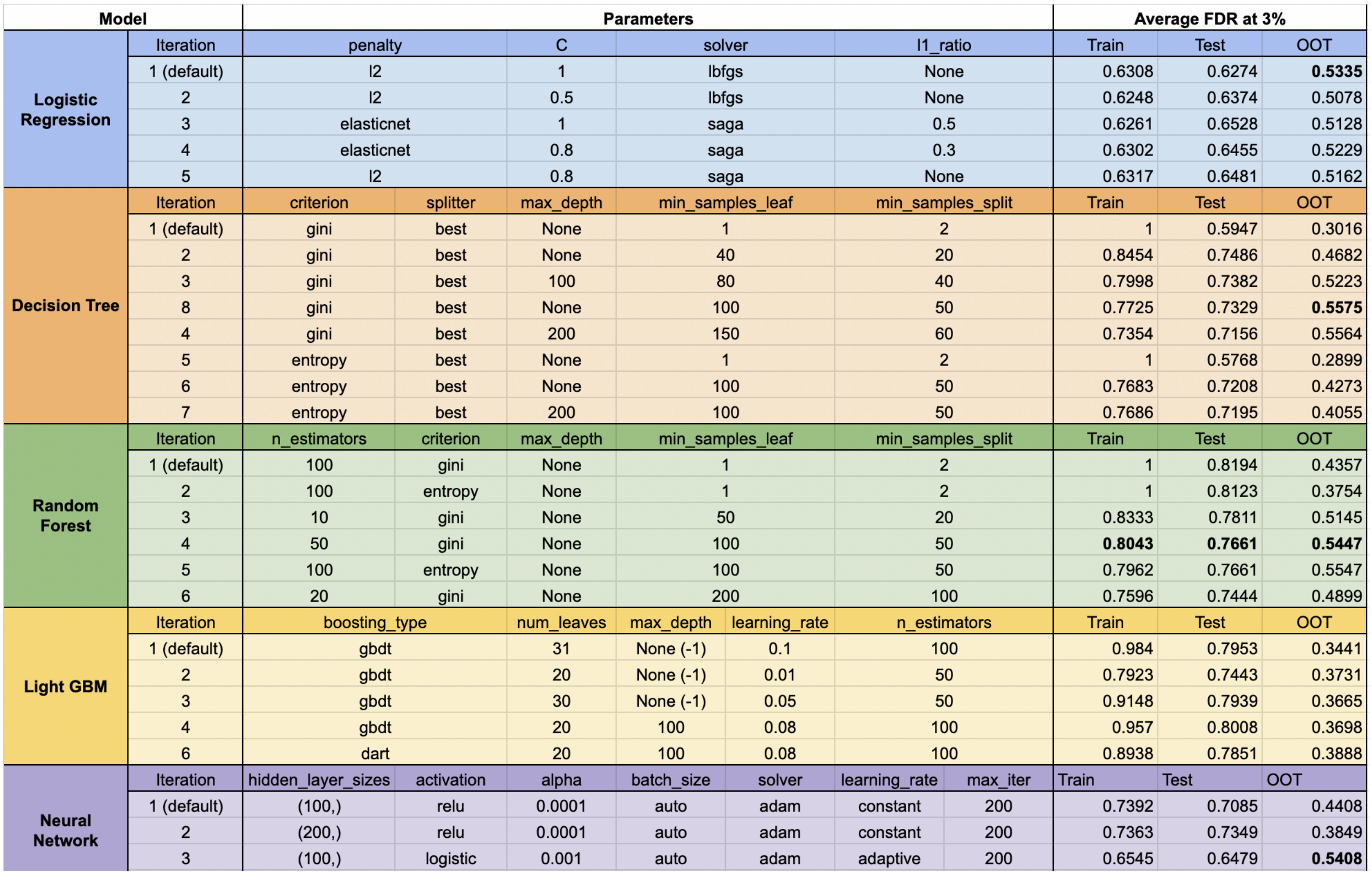
Model Fitting

Sheetal Srivastava

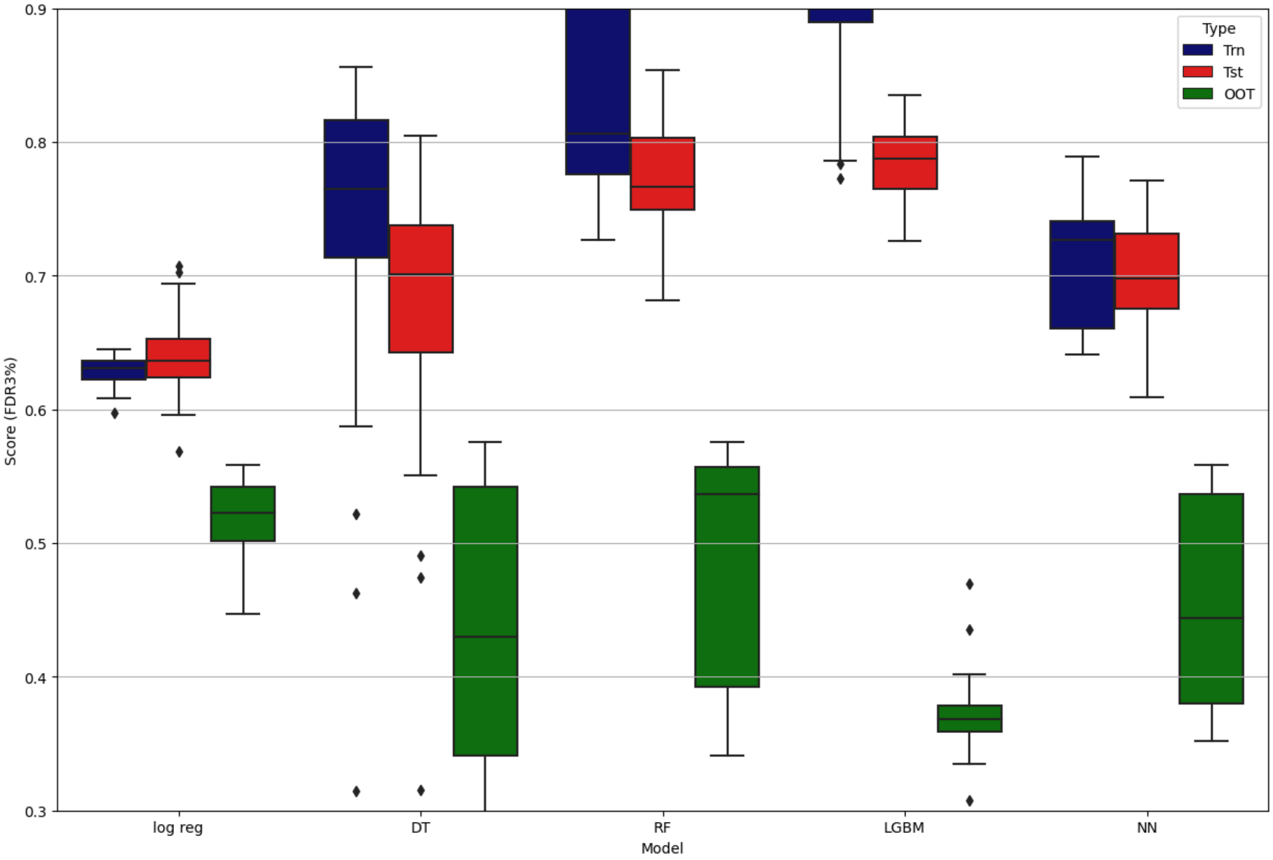
A. Model Exploration (hyperparameter tuning) :



Model		Parameters							Average FDR at 3%		
Logistic Regression	Iteration	penalty		C	solver		l1_ratio		Train	Test	OOT
	1 (default)	l2		1	lbfgs		None		0.6308	0.6274	**0.5335**
	2	l2		0.5	lbfgs		None		0.6248	0.6374	0.5078
	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	l2		0.8	saga		None		0.6317	0.6481	0.5162
Decision Tree	Iteration	criterion	splitter	max_depth	min_samples_leaf		min_samples_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	OOT
	1 (default)	100	gini	None	1		2		1	0.8194	0.4357
	2	100	entropy	None	1		2		1	0.8123	0.3754
	3	10	gini	None	50		20		0.8333	0.7811	0.5145
	4	50	gini	None	100		50		**0.8043**	**0.7661**	**0.5447**
	5	100	entropy	None	100		50		0.7962	0.7661	0.5547
	6	20	gini	None	200		100		0.7596	0.7444	0.4899
Light GBM	Iteration	boosting_type		num_leaves	max_depth	learning_rate	n_estimators		Train	Test	OOT
	1 (default)	gbdt		31	None (-1)	0.1	100		0.984	0.7953	0.3441
	2	gbdt		20	None (-1)	0.01	50		0.7923	0.7443	0.3731
	3	gbdt		30	None (-1)	0.05	50		0.9148	0.7939	0.3665
	4	gbdt		20	100	0.08	100		0.957	0.8008	0.3698
	6	dart		20	100	0.08	100		0.8938	0.7851	0.3888
Neural Network	Iteration	hidden_layer_sizes	activation	alpha	batch_size	solver	learning_rate	max_iter	Train	Test	OOT
	1 (default)	(100,)	relu	0.0001	auto	adam	constant	200	0.7392	0.7085	0.4408
	2	(200,)	relu	0.0001	auto	adam	constant	200	0.7363	0.7349	0.3849
	3	(100,)	logistic	0.001	auto	adam	adaptive	200	0.6545	0.6479	**0.5408**

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

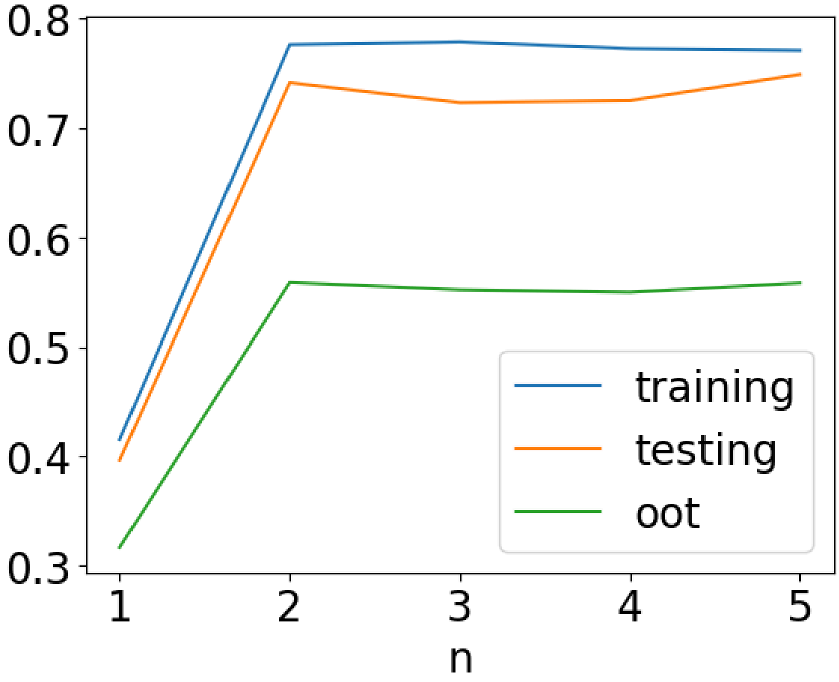
B. Model Comparison (boxplot) :



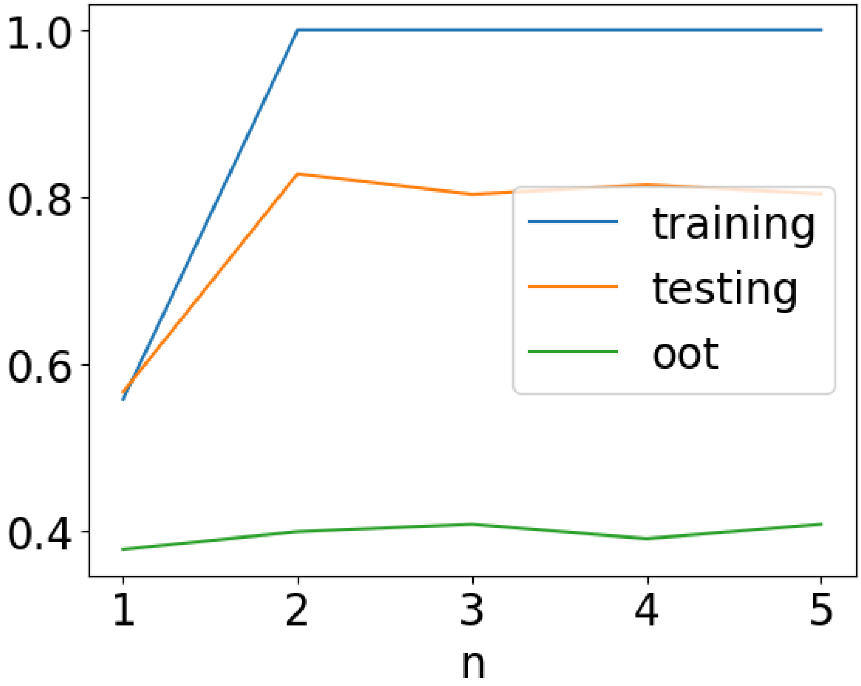
→ Once again, Random Forest seems to be the best out of all models, with the average highest scores for all three datasets - Train, Test, and OOT.

C. Performance vs Complexity

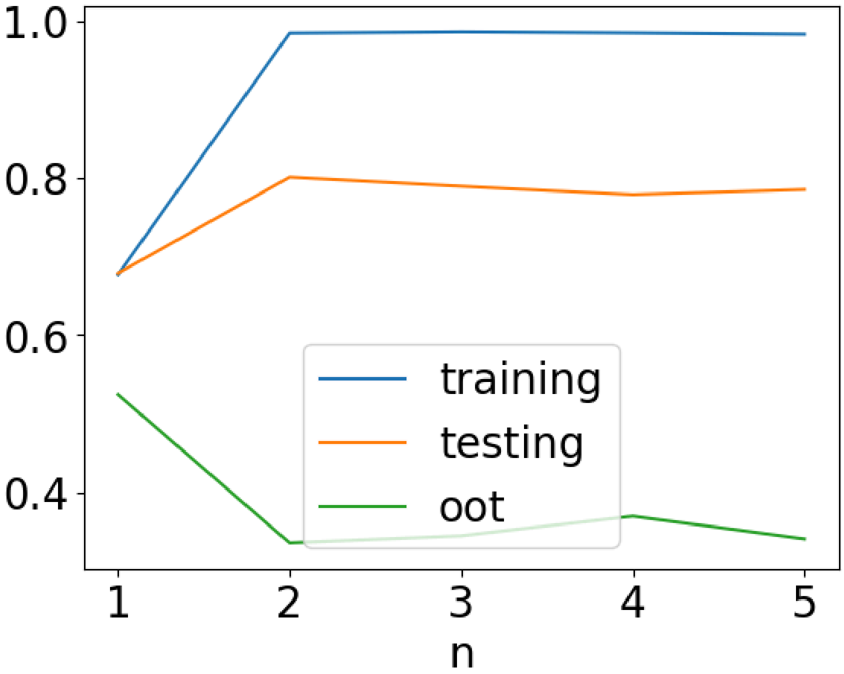
a. Decision Tree



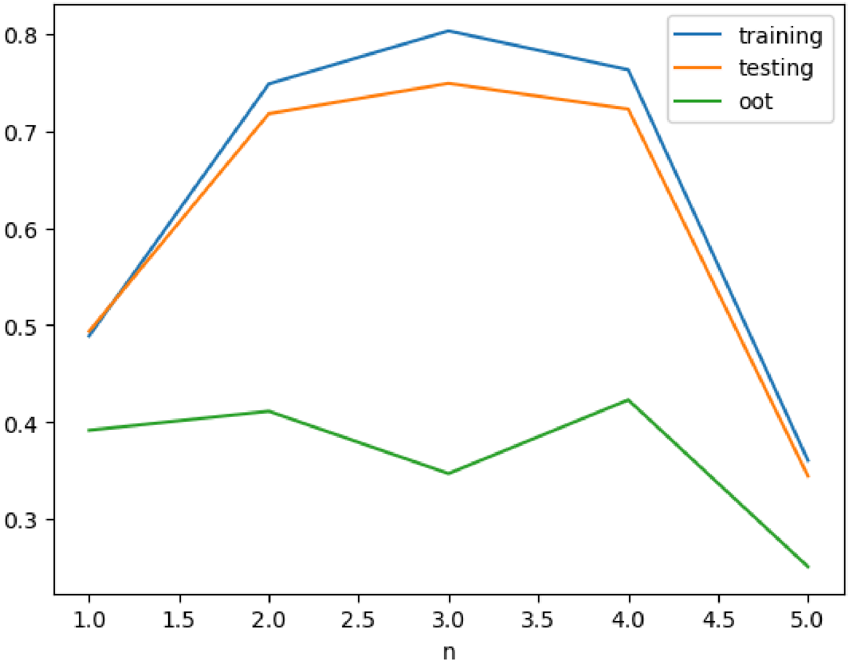
b. Random Forest



c. Light GBM



d. Neural Network



→ 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.
→ 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.
→ 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.