پروژه اول درس یادگیری ماشین (پیش پردازش و EDA)

استاد : دکتر رضا رمضانی
طراح سوالات : علی معینیان

سلاااام 🧡 به اولین ماجراجویی خودتون در دنیای یادگیری ماشین خوش اومدید.
این پروژه فقط یک تمرین ساده نیست، بلکه دفترچه راهنمایی برای شماست در ماجراجویی‌های آیندتون. هرجا توی دنیای علوم داده و یادگیری ماشین بخواید کاری کنید، اولین قدم‌هاتون همین‌هایی هست که قراره ببینید.
پس عمیق یاد بگیرید، از مطالبی که داخل کارگاه بیان شده استفاده کنید و لذت ببرید.
اگه هم جایی رو مشکل داشتید، مثل همیشه در تلگرام بهم پیام بدید تا راهنماییتون کنم. 😀

خب داستان ما از یه پست داخل لینکدین شروع میشه از آقای Mark Eltsefon که Staff Data Scientist هستند در کمپانی Meta و قبلا هم همین پوزیشن کاری رو در تیک تاک داشتند. ایشون میگن که توی 5 سال اول کارم مهم ترین درس هایی که گرفتم این بوده:

اول دیتا و بعد مدل

قبل از اینکه توی انتخاب مدل و مدل‌سازی و آموزش مدلت غرق بشی حواست باشه که اصلا داده‌هایی که در پروژه داری درست و تمیز هستند یا نه ؟

انتخاب متریک درست

هر مدلی که بزنی، بدون داشتن متریک‌های درست هیچ ارزشی نداره. یه مدل ممکنه توی فکرت یا روی کاغذ خیلی عالی به نظر برسه اما اگر به متریک‌هایی که برای سنجیدنش نیاز داره فکر کنی میبینی که همیشه نباید Accuracy رو در نظر بگیری و داستان پیچیده تر از این حرف هاست.

همیشه کار رو با ساده ترین مدل شروع کن

قبل از اینکه کارت رو بخوای پیچیده و سنگین کنی و با جدیدترین مدل‌ها کار کنی، یه Baseline ساده بزن. لازم نیست از همون اول مدل‌های خیلی پیچیده رو بری سراغشون. خیلی وقت ها راه حل‌های کلاسیک در سریع ترین زمان و با کمترین هزینه تورو به نتیجه مطلوبت میرسونند.

تحلیل اکتشافی داده ها یا همون EDA واجب ترین قسمته

شما میخوای بری مسافرت بدون اطلاع از مقصد و هتل و جاهایی که میخوای بری اصلا ممکنه بتونی مسافرت خوبی رو تجربه کنی؟ فعععکککککر نکنم 😉
دنیای علوم داده هم همینطوره! این دنیا بر اساس داده‌هایی که دارید بنا شده و شما باید قبل از هر کاری، یک درک عمیق و دقیق از داده‌هاتون داشته باشید. EDA دقیقا همون مرحله‌ای محسوب میشه که زیاد قراره باهاش سروکار داشته باشید.

و اما مهم ترین نکته در این پروژه : هر کاری که قراره انجام بدید، نیاز به تحلیل داره ! اینکه صرفا دو تا نمودار و دو تا عدد رو در نتیجه ارسال کنید به هیچ عنوان قابل قبول نیست. در اصل قراره با هم تمرین کنیم تا بتونیم داده‌ها رو بفهمیم و تحلیلشون کنیم.

قسمت اول: سوالات تشریحی

یک: شما به عنوان دانشمند داده در تیم آنالیز پلتفرم نماوا در حال کار هستید. وظیفه اصلی شما در پروژه اولتون طراحی یک سیستم پیشنهاد دهنده‌ی فیلم هست. توضیح بدید که برای انجام این پروژه چگونه از چرخه حیات علم داده (یادگیری ماشین) استفاده می‌کنید؟ (اگر شکلش رو هم بکشی دعات میکنیم 😁)

دو: بسیاری از دیتاست‌هایی که با آنها کار می‌کنیم، دیتاست‌های تمیز شده و بدون چالش هستند و چالش زیادی برای بهتر شدن کیفیت آن دیتاست وجود ندارد. دیتاست‌های آماده و تمیز شده را با دیتاست‌های واقعی که پر از داده‌های گم شده و ... هستند مقایسه کنید و در یک جدول این تفاوت‌ها را بررسی کنید. در Kaggle دو دیتاست با این ویژگی پیدا کنید، یکی تمیز شده و بدون چالش و دیگری یک دیتاست واقعی پر از چالش

سه: از معروف ترین دیتاست‌های زبان فارسی که کارهای مختلفی روی آن انجام شده است، دیتاست کامنت‌ها و محصولات دیجی کالاست :

https://www.kaggle.com/datasets/radeai/digikala-comments-and-products/data?select=digikala-comments.csv

وارد لینک بالا شوید و دیتاست را فقط بررسی کنید (اصلا نیاز نیست چیزی دانلود کنید). در قسمت Column ببینید چه ستون‌هایی وجود دارد. با اینکه منبع این دیتاست خود دیجی کالاست، اما به نظر شما با واقعیت دیتاهای دنیای واقعی تشابهی دارد؟ با پارامترهایی که خود Kaggle برای شما فراهم کرده، این دیتاست را تحلیل کنید.

چهار: فرض کنید شما یک ویژگی Nominal با کاردینالیتی بالا (High Cardinality) دارید. مثلا کد پستی کل افراد ساکن اصفهان که همگی مقدار منحصر به فردی دارند.

- چرا استفاده از Label Enconding برای این ویژگی اشتباه است و چه اطلاعات غلطی را به مدل خواهد داد؟
- به نظر شما میتوانیم از One-Hot Encoding استفاده کنیم؟
- چه ترکیبی از روش های Feature Engineering و Dimensionality Reduction را پیشنهاد می‌کنید تا اطلاعات این ویژگی تا حد امکان حفظ شود، اما ابعاد مدل قابل مدیریت باقی بماند؟

پنج: دو روش Z-Score و IQR را در نظر داشته باشید. اگر در یک دیتاست فرضی، ویژگی داشته باشیم با چولگی شدید، کدام روش برای شناسایی داده های پرت بهتر عمل میکند و مقاوم تر است؟
سناریوهای مختلف ممکن را بررسی کنید و در پاسخ بنویسید.

شش: دو روش محبوب Min-Max Scaling و Z-Score را در نظر بگیرید. کدام یک از این دو روش نسبت به داده‌های پرت یا همان Outlier حساسیت بیشتری دارند ؟چرا ؟ (به فرمول هر کدام دقت کنید)

هفت: فرض کنید یک مجموعه داده بسیار نامتوازن (Imbalanced) درباره تشخیص کالاهای تقلبی دارید که 98% سالم و 2% تقلبی هستند.
اگر بخواهید یک نمونه کوچک برای تست مدل جدا کنید، چه مشکلاتی ممکن است پیش بیاید؟
چرا استفاده از Random Sampling ممکن است منجر به ساخت مجموعه تست بلا استفاده شود؟ چه راه حل هایی موجود است؟

هشت: تفاوت اساسی بین PCA و LDA در کاهش ابعاد چیست؟ چرا PCA یک روش بدون نظارت است اما LDA با نظارت؟

نه: تفاوت‌های اساسی بین Feature Selection و Feature Extraction چیست؟

ده: دو روش Box-Cox Transformation و Min-Max Scaling را در نظر بگیرید. تفاوت اساسی در هدف این دو روش چیست؟
سناریویی را توصیف کنید که در آن، استفاده از Min-Max Scaling به تنهایی بر روی یک ویژگی با چولگی شدید منجر به عملکرد ضعیف مدلی مانند KNN می‌شود ؟
چرا اعمال Cox-Box می‌تواند بهتر باشد؟ (به نحوه اثرگذاری این دو تبدیل بر تراکم داده ها و محاسبه فاصله فکر کنید)

قسمت دوم : سوالات عملی

یک: هدف من از این قسمت آشنایی شما با نحوه نگارش نوت بوک های EDA و تحلیل دیتاهاست.
به نوت بوک Spaceship Titanic مراجعه کنید:
https://www.kaggle.com/code/shadesh/eda-data-preprocessing-and-ml-model-training

دیتاست استفاده شده در این نوت‌بوک از معروف‌ترین دیتاست های موجود در دنیای علوم داده است. همانند کارگاه، شروع به تحلیل این نوت‌بوک کنید و تحلیل‌های خود را بنویسید (فقط نیاز به تحلیل داریم - نیم نگاهی به کدها هم داشته باشید)

دو: لینک زیر، دیتاست خودروهاست:
https://www.kaggle.com/datasets/CooperUnion/cardataset

اول، بررسی کوتاهی از دیتاست موجود داشته باشید (همانند کارگاه).
به ضمیمه نوت بوکی آماده کرده‌ام برای تحلیل این دیتاست به نام Car_EDA که در اختیار شما قرار خواهد گرفت. شما باید با اجرای این نوت بوک و ایجاد تغییرات مورد نیاز (مثل وارد کردن دیتاست و جاهایی که کامنت قرار داده شده) خروجی ها را به دست آورده و تحلیل کامل آن را در سلول مارک داون پایین هر خروجی بنویسید (یا در فایل داکیومنت خود)
هدف از این تمرین آشنایی عمیق‌تر شما با کدها و برخی چالش‌های تحلیلی است.

سه: بیماری های قلبی-عروقی دسته‌ای از بیماری‌های هستند که در آنها قلب یا رگ‌های خونی مرتبط با قلب درگیر هستند. این بیماری اصلی‌ترین عامل مرگ و میر در سراسر جهان است.
شما به عنوان دانشمند داده بیمارستان قلب شهید چمران باید بتوانید تحلیل کاملی از دیتاست CVD که در فایل‌های ضمیمه در اختیار شماست داشته باشید (تمرکز ویژه‌ای روی Feature Engineering داشته باشید).
نوت بوک شما باید دارای این قسمت ها باشد :
Import libraries - Check the Shape of the Dataset - Preview of Dataset - Dataset Description - Data Types - Statistical properties of Dataset- Univarient Analysis - Analysis of Target feature variable - Bivariate analysis - Estimate correlation coefficients - Multi varient analysis - Dealing with Missing values - Outlier detection

چهار: تا به اینجا مرحله به مرحله با هم پیش رفتیم و به درک خیلی خوبی از EDA رسیدیم.
در تمارین قبلی سعی شده بود تا به شما در حل تمرین کمک شود. اما در آخرین قسمت، همه چیز با شماست. از خواندن دیتاست تا شناخت متغیرهای آن و کار و نتیجه گیری.

شما به عنوان دانشمند داده در شرکت آمازون آمریکا مشغول به کار هستید. دیتاست Amazon برای شما آماده شده است. این دیتاست شامل داده‌های محصولات آمازون از نظر امتیاز و نظرات است. وظیفه شما کاوش در این داده هاست. گام به گام، داده‌ها را پاک سازی کرده، از صحت و سازگاری آنها مطمئن شوید. سپس داده‌ها را با استفاده از آمار توصیفی تحلیل کنید و با استفاده از نمودارها به تجسم و تحلیل ابعاد مختلف این دیتاست و کشف الگوها بپردازید. راستی Outlier و ... فراموش نشه.

راستی برای مطالعه بیشتر اگر دوست داشتید میتوانید سراغ ابزارهایی بروید که فرایند EDA را ساده تر از قبل کرده‌اند. لینک هایی برای مطالعه بیشتر در این قسمت ارائه شده است:

https://github.com/cmudig/AutoProfiler?tab=readme-ov-file
https://github.com/exploripy/exploripy
https://pypi.org/project/sweetviz/
https://pypi.org/project/quickda/