Lindsay Sheppard

DSC 609: Machine Learning

Final Project

**Introduction**

Cancer is a disease in which cells mutate and divide out of control. Aside from skin cancer, breast cancer is the most common cancer in women in the United States (CDC, 2017). In 2019, an estimated 268,600 new cases of invasive breast cancer will be diagnosed in women, as well as 2,670 cases in men (Breastcancer.org, 2019). The use of machine learning techniques in medicine is expanding as researchers seek to find ways to detect and treat cancer early to improve survivability. Classification algorithms such as random forest, support vector machine, and artificial neural networks have been used successfully for the early detection of other medical illnesses (Vala, 2018). The report aimed to build these classification models as a way to potentially detect breast cancer in new patients based on certain characteristics of classified biological breast tumor samples

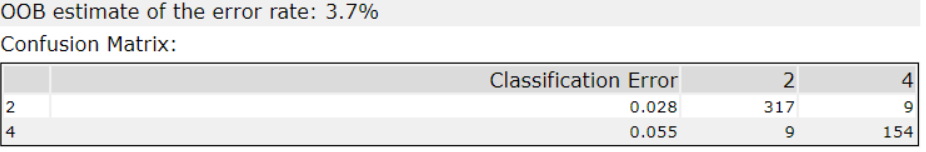
**Data**

Alteryx Designer version 2018.3.5.52487 investigated a University of Wisconsin Hospitals breast cancer dataset titled “Wisconsin Breast Cancer Database”, published originally by Wolberg and Mangasarian (1990). The dataset contained 699 records and 16 missing attribute values, according to the metadata text file. The target variable, *class*, included two values: 2 indicated “benign” and 4, “malignant”. The remaining variables were as follows: *id\_number, clump\_thickness, uniformity\_cell\_size, uniformity\_cell\_shape, marginal\_adhesion, single\_epithelial\_cell\_size, bare\_nuclei, bland\_chromatin, normal\_nucleoli,* and *mitoses*.Each variable except *class* and *id\_number*, the sample code number, contained values between 1 and 10.

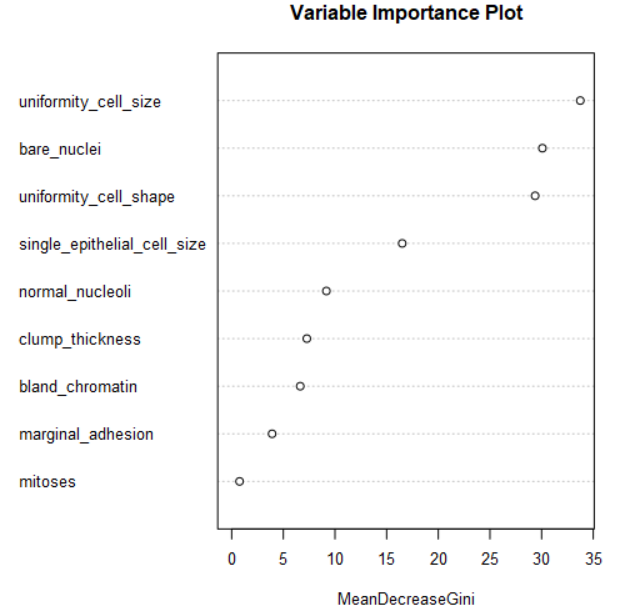
The Auto Field tool converted each variable from strings to bytes, with the exception of *bare\_nuclei*. The field *bare\_nuclei* contained the 16 missing values discussed by the metadata file, which is why the Auto Field tool left it a string data type. As it was only 2% of the data, the missing values were replaced using an imputed mean; a Formula tool converted the “?” values to nulls, then *bare\_nuclei* was changed to a byte and the *class* column to a string, as it was actually binary categorical. The Imputation tool imputed the mean of the column where nulls were present. The average of *bare\_nuclei* was approximately 3. The dataset was split 70/30 into estimation and validation sets, respectively.

**Random Forest**

The Forest Model tool was connected to the estimation set. Scaling is not necessary for random forest. The model was named FM\_Class and the target variable set to *class* with all other variables except *id\_number* as predictors. The default number of trees (500) remained, as well as all other defaults, i.e. the overall size of each model tree was not limited, with the minimum number of records allowed in a tree node equal to 5. Prior to testing the model, it reported a low out-of-box error of 3.7% and provided the confusion matrix shown in Figure 1. The matrix showed that the model correctly classified 317 values in class 2 and 154 values in class 4. The variable importance plot (Figure 2) determined that the most important variable was *uniformity\_cell\_size*, followed by *bare\_nuclei*, and *uniformity\_cell\_shape*. The model was validated; the results of all models against the validation set and a discussion of each model can be found in the Comparison section of the report.



**Figure 1**: Random forest results on estimation set



**Figure 2**: Variable importance plot, random forest

**Support Vector Machine (SVM)**

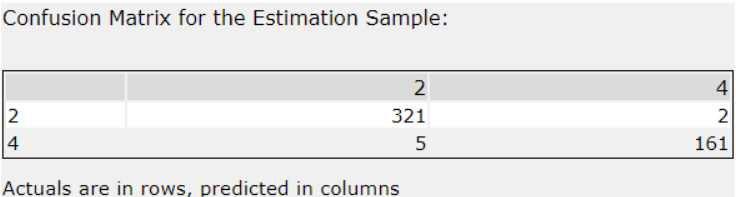
SVM was performed on the data using the C-classification method with no model customization and default cost and gamma parameters equal to 1 and 0.11, respectively. Results on the training data suggested the model performed quite well (Figure 3). Based on the confusion matrices alone, it appears that SVM performed better than the random forest model, predicting 319 values in class 2 and 160 values in class 4 correctly.



**Figure 3**: Support vector machine result on estimation set

**Artificial Neural Network (ANN)**

The ANN model was titled ANN\_Class. Default settings were maintained for the Neural Network model; the number of nodes in the hidden layers were kept at 10. These settings created a 9-10-1 network with 111 weights, with c = 0.7, decay = 0.1, and max iterations = 100. The confusion matrix (Figure 4) suggested that on the training data, the ANN performed the best with 321 correctly predicted class 2 values and 161 correct class 4 values.

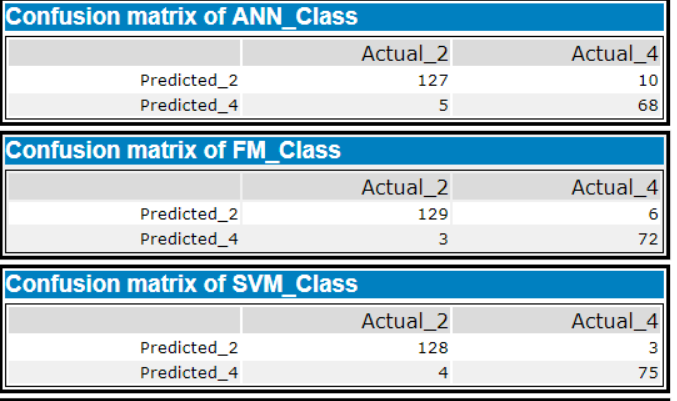
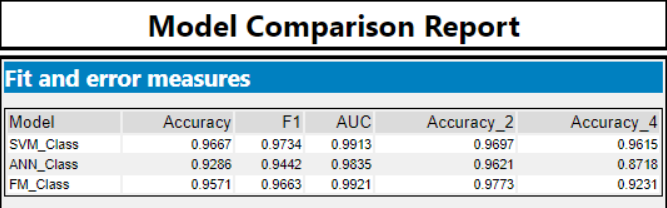
  
**Figure 4**: Artificial neural network result on estimation set

**Comparison**

Based on the results from the estimation set alone, the models appeared to perform similarly with ANN perfoming the best, then SVM, then the forest model. When validated and compared (Figure 5), SVM performed best, with an overall accuracy of 96.7% compared to the forest model’s 95.7% and ANN’s 92.9%, emphasizing the importance of validating each model. Still, the user must be careful to ensure there is no bias to accurately interpret the confusion matrices outcomes. It appears that because of the similarity in the outcomes between the models, this may not be the case.

SVM also reported the highest accuracy in classifying malignant tumors (Class 4), at 96.2% and classified benign tumors well (Class 2, 96.7%). The forest model reported the highest accuracy in classifying Class 2 (97.7%). ANN was almost as strong at predicting Class 2 (96.2%) but not Class 4 (87.2%). It is important to have relatively high accuracy for both classes, but especially for Class 4, as malignant tumor detection is the goal. For this reason, ANN may not be chosen for this particular data.

However, the models were run using Alteryx’s defaults with no additional tuning. With careful adjustments, ANN could outperform the other models due to its backpropagation technique, varying activation functions, and other adjustable parameters. Bias is minimized through the bias parameter, similar to an intercept in a linear equation. It adjusts the output and weighted sum of the inputs to a node, providing a constant that assists the model in fitting the data well. Still, drawbacks to ANN are its preference for larger data sets, its ‘black box’ nature, and its computational expense (Donges, 2018). The results are not as easy to interpret without knowing the inner workings of the model, providing the least clear understanding.



**Figure 5:** Comparison of the three models using validation

There are ethical considerations as well. Performance is open-ended, which means ANN may learn information that is not desired. The model may discriminate; a network may develop negative evaluations depending on a patient’s race, sex, or related information. Additionally, a model may assist with diagnoses, but is no replacement for patient-doctor interaction and the ethics of machines providing diagnoses to patients who require human aspects of sensitivity has been hotly debated (Digital, 2016).

Likewise, forest models are easy to interpret, can handle different data types, are robust to noise, do not require scaling, and are quite fast (Liberman, 2017). Still, bias may present in the model if it systematically under or over predicts the target variable. There is a “bias-variance tradeoff”, but ideally, a user will choose a model that minimizes both. With the default settings, the forest model may or may not overfit without a limit to the size of the model tree and requires further investigation through multiple runs and adjustments. The best approach to avoid overfitting is to optimize a parameter that limits the number of features chosen to grow each tree from the bootstrapped data (Equilibrium, 2014). Choosing a set of hyperparameters that mitigates the bias-variance tradeoff to minimize error on the validation set is the goal.

Similar ethical concerns to the ANN model arise anytime one is working with patients in a medical setting. The random forest model may be easy to use and interpret, but a patient should be aware that any model is subject to the biases aforementioned. Receiving a diagnosis that is machine-derived may be accurate, yet a pathologist or similarly qualified individual should always compare results prior to diagnosing patients with breast cancer. The results of the model in the wrong hands could be life-threatening for cancer patients.

SVM’s are not as easy to interpret as forest models but are more transparent than ANN. Alteryx provides many visualizations that depict the characterization of each data point with the decision boundaries between classes. SVM may be the ideal choice for the data based on its performance with the validation set. It is a powerful classifier that works well even with high-dimensional and non-linearly separable data (Gillian, n.d.). It contains a regularization parameter that may avoid over-fitting, uses a kernel trick that allows for greater understanding of the problem through kernel engineering, and is defined by a convex optimization problem for which there are efficient methods (Marsupial, 2012). However, it is wise to perform multiple runs with the model with minor adjustments to improve accuracy. A main disadvantage of SVM is the several key parameters that must be set properly to achieve the best classification results (Gillian, n.d.). Kernel models can be quite sensitive to overfitting and subsequent optimistic or selection bias; to avoid this, model selection must be treated as an integral part of the model fitting process (Cawley & Talbot, 2010). Additionally, choosing the kernel function appropriately is also tricky (Torpey, 2016).

**Conclusion**

For the Wisconsin Breast Cancer dataset, the random forest, SVM, and ANN models performed well with overall accuracy in the low to mid 90s; with careful fine-tuning any of the models may classify malignant and benign breast cancer tumors accurately. SVM performed the best despite its several hyperparameters that require optimization. Alteryx is powerful software and generates model results based on research-backed, well-performing hyperparameters.

Receiving a cancer diagnosis is difficult and the process should be handled with respect to the patient and his or her dignity. Machine learning algorithms have been excellent in predicting disease based on certain input variables, but it is important that the practice of medicine does not become streamlined into a computerized process without a human element. Emotional support cannot be delivered adequately by a machine, nor can a definitive diagnosis. The models should be used in conjunction with, and not a substitute for, healthcare provider interpretation and in a way that minimizes patient harm.

**References**

Breastcancer.org. (2019, February 23). US Breast Cancer Statistics. Retrieved from https://www.breastcancer.org/symptoms/understand\_bc/statistics

Cawley & Talbot. (2010, July). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research 11*(2010): 2079-2107. Retrieved from <http://jmlr.csail.mit.edu/papers/volume11/cawley10a/cawley10a.pdf>

Centers for Disease Control and Prevention (CDC) (n.d.). Breast cancer: Basic information. Retrieved from https://www.cdc.gov/cancer/breast/basic\_info/index.htm

Digital diagnosis: intelligent machines do a better job than humans. (2016, January 17). Retrieved from https://theconversation.com/digital-diagnosis-intelligent-machines-do-a-better-job-than humans-53116

Donges, N. (2018, April 17). Pros and cons of neural networks. *Towards Data Science.* Retrieved from https://towardsdatascience.com/hype-disadvantages-of-neural-networks-6af04904ba5b

Equilibrium, B. (2014). Re: Random forest – how to handle overfitting [Msg 2]. Retrieved from https://stats.stackexchange.com/questions/111968/random-forest-how-to-handle-overfitting

Gillian, N. (n.d.). SVM. Retrieved from http://www.nickgillian.com/wiki/pmwiki.php/GRT/SVM

Liberman, N. (2017, January 26). Decision trees and random forests. *Towards Data Science.* Retrieved from https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

Marsupial, D. (2012, March 10). Re: Advantages and disadvantages of SVM [Msg 2]. Retrieved from https://stats.stackexchange.com/questions/24437/advantages-and-disadvantages-of-svm

Torpey, D. (2016, September 28). Re: What are the disadvantages of SVM algorithms? [Msg 2]. *Quora.* Retrieved from https://www.quora.com/What-are-the-disadvantages-of-SVM-algorithms

Vala, A. (2018, March 7). Reinventing healthcare: Early detection of Alzheimer’s disease with deep learning. *Strata Business Summit.* Retrieved from https://conferences.oreilly.com/strata/strata ca-2018/public/schedule/detail/64259

Wolberg, W.H. & Mangasarian, O.L. (1990, December). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *PNAS 87* (23): 9193-9196. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC55130/