تخمین مساحت خانه ها

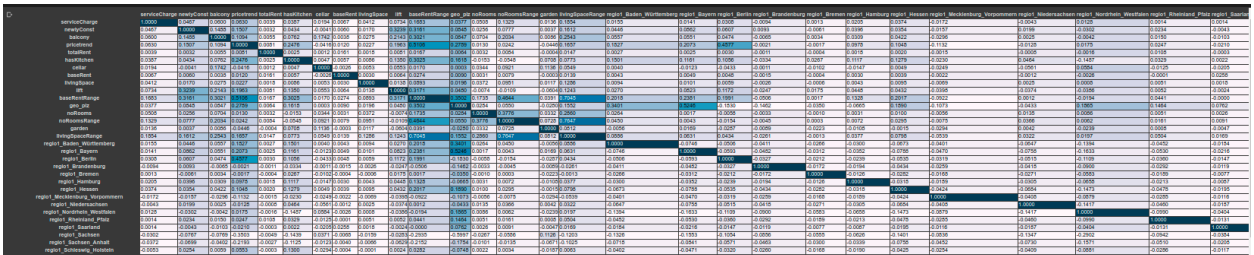
تمرین ۱ درس داده کاوی قسمت ۲

علی صالح
۹۷۲۲۲۰۵۳

پیش پردازش داده‌ها

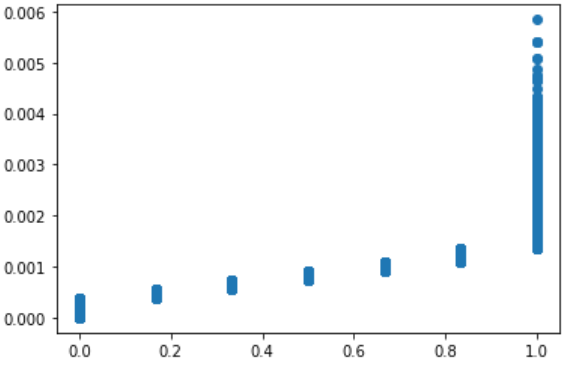
- ستون های غیر عددی که بیش از ۲۰ تا کتگوری هستند را حذف می‌کنیم.
- ستون های عددی که بیش از ۵۰ هزار مقدار نال دارند را حذف می‌کنیم.
- ستون های غیر عددی که بیش از ۱۵۰ هزار مقدار نال دارند را حذف می‌کنیم.
- ستون های **date scoutId picturecount** را حذف میکنیم. (به نظر ربطی به مساحت خانه ندارند.
- ستون **geo_bln** دقیقا همون **regio1** است پس این ستون را هم حذف می‌کنیم
- ستون های کتگوریکال را به صورت **one hot** انکود می‌کنیم.
- ستون ای عددی که همچنان مقدار نال دارند را با میانگین ستون پر می‌کنیم.
- داده ها را به روش **MinMax** اسکیل می‌کنیم.

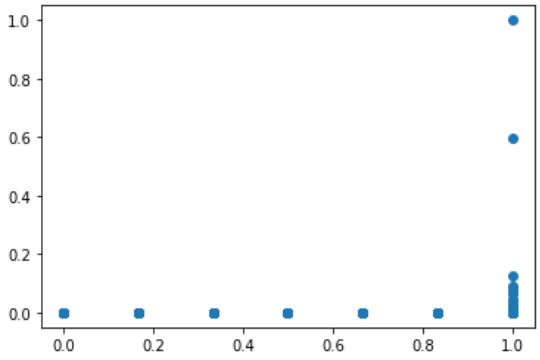
ماتریس کورلیشن را رسم می‌کنیم و متوجه میشویم در بین فیچر های گسسته فیچر **livingSpaceRange** و در بین فیچر های پیوسته **serviceCharge** بیشترین کورولیشن را با **livingSpace** دارند.



Living space range:

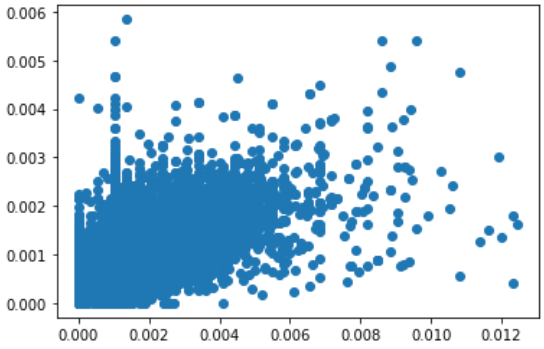
نمودار **livingSpaceRange** و **livingSpace** را می‌کشیم و به نمودار راست می‌رسیم. داده های نویز را حذف می‌کنیم و به نمودار چپ می‌رسیم که به نظر مناسب تر برای رگرسیون خطی می‌آید.

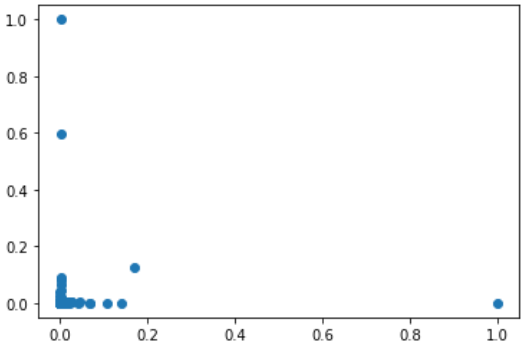


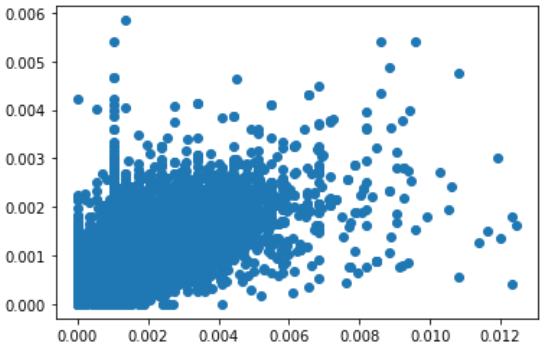


Service Charge

نمودار **serviceCharge** و **livingSpace** را می‌کشیم و به نمودار راست می‌رسیم. داده های نویز را حذف می‌کنیم و به نمودار چپ می‌رسیم که به نظر مناسب تر برای رگرسیون خطی می‌آید.



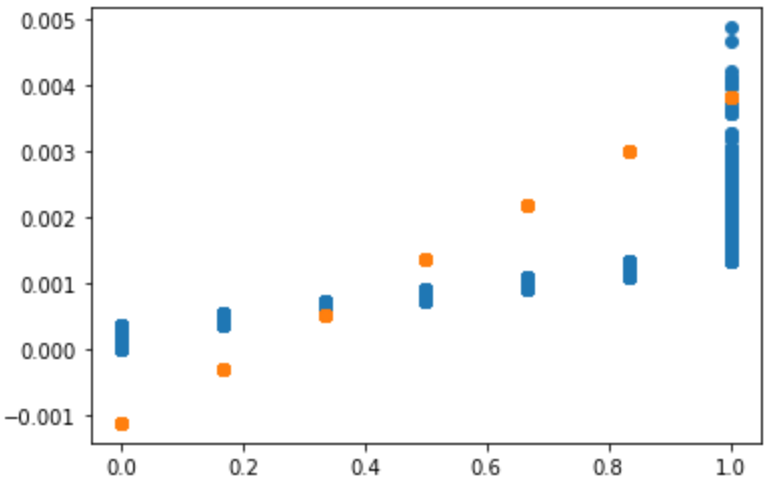




livingSpaceRange

ابتدا برای این فیچر را به تنهایی به مدل رگرسیونمان می‌دهیم.

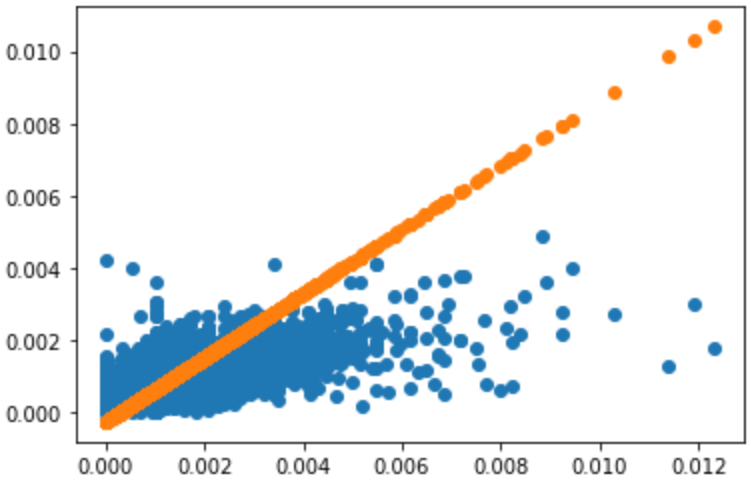
لرنینگ ریت را ۰.۰۵ می‌گذاریم و به اندازه ۱۰۰۰ ایپاک اجرا می‌کنیم. نتیجه حاصل:



serviceCharge

ابتدا برای این فیچر را به تنهایی به مدل رگرسیونمان می‌دهیم.

لرنینگ ریت را ۰.۰۵ می‌گذاریم و به اندازه ۱۰۰۰ ایپاک اجرا می‌کنیم. نتیجه حاصل:



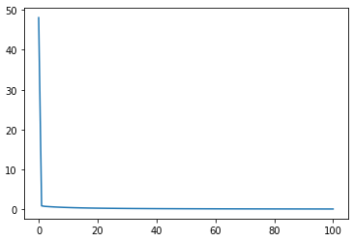
مدل رگرسیون

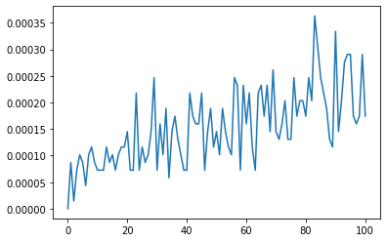
مدل رگرسیون ورژن ۲ نسبت به مدل قبلی چند قابلیت بیشتر دارد

- موقع تعریف accuracy_rate (مثلا ۰.۱ یعنی اختلاف یک دهم در محاسبه ی دقت درست در نظر گرفته شود) گرفته می‌شود و دقت محاسبه می‌شود.
- دقت و خطا در هر ایپاک محاسبه شده و همانجا چاپ می‌شود.
- دقت و خطای هر ایپاک در متغیر history ذخیره می‌شود.

با مدل رگرسیون ورژن۲ یک بار دیگر آزمایش all features را انجام می‌دهیم و به وسیله متغیر histoy نمودار های زیر می‌رسیم.

که به وضوح نشان می‌دهد داده های ما با این ابعاد برای رگرسیون خطی مناسب نیست. (راست دقت و چپ خطا)





Multi Process

با استفاده از مولتی پروسس دوباره از پری پروسسینگ ران می‌گیریم

اندازه ران تایم: ۰.۱۴ ثانیه

متاسفانه موفق به استفاده از dask و pyspark نشدم

به این دلیل که برای یک مدل با ۷۶ فیچر امکان درست کردن اسکیما نداشتم.

اگه راه دیگه ای هست لطفا راهنمایی کنید :)))