به نام خدا

تمرین شماره ۲

مقدمه:

این تمرین مربوط به یک دیتاست از یکی از بزرگترین پلتفرم های کشور آلمان در حوزه املاک است که از آگهی های استخراج شده از این پلتفرم بدست آمده اند. در ادامه گزارش مربوط به ۵ تسک اصلی که در این تمرین خواسته شده است ارایه می شود:

تسک ۱:

در این سوال ابتدا پس از فراخوانی دیتاست اطلاعات کلی دیتا ست را نگاه میکنیم سپس ستون هایی که بیش از ۵۰ درصد آن ناموجود است را مشخص کرده و حذف میکنیم. در مرحله بعد ستون هایی عددی که مقادیر ناموجود دارند، این مقادیر ناموجود را با میانگین داده های معلوم همان ستون مشخص میکنیم. در مرحله بعد داده های پرت را تشخیص داده و حذف میکنیم. داده پرت در اینجا منظورمان همان داده های هر ستون است که مقدار آن یا از میانگین به اضافه سه برابر انحراف معیار بیشتر است یا کمتر زیرا تقریبا ۹۹ درصد داده های در این بازه میباشند.

در مرحله بعد ستون های عددی که بعضی از مقادیر آنها ناموجود است را با بیشترین فراوانی از همان ستون پر میکنیم و در مراحل بعد ستون هایی که به نظر کاربرد خاصی در پردازش ندارند را حذف میکنیم. سپس در ستون های heatingType و condition و typeofflat به وسیله توابع edite_ heatingType و edite_condition و edite_typeofflat به ترتیب ۴ و ۳ و ۲ داده ای که کمترین فراوانی را دارند با هم ادغام میکنیم و به یک داده تبدیل میکنیم در نهایت به انکدر one_hot_endoder روی داده های غیر عددی اجرا میکنیم تا به ستون عددی تبدیل شوند و ستون های غیر عددی را از دیتاست اصلی حذف کرده و و دیتاست اصلی را با دیتاست حاصل از انکدر ادغام میکنیم.

به نام خدا

تسک ۲:

در این مرحله نمودار بین قیمت و هر یک از فیچرهای دیتاست را رسم میکنیم البته چون اکثر این فیچرها به صورت ۰ و ۱ هستند شاید ارتباط خطی در این نمودار های مشاهده نشود اما مثلا مشاهده میکنیم که totlRent و baseRent با هم رابطه خطی دارند.

تسک ۳:

در مرحله بعد baseRent را به عنوان خروجی و سایر ویژگی های دیتاست را به عنوان پارامتر ورودی در نظر میگیریم و سپس داده ها را به دو قسمت train و test تبدیل میکنیم که داده های تست ۲۰ درصد داده های اصلی میباشند سپس از رگرسیون خطی برای پیش بینی استفاده میکنیم و میبینیم که دقت یادگیری مدل ۰.۸۲۳۷ و دقت پیش بینی ان ۰.۸۲۳۸ میباشد.

تسک ۴:

در این مرحله یک ستون جدید به دیتاست اضافه میکنیم (ستون z) و مقدار آن را برابر صفر قرار میدهیم تا زمانی که بخواهیم در دیتاست groupby انجام دهیم تنها یک گروه داشته باشیم. سپس بر حسب z ، groupby میکنیم و یک تابع مینویسیم که در دیتاست ستون هایی که بیش از ۵۰ درصد مقادیر آن ناموجود است را حذف کنیم و ستون های عددی با مقادیر ناموجود را با میانگین مقادیر موجود پر کنیم و ستون های غیر عددی که مقادیر ناموجود دارند را با بیشترین فراوانی از همان ستون پر کنیم سپس به وسیله multiprocessing تابع را روی دیتا اجرا کرده و زمان را ثبت میکنیم و بعد از آن مشاهده میکنیم که مقدار ناموجود در دیتاست نیست. زمان پردازش : ۹ ثانیه

تسک ۵:

در این مرحله برای قسمت dask همان تابع مرحله قبل را به صورت دستور برای dask مینویسیم و اجرا میکنیم و زمان را ثبت میکنیم در این مرحله نیز دیتا مقادیر ناموجودی ندارد.
زمان پردازش : ۲۹ ثانیه