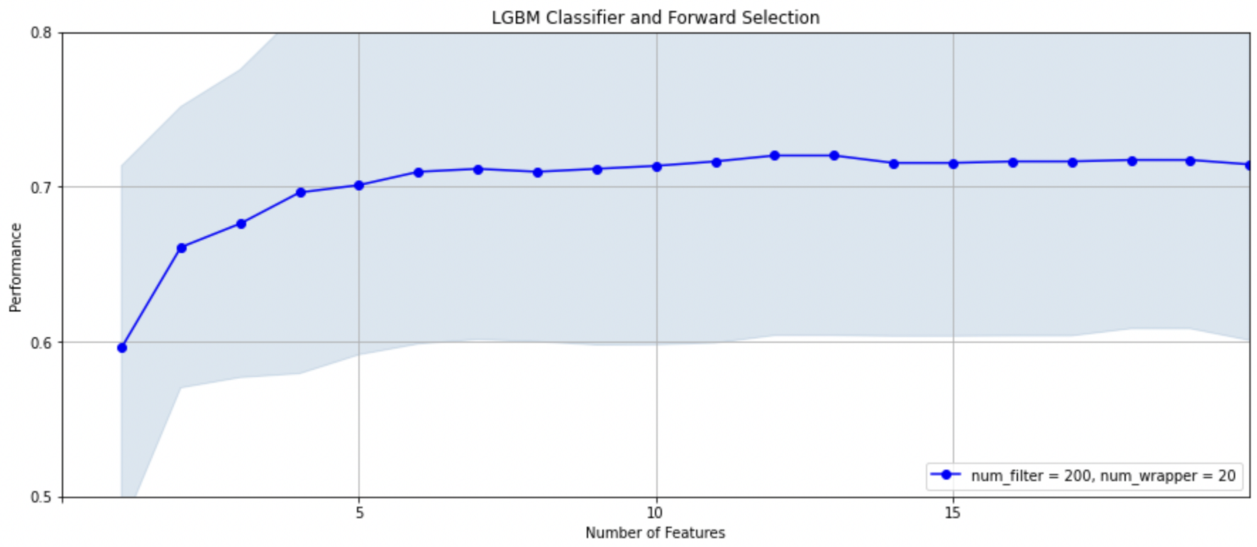
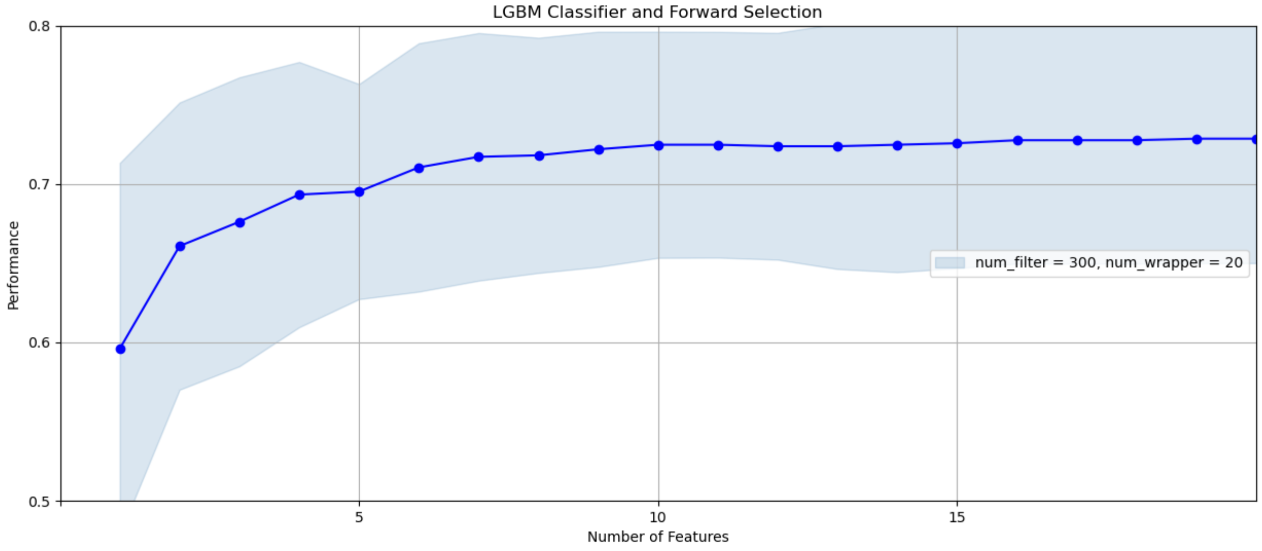
Homework 3 - Explore Feature Selection

Few explorations plots

LGBM Classifier with Forward Selection
num_filter = 200, num_wrapper = 20

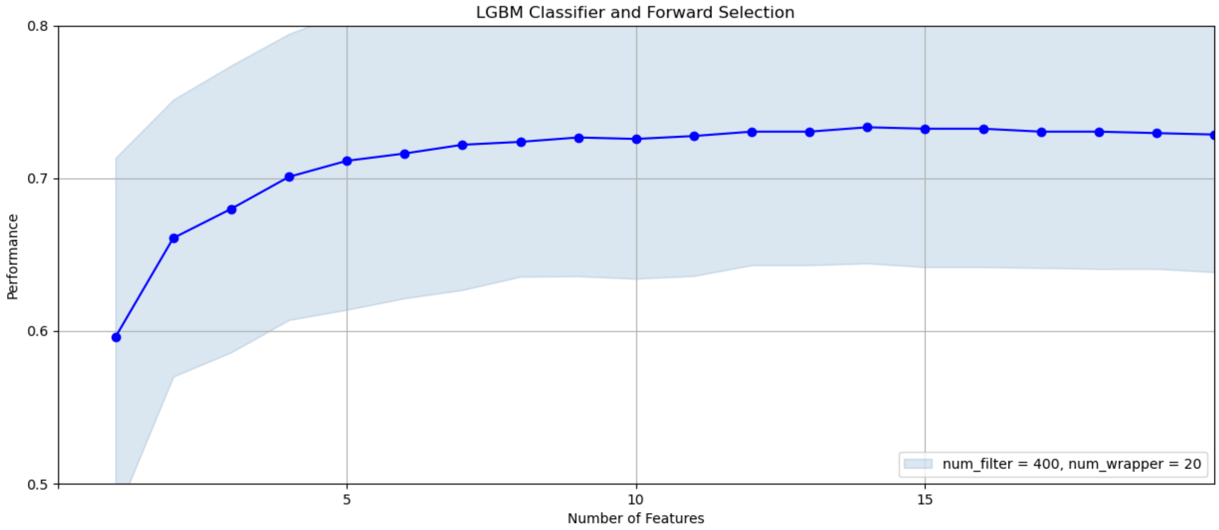


LGBM Classifier with Forward Selection
num_filter = 300, num_wrapper = 20

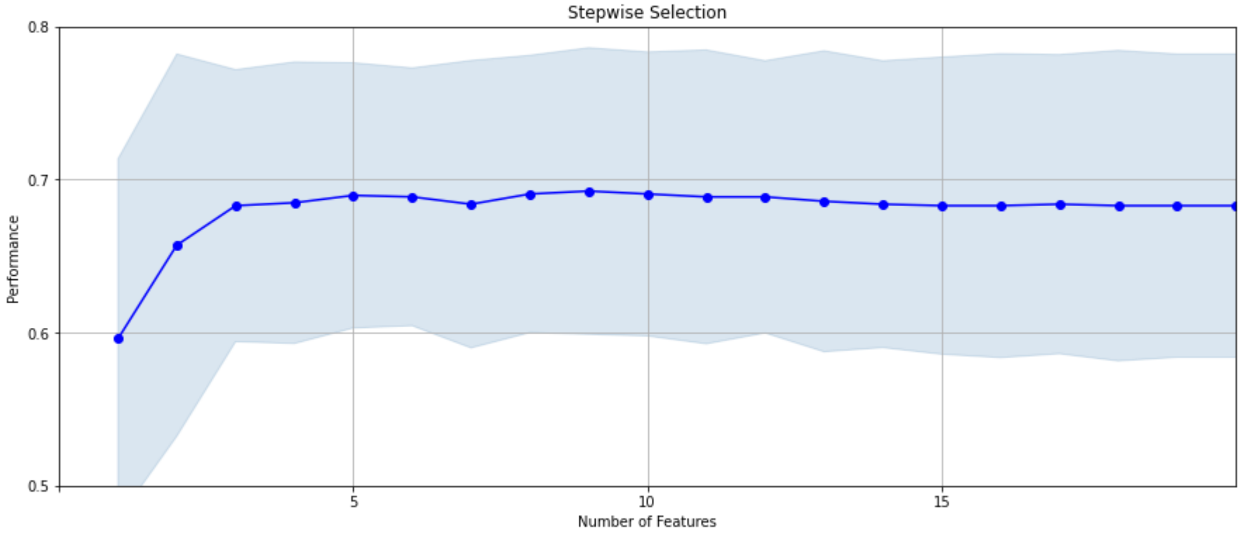


Akansha Lalwani (A59019733)

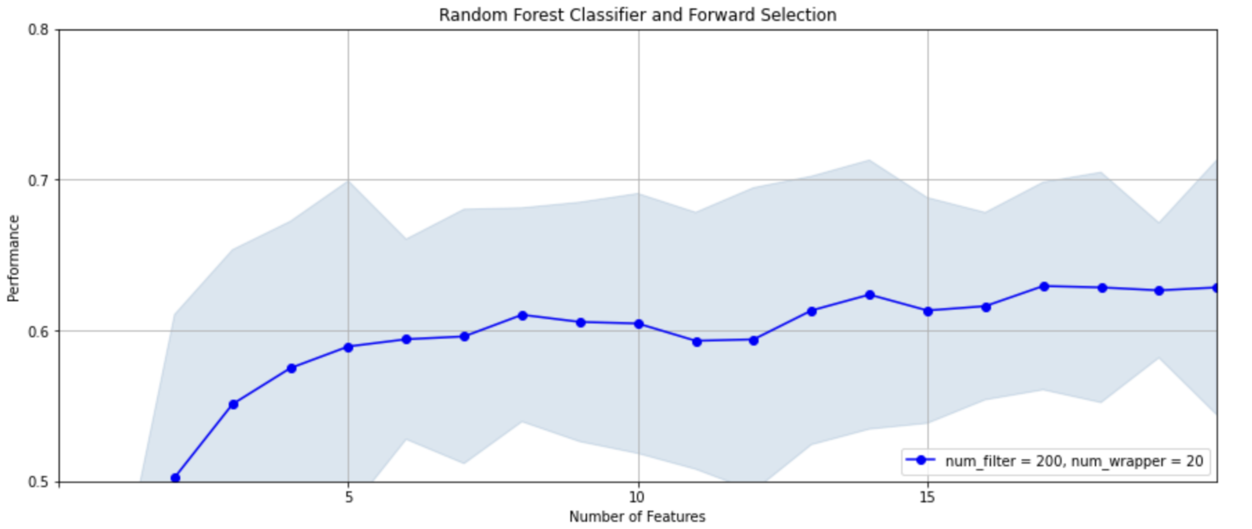
LGBM Classifier with Forward Selection
num_filter = 400, num_wrapper = 20



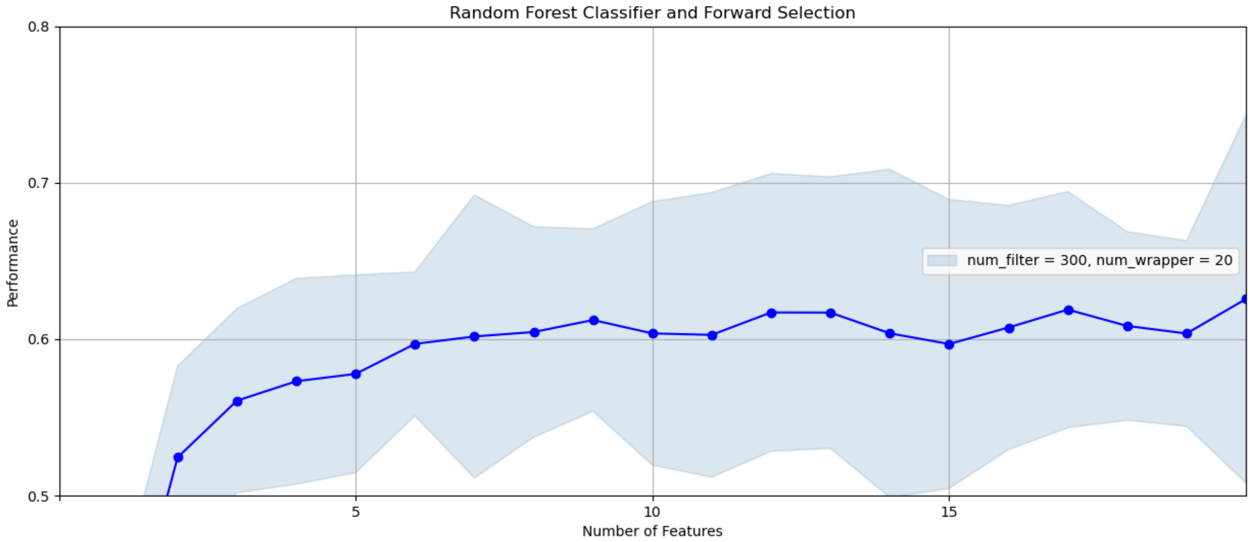
LGBM Classifier with Backward Selection
num_filter = 100, num_wrapper = 20



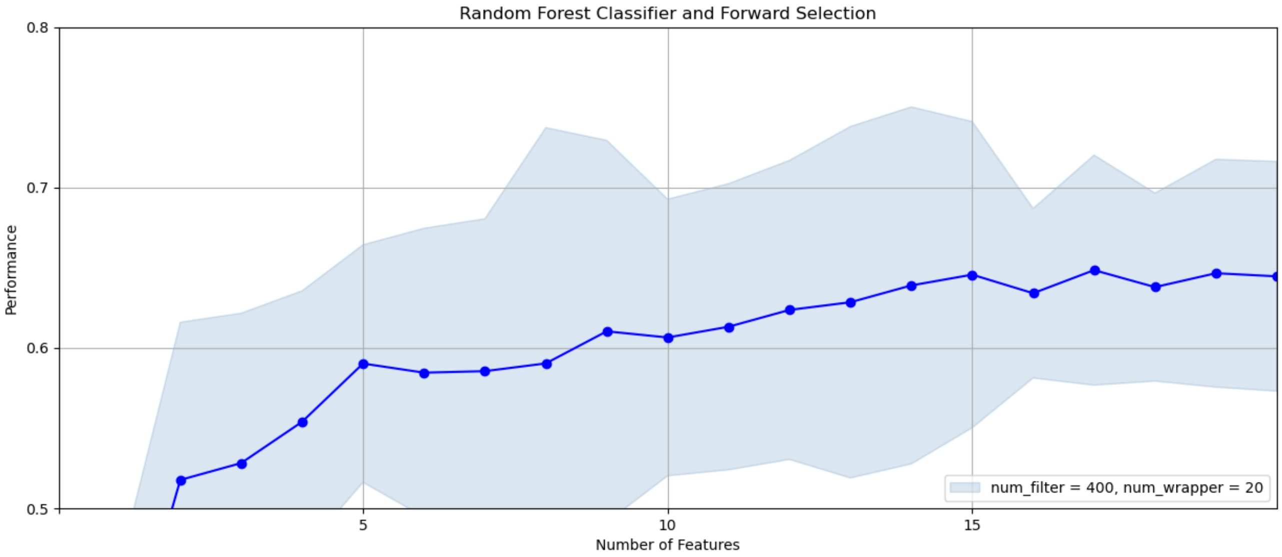
Random Forest Classifier with Forward Selection
num_filter = 200, num_wrapper = 20



Random Forest Classifier with Forward Selection
num_filter = 300, num_wrapper = 20



Random Forest Classifier with Forward Selection
num_filter = 400, num_wrapper = 20



Reasoning for my final selection

- I looked at how the performance increased as the # variables reduced. I looked at the results from forward selection, simple nonlinear wrapper. Looking at the plot, I decided to keep **20 variables** for my modeling **(num_wrapper = 20)**

- I chose **LGBM** as it has no stochastic nature so I always got the same result when I run it multiple times. This is also true when I do a backward selection with LGBM, I always got the same result.

- **Random Forest Classifier** has a stochastic nature so I get a different result each time I run it. I ran it 3 different runs and observed different results each time. None of these look as good as the FS LGBM.

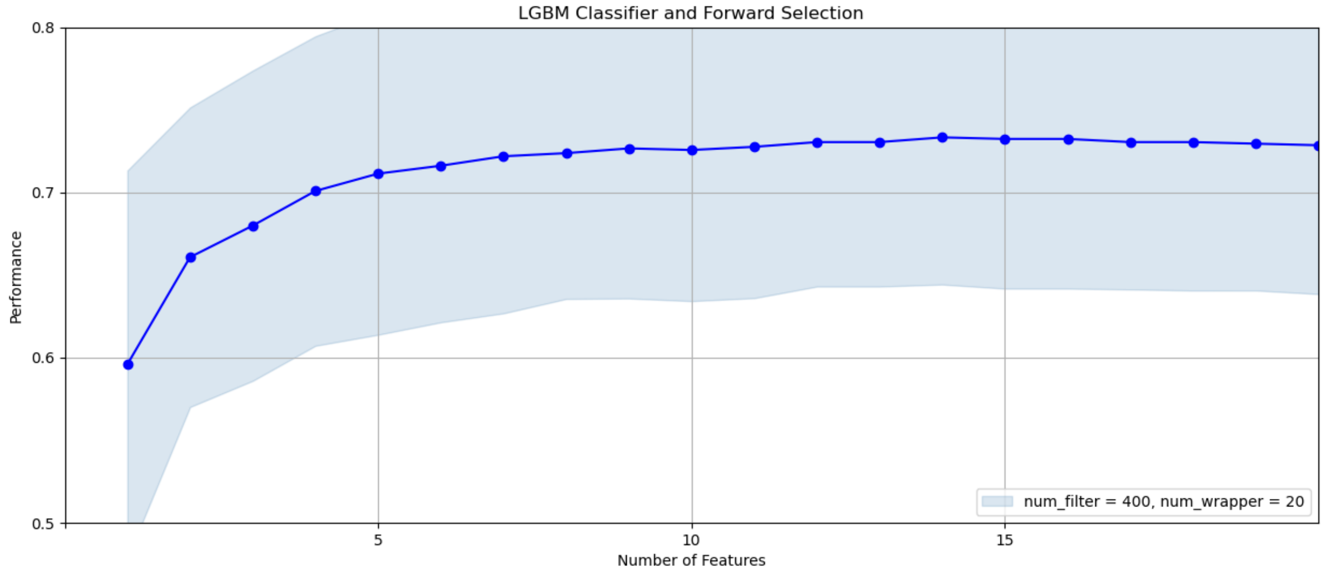
- For **Forward Selection** Experiments - I compared the results from runs with different num_filter values (200, 300, and 400)

- As the wrapper was able to consider a larger set of possible variables (num_filter), the wrapper performance got better until it reached saturation point at around **0.73**

- I observed the best wrapper performance at **num_filter = 400 (about 20% of candidate variables)** and **num_wrapper = 20**, so I decided to use that as my final model. I also chose to use LGBM for my modeling, as it is deterministic and always produces the same result when run multiple times.

Best Model

LGBM Classifier with Forward Selection
num_filter = 400, num_wrapper = 20



List of 20 final variables for modeling

wrapper order	variable	filter score
1	card_merch_total_14	0.630048056206397
2	card_zip3_max_14	0.6295145774877300
3	card_zip3_count_7	0.3878602480787210
4	Merchnum_desc_total_1	0.5284448838378500
5	Merchnum_desc_max_1	0.5236942252906410
6	Merchnum_desc_med_3	0.429393431315388
7	card_zip3_variability_max_3	0.3858683763465740
8	zip3_variability_avg_3	0.4050143980000460
9	merch_zip_total_14	0.44001853964965400
10	merch_zip_max_3	0.5144809550610280
11	Card_Merchnum_desc_total_60	0.5950188274379950
12	state_des_med_3	0.4255449473374220
13	Merchnum_desc_total_7	0.5171234063087220
14	merch_zip_max_1	0.522152980671316
15	card_merch_total_30	0.6154610858354430
16	Card_Merchnum_desc_total_30	0.6062795829714260
17	Card_Merchnum_Zip_total_30	0.6129313921901130
18	Card_Merchnum_Zip_total_14	0.6274209198898390
19	state_des_total_14	0.4908715461547040
20	Merchnum_desc_max_3	0.5168078939090770