بخش دو تمرین دو

در این تمرین با کمک ۳ فیچری که در صورت سوال گفته شد، الگوریتم رگرسیون خطی را اجرا می کنیم. چون فیچر Heating type کتگوریکال است، از one hot encoding استفاده می کنیم که منجر می شود تعداد فیچر های ورودی به الگوریتم برابر ۱۵ بشود. همچنین فیچر Living space نیز به عنوان لیبل ها در نظر گرفته می شود.

داده ها را به نسبت ۸ به ۲ برای ترین و تست تقسیم می کنیم و قبل از این که داده ها را به الگوریتم بدهیم، داده ها را با کمک standard scaler نرمال سازی می کنیم.

مقدار mse بدون استفاده از پکیج (استفاده از الگوریتم پیاده سازی شده):

```
MSE on the test set: 1125.6715078175303
```

مقدار mse با استفاده از پکیج:

```
MSE on the test set: 1125.7051500327655
```

تفاوت کم بین مقدار mse الگوریتم پیاده سازی شده و پکیج استفاده شده نشان دهنده این است که الگوریتم به خوبی پیاده سازی شده است.

همان فیچر هایی که به مدل رگرسیون خطی به عنوان ورودی داده شده بود را به عنوان ورودی به Ridge regression و Lasso regression نیز می دهیم.

مقدار mse با استفاده از پکیج Ridge regression :

```
10 Fold:Mean MSE: 87382.327
```

مقدار mse با استفاده از پکیج Lasso regression :

```
10 Fold:Mean MSE: 86213.890
```

رگرسیون ridge و lasso به دلیل وجود ترم محدود کننده در شرط بهینه سازی اجازه نمی دهند که وزن ها از حدی بیشتر شوند. به این دلیل روی داده های غیر خطی خطا بالا می رود و مقدار وزن ها از حدی بیشتر نمی شود که این منجر به افزایش مقدار mse می شود و در این دیتاست مدل underfit شده است. اما مزیت مدل رگرسیون ridge و lasso به تفسیرپذیری بالای آن ها می باشد.

در واقع وجود ترم محدود کننده در شرط بهینه سازی موجب این می شود که خیلی از فیچر ها صفر می شوند و می تواند فیچر های خوب مدل را به ما نشان دهد. به طور کلی یک trade off بین پیچیدگی مدل و تفسیرپذیری مدل وجود دارد و هرچقدر مدل پیچیده تر باشد معمولا خطای کمتری دارد و مدل ها با تفسیرپذیری بالا خطای زیادی دارند.

امیرحسین باباجانی ۹۷۲۲۲۰۰۹