به نام خدا

نمونه کار پروژه تشخیص کارت اعتباری جعلی – (Credit Card Fraud Detection)

خلاصه پروژه تشخیص کارت اعتباری جعلی (Credit Card Fraud Detection)

این پروژه هدف تشخیص کارت اعتباری جعلی، با استفاده از الگوریتم مدل XGBoost ، SGDClassifier ، RandomForestClassifier و Kernel Approximation توسعه داده شده است.

به دلیل دیتاست بزرگ و نامتوازن از XGBoost استفاده شد.

این الگوریتم یادگیری ماشین XGBoost نتیجه نهایی Test و Validation عملکرد خوبی ارائه شد.

نتایج نهایی مدل با استفاده از XGBoost

Validation metrics

Metric	Value
Accuracy	99.96%
Precision	96.83%
Recall	79.22%
f1 score	87.14%

:Classification report

```
------------------- Validation sets metrics -------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     42644
           1       0.97      0.79      0.87        77

    accuracy                           1.00     42721
   macro avg       0.98      0.90      0.94     42721
weighted avg       1.00      1.00      1.00     42721
```


DO NOT COPY

Test metrics

Metric	Value
Accuracy	99.97%
Precision	91.07%
Recall	86.44%
f1 score	88.70%

```
------------------- Test sets metrics -------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     42663
           1       0.91      0.86      0.89        59

    accuracy                           1.00     42722
   macro avg       0.96      0.93      0.94     42722
weighted avg       1.00      1.00      1.00     42722
```

مقدمه (introduction)

در این پروژه هدف طراحی، آموزش و ارزیابی یک سیستم یادگیری ماشین برای تشخیص تراکنش‌های تقلبی کارت اعتباری است. داده‌های مورد استفاده شامل حدود ۲۸۴ هزار تراکنش هستند که به‌شدت نامتوازن بوده و تنها حدود 0.2٪ از آن‌ها نمونه‌های تقلبی می‌باشند. این نامتوازن‌بودن، طراحی مدل را چالش‌برانگیز کرده و نیازمند استفاده از روش‌های پیش‌پردازش، معیارهای مناسب، و مدل‌هایی مقاوم در برابر عدم توازن داده است.

داده ها (Dataset)

شامل اطلاعات تراکنش‌های مالی واقعی است. این دیتاست شامل:

- 284,807 رکورد تراکنش
- 31 ویژگی (Feature) شامل:
- 28 ویژگی به کامپوننت ریاضی تبدیل شده و به صورت PCA اجرا کرده اند. (از V1 تا V28)
- ویژگی Amount (مبلغ تراکنش)
- ویژگی Class (برچسب هدف)
- ویژگی Time حذف شد زیرا در مدل‌سازی تأثیر مطلوب نداشت.

مدل انتخاب شده (Model Selection)

چندین مدل بررسی و مقایسه شد از جمله:

1- RandomForestClassifiier
2- Kernel Approximattion + SGDClassifier
3- SGDClassifier (Linear Models)
4- XGBoost (Extreme Gradient Boosting)

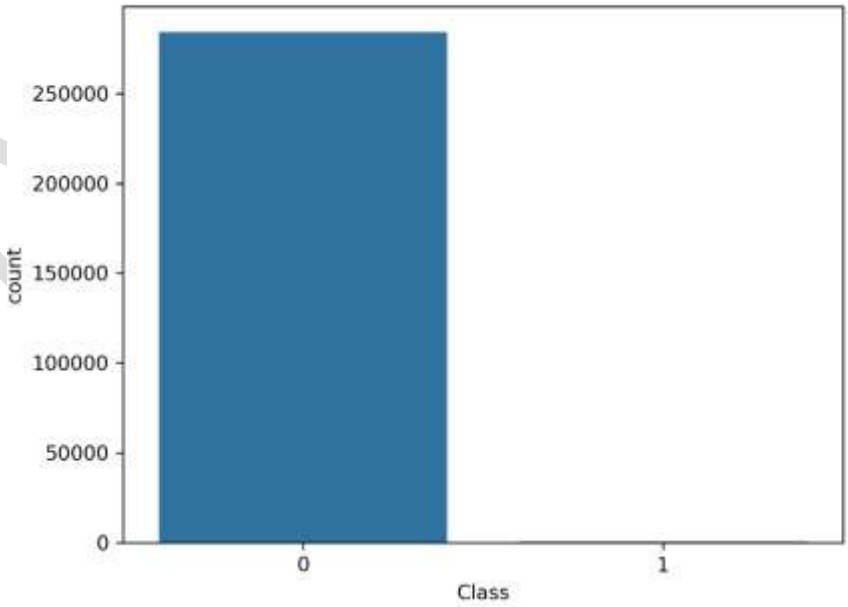
در نهایت، با وجود عملکرد قابل قبول برخی مدل ها، مدل XGBoost درمجموع بهترین نتایج ارائه داد. یکی از دلایل انتخاب XGBoost، میتونه بما کمک میکنه که بتونه دیتاست بزرگ و نامتوازن پیچیده عملکرد قوی کار کنه و همچنین این مدل دارای انطعاف پذیرتر در دیتاست بزرگ مناسب است چون از نظر دقت ، تشخیص و... عملکرد مدل و tuning دارای سرعت بالا بهتری داشت.

ارزیابی مدل (Model Evaluation)

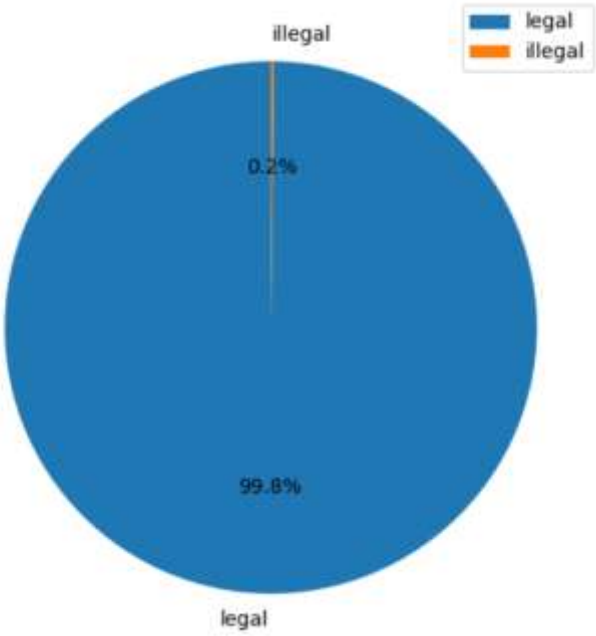
ارزیابی سه مرحله انجام شد:

1- Training set
2- Validation set
3- Test set

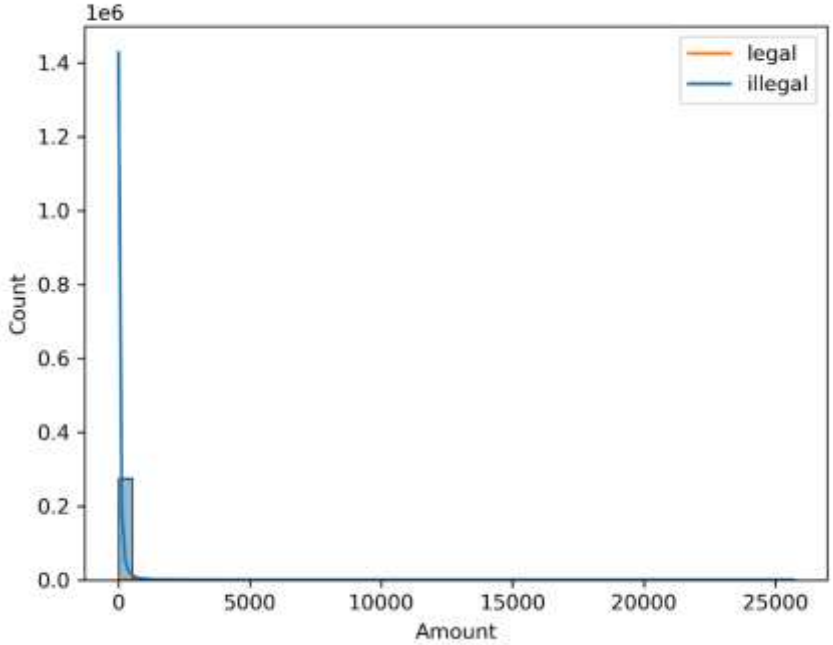
نمودار ها (Visualizations)



این نمودار نشان میدهد که داده های خام به صورت نمودار نمایش میدهد.



این نمودار به صورت دایره ای دو کلاس قانونی و غیر قانونی نمایش میدهد.



این نمودار توزیع مقدار تراکنش‌ها (Amount) را برای هر دو کلاس **قانونی (0)** و **تقلبی (1)** نشان می‌دهد.

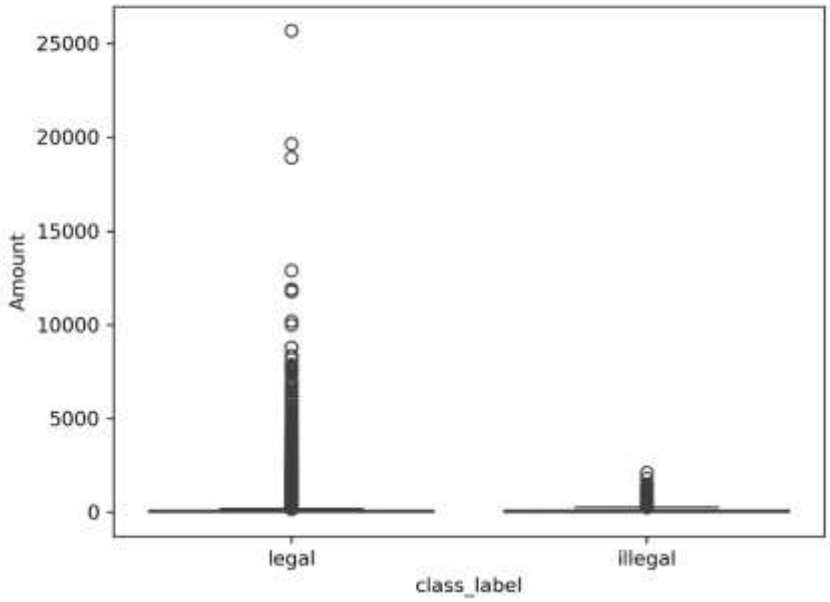
KDE (Kernel Density Estimation) منحنی تخمینی چگالی است که الگوی کلی توزیع را بهتر نمایش می‌دهد.

مشاهده می‌شود که:

- تراکنش‌های قانونی در محدوده‌های مختلف Amount پراکندگی بیشتری دارند.

- تراکنش‌های تقلبی معمولاً مقادیر کوچک‌تری دارند و تراکم آن‌ها در بازه‌های پایین‌تر Amount بیشتر است.

این اختلاف توزیع به مدل کمک می‌کند که ویژگی Amount اهمیت بالایی در تشخیص تقلب داشته باشد.



نمودار Boxplot توزیع آماری ویژگی Amount را به تفکیک کلاس‌ها نمایش می‌دهد.

در این نمودار موارد زیر قابل مشاهده است:

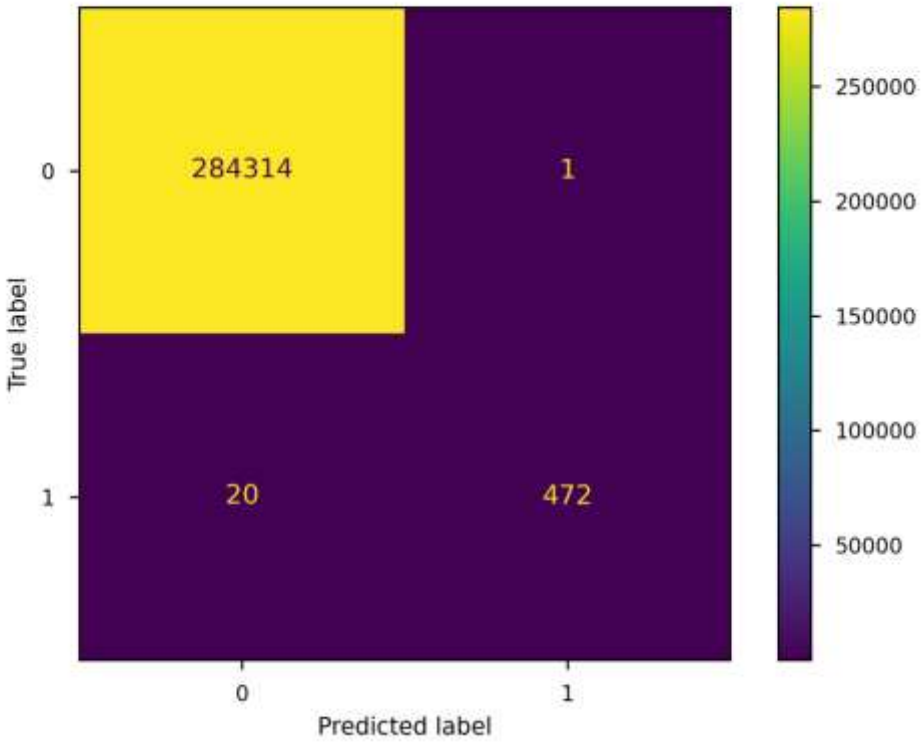
- **میانه (Median)** تراکنش‌های تقلبی بسیار کمتر از تراکنش‌های قانونی است.
- IQR (بازه بین چارکی) در کلاس قانونی گسترده‌تر است، یعنی پراکندگی مقادیر Amount در تراکنش‌های قانونی بیشتر است.
- تعداد **outlier** ها (نقاط پرت) در کلاس قانونی بسیار بیشتر دیده می‌شود، چون تعداد نمونه‌های آن زیاد است.
- تراکنش‌های تقلبی مقدار پول کمی جابجا می‌کنند، که یکی از الگوهای رایج در تقلب‌های مالی است.

مقایسه مدل ها

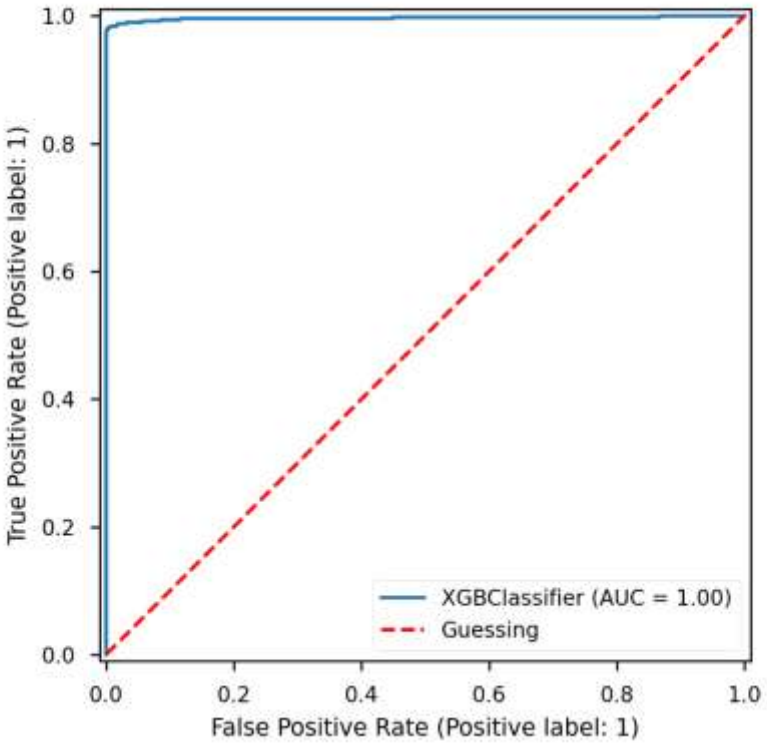
مدل	Accuracy	Precision	Recall	F1	ROC-AUC	نتیجه
RandomForest	99.96%	97.40%	76.53%	85.71%	88.26%	عملکرد خوب اما کند
Kernel Approximation	99.92%	84.27%	64.66%	73.17%	82.32%	سرعت بالا اما f1 پایین
SGDClassifier	94.65%	2.86%	92.68%	5.54%	93.67%	مناسب نبود.
XGBoost	99.96%	98.73%	79.59%	88.14%	89.80%	عملکرد عالی

نتایج اصلی در مدل XGBoost

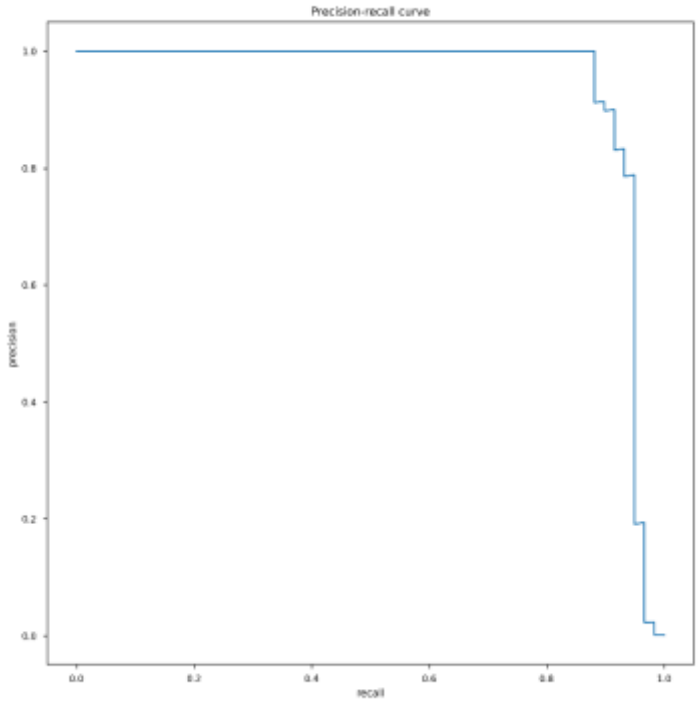
```
------------------- Validation sets metrics -------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     42644
           1       0.97      0.79      0.87        77

    accuracy                           1.00     42721
   macro avg       0.98      0.90      0.94     42721
weighted avg       1.00      1.00      1.00     42721

Accuracy: 99.96%
Precision: 96.83%
Recall: 79.22%
f1 score: 87.14%
confusion matrix:
[[42642     2]
 [   16    61]]
roc score: 89.61%

------------------- Test sets metrics -------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     42663
           1       0.91      0.86      0.89        59

    accuracy                           1.00     42722
   macro avg       0.96      0.93      0.94     42722
weighted avg       1.00      1.00      1.00     42722

Accuracy: 99.97%
Precision: 91.07%
Recall: 86.44%
f1 score: 88.70%
confusion matrix:
[[42658     5]
 [    8    51]]
roc score: 93.21%
{'accuracy': 0.999695707129816,
 'precision': 0.9107142857142857,
 'recall': 0.864406779661017,
 'f1': 0.8869565217391304}
```



مدل تعدادی زیادی از تراکنش قانونی را درست شناسایی کرده (284314 نمونه) و تنها 1 نمونه تراکنش قانونی را اشتباه تشخیص داده است. همچنین، مدل بیشتر تراکنش تقلبی را شناسایی کرده (472 نمونه) و 20 تراکنش تقلبی از دست رفته است.



- منحنی ROC مدل نشان دهنده عملکرد بسیار ایده آل است. مقدار AUC = 1.00 به این معنی است که مدل تقریباً به‌طور کامل توانایی تفکیک بین تراکنش‌های قانونی و تقلبی را دارد.
- خط قرمز (Guessing Line) مدل تصادفی را نشان می‌دهد؛ فاصله زیاد منحنی آبی از آن بیانگر قدرت بالای مدل است.
- در مسائلFraud Detection ، مقدار AUC بالا نشان می‌دهد که مدل **در طیف گسترده‌ای از Thresholdها پایدار و قابل اعتماد است.**



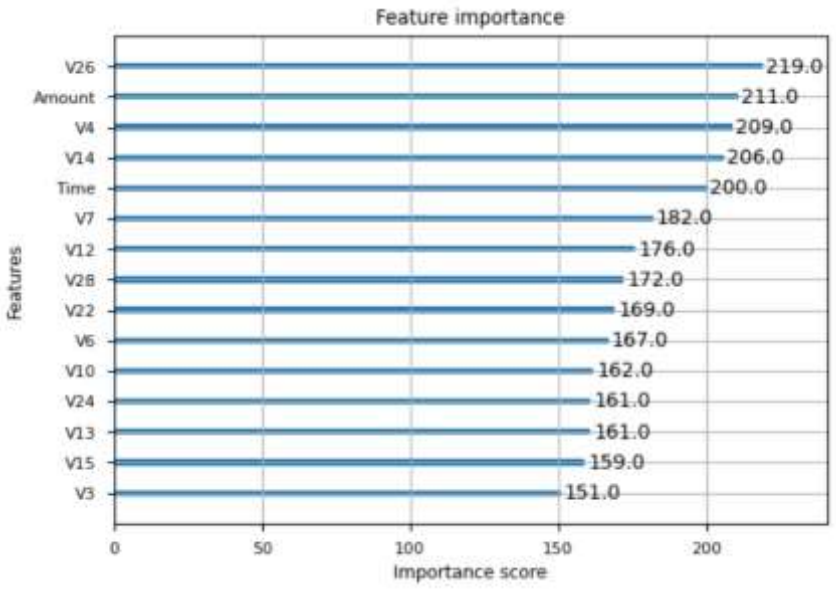
این نمودار رابطه بین Precision (دقت تشخیص تراکنش‌های تقلبی) و Recall (توانایی مدل در شناسایی تمام تقلب‌ها) را در سطوح مختلف آستانه تصمیم نشان می‌دهد.

مدل XGBoost تقریباً در تمام نقاط منحنی Precision-Recall بسیار نزدیک به مقدار ۱ است.

این رفتار نشان می‌دهد که مدل حتی هنگام افزایش (Recall که معمولاً باعث کاهش Precision می‌شود)، همچنان قدرت حفظ دقت بالا را دارد.

چنین عملکردی برای دیتاست‌های بسیار نامتوازن (Fraud Detection) بسیار ارزشمند است، زیرا تعداد تراکنش‌های تقلبی بسیار کم است و مدل معمولاً با افزایش Recall دچار افت Precision می‌شود، اما اینجا افت بسیار محدود است.

نتیجه: مدل حتی در شرایط نرخ تقلب بسیار پایین هم قادر به تشخیص مؤثر کلاس مثبت (تقلب) است.



این نمودار میزان اهمیت هر ویژگی را در مدل XGBoost نشان می‌دهد.

تحلیل فنی:

- ویژگی‌های اصلی مؤثر در تشخیص تقلب شامل V26، Amount، V4، V14، V7، V12، V28، V22، V6 هستند.
- این اهمیت‌ها بر اساس تعداد دفعات استفاده ویژگی‌ها در درخت‌های تصمیم در فرآیند Boosting محاسبه شده‌اند.
- اهمیت زیاد ویژگی Amount نشان می‌دهد که مقدار تراکنش نقش مهمی در تشخیص الگوهای غیرطبیعی دارد.
- ویژگی‌های V1 تا V28 که از PCA در بانک داده اصلی ساخته شده‌اند، رفتارهای پنهان کاربران را مدل می‌کنند و باعث می‌شوند XGBoost الگوهای غیرخطی پیچیده تقلب را شناسایی کند.

نتیجه: مدل توانسته است ویژگی‌های کلیدی را از میان بیش از ۳۰ ورودی به‌صورت مؤثر استخراج و بر اساس آن‌ها تصمیم‌گیری کند.

DO NOT COPY

تکنولوژی و ابزارهای استفاده شده

- Python
- Scikit learn
- Numpy
- Pandas
- Matplotlib
- Seaborn
- SMOTE
- Jupyter notebook

نتیجه گیری (Conclusion)

در این پروژه، با هدف تشخیص تراکنش‌های تقلبی، یک مدل طبقه‌بندی مبتنی بر XGBoost توسعه داده شد. با توجه به نامتوازن بودن شدید داده‌ها (کمتر از 0.2٪ تراکنش‌ها تقلبی هستند)، تمرکز اصلی بر روی شاخص‌های کلیدی مانند Precision، Recall، F1-score، ROC-AUC و تحلیل Precision-Recall Curve بوده است.

ارتباط با ما

Email: www.mohamadnazari771998@gmail.com

Phone number: 09397196427

GitHub: https://github.com/MohammadNazari98/MyPortfolio

DO NOT COPY