Name: Joyce Xinyi Jiang

Subject: HW 5 - Building Preliminary Models

There are 7 models in total that are explored in this summary report. Different number of variables and parameter values are utilized to explore model performance, especially to observe its effect on over-fitting or under-fitting. The baseline logistic regression yielded a result of 0.48, in regards to keeping FDR at 3%. The at-best result for train and test set are around 0.52-0.53, while certain overfitting examples pushed train sets to 0.54, but nothing beyond 0.55. The OOT sets performance averages around 0.49-0.50.

Model		Dataset	Parameters						Average FDR at 3%			
	Iteration	NVARS	max_iter	penalty	c	solver	l1_ratio		Train	Test	OOT	
Logistic Regression	1	10	20	NA	NA	NA	NA		0.4889	0.4858	0.4730	
	2	10	20	l2	1	lbfgs	None		0.4878	0.4886	0.4733	
	3	15	20	l2	1	saga	None		0.4797	0.4783	0.4653	
	4	10	20	l1	0.5	saga	None		0.4889	0.4795	0.4711	
	5	15	20	elasticnet	0.5	saga	0.4		0.4840	0.4770	0.4679	
	Iteration	NVARS	max_depth		min_samples_split		min_samples_leaf		Train	Test	OOT	
Single Decision Tree	1	15	5		50		30		0.4765	0.4809	0.4530	UNDER-FITTING
	2	10	10		40		24		0.5289	0.5238	0.5041	
	3	10	20		30		14		0.5358	0.5221	0.5017	
	4	15	25		20		8		0.5434	0.5173	0.5008	OVER-FITTING
	5	15	30		5		4		0.5412	0.5221	0.4985	OVER-FITTING
	Iteration	NVARS	max_depth	min_samples_split	min_samples_leaf	max_features	bootstrap	n_estimators	Train	Test	OOT	
Random Forest	1	10	2	50	30	4	TRUE	3	0.4402	0.4365	0.4120	UNDER-FITTING
	2	10	10	40	24	5	TRUE	15	0.5250	0.5219	0.5036	
	3	10	20	30	14	6	TRUE	40	0.5276	0.5256	0.5021	
	4	15	20	20	10	12	TRUE	70	0.5417	0.5211	0.5013	OVER-FITTING
	5	15	30	5	4	15	TRUE	100	0.5431	0.5213	0.5013	OVER-FITTING
	Iteration	NVARS	hidden_layer_size	activation	alpha	learning_rate	solver	learning_rate_init	Train	Test	OOT	
Nueral Net (NN)	1	10	5	logistic	0.1	constant	adam	0.01	0.5023	0.5026	0.4834	
	2	15	5	relu	0.1	adaptive	lbfgs	0.01	0.5232	0.5212	0.5011	
	3	15	20, 20, 20	logistic	0.01	constant	sgd	0.001	0.4785	0.4678	0.4421	UNDER-FITTING
	4	10	20, 20, 20	relu	0.001	adaptive	lbfgs	0.001	0.5282	0.5250	0.5048	
	5	15	10, 10	relu	0.001	constant	lbfgs	0.0001	0.5304	0.5191	0.5063	OVER-FITTING
	Iteration	NVARS	num_leaves			n_estimators			Train	Test	OOT	
LightGBM (Boost)	1	10	2			20			0.509	0.518	0.489	
	2	15	4			100			0.529	0.520	0.504	
	3	10	6			300			0.527	0.527	0.507	
	4	15	8			700			0.532	0.523	0.509	OVER-FITTING
	5	10	10			1000			0.532	0.532	0.506	
	Iteration	NVARS	max_depth	n_estimators	tree_method	subsample	eta	eval_metrics	Train	Test	OOT	
XGBoost	1	15	2	20	auto	1	0.3	logloss	0.5162	0.5183	0.4939	
	2	10	3	100	exact	0.8	0.2	logloss	0.5273	0.5300	0.5069	
	3	15	4	300	approx	0.8	0.3	logloss	0.5364	0.5237	0.5037	
	4	10	10	700	auto	0.8	0.2	logloss	0.5439	0.5109	0.4942	OVER-FITTING
	5	15	30	100	auto	1	0.3	logloss	0.5273	0.5241	0.5124	
	Iteration	NVARS	bootstrap_type	max_depth	iterations	l2_leaf_reg	verbose	random_state	Train	Test	OOT	
CatBoost	1	10	Bayesian	2	5	3	0	none	0.4652	0.4701	0.4536	UNDER-FITTING
	2	15	MVS	5	10	6	0	8	0.5017	0.4989	0.4785	
	3	10	Bayesian	8	45	8	0	10	0.5182	0.5180	0.4955	
	4	15	Bayesian	10	100	12	0	8	0.5234	0.5217	0.4985	
	5	10	MVS	15	30	14	0	3	0.5214	0.5214	0.4988	