Stellar Identification

Nicholas Morris

Machine Learning

Wrangling

Removing Unnecessary Data

Removing ID columns that took on unique values for every row or took on a constant value for each row.

Feature Engineering

Transformations

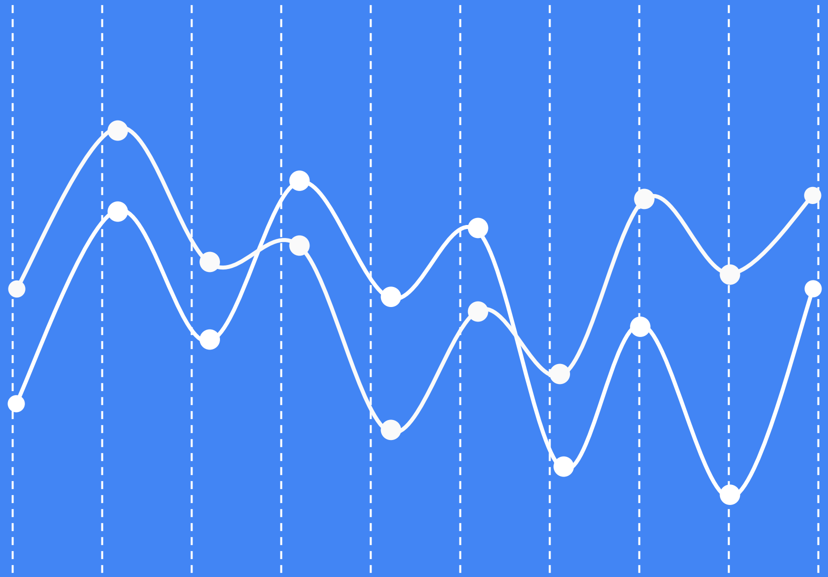
Attempting feature transformations such as Atwood Numbers, Binning, Reciprocals, and Interactions.

Modeling

Predictions

Training logistic regression, XGBoost, and deep learning neural network models. Evaluating performance. Computing feature drift to signal retraining.

Wrangling



Dataset

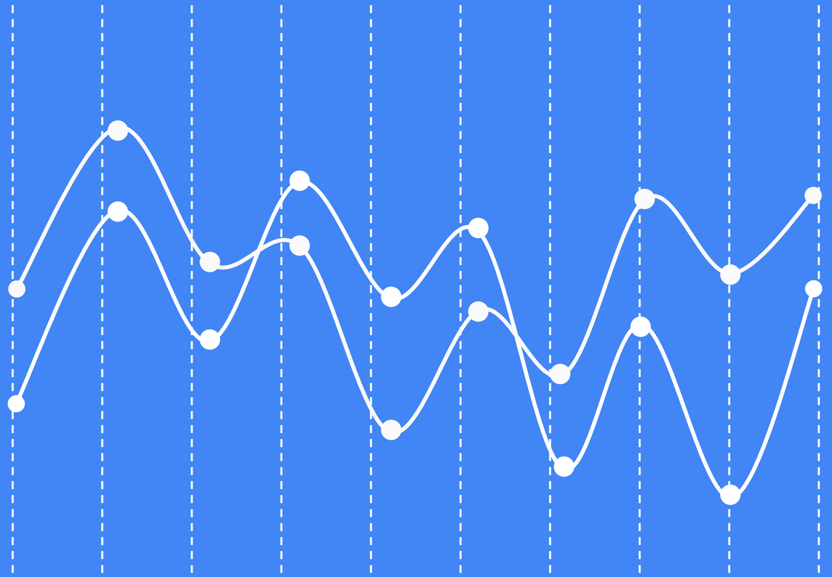
Below is two random rows of the data. There are 100,000 total stellar objects. The target we are predicting is class. [Link to the dataset]

alpha	delta	u	g	r	i	z	run_ID	cam_col	field_ID	class	redshift	plate	MJD	fiber_ID
13.161838	20.329558	23.96359	24.41554	21.92542	20.89438	19.96155	7923	5	264	GALAXY	0.762486	7619	56900	817
200.572696	38.674799	21.25550	21.08849	20.97834	20.77301	20.56856	3900	3	456	QSO	0.766414	8845	58159	423

Removing ID's

The dataset is ready for machine learning without any necessary preprocessing. The one step that was taken is removing three ID column: obj_ID, rerun_ID, and spec_obj_ID. This was done because the ID column took on unique values for every single row or took on a constant value for every row.

Feature Engineering



Atwood Numbers

An Atwood Number is a calculation that shows the relative change between two variables. The formula for two variables x and y is:

$(x - y) / (x + y)$

This calculation was done on all pairs of non-binary variables; but did not improve model performance, so, it was left out of the final model.

Binning

Binning is when a non-binary variable is grouped into histogram bins, and represented as binary variables.

Binning did not improve model performance, so, it was left out of the final model.

Reciprocals

A reciprocal is when a non-binary variable x is calculated as 1 / x.

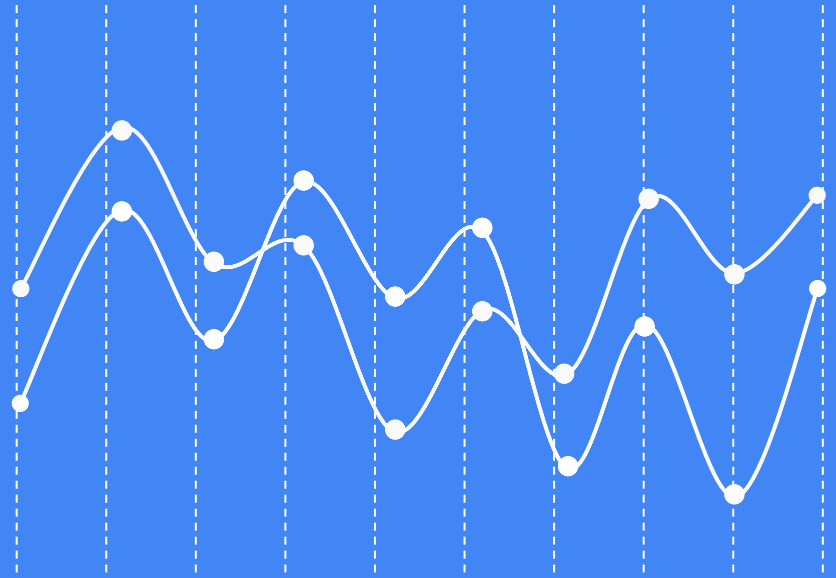
Reciprocals did not improve model performance, so, it was left out of the final model.

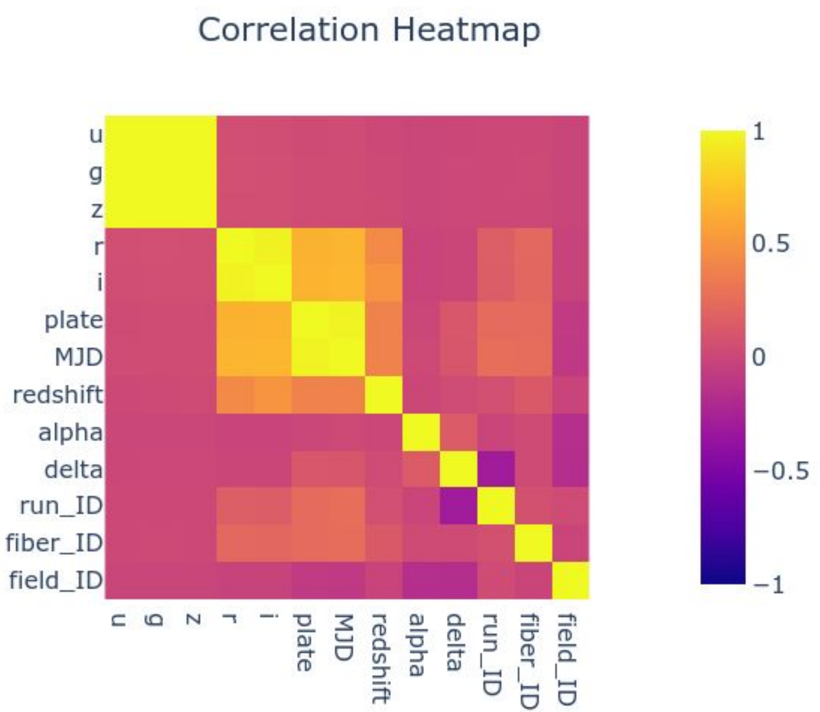
Interactions

An interaction is when two variables x and y are calculated as x * y. Reciprocals were fed into this calculation to generate x / y as well.

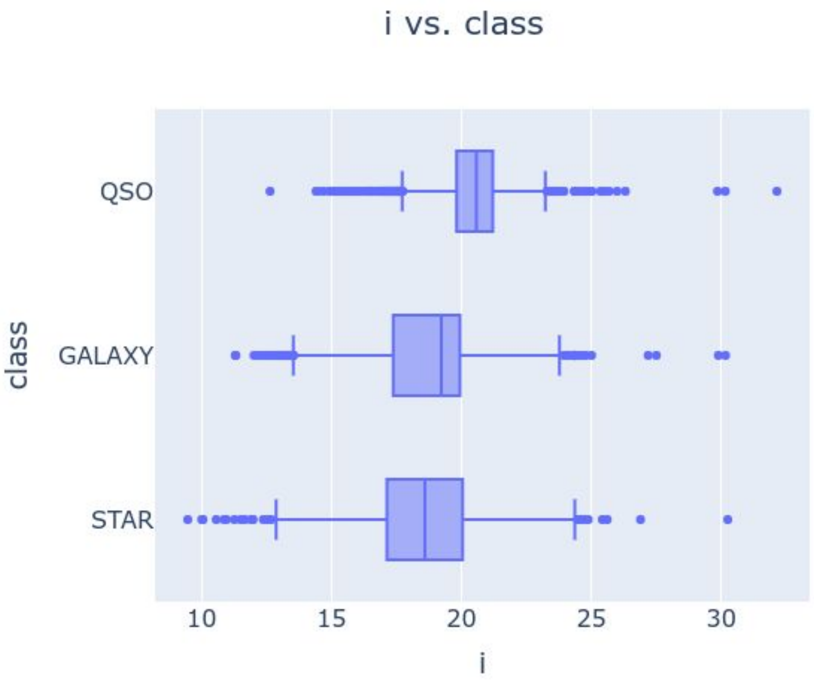
Interactions did not improve model performance, so, it was left out of the final model.

Data Exploration

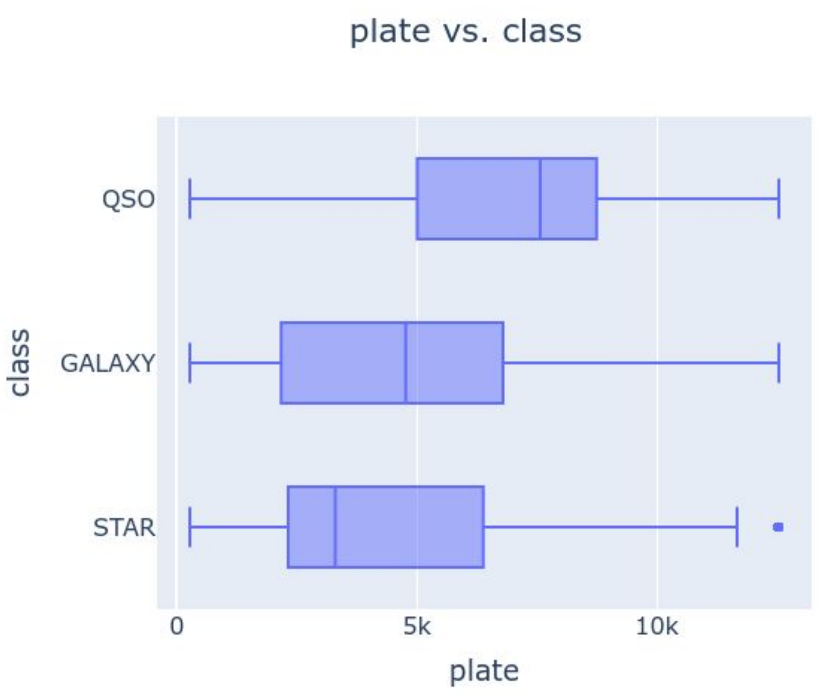




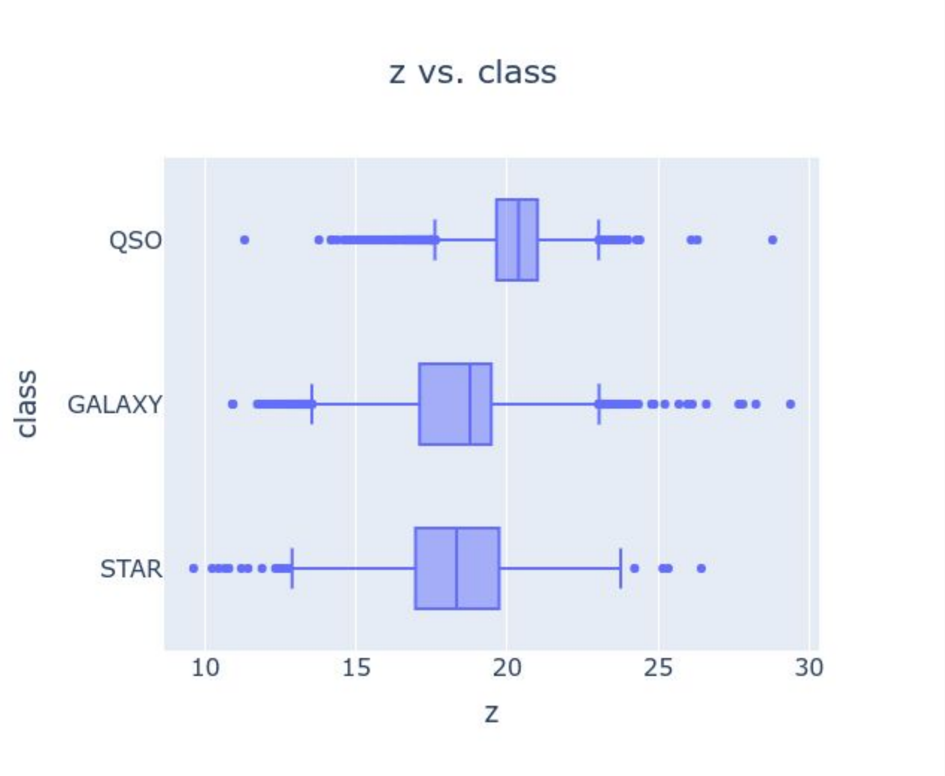
There are two zones in the heatmap where there's strong correlations. The first is between u, g, and z which are the ultraviolet, green, and infrared filters respectively. The next is between r, i, plate, MJD, and redshift which are the red filter, the near infrared filter, plate ID, modified julian date, and redshift value based on an increase in the wavelength respectively.



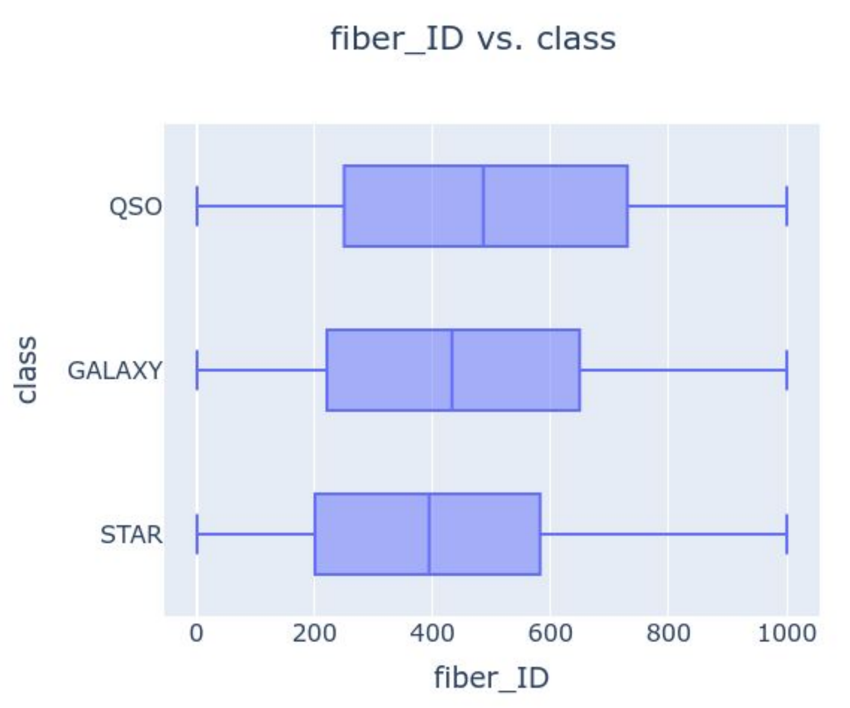
We can see that as the near infrared filter increases the likelihood of a particular stellar object changes from Star to Galaxy to Quasar when looking at the median line.



We can see that as the plate ID increases the likelihood of a particular stellar object changes from Star to Galaxy to Quasar when looking at the median line.

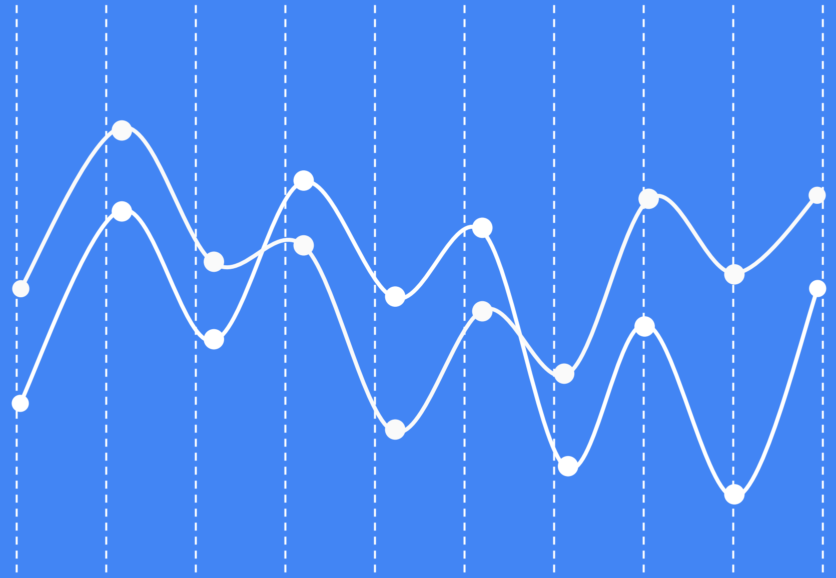


We can see that as the infrared filter increases the likelihood of a particular stellar object changes from Star to Galaxy to Quasar when looking at the median line.



We can see that as the fiber ID increases the likelihood of a particular stellar object changes from Star to Galaxy to Quasar when looking at the median line.

Modeling



Model Parameters

Logistic Regression

Library: scikit-learn
Penalty: L1
Number Of Alphas: 16
Cross Validation Folds: 3
Tolerance: 1e-4
Max Iterations: 100

XGBoost

Library: xgboost
Boosting Rounds: 100
Learning Rate:
0.001, 0.01, 0.1
Max Depth:
5, 7, 10, 14, 18
Min Child Weight: 1
Column Sampling: 0.8
Row Sampling: 0.8
Cross Validation Folds: 3

Neural Network

Library: Tensorflow
Epochs: 500
Learning Rate:
0.0001, 0.001, 0.01
Batch Size: 16
Layers: 10
Nodes Per Layer:
32, 64, 128, 256, 512
Solver: Adam
Cross Validation Folds: 3

Model Comparison

Logistic Regression

Accuracy: 0.95
F1: 0.94
In Control: 99%

Model Indicators:

1. redshift
2. i
3. r
4. u
5. z

XGBoost

Accuracy: 0.98
F1: 0.98
In Control: 98.4%

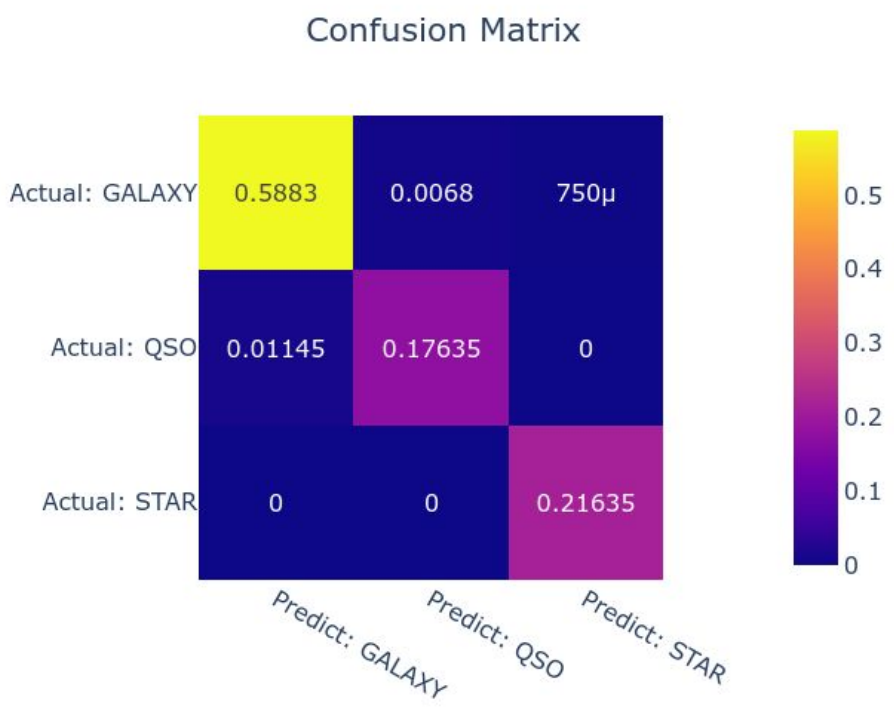
Model Indicators:

1. redshift
2. g
3. plate
4. z
5. u

Neural Network

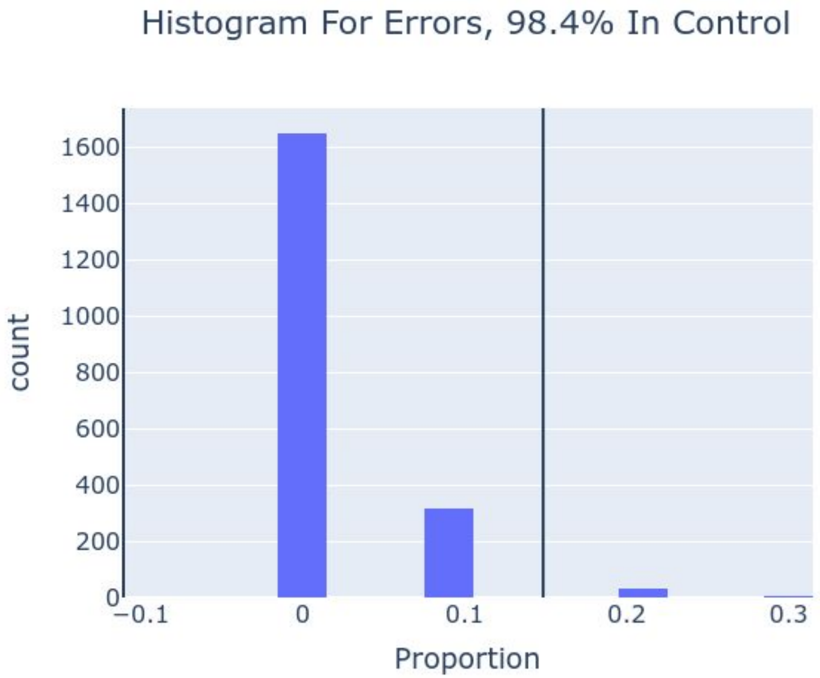
Accuracy: DNF
F1: DNF
In Control: DNF

Model Indicators:
DNF



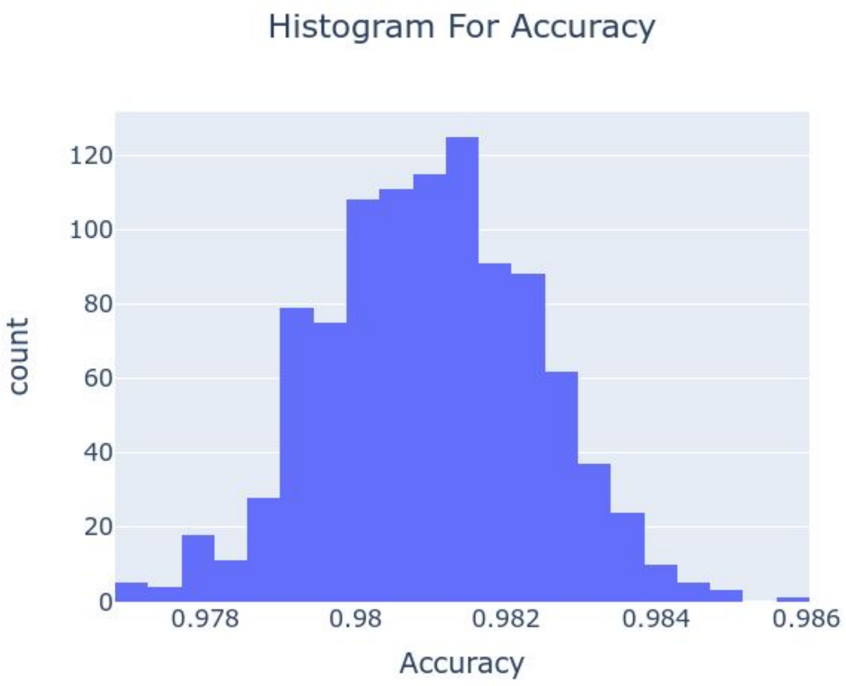
These predictions come from the XGBoost model. These predictions are done on 20% of the data that the model did not see during training. Only 2% of the predictions are wrong, and 98% of the predictions are correct. There's a slightly stronger tendency to predict a Quasar as a Galaxy compared to other wrong predictions.

A snapshot of the predictions show that most of them are on target. There’s one where a Quasar was predicted as a Galaxy.



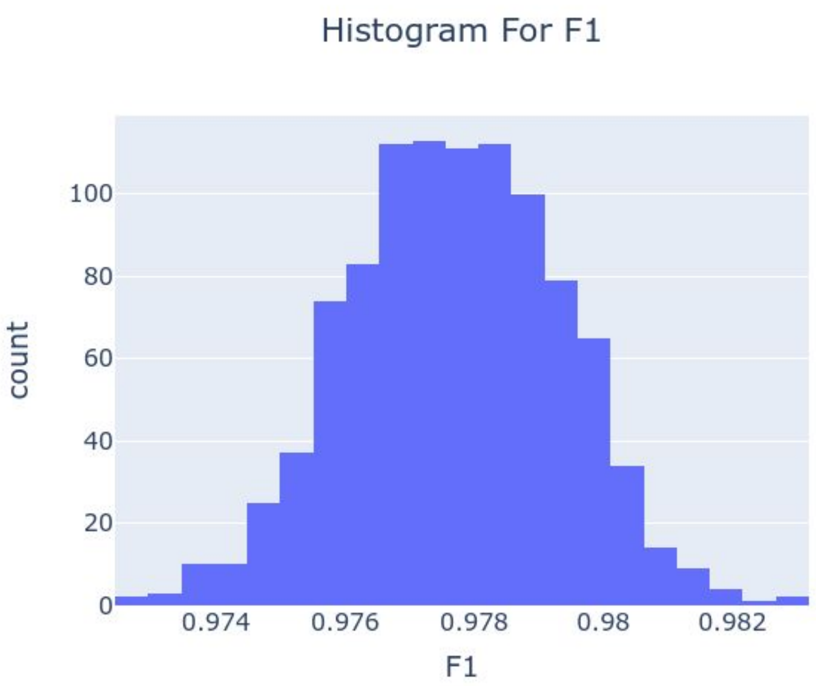
The errors are the fraction of 10 predictions that were wrong.

The errors are most likely to be 0%. Control limits were computed on the errors and we can see that the prediction error is mostly under control.



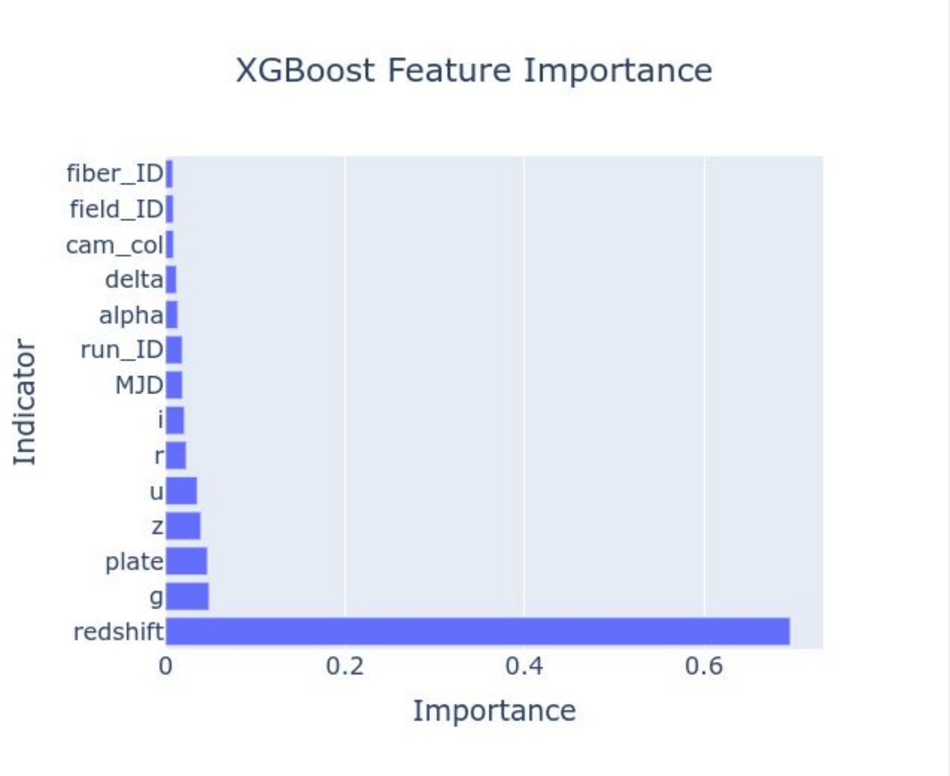
The prediction error was resampled 1000 times at a 50% sampling rate with replacement. Then Accuracy was computed on each sample to get a distribution.

Accuracy has a tight range between 0.978 and 0.986, which is good. Accuracy has a bell shape, which is good.

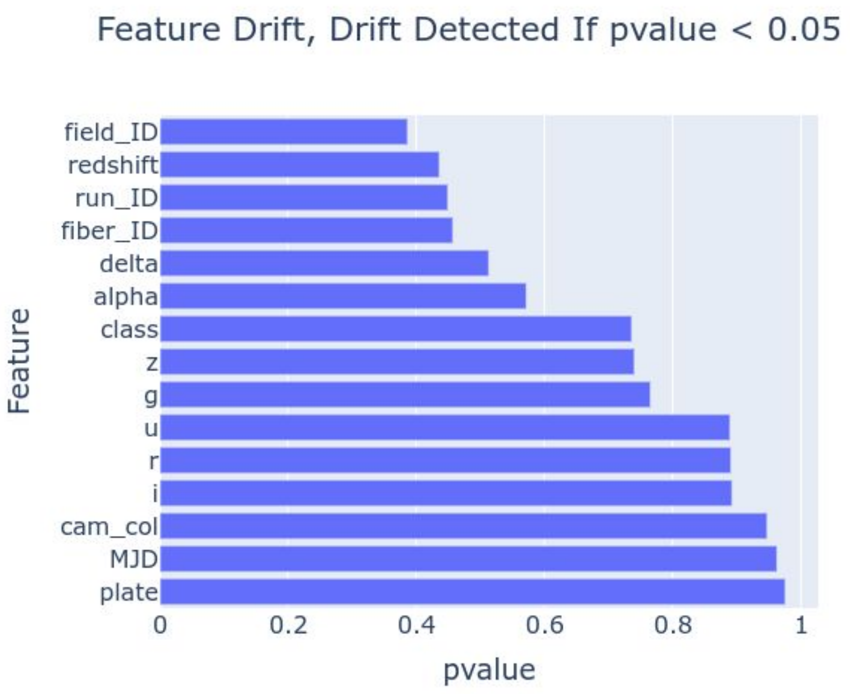


The prediction error was resampled as previously mentioned to get a distribution for F1. F1 is a combination of Precision and Recall. Precision tells us how well the model doesn't label a stellar object as another one. Recall tells us how well the model labels all stellar objects.

F1 has a tight range between 0.974 and 0.982, which is good. F1 has a bell shape, which is good.

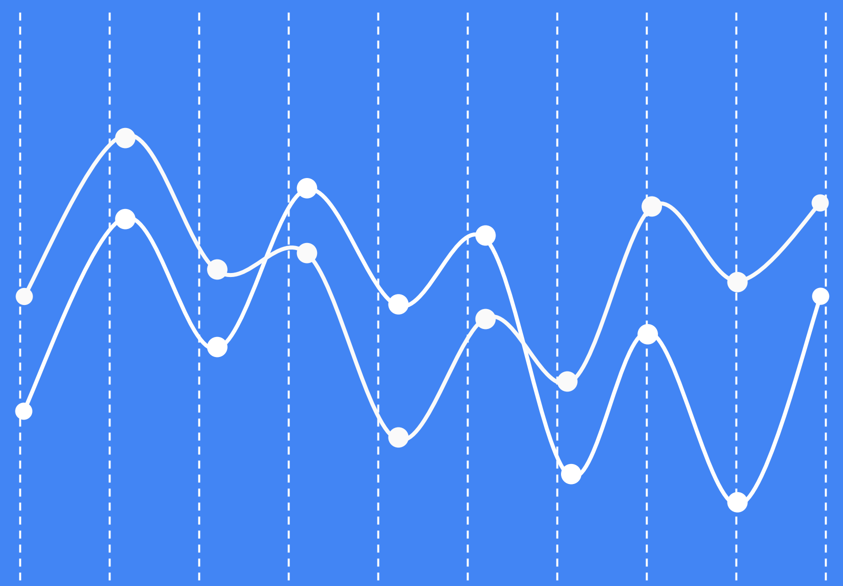


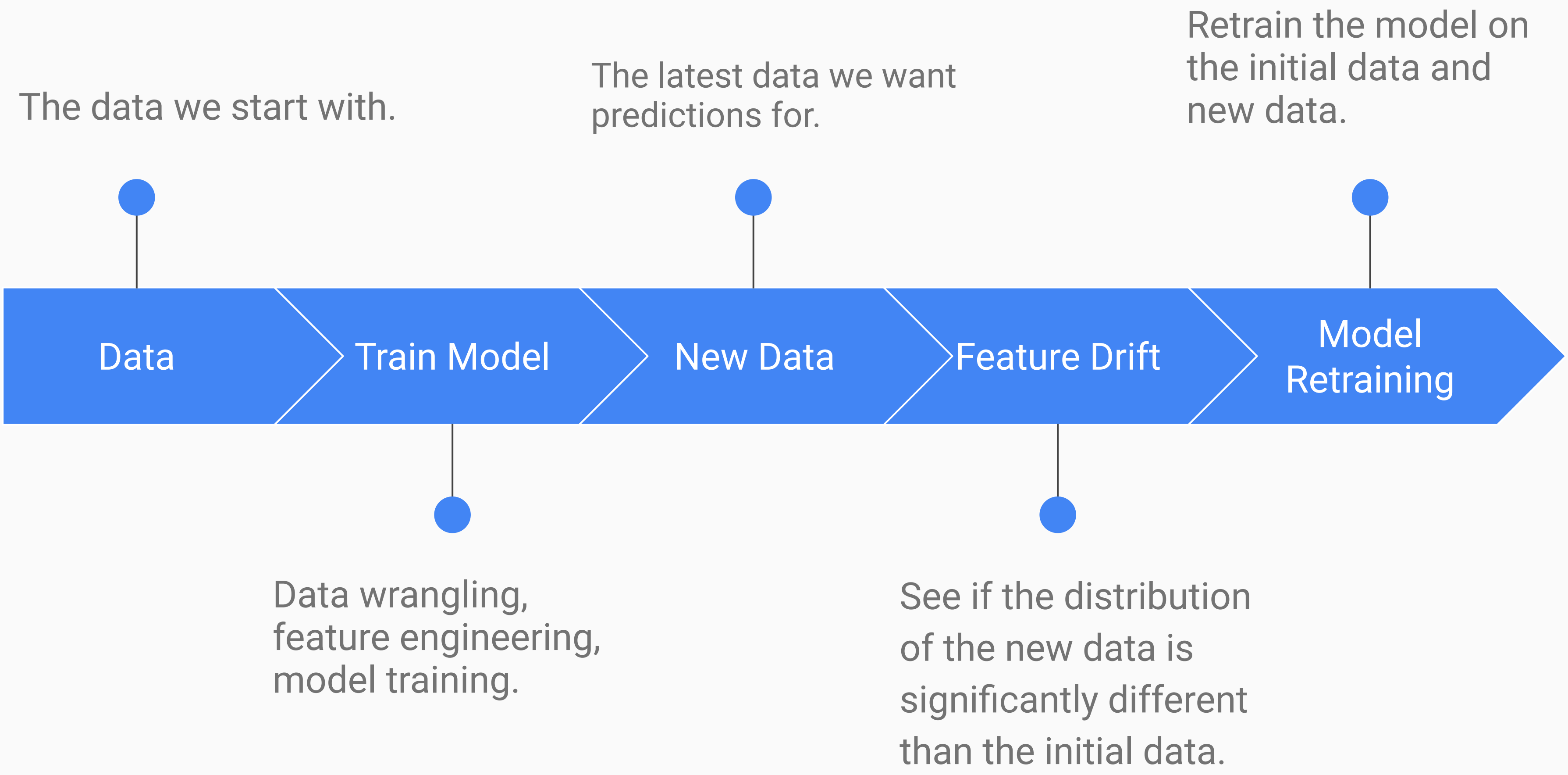
The importance of all 14 features is shown to the left. Redshift is vastly more important than the rest of the features for predicting stellar objects.



A Kolmogorov-Smirnov test was performed for each column in the data to see if the distribution of the testing data is the same as the training data. If the testing data does not share the same distribution as the training data, then there is a drift, which signals for model retraining. All of the columns do not experience a drift, which is good.

Deployment




The data we start with.
Data
Train Model
Data wrangling, feature engineering, model training.
The latest data we want predictions for.
New Data
Feature Drift
See if the distribution of the new data is significantly different than the initial data.
Model Retraining
Retrain the model on the initial data and new data.

Thank You