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.
- 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.

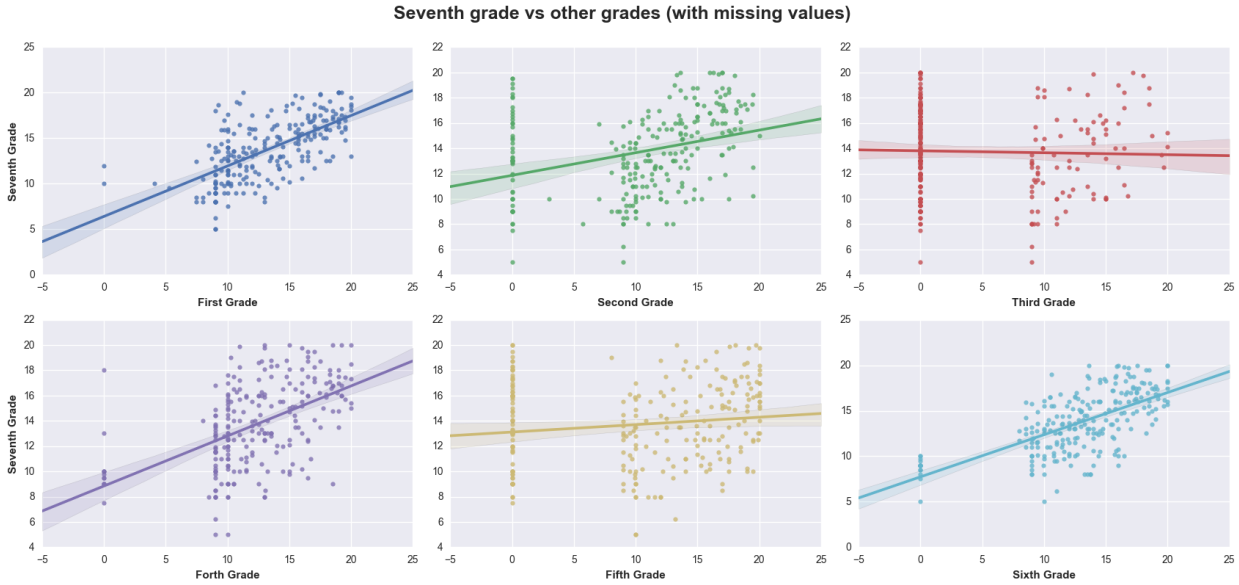
L1 Norm به صورت زیر تعریف می‌شود.

$$\|x\|_1 = \sum_{i=1}^{n} x_i$$

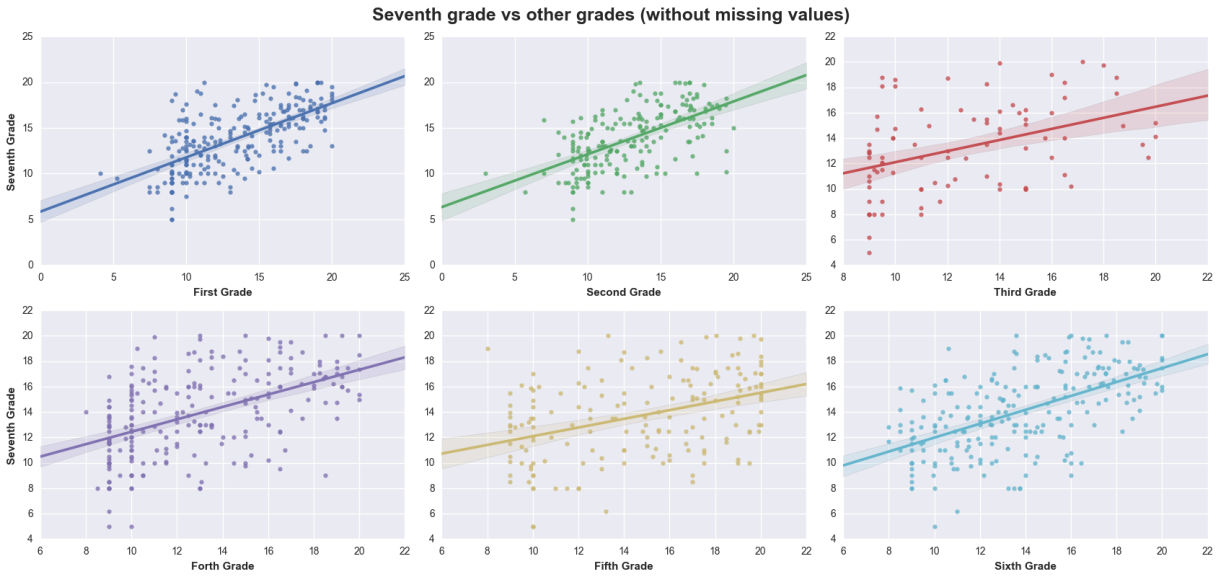
- L2 Norm به صورت زیر تعریف می‌شود. Ridge Regression روشی است که در آن سعی می‌شود مدل را به طوری آموزش داد که مجموع مربع ضرایب (L2 Norm) از حدی بیشتر نشود. در نتیجه استفاده از این روش، Regularization به این شکل انجام می‌شود که اندازه ضرایب کاهش می‌یابد پس تغییرات تابع نرم‌تر می‌شود.

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}}$$

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 $p = \frac{p1+p2+p4+p5+p6}{5}$ 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.

در نهایت با استفاده از متد PCA داده‌ها به پنج بعد کاهش پیدا کردند.

بخش چهارم

نتایج مدل‌های انتخاب شده به ترتیب زیر است.

مدل	میانگین RMSE برای 10-Fold CV
Linear Regression	۲,۳۷۲۹۱۳
Lasso Regression (alpha: 0.001)	۲,۳۷۲۹۳۶
Lasso Regression (alpha: 0.003)	۲,۳۹۲۷۸۸
Lasso Regression (alpha: 0.01)	۲,۳۷۳۳۲۳
Lasso Regression (alpha: 0.03)	۲,۳۷۴۲۸۴
Lasso Regression (alpha: 0.1)	۲,۳۷۸۹۲۳
Lasso Regression (alpha: 0.3)	۲,۳۱۵۷۱۰
Gradient Boosting	۲,۳۲۵۹۱۴
AdaBoost	۲,۴۹۴۰۳۲
Random Forest	۲,۴۱۵۸۸۰

بخش دوم

کنترل کردن داده‌های پرت

برای کنترل کردن داده‌های پرت فرض را برا این می‌گذاریم که ۰,۵ درصد داده‌ها داده پرت هستند. برای انجام این کار مقدار ویژگی را برای داده ۹۹,۵ درصد پیدا می‌کنیم پس از یافتن این مقدار داده‌هایی که مقدار ویژگیشان بیشتر از این مقدار است را به این مقدار تغییر می‌دهیم. برای ویژگی هدف، TotalBsmtSF و GarageArea این روش اعمال شده است.

پر کردن داده‌های ناقص

در مقابل ویژگی‌های مختلف استراتژی‌های مختلفی برای پر کردن داده‌های ناقص در نظر گرفته شد.

استفاده از ویژگی‌های دیگر

برای پر کردن داده‌های گم شده در ویژگی LotArea که از این روش استفاده شده است از ویژگی LotFrontage استفاده می‌کنیم. ابتدا همبستگی این دو ویژگی را به دست می‌آوریم که حدود ۰,۴۲ است. اگر ویژگی جدیدی به نام SqrtLotFrontage بسازیم که ریشه دوم ویژگی LotFrontage باشد و همبستگی آن را با ویژگی LotArea به دست آوریم مقداری حدود ۰,۶۰ به دست می‌آید که همبستگی بیشتری است. پس برای داده‌های گم شده مقدار LotArea را برابر با SqrtLotFrontage قرار می‌دهیم.

مقدار پایه None یا صفر

ویژگی‌هایی وجود دارند که برای خانه‌های امکان وجود ندارند مثلا خانه‌های وجود دارند که اصلا زیرزمین ندارند تا برای آن‌ها این مقادیر را گزارش دهیم. برای ویژگی‌هایی که مربوط به ویژگی‌های از این دست هستند در صورتی که از نوع عددی باشند مقدار صفر می‌دهیم در غیر این صورت None را به آن‌ها نسبت می‌دهیم. برای مثال MasVnrType و MasVnrArea از این دست ویژگی‌ها هستند.

نسبت دادن مد

برای برخی ویژگی‌ها مد داده‌های موجود برای داده‌های ناقص در نظر گرفته می‌شود. یکی از این ویژگی‌های Electrical است.

تغییر دادن مقدارهای دسته‌ای به عددی

ویژگی‌هایی وجود دارند که مقادیر آن‌ها به صورت دسته‌ای ثبت شده است برای مثال ویژگی MasVnrType از این نوع است که دارای دسته‌های BrkFace، Stone، None و BrkCmn است. برای تبدیل این ویژگی‌ها به داده‌های عددی ابتدا ویژگی‌های جدیدی به تعداد انواع موجود می‌سازیم. یعنی برای این مثال چهار ویژگی جدید می‌سازیم و مقدار همه آن‌ها را برابر صفر قرار می‌دهیم. حال مقدار ویژگیی را که مقدار ویژگی اصلی داده برابر آن بود را برابر یک قرار می‌دهیم. یعنی اگر MasVnrType برای یک داده برابر Stone بود مقدار ویژگی‌های جدید آن به ترتیب صفر، یک، صفر و صفر خواهد بود. در مرحله بعدی ویژگی اصلی از داده‌ها حذف می‌شود و ویژگی‌های جدید به آن اضافه می‌شوند. برای سادگی یک حالت پایه برای ویژگی در نظر گرفته می‌شود و آن ویژگی به داده‌ها اضافه نمی‌شود. یعنی برای مثال BrkFace حالت پیشفرض در نظر گرفته می‌شود یعنی یا یکی از ویژگی‌های دیگر مقدار یک خواهند داشت یا اگر همه صفر بودند پس این ویژگی یک بوده است.

انتخاب ویژگی‌ها

برای انتخاب ویژگی‌ها از روش‌های متدهای موجود در کتابخانه scikit استفاده شد. برای این کار از متدهای انتخاب Variance Threshold و Select K Best استفاده شد. ابتدا بهترین عملکرد Select K Best برای Kهای ۱۵۰، ۱۷۵ و ۱۹۰ بررسی شد و با بهترین K ممکن که ۱۷۵ بود با Variance Threshold مقایسه شد. یک روش دیگر نیز با منتخب این دو روش مقایسه شد. برای PCA نیز از کتابخانه بالا استفاده شد و برای انتخاب بهترین تعداد ابعاد تعداد ۱۵۰ و ۲۰۰ مقایسه شدند. در نهایت ۲۰۰ انتخاب شد. بین PCA با خروجی ۲۰۰ بعد و Select K Best با K برابر ۱۷۵، PCA انتخاب شد.

انتخاب مدل

برای انتخاب مدل از مدل‌های زیر استفاده شد که عملکردشان به صورت زیر گزارش می‌شود. این عملکرد در شرایطی است که عمل انتخاب ویژگی انجام نشده باشد.

مدل	میانگین RMSE برای 10-Fold CV
Linear Regression	۰,۱۳۸۲۸۳
Lasso Regression (alpha: 0.001)	۰,۱۱۸۹۵۸
Lasso Regression (alpha: 0.003)	۰,۱۲۸۲۱۹
Lasso Regression (alpha: 0.01)	۰,۱۳۴۹۹۶
Lasso Regression (alpha: 0.03)	۰,۱۴۳۸۶۳
Lasso Regression (alpha: 0.1)	۰,۱۶۵۴۲۹
Lasso Regression (alpha: 0.3)	۰,۱۷۱۹۸۶
Gradient Boosting	۰,۱۲۶۳۳۰
AdaBoost	۰,۱۷۳۹۸۵
Random Forest	۰,۱۴۸۶۸۴

بعد از عمل انتخاب ویژگی‌ها که بهترین آن‌ها PCA با خروجی ۲۰۰ ویژگی بود نتایج به دست آمده به صورت زیر هستند.

مدل	میانگین RMSE برای 10-Fold CV
Linear Regression	۰,۱۲۴۶۷۰
Lasso Regression (alpha: 0.001)	۰,۱۲۲۰۱۸
Lasso Regression (alpha: 0.003)	۰,۱۲۶۴۵۷
Lasso Regression (alpha: 0.01)	۰,۱۵۰۹۹۳
Lasso Regression (alpha: 0.03)	۰,۱۸۹۲۱۸
Lasso Regression (alpha: 0.1)	۰,۲۷۴۹۰۴
Lasso Regression (alpha: 0.3)	۰,۳۹۵۸۰۳
Gradient Boosting	۰,۱۴۹۹۵۹
AdaBoost	۰,۱۹۸۵۶۵
Random Forest	۰,۱۸۵۲۱۴

نتیجه

در نهایت خروجی Lasso Regression با آلفا برابر ۰,۰۰۱ و بدون اعمال کاهش بعد در سایت Kaggle ثبت شد و با امتیاز ۰,۱۲۳۲۸ رتبه ۸۶۶ کسب شد.