Creator **Olusanmi Hundogan** Created **Jan 12, 2024, 00:15** Last updated **Jan 12, 2024, 01:03**

Files

- notebooks/analyze_missing_value_imputers_kcluster.py
- experiments/experiment_missing_value_imputers_kcluster.py
-

Motivation

This experiment acts as a preliminary study for the experiment which compares different imputation methods. Currently we use a binning strategy to impute missing values in the dataset. For that purpose we select a feature as the basis for bucketing the other features. We use the mean/median of each bucket to compute the missing values of other features based on the selected feature's bucket assignment. For categorical features we compute the median per feature and bucket. For numerical features, we discretize the range into buckets/bins first and use these buckets as categorical variable. The search space of the number of bins for discretizing the numerical features and the feature chosen as reference feature is too large to include in the main experiment.

To narrow the search space we conduct this preliminary study to choose the best hyperparameters for this task.

For this experiment we tested two imputers.

- CategoricalStatisticsImputer: Uses a categorical variable to compute the median value for each category and each feature.
- NumericalStatisticsImputer: Uses a numerical variable to compute the median value for each category and each feature.

Design

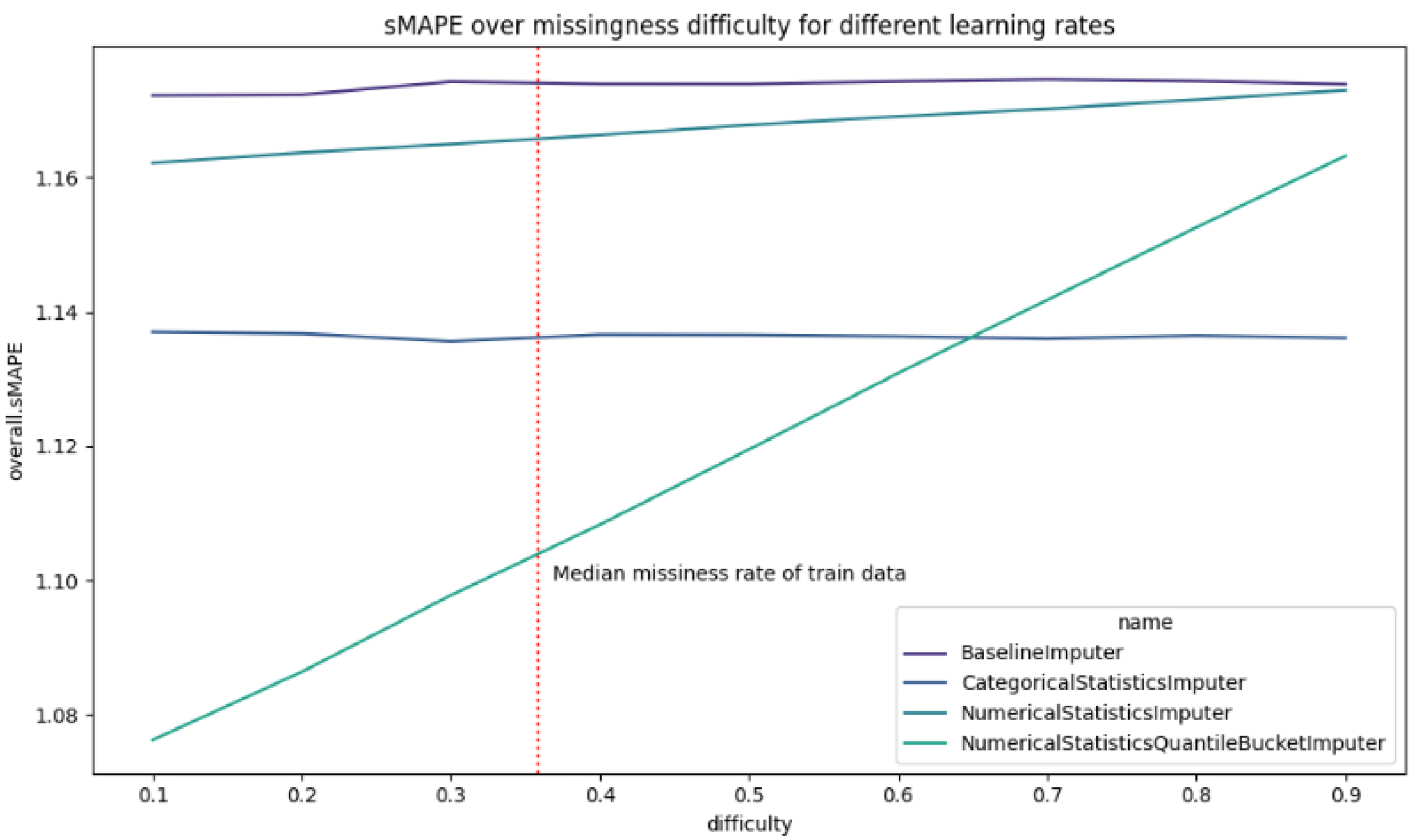
Within the experiment we explored different number of buckets and features. The number of buckets were [3, 5, 7, 9, 11]. As numerical features we use 'ft_numc_total_assets', 'ft_numc_revenue', 'ft_numc_net_income', 'ft_numc_inventories', 'ft_numc_roe', 'ft_numc_ppe', 'ft_numc_equity', 'ft_numc_total_liabilities', 'ft_numc_common_stock', 'ft_numc_cash_and_cash_equivalents'. As categorical variables we used "ft_catm_country_code", "ft_catm_industry_name", "ft_catm_sector_name". We also explore two different ways of discretizing the numerical features. First, by taking the range of the data and splitting the bin thresholds evenly or by using the quantile ranges. The ladder ensures that each bin has the same amount of data points.

We evaluate the imputer by removing values from a dataframe gradually and measuring the sMAPE of reconstructing the values. We test difficulties from 0.1 to 0.9 in even steps. A difficulty of 0.1 corresponds to the artificial removal of 10% of known values. (The number is not completely accurate because the 10% applies to the dataframe, but the dataframe already contains missing values.)

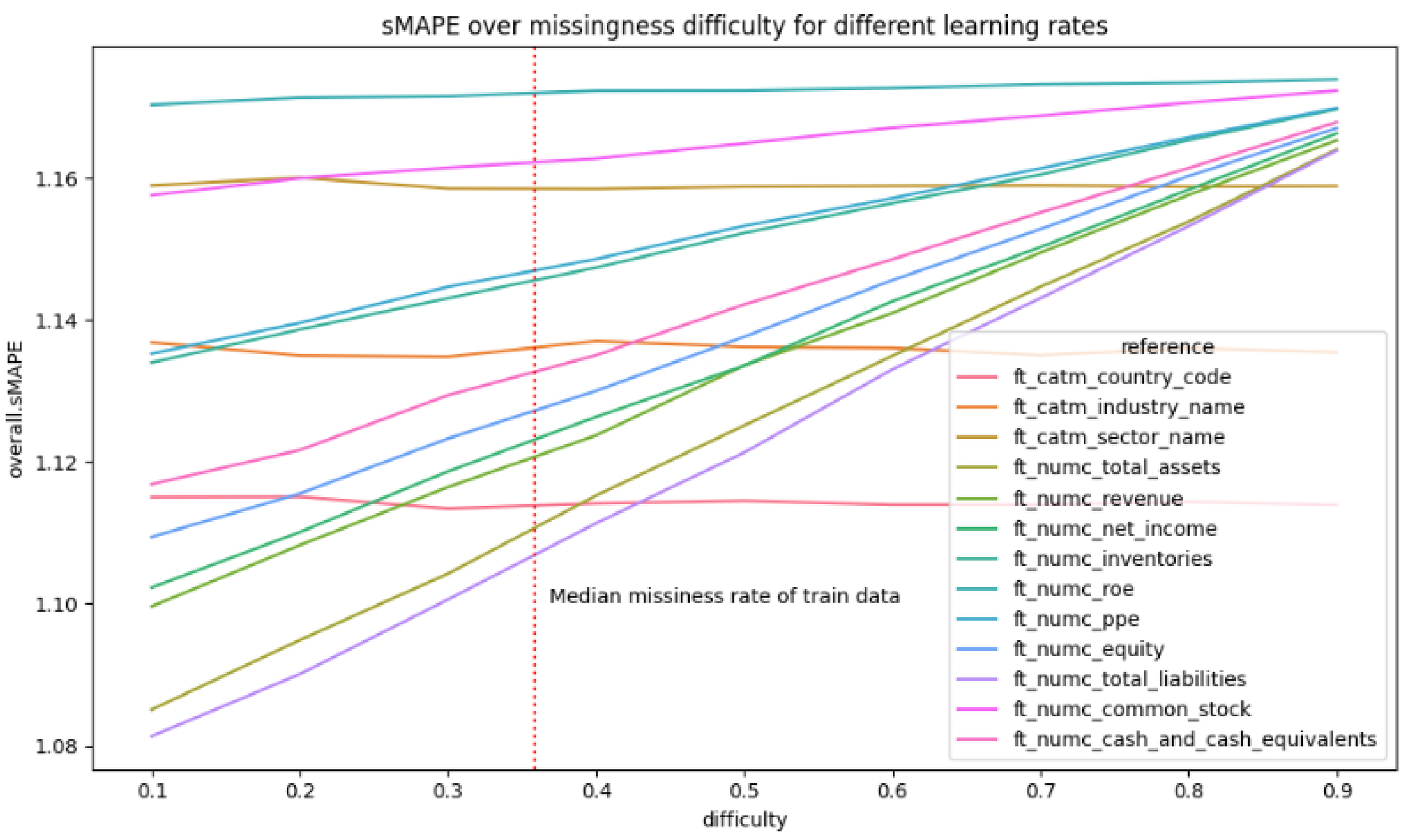
Fruthermore, we only impute on features that have a rate of missingness below 30% across the data set. After applying this feature selection 50% of the data rows have a missingness of below 10%.

Results and Insight

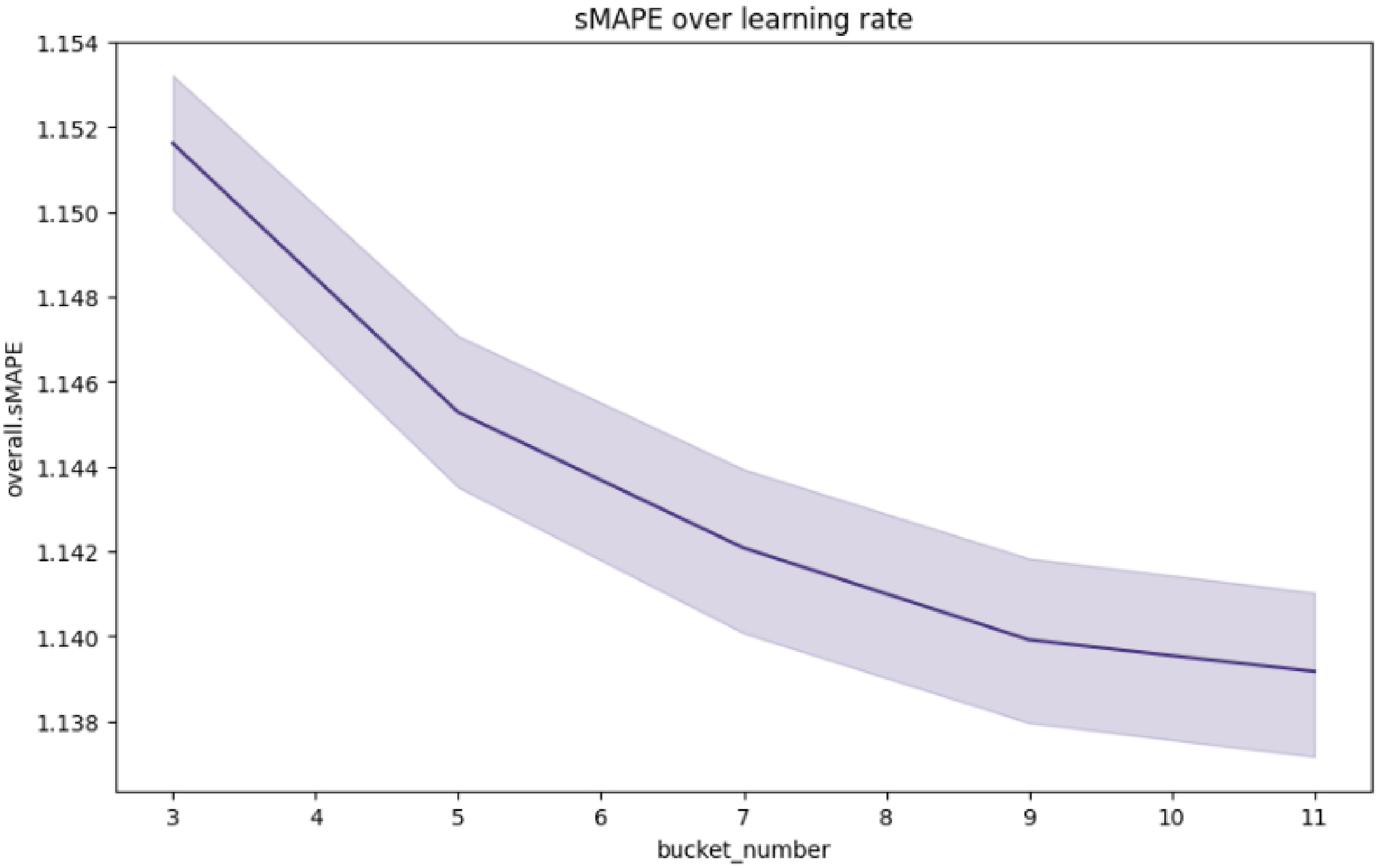
The plot shows all the different configurations tested across all difficulty levels. The results reveal a strong effect of the binning strategy on the performance. It also shows that the categorical imputers and the non-quantile imputers retain a stable performance regardless of the feature difficulty. The quantile based numeric imputer decreases in performance with increasing difficulty. However, it also strongly outperforms the other methods.



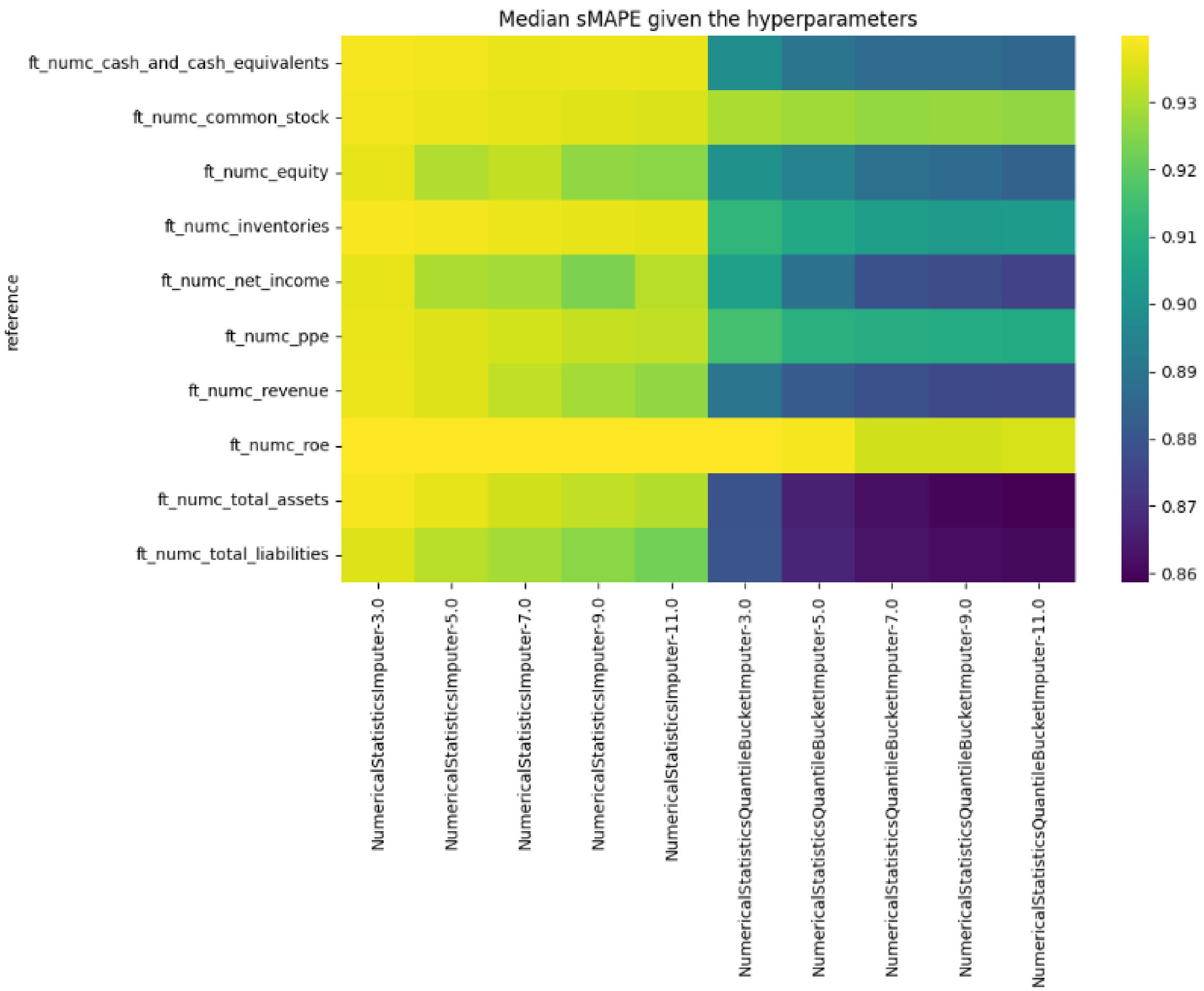
In this plot we can see that the feature performances differ widely from eachother. The best performing features are "ft_numc_total_liabilities" and "ft_numc_total_assets".



The bucket number is a strong factor influencing the the model performance. I does not seem to have converged at 11.



This figure shows the median smape values given heach configuration . The results show again that total assetsand total liabilities perform better than other features. Roe and common stock do not perform well and revenue performs within the middle range. We also see the difference between the normal binning method and the quantile binning method . Lastly, we see how the performance improves for the quantile methods if the bucket number is increased.



Decision

Update : 09.01.2024

For the final experiment we choose the bucket number to be 11 for the numerical statistics imputer with the features

- ft_numc_revenue Becausethat is the imputer currently in use
- ft_numc_total_liability: Becauseit performed exceptionally well
- ft_numc_total_assets Becauseit performed exceptionally well

For the categorical imputers we use all three categorical variables again for the final study.