פרויקט בלמידת מכונה – חלק א'

Contents

Data collection and Sensing

- Data Collection מייצג את העולם האמיתי אותו אנו רוצים ללמוד, בפרויקט זה מדובר באוסף של מודעות דרושים.
- סוג ה Sensing שבוצע לדאטה הינו חישה סטטית. סט המאפיינים של כל הסאמפלים בעלי ערכים קבועים שאינם משתנים בזמן ולכן ניתן להסיק כי מדובר בחישה סטטית ולא דינמית כפי שניתן לראות מקובץ הנתונים לפרויקט. כלומר, נקבל את אותם הנתונים בכל רגע נתון.
- סוג חישה שלא בוצע על הדאטה הינו חישה דינמית. לראות עינינו, ביצוע חישה דינמית עשוי לעזור למשימת הלימוד מכיוון שהוא מערב ניטור בזמן אמת אחר ההתנהגות של המשתמשים בתגובה למודעה, כולל איך הם מתקשרים עם המודעה והאתר אליו מובילה.
- מספר דוגמאות למידע שנאסף מחישה מסוג זה הינו: מספר צפיות במודעה, השעות בהן מפרסמים את המודעה, מספר תגובות שהתקבלו אלו דוגמאות למידע שעשוי לסייע במשימת הלימוד אם המודעה הינה מזויפת או אמיתית.
- קטגוריית הלמידה הינה Supervised Learning כי כל הסאמפלים מתויגים.
- סוג משימת הלמידה שלנו הינו משימת סיווג מסוג Binary Classification מכיוון שבמקרה של משתנה מטרה בינארי המטרה הינה לסווג סאמפלים חדשים לאחד מבין שתי המחלקות ולכן משימת למידה זו הינה הבחירה המתאימה ביותר. מדובר בסיווג מודעות חדשות לקבוצת המודעות המזויפות ואז הערך יהיה 1, אחרת הערך יהיה 0. סוג נוסף של משימת למידה אפשרית על הדאטה הינה Supervised Anomaly Detection בה ניתן להשתמש באלגוריתמים שמסתמכים על מחלקת הרוב כך שיהיה ניתן לאתר חריגות שיאפשרו זיהוי פרסומי משרות אשר שונים מאוד מרוב הנתונים שבמקרה שלנו משתייכים למודעות אמיתיות. המחלקה של מודעות אמיתיות מהווה את רוב המידע שנאסף ולכן ישנה בעייתיות בלמידה ממחלקת המיעוט שמייצגת את המודעות המזויפות לאור העובדה שרק 700 מתוך 14,000 מהסאמפלים שנאספו מתויגים כמודעות כאלה.

Dataset Creation

Exploratory data analysis

1) **Title**:

- ביצענו ניתוח של 10 המילים הנפוצות ביותר במודעות אמיתיות ומזויפות, והשתמשנו בענן מילים. במודעות האמיתיות המילים הנפוצות יותר ממוקדות יותר לדרישות המשרה עצמה, אם דרוש מהנדס, מנהל, איש מכירות וכו'. לעומת זאת במודעות המזויפות המילים הנפוצות הן כלליות יותר ולא ניתן להסיק מהן על המשרה הדרושה באופן ברור.
- חישבנו את ממוצע המילים במודעות מזויפות לעומת אמיתיות, הפרש המילים יצא זניח יחסית – 1.79 מילה יותר במודעה מזויפת. אורך הכותרת לא מסייע בסיווג המודעה.

2) **Location**:

- סיננו את 10 המדינות עם הכי הרבה פרסומים של מודעות, וחישבנו את היחס של מספר מודעות אמיתיות לכל מודעה מזויפת. יש שוני גבוה בין המדינות מבחינת יחס מספר המודעות המזויפות שמפורסמות דרכן, ולכן כדאי לשמור על משתנה זה.

	country	real_count	fake_count	real_to_fake_ratio	total_count
0	US	7950	593	13.406408	8543
1	GB	1898	20	94.900000	1918
2	GR	749	0	inf	749
3	CA	347	10	34.700000	357
4	DE	313	0	inf	313
5	NZ	263	0	inf	263
6	IN	206	3	68.666667	209
7	AU	131	33	3.969697	164
8	NL	106	0	inf	106
9	PH	99	1	99.000000	100

- ביצענו ניתוח דומה על בסיס היבשת של כל מודעה, נסיק כי ישנה השפעה רבה ליבשת על סיווג המודעה. מספר היבשות מועט לעומת מספר המדינות נשקול להוסיף מאפיין זה.

3) **Department**:

- ניסיון לצמצם את כמות הקטגוריות לא צלח, גם לאחר צמצום נשארו עם מעל 600 קטגוריות שונות. לאור כך ולאור הערכים החסרים הרבים נשקול להסיר מאפיין זה.

4) **Salary**:

- מספר הערכים החסרים במשתנה זה הוא יחסית גדול (~84%), לכן ייתכן ונרצה להסיר משתנה זה או להשקיע מחשבה רבה בהשלמת ערכיו החסרים.
- קיימים ערכים חריגים רבים ולכן לצורך הניתוח ערכי המשכורת הוגבלו עד ל-$300,000 (81 רשומות הוסרו בעקבות כך לצרכי בדיקה בלבד, לא מה- **data frame** המקורי). מהמשכורת הממוצעת ואורך טווח המשכורות הממוצע בין מודעות אמיתיות למזויפות קשה לקבוע בבירור הבדל ביניהן.
- בוצעה דיסקרטיזציה על ערכי המאפיין אשר לא עולה ממנה מסקנה חד משמעית בנוגע לקשר בין דרגות השכר (נמוכה, בינונית, גבוהה, גבוהה מאוד) לבין סיווג המודעה.

5) **Company Profile**:

- אין הבדל במילים הנפוצות במאפיין זה בין מודעות מזויפות לאמיתיות, וההפרש בין כמות המילים הממוצעת במאפיין זה בין סוגי המודעות אינו גבוה במיוחד. לכן נשקול להסירו.
- מניתוח הסנטימנט עולה כי מודעות אמיתיות בעלות סנטימנט חיובי יותר ממזויפות, אך לא בצורה דרסטית מאוד. השוני בסנטימנט הוא לא גבוה מספיק שנחליט להשאיר מאפיין זה.

6) **Description**:

- הבחנה על סמך מאפיין זה לאור המילים השכיחות כנראה תהיה לא מוצלחת, יש דמיון גבוה בין סוגי המודעות.

הבחנה בין סוגי המודעות לפי אורך מאפיין זה תהיה כנראה לא מוצלחת, אין הבדל משמעותי. כנ"ל לגבי ניתוח סנטינמט. ייתכן ונרצה להסיר מאפיין זה.

7) **Requirements**:

- יהיה קשה לסווג את סוג המודעה עפ"י המילים הנפוצות במאפיין זה או עפ"י אורכו.
- יש קשר בין המאפיין לבין המאפיינים **'required_experience / education'** אבל הוא אינו חזק מספיק ברמה שנוותר על המאפיין **requirements** לאור דמיון עם המאפיינים האחרים.

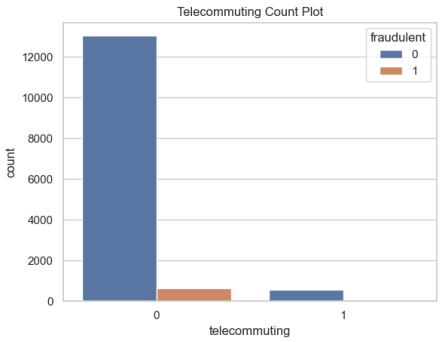
```
Percentage of records where 'required_experience' appears in the 'requirements': 13.49%
Percentage of records where 'required_education' appears in the 'requirements': 15.84%
```

8) **Benefits**:

- יהיה קשה לסווג את סוג המודעה עפ"י המילים הנפוצות במאפיין זה או עפ"י אורכו.
- עפ"י ניתוח הסנטימנט, כמעט ואין הבדל בסנטימנט בין מודעות אמיתיות למזויפות ולכן סביר להניח שמאפיין זה פחות יתרום בסיווג המודעות.

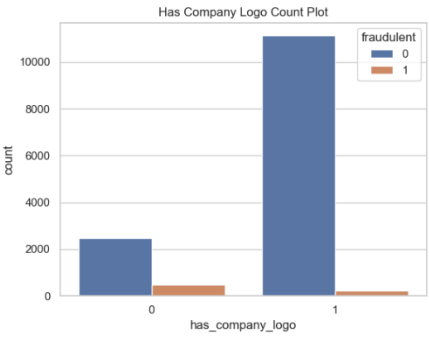
9) **Telecommuting**:

- בהתאם לגרף המצורף, ניתן להסיק כי מדובר בסט נתונים שאינו מאוזן לאור העובדה כי אין ערכים חסרים והמודל רואה בעיקר מודעות של משרות אמיתיות שמאופיינות לרוב בעבודה מרחוק, זאת למרות שעבור מודעות מזויפות הקטגוריה השכיחה הינה לא לעבודה מרחוק. לכן, קיים חשש כי המודל עלול ללמוד סטייה זו בנתונים. בהתאם לכך, חשוב לציין כי הנתונים הללו אינם מייצגים ולכן המסקנה שקיבלנו מהגרף לפיה מודעות של משרות מזויפות אינן מאופיינות בעבודה מרחוק אינה חד משמעית.



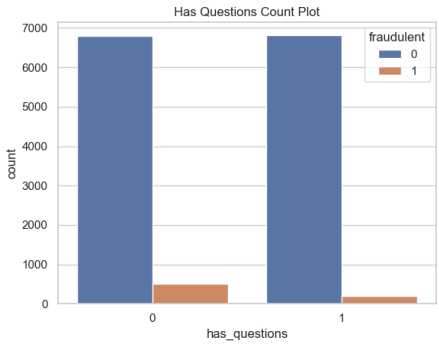
10) **Has company logo**:

- ניתן להסיק מהגרף המצורף כי מודעות של משרות מזויפות אינן כוללות לוגו, אולם חשוב לציין כפי שנאמר קודם לכן כי סט הנתונים אינו מאוזן ולכן מסקנה זו אינה חד משמעית.



11) **Has Questions**:

- משמעות המשתנה בהקשר של משימת הלימוד הינה כי עבור מודעה של משרה לא מתויגת, ככל הנראה תשויך למשרה ללא שאלות מיון מכיוון שבעוד שעבור מודעות אמיתיות הסיכויים שווים, עבור מודעות מזויפות הסיכוי גבוה יותר שלא יכללו במודעה.
- ניתן להסיק מהגרף המצורף כי עבור מודעות של משרות אמיתיות הסיכוי לקבל שאלות מיון הינו די מאוזן. לעומת זאת, עבור מודעות של משרות מזויפות ניכר כי ההעדפה הינה לא להוסיף שאלות מיון.

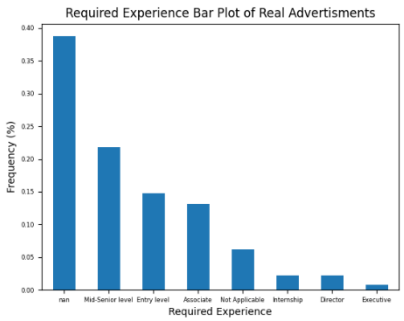


12) **Employment Type**:

- אחוז הערכים החסרים במאפיין זה במודעות אמיתיות הינו 27.8%, בעוד שעבור מודעות מזויפות מדובר ב18.9%. כפי שניתן לראות במאפיין של description ישנן מילים שבאמצעותן ניתן לדעת באופן עקיף או ישיר מהו סוג ההעסקה ולכן נשקול בהמשך לחפש מילות מפתח באמצעותן נוכל להשלים ערכים חסרים במאפיין זה.
- ניתן לראות לפי הגרף כי הקטגוריה השכיחה ביותר של סוג ההעסקה הן עבור מודעות אמיתיות והן עבור מודעות מזויפות הינה משרה מלאה. כלומר, מאפיין זה לא מספק מידע אינפורמטיבי ולאור העובדה כי הסט נתונים אינו מאוזן ככל הנראה נבחר בהמשך לא להשתמש במאפיין זה.

13) **Required Experience**:

- אחוז הערכים החסרים במודעות מזויפות ואמיתיות הינו יחסית גבוה ועומד על 50% במודעות מזויפות לעומת 38% במודעות אמיתיות.
- ניכר כי בעוד שעבור מודעות שכוללות דרישות ניסיון מתקבל כי עבור מודעות אמיתיות נדרשת רמה בינונית בכירה, ואילו עבור מודעות מזויפות נדרשת רמה התחלתית בלבד. כלומר, מסקנה אפשרית שעשויה לעלות מהנתונים הינה כי מודעות מזויפות מיועדות לקהל צעיר יותר, פחות מנוסה שייתכן כי יותר קל לרמות אותו.



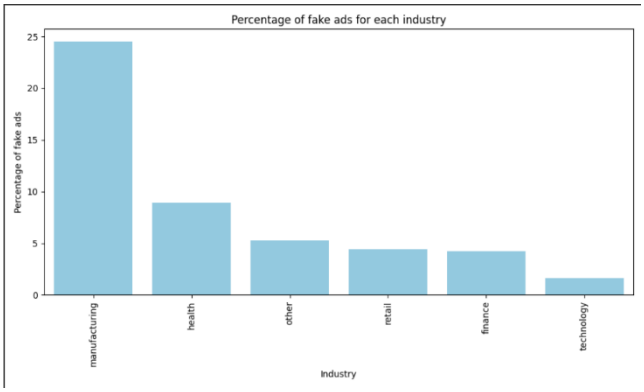
14) **Required Education**:

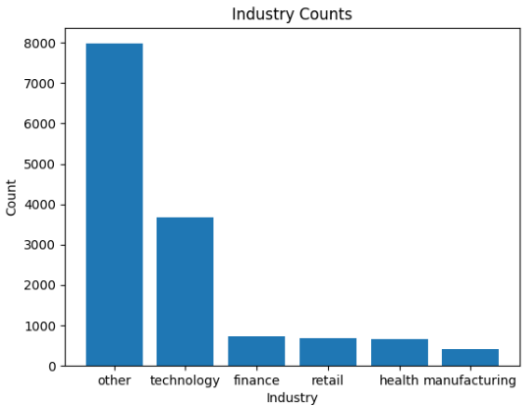
- אחוז הערכים החסרים במודעות מזויפות ואמיתיות הינו יחסית דומה (במזויפות 45% לעומת אמיתיות 52%), כלומר בשניהם מדובר בערך במחצית מהמודעות שלא מכילות בהן את המאפיין של דרישת השכלה למשרה. מספר הערכים הייחודים עבור מאפיין זה הינו 13, כאשר חלקם שקולים לערכים הייחודיים של מאפיין דרישות הניסיון התעסוקתי.
- ניכר כי בעוד שעבור מודעות שכוללות דרישות השכלה מתקבל כי עבור מודעות אמיתיות נדרש תואר ראשון, ואילו עבור מודעות מזויפות נדרש תעודת בגרות בלבד. מסקנה אפשרית זו מתיישבת עם המסקנה שהתקבלה מפיצ'ר דרישות ניסיון תעסוקתי.

15) **Industry**:

- צמצמנו את מספר הערכים הייחודיים מ-128 ל-6 קטגוריות עליהן ביצענו עיבוד על מנת שנוכל להסיק מסקנות רלוונטיות על הנתונים.

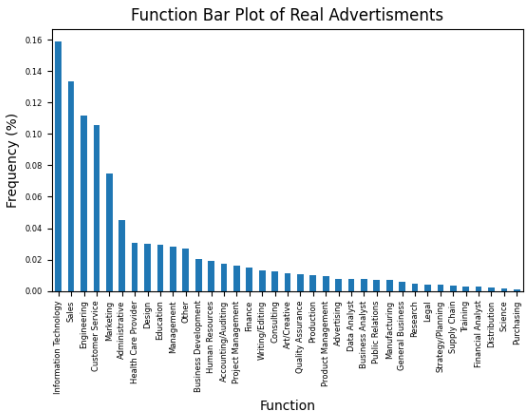
- בקטגוריה **Manufacturing** יש אחוז גבוה של מודעות מזויפות, ייתכן שזה נובע ממספר נמוך יותר של דגימות. לעומת זאת בקטגוריה **Technology** יש הרבה דגימות ואחוז נמוך של מודעות מזויפות, ייתכן שמאפיין זה יכול להעיד על סיווג המודעה.
- בשאר הקטגוריות אחוז המודעות המזויפות יחסית דומה ונמוך.

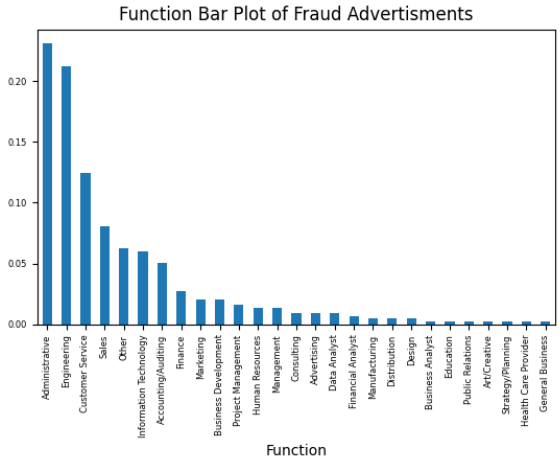




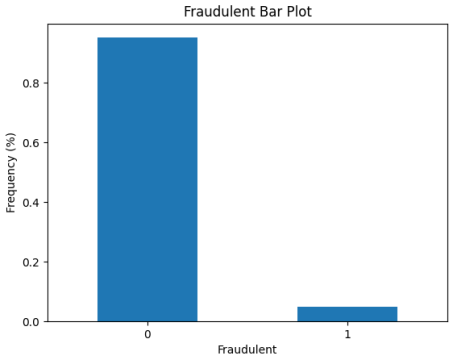
16) **Function**:

- מספר הערכים הייחודיים הינו 37. מדובר במספר גדול של ערכים מצד אחד, אולם מצד שני כן ניתן לייצג גרפית ולהגיע למסקנות רלוונטיות בשלב זה. אולם, בהמשך נשקול לבצע איחוד ערכים למספר קטגוריות מצומצם יותר.
- ניכר כי בעוד שעבור מודעות אמיתיות הפונקציה הנפוצה ביותר הינה טכנולוגיית מידע, ואילו עבור מודעות מזויפות מדובר בתפקיד אדמיניסטרטיבי. מסקנה אפשרית העולה מכך הינה כי בעולם האמיתי קיים ביקוש גבוה יותר למשרות טכנולוגיות, ואילו מודעות מזויפות מוכוונות דווקא לאוכלוסייה יותר כללית ופחות מקצועית. ייתכן כי יש קשר לכך שמדובר בקהל יעד רחב יותר עבור המודעות המזויפות כך שהסיכויים להצלחה בהונאה יהיו גדולים





17) **Fraudulent**:



- בהתאם לגרף המצורף, ניתן להסיק בראש ובראשונה כי מדובר בסט נתונים שאינו מאוזן לאור העובדה כי אין ערכים חסרים והמודל רואה בעיקר מודעות של משרות אמיתיות ולכן עלול ללמוד סטייה זו בנתונים. בהתאם לכך, חשוב לציין כי הנתונים הללו אינם מייצגים ולכן כל אחת מהמסקנות שנכתבו לעיל אינן חותכות.

Pre-Processing

1) חזרתיות – השמטנו מסט הנתונים המקורי את המאפיין job_id מכיוון שהינו חח"ע לכל סאמפל. בעזרת פונק' duplicated מצאנו כי ישנה חזרה של 186 רשומות ב data collection. מבדיקה באינטרנט מצאנו כי 20% מהמודעות דרושים הן מזויפות, לעומת 5% מודעות מזויפות במאגר שלנו. לכן בחרנו להסיר רק מודעות שסווגו כאמיתיות. לאחר בדיקת מספר הרשומות הכפולות, השתמשנו בפונק' drop_duplicate על מנת להסירן.

```
Number of records before removing duplicates: 14304
Number of records after removing duplicates: 14123
```

2) איזון נתונים – על מנת לאזן את הנתונים בחרנו לבצע up sampling עד למצב בו אחוז המודעות המזויפות יהיו 20% מסך המודעות, בדומה למצב המוערך מהאינטרנט. לצורך כך השתמשנו במתודה resample מהחבילה sklearn.

```
Number of records before removing duplicates: 14304
Number of records after removing duplicates: 14123
fake ads percentage before upsampling:  4.991857254124478 %
Number of records before upsampling: 14123
Number of records after upsampling: 16772
fake ads percentage after upsampling:  19.99761507274028 %
```

3) ערכים חסרים:

- **Location** – השלמנו ערכים חסרים עפ"י השכיחות של המדינות ברשומות עם הערכים המלאים.
- **Salary** – לקחנו את הערך הממוצע של העמודה 'mean_salary' של כל המשכורות שהן מתחת ל-1,000,000 והצבנו בכל הערכים החסרים.
- **Company Profile / Description / Requirements / Benefits** – הצבנו את 50/100 הערכים הנפוצים ביותר בעמודה זו בכל הערכים החסרים.
- **Employment_Type / Required Experience / Required Education** – קודם כל התחלנו בחיפוש השם של כל אחת מהקטגוריות בפיצ'רים הטקסטואליים הבאים: **description/requirements/benefits**. זאת מכיוון ששמנו לב כי לרוב ייתכן כי יופיע שם אחת הקטגוריות של כל אחד מהפיצ'רים הנ"ל כך שנוכל להשלים באופן ישיר את הערך המתאים. אחרת, פנינו להשלמת הערך על ידי השמת הקטגוריה השכיחה ביותר

בהתאם לסוג המודעה – אמיתית או מזויפת, בחרנו בשיטה זו מכיוון שלא מצאנו אינדיקציה אחרת להשלים את אותם ערכים חסרים.

- **Industry & Function** – השלמנו ערכים חסרים ברשומה כלשהי מערך במאפיין השני אם הוא קיים. לאחר צמצום הקטגוריות בכל מאפיין, השלמנו את שאר הערכים החסרים על פי שכיחות הופעת כל קטגוריה בכל מאפיין.

4) <u>המרה של סוגי משתנים</u>:

- **Text Features** – עבור מאפיינים אלו פירקנו אותם למילים, סיננו מילים מקשרות והוצאנו את השורש עבור כל מילה. ביצענו זאת באמצעות stopwords ו- PorterStemmer מספריית NLTK. ביצענו עיבוד זה על המאפיינים: Title, Company Profile, **Description**, **Requirements**, **Benefits**.

```
0                          eroad applic form
1                        applic secur consult
3                            digit copywrit
4                               php develop
5    substanc abus counselor greenvil center
Name: title, dtype: object
```

- **Location** – השארנו את קוד המדינה בלבד (2 אותיות) תחת העמודה 'country'.
- **Salary** – במקום טווח המשכורת, נציב את הממוצע בין שני קצוות הטווח.
- **Required_education / experience** – צמצמנו את מספר הערכים ל-6 קטגוריות. הקטגוריות שקיבלנו מכילות את מרבית הקטגוריות הקיימות בפיצ'ר המתאר את רמת הניסיון הנדרש למשרה. במציאות, סביר להניח כי קיים קשר בין רמת ההשכלה לרמת הניסיון הנדרשת ולכן לאור העובדה כי מדובר בשני פיצ'רים דומים ובעלי אותה השפעה על משתנה המטרה בחרנו לבצע איחוד פיצ'רים זה.
- **Industry & Function** – ביצענו איחוד של קטגוריות בכל מאפיין כך שנשאר עם 6 קטגוריות בסופו של דבר. האיחוד התבצע על סמך מילות מפתח שחזרנו על עצמן
- הדפסת דגימה מה **dataset** בסיום השלב:

```
                                     title  ...    mean_salary
0                          eroad applic form  ...   67428.042239
1                        applic secur consult  ...   67428.042239
3                            digit copywrit  ...   67428.042239
4                               php develop  ...   67428.042239
5     substanc abus counselor greenvil center  ...   67428.042239
...                                      ...  ...            ...
6741   develop databas administr pittsburgh pa  ...   67428.042239
274                    cruis staff want urgent  ...   67428.042239
10188                             seo analysi  ...   67428.042239
854                          administr assist  ...      17.000000
5816                          ic e technician  ...  105000.000000

[16772 rows x 17 columns]
```

<u>Segmentation</u> – החלטנו לוותר על שלב זה משום שהמאפיינים ב- Data Collection שלנו לאחר שלב ה Pre-Processing הם ברמת גרעיניות מספיק טובה לצורך בניית המודל. אין לנו מאפיינים מורכבים כמו: קבצים, תמונות, סרטונים שנצטרך לבצע עבורם את שלב זה.

Feature Extraction

5) חילוץ מפיצ'רים מסוג Text:

- **Title**, **Company Profile**, **Description**, **Requirements**, **Benefits** – עבור כל אחד מהמאפיינים האלו נשתמש בשיטת **TF-IDF** לחילוץ מאפיינים נוספים. נבצע זאת באמצעות שימוש בספרייה sklearn. בחנו עבור כל מאפיין את הטקסטים שהוא מכיל כחלק ממודעה אמיתית או מזויפת. שיטה זו נותנת ציון לכל מילה בטקסט, המילים האינפורמטיביות ביותר היו אלו עם ההפרש הגבוה ביותר בין הציון שהן קיבלו במודעות מזויפות לבין מודעות אמיתיות (אותה מילה, תחת אותו מאפיין, שמופיעה גם במודעות מזויפות וגם באמיתיות). סביר להניח שמילים אלו יסייעו לנו לסווג בין סוגי המודעות, לכן עבור כל מאפיין נבחר את 100 המילים האינפורמטיביות ביותר ונצרף אותן כמאפיינים חדשים שחולצו.

6) חילוץ פיצ'רים מתוך ידע והיכרות עם הנתונים:

- **Continent** – מאפיין חדש שנבנה על סמך המאפיין Country באמצעות הספרייה **Country_Converter** שיודעת לזהות את היבשת של כל מדינה עפ"י קוד המדינה.
- **Qualification** – שמנו לב כי הערכים במאפיינים Education /Required Experience דומים מאוד, לכן החלטנו לחלץ פיצ'ר מאוחד המשלב את שניהם. בדקנו עבור כל רשומה רמות ההשכלה והניסיון אשר נופלות תחת אותה קטגוריה ושייכנו כל סאמפל לקטגוריה המאוחדת המתאימה לו. במציאות, סביר להניח כי קיים קשר בין רמת ההשכלה לרמת הניסיון הנדרשת ולכן לאור העובדה כי מדובר בשני פיצ'רים דומים ובעלי אותה השפעה על משתנה המטרה ואף מסקנות זהות כפי שניתן לראות בשלב ה-EDA בחרנו לבצע איחוד פיצ'רים זה.
- **Has_Salary** – מתבסס על המאפיין Salary_Range שיצרנו, בכל רשומה בה ישנו ערך של משכורת, אז המאפיין Has_Salary יקבל את הערך 1. אחרת יקבל את הערך 0.
- **Has_Department** - מתבסס על המאפיין Department, בכל רשומה בה ישנו ערך בעמודה Department, המאפיין Has_Department מקבל את הערך 1, אחרת 0.
- הדפסת דגימה מה- dataset בסיום השלב:

```
                                        title  ... benefits_see
0                            eroad applic form  ...          0.0
1                         applic secur consult  ...          0.0
2                               digit copywrit  ...          0.0
3                                  php develop  ...          0.0
4      substanc abus counselor greenvil center  ...          0.0
...                                        ...  ...          ...
16767  develop databas administr pittsburgh pa  ...          0.0
16768                  cruis staff want urgent  ...          0.0
16769                              seo analysi  ...          0.0
16770                          administr assist  ...          0.0
16771                           ic e technician  ...          0.0

[16772 rows x 521 columns]
```

Feature Representation

- מאפיינים מסוג טקסט – בשלב ה Feature Extraction השתמשנו בשיטת ה **TF-IDF** על מנת לחלץ מאפיינים נוספים ממאפיינים קיימים מסוג טקסט. המאפיינים הנוספים שחולצו הם מסוג נומרי, לכן בשלב זה הסרנו את המאפיינים המקוריים שהם מסוג טקסט משום שהם כבר מיוצגים על ידי מאפיינים אחרים.
- מאפיינים קטגוריאליים – השתמשנו בשיטת One Hot Encode שלוקחת כל מאפיין קטגוריאלי, ועבור כל ערך שהוא מקבל היא יוצרת משתנה בינארי חדש.
- לאחר המרת כל המאפיינים לנומריים, ביצענו נרמול של הערכים בשיטת Min-Max Scaling כך שכלל הערכים יהיו בטווח בין 0 ל- 1, אחרת יכולה להיווצר הטייה מסוימת במודל העתידי כתוצאה ממאפיינים עם ערכים גדולים יותר.
- הדפסת דגימה מה **dataset** בסיום השלב:

```
       telecommuting  ...  qualification_Unspecified
0                0.0  ...                        0.0
1                0.0  ...                        0.0
2                0.0  ...                        0.0
3                0.0  ...                        0.0
4                0.0  ...                        0.0
...              ...  ...                        ...
16767            0.0  ...                        0.0
16768            0.0  ...                        0.0
16769            0.0  ...                        0.0
16770            0.0  ...                        0.0
16771            0.0  ...                        0.0

[16772 rows x 641 columns]
```

Feature Selection

- כפי שנלמד בהרצאה, לא כל הפיצ'רים אינפורמטיביים ובייחוד לאור העובדה כי מביצוע השלבים הקודמים התקבלו לנו 641 פיצ'רים. נרצה לספק לאלגוריתם הלמידה סט של פיצ'רים קטן אך אינפורמטיבי ומקיף.
- בחרנו לבצע את שלב בחירת הפיצ'רים על ידי שימוש בהערכה כמותית על פני איכותנית מכיוון שהערכה איכותנית דורשת התערבות מומחה וכמו כן מדובר בוויזואליזציה של נתונים ובהיקף כזה של פיצ'רים לראות עינינו זו לא הערכה אפקטיבית או אפשרית.
- כחלק מאסטרטגיית הערכה כמותית השתמשנו ב-Wrappers procedures: בחרנו להשתמש ב- **(**RFE**)** Recursive Feature Elimination לביצוע תהליך של backward stepwise selection. בקוד שלנו בחרנו ברגרסיה לוגיסטית כמעריך **ב** RFE מכיוון שהמודל שלה פשוט וניתן לפרשנות וגם לאור העובדה שמודל זה מתאים במיוחד לבעיות סיווג בינארי, כמו הבעיה שלנו שבה משתנה היעד (הונאה או לא) הוא משתנה בינארי.
- **Gain Ration (GR)** – שיטה נוספת בה השתמשנו מתבסס על המדד של Information Gain (IG)**,** ההפחתה הצפויה בין האנטרופיה במידע. עפ"י שיטה זו נעדיף את הפיצ'רים שקיבלו IG גבוה. החלטנו לבחור את 100 הפיצ'רים הכי אינפורמטיביים.
- **K-Best** – בשיטה השלישית בה בחרנו להשתמש כל מאפיין מקבל ציון בהתאם לטבלת **ANOVA**, ובחרנו את 100 המאפיינים המדורגים ראשונים.
- הדפסת דגימה מה- dataset בסיום השלב:

```
   benefits_career  benefits_day  ...  title_sale  fraudulent
0              0.0           0.0  ...         0.0         0.0
1              0.0           0.0  ...         0.0         0.0
2              0.0           0.0  ...         0.0         0.0
3              0.0           0.0  ...         0.0         0.0
4              0.0           0.0  ...         0.0         0.0

[5 rows x 181 columns]
```

Dimensionality Reduction

- בשלב זה השתמשנו בשיטת PCA על מנת להפחית את מספר המאפיינים וכתוצאה מכך לשפר את פשטות המודל וזמן הרצתו. בפועל כמות המשתנים לא הצטמצמה כתוצאה משימוש בשיטה, אלא רק הערכים של כל מאפיין השתנו. בשלב הבא נוכל לבחון את המודל לפני ה PCA ולאחריו ולהשוות.

	benefits_career	benefits_day	...	title_sale	fraudulent
0	0.043117	0.013569	...	0.022077	0.0
1	-0.003224	-0.001889	...	-0.003500	0.0
2	-0.006032	0.018584	...	-0.020609	0.0
3	-0.002773	0.010210	...	0.020350	0.0
4	0.015025	0.022229	...	0.032164	0.0
...	...	...	...	...	...
16767	-0.003461	0.009266	...	0.027985	1.0
16768	0.004147	0.016915	...	0.023697	1.0
16769	-0.011143	0.026550	...	0.041689	1.0
16770	-0.009836	0.014522	...	0.022711	1.0
16771	0.042189	0.012553	...	0.051979	1.0

[16772 rows x 181 columns]

Model Training

- בחרנו בשיטת ולדיציה k-fold לנתונים. על סמך החומר שלמדנו בכתה בהרצאה 3, נלמד כי שיטת leave one out מתאימה למערך נתונים קטן מאוד ולכן אינה רלוונטית עבור המערך שלנו שכולל סדר גודל של כ-16,000 סאמפלים לאחר ביצוע שלב יצירת סט הנתונים. בחרנו בשימוש בשיטת k-fold מכיוון שמדובר במודל יקר מבחינה חישובית לאימון, שיטה זו מספקת אומדן אמין יותר של ביצועי המודל מאשר שיטת holdout.
- תהליך הולידציה שנבצע תוך שימוש בשיטה של k-fold, מתבצע באופן הבא:

  קודם כל מחלקים את הנתונים לtest data ו"אחר" (training set and validation set).

  מחלקים את ה"אחר" ל-**k** קיפולים (תתי קבוצות) השוות בגודלן, כאשר **k-1** קיפולים הם לטובת האימון (training set) והקיפול שנותר לטובת הולידציה (validation set). חוזרים על ההערכה עד שנכסה את כל **k** האפשרויות השונות. כלומר, אנו מאמנים את הדגם **k** פעמים, בכל פעם משתמשים בקפל אחר כסט הולידציה והקיפולים הנותרים כסט האימון.

  אנו מחשבים את מדד הביצועים עבור כל קיפול. לבסוף, מחשבים את ממוצע הביצועים של כלל הקיפולים והוא מהווה המדד להערכת ביצועי המודל.

  נבחר להעריך את המודל שלנו באמצעות מטריקה של AUC מכיוון שמדד זה משמש לעיתים קרובות בבעיות סיווג בינארי כפי שנתון בפרויקט שלנו. מדד AUC נחשב למדד חזק מכיוון שהוא אינו מושפע ממערכי נתונים לא מאוזנים ומספק הערכה מקיפה יותר של ביצועי המודל בניגוד למדדי דיוק או שגיאה. הרי בפרויקט שלנו כפי שנאמר קודם לכן הנתונים אינם מאוזנים ולכן מצאנו כי מדד AUC מהווה הבחירה האופטימלית עבורנו.

נספח 1 – EDA

1. **Title** – הדפסת 10 המילים הנפוצות ביותר במאפיין זה לפי סוג מודעה.

```
Most common words in fake titles:
[('entry', 85), ('data', 79), ('engineer', 63), ('assistant', 63), ('home', 57), ('manager', 56), ('payroll', 56), ('positions', 50), ('customer', 48), ('clerk', 47)]
Most common words in real titles:
[('manager', 1703), ('developer', 1409), ('engineer', 1217), ('sales', 1011), ('senior', 747), ('customer', 712), ('service', 661), ('english', 631), ('teacher', 623), ('marketing', 598)]
```

2. **Location** – הדפסת פרטים בסיסיים על המאפיין: כמות ערכים חסרים ומספר ערכים יחודיים:

```
Amount of null values: 281
Percentage of null values: 0.019644854586129756 %
Unique countries: 86
```

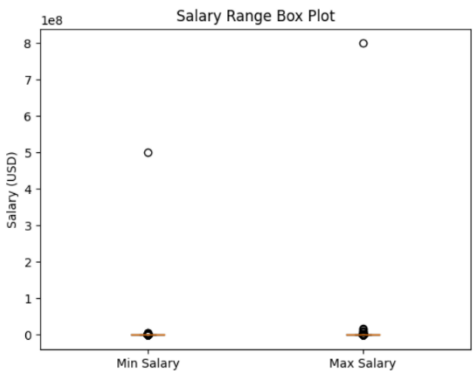
3. **Department** – צילום של אופן ביצוע איחוד הקטגוריות:

```
# Attempt to discretize 'department'
df_copy['department'].fillna('', inplace=True)
for i, department in enumerate(df_copy['department']):
    if any(substring in department.lower() for substring in ['sale',
                                                              'customer',
                                                              'marketing', 'cs', 'member', 'commercial', 'retail',
                                                              'support']):
        df_copy.loc[i, 'department'] = 'Sales'
    elif any(substring in department.lower() for substring in ['data','information','qa', 'it', 'tech','engineer', ' \
            ''research','development', 'r&d', 'product']):
        df_copy.loc[i, 'department'] = 'Tech'
    elif any(substring in department.lower() for substring in ['creative','design', 'media']):
        df_copy.loc[i, 'department'] = 'Creative'

    elif any(substring in department.lower() for substring in ['operations','admin', 'management', 'account',
                                                                'finance']):
        df_copy.loc[i, 'department'] = 'Operations'

print(df_copy['department'].nunique())
```

4. **Salary** – צילום של **box plot** הראשון לפני הסרת הערכים החריגים:



5. **Employment Type** – הדפסת השכיחות של סוג התעסוקה עפ"י סיווג המודעה:

```
The Freqyency of each Employment Type in fraud Advertisments is:  employment_type
Full-time    0.568794
NaN          0.278014
Part-time    0.083688
Contract     0.046809
Other        0.019858
Temporary    0.002837
Name: proportion, dtype: float64
The Freqyency of each Employment Type in real Advertisments is:  employment_type
Full-time    0.653945
NaN          0.189132
Contract     0.088242
Part-time    0.041915
Temporary    0.014486
Other        0.012280
Name: proportion, dtype: float64
The number of unique values in the feature employment type is: 5
```

6. **Required Experience** – הדפסת השכיחות של קטגורית הניסיון עפ"י סיווג המודעה:

```
The Frequency of each Required Experience Type in fraud Advertisments is:  required_experience
NaN                 0.500709
Entry level         0.207092
Mid-Senior level    0.129078
Not Applicable      0.073759
Associate           0.048227
Director            0.017021
Executive           0.014184
Internship          0.009929
Name: proportion, dtype: float64
The Frequency of each Required Experience Type in real Advertisments is:  required_experience
NaN                 0.387455
Mid-Senior level    0.218619
Entry level         0.147437
Associate           0.131701
Not Applicable      0.062284
Internship          0.022208
Director            0.021987
Executive           0.008309
Name: proportion, dtype: float64
The number of unique values in the feature required experience is: 7
```

7. **Required Education** – הדפסת השכיחות של קטגורית ההכשרה עפ"י סיווג המודעה:

```
The three most common Required Education Types in fraud Advertisments are:  required_education
NaN                          0.520567
High School or equivalent    0.197163
Bachelor's Degree            0.114894
Name: proportion, dtype: float64
The three most common Required Education Types in real Advertisments are:  required_education
NaN                          0.450254
Bachelor's Degree            0.297522
High School or equivalent    0.111185
Name: proportion, dtype: float64
The number of unique values in the feature required education is: 13
```

8. **Description**:

- הבחנה על סמך מאפיין זה לאור המילים השכיחות:

- הדפסת ההבחנה בין סוגי המודעות לפי אורך מאפיין זה:

```
Average description length of fake ads: 1146.8893617021276 words
Average description length of real ads: 1173.0768495366965 words
Words difference: 26.187487834568856 words
Average sentiment score for real ads: 0.85
Average sentiment score for fake ads: 0.80
```