**Keep It Simple Stupid: Lesson in Model Selection**

I was given a project at General Assembly where I was given physical breast tumor data, and was asked to predict whether the tumor was malignant or benign. I was specifically asked to compare two different modeling techniques in this assignment, and explain why I picked one over the other. The two techniques I started with were logistic regression and random forest. Logistic regression was the ultimate winner in this case because random forest in the end didn’t offer any more predictive power and would have been tougher to explain the inferences to the mock client.

**Model Creation**

*Heatmap Image*

After scaling the data, created a heatmap, which showed I had little wiggle room in feature selection for my logistic regression. Pretty much everything was highly correlated except a few variables and using the majority of them would mess up the results. I ended up picking three variables (radius mean, concavity worst, and symmetry worst), and added an interaction term. The results were the following:

***Logistic Training Set Results***

**Accuracy:** 96%

**Sensitivity:** 94%

**Specificity:** 97%

***Logistic Test Set Results***

**Accuracy:** 94%

**Sensitivity:** 89%

**Specificity:** 97%

I was pretty happy with these results, but needed to try another model according to the assignment. Random forest was an algorithm that I used pretty frequently, and I didn’t have to worry about independence between my variables, so I gave that a shot. The results were slightly better in the training set, but the test set results matched that of the logistic regression:

***Random Forest Training Set Results***

**Accuracy:** 98%

**Sensitivity:** 95%

**Specificity:** 99%

***Random Forest Test Set Results***

**Accuracy:** 94%

**Sensitivity:** 89%

**Specificity:** 96%

**Model Selection**

Logistic regression was the clear winner in this case because random forest it was simpler in the following ways:

1. **Logistic Regression is a Simpler Algorithm**

Most people familiar with limited exposure to data science know that linear regression is drawing a line through a dataset. Well then, it’s not that far of a leap explaining that logistic regression is the same thing except the target is bound by 0 and 1, and the result is the probability that the record is the target class you are trying to predict.

This is versus a random forest, which is an ensemble method. I think most people could grasp the concept of a single decision tree. But explaining why you create multiple trees in order to prevent overfitting and how they are combined is a little tougher. Between the two algorithms, I would prefer to explain a logistic regression to a client.

1. **Logistic Regression Offers Simpler Inferences**

Just like linear regression, you can get the coefficients from a logistic regression. We can see here that an increase in concavity worst results in a higher probability that the tumor is malignant.

*Logistic Image*

*Note: I know that radius mean and the interaction term are not significant, but removing them affected the predictive power of the logistic regression.*

This is versus a random forest where you can only extract which features are important. You can’t say how they correlate with the target because a spit for a variable in one tree may contradict a split in another tree.

Keep it simple stupid is an important lesson to remember in a world where neural networks and other complex algorithms are the hype these days. Although they may be powerful, if the same results can be achieved with something simpler, you might get the added bonus of being able to explain it and its results easier to the client.

If you would like to see jupyter notebook associated with this project, check out my portfolio page.