**Introduction:**

**Random forests** are an [ensemble learning](http://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](http://en.wikipedia.org/wiki/Statistical_classification), [regression](http://en.wikipedia.org/wiki/Regression_analysis) and other tasks, that operate by constructing a multitude of [decision trees](http://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](http://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of [overfitting](http://en.wikipedia.org/wiki/Overfitting" \o "Overfitting) to their training set.

**Random Forest Regressor:**

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

**Jist:**

For the data set that has been provided for the Seattle Grace Hospital by the Chief Oncologist , a machine learning model is build that can predict if a tumor is malignant or benign.

The Random Forest method has been applied for the better results when compared to the linear Regression and the Decision tree.

The prediction is made by considering the categorical values that are important and by removing the unimportant values from the data set that has been provided.

I feel the *Unnamed:0* and *id number* are not important in the Breast Cancer Prediction. So, I have dropped these two attributes. The reason for this is, the id number is no way useful to predict if a tumor is malignant or benign.

**Most Important:**

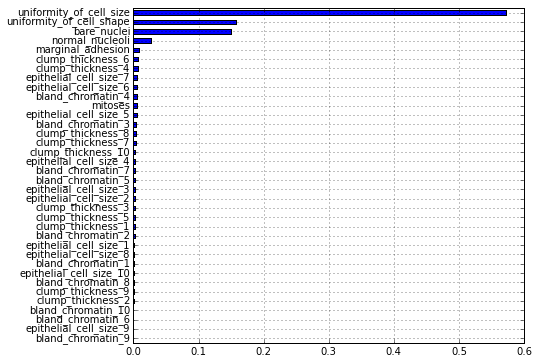
I feel the categorical variables that I have considered are the most important. They are 'clump\_thickness', 'bland\_chromatin' and 'epithelial\_cell\_size'.

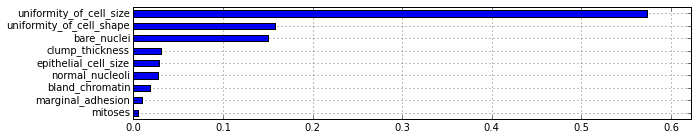
I have taken these categorical variables to build the model to obtain the highest C-stat value to check the model consistency.

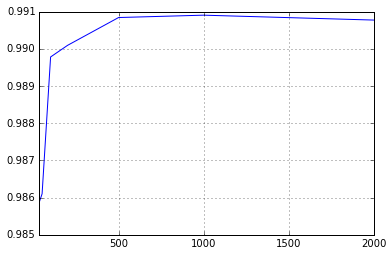
For the categorical variables that I have taken, the C-stat value obtained after trying it with different parameters is

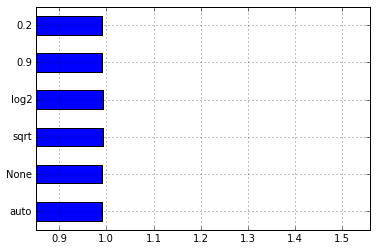
0.992185942851

**Graphical Representation**: Random Forest Regressor



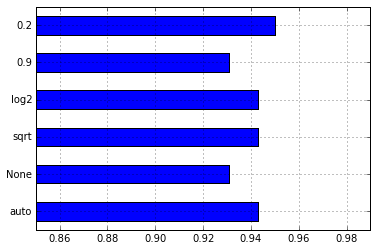
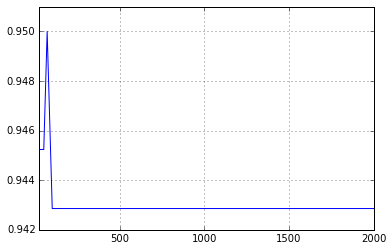






**Random Forest Classifier:**

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.



Using Random forest classifiers, I have built the model by taking the same categorical variables that I have taken for the Regressor model. The C-stat value that I have obtained is 0.95 which means the model that I have built is 95% consistent.

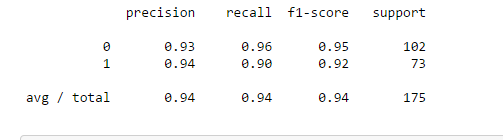
There is a 5% chance that my model might miss the malignancy of tumor.

After comparison with various categorical variable, I found that epithelial cell size and bare nuclei plays a major role in building the best model to predict the malignancy.

Model Comparison:

**Precision**: Fraction of correctly identified examples of a class. The ratio of true positives to all the positives

**Recall:** Fraction of observations classified in the class that was correctly classified.



**Conclusion:** When both the models are compared i.e., the Random Forest Regressor and Random Forest Classifier, I feel that the model obtained with the Random Forest Regressor is the best model for the prediction of the Breast Cancer. As there is very less chance of false predictions with the model that has been build with the Random Forest Regressor. The model is 99% successful in predicting whether the tumor is malignant or benign.

**Source:**

http://en.wikipedia.org/wiki/Random\_forest

http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html