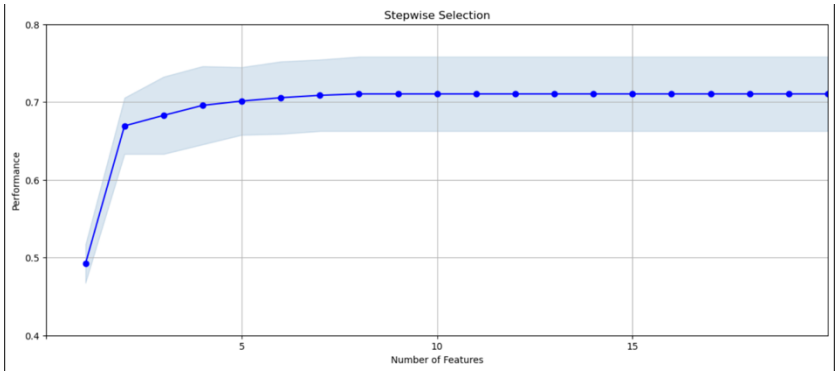
- Results Table:

	Original	1st	2nd	3rd	4th
Backward/Forward	Forward	Forward	Forward	Forward	Forward
Classifier	LightGBM	LightGBM	LightGBM	LightGBM	LightGBM
num_filter	200	1330*0.2 = 266	1330*0.1 = 133	200	1330*0.2 = 266
num_wrapper	20	30	20	35	25
balance	0	0	0	0	0
detect_rate	0.03	0.03	0.03	0.03	0.03
Saturation at:	5	10	7	7	10
Avg. performance	0.71	0.72 – 0.73	0.71	0.71	0.73

Note: I've tried some other approaches like backward, random forest and so on, but the results either not showed, or below 0.7, so I didn't put those into my final report. My computer memory is 8GB with M1, so it's bit hard to run larger set.

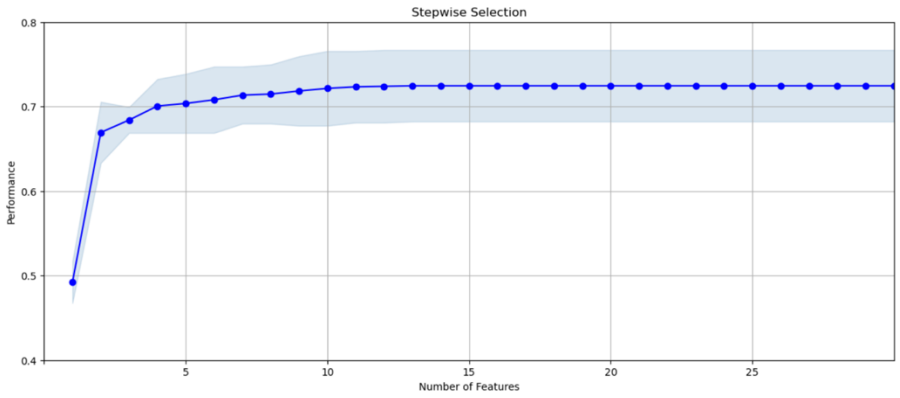
Original:

- A forward feature selection approach was employed, using LightGBM as the primary classifier. The objective was to sift through a dataset with a large number of variables to identify those that significantly impact model performance.
- Evaluating the model's performance with varying numbers of features, it was observed that the model reached a saturation point upon the addition of 5 features. At this juncture, the average performance was 0.71, meeting the performance threshold required for the project.
- The selection process prioritized diversity across entity types, time scales, and quantities within the top variables.
    - Variables encompassing different transaction entities were chosen, including those pertaining to cardholder activity and merchant descriptions.
    - Time scales were represented through variables capturing short-term (such as within 7 days) and long-term (30 days and beyond) transaction metrics.
    - Quantitative diversity was considered by incorporating a variety of measures such as transaction frequency, total transaction amounts, and ratios indicative of deviations from typical behavior.
    - This mixed approach not only enhances multidimensional insights into the data but also aids in capturing complex patterns that may indicate fraudulent activities.



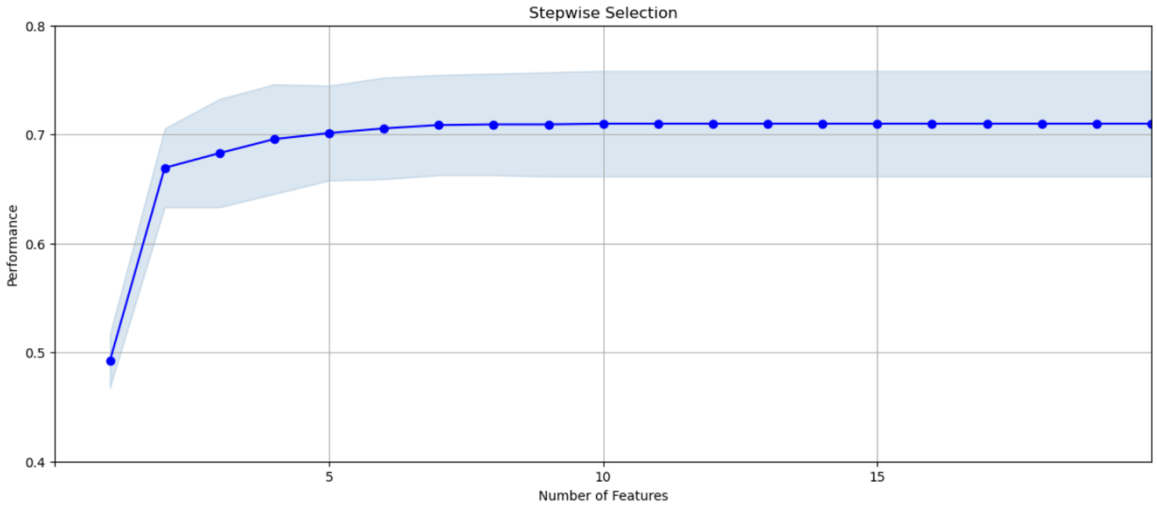
	wrapper order	variable	filter score
0	1	Cardnum_unique_count_for_card_state_1	0.476067
1	2	Card_Merchdesc_total_7	0.324631
2	3	Cardnum_count_1_by_30_sq	0.428229
3	4	Cardnum_max_14	0.318826
4	5	Card_dow_vdratio_0by7	0.467961
5	6	card_state_max_7	0.329132
6	7	card_zip_count_1_by_60_sq	0.314822
7	8	merch_state_total_7	0.284715
8	9	Cardnum_unique_count_for_card_state_3	0.466410
9	10	Cardnum_actual/toal_1	0.459715
10	11	Cardnum_unique_count_for_card_state_7	0.445967
11	12	Card_dow_count_14	0.443405
12	13	Cardnum_unique_count_for_card_zip_7	0.438242
13	14	Cardnum_unique_count_for_Merchnum_7	0.436938
14	15	Cardnum_day_since	0.432169
15	16	Card_dow_day_since	0.432169
16	17	Cardnum_unique_count_for_card_state_14	0.423642
17	18	Cardnum_actual/max_1	0.418902
18	19	Cardnum_unique_count_for_card_zip_14	0.416456
19	20	Cardnum_unique_count_for_Merchnum_14	0.415871

- 1st:
  - A forward feature selection method was applied with LightGBM as the classifier. The goal was to filter through a dataset rich in variables to pinpoint those with a substantial effect on the model's performance.
  - The model's evaluation, conducted across varying numbers of features, indicated a saturation point after incorporating 10 features. This resulted in an average performance ranging between 0.72 and 0.73, surpassing the project's performance benchmark.
  - The selection emphasized a wide-ranging mix in terms of entity types, timeframes, and quantitative aspects among the top variables.
    - Chosen variables represented diverse transactional entities, including aspects related to cardholder actions and merchant details.
    - The timeframes were captured via metrics for short-term (within a 7-day window) and more extended periods (beyond 30 days).
    - A breadth of quantitative dimensions was incorporated, covering transaction frequencies, aggregate transaction volumes, and ratios that might signal deviations from normative patterns.
  - This holistic strategy not only provided a multi-faceted view of the data but also helped identify intricate patterns potentially indicative of fraudulent activity.



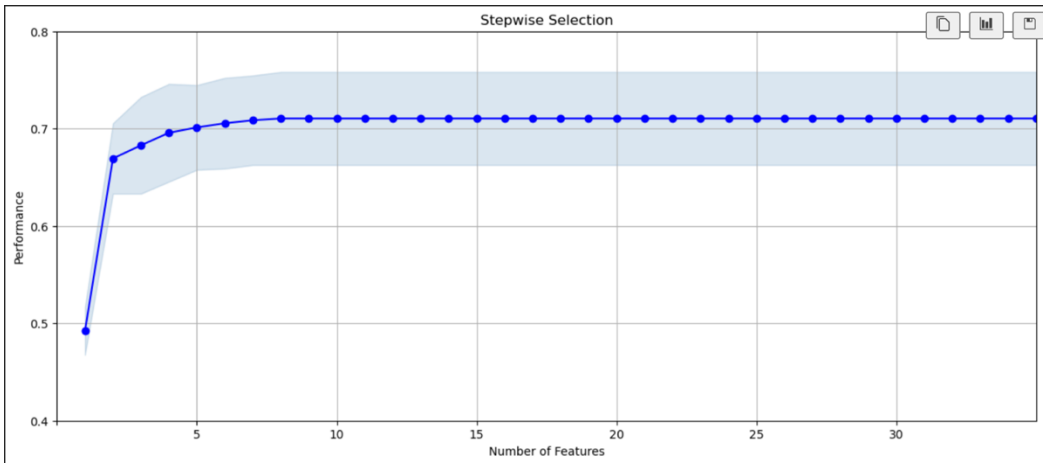
	wrapper order	variable	filter score
0	1	Cardnum_unique_count_for_card_state_1	0.476067
1	2	Card_Merchdesc_total_7	0.324631
2	3	Card_Merchdesc_vdratio_0by7	0.268933
3	4	Cardnum_count_1_by_30_sq	0.428229
4	5	state_des_total_3	0.315540
5	6	Cardnum_max_7	0.410589
6	7	Card_dow_unique_count_for_merch_state_1	0.447357
7	8	Cardnum_count_7	0.526897
8	9	Card_dow_vdratio_0by14	0.479086
9	10	card_state_max_14	0.305946
10	11	Cardnum_unique_count_for_card_state_60	0.343111
11	12	Cardnum_unique_count_for_Merchnum_1	0.472017
12	13	card_zip_total_60	0.302130
13	14	Cardnum_total_14	0.494375
14	15	Card_dow_max_7	0.486177
15	16	Cardnum_variability_max_0	0.484245
16	17	Card_dow_count_7	0.482384
17	18	Cardnum_actual/toal_0	0.479550
18	19	Cardnum_variability_max_1	0.477836
19	20	Card_dow_total_30	0.474759
20	21	Card_dow_max_14	0.470975

- 2nd:
  - The forward feature selection was carried out using LightGBM as the classifier, intending to isolate impactful variables from a dataset with a vast array of options.
  - Analysis of the model's performance in relation to the number of features utilized highlighted a saturation point reached at the addition of 6 features, with performance leveling at 0.71.
  - The selection process was calibrated to encompass a wide range of entity types, time dimensions, and quantitative metrics within the most influential variables.
    - The chosen variables offer a comprehensive view of cardholders' behavioral patterns and merchants' transactional details.
    - The time-related variables span from short-term windows (within 7 days) to more extended periods (beyond 30 days), capturing immediate and long-standing transactional behaviors.
    - The selection includes a variety of measures such as the frequency of transactions, total transaction amounts, and ratios that signal deviations from typical behaviors, providing a broad spectrum for detecting potential anomalies.
  - This methodical approach does more than just expand the analytical scope of the data; it also contributes significantly to the identification of complex, nuanced patterns that could be indicative of fraudulent activities.



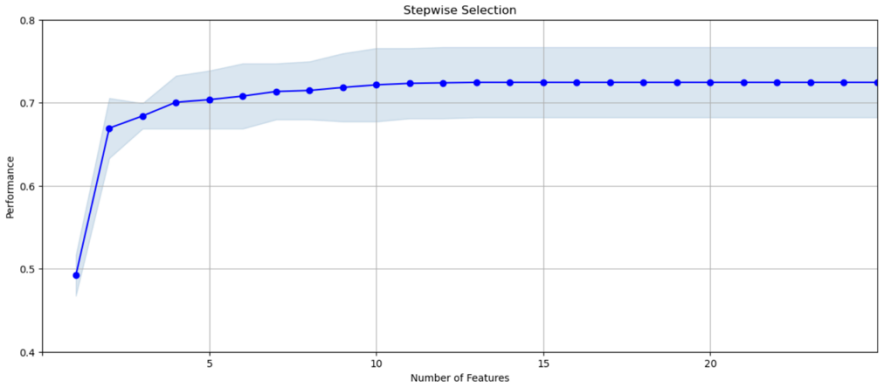
	wrapper order	variable	filter score
0	1	Cardnum_unique_count_for_card_state_1	0.476067
1	2	Card_Merchdesc_total_7	0.324631
2	3	Cardnum_count_1_by_30_sq	0.428229
3	4	Cardnum_max_14	0.318826
4	5	Card_dow_vdratio_0by7	0.467961
5	6	card_state_max_7	0.329132
6	7	card_zip_count_1_by_60_sq	0.314822
7	8	Cardnum_actual/toal_0	0.479550
8	9	Card_dow_max_14	0.470975
9	10	state_des_total_3	0.315540
10	11	Cardnum_total_14	0.494375
11	12	Card_dow_total_30	0.474759
12	13	Cardnum_unique_count_for_card_state_3	0.466410
13	14	Cardnum_unique_count_for_card_zip_3	0.464311
14	15	Cardnum_unique_count_for_Merchnum_3	0.460748
15	16	Cardnum_actual/toal_1	0.459715
16	17	Cardnum_unique_count_for_card_state_7	0.445967
17	18	Cardnum_actual/max_0	0.445726
18	19	Card_dow_unique_count_for_merch_state_1	0.447357
19	20	Cardnum_count_14	0.445443

- 3rd:
  - The feature selection was guided using a forward approach, with LightGBM serving as the classifier. The aim was to traverse through the data, rich in variables, to select those with a notable impact on model performance.
  - Performance evaluation at various feature levels indicated that the model reached saturation with the inclusion of 7 features. At this point, the performance stabilized at 0.71.
  - The selection sought to ensure diversity in terms of entity types, time dimensions, and quantitative metrics among the upper echelon of variables.
    - A blend of variables related to cardholders' activity patterns and merchants' transactional characteristics was selected.
    - Temporal aspects were covered by short-term (within 7 days) and extended-period (beyond 30 days) transactional variables.
    - The assortment included measures of transaction frequency, cumulative transaction amounts, and ratios reflective of behavioral anomalies.
  - This amalgamated approach not only broadened the perspective on data but also aided in pinpointing complex patterns potentially indicative of fraudulent transactions.



	wrapper order	variable	filter score
0	1	Cardnum_unique_count_for_card_state_1	0.476067
1	2	Card_Merchdesc_total_7	0.324631
2	3	Cardnum_count_1_by_30_sq	0.428229
3	4	Cardnum_max_14	0.318826
4	5	Card_dow_vdratio_0by7	0.467961
5	6	card_state_max_7	0.329132
6	7	card_zip_count_1_by_60_sq	0.314822
7	8	merch_state_total_7	0.284715
8	9	Cardnum_unique_count_for_card_state_3	0.466410
9	10	Cardnum_actual/toal_1	0.459715
10	11	Cardnum_unique_count_for_card_state_7	0.445967
11	12	Card_dow_count_14	0.443405
12	13	Cardnum_unique_count_for_card_zip_7	0.438242
13	14	Cardnum_unique_count_for_Merchnum_7	0.436938
14	15	Cardnum_day_since	0.432169
15	16	Card_dow_day_since	0.432169
16	17	Cardnum_unique_count_for_card_state_14	0.423642
17	18	Cardnum_actual/max_1	0.418902
18	19	Cardnum_unique_count_for_card_zip_14	0.416456
19	20	Cardnum_unique_count_for_Merchnum_14	0.415871
20	21	Cardnum_count_1_by_14	0.413860

- 4th:
  - A forward selection approach with LightGBM as the classifier was conducted. The process aimed to navigate through a dataset abundant with variables to identify those significantly influencing model performance.
  - An evaluation of the model's performance with a varying number of features pinpointed saturation at the addition of 10 features. The performance measured at this point was 0.73.
  - The approach emphasized the selection of variables that were diverse in terms of entity types, temporal frames, and quantitative metrics.
    - Selected variables covered different aspects of transaction entities, ranging from cardholder activities to merchant transaction details.
    - Time dimensions were reflected by variables that captured both short-term (within 7 days) and prolonged (beyond 30 days) transaction activities.
    - The selection incorporated a range of measures including transaction frequency, total transaction volumes, and ratios indicating deviations from standard patterns.
  - This diversified approach was designed to deepen the understanding of the data and to assist in detecting complex patterns that may suggest fraudulent activities.



	wrapper order	variable	filter score
0	1	Cardnum_unique_count_for_card_state_1	0.476067
1	2	Card_Merchdesc_total_7	0.324631
2	3	Card_Merchdesc_vdratio_0by7	0.268933
3	4	Cardnum_count_1_by_30_sq	0.428229
4	5	state_des_total_3	0.315540
5	6	Cardnum_max_7	0.410589
6	7	Card_dow_unique_count_for_merch_state_1	0.447357
7	8	Cardnum_count_7	0.526897
8	9	Card_dow_vdratio_0by14	0.479086
9	10	card_state_max_14	0.305946
10	11	Cardnum_unique_count_for_card_state_60	0.343111
11	12	Cardnum_unique_count_for_Merchnum_1	0.472017
12	13	card_zip_total_60	0.302130
13	14	Cardnum_total_14	0.494375
14	15	Card_dow_max_7	0.486177
15	16	Cardnum_variability_max_0	0.484245
16	17	Card_dow_count_7	0.482384
17	18	Cardnum_actual/toal_0	0.479550
18	19	Cardnum_variability_max_1	0.477836
19	20	Card_dow_total_30	0.474759
20	21	Card_dow_max_14	0.470975

Overall, after comparing the mix set of features:

- Original Iteration:
    - Features a combination of card state counts, merchant totals, count ratios, maximum transaction values, day of the week ratios, and card ZIP counts.
- 1st Iteration:
    - Maintains a diverse range of features from the original set and introduces additional time-based variables like Card_dow_max_14 and state_des_total_3, enhancing the temporal analysis. It also enriches the dataset with various transaction amount ratios such as Cardnum_actual/total_0 and Cardnum_actual/max_0, which help capture the nuances of transactional behavior across different periods.
- 2nd Iteration:
    - Similar to the original but with the addition of variables like Cardnum_count_1_by_14, providing an expanded scope of time-based variables.
- 3rd Iteration:
    - Preserves a consistent type of variables, with a focus on capturing both merchant-related and cardholder-related behaviors, as seen in the original iteration.
- 4th Iteration:
    - Introduces new types of variables, including card-merchant ratio variables across different time frames and detailed state descriptions, potentially offering fresh insights into transaction behavior.

In reviewing the top ~10 variables from each iteration, the original iteration already provides a mix of entity types and time scales, which is crucial for generalization. However, it is the 4th Iteration that stands out with the introduction of distinct variable types, such as the card_merch_vdratio_0by14 and specific day counts like Card_dow_max_14, which may capture unique patterns in the data.

The choice between the original and the 4th Iteration would hinge on which set of features encompasses the broadest spectrum of behaviors while avoiding redundancy. The inclusion of unique variable types in the 4th Iteration could offer an advantage in capturing diverse patterns, which is critical for a model's ability to generalize across various data scenarios.

	wrapper order	variable	filter score
0	1	Cardnum_unique_count_for_card_state_1	0.476067
1	2	Card_Merchdesc_total_7	0.324631
2	3	Card_Merchdesc_vdratio_0by7	0.268933
3	4	Cardnum_count_1_by_30_sq	0.428229
4	5	state_des_total_3	0.315540
5	6	Cardnum_max_7	0.410589
6	7	Card_dow_unique_count_for_merch_state_1	0.447357
7	8	Cardnum_count_7	0.526897
8	9	Card_dow_vdratio_0by14	0.479086
9	10	card_state_max_14	0.305946
10	11	Cardnum_unique_count_for_card_state_60	0.343111
11	12	Cardnum_unique_count_for_Merchnum_1	0.472017
12	13	card_zip_total_60	0.302130
13	14	Cardnum_total_14	0.494375
14	15	Card_dow_max_7	0.486177
15	16	Cardnum_variability_max_0	0.484245
16	17	Card_dow_count_7	0.482384
17	18	Cardnum_actual/toal_0	0.479550
18	19	Cardnum_variability_max_1	0.477836
19	20	Card_dow_total_30	0.474759
20	21	Card_dow_max_14	0.470975