**‘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 anlysis 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 first ran a linear regression analysis on the data.

Linear regression is defined as-

*“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*

Secondly, we ran the Random Forest Regression analysis. Random Forest Analysis is defined as

**Data Mining**

**Data Analysis Techniques**

**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 seem 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>