**זיהוי הונאות בכרטיסי אשראי**

**Credit card fraud detection**

המכללה האקדמית להנדסה עזריאלי ירושלים

המחלקה להנדסת תוכנה

פרויקט מסכם ב-למידה חישובית

**מאת**: מחמד חדר זעתרי. ת״ז: 314929811

**מאת**: מועתז תמימי. ת״ז: 319019378

**המרצה**: ד״ר יהודה חסין

**מקדמה**

אנחנו הולכים להשתמש בשלושה מודלים לזיהוי הונאות בהעברת כספים דרך כרטיסי האשראי, התכונות scaled , גם השמות שלהם אינם מוצגים בשמם האמתי מסיבות הפרטיות, למרות זאת  אנחנו עדיין יכולים לעשות אנליזות על הנתונים האלה.

**הבעיה**

בחיי יום יום רוב האנשים משתמשים בכרטיסי אשראי כדי לשם כסף לאחרים או כדי לקנות מחנויות, רוב ההעברות האלה הן העברות שבעל הכרטיס הוא שהעביר אותם לצד השני, אבל יש מקרים שלא בעל הכרטיס שמעביר את הכסף אלא צד שלישי שהוא שעושה את ההונאה, חברות האשראי רוצים להגן על הלקוחות שלהם בזה שלא יאפשרו העברות לא תקינות ויימנעו את ההונאה.

**פתרון הבעיה**

אנחנו הולכים לנתח את הנתונים שיש לנו כדי למצוא את העברות ההונאה, ואחרי הניתוח את נתונים צריך לאמן כמה מודלים עם פרמטרים שונים ולקחת את המודל עם הכי פחות שגיאה ולהשתמש בו כדי לזהות הונאות.

* **Logistic regression**

הוא מודל סיווג בינארי שבו ההסתברות המותנית של אחד משני המימושים האפשריים של משתנה הפלט נחשבת שווה לשילוב ליניארי של משתני הקלט, שהשתנו על ידי הפונקציה הלוגיסטית.

היא אחד האלגוריתמים הפשוטים ביותר של למידת מכונה וקלה ליישום אך מספקת יעילות אימון נהדרת במקרים מסוימים. גם בשל סיבות אלה, אימון מודל עם אלגוריתם זה אינו דורש כוח חישוב גבוה..

A picture containing text, clock, gauge

Description automatically generated

* Support Vector machine

הוא אלגוריתם למידת מכונה המנתח נתונים לניתוח סיווג. היא שיטת למידה מפוקחת שמסתכלת על נתונים וממיין אותו לאחת משתי קטגוריות. SVM מפיק

פלט של הנתונים ממוינים עם השוליים בין השניים רחוקים ככל האפשר זה מזה.

A picture containing diagram

Description automatically generated

* Random forest Classifier

הוא מסווג המכיל מספר עצי החלטה על תת-קבוצות שונות של ערכת הנתונים הנתונה ולוקח את הממוצע כדי לשפר את הדיוק החזוי של ערכת נתונים זו

**הנתונים**

* האתר שמצאנו בו את הנתונים: <https://www.kaggle.com/>
* אתר הנתונים שעבדנו עליהם: <https://www.kaggle.com/mlg-ulb/creditcardfraud>

|  |  |  |
| --- | --- | --- |
| מספר התכונות | מספר הדגימות | מספר הדגימות הן הונאה |
| 31 | 284,804 | 492 , 0.172% מכל הדגימות |

הגרף הבא מיישג ביזור המחלקות :

Shape, square

Description automatically generated

--מהנתון האחרון יוצא לנו **שהנתונים** **אינם מאוזנים**--

**התמודדות עם הנתונים**

בעיות בנתונים שלנו:

1. אי איזון במספר המחלקות (0 , 1)של עמודת ה (y).
2. Scaled data

איך התמודדנו עם הבעות האלה:

* **OverSampling** : זה אומר העלאת מספר הדגימות של ההעברות הלא תקינות למספר ששווה למספר הדגימות של ההעברות התקינות. על ידי שכפול הנתונים ובכך ניצור מאגר נתונים שבו נשתמש כדי לאמן את המודלים שלנו כדי להגיע לאחוזי הצלחה רבה יותר.
* **Synthetic Minority Oversampling Technique (SMOTE):** זאת הגישה הנפוצה ביותר להתמודד עם אי איזון הנתונים היא כמו השיטה למעלה אבל היא לא משכפלת אלא יוצרת נתונים חדשים על פי הנתונים שיש לנו.

**Hyperparameter**

כדי להשיג את הפרמטרים הטובים ביותר עבור המודלים שלנו השתמשנו בפונקציית GridSearchCV, הפונקציה הזו מאמנת את המודל ובוחנת את ה score עובר כל קומבינציה של פרמטרים ומחזירה את הפרמטרים הטובים ביותר.

|  |  |  |
| --- | --- | --- |
| F-Score | C | Name |
| 0.7296137339055795 | Default | Logestic Regressions |
| 0.7296137339055795 | 1 |
| 0.10108303249097474 | Default | Logestic Regressions  OverSampling |
| 0.10140845070422536 | 0.1 |
| 0.09809264305177112 | Default | Logestic Regressions  SMOTE |
| 0.09809264305177112 | 10 |

**מטרצית הConfusion**:

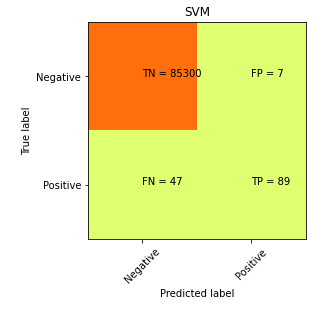
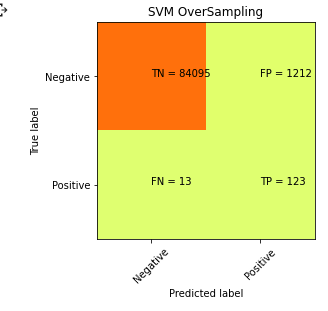
**Chart

Description automatically generated**Chart

Description automatically generatedרצינו לדעת היעילות לכל אחד מהפונקציות האלה לכן מצאנו מטריציה שאומרת לנו עד כמה המודל יעיל בפונקציה הזאת וממנה מצאנו את הערכים למעלה.

**Chart

Description automatically generated**

**Chart

Description automatically generated**

**Overfitting:**

בעיה נפוצה אם המודל נותן תוצאות טובות לנתונים שאימנו (Training) אבל לא ל(test) בדקנו מודל ה RandomForestClassifier מצאנו ה LOSS ב אימון ובדיקה ויצא שאין הבדל גדול אבל כש השתמשנו בשיטה OverSampling

ההפרש ירד עוד זה אומר שהמודל נותן תוצאות טובות לנתונים לא ידועים .

גרף הבא מתאר ה מודל בלי ועם השיטה:

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

**העמודות החשובות ביותר**

הפעלנו את האלגוריתם של (MI) mutual info ויצא להו שהתכונות האלה הם הכי חשובים לפי הסדר

|  |
| --- |
| V 14 |
| V17 |
| V12 |
| V10 |
| V11 |

**סיכום:**

Table

Description automatically generated

תוצאת ה f-score הטובה ביותר היא של המודל RandomForest עם שיטה האיזון OverSampling . בהפתעה גם קיבלנו תוצאה לא רעה של המודל RandomForest בלי לעשות איזון לנתונים שלנו.

ביחס לזמן הרצה המודלים גם ביחס לתוצאות ה f-score הטובות היינו ממליצים על המודל RandomForest הרגיל בלי שיטות איזון.

ה SVM עם OverSampling הוא המודל שלקח המון זמן באימון הנתונים.

למחברת הcolab : https://colab.research.google.com/drive/1PLhKvzm97OkibN86IYG2xHdPgNiokegF?usp=sharing