COBB2010: Foundations of Computational Biology 12/7/2020

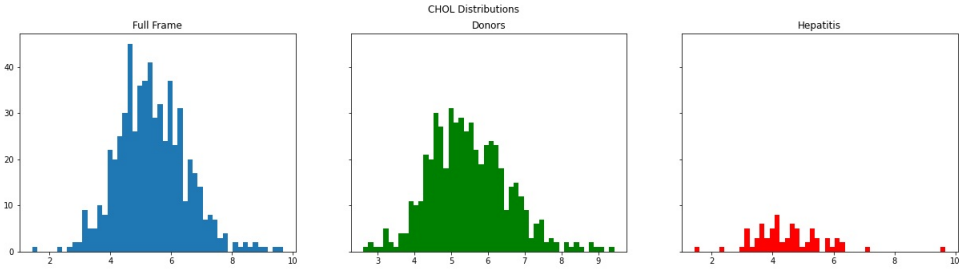
Final Project

Name: Kamaru-Deen Lawal

Email: kal246@pitt.edu

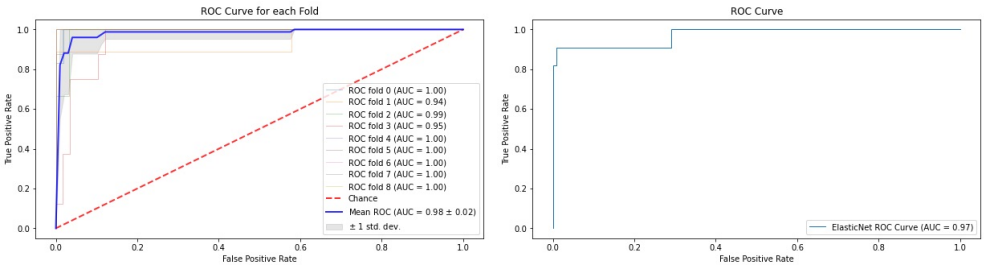
The goal of this project is to examine the relationship between Hepatitis C, and a series of fourteen attributes. Hepatitis C is a liver infection spread via contact with blood from an infected person. The disease is most commonly spread as a result of sharing needles or other equipment to prepare and inject drugs. It can also be transmitted as a result of having unprotected sexual intercourse. Researchers have put a considerable amount of work into building predictive models that can determine whether or not a patient is positive for Hepatitis C based on standard blood/urine tests. We hope to extend this research by focusing specifically on interpretable models.

We began by looking for Hepatitis C datasets. After a bit of searching we found a dataset assembled by Ralf Lichtinghagen at The Institute of Clinical Chemistry in Hanover. The dataset consists of 615 patients with twelve features (11 quantitative, 1 categorical). The target variable takes on five varying levels of Hepatitis C. The initial goal was to build a multi-class classifier based on the features, but because of severe class imbalance we ended up analyzing a much simpler binary classification problem. We did this by consolidating the values in the target column based on whether or not a patient could become a blood donor or not. Before doing any external research outside of our dataset we began by examining the distributions of the quantitative features in the dataset conditioned on whether or not patients were positive or negative for Hepatitis C. An example plot for the cholestorol feature can be seen below.

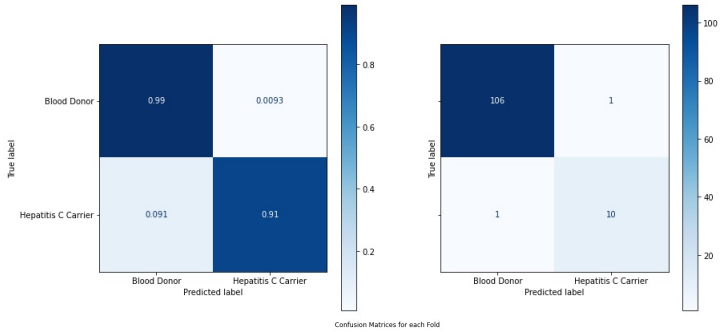


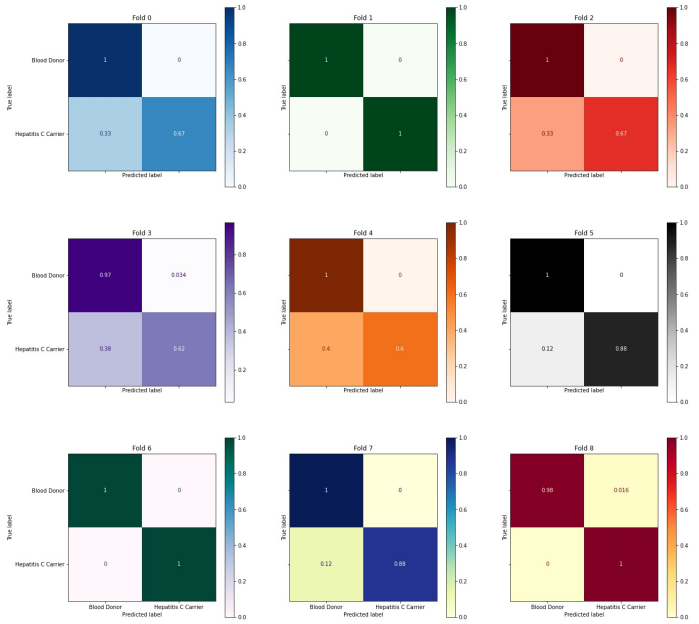
For each of the features we examined the distributions, and attempted to draw some minor conclusions about the given feature along with how the feature might impact the model. A lot of the external information we found on the feature set ended up aligning with the conclusions we were able to draw about our specific dataset. An example of such a conclusion is that Hepatitis C carriers tend to have a significantly higher mean and variance for Creatinine than patients that are negative for Hepatitis C. After a thorough examination of the dataset we built our model. Namely, we built a Logistic Regression classifier with an Elastic-Net penalty. We evaluated the model in two separate ways. The first method of evaluation was via a simple train/test split. The second method of evaluation was using k-Fold cross validation.

Overall the model ended up performing well out of the box with an AUC of.97 for the train/test split method of evaluation, and a Mean AUC of .98 for the k-Fold CV method of evaluation. The small size of the dataset coupled with the extreme class imbalance for the target variable made kFCV the ideal method for gathering a true assessment of the model. Below are the AUC plots for both methods of evaluation.



In addition we examined normalized and unnormalized confusion matrices for the simple train/test split along with normalized confusion matrices for each of the folds. The aforementioned plots can be seen below.





Overall the results that we managed to gather from the project seem promising with a wide range of directions for future work. One fairly obvious path for future work would be to place a higher emphasis on optimizing parameters for the model. The model ended up performing fairly well out of the box, so changes were not necessary. But a wide range of tools exist to help perform hyper-parameter optimization in Python. Another path for future work would be to place a higher emphasis on the resulting output weights for each feature in the Elastic-Net Model. The whole point of making use of interpretable methods is to find some type of causal relationship between the input features and the output target, so actually exploring the model's output weights seems like a great way to begin making sense of what leads to Hepatitis C.

Citations

- Gareth James, Daniela Witten,Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.
- Kuhn, Max., and Kjell Johnson. Applied Predictive Modeling. New York: Springer, 2013.