Chris Johanson
Capstone 1 Final Report

As technologies and capabilities for collecting and storing data increase, there is a parallel need to make sense of it all. Computing power is continuing to become less expensive and more powerful and efficient. The accessibility of both large-scale data and computer power gives way to a multitude of learning and research opportunities across a wide array of disciplines. The field of astronomy is certainly no exception. The focus of this report is the investigation of pulsar stars in the field of astronomy and how machine learning algorithms play a role in the discovery process.

A pulsar star is a type of neutron star that spins rapidly and produces radio energy emissions in the form of beams which are ejected from its magnetic poles. Pulsars also rotate at a somewhat regular interval, so the combination of the emissions and rotation patterns are detectable via radio telescopes here on Earth.The data set being used for this project consists of astronomical radio data collected by the Parkes High Time Resolution Universe Legacy Survey and was accessed via Kaggle. Regarding signals of radio emissions, the Parkes High Time Resolution Universe Legacy Survey scanned across a certain band of radio frequencies. Pulsars have slightly different emission patterns which are unique to the individual star, similar to the way all humans have fingers but the fingerprints uniquely identify an individual. If these specific emission patterns were the only detectable patterns, finding pulsars would not be so difficult. However, given the large amount of background noise and interstellar signal disruptions, it is extremely difficult to find the proverbial needle in the haystack. The computational power of machine learning algorithms enable researchers to comb through large amounts of data with a reasonable degree of accuracy.

The implications of pulsar classification algorithms impact multiple fields within astronomy. Information about the nature and behavior of pulsar stars sheds light on topics such as: the density of matter (pulsars are second to black holes in terms of being the most dense objects we have detected thus far in the universe), Einstein's general relativity, and gravitational waves generated by supermassive black holes merging in the early universe.

Regarding the dataset, it was a relatively straightforward dataset consisting of 8 features and 1 target column. The 8 features are the following:

1. Mean of the integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.
4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.
9. Target class

Features 1 through 4 are statistical measurements of the integrated profile. As explained above, pulsars produce different emission patterns. This emission pattern varies with each rotation and is normally too weak to detect on its own. An integrated profile is the combining, or folding, of

these weaker signals in order to increase the signal quality. Ultimately, the integrated profile is what gives a pulsar star its unique identity and the increased signal quality makes detection easier. Features 5 through 8 are statistical measurements of the dispersion measure-signal to noise ratio, or DM-SNR curve. Dispersion measure represents the disruptive effect that the interstellar medium (matter/particles that exist between the star systems in a galaxy) has on signals generated by pulsars. The signal to noise ratio is the "strength of the signal compared to random noise" (4. Observations of Pulsars). When combined, the dispersion measure and signal to noise ratio increase the ability to make the necessary measurements when searching for pulsars.

A brief overview of the data demonstrated that there are, as stated above, 9 rows in total and 17,898 rows. Each row is, of course, an individual observation. To find out how many pulsars and non-pulsars there are in the dataset, the data was divided into those two groups using the target class column. Of the 17,898 rows, 16,529 (~90.8%) are not pulsars and 1,639 (~0.09%) are pulsars. In terms of data types, the target class consists of an int64 while the remaining are float64. In addition, there are 0 null values.

In terms of cleaning and wrangling the data, the process was simple and painless. The only issue that was encountered was with the columns names of the 8 features. There was a white space at the beginning of the column names and was removed with the following code:

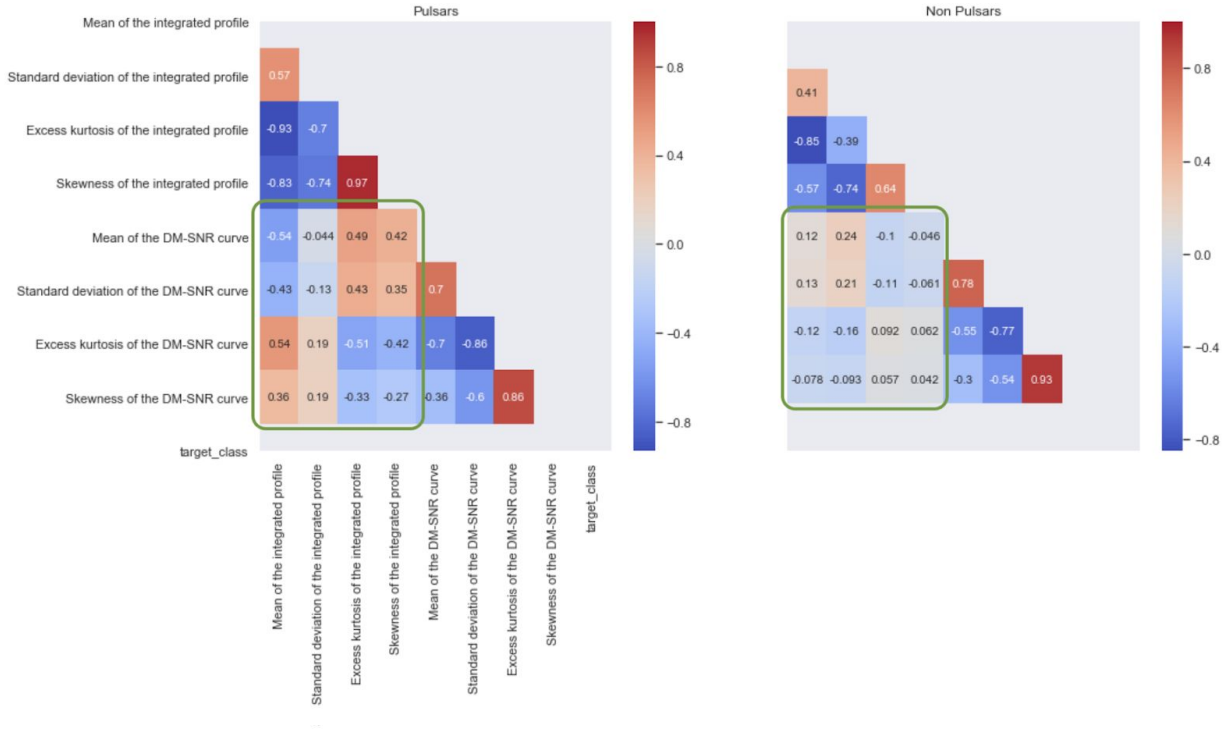
```
#Fix the white space in front of all the column names (except for target_class)
clean_column_names = []

for name in df.columns:
    clean_name = name.lstrip()
    clean_column_names.append(clean_name)

df.columns = clean_column_names
df.columns
```

```
Index(['Mean of the integrated profile',
       'Standard deviation of the integrated profile',
       'Excess kurtosis of the integrated profile',
       'Skewness of the integrated profile', 'Mean of the DM-SNR curve',
       'Standard deviation of the DM-SNR curve',
       'Excess kurtosis of the DM-SNR curve', 'Skewness of the DM-SNR curve',
       'target_class'],
      dtype='object')
```

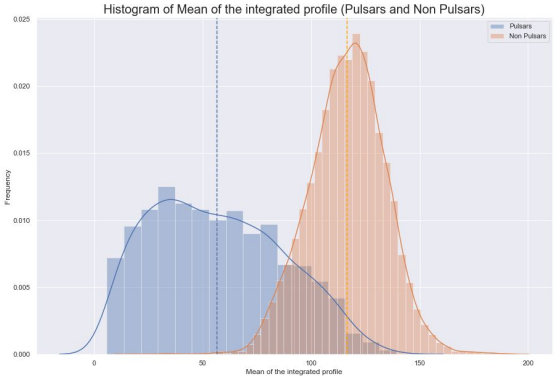
The exploratory analysis was straightforward. Being a classification problem, the first step was to take the entire dataset and split it in two according to the target class feature, where a "0" indicates that an observation is not a pulsar star and a "1" indicates that an observation is a pulsar star. Next, a correlation matrix heatmap was generated for both groups and plotted next to each other to investigate the relationships between each pair of variables:

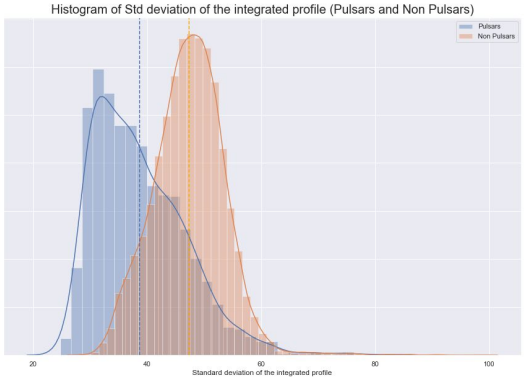


Comparing “Non Pulsars” and “Pulsars”, there are some interesting patterns. First, there are some relationships between variables that do not change, mostly occurring outside of the green rectangle seen in the visualization above. A lack of change is not surprising because the relative stagnancy of these elements is likely due to the fact that said elements either measure the integrated profile against itself or the DM-SNR curve against itself, as opposed to measuring an aspect of the integrated profile against the DM-SNR curve (or vice versa). Second, there are indeed noteworthy changes among the elements that fall within the highlighted green box. That is to say, elements comparing some aspect of the integrated profile and DM-SNR curve to each other. A number of these elements are rather neutral in the “Non Pulsars” group but become more polarized in the “Pulsars” group, for example the excess kurtosis of the DM-SNR curve with the mean of the integrated profile and the mean of the DM-SNR curve with the excess kurtosis of the integrated profile. It is too early to say as of right now, but these patterns merit further investigation when compiling the classification algorithm.

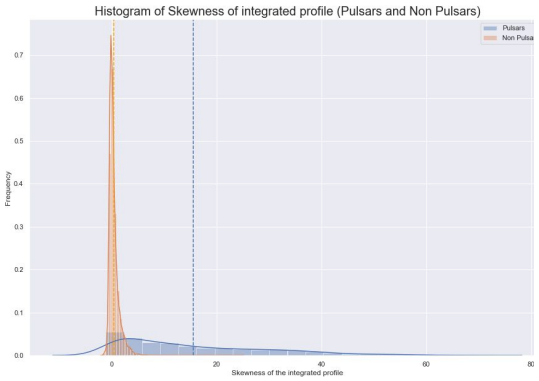
The exploratory analysis was concluded by dividing all 8 features into two groups: integrated profile and DM-SNR curve. The four statistical measurements (mean, standard deviation, excess kurtosis, and skewness) were compared between the two groups so the mean of the integrated profile could be compared with the mean of the DM-SNR curve, and so on. For all of the individual features, a dual histogram was plotted that included both cases of the two groups. Then, a two sample t-test was performed to make sure the result was statistically significant.

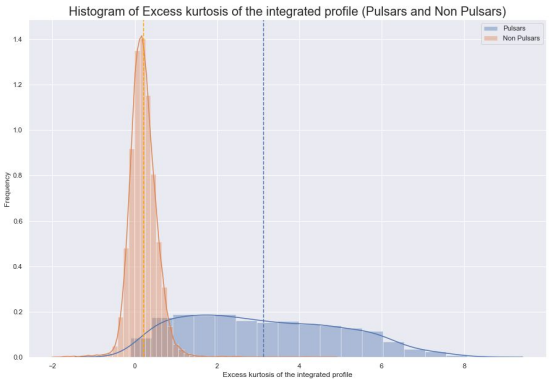
Starting with the mean of the integrated profile and the standard deviation of the integrated profile, the following dual histograms were generated:



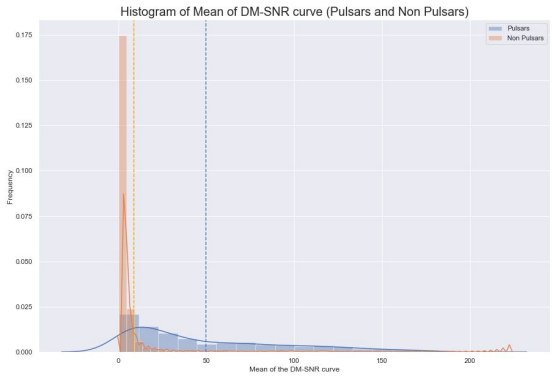


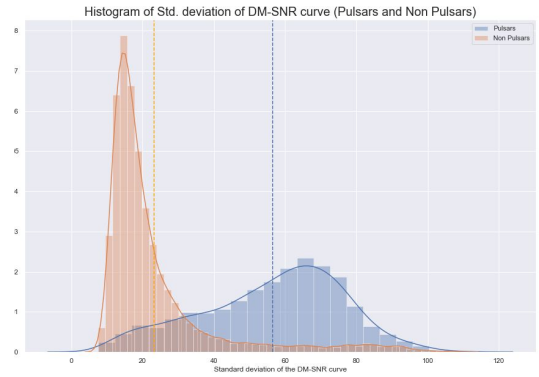
The two sample t-test for the mean of the integrated profile returned a p-value of 0.0, which statistically confirms that the two independent samples have different expected values. In terms of the standard deviation of the integrated profile, The corresponding two sample t-test returned a p-value of 3.99272404103977e-273, statistically confirming that the two independent samples have different expected values. Next are the dual histograms for the excess kurtosis of the integrated profile and the skewness of the integrated profile:



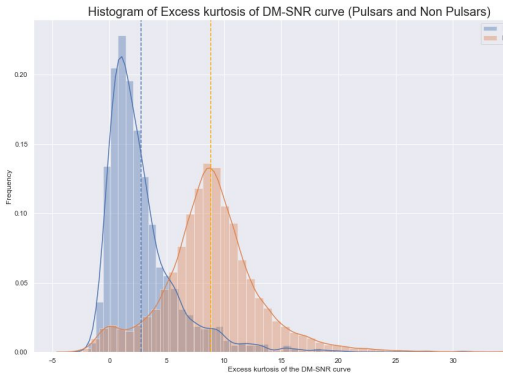


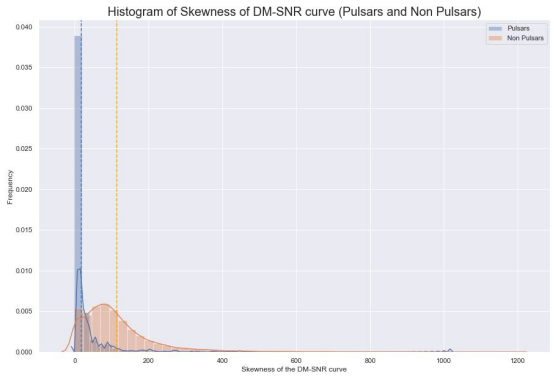
The corresponding two sample t-test for the standard deviation of the integrated profile returned a p-value of 3.99272404103977e-273, statistically confirming that the two independent samples have different expected values. Regarding the two sample t-test for the excess kurtosis of the integrated profile resulted in a p-value of 0, which statistically confirms that the two independent samples have different expected values. With the statistical significance established for the integrated profile, it was then time to move on to the dispersion measure-signal to noise ratio (DM-SNR) curve. The dual histograms for the mean of the DM-SNR curve and the standard deviation of the DM-SNR follow:





The two sample t-test for the mean of the DM-SNR curve returned a p-value of 2.7586700453147315e-213, which statistically confirms that the two independent samples have a different expected value. Then, the two sample t-test for the standard deviation of the DM-SNR curve returned a p-value of 0, also statistically confirming that the two independent samples have different expected values. Finally, the excess kurtosis and skewness of the DM-SNR curve were examined with the resulting dual histograms displayed below:





The two sample t-test for the excess kurtosis of the DM-SNR curve returned a p-value of 0, which statistically confirms that the two independent samples have different expected values and the two sample t-test for the skewness of the DM-SNR curve returned a p-value of 0, as well, statistically confirming that the two independent samples have different expected values.

Exploratory analysis and statistical testing have brought forward a number of insights. First, the correlation matrix heatmap aided in highlighting interesting relations, with the most changes in correlation occurring when an integrated profile feature is compared to a DM-SNR feature. Said correlations will be strongly considered when a classification algorithm is constructed. Second, when the four statistical measurements (mean, standard deviation, excess kurtosis, and skewness) were compared between the integrated profile and the DM-SNR, the two sample t-test confirmed that the differences were statistically significant. The insights gained from the crucial exploratory analysis will play an meaningful role in constructing a classification algorithm that can be used to determine whether or not observations are a pulsar star.

Chris Johanson

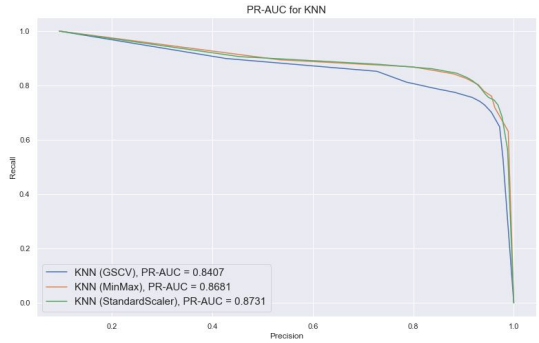
Capstone 1 In-Depth Analysis

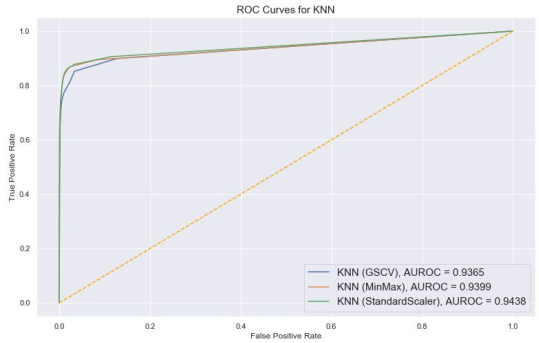
For the machine learning portion of the project, several models were created across five different classifiers with the goal of finding an effective classification model. For each type of classifier, multiple models were created by applying different techniques to test for optimization. Some preprocessing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

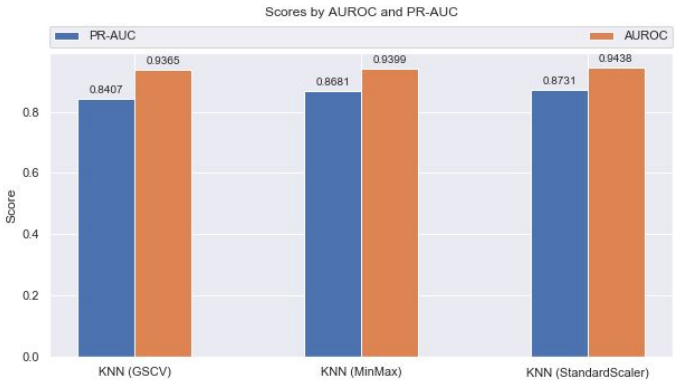
Regarding evaluation metrics, mainly two were chosen: ROC curves and Precision-Recall curves. For each curve, the area under the curve (AUC) was also calculated. The ROC and AUROC were chosen to observe the false positive rates and true positive rates. However, the dataset was imbalanced so the Precision-Recall 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.

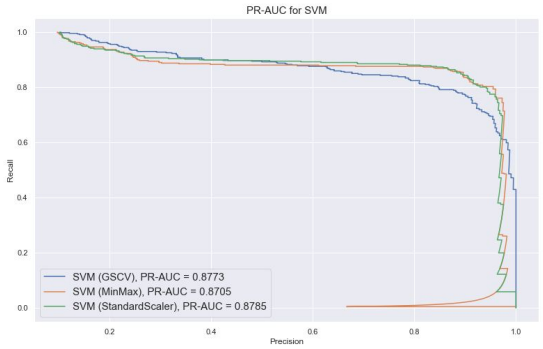
The K-Nearest Neighbors (KNN) portion of the analysis involved three models that were grid-search cross validated to tune the hyperparameters. KNN (GSCV), the first model, was tested using raw data. The following two models, KNN (MinMax) and KNN (StandardScaler) used scaled data via MinMaxScaler and 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.

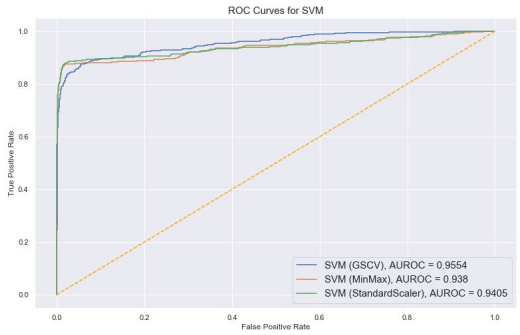


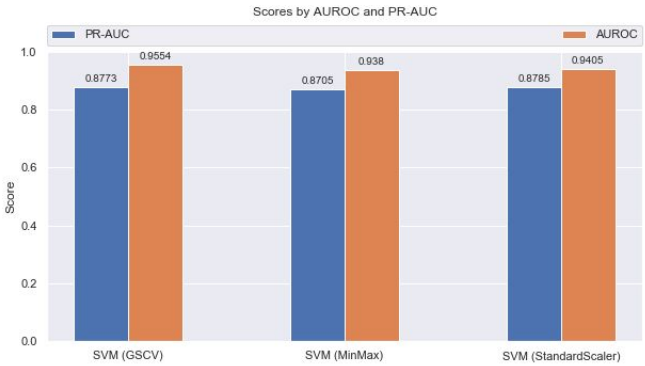




The Support Vector Machine (SVM) portion of the analysis used the default Radial Basis Function. Similar to the KNN portion, three models were tested. The first was a grid-search cross validated SVM using raw data, while the other two were also grid-search cross validated. However, the second and third 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.

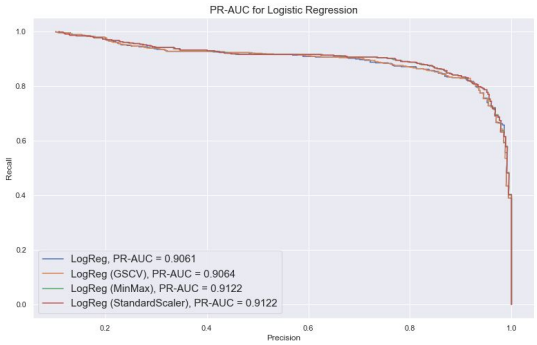


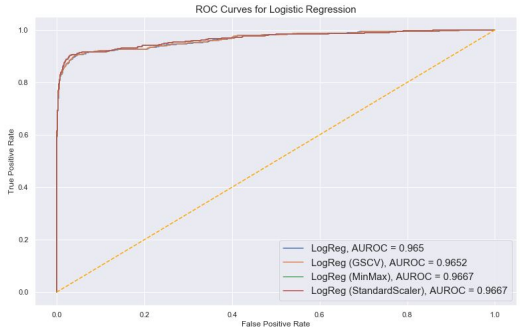




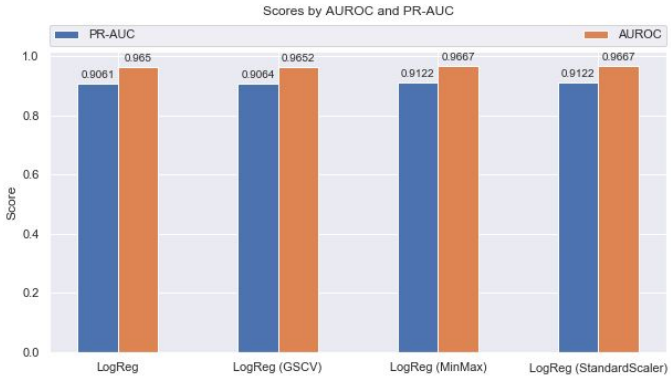
The Logistic Regression portion included four separate models. The first model was “out of the box,” meaning there was no hyperparameter tuning and raw data was used. Next, the second model was grid-search cross validated to tune hyperparameters, with raw data also being

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

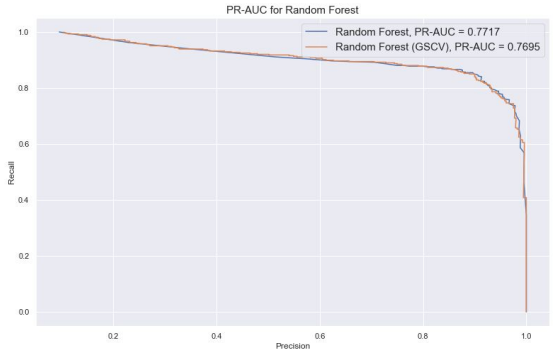


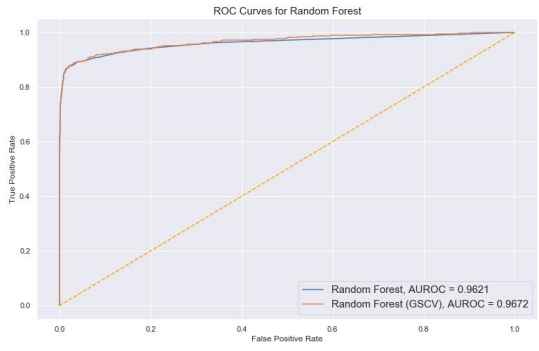


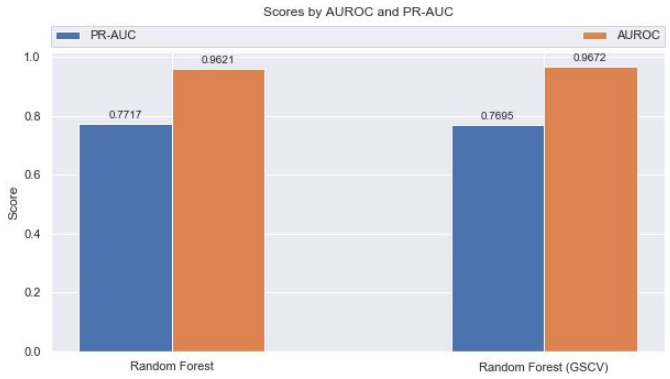
arbitrarily. chosen.



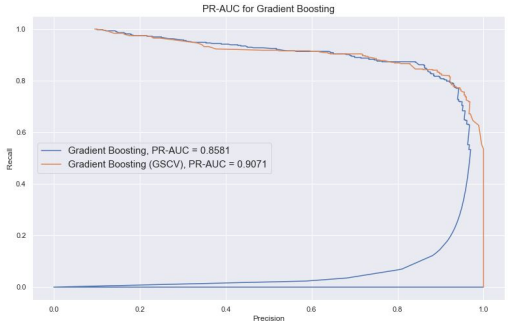
The Random Forest portion included two models: the first was “out of the box” and the second used grid-search cross validation for hyperparameter tuning. Regarding the scores, both models had similar AUROC and PR-AUC scores, but the tuned model did have a slightly better performance.

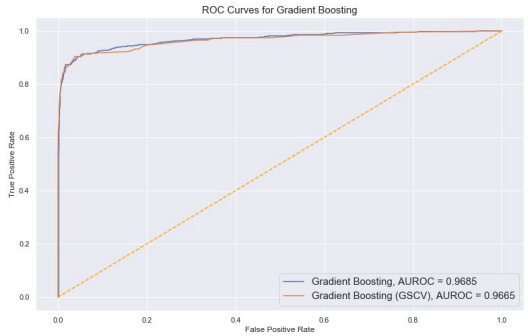


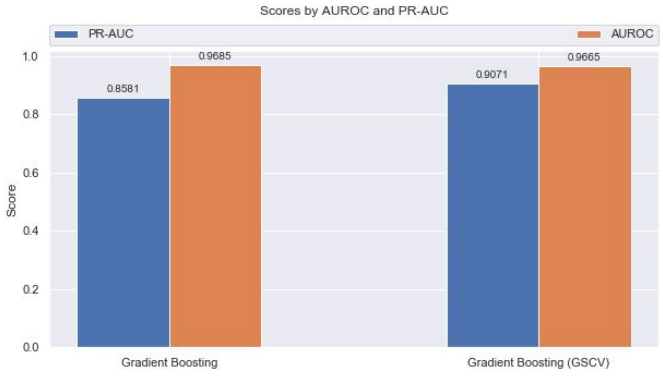




Gradient Boosting was included in the analysis to continue exploring ensemble decision tree classifiers. Like the Random Forest section, two models were included. The first was “out of the box” and the second was tuned using grid-search cross validation. The AUROC scores for the two models were nearly 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.

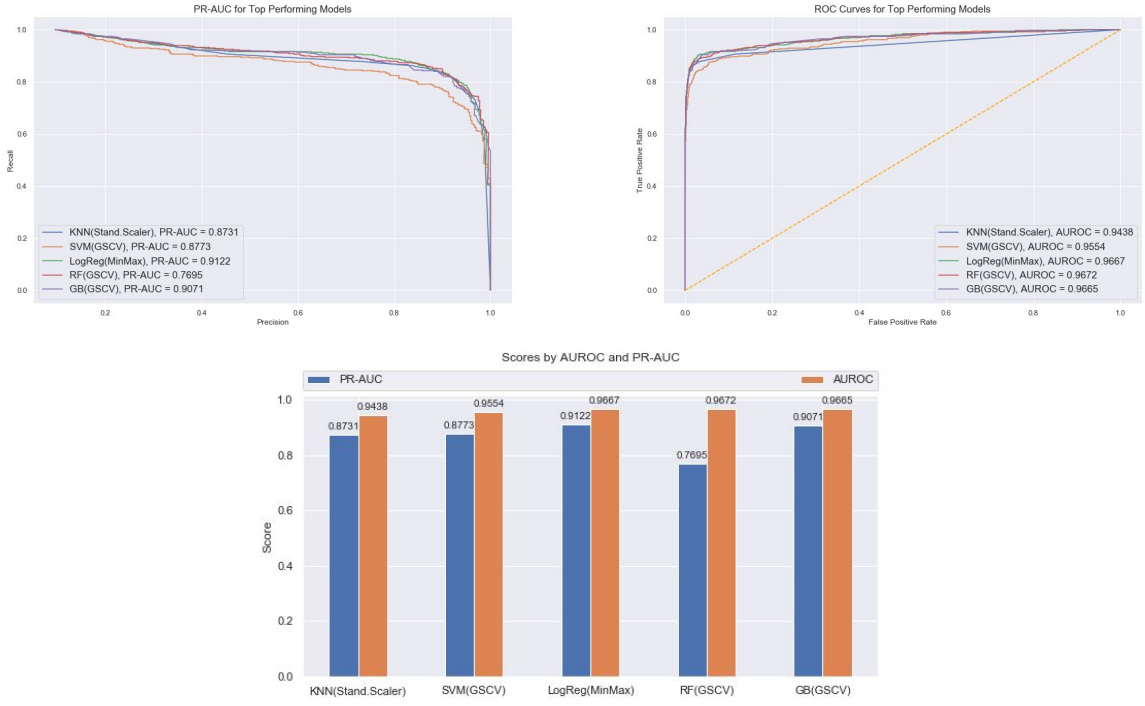






The final portion of the analysis consists of a comparison of each classifier's top-performing model. The models are KNN with standard scaled data, SVM with raw data, logistic regression with min-max scaled data, the hypertuned Random Forest, and hypertuned Gradient Boosting. The top performing model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



In summary, pulsar stars play an important role within the field of astronomy and are particularly dynamic as a subject to research because of the impact they have on a variety of astronomical research topics. Stemming from the accessibility of relatively inexpensive, powerful, and efficient computing power paired with large-scale data sets, machine learning has been proven to be an effective tool for making predictions regarding the discovery of pulsar stars. Exploring and understanding the dataset was a crucial step in designing the machine learning process. Considering multiple algorithms for testing and comparison was particularly useful and could potentially serve as a starting point for future research to build upon. Given the propensity for technical advancements in computer science, not only is machine learning a valuable tool for today's analytics, one can easily imagine a future where machine learning practices continue to flourish in research and business.