نام ونام خانوادگی : سارا رضایی
شماره دانشجویی: ۹۸۲۲۲۰۴۳
عنوان : گزارش و تحلیل دیتاست دوم تمرین ۲ درس مبانی یادگیری ماشین

مقدمه

داده های مورد بررسی در این تمرین داده های یکی از بزرگترین آژانس های املاک آلمان است. این دیتاست فقط شامل خانه های اجاره ای است.

مجموعه داده شامل بسیاری از ویژگی های مهم، مانند اندازه منطقه زندگی، اجاره، هر دو اجاره پایه و همچنین اجاره کل ، مکان (خیابان و شماره خانه ، کد پستی ، ایالت ، منطقه و ...)، نوع انرژی وفیچر های دیگر می باشد.

در این دیتاست تلاش برای مدلسازی برای پیش بینی قیمت اجاره با کمک فیچر های انتخاب شده و مدل های linear regression هم با پکیج و هم پیاده سازی from scratch انجام خواهد شد. در آخر با کمک پکیج مدل های رگرسیون lasso , ridge نیز مدل سازی انجام شد.

متد ها و مدل ها

از بلاک اول تا قبل از هدینگ تسک ۱ لود و بارگذاری داده ها از کگل انجام شده و shape و head و info دیتافریم برای مشاهده ی کلی داده ها چاپ شد. همچنین درصد داده های null در هر ستون محاسبه شده است و به ترتیب صعودی نمایش داده شده است و ستون هایی که درصد داده های null در آنها بیشتر از ۳۰درصد است از دیتافریم حذف شده اند.

در قسمت بعدی سطر هایی که اطلاعات خانه ها بدون اجازه نهایی ثبت شده اند از دیتافرم حذف شده اند زیرا در مدلسازی نمیتوان از آنها استفاده کرد.

پس از آن ستون هایی که شامل داده هایی هستند که اطلاعات غیرمفید یا غیرقابل استفاده دارند(مثلا داده هایی که نیاز به متن کاوی دارند) حدف شده اند. و سطر های تکراری(dublicated) شناسایی و حذف شده اند.

درقسمت بعدی داده هایی که null باقی مانده اند به این ترتیب پر شده اند: داده های عددی با میانگین داده های ستون خودشان پر شدند. و داده های categorical با پرتکرار ترین داده ی ستون خودشان پر شدند.

در نهایت داده های پرت در ستون داده های عددی با تکنیک IQR شناسایی و حذف شدند.

تسک اول

پیاده سازی یک مدل رگرسیون خطی بدون استفاده از پکیج

در این قسمت با سه فیچر 'telekomUploadSpeed', 'serviceCharge' ,'heatingType' مدلسازی انجام شده است.

ابتدا فیچر 'heatingType' یک فیچر categorical است با کمک پکیج ordinal encoder کدگذاری شده است و فیچر های 'telekomUploadSpeed', 'serviceCharge' که داده ی عددی هستند با کمک پکیج standard scaler هم مقیاس شده اند.

در نهایت مقادیر X_train,X_test,y_train,y_test جدا شده اند.

در قسمت بعدی پیاده سازی رگرسیون خطی را داریم که در آن گرادیان به این شکل آپدیت میشود که ابتدا با یک w شروع میشود و هربار برای کاهش cost function این w آپدیت میشود.

همچنین همانطور که در تمرین خواسته شده تابع mean square error نیز بدون استفاده از پکیج پیاده سازی شده است.

در قسمت اول با train کردن مدل رگرسیون خطی ای که بدون استفاده از پکیج پیاده سازی شده است مقدار **MSE = 67288.178 و MAE = 196.39** میباشد.

تسک دوم

پیاده سازی یک مدل رگرسیون خطی با استفاده از پکیج sklearn

با همان ۳ فیچر قبلی مدل آموزش داده شد و مقدار **MSE = 67288.179 و MAE = 196.39** میباشد. که بسیار نزدیک و تقریبا مساوی با حالت بدون استفاده از پکیج میباشد.

تسک سوم

پیاده سازی ۲ مدل رگرسیون lasso, ridge

در این حالت هم با توجه به اینکه از همان ۳ فیچر قبلی استفاده شد مقادیر mse , mae بسیار نزدیک به حالات قبلی میباشد.