Machine learning project:

part 1

מגישים:

ענבר דובדבני: 206104028

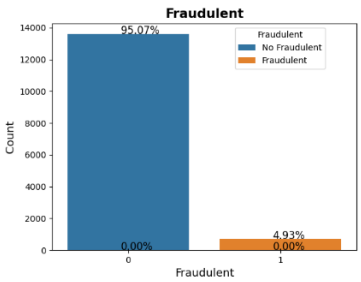
אמיר יטיב: 207128513

Data collection and Sensing

- Data collection הוא השלב הראשון בתהליכים ומודלים של מערכת ML. בשלב זה נעשה איסוף של ישויות מהעולם האמיתי. נרצה שה- Data collection יהיה אמין, מייצג, לא חזרתי ובעל אפשרות לליבלינג אם אפשרי. במקרה שלנו, נעשה איסוף רחב של מידע על הודעות דרושים מהאינטרנט מכל הסוגים ומהווה בסיס להמשך התהליך. Sensing הוא השלב השני בתהליך ובו אנו הופכים את ה samples למידע גולמי. במקרה שלנו איסוף הנתונים נעשה בצורה סטאטית ומוגדר כך שנתון כלשהו אינו משתנה מרגע לרגע. את המידע על כל ישות (ת.ז של מודעה) אנו אוספים באופן נפרד ואותו אנחנו מחלקים למאפיינים.
- ה- sensing שבוצע על הדאטה שלנו הוא מסוג סטטי מפני שכלל הדאטה שקיבלנו אינו משתנה בזמן.
- נציע לבצע sensing דינאמי (בכל רגע הנתונים משתנים) על כמות המשתמשים שראו כל מודעת עבודה. נרצה לדעת לאורך זמן את כמות המשתמשים שנחשפו למודעה. אנחנו יוצאים מנקודת הנחה כי מודעות עבור משרה מסוימת תהיה חשופה לזמן מוגבל- עד הרגע שלא יזדקקו לה יותר (מציאת עובדים) ואילו מודעות זדוניות ימשיכו להיות חשופות גם הרבה אחרי שפורסמו. כתוצאה מכך כל עוד מספר הנחשפים גדל לאורך זמן עולה החשד שהמודעה הינה מזויפת.
- משימת הלמידה הינה ביצוע סיווג (classification) מודעת הדרושים (הונאה או תמימה) על סמך מאפיינים שונים המופעים במודעה אשר להם תימצא השפעה על אופן הסיווג. משימת הלמידה נכנסת תחת למידה מונחית (supervised Learning) היות ואת הסיווג אנו מבצעים בצורה בינארית- מזויף (1) או תמים (0). ניתן לבצע משימת למידה של Reinforcement Learning שכן בסוג למידה זה אנו מעריכים אם המודל שבנינו הוא טוב על סמך סיווג התצפיות במציאות. ובמקרה שלנו כאשר לנתונים יש שתי קטגוריות בלבד, Reinforcement Learning שווה ערך ללמידה מפוקחת סטנדרטית.

Dataset Creation

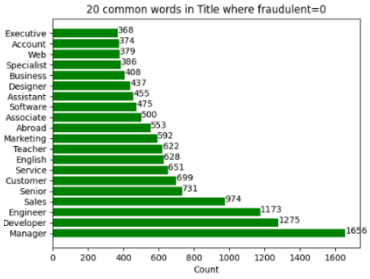
Exploratory data analysis

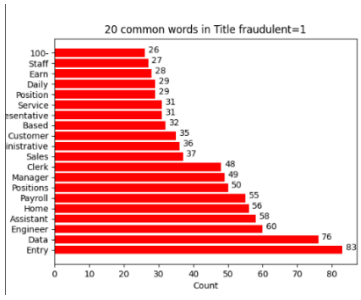


Fraudulent- ניתן לראות כי סט הנתונים איננו מאוזן כלל מכיוון שקיים מספר רשומות רב עבור משרות אמיתיות (13,600) לעומת משרות הונאה (705)

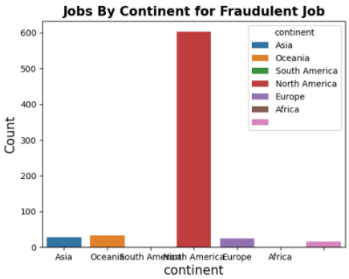
Title- בחרנו להראות את עשרים המילים הנפוצות ביותר ב"כותרת המשרה" על מנת לראות אם מילים מסוימות יכולות להעיד לנו על משרה שהיא הונאה בהשוואה. ניתן לראות כי יש חפיפה קטנה בין מספר מילים (רק engineer,

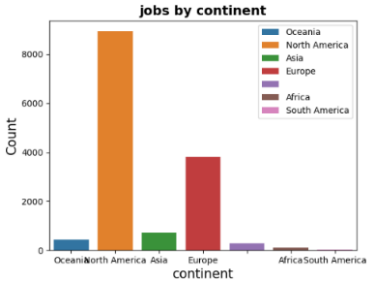
manager ו – sales נמצאות בשניהם וכל השאר שונות) ולכן מילים אלו לא יתרמו לנו בהמשך. בנוסף ישנן מילים אשר מופיעים במודעות הונאה שאנו מעריכים כי יהיו אטרקטיביות עבור מחפשי המשרות (כמו home ו- payroll).





<u>Location</u>-פילוג היבשות המופיעות בנתונים. ניתן לראות כי הערכים בפיצ'ר זה אינם מאוזנים. נרצה לבדוק באילו יבשות נמצאות רוב המשרות ואילו מתוכן הן משרות הונאה. ניתן לראות כי רוב המשרות מוצעות באמריקה הצפונית וכי בהסתכלות על משרות ההונאה רובן גם הן נמצאות באמריקה הצפונית בהפרש ניכר מהשאר.





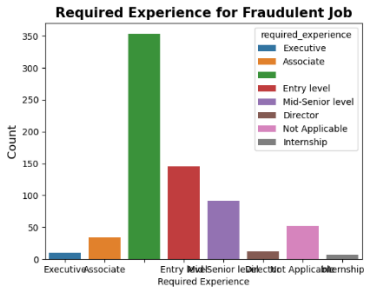
-

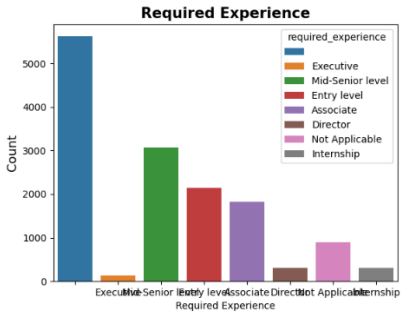
<u>Requirements</u>- <u>עשרים במילים הנפוצות בדרישות</u>

בחרנו להראות את עשרים המילים הנפוצות ביותר בדרישות על מנת לראות אם מילים מסוימות יכולות להעיד לנו על משרה שהיא הונאה. ניתן לראות כי המלים שקיבלו את הערך הגבוה ביותר כמו experience, work, skills, years מופיעות בשני סוגי המודעות והדבר יקשה לנו ללמוד מכך. נוסף על כך עשינו בדיקה על כמות המילים המופיעות בפיצ'ר (נספח 1) וזאת על מנת לראות קשרים שנציין בהמשך. נראה את הנתונים טבלה:

ערך\מודעה	כמות מילים במודעה אמיתית	כמות מילים במודעת הונאה
ערך ממוצע -	78.3	58.5
כמות ערכים ריקים -	2059 (%15.1)	471 (%66.8)
ערך ממוצע ללא ערכים ריקים -	92.2	71

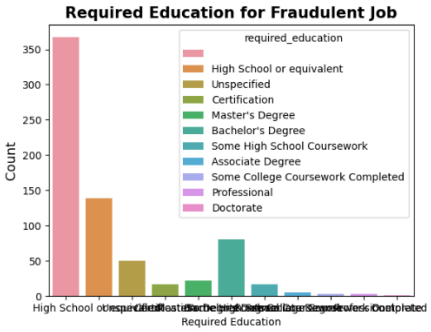
Required Experience- ניתן לראות כי הערכים בתוך פיצ'ר זה אינם מאוזנים. רוב המשרות לא מכילות ערכים בפיצ'ר זה והמספר של משרות ההונאה עבור ערכים אלו הוא הגבוה ביותר בהתאם בגרף השמאלי. היחס הזה נשאר גם עבור ערכים אחרים בפיצ'ר כמו Mid-Senior level ו- Intemship ויתכן שלא יתרום לנו בהמשך.

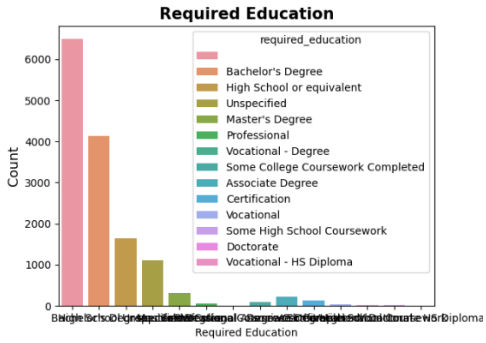




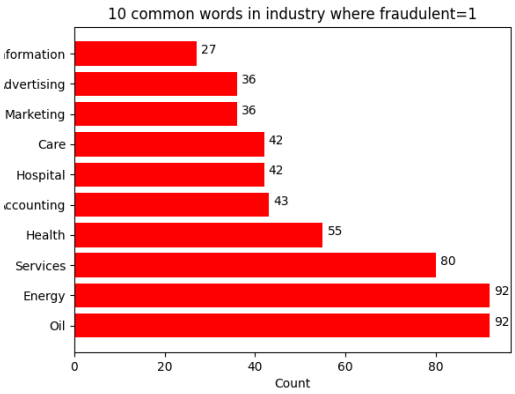
Required Education- ההשכלות הנדרשות לעבודות

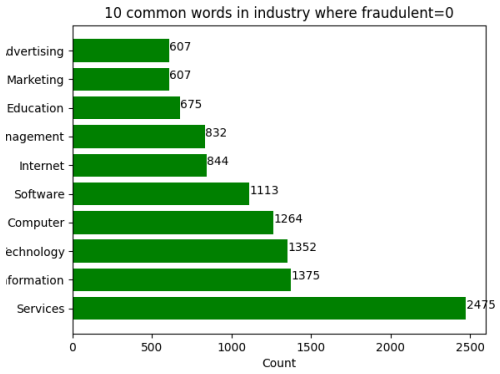
ניתן לראות כי הערכים בתוכף פיצ'ר זה אינם מאוזנים. רוב המשרות לא מכילות ערכים בפיצ'ר זה והמספר של משרות ההונאה עבור ערכים אלו הוא הגבוה ביותר משמעותית משאר המשרות. ניתן לראות שערכי Bachelor's Degree, Hight School or equivalent ו- Unspecified שומרים על יחס המקומות שלהם בין שני הגרפים.



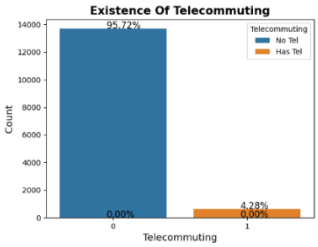


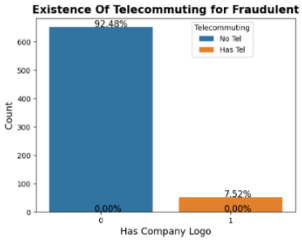
industry-בחרנו לבדוק מהן עשר המילים שמופיעות הכי הרבה בפיצ'ר industry על מנת שיכוון אותנו לתעשיות שעשוי להיות בהם כמות גדולה יותר של משרות הונאה. ואכן, שבע מהמילים הנפוצות ביותר במשרות הונאה אינן מופיעות כלל ברשימת עשרת המילים הנפוצות ביותר במשרות אמיתיות, מה שעלול לרמז לנו על הונאה.



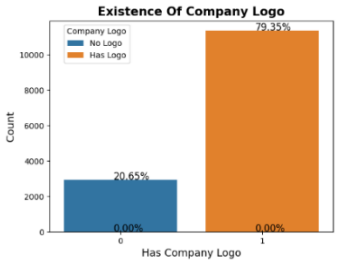


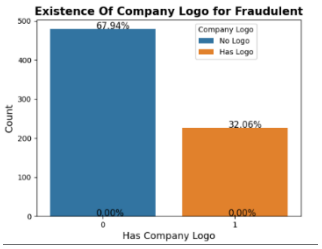
Telecommuting- בכמה מהעבודות ניתן לעבוד מרחוק. ניתן לראות כי ערכי הפיצ'ר כלל אינם מאוזנים. נראה כי יש מעט משרות באופן כללי שמבצעות עבודה מרחוק וישנם פחות מ-100 שהן משרות הונאה.





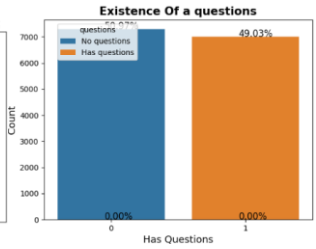
Has Company Logo- ניתן לראות כי כמות המשרות שיש בהן שאלון ואין בהם שאלון הוא אינו מאוזן. ברוב המשרות ניתן לראות כי יהיה לוגו לחברה ועבור משרות הונאה בערך ל2/3 לא יהיה לוגו ול1/3 יהיה.

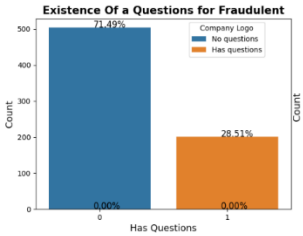




Has questions-

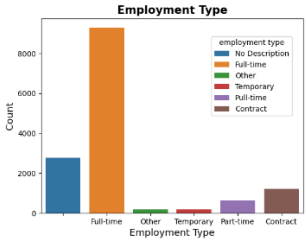
ניתן לראות כי כמות המשרות שיש בהן שאלון ואין בהם הוא יחסית מאוזן. לעומת זאת עבור משרות ההונאה ניתן לראות כי בערך ל2/3 לא יהיה שאלון ול-1/3 יהיה שאלון. נשים לב כי ישנו יחס דומה עבור משרות ההונאה בין פיצ'ר זה לפיצ'ר Has Company Logo שעלול להעיד על קשר בהמשך.

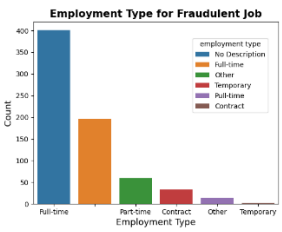




Employment Type-

ניתן לראות הגרף אינו מאוזן וכי רוב המשרות הן Full-time הן בפער משמעותית יותר גדול משאר סוגי המשרות לכן גם מספר משרות ההונאה יהיה גדול יותר עבור סוג עבודה זה אך רק פי 2 מהמקום השני שהוא משרות בן לא הוזן סוג העבודה.

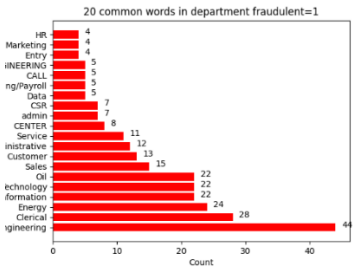


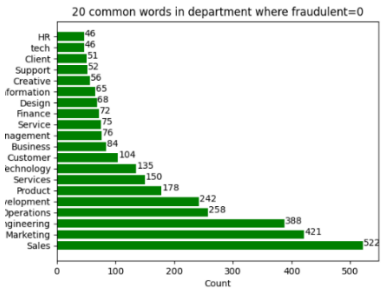


ערך\מודעה	כמות המילים במודעה אמיתית	כמות המילים במודעת הונאה
ערך ממוצע -	171.7	159.2
כמות ערכים ריקים -	0	1
ערך ממוצע ללא ערכים ריקים -	171.7	159.4

Description- בחרנו להראות את עשרים המילים הנפוצות ביותר ב"תיאור משרה" על מנת לראות אם מילים מסוימות יכולות להעיד לנו על משרה שהיא הונאה. כמעט חצי מהמילים מופיעות רק במשרות הונאה או רק במשרות אמיתיות ואילו השאר מופיעות בשני הגרפים. עובדה זו יכולה לעזור לנו להחליט אילו מילים יתרמו לנו לסווג ואילו לא יועילו (אלו שחוזרות על עצמן). נוסף על כך עשינו בדיקה על כמות המילים המופיעות בפיצ'ר (נספח 2) וזאת על מנת לראות קשרים שנציין בהמשך. נראה את הנתונים טבלה הבאים:

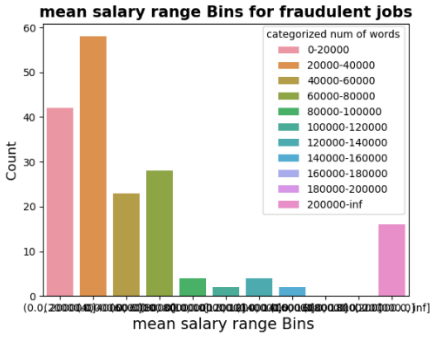
Department- בחרנו להראות את עשרים המילים הנפוצות ביותר ב"מחלקה" על מנת לראות אם מילים מסוימות יכולות להעיד לנו על משרה שהיא הונאה. נראה שבאופן כללי אין הרבה מילים במשרות שהן הונאה (267) עבור הסלקציה שעשינו לעומת משרות אמיתיות (3089). על כן מילים שמופיעות רק במשרות הונאה (clerical) עשויות לתרום לנו. (נבדוק את הנתון הזה בשלב של הfeature selection)





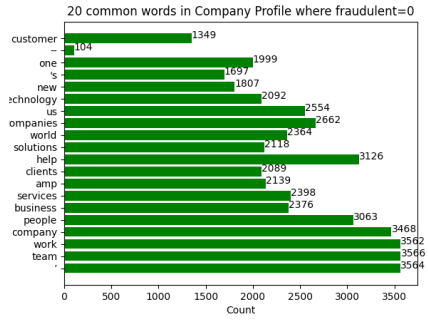
Salary Range- בחרנו להראות את ממוצע טווח המשכורות. ראשית ניתן לראות כי סקאלת המספרים נמוכה מאוד ביחס לכמות הרשומות שקיבלנו ומראה כי עבור פיצ'ר זה הרבה מהרשומות לא קיבלו טווח משכורת והן אינן מוצגות בגרף זה. כדי שיהיה לנו נוח לראות את ערכי פיצ'ר זה בחרנו להראות את ממוצע הטווחים ולמקד יותר את הערכים בפיצ'ר זה. ניתן לראות כי ככל שממוצע המשכרות עולה השכיחות יורדת עבור שני הגרפים, למעט משרות בהן ממוצע המשכורות הוא מעל 200,000 אותם בחרנו לייצג כקטגוריה אחת. כתוצאה מכך אנו מניחים כי עבור משרות שונות ממוצע המשכורות עלול להיות שונה, היות

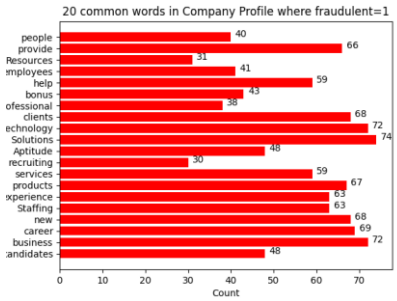
ולא אנחנו ביצענו את שלב ה-Data Collection ולכן לא נוכל לדעת באיזה סוג טווח משכורות מדובר (שנתית/שבועית/חודשית) ובאיזה מטבע (שקל/יורו/דולר).



Company Profile- בחרנו להראות את עשרים המילים הנפוצות ביותר ב"פרופיל החברה" על מנת לראות אם מילים מסוימות יכולות להעיד לנו על משרה שהיא הונאה. נראה שבאופן כללי אין הרבה מילים במשרות הונאה ביחס למשרות אמיתיות. על אף עובדה זאת, ישנם מספר מילים שניתן יהיה לראות אותם רק במשרות הונאה כמו candidates ו- career שאנו מעריכים כי יהיו אטרקטיביים עבור מחפשי המשרות. נוסף על כך עשינו בדיקה על כמות המילים המופיעות בפיצ'ר (נספח 3) וזאת על מנת לראות קשרים שנציין בהמשך. נראה את הנתונים טבלה הבאים.

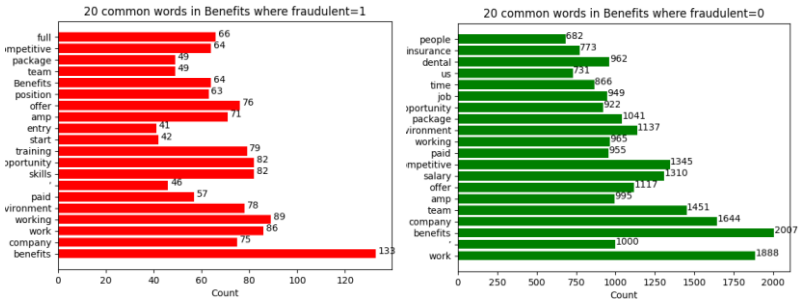
ערך\מודעה	כמות המילים במודעה אמיתית	כמות המילים במודעת הונאה
ערך ממוצע -	95.5	30.9
כמות ערכים ריקים -	(%16.1)2200	(%68.7)484
ערך ממוצע ללא ערכים ריקים -	114	98.42



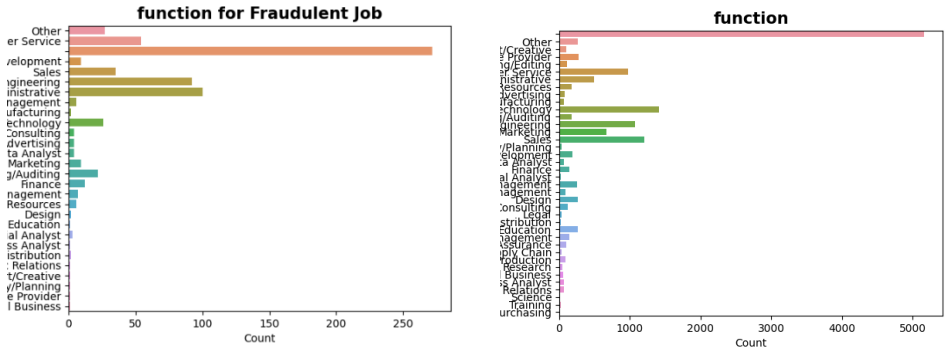


Benefits- בחרנו להראות את עשרים המילים הנפוצות ביותר ב"הטבות" על מנת לראות אם מילים מסוימות יכולות להעיד לנו על משרה שהיא הונאה. נראה שבאופן כללי אין הרבה מילים במשרות הונאה ביחס משרות אמיתיות. על אף עובדה זאת, ישנם מספר מילים שניתן יהיה לראות אותם רק במשרות הונאה כמו skills ו- opportunity שאנו מעריכים כי יהיו אטרקטיביים עבור מחפשי המשרות. נוסף על כך עשינו בדיקה על כמות המילים המופיעות בפיצ'ר (נספח 4) וזאת על מנת לראות קשרים שנציין בהמשך. נראה את הנתונים בטבלה

ערך\מודעה	כמות המילים במודעה אמיתית	כמות המילים במודעת הונאה
ערך ממוצע -	29.9	30.5
כמות ערכים ריקים -	5462 (%40.1)	291 (%41.2)
ערך ממוצע ללא ערכים ריקים -	50.1	51.1



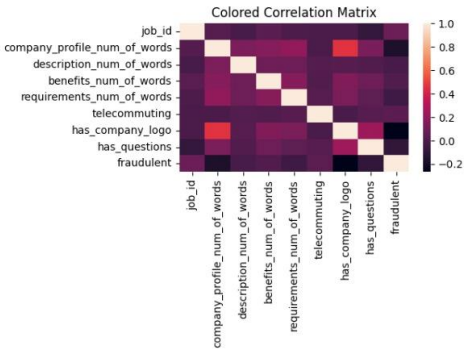
Function- רצינו לראות מהם המחלקות הנפוצות ביותר במשרות באופן כללי ומהן המחלקות הנפוצות ביותר במשרות הונאה. ניתן לראות כי הערכים בפיצ'ר זה אינם מאוזנים וכי רוב המשרות הן ללא תיוג של מחלקה כלל, וכתוצאה מכך גם מספר משרות ההונאה בקטגוריה זו. עם זאת ניתן לראות כי משרות ההונאה נמצאות במספר מצומצם יותר של מחלקות.



-

קשרים בין משתנים

בנוסף יצרנו טבלת קורלציה המראה את חוזק הקשר בין כל שני משתנים נומרים שבבעלותנו. ניתן לראות שאין קשרים חזקים בין המשתנים אך יש קשר אחד שמתבלט – מספר המילים בפרופיל החברה והאם יש לוגו לחברה. קשר זה יכול ללמד אותנו על מהימנות המודעה. אנו מאמינים שאם גם חסר לוגו החברה וגם מספר המילים שמתאר את החברה הוא נמוך יחסית זאת סיבה טובה לחשוד שהמודעה הינה מודעת הונאה.



Pre-processing

בשלב זה של התהליך נפרט על המניפולציות בנתונים שביצענו וזאת שנוכל להשתמש בהם טוב יותר בשלבים הבאים.

כפילות של נתונים

על מנת לאתר כפילויות הסתכלנו על שילוב של פיצ'רים מסוימים כמו location, title ו- company profile שבעיננו שילוב הנתונים ביניהם מהווה ייחודיות למודעת דרושים. היות ולא מצאנו אף שילוב של השלושה שחוזר על עצמו לא מצאנו כפילויות בנתונים.

נתונים חסרים

פיצ'ר	location	department	salary range	company profile
מספר חוסרים	281	9226	11,991	2684
אחוז חוסר מכלל הרשומות	1.964%	64.499%	83.829%	18.763%
פיצ'ר	**description**	**requirements**	**benefits**	**employment type**
מספר חוסרים	1	2181	5744	2768
אחוז חוסר מכלל הרשומות	0.007%	15.247%	40.156%	19.351%
פיצ'ר	**required experience**	**required education**	**industry**	**function**
מספר חוסרים	5622	6490	3929	5164
אחוז חוסר מכלל הרשומות	39.303%	45.371%	27.467%	36.101%

ניתן לראות כי לפי הטבלה ניכר כי יש חוסרים רבים בנתונים. הערכים החסרים מהווים כ-78.404% מכלל הנתונים. היות וזהו אחוז גבוה מאוד, בחרנו שלא להשלים את החוסרים ולקבל אותם כחלק מהנתונים. נציין עם זאת כי קיימת אפשרות של השלמת נתונים על פי הסקה מרשומות אחרות (לדוגמא להשליך מה department אל industry או מ "title" על "location") אך אנו חוששים מהאפשרות שלא תהיה השלמה נכונה עבור כלל הרשומות ועלולה להטות את תוצאות המודל. בנוסף נציין כי חשבנו להתייחס לרשומה שבה יש חוסר יחיד בפיצ'ר description וגלינו כי עבור משרה זאת ישנם 9 ערכים חסרים והיא מתויגת כמשרה מזויפת. היות וכאמור הנתונים שלנו אינם מאוזנים וכי יש רוב גדול של משרות אמיתיות, ראינו לנכון להשאיר רשומה זו.

המרות של ערכים

ביצענו דיסקרטיזציה לפיצ'ר "טווח המשכורת". לאחר שהמרנו את טווחי הערכים לממוצע הטווח, חילקנו את התוצאה לטווחים של משכורת ממוצעת כך שכל טווח הוא בגודל 20,000 יחידות, והטווחים הם מוצגים באופן מצבר.

פרופורציה הנתונים

לאחר בדיקה נראה שמספר המשרות המזויפות קטנה משמעותית ממספר המודעות האמיתיות במקור הנתונים ומהווה רק 4.93% מכלל המודעות. לא הצלחנו למצוא מידע מספיק מהימן מספיק ברשת שיספק לנו מהו אחוז משרות ההונאה במציאות ולכן אנו רואים את מקור הנתונים כמייצג כמו שהוא את המציאות.

Segmentation

אנו מבצעים את השלב הזה בכדי למקד למחשב את הנתונים לנתונים שאיתם יהיה לו נוח יותר לעבוד. הנתון היחידי שהחלטנו לבצע עבורו את שלב זה הוא "location" שעבורו המרנו

את המיקומים שבידינו ממדינו וערים – ליבשות. ביצענו את ההמרה הזו כיוון שזה יכול לעזור למקד את המאמצים ולא "ללמוד" את ההיתכנות של כל מדינה בפני עצמה.

Feature extraction

בשלב ה feature extraction אנו מחלצים מהנתונים שבידינו נתונים חדשים אשר יכולים ללמד אותנו על פונקציית המטרה.

בנתונים שקיבלנו יש מספר רב של אלמנטים טקסטואליים אשר נרצה לבצע עליהם פעולות שיעזרו לנו לנתח האם המודעה היא מודעת הונאה או לא. בשביל שיהיה לנו קל יותר לעבוד עם הנתונים האלו החלטנו להפוך את הנתונים הטקסטואליים (הארוכים מבניהם) לנתונים נומריים. עבור הפיצ'רים "company profile" "description" "requirements" ו-"benefits" ספרנו את כמות המילים שקיימות בכל תיאור, חישבנו ממוצע וספרנו את כמות הערכים החסרים (נספחים). עבור הפיצ'רים הטקסטואליים "company profile" "benefits" "department" "description" "industry" "requirements" ו-"title" הוצאנו את המילים עם החזרתיות הגבוהה ביותר במטרה למצוא בסיס שחוזר רק במודעות אמיתיות ולא בהונאות או ההיפך. פעולה נוספת שעשינו על מנת לחלץ נתונים היא על הפיצ'ר "salary". בפיצ'ר זה קיבלנו טווח ערכים שהמשכורת יכולה לנוע בהם. לטווח זה הוצאנו ממוצע, בכדי לקבל ערך יחיד שיותר נוח לעבוד איתו, וחילקנו ערך זה לקטגוריות לפי קפיצה קבועה (כמו שנאמר בשלב הדיסקרטיזציה).

Feature representation

שלב זה נועד לקבע את הסקאלה שאותה המחשב יכול לקבל עבור כל הפיצ'רים. שלב זה חשוב מכיוון שלמחשב יש נטייה לייחס חשיבות גדולה יותר למספרים גדולים יותר ואנו נרצה להימנע מנטייה זו. הסוג הראשון הוא נתונים בינארים ("has questions" " has company logo" ו-"telecommuting") . אותם השארנו כמו שהם.

הסוג השני של נתונים שיש בידינו הם נתונים מספריים שהוצאנו מטקסטים רבי מלל- **כמות מילים שנמצאות** ב: "company profile" "description" "requirements" ו-"benefits".

בשביל לנרמל נתונים אלה, חילקנו את עמודות אלו בערך המקסימאלי.

סוג הנתונים השלישי שבידנו הוא המילים הנפוצות ביותר לתיאור כל פיצ'ר במשרה.

בידינו שני רשימות של מילים – המילים שמופיעות הכי הרבה במודעות אמיתיות ומילים שמופיעות הכי הרבה בהונאות. ספרנו את כמות המילים שמופיעות בכל אחת מהרשימות (מילים אשר קיימות בשתיהן לא נספרו). בכדי לנרמל את כמות המילים נסכום את שני המספרים שקיבלנו ונחלק את כמות המילים שנמצאות ברשימה המודעות האמיתיות בסכום.

הסוג הרביעי הוא פיצ'רים מסוגים קטגוריאליים כמו employment type ו- required experience אותם השארנו כמו שהם.

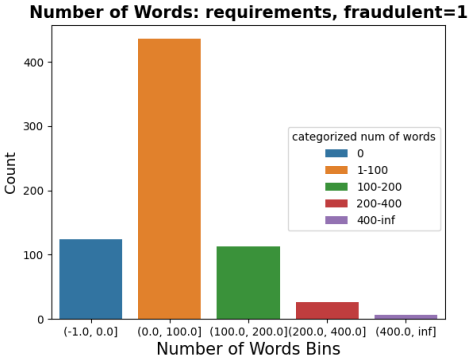
Feature Selection:

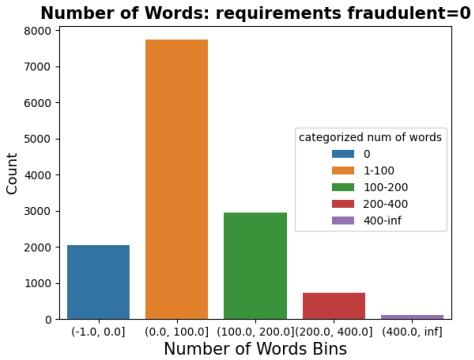
Model training:

על פי השיטות שנלמדו בכיתה לפי דעתנו כדאי להשתמש בשיטת הוולידציה K-fold. בחרנו בשיטה זו כיוון שהיא עוברת על כל הנתונים ומספר הריצות שלה יהיה לבחירתנו. שיטת הholdout נותנת לנו לפצל את הנתונים לשני קבוצות לבחירתנו – לימוד ומבחן. לא בחרנו בשיטה זו כיוון שבדרך זו אנחנו יכולים לאבד חלק מתהליך הלימוד אם נפריד נתונים שהם רלוונטים ללמידה שלנו, בשיטת הK-fold אנו משתמשים בכל הנתונים (אמנם באמת לא בכל ריצה) ונמנעים ממחסור במידע. שיטת הleave one out לא כל כך ריאלית כמות נתונים גדולה כמו שלנו כיוון שזמן הריצה שלה יהיה גבוה במיחד(מכיוון שעל כל שורה עוברים ככמות השורות פחות אחד).

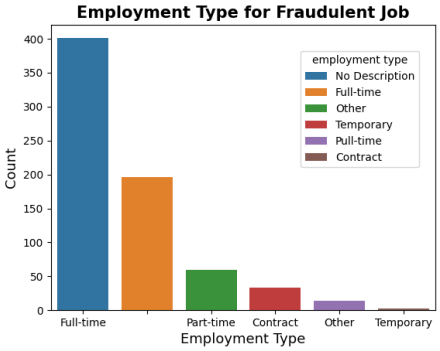
נספחים

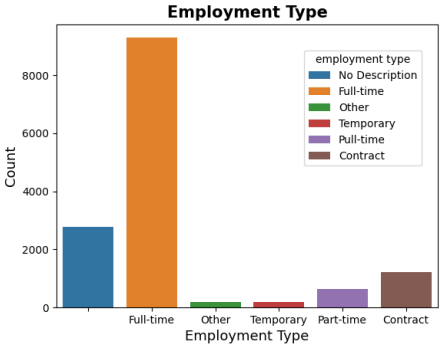
מספר מילים בפיצ'ר "Requirements "



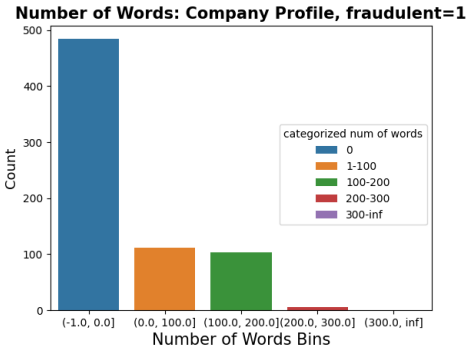


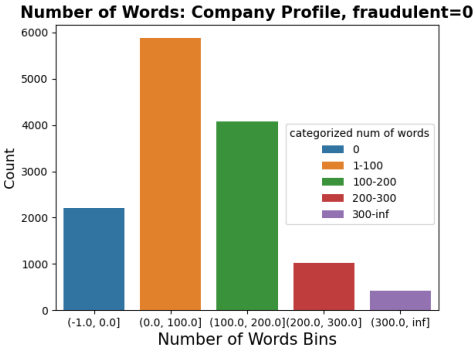
מספר מילים בפיצ'ר " Description "





מספר מילים בפיצ'ר " company profile "





מספר מילים בפיצ'ר "benefits "