

נושאים מתקדמים בלמידת מכונה

דוח מסכם

חיזוי תוצאות ביקורות תברואה במוסדות מזון

שמות הסטודנטיות:

מרין בן חמו -209108604

עדי גרינפלד-322875949

מבוא

ביקורות תברואה במוסדות מזון מהוות כלי מרכזי בשמירה על בריאות הציבור, שכן ליקויים תברואתיים עלולים לגרום לסיכונים בריאותיים משמעותיים. רשויות הפיקוח מבצעות אלפי ביקורות מדי שנה, אולם משאבי הפיקוח מוגבלים, ולכן קיימת חשיבות ליכולת לזהות מראש מוסדות בעלי סיכון גבוה לכישלון בביקורת. יכולת חיזוי מוקדמת עשויה לסייע בתיעדוף ביקורות ובהקצאה יעילה יותר של משאבי הפיקוח.

בפרויקט זה נבחנת היכולת לחזות את תוצאות ביקורות התברואה במוסדות מזון בעיר שיקגו באמצעות שיטות של למידת מכונה. על בסיס נתונים היסטוריים של ביקורות, נבנה תהליך סיווג שמטרתו לנבא האם מוסד מזון צפוי לעבור או להיכשל בביקורת תברואה. בנוסף, מבוצע ניתוח למידה לא-מונחית לצורך זיהוי דפוסים, אשכולות וחריגים בנתונים, אשר עשויים להעיד על רמות סיכון שונות.

מטרות הפרויקט הן:

1. יישום תהליך מלא של למידת מכונה, החל מעיבוד מקדים של הנתונים, דרך בניית מודלים והערכתם.
2. בחינת ביצועי מודלים שונים לחיזוי כישלון בביקורת, תוך דגש על זיהוי מוסדות בסיכון.
3. הפקת תובנות בעלות משמעות מעשית עבור גורמי פיקוח ורגולציה.

הדאטה והמאפיינים

תיאור הדאטה

בפרויקט זה נעשה שימוש בדאטה פומבי של ביקורות בטיחות מזון בעיר שיקגו, מתוך מאגר Food Inspections באתר data.gov. הדאטה כולל 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, וכוללת מידע על מיקום העסק, סוג, סוג הביקורת ותוצאתה.

מטרת העבודה היא ניבוי כישלון בביקורת, אשר הוגדר כמשתנה בינארי (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 לחיזוי, או כאלה שאינן זמינות בזמן אמת, הוסרו על מנת להימנע מדליפת מידע ולשמור על מודל פשוט ובעל יכולת הכללה טובה.

עיבוד מקדים של הנתונים

בוצע עיבוד מקדים של הנתונים שכלל טיפול בערכים חסרים, באמצעות השלמת הערך השכיח במאפיינים קטגוריאליים והחציון במאפיינים מספריים.

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

שיטות ומודלים

בפרויקט זה יושמה למידת מכונה לצורך פתרון בעיית סיווג בינארית, שמטרתה חיזוי כישלון בביקורת תברואה (Fail) לעומת מעבר ביקורת (Pass). מאחר שפספוס מקרי כשל עלול להיות בעל השלכות משמעותיות, הושם דגש על זיהוי נכון של מחלקת ה-Fail.

הגדרת הבעיה ואסטרטגיית הערכה

הבעיה הוגדרה כבעיית סיווג בינארית. בשל חוסר איזון יחסי בין המחלקות, נבחרו מדדי הערכה המתמקדים באיכות הזיהוי של מחלקת ה-Fail, ובראשם Recall ו-F1-score, לצד ROC-AUC למדידת יכולת ההבחנה הכללית של המודלים. מדד Accuracy שימש כמדד משלים בלבד.

הנדסת מאפיינים

בנוסף למאפיינים הגולמיים בדאטה, בוצעה הנדסת מאפיינים במטרה לשקף טוב יותר את רמת הסיכון של כל ביקורת. מתוך תיאור ההפרות הופקו מאפיינים כמותיים המייצגים את מספר ההפרות וחומרתן, וכן מדד חומרה מצטבר. כמו כן הופקו מאפייני זמן מתאריך הביקורת ואינדיקציות להקשר הביקורת, כגון ביקורת בעקבות תלונה. מאפיינים אלו נבחרו מתוך הנחה שהם משפיעים על תוצאת הביקורת.

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 ואינטראקציות
- Random Forest ו-Gradient Boosting כמודלי Ensemble מתקדמים
- Neural Network לצורך הערכת יכולת ייצוג לא-לינארית בדאטה טבלאי

בחירה זו מאפשרת השוואה בין גישות שונות ללמידה מפקחת

קביעת סף החלטה (Threshold)

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

ניסויים ותוצאות

הניסויים בוצעו תחת תנאים אחידים לכל המודלים, תוך חלוקת הנתונים לסט אימון וסט בדיקה באמצעות Stratified Split לשמירה על יחס המחלקות, והגדרת random_state קבוע לצורך שחזור התוצאות.

סף החלטה של 0.35 הוחל רק על מודלים המחזירים הסתברויות. סף זה נבחר במטרה לשפר את זיהוי מקרי הכשל, והוא משפיע ישירות על האיזון בין Recall ל-Precision, ולכן מהווה פרמטר מרכזי בהערכת הביצועים במודלים אלו.

מדדי הערכה ומשמעותם

הערכת המודלים בוצעה באמצעות המדדים Accuracy, Precision, Recall, F1-score ו-ROC-AUC. מאחר שמטרת הפרויקט היא זיהוי מקרי כשל, הושם דגש על Recall ו-F1 עבור מחלקת ה-Fail. Recall. מודד את שיעור מקרי הכשל שהמודל מצליח לזהות, בעוד F1-score מספק איזון בין זיהוי כשלים לבין צמצום אזהקות שווא. 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 ממקרי הכשל, עם ערכי Recall הנעים בטווח של כ-0.64–0.80. עם זאת, זיהוי מלא של כלל מקרי הכשל אינו מושג, דבר המעיד על מגבלות הדאטה, רעש ומידע חסר שאינם ניתנים ללכידה מלאה באמצעות המאפיינים הקיימים. בנוסף, ניכרת רגישות לבחירת סף ההחלטה במודלים שבהם בוצע כיוונון, אשר השפיעה על האיזון בין Recall ל-Precision.

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 יעילה 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 ה-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-מפוקחת

לצורך בחינת מבנה הדאטה וזיהוי דפוסים שאינם תלויים בתוויות היעד, בוצע ניתוח למידה לא-מפוקחת על מדגם של 20,000 תצפיות, לאחר ביצוע תקנון באמצעות StandardScaler. לצורכי ויזואליזציה והפחתת ממדיות יושם PCA, כאשר שני הרכיבים הראשיים הראשונים מסבירים כ-40% מהשונות הכוללת בנתונים.

אלגוריתם KMeans יושם על הייצוג המופחת, כאשר מספר האשכולות נבחר על בסיס ניתוח Elbow ו-Silhouette והצביע על ארבעה אשכולות מרכזיים. לצורך זיהוי חריגים ומבני צפיפות יושם אלגוריתם DBSCAN, לאחר בחירת הפרמטרים באמצעות k-distance plot, אשר זיהה תשעה אשכולות וכן כ-4.7% תצפיות חריגות.

ממצאים אלו מצביעים על קיומם של דפוסי התנהגות שונים במוסדות המזון, לצד קבוצת תצפיות חריגה שאינה משתלבת בדפוסים המרכזיים, ועשויה לייצג מקרי סיכון ייחודיים. ניתוח זה מדגיש את התרומה המשלימה של למידה לא־מפוקחת להבנת הדאטה מעבר לחיזוי המפוקח.

סיכום ודיון

בפרויקט זה יושם תהליך מלא של למידת מכונה לניתוח נתוני ביקורות תברואה במוסדות מזון בעיר שיקגו. העבודה שילבה למידה מפוקחת לצורך חיזוי מקרי כשל בביקורת, יחד עם למידה לא־מפוקחת שנועדה להבין את מבנה הנתונים ולזהות דפוסים וחריגים. שילוב זה אפשר הסתכלות משלימה על הנתונים הן מנקודת מבט חיזויית 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 ללכוד דפוסים מורכבים. במקביל, הניתוח הלא־מפוקח הדגיש קיומם של אשכולות מובחנים לצד תצפיות חריגות, אשר עשויות להעיד על מקרים חריגים בנתונים, שאינם מזהים במלואם באמצעות מודלי סיווג בלבד.

חלוקת תפקידים

העבודה בוצעה בשיתוף פעולה מלא בין חברות הצוות, כאשר שלבי הפרויקט בוצעו במפגשים משותפים. שלבי העבודה כללו עיבוד וניתוח הנתונים, הגדסת מאפיינים, בניית מודלים מפוקחים, ביצוע ניתוח למידה לא־מפוקחת, ניתוח התוצאות, הכנת המצגת, וכתיבת הדוח הסופי. כל חברות הצוות נטלו חלק פעיל בכל שלבי הפרויקט, והאחריות התחלקה באופן שווה לאורך העבודה.

כיוונים עתידיים

כיוונים אפשריים להמשך כוללים הרחבת הדאטה באמצעות מקורות מידע נוספים, כיוון מתקדם של פרמטרים במודלים המפוקחים, ושילוב כלים נוספים לניתוח והבנה של הנתונים. בנוסף, ניתן להעמיק את השימוש בלמידה לא־מפוקחת לצורך זיהוי דפוסים נוספים וחריגים 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 מאתר data.gov באמצעות קריאה דינמית בזמן הרצת הקוד. מכיוון שהמאגר מתעדכן באופן שוטף, ייתכנו 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.