



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

پروژه
علوم کامپیوتر

تشخیص تقلب روی کردیت کاردهای بانکی اروپا

نگارش
مهدی عباسعلی پور

استاد راهنما
جناب آقای دکتر قطعی

استاد مشاور
جناب آقای یوسفی مهر

آذرماه ۱۴۰۲

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

چکیده

در این پروژه هدف بر این است تا دیتاست کردیت کارت های بانک های اروپایی در ابتدا پیش پردازش شوند و سپس با استفاده از روش کی-میانگین به عنوان الگوریتم خوشه بندی و الگوریتم رگرسیون منطقی به عنوان الگوریتم طبقه بندی برای این دیتاست مدل ساخته شود . در نهایت دو مدل ساخته شده ارزیابی و نتایج با هم مقایسه شود .

واژه‌های کلیدی:

کردیت کارت های بانک های اروپایی، خوشه بندی، طبقه بندی، رگرسیون منطقی، کی میانگین

فهرست مطالب

عنوان

صفحه

۱	معرفی داده های تراکنش های قلبی	۲
۱-۱	مقدمه	۳
۲-۱	طرح مسئله	۳
۳-۱	لینک گیت هاب کد	۴
۲	پیاده سازی	۵
۱-۲	پیش پردازش	۶
۲-۲	کی میانگین	۶
۳-۲	رگرسیون منطقی	۶
۴-۲	جمع بندی	۶
۳	نتایج و ارزیابی مدل های آموزش دیده	۷
۱-۳	خوشه بندی	۸
۲-۳	طبقه بندی	۸
۳-۳	مقایسه ی مدل ها	۹
	مراجع	۱۱

شکل	فهرست تصاویر	صفحه
۱-۱	تشخیص تراکنش های قلبی	۳
۱-۳	نمودار scatter برای خوشه بندی	۸
۲-۳	ماتریس آشفته برای kmeans	۹
۳-۳	نمودار scatter برای طبقه بند رگرسیون منطقی	۱۰
۴-۳	ماتریس آشفته برای طبقه بندی با رگرسیون منطقی	۱۰

فصل اول

معرفی داده های تراکنش های قلبی

۱-۱ مقدمه

تشخیص تراکنش های تقلبی توسط شرکت های کارت اعتباری اهمیت زیادی دارد تا از مشتریان هزینه ای برای اقلامی که خریداری نکرده اند دریافت نشود. در [۱] داده های مربوط به تراکنش های بانک های اروپایی که در سپتامبر ۲۰۱۳ جمع آوری شده است ، می باشد .



شکل ۱-۱: تشخیص تراکنش های تقلبی
[۲]

۲-۱ طرح مسئله

دیتاست شامل مجموعاً ۲۸۴۸۰۷ است که شامل ۴۹۲ مورد کلاهبرداری است . بنابراین داده ها به شدت به سمت داده های غیرتقلبی بایاس می باشد . داده ها مجموعاً شامل ۳۰ ویژگی می باشند که به دلیل محرمانگی تنها دو ویژگی زمان و مبلغ تراکنش روشن هستند و ۲۸ ویژگی دیگر به صورت v_1, v_2, \dots, v_{28} نام گذاشته شده اند . این ۲۸ ویژگی با اعمال PCA استخراج شده اند . در ادامه قصد داریم تا به وسیله ی الگوریتم های یادگیری ماشین دو مدل را به منظور تشخیص تراکنش های تقلبی با استفاده از دیتاست مذکور آموزش دهیم .

۳-۱ لینک گیت هاب کد

با مراجعه به https://github.com/mahdialipoo/Project5_AI می توانید کد مربوط به پیاده سازی پروژه را مشاهده نمایید .

فصل دوم

پیاده سازی

در این بخش پس از پیش پردازش داده ها به پیاده سازی دو مدل خوشه بندی (Kmeans) و طبقه بندی (logistic regression) می پردازیم .

۱-۲ پیش پردازش

پس از بارگیری داده ها از سایت کگل مشاهده می شود که دیتا شامل هیچ گونه ویژگی از دست رفته نمی باشد . بررسی ماتریس همبستگی هم نشان داد که وابستگی زیادی بین ویژگی ها موود نمی باشد که بتوان برخی از آن ها را حذف نمود . زمان تراکنش عاملی برای بررسی صحت تراکنش در نظر گرفته نمی شود بنابراین آن را حذف می کنیم . ویژگی ها به صورت اعداد اعشاری می باشند ۸۰ درصد داده ها برای آموزش اختصاص داده می شوند .

۲-۲ کی میانگین

پس از پیش پردازش داده ها با استفاده از کتابخانه ی sklearn مدل کی میانگین را بر روی داده ها آموزش می دهیم .

۳-۲ رگرسیون منطقی

این الگوریتم سعی دارد تا با عبور دادن ابرصفحه در فضای ورودی ها داده ها را به دو کلاس به صورت باینری تقسیم نماید .

۴-۲ جمع بندی

در این فصل مروری کلی بر مراحل آموزش و ساخت مدل (با پارامتر های مدل به صورت پیش فرض) در کتابخانه ی sklearn [۳] داشتیم .

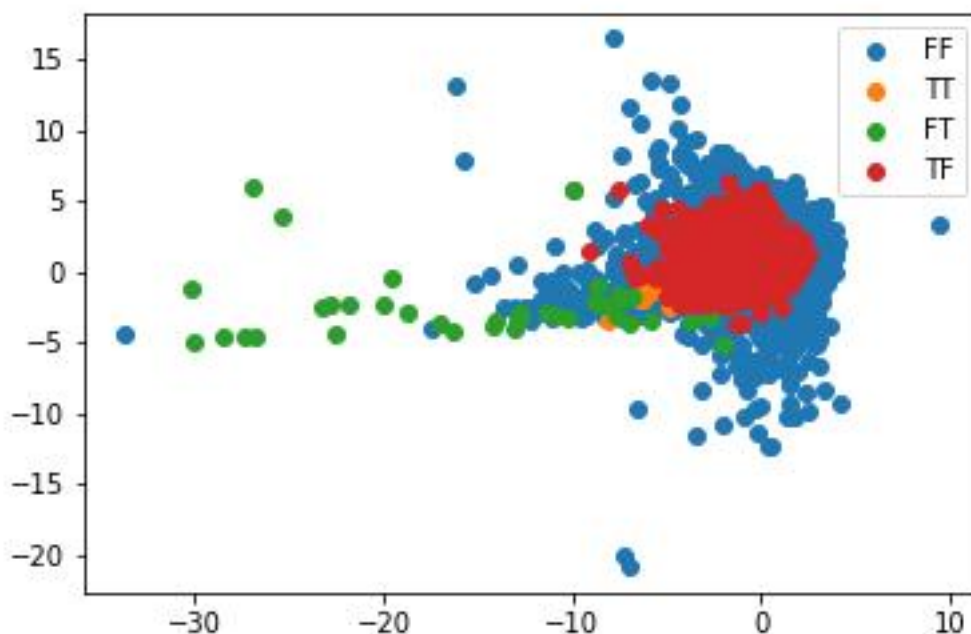
فصل سوم

نتایج و ارزیابی مدل های آموزش دیده

داده های دیتاست مذکور به شدت بایاس می باشند بنابراین پیشنهاد شده است تا به جای معیار صحت از معیار AUPRC^۱ به معنای مساحت زیر منحنی دقت - پوشش ارزیابی شود .

۱-۳ خوشه بندی

طبق معیار AUPRC امتیاز مدل بر روی داده های تست ۰۰۱۵.۰ بود . در ۱-۳ و ۲-۳ می توانید نمودار scatter برای دو ویژگی دلخواه ۳ و ۶ و همین طور ماتریس آشفتگی را مشاهده کنید .

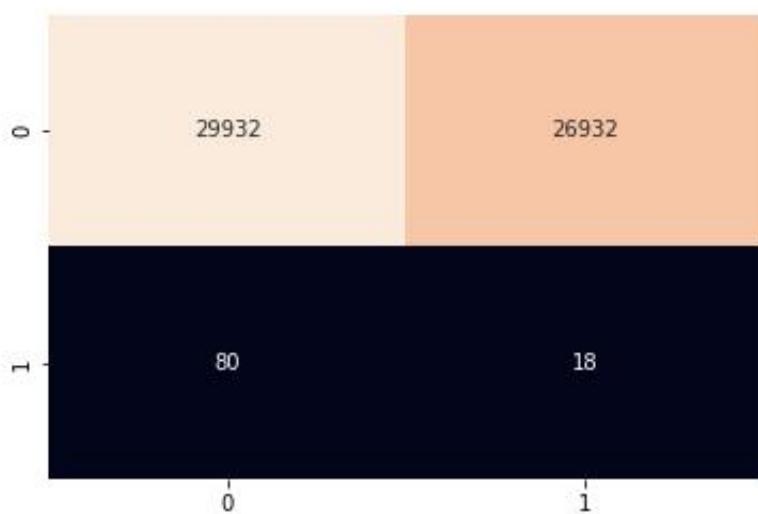


شکل ۱-۳: نمودار scatter برای خوشه بندی

۲-۳ طبقه بندی

طبق معیار AUPRC امتیاز مدل بر روی داده های تست ۹۹.۰ بود . در ۳-۳ و ۴-۳ می توانید نمودار scatter برای دو ویژگی دلخواه ۳ و ۶ و همین طور ماتریس آشفتگی را مشاهده کنید .

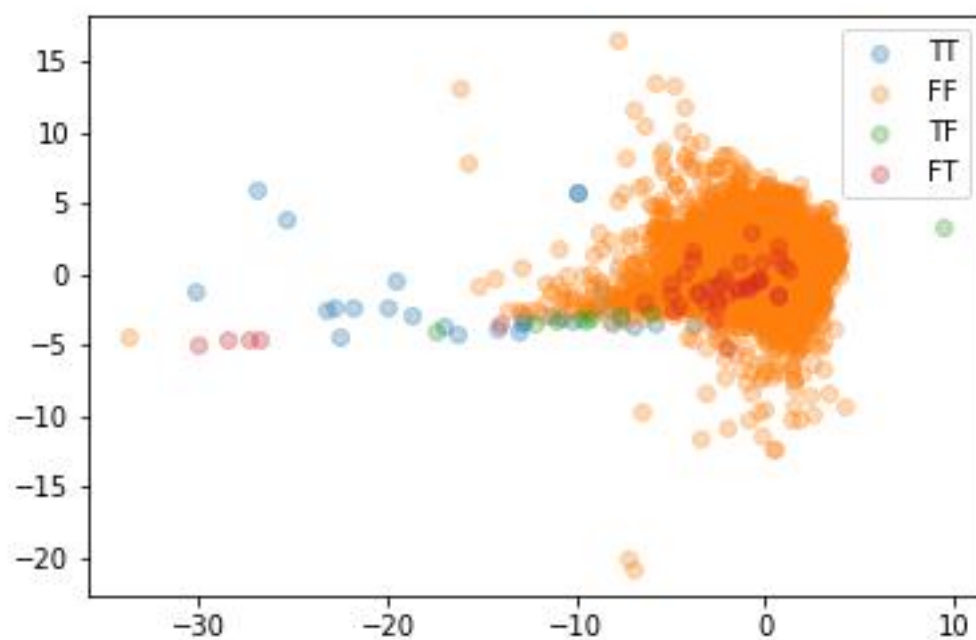
^۱Area Under the Precision-Recall Curve



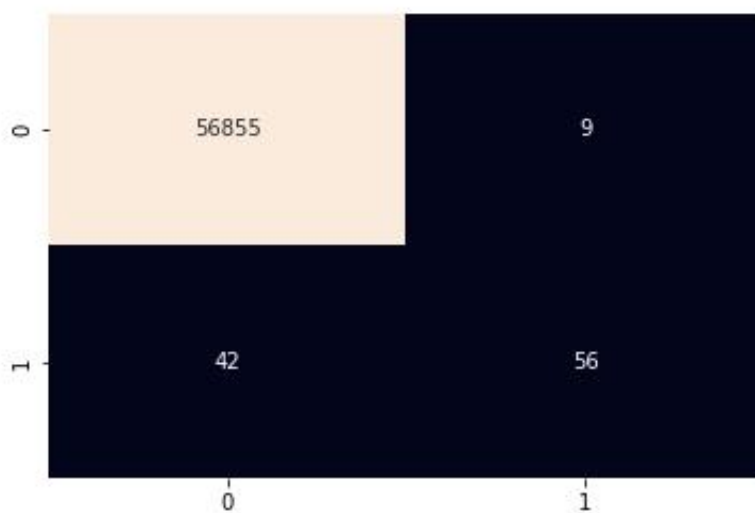
شکل ۳-۲: ماتریس آشفتگی برای kmeans

۳-۳ مقایسه ی مدل ها

باتوجه به امتیازات بدست آمده از مدل ها بر روی داده های تست ، مدل رگرسیون منطقی



شکل ۳-۳: نمودار scatter برای طبقه بند رگرسیون منطقی



شکل ۳-۴: ماتریس آشفتگی برای طبقه بندی با رگرسیون منطقی

مراجع

- [1] kaggle. Credit card fraud detection. <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>, 2013.
- [2] Matheson, Rob. Reducing false positives in credit card fraud detection. <https://news.mit.edu/2018/machine-learning-financial-credit-card-fraud-0920>, 2018.
- [3] scikit learn. scikitlearn. <https://scikit-learn.org/>, 2023.