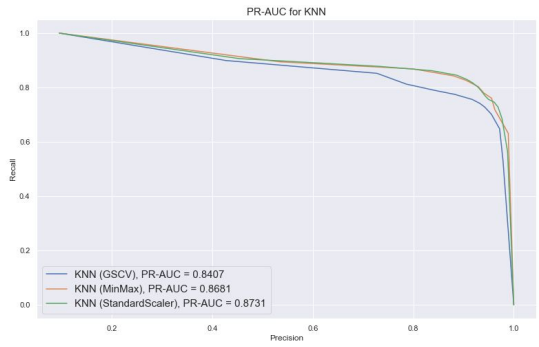
Chris Johanson
Capstone 1 In-Depth Analysis

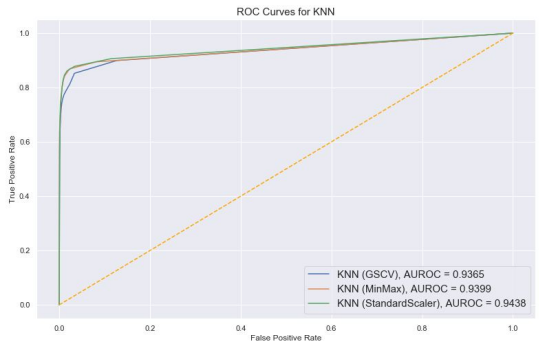
For the machine learning portion of the project, several models were created across five different classifiers with the goal of finding an effective classification model. For each type of classifier, multiple models were created by applying different techniques to test for optimization. Some preprocessing techniques were involved, such as scaling the data using MinMaxScaler and StandardScaler. In addition, hyperparameter tuning with grid-search cross validation was utilized.

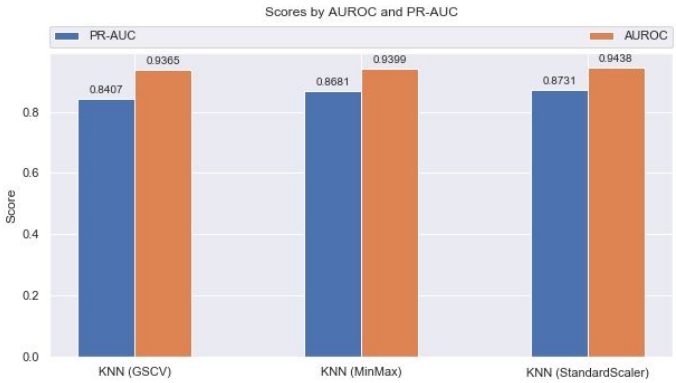
Regarding evaluation metrics, mainly two were chosen: ROC curves and Precision-Recall curves. For each curve, the area under the curve (AUC) was also calculated. The ROC and AUROC were chosen to observe the false positive rates and true positive rates. However, the dataset was imbalanced so the Precision-Recall curve was included to account for evaluating with said imbalance in mind. The AUROC and PR-AUC were plotted on a grouped bar chart in order to easily compare the two values for a given model, as well as comparing multiple models at a time.

With the goal of finding the most effective classification of the model, the in-depth analysis will culminate in a comparison of the top performers for each classification. Such a comparison will be valuable in terms of knowing which model will work best for the issue of identifying pulsar stars.

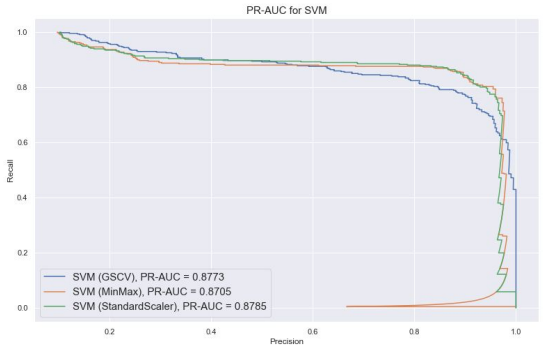
The K-Nearest Neighbors (KNN) portion of the analysis involved three models that were grid-search cross validated to tune the hyperparameters. KNN (GSCV), the first model, was tested using raw data. The following two models, KNN (MinMax) and KNN (StandardScaler) used scaled data via MinMaxScaler and StandardScaler, respectively. Of the three models, KNN (StandardScaler) had the best AUROC and PR-AUC scores. It is important to note that KNN (MinMax) has very similar, slightly lower scores. Since the scores are so close, it is well within the realm of possibility that, were the tests run again, KNN (MinMax) could outperform KNN (StandardScaler) by the same narrow margin.

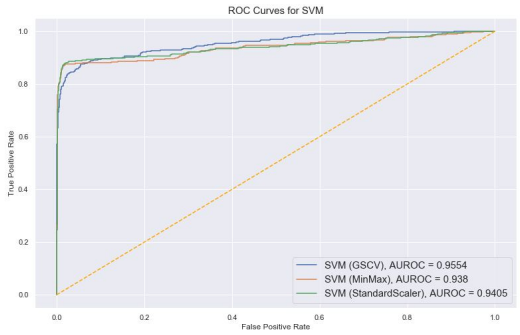


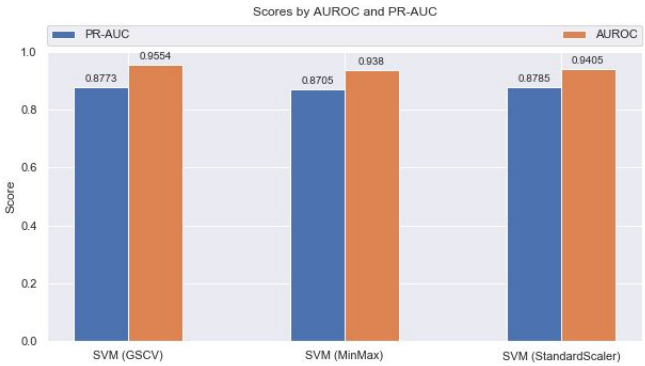




The Support Vector Machine (SVM) portion of the analysis used the default Radial Basis Function. Similar to the KNN portion, three models were tested. The first was a grid-search cross validated SVM using raw data, while the other two were also grid-search cross validated. However, the second and third models used data that was preprocessed with MinMaxScaler and StandardScaler, respectively. The SVM with raw data had the highest AUROC and second-highest PR-AUC. With the highest PR-AUC only having a lead of .0012, the SVM with raw data was the best choice.

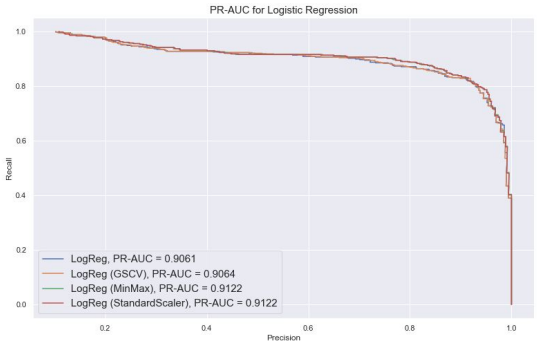


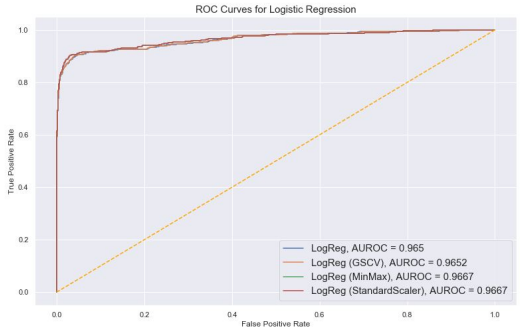




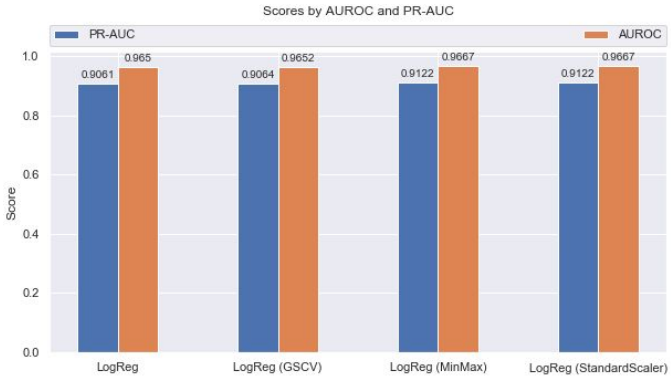
The Logistic Regression portion included four separate models. The first model was “out of the box,” meaning there was no hyperparameter tuning and raw data was used. Next, the second model was grid-search cross validated to tune hyperparameters, with raw data also being

used. Models three and four differed from the previous two in that they used scaled data, MinMaxScaler and StandardScaler, respectively. Across the board, the scores were rather similar. The models that used scaled data tied for the top score, so the MinMax model was

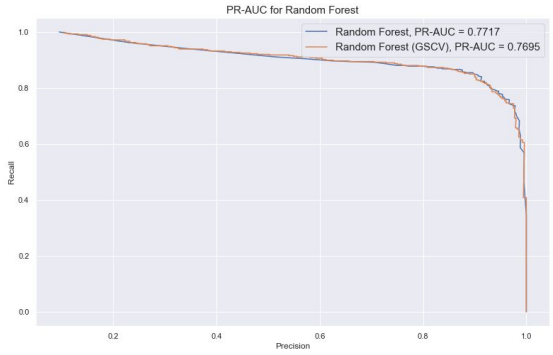


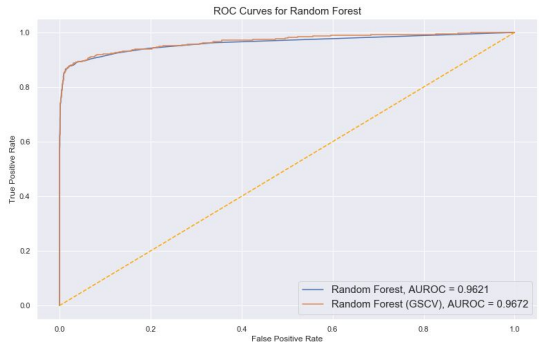


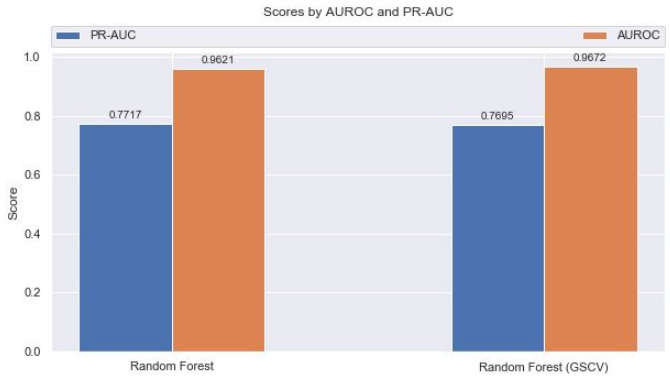
arbitrarily. chosen.



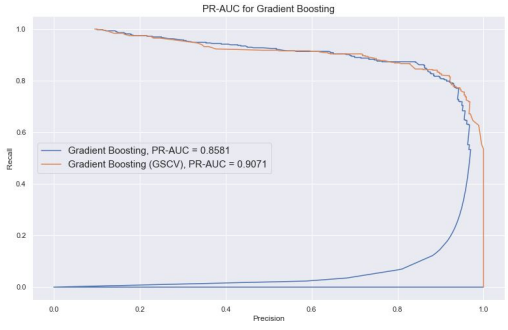
The Random Forest portion included two models: the first was “out of the box” and the second used grid-search cross validation for hyperparameter tuning. Regarding the scores, both models had similar AUROC and PR-AUC scores, but the tuned model did have a slightly better performance.

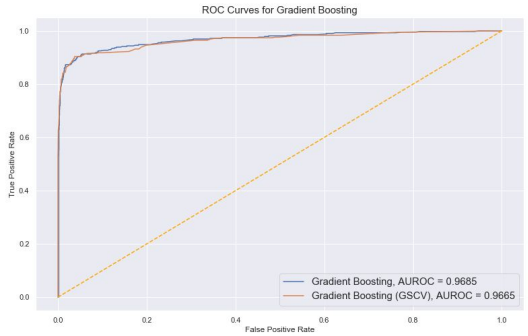


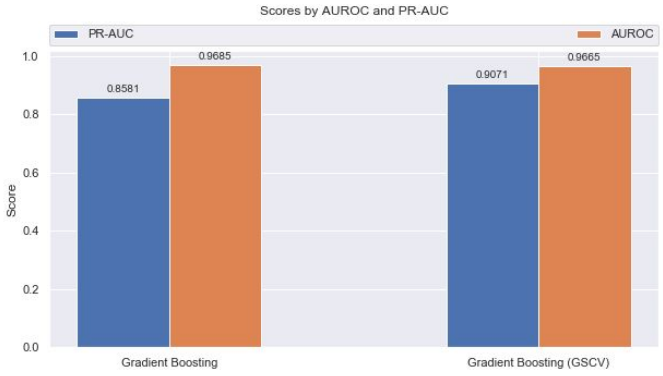




Gradient Boosting was included in the analysis to continue exploring ensemble decision tree classifiers. Like the Random Forest section, two models were included. The first was “out of the box” and the second was tuned using grid-search cross validation. The AUROC scores for the two models were nearly identical. PR-AUC showed a significant difference between the two models with the tuned model’s PR-AUC being .05 ahead of its untuned counterpart. Of course, the tuned model was chosen.







The final portion of the analysis consists of a comparison of each classifier's top-performing model. The models are KNN with standard scaled data, SVM with raw data, logistic regression with min-max scaled data, the hypertuned Random Forest, and hypertuned Gradient Boosting. The top performing model was the logistic regression model, which used

min-max scaled data, with the second highest AUROC and the highest PR-AUC. It should be noted that the Gradient Boosting model was a close second, but the difference (.51) between the two PR-AUC scores clearly demonstrate that, of the classifiers this analysis has investigated, logistic regression has been the best performer for determining pulsar stars from astronomical radio data.