CS50x Iran

تورنومنت نهایی!

Apex

بهار ۱۴۰۲

چالش پیش‌بینی قیمت خانه با تلفیقی از داده های تصویری و جدولی

توسعه یک مدل یادگیری ماشین که با استفاده از داده های تصویری و
جدولی، قیمت فروش و مساحت خانه را به درستی پیش بینی کند.

از دیتاست House Price Dataset موجود در [این لینک] استفاده کنید.

دیتاست جدولی (ذخیره شده در قالب یک جدول) ویژگی‌های مختلف یک خانه، به عنوان مثال تعداد اتاق‌ها، منطقه و مکان را با داده های تصویری مربوط هر یک از این خانه‌ها، ترکیب می‌کند. این مجموعه به طور کلی شامل ۱۵۵۰۰ نمونه داده است که ۱۵۰۰۰ نمونه برای آموزش و ۴۰۰ نمونه برای تست دارد.

جزئیات مجموعه داده

داده‌های جدولی در یک فایل CSV با چندین ستون ارائه می‌شود که ویژگی‌های خانه‌های مختلف مانند مساحت، تعداد تختها، مکان و غیره را نشان می‌دهد.

داده های تصویری حاوی تصاویر JPEG ارائه می‌شوند که نام فایل تصویر مربوط به شناسه منحصر به فرد (image_id) در داده های جدولی است.

جزئیات مجموعه داده

فایل دیتاستی که دانلود میکنید شامل سه پوشه است.

پوشه اول (trian) مربوط به دیتا یادگیری مدل است. پوشه دوم (test) مربوط به ارزیابی مدل شما در زمان توسعه است. هر کدام از این دو پوشه شامل عکس و یک فایل csv میباشد. که در فایل csv اطلاعات خانه و متناظر با آن عکس خانه در این مسیر وجود دارد.

و پوشه آخر (test_final)!!!!

جزئیات مجموعه داده

پوشه test_final مربوط به ارزیابی مدل شماست.

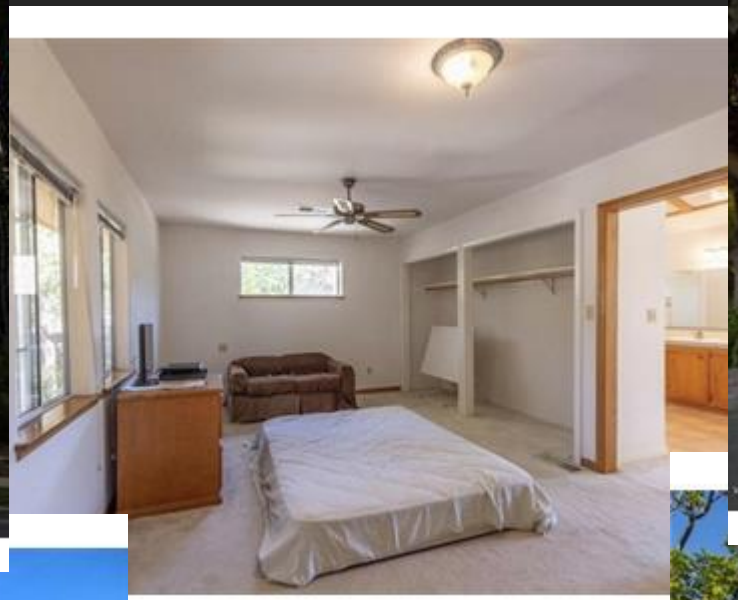
بعد از توسعه و یادگیری، مدل شما باید خروجی مدل را به اعضا دیتا که در این پوشه وجود دارد ارزیابی کنید.

اگر به فایل csv این مسیر دقت کنید، بعضی از آنها در ستون قیمت (price) و مساحت (sqft) خالی هستند.

وظیفه شماست که آنها را تکمیل کنید.

جزئیات مجموعه داده

image_id	Id عکس متناظر در پوشه
street	نام خیابان
citi	نام شهر
n_citi	کد شهر
bed	تعداد تخت
bath	تعداد دستشویی
sqft	مساحت خانه
price	قیمت خانه



شرکت کنندگان باید مدلی بسازند که قیمت و مساحت خانه را براساس هر دو داده تصویری و داده های جدولی پیش بینی کند.

معیار ارزیابی این رقابت، ریشه میانگین مربعات خطا (RMSD) بین خروجی پیش‌بینی شده مدل شما و قیمت‌های فروش واقعی خواهد بود. مقادیر کمتر RMSD عملکرد بهتر مدل را نشان می‌دهد.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

فایل های ارسالی شما براساس معیار ریشه میانگین مربعات خطا (RMSD) محاسبه شده بین قیمت‌های پیش‌بینی شده برای خانه توسط مدل شما و قیمت‌های فروش واقعی ارزیابی می‌شوند. فرآیند ارزیابی به صورت خودکار در پلتفرم مسابقه انجام خواهد شد.

برای نشان دادن فرآیند ارزیابی، اجازه دهید یک فایل ارسال شده فرضی توسط شرکت‌کننده‌ای به نام «امیر» را برای مجموعه داده آزمایشی زیر در نظر بگیریم. فایل امیر حاوی پیش بینی قیمت فروش ۱۰۰ خانه است.

فایل ارسالی امیر باید در قالب دو فایل CSV، با ستون های زیر باشد:

ID شناسه منحصر به فرد برای هر خانه در مجموعه داده تست

predicted_price قیمت فروش پیش بینی شده توسط مدل امیر برای هر خانه

predicted_sqft قیمت فروش پیش بینی شده توسط مدل امیر برای هر خانه

پلتفرم مسابقه، قیمت‌های پیش‌بینی شده توسط مدل امیر را با قیمت فروش واقعی خانه‌های موجود در مجموعه داده تست مقایسه می‌کند.

شرکت‌کنندگان باید پیش‌بینی‌های خود را در مورد مجموعه داده تست در قالب یک فایل CSV ارسال کنند.

فایل ارسالی باید شامل شناسه منحصربه‌فرد، قیمت و مساحت‌های پیش‌بینی شده خانه و هر ستون اضافی مورد نیاز برای ارزیابی باشد.

توجه داشته باشید:

برای اطمینان از ارزیابی دقیق، مهم است که پیش‌بینی‌هایتان را در قالب یک CSV با ساختار خواسته شده ارسال کنید.

عدم استفاده از داده های خارجی و مدل های از پیش آماده: شرکت کنندگان مجاز به استفاده از مجموعه داده های خارجی یا مدل های از پیش آموزش دیده (pretrained) برای آموزش مدل های خود نیستند! هدف از این مسابقه، تشویق شرکت کنندگان به کار با مجموعه داده ارائه شده و توسعه مدل با استفاده از تکنیک های یکپارچه سازی داده های تصویری و جدولی است.

اصالت: شرکت کنندگان باید کد برنامه را خودشان نوشته باشند و از هر گونه کپی یا استفاده از کدهای دیگران خودداری کنند هر کتابخانه یا منبع خارجی استفاده شده نیز باید به درستی ارجاع داده شود.

ارسال های تیمی: شرکت کنندگان می توانند به صورت انفرادی یا به عنوان بخشی از یک تیم شرکت کنند. در صورت ارسال تیمی، هر تیم می تواند حداکثر سه عضو داشته باشد.

به اشتراک گذاری کد و همکاری: ما شرکت کنندگان را تشویق میکنیم تا از یکدیگر یاد بگیرند و در بحث های سالم و مفید با هم به تبادل اطلاعات بپردازند. با این حال، اشتراک گذاری کد یا همکاری با سایر شرکت کنندگان خارج از تیم ثبت نام شده در مسابقه اکیداً ممنوع است. نقض این قانون ممکن است منجر به رد صلاحیت شود.

فایل های ارسالی باید در قالب CSV و طبق ساختار و نام تعیین شده ارسال شوند. توجه نکردن به هرکدام از این موارد موجب رد شدن فایل شما خواهد شد.

فایل ارسالی باید در قالب زیر باشد:

team_price.csv در قسمت team باید نام تیم قرار بگیرد.

team_sqft.csv درقسمت team باید نام تیم قرار بگیرد.

شرکت کنندگان باید از صحیح بودن فایل های ارسالی خود اطمینان حاصل کنند. توصیه می‌کنیم که قبل از ارسال نهایی، قالب، نام ستون ها و سازگاری داده ها را دوباره بررسی کنید.

فایل های ارسالی که حاوی پیش بینی تمامی داده های تست نباشند و پیش بینی قیمت برای یک یا تعدادی از خانه ها در این فایل وجود نداشته باشد یا داده ای که پیش بینی برای آن صورت گرفته با داده اصلی مغایرت داشته باشد، پذیرفته نمی‌شوند. پس حتما اطمینان حاصل کنید که یک قیمت برای برای همه خانه های در مجموعه داده تست پیش بینی شده باشد.

فایل ارسال شده پس از مهلت تعیین شده پذیرفته نخواهد شد. شرکت‌کنندگان باید پیش‌بینی‌های خود را زودتر ارسال کنند تا از هرگونه مشکل فنی یا تاخیر لحظه آخری جلوگیری شود.

CS50x Iran

Harvard's Computer Science 50x Iran