Exploring the Universe with Supervised Machine Learning: Analysing Exoplanetary Atmospheres

Project Midway Presentation

By Ayush Singhal , Gaurav Shukla

Introduction

When stellar light passes through a planet's atmosphere, molecules in the atmosphere can absorb or re-emit different light wavelengths, which leaves a characteristic fingerprint on the light that reaches us. By measuring the change in the dips (transit depth) as a function of wavelength/frequency of light, we can work out which molecules or clouds absorb photons in the atmosphere and understand the planet's chemistry, temperature, cloud coverage, wind speeds, and climate.

One of the main challenges of studying exoplanetary atmospheres is the complexity of the planetary models required to understand the complex processes happening in their atmospheres, including chemistries, clouds, and dynamics. To overcome the challenges of analyzing spectral data from exoplanetary atmospheres, machine learning (ML) techniques can be used. By using ML algorithms to classify and characterize exoplanetary atmospheres based on their spectral features, we can obtain more reliable and comprehensive results than traditional manual inspection and interpretation methods. ML techniques can also help identify potential candidates for further study and determine which exoplanets may have the necessary conditions for life to exist.

Literature Review (MN18)

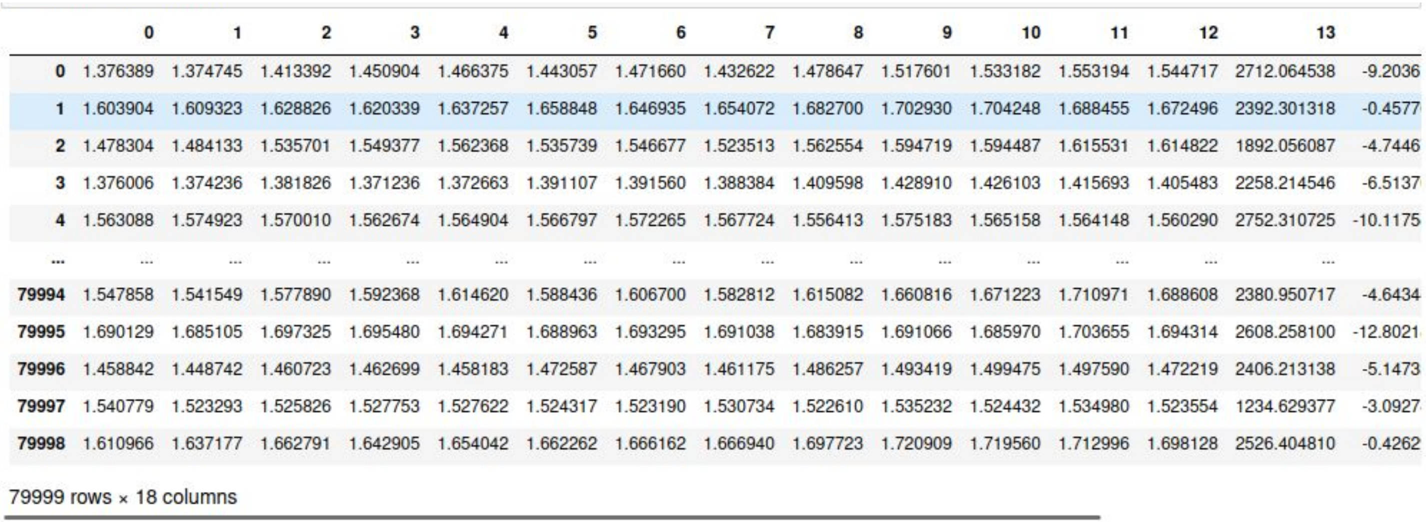
The introduction of this paper discusses the importance of studying exoplanet atmospheres and the challenges associated with analyzing the data obtained from observations. It highlights the need for new methods that can accurately and efficiently retrieve atmospheric parameters from exoplanet data. The paper proposes a new method that uses machine learning algorithms to analyze exoplanet data and retrieve atmospheric parameters, and it presents the results of testing this method on simulated data. The introduction also provides an overview of the structure of the paper and the contributions of the research.

The paper uses simulated data to test the new method for analyzing exoplanet atmospheres using machine learning algorithms. The authors generated a grid of atmospheric models with varying parameters, such as the presence and abundance of different molecules, and used this grid to train the machine learning algorithm. They then used the algorithm to retrieve atmospheric parameters from the simulated data and compared the results to the true parameters used to generate the data.

They trained their model on 80,000 synthetic spectra and used it to analyze 20,000 more synthetic spectra. They found that the outcomes of the retrievals converged when the number of trees used exceeded 100. They also tested the retrieval outcomes with different levels of assumed noise floors, which represent the uncertainty in the transit depths of the data points in the synthetic WFC3 spectra. They found that the variance associated with the true versus predicted values of the parameters decreased when the assumed noise floor was lower. Overall, these tests demonstrate the robustness of the authors' implementation of the random forest method for analyzing the synthetic spectra.

Paper	DataSet	Model	Limitations
Márquez-Neila, Pablo et al. (2018)	Data based on a simple analytical model derived in Heng & Kitzmann (2017) which provides a range of atmospheric parameters and corresponding transit spectra for testing and validating inversion methods.	Random Forest	● Simulated Data ● Limited set of parameters ● Biases in the training data
Munsaket, Patcharawee et al. (2021)	Simulated photometric observational data of hot Jupiters generated by the PLanetary Atmospheric Transmission for Observer Noobs (PLATON) software	Random Forest	● Simulated Data ● Focus on hot Jupiters ● Using just RF and no comparison
Nixon, Matthew C. and Nikku Madhusudhan.(2020)	The AURA forward model (Pinhas et al. 2018), is used to generate synthetic spectra in the wavelength range of WFC3	Random Forest	● The paper only considers two hot Jupiter exoplanets ● Their approach struggles to deal with higher dimensional parameter spaces efficiently.

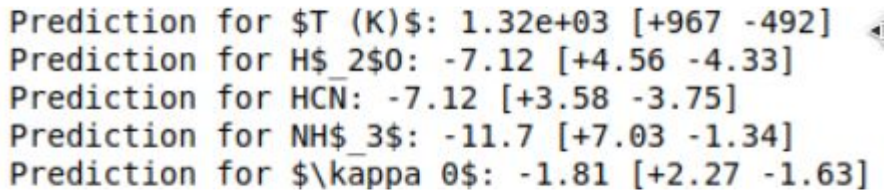
Simulated Data (100,000)(80% split for training and 20% for testing)



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
0	1.376389	1.374745	1.413392	1.450904	1.466375	1.443057	1.471660	1.432622	1.478647	1.517601	1.533182	1.553194	1.544717	2712.064538	-9.2036
1	1.603904	1.609323	1.628826	1.620339	1.637257	1.658848	1.646935	1.654072	1.682700	1.702930	1.704248	1.688455	1.672496	2392.301318	-0.4577
2	1.478304	1.484133	1.535701	1.549377	1.562368	1.535739	1.546677	1.523513	1.562554	1.594719	1.594487	1.615531	1.614822	1892.056087	-4.7446
3	1.376006	1.374236	1.381826	1.371236	1.372663	1.391107	1.391560	1.388384	1.409598	1.428910	1.426103	1.415693	1.405483	2258.214546	-6.5137
4	1.563088	1.574923	1.570010	1.562674	1.564904	1.566797	1.572265	1.567724	1.556413	1.575183	1.565158	1.564148	1.560290	2752.310725	-10.1175
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
79994	1.547858	1.541549	1.577890	1.592368	1.614620	1.588436	1.606700	1.582812	1.615082	1.660816	1.671223	1.710971	1.688608	2380.950717	-4.6434
79995	1.690129	1.685105	1.697325	1.695480	1.694271	1.688963	1.693295	1.691038	1.683915	1.691066	1.685970	1.703655	1.694314	2608.258100	-12.8021
79996	1.458842	1.448742	1.460723	1.462699	1.458183	1.472587	1.467903	1.461175	1.486257	1.493419	1.499475	1.497590	1.472219	2406.213138	-5.1473
79997	1.540779	1.523293	1.525826	1.527753	1.527622	1.524317	1.523190	1.530734	1.522610	1.535232	1.524432	1.534980	1.523554	1234.629377	-3.0927
79998	1.610966	1.637177	1.662791	1.642905	1.654042	1.662262	1.666162	1.666940	1.697723	1.720909	1.719560	1.712996	1.698128	2526.404810	-0.4262

79999 rows × 18 columns

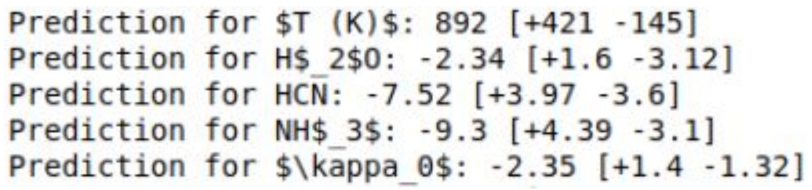
We used the dataset used by the authors of the MN18 paper for initial experimentation. The dataset of 100,000 noisy synthetic spectra was generated by using the forward model of Heng & Kitzmann (2017). The spectra were generated in the wavelength range 0.8 - 1.7 μm, and five parameters described each spectrum: temperature (T), volume mixing ratios of water (X_{H2O}), ammonia (X_{NH3}), and hydrogen cyanide (X_{HCN}), and a constant cloud opacity (k_o). The values of the parameters were chosen randomly from a uniform or log-uniform distribution.



```
Prediction for $T (K)$: 1.32e+03 [+967 -492]
Prediction for H$_2$O: -7.12 [+4.56 -4.33]
Prediction for HCN: -7.12 [+3.58 -3.75]
Prediction for NH$_3$: -11.7 [+7.03 -1.34]
Prediction for $\kappa 0$: -1.81 [+2.27 -1.63]
```

HD 209458b

Using the model trained on the dataset from the MN18 paper to predict the atmospheric composition of the planet HD209548b.

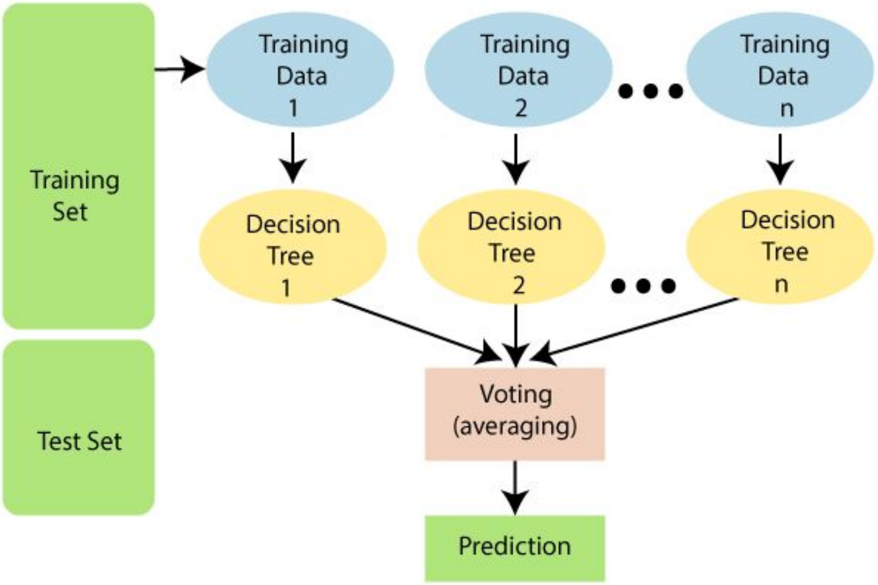


```
Prediction for $T (K)$: 892 [+421 -145]
Prediction for H$_2$O: -2.34 [+1.6 -3.12]
Prediction for HCN: -7.52 [+3.97 -3.6]
Prediction for NH$_3$: -9.3 [+4.39 -3.1]
Prediction for $\kappa_0$: -2.35 [+1.4 -1.32]
```

WASP 12-b

Using the model trained on the dataset from the MN18 paper to predict the atmospheric composition of the planet WASP12-b.

Random Forest Algorithm



Training Set

Test Set

Training Data 1

Training Data 2

Training Data n

Decision Tree 1

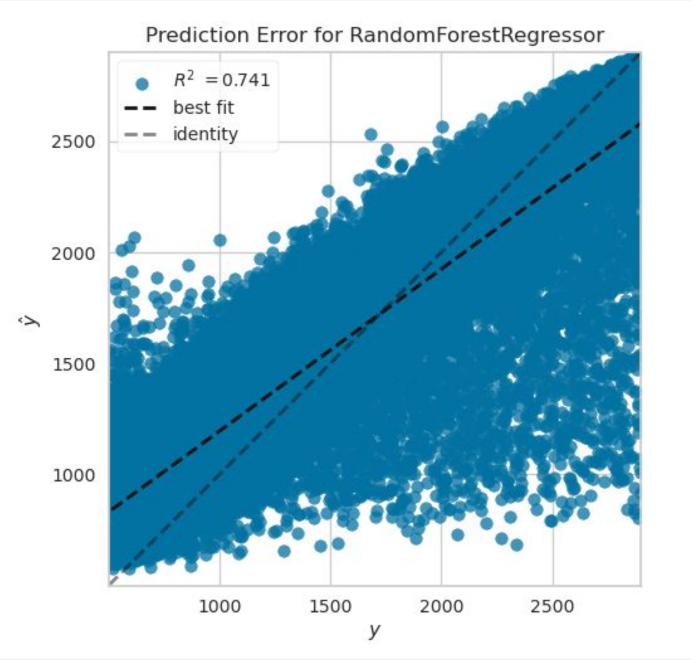
Decision Tree 2

Decision Tree n

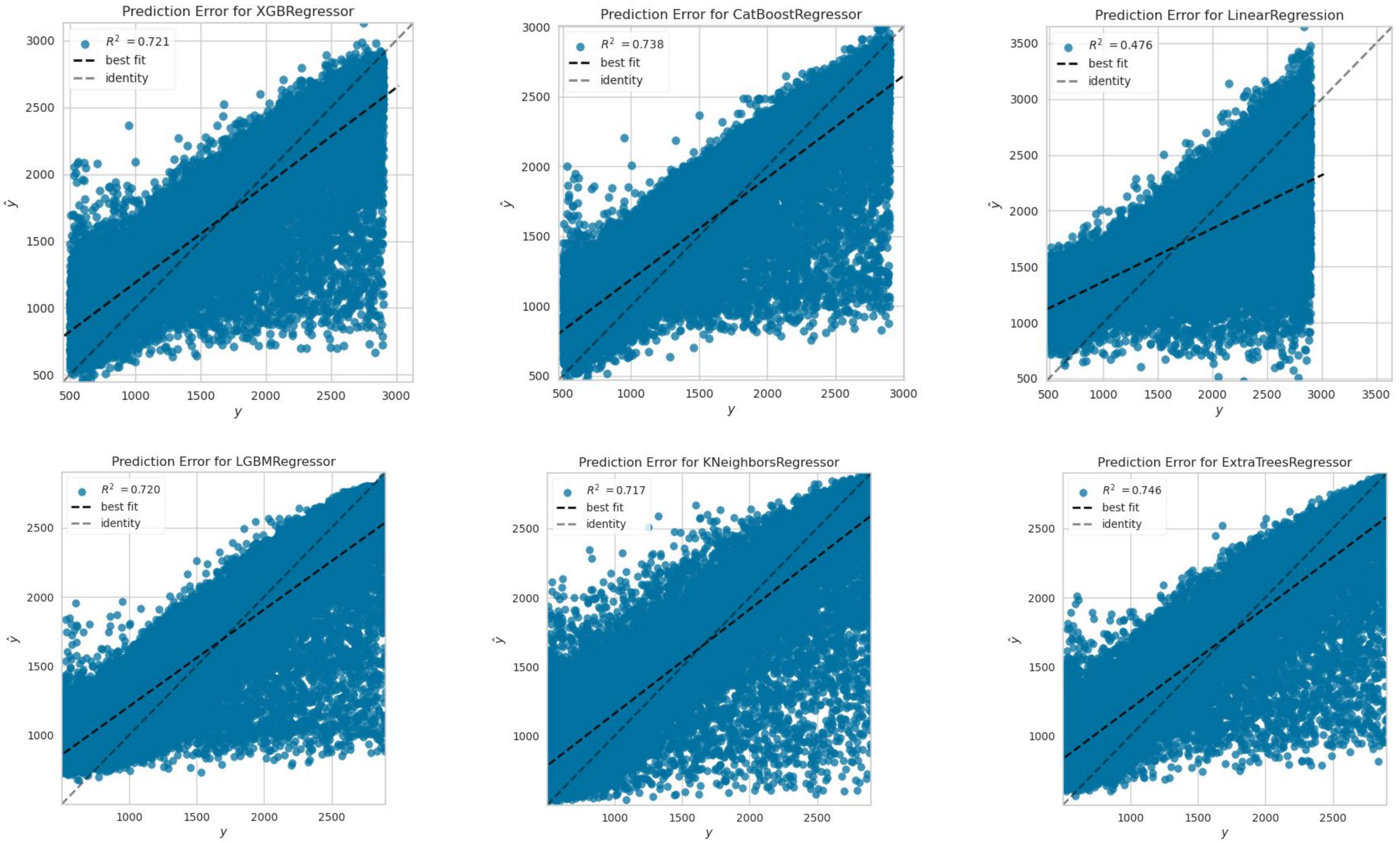
Voting (averaging)

Prediction

Trained a Random forest using the generated data from the models and to predict the planetary parameters(eg. Temperature) from the observed spectra of a planet.

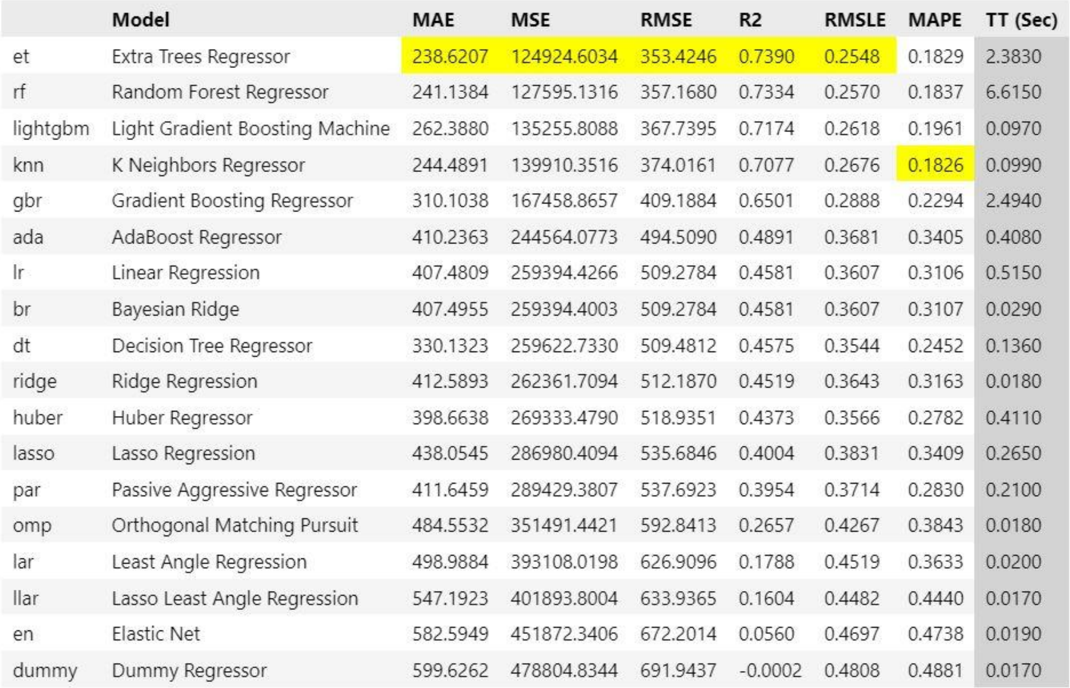


True (y) versus Predicted (ŷ) values of the temperature by different Regression models



PyCaret

The Pycaret package was used to find the best algorithm for a regression problem and it was determined that the most suitable algorithms for the given dataset are Extra Trees Regressor and Random Forest Regressor, this suggests that the data has complex relationships and the chosen algorithms are capable of handling such complexity



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	238.6207	124924.6034	353.4246	0.7390	0.2548	0.1829	2.3830
rf	Random Forest Regressor	241.1384	127595.1316	357.1680	0.7334	0.2570	0.1837	6.6150
lightgbm	Light Gradient Boosting Machine	262.3880	135255.8088	367.7395	0.7174	0.2618	0.1961	0.0970
knn	K Neighbors Regressor	244.4891	139910.3516	374.0161	0.7077	0.2676	0.1826	0.0990
gbr	Gradient Boosting Regressor	310.1038	167458.8657	409.1884	0.6501	0.2888	0.2294	2.4940
ada	AdaBoost Regressor	410.2363	244564.0773	494.5090	0.4891	0.3681	0.3405	0.4080
lr	Linear Regression	407.4809	259394.4266	509.2784	0.4581	0.3607	0.3106	0.5150
br	Bayesian Ridge	407.4955	259394.4003	509.2784	0.4581	0.3607	0.3107	0.0290
dt	Decision Tree Regressor	330.1323	259622.7330	509.4812	0.4575	0.3544	0.2452	0.1360
ridge	Ridge Regression	412.5893	262361.7094	512.1870	0.4519	0.3643	0.3163	0.0180
huber	Huber Regressor	398.6638	269333.4790	518.9351	0.4373	0.3566	0.2782	0.4110
lasso	Lasso Regression	438.0545	286980.4094	535.6846	0.4004	0.3831	0.3409	0.2650
par	Passive Aggressive Regressor	411.6459	289429.3807	537.6923	0.3954	0.3714	0.2830	0.2100
omp	Orthogonal Matching Pursuit	484.5532	351491.4421	592.8413	0.2657	0.4267	0.3843	0.0180
lar	Least Angle Regression	498.9884	393108.0198	626.9096	0.1788	0.4519	0.3633	0.0200
llar	Lasso Least Angle Regression	547.1923	401893.8004	633.9365	0.1604	0.4482	0.4440	0.0170
en	Elastic Net	582.5949	451872.3406	672.2014	0.0560	0.4697	0.4738	0.0190
dummy	Dummy Regressor	599.6262	478804.8344	691.9437	-0.0002	0.4808	0.4881	0.0170

Future Plans

- We want to use the Dataset from the Ariel ML Data Challenge which is generated with Alfnoor, which combines the open source TauREx 3 atmospheric modelling suite with the official Ariel instrument simulator ArielRad to produce large-scale simulations of atmospheres.
- Use Extra trees regressor because its faster,less compute heavy and best suites the type of dataset we are using. Try to explore its performance on wider range of exoplanetary atmospheres and try using it on real observational data.
- Explore the possibility to applying neural networks if time permits.
- Developing a more comprehensive and flexible framework for exoplanetary atmospheric parameter retrieval using machine learning techniques.

References

- Márquez-Neila, Pablo et al. “Supervised machine learning for analysing spectra of exoplanetary atmospheres.” *Nature Astronomy* 2 (2018): 719-724.

- Nixon, Matthew C. and Nikku Madhusudhan. “Assessment of supervised machine learning for atmospheric retrieval of exoplanets.” *Monthly Notices of the Royal Astronomical Society* 496 (2020): 269-281.

- Munsaket, Patcharawee; Awiphan, Supachai; Chainakun, Poemwai; Kerins, Eamonn "Retrieving exoplanet atmospheric parameters using random forest regression. " Journal of Physics: Conference Series, Volume 2145, Siam Physics Congress 2021 (SPC 2021) 24-25 May 2021 Thailand

- pycaret.org. PyCaret, April 2020. URL https://pycaret.org/about. PyCaret version 1.0.0.