گزارش کار : الهه قادری

LogesticRegression for prediction loan default algorithm

	precision	recall	f1-score	support
0	0.80	1.00	0.89	1189
1	0.75	0.03	0.06	301
accuracy			0.80	1490
macro avg	0.78	0.51	0.47	1490
weighted avg	0.79	0.80	0.72	1490

جدول بالا **جدول طبقه‌بندی** (Classification Report)مربوط به عملکرد یک مدل یادگیری ماشین در **پیش‌بینی دو کلاس** را نشان می‌دهد.

ستون‌ها:

Precision: نشان می‌دهد که چه تعداد از **پیش‌بینی‌های مثبت** مدل **درست** هستند.

Recall: نشان می‌دهد که چه تعداد از **موارد مثبت واقعی** توسط مدل **شناسایی** شده‌اند.

F1-Score : میانگین وزنی دقت و بازخوانی است که **تعادلی** بین این دو معیار ایجاد می‌کند.

Support :تعداد کل نمونه‌ها در یک کلاس خاص را نشان می‌دهد.

سطرها:

صفر : مربوط به کلاس **منفی** (غیر پرداخت کننده)

یک : مربوط به کلاس **مثبت** (پرداخت کننده)

مدل در **شناسایی موارد منفی** (غیر پرداخت کننده) با دقت ۸۰٪ و F1-Score 89% عملکرد **خوبی** دارد.

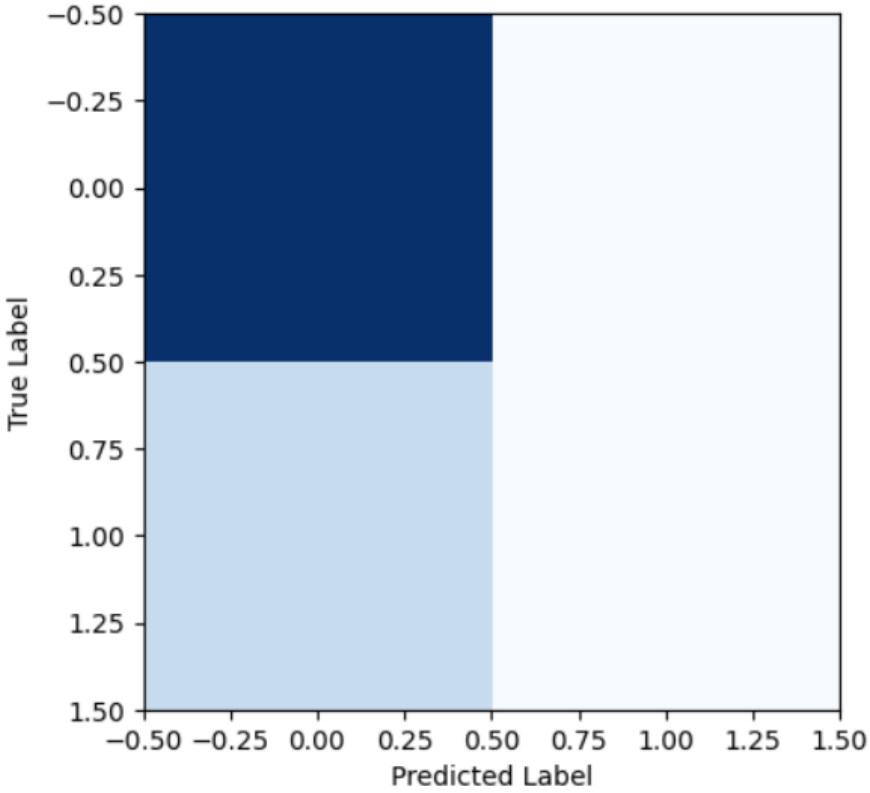
مدل در **شناسایی موارد مثبت** (پرداخت کننده) با دقت ۷۵٪ و F1-Score 6% عملکرد **ضعیفی** دارد.

دلایل احتمالی عملکرد ضعیف در شناسایی موارد مثبت:

تعداد کم نمونه‌های مثبت در مجموعه داده، **عدم تعادل** بین تعداد نمونه‌های مثبت و منفی،**انتخاب نامناسب ویژگی‌ها** برای مدل، **تنظیم نامناسب پارامترهای** مدل

راهکارهای احتمالی برای بهبود عملکرد:

جمع‌آوری داده‌های بیشتر برای کلاس مثبت ، **استفاده از تکنیک‌های نمونه‌گیری** برای تعادل تعداد نمونه‌ها در دو کلاس، **انتخاب ویژگی‌های مناسب‌تر** برای مدل، **تنظیم دقیق پارامترهای** مدل



نمودار بالا **Confusion Matrix را نشان می دهد** برای ارزیابی عملکرد الگوریتم‌های یادگیری ماشین در **مسائل طبقه‌بندی** است. این ماتریس نشان می‌دهد که الگوریتم چه تعداد نمونه را به درستی و چه تعداد را به اشتباه طبقه‌بندی کرده است.

محور افقی نشان‌دهنده **برچسب‌های پیش‌بینی‌شده** توسط الگوریتم است.

محور عمودی نشان‌دهنده **برچسب‌های واقعی** نمونه‌ها است.

هر سلول در ماتریس نشان‌دهنده تعداد نمونه‌هایی است که به یک برچسب خاص **پیش‌بینی** و **طبقه‌بندی** شده‌اند.

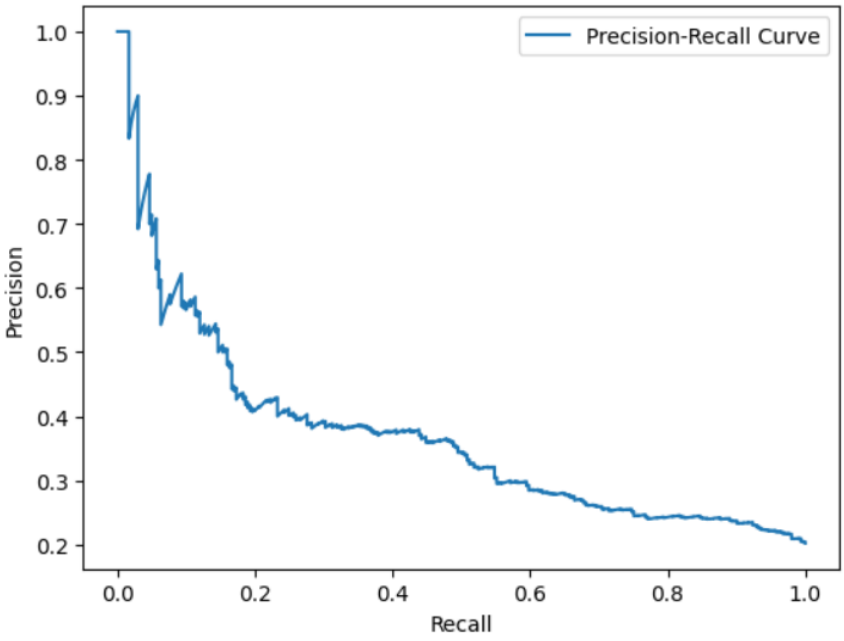
با استفاده از ماتریس درهم‌ریختگی، می‌توانید معیارهای مختلفی را برای ارزیابی عملکرد الگوریتم خود محاسبه کنید:

- **دقت :(Accuracy)** نسبت نمونه‌های **به درستی طبقه‌بندی‌شده** به کل نمونه‌ها.

- **دقت :(Precision)** نسبت نمونه‌هایی که **به عنوان یک برچسب خاص پیش‌بینی شده‌اند** و **واقعاً آن برچسب را دارند** به **کل نمونه‌هایی که به عنوان آن برچسب پیش‌بینی شده‌اند**.

- **توان یادآوری :(Recall)** نسبت نمونه‌هایی که **واقعاً یک برچسب خاص دارند** و **توسط الگوریتم به عنوان آن برچسب پیش‌بینی شده‌اند** به **کل نمونه‌هایی که واقعاً آن برچسب را دارند**.

- **F1-Score:** میانگین هموزنی از دقت و توان یادآوری.



نمودار Precision-Recall مربوط به عملکرد یک مدل یادگیری ماشین در **تشخیص دو کلاس** را نشان می‌دهد.

محورها:

- **Recall**نشان‌دهنده **نسبت موارد مثبت** که به درستی **مثبت** پیش‌بینی شده‌اند.
- **Precision**نشان‌دهنده **نسبت موارد پیش‌بینی شده مثبت** که **در واقع** مثبت هستند.

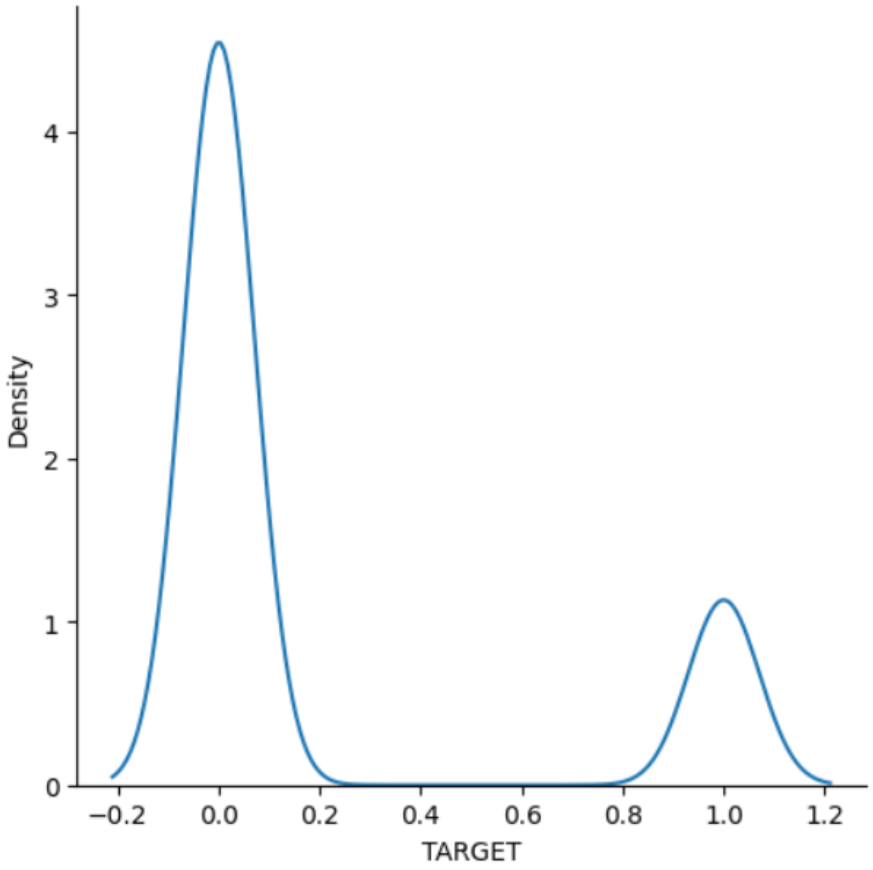
منحنی:Precision-Recall

منحنی Precision-Recall نشان‌دهنده **عملکرد مدل** در **تعادل بین دقت و بازخوانی** است.

منحنی ایده‌آل در گوشه سمت راست بالا قرار دارد و نشان‌دهنده **دقت و بازخوانی ۱۰۰ %**است.

منحنی تصادفی در خط مورب قرار دارد و نشان‌دهنده **عملکرد تصادفی** مدل است.

- از منحنی Precision-Recall می‌توان برای **انتخاب آستانه** مناسب برای پیش‌بینی‌های مدل استفاده کرد..



نمودار چگالی احتمال (Probability Density Function - PDF) یک متغیر تصادفی پیوسته است.

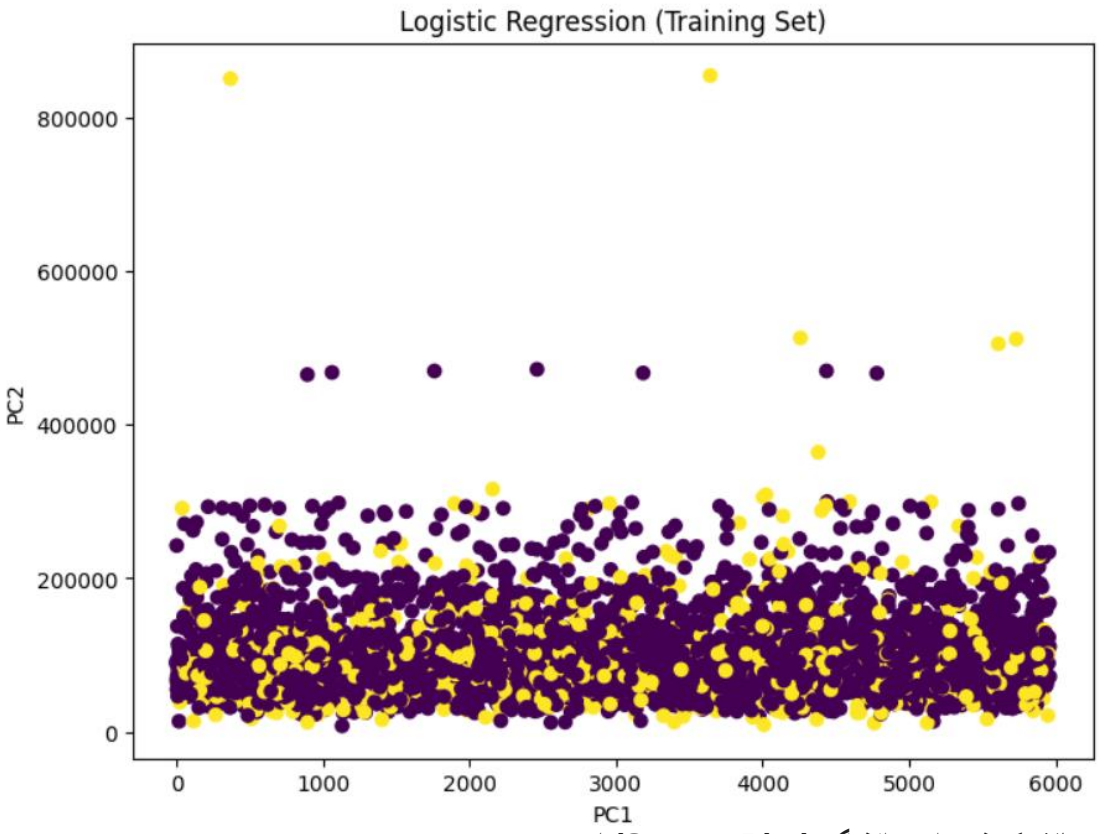
محور افقی :مقادیر متغیر تصادفی را نشان می‌دهد .**محور عمودی :**چگالی احتمال را در هر مقدار نشان می‌دهد.

در این نمودار:

منحنی صاف و هموار نشان می‌دهد که متغیر تصادفی در **محدوده‌ای خاص** بیشتر احتمال وقوع دارد.

نقطه اوج منحنی نشان می‌دهد که **بیشترین احتمال** وقوع متغیر تصادفی در **کدام مقدار** است.

مساحت زیر منحنی در هر بازه، **احتمال وقوع** متغیر تصادفی در آن بازه را نشان می‌دهد



نمودار بالا یک **نمودار پراکندگی (Scatter Plot)** است.

محورها:

محور افقی :مقادیر PC1 را نشان می‌دهد.

محور عمودی :مقادیر PC2را نشان می‌دهد.

نقاط:

هر نقطه در نمودار، **یک نمونه** را نشان می‌دهد.

موقعیت هر نقطه بر روی نمودار، **مقادیر PC1 و C2** مربوط به آن نمونه را نشان می‌دهد.

همبستگی:

برای همبستگی مثبت:

نقاط در نمودار **به صورت مورب بالا به سمت راست** پراکنده شده‌اند.

با افزایشPC1 ، PC2نیز **افزایش** می‌یابد.

برای همبستگی منفی:

نقاط در نمودار **به صورت مورب پایین به سمت راست** پراکنده شده‌اند.

با افزایشPC1 ، PC2**کاهش** می‌یابد.

برای عدم همبستگی:

نقاط در نمودار **به صورت تصادفی** پراکنده شده‌اند.

بین PC1 و C2 **هیچ رابطه‌ای** وجود ندارد.

در نمودار بالا همبستگی مثبت بین PC1 و PC2 وجود دارد.

با افزایشPC1 ، PC2نیز **افزایش** می‌یابد.

پراکندگی نقاط نشان می‌دهد که این رابطه **خطی** نیست.

الگوریتم های دیگر از جمله kneighbor,randomforest,svc را نیز میتوان به همین شیوه الگوریتم logeticregressionتحلیل و ارزیابی کرد

رفرنس های مورد استفاده برای کدنویسی پیش بینی مدل:

https://github.com/sonarsushant/Loan-Defaulter-Prediction/blob/master/Model_Training_and_Evaluation.ipynb

https://github.com/harishpuvvada/LoanDefault-Prediction/blob/master/Loan_Default_Prediction_Final.ipynb

https://github.com/the-ogre/LoanDefaulterPrediction/blob/main/Bank_Loan_Defaulter_Prediction_.ipynb

https://github.com/naveen-chauhan/Loan-Prediction-Classification/blob/master/Loan%2BPrediction.ipynb

https://github.com/shrikant-temburwar/Loan-Prediction-Dataset/blob/master/LoanPrediction.ipynb