

گزارش درس داده‌کاوی

**«تمرین دوم»**

گردآورنده: سعید دادخواه

استاد: دکتر ناظرفرد

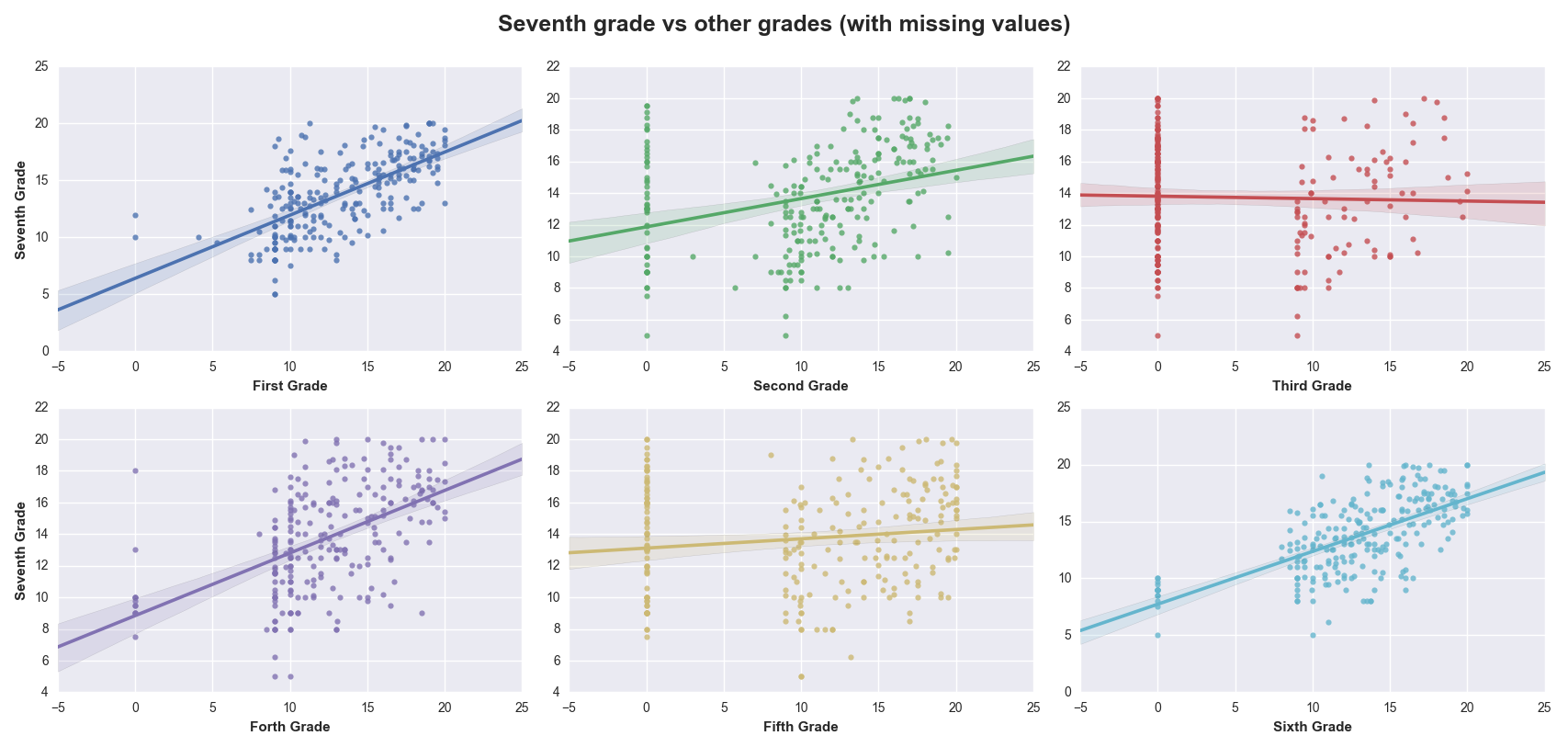
بهمن ۱۳۹۵

# بخش اول

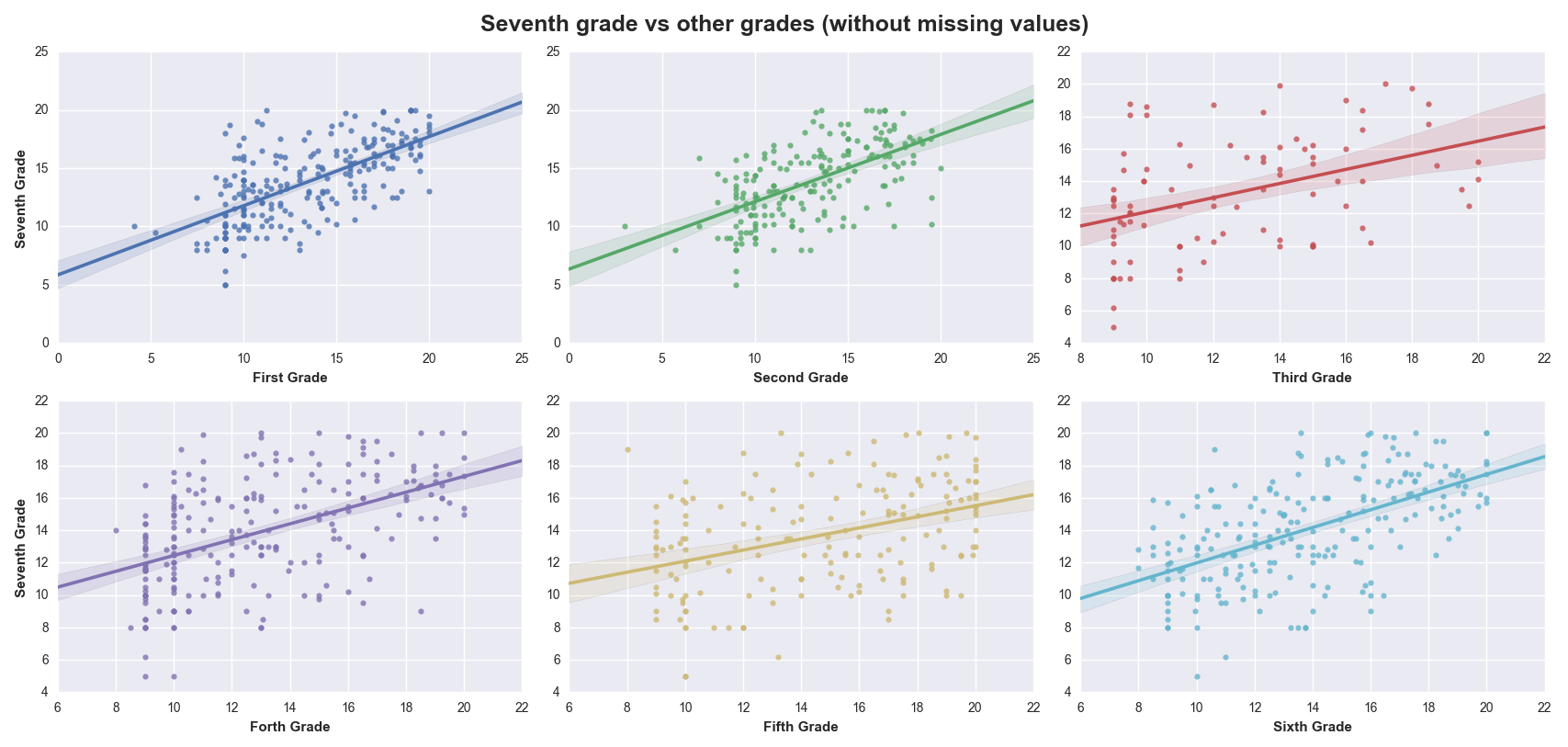
## قسمت ۱

* K-fold cross validation روشی برای ارزیابی مدل است. در این روش داده‌های آموزشی به K قسمت مساوی تقسیم می‌شوند و مدل K بار آموزش داده می‌شود. در هربار آموزش یک قسمت از داده‌های آموزشی جدا می‌شوند و با باقی داده‌ها مدل آموزش داده می‌شود و پس از آموزش با داده‌های جدا شده مدل آزمایش می‌شود. بعد از تکرار این کار با همه دسته‌ها K نتیجه از عملکرد مدل وجود دارد که می‌توان با استفاده از این مقادیر اطلاعات خوبی راجع به عملکرد مدل مانند میانگین و انحراف معیار و در نتیجه عملکرد کلی آن به دست آورد.
* MSE، MAE و RMSE معیارهایی برای اندازه‌گیری خطا هستند. MAE میانگین اختلاف مقدار به دست آمده و مقدار مورد انتظار را محاسبه می‌کند. MSE برای افزایش دقت خطاها را به نحوی محاسبه می‌کند که با افزایش اختلاف بین مقدار محاسبه شده و مقدار مورد انتظار خطا بیشتر در نظر گرفته شود، برای این کار احتلاف‌ها به توان دو می‌رسند سپس میانگین آن‌ها محاسبه می‌شود. مشکل MSE این است که واحد آن با واحد مقدار اندازه‌گیری شده یکی نیست و توان دوم آن است، برای درک بهتر خطا RMSE ریشه دوم MSE را در نظر می‌گیرد.
* Covariance و Correlation دو مفهوم آماری هستند که ارتباط دو متغیر تصادفی را مشخص می‌کنند. در واقع Correlation از تقسیم Covariance بر انحراف معیار دو متغیر تصادفی به دست می‌آید که نشان دهنده عملکرد حدودا یکسان این دو متغیر است. Correlation مقادیر بین منفی یک و یک را به خود می‌گیرد. اگر این مقدار برای دو متغیر تصادفی در حدود یک، منفی یک یا صفر باشد به ترتیب نشانگر وابستگی مستقیم، وابستگی معکوس یا عدم وابستگی دو متغیر است.
* Regression toward the mean پدیده و فرضی آماری است که بیان می‌کند اگر یک نمونه دارای مقدار بالایی در یکی از مقادیر اندازه‌گیری شده‌اش باشد مقدار دیگرش به میانگین نزدیک‌تر خواهد بود.
* L1 Norm به صورت زیر تعریف می‌شود. LASSO Regression روشی است که در آن سعی می‌شود مدل را به طوری آموزش داد که مجموع اندازه ضرایب (L1 Norm) از حدی بیشتر نشود. در نتیجه استفاده از این روش برخی از ضرایب صفر می‌شود؛ در نتیجه علاوه بر Regularization عمل Feature Selection نیز انجام می‌شود.
* L2 Norm به صورت زیر تعریف می‌شود. Ridge Regression روشی است که در آن سعی می‌شود مدل را به طوری آموزش داد که مجموع مربع ضرایب (L2 Norm) از حدی بیشتر نشود. در نتیجه استفاده از این روش، Regularization به این شکل انجام می‌شود که اندازه ضرایب کاهش می‌یابد پس تغییرات تابع نرم‌تر می‌شود.

## بخش دوم



با توجه به تصویر بالا و در نظر گرفتن داده‌های ناقص که در سمت چپ نمودارها روی مقدار صفر نمره اول تا ششم قرار دارند، می‌توان گفت رابطه‌ای میان نمره هفتم و نمره سوم و پنجم به سختی مشاهده می‌شود ولی در مورد نمره‌های دیگر می‌توان رابطه مشخص‌تری را مشاهده کرد. نکته اصلی در مورد رابطه نمره‌ها حضور داده‌های صفر است. بعد از حذف کردن داده‌های ناقص به نمودارهای زیر می‌رسیم که روابط بیشتری بین نمره‌ها را نمایش می‌دهند.



## بخش سوم

راه حل پیشنهادی به این شکل است که برای هر دسته از نمره‌ها با استفاده از داده‌های موجود توزیع t-student (یا برای صرف نظر کردن از پیچیدگی‌ها توزیع نرمال) را به دست آوریم. فرض کنید نمره i-ام و j-ام یک نمونه در دسترسی نباشد و هدف پر کردن اطلاعات i و j آن نمونه باشد. ابتدا موقعیت نمره‌های مشخص شده آن نمونه را در توزیع‌های همان نمره به این صورت پیدا می‌کنیم که چند درصد از جامعه نمره کمتر (یا بیشتر) از او می‌گیرند. حال میانگین درصدهای به دست آمده را حساب می‌کنیم. حال برای پر کردن نمره i-ام به این گونه عمل می‌کنیم که نمره‌ای را می‌یابیم که درصدی از نمرات کمتر (یا بیشتر) از آن هستند میانگین مقادیر یافته شده است و آن نمره را با این مقدار پر می‌کنیم. برای مثال نمونه‌ای نمره سوم نامشخص دارد و نمره اول تا ششم به استثنای سوم به ترتیب p1، p2، p4، p5 و p5 باشند. تعریف می‌کنیم و نمره را برابر مقدار نمره‌ای قرار می‌دهیم که درصد p در توزیع همان نمره مشخص می‌کند. البته در نهایت باید نمراتی که کمتر از صفر و یا بیشتر از بیست هستند را به این مقادیر تغییر دهیم.

## بخش چهارم

نتایج مدل‌های انتخاب شده به ترتیب زیر است.

|  |  |
| --- | --- |
| مدل | میانگین RMSE برای 10-Fold CV |
| Linear Regression | ۲.۳۸۴۴۶۵ |
| Lasso Regression (alpha: 0.001) | ۲.۳۸۴۴۸۶ |
| Lasso Regression (alpha: 0.003) | ۲.۳۸۴۵۰۸ |
| Lasso Regression (alpha: 0.01) | ۲.۳۸۴۳۲۶ |
| Lasso Regression (alpha: 0.03) | ۲.۳۸۳۵۷۴ |
| Lasso Regression (alpha: 0.1) | ۲.۳۷۷۶۸۷ |
| Lasso Regression (alpha: 0.3) | ۲.۳۶۳۹۵۹ |
| Gradient Boosting | ۲.۵۲۰۴۲۲ |
| AdaBoost | ۲.۴۸۴۶۲۱ |
| Random Forest | ۲.۴۸۸۱۹۸ |