دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک 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


بسم الله الرحمن الرحيم

چکیده

در این پروژه هدف بر این است تا دیتاست کردیت کارت های باناک های اروپایی در ابتدا پیش پردازش شوند و سپس با استفاده از روش نزدیک ترین همسایگی به عنوان الگوریتم خوشه بندی و الگوریتم بردار پشتیبان 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 کارد های بانک های اروپایی، خوشه بندی، طبقه بندی،بردار پشتیبان، نزدیک ترین همسایگی

فهرست مطالب

عنوان صفحه

فهرست تصاویر

شکل صفحه

فصل اول

معرفی داده های تراکنش های تقلبی

۱-۱ مقدمه

تشخیص تراکنش های تقلبی توسط شرکت‌های کارت اعتباری اهمیت زیادی دارد تا از مشتریان هزینه‌ای برای اقلامی که خریداری نکرده‌اند 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]

۱-۲ طرح مسئله

دیتاست شامل مجموعا ۲۸۴۸۰۷ است که شامل ۴۹۲ مورد کلاهبرداری است . بنابراین داده ها به شدت به سمت داده های غیرتقلبی بایاس می باشد . داده ها مجموعا شامل ۳۰ ویژگی می باشتد که به دلیل محرمانگی تنها دو ویژگی زمان و مبلغ تراکنش روشن هستند و ۲۸ ویژگی دیگر به صورت v1,v2,...,v28 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 بهمنظور تشخیص تراکنش های تقلبی با استفاده از دیتاست مذکور آموزش دهیم .

۱-۳ 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 (نزدیک ترین همسایگی) و طبقه بندی (بردار پشتیبان) می پردازیم .

۲-۱ 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 وییگی 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 نزدیک ترین همسایگی

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

۲-۳ بردارپشتیبان

در ابتدا داده ها را به صورت استاندارد نرمال می کنیم . برای این منظور با استفاده از کتابخانه ی sklearn یک لایه نرمال کننده را به لایه ی SVM متصل می نماییم و بعد مدل را آموزش می دهیم .

۲-۴ جمع بندی

در این فصل مروری کلی بر مراحل آموزش و ساخت مدل و تنظیم پارامتر های مدل در کتابخانه ی sklearn [۳] داشتیم .

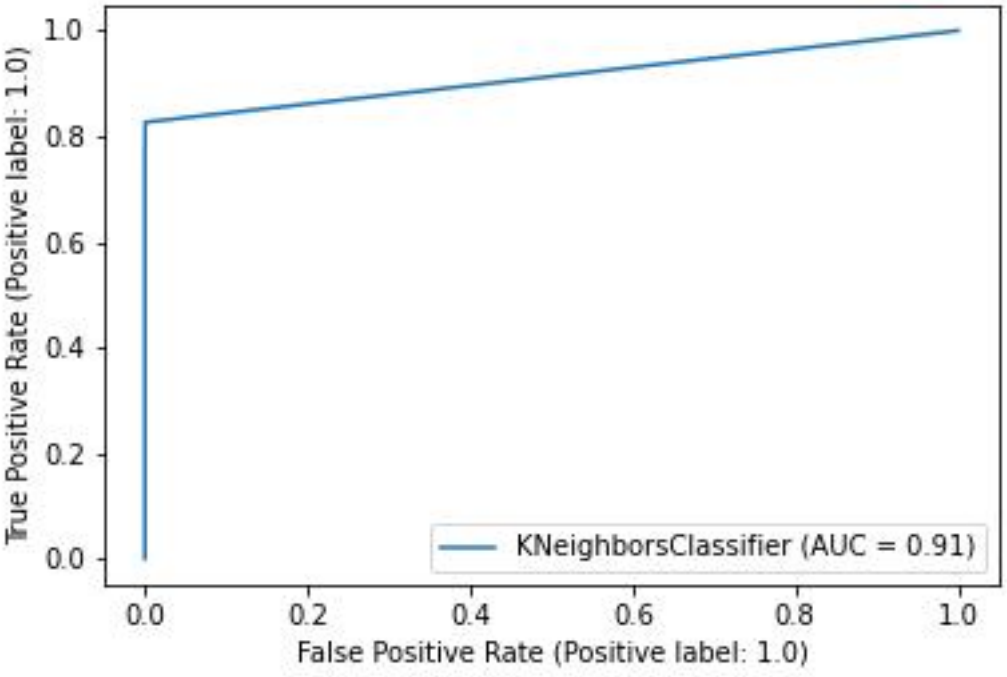
فصل سوم

نتایج و اررزیابی 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 [1] به معنای مساحت زیر منحنی دقت - پوشش ارزیابی شود .

۳-۱ مدل نزدیک ترین همسایگی

آموزش مدل نزدیک ترین همسایگی در زمان پایینی انجام پذیرفت . طبق معیار AUPRC امتیاز مدل بر روی داده های تست ۷۹.۰ بود . در ۳-۱ و ۳-۲ می توانید نمودار Precision-Recall و همین طور ماتریس آشفتگی را مشاهده کنید .

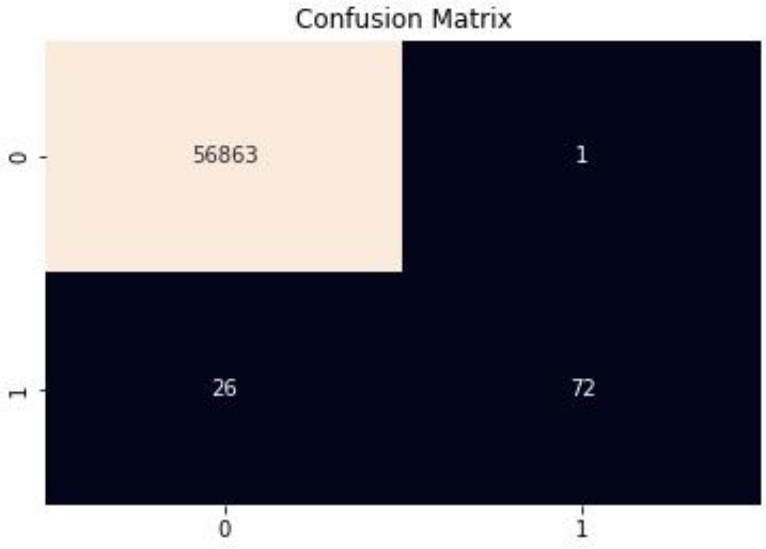


شکل ۳-۱: نمودار ROCبرای خوشه بندی نزدیک ترین همسایگی

۳-۲ مدل بردار پشتیبان

آموزش مدل بردارپشتیبان در مقایسه با نزدیک ترین همسایگی زمان بیشتری برد. طبق معیار AUPRC امتیاز مدل بر روی داده های تست ۷۱.۰ بود . در ۳-۳ و ۳-۴ می توانید نمودار Precision-Recall و همین طور ماتریس آشفتگی را مشاهده کنید .

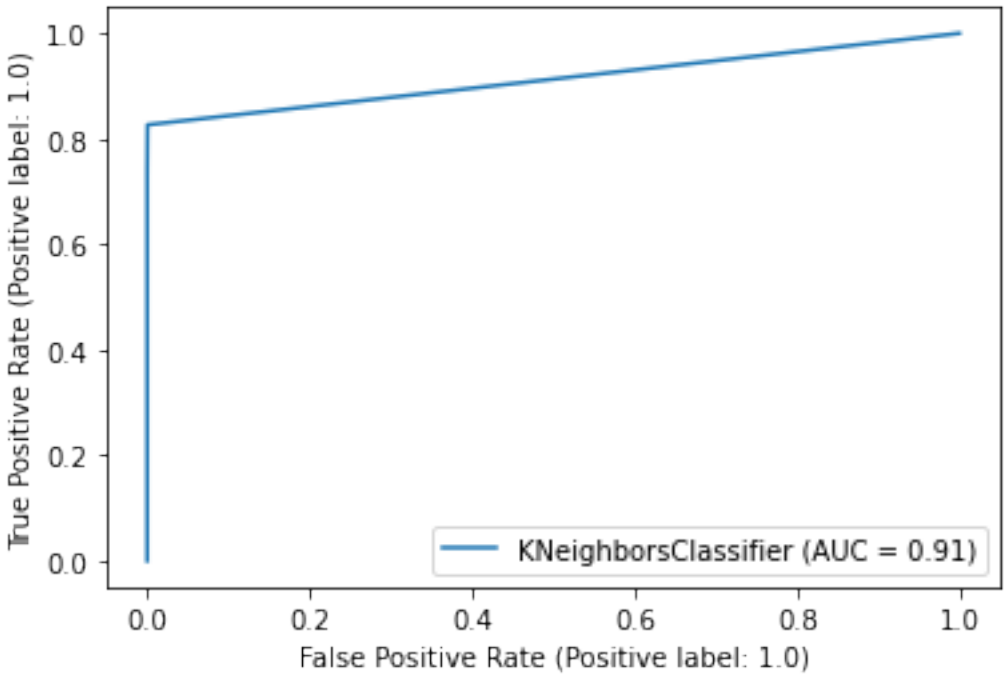
[1] Area Under the Precision-Recall Curve



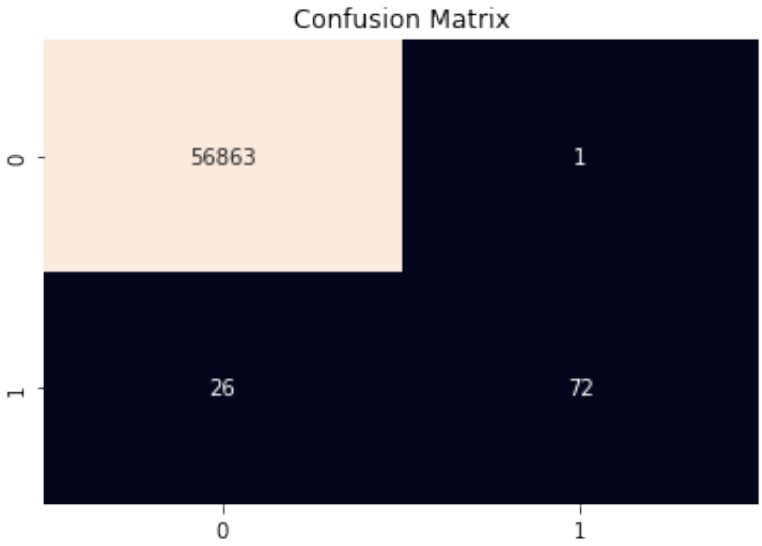
شکل ۳-۲: ماتریس آشفتگ برای نزدیکترین همسایگی

۳-۳ مقایسه ی مدل ها

باتوجه به امتیازات بدست آمده از مدل ها بر روی داده ای تست نزدیک ترین همسایگی نتایج نسبتا بهتری داشته است .



شکل ۳-۳: نمودار ROCبرای طبقه بندی بردارپشتیبان



شکل ۳-۴: ماتریس آشفتگ برای بردار پشتیبان

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.