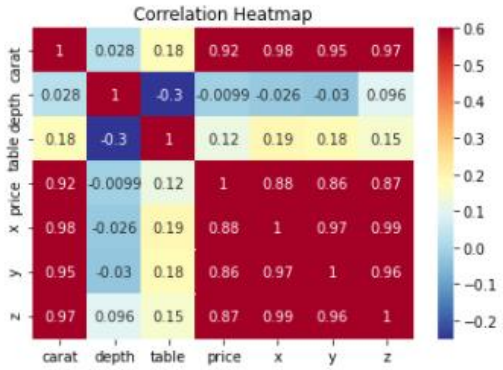
Name: Betty (Zihui) Qin

Subject: Data Science Project

1. Feature Engineering

1.1 Reduce Multicollinearity

Below is a correlation heatmap from all columns from the training set:



As shown above, there is high correlation between caret and x,y,z columns. Therefore, I decided to remove columns x,y,z.

1.2 Ordinal Encoding

Then based on the industry knowledge, Cut column is encoded with 'Very Good' as 3, 'Premium' as 5, 'Good' as 2, 'Fair' as 0, 'Ideal' as 4 ; Color column is encoded with 'E' as 5, 'H' as 2, 'D' as 6 , 'F' as 4, 'G' as 3, 'I' as 1, 'J' as 0; clarity columns is encoded with 'SI2': as 1, 'VS2' as 3, 'SI1' as 2, 'VVS1' as 6, 'VS1' as 4, 'VVS2' as 5, 'IF' as 7, 'I1' as 0.

2. Model Selection

I used the H2o AutoML package, which chooses the best model in the 20 models, it compared. The best model is a stacked ensembled model. The model metric comparison as below:

model_id	rmse	mse	mae	rmsle	mean_residual_deviance
StackedEnsemble_AllModels_1_AutoML_1_20221104_140150	540.052	291657	277.595	0.104538	291657
StackedEnsemble_BestOfFamily_1_AutoML_1_20221104_140150	542.663	294484	280.898	0.108618	294484
GBM_5_AutoML_1_20221104_140150	550.024	302527	289.677	nan	302527
GBM_3_AutoML_1_20221104_140150	557.798	311138	294.659	0.119565	311138
GBM_grid_1_AutoML_1_20221104_140150_model_1	560.648	314326	285.281	0.103772	314326
GBM_2_AutoML_1_20221104_140150	560.885	314593	305.721	nan	314593
GBM_1_AutoML_1_20221104_140150	567.022	321514	306.141	nan	321514
GBM_grid_1_AutoML_1_20221104_140150_model_4	569.745	324610	307.876	0.123899	324610
DRF_1_AutoML_1_20221104_140150	588.934	346843	306.114	0.12147	346843
GBM_4_AutoML_1_20221104_140150	592.38	350914	336.369	nan	350914
XRT_1_AutoML_1_20221104_140150	608.88	370735	317.616	0.128205	370735
DeepLearning_1_AutoML_1_20221104_140150	661.134	437098	373.225	0.164818	437098
GBM_grid_1_AutoML_1_20221104_140150_model_2	674.453	454887	382.729	nan	454887
GBM_grid_1_AutoML_1_20221104_140150_model_3	684.686	468795	382.809	nan	468795
DeepLearning_grid_1_AutoML_1_20221104_140150_model_2	718.863	516764	445.157	nan	516764
GBM_grid_1_AutoML_1_20221104_140150_model_5	767.935	589724	431.936	nan	589724
DeepLearning_grid_2_AutoML_1_20221104_140150_model_2	819.873	672192	565.753	0.309035	672192
DeepLearning_grid_1_AutoML_1_20221104_140150_model_1	862.364	743672	522.152	nan	743672
DeepLearning_grid_3_AutoML_1_20221104_140150_model_2	901.395	812513	616.413	0.310658	812513
DeepLearning_grid_2_AutoML_1_20221104_140150_model_1	943.649	890473	610.511	0.288323	890473
DeepLearning_grid_3_AutoML_1_20221104_140150_model_1	1115.62	1.24462e+06	807.001	0.438338	1.24462e+06
GLM_1_AutoML_1_20221104_140150	4006.36	1.60509e+07	3049.64	1.13076	1.60509e+07

[22 rows x 6 columns]