Stellar Classification

Presented by Parker Stratton
13 February 2022

Project Task and Purpose

Task: Prepare stellar classification dataset for application of Machine Learning models and conduct exploratory analysis of the dataset.

Purpose: Ensure data is ready to be utilized by Machine Learning algorithms, without any leakage. Provide deeper understanding of the data and its interrelated aspects that may improve modeling.

Dataset Summary

Stellar Classification Background:

- Celestial objects that emit EM radiation have “finger-prints” due to the wavelengths they absorb
- Celestial objects moving toward or away from the observer at high enough relative speeds experience red shift (elongation) or blue shift (compression) of their perceived EM emission – similar to the Doppler effect on sound
- By looking at a celestial object’s EM “finger-prints” and red/blue shift, a determination of the objects type can be made

Dataset

100,000 observations from the Sloan Digital Sky Survey (SDSS)

```
[18]  1 df = pd.read_csv('/content/star_classification.csv')
      2 df.head()
```

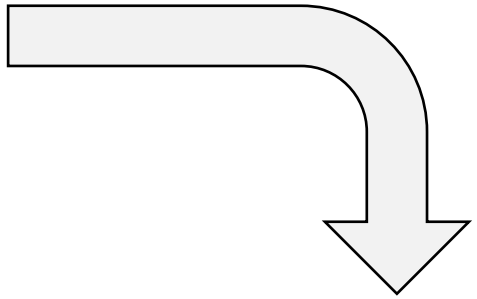
	obj_ID	alpha	delta	u	g	r	i	z	run_ID	rerun_ID	cam_col	field_ID	spec_obj_ID	class	redshift	plate	MJD	fiber_ID
0	1.237661e+18	135.689107	32.494632	23.87882	22.27530	20.39501	19.16573	18.79371	3606	301	2	79	6.543777e+18	GALAXY	0.634794	5812	56354	171
1	1.237665e+18	144.826101	31.274185	24.77759	22.83188	22.58444	21.16812	21.61427	4518	301	5	119	1.176014e+19	GALAXY	0.779136	10445	58158	427
2	1.237661e+18	142.188790	35.582444	25.26307	22.66389	20.60976	19.34857	18.94827	3606	301	2	120	5.152200e+18	GALAXY	0.644195	4576	55592	299
3	1.237663e+18	338.741038	-0.402828	22.13682	23.77656	21.61162	20.50454	19.25010	4192	301	3	214	1.030107e+19	GALAXY	0.932346	9149	58039	775
4	1.237680e+18	345.282593	21.183866	19.43718	17.58028	16.49747	15.97711	15.54461	8102	301	3	137	6.891865e+18	GALAXY	0.116123	6121	56187	842

Dataset from: https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17

Data Cleaning

```
[9]   1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 18 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   obj_ID       100000 non-null  float64
 1   alpha        100000 non-null  float64
 2   delta        100000 non-null  float64
 3   u            100000 non-null  float64
 4   g            100000 non-null  float64
 5   r            100000 non-null  float64
 6   i            100000 non-null  float64
 7   z            100000 non-null  float64
 8   run_ID       100000 non-null  int64
 9   rerun_ID     100000 non-null  int64
 10  cam_col      100000 non-null  int64
 11  field_ID     100000 non-null  int64
 12  spec_obj_ID  100000 non-null  float64
 13  class        100000 non-null  object
 14  redshift     100000 non-null  float64
 15  plate        100000 non-null  int64
 16  MJD          100000 non-null  int64
 17  fiber_ID     100000 non-null  int64
dtypes: float64(10), int64(7), object(1)
memory usage: 13.7+ MB
```



```
[34]   1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 99999 entries, 0 to 99999
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   u         99999 non-null  float64
 1   g         99999 non-null  float64
 2   r         99999 non-null  float64
 3   i         99999 non-null  float64
 4   z         99999 non-null  float64
 5   class     99999 non-null  object
 6   redshift  99999 non-null  float64
dtypes: float64(6), object(1)
memory usage: 6.1+ MB
```

Cleaning Actions:

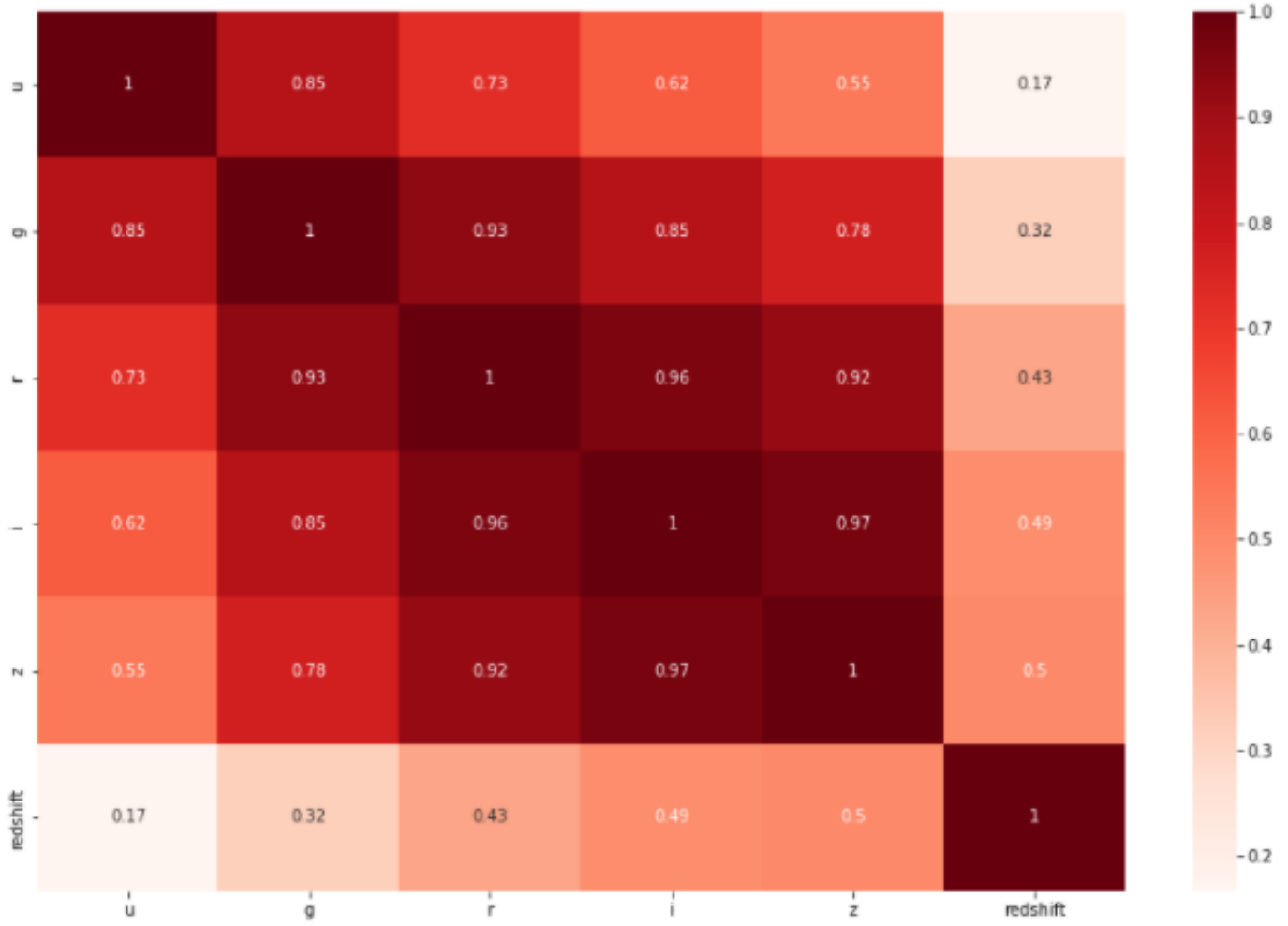
- Missing data
- Duplicates
- Dropping unnecessary rows
- Checking for and handling inconsistent and unrealistic data

Key Take Away:

- Data cleaning produced a modeling ready data frame with loss of only a single row and streamlined features relevant to classification predictions

Exploratory Analysis

```
[20]  1 corr = df.corr()
      2 plt.subplots(figsize = (15, 10))
      3 sns.heatmap(corr, cmap = 'Reds', annot = True);
```



Exploratory Visualizations

- Univariate Analysis
- Red Shift feature exploration
- Target class frequencies
- Feature Correlation

Key Take Aways

- High level of correlation across the features (except redshift)
- Target classes unbalanced

Conclusion

Challenges

• No significant challenges so far, however application of machine learning models may introduce unforeseen challenges with prediction

Of note:
This data set only covers the difference between galaxies, quasars, and stars. While important distinctions, this is certainly not all encompassing of the vast variety of celestial objects. With additional data, similar ML applications can be applied for more specific classification with more far reaching benefits to the scientific community

Summary

• Data is clean

• Exploratory analysis complete

Way Forward

• Development, training, and evaluating classification prediction models with cleaned data

• Intended Outcome: Enable accurate automated classification of celestial objects based on their EM emissions and redshift