We all know infamous bias-variance trade-off and how important it is. One of the steps in data-driven modelling where we address this trade-off is variable selection where we try to identify the significant variables for our model and eliminate the rest of them. By doing so, we make sure we determine the “optimum” variables sets that correspond to the optimum complexity of our model addressing bias-variance trade-off.

There are a number of algorithms out there that are used for variable selection: forward stepwise elimination, backward stepwise elimination, lasso, ridge and so on. One of my favorite approaches is forward stepwise elimination where you start with finding the variable that maximizes the accuracy and then you keep increasing the number of variables as long as accuracy of the model increases (i.e., greedy-search optimization). There are, however, issues with this approach. The first one is that interaction of the variables are overlooked. In other words, it does not check the all the combinations. It evaluates them one by one. However, as stated in the Elements of Statistical Learning, a variable that may not work well standalone can lead to higher accuracies with the interaction of another variable. The second issues is that which prediction algorithm to use while selecting the variables. Conventionally, linear regression is preferred (e.g., forward stepwise regression) because its computationally cheap. These two issues are motivation behind this blog. To sum up, this blog has two objectives:

1. We know that selection of ML algorithm should play a role in variable selection. The first objective is try to quantify/validate how important it is.
2. The second objective is to check if stepwise approach (i.e., greedy search) is good approximation global solution found by brute force.

**Data Sets**

Here I wanted to introduce the datasets I used for the experiments (all the code and data can be found here). I used two datasets: one for regression and one classification. The dataset used for regression is about predicting housing prices in Boston and has 13 explanatory variables. The dataset used for classification is about predicting the onset of diabetes within 5 years in Pima Indians. I got these datasets from this blog entry. Here are the index of the each explanatory variables and its explanations.

Boston Housing Price Dataset:

1. CRIM: per capita crime rate by town.
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of nonretail business acres per town.
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
5. NOX: nitric oxides concentration (parts per 10 million).
6. RM: average number of rooms per dwelling.
7. AGE: proportion of owner-occupied units built prior to 1940.
8. DIS: weighted distances to five Boston employment centers.
9. RAD: index of accessibility to radial highways.
10. TAX: full-value property-tax rate per $10,000.
11. PTRATIO: pupil-teacher ratio by town.
12. B: 1000(Bk – 0.63)^2 where Bk is the proportion of blacks by town.
13. LSTAT: % lower status of the population.

**Pima data set**

1. Number of times pregnant.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg).
4. Triceps skinfold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (weight in kg/(height in m)^2).
7. Diabetes pedigree function.
8. Age (years).

**Studying the impact of variable selection**

For determining the optimum variable set, I used five fold validation and the error calculated based on “out-of-fold” predictions was used in decision making.

For regression datasets, I used linear regression, svm with rbf kernel, decision trees and neural network. I tried all combinations of the explanatory variables which results in fitting 2n number of models (n is number of explanatory variables). Figure 1 presents the results for regression problem. Each red point in the figure refers to a single combination given number of variables. As clearly seen from the figures, even the size of optimum variable set varies across the algorithm. While ANN selected 9 variables for the optimum set, SVM\_rbf selected only 2. The list of the selected variables is given below. It varies significantly between the algorithms. For example, 13th variable is selected by all the algorithms other than the linear regression which can be interpreted (by stretching it a lot) that its effect is non-linear on the predictor. Another example is that the second variable was selected by linear regression and ANN but not by SVM and decision tree.

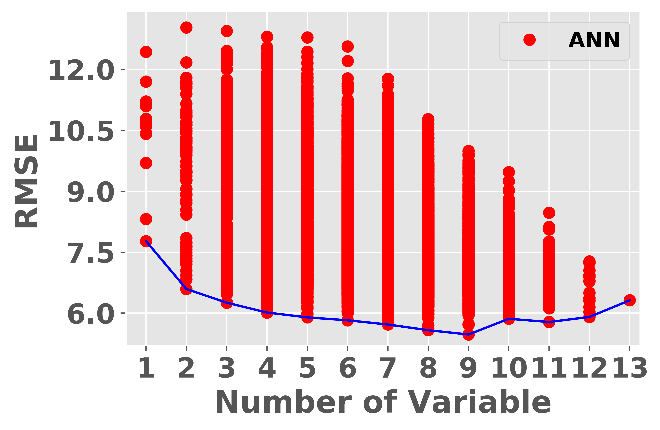
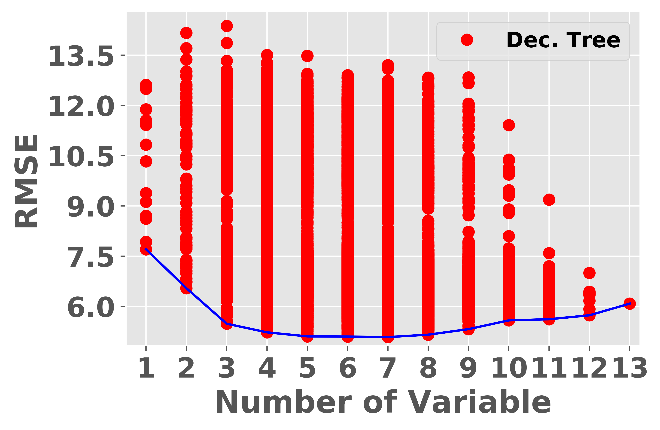
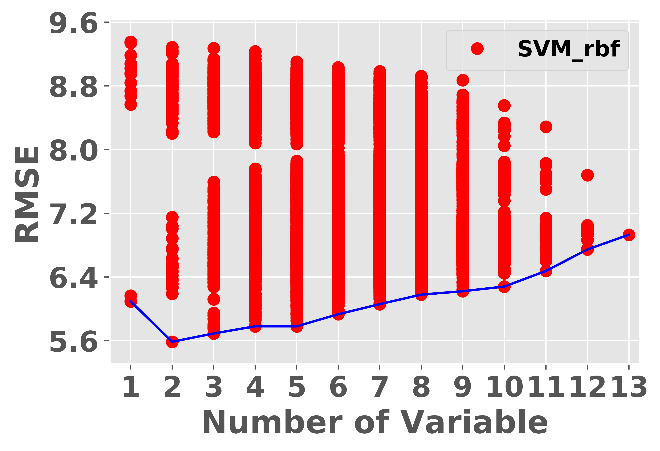
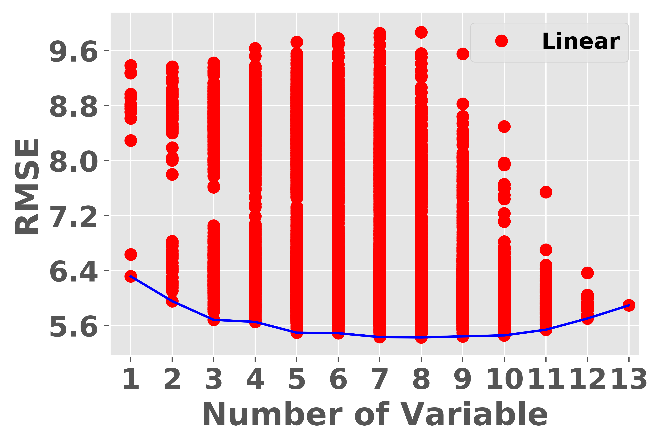
Selected Variables:

Linear Regression: [2, 6, 7, 8, 9, 10, 11, 12]

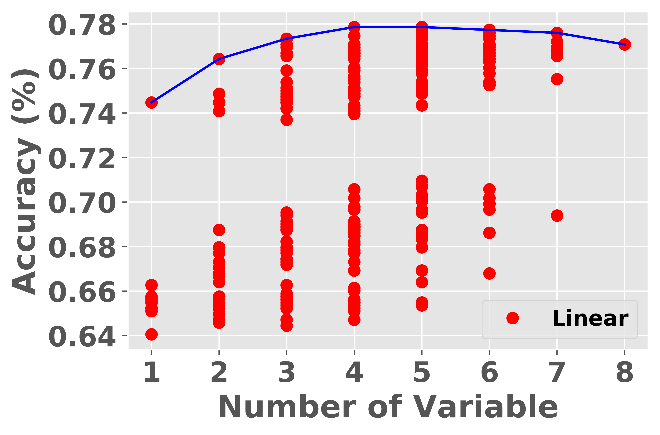
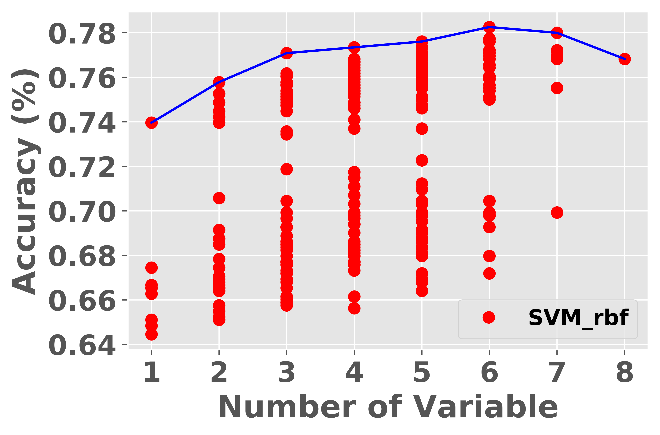
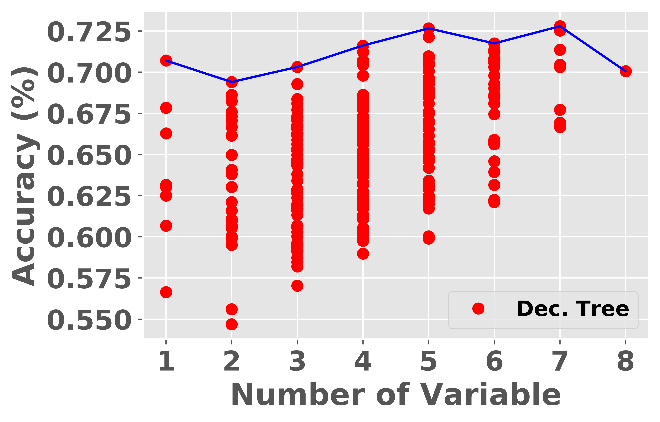
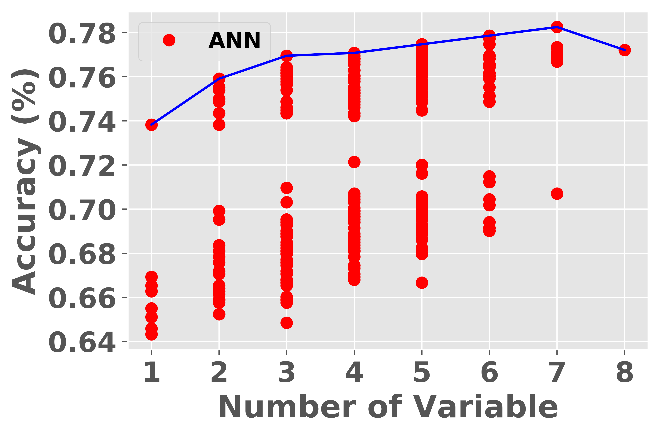
SVM\_ref: [6, 13],

Dec\_tree: [1, 4, 5, 6, 9, 10, 13],

ANN = [1, 2, 4, 5, 6, 8, 9, 11, 13],



The results for classification problem are given below. As seen from the graphs, similar observations can be made for this dataset too. It is interesting to see that ANN selected the highest number of variables for both problems. **But it is very important to note that** I ran these models without parameter tuning. I just used default values. Tuning may change the trends in the data. Maybe a topic for another blog entry 😊.



**Studying the impact of variable selection**

In this part, the stepwise approach is compared by the brutal approach. In brutal approach, optimum set was determined though all combinations of the explanatory variables (complexity 2n). In stepwise approach, the optimum variable set was found in a greedy-way (complexity n). The results are presented below. It turns out stepwise solution is not a bad approximation to the brute force approach, that can be considered as the best solution. ANN especially found the same set for both brute force and stepwise cases which made it the most consistent ML algorithm for variable selections



**Summary and Conclusions**

In this blog, I experimented the impact of ML algorithm selection and stepwise search (i.e., greedy search) on variable selection. It is important to state that all the conclusions and observations are specific to two databases I used and they may not generalize. Furthermore, I did not do hyperparameter optimization which may change the course of the results.

First observation is that selection of ML algorithm plays a significant role in variable selection. Conventionally, linear regression is selected for variable selection because it is computationally cheap. However, the optimum set differs quite significantly w.r.t. the results. Generally, non-linear models e.g., ANN and decision trees showed similar results. The second main observation is that the stepwise regression is not a bad substitute to the brute force search. Especially, the variable set that was found by ANN was exactly the same for stepwise and brute force.