Model Fitting

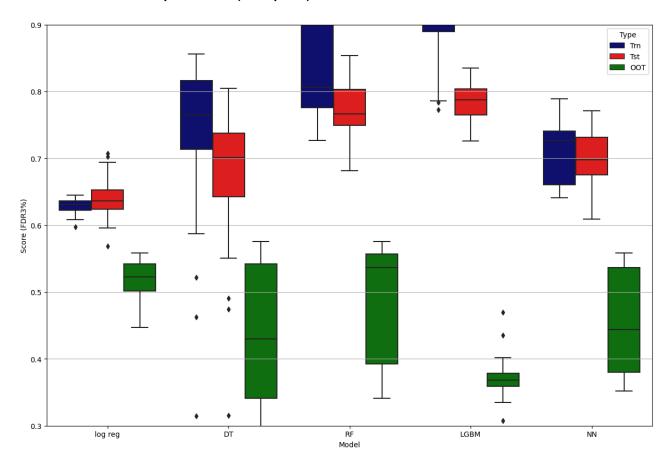
Sheetal Srivastava

A. Model Exploration (hyperparameter tuning):

Model		Parameters							Average FDR at 3%		
Logistic	Iteration	penalty		С	solver		I1_ratio		Train	Test	ООТ
	1 (default)	12		1	lbfgs		None		0.6308	0.6274	0.5335
	2	12		0.5	Ibfgs		None		0.6248	0.6374	0.5078
Regression	3	elasticnet		1	saga		0.5		0.6261	0.6528	0.5128
	4	elasticnet		0.8	saga		0.3		0.6302	0.6455	0.5229
	5	12		0.8	saga		None		0.6317	0.6481	0.5162
Decision Tree	Iteration	criterion	splitter	max_depth	min_samples_leaf min_samples_split		les_split	Train	Test	OOT	
	1 (default)	gini	best	None	1		2		1	0.5947	0.3016
	2	gini	best	None	40		20		0.8454	0.7486	0.4682
	3	gini	best	100	80		40		0.7998	0.7382	0.5223
	8	gini	best	None	100		50		0.7725	0.7329	0.5575
	4	gini	best	200	150		60		0.7354	0.7156	0.5564
	5	entropy	best	None	1		2		1	0.5768	0.2899
	6	entropy	best	None	100		50		0.7683	0.7208	0.4273
	7	entropy	best	200	100		50		0.7686	0.7195	0.4055
Random Forest	Iteration	n_estimators	criterion	max_depth	min_samples_leaf min_samples_split		Train	Test	ООТ		
	1 (default)	100	gini	None		1		2		0.8194	0.4357
	2	100	entropy	None	1		2		1	0.8123	0.3754
						50		20		0.0120	
	3	10	gini	None	5	50	20		0.8333	0.7811	0.5145
	3 4	10 50	gini gini	None None		00 00	20 50		0.8333 0.8043		
					1					0.7811	0.5145
	4	50	gini	None	1	00	50		0.8043	0.7811 0.7661	0.5145 0.5447
	4 5	50 100	gini entropy gini	None None	1	00	50 50))	0.8043 0.7962	0.7811 0.7661 0.7661	0.5145 0.5447 0.5547
	4 5 6	50 100 20	gini entropy gini	None None None	1 1 2	00 00 00	50 50 100	o nators	0.8043 0.7962 0.7596	0.7811 0.7661 0.7661 0.7444	0.5145 0.5447 0.5547 0.4899
Forest	4 5 6 Iteration	50 100 20 boosting_t	gini entropy gini	None None None num_leaves	1 1 2 max_depth	00 00 00 learning_rate	50 50 100 n_estim	o Diators	0.8043 0.7962 0.7596 Train	0.7811 0.7661 0.7661 0.7444 Test	0.5145 0.5447 0.5547 0.4899 OOT
	4 5 6 Iteration 1 (default)	50 100 20 boosting_t gbdt	gini entropy gini	None None None num_leaves 31	1 2 max_depth None (-1)	00 00 00 learning_rate 0.1	50 50 100 n_estim	nators	0.8043 0.7962 0.7596 Train 0.984	0.7811 0.7661 0.7661 0.7444 Test 0.7953	0.5145 0.5447 0.5547 0.4899 OOT 0.3441
Forest	4 5 6 Iteration 1 (default) 2	50 100 20 boosting_t gbdt gbdt	gini entropy gini	None None num_leaves 31 20	1 1 2 max_depth None (-1) None (-1)	00 00 00 learning_rate 0.1 0.01	50 50 100 n_estim 100 50	nators	0.8043 0.7962 0.7596 Train 0.984 0.7923	0.7811 0.7661 0.7661 0.7444 Test 0.7953 0.7443	0.5145 0.5447 0.5547 0.4899 OOT 0.3441 0.3731
Forest	4 5 6 Iteration 1 (default) 2 3	50 100 20 boosting_t gbdt gbdt gbdt	gini entropy gini	None None num_leaves 31 20 30	1 1 2 max_depth None (-1) None (-1)	00 00 00 learning_rate 0.1 0.01 0.05	50 50 100 n_estim 100 50	nators	0.8043 0.7962 0.7596 Train 0.984 0.7923 0.9148	0.7811 0.7661 0.7661 0.7444 Test 0.7953 0.7443 0.7939	0.5145 0.5447 0.5547 0.4899 OOT 0.3441 0.3731 0.3665
Forest	4 5 6 Iteration 1 (default) 2 3 4	50 100 20 boosting_t gbdt gbdt gbdt gbdt gbdt	gini entropy gini	None None num_leaves 31 20 30 20	1 1 2 max_depth None (-1) None (-1) 100	000 000 learning_rate 0.1 0.01 0.05 0.08	50 50 100 n_estim 100 50 50	nators	0.8043 0.7962 0.7596 Train 0.984 0.7923 0.9148 0.957	0.7811 0.7661 0.7661 0.7444 Test 0.7953 0.7443 0.7939 0.8008 0.7851	0.5145 0.5447 0.5547 0.4899 OOT 0.3441 0.3731 0.3665 0.3698
Forest Light GBM	4 5 6 Iteration 1 (default) 2 3 4 6	50 100 20 boosting_t gbdt gbdt gbdt gbdt gbdt dart	gini entropy gini ype	None None None num_leaves 31 20 30 20 20	1 1 2 2 max_depth None (-1) None (-1) 100 100	000 000 000 learning_rate 0.1 0.01 0.05 0.08 0.08	50 50 100 n_estim 100 50 50 100	nators 0	0.8043 0.7962 0.7596 Train 0.984 0.7923 0.9148 0.957 0.8938	0.7811 0.7661 0.7661 0.7444 Test 0.7953 0.7443 0.7939 0.8008 0.7851	0.5145 0.5447 0.5547 0.4899 OOT 0.3441 0.3731 0.3665 0.3698 0.3888
Forest	4 5 6 Iteration 1 (default) 2 3 4 6 Iteration	50 100 20 boosting_t gbdt gbdt gbdt gbdt gbdt dart hidden_layer_sizes	gini entropy gini ype	None None None num_leaves 31 20 30 20 20 alpha	1 1 2 max_depth None (-1) None (-1) 100 100 batch_size	000 000 learning_rate 0.1 0.01 0.05 0.08 0.08 solver	50 50 100 n_estim 100 50 50 100 100 learning_rate	nators)) max_iter	0.8043 0.7962 0.7596 Train 0.984 0.7923 0.9148 0.957 0.8938 Train	0.7811 0.7661 0.7661 0.7444 Test 0.7953 0.7443 0.7939 0.8008 0.7851 Test	0.5145 0.5447 0.5547 0.4899 OOT 0.3441 0.3731 0.3665 0.3698 0.3888 OOT

 $[\]rightarrow$ Random Forest seems to give the best scores on all three datasets - train, test, and Out of time (OOT).

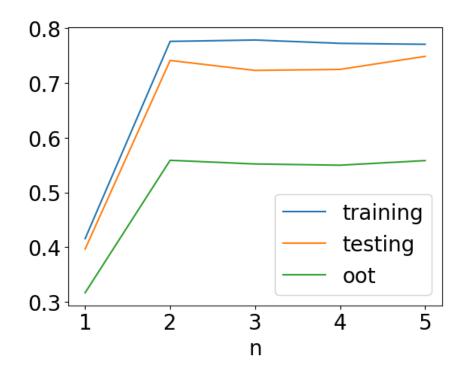
B. Model Comparison (boxplot):



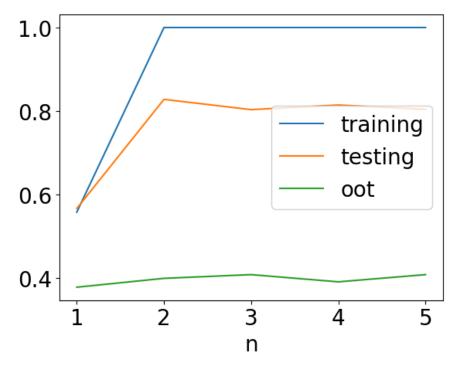
 \rightarrow Once again, Random Forest seems to be the best out of all models, with the average highest scores for all three datasets - Train, Test, and OOT.

C. Performance vs Complexity

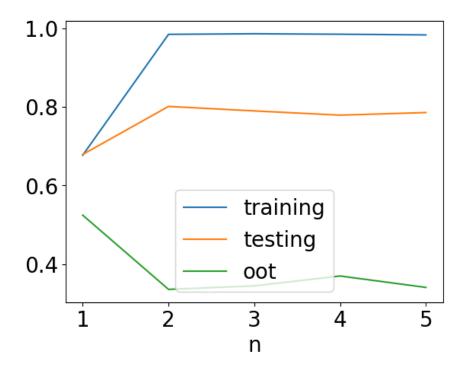
a. Decision Tree



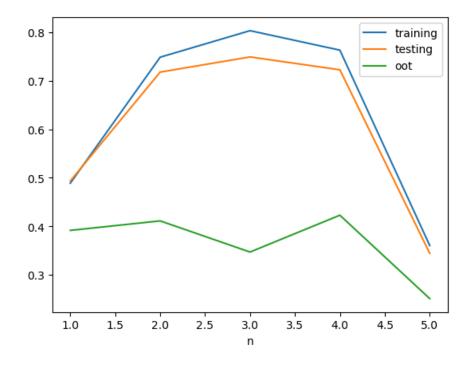
b. Random Forest



c. Light GBM



d. Neural Network



- \rightarrow As we can observe from the plots, the performance of the model on the OOT data starts decreasing as the model becomes more complex and starts overfitting.
- → For the first two models (Decision Tree and Random Forest) we see the performance stagnate after 5 iterations. However if we were to increase the number of iterations (and hence complexity) further we would see a decrease in the performance numbers on the OOT data.
- → For Random Forest and Light GBM we can see that the model starts overfitting in just 2 iterations itself. The performance on training data is already 1, i.e., 100% accuracy.
- \rightarrow In the case of the Neural Network model we can observe the effect of complexity a little better. Increasing the model complexity up to a certain level improves the performance of the model, however beyond a point the model starts overfitting.
- \rightarrow For NN models, we increase the complexity by increasing the number of neurons in the hidden layer. For the remaining 3 models, we increase the complexity by increasing the depth of the tree or the number of leaf nodes.