



### مقدمه

در این پروژه، هدف ما تخمین قیمت خودروها در آگهی های سایت دیوار است. می خواهیم به کمک کتابخانه ی scikit learn و مدل های مختلف پیاده سازی شده در آن، سعی کنیم برای هر مدل های پارامترهای بهینه را بیابیم و در نهایت نتایج مدل ها را مقایسه کنیم.

### معرفی مجموعه ی داده

مجموعه داده تعدادی آگهی درج شده برای فروش خودرو در سایت دیوار است که در فرمت CSV در اختیار شما قرار گرفته است. در این داده، عنوان آگهی، توضیحات مربوط به خودرو، نوع خودرو (خودروی سبک یا سنگین)، زمان درج آگهی، تعداد تصاویر بارگذاری شده در کنار آگهی، کارکرد خودرو و سال تولید آن مشخص شده است. در نهایت یک ستون مربوط به قیمت درج شده در آگهی وجود دارد که همان هدف مدل خواهد بود.

### فاز اول: پیش پردازش و مشاهده<sup>۱</sup>ی داده

در فاز اول باید داده های خام ورودی را به مجموعه ای از ویژگی های قابل پردازش تبدیل کنید. به این منظور لازم است مقادیر خالی ستون ها را حذف کنید، داده هایی که دسته ای<sup>۲</sup> هستند را با روش های مرسوم به داده هایی قابل استفاده برای مدل تان استفاده کنید. همچنین ستون های حاوی اطلاعات متنی داخل مجموعه داده را برای تحلیل های بعدی پیش پردازش کنید. برای این کار می توانید از کتابخانه ی [هضم](#)<sup>۳</sup> استفاده کنید یا خودتان موارد مورد نیازتان را پیاده سازی کنید. شما باید توضیحات و تیتروایی که موجود است را تا حد ممکن Normalize کنید. (روش های ممکن، شامل حذف کلمات پرتکرار یا همان stop words، تبدیل کلمات به ریشه آنها و ... است.) همچنین لازم است رابطه ی بین فیچرهایی که دارید و ستون هدف را بررسی کنید. به این منظور information gain را برای ویژگی ها محاسبه کنید و نمودار gain بر

<sup>۱</sup> Preprocess and visualization

<sup>۲</sup> categorical

<sup>۳</sup> <https://github.com/sobhc/hazm>

حسب ویژگی ها را رسم کنید. (برای محاسبه میتوانید از متد `mutual_info_regression` در کتابخانهی SciKit-Learn استفاده کنید. بخش مربوط به regression در این [لینک](#) می تواند به شما کمک کند.)

**پرسش ۱)** این نمودارها چه اطلاعاتی برای استفاده از فیچرها در ادامه کار میدهند؟ به طور خلاصه توضیح دهید. هرکاری که برای آماده کردن ستون ها برای مدل هایتان انجام داده اید را حتما در گزارش ذکر کنید.

**پرسش ۲)** برای تبدیل داده های دسته ای به اطلاعات قابل استفاده برای مدلتان راه های متفاوتی وجود دارد. میتوانید در این [لینک](#) تعدادی از آن ها را مشاهده کنید. شما کدام روش را انتخاب کردید؟ چرا؟

**پرسش ۳)** برای پردازش داده های متنی، شما میتوانید تعداد مشخصی از کلمات پرتکرار متن را به عنوان فیچر در نظر بگیرید و برای هر سطر داده، مشخص کنید این کلمات چه تعداد بار تکرار شده اند (count vectorizer) یا از معیاری مثل tf-idf استفاده کنید که علاوه بر تعداد کلمات، تعداد بار تکرار یک کلمه را در هم در نظر میگیرد، به عنوان مثال اگر کلمه ای زیاد در عبارت ها تکرار شود، ارزش آن را کم میکند چرا که وقتی در همه ی عبارت ها است داشتن یا نداشتن آن کلمه بار اطلاعاتی کمتری دارد. در این [لینک](#) میتوانید بیشتر در مورد این دو مدل مطالعه کنید. در یادگیری مدل هایتان (حداقل در یکی از مدل ها) استفاده از تعداد کلمات مختلف و معیارهای متفاوت را امتحان کنید و نتایج را ذکر کرده و توضیح دهید.

**پرسش ۴)** در برخی ستون ها مقادیر از دست رفته<sup>4</sup> وجود دارد، برای حل این مساله روش های متفاوتی وجود دارد، دو تا از این روش ها را توضیح دهید و بگویید شما از کدام یک از آن ها استفاده کردید. میتوانید در این [لینک](#) تعدادی از این روش ها را ببینید.

## فاز دوم: پیش بینی قیمت

در مرحله ی اول داده ای که در اختیارتان قرار گرفته را به دو بخش یادگیری و تست تقسیم کنید. سپس با استفاده از کتابخانه ی scikit-learn با مدل های KNN، Linear Regression و Decision Tree مدلتان را با داده هایی که برای یادگیری جدا کردید train کنید و سپس به کمک مدل، ستون قیمت را برای داده های تست خودتان تخمین بزنید. برای ارزیابی تخمین هایتان از معیار Root Mean Square Error و Mean Square Error استفاده کنید. (هر دو معیار را برای هر مورد ذکر کنید) فرمول هر کدام از این خطا ها در زیر آمده است ولی برای مطالعه ی بیشتر میتوانید از این [لینک](#) هم استفاده کنید. میزان خطای rmse مطلوب برای این پروژه 10,000,000 میباشد.

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

برای هر کدام از مدل ها جست و جو کنید که هایپارامترهای ذکر شده در زیر چه هستند و آن ها را تغییر دهید و مقدار بهینه را برای مدلتان پیدا کنید، مقدار بهینه یعنی حالتی که loss مدل کمینه باشد ولی overfitting رخ ندهد.

KNN: `n_neighbors`

Decision tree: `max_depth`, `min_samples_split`

---

<sup>4</sup> Missing values

در صورتی که اجرای الگوریتم  $KNN$  زمان زیادی میگیرد، میتوانید این الگوریتم را روی بخش کمتری از داده اجرا کنید. همچنین آزمودن حداقل ۳ مقدار برای هایپر پارامتر ذکر شده کافی است.

**پرسش ۵)** داده‌هایتان را به چه نسبتی برای یادگیری و تست تقسیم کرده‌اید؟ چرا؟ برای linear regression یکبار ۹۸ درصد داده را برای یادگیری استفاده کنید و ۲ درصد را برای تست، نتایج را با حالتی که خودتان تقسیم کردید مقایسه کنید و توضیح دهید. اگر ۴۰ درصد را برای یادگیری استفاده کنید چطور؟ نام این پدیده‌ها چیست؟ در مورد آن‌ها کمی توضیح دهید.

**پرسش ۶)** اگر max\_depth را برای درختان بسیار زیاد کنید، چه اتفاقی می‌افتد؟ اگر کم کنید چطور؟ نمودار تغییرات هر دو معیار خطا را برای داده‌های تست و یادگیری در این بازه رسم کنید. (سعی کنید حداقل ۷ مقدار را برای max depth امتحان کنید و در نمودار نمایش دهید.) نمودار را تحلیل کنید و بگویید از آن چه برداشتی دارید؟

## فاز سوم: استفاده از مدل‌های تجمیعی

در این بخش برای پیش‌بینی قیمت از مدلی استفاده میکنیم که از یادگیری گروهی استفاده می‌کند به عبارتی از تجمیع نتایج حاصل از تعدادی مدل، پیش‌بینی نهایی را انجام می‌دهند.

به عنوان اولین مدل در این بخش از random forest استفاده کنید. در این مدل، تعدادی decision tree ساخته میشود که هرکدام جداگانه و با فیچرهای متفاوت آموزش می‌بینند، سپس برای تخمین نهایی بین نتایج درخت‌ها نوعی رای‌گیری میشود. در مورد حداقل دو تا از هایپرپارامترهای این مدل مطالعه کنید و تاثیر تغییر این هایپرپارامترها را روی نتایجتان بسنجید. حتما نتایج را دقیق ذکر کنید یا با رسم نمودار نشان دهید (میتوانید از این [لینک](#) استفاده کنید). برای این کار می‌توانید از متد gridsearchCV کتابخانه‌ی scikit learn استفاده کنید. این [لینک](#) می‌تواند به شما در استفاده از این کتابخانه کمک کند.

**پرسش ۷)** نتایج این مدل را با مدل decision tree مقایسه کنید. در مورد bias و variance و ارتباط بین آن‌ها در این [لینک](#) مطالعه کنید و بگویید به نظر شما از نظر هر کدام از دو مورد bias و variance یک مدل تنها (decision tree) بهتر عمل میکند یا یک مدل تجمیعی (random forest). آیا نتایجی که به دست آوردید با نظرتان مطابقت دارد؟

## نکات پایانی

- دقت کنید که هدف پروژه تحلیل نتایج است بنابراین از ابزارهای تحلیل داده بطور مثال نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. اگر در جایی ذکر شده مقایسه‌ای انجام دهید، حتما نتایج را دقیق ذکر کنید و سپس آن‌ها را تحلیل و مقایسه کنید.
- در همه‌ی بخش‌ها مجازید از متدهای کتابخانه‌ی scikit learn استفاده کنید ولی باید اطلاعات لازم در مورد هر کاری که انجام میدهید را داشته باشید، در هنگام تحویل ممکن است در مورد هرکدام از شما سوال پرسیده شود. (به عنوان مثال در مورد اینکه هر مدل چگونه پیش‌بینی را انجام میدهد، خطا چطور محاسبه میشود، هر هایپر پارامتر چه چیزی را تغییر میدهد و...)
- نتایج و گزارش خود را در یک فایل فشرده با عنوان AL\_CA4\_<#SID>.zip تحویل دهید. محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. توضیح و نمایش خروجی‌های خواسته شده بخشی از نمره این تمرین را تشکیل می‌دهد. از نمایش درست خروجی‌های مورد نیاز در فایل html مطمئن شوید.

- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت توسط ایمیل با طراحان در ارتباط باشید.
- هدف از تمرین، یادگیری شماست. لطفا تمرین را خودتان انجام دهید.