گزارش بخش پردازش و کنترل کیفیت داده

بررسی کیفیت داده‌ها در پروژه تشخیص تقلب کارت اعتباری

در این گزارش، نتیجه بررسی‌هایی را که روی ساختار دیتاست، داده‌های ناقص، نوع داده‌ها، ستون هدف، مقادیر نامعتبر، رکوردهای تکراری و داده‌های پرت انجام دادم، ارائه کرده‌ام.

جمع‌بندی کلی: دیتاست با موفقیت بارگذاری و بررسی شد. داده‌ها شامل 284,807 ردیف و 31 ستون هستند. هیچ مقدار ناقص، مقدار نامعتبر در ستون هدف، مقدار منفی غیرمنطقی، مقدار بی‌نهایت یا مقدار غیرعددی مشاهده نشد. در عین حال، 1,081 ردیف تکراری و وجود داده پرت در 29 ستون شناسایی شد. در این مرحله هیچ داده‌ای حذف یا تغییر داده نشد.

نقش من در تیم

پردازش و کنترل کیفیت داده

پروژه

تشخیص تقلب با استفاده از یادگیری ماشین

برنچ گیت

preprossing

فایل نوت‌بوک

notebooks/data_quality_inspection.ipynb

تاریخ تهیه گزارش

۱۳ ژوئیه ۲۰۲۶

وضعیت اجرا

اجرای کامل و موفق

تهیه‌کننده: نلیسا - تیم Data Processing Engineering

۱. هدف و محدوده بررسی

هدف من در این مرحله، ارزیابی کیفیت دیتاست قبل از ورود به مراحل مدل‌سازی بود. برای این کار، فایل اصلی را از مسیر data/raw/creditcard.csv خواندم و فقط کیفیت و صحت ساختار آن را بررسی کردم. در این نوت‌بوک هیچ عملیات مدل‌سازی، مقیاس‌بندی، جایگزینی داده‌های ناقص، تقسیم داده به آموزش و آزمون یا حذف رکوردها انجام نشده است.

0	31	284,807
مقادیر ناقص	تعداد ستون‌ها	تعداد ردیف‌ها

۲. ساختار دیتاست

ساختار مورد انتظار دیتاست شامل ستون‌های زیر است:

گروه ستون	توضیح
Time	ویژگی مربوط به زمان سپری‌شده تراکنش.
V1 - V28	بیست‌وهشت ویژگی عددی ناشناس‌سازی‌شده که از تبدیل مؤلفه‌های اصلی به دست آمده‌اند.
Amount	مبلغ تراکنش.
Class	ستون هدف دودویی؛ مقدار 0 برای تراکنش عادی و مقدار 1 برای تراکنش تقلبی.

۳. روش انجام بررسی

برای ارزیابی کیفیت داده، مراحل زیر را انجام دادم:

- تعداد، نام و ترتیب ستون‌ها را با ساختار مورد انتظار مقایسه کردم.
- تعداد و درصد داده‌های ناقص را برای تمام ستون‌ها محاسبه کردم.
- نوع داده هر ستون و وجود مقادیر غیرعددی را بررسی کردم.
- ستون Class را بررسی کردم تا فقط شامل مقادیر 0 و 1 باشد.
- مقادیر منفی در Time و Amount، مقادیر بی‌نهایت و نویزهای عددی را بررسی کردم.
- رکوردهای کاملاً تکراری را شناسایی کردم، اما حذف نکردم.
- داده‌های پرت را با روش IQR شناسایی کردم، اما آن‌ها را حذف نکردم.

اجرای نوت‌بوک با پیام Data quality inspection completed successfully. به پایان رسید و بررسی نهایی وجود ۳۱ ستون، دودویی بودن ستون هدف و نبود مقادیر بی‌نهایت را تأیید کرد.

۴. نتایج نهایی

مورد بررسی	نتیجه	برداشت من
تعداد ردیف‌ها	284,807	تمام رکوردهای دیتاست با موفقیت بارگذاری شدند.
تعداد ستون‌ها	31	ساختار مورد انتظار کامل است.
داده‌های ناقص	0	در حال حاضر نیازی به جایگزینی یا حذف داده ناقص نیست.
مقادیر نامعتبر ستون هدف	0	ستون Class فقط مقادیر صفر و یک دارد.
مقادیر منفی در زمان	0	مقدار منفی غیرمنطقی در ستون Time وجود ندارد.
مقادیر منفی در مبلغ	0	مقدار منفی غیرمنطقی در ستون Amount مشاهده نشد.
مقادیر بی‌نهایت	0	هیچ مقدار مثبت یا منفی بی‌نهایت وجود ندارد.
مقادیر غیرعددی	0	تمام ستون‌های مورد انتظار عددی هستند.
ردیف‌های تکراری اضافه	1,081	این ردیف‌ها شناسایی شدند، اما برای جلوگیری از حذف اشتباه، نگه داشته شدند.
ستون‌های دارای داده پرت	29	داده‌های پرت گزارش شدند و هیچ‌کدام حذف نشدند.

۵. نکات مهمی که از بررسی به دست آمد

- دیتاست از نظر ساختار، کامل و برای ادامه فرایند یادگیری ماشین مناسب است.
- توزیع ستون هدف بسیار نامتوازن است: 284,315 تراکنش عادی و فقط 492 تراکنش تقلبی وجود دارد.
- Accuracy به‌تنهایی معیار مناسبی برای ارزیابی مدل این پروژه نیست، چون کلاس تقلب بسیار کم‌تعداد است.
- رکوردهای تکراری باید قبل از حذف با منطق کسب‌وکار یا منبع تولید داده بررسی شوند.
- در مسئله تشخیص تقلب، بعضی داده‌های پرت ممکن است دقیقاً همان الگوهای مهم تقلب باشند؛ بنابراین حذف خودکار آن‌ها می‌تواند اطلاعات ارزشمند را از بین ببرد.

تصمیم نهایی درباره حذف رکوردهای تکراری و نحوه برخورد با داده‌های پرت باید با هماهنگی تیم تحلیل داده و مدل‌سازی انجام شود. در این مرحله من فقط آن‌ها را شناسایی و گزارش کردم.

۶. پیشنهادهای من برای مرحله بعد

- نسخه خام دیتاست بدون تغییر نگهداری شود و داخل مخزن گیت قرار نگیرد.
- قبل از حذف ردیف‌های تکراری، علت تکرار آن‌ها بررسی شود.
- هر نوع مقیاس‌بندی، جایگزینی یا نمونه‌برداری مجدد فقط بعد از تقسیم داده به آموزش و آزمون انجام شود تا نشت داده رخ ندهد.
- برای ارزیابی مدل از معیارهایی مانند Precision، Recall، F1-score، PR-AUC، ROC-AUC و ماتریس درهم‌ریختگی استفاده شود.

۷. نتیجه‌گیری

بر اساس بررسی‌هایی که انجام دادم، دیتاست از نظر تعداد ستون‌ها، نوع داده‌ها، کامل بودن مقادیر و اعتبار ستون هدف وضعیت مناسبی دارد. هیچ داده ناقص، مقدار بی‌نهایت، مقدار غیرعددی یا مقدار منفی غیرمنطقی وجود ندارد. تنها مواردی که به تصمیم تیمی نیاز دارند، 1,081 ردیف تکراری و داده‌های پرت شناسایی‌شده در 29 ستون هستند. در این مرحله برای حفظ اطلاعات اصلی، هیچ ردیف یا مقداری را حذف یا تغییر ندادم.

این گزارش بر اساس اجرای موفق نوت‌بوک notebooks/data_quality_inspection.ipynb در برنچ preprossing تهیه شده است.