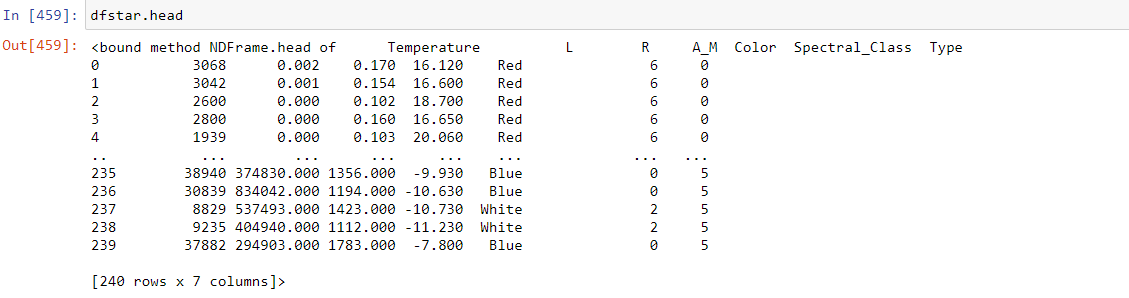
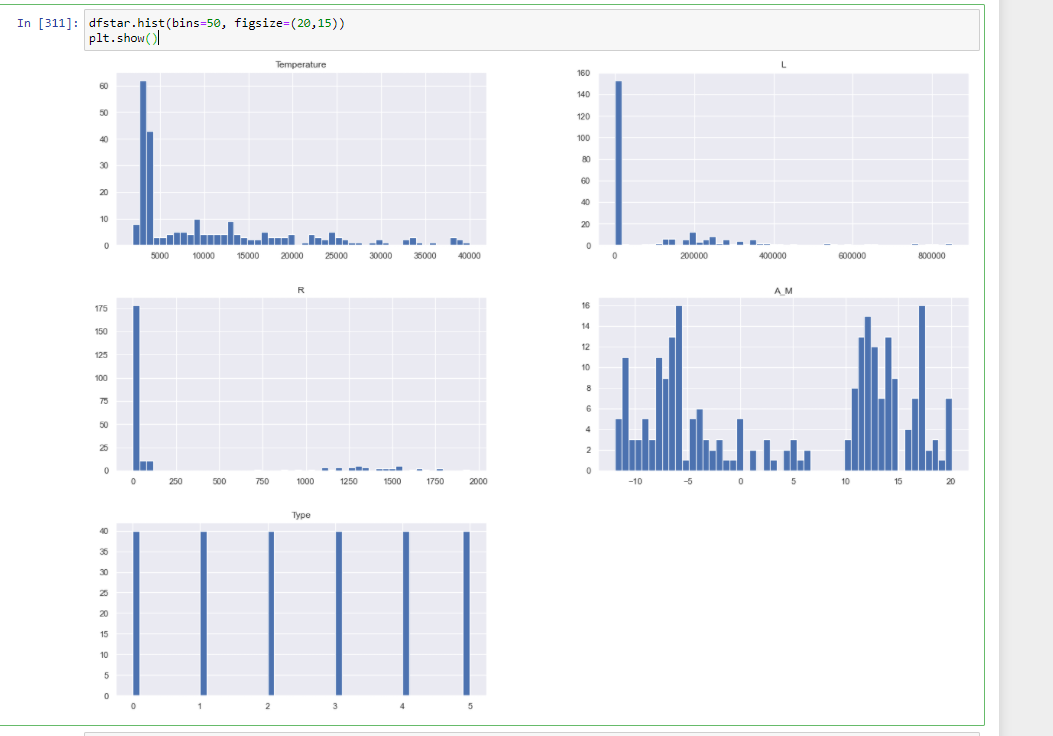
John Robinson

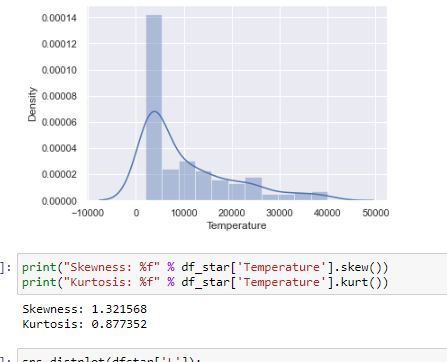
Machine Learning Final Project

Star Type Classification Data Analysis

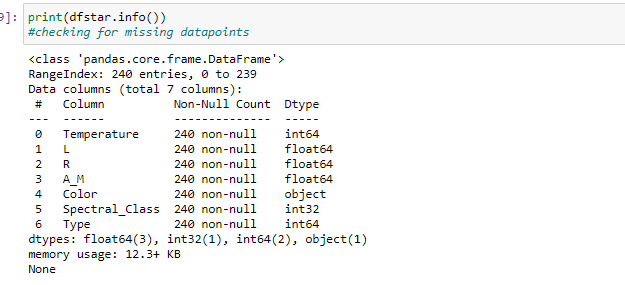
* The domain I chose for my project was based on a dataset I found intriguing. It is a NASA star classification analysis. I wanted to explore the dataset and see what interesting things I could find on the data in terms of correlation and connection between the datapoints using machine learning to do so. First thing is to establish what the dataset is, it consists of Temperature', 'L', 'R', 'A\_M', 'Color', 'Spectral Class', 'Type'. The Temperature is self-explanatory but it is listed in kelvin, L is the luminosity of the star which based off of the average luminosity of the sun measured at about 3.828 x 10^26 Watts so that’s what one is in that column. Next is R which was Radius of the star relative to our sun where average radius of the sun is 6.9551 x 10^8 m. A\_M is absolute magnitude which is the measure of the luminosity of a celestial objects. Color is just the color of the star, spectral class is just a another categorical data column can be O,B,A,F,G,K,M. Lastly is type which denotes at which point of life the star is in or its situation with a number to represent the type so Red Dwarf is 0 Brown Dwarf is 1 a White Dwarf is 2, a Main Sequence is 3,Super Giants are 4 and Hyper Giants are 5.



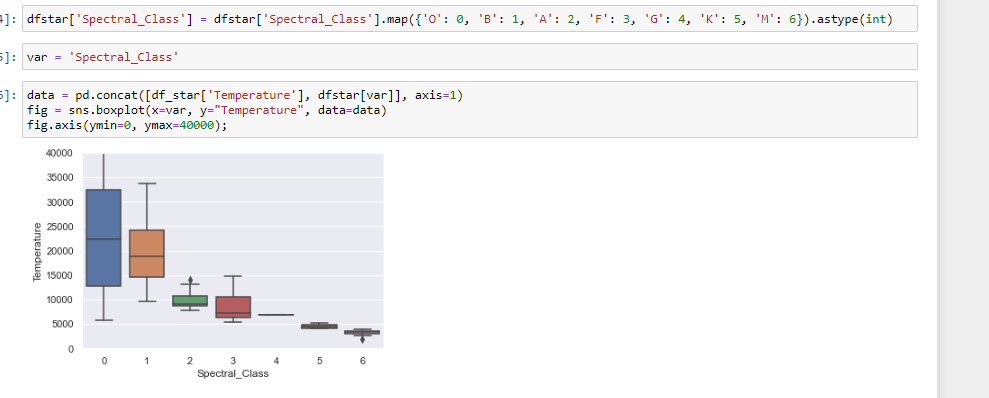
First thing I did after choosing a dataset was go through the data and decide which parts of the data were the most important to a comprehensive data exploration and classification of the data. I decided to explore temperature first as I thought it would be strongly correlated to the other datapoints. I first plotted the data and checked skewness and kurtosis in the data to see if my initial thoughts were correct.

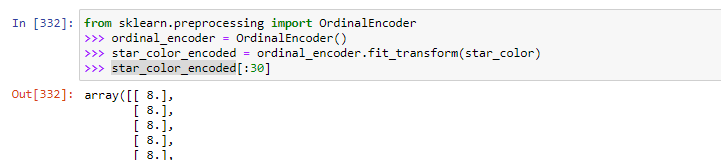


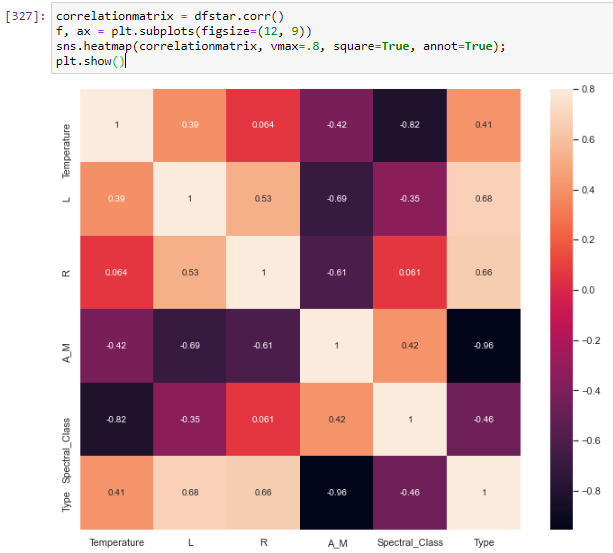
Temperature had a pretty good skewness and kurtosis so I decided to move forward with it being the major datapoint, but also checking rest of the data but none were as good as temp. Before moving forward I did some data cleaning making sure there was no missing datapoints.



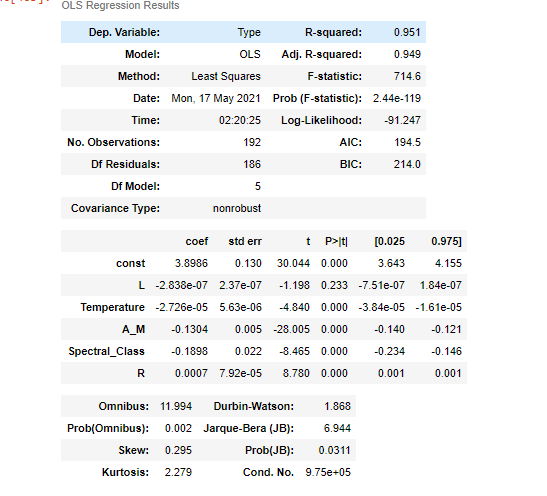
Then transferred the categorical columns to numerical categories to be analyzed, so I converted color and spectral class.

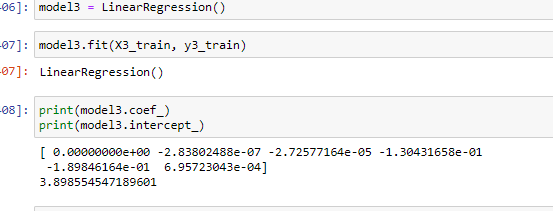


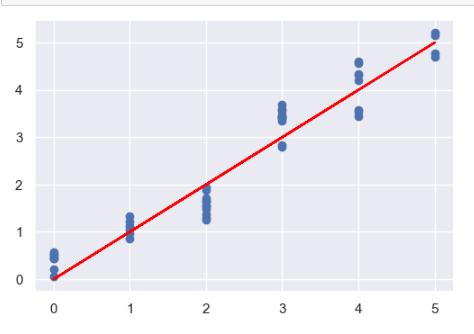
Once all the data was cleaned up I selected my models moving forward. I started with a correlation matrix seeing several data points of correlation but the strongest ones within the Type category. Specifically, a correlation with Luminosity and Radius.



My initial investigation into temperature was not as supported as the Type variable. I did dive deeper into temperature, but its statistics were as relevant so Types results. Next, I began training the dataset with twenty percent reserved for testing and got a decent P value with only irrelevant datapoint in L as anything under 0.05 is relevant data.

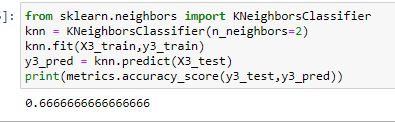


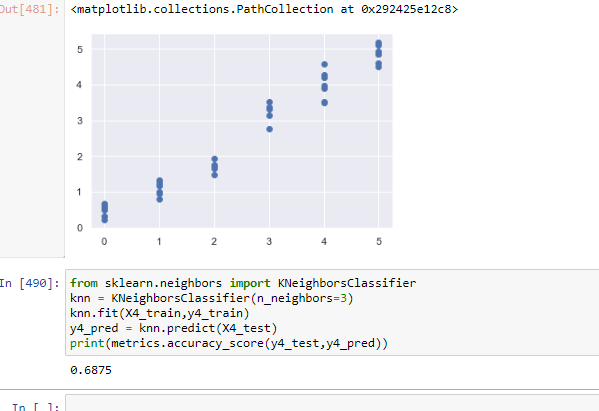
Next I began selecting models to train with the dataset. I performed several types of regression and classification. I started with simple linear regression and got decent results



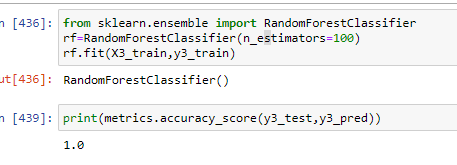
I applied several classifiers

Knn, the best result I got for the number k was 2 but still low accuracy but probably due to spread out data. Had some better results removing L which was deemed irrelevant but KNN doesn’t seem to be a good classifier a dataset organized such as this one all split up.

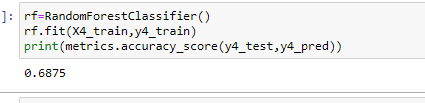




Random forest, unusually small dataset might be causing accuracy of 1 but unsure if error or not.



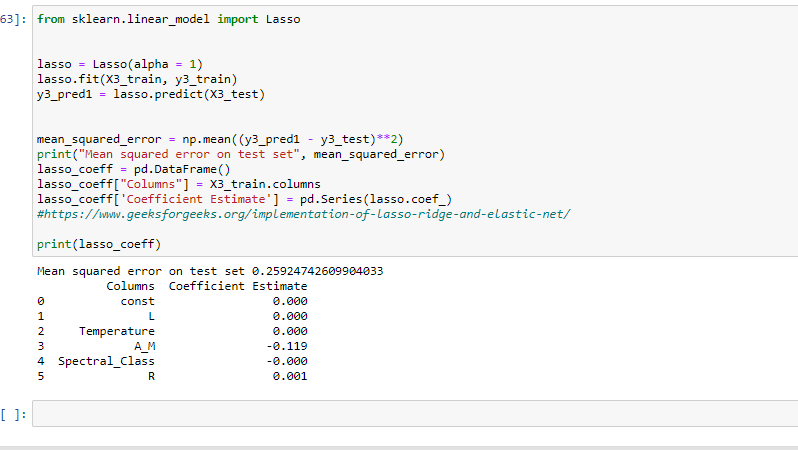
Removing L from the dataset due to its P value reducing the accuracy to similar that of the knn, unforunatly all the



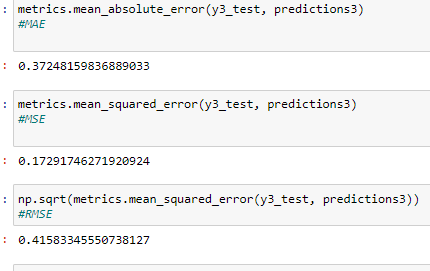
I got the idea to apply some different not previously discussed classifiers from another Kaggle user called Ada boost and gradient boost. The boost functions had a somewhat similar issue as my other classifiers, either 100% or a low end accuracy closer to 60.



Lastly I applied lasso regression



Below MAE, MSE and RMSE found from linear regression, just comparing to lasso regression



Sources

Dataset: https://www.kaggle.com/brsdincer/star-type-classification

<https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python#Out-liars>!

<https://www.kaggle.com/arpithaananth/nasa-star-type-classification-eda>

<https://www.geeksforgeeks.org/implementation-of-lasso-ridge-and-elastic-net/>