Jeff Blackburn

12305379

Data Analytics – CS7DS1

Data Imputation Methods

23rd November 2018

1. Introduction

The aim of this project was to predict a response variable from a dataset with many missing values. The response variable could be anything, but for the nature of this project it has been assumed to be "Does a patient have cancer?". Therefore, the goal is to predict whether a patient has cancer or not using an incomplete dataset.

Working with an incomplete dataset is a very common issue in industry and one that any good data scientist must be able to deal with. There are many different types of missing data, without knowing what the variables in this dataset are it's impossible to decide what type of missing data is present. Knowing the reason for missing data can be very useful when understanding how to deal with. There are four primary types of missing data.

Missing completely at random (MCAR) - whether or not data is missing is completely independent of other information in the data. If it's possible to predict what data will be missing then it is not MCAR. It's rarely accurate to assume data is MCAR.

Missing at random (MAR) - Assumes that we can predict the values of missing data based on the other information present. This is the assumption that will be made during this project, advanced imputation techniques will be used to predict the missing data.

Missing not at random - Data is likely missing for a reason. A common example of this is people with low incomes (or extremely high incomes) are less likely to answer a survey about income.

Structurally Missing - Data is missing for a logical reason. Certain variables may not be applicable to candidates.

A description of the dataset can be found in Table 1. Each continuous X variable mapped to a corresponding categorical Y Value. There was a also a "Group" variable present, this was taken to be gender where Female = 0 and Male = 1.

Variable Name	**Data Type**
Response (Cancer / Not Cancer)	*Categorical*
Group (Gender)	*Categorical*
X1	*Continuous*
X2	*Continuous*
X3	*Continuous*
X4	*Continuous*
X5	*Continuous*
X6	*Continuous*
X7	*Continuous*
Y1	*Categorical*
Y2	*Categorical*
Y3	*Categorical*
Y4	*Categorical*
Y5	*Categorical*
Y6	*Categorical*
Y7	*Categorical*

Table 1. Dataset Description

2. Exploratory Analysis of Data

2.1 Missing Data

Data was assumed to be missing at random. The spread of missing data across variables however was very uneven, please see Figure 1. It can be seen that X3,Y3 and X2,Y2 have significantly more missing values than other variables, with over 40% of the data missing. Whereas, there is a complete dataset for Response, Group and X4,Y4. Figure 1 also demonstrates the patterns of missing data, with red cells denoting missing data and navy cells denoting available data. From this it can be seen that only 44% of entries are complete, whilst 30% are missing both X3,Y3 and X2,Y2.

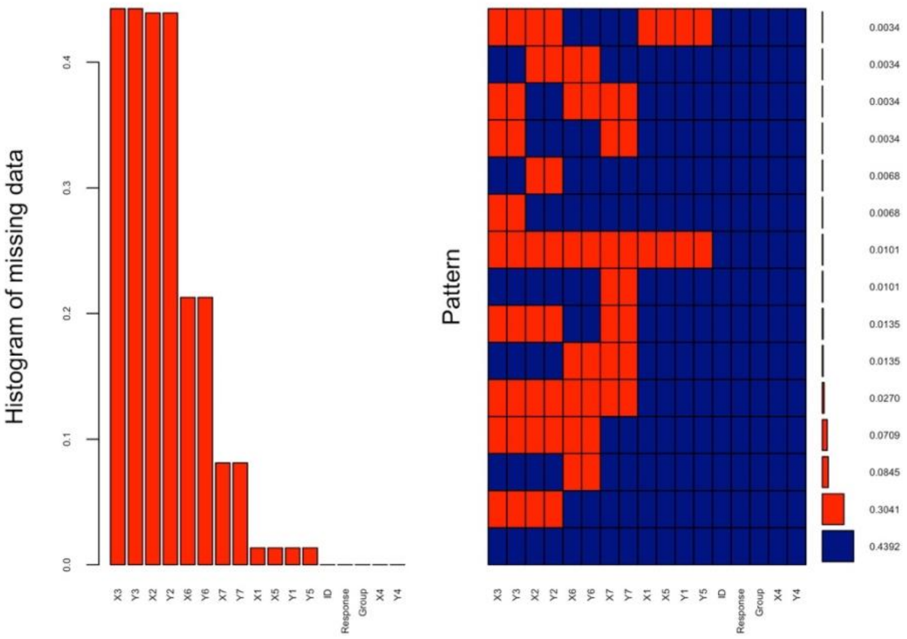
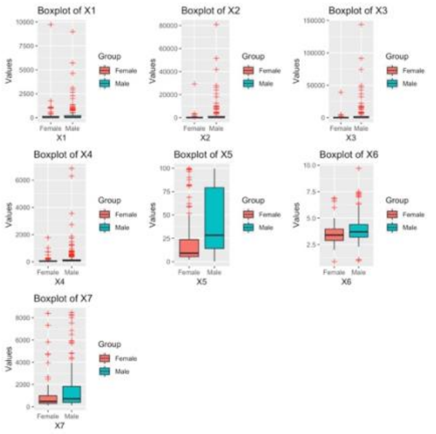


Figure 1. Breakdown of missing values

2.2 Variance & Shape of Data

It can be seen from Figure 2 and Figure 3 that there are a large number of outliers in the dataset. This is important to note as it may cause some problems during prediction, depending on what model is selected. Methods such as top-coding may be useful.



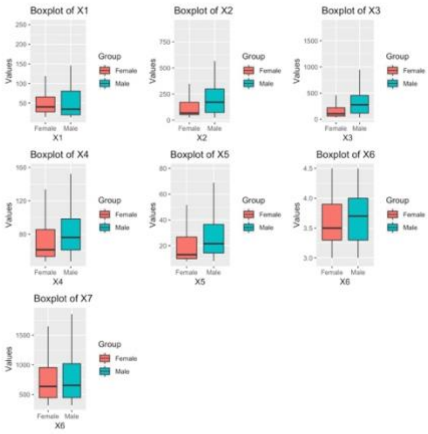


Figure 2. Boxplots with respect to Gender

Figure 3. Boxplots with respect to Gender, not plotting outliers

It is clear from Figure 4, Figure 5 and Figure 6 that a large number of the variables in this dataset are very skewed - X1,X2,X3,X4,X7. The skewedness of each variable is consistent across genders. It is important to note the shape of the distribution when considering imputation. It is generally unwise to use measures such as the mean on heavily skewed datasets with big outliers and it may be prudent to work with the median instead. It is also important to note the different scales that the variables are active on, with X1 ranging from 0-10,000 whereas X6 ranges from 0-10. Standardization is a useful solution for datasets with variables that have different ranges.

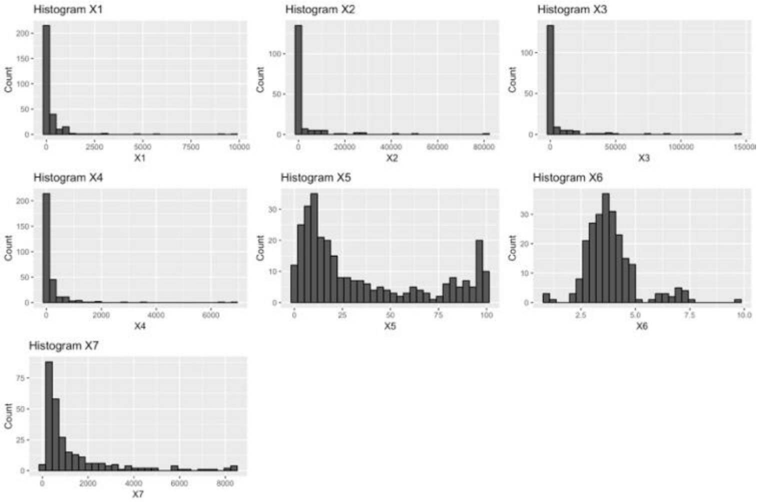


Figure 4. Histogram of X variables

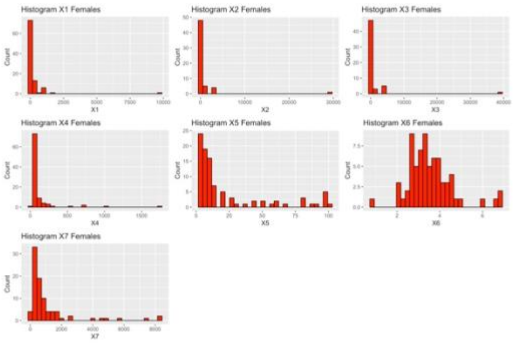


Figure 5. Histogram of X variables for Females

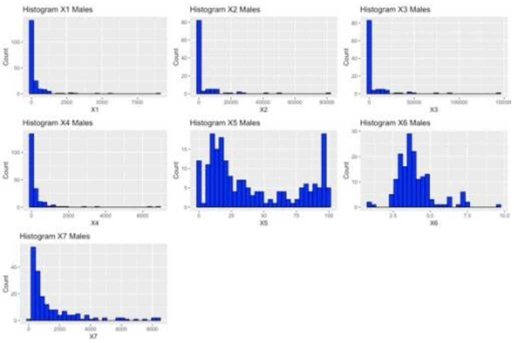


Figure 6. Histogram of X variables for Males

2.3 Relationship of X,Y Pairs

Information had been given with the dataset stating there was a relationship between X,Y pairs, that is Y1 mapped to X1, Y2 mapped to X2 etc. Figure 7 demonstrates that there was a relationship with a clear cut-off point for all pairs apart from X5,Y5. An example of this can be seen with X1,Y1 – once X1 is greater than circa 35, the value of Y1 changes from 0 to 1. This means that all the information contained in the Y variables is held by the X variables. It may be possible to ignore all Y variables apart from Y5 for prediction. Gender appears to have no effect on this relationship.

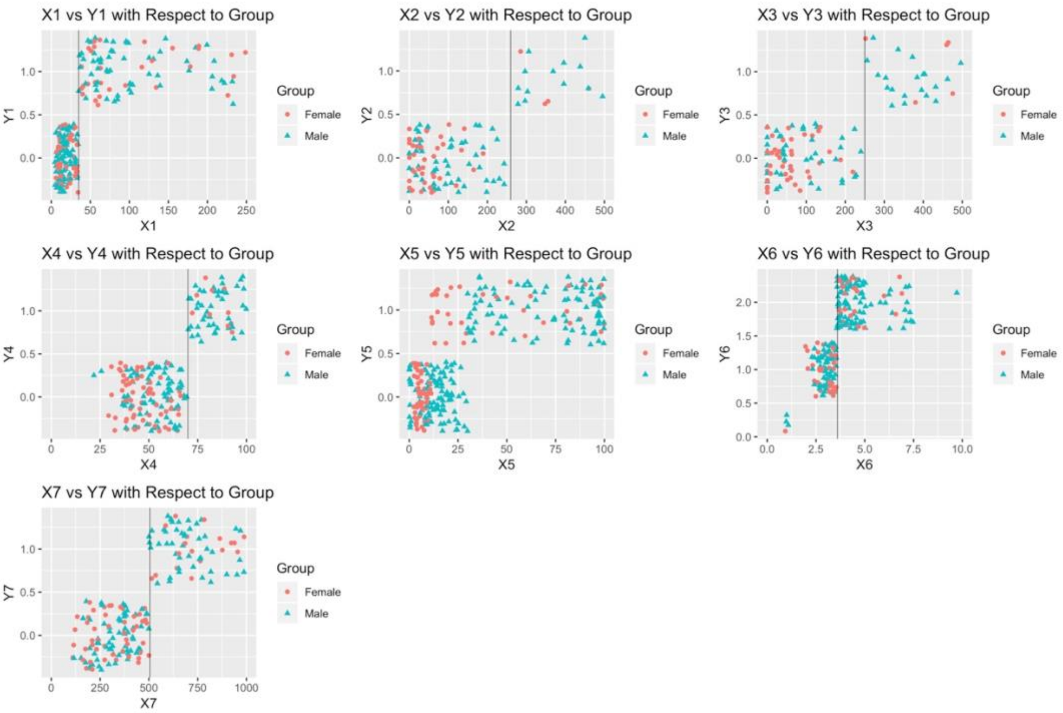


Figure 7. Relationship of X and Y (noise added) variables with appropriate cut-off points

2.4 Correlation of X variables

Figure 8 demonstrates the strong relationship between the X1,X2 and X3 variables, correlation values circa 0.9 and above. This relationship means it may be possible to perform predictions without X2 and X3 reducing the complexity of the model. There is only 296 entries in this dataset, so a reduction in dataset complexity is not a necessity but with a larger dataset it may be useful.

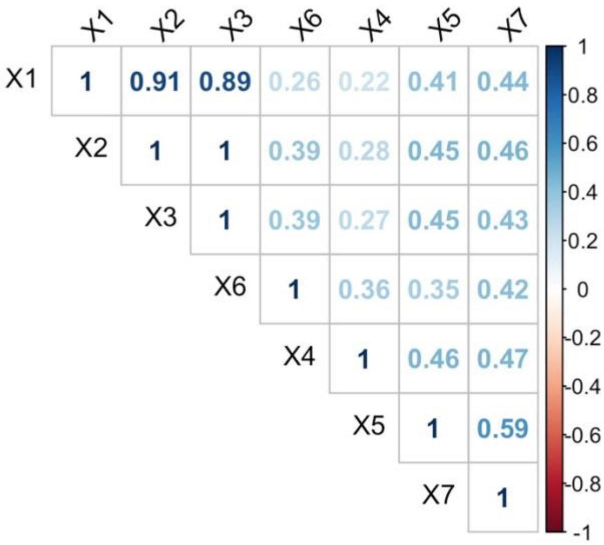


Figure 8. Correlation matrix of X variables

2.5 Response Variable Analysis

It is not clear from simple visualization (Figure 9) which X variable has the strongest relationship with the response variable. The relationships of X and response will now be explored using predictive models.

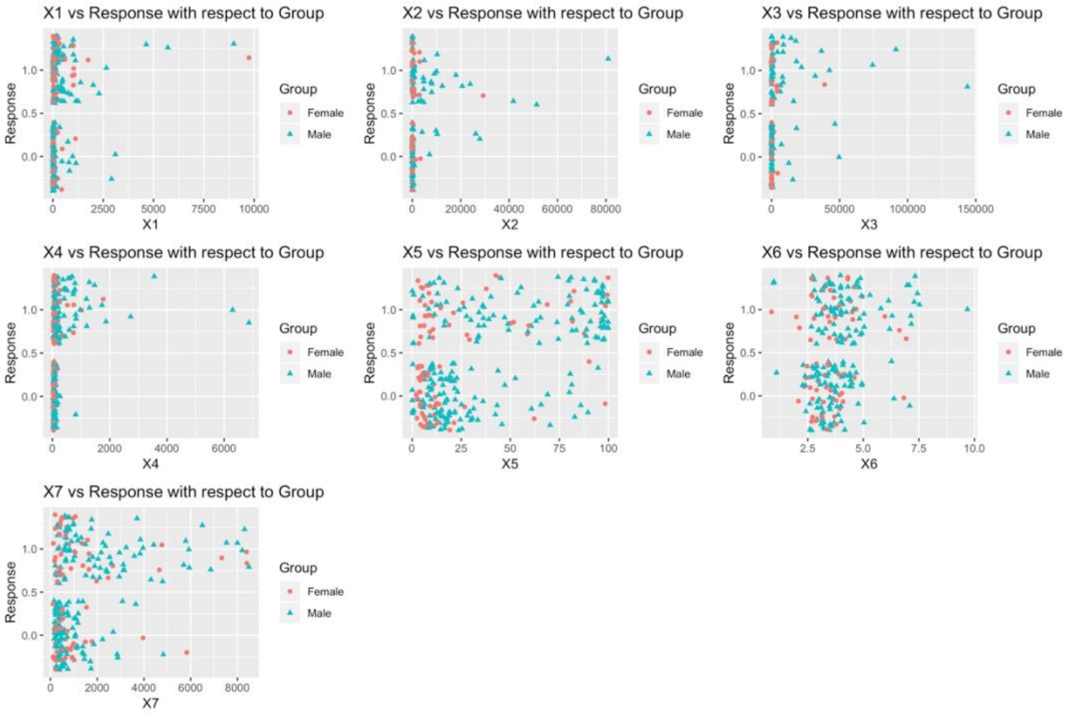


Figure 9. Relationship of X Variables and Response

3. Imputation Methods & Predictive Analytics

Several methods of data imputation with varying complexity will be used to replace the missing data during this project - median imputation, predictive mean matching and k-nearest neighbours. The performance of these imputation methods will be compared to a baseline but also using three predictive models – Decision Tree, Logistic Regression and Random Forest. To reiterate, the goal of these models is to predict whether a patient has cancer or not. The performance of these models will be based on four metrics: accuracy, error, precision and recall.

3.1 Median Imputation

This is a quick approach to replacing missing data values. Similar to working with only complete cases, median imputation can introduce errors and bias to the data. It involves replacing all missing values with the median of the dataset. Median imputation has been used rather than mean imputation due to the skewed nature of the dataset and the large number of outliers.

3.2 MICE - Predictive Mean Matching

The MICE package in R, provides many methods of data imputation. In this project predictive mean matching (PMM) has been used. The imputed values are "borrowed" from the real values. This means that the imputed data will be skewed like the real data and also bounded between the minimum and maximum values of the real data. The performance of this imputation method has not been widely compared to other methods that are more commonly used.

3.3 K-Nearest Neighbour

The K-Nearest Neighbour (KNN) method is based on the assumption that a missing data value can calculated based on the values of the 'k' closest points. The value of k is determined by the user and is a hyperparameter than must be tuned from model to model. It can be used for all different types of data and is a commonly used method.

3.4 Baseline Accuracy

Before imputing any data, it is important to have a baseline to compare results with. The initial baseline will be a decision tree using all available data points. It can be seen in Table 2 that this produced an accuracy of 0.68. Top-coding the outliers in X1,X2,X3,X4 and X7 improved the accuracy of the decision tree to 0.73. Standardization was then performed across all X variables to ensure they were all on the same scale, this improved model accuracy to 0.75. Removing variables X2 and X3 (due to their high correlation with X1) increased accuracy to 0.76. It can be seen that before any imputation of missing variables, it is possible to drastically increase model accuracy – using methods such as top-coding, standardization and feature selection.

Method (Decision Tree)	Accuracy	Error	Recall	Precision
Baseline	0.68	0.32	0.6	0.78
Top-coding	0.73	0.27	0.75	0.7
Top-coding and Standardization	0.75	0.25	0.8	0.69
Top-coding, Standardization and Feature Selection	0.76	0.24	0.83	0.69

Table 2. Model Performance before Data Imputation

3.5 Decision Tree Model

The first model used for comparing different imputation methods was the Decision Tree (as used in the baseline test). The results for different imputation methods can be seen in Table 4. KNN repeatedly performed better than either median or MICE imputation. The best value of k within the KNN test was found to be 3. This also outperformed all methods used during baseline testing with an accuracy of 0.78. This demonstrates that imputation of missing data variables can improve overall predictive performance. Median and MICE imputation performed similarly, however MICE had a higher recall with Median offering a better precision.

Imputation Method	Accuracy	Error	Recall	Precision
KNN imputation (k=3)	0.78	0.22	0.84	0.71
KNN imputation (k=5)	0.73	0.27	0.79	0.64
KNN imputation (k=10)	0.76	0.24	0.8	0.73
Median Imputation	0.73	0.27	0.76	0.7
Mice Imputation	0.73	0.27	0.81	0.64

Table 3. Decision Tree Performance after Data Imputation

3.6 Logistic Regression Model

Logistic regression models performed poorly when compared to the baseline. See Table 4. for results. None of the chosen methods of imputation produced a better model than the decision tree model used in baseline. However, without imputation it is not possible to run a logistic regression model. KNN with a k value of 3 was found to be strongest of the imputation methods use, with MICE imputation and KNN with K=10 proving the weakest. Median imputation produced a high recall value but a very low precision value.

Imputation Method	Accuracy	Error	Recall	Precision
KNN imputation (k=3)	0.74	0.26	0.81	0.66
KNN imputation (k=5)	0.73	0.27	0.82	0.66
KNN imputation (k=10)	0.72	0.28	0.78	0.68
Median Imputation	0.72	0.28	0.86	0.56
Mice Imputation	0.72	0.28	0.79	0.64

Table 4. Logistic Regression Performance after Data Imputation

3.7 Random Forest

All of the random forest imputation methods outperformed or equalled the baseline tests. Again, KNN imputation with a K value of 3 was found to be the best performing method of imputation producing an accuracy of 0.84. The lowest accuracy was found with the MICE imputation of 0.75, matching the baseline.

Imputation Method	Accuracy	Error	Recall	Precision
KNN imputation (k=3)	0.84	0.16	0.88	0.79
KNN imputation (k=5)	0.76	0.24	0.84	0.68
KNN imputation (k=10)	0.8	0.2	0.86	0.73
Median Imputation	0.76	0.24	0.81	0.7
Mice Imputation	0.75	0.25	0.82	0.68

Table 5. Random Forest Performance after Data Imputation

Random forest is a very powerful tool and offers the ability to show the most important variables during prediction through the mean decrease in GINI index. It can be seen in Figures 10 – 14 that the X variables were consistently more important for predicting the response than the Y variables. This was expected at the outset as all information contained in Y was also contained in X. It may be possible to reduce the complexity of a model through feature selection, removing the Y variables. It was also interesting to see that Gender / Group had little to no effect on the response. These figures also show the type of error on the response variable. It can be seen that all of the random forest models had a greater Recall than Precision, this means they were more accurate at predicting the cases of no cancer than the cases with cancer.

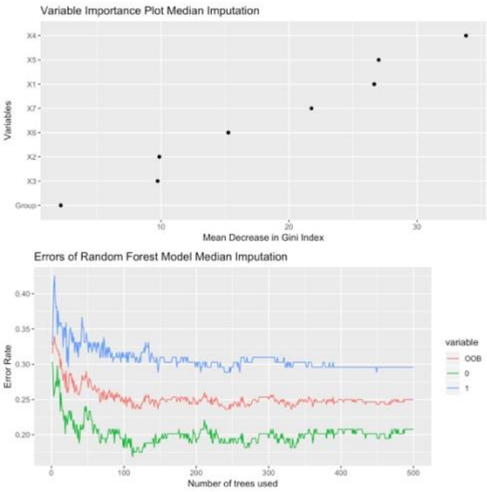


Figure 10. Random Forest Median Imputation

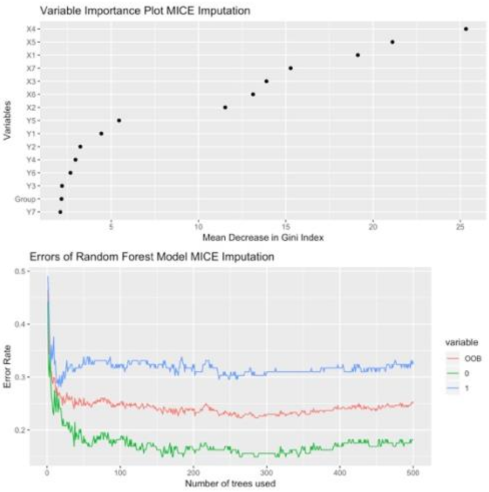


Figure 11. Random Forest MICE - PMM Imputation

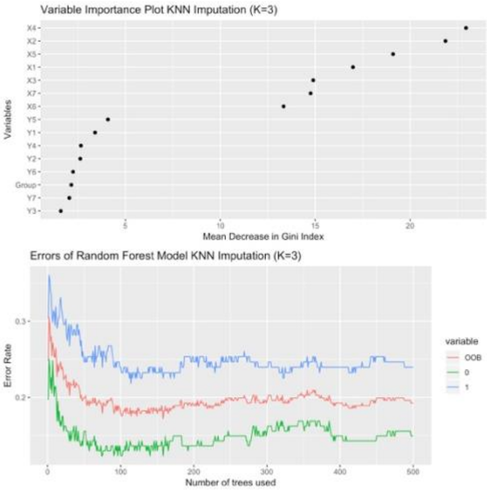


Figure 12. Random Forest K-Nearest Neighbours (K=3)

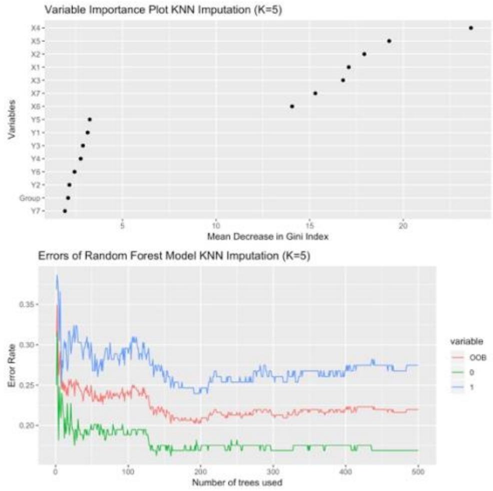


Figure 13. Random Forest K-Nearest Neighbours (K=5)

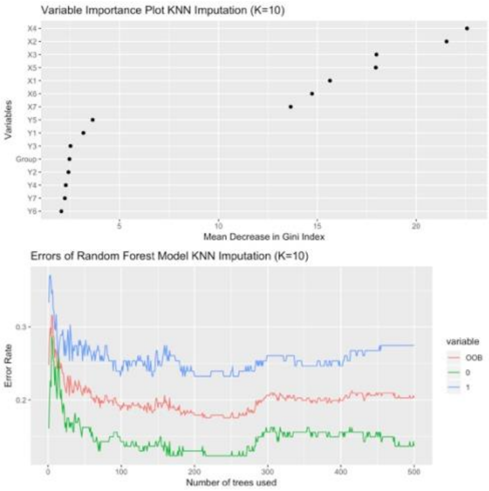


Figure 14. Random Forest K-Nearest Neighbours (K=10)

4. Conclusion

In summary, imputing the missing data values generally led to an increase in the predictive accuracies of the models when compared to the baseline. Pre-processing of data before imputation was a useful method to improve prediction accuracy – methods such as top-coding, standardization and feature selection were particularly useful in this example. These pre-processing measures improved a baseline accuracy from 0.68 to 0.76.

The K-Nearest Neighbours method with a k-value of 3 consistently performed best for this dataset when compared to median and predictive mean matching imputation methods. It is possible that the performance of KNN was due to the size and shape of the dataset. It would be incorrect to assume that it will always outperform other methods of data imputation.

When comparing the prediction models, random forest outperformed the decision tree and logistic regression models across all imputation methods. The random forest achieved a maximum accuracy of 0.84 compared to 0.78 for the decision tree and 0.74 for logistic regression.

For future work with this dataset it would be recommended to attempt to find out what types of missing data are present in this dataset. The assumption was made that the data is Missing at Random, however if this is not the case then it would be irresponsible to use the imputation methods outlined.

5. Further Analysis

5.1 Individual Variable Analysis – Prior to Imputation

Using each individual variable a decision tree model was built and the performance evaluated. Group was then added to each individual variable and another decision tree model built. The predictive performance of these models were measured using k-fold (k=10) cross-validation and the results can be seen in Table 6. The effects of Group aren't conclusive with some increases in performance but also some decreases. The best accuracy was found using X4 to build the decision tree.

Variables	Accuracy	Error	Recall	Precision
X1	0.68	0.32	0.7	0.66
X2	0.56	0.44	0.71	0.39
X3	0.57	0.43	0.81	0.31
X4	0.74	0.26	0.87	0.58
X5	0.72	0.28	0.83	0.61
X6	0.59	0.41	0.77	0.4
X7	0.63	0.37	0.73	0.51
Y1	0.71	0.29	0.68	0.72
Y2	0.62	0.38	0.89	0.3
Y3	0.6	0.4	0.83	0.25
Y4	0.67	0.33	0.64	0.67
Y5	0.72	0.28	0.74	0.68
Y6	0.57	0.43	0.44	0.72
Y7	0.64	0.36	0.52	0.78
Group & X1	0.72	0.28	0.8	0.63
Group & X2	0.59	0.41	0.82	0.35
Group & X3	0.57	0.43	0.8	0.31
Group & X4	0.75	0.25	0.89	0.59
Group & X5	0.72	0.28	0.8	0.62
Group & X6	0.6	0.4	0.8	0.39
Group & X7	0.63	0.37	0.68	0.6
Group & Y1	0.71	0.29	0.68	0.72
Group & Y2	0.62	0.38	0.9	0.31
Group & Y3	0.61	0.39	0.55	0.68
Group & Y4	0.67	0.33	0.64	0.69
Group & Y5	0.72	0.28	0.75	0.69
Group & Y6	0.6	0.4	0.54	0.67
Group & Y7	0.65	0.35	0.56	0.75

Table 6. Evaluation of individual variables using Decision Tree

5.2 Individual Variable Analysis – After Imputation

Using the KNN method (K=3) the missing variables in the dataset were imputed. A decision tree model was then built for each individual variable with and without group as before. The predictive performance of these models were measured using k-fold (k=10) cross-validation and the results can be seen in Table 7. It can be seen that in general imputing the data improved the overall performance of the model. The average change in accuracy was +0.04 within a corresponding decrease in error.

Recall was not found to change however there was a large change of +0.1 in the precision. As expected the use of individual variables does not get close to the performance power of the random forest model used in previous sections.

Variables	Accuracy	Change in Accuracy	Error	Change in Error	Recall	Change in Recall	Precision	Change in Precision
X1	0.68	0	0.32	0	0.74	0.04	0.63	-0.03
X2	0.79	0.23	0.21	-0.23	0.79	0.08	0.8	0.41
X3	0.72	0.15	0.28	-0.15	0.73	-0.08	0.72	0.41
X4	0.75	0.01	0.25	-0.01	0.88	0.01	0.6	0.02
X5	0.73	0.01	0.7	0.42	0.85	0.02	0.61	0
X6	0.62	0.03	0.38	-0.03	0.75	-0.02	0.51	0.11
X7	0.68	0.05	0.32	-0.05	0.74	0.01	0.6	0.09
Y1	0.71	0	0.29	0	0.67	-0.01	0.72	0
Y2	0.69	0.07	0.31	-0.07	0.88	-0.01	0.49	0.19
Y3	0.69	0.09	0.31	-0.09	0.75	-0.08	0.64	0.39
Y4	0.67	0	0.33	0	0.66	0.02	0.69	0.02
Y5	0.72	0	0.28	0	0.76	0.02	0.7	0.02
Y6	0.62	0.05	0.38	-0.05	0.56	0.12	0.69	-0.03
Y7	0.66	0.02	0.34	-0.02	0.55	0.03	0.77	-0.01
Group & X1	0.67	-0.05	0.33	0.05	0.71	-0.09	0.62	-0.01
Group & X2	0.77	0.18	0.23	-0.18	0.77	-0.05	0.79	0.44
Group & X3	0.72	0.15	0.28	-0.15	0.67	-0.13	0.77	0.46
Group & X4	0.74	-0.01	0.26	0.01	0.87	-0.02	0.6	0.01
Group & X5	0.72	0	0.28	0	0.8	0	0.63	0.01
Group & X6	0.63	0.03	0.37	-0.03	0.74	-0.06	0.52	0.13
Group & X7	0.66	0.03	0.34	-0.03	0.67	-0.01	0.65	0.05
Group & Y1	0.71	0	0.29	0	0.69	0.01	0.75	0.03
Group & Y2	0.69	0.07	0.31	-0.07	0.89	-0.01	0.49	0.18
Group & Y3	0.69	0.08	0.31	-0.08	0.75	0.2	0.62	-0.06
Group & Y4	0.67	0	0.33	0	0.65	0.01	0.69	0
Group & Y5	0.72	0	0.28	0	0.75	0	0.68	-0.01
Group & Y6	0.62	0.02	0.38	-0.02	0.56	0.02	0.67	0
Group & Y7	0.66	0.01	0.34	-0.01	0.56	0	0.76	0.01
Average		0.04		-0.04		0.00		0.10

Table 7. Evaluation of individual variables using Decision Tree after data imputation