בועז גורביץ' - 325813970, אראל דקל - 326064888

עבודה בכריית מידע וייצוג מידע

בעבודה זו נדרשנו לחקור מערך נתונים המחזיק מידע אודות בתי עסקים שונים, פרמטרים שונים המגדירים את אותו בית עסק והאם הוא נסגר או לא. בחלק הראשון נשאף לקחת את את 41 העמודות שנמצאות במערך הנתונים הנ"ל ולצמצם אותן ככל שנוכל על ידי מחיקת עמודות לא רלוונטיות או עמודות שחוזרות על עצמן. בנוסף לכך, נקח שורות שבהן ערכים ריקים או אינם הגיוניים ונבצע מניפולציות שונות עליהם. מטרת חלק זה היא לקחת הרבה מידע שקיבלנו במערך הנתונים, לנקות אותו מערכים חריגים וריקים ולדאוג לכך שמודל למידת המכונה אשר ניצור בחלק השני של הפרויקט לא יעבוד קשה ממה שהוא צריך לעבוד.

תחילה, נעלה מספר ספריות לסביבת העבודה שלנו על מנת שנוכל לעבוד עם מערך הנתונים בצורה נוחה:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
```

נשתמש ב-padnas על מנת לערוך את הטבלה שמחזיקה את הנתונים שלנו בצורה נוחה. נשתמש ב-numpy על מנת לבצע פונקציות מתמטיות בצורה נוחה יותר ומהירה יותר. נשתמש ב-seaborn ו-matplotlib על מנת ליצור גרפים ולהציג את הנתונים בצורה ויזואלית. ונשתמש ב-sklearn בחלק זה כדי לבצע את אלגוריתם PCA שבמילים פשוטות מוריד את מימדי הנתונים.

הצגת המידע

נקח את המידע שהעלו ונציג את העמודות שלו. עבור כל עמודה יוצגו שם העמודה, מספר הערכים שאינם ריקים ומה סוג האובייקט שאמור להיות בכל עמודה (int64, object וכדומה). נוכל לראות כי יש לנו 41 עמודות ונשים לב כי ישנן עמודות שלא כל הערכים בהן מלאים. לאחר מכן נציג את 5 השורות הראשונות של הטבלה כדי לקבל הבנה של צורת ההצגה של המידע בטבלה, כלומר, אילו ערכים מצופים להיות מוצגים בעמודות שונות.

הצגת סטטיסטיקות

נציג את טבלת ה-describe שבה נוכל למצוא את המדדים הסטטיסטים הכי בסיסיים עבור עמודות נומריות (ולא קטגוריאליות). בטבלה זו נוכל למצוא את מספר הפעמים שמופיע ערך לא ריק, את הממוצע, סטיית תקן, ערכים מינימלים ומקסימליים ואחוזונים שונים של המידע. כך נוכל לקבל מידע אודות ההתפלגויות השונות של כל אחת מהעמודות ובעזרת זה נוכל לקבוע אילו ערכים הם חריגים ולא רלוונטים עבורנו.

כעת שוב, נבחר בעמודות הנומריות ונציג את הצידוד הסטטיסטי של כל אחת מהן. צידוד הוא מדד המציג חוסר הסימטריה של פונקציית צפיפות או התפלגות של משתנה מקרי ממשי. למשתנה בעל צידוד חיובי יש נטייה לקבל ערכים גבוהים ורחוקים מן הממוצע, יותר מאשר ערכים נמוכים הרחוקים ממנו, ובגרף של פונקציית הצפיפות ניכר שהזנב הימני ארוך ועבה יותר מן השמאלי. כאשר הצידוד שלילי, הצדדים מתהפכים. בפונקציה סימטרית ביחס לתוחלת, הצידוד הוא אפס. מיד נוכל לשים לב כי שבעמודת ה-total_funding יש צידוד של כ-26.1 מה שמעיד לנו כי הרהב עסקים גייסו מעט מאוד כספים ביחס לממוצע או במילים אחרות יש מעט חברות שהשקיעו סכומי עתק בעסקים שלהם ובכך השפיעו רבות על הממוצע וגרמו לערכים לא הגיוניים (או לא סבירים) במערך הנתונים שלנו. בהמשך נטפל בבעיה זו.

מחיקת עמודות

בחלק זה נתעסק במחיקת עמודות שאין רלוונטיות או שחוזרות על עצמן. למדנו בכיתה כי הפקטור המכריע בזמן ריצת אלגוריתם של למידת מכונה הוא מספר העמודות ולכן כל עמודה שנוריד תשפיע רבות על זמן הריצה של אלגוריתמים שנריץ בחלק השני של הפרויקט.

קודם כל נשים לב שיש לנו עמודה ששמה "category" וישנן עמודות נוספות שעונות לשאלה "is_typeOfCategory" ברור לנו כי אין צורך בחזרתיות הזו ומספיקה רק עמודות הקטגוריות שמחזיקה את כל הקטגוריות של העמודות החזרתיות ועוד מספר קטגוריות (שבהמשך נצמצם למספר קטן יותר של קטגוריות).

באופן דומה, יש לנו עמודת "state_code" שמכילה את הקוד של המדינה שבה העסק נמצא. נשים לב כי ישנן עמודות נוספות השאלות את השאלה "is_statecode" ולכן נוכל למחוק גם אותן מאותה הסיבה.

נמחק את עמודות "first_funding_date" ו-"last_funding_date" מכיוון שיש לנו את ערך ה-"foundation_date" ויש לנו את עמודות "first_funding_age" ו-"last_funding_age" ומכיוון שנוח לנו יותר לעבוד עם age ולא נאבד מידע כתוצאה מכך.

נוריד את עמודות ה-"name" ו-"id" כיוון שאינן תורמות לנו מידע נוסף על בתי העסקים.

ה-"id" מתואר באופן אוטומטי על ידי ה-pandas ובהמשך כאשר נרצה להפוך את עמודת ה-"name" לערכים נומרים נקבל כ-700 קטגוריות שונות עם התפלגות אחידה, כלומר כל ערך של שם מופיע רק פעם אחת ואינו תורם כלל להבנת ההצלחה של עסק.

לאחר מעט מחשבה, החלטנו גם להוריד את עמודות ה-"state_code", "city", "zip_code" מהסיבה הפשוטה שיש לנו את עמודות ה-"latitude", "longitude" והן מראות לנו את המיקום המדויק של כל עסק בשתי עמודות במקום בעזרת 3 עמודות. באופן כזה נוכל להציג גרפים ונוכל לראות התקבצויות שונות של עסקים והאם ישנם קשרים בין מיקום גיאוגרפי של עסק לבין הצלחתו.

ויזואליזציה של המידע

נכתוב קוד לבניית טבלת גרפים 4 על 4, כאשר בכל גרף נציג את ההתפלגות של עמודה נומרית אחרת. מתוך הגרפים הללו נוכל להסיק על הרבה ערכים חריגים שלא נהייה מעוניינים בהם ונוכל לראות הרבה קורלציות וגרפים דומים בהן עמודות שונות.

ניתן לראות בגרף של העמודה total_funding שיש ערכים חריגים מאוד ולכן הגרף נראה כמו פונקצית דלתא. אם נמחק את הערכים החריגים נוכל לראות התפלגות יותר רחבה ולהסיק מסקנות.

בגרפי עמודות הגלאים ניתן לשם לב כי יש לנו ערכים שליליים וערכים חריגים שגדולים מ-20. ולכן בפרקי ההמשך נדאג למחוק את הערכים הנ"ל כדי לוודא שערכים חריגים לא משפיעים לנו על התוצאות הסופיות.

בגרף connections נראה כי רוב הערכים קטנים מ-50 ואז יש גבעה קטנה סביב ה-60, ולכן נחליט למחוק אותה שכן היא יכולה לפגוע בממוצע ובמודל הסופי.

בגרף funding_rounds נראה שרוב הערכים קטנים מ-8 ואז יש גבעה קטנה ב-10, ולכן נמחק אותה מכיוון שנעדיף להוריד ערכים שיגרמו לסטייה גדולה.

בגרף milestones נמחק את הערכים שגדולים מ-5 כיוון שהם כמה בודדים שכן רוב הערכים קטנים מ-5 ולכן נעדיף להסתכל רק עליהם בלבד.

בגרף avg_group_size נשים לב כי יש כמה ערכים בודדים שגדולים מ-10 ולכן נחליט למחוק אותם כי הם יכולים לגרום לסטייה גדולה ואיננו מעוניינים בהם.

הפחתת נתונים ותיקון נתונים

לאור מה שראינו בחלק הויזואליזציה של המידע, נוריד בחלק זה את כל הערכים הלא רצויים שמצאנו בגרפי ההתפלגות של העמודות, תחילה נוריד את הערכים החריגים של עמודות: connections, avg_group_size, total_funding ו milestons.

כעת נעבור לעמודות הגילאים בהם שמנו לב שכאשר בדקנו את המידע על כל העמודות, העמודות first_milestone_age ו last_milestone_age היו עם ערכים חסרים, ולכן מילאנו את הערכים החסרים ע"י התפלגות נורמלית של כל עמודה והכרחנו אותם להיות בין 0 ל-20. בנוסף, נמחק את כל ערכי הגילאים אשר קטנים מ-0 וגדולים מ-20, כיוון שראינו בגרפים שהם ערכים חריגים.

כעת, נרצה להסתכל על בעיה שיש סטארטאפים שלא ביצעו שום גיוס כספים מסוג A,B,C או D, ולמרות שראינו כי כולם ביצעו לפחות סבב גיוס כספים אחד, ולכן החלטנו שמדובר בבעיה כי הם היו חייבים לבצע סבב כלשהו, ולכן החלטנו להכניס כאילו שהם ביצעו את סבב A מכיוון שלאור מה שקראנו הוא מהווה את הכי פחות משמעות ואנחנו מאמינים שזה לא יגרום לסטייה רבה מהנתונים מקוריים. בנוסף לכך, נמחק את הערכים החריגים בעמודת funding_rounds שראינו בגרף ההתפלגות של העמודה.

נציג גרף scatterplot אשר יתאר לנו את מיקומי הסטארטאפים על פי מיקום ויסמן בירוק אם הצליחו ובאדום את נסגרו, שמנו לב כי יש לנו ערך חריג בודד אשר לא נמצא בסביבת שאר הערכים ולכן החלטנו להוריד אותו ולצייר את המפה מחדש ובמפה אף שמנו לב שהנקודות מזכירות את המפה של ארצות הברית כמצופה.

כעת, שמנו לב שיש הרבה ערכים בשדה הקטגוריה של הסטארטאפ ולכן החלטנו לשנות אותו ככה שיכיל רק את 8 הערכים הכי שכיחים ואת השאר יחליף לערך "other".

קורלציה בין עמודות

בחלק זה נבנה heatmap אשר תייצג לנו קורלציות בין כל העמודות, נצבע אותה בפורמט red&gray בשביל שנוכל לראות בבירור את כל הזוגות שיש להן קורלציה חזקה. ראינו הרבה קורלציות חזקות אבל רצינו להתייחס במיוחד לקורלציה שיש בין כל הגילאים, לקורלציה בין funding_rounds ו total_funding, ולקורלציה בין milestones ו connections.

ראשית, הראנו גרף שקישר בין ההצלחות למספר הקשרים (connections) וראינו כי אכן ככל שיש יותר קשרים יש יותר סיכוי להצלחה.

נציג גרפים של histplot על מספר הקשרים כתלות במספר ה milestones. נשים לב שככל שהעסק עבר יותר milestones כך באופן טבעי ניתן לראות כי באופן ממוצע ישנם יותר קשרים.

כעת נעשה boxplot של סכום הכסף שנאסף כתלות במספר סבבי גיוס הכספים. נשים לב גם כאן כי ככל שהיו יותר סבבי גיוס כך גם הכספים שנאספו במהלך כל הסבבים יותר גדול.

כעת נציג מספר גרפי התפלגות של ארבעת הגילאים first_funding, last_funding, first_milestone, last_milestone, עם הפרדה בין הצלחה וסגירה של עסק. נשים לב כי התפלגות בגרפים שעשינו אכן מאוד דומה, וזה משקף את הקורלציה בין העמודות.

המרה ונרמול המידע והפעלת PCA

ראשית, נצטרך להמיר את העמודות הלא נומריות לערכים נומרים. על עמודת הTarget ביצעו המרה לערכים של 0 או 1 שכן היא עמודה בינארית. על עמודת הfoundation_date ביצענו המרה למספר שלם ע"י pd.to_datetime לאחר מכן עשינו casting ל-int. אנחנו יכולים לעשות פעולה זו מכיוון שבמחשבים מוגדר תאריך התחלה אוניברסלי שממנו אנחנו יכולים לחשב ערך יחסי מסוים כדי לכמת את התאריכים שנמצאים בעמודה זו. את עמודת הcategory שהכילה בתוך ערכיה 9 מחרוזות שונות החלפנו את זה כך שתכיל מספרים מ-0 ל-8 במקום.

לאחר מכן, ביצענו נרמול min-max בין 0 ל-1 לכל עמודות הטבלה, מכיוון שהיה מימוש נוח, והקל על השלב הבא של ה PCA. את הנירמול עשינו על מנת לשים את כל הערכים על אותה סקאלה.

ביצענו את אלגוריתם ה PCA עם רגישות של 0.99, ויצרנו גרף שתיאר את הPCA וצבענו אותו בצבעים שהיו נוחים לראות את ההבדל, שמנו לב כי הגרף היה ממש נחמד ויכולנו לראות את clustering.