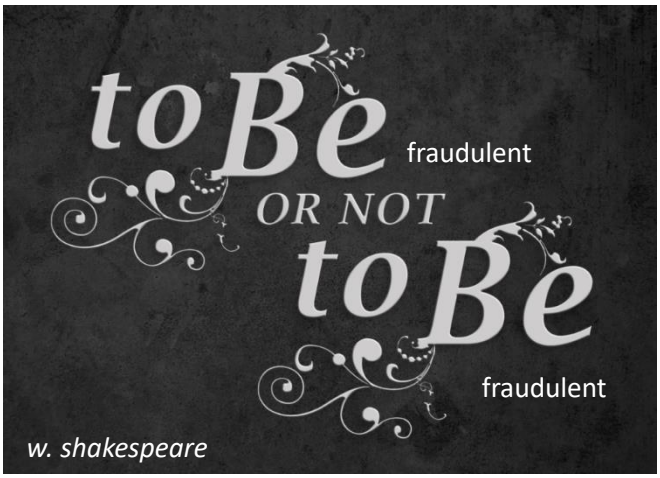
# פרויקט לימוד מכונה

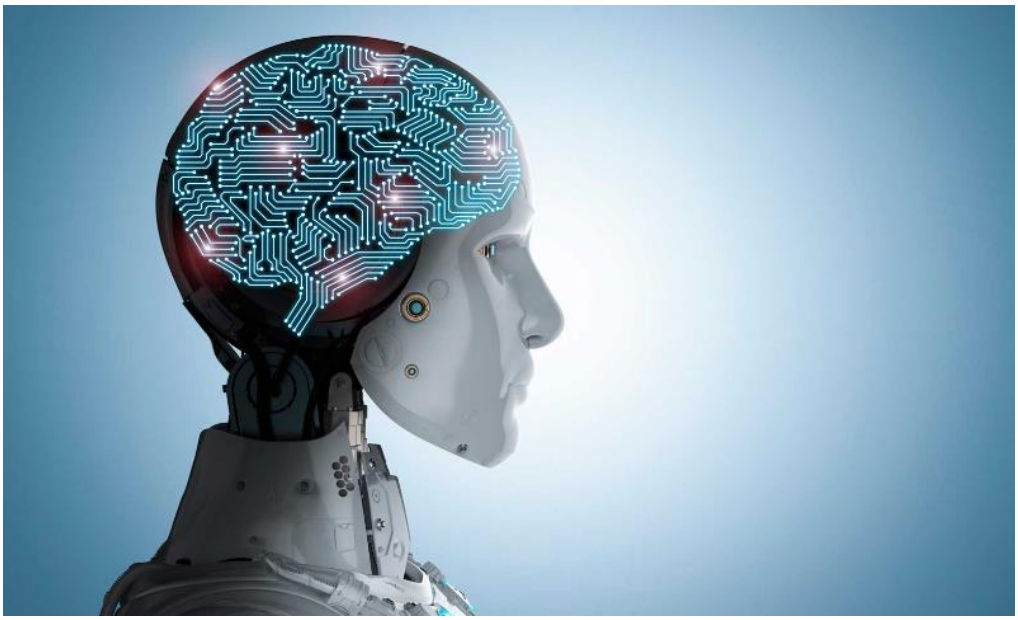
תאריך הגשה: 16/04/2023

מספר קבוצה: 30

מגישים: דור קורנט - 318871282

גיא ברון - 205984503



**תוכן עניינים:**

## Data collection and Sensing:

1. Data collection זהו תהליך של איסוף נתונים ממקורות שונים. ב- Data collection נאספים סמפלים מאותו התחום והם מייצגים את העולם האמיתי אותו אנחנו רוצים ללמוד. ה- Sensing שבוצע על הדאטה הוא Static Sensing מאחר והנתונים על הצעות העבודה הם נתונים קבועים ואינם "התנהגותיים" (דינאמיים) כלומר אינם יכולים להשתנות בכל רגע נתון.
2. מאחר ובוצע Static Sensing על הדאטה, נציע להשתמש ב- Dynamic Sensing. למשל, כמות האתרים והמקומות בהם פורסמה המודעה. על כן, במידה ואנו רואים שהצעת עבודה מסוימת מתפרסמת באופן עקבי במקומות שונים ברחבי באינטרנט יותר מהממוצע או מהמקובל למשרות דומות. נחשוד שמדובר במודעה מזויפת.
3. קטגורית משימת הלמידה היא Supervised Classification. מאחר ואנו יודעים מהם הלייבלים שלנו (חשוד כהונאה או לא חשוד) אנו יודעים כי משימת הלמידה הינה מונחית. בנוסף, המשימה הינה משימת סיווג מאחר ואנחנו רוצים לחלק את הסמפלים ל-2 קטגוריות בלבד. מאחר ואנחנו מקבלים את הסמפלים כאשר הם מסווגים ל"תקין" ו"לא תקין" (fraudulent) משימת הלמידה תהיה מסוג Supervised Anomaly Detection. מכיוון שאין מספיק סמפלים המסווגים כלא תקינים, נוכל ללמוד מהסמפלים התקינים על התנהגותם והסיבה לכך, ומכאן להסיק על התנהגות שאינה תקינה.

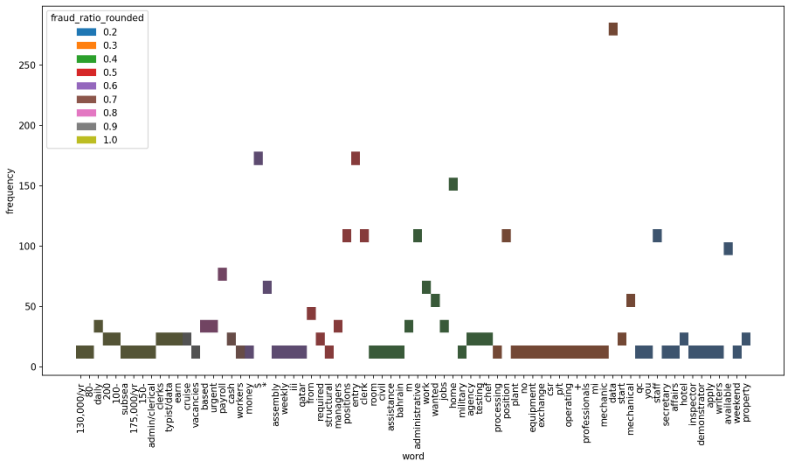
## Dataset Creation:

## Exploratory Data Analysis:

**Title:** בדקנו האם קיימות מילים חשודות, אשר חוזרות על עצמן במודעות שסווגו כשקריות:

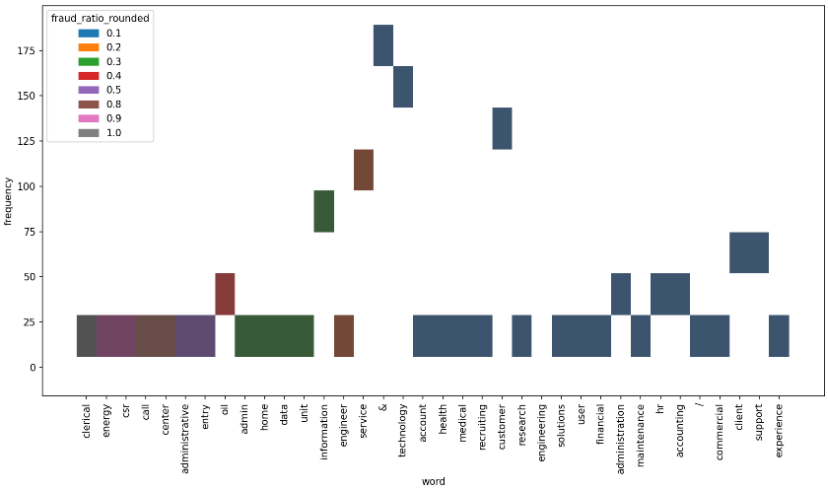
בחרנו להציג מילים שהופיעו מעל 6 פעמים ומעל ל-20% הופעות במודעות כוזבות.

ניתן לראות כי ישנה חזרתיות רבה על מילים מתחום הפיננסים, כגון: "money" ,"weekly" ,"cash", "earn", "daily","130,000/yr" ,"175,000/yr"



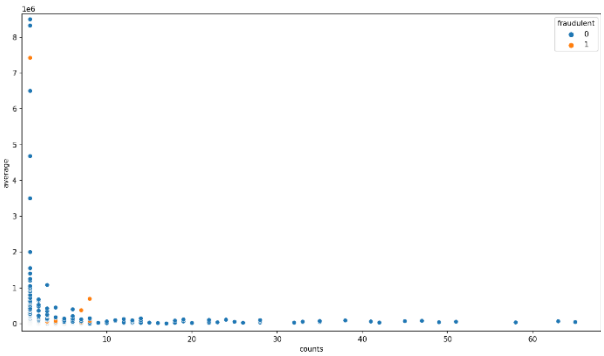
**Department:** מקומות עבודה שונים נוטים לתת שמות שונים ועל כן ישנם המון מחלקות. בדקנו האם קיימות מילים אשר נוטות לחזור על עצמן במודעות שסווגו שקריות.

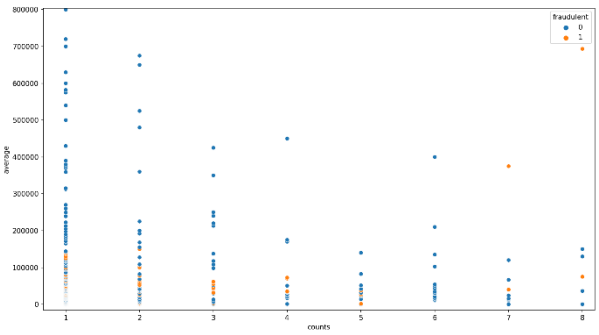
בחרנו להראות מילים באשר הופיעו מעל 5 פעמים והופיעו במודעות כוזבות במעל 5% מהשימוש בהם .




הערה : המילה engineering חזרה 450 פעמים, אך נחתכה מהתמונה כדי לאפשר תצוגה טובה יותר על שאר חלקי הגרף.

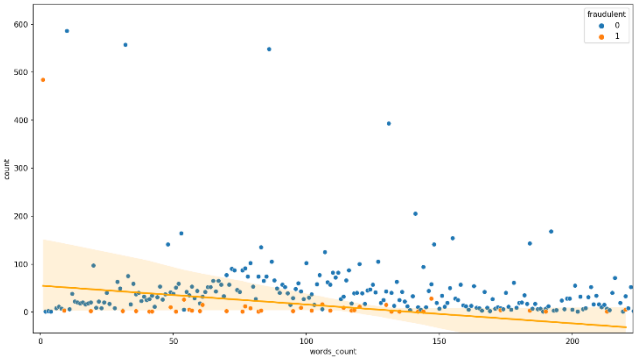
**Salary range**: ביצענו ממוצע לכל משרה והצגנו את התוצאה על הגרף בנוסף התפלגות מס' הפעמים שהממוצע הופיע. לבסוף סיווגנו את הממוצעים למודעות שסווגו כמודעות מזויפות ואמיתיות:

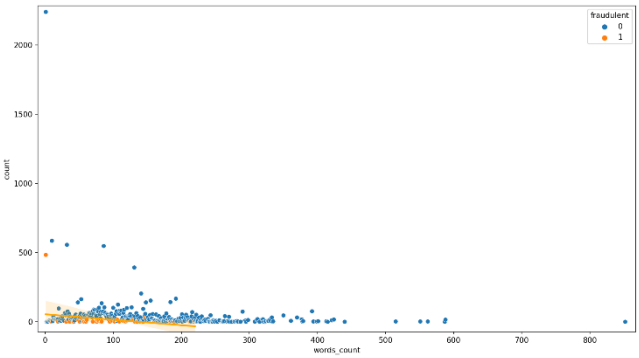



הערה: הגרף הימני הינו תקריב של הגרף השמאלי על מנת לראות את התפלגות הנתונים בשטח הצפוף בצורה ברורה יותר. ניתן לראות כי ישנו טווח ברור בו מתפלגים הערכים במודעות כוזבות, מלבד ערך חריג אחד של 7m.

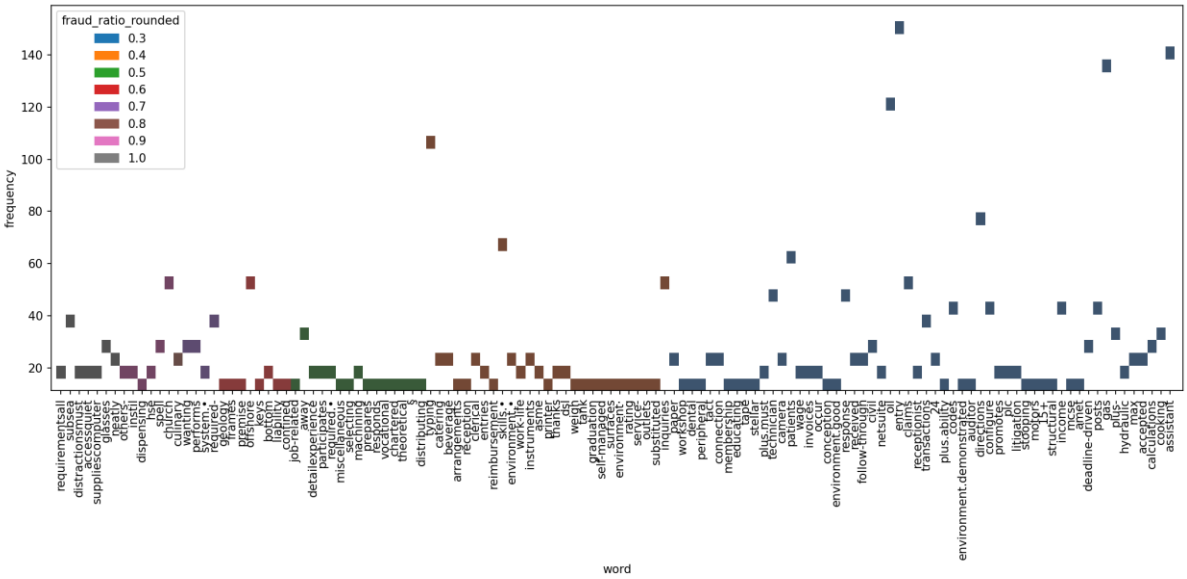
**Company profile:** בחרנו לספור את המילים בכל סמפל ולבדוק מהו מספר המילים במודעות כוזבות ובמודעות אמת. בציר y ספרנו את כמות המודעות עם מספר המילים שמופיעות בציר x.



בחרנו להוסיף עקומה לינארית כדי לבדוק האם קיים קשר בין מספר המילים למספר המופעים. בנוסף ראוי לציין כי ישנו תחום ערכים ברור בו מתכנסים מספר מילים למודעות כוזבות

**<u>Requirements</u>:** רשימת דרישות לתפקיד, בחרנו לבדוק האם יש מילים שחוזרות על עצמן במודעות כוזבות, ובאיזה תדירות. הגרף הבא מציג מילים שהופיעו במעל 25% מודעות כוזבות ומעל 10 הופעות

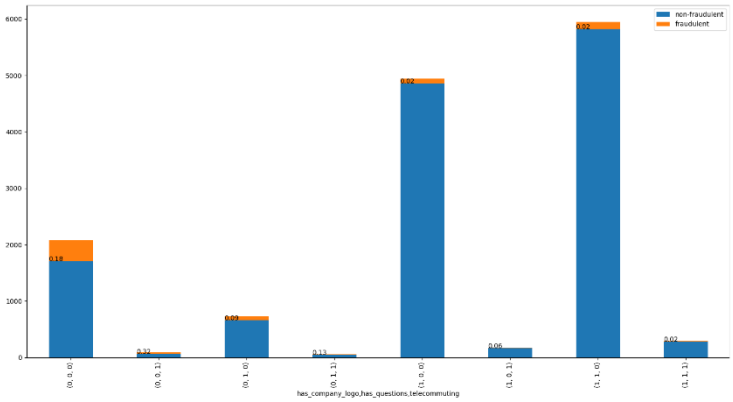



**<u>:Telecommuting, Has company logo , Has questions</u>**

בקובץ הנתונים ישנם שלושה פיצ'רים בינאריים מלבד פיצ'ר המטרה:

1. <u>Telecommuting-</u> משתנה בינארי המעיד על נכונות לעבוד מרחוק.
2. <u>Has company logo-</u> משתנה בינארי המעיד האם לוגו החברה מוצג במודעה.
3. <u>Has questions -</u> פיצ'ר המציין האם במודעה מופיעות שאלות מיון

**<u>כעת נציג אגריגציה בין שלושת המשתנים למול משתנה המטרה:</u>**



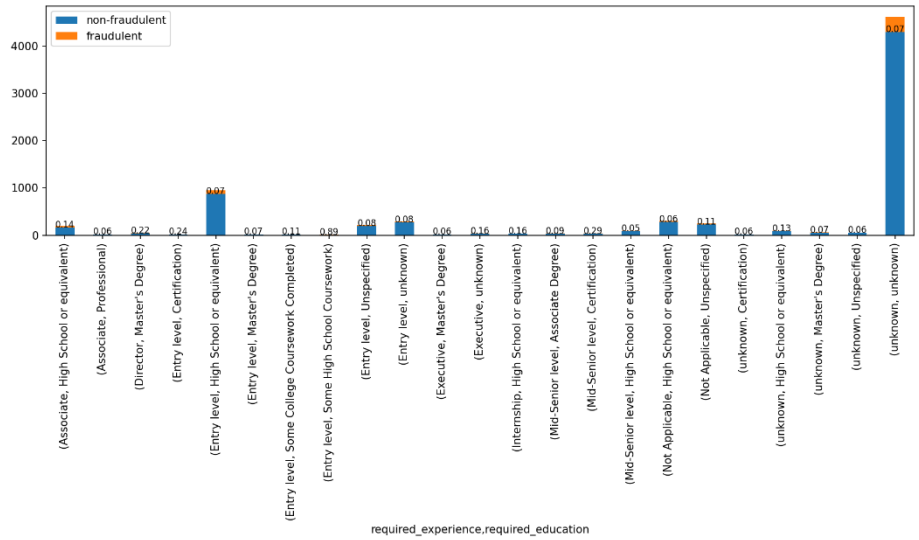

ראשית בהינתן שאין לוגו לחברה, אין שאלות נוספות ויש אפשרות לעבוד מהבית, בסבירות של 32% מדובר במודעה מזויפת, אחוז גבוהה מאוד ביחס לשאר הקומבינציות שהוצגו עד כה.

במידה שישנו לוגו לחברה, הפרמטרים המעידים על שאלות נוספות ואפשרות

לעבוד מהבית אינם משנים את ההסתברות למודעה מזויפת כך שחיבור ביניהם כמעט ולא תורם. <u>הערה</u>: בחלק הנספחים צרפנו התפלגויות נוספות של המשתנים

**<u>Required experience & Required education :</u>** מדובר בפיצ'רים קטגוריאליים

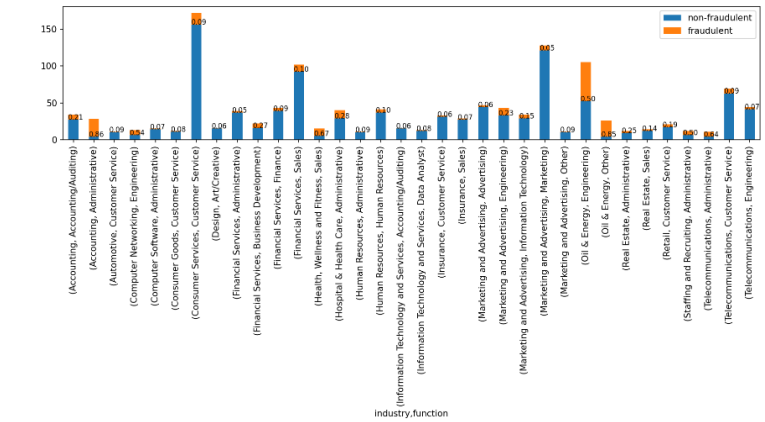
המעידים על הניסיון הדרוש, ומקום הלימוד שהמודעה דורשת. נבחן את השילובים בגרפים מסוג bar ונחפש מגמות המעידות על פרסום מודעה חריגה.




<u>הערה</u>: בחרנו להציג רק אגרגציות בעלות 10 סמפלים או יותר בעלות סיכון של5% ומעלה להיות מודעת כזב.

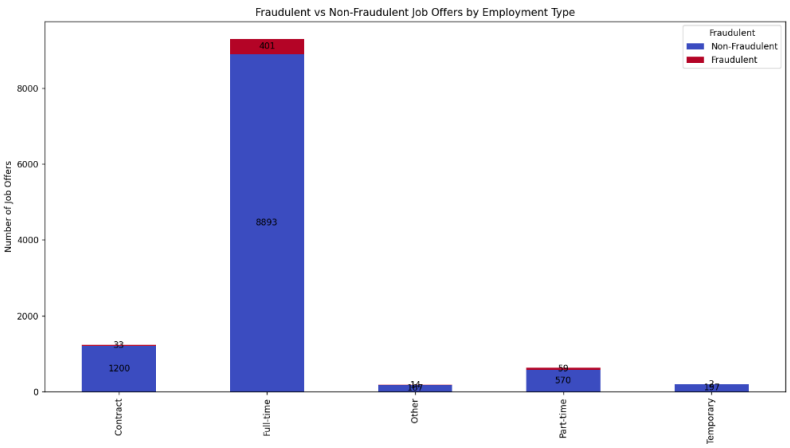
בנספחים ישנם התפלגויות מצומצמות יותר.

**<u>Industry & function:</u>** בחרנו להציג אגרגציה רק במקרים בהם אחוז הסיכוי להיות מודעה כוזבת עולה על 5% ובנוסף ישנם מעל ל10 סמפלים מסוג זה.



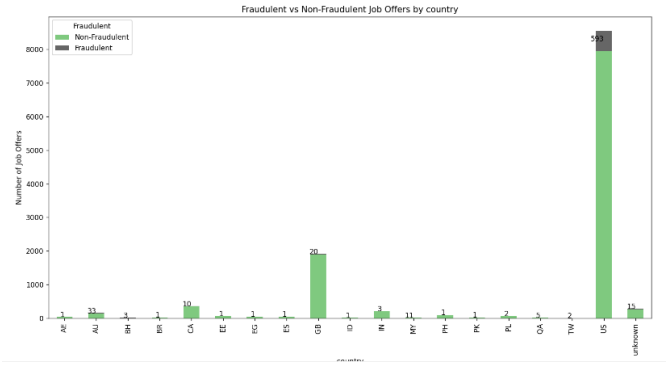

<u>הערה</u>: בנספחים מוצגים התפלגויות כל משתנה בנפרד.

**<u>Employment type:</u>** זהו משתנה קטגוריאלי המציג את סוג המשרה. בגרף הבא ניתן לראות את כמות הפעמים שסוג המשרה מופיע ומתוכן את כמות הפעמים שהמשרה הינה מזויפת ולא מזויפת:

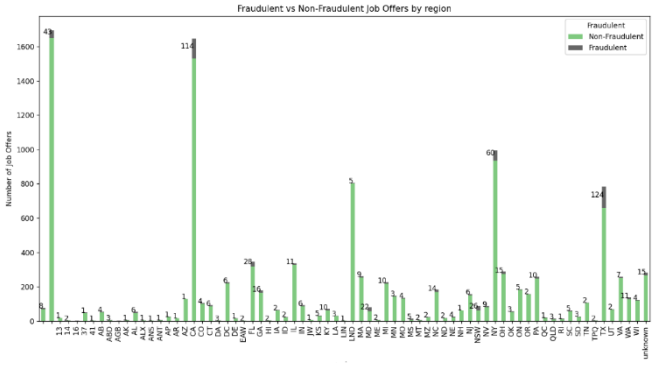


**Location:** את פיצ'ר זה חילקנו ל-3 פיצ'רים. בהצגת הגרפים החלטנו להראות רק את המופעים המופיעים לפחות פעמיים וגם שמופיעים במודעות כוזבות לפחות פעם אחת. לכל אחת מהעמודות החדשות נציג גרף המראה את כמות המופעים הכוללת ואת כמות הפעמים שהמדינה הופעה במודעות שקר לעומת מספר הפעמים שהופיעה במודעות אמת.

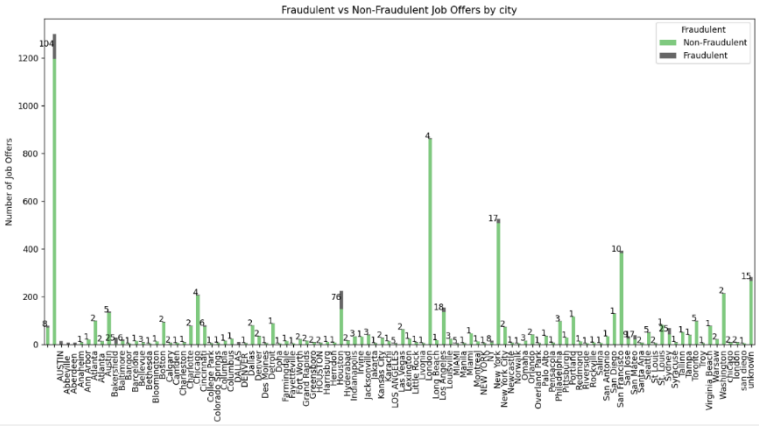
:Country




:Region




:City

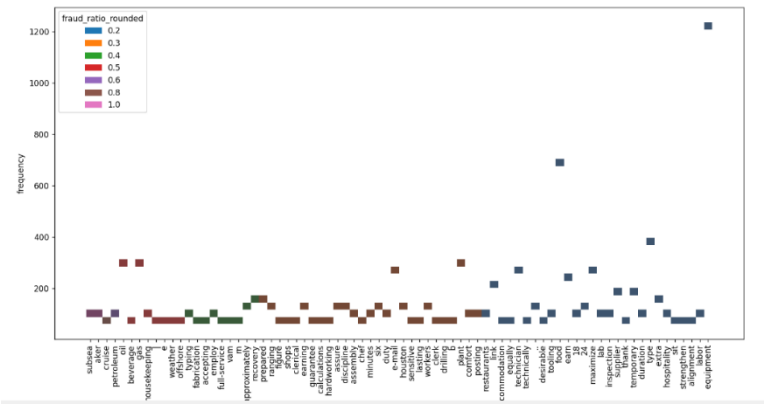


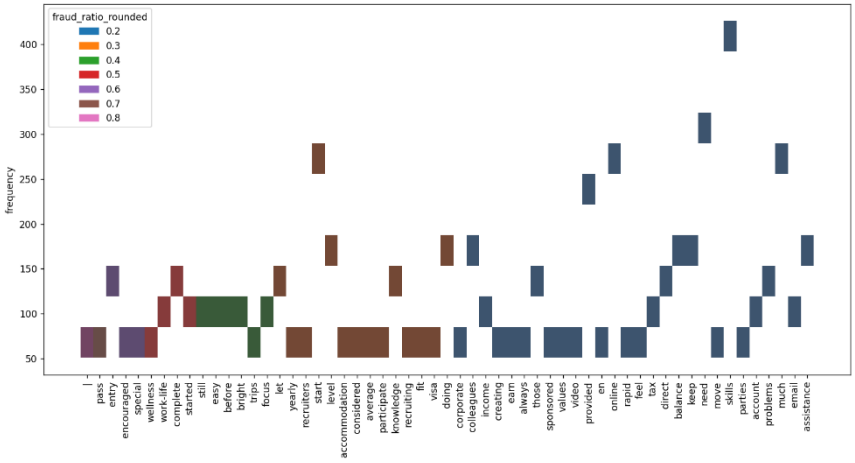

הערות:

1. ישנן בטבלאות מידע נוסף אך הצגנו רק את המידע הרלוונטי עם אחוזי ה-fraudulent הגבוהים ביותר.

2. טבלת addit הכוללת מידע נוסף מוצגת בנספחים.

**Description:** בגרף זה נציג את המילים שחוזרות על עצמן הכי הרבה פעמים במודעות מזויפות כאשר ציר ה-y מסמל את מספר הפעמים שהמילה מופיעה בשדה זה והצבע מעיד על היחס בין מספר הפעמים שהמילה מופיעה במודעות שקר לבין מספר הפעמים שהמילה מופיעה באופן כללי (כפי שנראה במקרא בצד שמאל למעלה). הגרף מסונן לפי מילים שמופיעות לפחות 60 פעמים ולהם מעל 20 אחוז מילים המופיעות במודעות מזויפות.






**Benefits**: בגרף זה נציג את המילים שחוזרות על עצמן הכי הרבה פעמים במודעות מזויפות כאשר ציר ה-y מסמל את מספר הפעמים שהמילה מופיעה בשדה זה והצבע מעיד על היחס בין מספר הפעמים שהמילה מופיעה במודעות שקר לבין מספר הפעמים שהמילה מופיעה באופן כללי (כפי שנראה במקרא בצד שמאל למעלה). הגרף מסונן לפי מילים שמופיעות לפחות 50 פעמים ולפחות מעל 10 אחוז מופיעות במודעות מזויפות.

- **Pre Processing**:

חזרתיות – אחד הדברים שנרצה לעשות הוא הסרת כפילויות מאחר ואנחנו לא רוצים לקבל מידע כפול על מודעות זהות שכן זה יכול לפגוע באמינות האלגוריתם. נגדיר כפילות בנתונים כך: השדות company profile, description, requirements צריכים להיות זהים מאחר והסבירות שיהיה פרופיל חברה, תיאור משרה ודרישות זהות לשתי משרות שונות הוא נמוך מאוד. (למטה, דוגמא מדוע התבסוסת על 2 שדות בלבד אינה מספיקה).

| | company_profile | description | requirements | bene |
|---|---|---|---|---|
| 30 | Established on the principles that fu... | We are currently recruiting for an exciting S... | Experience in fragrance and sales. | Bonuses are available. |
| 31 | Established on the principles that fu... | We are currently recruiting for an exciting S... | Experience in promotional work fragrance an... | Bonuses may be given. |

השלמת חוסרים – כרגע החלטנו למלא חוסרים על ידי החלפת שדה חסר במילה unknown אך אם במידת הצורך בהמשך הפרויקט נצטרך להשלים חוסרים נעשה זאת על ידי השיטה הבאה: אם נצטרך למלא חוסרים מעל 60% נמחק את הרשומה מאחר ולא ניתן ללמוד ממנה. אחרת, נשלים את החוסרים לפי מודעה דומה אחרת. הפיצ'רים עליהם נסתכל הם: job title, employment type, industry, required education, industry, function, required education. נחפש את המודעה שלה המספר הגבוה ביותר של פיצ'רים דומים ונשלים לפיה.

Data type conversions – אנחנו נרצה להמיר ערכים רציפים וטקסטואליים לערכים קטגוריאליים מאחר ויהיה לנו יותר קל לעבוד איתם ולסווג כל הצעת עבודה לקטגוריה מסוימת. העמודות להן נרצה לעשות קטגוריזציה הן עמודות כמו description, benefits, company profile, requirements שהם עמודות המכילות מלל חופשי ובמצב כזה קשה להסיק עליו מסקנות. נרצה לעשות עליהם מניפולציות מסוימות כמו לסכום את מספר המילים ולסמן מילים מסוכנות

Manage the data efficiently – בשלב ה- EDA ראינו כי פירוק העמודה location ל3 פיצ'רים שונים אשר כל אחד בפני עצמו מבדיל את הקלאס בצורה קונקרטית וברורה יותר.
Proportions in the data – לא ניתן לקבוע כרגע אם ה- data מאוזן או לא. ראינו ש- 5% מהמודעות הן מודעות שקר ו- 95% הן הודעות אמיתיות, אך איננו יודעים אם זה מייצג את העולם האמיתי ולכן לא נעשה דבר בשלב זה.

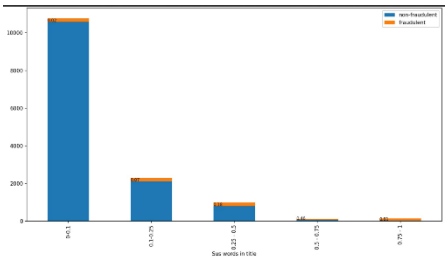
הערה: לא נבצע את שלב ה- Segmentation מהסיבה שאנו סבורים שהוא אינו הכרחי למודל זה. הצעות עבודה מזויפות יכולות ללבוש צורות רבות ושונות ועשויות שלא לעקוב אחר דפוס או מבנה עקביים, מה שעלול להקשות על יישום ה- segmentation. ה- segmentation יהיה יעיל כאשר צריכים לחלץ נתונים מתמונה או קול מהקלטה.

- **Feature Extraction**:

1. Sus words in strings - אנחנו רוצים לסמן "מילים מסוכנות" בעזרת n-grams, כלומר לבדוק כמה מילים מסוכנות מופיעות בשדה זה. נרצה להכיל את שיטה זו על שדות טקסטואליים, כגון title, department, description, requirements, benefits (מדובר ב-5 פיצ'רים, אחד לכל שדה).
2. Sus country/ region /city - נחלק את ה- location לשלוש עמודות: country, region, city. כל עמודה תייצג פיצ'ר. נסמן אילו אזורים מופיעים הכי הרבה במודעות כוזבות ונגדיר אותם כ-"אזורים חשודים".
3. N-words in string - ב- company profile נרצה לבדוק את מספר המילים. נחלק אותם לקטגוריות מ- 0 עד 100, מ-100 עד 200 ומ-200 ואילך, כפי שראינו ב-EDA. נרצה את המידע הזה גם לגבי description, requirements מאחר וגם הם בנויים ממחרוזות ארוכות (כלומר 3 פיצ'רים, אחד לכל שדה).
4. Industry & function category - נעשה קטגוריות של קומבינציות שונות של ערכים מ- industry ו- function. ראינו ב-EDA שישנן קומבינציות בהם אחוז גבוה יותר להופיע במודעות שקר ולכן נוכל לעשות קטגוריזציה כזו.
5. Education and experience mismatch – לכל שדה של ניסיון דרוש ושל השכלה דרושה ניתן ציון ונחשב את ההפרש בין הציונים לכל הצעת עבודה. ככל שההפרש גבוה יותר כך יש פחות תאימות בין הניסיון לבין ההשכלה. נעשה זאת מכיוון שחוסר תאימות בניהם יכול להיות להעלות חשד שמדובר במודעה מזויפת.

- **Feature Representation**:

1. Sus words in strings - נגדיר לכל משפט איזה מבין 3 קטגוריות הוא נמצא, במידה וקיימת בו מילה בעלת סיכון של 50% ומעלה להיות חשודה יוגדר בסיכון גבוהה, בין 10% ל-50% יוגדר בסיכון בינוני, ואחרת בסיכון נמוך. המילים החשודות יהיו מילים אשר הופיעו במעל 10 מודעות שסווגו כמודעות כוזבות.



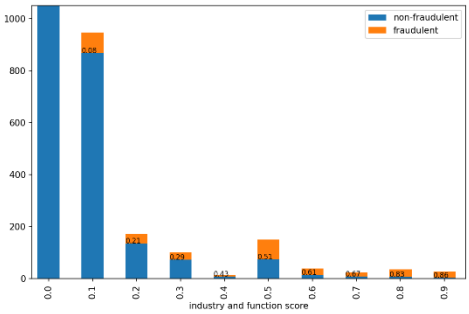

2. Sus country/ region/ city – פיצ'ר זה יכיל ערכים בינאריים, כאשר 1 מסמל אזור מסוכן (מסוכן = מופיע באחוז גבוה במודעות שקר) ו-0 מסמל אזור שאינו מסוכן.

| | job_id | sus_country | sus_region | sus_city |
|---|---|---|---|---|
| 0 | 7851 | 0 | 0 | 0 |
| 1 | 332 | 0 | 0 | 0 |
| 2 | 5445 | 1 | 1 | 0 |
| 3 | 7874 | 0 | 0 | 0 |
| 4 | 15471 | 0 | 0 | 0 |
| 5 | 10902 | 0 | 0 | 0 |
| 6 | 7681 | 0 | 0 | 0 |
| 7 | 2830 | 0 | 0 | 0 |
| 8 | 844 | 0 | 0 | 0 |
| 9 | 15099 | 0 | 0 | 1 |
| 10 | 9436 | 0 | 0 | 1 |

3. N-words in string – בפיצ'ר זה נרשום באיזה קטגוריה נמצא מספר המילים המופיעות במחרוזת, כאשר הקטגוריות הן כפי שציינו: מ-0 עד 100 (less than 100), מ-100 עד 200 (between 100 to 200) ומעל 200 (above 200).

| | job_id | n-words in description |
|---|---|---|
| 0 | 7851 | less than 100 |
| 1 | 332 | less than 100 |
| 2 | 5445 | above 200 |
| 3 | 7874 | above 200 |
| 4 | 15471 | above 200 |
| 5 | 10902 | less than 100 |
| 6 | 7681 | above 200 |
| 7 | 2830 | above 200 |
| 8 | 844 | between 100 and 200 |
| 9 | 15099 | between 100 and 200 |

4. Industry & function category – מאחר והשילובים בין שני השדות הללו מניבים כ-30 קומבינציות, נחלק את הקומבינציות לקבוצות על פי דרגת הסיכון שלהם. הערך של פיצ'ר זה יהיה מספרי, כאשר הערך 0 ייצג קטגוריות ברמת סיכון בין 0 ל-9 אחוזים, הערך 1 ייצג קטגוריות ברמת סיכון בין 10-19 אחוזים וכן הלאה.




5. Education experience mismatch – תחילה ניתן ציון לכל אחד מהערכים מ-0 עד 5, כאשר ככל שהציון נמוך יותר כך ההשכלה הנדרשת תהיה נמוכה יותר והניסיון הנדרש יהיה נמוך יותר. לאחר מכן נעשה הפרש בערך מוחלט בין שני הציונים של ההשכלה והניסיון. הערך של ההפרש גם כן יהיה בין הערכים 0 ל-5, כאשר ציון 0 אומר שישנה תאימות גבוהה בין שני הציונים וככל שהמספר גבוהה יותר כך התאימות יותר נמוכה.

| | job_id | exp_weight | edu_weight | edu_exp_mismatch | fraudulent |
|---|---|---|---|---|---|
| 0 | 7851 | 0 | 0 | 0 | 0 |
| 1 | 332 | 0 | 0 | 0 | 0 |
| 2 | 5445 | 4 | 0 | 4 | 1 |
| 3 | 7874 | 3 | 3 | 0 | 0 |
| 4 | 15471 | 0 | 0 | 0 | 0 |
| 5 | 10902 | 0 | 3 | 3 | 0 |
| 6 | 7681 | 0 | 0 | 0 | 0 |
| 7 | 2830 | 0 | 3 | 3 | 0 |

- **Feature Selection**:

בשלב זה, השתמשנו בשיטת ה- Fischer score לחישוב ציון לכל הפיצ'רים. ככל שהציון גבוהה יותר הפיצ'ר מחלק טוב יותר את המחלקות השונות במשתנה המטרה, כאשר בבחירת הפיצ'רים נרצה לבחור את הפיצ'רים בעלי הציון הכי גבוה. ראוי לציין כי שיטה זו אינה עובדת בצורה טובה עם הפיצ'רים הטקסטואליים המקוריים אך בכל זאת בחרנו

להשתמש בשיטה זו. הסיבה לכך היא שיצרנו פיצ'רים קטגוריאליים לכל משתנה טקסטואלי ולא נשתמש בפיצ'רים הטקסטואליים בצורתם המקורית

הפיצ'רים הטקסטואליים בהם לא נשתמש: department, description, benefits, company_profile, requirements.

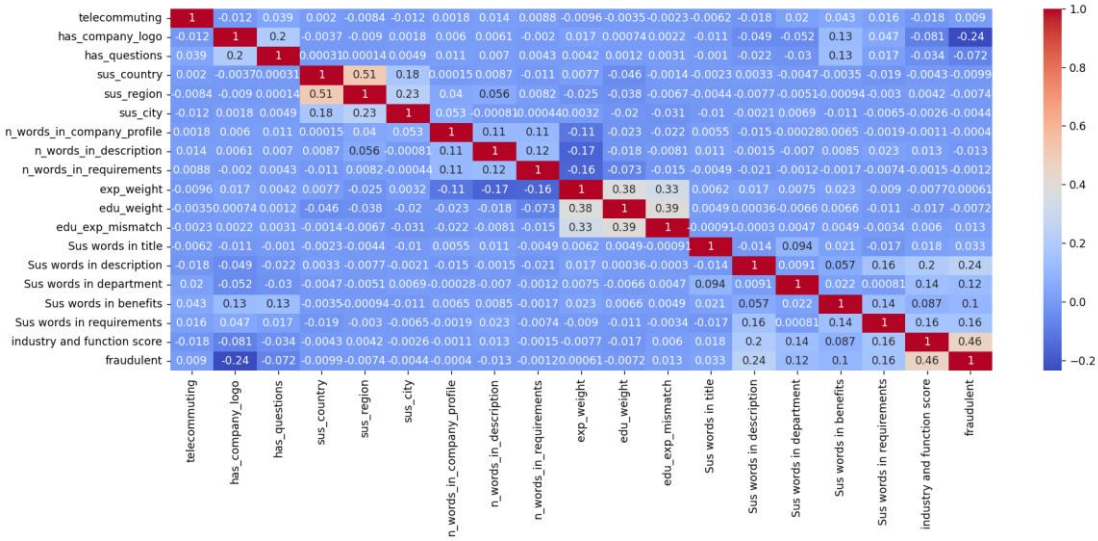
**הניקוד:**

```
Fischer score for sus_country: 0.38657109396271083
Fischer score for sus_region: 0.41926644030686244
Fischer score for sus_city: 2.98351254481507
Fischer score for n_words_in_company_profile: 0.12867333507947937
Fischer score for n_words_in_description: 0.03925719451826053
Fischer score for n_words_in_requirements: 0.01303567712479308
Fischer score for exp_weight: 0.03647527219263489
Fischer score for edu_weight: 0.021103368750091077
Fischer score for edu_exp_mismatch: 0.03829322528334339
Fischer score for Sus words in title: 0.011909024105086395
Fischer score for Sus words in description: 0.46301028710483816
Fischer score for Sus words in department: 0.17417391390987808
Fischer score for Sus words in benefits: 0.25283327888232643
Fischer score for Sus words in requirements: 0.32273098363338903
Fischer score for industry and function score: 0.1595606237460531
```

```
Fischer score for title: 0.0022379233142450373
Fischer score for location: 0.004310945806075221
Fischer score for department: 0.004420599991391179
Fischer score for salary_range: 0.004279440197659252
Fischer score for company_profile: 0.005833431624559669
Fischer score for description: 0.0018371108368801215
Fischer score for requirements: 0.0018584362543660803
Fischer score for benefits: 0.0024118193004043055
Fischer score for telecommuting: 0.00178081298121449
Fischer score for has_company_logo: 1.2333209336415096
Fischer score for has_questions: 0.11471918867336597
Fischer score for employment_type: 0.017490677042717544
Fischer score for required_experience: 0.02338851497834166
Fischer score for required_education: 0.01460806806968584
Fischer score for industry: 0.018228084816013082
Fischer score for function: 0.016705648096568004
Fischer score for country: 0.011072278656585749
Fischer score for region: 0.004848351321972796
Fischer score for city: 0.004643489676598648
```

<u>פיצ'רים נוספים שהחלטנו לוותר עליהם, מעבר לעובדה שקיבלו ציון נמוך ב- Fischer scoring:</u>

- Job_id – מזהה חד חד ערכי שאינו תורם למשימת הלמידה.
- Salary range – החלטנו לוותר על פיצ'ר זה מאחר וראינו כי יש מעל 80 אחוז הצעות עבודה להם חסר שדה זה. בנוסף, הערכים בשדות אלו לא באותו קנה מידה בכל סמפל ורובם אינם מוצגים באותו מטבע (פרטים אשר לא הצלחנו לאתר).
- Title – המרנו אותו לפיצ'ר קטגוריאלי התורם יותר למשימת הלמידה (sus words in title)
- Location – פיצלנו אותו ל-3 עמודות ולכן אין בו שימוש יותר. ויתרנו גם על שלושת העמודות המייצגות את location (Country, region, city) מאחר והמרנו אותם למשתנים קטגוריאליים וכתוצאה מכך אין לנו שימוש בהם בצורתם המקורית.
- ניתן לראות כי "has questions" , "exp weight" ו"Edu exp mismatch"" אינם קורלטיביים עם משטנה המטרה ולכן החלטנו להוריד אותם. (ל"has company logo" יש ציון Fischer scoring גבוהה ולכן נשאר)
- מלבד זאת עולה כי אין רוב המשתנים אינם קורלטיביים או קורלטיביים בצורה נמוכה אחד עם השני ולכן נבחר להשאיר אותם.




- **Dimensionality Reduction:**

בחרנו שלא להוריד את מימד הפיצ'רים ממספר סיבות:

ראשית, נותרנו עם 13 פיצ'רים בסה"כ למול 14000 רשומות כך שאין צורך ממשי בהורדת מימד, זמן הריצה של האלגוריתם טוב, בנוסף אין לנו צורך לבצע הכמסה (אנו לא מסתירים את הפ'יצרים בעבודה זו) ומכיוון שבתהליך זה אנו מאבדים מידע והשימוש במידע הוא פנימי אין לנו צורך להסתיר אותו.

**<u>Model Training – Validation:</u>**

- בחרנו לבצע ולידציה בעזרה K-fold מכיוון שהוא נחשב לשיטת הוולידציה הטובה ביותר להתמודדות מול over fitting בכך שלוקח בכל פעם קבוצות אחרות ל test, training validation, ודוגם בצורה שונה מספר פעמים שנרצה. שיטה זאת עדיפה על הholdout בה קשה יותר להוריד את השונות שנגרמה כתוצאה מחלוקה אקראית של הדאטה (דרוש לעשות "ערבול" לנתונים ולדגום שוב, אין הכרח שנדגום קבוצות שונות מספיק). ובנוסף שיטה זאת עדיפה על leave one out כיוון שיש לנו מספיק דאטה והוצאה של סמפל אחד לא תספיק בשביל לגרור מסקנות.
- המטריקה שנשתמש בה היא מטריקה הנקראת Precision. מטריקה זו מתאימה לשימוש בשני קלאסים או לסיווג של קלאס בינארי, כמו במקרה שלנו. Precision מראה את הפרופורציה בין 2 סיווגים שונים: true positive, שאלו סמפלים אשר סווגו ל-fraudulent והם באמת fraudulent ו- false positive, שאלו סמפלים אשר סווגו כ- fraudulent אך לא היו אמורים להיות מסווגים לשם. תהליך הוולידציה יהיה חלוקת הדאטה סט לקבוצות של k סמפלים. בכל קבוצה ישנם k-1 סמפלים עליהם עושים training. לאחר ביצוע ה- training עושים הערכה לפי המדדים של precision לכל קבוצה ומחשבים את הממוצע בניהם. נוכל לשחק עם ה- k בכדי לבדוק עבור איזה k נקבל ערכים טובים ככל שניתן.

**נספחים:**

1. התפלגויות של המשתנים

Telecommuting:

| telecommuting | non-fraudulent | fraudulent | rate |
|---|---|---|---|
| 0 | 13040 | 652 | 0.047619 |
| 1 | 559 | 53 | 0.086601 |

Has company logo

| has_company_logo | non-fraudulent | fraudulent | rate |
|---|---|---|---|
| 0 | 2475 | 479 | 0.162153 |
| 1 | 11124 | 226 | 0.019912 |

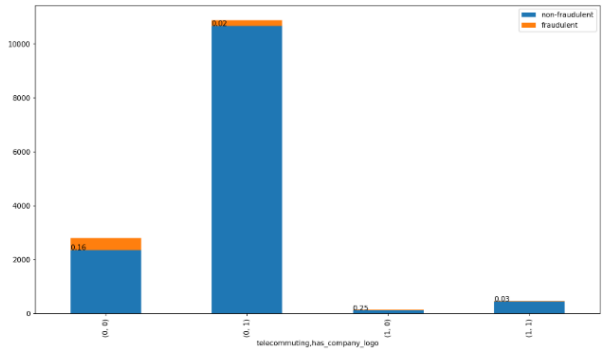
Has questions

| has_questions | non-fraudulent | fraudulent | rate |
|---|---|---|---|
| 0 | 6787 | 504 | 0.069126 |
| 1 | 6812 | 201 | 0.028661 |

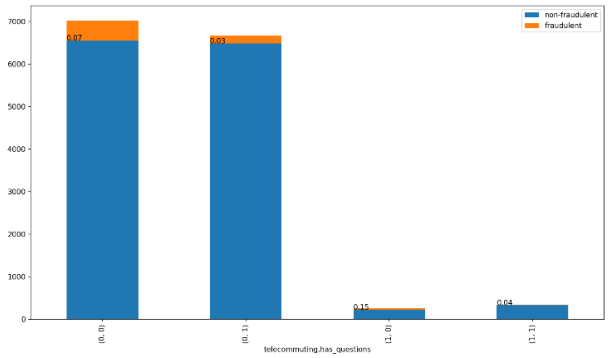
בפלט הבא יוצג גרף מסוג בר, ובו בציר X יוצגו המשתנים המסבירים בצורה בינארית (1,0) המייצגים את הקומבינציה ביניהם. על גבי ציר הY יוצגו כמות המופעים של קומבינציה זו (סמפלים מסוג זה ) המפולגים ל2 צבעים. כחול המעיד על סמפלים שסווגו כמודעות אמת, וכתום המעיד על סמפל שסווג כמודעת שקר. בחלקו העליון והכתום של הבר מוצג אחוז המודעות שסווגו כמזויפות בקומבינציה זו.

4. Has company logo & Telecommuting **:**

ניתן לראות כי **25%** מהחברות שהוצגו עם אפשרות לעבוד מהבית וללא לוגו זוהו כמודעות שקריות.

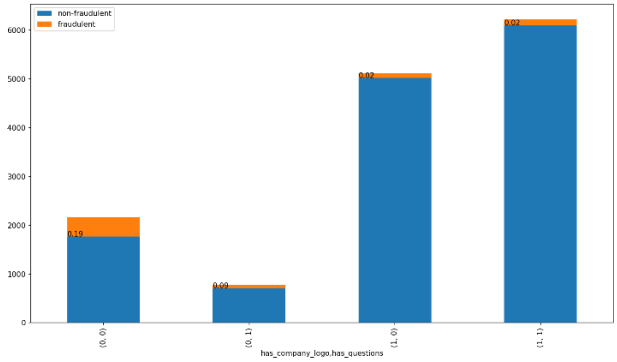


## Telecommuting& Has questions :




ניתן לראות כי 15% מכלל המודעות שאפשרו לעבוד מרחוק אך לא כללו שאלות סיווג נוספות תועדו כמודעות מזויפות. נתון זה מעניין מכוון שכל פיצ'ר בפני עצמו מאפשר סיווג של עד 8% לכל היותר, כך שבהצלבתן הצלחנו להכפיל את אחוז הסיכוי לסווג נכונה את הסמפל.

## 5. Has company logo & Has questions :




ניתן לראות כי 19% מכלל המודעות שלא הציגו לוגו ולא אפשרו לעבוד מהבית סווגו כמזויפות. בנוסף ראוי לציין כי מודעות שהוצגו עם לוגו לא שינו את התפלגות התוצאות בהוספת הנתון החדש, (2% בגרף זה לעומת 2% בגרף של מודעות עם לוגו) כך שאין בחיבור זה מידע ממשי.

## Required experience & Required education

**required experience:**

| fraudulent | 0 | 1 | rate |
|---|---|---|---|
| required_experience | | | |
| Associate | 1791 | 34 | 0.018630 |
| Director | 299 | 12 | 0.038585 |
| Entry level | 2005 | 146 | 0.067875 |
| Executive | 113 | 10 | 0.081301 |
| Internship | 302 | 7 | 0.022654 |
| Mid-Senior level | 2973 | 91 | 0.029700 |
| Not Applicable | 847 | 52 | 0.057842 |
| unknown | 5269 | 353 | 0.062789 |

| required_experience | |
|---|---|
| Associate | 1825 |
| Director | 311 |
| Entry level | 2151 |
| Executive | 123 |
| Internship | 309 |
| Mid-Senior level | 3064 |
| Not Applicable | 899 |
| unknown | 5622 |

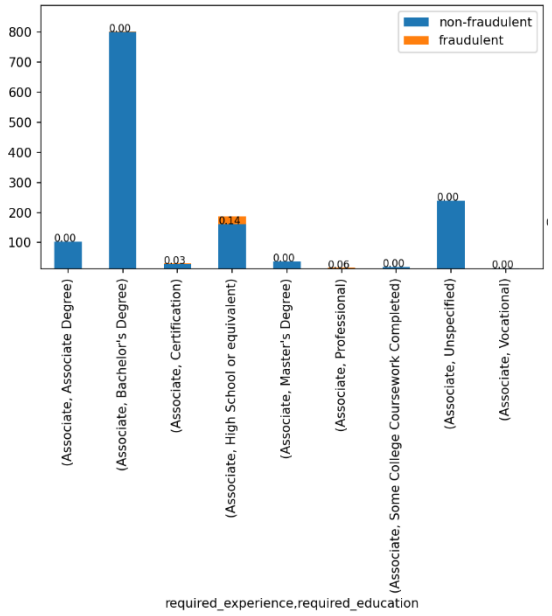
**::required education**

| fraudulent | 0 | 1 | rate |
|---|---|---|---|
| required_education | | | |
| Associate Degree | 214.0 | 5.0 | 0.022831 |
| Bachelor's Degree | 4046.0 | 81.0 | 0.019627 |
| Certification | 124.0 | 17.0 | 0.120567 |
| Doctorate | 15.0 | 1.0 | 0.062500 |
| High School or equivalent | 1512.0 | 139.0 | 0.084191 |
| Master's Degree | 301.0 | 22.0 | 0.068111 |
| Professional | 60.0 | 3.0 | 0.047619 |
| Some College Coursework Completed | 86.0 | 3.0 | 0.033708 |
| Some High School Coursework | 5.0 | 17.0 | 0.772727 |
| Unspecified | 1063.0 | 50.0 | 0.044924 |
| Vocational | 39.0 | NaN | NaN |
| Vocational - Degree | 5.0 | NaN | NaN |
| Vocational - HS Diploma | 6.0 | NaN | NaN |
| unknown | 6123.0 | 367.0 | 0.056549 |

| required_education | |
|---|---|
| Associate Degree | 219 |
| Bachelor's Degree | 4127 |
| Certification | 141 |
| Doctorate | 16 |
| High School or equivalent | 1651 |
| Master's Degree | 323 |
| Professional | 63 |
| Some College Coursework Completed | 89 |
| Some High School Coursework | 22 |
| Unspecified | 1113 |
| Vocational | 39 |
| Vocational - Degree | 5 |
| Vocational - HS Diploma | 6 |

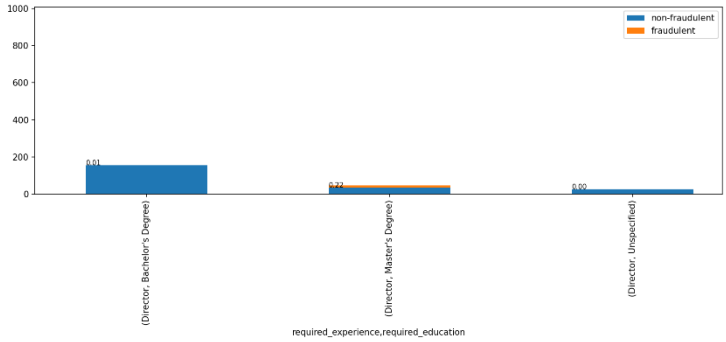
### 1. Associate:

ניתן לראות כי למרות שישנה הסתברות של 2% כי מודעה מסוג זה תהייה עוינת, לרוב החשש אינו מבוסס. יתרה מכך מתוך 34 המודעות הכוזבות שהופיעו תחת Associate כ27 מתוכם הגיעו מהשילוב עם high School ובכך מעלים את ההסתברות ל15% מודעה כוזבת תחת שילוב זה.



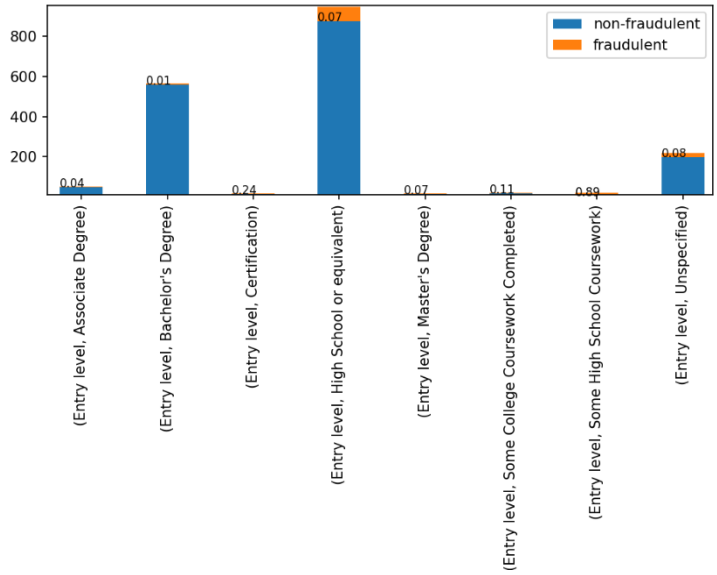

### 2. Director:

למרות שכ4% ממודעות Director מסווגות כחשודות, ניתן לראות שהסיווג מגיע רק ממודעות מסוג Master's Degree. שילוב זה מעניין במיוחד כיוון של Master's Degree כולו ישנה הסתברות של 8% סה"כ להיות מודעה כוזבת, ואילו שילובם מביא אותנו ל22%.

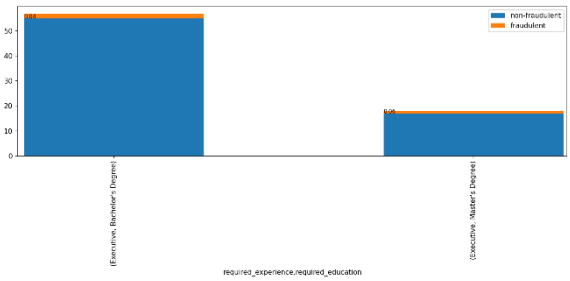



### 3. Entry level:

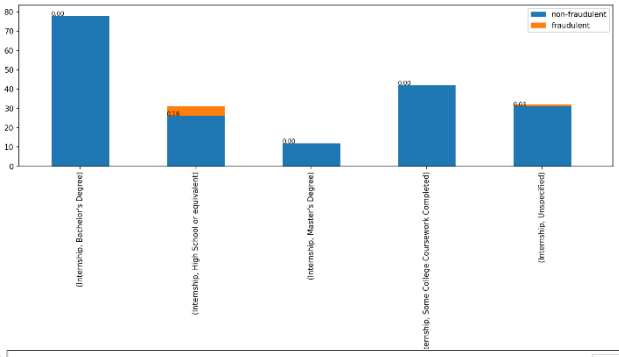
ניתן לראות כי למרות שסוג העבודה "entry level" סווג כבעל 69% מודעות כוזבות, השילוב שלו עם "high school" מגיע ל89%! סיכוי למודעה כוזבת. בנוסף לעלייה משמעות עד ל24% עם "certification" כאשר ב2 המקרים, הסיכוי להיות מודעה כוזבת נמוך יותר ללא האגרגציה. בנוסף שילוב זה יחד עם "bachelor degree" מוריד כמעט ל0 את הסיכוי למודעה כוזבת, (בהתחשב בעובדה שיש כמעט 600 סמפלים מסוג זה).



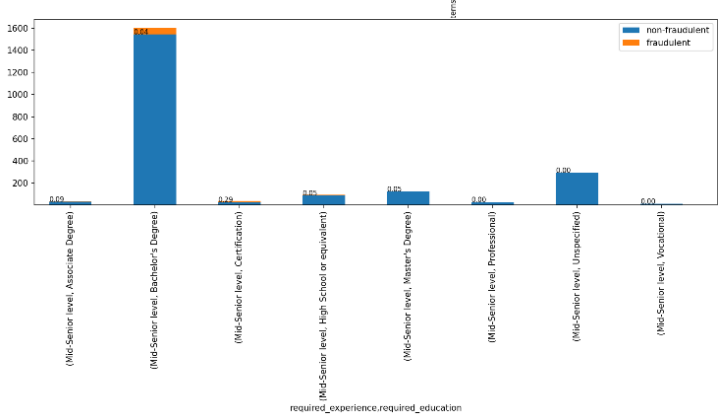
4. **Executive:** ניתן לראות כי כל 8% הסיכוי להיות מודעה כוזבת, מגיע מהשילוב של Bachelor's Degree וMasters Degree



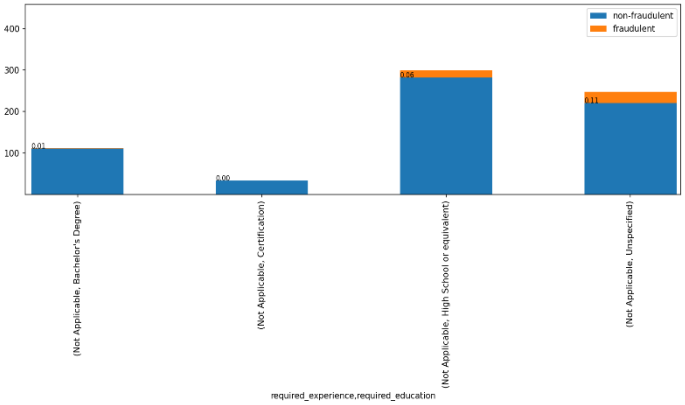

5. **Internship:** ניתן לראות כי למרות שאחוז המודעות הכוזבות מסוג זה נמוך מאוד (רק 2%) הוא מגיע כמעט כולו מHigh school בעוד אין עדויות להתנהגות שכזו מצד שאר הלייבלים




6. **Mid-Senior level:** התפלגות זו מכילה 3% של מודעות כוזבות, ניתן לראות כי 30% מ Certification מסווגות ככוזבות, זה מעניין בעיקר כי רק 12% אחוז מכלל ה Certification סווגו ככוזבות ולכן השילוב ביניהם מעניין מאוד.




7. **Not Applicable:** ההתפלגות עצמה מכילה 5.7% שגיאה, ניתן לראות כי כולה מגיע או מHigh School או מ Unspecified .



:Industry

| | non-fraudulent | fraudulent | rate |
|---|---|---|---|
| industry | | | |
| Accounting | 71.0 | 43.0 | 0.377193 |
| Biotechnology | 28.0 | 4.0 | 0.125000 |
| Business Supplies and Equipment | 12.0 | 3.0 | 0.200000 |
| Computer & Network Security | 37.0 | 2.0 | 0.051282 |
| Computer Hardware | 23.0 | 2.0 | 0.080000 |
| Computer Networking | 25.0 | 8.0 | 0.242424 |
| Consumer Services | 267.0 | 23.0 | 0.079310 |
| Electrical/Electronic Manufacturing | 53.0 | 3.0 | 0.053571 |
| Entertainment | 51.0 | 5.0 | 0.089286 |
| Environmental Services | 36.0 | 3.0 | 0.076923 |
| Health, Wellness and Fitness | 91.0 | 13.0 | 0.125000 |
| Hospital & Health Care | 354.0 | 42.0 | 0.106061 |
| Hospitality | 62.0 | 13.0 | 0.173333 |
| Human Resources | 85.0 | 5.0 | 0.055556 |
| Information Services | 23.0 | 2.0 | 0.080000 |
| Leisure, Travel & Tourism | 45.0 | 19.0 | 0.296875 |
| Management Consulting | 105.0 | 6.0 | 0.054054 |
| Marketing and Advertising | 607.0 | 36.0 | 0.055988 |
| Mechanical or Industrial Engineering | 23.0 | 4.0 | 0.148148 |
| Media Production | 36.0 | 3.0 | 0.076923 |
| Medical Devices | 12.0 | 1.0 | 0.076923 |
| Oil & Energy | 148.0 | 92.0 | 0.383333 |
| Outsourcing/Offshoring | 16.0 | 1.0 | 0.058824 |
| Real Estate | 124.0 | 16.0 | 0.114286 |
| Staffing and Recruiting | 99.0 | 8.0 | 0.074766 |
| Telecommunications | 247.0 | 18.0 | 0.067925 |
| Transportation/Trucking/Railroad | 41.0 | 3.0 | 0.068182 |

:function

| | non-fraudulent | fraudulent | rate |
|---|---|---|---|
| function | | | |
| Accounting/Auditing | 150.0 | 22.0 | 0.127907 |
| Administrative | 393.0 | 100.0 | 0.202840 |
| Advertising | 69.0 | 4.0 | 0.054795 |
| Customer Service | 920.0 | 54.0 | 0.055441 |
| Data Analyst | 65.0 | 4.0 | 0.057971 |
| Distribution | 19.0 | 2.0 | 0.095238 |
| Engineering | 974.0 | 92.0 | 0.086304 |
| Finance | 128.0 | 12.0 | 0.085714 |
| Financial Analyst | 24.0 | 3.0 | 0.111111 |
| Other | 233.0 | 27.0 | 0.103846 |

Location: לכל עמודה נציג טבלה המראה את שם המיקום ואת אחוז הפעמים שהופיע בהודעה כוזבת:

:Country

| fraudulent | Non-Fraudulent | Fraudulent | Total |
|---|---|---|---|
| country | | | |
| US | 93.058645 | 6.941355 | 8543.0 |
| GB | 98.957247 | 1.042753 | 1918.0 |
| CA | 97.198880 | 2.801120 | 357.0 |
| unknown | 94.661922 | 5.338078 | 281.0 |
| IN | 98.564593 | 1.435407 | 209.0 |
| AU | 79.878049 | 20.121951 | 164.0 |
| PH | 99.000000 | 1.000000 | 100.0 |
| PL | 96.491228 | 3.508772 | 57.0 |
| EE | 98.181818 | 1.818182 | 55.0 |
| ES | 97.916667 | 2.083333 | 48.0 |
| EG | 97.777778 | 2.222222 | 45.0 |
| AE | 97.500000 | 2.500000 | 40.0 |
| BR | 96.296296 | 3.703704 | 27.0 |
| PK | 95.652174 | 4.347826 | 23.0 |
| MY | 35.294118 | 64.705882 | 17.0 |
| QA | 61.538462 | 38.461538 | 13.0 |
| ID | 90.000000 | 10.000000 | 10.0 |
| BH | 57.142857 | 42.857143 | 7.0 |
| TW | 50.000000 | 50.000000 | 4.0 |

:region

| fraudulent | Non-Fraudulent | Fraudulent | Total |
|---|---|---|---|
| region | | | |
| AGB | NaN | 100.000000 | 2.0 |
| 16 | NaN | 100.000000 | 2.0 |
| DA | 25.000000 | 75.000000 | 4.0 |
| EAW | 33.333333 | 66.666667 | 3.0 |
| LIN | 50.000000 | 50.000000 | 2.0 |
| ANS | 50.000000 | 50.000000 | 2.0 |
| HI | 50.000000 | 50.000000 | 4.0 |
| 41 | 50.000000 | 50.000000 | 2.0 |
| TPQ | 50.000000 | 50.000000 | 4.0 |
| 14 | 50.000000 | 50.000000 | 4.0 |
| ABD | 57.142857 | 42.857143 | 7.0 |
| MS | 61.538462 | 38.461538 | 13.0 |
| NSW | 70.454545 | 29.545455 | 88.0 |
| ME | 71.428571 | 28.571429 | 7.0 |
| MT | 71.428571 | 28.571429 | 7.0 |
| MD | 72.500000 | 27.500000 | 80.0 |
| QLD | 76.923077 | 23.076923 | 13.0 |
| ALX | 80.000000 | 20.000000 | 5.0 |
| ANT | 83.333333 | 16.666667 | 6.0 |
| AK | 83.333333 | 16.666667 | 6.0 |
| NE | 84.000000 | 16.000000 | 25.0 |
| TX | 84.203822 | 15.796178 | 785.0 |
| KS | 84.848485 | 15.151515 | 33.0 |
| KY | 86.111111 | 13.888889 | 72.0 |
| JW | 87.500000 | 12.500000 | 8.0 |
| AL | 88.888889 | 11.111111 | 54.0 |
| SD | 88.888889 | 11.111111 | 27.0 |
| ND | 89.473684 | 10.526316 | 19.0 |

:city

| fraudulent | Non-Fraudulent | Fraudulent | Total |
|---|---|---|---|
| city | | | |
| City of Industry | NaN | 100.000000 | 2.0 |
| San Mateo | NaN | 100.000000 | 2.0 |
| Kuala lumpur | NaN | 100.000000 | 3.0 |
| LAS VEGAS | NaN | 100.000000 | 2.0 |
| Moravia | NaN | 100.000000 | 2.0 |
| CHICAGO | NaN | 100.000000 | 3.0 |
| COLUMBUS | NaN | 100.000000 | 2.0 |
| BALTIMORE | NaN | 100.000000 | 2.0 |
| Taipei | NaN | 100.000000 | 2.0 |
| DALLAS | NaN | 100.000000 | 10.0 |
| RTP | NaN | 100.000000 | 2.0 |
| Deweyville | NaN | 100.000000 | 3.0 |
| Gold coast | NaN | 100.000000 | 2.0 |
| Accord | NaN | 100.000000 | 2.0 |
| Absarokee | NaN | 100.000000 | 2.0 |
| Abilene | NaN | 100.000000 | 2.0 |
| Aberdeen | NaN | 100.000000 | 7.0 |
| Abbeville | NaN | 100.000000 | 7.0 |
| Abbeville | NaN | 100.000000 | 2.0 |
| AUSTIN | NaN | 100.000000 | 14.0 |
| ATLANTA | NaN | 100.000000 | 3.0 |
| Groveport | NaN | 100.000000 | 2.0 |
| Bakersfield | 7.407407 | 92.592593 | 27.0 |
| LOS ANGELES | 16.666667 | 83.333333 | 6.0 |
| MIAMI | 16.666667 | 83.333333 | 6.0 |
| MANHATTAN | 33.333333 | 66.666667 | 3.0 |
| Tamarac | 50.000000 | 50.000000 | 2.0 |
| DETROIT | 50.000000 | 50.000000 | 2.0 |
| Clarksville | 50.000000 | 50.000000 | 2.0 |
| Raleigh | 50.000000 | 50.000000 | 2.0 |
| Morristown | 50.000000 | 50.000000 | 2.0 |
| Stratford | 50.000000 | 50.000000 | 2.0 |
| Menomonie | 50.000000 | 50.000000 | 2.0 |
| NY | 50.000000 | 50.000000 | 16.0 |
| Farmington Hills | 50.000000 | 50.000000 | 2.0 |
| Immingham | 50.000000 | 50.000000 | 2.0 |
| West Chester | 50.000000 | 50.000000 | 2.0 |
| Yankton | 50.000000 | 50.000000 | 2.0 |
| Plattsburgh | 50.000000 | 50.000000 | 2.0 |
| fort lauderdale | 50.000000 | 50.000000 | 2.0 |
| San Antonio | 50.000000 | 50.000000 | 2.0 |
| newyork | 50.000000 | 50.000000 | 2.0 |
| Long Island | 50.000000 | 50.000000 | 2.0 |
| Oneonta | 50.000000 | 50.000000 | 2.0 |
| PHILADELPHIA | 50.000000 | 50.000000 | 2.0 |
| San Mateo | 54.054054 | 45.945946 | 37.0 |
| Sydney | 62.686567 | 37.313433 | 67.0 |
| Houston | 65.919283 | 34.080717 | 223.0 |
| McAllen | 66.666667 | 33.333333 | 3.0 |
| St Louis | 66.666667 | 33.333333 | 6.0 |
| Reno | 66.666667 | 33.333333 | 3.0 |
| Visalia | 66.666667 | 33.333333 | 3.0 |
| Rochester | 66.666667 | 33.333333 | 3.0 |
| Buffalo | 66.666667 | 33.333333 | 3.0 |
| tampa | 66.666667 | 33.333333 | 3.0 |
| BOSTON | 66.666667 | 33.333333 | 3.0 |
| Trenton | 66.666667 | 33.333333 | 3.0 |
| Middletown | 66.666667 | 33.333333 | 3.0 |

Addit: עמודה המיועדת למודעות שכתבו מידע נוסף מעבר לחלק של העיר, לא ניתן ללמוד ממנה מידע נוסף

| fraudulent | Non-Fraudulent | Fraudulent | Total |
|---|---|---|---|
| addit | | | |
| CA / Mt. Poso | NaN | 100.000000 | 2.0 |
| CA | 66.666667 | 33.333333 | 12.0 |
| | 95.099972 | 4.900028 | 14204.0 |