

Catboost For Fraud Detection In Financial Transactions

2021 IEEE International Conference on Consumer Electronics and Computer Engineering

چکیده

- کلاهبرداری مالی یک تهدید بزرگ رو به رشد و با پیامدهای شدید در صنعت مالی است.
- یادگیری ماشینی نقش فعالی در تشخیص تقلب در معاملات مالی ایفا می کند .
- این پژوهش یک روش یادگیری ماشین مبتنی بر CatBoost را برای تشخیص تقلب معرفی می کند.
- یکی دیگر از کارهای مهم، استفاده از فشرده سازی حافظه برای ردیابی سرعت است.

مقدمه

- یکی از چالش های عمده ای که ارائه دهندگان خدمات مالی با آن مواجه هستند، کشف تقلب است.
 - تشخیص تقلب مالی عمدتاً ناشی از فقدان کشف دانش و بینش کامل در مورد ماهیت یا الگوهای معاملات انجام شده و روند آنها است.
 - یادگیری ماشین روشی قابل توجه برای یادگیری دانش از مجموعه داده های بزرگ است و یکی از راه حل های رایج فعلی برای کشف تقلب و جلوگیری از تقلب مالی می باشد.
 - رویکردهای زیادی برای حل این مشکل پیشنهاد شده است که از جمله می توان:
- Support Vector Machine
 - Naive Bayes
 - Logistic Regression
 - K-Nearest Neighbor
 - Random Forests
 - Data Mining
 - Light Gradient Boosting Machine

روش

- یک روش جدید مبتنی بر CatBoost برای تشخیص تقلب بر روی مجموعه داده های تقلب در مقیاس بزرگ IEEE-CIS که در پلتفرم Kaggle
- شناسایی هر تراکنش در مجموعه داده را به عنوان یک رویداد تقلب یا غیر تقلب
- فشرده سازی حافظه برای داده های تراکنش خام به منظور بهبود سرعت تشخیص
- مهندسی ویژگی، یک تکنیک اساسی برای انتخاب بیشتر متغیرهای مرتبط برای تشخیص
- ساخت انواع ویژگی های جدید با فرآیند طراحی ویژگی با توجه به ویژگی های اصلی در مجموعه داده
- پیاده سازی مدل CatBoost کارآمد با ویژگی های استخراج شده به عنوان ورودی

مجموعه داده

- مجموعه داده تقلب IEEE - CIS توسط تراکنش های تجارت الکترونیک دنیای واقعی وستا ارائه شده است و شامل طیف گسترده ای از ویژگی ها از نوع دستگاه تا ویژگی محصول است. می توان آن را به چهار بخش تقسیم کرد:
- جدول تراکنش آموزشی
- جدول هویت آموزشی
- جدول تراکنش تست
- جدول شناسایی تست

مجموعه داده

Table 1. Details of transaction table

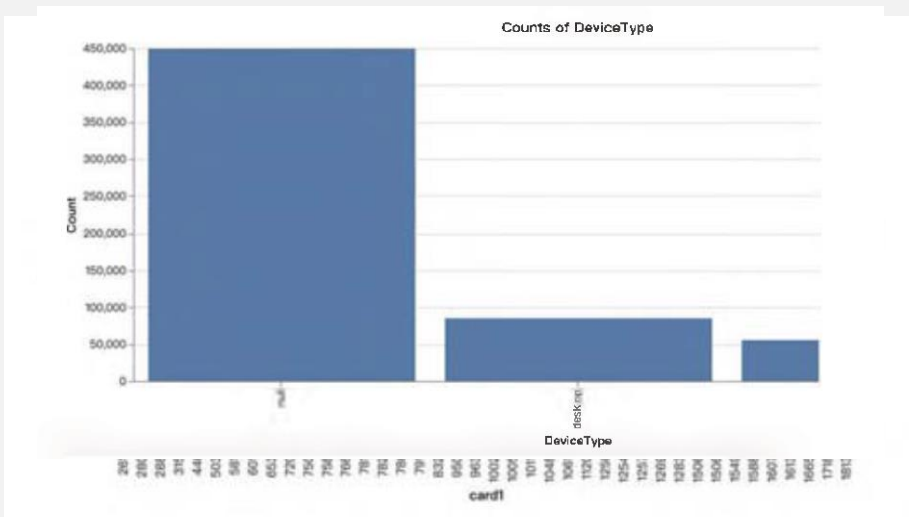
Variables	Description
TransactionDT	timedelta from a given reference datetime
TransactionAMT	transaction payment amount in USD
ProductCD	product code
card1-card6	payment card information
addr1-addr2	address
Dist	distance
Px and (Rx)	purchaser and recipient email domain
C1-C14	counting
D1-D15	timedelta
M1-M9	match
Vxxx:	Vesta engineered rich features
IsFraud	1 is fraudulent and 0 is non-fraudulent

پیش پردازش

- حذف نقاط پرت در داده های خام، برای کاهش نویز
- استفاده از تکنیک های فشرده سازی متفاوت برای انواع مختلف ستون های داده پردازش شده
- برای ستون های دسته بندی با کاردینالیتیه کم، الگوریتم را مجبور می کنیم از یک جدول نگاشت مجازی استفاده کند که در آن همه مقادیر منحصر به فرد از طریق یک عدد صحیح به جای اشاره گر نگاشت می شوند.

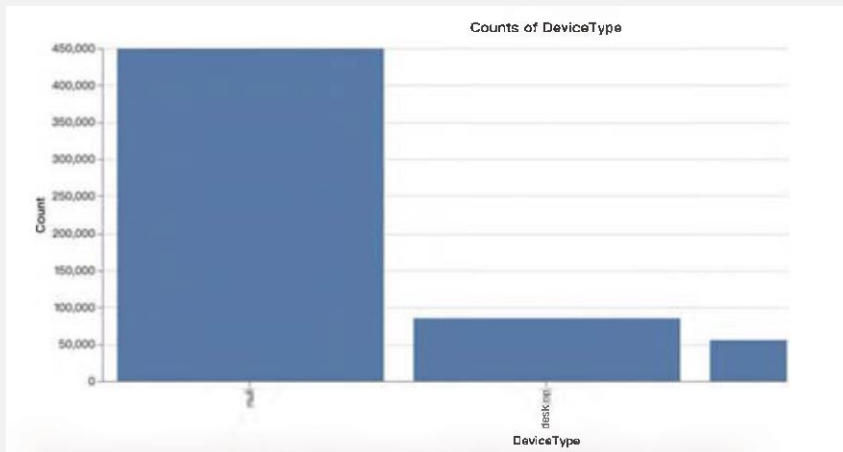
مهندسی ویژگی

- ویژگی های اصلی با توجه به برخی ویژگی ها در کل مجموعه داده , به عنوان تعداد card1 نشان داده شده در شکل ۱ و تعداد DeviceType نشان داده شده در شکل ۲
- برخی از ویژگی های داده شدید مانند حداقل، حداکثر.
- ویژگی های مبلغ تراکنش سالانه، ماهانه، هفتگی و روزانه با توجه به ویژگی های زمانی.
- چند ویژگی آماری مانند میانگین، واریانس، مجموع، صدک ها و ...



مهندسی ویژگی

- افزایش ویژگی ها ممکن است باعث بیش برآزش شود و منابع آموزشی بیشتری مصرف کند. برای جلوگیری از برآزش بیش از حد و سرعت بخشیدن به تمرین، از تحلیل همبستگی برای حذف برخی ویژگی ها با ضریب همبستگی بالای ۰.۹۵ استفاده می شود.
- که حذف ویژگی ها ممکن است باعث از دست رفتن داده ها شود که یک اتفاق رایج است و عملکرد الگوریتم را در سطوحی تحت تأثیر قرار می دهد، ما برای مقابله با مقادیر از دست رفته با پر کردن مقادیر گم شده با -۹۹۹، از انتساب استفاده می کنیم.



مدل تشخیص تقلب مبتنی بر CATBOOST

- CatBoost، پیاده سازی Gradient Boosting و استفاده از درخت های تصمیم باینری به عنوان پیش بینی کننده های پایه، یک الگوریتم یادگیری ماشینی قدرتمند است و به نتایج پیشرفته ای در انواع وظایف عملی دست می یابد.

مدل تشخیص تقلب مبتنی بر CATBOOST

- برای ویژگی های طبقه بندی شده، همانطور که در مجموعه داده های تقلب IEEE-CIS دیده می شود، CatBoost یک الگوریتم خلاقانه برای پردازش آنها پیشنهاد می کند. در مورد ویژگی های با کاردینالیت بالا، برخی از رایج ترین روش ها برای رمزگذاری داده های طبقه بندی شده مانند رمزگذاری one-hot منجر به تعداد بسیار زیادی ویژگی جدید می شوند. یکی دیگر از روش های محبوب، گروه بندی دسته ها بر اساس آمار هدف است.
- CatBoost از اصل مرتب سازی استفاده می کند و به نام Target-Based نامیده می شود.

آزمایش

$$\text{TPR} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{FPR} = \frac{FP}{FP + TN} \quad (2)$$

$$\text{accuracy (ACC)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$


Table 2. The experimental results of our algorithm with other machine learning algorithms.

Models	AUC-ROC Score	Accuracy
Naïve Bayes	0.847	0.925
SVM	0.918	0.950
CatBoost	0.971	0.983

نتیجه گیری

این پژوهش به بررسی چالش‌های کشف تقلب در صنعت مالی می‌پردازد و یک روش یادگیری ماشینی مبتنی بر CatBoost را برای بهبود کارایی تشخیص پیشنهاد می‌کند. نتایج آزمایشی بر روی مجموعه داده تقلب IEEE - CIS که توسط تراکنش‌های تجارت الکترونیک دنیای واقعی وستا ارائه شده است و دارای طیف گسترده‌ای از ویژگی‌ها است، نشان می‌دهد که CatBoost در مقایسه با سایر روش‌های یادگیری ماشین پیشرفت قابل توجهی در عملکرد تشخیص تقلب کسب کرده است.

با تشکر از توجه شما

 Mohammadpour.b1601@Semnan.ac.ir