Applications of Machine Learning in Imputation

Methodology

2019

Contents

Chapter 1

Introduction

Editing and imputation are both methods of data processing. Editing refers to the detection and correction of errors in the data, whilst imputation is a method of correcting errors in a dataset (Ton de Waal, 2011). This document presents findings from work carried out at the Office for National Statistics on the use of machine learning in imputation. The chapters address the following questions:

1) What is imputation?

2) What is machine learning?

3) Why use machine learning?

4) How XGBoost works?

5) Methods used for the investigation

6) Results of the investigation

7) Conclusions and future direction

Chapter 2

What is Imputation?

Editing and imputing are both methods of data processing. Editing refers to the detection and correction of errors in the data. Whilst imputation is a method of correcting errors and estimating and filling in missing values in a dataset. Where there are errors in the dataset, and when these are considered to add no value in the correction process, these values are set to missing and are imputed with a plausible estimate. Alternatively, missing values may already exist in the data, and imputation may be carried out to produce a complete dataset for analysis.

This research project evaluated the use of machine learning methods for imputation. In order to provide a context for using machine learning in the imputation process, the reader is presented with:

- A rationale for carrying out imputation
- An introduction to the methods of imputation

2.1 Why is imputation carried out?

Missingness and erroneous values can impact the quality of data. A large volume of incorrect and/or missing values increase the risk of the product failing to capture the target concept or target population. That is, ommissions (introduced in collection or processing) may result in certain sub-groups of the target population from being excluded in the analysis dataset, and in turn increasing the risk of biased estimates, reducing the power of inferential statistics and increasing the uncertainty of estimates and inferences derived from the data. Similarly, errors in a dataset may impact the degree to which the final estimate or output represents the reality it was designed to capture.

Correcting erroneous responses and filling in missing values helps manage the quality of data. A complete dataset can improve the accuracy and reliability of estimates, and help maintain the consistency of counts across published tables. Moreover, a dataset with fewer errors and more units may more accuately capture the underlying distribution of the variable of interest. Selecting a method for estimating values in a datset is generally advised by the nature of errors or missingness in the data, and the output desired from the analysis dataset.

2.2 Methods

An imputation process of a dataset can be broken down into the following three phases:

1) Review, whereby data is examined for potential problems; identifying instances of missingness and erroneous values

2) Select, whereby data is identified for further treatment. Of the potential problems identified in the review phase, a method is applied to determine which of these erroneous or missing cases need to be treated
3) Amend, whereby changes are applied to the data identified in the select phase by either correcting errors/ filling in missing values

The focus of this project was in applying Machine Learning methods to amend values in a dataset. That is, it was of interest to compare existing approaches, of treating missing or erroneous values by estimating replacement figures, to machine learning methods. Methods of variable amendment can be grouped into one of the following categories:

- interactive treatment
- deductive imputation
- model based imputation
- donor based imputation

The mechanisms for a given imputation method could be deterministic or stochastic. The former refers to instances where repeated trials of the same method yield identical output. Whereas the latter refers to instances where there is element of randomness; repeated iterations will produce different results.

2.2.1 Interactive treatment

Interactive treatment refers to a class of methods whereby the data are adjusted by a human editor by either re-contacting the respondent/ data provider, replacing values from another variable/ data source, or creating a value based on subject matter expertise.

2.2.2 Deductive imputation

Deductive imputation uses logic or an understanding about the relationship between variables and units to fill in missing values. Examples include deriving a value as a function of other values, adopting a value from a related unit, and adopting a value from an earlier time point. Generally, this method is used when the true value can be derived with certainty or with a very high probability.

2.2.3 Model based imputation

Model based imputation refers to a class of methods that estimate missing values using assumptions about the distribution of the data, which include mean and median imputation. Or assumptions about the relationship between auxiliary variables (or x variables) and the target y variable to predict missing values.

2.2.4 Donor based imputation

Donor based imputation adopt values from an observed unit, which are then used for the missing unit. For each recipient with a missing value for variable y, a donor is identified that is similar to the recipient with respect to certain background characteristics (often referred to as matching variables) that are related to the target variable y. Such methods are relatively easy to apply when there are several related missing values in one record, and if the intention is to preserve the relationship between variables.

Chapter 3

What is Machine Learning?

Machine learning is the field of study that enables a program to learn from its experience of iterating through a task multiple times. A performance measure is generally specified by the programmer, which is used to evaluate how well the machine has carried out the task at each iteration. Learning of the task is evidenced by its improvement against the performance measure.
The different types of machine learning systems can be categorised with respect to:

- Whether or not they are trained with human supervision
- Whether or not they can learn incrementally or on the fly
- Whether they work by comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model

3.1 Supervision

Machine learning systems can vary with regards to the degree of supervision. The major types of supervision:

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

3.1.1 Supervised learning

Supervised learning is the specification of the desired output. That is, the data used to train the model includes the solutions (which are referred to as labels), which the machine learning system attempts to estimate. The desired solutions specified in the machine learning algorithm are referred to as labels.

3.1.2 Unsupervised learning

Unsupervised learning uses training data that is unlabelled. In this class of machine learning systems, the outcome/ desired solutions are not specified in the machine learning algorithm.

3.1.3 Semi-supervised learning

Machine learning systems that use partially labelled data are categorised as utilising semi-supervised learning.

3.1.4 Reinforcement learning

Reinforcement learning involves the use of rewards or penalties to train the machine in identifying the appropriate action for a given situation. The learning system, which is referred to as an agent, observes the environment, selects and performs actions, and gets a response in the form of a reward or penalty. After multiple iterations, it identifies the best strategy, referred to as a policy, that results in the most reward over time.

3.2 Batch and Online learning

Another criterion for classifying machine learning systems is the way in which the algorithm learns. That is, whether the learning takes place at once or if it happens in increments.

3.2.1 Batch learning

Batch learning uses all the available data to train the machine learning system. This is generally time consuming and computationally expensive, and as a result is carried out offline. Whilst in production, the system is no longer learning, and is simply applying what it has learnt from the full set of training data.

Any changes to the data generating mechanism (GDM) will mean that a new system would need to be trained, from scratch on the full set of data (that includes data points before and after changes to the GDM).

3.2.2 Online learning

Online learning trains the system incrementally through sequential input of data. Data can be delivered individually or in small groups, referred to as mini-batches. As each learning step is relatively fast and cheap, the system can learn about new data whilst in production, as it arrives. It is an ideal approach for when the velocity of new data is high, and when there is a need to adapt to changes rapidly or autonomously.

3.3 Approaches to generalisation

Machine learning systems can also be categorised with regards to how the systems generalise. That is, there are different approaches to using the training data to develop a system that can then be generalised to new cases. The two main approaches are instance-based learning and model-based learning.

3.3.1 Instance-based learning

Instance-based learning identifies all instances of a given feature in the training data and uses a similarity measure to generalise to new cases.

3.3.2 Model-based learning

Model-based learning uses features in the training data to predict the outcome/ variable of interest; the model used to specify the relationship between the predictor(s) and outcome(s) are then generalised on new cases.

Chapter 4

Why use Machine Learning?

It was of interest to explore the utility of Machine Learning to directly impute for missing values in datasets. More specifically, the Methods Division was interested in examining whether Machine Learning models can improve the timeliness, reliability and accuracy of the imputation process in social survey data. Figure 1 presents the imputation pipeline for social survey data. Prior to imputation, units and values are reviewed, and those that are missing and should be routed to the item in question, are selected (i.e. flagged) for imputation. Data is then further processed by the Social Survey Division before imputed and observed data are compiled in an analysis dataset, used for publishing Official Statistics estimates.

The intention was to use a machine learning system to impute flagged missing values. This model based approach for imputation may reduce the data processing time and improve the precision and reduce the variance of estimates. The current approach, which utilises nearest neighbour donor imputation involves the following:

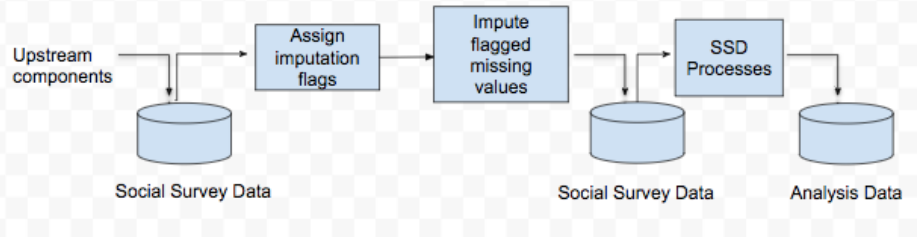

Figure 4.1: Figure 1. Imputation pipeline in social survey data.

- Setting up specification files for each variable and imputation group combination
- Iterating through weights for matching variables so that all missing values are imputed

Designing the selection criteria for donors can be time consuming as it requires analysts to identify matching variables (MV), along with weights for each MV. Teams currently use subject matter expertise in designing the donor imputation strategy for each variable. As this process is not data driven, it introduces an element of subjectivity and does not guarantee that matching variables selected are the best predictors of the variable of interest. In contrast, a data driven approach would be reproducible and identify the best predictors, in the dataset, to estimate missing values. Moreover, applying the machine learning system may offer a more parsimonious approach as fewer input paramters and files would be required in executing imputation.

The Methodology Division was interested in whether a Machine Learning System would perform better compared to the current imputation process with regards to:

- Timeliness: Would the ML system reduce processing time and by how much?
- Accuracy & Reliability: How do the two methods compare with respect to the bias and variance of estimates?
- Interpretability: What advantages and challenges do the ML system present with regards to making the imputation methods transparent?

At present, the following Machine Learning library was used in the investigation:

- XGBoost

Chapter 5

XGBoost

XGBoost is a set of open source functions and steps, referred to as a library, that use supervised ML where analysts specify an outcome to be estimated/ predicted. The XGBoost library uses multiple decision trees to predict an outcome.

The ML system is trained using batch learning and generalised through a model based approach. It uses all available data to construct a model that specifies the relationship between the predictor and outcome variables, which are then generalised to the test dataset.

XGBoost stands for eXtreme Gradient Boosting. The word "extreme" reflects its goal to push the limit of computational resources. Whereas gradient boosting is a machine learning technique for regression and classification problems that optimises a collection of weak prediction models in an attempt to build an accurate and reliable predictor.

In order to build a better understanding of how XGBoost works, the documentation will briefly review:

- Decision trees: How decision trees play a role in XGBoost?
- Boosting: What is it?

The final section of this chapter provides a step by step guide on building models using XGBoost; the reader is encouraged to use this code to predict an outcome variable using available auxiliary variables.

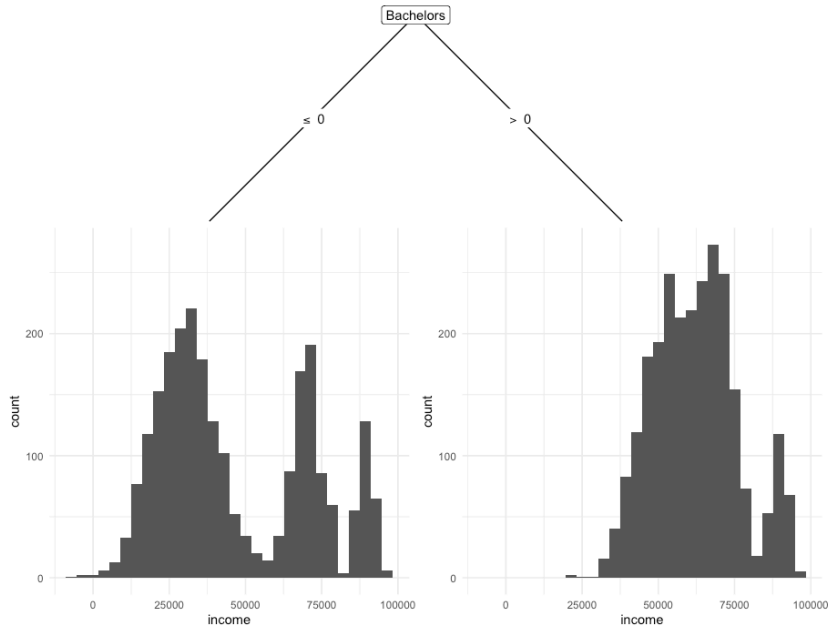

Figure 5.1: Figure 2. Decision tree that splits units in a dataset based on whether individual has a Bachelor's degree or not, in order to predict Income. The tree shows that those with a Bachelors degree (> 0) on average earn more than than those wihtout a Bachelor's degree (< 0).

5.1 Decision trees

Decision trees can be used as a method for grouping units in a dataset by asking questions, such as "Does an individual have a Bachelor's degree?". In this example, two groups would be created; one for those with a Bachelor's degree, and one for those without. Figure 2 provides a visual depiction of this grouping in an attempt to explain Income.

Each subsequent question in a decision tree will produce a smaller group of units. This grouping is carried out to identify units with similar characteristics with respect to an outcome variable. The model in Figure 3 attempts to use University qualifications to predict Income.

The following characteristics are true of decision trees:

- A single question is asked at each decision node, and there are only two possible choices. With the example in Figure 3, the questions include 1) Does individual have a pHD, 2) Does individual have a Master's and 3)

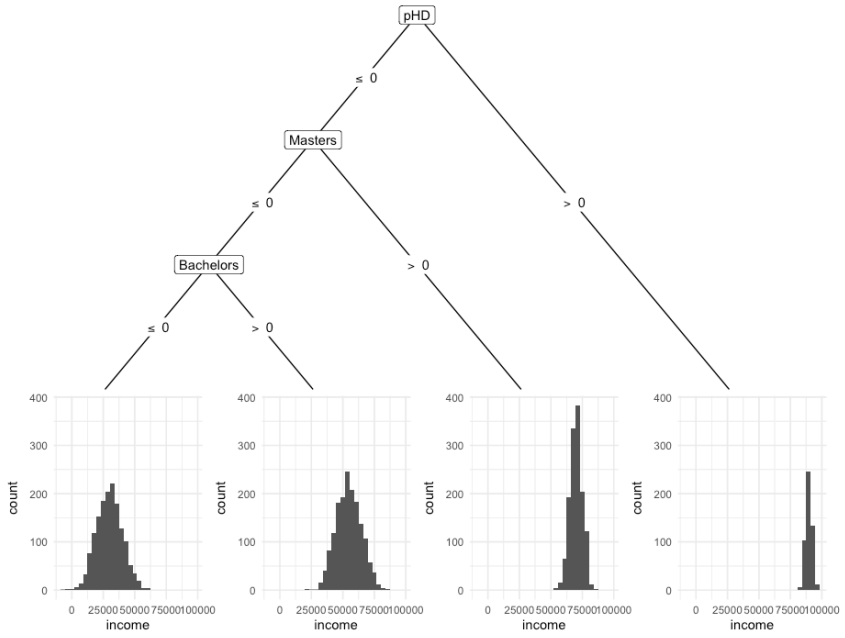

Figure 5.2: Figure 3. Decision tree that splits units in a dataset based on whether individual has a Bachelor's degree (yes/no), a Master's degree and pHD (yes/no), in order to predict Income. The tree shows that those with a higher qualification tend to earn more.

Does individual have a Bachelor's degree.

- At the bottom of every decision tree, there is a single possible decision. Every possible decision will eventually lead to a choice. Some decisions will lead to a choice sooner. The model in Figure 3 attempts to predict Income using University Qualifications. Each node presents a question about whether an individual possesses a given qualification. The end nodes present the distribution of income for individuals with the specified qualifications. As a result, the choices would be the expected value of Income for an individual, given the qualificaitons obtained.

Decision trees are a learning method that involve a tree like graph to model either continuous or categorical choice given some data. It is composed of a series of binary questions, which when answered in succession yield a prediciton about data at hand. XGBoost uses Classification and Regression Trees (CART), which are presented in the above examples, to predict the outcome variable.

5.1.1 Boosting

A single decision tree is considered a weak/ base learner as it only slightly better than chance at predicting the outcome variable. Whereas strong learners are any algorithm that can be tuned to achieve peak performance for supervised learning. XGBoost uses decision trees as base learners; combining many weak learners to form a strong learner. As a result it is referred to as an ensemble learning method; using the output of many models (i.e. trees) in the final prediction.

The concept of combining many weak learners to produce a strong learner is referred as boosting. XGBoost will iteratively build a set of weak models on subsets of the data; weighting each weak prediction according to the weak learner's performance. A prediction is derived by taking the weighted sum of all base learners.

5.1.2 Building models with XGBoost

In the training data, a target variable y_i is specified, whilst all other features x_i are used as predictors of the target variable. A collection of decision trees are used to predict values of y_i using x_i. Individually, each decision tree, would be a weak predictor of the outcome variable. However, as a collective, the decision trees may enable analysts to make accurate and reliable predictions of y_i. As a result, the method for predicting the target variable using x_i in XGBoost is referred to as decision tree ensembles. The steps below demonstrate how XGBoost was used to build a model, to predict income, using Univeristy Qualifications.

1) Load the following packages

```
library(caret)

library(xgboost)
```

2) Load the dataset and remove the identifer

```
# Load data
load("data/Income_tree.RData")

# Remove identifier
Income <- Income[,-1]
```

3) Split the dataset into training and test

```
# Split data into training and test
set.seed(5)

s <- createDataPartition(Income$income, p = 0.8, list=FALSE)

training <- Income[s,]

test <- Income[-s,]
```

4) Convert the data into DMatrix objects, which is the recommended input type for xgboost

```
# Convert the data to matrix and assign output variable
train.outcome <- training$income

train.predictors <- sparse.model.matrix(income ~ .,
                                        data = training
)[, -1]

test.outcome <- test$income

test.predictors <- model.matrix(income ~ .,
                                data = test
)[, -1]

# Convert the matrix objects to DMatrix objects
dtrain <- xgb.DMatrix(train.predictors, label=train.outcome)

dtest <- xgb.DMatrix(test.predictors)
```

5) Train the model

```
# Train the model
model <- xgboost(
  data = dtrain, max_depth = 2, eta = 1, nthread = 2, nrounds = 10,
  objective = "reg:linear")
```

6) Test the model

```
# Test the model
pred <- predict(model, dtest)

# Evaluate the performance of model
RMSE(pred,test.outcome)
```

7) Examine the importance of each feature in the model

```
# Examine feature importance
importance_matrix <- xgb.importance(model = model)

print(importance_matrix)

xgb.plot.importance(importance_matrix = importance_matrix)
```

8) Plot the individual trees in the model

```
# Plot the trees
# Tree 1
xgb.plot.tree(model = model, tree=0)
# Tree 2
xgb.plot.tree(model = model, tree=1)
# Tree 3
xgb.plot.tree(model = model, tree=2)
```

Chapter 6

Methods

The project evaluated the machine learning methods using:

1) The Census Teaching File, an open dataset containing 1% of the person records from the 2011 Census in England & Wales.

2) Survey Data

6.1 Census Teaching File

The Census Teaching File was downloaded from the ONS website as a CSV file named "CensusTeachingFile", and was read into R using the following line of code. The dataset consisted of 569,741 individuals and 18 categorical variables from the 2011 Census population.

```
# Read CSV into R
CensusRaw <- read.csv(
  file = "Data/CensusTeachingFile.csv", skip = 1,
  header = TRUE, sep = ","
)
```

The code below specifies the packages used in the preparation, study and build of machine learning systems using the Census Teaching File.

```
library(tidyverse)
library(mice)
library(reshape2)
library(GGally)
```

```
library(Matrix)
library(xgboost)
library(caret)
library(DiagrammeR)
library(MLmetrics)
library(rpart)
library(scales)
library(knitr)
library(kableExtra)
library(DescTools)
```

The steps below present the methods used to compare the performance of model based imputation (using XGBoost) with that of donor based imputation (using CANCEIS). Code chunks are provided to demonstrate how each step was carried out; with the imputable variable, economic activity, as the example.

1) Variables in the dataset were renamed and recoded so that:

- Variable names were consistent with Google's R style guide
- The response categories for all variables were numeric

```
# Rename variables
Census <- plyr::rename(CensusRaw, c(
  "Person.ID" = "person.id",
  "Region" = "region",
  "Residence.Type" = "residence.type",
  "Family.Composition" = "fam.comp",
  "Population.Base" = "resident.type",
  "Sex"="sex",
  "Age"="age",
  "Marital.Status" = "marital.status",
  "Student"="student",
  "Country.of.Birth" = "birth.country",
  "Health"="health",
  "Ethnic.Group" = "ethnicity",
  "Religion"="religion",
  "Economic.Activity" = "econ.act",
  "Occupation" = "occupation",
  "Industry" = "industry",
  "Hours.worked.per.week" = "hours.worked",
  "Approximated.Social.Grade" = "social.grade"
))

# Recode variables (dataset is mutated in order to recode variables)
```

```
Census <- Census %>% mutate_if(is.factor, as.character)

# Recode the Region variable so that it is numeric
Census$region[Census$region == "E12000001"] <- 1
Census$region[Census$region == "E12000002"] <- 2
Census$region[Census$region == "E12000003"] <- 3
Census$region[Census$region == "E12000004"] <- 4
Census$region[Census$region == "E12000005"] <- 5
Census$region[Census$region == "E12000006"] <- 6
Census$region[Census$region == "E12000007"] <- 7
Census$region[Census$region == "E12000008"] <- 8
Census$region[Census$region == "E12000009"] <- 9
Census$region[Census$region == "W92000004"] <- 10

Census$residence.type[Census$residence.type == "C"] <- 1
Census$residence.type[Census$residence.type == "H"] <- 2

Census$student[Census$student == 1] <- 0
Census$student[Census$student == 2] <- 1

Census <- Census %>% mutate_if(is.character, as.numeric)

Census$person.id <- as.character(Census$person.id)

Ht <- table(Census$hours.worked)
Census$hours.cont <- ifelse(Census$hours.worked == 1, runif(
  1:Ht[names(Ht) == 1],
  1, 15
),
ifelse(Census$hours.worked == 2, runif(1:Ht[names(Ht) == 2], 16, 30),
  ifelse(Census$hours.worked == 3, runif(1:Ht[names(Ht) == 3], 31, 48),
    ifelse(Census$hours.worked == 4, runif(1:Ht[names(Ht) == 4], 49, 60),
      Census$hours.worked
    )
  )
)
)

save(Census, file = "data/Census.Rda")
```

A preview of the dataset is provided below.

person.id

region

residence.type

fam.comp

resident.type

sex

age

marital.status

student

birth.country

health

ethnicity

religion

econ.act

occupation

industry

hours.worked

social.grade

hours.cont

7394816

1

2

2

1

2

6

2

1

1

2

1

2

5

8

2
-9
4
-9.00000
7394745
1
2
5
1
1
4
1
1
1
1
1
2
1
8
6
4
3
56.92578
7395066
1
2
3
1
2
4
1

1

1

1

1

1

1

6

11

3

4

36.79397

7395329

1

2

3

1

2

2

1

1

1

2

1

2

1

7

7

3

2

36.28146

7394712

1
2
3
1
1
5
4
1
1
1
1
2
1
1
4
3
2
41.81308
7394750
1
2
2
1
1
6
2
1
1
2
1
1

1

9

2

3

3

39.98968

2) For the purposes of training and testing a machine learning system, the data was divided into training and test datasets using the following code.

```
# Randomly select 80% of Census units and split into Train and Test data
set.seed(5)
Census80 <- sample(1:nrow(Census), 0.8 * nrow(Census), replace = FALSE)
Census20 <- setdiff(1:nrow(Census), Census80)

Census.train <- Census[Census80, ]
Census.test <- Census[Census20, ]

save(Census.train, file = "data/Census.train.Rda")
save(Census.test, file = "data/Census.test.Rda")
```

The intention was to build models to predict a selection of variable using training data, which had no missingness. This model would then be evaluated with respect to its accuracy and generalisability using a test dataset, which would have missingness. The Census Teaching File was a complete dataset. As a result, missingness was simulated in the test dataset, and the imputation models (derived for each variable) were evaluated with regards to how well they predicted the true values.

Models were tested for the following variables:
- Economic activity (a multi-class variable) - Hours worked (a derived continuous variable)
- Social Grade (a multi-class variable)
- Student status (a binary variable) A more detailed description of all the variables can be found here.

3) The distribution of the imputable variable was studied

```
# Study the variable: How many units in each category?
EAt <- table(Census$econ.act)

EAt
```

```
# What is the distribution of the variable: Remove NCR and plot to look at distribution
g <- ggplot(Census[!Census$econ.act == -9, ], aes(econ.act))

g + geom_bar() + scale_x_discrete(
  name = "Economic Activity",
  breaks = pretty_breaks()
)
```

4) The dataset was cleaned for model training; the personal identifier and the categorical hours worked variable were removed. Moreover, units that were classified as no code required for the imputable variable were removed from the training and test datasets

```
# Tidy/treat the training and test datasets
# Remove units with NCR codes for variable & Remove the personal identifier
Census.train.tidy <- Census.train[!Census.train$econ.act == -9, c(-1, -17)]

Census.test.tidy <- Census.test[!Census.test$econ.act == -9, c(-1, -17)]
```

4) Missingness was simulated in the test dataset

```
# Simulate missingness in test data & Convert all missing responses to -999
Census.test.tidy.amp <- ampute(Census.test.tidy, prop = 0.7)

Census.test.tidy.miss <- Census.test.tidy.amp$amp

# Study the test dataset with missingness: How much missingness per variable?
NumberMissing <- sapply(Census.test.tidy.miss, function(y) sum(length(
    which(is.na(y))
  )))

TestNumberMissing <- data.frame(NumberMissing)

# Convert missing cases to -999
Census.test.tidy.miss[is.na(Census.test.tidy.miss)] <- -999

# Save dataset with missingenss
save(Census.test.tidy.miss, file = "data/EconAct/Census.test.tidy.miss.Rda")
```

5) A model was built using the training data

```
# Train the model
# Convert the data to matrix and assign output variable
train.outcome <- Census.train.tidy$econ.act

train.predictors <- sparse.model.matrix(econ.act ~ .,
  data = Census.train.tidy
)[, -1]

test.outcome <- Census.test.tidy.miss$econ.act

test.predictors <- model.matrix(econ.act ~ .,
  data = Census.test.tidy.miss
)[, -1]

# Convert the matrix objects to DMatrix objects
dtrain <- xgb.DMatrix(train.predictors, label = train.outcome)

dtest <- xgb.DMatrix(test.predictors, missing = -999)

# Train a model using training set
trainEA_v1 <- xgboost(
  data = dtrain, max_depth = 2, eta = 1, nthread = 2, nrounds = 10,
  objective = "multi:softmax", num_class = 10, missing = -999
)

# Examine feature importance
importance_matrix <- xgb.importance(model = trainEA_v1)

print(importance_matrix)

xgb.plot.importance(importance_matrix = importance_matrix)

# Save model
xgb.save(trainEA_v1, "XGBoost/xgboost.econAct")
```

6) The model was used to predict values in the test dataset

```
# Test the model
predicted <- predict(trainEA_v1, dtest, missing = -999, na.action = na.pass)

# Save predicted values
save(predicted, file = "data/EconAct/XGBoost/predicted.RData")
```

7) Donor based imputation was carried out on the test data (with missingness):

i) CANCEIS: One round of CANCEIS selected matching variables based on a correlation matrix. Variables that had a correlation coefficient of |0.4| or greater were included as matching variables in the CANCEIS imputation specification. All variables were given the same weight.

ii) Mixed Methods: One round of CANCEIS selected matching variables based on the feature importance figures from the XGBoost model. The six most important variables from the feature importance output were selected as matching variables; with more important variables assigned a larger weight.

```
# Impute values using CANCEIS
# Create CANCEIS input file with imputable and matching variables
CANCEIS.input <- Census.test.tidy.miss[, c(
  "econ.act", "student", "industry",
  "age", "occupation", "social.grade"
)]

CANCEIS.input$canceis.id <- 1:nrow(CANCEIS.input)
CANCEIS.input <- CANCEIS.input[, c(
  "canceis.id", "econ.act", "student", "industry",
  "age", "occupation", "social.grade"
)]

write.table(CANCEIS.input,
  file = "data/EconAct/CANCEIS/xxxUNIT01IG01.txt", sep = "\t",
  row.names = FALSE, col.names = FALSE
)
```

```
# Impute values using CANCEIS (with XGBoost to advise selection of MVs)
# Create CANCEIS input file with imputable and matching variables
CANCEISXG.input <- Census.test.tidy.miss[, c(
  "econ.act", "hours.cont", "age", "student",
  "sex", "health", "industry"
)]

CANCEISXG.input$canceis.id <- 1:nrow(CANCEISXG.input)
CANCEISXG.input <- CANCEISXG.input[, c(
  "canceis.id", "econ.act", "hours.cont", "age", "student",
  "sex", "health", "industry"
)]

write.table(CANCEISXG.input,
  file = "data/EconAct/MixedMethods/xxxUNIT01IG01.txt", sep = "\t",
```

```
  row.names = FALSE, col.names = FALSE
)
```

8) The two rounds of donor based imputation (using CANCEIS) and the (XGBoost) model based imputation were compared using either root mean squared error, absolute error (for continuous variables) and the confusion matrix (for categorical variables).

   i) First, information and data from the previous steps was loaded into working memory

   ii) Next, Each of the methods was evaluated with respect to the performance measure, and compared to either median (for continuous) or mode (for categorical) variables

```
# Load datasets
# Test and Training data
load("data/Census.train.Rda")

load("data/Census.test.Rda")

# Load dataset with missingenss
load("data/EconAct/Census.test.tidy.miss.Rda")

# Create test.tidy and train.tidy datasets (Remove units with NCR codes for variable
# & Remove the personal identifier)
Census.train.tidy <- Census.train[!Census.train$econ.act == -9, c(-1,-17)]

Census.test.tidy <- Census.test[!Census.test$econ.act == -9, c(-1,-17)]

# Read in CANCEIS input and output
CANCEIS.test.in <- read.table("data/EconAct/CANCEIS/xxxUNIT01IG01.txt",
                              header = FALSE,
                              col.names = c(
                                "canceis.id", "econ.act", "student",
                                "industry", "age", "occupation",
                                "social.grade"
                              )
)[, -1]

CANCEIS.test.out <- read.table("data/EconAct/CANCEIS/XXXUNITIMP01IG01.txt",
                               header = FALSE,
                               col.names = c(
                                 "canceis.id", "econ.act", "student",
                                 "industry", "age", "occupation",
```

```
                                  "social.grade"
                                )
)[, -1]

# Read in CANCEISXG input and output
CANCEISXG.test.in <- read.table("data/EconAct/MixedMethods/xxxUNIT01IG01.txt",
                                header = FALSE,
                                col.names = c(
                                  "canceis.id", "econ.act", "hours.cont",
                                  "age", "student",
                                  "sex", "health", "industry"
                                )
)[, -1]

CANCEISXG.test.out <- read.table("data/EconAct/MixedMethods/XXXUNITIMP01IG01.txt",
                                 header = FALSE,
                                 col.names = c(
                                   "canceis.id", "econ.act", "hours.cont",
                                   "age", "student",
                                   "sex", "health", "industry"
                                 )
)[, -1]
# Load predicted values from XGBoost
load("data/EconAct/XGBoost/predicted.RData")

# Load model
trainEA_v1 <- xgb.load("XGBoost/xgboost.econAct")
```

```
# Evaluate performance of XGBoost model
# Compare versions of the outcome variable (Actual, Predicted, Missing)
actuals <- Census.test.tidy$econ.act

missing <- Census.test.tidy.miss$econ.act

compareVar <- tibble(
  Actuals = actuals, Predictions = predicted,
  Missing = missing
)

compareMissing <- compareVar[compareVar$Missing == -999, ]

compareMissing$indicator <- ifelse(compareMissing$Actuals ==
                                     compareMissing$Predictions,"Correct", "Wrong")

counts <- table(compareMissing$indicator)
```

```
barplot(counts, main = "Accuracy of predictions", xlab = "Outcome")

# Using Confusion Matrix to evaluate predictions
confusionML <- confusionMatrix(
  as.factor(compareVar$Actuals),
  as.factor(compareVar$Predictions)
)

qplot(Actuals, Predictions,
      data = compareVar,
      geom = c("jitter"), main = "predicted vs. observed in test data",
      xlab = "Observed Class", ylab = "Predicted Class"
) + scale_x_discrete(limits=c("1","2","3","4","5","6","7","8","9")
) + scale_y_discrete(limits=c("1","2","3","4","5","6","7","8","9"))

ggsave("images/EAXGqplot.png")

# Evaluate performance of CANCEIS
# Compare predicted and actuals
actuals.CANCEIS <- Census.test.tidy$econ.act

missing.CANCEIS <- CANCEIS.test.in$econ.act

predicted.CANCEIS <- CANCEIS.test.out$econ.act

compare_var_CANCEIS <- tibble(
  Actuals = actuals.CANCEIS, Predictions =
    predicted.CANCEIS, Missing = missing.CANCEIS
)

compare_missing_CANCEIS <- compare_var_CANCEIS[
  compare_var_CANCEIS$Missing == -999, ]

compare_missing_CANCEIS$indicator <- ifelse(
  compare_missing_CANCEIS$Actuals ==
    compare_missing_CANCEIS$Predictions,
  "Correct", "Wrong"
)

counts_CANCEIS <- table(compare_missing_CANCEIS$indicator)

barplot(counts_CANCEIS, main = "Accuracy of predictions", xlab = "Outcome")

# Using Confusion Matrix to evaluate predictions
confusion_CANCEIS <- confusionMatrix(
```

```
  as.factor(compare_missing_CANCEIS$Actuals),
  as.factor(compare_missing_CANCEIS$Predictions)
)

qplot(Actuals, Predictions,
      data = compare_missing_CANCEIS,
      geom = c("jitter"), main = "predicted vs. observed in validation data",
      xlab = "Observed Class", ylab = "Predicted Class"
) + scale_x_discrete(limits=c("1","2","3","4","5","6","7","8","9")
) + scale_y_discrete(limits=c("1","2","3","4","5","6","7","8","9"))

ggsave("images/EACANCEISqplot.png")

# Evaluate performance of CANCEISXG
# Compare predicted and actuals
actuals.CANCEISXG <- Census.test.tidy$econ.act

missing.CANCEISXG <- CANCEISXG.test.in$econ.act

predicted.CANCEISXG <- CANCEISXG.test.out$econ.act

compare_var_CANCEISXG <- tibble(
  Actuals = actuals.CANCEISXG, Predictions =
    predicted.CANCEISXG, Missing = missing.CANCEISXG
)

compare_missing_CANCEISXG <- compare_var_CANCEISXG[
  compare_var_CANCEISXG$Missing == -999, ]

compare_missing_CANCEISXG$indicator <- ifelse(
  compare_missing_CANCEISXG$Actuals ==
    compare_missing_CANCEISXG$Predictions,
  "Correct", "Wrong"
)

counts_CANCEISXG <- table(compare_missing_CANCEISXG$indicator)

barplot(counts_CANCEISXG, main = "Accuracy of predictions", xlab = "Outcome")

# Using Confusion Matrix to evaluate predictions
confusion_CANCEISXG <- confusionMatrix(
  as.factor(compare_missing_CANCEISXG$Actuals),
  as.factor(compare_missing_CANCEISXG$Predictions)
)
```

```
qplot(Actuals, Predictions,
      data = compare_missing_CANCEISXG,
      geom = c("jitter"), main = "predicted vs. observed in validation data",
      xlab = "Observed Class", ylab = "Predicted Class"
) + scale_x_discrete(limits=c("1","2","3","4","5","6","7","8","9")
) + scale_y_discrete(limits=c("1","2","3","4","5","6","7","8","9"))

ggsave("images/EACANCEISXGqplot.png")

# Impute values using mode imputation
# Create a vector of imputable variable excluding missing values
mode.dat <- Census.test.tidy.miss[
  Census.test.tidy.miss$econ.act != -999, ]

mode.val <- Mode(mode.dat$econ.act)

# Compare predicted and actuals
actuals.mode <- Census.test.tidy$econ.act

missing.mode <- Census.test.tidy.miss$econ.act

predicted.mode <- ifelse(
  Census.test.tidy.miss$econ.act == -999, mode.val,
  Census.test.tidy.miss$econ.act)

compare_var_mode <- tibble(
  Actuals = actuals.mode, Predictions =
    predicted.mode, Missing = missing.mode
)

compare_missing_mode <- compare_var_mode[
  compare_var_mode$Missing == -999, ]

compare_missing_mode$indicator <- ifelse(
  compare_missing_mode$Actuals ==
    compare_missing_mode$Predictions,
  "Correct", "Wrong"
)

counts_mode <- table(compare_missing_mode$indicator)

barplot(counts_mode, main = "Accuracy of predictions", xlab = "Outcome")
```

Chapter 7

Results

7.1 Summary statistics

Summary statistics were produced to review the pattern of responses of individuals included in the dataset. Please note, that all comments that refer to respondents reflect only the respondents included in the 2011 Census Teaching File, and not descriptive statistics for the Census Population as a whole. Summary statistics produced using the dataset show that:

- The majority of respondents resided in the South East and London. A relatively small proportion reside in North East and Wales.
- Almost all respondents resided in a non-communal establishment
- The majority of respondents were living in a family that was composed of a married or same-sex civil partnership couple
- Almost all respondents were usual residents at the collection address during time of collection
- There were a similar number of male and female respondents in the dataset
- The majority of respondents were aged 0 to 15
- The majority of respondents were single and had never married or registered for a same-sex civil partnership
- The majority of respondents were not school children nor were they in full time study
- The majority of respondents were born in the United Kingdom

- The majority of respondents reported as being in very good health at time of collection
- With respect to Ethnicity, the majority of respondents identified as White
- With respect to religion, the majority of respondents identified as Christian. The second largest group were those that stated they had no religion.
- The two most prevalent categories for economic activity were employee and retired
- Of those that were eligible to answer the Occupation item, the majority were either in a Professional or Elementary occupation
- Of those that were eligible to answer the Industry item, the majority were employed in the Wholesale and retail trade industry
- Of those that were eligible to answer the hours worked item, the majority worked between 31 and 38 hours per week
- Of those elgible to answer the Social Grade item, the majority would be classed into the Supervisory, Clerical, and Junior Managerial social group

```
# Study the data: 1) How many units, 2) How many attributes? and
# 3) How many missing units?
str(Census)

summary(Census)

sapply(apply(Census[, c(-1, -19)], 2, table), function(x) x / sum(x))

missing_values <- sapply(Census, function(y) sum(length(which(is.na(y)))))

missing_values <- data.frame(missing_values)

# Look for relationship between variables
ggcorr(Census[, -1],
       nbreaks = 8, palette = "RdGy",
       label = TRUE, label_size = 3, label_color = "white"
)

ggsave("images/cor_all.png")
```

Bar charts were used to review the distribution of responses for each categorical variable in the complete, training and test datasets (see Figures 2 to 4). As expected, there was a similar response pattern for a each variable between the

complete, training and test datasets. The correlation matrix in Figure 5 presents the relationship between the variables in the complete dataset.

```
# Compare the test and training datasets
str(Census.train)

str(Census.test)

summary(Census.train)

summary(Census.test)

sapply(apply(Census.train[, c(-1, -19)], 2, table), function(x) x / sum(x))

sapply(apply(Census.test[, c(-1, -19)], 2, table), function(x) x / sum(x))

# Plot distribution of variables
melt.Census <- melt(Census)

head(melt.Census)

ggplot(data = melt.Census, aes(x = value)) +
  geom_bar() +
  facet_wrap(~variable, scales = "free")

ggsave("images/dist_all.png")

# Plot distribution of variables
melt.Census.train <- melt(Census.train)

head(melt.Census.train)

ggplot(data = melt.Census.train, aes(x = value)) +
  geom_bar() +
  facet_wrap(~variable, scales = "free")

ggsave("images/dist_train.png")

melt.Census.test <- melt(Census.test)

head(melt.Census.test)

ggplot(data = melt.Census.test, aes(x = value)) +
  geom_bar() +
  facet_wrap(~variable, scales = "free")
```

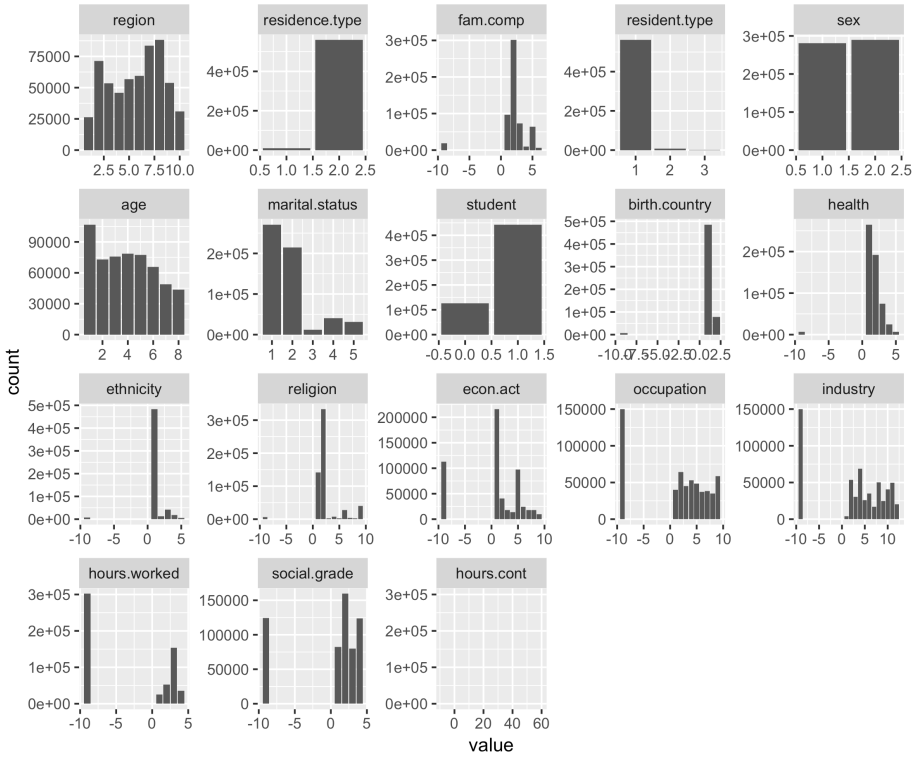

Figure 7.1: Figure 2. Distribution of responses for variables in complete Census Teaching File

```
ggsave("images/dist_test.png")
```

7.2 Comparison of imputation methods

The results for the different imputation methods are presented for each imputable variable. Performance measures were selected based on the type of imputable variable used (i.e. categorical or continuous). Please refer to the links below for guidance on interpreting the performance measures:

- Root mean squared error and mean absolute error (for continuous variables)
- Confusion matrix (for categorical variables)

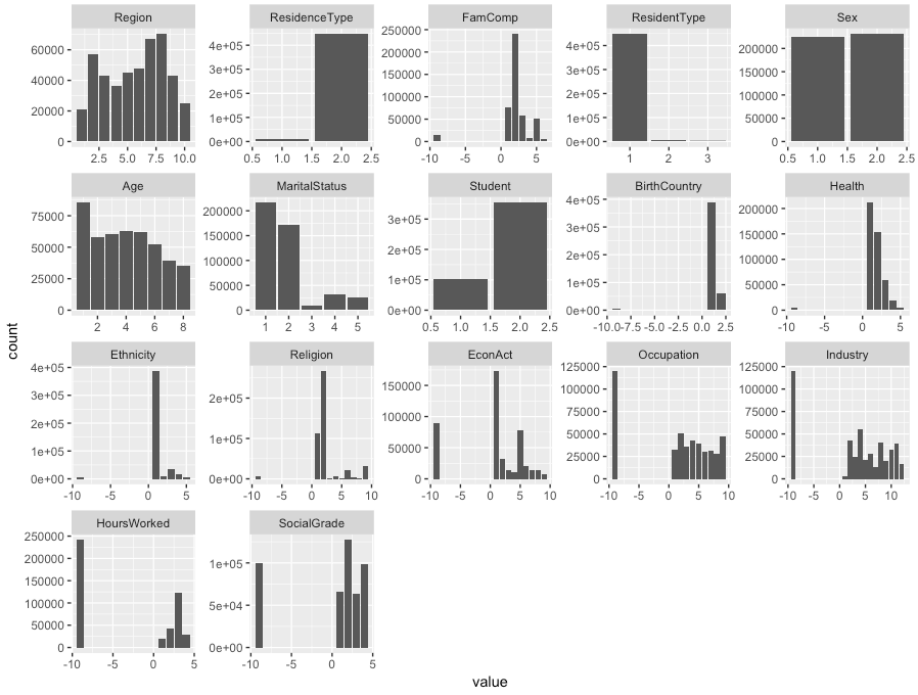

Figure 7.2: Figure 3. Distribution of responses for variables in training dataset

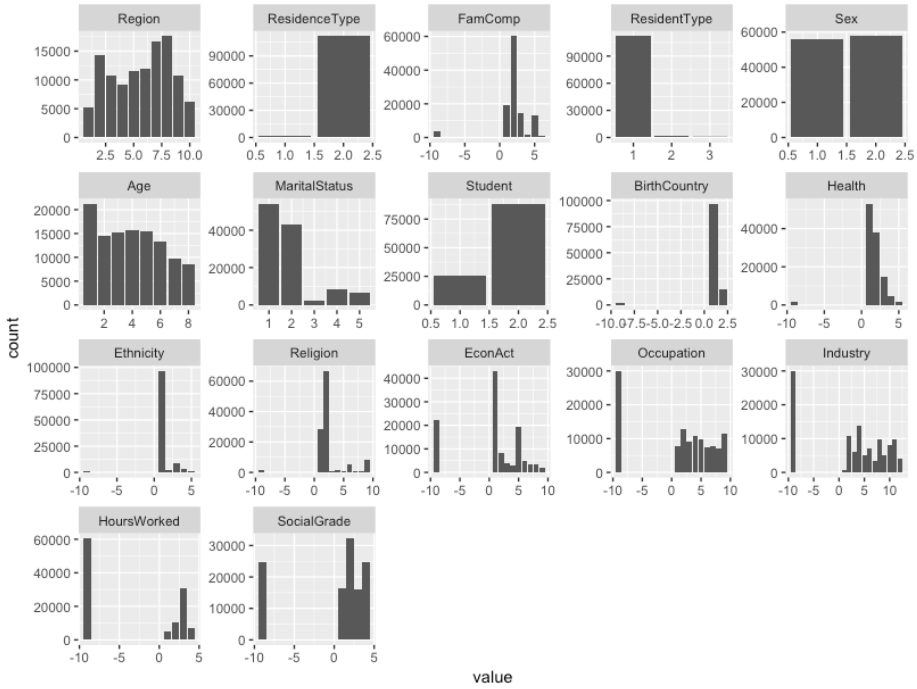

Figure 7.3: Figure 4. Distribution of responses for variables in test dataset

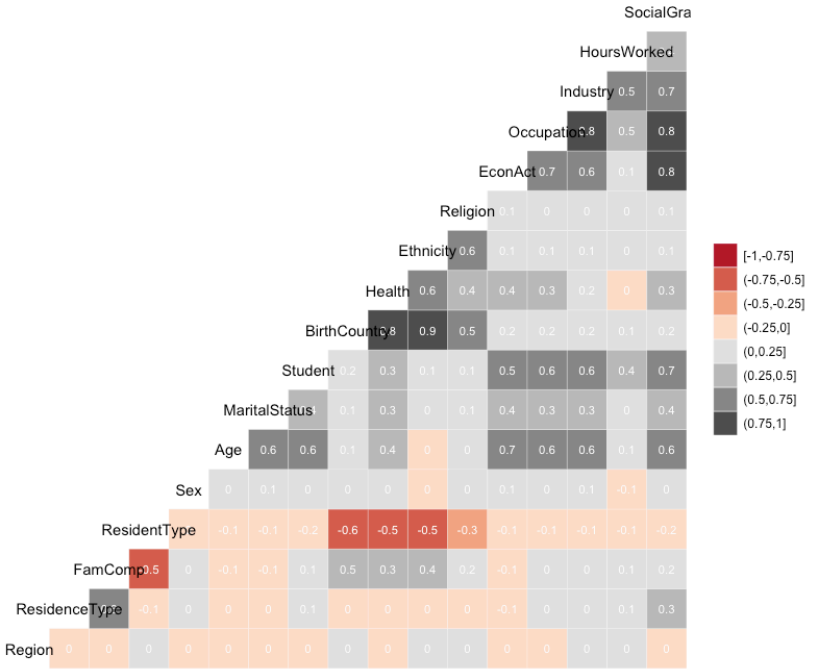

Figure 7.4: Figure 5. Correlation matrix of variables in the Census dataset

For the categorical imputable variables, plots of the Observed vs Predicted are provided to give an indication of which categories each method predicted relatively well.

7.2.1 Economic Activity

The results shows that:

- XGBoost predicted economic activity with greater accuracy relative to donor and mode imputation

- Compared to donor based methods, the XGBoost model appeared to have greater sensitivity for the different classes of economic activity. That is, for any given class of economic activity, the model based approach was more likely to predict the correct response relative to donor based methods.
- The Mixed Methods model was the least accurate imputation method for the multi-class variable, economic activity.

XGBoost

CANCEIS

MixedMethods

Mode

Accuracy

0.6956778

0.5448752

0.2934148

0.4944238

Kappa

0.5640313

0.3329717

0.1218344

NA

7.2.2 Hours worked

The results shows that:

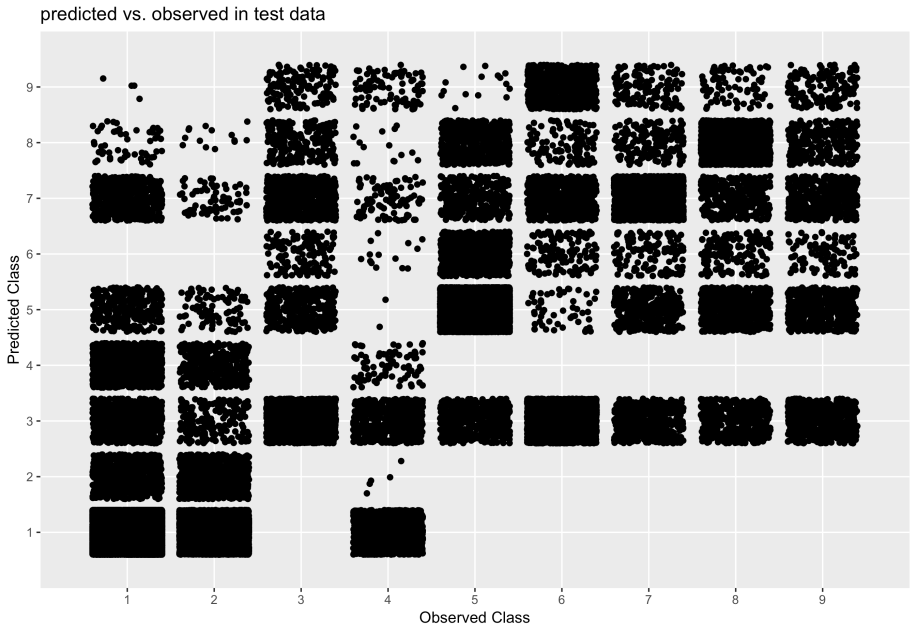

Figure 7.5: Performance of XGBoost in predicting economic activity

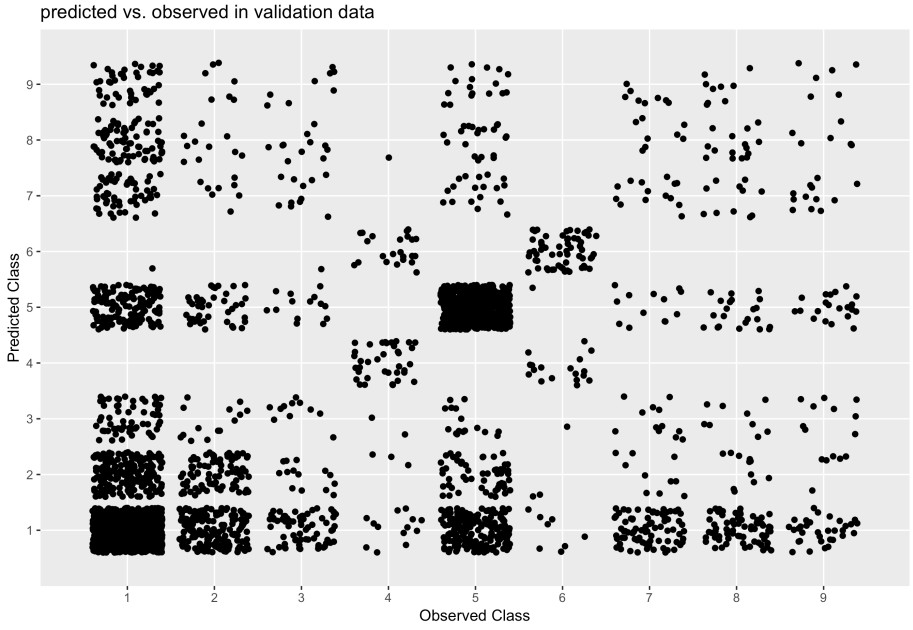

Figure 7.6: Performance of CANCEIS in predicting economic activity

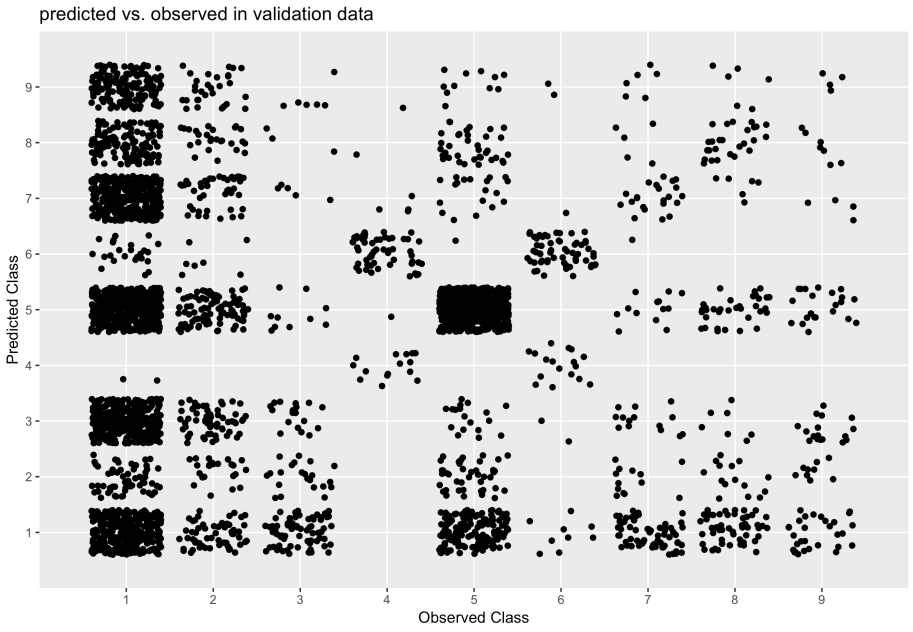

Figure 7.7: Performance of Mixed Methods approach in predicting economic activity

- XGBoost predicted hours worked with greater accuracy relative to donor and mode imputation
- Median imputation and the Mixed Methods approach had a similar level of accuracy
- Donor based imputation had the lowest level of accuracy

XGBoost

CANCEIS

MixedMethods

Median

MAE

9.31

13.29

12.63

10.64

RMSE

11.83

16.92

16.43

16.43

7.2.3 Social Grade

The results shows that:

- All three approaches performed with similar degree of accuracy, whilst out-performing mode imputation
- Compared to donor based methods, the XGBoost model appeared to have greater sensitivity for the different classes of social grade. That is, for any given class of social grade, the model based approach was more likely to predict the correct response relative to donor based methods.

XGBoost

CANCEIS

MixedMethods

Mode

Accuracy

0.6425706

0.6230961

0.6471891

0.3447798

Kappa

0.5116984

0.4873809

0.5197337

NA

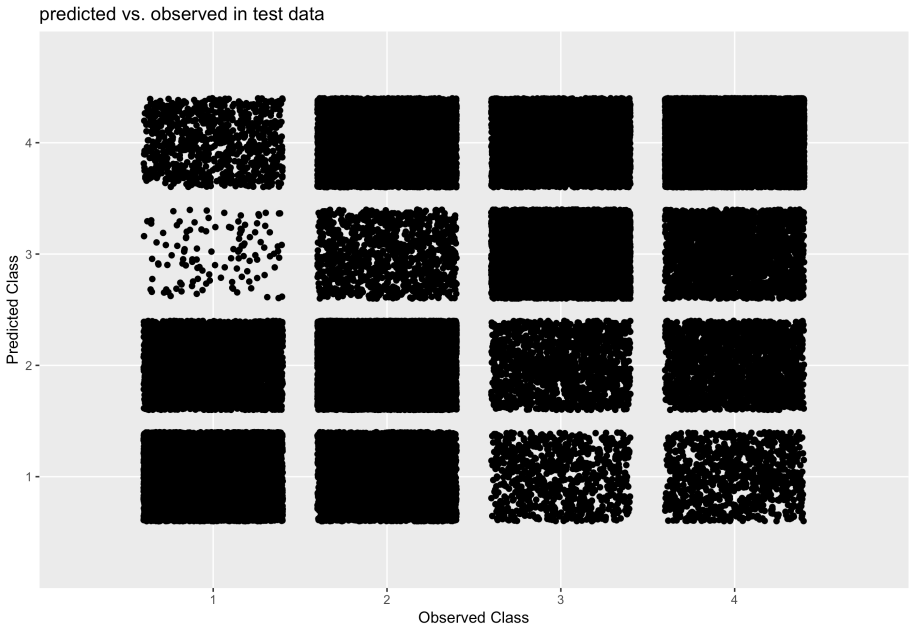

Figure 7.8: Performance of XGBoost in predicting social grade

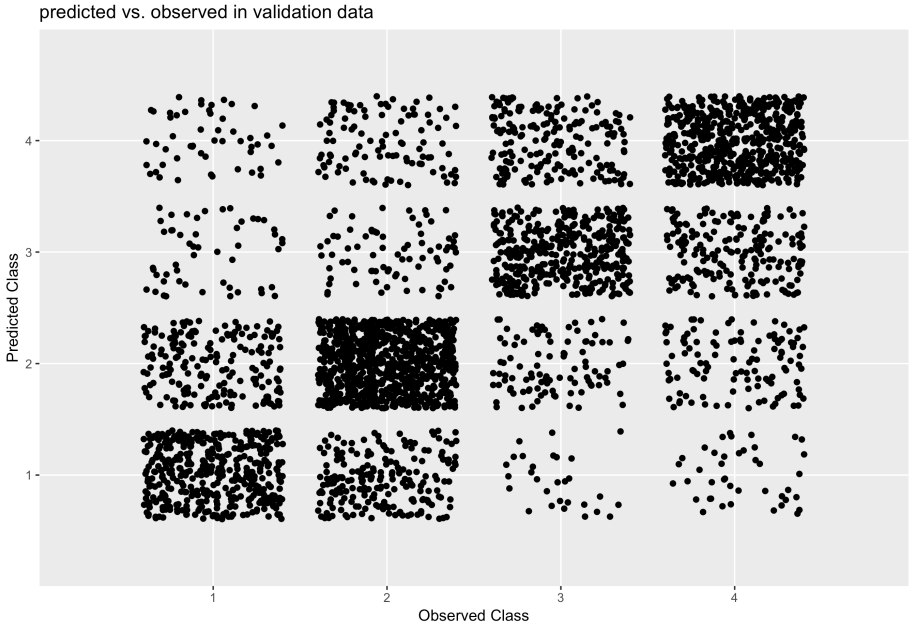

Figure 7.9: Performance of CANCEIS in predicting social grade

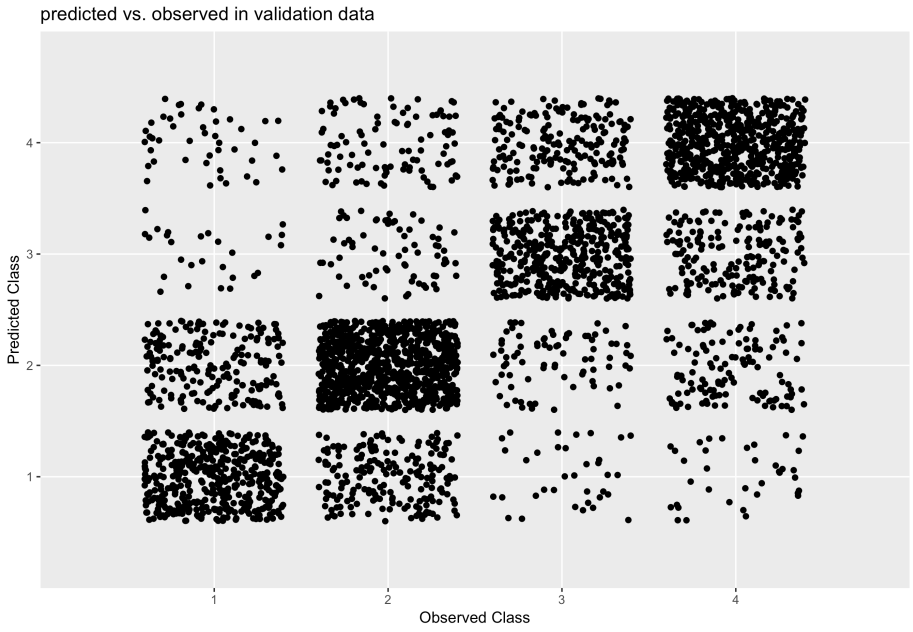

Figure 7.10: Performance of Mixed Methods approach in predicting social grade

7.2.4 Student

The results shows that:

- All three approaches performed with similar degree of accuracy, whilst out-performing mode imputation
- Compared to donor based methods, the XGBoost model appeared to have greater sensitivity for the different classes of student status. That is, for both students and non-students, the model based approach was more likely to predict the correct response relative to donor based methods.

XGBoost

CANCEIS

MixedMethods

Mode

Accuracy

0.9486349

0.9423158

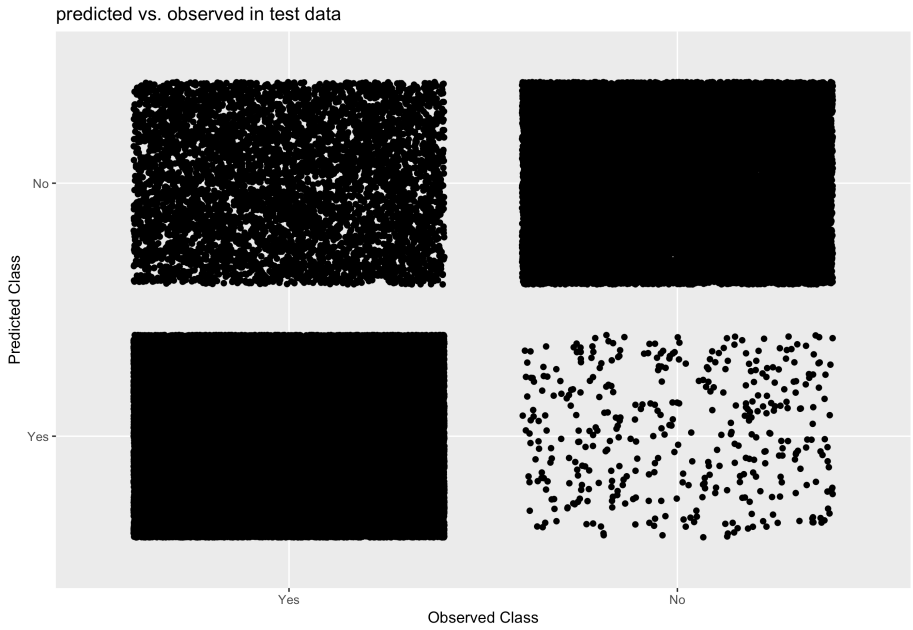

Figure 7.11: Performance of XGBoost in predicting student status

0.9461053

0.856

Kappa

0.8613193

0.7668641

0.7800476

NA

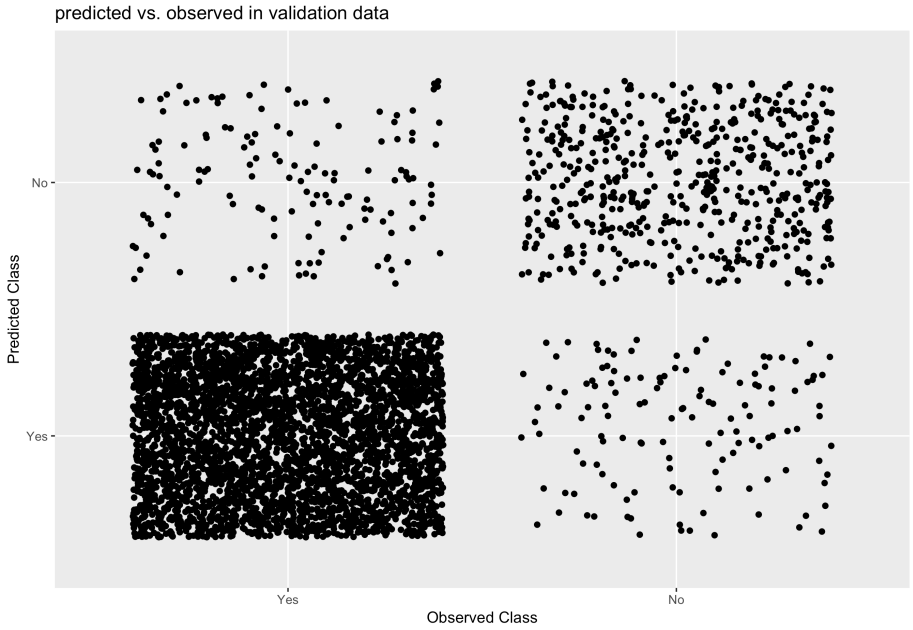

Figure 7.12: Performance of CANCEIS in predicting student status

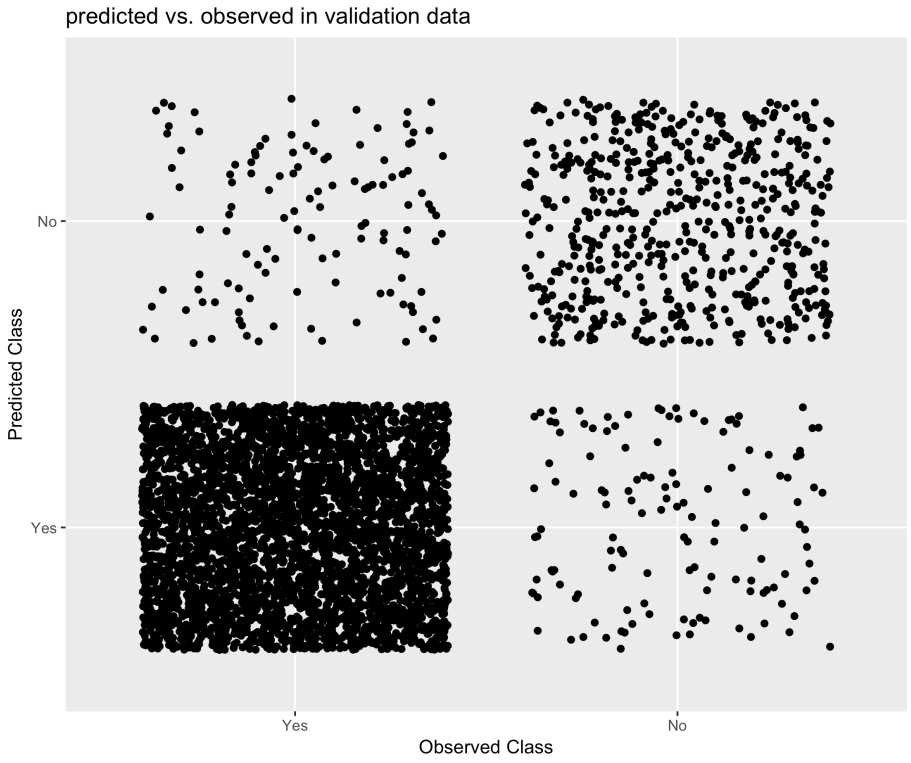

Figure 7.13: Performance of Mixed Methods approach in predicting student status

Chapter 8

Next steps

Preliminary results indicate that XGBoost performs well, relative to donor based methods, in univariate imputation. The models were especially effective at predicting the different classes of multi-class and binary variables. The ability to program an end to end imputation process is an added advantage of XGBoost; it reduces the time taken to implement an imputation method, and presents clients with the option of automating the imputation process. Current donor based methods utilise either closed, or proprietary code, which cannot easily be integrated into open source platforms.

The intention is to build on this work, and examine the efficacy of machine learning systems on household survey imputation by carrying out the following:

- Compare XGBoost, as well as Deep Learning methods (such as Autoencoders) to current donor based methods on ONS household survey data
- Study how changes to XGBoost hyper-parameters can influence the performance of the imputation model. As the current investigation, using Census data, did not iterate through different combinations of hyper-parameters, it would be of interest to see if this improves the performance of this ML system in household survey imputation.
- If time permits, it would be of interest to explore the use of XGBoost as a method to advise donor selection. That is, how the selection and importance of features could be used to identify matching variables and weights respectively.

Chapter 9

Resources

9.1 R scripts

All R scripts pretaining to this investigation are saved in the folder "R". If you are intending to run any of the code, please load all the packages first, which can be found in the script "WF1_package_load.R".

The program "WF2_data_prep.R" cleans and edits the data for the investigation, whilst the script "WF3_data_study.R" studies the dataset. Scripts with the prefix "WFM" refer to programs that produce the XGBoost models for each imputable variable, and create the CANCEIS input files. Do not run these scripts unless you would like to replicate the investigation.

If you would simply like to view the results, please load the packages and use the following scripts to compare the results of the different imputation methods:

- WFZ_EconAct_CompareImp.R: Comparing methods for the variable, economic activity
- WFZ_HoursCont_CompareImp.R: Compareing methdos for the variable, hours worked (continuous)
- WFZ_SocialGrade_CompareImp.R: Comparing methods for the variable, social grade
- WFZ_Student_CompareImp.R: Comparing methods for the variable, student status

9.2 Data

The folder named “data” includes:

- full Census Teaching File
- test and training data
- the predicted values from each imputation method and variable

9.3 XGBoost

The models produced for each imputable variable can be found in the following folder (located in the main directory):
- XGBoost

9.4 Donor imputation

The CANCEIS specifications used for the two rounds of donor based imputation can be found in the following folders (located in the main directory):
- CANCEIS
- MixedMethods

Bibliography

Ton de Waal, Jeroen Pannekoek, S. S. (2011). *Handbook of Statistical Data Editing and Imputation*. John Wiley Sons Inc, Hoboken, New Jersey, 1st edition. ISBN 978-0-470-54280-4.