گزارش کد:

در ابندا که کتابخانه ها را قرار میدهیم و به درایو وصل میشویم تا دیتاست ر لود کنیم.

دیتا را از نظر اینکه خالی دارد در جایی یا خیر چک میکنیم که ندارد.

برای داده های پرت میتوانیم از نمودار جعبه ای استفاده کنیم و از انجایی که مفدار فیچر ها زیاد است به کشیدن یکی بسنده کردیم ولی با استفاده از امکانات نرم افزار اکسل مینیمم و ماکسیمم هارا بررسی کردیم و همه فیچر ها از این نظر مشکلی نداشتند.

حالا باید داده های کتگوریک را به عدد تبدل کنیم. برای grade و Employment Duration که ترتیب دارند با مپ کردن به اعداد این کار را میکنیم. برای انهایی که دو حالت دارند .astype('category').cat.codes میکنیم . برای Verification Status از one hot انکدینگ استفاده میکنیم و بقیه را هم چون خیلی اهمیت ندارند حرف میکنیم.

حالا با استفاده از پیدا کردن همبستگی یک سری دیگه از فیچر هارای کم اهمیت را حذف میکنیم.

پس از مرتب شدن دیتاست باید مفادیر را نرمالایز کنیم پس از standard scaler استفاده میکنیم.

مقدار ستون مورد پیش بیتی با جواب ها مثبت و منفی ر محاسبه مکنیم تا از وضعیت دیتاست مطلع باشیم. با توجه به مقدار بدست امده دیتاست به شدت نامتعادل است.

برای رفع این مشکل داده را متغادل میکنیم و از هر دو قسمت 6000 تا بر میداریم و شافل میکنیم و دیتاست جدیدی اجاد میکنم.

داده های تست و اموزش را درست میکنیم.

با سه الگوریتم مدل را اموزش داده ایم و همچنین برای هر الگوریتم این اموزش را برای هم کل داده ها و هم داده های متعادل شده انجام دادیم.

چون در سوال گفته نشده از چند الگوریتم استفاده کنید ماهم تنها یه شرح یکی از انها می پردازیم و یا یکی هم منحنی یادگیری میکشیم.

Logistic Regression:

رگرسیون لجستیک یکی از محبوب‌ترین الگوریتم‌های یادگیری ماشینی است که تحت تکنیک یادگیری نظارتی قرار می‌گیرد. برای پیش‌بینی متغیر وابسته طبقه‌بندی با استفاده از مجموعه معینی از متغیرهای مستقل استفاده می‌شود.

رگرسیون لجستیک خروجی یک متغیر وابسته مقوله ای را پیش بینی می کند. بنابراین نتیجه باید یک مقدار مقوله ای یا گسسته باشد. می تواند بله یا خیر، 0 یا 1، درست یا نادرست و غیره باشد، اما به جای دادن مقدار دقیق 0 و 1، مقادیر احتمالی را می دهد که بین 0 و 1 قرار دارند.

رگرسیون لجستیک بسیار شبیه به رگرسیون خطی است به جز نحوه استفاده از آنها. از رگرسیون خطی برای حل مسائل رگرسیون استفاده می شود، در حالی که از رگرسیون لجستیک برای حل مسائل طبقه بندی استفاده می شود.

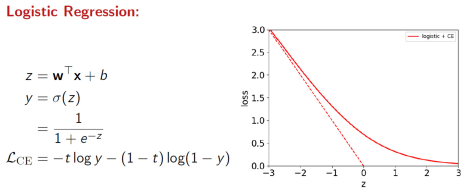
در رگرسیون لجستیک، به جای برازش یک خط رگرسیون، یک تابع لجستیک به شکل "S" را قرار می دهیم که دو مقدار حداکثر (0 یا 1) را پیش بینی می کند.

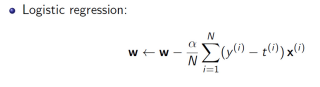
تابع سیگموئید یک تابع ریاضی است که برای ترسیم مقادیر پیش بینی شده به احتمالات استفاده می شود.

هر مقدار واقعی را به مقدار دیگری در محدوده 0 و 1 نگاشت می کند.

مقدار رگرسیون لجستیک باید بین 0 و 1 باشد که نمی تواند از این حد فراتر رود، بنابراین منحنی مانند فرم "S" را تشکیل می دهد. منحنی شکل S را تابع سیگموئید یا تابع لجستیک می نامند.

در رگرسیون لجستیک، از مفهوم مقدار آستانه استفاده می کنیم که احتمال 0 یا 1 را تعریف می کند. مانند مقادیر بالای مقدار آستانه به 1 میل می کند و مقدار زیر مقادیر آستانه به 0 میل می کند.





برای منحنی یادگیری هم از learning\_curve استفده میکنیم و میانگین و احراف معیار را برای داده های تست و اموزش محاسبه میکنیم تا با استفده از انها fill\_between را ایجاد کنیم.

تفسیر نمودار در کد امده است. (سوال دوم)

سوال اول:

از انجا که دیتاست نا متعادل است استفاده از accuracy نتیجه یالایی خواهد داد اما این نتیجه کاذب است چون ضعیف ترین مدل هم که شانسی 0 و 1 بدهد بالاخره تعداد زیادی از داده ها را درست جواب میدهد. بنابر این از roc\_auc استفاده میکنیم که برای دده های نامتعدلی است که اهمیت 0 یا 1 شدن برای مان یکسان است. با اینحال سه شاخص ارزیابی را برای همه مدل ها استفاده کردیم و انچه توصیه میشود را مشخص کردیم.