**گزارش بخش سه تمرین دوم**

**محمد ویس مصطفی پور 97222085**

**پردازش داده ها:** ابتدا تهی بودن یا نبودن خانه ها را چک میکنیم، داده تهی نداریم. سپس چک میکنیم که داده های ها غیرعددی داریم و پی میبریم که خیر نداریم. پس سراغ اسکیل کردن داده و بعد جدا کردن بخش تست و ترین می رویم.

**سوال اول:** از طریق پکیج sklearn یک رگرسیون لاجستیک میسازیم . تارگت آن را محدوده قیمت و ویژگی های دیگر آن را ورودی آن قرار میدهیم. و آن را روی مدل فیت میکنیم.

Precision برابر 0.67

Recall برابر با 0.65

F1-score برابر با 0.71 می باشد

**سوال دوم:** visualize را انجام میدهیم و میبینیم که کلاس های ما کاملا متوازن هستند و هر کدام برابر با 500 تا هستند.

**سوال سوم:** پس از انجام اینکار میبینیم که تعداد داده های کلاس صفر برابر 500 و تعداد داده های کلاس یک برابر 1500 میشود.

**سوال چهارم:** ابتدا با کتابخانه sklearn مدل logistic regression را با داده های که از سوال قبلی به دست آمدند ترین میکنیم و امتیاز های زیر بدست می آید.

برای کلاس صفر :

Precision برابر 0.65

Recall برابر با 0.69

F1-score برابر با 0.68 می باشد

برای کلاس یک:

Precision برابر 0.71

Recall برابر با 0.7

F1-score برابر با 0.69 می باشد

**سوال پنجم:**

وقتی که داده ها بالانس نباشند مدل ترین شده به سمت داده های اکثریت بایاس پیدا میکند. همچنین امکان generalize کردن در مدل پایین می آید و نمونه های مختلف را نمیتواند تشیص دهد. راه حل های مختلفی برای حل نامتوازن بودن داده ها وجود دارد که چند تا را توضیح میدهیم:

1- ساختن داده های جدید از روی داده های قبلی بوسیله شبکه ی GAN
2- حذف کردن داده های اکثریت تا جایی که دیتاست متوازن شود یا به اصطلاح Undersampling
3- اضافه کردن داده های جدید به کلاس اقلیت با استفاده از روشِ SMOTE
4- کپی کردن داده های کلاس اقلیت و اضافه کردن دوباره آن به دیتا ست upsampling

ما از روش چهارم یا همان آپ سمپلینگ برای رفع این مشکل استفاده میکنیم

**سوال ششم:** از روش forward selection استفاده میکنیم و best feature ها را استخراج می کنیم و به فیچر های زیر میرسیم:

**“FC” , “Int_memory” , “Ram”, “N_dep”, “Mobile_wt”, “dual_sim”**

**سوال هفتم:** لاجستیک رگرشن را روی فیچر های بدست آمده از سوال قبلی اعمال میکنیم و به امتیازهای زیر میرسیم:

Precision برابر 0.68

Recall برابر با 0.69

F1-score برابر با 0.66 می باشد

**سوال هشتم و نهم:** این کار را انجام میدهیم و به pca مقدار شش میدهیم

پس از آنکه x های مورد نظر را توسط Pca پیدا کردیم با استفاده از ان ها مدل خود را فیت میکینم، نتایج نهایی دقت پایین تری دارند زیرا pca به صورت رندوم نتایج را انتخاب میکند

**سوال دهم:** تابع بکوارد سلکشن را پیاده سازی میکنیم. و ویژگی های های مهم را بدست می آوریم سپس لاجستیک رگرشن را روی آن پیاده سازی می کنیم و نتایج زیر بدست می آید که میبینم دقت چندان بالایی ندارد.

Precision برابر 0.33

Recall برابر با 0.35

F1-score برابر با 0.31 می باشد

**سوال یازده:** در پایان هم برای سوال ،11خواسته شده که fold-kرا پیاده سازی کنیم. مانند بخش 1عمل میکنیم و مدل لاجستیک خود را با 5-foldو 10-foldبرروی تمامی فیچرها اعمال کردهایم و accuracyرا ن ری مشاهده میکنیم.