گزارش تمرین اول مبانی یادگیری ماشین

محمد ویس مصطفی پور 97222085

صورت مساله داده شده: در این تمرین به ما داده های حدود دویست و شصت هزار خانه داده شده و قرار است مدل هوش مصنوعی پیاده سازی کنیم که بر اساس ویژگی های مختلف داده شده، متراژ اتاق پذیرایی یا به اصطلاح living space را پیش بینی کنیم.

داده های مساله: به ما داده های 268850 خانه مختلف در آلمان داده شده است و برای هر کدام از خانه ها 49 فیچر تعریف شده است. که البته ما قرار از 48 تای آن استفاده کنیم و در نهایت living space را بدست آوریم

	regio1	serviceCharge	heatingType	telekomTvOffer	telekomHybridUploadSpeed	newlyCor
0	Nordrhein_Westfalen	245.00	central_heating	ONE_YEAR_FREE	NaN	False
1	Rheinland_Pfalz	134.00	self_contained_central_heating	ONE_YEAR_FREE	NaN	False
2	Sachsen	255.00	floor_heating	ONE_YEAR_FREE	10.0	True
3	Sachsen	58.15	district_heating	ONE_YEAR_FREE	NaN	False
4	Bremen	138.00	self_contained_central_heating	NaN	NaN	False

سوالات تمرین: در این بخش لازم است پیش پردازش لازم بر روی داده ها را انجام دهیم و سپس و feature engineering های لازم را اعمال میکنیم. سپس به سوالات داده شده در تمرین پاسخ میدهیم.

پیش پردازش داده ها:

1- **مدیریت null :** فیچرهایی که بیشتر از 50 درصد آن ها null باشد را حذف میکنیم. سپس داده های تهی باقی مانده را اگر داده عددی باشن با میانگین اعداد آن فیچر پر می کنیم در نهایت همه داده های عددی را نرمالایز و سپس بین صفر و یک اسکیل سازی میکنیم

2- **حذف ویژگی های بدردنخور:** تعدادی از ویژگی ها مانند تاریخ، شماره خانه یا ویژگی دیگری مثل "توضیحات" نه تنها راهگشا نیستند بلکه بدتر مدل را سردرگم میکنن، پس آن ها را حذف میکنیم

3- **کوچک سازی ویژگی های غیرعددی:** ویژگی های کتگوریکالی را که نیاز به کوچک سازی دارند به چند بخش اصلی و یک بخش Other تقسیم میکنیم

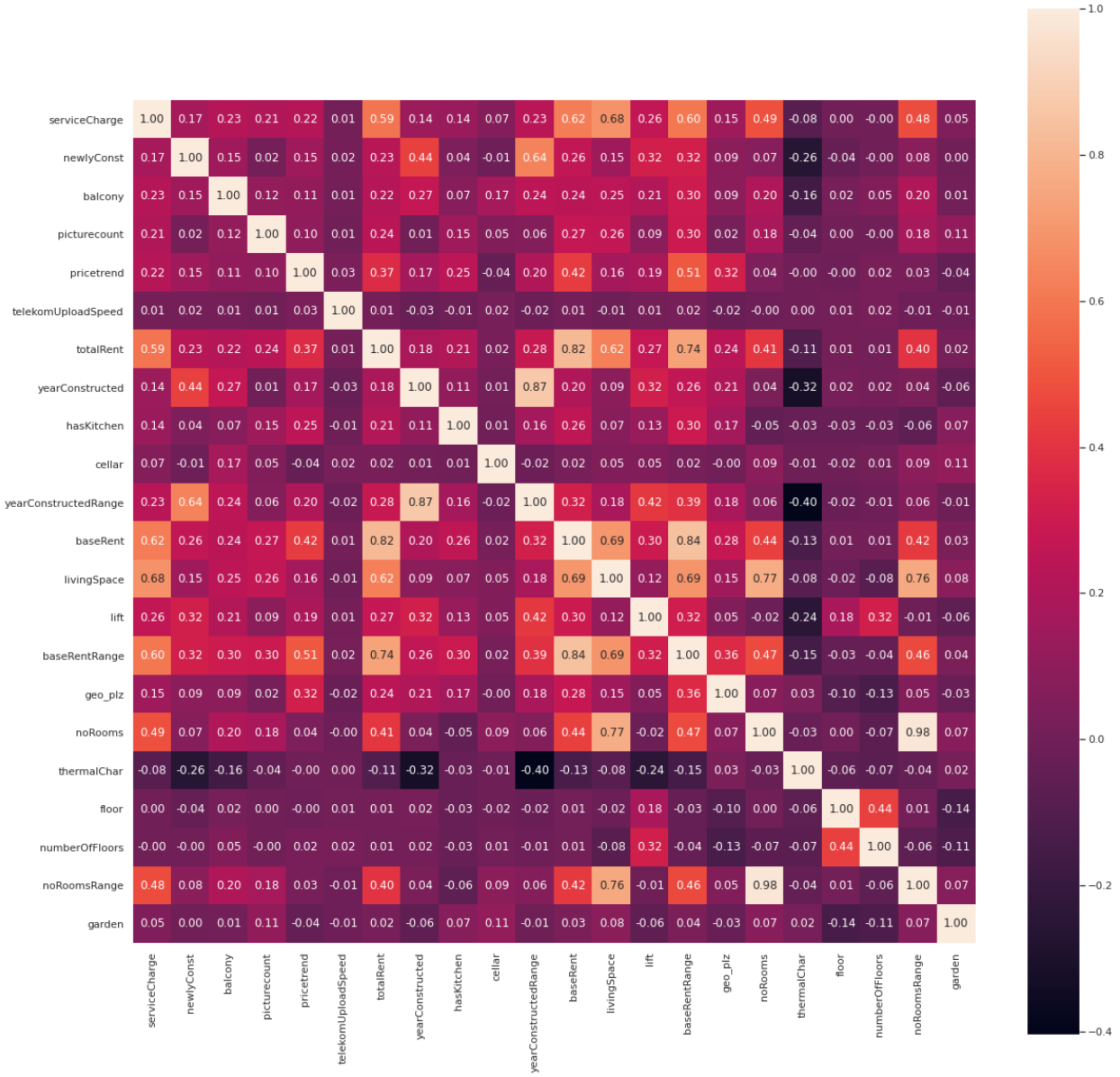
4- **مدیریت null بخش کتگوریکال:** داده های تهی کتگوریکال را با مد داده ها پر میکنیم

5- **One hot encoding داده ها:** داده های کتگوریکال را وان هات انکودینگ میکنیم

6- **جداکردن داده های test & train:** داده ها را به دو بخش ترین (نود درصد) و تست (10 درصد) تقسیم می کنیم.

بخش اول: کدام فیچرها بیشترین کورولیشن خطی را با هدف دارا میباشند؟

نمودار کورولیشن آن را رسم می کنیم و بررسی میکنیم که کدام ویژگی بیشترین کورولیشن خطی را با ویژگی مساحت داراست. نمودار زیر را ببینید:



خب همانطور که در نمودار هم پیداست دو ویژگی BaseRoom و BaseRent بیشترین کورولوشین خطی را با مساحت دارند.

پیاده سازی رگرسیون: از الگوریتم زیر برای پیاده سازی رگرسیون خطی در این سوال استفاده می کنیم:

$$\hat{y} = w_1 * x_1 + w_2 * x_2 + \cdots + w_{61} * x_{61} + b$$

$$error^i = \frac{1}{2}({y_{train}}^i - \hat{y}^i)$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1} N(error^i)^2$$

$$\text{MSE} = \frac{1}{N}((error^0)^2 + (error^1)^2 + \cdots + (error^N)^2)$$

$$\text{MSE} = \frac{1}{N}((y^0_{train} - (w_1 * x_1^0 + \cdots + w_{61} * x_{61}^0 + b))^2 + \cdots + ((y^N_{train} - (w_1 * x_1^N + \cdots + w_{61} * x_{61}^N + b))^2)$$

$$\frac{\partial \text{MSE}}{\partial w_1} = \frac{-2}{2N}({error^0} * {x_1}^0 + {error^1} * {x_1}^1 + \cdots + {error^N} * {x_1}^N) = \frac{-2}{2N}(\sum_{i=1}^{N} error^i * x_1^i)$$

$$\frac{\partial \text{MSE}}{\partial w_2} = \frac{-2}{2N}({error^0} * {x_2}^0 + {error^1} * {x_2}^1 + \cdots + {error^N} * {x_2}^N) = \frac{-2}{2N}(\sum_{i=1}^{N} error^i * x_2^i)$$

$$\vdots$$

$$\frac{\partial \text{MSE}}{\partial w_{61}} = \frac{-2}{2N}({error^0} * {x_{61}}^0 + {error^1} * {x_{61}}^1 + \cdots + {error^N} * {x_{61}}^N) = \frac{-2}{2N}(\sum_{i=1}^{N} error^i * x_{61}^i)$$

$$\frac{\partial \text{MSE}}{\partial b} = \frac{-2}{2N}(\sum_{i=1}^{N} error^i)$$

کد مربوطه را و نتایج آن اتچ میشود.

PCA : پس از استفاده از pca تعداد ویژگی ها در نهایت به 17 تا تقلیل پیدا میکند

نتایج مدل:

مدل اول: مدل اول که خودمان به صورت دستی ساختیم دارای خطای 0.23 و با تلرانس 0.2 دارای دقت 61 درصد بود که در مقایسه با مدل های پیش ساخته دقت پایین ترین داشت

مدل دوم PCA : با PCA به خطای نهایی 0.046 و دقت 62 درصد دست پیدا کردیم که در مقایسه با مدل های از پیش ساخته شد دقت پایین تری دارد.