Experiment

Creator **Olusanmi Hundogan** Created **Jan 12, 2024, 01:39** Last updated **Jan 27, 2024, 16:09**

Files

- experiment_missing_value_imputers_downstream_task.ipynb
- experiment_missing_value_imputers_downstream_task_validation.ipynb
- analyze_missing_value_imputers_downstream_task.ipynb

Motivation

As many of the machine learning models cannot work with missing data fields, we employ an imputer as a crucial step in the pipeline. To find the best pipeline we evaluated several imputation strategy purely on the act of imputation by removing known values and attempting to re-construct them using the imputers. For this study, we use a subset of these imputers to decide which imputer improves the model predictions the most. For this purpose we compared a number of confiugrations to select the best imputer for the model.

The configurations were:

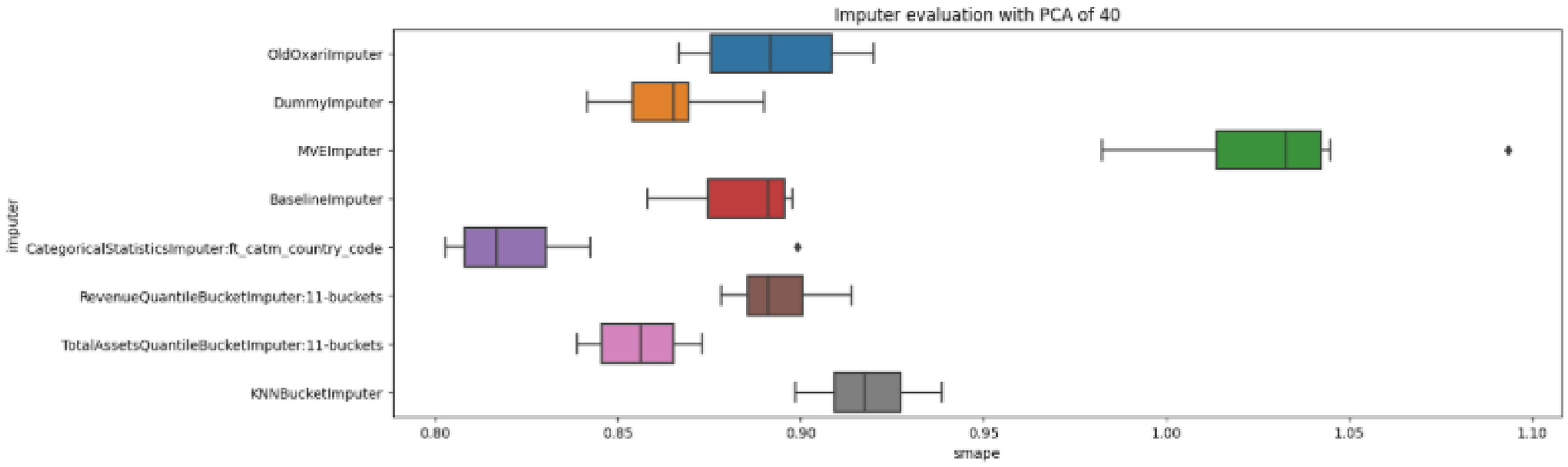
- BaselineImputer: Imputes values based on a simple heuristic
- DummyImputer: Imputes values with random numbers
- MVEImputers: Column-wise RegressionImputers
  - Decision Tree
- OldOxariImputer: Same as MVEImputater but with RandomForest as regressor
- K-Bucket Imputers: Imputations based on an instance based K algorithm
  - KNN (K=9)
- Categorical StatisticsImputer: Computes imputation values based on categories
  - Country
  - Industry
- Numerical StatisticsImputer: Similar to categorical but based on discretized numerical values
  - TotalAssets(num_buckets=11)
  - Revenue(num_buckets=11)

Design

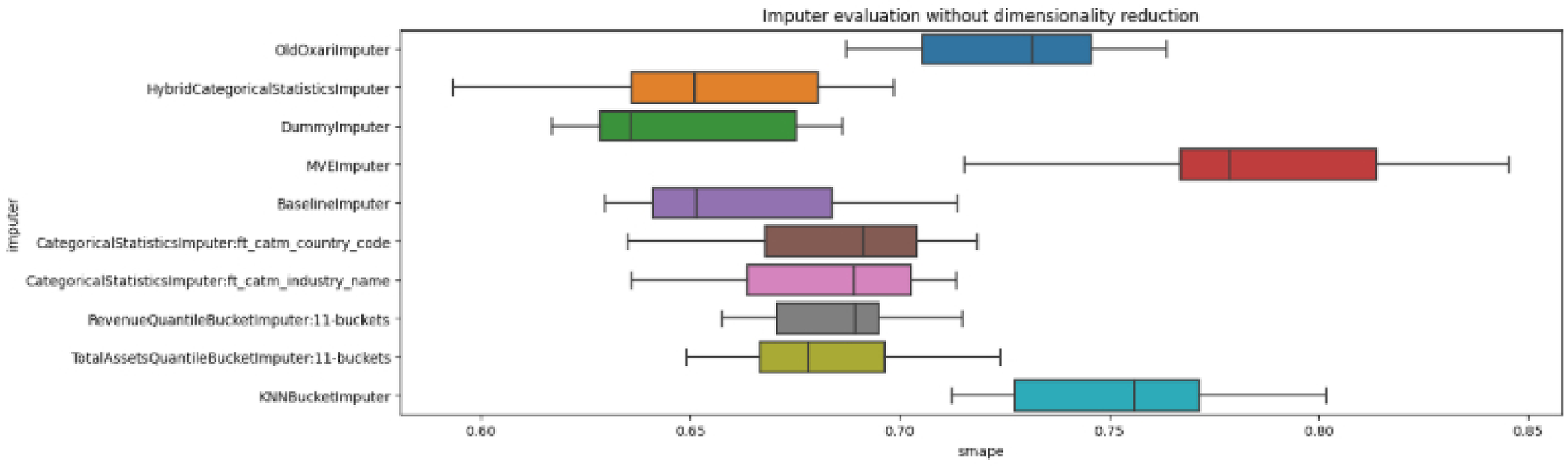
For this experiment, we train an MMA model for each configuration and evaluate the sMAPEvalue on scope 1. We only use 50% of the dataset and ran the experiment for 10 repetitions each. We ran the experiment with a PCAof 40 components and without dimensionality reduction. The PCArun lacked Revenueand ft_catm_industry_name configurations.

Result and Insights

With PCAenabled the CategorycalStatisticsImputerperforms the best. The trends seemsto indicate that more complex approaches yield worse sMAPEresults. Therefore, the MVE imputer performs the worst in this experiment. Afterwards the KNNImputer and then the OldOxariImputer.



The same observation is true without using any feature reduction scheme. The MVEImputer clearly worsens the prediction. Surprisingly even the DummyImputer performs well in this scenario as second best approach. The best imputer appears to be the CategoricalStatisticsImputer which uses the industry_name.



Decision

Update 22.01.2024

It appears weird that the dummy imputer performs so well. We should investigate whether there is a problem in the pipeline when the imputers impute. Other than that results indicate that the categorical features imputer performs the best. Hence, we should use that one in the future. Especiallywith the industry name as a grouping factor. Theseresults might be subject to change after acquiring more feature data from twelve data.