

נושאים מתקדמים בلمידת מכונה

דו"ח מסכם

היזוי תוצאות ביקורת תברואה במוסדות מזון

שמות הסטודנטיות:

מרין בן חמו - 209108604

עדי גリンפלד - 322875949

מבוא

ביקורת תברואה במוסדות מזון מהוות כלי מרכזי בשימירה על בריאות הציבור, שכן ליקויים תברואתיים עלולים לגרום לסיכונים בריאותיים ממשמעותיים. רשות הפיקוח מבצעות אלף ביקורות מדי שנה, אולם משאבי הפיקוח מוגבלים, ולכן קיימת חשיבות לכolumbia להזות מראש מוסדות בעלי סיכון גבוה לכישלון בבדיקה. יכולת היזוי מוקדמת עשויה לסייע בתיעוד בבדיקות ובהקאה יعلاה יותר של משאבי הפיקוח.

בפרויקט זה נבחנת היכולת להזות את תוצאות ביקורת התברואה במוסדות מזון בעיר שיקגו באמצעות שיטות של למידת מכונה. על בסיס נתונים היסטוריים של ביקורות, נבנה תהליך סיוג שמטרתו לנבא האם מוסד מזון צפוי לעבור או להכשל בבדיקה תברואה. בנוסף, מבוצע ניתוח למידה לא-דומינית לצורך זיהוי דפוסים, אשכולות וחיריגים נתונים, אשר עשויים להעיד על רמות סיכון שונות.

מטרות הפרויקט הן:

1. יישום תהליך מלא של למידת מכונה, החל מעיבוד מקדים של הנתונים, דרך בנייתמודלים והערכתם.
2. בוחנת ביצועי מודלים שונים להיזוי כישלון בבדיקה, תוך דגש על זיהוי מוסדות בסיכון.
3. הפקת תובנות בעלות משמעות עבור גורמי פיקוח ורגולציה.

הדעתה והמאפיינים

תיאור הדעתה

בפרויקט זה נעשה שימוש בדעתה פומבי של ביקורות בטיחות מזון בעיר שיקגו, מתוך מאגר Food Inspections data.gov. הדעתה כוללת בערך 300,554 ביקורות שבוצעו במוסדות מזון לאורך מספר שנים, וכל ביקורת מתוארת באמצעות 17 מאפיינים. כל שורה מייצגת ביקורת אחת לעסק מסוים בתאריך נתון, כולל מידע על מקום העסק, סוג, סוג הביקורת ותוצאתה.

מטרת העבודה היא ניבוי כישלון בבדיקה, אשר הוגדר כמשתנה ביןארי (*target_fail*), כאשר ערך 1 מייצג ביקורת שנכשלה וערך 0 מייצג ביקורת שעברה.

בחירת מאפיינים

בשלב זה נבחרו מאפיינים בעלי פוטנציאל תחזיתי, הכוללים מידע גיאוגרפי (Latitude, Longitude), מאפיינים תפעוליים כגון סוג העסק, רמת הסיכון וסוג הביקורת, וכן מאפייני זמן שהופקו מתאריך הביקורת.

עמודות שאינן תורמות להיזוי, או כאלה שאינן זמינים בזמן אמיתי, הוסרו על מנת להימנע מדיליפה מידע ולשמור על מודל פשוט ובבעל יכולת הכללה טובה.

שימוש מקדים של הנתונים

בוצע שימוש מקדים של הנתונים شامل טיפול בערכים חסרים, באמצעות השלמת הערך השכיח במאפיינים קטגוריאליים והחציוון במאפיינים מספריים.

בנוסף, עמודת תאריך הביקורת פוצלה למאפייני זמן, והמאפיינים הקטגוריאליים הומרו לייצוג מספרי באמצעות One-Hot Encoding. בעקבות ריבוי המאפיינים שנוצר, בוצע אוסף מאפיינים באמצעות הסרת מאפיינים נדירים, במטרה להפחית ממדיות ולשפר את ביצועי המודלים.

שיטות ומודלים

בפרויקט זה יושמה למידת מכונה לצורך פתרון בעית סיווג ביןארית, שמטרתה חיזוי כישלון בבדיקה תברואה (Fail) לעומת בקורת (Pass). לאחר שפספוס מקרי כשל עלול להיות בעל השלכות משמעותיות, הושם דגש על זהות נכון של מחלוקת ה-Fail.

הגדרת הבעיה ואסטרטגיית הערכה

הבעיה הוגדרה כבעית סיווג ביןארית. בשל חוסר איזון יחסי בין המחלקות, נבחרו מדדי הערכה המתמקדים באיכות הזיהוי של מחלוקת ה-Fail, ובראשם Recall ו-F1-score, לצד ROC-AUC למדידת יכולת ההבנה הכללית של המודלים. ממד Accuracy שימש כמדד משלים בלבד.

הנדסת מאפיינים

בנוסף למאפיינים הגולמיים בדעתה, בוצעה הנדסת מאפיינים במטרה לשפר טוב יותר את רמת הסיכון של כל ביקורת. מתוך תיאור הഫרות הופקו מאפיינים כמו מיצגים מתאריך מספר ההפרות וחומרתן, וכן ממד חומרה מצטבר. כמו כן הופקו מאפייני זמן מתחום הביקורת ואינדיקציות להקשר הביקורת, כגון ביקורת בעקבות תלונה. מאפיינים אלו נבחרו מתוך הנחה שהם משפיעים על תוצאה הבדיקה.

בחירת המודלים

לצורך ניתוח הבעיה נבחרו מספר מודלים בעלי רמות מורכבות שונות:

- Logistic Regression כמודל בסיס לינארי ופרשני
- Decision Tree כמודל לא-לינארי הלומד חוקים אינטראקטיביים
- Gradient Boosting ו-Ensemble Random Forest מתקדמים
- Neural Network לצורך הערכת יכולת ייצוג לא-לינארית בדעתה טבלאי

בחירה זו מאפשרת השוואת היעילות בין גישות שונות למידה מפוקחת

קביעת סף החלטה (Threshold)

עבור מודלים Gradient Boosting ו- Logistic Regression, Random Forest והוגדר סף החלטה של 0.35 במטרה לשפר את זיהוי מקרי ה-Fail. לעומת זאת, מודלים Neural Network ו- Decision Tree הופעלו באמצעות predict() עם הגדרות ברירות מחדל, וכך לא הוחל בהם סף מותאם.

ניסויים ותוצאות

הניסויים בוצעו תחת תנאים אחידים לכל המודלים, תוך חלוקת הנתונים לsett אימון וsett random_state במטרה לשמרם על יחס המחלקות, והגדרת קבוע לצורך שחרור התוצאות.

סף ההחלטה של 0.35 הוחל רק על מודלים המחזירים הסתברויות. סף זה נבחר במטרה לשפר את זיהוי מקרי ה-FAIL, והוא משפיע ישירות על האיזון בין Recall ל-Precision, וכך גם הוא פקטור מרכזי בהערכת הביצועים במודלים אלו.

מדדדי הערכה ומשמעותם

הערכת המודלים בוצעה באמצעות המՃדים Accuracy, Precision, Recall, F1-score ו-ROC-AUC. לאחר שמטרת הפרויקט היא זיהוי מקרי כשל, הושם דגש על Recall ו-F1 עבור מחלוקת ה-Fail. מודד את שיעור מקרי ה-FAIL שהמודול מצליח לזהות, בעוד F1-score מספק איזון בין זיהוי כשלים לבין צמצום אזעקות שווה.

מדד ROC-AUC שימש להערכת יכולת ההבחנה הכללית בין המחלקות ללא תלות בסף ההחלטה.

ניתוח תוצאות, מגבלות ותוצאות בלתי צפויות

מהשוואת תוצאות המודלים עולה כי המודלים מצליחים לזהות חלק משמעותי מהריגול, עם ערך Recall הנעים בטוחה של כ-0.64–0.80. עם זאת, זהה מלא של כלל מקרי הכשל אינו מושג, דבר המעיד על מגבלות הדאטה, רעש ומידע חסר שאינם ניתנים לכיצדה מלאה באמצעות המאפיינים הקיימים. בנוסף, ניכרת רגישות לבחירת סף ההחלטה במודלים שבהם בוצע כיווננו, אשר השפעה על האיזון בין Precision ל- Recall .

השוואת ביצועי אלגוריתמים

מודלים שהוצגו כבעלי הביצועים הטובים ביותר בזיהוי מקרי כשל היו Random Forest ו-Gradient Boosting, שכן מדובר במודלים מסוג Ensemble המבוססים על שילוב של מספר לומדים, המאפשר לכידה יעליה של דפוסים לא-לינאריים בדאטה טבלאי. מודלים פשוטים יותר, כגון Logistic Regression, הציגו ביצועים יציבים אך מוגבלים. מודל Neural Network הציג יכולת למידה טובה, אך לא השיג יתרון על פני מודלי הד-ENSEMBLE במסגרת הניסוי שנערך.

	Model	Eval Mode	Threshold	Accuracy	ROC-AUC	Precision (Fail)	Recall (Fail)	F1 (Fail)
0	Random Forest	predict_proba() + threshold	0.35	0.88	0.93	0.76	0.80	0.78
1	Gradient Boosting	predict_proba() + threshold	0.35	0.87	0.92	0.77	0.77	0.77
2	Decision Tree	predict()	default	0.88	0.92	0.84	0.70	0.77
3	Neural Network	predict()	default	0.87	0.91	0.86	0.64	0.73
4	Logistic Regression	predict_proba() + threshold	0.35	0.87	0.90	0.78	0.72	0.75

ניתוח למידה לא-empokhet

לצורך בחינת מבנה הדאטה וזיהוי דפוסים שאינם תלויים בתוויות היעד, בוצע ניתוח למידה לא-empokhet על מדגם של 20,000 תצפיות, לאחר ביצוע תקנון באמצעות StandardScaler. לצורך וייזואליזציה והפחיתת מדדיות יושם PCA, כאשר שני הרכיבים הראשיים הראשונים מסבירים כ-40% מהשונות הכוללת בתווונים.

אלגוריתם KMeans יושם על היצוג המופחת, כאשר מספר האשכולות נבחר על בסיס ניתוח Elbow והצבע על ארבעה אשכולות מרכזיים. לצורך זה היו חריגים ומבני ציפויו יושם אלגוריתם DBSCAN, לאחר בחירת הפרמטרים באמצעות - k-distance plot, אשר זיהה תשעה אשכולות וכן כ- 4.7% ציפוי חריגות.

מצאים אלו מצביעים על קיומם של דפוסי התנהגות שונים במוסדות המזון, לצד קבוצת ציפוי חריגה שאינה משתלבת בדפוסים המרכזיים, ועשוייה לייצג מקרי סיכון יהודים. ניתוח זה מדגיש את התרומה המשלימה של למידה לא-מפוקחת להבנת הדאטה מעבר להיזוי המפוקח.

סיכום ודיון

בפרויקט זה יושם תהליך מלא של למידת מכונה לנתחה נתוני ביקורת תברואה במוסדות מזון בעיר שיקגו. העבודה שילבה למידה מפוקחת לצורך חיזוי מקרי כשל בבדיקה, יחד עם למידה לא-מפוקחת שנועדה להבין את מבנה הנתונים ולזהות דפוסים וחריגים. שילוב זה מאפשר הסתכלות משלימה על הנתונים הן מנקודת מבט חיזوية והן מנקודת מבט מבנית.

תוצאות הלמידה המפוקחת הראו כי מודלים מסוג Ensemble הציגו ביצועים טובים יותר בזיהוי מקרי כשל, בעוד שמודלים פשוטים יותר סיפקו יכולת פרשנות אך היו מוגבלים ביכולתם לכוד דפוסים מורכבים. במקביל, הניתוח הא-מפוקח הדגיש קיומם של אשכולות מוחנים לצד ציפוי חריגות, אשר עשויה להעיד על מקרים חריגים בנתונים, שאינם מזוהים במלואם באמצעות מודלי סיווג בלבד.

חלוקת תפקידים

העבודה בוצעה בשיתוף פעולה מלא בין חברות הוצאות, כאשר שלבי הפרויקט בוצעו במפגשים משוחפים. שלבי העבודה כללו עיבוד וניתוח הנתונים, הנדסה מאפיינית, בניית מודלים מפוקחים, ביצוע ניתוח למידה לא-מפוקחת, ניתוח התוצאות, הכנת המיצגת, וכתיבת הדוח הסופי. כל חברות הוצאות נטו חלק פעיל בכל שלבי הפרויקט, והאחריות התחלקה באופן שווה לאורך העבודה.

כיוונים עתידיים

כיוונים אפשריים להמשך כוללים הרחבת הדאטה באמצעות מקורות מידע נוספים, כיוונו מתקדם של פרמטרים במודלים המפוקחים, ושלוב כלים נוספים לניתוח והבנה של הנתונים. בנוסף, ניתן להעמיק את השימוש בלמידה לא-מפוקחת לצורך זיהוי דפוסים נוספים וחיריגים בנתונים.

[קישור ל-GitHub-של הפרויקט](#)

הערה לגבי מקור הנתונים

מאחר שקובץ ה-CSV המקורי גדול מאוד, לא ניתן היה להעלות אותו ל-GitHub, גם לא בפורמט דחוס (ZIP). לכן, הנתונים בפרויקט נטענים ישירות לאתר data.gov באמצעות קריאה דינמית בזמן הרצת הקוד. מכיוון שהמادر מתעדכן באופן שוטף, יתכנו הבדלים קלים במספר הרשומות או בהתפלגות הנתונים בין הרצות שונות של המחברת. עם זאת, שינויים אלו אינם צפויים להשפיע באופן מהותי על המסקנות המרכזיות של הפרויקט.