|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 |

* Results Table:

**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.

A graph showing a line of a selection

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generated

* 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.

A graph showing a number of features

Description automatically generated

A screenshot of a graph

Description automatically generated

* 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.
  + A graph showing a line of a graph

    Description automatically generated with medium confidenceThis 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.

A screenshot of a graph

Description automatically generated

* 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.

A graph with a line in the middle

Description automatically generated

A screenshot of a graph

Description automatically generated

* 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.
  + A graph showing a line of a graph

    Description automatically generated with medium confidenceThis diversified approach was designed to deepen the understanding of the data and to assist in detecting complex patterns that may suggest fraudulent activities.

A screenshot of a black screen

Description automatically generated

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.

A screenshot of a black screen

Description automatically generatedThe 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.