השדות העיקריים שיגדירו באופן חד חד ערכי מודעה במקום המזהה הייחודי. במקרים בהם עבור מספר מודעות השדות הללו היו זהים השארנו רשומה אחת בלבד כזו.השדות שמגדירים רשומה באופן חד חד ערכי :

Title, location, company profile, description, requirements, employment type, Industry, function

בחרנו בשדות אלה מתוך הנחה שצירוף של שדות אלה מספק על מנת לזהות שאכן מדובר באותה המשרה.

**טיפול בערכים חסרים – Missing Values -** בשלב זה התמודדנו עם הערכים החסרים בדאטה. עבור השדות הקטגוריאלייםlocation, department, salary range) employment type, required experience, required education, Industry and function) השלמנו את הערכים החסרים במילה Unknown שתהווה קטגוריה נוספת. עבור השדות הטסקטואליים (Company profile, description, requirements and benefits) המרנו את הערכים החסרים לסטרינגים ריקים.

**ניהול הדאטה בצורה יעילה - Manage data efficiently -** בשלב זה איחדנו קטגוריות עבור המשתנים הקטגוריאליים בהם זיהינו ריבוי קטגוריות בהתאם לידע קודם ומחקר באינטרנט לגבי סיווג הקטגוריות שברשותנו לקטגוריות רחבות יותר.

**Location** - את משתנה קטגוריאלי זה צמצמנו מ 2749 קטגוריות ל – 5 קטגוריות מרכזיות לפי אזורים בעולם: Americas, EMEA, Africa, APAC, Unknown. עבור כל קטגוריה יצרנו משתנה דמי שמקבל 1 במידה והרשומה שייכת לאזור.

**Salary range** – שדה זה הכיל 782 טווחי משכורות שונים סיווגנו כל טווח לאחת מ 4 קטגוריות : small, medium, large, and Unknown. עבור כל קטגוריה יצרנו משתני דמי שיקבלו 1 אם טווח השכר עונה לתנאים שהגדרנו לכל קטגוריה.

**Employment type** – עבור 5 הקטגוריות של משתנה זה יצרנו משתני דמי שמקבלים 1 במידה והרשומה שייכת לקטגוריה.

**Required Experience -** עבור 7 הקטגוריות של משתנה זה יצרנו משתני דמי שמקבלים 1 במידה והרשומה שייכת.

**Required Education** – צמצמנו את מספר הקטגוריות של משתנה זה מ13 ל 7 הקטגוריות הבאות:

Unknown, Higher education, High school Education, Professional Education, Collage Education, Associate Education, Certification Education והוספנו משתנה דמי עבור כל קטגוריה.

**Industry** – המשתנה הכיל 129 ערכים שונים, קיבצנו את הערכים ל11 קטגוריות והוספנו משתני דמי לכל קטגוריה.

**Function -** המשתנה הכיל 38 ערכים שונים, קיבצנו את הערכים ל10 קטגוריות והוספנו משתני דמי לכל קטגוריה.  
**איזון הדאטה – Imbalanced Data -** כפי שזיהינו בתחילת שלב ה EDA הדאטה שלנו לא מאוזן באופן משמעותי, פחות מחמישה אחוזים ממודעות הדרושים בדאטה הן מודעות מזויפות. בדקנו האם פילוח המודעות בדאטה תואם את המצב במציאות ובחיפוש שערכנו מצאנו כי מחקר שנערך על ידי אתר חיפוש העבודה Flex Jobs בין 20-30% ממודעות הדרושים בלוחות דרושים פופלריים הן הונאה.על סמך נתון זה וכדי שסט הנתונים שלנו יתאם את המציאות החלטנו לאזן את הדאטה באמצעות שיטת Upsampling בה מחוללים רשומות נוספות עבור המחלקה הקטנה. חוללנו 2437 רשומות של מודעות מזויפות וכעת הדאטה תואם באיזונו את מצב העולם האמיתי.

A picture containing calendar

Description automatically generated**סט בדיקה וסט אימון – Test and Training set perspective -** סט האימון הוא סט הנתונים אותו המודל שלנו ילמד ועליו יתאמן. לאחר אימון המודל סט הבדיקה ישמש לבדיקת דיוק המודל. אם סט האימון קטן מידי ייתכן שהוא לא יספיק למודל כדי ללמוד ביעילות. מצד שני, אם סט הבדיקות יהיה קטן מידי המדדים שנחשב בעזרתו ובין היתר דיוק המודל יהיו לא אמינים.בבדיקה שערכנו מצאנו כי מומלץ להשתמש ב-80% מסך הנתונים כסט אימון וב-20% הנותרים לבדיקות ואכן כך חילקנו.

**Segmentation -** בחרנו לדלג על שלב הסגמנטציה משתי סיבות עיקריות – תחילה כמות הנתונים וכמות הפיצ'רים בדאטה שקיבלנו היא לא גדולה מאוד לכן נוכל לבחון את הפיצ'רים הכי רלוונטיים גם מבלי לפלח את הנתונים לתתי קטגוריות. שנית, בשלב ה EDA לא מצאנו קשרים חד משמעיים בין הפיצ'רים לכן סביר שתהליך סגמנטציה לא יספק לנו תובנות או יתרונות חדשים.

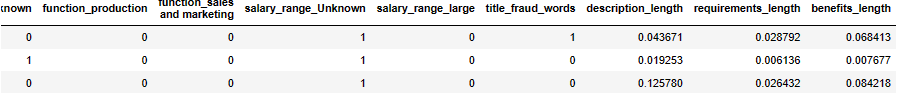
Graphical user interface, text, application

Description automatically generated **Feature Extraction-** בשלב זה הגדרנו פיצ'רים נוספים שאנחנו משערות שיוכלו לעזור למודל שלנו במשימת הלמידה. עבור כל אחד מהשדות הטקסטואליים (company profile, Description, Requirements, Benefits) הוספנו פיצ'ר שמייצג את אורך השדה ופיצ'ר נוסף שמייצג את כמות המילים בשדה. נוסף על כך עבור השדה title הוספנו פיצ'ר שמקבל 1 במידה והכותרת מכילה מילה מרשימת המילים שמופיעות רק בכותרת של מודעות מזויפות ו0 אחרת.

**A screenshot of a computer

Description automatically generatedFeature Representation -** כלל הפיצ'רים הנומריים שעומדים לרשותנו כעת הם בטווח ערכים של 0-1 , מלבד הפיצ'רים שהוספנו של אורך ומספר מילים עבור השדות הטקסטואליים. כדי לנרמל את הפיצ'רים החדשים לטווח הערכים הרצוי חילקנו את ערכי השדה בערך המקסימלי של אותו הפיצ'ר. כלומר את ערכי אורך הטקסט חילקנו באורך הטקסט המקסימלי ואת כמות המילים חילקנו בכמות המילים המקסימלית.

**Feature Selection -** בשלב זה הסרנו פיצ'רים שאנחנו משערות שלא יוכלו לעזור לנו במשימה הלמידה. הסרנו פיצ'רים בעלי כמות ערכים חסרים גבוהה מאוד, פיצ'רים אשר ניכרת בהם חוסר אחידות בData וריבוי קטגוריות אשר לדעתנו לא יתרמו לנו למשימת הלמידה. בנוסף הסרנו פיצ'רים שלהם הוצאנו משתני דמי קודם לכן בתהליך. בנוסף, בשלב זה ביצענו בדיקת קורלציה בין כל 2 פיצ'רים שונים ובין הפיצ'רים לבין משתנה המטרה שלנו – ‘fraudulent’. בדקנו האם קיימים זוגות של פיצ'רים בעלי קורלציה גבוהה מ0.5 ואם כן בדקנו את הקורלציה שלהם מול משתנה המטרה fraudulent ובחרנו להסיר את הפיצ'ר בעל הקורלציה הנמוכה יותר מול משתנה המטרה. כך שבסוף שלב זה נותרנו עם 79 פיצ'רים.



**Dimensionality Reduction** - בשלב זה מבצעים הורדת מימד. במקרה שלנו כמות הפיצ'רים עומדת על 79 בסיום שלב ה feature selection, זוהי כמות לא גדולה של פיצ'רים ביחס לכמות הנתונים שיש לנו ב Data ולכן בחרנו לדלג על שלב זה ולא לבצע סינון נוסף.

**Model Training -** שיטת הולדיציה שבחרנו היא K-Fold splitting שיטה זו מספקת הערכה יותר אמינה מהשיטות האחרות לביצועי המודל מכיוון שהיא משתמשת בכל הדאטה גם לאימון וגם לולידציות, בנוסף היא מפחיתה את השונות של הערכת הביצועים על ידי שימוש בממוצע הערכים מהאיטרציות השונות.

תהליך הולדיציה שנבצע על פי שיטת K-Fold כולל חלוקה של הדאטה ל K חלקים שווים,

בכל איטרציה המודל מתאמן על K-1 קבוצות ומתאמן על הקבוצה שנותרה. נחזור על התהליך K פעמים באופן

כזה שכל סט נתונים משמש לולידציה פעם אחת. בכל איטרציה נחשב את מטריקת AUC.