עבודה בלמידה

לא מפוקחת

מגיש 1 : נעם פרץ

ת.ז : 206949398

מגיש 2 : אבירם חדד

ת.ז : 200791945

מבוא:

לכל חברה אשר מקיימת קשר עם העולם העסקי שבחוץ ורוצה להציג את הדוחות הכספיים מתקיים קשר שותף עם רואי חשבון אשר מסתכם בסופו של דבר לדוח ביקורת שנתי של רואי החשבון על החברה.

הדאטה שלנו מציגה למעשה כ-21000 דוחות ביקורת של רואי חשבון בסרביה שכוללים בתוכו משתנים קטגוריאליים ועל סוג המבקר וסוג החברה ומשתנים מספריים של דוחות הביקורת.

בתחילה נבצע הורדת מימד לפי PCA ל-2 ו-3 מימדים על מנת לראות האם ניתן לבצע הורדת מימד לדאטה מבלי לאבד יותר מידי מידע. בנוסף נבצע הורדת גם בשיטת t-SNE ונבדוק את ההבדל בין שיטה זו ל-PCA.

לאחר מכן נרצה לבצע את האשכולים הבאים:

1. K-means.
2. Agglomerative Clustering.
3. GMM.

ולבחון האם קיימים מאפיינים שיכולים להעיד על דוח ביקורת חיובי/שלילי/נייטרלי.

בנוסף הדאטה שלנו מחולקת לדוחות שנוצרו ע"י מבקרים מארבעת החברות הגדולות (BIG4) ומחברות אחרות ונרצה לבחון האם קיים איזשהוא קשר בין מוצא הדוח לבין הדוח עצמו.

דבר אחרון שיעניין אותנו זה לראות האם שינוי מבני של החברות היו יכולות להיחזות כאשכול בפני עצמו או לחילופין האם ה- BIG4 ידעו לזהות טוב יותר את המקרים הנ"ל.

ניקוי וסידור הנתונים:

הסתכלות ראשונית:

דבר ראשון, נבדוק את הנתונים שלנו, המסודרים בטבלת csv.

פתחנו פרויקט פייתון שיעזור לנו להשתמש בנתונים ולערוך אותם.

בתחילה בדקנו את גודל הדאטה שלנו:

התוצאה הנתונה היא : (22394, 411) – כלומר, 22394 שורות, 411 פיצ'רים שונים.

לאחר מכן, בדקנו ממה מורכבים המשתנים של הדאטה:

קיבלנו שהמשתנים שלנו הם מסוג int,float,object.

ניקוי נתונים :

בשלב זה רצינו לסנן דאטה שחסרה באופן מהותי. לכן בדקנו עבור אילו פיצ'רים חסרים מעל 20 אחוז מהנתונים :

והתוצאה היא :

```
['Auditor_switch' 'Auditor_name' 'Auditor_id' 'Big4' 'Audit_opinion'
 'Audit_opinion_1_code']
```

ניתן לראות שרק לחלק קטן מאוד מהפיצ'רים חסרים מעל 20 אחוזים מהנתונים! נתבונן על הפיצ'רים שיצאו לנו :

Auditor_switch:

פרמטר שאומר האם המבקר התחלף בשנה הזו או לא. בעמודה זו חסרים כמעט 58 אחוז מהנתונים. בגלל התלות של המשתנה הזה במשתנה auditor_name(אפשר לראות האם התחלף או לא), החלטנו למחוק אותו לגמרי.

Auditor_name:

פרמטר שנותן לנו את שם המבקר בשורת נתונים. בעמודה זו חסרים 40 אחוז מהנתונים, אך בגלל שהיא קטגורית ולא מספרית, החלטנו להשאיר אותה, ולשים במקומות הריקים פשוט "Unknown".

Auditor_id:

פרמטר שנותן לנו מספר מזהה של המבקר בשורת נתונים. בעמודה זו חסרים 40 אחוז מהנתונים. בגלל שאין משמעות למספר, והוא רק זהות, החלטנו לתת לכל המקומות הריקים id=0.

Big4:

פרמטר שנותן 1 אם חברת הביקורת היא מ4 הגדולות ו0 אחרת. החלטנו לתת ערך -1 לכל אחד מהמקומות הריקים.

Audit_opinion:

פרמטר שמראה מה חושבת חברת הביקורת על החברה הפיננסית, בעלת 4 קטגוריות שונות – disclaimer, adverse, qualified, unqualified. החלטנו לתת ערך "Unknown" לכל אחד מהמקומות הריקים.

Audit_opinion_1_code:

פרמטר שנותן 0 אם הaudit_opinion הוא unqualified ו1 אחרת. בגלל התלות שלו במשתנה השני, החלטנו להוריד אותו לגמרי.

המשך ניקוי נתונים :

מהמשך הסתכלות על הנתונים, ראינו שחסר עוד מעל אחוז של מידע בעמודות CEO_id, Year_Of_Establishment, וIndustry_Code. מהסתכלות על המידע בטבלת אקסל, ראינו שהמידע של Year_Of_Establishment, Industry_Code חסר בדיוק באותם Instances בדאטא, ולכן החלטנו להוריד אותם.

בנוסף, שמנו במקומות הריקים בCEO_id 'Unknown' בחלקים החסרים.

לאחר מכן, ראינו שחסרים נתונים גם בהמון עמודות AOP, וראינו שהאחוז שחסר הוא מאוד נמוך. לכן, בדקנו את האחוז ע"י הפקודה הבאה :

```
print(data.isnull().values.ravel().sum() / data.shape[0])
```

קיבלנו שהאחוז שיש בו nulls הוא 0.4 אחוז, ולכן החלטנו לזרוק את כל הדאטא הבעייתי על ידי הפקודה :

```
data = data.dropna()
data.to_csv('AuditData.csv')
```

כעת, יש לנו דאטא נקי מתאים ריקים.

חקירת הדאטא וויזואליזציה:

בשלב זה רצינו להתחיל לראות איך הדאטא נראית לפני שאנחנו עושים איזשהו עיבוד על הדאטא. תחילה יש לשים לב כי הדאטא מחולקת באופן גס ל-3 קטגוריות:

1. נתונים על הדוח ביקורת של רואה החשבון (כל מה שמוגדר כ-AP, שנת הדוח, האם כותב הדוח היה חלק מה- Big4, ודעתו של רואה החשבון- Audit Opinion)
2. נתונים כללים על החברה עליה התבצע הדוח (שם החברה, האם היא מוגדרת כארגון/בע"מ וכו', האם היא פשטה את הרגל או ביצעה שינוי מבנה חברתי ושנת הקמה של החברה)
3. נתונים כללים על הדאטא (מי כתב את הדוח ביקורת, מספרי זיהוי שונים)

בתחילה בדקנו את ההתפלגות של הפיצ'רים הבאים:

1. Year- קיבלנו כי ההתפלגות כמעט יוניפורמית לגמרי.
2. Audit opinion- כפי שמצופה לראות בדוחות ביקורת, החלוקה של חוות הדעת מתחלקת באופן לא שווה כאשר יש מעט מאוד חוות דעת שליליות או הימנעות מחוות דעת, הרוב המוחלט הן חוות דעת "נקיות" או כאלה שיש להם סייג כלשהו. בנוסף, ניתן לראות שיש כ-40% מהדאטא שלא נתונה.
3. Big 4- ניתן לראות כי כ-20% מהדוחות ביצעו ה- big 4 (1) כ-40% שלא ידוע (1-), וכ-40% שביצעו מבקרים אחרים.
4. סוגי חברות- רוב החברות הן בע"מ ותאגיד. ישנם עוד מספר מועט מאוד של סוגי חברות אך הסקת המסקנות שלנו לגביהן תהיה בערבון מוגבל אם בכלל.
5. פשיטת רגל- 99.9% מהחברות לא פשטו רגל (רק 11 חברות פשטו רגל). כיוון שהחלוקה כאן כל כך לא מאוזנת החלטנו להוריד את 11 הדגימות הללו ולהתעלם מהפיצ'ר הזה בהמשך.
6. שינוי פנים מבני של החברה- אמנם גם כאן הדאטא כלל לא מאוזנת אבל יש כ-5% שביצעו שינוי מבנה. כיוון שזה יכול להעיד וללמד הרבה על החברה, החלטנו להשאיר את הפיצ'ר הזה.

לאחר מכן בדקנו פיצ'רים האחד ביחס לשני:

1. Year VS Audit opinion- במבט קצת יותר רחב, ניתן לראות שרוב חוות הדעת שאינן ידועות הן בשנים 08-09 ומהשנים האלה והלאה חסרות מעט מאוד חוות דעת, והן מתפלגות כפי שאנחנו מצפים לראות.
2. Big 4 VS Audit opinion- נתון מעניין שניתן לראות כאן הוא ההבדל ביחס בין חוות דעת "קיצוניות" – Disclaimer או Adverse - בין ארבעת החברות הגדולות לבין השאר. כמו כן, כל הדוחות שלא ידוע מי ביצע אותם, גם לא ידוע מה הייתה חוות הדעת של מבקר הדוח.
3. ארבעת הגדולים מול סוגי החברות- כפי שניתן לראות מהגרף, רוב החברות ש4 הגדולות מבקרות הן חברות בע"מ, ומעט תאגידים, באופן יחסי לאחוז הכללי של התאגידים בדאטא.

לסיום, רצינו לבדוק האם שינוי מבנה הייתה אולי פונק' של שנה קשה במשק ולכן בדקנו גם זאת אך בגרף לא קיבלנו איזושהיא חלוקה מיוחדת ולכן לא יכולנו להסיק את המסקנה הנ"ל.

לאחר שהתבוננו בדאטא, רצינו לחפש קורלציות שונות בין הפיצ'רים. על מנת לבצע את זה, הורדנו את כל הפיצ'רים שקשורים לזיהוית הורדנו את הפיצ'רים שקיימים באופן בדיד.

קיבלנו מטריצת קורולציה ענקית בגודל של 400 על 400 שמוצגת כנספח בקובץ נפרד.

לאחר ששמרנו את המטריצה ביצענו עליה חיפוש עבור קורלציות גבוהות במיוחד:

קיבלנו כי עבור 25 מאפיינים יש קורלציה של מעל 0.8 עבור יותר מ-10 מאפיינים אחרים. בנוסף ישנם מאפיינים שהקורלציה בינם לבין מאפיינים אחרים היה ממש אחד וגם אותם חיפשנו וגילינו שיש 10 כאלה. לבסוף בדקנו לאלו מאפיינים הקורלציה הכי גבוהה לאיזשהוא מאפיין אחר היא מקס' 0.1 כך שנוכל להגיד שהם עומדים בפני עצמם וקיבלנו 23 כאלה. את כל אלו שמרנו בקובץ שמופיע גם הוא כנספח נפרד.

ניתוח נתונים:

הורדת מימד:

על מנת לבצע ניתוח טוב ביצענו עוד שלושה שלבים מקדימים לפני כל אשכול:

1. חילקנו את הדאטה לפי דוחות הביקורת כך שכל ה-Unknown לא ישתתפו באשכול.
2. נרמלנו את הדאטה.
3. השתמשנו ב- one hot encoder עבור הפיצ'רים: audit opinion, legal form. בשלב בויזואליזציה, החזרנו אותם לצורה המקורית שלהם.

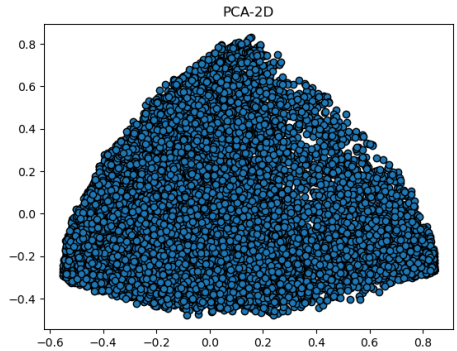
בתחילה ביצענו הורדת מימד בעזרת PCA על מנת לראות כמה מימדים כבר יתנו לנו תוצאה טובה של הצגת הדאטה. ביצענו בדיקה של הערכים העצמיים ובדקנו מתי אנחנו מקבלים שסכום הערכים העצמיים מהווה 90% מהדאטה:

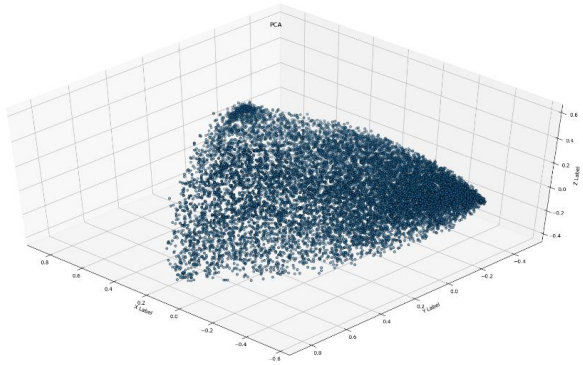
eighen value that represent 90% of the data = 19

and the ratio is 0.9019940148704854

אמנם זה הרבה מאוד אך כשבדקנו ע"י ה- explained variance ratio ראינו כי כבר שלושת הע"ע הראשונים מכילים קצת מעל 60% מהדאטה.

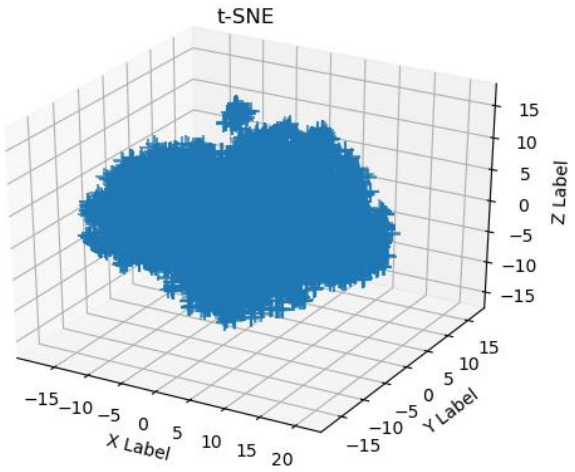
בנוסף, כאשר הוצאנו Plot של הדאטה ב-2D וב-3D ראינו שהתוצאה נראית טוב:





ממש ניתן לראות שהדאטה מרוכזת ומתקבלת תוצאה יפה מספיק כבר עבור המימדים הללו. כלומר הצגת הדאטה באשכולים ב-3 מימדים אכן תשקף יפה את התוצאה.

לסיום רצינו לבדוק גם איך נראית הורדת מימד בשיטת t-SNE:



גם כאן בסה"כ התוצאה יפה אך כיוון שהשיטה היא משמעותית יקרה יותר ולא ראינו שהיא מוסיפה הרבה מידע ביחס ל- PCA בהצגה של האשכולים החלטנו להשתמש ב- PCA.

אשכול :

אשכול K-means:

בתחילה, ביצענו K-means במס' מרכזים שונים על מנת לראות מהי החלוקה שתיתן את התוצאה הטובה:

קיבלנו כי חלוקה של 3 מרכזים ייתן לנו את ההתאמה הטובה ביותר עבור הדאטה שלנו. כמו כן אותה תוצאה יצאה לנו בדיוק עבור ה- ELBOW SCORE (ה"מרפק" מתקבל ב-3 קלאסטרים).

להלן החלוקה לפי האשכולים שקיבלנו:

Kmeans - cluster 0 = 3695

Kmeans - cluster 1 = 6143

Kmeans - cluster 2 = 3346

בגלל שהDATA שלנו בעל הרבה ממדים, לא יכולנו להראות אותו כמו שהוא בצורה מקורית, ולכן השתמשנו ב-PCA עבור ויזואליזציה.

מבחינה ויזואלית, ניתן לראות שהחלוקה יפה וברורה, כמו כן האשכולות שקיבלנו נותנים חלוקה די מאוזנת.

לאחר ביצוע האשכול רצינו לראות איך האשכולים מפלגים את הדאטה שלנו בפיצ'רים השונים, כמו שביצענו כאשר עשינו ויזואליזציה של הדאטה:

1. Audit opinion- כבר כאן ניתן לראות שמעל 50% מכמות הדוחות שהוגדרו כ- Unqualified נמצאות באשכול מס' 1. לעומתו שאר החלוקה די מאוזנת, כלומר אנחנו יכולים להניח מכאן שבדוחות ישנם פיצ'רים שמאפשרים לחזות כאשר הדוח הוא עם הסתייגות. זה אכן הגיוני שדווקא בסוג הזה של הדוחות תתקבל חלוקה משמעותית שכן סוג זה מהווה את רוב הדאטה.
2. Big 4- גם כאן, מעל 50% מדוחות ה- BIG 4 נמצאים באשכול 1 מה שמפתיע שכן ראינו בשלב הויזואליזציה שרוב ה- Adverse,Disclaimer הם תחת ה- Big 4 והם נמצאים בעיקר באשכולים 0 ו-2. כלומר באשכול 1 רוב ה-Big 4 למעשה מאשכלים את הדוח כ- Unqualified/quakified.
3. שינוי פנים מבני של החברה- כאן כבר קיבלנו אפילו עוד יותר מובהקות של חלוקה, כאשר הרוב המוחלט של שינויי המבנה של החברה נמצאים באשכול מס' 2. זה גם האשכול שבו חוות הדעת היא שלילית באופן מובהק היא בכמות הגדולה ביותר. כלומר, אכן חוות הדעת שלילית של רואה החשבון יכול להעיד על כך שבחברה כרגע בנויה לא נכון ויכול להיות שהיא תצטרך לבצע איזה שינוי מבני בפנים החברה.
4. לבסוף ביצענו גם בדיקת חלוקה של שנים וסוגי חברות לפי הקלאסטרים אך לא קיבלנו תוצאה מעניינת.

לסיום האשכול, ביצענו חישוב BIC:

וקיבלנו:

The BIC score of k_means with 3 centroids is: 12583148.482219802

אמנם התוצאה גבוהה אך החלורה של האשכול שהתקבלה עדיין נראית טוב ואנו מניחים שזה קשור לרעשים שיש בדאטה שהאלגוריתם לא יודע לסנן.

אשכול Agglomerative Clustering:

בשלב הבא ניסינו לבצע אשכול היררכי:

גם כאן, בחנו מהו מספר האשכולות האופטימלי באותו אופן כמו שביצענו באשכול הקודם :

גם כאן קיבלנו שחלוקה של 3 אשכולות היא החלוקה המיטבית שלנו.

בדיוק כמו באשכול הקודם גם כאן ביצענו את החלוקה לפי 3 אשכולות, השתמשנו ב- PCA והצגנו גרף 3D של החלוקה. בנוסף הדפסנו את כמות האובייקטים שיש בכל אשכול:

גם כאן החלוקה מבחינה ויזואלית נראית יחסית טובה עם אזורים די ברורים.

מס' האובייקטים בכל אשכול:

Agglomerative - cluster 0 = 7241

Agglomerative - cluster 1 = 3215

Agglomerative - cluster 2 = 2728

לאחר מכן, גם בחלוקה בזאת רצינו לראות איך הפיצ'רים מתפלגים בתוך האשכולות:

1. Audit opinion- גם כאן בדומה לאשכול הקודם, כמות הדוחות המירבית שהוגדרו כ- Unqualified נמצאות באשכול ספציפי. שאר הדעות של המבקר הן בחלוקה מאוזנת.
2. שינוי פנים מבני של החברה- גם כאן בדומה ל- K-means רוב השינוי המבני קיים באשכול אחד אבל עם חלוקה מעט פחות מובהקת.
3. Big 4- פה בשונה מ-K-means החלוקה של ה- Big 4 מול דעת המבקר היא לא הפוכה אבל גם כן מפתיעה ביחסים שלהם. ולבסוף בשאר הפיצ'רים שוב כמו באשכול הקודם לא קיבלנו איזושהיא מובהקות.
4. לבסוף ביצענו גם בדיקת חלוקה של שנים וסוגי חברות לפי הקלאסטרים אך כמו באלגוריתם הקודם, לא קיבלנו תוצאה מעניינת.

לסיכום, החלוקה כאן הייתה מאוד דומה לחלוקה ב-K-means עם מספרים טיפה שונים.

לסיום, ביצענו דנדוגרם גרף.

ממנו ניתן לראות שחלוקה ל-2 לדוג' הייתה נותנת לנו תוצאה מאד לא מאוזנת. מאידך חלוקה ל-4 כנראה הייתה מחלקת את אשכול 0 ל-2 חלקים שככל הנראה נורא דומים ע"פ ה- Silhouette score.

אשכול GMM:

לאשכול אחרון, רצינו לבדוק האם ניתן לקבל אשכול ללא הנחות על האשכולים שייתן לנו תוצאות טובות יותר מ-K-means. גם כאן בדקנו עבור מס' אשכולים שונים וביצענו את ה- Silhouette score.

כאן כבר קיבלנו שכמות האשכולות הכי גבוה שמתאים לנו הוא 2. בנוסף, הערך בגרף הוא מאוד נמוך גם עבור $number\ of\ cluster = 2$. בנוסף מבחינה ויזואלית החלוקה אצלנו לא חדה מספיק והאשכולים מעורבבים בתוך עצמם. גם ה- BIC המתקבל הוא:

The BIC score of GMM with 3 centroids is: -50515136.91084285

אנחנו מניחים שהסיבה היא ש-GMM מלכתחילה לוקח בחשבון את כל מטריצת ה-Covariance וכנראה שהמטריצה רועשת מידי.

לבסוף גם הגרפים עצמם שהתקבלו בין הפיצ'רים לאשכולים לא נתן לנו מידע מיוחד.

סיכום:

לסיכום ניתן לומר כי הדאטה שלנו מתחלקת די יפה ומאפשרת אשכול טוב הן היררכי והן K-means. קיבלנו שישנם קשרים גבוהים בין סוג הדוח לבין המאפיינים האחרים וניתן לומר ע"פ האשכולים שניתן לחזות במידה מסוימת את סוג הדוח. כמו כן ראינו כי שינוי מבני של החברה גם הוא מאושכל לאשכול ספציפי כמעט כולו יחד מה שמלמד על כך שמאפיינים דומים קיימים גם אצלו.