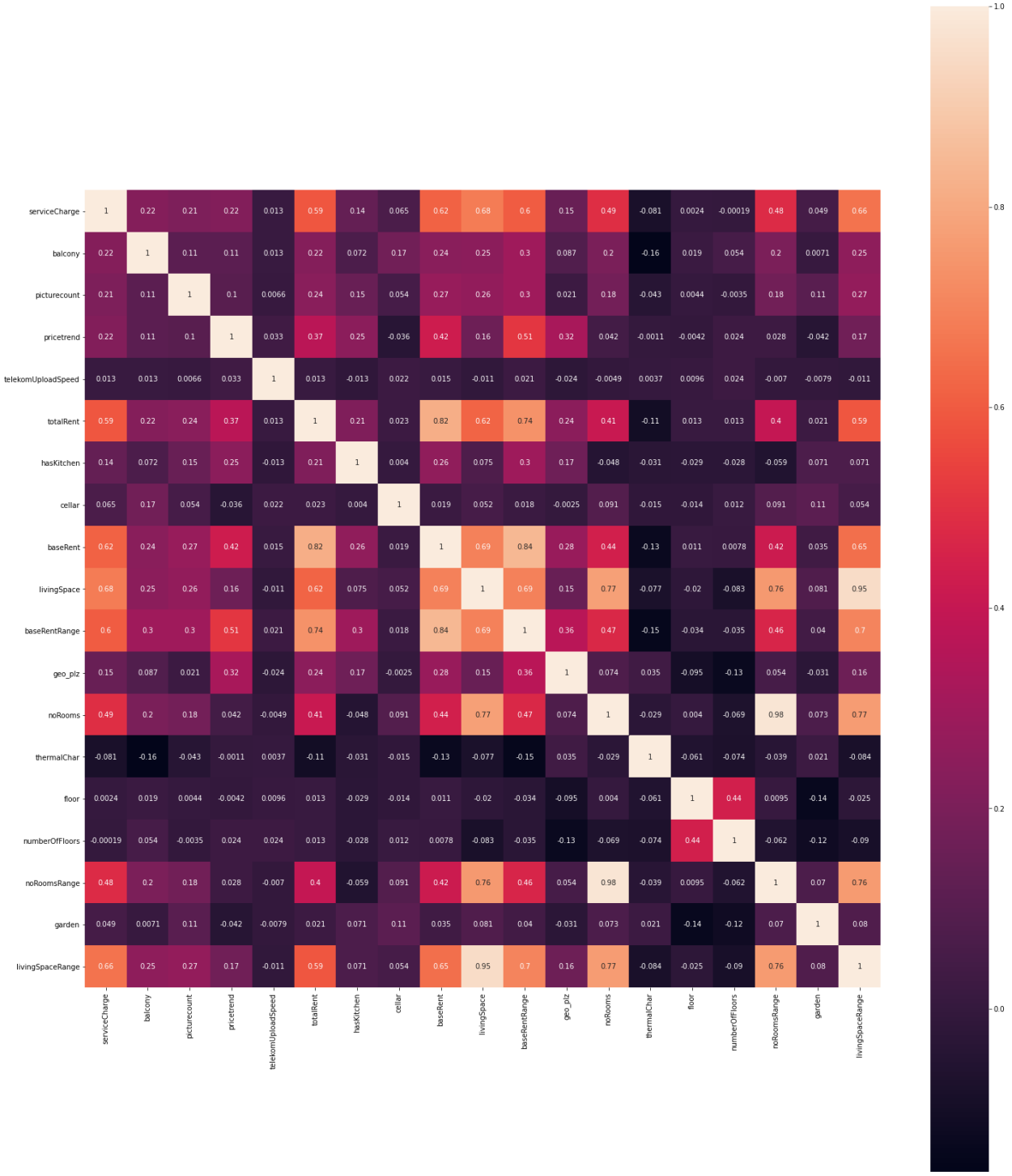
گزارش پروژه اول یادگیری ماشین:

شرح مسئله:
چالش این مسئله در پیاده سازی یک مدل رگرسیون خطی بود برای یک مجموعه داده از مشخصات خانه که متغیر وابسته ما یا همان متغیری که میخواستیم روش پیش بینی یا prediction انجام شود "مساحت یا فضای خانه" بود و بقیه متغیرها متغیر های آزاد ما بودند که میخواستیم از طریق آنها مساحت خانه را بدست آوریم.

کار با داده ها:
در اول کار داده ها را از سایت کگل خواندیم سپس اطلاعات کلی را از داده ها در آوردیم که مثلا از هر ستون چند داده null وجود دارد. سپس اون ستون هایی رو که بیشتر از 50% داده ها در آن null بود را حذف کردیم تا حجم داده ها کاهش یابد. سپس یک سری دیگر از ستون های نامربوط به هدف مسئله حذف شدن تا باز هم کاهش تعداد داده داشته باشیم.بعد با میانگین گرفتن از داده های عددی در هر ستون سعی کردیم که داده هایی که از جنس عدد اما null بودند توی اون ستون را پر کنیم.در مرحله بعد داده های پرت عددی رو حذف کردیم یعنی اونایی که خیلی از میانگین داده ها فاصله داشتند و در آخر داده های عددی رو normalize کردیم که بین 0 و 1 قرار بگیرند.تا این مرحله کار با داده های عددی تمام شد و در مرحله بعدی باید داده های null در داده های categorical را پر کنیم که اینجا از این استراتژی استفاده شد که اونایی که null هستند احتمالا از اون دسته ای هستند که بیشترین فراوانی را دارد. در قسمت بعدی سعی کردیم اون ستون هایی که داده های خاص زیاد دارند تا حدودی ادغام کرد و آنهایی که خیلی زیاد هستند کلا حذف شوند چون برای اینکه آن ها را عددی کنیم بسیار تعداد ستون هایمان زیاد میشود. در قسمت بعدی correlation matrix را رسم کردیم که نشان میدهد رابطه داده ها رو نسبت بهم که living space range بیشترین رابطه را با مساحت خانه دارد و telekom Upload Speed هم کمترین

ارتباط را دارد.



در آخر هم داده های categorical به داده های عددی تبدیل شدند و x و y هم جدا شدند.سپس داده ها به test و train تقسیم شدند تا عمل یادگیری روی داده های train اجرا شده و مدل نهایی ما روی داده های test ارزیابی شود.

مدل ها:

مدل 1) در این مدل نرخ یادگیری 0.01 و epoch را 500 در نظر گرفتیم سپس خود رگرسیون را پیاده سازی و روی داده ها اجرا کردیم که loss در آخر به 328 رسید سپس روی داده های test ارزیابی صورت گرفت که و به این شکل در نظر گرفته شد که اگه test*0.85 <predict> test*1.15 برقرار باشد یعنی مدل ما جواب درست داده است یعنی یک بازه 30% حول جواب واقعی. که در این مدل درستی 60% داشتیم.

مدل 2)در این مدل تعداد epoch را نسبت به مدل قبلی 2 برابر کردیم یعنی 1000 تا ببینیم ایندفعه مدل ما روی داده چه عملکردی دارد که این دفعه loss به 190 رسید و درستی ما به 74% رسید که نسبت به قبل بهتر بود.

مدل 3) در این مدل نرخ یادگیری را 0.1 کردیم و دیدم که بعد از اجرای 20 ایپاک متوقف شدیم چون هر دفعه loss بیشتر شد و این یعنی از نقطه optimum رد شدیم چون مقدار گام در هر مرحله که همان نرخ یادگیری مشخص میکند زیاد بود.منطقا در این مدل درستی 0 داشتیم.

مدل 4)در این مدل رگرسیون را روی یک متغیر اجرا کردیم طبق cor matrix روی متغیر living space range اجرا کردیم با نرخ یادگیری 0.01 و 500 epoch که به loss 456 رسیدیم و درستی 47% داشتیم

مدل 5)در این مدل مثل مدل قبلی روی یک متغیر living space range اجرا کردیم اما با epoch 5000 که این بار مقدار loss ما 102 شد و درستی به 87% رسید که از بهترین مدل های ما بود.

مدل 6) در این مدل روی متغیر بی ربط به مساحت خانه یعنی متغیر floor رگرسیون را اجرا کردیم با نرخ 0.01 و epoch 5000 که این دفعه loss ما 976 شد و درستی 32% داشتیم که درستی خوبی نبود چون متغیر هم ربطی به هدف ما نداشت.

مدل 7) در این مدل از پکیج sklearn استفاده شد برای رگرسیون خطی که loss 66 داشتیم و درستی 89% که درستی خیلی خوبی بود.

مدل 8) در این مدل از pca 0.9 استفاده کردیم برای کاهش حجم داده ها با همان نرخ 0.01 و epoch 500 که loss 329 و درستی 60% داشتیم که همان نتیجه مدل خودمان شد اما با داده های حجم کمتر

مدل 9) ایندفعه pca 0.8 استفاده کردیم با همان نرخ یادگیری و epoch قبلی که این دفعه loss 351 داشتیم و به درستی 59% رسیدیم.

مدل 10) در آخرین مدل هم از pca 0.98 استفاده کردیم به loss 338 رسیدیم و همچنین درستی 59%