An Overview of the Project

- We have ~ 20,000 DESI spectra
- For each source there are sets of coefficients that describe the spectra
- I want to do a project that incorporated machine learning models

Utilize these dimensionally reduced coefficients with machine learning regression models to predict physical parameters

Dimensionality Reduction: Less is more

Principal Component Analysis (PCA)

- We have 5 PCA coefficients to consider
- Order matters for PCA

What number of PCA coefficients produces the best predictions?

Uniform Manifold Approximation and Projection (UMAP)

- We have 2, 3, and 5 UMAP coefficients
- Order does NOT matter

How do PCA and UMAP coefficients compare when it comes to predicting parameters?

The Dynamic Duo of Predictive Modeling

Random Forest Regressor

A popular supervised machine learning algorithm that combines multiple decision trees to make predictions

XGBoost Regressor

A newer algorithm, it builds decision trees in a sequential manner, with the following tree learning from the errors of the previous tree

Parameters to Regress

Using PROVABGS, we can calculate

PRObabilistic Value-Added Bright Galaxy Survey (PROVABGS)

- Mass weighted age
- Average star formation rate (SFR) over 1 Gigayear

Evaluation Metrics

Normalized Average Absolute Deviation (NAAD)

- Quantifies the average deviation of a set of data points from their mean value

Root Mean Squared Error (RMSE)

- It measures the difference between the predicted and actual values of the target variable

Spearman Correlation Coefficient (SCC)

- ranges from -1 to +1
  - -1 indicates a perfect negative correlation
  - 0 indicates no correlation
  - +1 indicates a perfect positive correlation

Results - Mass Weighted Age

RFR	PCA(1)	PCA(2)	PCA(3)	PCA(4)	PCA(5)	UMAP (2)	UMAP (3)	UMAP (5)
NAAD	0.0687	0.0685	0.0662	0.0639	0.0679	0.0640	0.0671	0.0647
RMSE	0.0870	0.0882	0.0849	0.0829	0.0888	0.0824	0.0833	0.0815
SCC	0.404	0.410	0.413	0.414	0.436	0.423	0.406	0.416

XGB	PCA(1)	PCA(2)	PCA(3)	PCA(4)	PCA(5)	UMAP (2)	UMAP (3)	UMAP (5)
NAAD	0.0656	0.0664	0.0637	0.0676	0.0614	0.0677	0.0678	0.0681
RMSE	0.0819	0.0853	0.0811	0.0866	0.0804	0.0861	0.0866	0.0871
SCC	0.384	0.421	0.412	0.420	0.452	0.343	0.410	0.411

Results - Average SFR

RFR	PCA(1)	PCA(2)	PCA(3)	PCA(4)	PCA(5)	UMAP (2)	UMAP (3)	UMAP (5)
NAAD	0.4947	0.4706	0.4538	0.4408	0.4228	0.4551	0.4514	0.4454
RMSE	0.5014	0.4777	0.4618	0.4463	0.4288	0.4621	0.4587	0.4554
SCC	0.291	0.416	0.445	0.497	0.578	0.451	0.492	0.510

XGB	PCA(1)	PCA(2)	PCA(3)	PCA(4)	PCA(5)	UMAP (2)	UMAP (3)	UMAP (5)
NAAD	0.4934	0.4851	0.4526	0.4435	0.4259	0.4638	0.4436	0.4336
RMSE	0.4979	0.4923	0.4601	0.4494	0.4393	0.4702	0.4525	0.4382
SCC	0.291	0.357	0.446	0.501	0.559	0.406	0.506	0.507

Future Work

- Trying auto-encoder coefficients

  How do predictions change as we introduce different coefficients?

- Testing TensorFlow

  To what extent do predictions change as we use different ML models?

- Regressing other parameters

  Which combination of coefficients and ML models produces the more accurate predictions?