Yelp-DS: Results Estimation Notes

1. The trade-off between prediction accuracy and model interpretability.
Some models are less flexible, or more restrictive, in the sense that they can produce just a relatively small range of shapes. Flexible approach can lead to complicated estimation that is difficult to understand how any individual predictor is associated with the response. Non-linear methods such as bagging, boosting and SVM with non-linear kernels are highly flexible approaches that are harder to interpret.

In our case, we are using some existed models, so we can choose the model which can produce the highest prediction accuracy, we should try a lot of models and explain the advantages and disadvantages of each one.

2. Regression versus classification problem.
Problems with a quantitative response are regression problems, while those involving a qualitative response are referred to as classification problems.

In our case, due to the fact that the challenge is quite open, we don't have to define the rating prediction problem as a classification problem. We can regard it as regression problem and try to define a formula to estimate the regression results (Here needs more thinking, how to assess the obtained results).

3. Measuring the quality of fit.
In the regression setting, the most commonly-used measure is the mean squared error (MSE), overfitting refers to the case in which a less flexible model would have yielded a smaller test MSE on the training data.

By adjusting the level of flexibility of the smoothing spline fit, we can produce many different fits to this data, can't we?

In our case, we mainly use SVM and Random Forests as our classifiers, how to visualize the fitness of each model is pretty meaningful. Besides, cross-validation can be used as a method for estimating test MSE using the training data.

4. The bias-variance trade-off.
Variance refers to the amount by which the model would change if we estimated it using a different training set. In general, more flexible statistical methods have higher variance. Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. Generally, more flexible methods result in less bias. As a general rule, as we use more flexible methods, the variance will increase while the bias will decrease. One should always keep the bias-variance trade-off in mind. Ideally, we need to select a statistically learning method that simultaneously achieves low variance and low bias.

Cross-validation is a way to test whether the variance of a model is stable, how to test the bias of model needs more thinking.

5. Cross-validation and The bootstrap.

Cross-validation and the bootstrap are two commonly used resampling methods.
Leave-one-out cross-validation can be used with any kind of predicting modeling, k-fold CV has a computational advantage to LOOCV, there is a bias-variance trade-off associated with the choice of k in k-fold cross-validation.

The bootstrap relies on random sampling with replacement, it can be used to quantify the uncertainty associated with a given estimator or statistical learning method.

In our case, we can both cross-validation and the bootstrap method to generate the training data and testing data, then we can measure the variance of the estimator.

6. Predictor subset selection.

Some methods for selecting subsets of predictors (predictor combination as a model):

(a) Best subset selection: fit a separate least squares regression for each possible combination of the P predictors, cannot be applied with very large P.
(b) Forward stepwise selection: begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. At each step the variable that gives the greatest additional improvement to the fit is added to the model.
(c) Backward stepwise selection: begins with the full least squares model containing all P predictors, and then iteratively removes the least useful predictor, one-at-a-time.

In the Random Forests case, it has a combination of many decision trees, which can be looked as individual predictors. We can combine many classifiers as a strong classifier, as bagging and boosting do.

7. Randoms forests.

Tree-based methods are simple and useful for interpretation, by aggregating many decision trees, using methods like bagging, random forests, and boosting, the predictive performance of trees can be substantially improved. Bagging, random forests, and boosting use trees as building blocks to construct more powerful prediction models.

Maybe it's possible to combine several classifiers as a boosted classifier, which can give a higher prediction rate.

8. R-square.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model

In our case, if we use regression model to carry out the rating prediction, we can apply R-square measure to estimate the model.