به نام خدا

گزارش تمرین 2

محمد سعید حیدری 400422075

مقدمه

در این گزارش اقداماتی که در جریان حل تسک‌های مشخص شده برای تمرین 2 انجام داده شده شرح و توضیحات و تحلیل‌های مربوط به آن‌ها نیز ارائه شده است.

Section 1: Remove missing values and outliers

در تسک اول شناسایی و حذف داده‌های میسینگ و داده‌های پرت در دستور کار بوده است. برای این منظور، ما ابتدا لیستی از متغیرهای حاوی تعداد زیاد مقادیر Nan را شناسایی و به صورت دستی آن‌ها را حذف کردیم. در زیر شکل 1 لسیتی از منغیرهای حذف شده را نمایش می‌دهد: همچنین با استفاده از دستور df1=df1.drop_duplicates به بررسی و حذف داده های مشابهت پرداخته ام

```
df1 = df.drop(labels=['typeOfFlat', 'streetPlain', 'street', 'petsAllowed', 'interiorQual',
       'petsAllowed', 'condition', 'geo_krs', 'houseNumber', 'yearConstructedRange', 'geo_bln',
       'firingTypes', 'noParkSpaces', 'scoutId', 'yearConstructed', 'totalRent', 'telekomUploadSpeed',
       'telekomHybridUploadSpeed', 'telekomTvOffer', 'telekomTvOffer',
       'thermalChar','floor', 'numberOfFloors', 'regio2', 'regio3',
       'description', 'facilities', 'heatingCosts', 'energyEfficiencyClass',
       'lastRefurbish', 'electricityBasePrice', 'electricityKwhPrice', 'date'] ,axis=1)
df1 = df1.dropna(axis=0, how='any')
df1 = df1.drop_duplicates()
```

شکل 1: لیستی از متغیرهای حذف شده

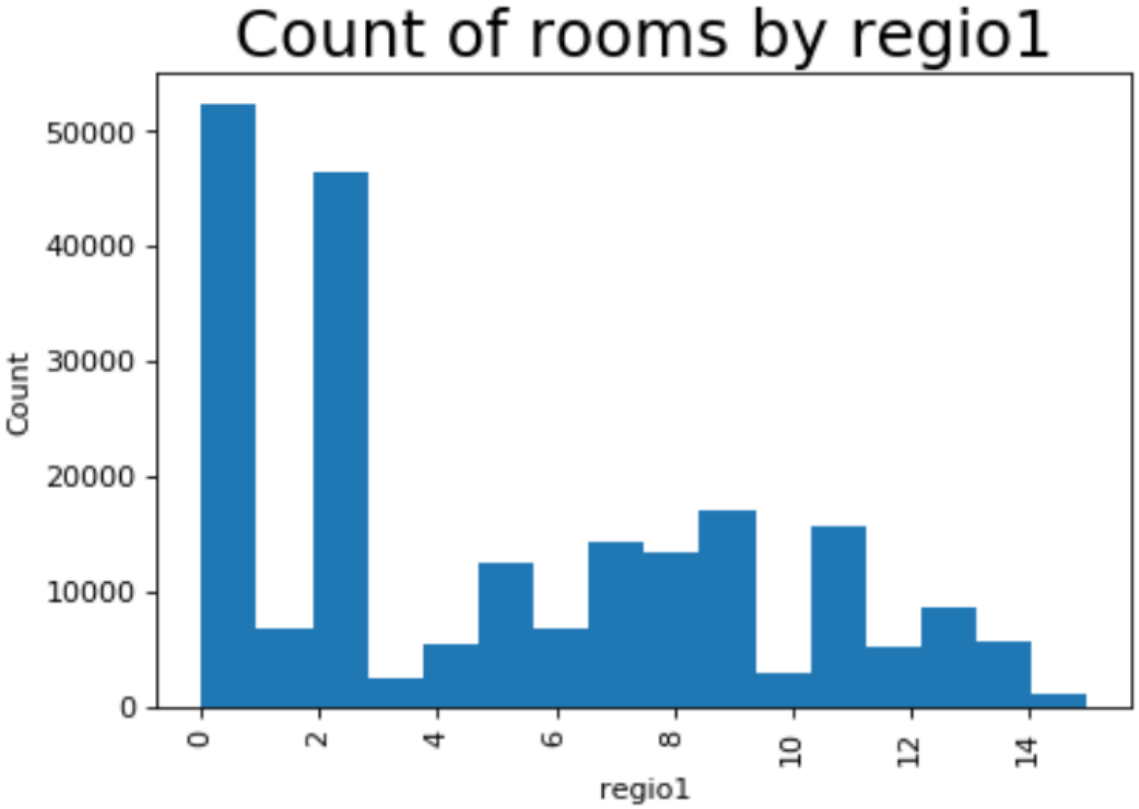
و سپس، با استفاده از دستور pd.dropna() سطرهای حاوی میسینگ را شناسایی و حذف کردیم. سپس، با استفاده از مفهوم نرمال بودن و فاصله نرمالیتی 3 برابری از طرفین میانگین، آن داده‌هایی که فاصله آن‌ها از میانگین کل داده‌ها بیش از 3 انحراف‌معیار فاصله داشت را به عنوان داده پرت شناسایی و حذف کردیم.

$$Outlier \quad if \ (|X_i - Mean|) > 3*Std \tag{1}$$

Section 2: Overal Information

در این تست نمایش اطلاعات کلی مربوط به دیتاست در دستور کار است.

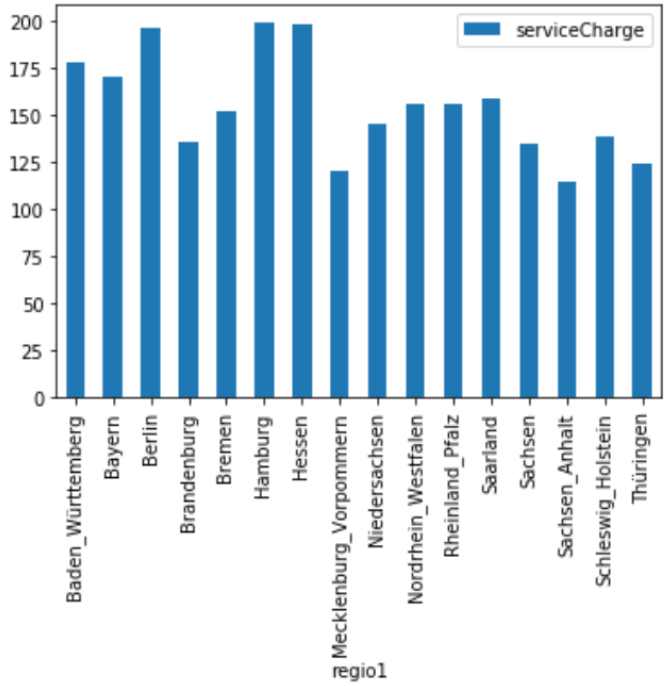
در ابتدا در شکل 2 تعداد خانه‌ها به تفکیک مناطق جغرافیایی داده شده‌اند.



شکل 2: فراوانی آگهی‌های منازل به تفکیک مناطق جغرافیایی

همانطور که در شکل بالا مشخص است، پراکندگی فراوانی آگهی در مناطق مختلف بالا است. در میان تمامی مناطق، و منطقه Nordrhein_westalen و Sachsen فراوانی بسیار بالایی دارند. در مقابل مناطقی مانند bermen و Saarland فراوانی پایینی دارند.

سپس، تلاش کردیم میانگین قیمت‌ها در هر یک از مناطق را محاسبه کرده و نمایش دهیم. شکل 3 میانگین قیمت منازل در نواحی مختلف جغرافیایی را نمایش داده است.



شکل 3: میانگین قیمت منازل در نواحی مختلف

آنطور که مشخص است، تغییرات قیمت در بین مناطق جغرافیایی چندان متفاوت نیست.

Section 3: Forecast charge

در این بخش تلاش شد با استفاده از ساختن یک مدل رگرسیون خطی قیمت خانه‌ها را برآورد کرده و سپس با تحلیل ضرایب رگرسیون نهایی، میزان اهمیت هر یک از متغیرهای مستقل نمایش داده شوند. برای این منظور ابتدا متغیرهای غیرعددی را به حالت کدهای عددی تبدیل کردیم. شکل 4 مقدار ضرایب تخصیصی به هر یک از متغیرهای مستقل را پیش‌بینی قیمت نمایش داده است.

```
print(MLR.coef_)
```

```
[-3.58299556e-17 -9.82070328e-16 -2.01939630e-16  3.77596978e-15
 -5.35178212e-15  1.00000000e+00 -1.57810456e-17 -4.00217887e-16
 -1.76045117e-16  9.90776655e-17  4.58396286e-17  3.40663524e-17
 -4.35156213e-16 -1.89464487e-17  5.22226387e-18 -1.55794856e-17
 -7.33018136e-17  4.68176220e-17]
```

```
print(X.columns)
```

```
Index(['regio1', 'serviceCharge', 'heatingType', 'newlyConst', 'balcony',
       'picturecount', 'pricetrend', 'hasKitchen', 'cellar', 'baseRent',
       'livingSpace', 'lift', 'baseRentRange', 'geo_plz', 'noRooms',
       'noRoomsRange', 'garden', 'livingSpaceRange'],
      dtype='object')
```

شکل 4: مقدار ضرایب تخصیص داده شده به هر یک متغیرهای مستقل در مدل رگرسیون

با توجه به ضرایب حاصله، اکثر متغیرها اهمیت برابری دریافت کرده‌اند و تنها متغیر picturecount ضریب بزرگتری دارد.

Section 4: Using multiprocessing in preprocessing

در این قسمت ما ابتدا زمان اجرای پیش‌پردازش را در حالت عادی محاسبه کردیم. سپس، پیش‌پردازش را در زمانی که از ابزار multiprocessing استفاده کردیم، انجام دادیم. جدول 1 زمان اجرا را در حالت استفاده و عدم استفاده از multiprocessing مقایسه کرده است.

جدول 1: مقایسه زمان اجرای پیش‌پردازش قبل و پس از استفاده از multiprocessing

multiprocessing	time
no	7.478757599999881
yes	6.570000005012844e-05

در جدول 1 مشخص است که زمان اجرا پس از استفاده از multiprocessing بسیار کاهش یافته است.

Section 5: Using dask in preprocessing

در این قسمت ما ابتدا زمان اجرای پیش‌پردازش را در حالت عادی محاسبه کردیم. سپس، پیش‌پردازش را در زمانی که از ابزار dask استفاده کردیم، انجام دادیم. جدول 2 زمان اجرا را در حالت استفاده و عدم استفاده از dask مقایسه کرده است.

جدول 2: مقایسه زمان اجرای پیش‌پردازش قبل و پس از استفاده از dask

dask	time
no	7.478757599999881
yes	0.00018379999892204069

در جدول 2 مشخص است که زمان اجرا پس از استفاده از dask بسیار کاهش یافته است. اگرچه میزان کاهش در استفاده از multiprocessing بوده است.

Section 6: Using dask in modeling

در این قسمت ما ابتدا زمان اجرای مدل‌سازی را در حالت عادی محاسبه کردیم. سپس، مدل‌سازی را در زمانی که از ابزار dask استفاده کردیم، انجام دادیم. جدول 3 زمان اجرا را در حالت استفاده و عدم استفاده از dask مقایسه کرده است.

جدول 3: مقایسه زمان اجرای پیش‌پردازش قبل و پس از استفاده از dask

dask	time
no	12.359803499999543
yes	0.00023830000282032415

در جدول 3 مشخص است که زمان اجرای مدلسازی پس از استفاده از dask بسیار کاهش یافته است.

Section 7: Using feature enginearing for improve modeliing

در این قسمت هدف بهبود کارایی مدل یادگیری با استفاده از مهندسی ویژگی است. برای این منظور، ابتدا ما دقت برآورد قیمت را بدون استفاده از مهندسی ویژگی ارزیابی کردیم. برای این منظور، ابتدا داده‌ها را نرمال کرده و سپس تمام متغیرهای غیرعددی را به حالت عددی تبدیل کردیم. سپس، 80 درصد داده‌ها را برای آموزش و 20 درصد را برای تست تخصیص دادیم.

سپس، یکبار بدون مهندسی ویژگی عمل تست مدل روی داده‌های تست را انجام داده و مقدار متغیرهای ارزیابی رگرسیون مانند Mean Absolute Error و Mean Squared Error و Root Mean Squared Error را حساب کردیم. سپس، با انجام عمل انتخاب ویژگی (یکی از رایج‌ترین اعمال مهندسی ویژگی) عمل بالا را تکرار کردیم.

روش مورد استفاده برای انتخاب ویژگی محاسبه mutual information بوده است. به صورتی که، ابتدا مقادیر mutual information متناظر با هر یک از ویژکیها را محاسبه کردیم. سپس، ویژگی هایی که مقدار mutual information آن ها یا همان توان تعریف کنندگی آن ها بیشتر از میانگین کل توان ها بود، انتخاب شدند. شکل 5 مقدار توان تعریف تخصیص داده شده به هر یک از متغیرهای مستقل را نمایش داده است:

```
print(mi)
print(X.columns)
```

```
[0.02855321 1.         0.01488216 0.00507625 0.00830658 0.01145925
 0.02274713 0.00616187 0.00234954 0.09949937 0.12702759 0.01094974
 0.06601473 0.07633813 0.035945   0.03436748 0.00167825 0.06866517]
Index(['regio1', 'serviceCharge', 'heatingType', 'newlyConst', 'balcony',
       'picturecount', 'pricetrend', 'hasKitchen', 'cellar', 'baseRent',
       'livingSpace', 'lift', 'baseRentRange', 'geo_plz', 'noRooms',
       'noRoomsRange', 'garden', 'livingSpaceRange'],
      dtype='object')
```

شکل 5: مقدار توان تعریف تخصیص داده شده به هر یک از متغیرهای مستقل

در میان کل ویژگی‌ها، ویژگی‌های 8، 9، 11، 12 و 16 انتخاب شدند.

جدول 4 مقدار معیارهای ارزیابی را قبل و بعد از انجام مهندسی ویژگی مقایسه کرده است.

ساخت مدل رگرسیونی برای پیش بینی تعداد مشتری:

```
In [14]: MLR = LinearRegression().fit(X, Y2)

In [15]: print(MLR.coef_)
         print(X.columns)

[-0.30977897  1.24290268 -1.44087565 -0.23274646 -0.16951436  0.08170939]
Index(['neighbourhood_group', 'latitude', 'room_type', 'minimum_nights',
       'calculated_host_listings_count', 'availability_365'],
      dtype='object')
```

شکل 6: ضرایب هر یک از متغیرهای ورودی در مدل رگرسیونی آموزش دیده جهت پیش‌بینی تعداد مشتری

همانطور که در شکل 6 مشخص است، از میان متغیرها مستقل موجود در دیتاست، latitude، room_type و calculated_host_listings_count_group دارای بیشترین تاثیر و اهمیت در ساختن مدل پیش بینی تعداد مشتریان بوده اند.

جدول 4: مقایسه نتایج رگرسیون پیش و پس از انتخاب ویژگی

Metric	Without features selection	With feature selection
Mean Absolute Error	40.38014638633405	41.165404769060174
Mean Squared Error	13571.857803163464	13651.464677802116
Root Mean Squared Error	116.49831673961415	116.83948252967451

آنطور که مشخص است با اعمال مهندسی ویژگی خطاها تغییر زیادی نکرده اند. البته طبیعتا کاهش تعداد ویژگی ها زمان انجام مدلسازی را کاهش می دهد

نتیجه گیری کلی : تمرکز این پروژه بیشتر روی پردازش و زمان پردازش و تاثیر مهندسی ویژگی می باشد. متغییر هایی مثل عرض جغرافیایی و منطقه جغرافیایی خیلی توی فیمت و تعداد تعداد مشتری ها تاثیر دارن. گزارش از تمامی مراحل به صورت استپ بای استپ در ژوپیتر نیز خدمتتان ارسال شده است جناب شریفی

پیروز و سرافراز باشید