Butool Abidi
INFO6105
AutoMLAssignment-1

1. **Is the relationship significant?**

In this project, predictor variables are stored in X dataframe and response variable is stored in Y dataframe. Using Automatic Machine Learning H2O AutoML to make prediction on life expectancy of a certain population. Various plots like Variable Importance Plot, Partial Dependence Plot, Individual Conditional Expectation Plot and SHAP Summary Plot is explained below to explain how each of our feature input affects our model prediction.

Predictor Variables (x)

['Country', 'Year', 'Status', 'Adult Mortality', 'infant deaths', 'Alcohol', 'percentage expenditure', 'Hepatitis B', 'Measles ', 'BMI ', 'under-five deaths ', 'Polio', 'Total expenditure', 'Diphtheria ', 'HIV/AIDS', 'GDP', 'Population', 'thinness 1-19 years', 'thinness 5-9 years', 'Income composition of resources', 'Schooling']

Response Variable (y)

Life_Expectancy

As per trained data H2oAutoMl generated 10 models and they were ranked by mean_residual_deviance, rmse, mse, mae, rmsle, and further generates two reports on train data and on validation data

Leaderboard

Leaderboard shows models with their metrics. When provided with H2OAutoML object, the leaderboard shows 5-fold cross-validated metrics by default (depending on the H2OAutoML settings), otherwise it shows metrics computed on the frame. At most 20 models are shown by default.

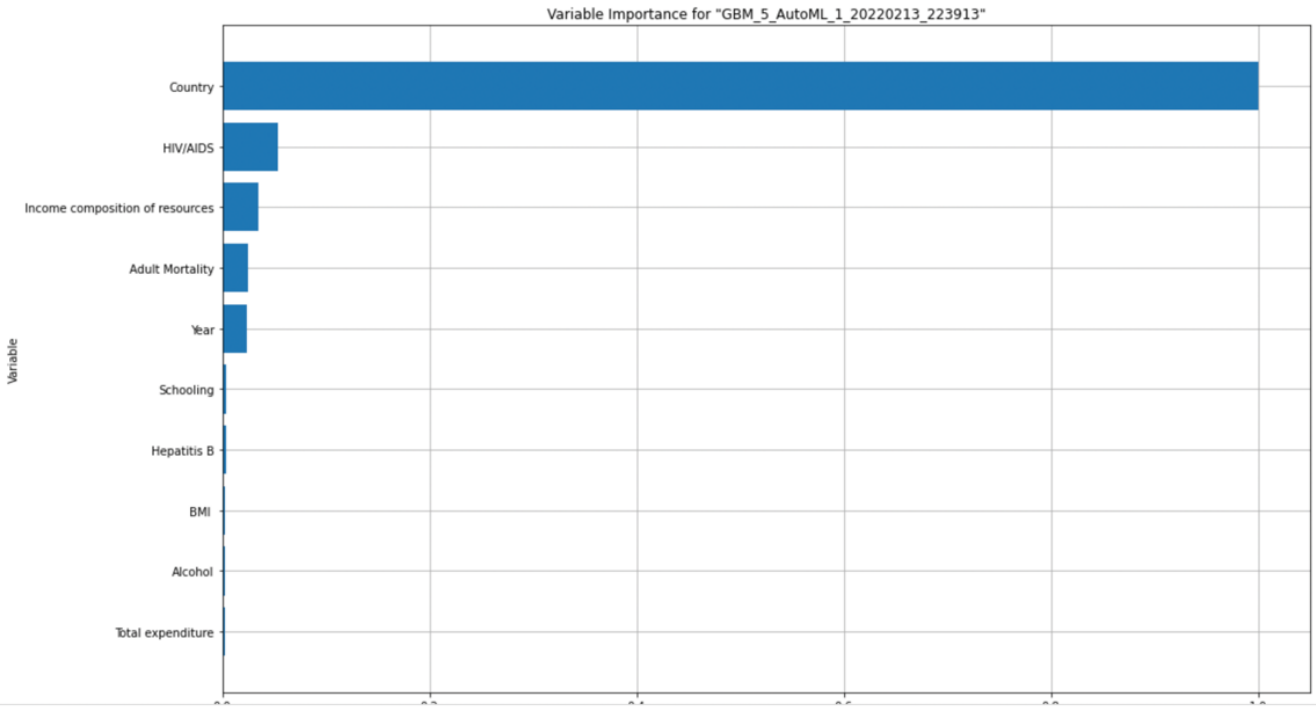
model_id	mean_residual_deviance	rmse	mse	mae	rmsle	training_time_ms	predict_time_per_row_ms	algo
GBM_5_AutoML_1_20220213_223913	1.92353	1.38691	1.92353	0.797835	0.0203709	2041	0.160268	GBM
GBM_2_AutoML_1_20220213_223913	2.05535	1.43365	2.05535	0.82179	0.0215672	3656	0.168384	GBM
GBM_3_AutoML_1_20220213_223913	2.09484	1.44736	2.09484	0.826807	0.0219908	2788	0.165629	GBM
GBM_4_AutoML_1_20220213_223913	2.17086	1.47338	2.17086	0.81388	0.0224966	1977	0.141341	GBM
DRF_1_AutoML_1_20220213_223913	2.71006	1.64623	2.71006	0.946883	0.0260461	12214	0.114997	DRF
XRT_1_AutoML_1_20220213_223913	2.9751	1.72485	2.9751	1.07094	0.0274396	6795	0.049092	DRF
GBM_1_AutoML_1_20220213_223913	3.02183	1.73834	3.02183	0.932133	0.0273351	4747	0.055951	GBM
GBM_grid_1_AutoML_1_20220213_223913_model_1	3.46324	1.86098	3.46324	1.1337	0.0297955	1418	0.077823	GBM
DeepLearning_1_AutoML_1_20220213_223913	11.3358	3.36687	11.3358	2.49305	0.0537636	1244	0.022673	DeepLearning
GLM_1_AutoML_1_20220213_223913	94.7513	9.73403	94.7513	8.02263	0.149389	904	0.072447	GLM

So, for the mentioned dataset top model was Gradient Boosting Machine whose rmse, mse, mae, rmsle, deviance were calculated in the training set. The smaller the values of errors the better the model fits.

Below show a variable importance plot, it predicted the most important variables and these variable were the main player of the model. It predicted top three variables for GBM model were Country, HIV/AIDS and Income Composition of resource.

Variable Importance

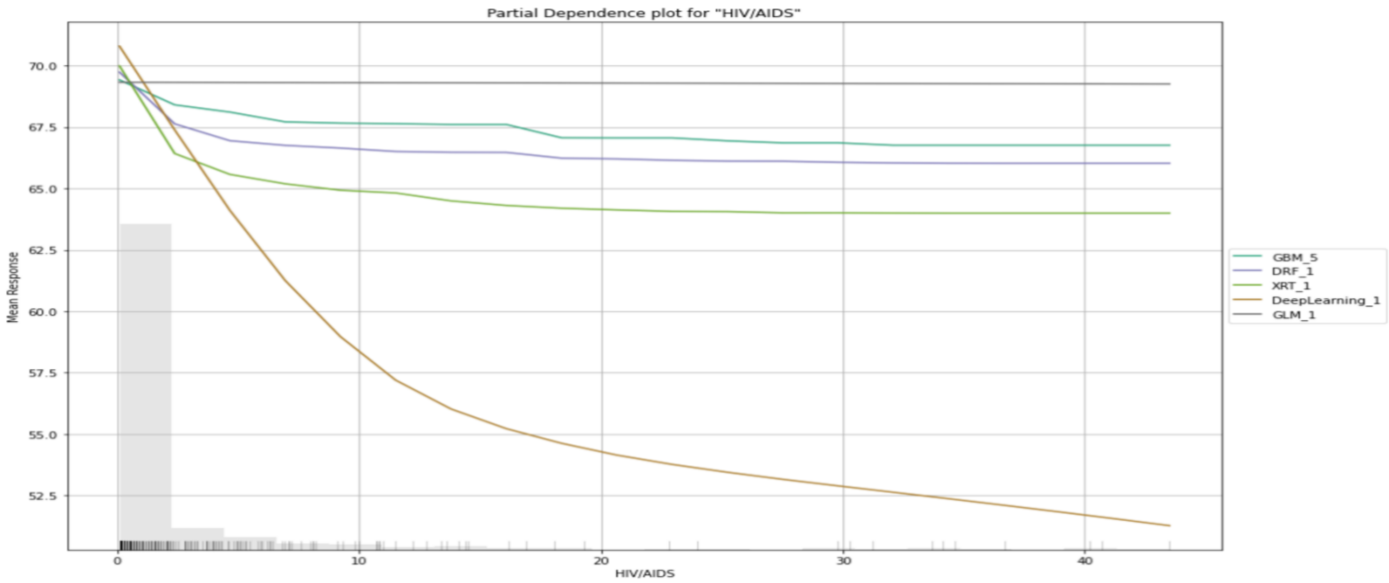
The variable importance plot shows the relative importance of the most important variables in the model.



✓ 0s completed at 5:41 PM

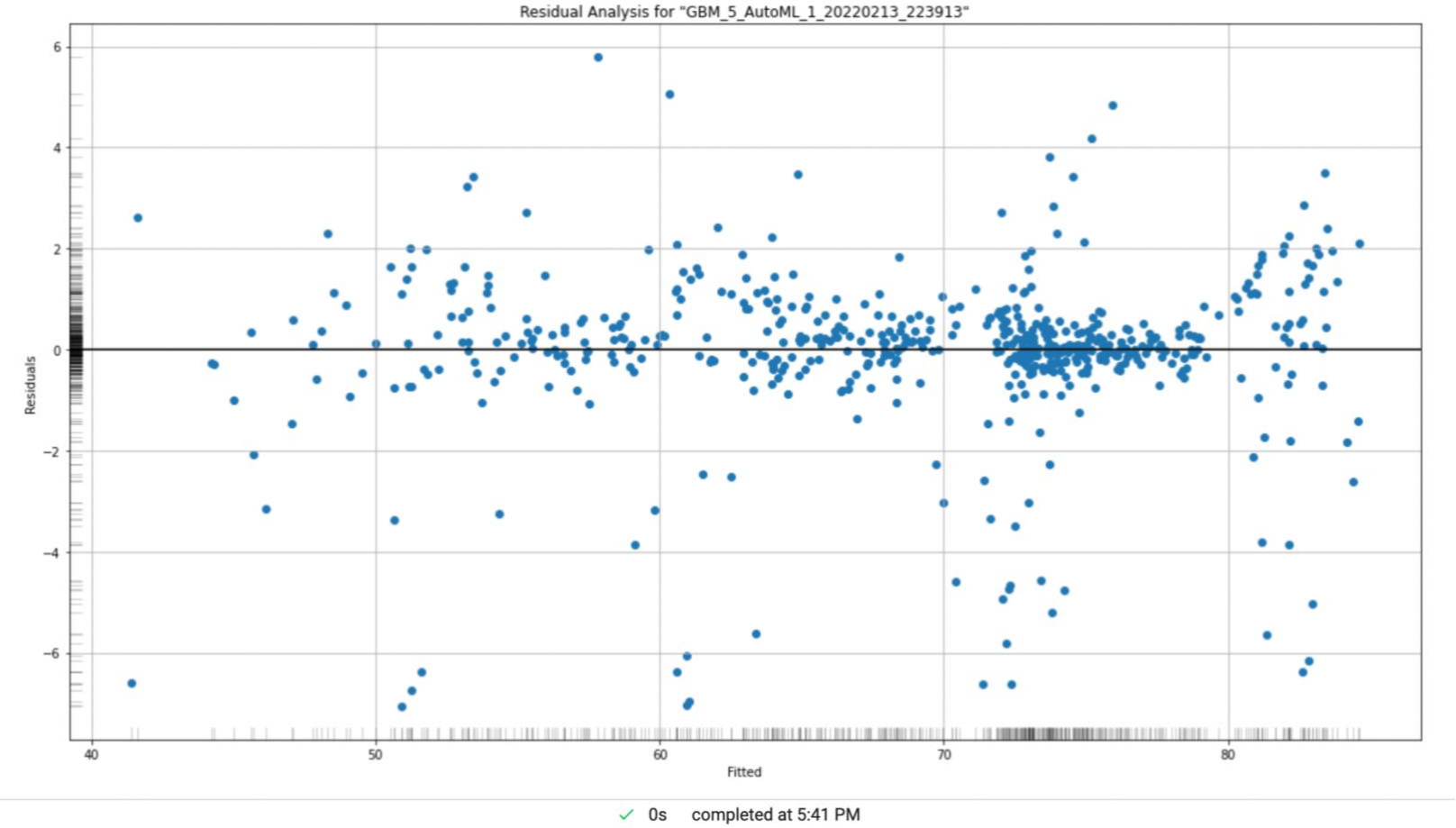
2. Are any model assumptions violated?

In this regression analysis, there is a linear relationship between the response and a predictor. These assumptions are essentially conditioned that should be met before illustration and conclusions or inferences to the model estimate. To show whether there is a linear or curvilinear relationship a partial dependence plot is show below. GLM_1 did not perform as well as other models. Although



it captures the correlation between a feature and the model's outcome, it has low predictive power and failed to capture the non-linear dynamic in these features.

Residual Analysis plots the fitted values vs residuals on a test dataset. Ideally, residuals should be randomly distributed. Patterns in this plot can indicate potential problems with the model selection, e.g., using simpler model than necessary, not accounting for heteroscedasticity, autocorrelation, etc. Note that if you see "striped" lines of residuals, that is an artifact of having an integer valued (vs a real valued) response variable.
The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual.
we see that our data points are randomly dispersed around the horizontal axis, meaning the residuals are consistent with random error, therefore our choice of a linear regression model is appropriate for the data.



3. Is there any multicollinearity in the model?

<u>There are 4 pairs of highly correlated variables:</u>

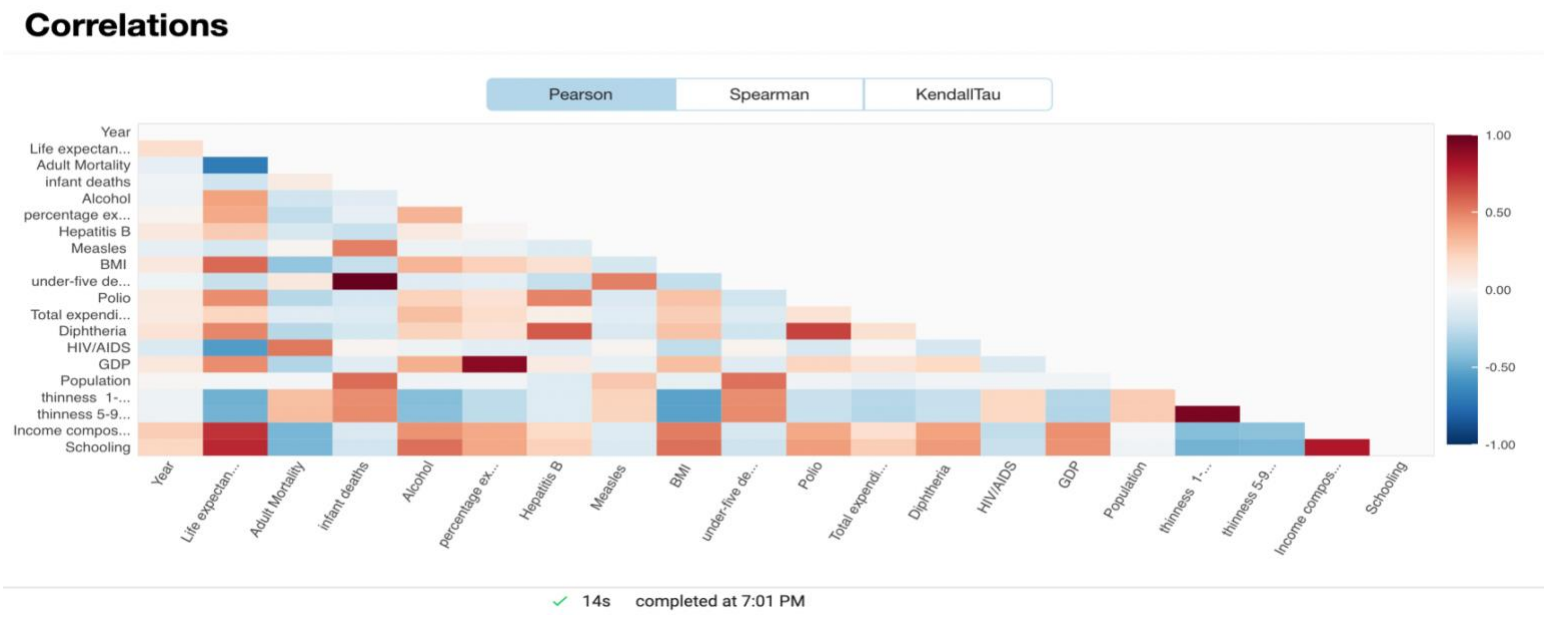
- infant_deaths and under-five_deaths
- percentage_expenditure and gdp
- thinness_1_to_19_years and thinness_5_to_9_years
- income_composition_of_resources and schooling

<u>Negative correlations with life_expectancy:</u>

- status_Developing
- adult_mortality
- infant_deaths
- under-five_deaths
- thinness_1_to_19_years
- thinness_5_to_9_years
- hiv/aids
- measle

<u>Positive correlations with life_expectancy:</u>

- GDP
- Total_expenditure
- percentage_expenditure
- income_composition_of_resources
- status_Developed
- schooling
- bmi, alcohol
- polio vaccine
- diphtheria vaccine
- hepatitis_b vaccine



4. In the multivariate models are predictor variables independent of all the other predictor variables?

No, in multivariate models there are predictor variables which are related to other predictor variables. These correlations can be negative as well as positive.

Examples:

a. percentage_expenditure and gdp
b. thinness_1_to_19_years and thinness_5_to_9_years
c. income_composition_of_resources and schooling

5. In in multivariate models rank the most significant predictor variables and exclude insignificant ones from the model.

In the table shown below illustrates an fine comparison between the importance of predictor variables in the models.

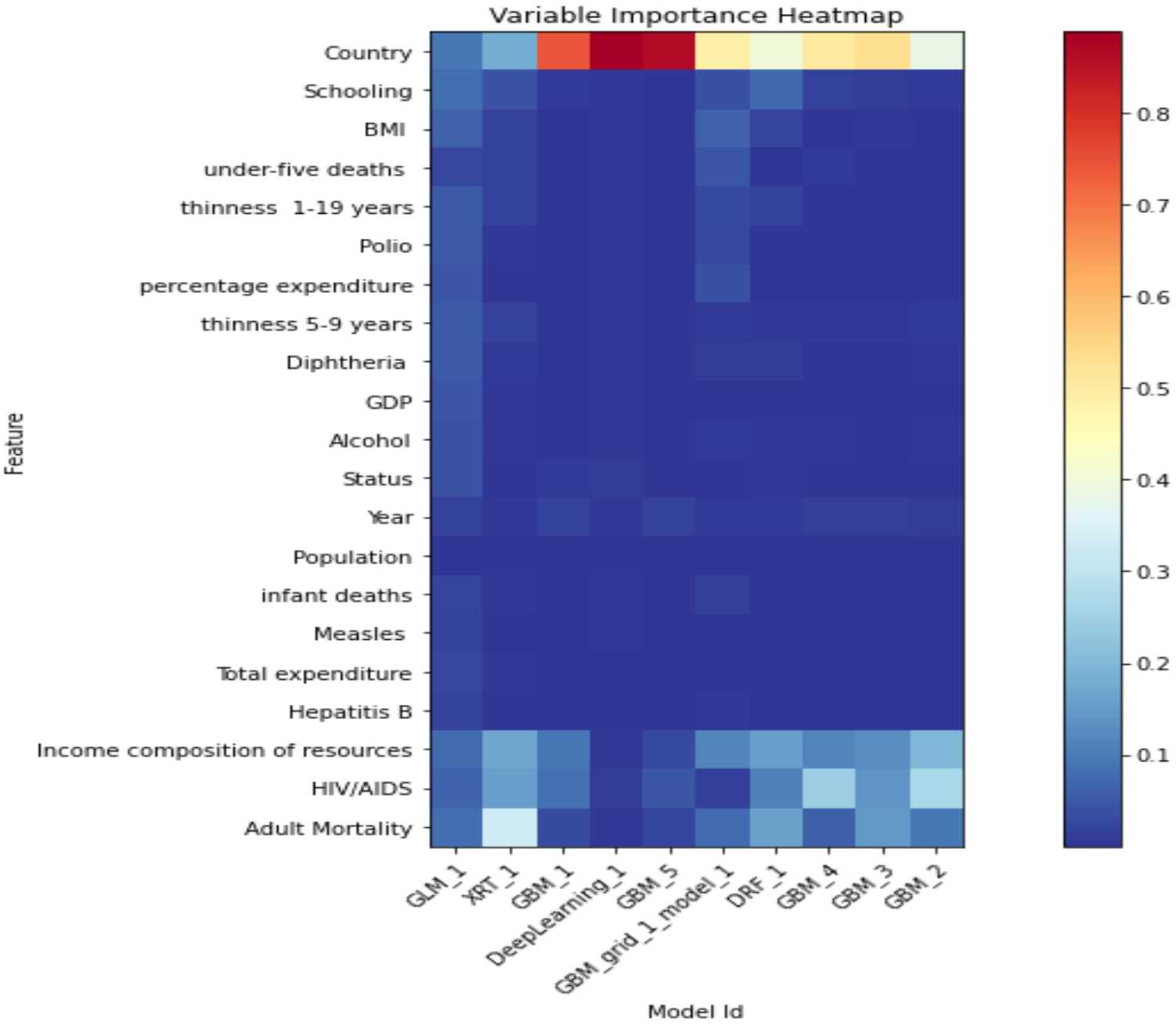
COUNTRY being the highest essential predictor variable followed by HIV/AIDS and then Income composition of resources in the first ranked model ie GBM_5

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
0	Country	825290.375000	1.000000	0.861134
1	HIV/AIDS	44024.191406	0.053344	0.045936
2	Income composition of resources	28707.910156	0.034785	0.029955
3	Adult Mortality	20579.714844	0.024936	0.021474
4	Year	19582.304688	0.023728	0.020433
5	Schooling	2987.936523	0.003620	0.003118
6	Hepatitis B	2800.889160	0.003394	0.002923
7	BMI	2119.259277	0.002568	0.002211
8	Alcohol	1982.320679	0.002402	0.002068
9	Total expenditure	1710.999512	0.002073	0.001785
10	thinness 5-9 years	1402.051025	0.001699	0.001463
11	thinness 1-19 years	1400.400024	0.001697	0.001461
12	GDP	1316.959106	0.001596	0.001374
13	percentage expenditure	870.778809	0.001055	0.000909
14	under-five deaths	861.030762	0.001043	0.000898
15	Population	764.198792	0.000926	0.000797
16	Diphtheria	748.970154	0.000908	0.000781
17	Polio	459.714264	0.000557	0.000480
18	infant deaths	426.068848	0.000516	0.000445
19	Measles	319.021088	0.000387	0.000333

✓ 14s completed at 7:01 PM

In the heat map shown is the comparison of highest ranked predictor variable amongst all the 10 models. COUNTRY being the most important predictor variable rates the highest in 8 out of 10 models. Following that HIV/AIDS is most essential component in 3 out of 10 models ie GBM_2, GBM_3, GBM_4 and similarly for income composition of resources.



6. Does the model make sense?

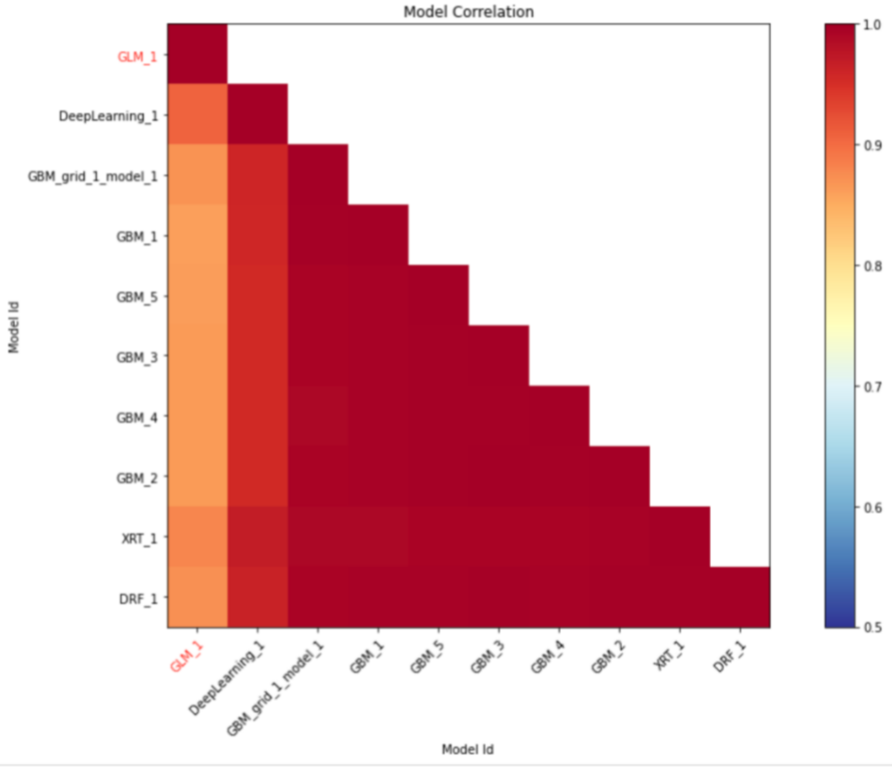
The GBM model implements boosting to yield accurate models. GBM sequentially builds regression trees on all the features of the dataset in a fully distributed way — each tree is built in parallel. All the prediction made by the model is close to accurate information. To predict the life expectancy of different populations and their own hierarchy of features results show how populations will age despite of social-economic issues or medical conditions. The model makes complete prediction with various input effects.

7. Does regularization help?

Yes, regularization does helps as it formulates over control of model being over-fitting. R^2 is the coefficient of determination. The higher the value, the higher quality the model is i.e., how the model generalizes to new data (validation set): is it overfitting or underfitting.
The plot shown below shows correlation between the models. It depicts that all the models are highly correlated with each other and produce similar predictions.

XGBoost (Extreme Gradient Boosting**)** implements boosting to yield accurate models. Boosting is an ensemble learning technique of building many models sequentially, with each new model attempting to correct for the deficiencies in the previous model. Compared to GBM, XGboost uses a more regularized model (DART) formalization to control over-fitting.

This plot shows the correlation between the predictions of the models. For classification, frequency of identical predictions is used. By default, models are ordered by their similarity (as computed by hierarchical clustering). Interpretable models, such as GAM, GLM, and RuleFit are highlighted using red colored text.



0s completed at 9:25 PM

8. Which independent variables are significant?

'Life_Expectancy' is the independent variable and it is significant as the more it increases it will depict lifestyle improvement. It is the key to assess population health. It draws the overall health and status of the community. When. Population is high it means that death of young children and infants is less along with elderly people are having long lives. Hence population is directly correlated with Life_Expectancy.

9. Which hyperparameters are important?

Hyperparameters are important as they directly showcase the behaviors of the training model. It overcasts the entire significance and performance of the designed model. Tuning of hyperparameters are essential for managing large datasets. Hyperparameter like country is the most important and significant as it showcases the trails and methods practiced in various countries resulting a significant note of life expectancy.

10. Coding professionalism?