**‘Predicting Breast Cancer’ Assignment**

**Data Analysis Report**

**By**

***Sai Yallapragada***

**For the consumption of**

**The Chief Oncologist**

***Seattle Grace Hospital***

**Executive Summary**

This report provides an in-depth analysis and recommendations regarding the reliability of the new *ScanWare Systems* developed by the *WestCoast Technologies Inc.* that identify the presence of malignant tumors in human breast tissues. The analysis and recommendations are requested by the *Seattle Grace Hospitals* which might consider using the new technology if it’s considered reliable in predicting the malignant tumors. The data used in this analysis is provided by the *Seattle Grace Hospitals* and recommendations are provided after exhaustive data analysis using widely used and industry standard data analysis techniques and methodologies. The model uses various regression techniques to check the accuracy of the models and settles on the Random Forest regression method.

After performing data analysis on the provided data we found that the new technology *ScanWare Systems* is highly reliable in predicting the presence of malignant tumors, as the data study shows high confidence level with an accurate fit and the model came out with high accuracy. Thus, the analysis recommends using the *ScanWare Systems* developed by the *WestCoast Technologies.*

**Introduction**

*WestCoast Technologies Inc.* is a technology company that came out with a new technology called *ScanWare Systems* that can scan a patient’s tumors and determine if its malignant or not. *Seattle Grace Hospital* is considering using this technology for their hospitals but wanted to test the accuracy of this new technology. They approached us and provided us with the data and wanted to find out the accuracy of this new technology with the help of the data. The data contains various variables that represent the physical characteristics of tumors and variable that shows if that tumor is malignant or not. We are going to perform data analysis on this data using various data regression methods to determine the accuracy of this data.

**Data Summary**

The provided data set contains 699 entries and the following variables-

* *id number*
* *clump\_thickness*
* *uniformity\_of\_cell\_size*
* *uniformity\_of\_cell\_shape*
* *marginal\_adhesion*
* *epithelial\_cell\_size*
* *bare\_nuclei*
* *bland\_chromatin*
* *normal\_nucleoli*
* *mitoses*
* *malignant*

For the purpose of this study we are going to consider all these variables as independent variables and only the malignant variable as dependent variable. The goal of this analysis is twofold-

1. To determine if the dependent variable malignant has any correlation to the provided variables, if so, can we measure the accuracy of this correlation? In other words, can we confidently say that the provided data shows an relation to the state of the tumor, whether it is malignant or not
2. If the model suggests that these independent variables together are indicative if the tumor is malignant or not, which of these variables(s) have strongest correlation to the malignant variable

**Data Analysis Process**

The process of analyzing the data involves the following. Without adjusting the data or adding any features to the prediction, we ran different types of regression tests to see if we are closer to a prediction

1**) Linear regression** – This is defined as follows-

*“In a cause and effect relationship, the****independent variable****is the cause, and the* ***dependent variable****is the effect.****Least squares linear regression****is a method for predicting the value of a dependent variable*Y*, based on the value of an independent variable*X*.” [1]*

What we are looking from the linear regression model is the Coefficient of determination or R2 value. R2 value is defined as

*“The****coefficient of determination****(denoted by R2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.*

* *The coefficient of determination ranges from 0 to 1.*
* *An R2 of 0 means that the dependent variable cannot be predicted from the independent variable.*
* *An R2 of 1 means the dependent variable can be predicted without error from the independent variable.*
* *An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R2 of 0.20 means that 20 percent is predictable; and so on.” [2]*

**The result of this linear regression is a coefficient of determination (R2) of 0.73**. This indicates that the model is a decent model although far from accurate.

2) Secondly, we ran the **Decision Tree Regression analysis**. Decision Tree regression is defined as –

“*Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with****decision nodes****and****leaf nodes****. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target.* ” *[3]*

**The result of this decision tree regression is a coefficient of determination (R2) of 0.73**. This indicates that the model is a decent model although far from accurate.

3) Next up, we ran the **Random Forest Regression analysis**. Random Forest Analysis is defined as –

*“Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of over fitting to their training set.” [4]*

**The result of this random forest regression is a coefficient of determination (R2) of 0.73**. This indicates that the model is a decent model although far from accurate.

So far, the three types of regression came out the same with raw data that was provided but the prospects seem very good considering that the Coefficient of determinations seem to indicate a decent model without any cleanup of the data or using advanced data techniques.

**Data clean up**

We ran an **Exploratory Data analysis (EDA)** on the provided data to see if it needs any clean up.

1. Step 1- check if there are any missing values for any of the 699 data as non-existent data can have some impact on the accuracy of the model. There are no rows that have missing data. So, we don’t need any data substitution of other techniques.
2. Step2- check if there are any categorical variables in the data set. Categorical variables are the variables that are not numeric in nature but can be a text based value and/or can have number of possibilities. Since data analyzing algorithms understand only numerical data, it is important for us to find if there are any categorical variables and come up with a solution of representing them numerically. In the provided dataset, there are no categorical data variables. So, we don’t need to perform any conversion.

**Data Model Tuning**

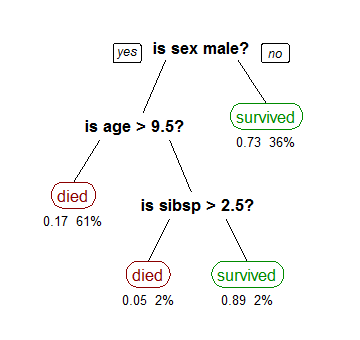
The process of fine tuning the model involves two steps

1) Step1- Run the industry famous Random Forest regression analysis and establish a bench mark.

We ran the Random Forest regression technique by passing the *n\_estimators* as 100 and *oob\_score* as ‘True’.

n\_estimators-This is nothing but the number of decision trees that we are using in this model. Decision trees can be described as

*“Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.” [5]*

*[5]*

oob\_score-In the random forest model, about one-third of the observations get left out in the regression. These left out observations are called out of bag (oob) observations and can be used for prediction the true value of observation.

In this step 1, we got a roc\_auc score of 0.990. This is very high. Roc\_auc score represents the accuracy of the model. The score of 0.990 is an excellent but we tried to see if we can improve this further

2) Step2 – Try various n\_estimators numbers to see if we can improve the roc\_auc score. Here are the results

30 trees

C-stat: 0.989848520539

50 trees

C-stat: 0.989354762724

100 trees

C-stat: 0.99082697639

200 trees

C-stat: 0.990224501259

500 trees

C-stat: 0.990165612713

1000 trees

C-stat: 0.990351338129

2000 trees

C-stat: 0.990210911595

Based on the above numbers, we settled on selecting 100 decision trees.

3) Step 3- Try various *features*(described below) numbers to see if we can improve the roc\_auc score. Here are the results

auto option

C-stat: 0.99082697639

None option

C-stat: 0.99082697639

sqrt option

C-stat: 0.989277754625

log2 option

C-stat: 0.989277754625

0.9 option

C-stat: 0.990056895396

0.2 option

C-stat: 0.99233995905

Based on the above numbers, we settled on selecting 0.2 features.

*Features* – it is nothing but the number of features to consider when looking for the best split.

4) Step 4- Try various number of leaves (explained below) to see if we can improve the roc\_auc score. Here are the results

1 min samples

C-stat: 0.99233995905

2 min samples

C-stat: 0.991284495099

3 min samples

C-stat: 0.992272010727

4 min samples

C-stat: 0.991193897335

5 min samples

C-stat: 0.992063635869

6 min samples

C-stat: 0.990387577235

7 min samples

C-stat: 0.991121419123

8 min samples

C-stat: 0.990786207396

9 min samples

C-stat: 0.990170142601

10 min samples

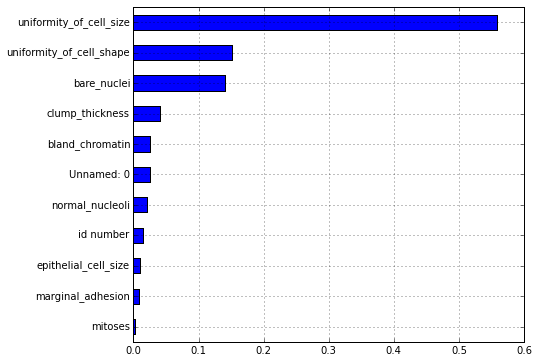
C-stat: 0.99055065321

Based on the above numbers, we settled on selecting 1 leaf.

*Leaves -* The minimum number of samples in newly created leaves.

**So, we have the final model all the above efficiency improvements. The final accuracy score is 0.9923, which is an excellent model and indicates that the model is highly accurate.**

As a final step, we ran a correlation test to see which of the variables provided in the data set has highest correlation towards predicting of the tumor is malignant or not. And the result is ‘uniformity of cell size’, as you can see from the below picture.

****

**Recommendations & Conclusion**

After the thorough analysis of the provided data, we highly recommend the usage of the *ScanWare Systems* developed by the *WestCoast Technologies Inc*. as this system seems to be highly accurate in predicting the presence of malignant tumors based on the data provided for this analysis. The study also concludes that the ‘uniformity of cell size’ seems to have major impact in determining if the tumor is cancerous or not.

**Summary – Q&A**

1. Were you successful in attempting to create a machine learning model to predict malignant tumors.

***Yes – the model clearly establishes that it’s an excellent machine model and can successfully predict malignant tumors****.*

1. The hospitals lawyers are VERY careful and are worried about the accuracy of your model.  You'll need to inform them of the risks of using your system, and possibly convince them of its safety.

***This model, like any model, carries some risk. But within the given parameters of the data science, this is certainly one of the top accurate models we can have using the Random Forest model.***

1. How good is your model?

***The model is very good. It is good to the point that the C-Score stands at 0.99, which is almost perfect****.*

1. How likely is it falsely predicting breast cancer?

***There is a chance of roughly 1 out of 100 cases where the model can falsely predict presence of malignant tumor***

1. How likely is it to miss a malignant case?

***There is a chance of roughly 1 out of 100 cases where the model can miss the presence of malignant tumor***

1. The Chief Oncologist needs to be convinced that the system is making 'realistic' choices and wants to understand which variables are the most important in predicting cancer.   Explain or show the Doctor which variables are most important.

***The system is making a truly realistic choice using the provided data. This can be substantiated by the data variable ‘uniformity of cell size’ which seems to have major impact in predicting the presence of malignant tumor. What this translates to is – the tumor being malignant or not seems to have a direct correlation to the uniformity of the cell sizes inside the tumor area.***

**Bibliography**

1. <http://stattrek.com/regression/linear-regression.aspx>
2. <http://stattrek.com/regression/linear-regression.aspx>
3. <http://en.wikipedia.org/wiki/Random_forest>
4. <http://www.saedsayad.com/decision_tree_reg.htm>
5. <http://en.wikipedia.org/wiki/Decision_tree_learning>