Report for **ID5059: Project 2**

Note: *Supporting Evidence, as mentioned in this report, are the related markdown documents.*

Imputation Section

One of the original client requests was to **evaluate the imputation methods for the data provided**. For reference, note that imputation is the '*act of converting an incomplete sample into a complete sample'* [1] . In order to answer the relevant questions about the performance of three imputation methods, our work focused on constructing a **simulation study**. This consists of three key steps, **1.** *Removing data at random*, **2.** *Choosing the imputation methods and* **3.** *Evaluating performance of the imputation methods.*

Before going into the details of the results, a general overview of the structure of the analysis is necessary. Note that we constructed a framework which gives a lot of flexibility to adjust the analysis as desired and could therefore be expanded if the client wishes to analyse more instances.

Step 1. We chose to analyse two main scenarios:

- Percentage of Data to Drop: {10%, 20%, 30%}
- Types of Data to Drop: {Numerical Only, Categorical Only, Both}

These choices were made to examine if there are specific combinations for which methods perform better than others. In practice these led to identifying methods that could become most appropriate in certain instances. For example, if one method performed best for numerical features and another performed best for categorical features, this could indicate that a hybrid approach is recommended, as we will see.

Step 2. Choosing the models has been dictated by the type of feature. For **numerical features** we chose a *Mean Imputer*, *Predictive Mean Matching* and *missForest* approach. Whilst for **categorical features** we used *Most Frequent Imputer*, *Bayesian Polytomous Regression* and a *missForest* approach. Not only these are the most popular imputation approaches, but they also all have fundamentally different assumptions. By exploring such a diverse range of methods, we wished to account for the possibility of certain assumptions not holding in practice (see *MICE Fitting and Function* and *MissForest Fitting and Function* in the *Supporting Evidence*).

Step 3. Finally, in terms of the measures of performance we focused on metrics which assess the deviation of the imputed values from the true values. In practice these are captured differently across feature types; *accuracy* for the categorical features and a *normalised root mean squared error* (NRMSE) for the numerical features. In simple terms, the *accuracy* captures the proportion of correctly imputed values. Whilst on the other hand, the *normalised root mean squared error* captures the squared error (difference between true and imputed) and normalises it in order to allow for the results to be meaningfully compared across the different scenarios.

After running the simulation for a total of 162 observations, our evaluation of the outcomes highlighted a **two key results**:

⇒ **MissForest is the best performing approach for the modelling of numerical data features** as highlighted by *Figure 1*. This performance is captured by the NRMSE. As

---

[1] Flexible Imputation of Missing Data - Second Edition - Stef Van Buuren

you can see the average of the NRMSE across all scenarios for the Random Forest approach is the smallest (0.87 *see Table 1*)

⇒ **Most Frequent Imputer is the best performing approach for the modelling of categorical data features** as highlighted by *Figure 2*. This is captured by the *accuracy* metric. Higher accuracy indicates better outcomes, and the most frequent imputer has performed best on average (0.65, *see Table 1).*

model	**Mean NRMSE**	**Mean Accuracy**
forest	**0.87**	0.41
mice	1.22	0.54
simple	1.00	**0.65**

Table 1*. Summary Statistics of Imputation Study*

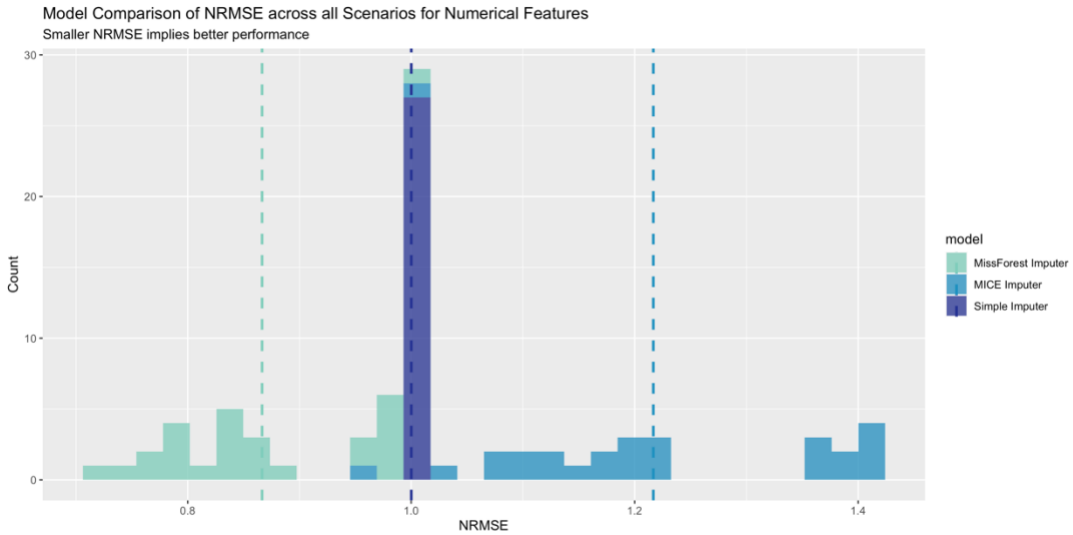

Figure 1*. Histogram plot of the NRMSE comparison across all Numerical Data Columns*

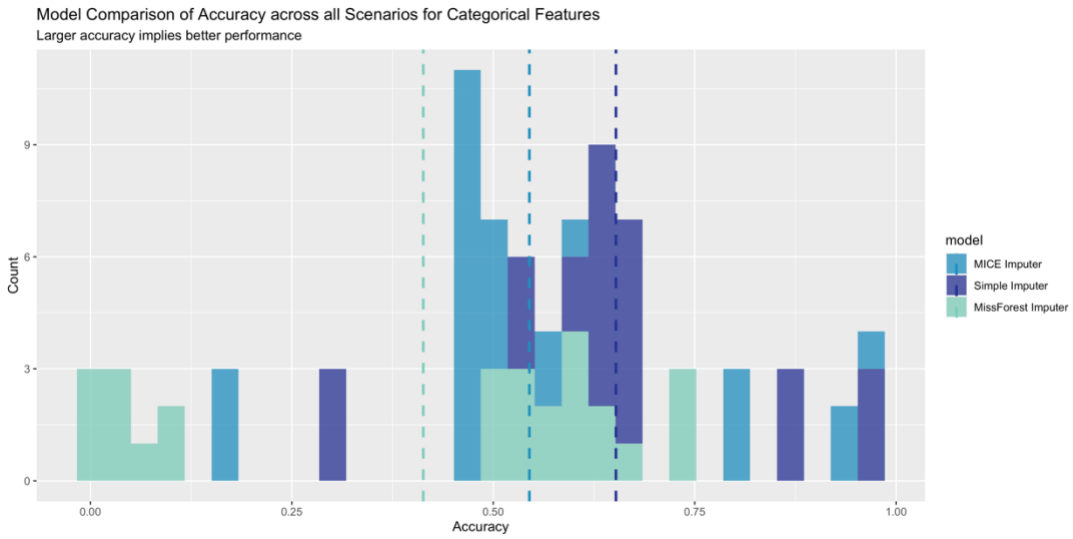

Figure 2*. Histogram plot of the Accuracy comparison across all Categorical Data Columns*

Therefore, our formal recommendation is a **hybrid approach** of the MissForest Imputation for the Numerical Features and a Most Frequent Imputer for the Categorical Features.

Please **note** that a few key simplifying assumptions have been taken in producing these results, and a further in-depth analysis would require additional time (see *Limitations and Areas for Development*, in the Supporting Evidence*)*.

Classification Section

The classification task set by the client required us to explore appropriate modelling approaches to suit the data provided. For reference, by classification we mean the modelling of a binary outcome, or *target*, which is a variable that takes values 0 or 1. Our efforts focused on utilising state-of-the-art unsupervised learning approaches to model all the data *features* which explain our *target* outcome.

Our main objectives to solve this binary classification problem are as follows:

1. Finding the best model predicting the *target*.
2. Identifying the most important *features* of the data.

For **objective 1** we chose and compared three modelling approaches of different nature. Some methods work better than others given certain *data assumptions*, therefore we have chosen a wide range of them to make a more comprehensive analysis as per the client's requirements. For **objective 2** our team focused on identifying the data which explains our *target* the most, so-called *feature importance,* to enable the client to optimise their data collection processes. Moreover, we produced a comparison of classification performance given datasets of smaller size. These comparisons highlight the client could choose to reduce dramatically the size of the data collected and yet retain a strong accuracy performance for the prediction. In the process of achieving these objectives, we assessed the performance of our models using a measure of *accuracy*, namely the proportion of values in the test set that are correctly predicted (as used in the Imputation Section). This is widely considered an efficient measure to assess the performance of model classifiers and has been used throughout this analysis.

1. Finding the best model predicting the *target*.

As per **objective 1**, in order to predict the target, we chose three models: Logistic Regression, Random Forest and the XGBoost Classifier. These models were chosen based on the research that was conducted, to find the most suitable models for the given dataset. All three models were used to predict the outcome for the test dataset. The accuracy scores for each were noted and compared. Based on the first set of scores, the XGBoost Classifier had the highest accuracy score of 88.56% i.e., correctly predicted the target values the most, followed in close second by Random Forest with an accuracy score of 88.22%. These performances were explored further by tuning certain parameters to find the models' optimal configuration. The outcome indicated XGBoost Classifier was still favoured.

2. Identifying the most important *features* of the data.

In accordance with **objective 2**, we ran a thorough analysis on all the features in the dataset to determine which features contributed the most in making a prediction. The initial dataset, which

we refer to as *Dataset 1*, has 30 features which consist of a total of 632 levels. For further explanation of the data preparation process, refer to *Section 2.1: One-Hot-Encoding* of the *Supporting Evidence*.

For the benefit of the client, we carried out a comparison of performance given models have been fitted to datasets of reduced size. The **first analysis** examines the effect of reducing the number of **levels**, whilst the **second analysis** examines the effect of reducing the number of **features**. This choice was made to account for different data collection methods, as it may not be feasible to selectively pick which levels to collect, hence it may be easier to focus on the features themselves. As this information was unknown, we analysed both cases. Note, the data dropped in either case was chosen on the basis of their importance with respect to the *target*.

These processes and the results obtained are explained below.

Model	Dataset used	Number of Levels	Accuracy Score
Random Forest	Dataset 1	632	88.22%
	Dataset 2	434	88.15%
	Dataset 3	312	88.14%
Logistic Regression	Dataset 1	632	87.736%
	Dataset 2	434	87.715%
	Dataset 3	312	87.401%
XGB Classifier	Dataset 1	632	88.56%
	Dataset 2	434	88.54%
	Dataset 3	312	88.479%

Table 2: model comparison with associated accuracy scores

To find out which levels are important, we used a technique using a *decision tree model*. In short, we used this model on the entire dataset and calculated the *importance score*, namely a score that indicates if the level is useful for predicting the target value. Hence, for the **first analysis**, each level that had an *importance score* of 0 was removed. This resulted in 208 levels to be dropped from the dataset, producing *Dataset 2*. To confirm that the levels removed did not heavily impact the effectiveness of our models we ran our models again with the reduced dataset to redetermine the accuracy scores of each. Note, as mentioned in objective 1, we optimised the model configuration (see *Fine-Tuning* in the *Supporting Evidence*) of three models. Once the best configuration was chosen, the aforementioned reduced dataset, *Dataset 2*, was used by all three models to predict the target value again. The resulting accuracy scores decreased by 0.07%, 0.01% and 0.02% respectively, which we deemed reasonable for the trade-off of reducing our level count by a third.

We carried out another size reduction, to check the further effect on the accuracy scores. This time retaining the levels explaining 95% of the *target.* This smaller dataset we refer to as *Dataset 3*, it consists of roughly half of the original data levels. Similar to the second set of accuracy score, there was a slight decrease in each score. However, they were still very close to the initial best accuracy score. Reference *Table 2* for more information on the scores.

This first analysis indicates that the data consists of a lot of levels which affect very little our ability to correctly predict the *target*. We, therefore, have shown the client could reduce significantly the size of their levels data collection (by about half) and still achieve an accuracy of over 88%.

To provide the client with further research on efficient data collection, we conducted the **second analysis** to examine the effect of reducing the number of features on the prediction.

When the 30 features were ordered in decreasing level of importance, we noted that the categorical feature cat16 had the highest relevance to the target value at 0.225 (about 23%). This was followed by other features such as cat5 and cat1 as being the most important features to our predictions. By splitting the features into 6 groups by level of importance we were able to plot our accuracy scores against the number of features we had. See Figure 3 for a visual representation of the results.

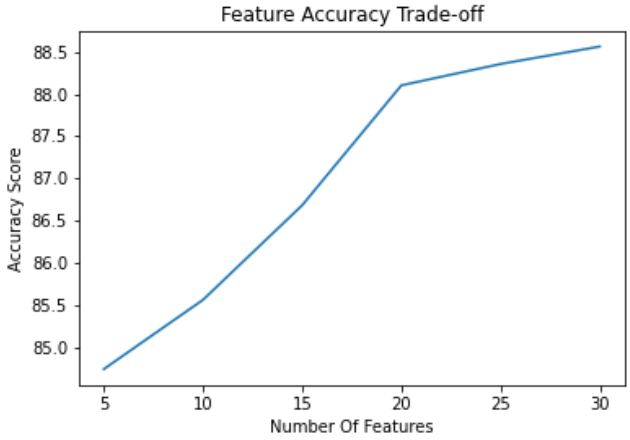

Figure 3: graph representing the feature-accuracy trade-off

This shows that the 20 most important features contribute to an accuracy score of 88.1% and that the 10 least important features only contributed 0.4% more accuracy combined. The latter are highlighted in red, see *Table 3* for reference.

Feature	Importance Score
cat16	0.2251
cont5	0.0686
cont1	0.0516
cont6	0.0497
cont4	0.0470
cont2	0.0466
cont3	0.0458

cont8	0.0448
cont9	0.0433
cont10	0.0426
cont0	0.0422
cont7	0.0417
cat7	0.0287
cat10	0.0260
cat8	0.0241
cat1	0.0227
cat18	0.0213
cat2	0.0174
cat0	0.0169
cat15	0.0157
cat14	0.0119
cat4	0.0117
cat6	0.0108
cat9	0.0093
cat3	0.0090
cat11	0.0088
cat17	0.0070
cat5	0.0046
cat12	0.0025
cat13	0.0023

Table 3: Feature Importance Scores

Conclusion

Through an extensive analysis of the available datasets and the testing of different classifiers, we found that the XGB model performed the best. This model's results proved to be extremely efficient due to its fast processing runtime and minimal maintenance requirements. While other models seemed to be less affected by the loss of information from the reduction in dimensionality (i.e., the levels), the consistency of our XGB classifier, regardless of the size of the dataset, had the best results overall.

By exploring the importance of each feature, we were able to conclude that although every feature played a role in influencing the performance of our model, we noticed that 10 of the 30 features, when combined, only offered a 0.4% improvement in accuracy. These variables are the categorical features cat3-cat6, cat9, cat11-14 and cat17.

We would recommend using this classifier for any future predictions. We would also propose that the aforementioned categorical features could be excluded as the cost of obtaining this type of information does not outweigh any benefit from their inclusion.