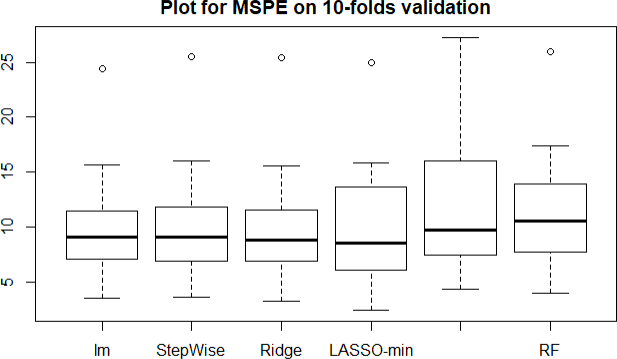
Model fitting

# Model fitting

One of the problems that we encountered after analysing the data, is that some predictors have a weak correlation with our response variable. So we considered using linear model, step-wise algorithm, random forest and LASSO regression to predict the values of Y. Moreover, we used cross-validation to evaluate our models, the performance of models were based on the results of the MSPE.



MSPE is shown in Figure X for each model. The boxplot represents that there were remarkable variability between each fold for all models. The standard deviation were large for all models. We observed that all models except

LASSO-lse gave similar predictions. In conclusion, we suggest that cross- validation isn't remarkabley useful because of the limited sample data and weak correlation between response variables and explanatory variables.

# Model evaluating

Due to the limited data sample, variability and overfitting were the main issues that our data set has. We decided to focus on 3 models that could handle these issues well. We used kaggle public leaderboard score as an unbiased estimator **for the final private leaderboard score.**

# Conclusion

In this competition, we realized that data preprocessing plays the most important role in predicting the result. However, model selection does not affect the prediction significantly in our analysis. The Limited data sample might be the reason. Furthermore, we believe that the correlation between

explanatory variables and the response variable is weak. We obtained our best score from the **simplest??** model. In the future competition, we will put more effort into data preprocessing and feature engineering.

我们是不是应该整个EDA，我看⼤部分都需要呢。

⾸先，有⼏个问题，我们是不是应该加⼀个feature selection什么的，再搞个correlation 的图，因为我们多次⽤到这个观点。

Random Forest是我们认为最好的模型，你觉得为什么它表现的最好？ Random forest哪些parameter的选择，要写进去不

除了sample size，weak correlation以外，还有哪些注意的点？ parameters tuning要解释不

还有⼀些，⽐如我们expand 了training data，这些点要不要也写进去

