Creator **Olusanmi Hundogan** Created **Jul 30, 2023, 19:47** Last updated **Mar 06, 2024, 11:43**

Files

- experiments/experiment_static_feature_selection.py
- notebooks/analyze_static_feature_selection.py
- notebooks/analyze_feature_set_initial_discovery.py

Motivation

After the integration of the 12data dataset information, we have over 100 features and many of them have unknown values or are highly correlated with each other. This number of features leads to a higher training time and is prone to overfitting. Therefore we need to investigate which subset of the feature set leads to comparable results.

Design

Selecting the Features

In order to select viable feature subsets, we utilise a number of feature selection approaches:

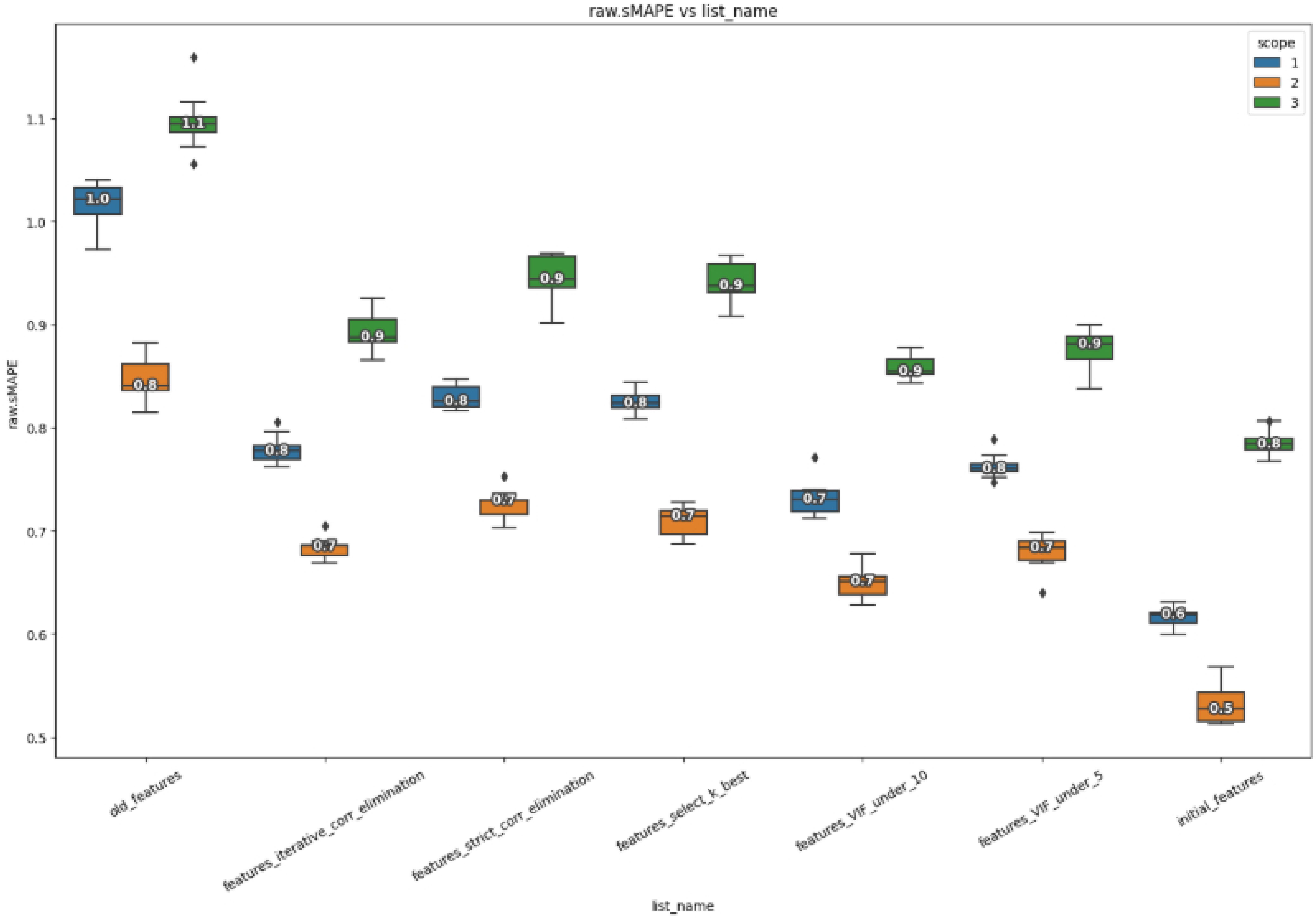
- old features (before 12data integration) -> 13 features
- initial features (after 12data integration) -> over 100 features
- iteratively eliminating the features with highest correlation -> 33 features
- keeping only features that do not correlate highly with other features (threshold 0.5) -> 8 features
- select K best features based on f-regression -> 10 features
- choose best features based on the VarianceInflation Factor (VIF) under 10 -> 39 features
- choose best features based on the VarianceInflation Factor (VIF) under 5 -> 29 features

Experimental Design

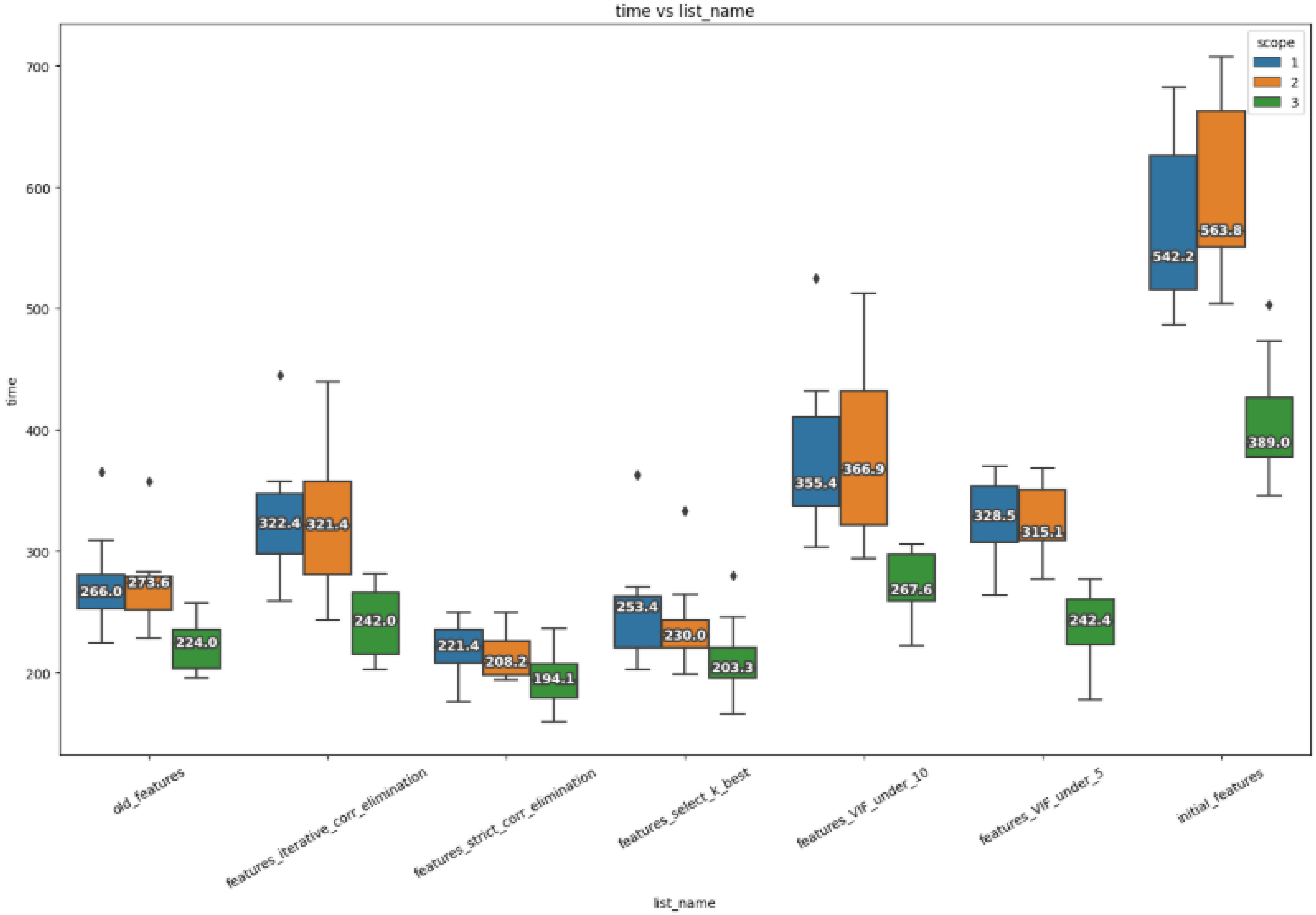
We ran the experiment 10 times for all scopes, each time we specifically select only the features in the subset categories. For each feature list we define a new pipeline. We measure the runtime, as well as the raw sMAPE values.

Results and Insights

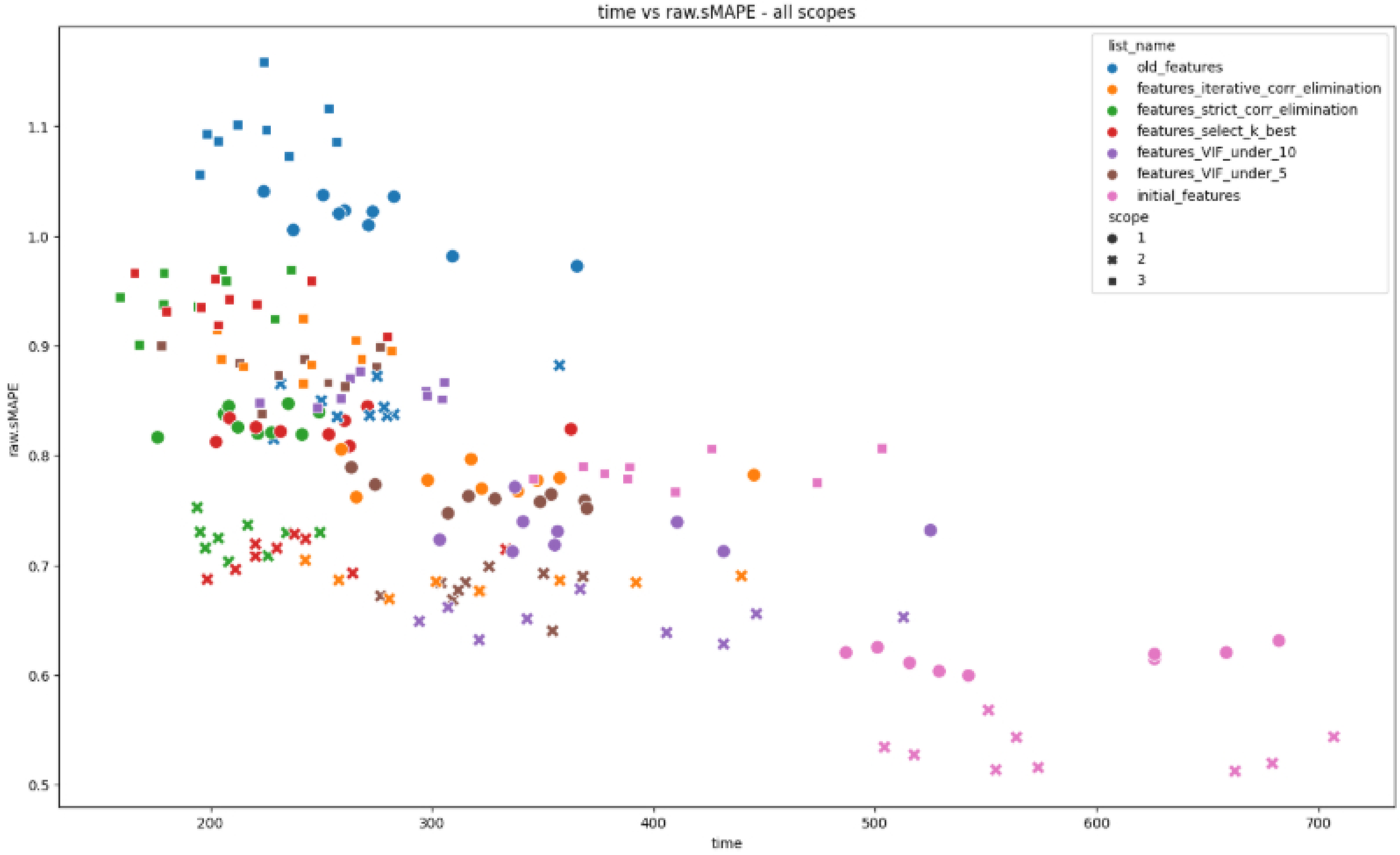
The raw sMAPE values are the smallest for the VIF_under_10 feature subset for all scopes compared to the other subsets, excluding the initial features. The initial_features subset was expected to perform the best. The worst subset is the old_features, which is surprising.

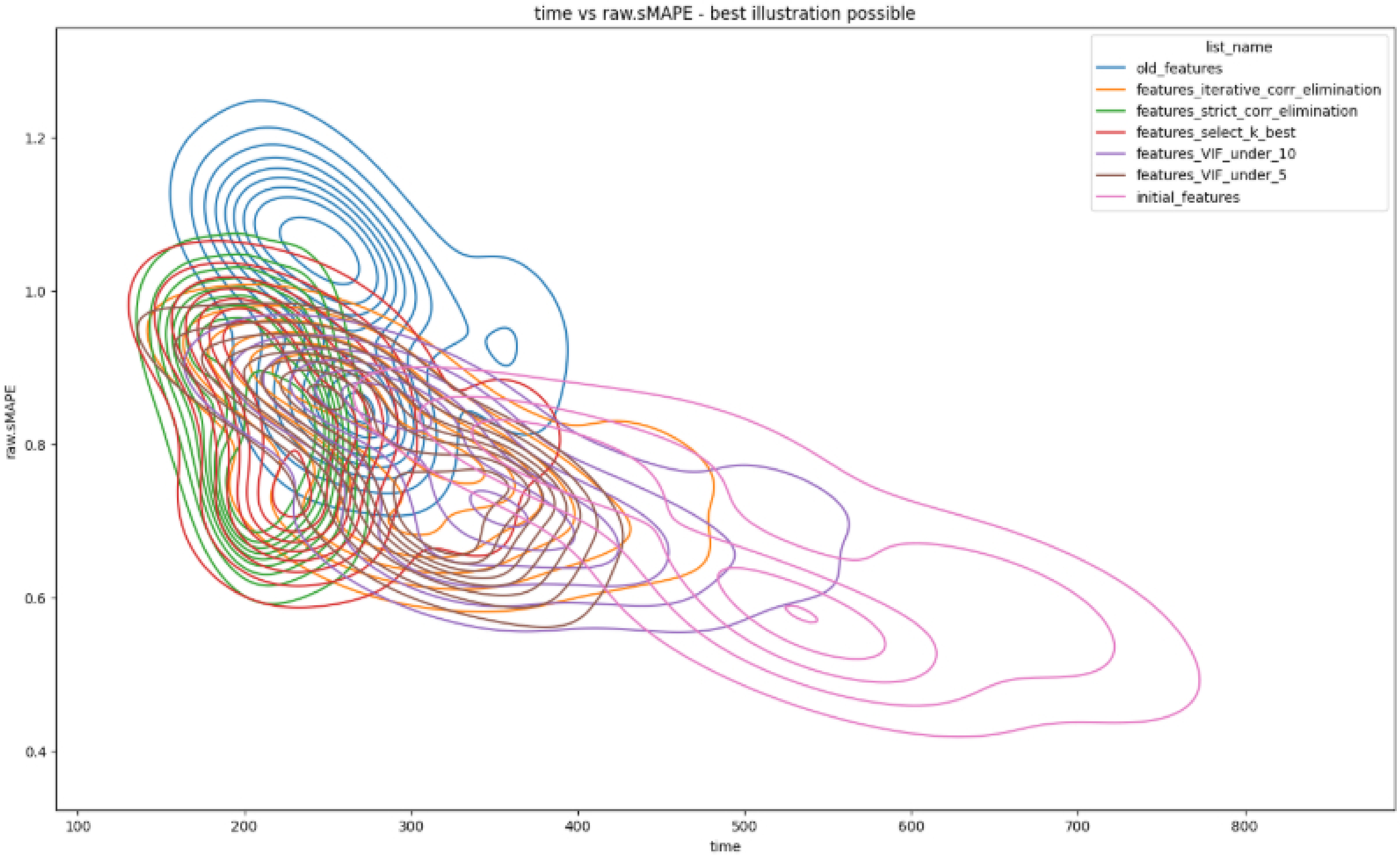


In terms of time, the smaller feature subsets expectedly perform the best and the ones which contain more features perform worse. VIF_under_10has a higher time variance, but performs reasonably well for its size.



The tradeoff between sMAPEand time is more pronounced in this plot. For all 3 scopes, we can see that only the VIF_under_10subset experiments are within the elbow area. In other casessome points lie in the elbow are for ceratin scopes, but the others are further away (ex. old_features).





Random Insights

For this experiemnt we only used numerical features, so we can get pretty good results even without categorical features. This means that, by adding those, the sMAPEvalue should decreaseeven more.

Decision

We decided to use the VIF_under_10feature subset for the live prediction casebecausethe user should not have to provide over 100 features to use for prediction.

We will explore the use of PCAfor subsequent experiments to maintain generalisability acrossall experiments.

For scope imputation, we will use all the features (initial_features subset) to make sure the imputed predictions are more accurate.

Update 08.08.2023

We realized that **selecting** VIF_under_10features is **antithetical to** the idea of being able to effortlessly **extending features through prefixing**. Hence, if we alwaysselect **ft_numc_x1**, **ft_numc_x2**, **ft_numc_x3** the model will alwaysuse those three even if we add **ft_numc_x_special** to the dataset. Without selection, the model would pick up these features automatically. With selection, we have to explicitly add them.

Two options are available to ensure features are reduced, while being able to extend them with automatic feature detection:

- We pivot from a select-features approach to a drop-features approach.
- We use PCAto reduce the amount of features.

Therefore, going forward, we will try to improve the PCA approach, as this would allow the use of all features, while reducing the amount of features used. Furthermore, any improvements and insights on experiments with PCA will likely carry over to another feature reduction approach.