# Milestone 4 Baseline models

We used the lasso regression (available in sci-kit learn) and xgboost’s implementation of the gradient boosting machine (gbm). For feature engineering, we one-hot-encoded all the categorical variables and filled in the missing values in 5 of the variables with -1. It should be noted that xgboost has built-in mechanisms to deal with missing values and all one has to do is to indicate to the algorithm how missing values are represented in the data.

A large part of this milestone is focused on finding the optimal hyperparameters of the algorithms. For lasso, it will be the alpha (the larger the alpha, the greater the amount of regularization and the lower the model complexity). For xgboost, there are many hyperparameters and the important ones are max\_depth (maximum depth of each), eta (the learning rate). Because there are a large of hyperparameters to tune, we used the optimization package, hyperopt.

To select the optimal hyperparameters, 5-fold cross validation is used. The best 5-fold rmspe of xgboost is 0.16208736 with a standard deviation of 0.097645484 whereas the best 5-fold rmspe of lasso is 0.491645700 with a standard deviation 0.07904774.

Figure 1 and Figure 2 show that for xgboost, we should probably increase the max\_depth in order to increase the performance. Figure 3 and Figure 4show that for lasso regression, smaller values of alpha give superior performance.

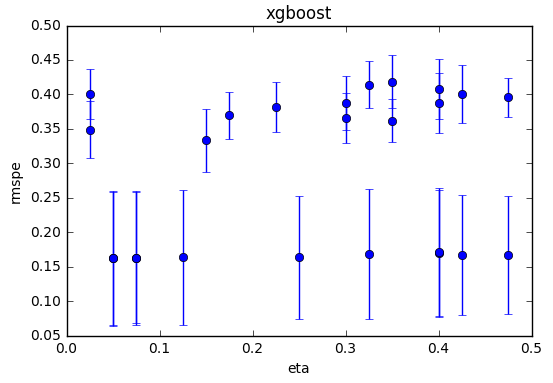


Figure 5-fold cross validated rmspe of xgboost with varying eta

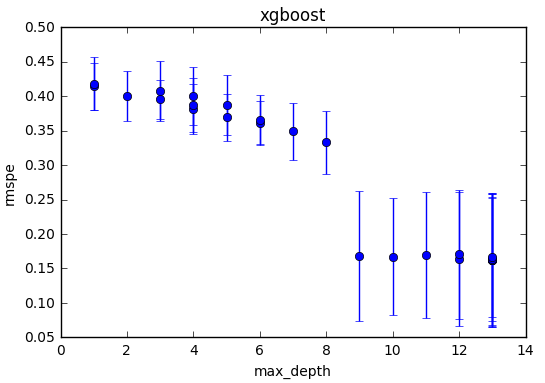


Figure 5-fold cross validated rmspe of xgboost with varying max\_depth

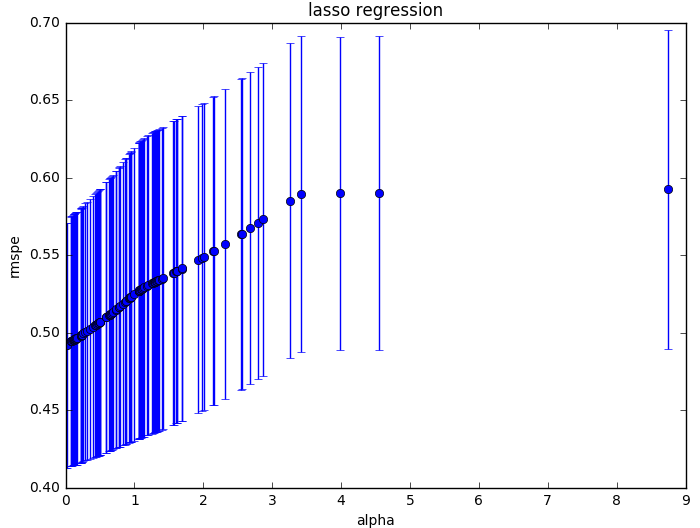


Figure 5-fold cross validated rmspe of lasso regression with varying 0 < alpha < 10

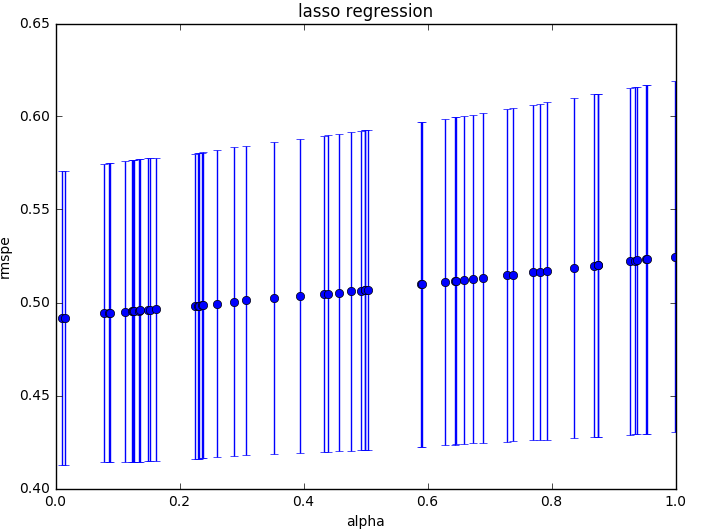


Figure Figure 3 5-fold cross validated rmspe of lasso regression with varying 0 < alpha < 1

# Milestone 5 Proposal of future work

Future work on this dataset can be divided into two parts, feature engineering and modelling.

## Feature engineering

For lasso regression, we can standardize the numerical variables so that each of them has a mean of zero and a standard deviation of one. This is useful as large coefficients are penalized in lasso and standardization will ensure that variables with a larger range of values are not unduly penalized.

For both xgboost and lasso regression, we can log the independent variable, sales and see if it improves the performance.

## Modeling

For xgboost, we can increase the max\_depth parameter as Milestone 4 seems to imply that max\_depth used is too low.

We can also combine the different models, lasso regression, random forests, and xgboost, by combining the predictions by using a meta-classifier, similar to that used in HW8. The two types of meta-classifiers that are appropriate for this dataset include simple averaging and weighted averaging .