Comparison of prediction performance using original data and expanded data

As mentioned earlier in introduction and data collection, we hypothesized that house price not only depends on intrinsic characteristics such as year built, square feet, number of bedrooms and bathrooms, but also depends on the environment it is located. Zipcode largely determines the environment but it is meaningless by itself. To test this hypothesis, we compared the prediction performance using the original data, which contains mainly intrinsic features, and the expanded data with environmental information included, which consists of traffic, crime, population, school and hospital. Four regression models were used here, linear regression, ridge regression, lasso regression and support vector regression. Two kernels were used for support vector regression, “linear” and “rbf”. Except “linear regression”, all other models were optimized with gridsearch and cross validation. Two metrics were used, mean squared error of predicted house price v.s. true house price, and score which is the coefficient of determination R^2 of the prediction.

As shown below, Linear regression and Ridge regression performed the best given their lowest errors and highest scores. The performance of Lasso regression and SVR with linear kernel followed. SVR with RBF kernel performed the worst. By comparing the two SVR models, we can clearly see that the radial basis function kernel, RBF, is not very suitable for our dataset while the linear kernel works much better. More interestingly, for all these linear-based algorithms, the expanded dataset worked better than the original dataset given the lower errors and higher scores. This provided a strong evidence to support our primary hypothesis in this paper that including environmental information will improve house price prediction.

