CS 5830 - Project 7

Chris Brown, Thomas Brower

INTRODUCTION

In these analyses, we investigated the suitability of Logistic Regression and Support Vector Machines in identifying fraudulent bills and classifying wine quality. Our dataset to fight against fake bills, we used a dataset that banks and authorities could easily replicate with existing tools. Our dataset for wine quality is more involved, but just as easily reproducible by wineries. It is important to be reproducible as that allows our target audiences to benefit from our machine learning models. These analyses can be used by their respective target audiences to combat fraud or to produce higher quality wine decreasing costs for fraud detection and increasing profit for the wineries.

DATASET

We acquired our Wine Quality dataset from Paulo Cortez, University of Minho, Guimarães, Portugal. The dataset can be downloaded here. The dataset is split into two sections that describe the red and white variants of the Portuguese "Vinho Verde" wine. The dataset has 1600 samples of red wine and 4900 samples of white wine. There are 11 physicochemical attributes one 1 sensory target. Due to privacy and logistic issues, the datasets do not contain grape types, wine brands, wine selling prices, etc. For this reason, there are many outliers in the datasets that cannot be explained purely using the provided data. The data is also unbalanced with there being many more normal wines than excellent or poor quality ones. To conduct our analyses with these datasets, we merged them into one set and transformed the target variable to a binary class of "good wine" or "bad wine." This decision boundary was chosen so that half of the wine was considered good and half considered bad. The attributes were left as is and were all continuous-valued.

We acquired our Bank Note dataset from the University of California. The dataset has 1372 samples of banknotes with 3 attributes describing the wavelet transformation of the image, 1 attribute describing the entropy of the image, and the classification being binary stating whether the banknote was real or fake. There was about 60% of the data that has been classified as a real banknote and the attributes were left as is with all the attributes being continuous-valued.

ANALYSIS TECHNIQUE

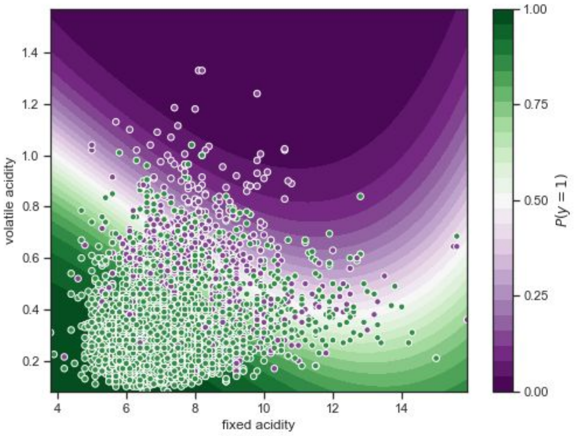
To find the best variables for predicting our target attributes, we plotted the decision boundaries between every attribute pair. Looking at the plots we determined the attributes that are good indicators based on how well they separated the output classes.

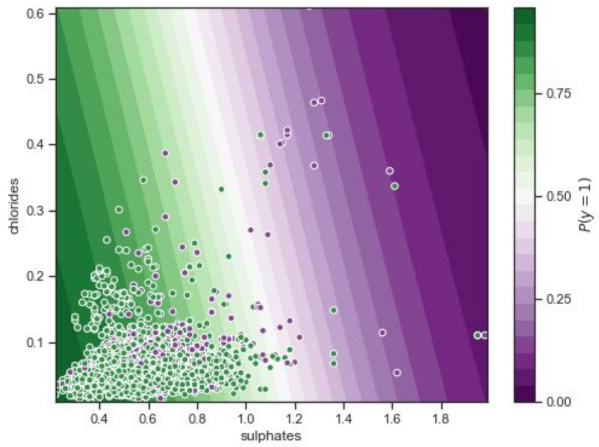
To compare the performance of Logistic Regression and Support Vector Machine variants, we performed 5-cross-fold validation on each classifier with a myriad of parameters. We compared the precision, recall, and F1 scores for each class in each classifier. While doing so, we also compared the runtime performance of each classifier. For the Support Vector Machines, we looked at linear, polynomial, and RBF kernels along with various gamma values and polynomial degrees. We also tested one-vs-rest and one-vs-one Support Vector Machines.

Altering the parameters on the Support Vector Machines allowed us to test the bias-variance tradeoff and to find how well models generalized on our data. To find the "best" models, we plotted each predictive attribute pair with high and low gamma values along with their decision boundaries to observe cases of over and under-fitting. To test the class weights we ran the data with balanced weights and with unbalanced weights and compared our recall, F1, and precision scores.

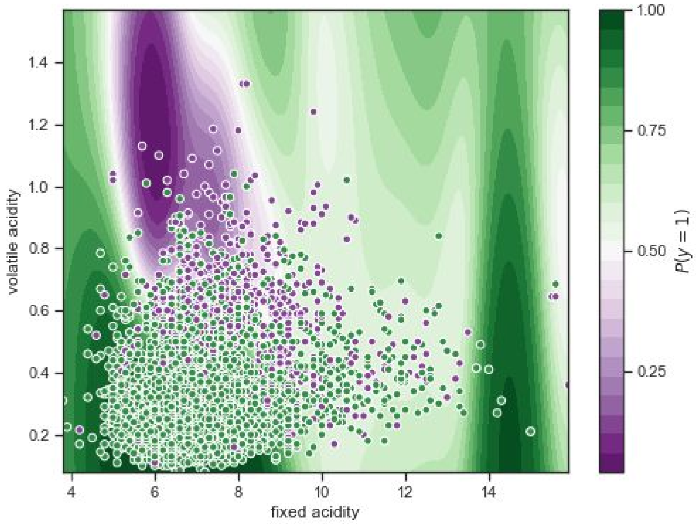
RESULTS

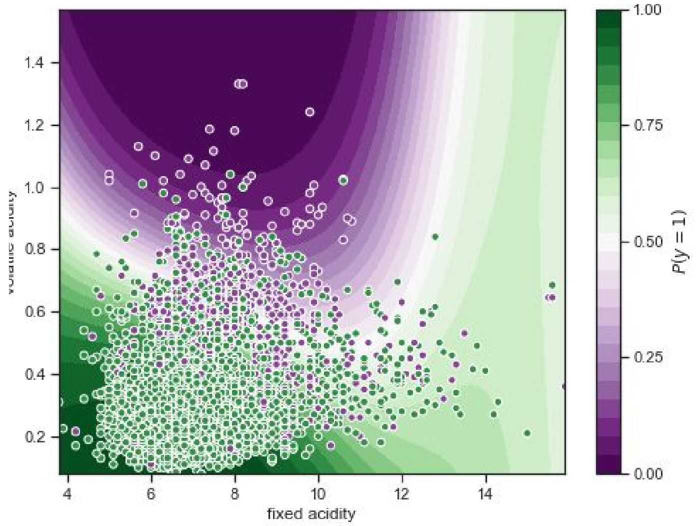
We found the best predictive variables of the wine quality to be the acidity of the wine (includes fixed, volatile, and citric) and the alcohol percentage. The worse predictors were the density, residual sugar, and chlorides in the wine. Including all attributes produces an F1 score of 0.4416 with bad wine and 0.9338 with good wine. Only keeping the acidity measures produces an F1 score of 0.5959 with bad wine and 0.9070 with good wine. The images below show that acidity measures have a somewhat clear separation of classes while there is no separation between other variables. Green dots show good wine quality while purple is bad wine quality.



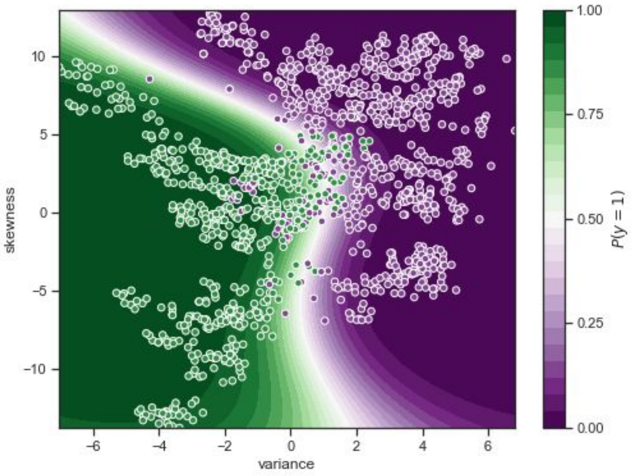


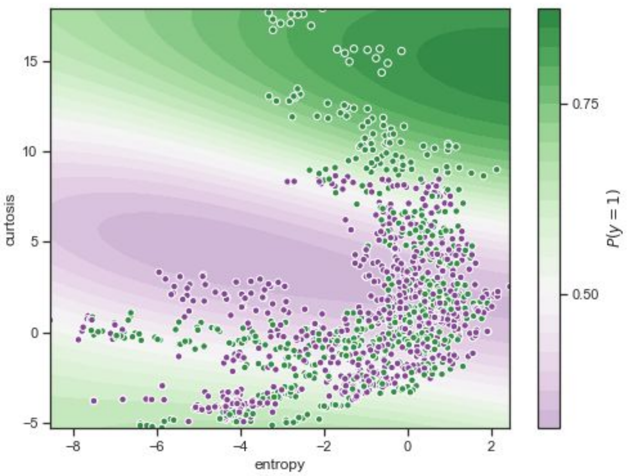
Comparing the performance of Logistic Regression and Support Vector Machines we find that all methods perform similarly with an RBF SVM performing the best. The Logistic Regression had an F1 score of 0.5702 with bad wine and 0.9149 with good wine. The linear SVM performed the same with a poly SVM performing slightly better. The RBF performed the best but greatly improved the F1 score of bad wine classification when the gamma value was decreased. At the same time, the RBF was the second slowest to construct with the poly SVM being the slowest. The Logistic Regression was by far the fastest to construct. We believe the lower gamma performs best as the model overfits to good wine quality so the model has great predictive capabilities with good wine, but terrible with bad wine. Reducing the variance improves overall performance. Below are two plots comparing (left) high gamma/variance and (right) low gamma/variance.



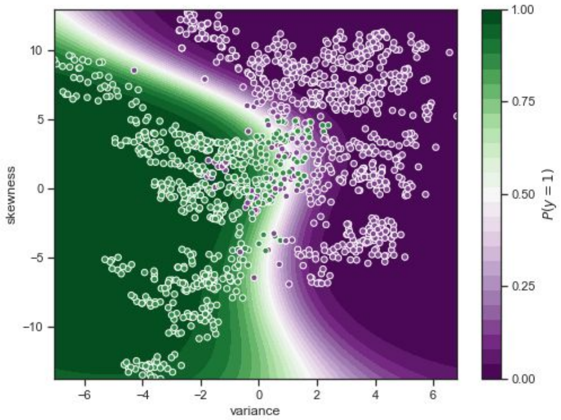


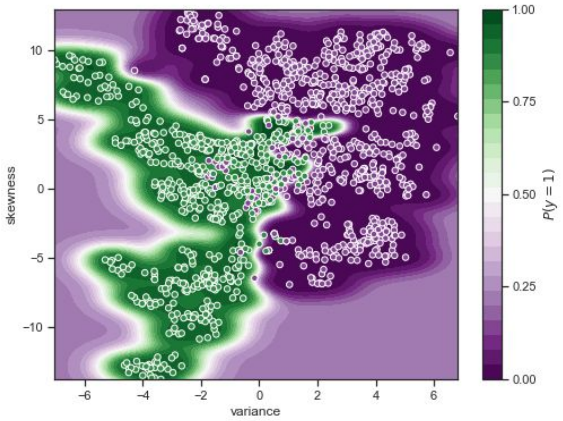
For the banknote data, we found the best predictive attributes to be the Variance and the skewness of the Wavelet transformed Image. The worst predictors were curtosis and entropy. The following graphs show a good and bad pair for the previously discussed attributes.





Similarly with the wine dataset, when comparing the performance of our Logistic Regression and Support Vector Machines we found that all methods perform similarly with an RBF SVM performing the best. It should be noted that with the banknote data, the increase in performance was only in increments of hundredths of a decimal. The Logistic Regression had an F1 score of 0.9874 for a good bill and 0.9851 for a bad bill. The linear SVM performed the same as the Logistic Regression and the Poly SVM performed slightly better. The RBF gave us the best results with an F1 score of 0.9966 for good bills and 0.996 for bad bills. The best results were when we used a small gamma value for all SVM, however, computing SVM took significantly longer to compute than the Logistic Regression. Reducing the variance, having a low gamma value, improves the overall performance of our model as is shown in the next two visualizations.





CONCLUSIONS

From our analyses on fake banknotes and predictive models, banks and authorities can identify fraudulent bills quicker and cheaper. Wineries can similarly use our analyses on wine quality to increase both their profit and products' quality. In the future, as criminals have access to better technology and tools, our models would need to be reassessed and updated to capture the new forms of fraudulent bills. Wineries would also need to do the same thing as public opinion on what constitutes good wine changes.

Slides

https://docs.google.com/presentation/d/1tKEs4WUnTaJXXUeTISGqS3GwL969A1R9yctqfz95LP0/edit?usp=sharing