John Ferrara

Assignment 3 – DATA 622

Overview & Assigned Readings

As mandated, for this assignment the data that was analyzed using decision tree models in Assignment 2 was used. However, in Assignment 3 the SVM modeling methodology was used, so as to compare the performance to that of the decision trees in Assignment 2. Comparing the results in both assignments allowed for the further understanding of the nuance in model category performance.

There were additional assigned readings and research instructions for Assignment 3 as well. The two assigned research papers examined COVID-19 diagnosis and infection using decision trees and SVM modeling. In the first paper,[1] decision trees were used with lab results from Brazilian patients. Input features included 18 different lab biomarker measurements on 600 different patients, with 520 were COVID negative and 80 were COVID positive. Sampling techniques like RUS and SMOTE were used on this imbalanced data set and the overall best modeling technique was random forest, it had the highest accuracy. In the second paper,[2] researchers also tried to predict COVID infection, but with SVM methodologies using 8 different symptoms (e.g., heart rate, temperature, cough, etc.) for 200 patients. The research categorized patients into three classes, essentially no covid, mild covid, severe covid. With a 70/30 training test split, the SVM models were best on this dataset when compared to k-nearest neighbors, naive Bayes, random forest, and AdaBoost.

Custom Article Research & Model Comparison

After researching three additional papers comparing SVM to decision trees, several looking at financial categorizations and text classifications were found. The first article, "A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees," takes a look at corporate credit ratings and examines which models best forecast them. The data looks at financial metrics for firms over time across three sectors attempting to categorize them in to 19 different credit rating classes. They used SVM in both one vs one and one vs all modes for the multi class problem, and

[1] Decision Tree Ensembles to Predict Coronavirus Disease 2019 Infection: A Comparative Study – (https://onlinelibrary.wiley.com/doi/10.1155/2021/5550344)
[2] A novel approach to predict COVID-19 using support vector machine (https://pmc.ncbi.nlm.nih.gov/articles/PMC8137961/)

their results showed that Bagged Decision Trees and Random Forests performed the best in accuracy for these sectors, while the SVM models generally performed worse.

The second article, “Comparing Support Vector Machines and Decision Trees for Text Classification,” compares SVM and decision tree modeling for text classification using the 20 Newsgroups dataset, which is composed of about 20,000 newsgroup documents partitioned across 20 different newsgroups. After split the data into 80-20 training-testing sets, vectorized the word data into numeric values using TFIDFVectorizer, and then training both models. The SVM model outperforms the decision tree across all the main metrics. The article explains that SVMs are well suited to high dimensional text features, while decision trees are simpler and more interpretable but not as strong here.

The third article, “Predicting of Credit Default by SVM and Decision Tree,” takes a look at SVM modeling and decision tree modeling to predict credit defaults using a UCI credit card dataset. The dataset contains data points on customer age, gender, education, marriage status, and loan amount, and examines whether people default on payments. In the data, roughly 75 percent made on time payments, while one quarter were late and defaulted. The modeling techniques showed decision tree models performing slightly better by having lower error rates when compared to SVM techniques.

SVM Modeling Work

Continuing on for work completed in Assignment 3, I performed about 57 different SVM models with varying parameters in order to identify the best model for the Assignment 2 data. Preparation methodology was kept identical to Assignment 2 with the exception of scaling the data, which is needed for SVM modeling. The best models used class weight balanced and RBF kernels. My top SVM model got a ROC AUC of about 0.7956, with the runner up obtaining 0.7952. The highest performing RBF Kernel SVM model has C value of 2 and a gamma (y) value of 0.02.

After the modeling was done for Assignment 3, i pulled in the top three decision tree models from Assignment 2. The like Random Forest and AdaBoost methods were best in that assignment. I compared their ROC AUC and PR AUC scores with the top three SVM models. When comparing the results from Assignment 2, to the top SVM models in this assignment, the scores are very similar. However, the Random Forest model is still slightly better. For example, the second highest scoring model from the decision tree cohort in Assignment 2 has a ROC AUC score of 0.7958, while the best SVM model has 0.7956. These are basically tied, but technically the Random Forest edge is a bit higher.

Conclusion

In conclusion, based on the combination of the five papers and my own modeling work, my recommendation for this dataset is to use a decision tree ensemble, like Random Forest, as the main model. The tree-based models slightly outperformed the SVM on my data, and they fit the pattern seen in the corporate credit rating and credit default papers, where decision tree methods did better on structured tabular data. SVM is still a strong model and seems especially good on high dimensional text and on certain symptom based problems, but for this particular classification problem and dataset, I would stick with Random Forest as the most accurate and practical choice, with SVM as a good backup or comparison model.