# Assessing the Risk Factors for Breast Cancer using Penalised Regression and Data Mining Methods

**Srishti Bhargava**

*180006373*

A project submitted in partial fullfillment of the requirements for the degree of

M.Sc. Applied Statistics and Data Mining

University of St. Andrews

Supervisor : Professor Michail Papathomas

16 August 2019

## Declaration

I, Srishti Bhargava, certify that this project report has been written by me, is a record of work carried out by me, and is essentially different from work undertaken for any other purpose or assessment.

## Contents

## List of Figures

## List of Tables

# 1 Abstract

The objective of this paper is to determine the impact of anthropometric measurements, lifestyle choices, reproductive factors and nutritional and dietary habits on the risk of developing breast cancer. The data used for the analysis in this paper was collected, in Italy, by the European Prospective Investigation into Cancer and Nutrition as a part of an ongoing prospective cohort study. 3484 women without breast cancer and 136 with breast cancer were analysed using logistic regression, penalised logistic regression, and data mining methods. The final significant relationships identified by determining the importance of predictors used in each model showed an increase in the risk of breast cancer for women who experienced menopause after the age of 49. An increased risk of breast cancer for women who had been pregnant and had lactated was also observed. Finally, when adjusted for the age at childbirth– physical activity, complications in pregnancy and hyperlipidaemia were also found to increase the risk of breast cancer.

## 2 Introduction

A rapid increase in the breast cancer rate is being observed globally. The World Health Organization (WHO) estimates that over 2.1 million new cases of breast cancer are being diagnosed in women each year. In 2018, over 600,000 women died because of breast cancer, making it a leading cause of cancer-related deaths in women (World Health Organization|International Agency for Research on Cancer- Global Cancer Observatory 2018). Thus the need to research the causes, impact of lifestyle choices and dietary habits in particular, of the disease is imperative. Limited number of cases to study and the instability of human decisions restrict power–the ability of statistical tests to detect differences when they are present (N Slimani et al. 2002). Therefore, despite decades of research in the sector, it has been hard to establish strong relationships between human choices and cancer development.

Observational studies are often used to establish connections between disease and causation. There are two primary methods of conducting such studies, cohort and case-control. Case-control studies recruit participants, from a given population, who have already been diagnosed with the disease of interest and categorise them as cases. Then, from the same population, people without the disease are recruited and categorised as controls. Data about these individuals is collected by retrospective recollection of risk factors under consideration. This method is ideal for situations where the disease is rare since they are quick to conduct, require few subjects and are relatively inexpensive. Drawbacks include methodological issues such as case and control selection, which inadvertently introduces some form of bias in the study, and recall bias – the tendency of a subject to inaccurately describe an event or omit important information. On the other hand, cohort studies involve following groups of people with similar characteristics over time to see if they develop the disease of interest. This can be done in two ways, retrospectively or prospectively. Retrospective studies, also known as historical studies, select subjects and collect information in present time about past occurrences. They provide the data collector with limited information due to gaps in recollection and missing data in records but are less expensive as compared to prospective cohorts. Prospective studies involve following the subject under observation from the present to the future. These are advantageous as the study can be thoroughly designed to observe the impact of specific factors but are hard to execute since they have long follow-up periods and are expensive as they require a large number of participants.

This report examines a prospective cohort study by WHO. The International Agency for Research on Cancer - WHO established the European Prospective Investigation into Cancer and Nutrition (EPIC) in 1990 to overcome the issue of low statistical power attached with research into risk factors of breast cancer. EPIC is an ongoing multicentre prospective cohort study, with over 521,000 participants enrolled from 23 centres in 10 western European countries (Riboli et al. 2002). An EPIC centre provided the data used for the analysis, conducted in this report, of risk factors in Italy. The centre used baseline health check-ups, questionnaires and interviews with regular follow-ups to collect and maintain information about factors including anthropometric measurements taken at the centre, lifestyle and reproductive histories, and dietary habits. Unavoidable drawbacks, such as missing values in certain factors, that probably arise from long follow-up periods associated with the study design, have impacted the data collection process (Song and Chung 2010). Details of these can be found in section 4.1.

Previous studies, analysing the distribution and determinants of breast cancer, focus on standardised regression methods. The methods are based on assumptions that may not reflect the true nature of interactions between the different risk factors and breast cancer (Lynch and Moore 2016). Thus, besides standard linear methods, machine learning approaches such as classification trees and random forests have also been applied in this study.

# 3 Literature Review

There are a large number of studies that look into the potential risk factors of breast cancer. This section categorises and summarises some of those studies which are relevant to the analysis conducted hereafter.

## 3.1 Anthropometric Measures

Anthropometric measures including height, weight, waist circumference, hip circumference, body mass index (BMI) and waist to hip ratio (WHR) are considered indicators of body type and composition. While they have been studied continuously as predictors of breast cancer, they have also been penalised for being imperfect reflections of the biological structures they represent (Bandera et al. 2016). Nevertheless, they provide valuable information, as summarized in the studies reviewed below, about associations between body structure and the risk of breast cancer.

A study within the EPIC cohort comprising of 69,116 women, out of whom 1135 were diagnosed with breast cancer, was conducted using Cox Proportional hazards modelling. This provided insight into the relationship between anthropometric measures and breast cancer for premenopausal and postmenopausal women separately. The study reported that taller women have a marginally higher risk of developing the disease regardless of their menopausal status. Weight, waist circumference, waist to hip ratio and body mass index were negatively correlated with breast cancer cases in postmenopausal women and positively correlated in premenopausal women. These results became insignificant after adjustment for body mass index. The waist to hip ratio's impact on the risk of breast cancer was highly uncertain and thus recommended as an area for further study. Nevertheless, the study concluded that obesity is a risk factor in the risk of breast cancer (Tehard and Clavel-Chapelon 2006).

Another study conducted in Carcow, Poland, analysed detailed anthropometric measurements of 487 women out of whom 193 suffered from breast cancer. This analysis used Student's t test, Mann-Whitney's test and logistic regression to identify the significance of each anthropometric measure. While BMI was negatively associated with the risk of breast cancer, the study reported an increase in the risk of cancer for women with a higher waist-to-hip ratio in premenopausal women. It concluded that obesity was a risk factor in the prevalence of breast cancer but could not establish strong relationships between specific anthropometric measures and breast cancer (Pacholczak, Klimek-Piotrowska, and Kuszmiersz 2016).

Research shows that complex relationships exist between anthropometric measures and other risk factors which impact the risk of breast cancer, including, but not limited to menopausal status, physical changes and age (Ziegler 2018). Thus, there is a need to explore these interactions.

### 3.2 Reproductive History

EPIC designed a comprehensive questionnaire to understand the impact of different aspects of the female reproductive system on the chance of developing breast cancer. Since breast cancer development is known to be affected by oestrogen, these factors are significant because they are closely linked with oestrogen production in the female body (Newcomb et al. 1998).

Three case-control studies conducted in Italy were combined to analyse 4,072 cases and 4,099 controls using multiple logistic regression to find the relative risks reproductive factors pose on breast cancer. The analysis found a direct relationship between age at menopause and the risk of developing breast cancer. Since many women were yet to experience menopause, the study grouped women–by age–into categories. The study reported that women who experienced early menopause have a lower risk of being diagnosed with breast cancer. The same study revealed that pregnancy and age at first birth were also indicators of breast cancer. Age at first birth was divided into categories since not all women gave birth. Women who gave birth after the age of 28 years had a higher chance of developing breast cancer than women who gave birth before the age of 22. Age at menarche was also identified as a significant predictor of breast cancer. A noted shortcoming of this study was the inability to identify significant interactions (E. et al. 1988).

A total of 4575 women with breast cancer and 4682 without breast cancer, from Atlanta, Detroit, Philadelphia, Los Angeles, and Seattle were interviewed in a population-based case-control study. This was done to establish a relationship between oral contraceptives and breast cancer. The data collected was analysed using conditional logistic regression. It showed a small increase in the risk of breast cancer with the use of birth control. This relationship is not well established, and hence, the impact is still under study (Hunter 2017).

A case-control study conducted in western New York State, including only participants that had given at least one live birth, examined lifetime lactation history in relation to breast cancer. In accordance with previous research, this study, consisting of 620 cases and 693 controls, established a weak link between breast cancer risk and lactation. Due to the small number of participants in this study, further research is required to verify the contribution of lactation as a potential risk factor of breast cancer (Freudenheim et al. 1997).

A report published by the World Cancer Research Fund stated that complications in pregnancy such as stillbirth, miscarriage or induced abortion have individually not been proven to increase the risk of breast cancer. An overall impact of any complication is seen as an interesting area of research (World Cancer Research Fund/American Institute for Cancer Research 2018).

### 3.3 Lifestyle Factors

Lifestyle decisions have been repeatedly shown to impact health. The extent of this impact is hard to measure as human life choices change from time to time. Thus, any impact of these factors on health would be calculated on the assumption that the individual under study continues the same routine that the individuals under study indicated.

A review of the literature on physical activity and the risk of breast cancer showed inconsistency. Although 15 out of 21 papers compared showed a decrease in the risk of breast cancer with increased physical activity, the magnitude and biological reasons for any impact were uncertain. Furthermore, most of these studies were case-control, in comparison to the fewer cohort studies in which the findings were more unstable (Friedenreich et al. 1998).

It has been proven that smoking has adverse effects on health. A multicentre population-based case-control study in the United States compared 4720 cases, with 4682 controls, with the help of logistic regression and revealed a higher risk of breast cancer for smokers than for non-smokers (Chu et al. 1990).

Many studies focus on proving the enormity of this impact, but cohort studies struggle to establish a link between the two. Any significant results are possibly the outcome of bias since smoking has an antiestrogenic effect, bound to impact the risk of breast cancer (Terry Pd Fau Rohan and Rohan 2002).

Alcohol consumption has been a debatable indicator of many health issues. Studies with contradicting findings, depending on design, have been published. A review of 53 epidemiological studies concluded that the risk of breast cancer depended a lot on the amount of alcohol consumed. Limited intake did not have much impact, whereas a linear relationship between moderate to high consumption and breast cancer risk was observed (Schuetze 2007).

Hypertension and hyperlipidaemia are common health problems and hence widely researched. A meta-analysis of the literature testing the association between cholesterol and breast cancer indicated a small yet statistically significant relationship between hyperlipidaemia and decreased risk of breast cancer (Touvier et al. 2015). The meta-analysis of hypertension studies indicated inconclusive results. Though there were statistically significant relationships between hypertension and increased risk of breast cancer, these results differ from study to study and with difference menopausal status (Han et al. 2017).

### 3.4 Nutritional and Dietary Information

Nutritional and dietary information is hard to measure as diet changes frequently. Nevertheless, research into these is vital to establish a link between nutrition and disease. Many studies, such as the one reviewed below, divide information into categories based on food groups defined by the Food and Agricultural Association of the United Nations.

A review of papers on diet and risk of breast cancer showed that the risk increases with an increase in consumption of red meat and fatty foods. The same study also indicated a mildly significant inverse relationship between consumption of fruits and vegetables and risk of breast cancer (Kotepui 2016). The same conclusion was made by another study on EPIC data from 8 countries into the impact of fruits and vegetables on cancer (P. Lagiou et al. 2005).

22 cohort studies and 5 case-control studies were reviewed to analyse the association between dairy products and breast cancer. The conclusion presents strong evidence for an inverse linear relationship between breast cancer and dairy consumption. This is attributed to the calcium content in dairy products, although a real causal relationship is yet to be established with certainty(Zang et al. 2015).

Italy based study into cereal and grain intake and selected cancers demonstrated an increase in the risk of breast cancer with a higher quantity of consumption(L. et al. 1999). However, most studies that look into dietary variables, including the one mentioned above, provide inconclusive evidence for the same.

Fatty acids–if not moderated–have been known to impact health. Oils and other such substances are a major source of fatty acids in a staple diet. Research into their impact on the probability of developing cancer has been inconclusive at best. A meta-analysis of 52 papers on this subject revealed a positive relationship between oil intake and breast cancer in premenopausal women. Overall, evidence was inconclusive as results varied across study type (Xia et al. 2015).

A case-control study in Italy that used multiple logistic regression to detect differences-using 2569 cases and 2588 controls-found a directly proportional relationship between the consumption of sweets, including sugar and desserts, and the risk of breast cancer (Tavani et al. 2006).

Literature indicates that coffee has a significant impact on health, although the magnitude and direction of the relationship is highly uncertain. In the case of breast cancer, a review of 26 studies pointed towards a weak inverse link that depended on the amount of coffee intake; higher coffee intake reduces the chance of developing breast cancer (Li et al. 2013).

### 3.5 Age

Age has been an established indicator of health in many studies; old age is associated with degeneration of tissues which impact the vital functions of organs (MacNee, Rabinovich, and Choudhury 2014). Though the result of a study on breast cancer epidemiology did not indicate an impact of age on the risk of developing breast cancer when adjusted for confounders such as menopausal status, the initial results were statistically significant (Mcpherson, Steel, and Dixon 2000). This result aligned with other research findings that show a higher risk of breast cancer development in older women.

Previous studies have therefore highlighted the importance of anthropometric measures, lifestyle factors, diet and nutrition and age on the risk of developing breast cancer. While the research has shown promising leads, there is still much to be learnt from the data through further exploration and analysis.

## 4 Data Exploration

The original dataset, obtained from Italy, comprised of 8840 subjects and 762 variables or risk factors. It was initially recorded in Italian and was translated to English. An initial look at the data revealed that many variables were summarizing specific details about measures of interest; these were discarded due to the negligible amount of information they provided. Some variables were combined into one either because they yielded similar information or because they provided new insights into the concepts as combinations.

### 4.1 Missing Values

Summary statistics revealed only 3 cases of breast cancer amongst men. Thus the study was limited to analysing risk factors for females only. Approximately 12 % of the total subjects had missing values in at least one variable. On further inspection, 218 subjects were found to have only anthropometric measurements and were thus discarded. Another 35 observations were removed as they had multiple missing values.

On plotting the missing values according to the recruitment date, no pattern was observed since the values may have been missing at random. This meant that the probability of a missing observation depended on observed values. Thus, instead of deletion, predictive mean matching was used to impute missing values using multivariate imputation by chained equations (MICE).

Unlike the traditional approach of imputing the mean value of the continuous variable, MICE can predict both continuous and categorical variables. The missing values are imputed multiple times to create different “complete” datasets, which help decrease uncertainty. Most methods of this variety follow a similar algorithm.

The first step is to generate imputations for the variables with missing values using Gibbs sampling. Gibbs sampling is a Monte Carlo Markov Chain method that determines the conditional probability of the variable with missing values keeping the remaining variables constant. It then generates plausible values for the missing observations. If more than one of the variables have missing values, the most recent imputation is used to complete all predictors before imputing the variable under focus. This method is applied to all variables with missing values (Buuren and Groothuis-Oudshoorn 2011).

In particular, the predicted value of a missing observation is calculated using predictive mean matching method by fitting the variable containing missing value-as response- in a linear model. For each missing value a set of cases is identified, that does not contain missing values, with predicted values close to the predicted values of the missing variables. From these cases, a random value is chosen and assigned the same value as the missing case. The process is then repeated. The variables imputed using predictive mean matching method were height, weight, waist ratio, hip ratio, age at menopause and physical activity level.

Although it is a general method that is not sensitive to distributional assumptions and can impute both categorical and continuous variables, it only imputes values within the range of the values already present in the dataset. This can lead to many values being repeated multiple times and hence introduce bias in the study.

### 4.2 Variable Selection

Out of the 762 variables measured in the dataset, 30 were chosen with the help of the literature reviewed in section 2. Some of these were manipulated to maximize the information gained from the risk factor. A category wise summary of variables under consideration is presented in this section.

#### 4.2.1 Anthropometric Measures

Anthropometric measures- taken at the time of recruitment- included height, weight, waist circumference and hip circumference. Body Mass Index (BMI) was computed using weight and height. Height measurement in centimetres was divided by 100 to convert the measurements to meters for the computation.

$$BMI = \frac{Weight(kg)}{(Height(cm) * 100)^2} \tag{1}$$

The waist to hip ratio was calculated by dividing the waist circumference by the hip circumference.

$$WHR = \frac{WaistCircumference}{HipCircumference} \tag{2}$$

The data was visualised to identify any differences in the range of values of the two groups. From figure 1 and appendix 11.1.1 plots it can be inferred that heavy outliers might be influencing or aliasing some relationships.

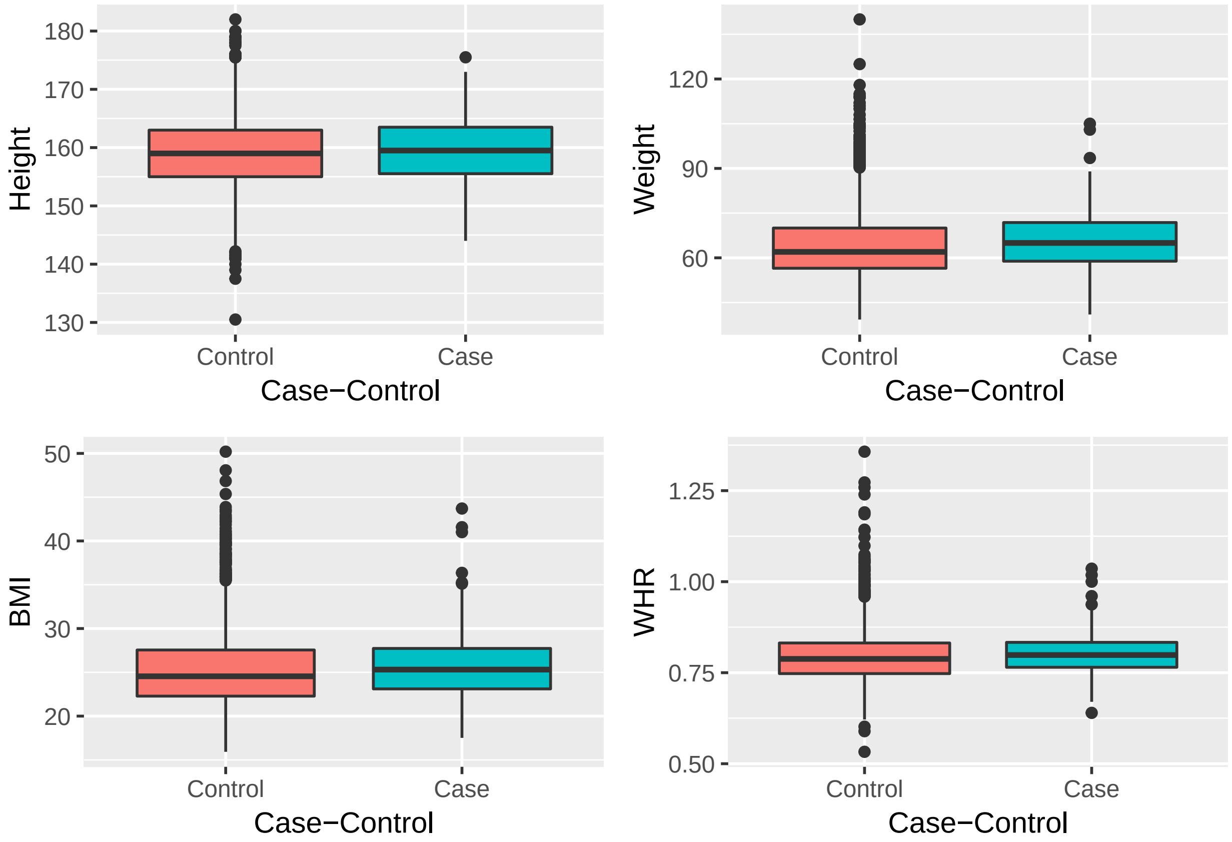


Figure 1: Boxplot comparing differences between case and control in anthropometric measures

#### 4.2.2 Reproductive factors

In this analysis, reproductive variables were combined together to eliminate missing values and/or because they provided similar information.

Table 1: Categorical division of age at menopause and percentage of data associated with each category

| Category | Age at Menopause | % |
|---|---|---|
| 0 | still mensturating | 48 |
| 1 | less than 45 years | 9 |
| 2 | 45 to 49 years | 18 |
| 3 | greater than 49 years | 25 |

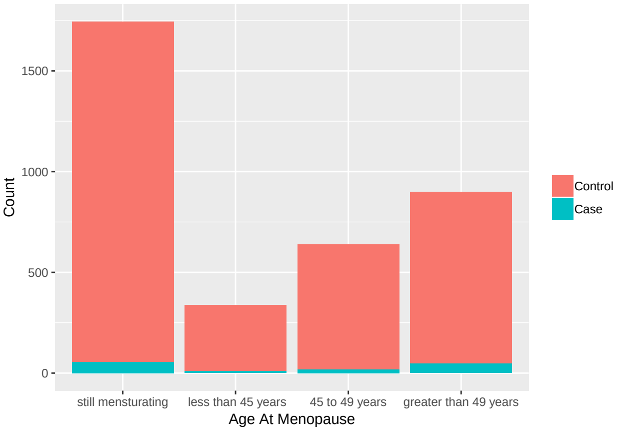


Figure 2: Bar graph of data associated with case and control according to age at menopause

Age at menopause and age at first childbirth were collapsed into categories based on a previous research paper, from Italy, containing datasets (E. et al. 1988) that were reviewed in the previous section. This was done because not all women in the study had experienced menopause or given birth. The categories created are summarized in the table and graphs below.

Table 2: Categorical division of age at first childbirth

| Category | Age at first birth | % |
|---|---|---|
| 0 | no children | 17 |
| 1 | less than 22 | 13 |
| 2 | 22 to 24 years | 20 |
| 3 | 24 to 28 years | 23 |
| 4 | more than 28 | 26 |

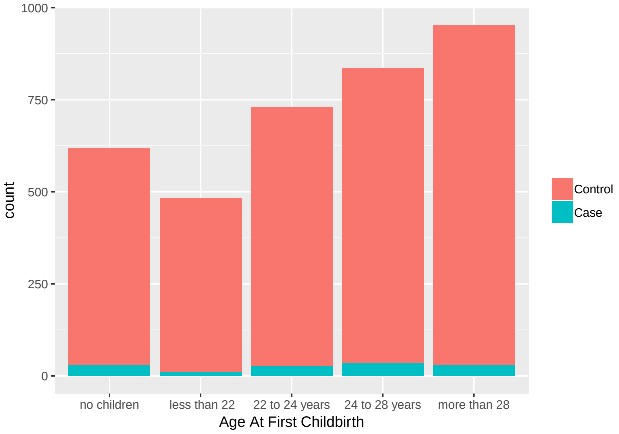


Figure 3: Bar graph of data associated with case and control according to age at first childbirth

Literature mentioned no difference in the relative risk posed by miscarriages, induced abortions and still births on breast cancer. They were included in the study since they affect the level of oestrogen in the body. In this data, each different complication had limited information available and many women had overlaps between two different complications. Thus they were clubbed together to analyse the overall impact of issues in pregnancy.

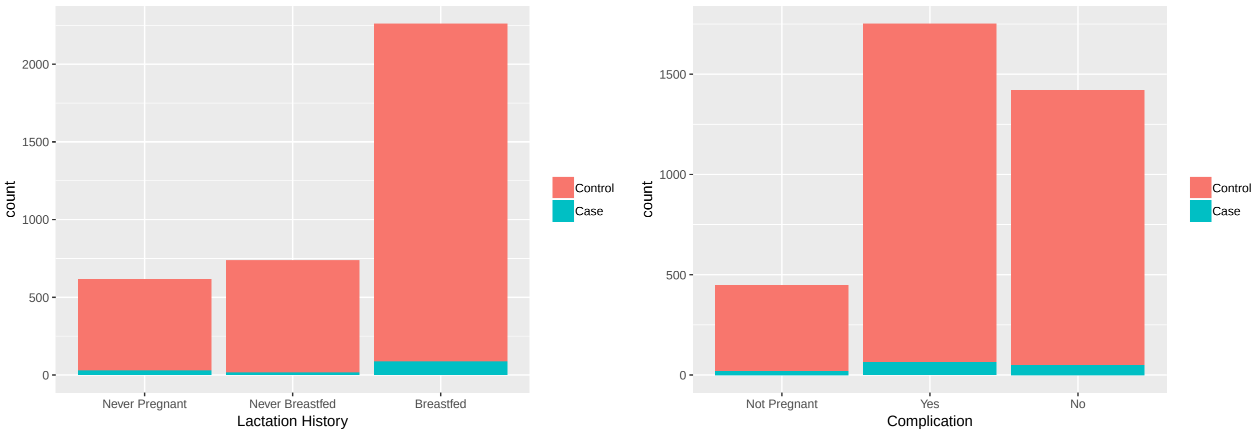


Figure 4: Bar graph of data associated with case and control according to complications in preganancy and childbirth

The dataset had separate variables to record lactation history for each pregnancy. This was combined into one variable with three categories; never been pregnant, not breastfed and breastfed.

The use of oral contraceptives had two separate variables associated with it. Both were merged to eliminate missing values.

Certain factors, such as age at menarche and use of hormone replacement therapy, were recorded in the dataset but could not be included in the analysis as they were collapsed into categories that were ambiguous. These factors have shown to heavily impact the risk of breast cancer and could therefore be potential confounders in the analysis (E. et al. 1988).

The final reproductive variables- under study- included menstrual status, age at menopause, whether or not the female had been pregnant, if any complication occurred with the pregnancy, whether or not the female had a child, age when the first child was born, use of oral contraceptives and lactation history.

#### 4.2.3 Lifestyle Factors

The lifestyle factors recorded include physical activity level, smoking status, alcohol consumption and impact of high cholesterol and high blood pressure. Other lifestyle variables that are said to impact the risk of breast cancer, such as the educational level, marital status and other such socioeconomic factors (Lundqvist et al. 2016), were also recorded but not considered in this study as they too were collapsed into undefined categories.

The EPIC study intricately designed the survey to include different levels of physical activity: occupational, leisure and exercise. Assessment included walking, cycling, gardening, light fitness regimes and vigorous physical exercise.

This holistic assessment is unlike most studies undertaken to establish a relationship between breast cancer and physical activity(Haftenberger et al. 2002). The level of physical activity level was classified into three categories; inactive, people who practiced less than 2 hours of leisurely physical activity; moderately active, people who engaged in more than 3 hours of physical activity; and the last group consisted of those who engaged in vigorous physical exercise.

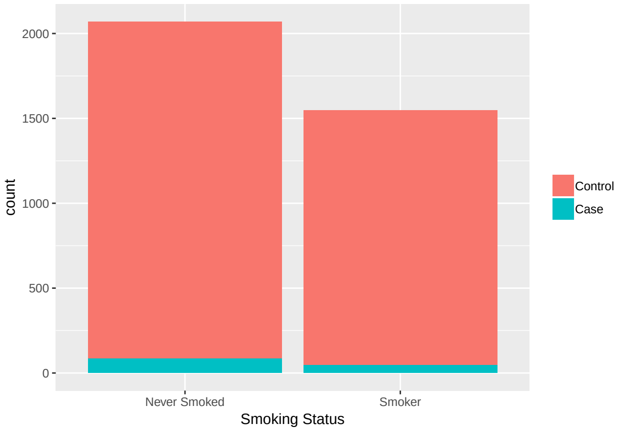


Figure 5: Bar chart comparing differences between case and control for smokers and non-smokers

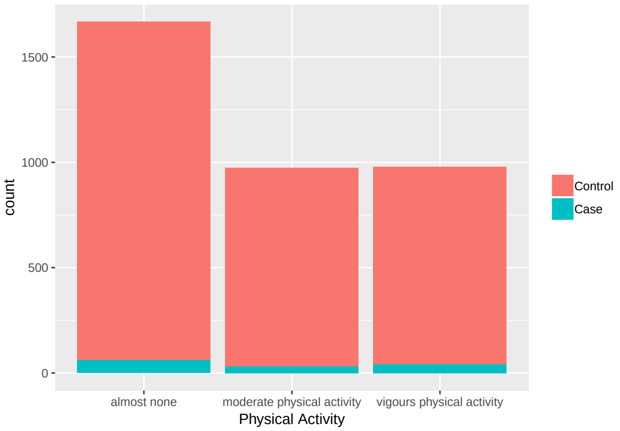


Figure 6: Bar graph of data associated with case and control according to physical activity level

Most studies, like the one reviewed above, do not indicate a difference between former and current smokers. Therefore, the categories in the data were collapsed to ever and never smokers.

Per day consumption of alcohol was calculated by direct summation of the amount of beer, spirits, red and white wine consumed.

The risks other illnesses impose on developing cancer are a fascinating area of research. Many illnesses were not included in this study due to the limited availability of cases; however, hyperlipidaemia and hypertension were considered as they are common.

#### 4.2.4 Nutritional and dietary information

One of the major goals of the EPIC study was to establish relationships between cancer and nutrition. For this purpose, dietary patterns in different countries were recorded according to cuisine. Quantity consumed was measured in portions per day and grams per day. This analysis uses the quantity in grams. Since the nutritional variables were very country specific it has been hard to establish any real relationships between diet and disease. Thus, there is a need to standardize the food measures into categories (Nadia Slimani et al. 2000). These categories were therefore determined using guidelines set by the Food and Agricultural Organization of the United Nations. In this case, the direct sums of quantity consumed per day have been classified into the following categories –

Table 3: Food items by classfied categories

| Category | Food Item |
|---|---|
| Cereal and Grains | Different types of pastas, Bread, Pizza, Rice etc |
| White Meat | Rabbit, Chicken, Veal |
| Red Meat | Pig, Beef, Liver, Offal |
| Seafoood | Different types of fish, Octupus, Prawns, etc. |
| Vegetables | Broccoli, Spinich, Pepper, Eggplant etc. |
| Dairy | All types of cheese, Milk, etc |
| Fruits | Oranges, Apples, Pear, Banana, etc. |
| Fatty Acids | All types of oils and butter eaten raw and used in cooking |
| Coffee | All different types of Coffee |

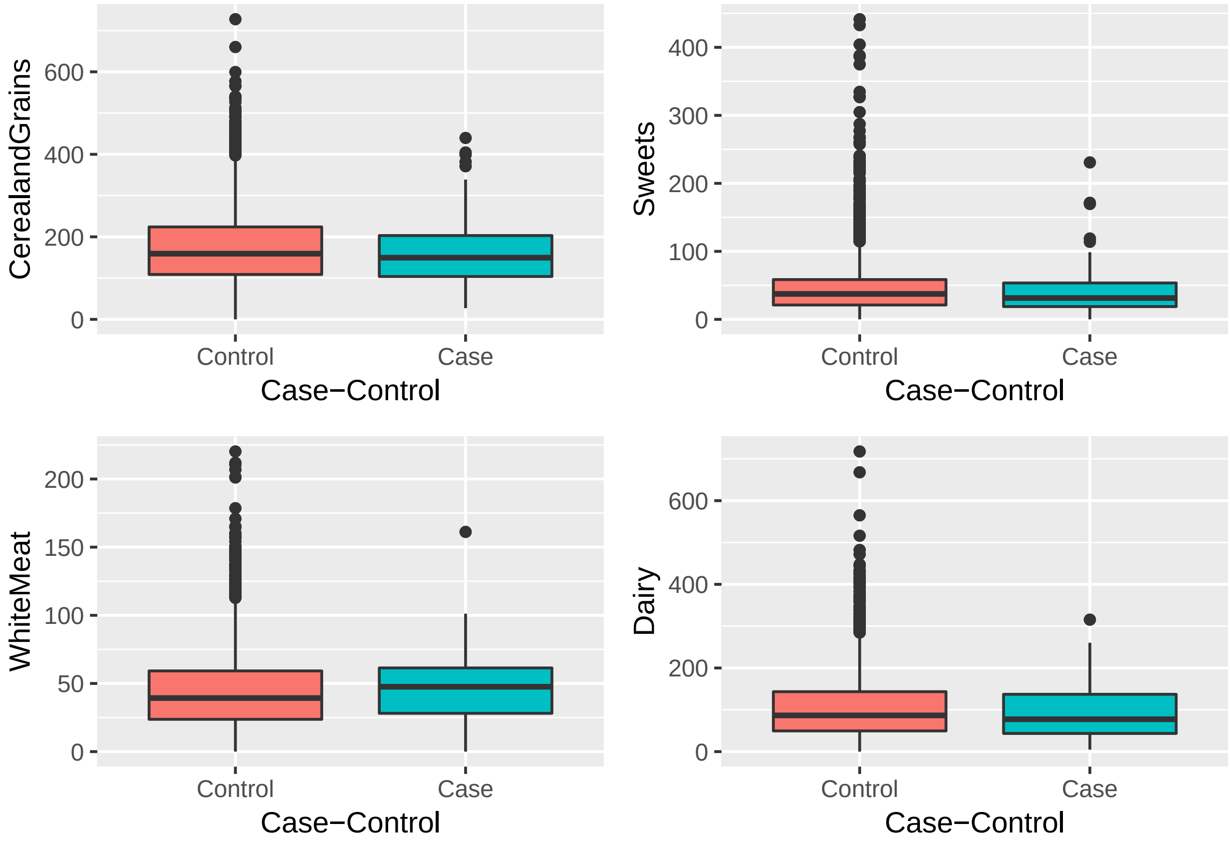


Figure 7: Boxplot comparing differences between case and control in dietary variables

# 5 Methodology

The final dataset of 3620 subjects, which comprised of 30 different factors with 3484 controls and 136 cases, was analysed. It is to be noted that 96% of the data was associated with controls while only 4% was associated with cases. Initially, a series of Mann Whitney U tests on the continuous variables and chi-square tests on the categorical variables were conducted to identify differences in the impact of each factor on the two groups: case and control.

The Mann Whitney U test hypothesizes that the two independent groups being tested belong to the same population, without assuming that the distribution of the population is normal. The method applied, tests the hypothesis by ranking all observations- of the variable being tested- in ascending order without acknowledging which group they belong to. In the next step, the ranks of the observations are averaged for each group separately and the difference between the two averages is calculated. Every hypothesis test has a corresponding p-value which is the probability of observing more extreme results than the current result. This means that if the mean rank of the two groups is different, then the p-value associated with the test statistic will be small [<0.05] and thus statistically significant.

The Chi Square statistic is used to test relationships between two categorical variables. A cross tabulation which summarises the number of observations in each pair of categories is computed. This test of independence compares the pattern in the crosstabulation with a pattern of responses that would have been observed if the variables were independent. A small p-value corresponding to a test statistic computed with the help of the equation 3 results in rejection of the null hypothesis that there is independence.

$$\chi^2 = \sum \frac{(f_0 - f_e)^2)}{f_e} \tag{3}$$

Here, $f_0$ is the observed values in the crosstabulation and $f_e$ is the expected values.

Research shows that the probability of finding a significant result increases with the increase in the number of tests being conducted; this is known as the problem of multiple testing (Sainani 2009).Caution needs to be observed while interpreting these results, as the chance of getting significant results is increased due to the number of tests being conducted. An easy fix to the problem of multiple testing is using the Holm's correction, which adjusts the level of significance of p-values according to the number of tests (n) being conducted by assigning a rank ($r_i$) in ascending order to the significance of each test as shown in the equation 4.

$$p - adjustedvalue_i = \frac{\alpha}{n - r_i + 1} \tag{4}$$

It is an advancement of the commonly used Bonferroni adjustment method that is simple to calculate but is low in statistical power(ETHZ 2012).

To find factors and their combinations that impact the risk of breast cancer, a two-method approach was adopted; the first was standardized linear regression and the second involved machine learning methods such as trees and random forests. The analysis was conducted in R-Studio(R Core Team 2018).

Every model is trying to differentiate between significant differences, observed in the outcome because of the factors under study, from the differences observed only due to individual variability. This process is often seen as separating signal from noise. The ideal model seeks to find the right complexity: to model only the signal and leave out the noise. Overly complex models exhibit low bias, since they have predictions close to the actual data they have been trained on. At the same time, they often tend to show high variance because predictions on unseen data are often wildly inaccurate. Overly simplistic models tend to exhibit high bias, as they often fail to capture the signal and they have low variance as predictions across datasets are similar. The goal is to find a model with low variance and low bias. The bias-variance trade-off had to be assessed, based on some error measurements, to achieve this parsimonious state. For this purpose, the data was divided into training and testing samples using proportional splitting: 80% of the data in the training set and 20% in the testing set. The training set was used to fit the model, and the validation set was used to measure the choice of error and evaluate model performance.

Binary outcomes are often modelled by calculating the probability that an observation belongs to each of the two given outcomes. This probability is then used to classify the observation into the two outcomes. Most often, the probability threshold for this classification is .5 for each outcome, although it can be altered according to the data requirements.

Since the outcome was binary, the confusion matrix 4 was used to calculate performance statistics. This table represents the outcome that was observed and the outcome that the model predicted.

Table 4: Confusion Matrix

| **Prediction** | **Truth** | |
|---|---|---|
| | Control | Case |
| Control | True Negative (TN) | False Positive (FP) |
| Case | False Negative (FN) | True Positive (TP) |

Performance measures that were used included

- Sensitivity – The number of times the model correctly predicted a case also known as the true positive rate.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

- Specificity - True negative rate; the number of times a model correctly predicted a control.

$$Specificity = \frac{TN}{FP + TN} \tag{6}$$

- Balanced Accuracy – Number of times the model was accurate in its prediction of both classes

$$Accuracy = \frac{Sensitivity + Specificity}{2} \tag{7}$$

- Precision – Rate at which the model predicted relevant cases out of all the cases

$$Percsion = \frac{TP}{TP + FP} \tag{8}$$

- Recall – Rate at which the model successfully identified cases

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

- F-measure – F1 score is a combination of precision and recall; it measures the accuracy of a model in predicting cases.

$$F - measure = (\frac{Recall^{-1} + Percision^{-1}}{2})^{-1} \tag{10}$$

- Receiver Operator Characteristic (ROC) - ROC curve is a plot of the sensitivity and specificity of the model at different probability thresholds. The area under the derived curve (AUC) is a measure of the accuracy of the model. It is a method of assessing the trade-off between the two measures and thus used for selecting the best threshold. The "best" threshold can be computed by various methods. For this analysis, pRoc (Robin et al. 2011) was used, which determines the best threshold by maximising sensitivity + specificity: also known as the Youden Index.

- AIC - Model selection, based on different parameters of the logistic model, was done by comparing Akaike's Information Criteria. AIC is a measure of the relative quality of the model. It is computed using the logliklihood and the number of parameters (npar) in the model. The logliklihood is explained in the section below.

$$AIC = -2 * loglikelihood * npar \tag{11}$$

- Generalized R-squared - generalized R-Squared was also computed for performance analysis of the regression models. This is calculated from the deviance estimates. Deviance is a numeric estimate of how different the selected model is from the saturated or perfectly fitted model, calculated with the help of the loglikelihood.

$$R^2 = \frac{1 - e^{D - D_N ULL}}{N} \tag{12}$$

*Note: D is the deviance of the model under consideration, $D_{NULL}$ is the deviance of the null model (without any covariates) and N is the number of observations used to build the model.*

While all measures are used in performance analytics the focus is on the measures that indicate that the model is good at identifying cases. Not all measures can be applied to all model types. The AIC and generalised R-Squared were used to measure the performance of the regression models. The other measures were applied to all models.

## 5.1 Linear Modelling Methods

### 5.1.1 Logistic Regression

The response under study had two possible outcomes, case and control, and hence models for proportions were considered. The number of cases was limited to the number of participants in the cohort study that had substantial data in each potential risk factor under investigation. Since participants in the study were selected randomly from the population, they are relatively independent of each other. Other things remaining constant, it can be assumed that the probability of observing cases will not change substantially. While this condition is not ideal, it helps in the construction of the logistic regression model that estimates the probability ($p_i$) of observing an outcome by finding the most likely value for the effect of each predictor/factor $\beta_i$ .

$$p_i = \frac{e^{\eta_i=\beta_0+\beta_i x_i}}{1 - e^{\eta_i=\beta_0+\beta_i x_i}} \tag{13}$$

The logistic regression model in R-Studio is fitted using an iterative algorithm to estimate the maximum-likelihood by solving the first order partial differential equation generated by equation 14 . This is computed on the scale of a link function, to ensure that probabilities stay in the range of the outcome [0,1]. There exist several link-functions, and use of each depends on the field of study and model performance. The log-link approach employed here solves the equation 14 to estimate the best values for the $\beta_i s$.

$$log(L) = \sum_{i=1}^{n}(y_i log(\frac{p_i}{1-p_i}) + n_i log(1-p_i) + log(\frac{n_i}{y_i})) \tag{14}$$

The use of the link function makes the interpretation of the coefficients complicated. Any statistically significant coefficient changes the log odds of observing a case instead of directly impacting the probability, i.e. it solves 15. Odds are calculated by dividing the probability of observing a case by the probability of not observing a case. The corresponding confidence intervals, which estimate the range of the true impact, are also computed as log odds. These log odds need to be exponentially transformed to estimate the magnitude of the impact a predictor has on the odds of observing a case. It is to be noted that an odds ratio (OR) >1 implies a negative relationship between the predictor, that is the predictor reduces the risk of breast cancer and the response whereas an odds ratio $<$ 1 implies the predictor increases the risk of breast cancer. If the OR is very close or equal to 1 there is no significant impact.

$$g(p_i) = log(\frac{p_i}{1-p_i}) = \eta_i = \beta_0 + \beta_i x_i \tag{15}$$

The significance of a predictor is determined by the use of Wald's z- test. To find these significant $\beta_i$, variable selection is performed using step selection which removes and adds variables to the model according to their impact on the AIC. All possible subsets selection which considers all possible models that can be generated by including and excluding predictors on the basis of the AICc, which is a small sample equivalent of the AIC, was also used. Elastic net regression is also used for variable selection, which is explained below.

The model was first built using all 30 predictors since they were carefully selected through the literature review. Since not all predictors were significant, step selection was performed to identify the most relevant results. All possible subsets for the stepped model were computed, but the best model, as indicated by this method had a lower adjusted r-square, thus a worse fit. Thus, the stepped model was retained.

A second approach to identify significant predictors was to use the penalised approach, explained below, since it automatically selects relevant covariates. At this stage, this method was applied for the sole purpose of identifying relevant covariates. It identified 17 significant predictors. These were then refitted using logistic regression, and the stepwise selection was performed again. This model had the lowest AIC and hence, was chosen as the best logistic predictor without interactions.

The next step was to identify relevant interactions between covariates. For this purpose, all possible interactions were explored. The built-in R function encountered issues while fitting this model. The algorithm did not converge with the pre-set number of the maximum number of iterations conducted while trying to find the optimal loglikelihood. This was increased to 100, and the algorithm converged, but the function still was not able to handle 659 covariates. Thus, covariate all interactions were explored using the penalised regression method explained below.

Literature indicated the importance of certain uncontrollable factors such as age, menopausal status and pregnancy. Many studies that are reviewed above concentrate on how these factors interact with others. Thus, models including only interactions of these factors were also constructed and penalised.

Logistic models are fitted, making certain assumptions about the data under consideration. These include linearity between the response and the predictor, independence of observations from each other, a specified mean-variance relationship and no collinearity between predictors. These are explained below.

In this case, instead of assuming linearity between the response and predictor directly, it is assumed on the scale of the link function. This can be assessed with the help of effect plots.

The mean-variance relationship can be assessed graphically as the model assumes the mean and variance as follows:

$$\mu = np \tag{16}$$

$$V(\mu) = np(1-p) \tag{17}$$

The graph should, therefore, show no patterns in order to meet this criterion.

The model also assumes independence of observations and hence the residuals, which can easily be assessed by the runs test and acf plots.

Finally, often multicollinearity between predictors is an issue while estimating the coefficients $\beta_i$ as they become unreliable for interpretation. This is because, if two predictors are correlated, their corresponding coefficients can provide the same probability estimate taking on different values. For example, say $x_1$ and $x_2$ are correlated, $\beta_1$ and $\beta_2$ are their estimated coefficients, they could take on the values 4 and 1 or 3 and 2 but would still provide the same final probability estimate. This issue was expected in modelling such a study, as many factors are related to each other.

### 5.1.2 Penalised Logistic Regression

To overcome any issue of collinearity a penalised regression approach was adopted to shrink the coefficients and simultaneously perform variable selection to find statistically significant predictors. This is done by introducing a shrinkage parameter ($\lambda$) in the likelihood estimation process.

$$-log(L) + \lambda[\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|] \tag{18}$$

*Note: when $\alpha = 0$, this equation is reduced to ridge regression and when $\alpha = 1$, it is lasso.*

There are three types of regularisation methods that can be used to estimate coefficients with this equation 18: the ridge method, the least absolute shrinkage and selection operator (LASSO) and the elastic net. Each method has certain advantages and disadvantages. The ridge method although provides stable results struggles to achieve parsimony. The lasso method provides variable selection but fails to identify groups of related covariates and struggles to deal with the presence of a large number of covariates. Finally, the elastic net, a combination of both ridge and lasso removes most limitations faced by both estimators. Although it sometimes struggles to perform well if it is not close to either the lasso or ridge method (elasticity depends on the value of alpha), it outperforms them in most situations (H. Zou and Hastie 2005) and thus is the method employed here.

The glmnet package in R is used to fit the models (Simon et al. 2011). This method uses quadratic approximation to find the optimum likelihood estimate. Quadratic approximation expands the likelihood (log(L)) part of the equation 18 with the help of Taylor series of the second order and iteratively finds the optimal solution for it. Then coordinate descent is used, which successively optimises the equation 18 over each parameter, keeping others fixed. This process is repeated until the penalty term in the objective function improves the likelihood at a negligible rate.

The best estimate for the penalty factor can be calculated using k-fold cross-validation. This method divides the available dataset into k parts, to train the model. Each part is left out once during the training process to evaluate the performance of the model. In this study, 3-fold cross validation was used to find the best value of the penalty factor using area under the ROC curve as the performance measure. The best estimate chosen is the minimum of all the lambdas computed.

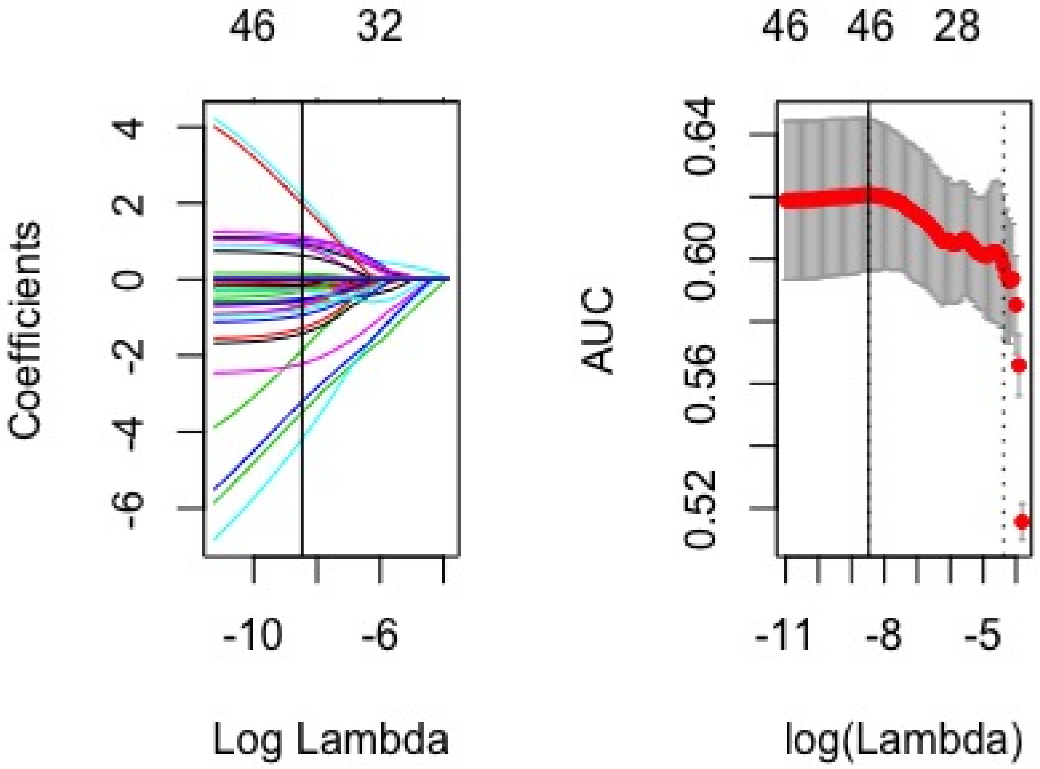


Figure 8: Variable Selection with penalised regression

The 659 covariates including all possible two-way interactions were simultaneously provided to the algorithm, initially keeping the elasticity as $\alpha 0.5$. The alpha level was adjusted multiple times, [$\alpha =$ 0.4,0.6,0.7] to get the method close to ridge and lasso. The most interpretable model was at $\alpha = 0.5$; it also had the best estimate for the generalised r-squared.

A significant drawback to this method is that although the variance of the estimate is small across datasets, it is at the cost of bias. Predictions vary vastly from dataset to dataset. The bias also makes confidence intervals redundant since the range of the actual estimate is small and hence gives unrealistic confidence in the results. Another problem is that this method can be sensitive to outliers.

### 5.2 Data Mining Methods

Data mining methods don't use assumption driven approaches. They are sensitive to class imbalance, which is one of the crucial statistical problems with cohort studies. The number of breast cancer cases in the population under study is bound to be lower than the number of controls. Thus, the classifier might have insufficient information about the minority class to achieve accuracy or achieve high accuracy rates by always predicting the majority class (Lemnaru and Potolea 2012). Although class imbalance has a different impact on different classifiers, it is hard to identify the magnitude of this effect. The extent of the impact depends on many factors such as the degree of imbalance, sample size and complexity of concepts under study (Japkowicz and Stephan 2002). Different methods of training models were employed to overcome this issue. They included oversampling, downsampling, cost-sensitive training, and SMOTE(Kabir and Ludwig 2019). The unimputed sample was used in this study to evaluate model performance.

- Oversampling and downsampling - Oversampling involves randomly sampling as many samples of cases as there are controls. Though the method is simple, it can often lead to overfitting. Downsampling is a method of randomly under-sampling the majority class to match the number of cases. Downsampling leads to the unavoidable loss of data.
- SMOTE – The synthetic minority oversampling technique looks for the k-nearest neighbour of all cases and randomly selects a reference point. It then computes the difference between the two neighbours and multiplies it by a random number between [0,1] and adds it to the initially chosen point, to keep the sample generated between the neighbours. In this way, it oversamples cases and under-samples controls. A potential drawback of this approach is the generation of the same number of synthetic points for each minority class without considering the neighbours already computed and thus increasing the probability of class overlap (Chawla et al. 2002).
- Cost-sensitive learning – An algorithm level approach towards classifying imbalanced data: by highlighting the class imbalance problem for the classifier. Class imbalance is highlighted by assigning a cost matrix to the corresponding confusion matrix (Domingos 1999).

The classifier is trained to minimize the total cost depending on the misclassification error.

Table 5: Cost Matrix

| **Prediction** | **Truth** | |
|---|---|---|
| | Control | Case |
| Control | 0 | C(FP) |
| Case | C(FN) | 0 |

$$TC = C(FN) * FN + C(FP) * FP \tag{19}$$

In this problem, C(FN) > C(FP) since it is more important to assess the risk each predictor imposes on the possibility of developing cancer. Controls predicted as cases can be indicators of a warranted change in lifestyle and diet. Nevertheless, caution is applied to limit the number of false positives.

The caret package was used to build all such models as it provides an easy way to choose model parameters. The package calls on rpart (Therneau and Atkinson 2018) to grow trees and RandomForests (Liaw and Wiener 2002) to construct the random forests. Optimal model parameters are identified using repeated k-fold cross-validation. Cross-validation is repeated by resampling from the data that is used to build the model. Thus, providing sufficient evidence to support the choice of model parameters.

### 5.2.1 Classification and Regression Trees

Classification and Regression Trees (CART) are built with the help of recursive binary partitioning. CART works by diving the data into subsets based on the given factors. The best factor to split on is determined by the accuracy with which it separates the binary outcome, that is, by measuring the impurity with either the Gini index 20 or entropy 21. Impurity is the measure of the degree of separation; the more distinct the separation, the purer the node.

$$Gini = 1 - \sum_{j=1}^{n} p_j \tag{20}$$

$$Entropy = -\sum_{j=1}^{n} p_j log(p_j) \tag{21}$$

Entropy is computationally costly as it uses logarithms and also tends to make more complex trees, whereas the Gini index is easier to compute and often less complex. Both measures have been observed equally competent in choosing splits for portioning the data (Raileanu and Stoffel 2000). Thus, the Gini index was selected for this analysis. After the initial split is chosen, the same process is repeated in each subset created, and more nodes and leaves are added to create a classification tree.

The tree grown is then pruned to decrease the complexity and avoid overfitting. Pruning reduces the complexity of the tree by penalizing the number of terminal nodes (T), used in building the tree, with the help of a cost complexity parameter ($\alpha$) and training error (R(T)).

$$R(T)_\alpha = R(T) + \alpha|T| \tag{22}$$

When no cost is imposed on the tree [$\alpha = 0$], it tries to imitate the training set. This leads to poor performance on the testing set. To avoid this issue and build a parsimonious model, the cost complexity ($\alpha$) is varied with the help of 5-fold cross-validation on the training set– repeated five times. The area under the ROC curve is used to set the best value of $\alpha$.

The trees were fitted using five different methods. The first two were focused on cost-sensitive methods to incorporate the importance of minimizing false negatives while training/tuning the tree. The first of these was the use of the generalized Gini index by incorporating a loss/cost function while choosing the best splits. For this analysis, the cost matrix had been specified according to the class imbalance [0.96,0.04]. The other approach involved altering the probability used to calculate the Gini index; this incorporates the disbalance in classes only by affecting the choice of splits. The probability was set to the imbalance ratio, which is .04 and 0.96. The final three approaches were the sampling methods – downsampling, upsampling and SMOTE.

An advantage of training the tree is that it uses surrogate splits to deal with the missing values. These surrogate splits are created by using each independent variable to create binary partitions for the variable with the missing value. The optimal split criteria and the misclassification rates are computed for each predictor, and the final prediction for the missing value is the prediction from the best surrogate variable.

The use of this technique allows all variables being studied to be considered simultaneously. It uses the goodness of split measures for each split criteria to determine the most relevant factors that help grow the tree. Since most of the training methods are focused on predicting cases, it can be inferred that the important variables are protentional risk factors. Disadvantages of the method include the difficulty in interpreting complex trees that are built as a result of cost-sensitive training. Although this makes interpreting the output as interactions difficult, the final goal is still achievable as the importance of each variable separately is also equally essential. Another drawback of this method is that it does not explore independent relationships of variables with the outcome, only complex interactions (Therneau and Atkinson 2018).

#### 5.2.2 Random Forests

An ensemble of classification or regression trees is known as a random forest. Essentially, the algorithm samples the observations with replacements and tries out a given number of variables (mtry) to pick the best split to begin growing the tree. Subsequently, the chosen attributes are sampled without replacement at each node to grow the tree. The depth of the tree is determined by the minimum number of observations required to split each node and number of terminal nodes, which is left unrestricted in the building process. In this way, n number of trees are grown. The prediction made by a tree individually is considered a vote. The final prediction is the prediction made by the majority of the trees.

How well the random forest performs depends on error rates which are determined by the performance of each tree in the forest. These two tend to move in opposite directions. The aim is to find the optimal level of trade-off between them.

Two methods determine the importance of variables used in the construction of the random forest. The first is averaging the decrease in Gini index induced by splitting the tree on the variable. The second method is by using the out of bag data – the data not used while resampling to calculate the misclassification error. It also computes the misclassification after sampling the attributes at each split. The difference between the two errors is averaged over all trees and normalized to calculate the importance.

While random forests come with several benefits such as providing the importance of each variable and being easy and fast to compute, the model is almost uninterpretable. It also struggles with the class imbalance in the data, and the output is hard to explain. Thus, the balanced and SMOTE sampling approaches are used to train random forests.

The balanced random forest uses stratified samples with replacement to implement the algorithm explained above. Balancing entails sampling as many controls as there are cases. It is usually considered a naïve way of fixing the imbalance problem but is shown to perform well in many instances (C. Chen, Liaw, and Breiman, n.d.).

To determine the number of variables to be tried at each node, several models were trained. The best model was selected with the help of the area under the ROC curve. The same was then repeated for the number of trees in the ensemble (Liaw and Wiener 2002).

# 6 Results

## 6.1 Initial Testing

The initial Mann Whitney U and Chi-square test indicated that weight, BMI, age, white meat consumption, sweet consumption, age at menopause and lactation history could be important indicators of the risk of breast cancer as seen from 6. These results became insignificant after adjusted for multiple testing. The results from the tests on the other variables can be found in the appendix. 11.2.

Table 6: Results of the Mann Whitney U and Chi-Square Tests

| | Test Statistic | p-value | Adjusted p-value |
|---|---|---|---|
| Weight | 207336.00 | 0.01 | 0.42 |
| Age | 207942.00 | 0.02 | 0.44 |
| BMI | 212907.50 | 0.05 | 1.00 |
| WhiteMeat | 210293.00 | 0.03 | 0.71 |
| Sweets | 261977.00 | 0.04 | 0.94 |
| AgeAtMeno | 9.32 | 0.03 | 0.71 |
| BreastFeed | 6.27 | 0.04 | 1.00 |

### 6.2 Logistic Regression

The best model identified the age at menopause, lactation history, cigarette smoking, sweet consumption, cereals and grains consumption and white meat consumption as the most significant contributors to the risk of developing breast cancer. It indicated a marginally higher risk of developing breast cancer for women who experienced menopause after the age of 49, than women who had not yet hit menopause. Women who had been pregnant and had not breastfed had lower odds of developing breast cancer than women who had not been pregnant. Although the other covariates were included in the model, they were not statistically significant. The odds ratios associated with these results with the confidence intervals are given in the table 7 below. When compared with the model built in the same way on the unimputed data, the same covariates were identified as significant. The raw output for this model can be found in the appendix 11.4.

Table 7: Effect of predictors on the odds ratio

| | Coffiecients | 2.5 | 97.5 |
|---|---|---|---|
| Baseline | 0.09 | 0.04 | 0.18 |
| | | | |
| **Age At Menopause** | | | |
| less than 45 years | 0.97 | 0.44 | 1.92 |
| 45 to 49 years | 0.71 | 0.36 | 1.29 |
| greater than 49 years | 1.64 | 1.05 | 2.54 |
| | | | |
| **Lactation History** | | | |
| Never lactated | 0.37 | 0.19 | 0.70 |
| Lactated | 0.55 | 0.35 | 0.87 |
| Smoker | 0.71 | 0.47 | 1.07 |
| White Meat | 1.01 | 1.00 | 1.01 |
| Sweets | 0.99 | 0.99 | 1.00 |
| Cereal and Grains | 1.00 | 1.00 | 1.00 |

The model performance was measured with the help of the ROC.

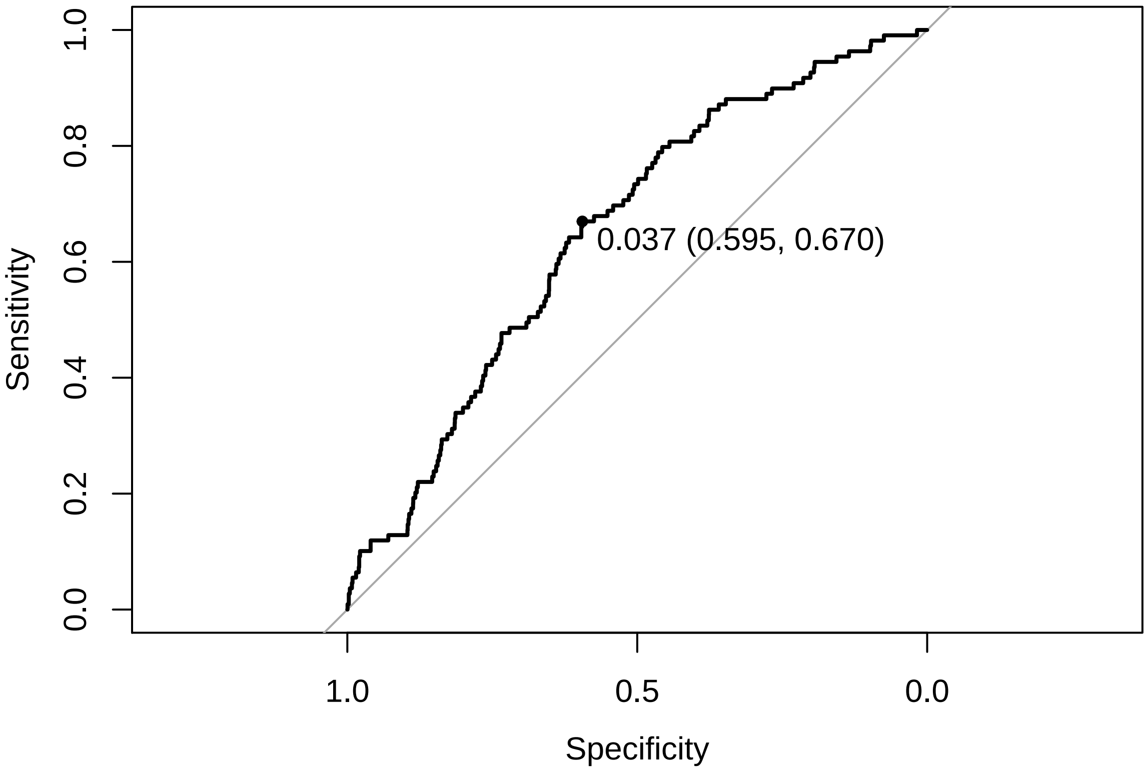


Figure 9: ROC curve for final model without interactions to determine best probability threshold and visualise model performance

Table 8: Training statistics for model without interaction

| Percision | Recall | F_Measure | Sensitivity | Specificity | Balanced_Accuracy |
|---|---|---|---|---|---|
| 0.06 | 0.63 | 0.11 | 0.63 | 0.62 | 0.63 |

Models with interactions were fitted but had unreliable coefficient estimates due to possible collinearity between predictors and thus were subjected to the penalised approach. These are summarized in the appendix 11.5.

### 6.3 Penalised Regression

The model exploring all possible interactions, fitted with the help of the regularization methods, identified many significant interactions and estimated their corresponding covariate. They included the interactions between pregnancy and seafood; age at menopause and complication; age at first childbirth and physical activity; age at first childbirth and high cholesterol; breastfeeding and sweet consumption; cereal and grain consumption and dairy consumption; complications in pregnancy and dairy consumption; and lastly cigarette smoking and sweet consumption. The model also identified the individual significance of weight and coffee consumption. The effect of each factor on the odds is summarized in table 9 and appendix 11.5.5. Some interactions were not practically significant or had little to no theoretical backing to them.

Table 9: Effect of predictors on the odds ratio

| | Odds |
|---|---|
| Baseline | 0.02 |
| **Age At Menopause - No Complication** | |
| less than 45 years | 8.75 |
| 45 to 49 years | 2.63 |
| **Age At Menopause - Complication** | |
| less than 45 years | 7.27 |
| 45 to 49 years | 0.03 |
| greater than 49 years | 0.58 |
| **Age At First ChildBirth - Moderate Physical Activity** | |
| less than 22 | 0.02 |
| 22 to 24 years | 0.51 |
| 24 to 28 years | 0.24 |
| more than 28 | 0.28 |
| **Age At First ChildBirth - Vigorous Physical Activity** | |
| less than 22 | 0.74 |
| 22 to 24 years | 0.39 |
| 24 to 28 years | 0.45 |
| more than 28 | 0.11 |
| **Age At First ChildBirth - Hyperlipidemia** | |
| less than 22 | 0.04 |
| 22 to 24 years | 0.75 |
| 24 to 28 years | 3.07 |
| more than 28 | 1.87 |

Given that a woman went through menopause before the age of 45 and has experienced either a miscarriage, induced abortion or given birth to a stillborn baby, from the table above it can be inferred that her odds of developing breast cancer have increased. Women who had children before they were 22 years old and faced a complication in pregnancy, had a higher chance of developing breast cancer than women who had never been pregnant. The odds of women who had given birth also decreased according to the level of physical activity. Finally, women who suffered from hyperlipidemia had a reduced risk of breast cancer if they gave birth before the age of 24 and an increased risk if they gave birth after 24 years. Since age at first birth was identified as an extremely significant interaction area, a model 11.5.4 was constructed including all possible interactions of this covariate that indicated similar relationships.

The model performance statistics are summarised below.

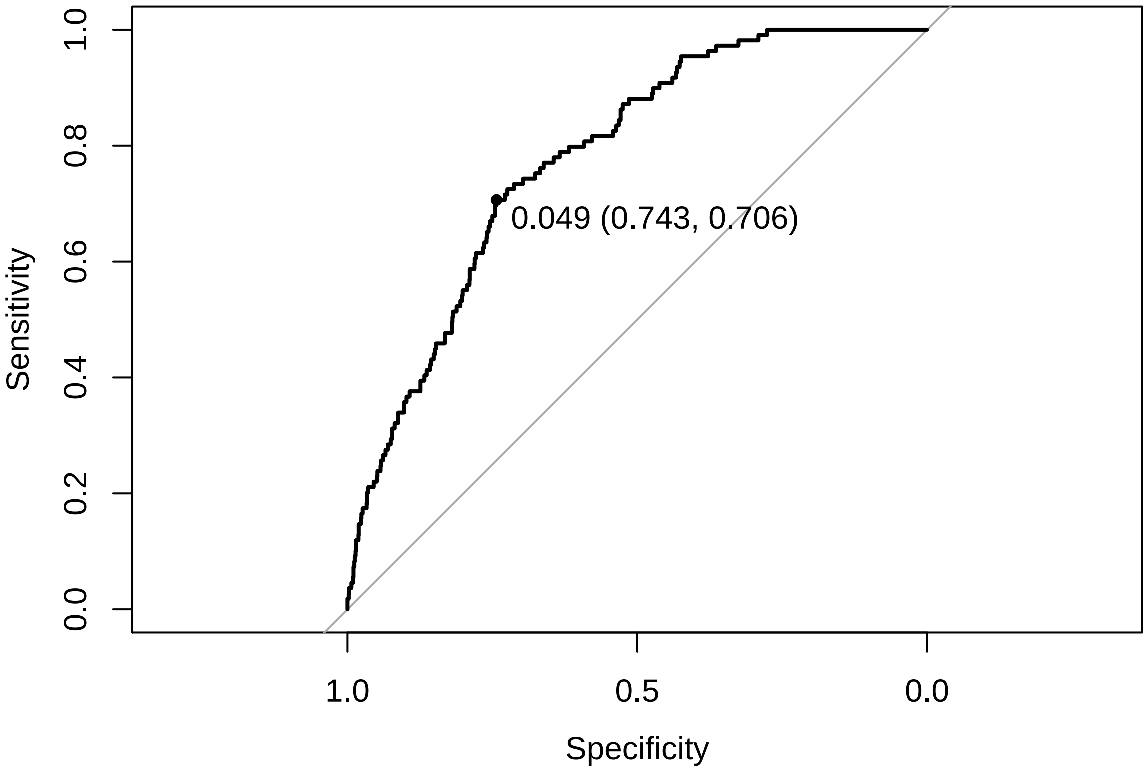


Figure 10: ROC curve for final interaction model to determine best probability threshold and visualise model performance

Table 10: Training statistics for Penalised Regression Model with all possible interactions

| Percision | Recall | F_Measure | Sensitivity | Specificity | Balanced_Accuracy |
|---:|---:|---:|---:|---:|---:|
| 0.1 | 0.71 | 0.17 | 0.71 | 0.74 | 0.72 |

Models were also constructed exploring interesting interactions as indicated by the literature. The first of these was the differences in pre and post-menopausal women. This model identified 18 significant covariates out of which were 5 significant interactions but not all of them were practically significant [output in appendix 11.5.1. Postmenopausal women had a lower probability of developing breast cancer than premenopausal women [OR:0.04]. There were indications that post menopause, women who had never lactated but had children had significantly higher odds [OR:5.3] of developing cancer than women who had never lactated because they had not been pregnant. Lastly, postmenopausal women who smoked cigarettes at some point in their lives had lower odds of developing cancer [OR:0.54] than those who had not.

The model identifying pregnancy and the risk of breast cancer found that for women who had been pregnant, if they suffered from hypertension, they had a higher odds of developing breast cancer [OR:5.51] than women who had been pregnant but not suffered from hypertension. Pregnant women who suffered from hyperlipidemia had lower odds of developing breast cancer than women who did not have hyperlipidemia [OR:0.29]. The other interactions were not practically significant [Appendix 11.5.2] .

### 6.4 Classication Tree

Table 11: CART: training statistics

| | Percision | Recall | F_Measure | Sensitivity | Specificity | Balanced_Accuracy |
|---|---|---|---|---|---|---|
| Loss | 0.23 | 0.95 | 0.38 | 0.95 | 0.88 | 0.92 |
| Prob | 0.08 | 1.00 | 0.14 | 1.00 | 0.53 | 0.76 |
| Downsample | 0.06 | 0.79 | 0.11 | 0.79 | 0.52 | 0.65 |
| Upsample | 0.08 | 0.82 | 0.15 | 0.82 | 0.63 | 0.73 |
| SMOTE | 0.13 | 0.28 | 0.18 | 0.28 | 0.92 | 0.60 |

The classification tree models were compared based on the statistics mentioned above, some seemed to be overfitted, and hence, the testing errors were compared instead. Although the models did not perform as well as expected on testing data, the model built by under-sampling the data had the best performance statistics. This tree had 25 leaves out of which 13 were terminal nodes.

Table 12: CART: testing statistics

| | Percision | Recall | F_Measure | Sensitivity | Specificity | Balanced_Accuracy |
|---|---|---|---|---|---|---|
| Loss | 0.04 | 0.15 | 0.06 | 0.15 | 0.84 | 0.50 |
| Prob | 0.03 | 0.41 | 0.06 | 0.41 | 0.53 | 0.47 |
| Downsample | 0.04 | 0.52 | 0.07 | 0.52 | 0.51 | 0.51 |
| Upsample | 0.05 | 0.48 | 0.09 | 0.48 | 0.62 | 0.55 |
| SMOTE | 0.07 | 0.15 | 0.09 | 0.15 | 0.92 | 0.53 |

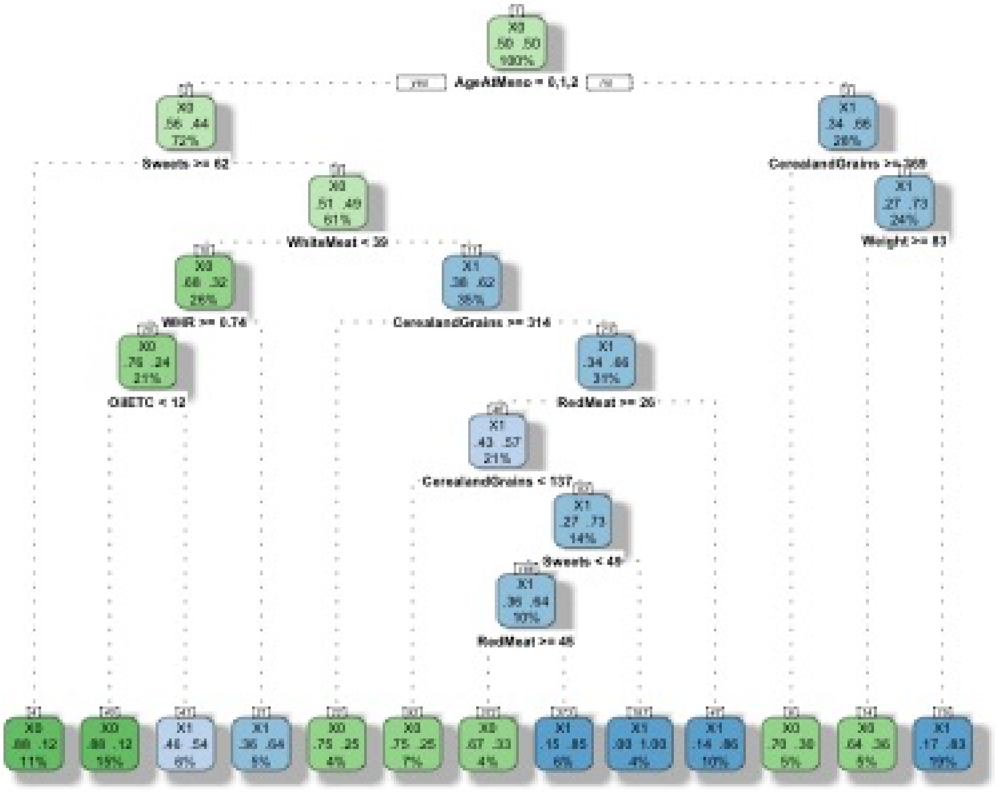


Figure 11: Classification Tree: Green nodes predict controls and blue nodes predict cases

From figure 11 it can be concluded that age at menopause, cereal and grain consumption, sweets consumption, white meat consumption, weight, waist to hip ratio, red meat consumption and fatty acids consumption are important predictors of breast cancer. The variable importance plot 12 also indicates the importance of hip circumference, body mass index, age at first birth, complications in pregnancy, age, vegetable consumption, seafood consumption, dairy consumption and menstrual status.

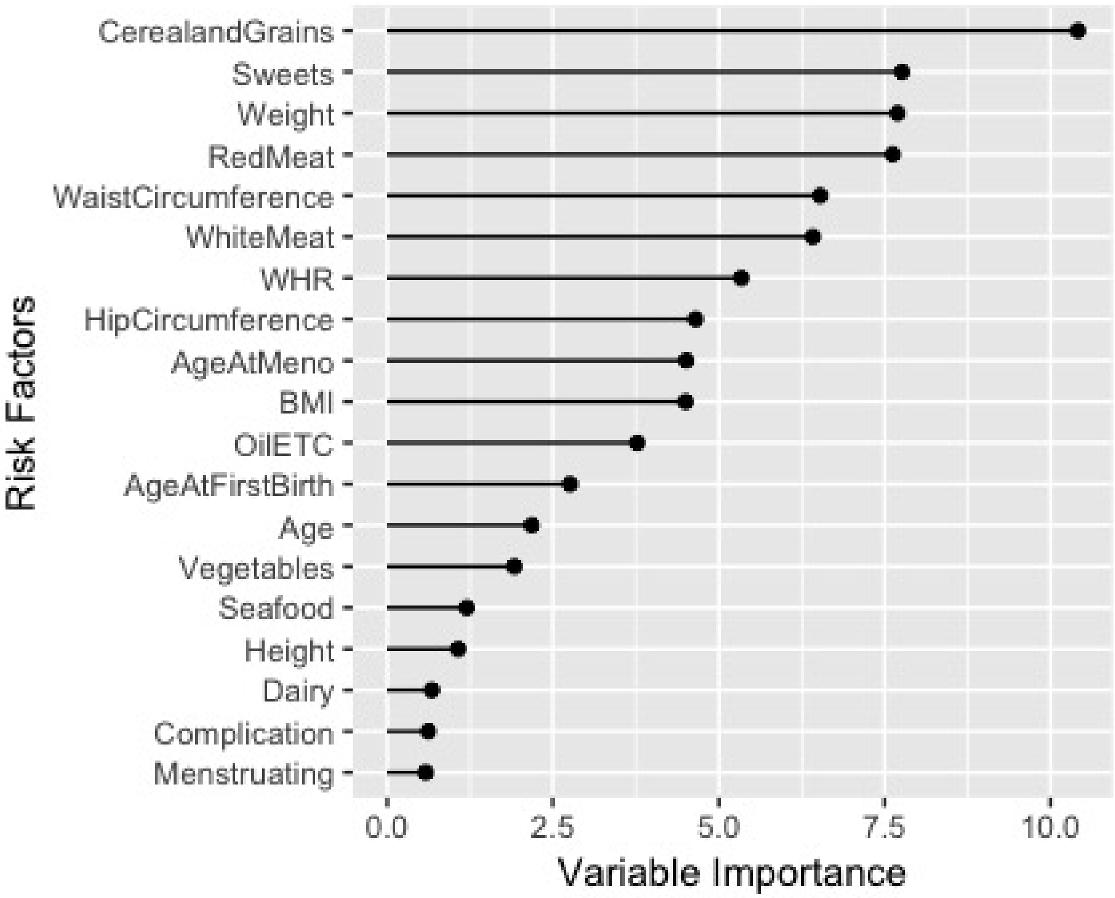


Figure 12: Classification Tree: Variable Importance

Each node in the tree in figure 12 indicates the probability of observing a case or control and the percentage of the total data associated with the split. The prediction is made by dropping a vector down the tree. The observation is classified into case or control according to the interaction between the different attributes. For example, a woman who experienced menopause after the age of 49 has a higher probability of developing breast cancer but this probability changes if she consumes more than 269g of cereal and grains.

### 6.5 Random Forest

The best random forest decided on the basis of the testing error had the following model measures statistics in tables 13 14. This was built using 1000 trees and mtry 1 with the help of stratified sampling. Variable importance of each factor was accessed with two methods. These are plotted in the figure 13.

Table 13: Random Forest: Training performance measurements

| | Percision | Recall | F_Measure | Sensitivity | Specificity | Balanced_Accuracy |
|---|---|---|---|---|---|---|
| Balanced 500 | 0.46 | 1 | 0.63 | 1 | 0.95 | 0.98 |
| Balanced 1000 | 0.27 | 1 | 0.42 | 1 | 0.89 | 0.95 |
| SMOTE 500 | 0.41 | 1 | 0.58 | 1 | 0.94 | 0.97 |
| SMOTE 1000 | 0.40 | 1 | 0.57 | 1 | 0.94 | 0.97 |

Table 14: Random Forest: Testing performance measurements

| | Percision | Recall | F_Measure | Sensitivity | Specificity | Balanced_Accuracy |
|---|---|---|---|---|---|---|
| Balanced 500 | 0.02 | 0.04 | 0.03 | 0.04 | 0.94 | 0.49 |
| Balanced 1000 | 0.05 | 0.15 | 0.07 | 0.15 | 0.88 | 0.51 |
| SMOTE 500 | 0.02 | 0.04 | 0.03 | 0.04 | 0.94 | 0.49 |
| SMOTE 1000 | 0.02 | 0.04 | 0.03 | 0.04 | 0.93 | 0.48 |

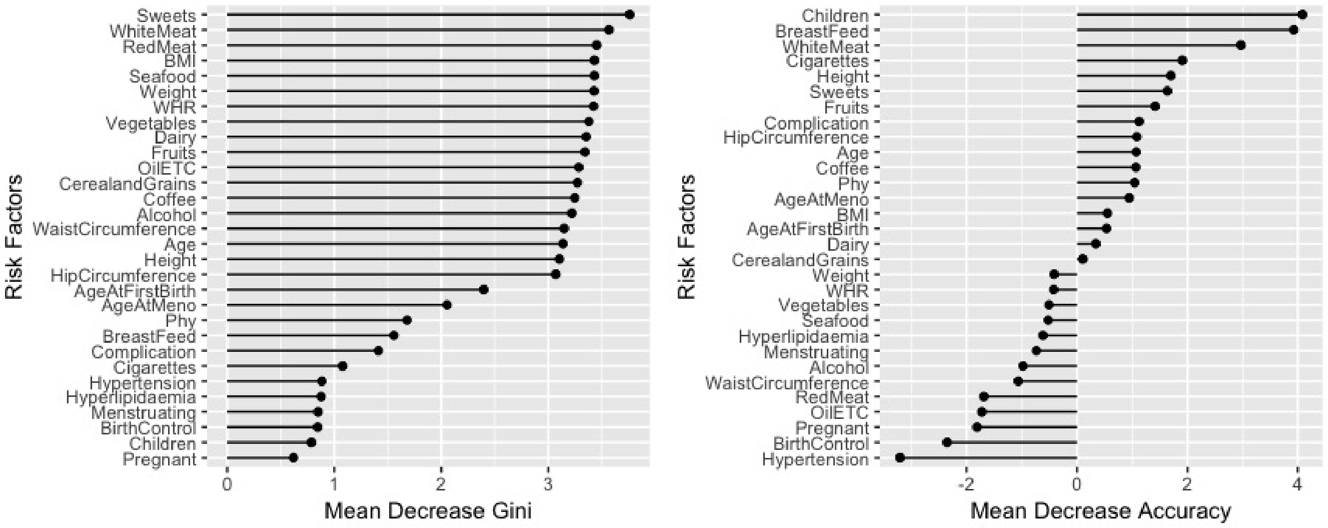


Figure 13: Random Forest: Variable Importance

The most significant predictors of breast cancer were selected based on the number of models that indicated their importance or how important their interactions with other variables were deemed. These included age at menopause, age at first childbirth, and lactation history.

# 7 Discussion

A systematic assessment of the final predictors is undertaken to assess the relevance of the findings in relation to existing research and how they hold across different models built for this study.

Literature advocates the vast difference between risk factors for breast cancer in premenopausal and postmenopausal women. Therefore it is expected that factors related to menopause will interact with other factors to impact the risk of breast cancer. Earlier research shows that women who go through menopause at a young age have a relatively low risk of developing breast cancer. A similar relationship is identified in this study as women who experienced menopause after the age of 49 had a significantly higher probability of developing breast cancer. Importance of this factor in determining the risk of breast cancer is reiterated in the construction of all logistic models. Data mining methods also indicated the importance of this relationship. The complexity of the interactions between the age at menopause and other variables was evident during the construction of the tree model. Thus an interesting area of further research would be to study different age groups separately.

While age at first childbirth was not significantly related to breast cancer risk stand-alone, its interactions seem to have a statistically significant impact.

Complications faced in pregnancy, hyperlipidaemia and physical activity level were identified as significant predictors in accordance to age at first childbirth. The interaction between age at first birth and complications faced during pregnancy could be attributed to the changes in oestrogen levels due to these conditions. Often pregnancy leads to weight gain that can be controlled by physical activity; thus, this relationship could potentially reflect that factor. Finally, hyperlipidaemia is found to decrease the risk of breast cancer, and pregnancy has found to increase the risk of hyperlipidaemia the models constructed could be identifying this complex relationship.

The importance of lactation history has been researched thoroughly, and the findings from this study align with previous research of an increased risk of breast cancer for women who have breastfed. They also align with the theory of breast cancer epidemiology being significantly different for women who have been pregnant and those who have not.

Finally, the models constructed also showed the slight inverse impact of sweet consumption and the direct impact of cereal and grain consumption. Anthropometric measures were also proven to be important by the data mining methods, but it is hard to interpret the direction of any impact these measures pose through such modelling methods.

A few shortcomings of this study include the low power of the tests conducted here due to the limited data attached to cases. Outliers are also a problem in the computation of the continuous variables. They have a high probability of aliasing relationships in the data. Outlier analysis would be an interesting area to explore further. Imputation of data could also introduce bias in the study since it makes strong assumptions about unknown subjects. Finally, many of the models constructed above, like most cohort studies, are bound to have poor generalization error.

# 8 Conclusion

To determine the factors that impact the risk of breast cancer, logistic and penalised regression and data mining approaches were used. Although the tests conducted are low in power due to the limited number of cases under study, conclusions made based on the information available aligned with previous research. The most significant relationships identified were those exploring the relationship between reproductive factors and breast cancer. Age at menopause has a direct relationship with breast cancer development. Age at first birth intricately interacts with other factors to determine the risk of breast cancer. Breastfeeding and pregnancy are positively related to breast cancer.

Among other relationships that should be explored with further research, is the increased consumption of cereal and grains and an increase in the risk of breast cancer; increased consumption of sweet products and a decrease in the risk of breast cancer; white meat consumption and decreased risk of breast cancer.

## 9 Acknowledgements

I want to thank my supervisor, Professor Michail Papathomas, for providing me with assistance in formulating my research and methodology for this dissertation.

I want to express my appreciation for my sister, Pragya Bhargava and my aunt Vidhi Bhargava, for spending endless time proofreading and editing with me. I also want to thank my parents for providing me with unconditional support throughout my masters.

Furthermore, I would like to acknowledge all my friends, especially Gigi, Brooke and Marc for always being there when I needed to run through my ideas and helping me with translations necessary for writing this paper.

Finally, I would like to thank the British Council, India, for funding my master's degree.

## 10 References

Bandera, Elisa V., Stephanie H. Fay, Edward Giovannucci, Michael F. Leitzmann, Rachel Marklew, Anne McTiernan, Amy Mullee, et al. 2016. "The use and interpretation of anthropometric measures in cancer epidemiology: A perspective from the world cancer research fund international continuous update project." *International Journal of Cancer* 139 (11): 2391–7. doi:10.1002/ijc.30248.

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. "{mice}: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (3): 1–67. https://www.jstatsoft.org/v45/i03/.

Chawla, Nitesh V., Kevin W. Bowyer, Lawarence O. Hall, and Philip W. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16 (1): 732–35. doi:10.1613/jair.953.

Chen, Chao, Andy Liaw, and Leo Breiman. n.d. "Using Random Forest to Learn Imbalanced Data." *Discovery*, no. 1999: 1–12.

Chu, Susan Y, Nancy E Stroup, Phyllis A Wlngo, and Nancy C Lee. 1990. "CIGARETTE SMOKING AND THE RISK OF BREAST CANCER." *Ameri-*

*can Journal of Epidemiology* 131 (2): 224–53. doi:https://doi.org/10.1093/oxfordjournals.aj

Domingos, Pedro. 1999. “MetaCost: A General Method for Making Classifiers Cost-Sensitive.” In *Proceedings of the Fifth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 155—–164. ACM. doi:http://doi.acm.org/10.1145/312129.312220.

E., Negri, La Vecchia C., Bruzzi P., Dardanoni G., Decarli A., Palli D., Parazzini F., and Del Turco M.R. 1988. “Risk factors for breast cancer: Pooled results from three italian case-control studies.” *American Journal of Epidemiology* 128 (6): 1207–15. http://www.embase.com/search/results?subaction=viewrecord{\&}from=ex

ETHZ. 2012. “Multiple Testing FWER.”

Freudenheim, Jo L, Paola Muti, Saxon Graham, James R Marshall, Rosemary Laughlin, John E Vena, Takuma Nemoto, and Kirsten B Moysich. 1997. “Lactation history and breast cancer risk.” *American Journal of Epidemiology* 146 (11): 932–38. doi:10.1093/oxfordjournals.aje.a009219.

Friedenreich, Christine M., Inger Thune, Louise A. Brinton, and Demetrius Albanes. 1998. “Epidemiologic issues related to the association between physical activity and breast cancer.” *Cancer* 83 (S3): 600–610. doi:10.1002/(sici)1097-0142(19980801)83:3+<600::aid-cncr2>3.3.co;2-0.

Haftenberger, M, AJ Schuit, MJ Tormo, H Boeing, N Wareham, HB Bueno-de-

Mesquita, M Kumle, et al. 2002. "Physical activity of subjects aged 50–64 years involved in the European Prospective Investigation into Cancer and Nutrition (EPIC)." *Public Health Nutrition* 5 (6b): 1163–77. doi:10.1079/phn2002397.

Han, Hedong, Wei Guo, Wentao Shi, Yamei Yu, Yunshuo Zhang, Xiaofei Ye, and Jia He. 2017. "Hypertension and breast cancer risk : a systematic review and meta- analysis." *Nature Publishing Group*. Nature Publishing Group, 1–9. doi:10.1038/srep44877.

Hunter, David J. 2017. "Oral Contraceptives and the Small Increased Risk of Breast Cancer." *New England Journal of Medicine* 377 (23): 2276–7. doi:10.1056/nejme1709636.

Japkowicz, Nathalie, and Shaju Stephan. 2002. "The Class Imbalance Problem - A Systematic Study." *Intelligent Data Analysis*, 429——449.

Kabir, Md Faisal, and Simone Ludwig. 2019. "Classification of Breast Cancer Risk Factors Using Several Resampling Approaches." *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 1243–8. doi:10.1109/ICMLA.2018.00202.

Kotepui, Manas. 2016. "Diet and risk of breast cancer." *Wspolczesna Onkologia* 20 (1): 13–19. doi:10.5114/wo.2014.40560.

L., Chatenoud, La Vecchia C., Franceschi S., Tavani A., Jacobs Jr. D.R., Parpinel M.T., Soler M., and Negri E. 1999. "Refined-cereal intake and risk of selected cancers in Italy." *American Journal of Clinical Nutrition*, 1–4.

Lagiou, Pagona, Jorn Olsen, Dimitrios Trichopoulos, Bernhard Watzl, Carla H. Van Gils, Petra H. M. Peeters, and Elio Riboli. 2005. "Consumption of vegetables and fruits and risk of breast cancer." *Journal of the American Medical Association* 293 (18): 2209–10. doi:10.1001/jama.293.18.2209-a.

Lemnaru, Camelia, and Rodica Potolea. 2012. "Imbalanced classification problems: Systematic study, issues and best practices." *Lecture Notes in Business Information Processing* 102 LNBIP (1): 35–50. doi:10.1007/978-3-642-29958-2_3.

Li, Xiu Juan, Zhao Jun Ren, Jian Wei Qin, Jian Hua Zhao, Jin Hai Tang, Ming Hua Ji, and Jian Zhong Wu. 2013. "Coffee Consumption and Risk of Breast Cancer: An Up-To-Date Meta-Analysis." *PLoS ONE* 8 (1). doi:10.1371/journal.pone.0052681.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://cran.r-project.org/doc/Rnews/.

Lundqvist, Adam, Emelie Andersson, Ida Ahlberg, Mef Nilbert, and Ulf Gerdtham. 2016. "Socioeconomic inequalities in breast cancer incidence and mortality in Europe - A systematic review and meta-analysis." *European*

*Journal of Public Health* 26 (5): 804–13. doi:10.1093/eurpub/ckw070.

Lynch, Shannon M., and Jason H. Moore. 2016. “A call for biological data mining approaches in epidemiology.” *BioData Mining* 9 (1). BioData Mining. doi:10.1186/s13040-015-0079-8.

MacNee, William, Roberto A Rabinovich, and Gourab Choudhury. 2014. “Ageing and the border between health and Disease.” *European Respiratory Journal* 44 (5): 1332–52. doi:10.1183/09031936.00134014.

Mcpherson, K, C M Steel, and J M Dixon. 2000. “Breast cancer — epidemiology , risk factors , and genetics Risk factors for breast cancer.” *British Medical Journal* 321 (September): 624–28.

Newcomb, Polly A, P Longnecker, E Robert, and Meir Stampfer. 1998. “Menstrual Factors in Relation to Breast Cancer” 7 (September): 783–89.

Pacholczak, Renata, Wiesława Klimek-Piotrowska, and Piotr Kuszmiersz. 2016. “Associations of anthropometric measures on breast cancer risk in pre- and postmenopausal women-a case-control study.” *Journal of Physiological Anthropology* 35 (1). doi:10.1186/s40101-016-0090-x.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https:

//www.r-project.org/.

Raileanu, Laura E, and Kilian Stoffel. 2000. "Theoretical Comparison between the Gini Index and Information Gain," no. 2100. http://citeseerx.ist.psu.edu/viewdoc/down

Riboli, E, KJ Hunt, N Slimani, P Ferrari, T Norat, M Fahey, UR Charrondière, et al. 2002. "European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection." *Public Health Nutrition* 5 (6b): 1113–24. doi:10.1079/phn2002394.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: an open-source package for R and S+ to analyze and compare ROC curves." *BMC Bioinformatics* 12: 77.

Sainani, Kristin L. 2009. "The Problem of Multiple Testing." *PM and R* 1 (12): 1098–1103. doi:10.1016/j.pmrj.2009.10.004.

Schuetze, Madlen. 2007. "Alcohol attributable burden of incidence of cancer in eight European countries based on results from prospective cohort study." *British Medical Journal* 342: 1–10. doi:https://doi.org/10.1136/bmj.d1584.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39 (5): 1–13. http://www.jstatsoft.

org/v39/i05/.

Slimani, N, R Kaaks, P Ferrari, C Casagrande, F Clavel-Chapelon, G Lotze, A Kroke, et al. 2002. "European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study: rationale, design and population characteristics." *Public Health Nutrition* 5 (6b): 1125–45. doi:10.1079/phn2002395.

Slimani, Nadia, Ute Ruth Charrondière, Wija Van Staveren, and Elio Riboli. 2000. "Standardization of food composition databases for the European prospective investigation into cancer and nutrition (EPIC): General theoretical concept." *Journal of Food Composition and Analysis* 13 (4): 567–84. doi:10.1006/jfca.2000.0910.

Song, Jae W., and Kevin C. Chung. 2010. "Obersvational Studies: Cohort and Case-Conrtol Studies." *Plastic and Reconstructive Surgery* 126 (6): 2234–42. doi:10.1097/PRS.0b013e3181f44abc.Observational.

Tavani, A, L Giordano, S Gallus, R Talamini, S Franceschi, A Giacosa, M Montella, and C. La Vecchia. 2006. "Consumption of sweet foods and breast cancer risk in Italy." *Annals of Oncology* 17 (2): 341–45. doi:10.1093/annonc/mdj051.

Tehard, B., and F. Clavel-Chapelon. 2006. "Several anthropometric measurements and breast cancer risk: Results of the E3N cohort study." *International*

*Journal of Obesity* 30 (1): 156–63. doi:10.1038/sj.ijo.0803133.

Terry Pd Fau Rohan, Thomas, and T Rohan. 2002. "Cigarette smoking and the risk of breast cancer in women: a review of the." *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 11 (10 Pt 1): 953–71. http://cebp.aacrjournals.org/content/cebp/11/10/953.full.pdf.

Therneau, Terry, and Beth Atkinson. 2018. *rpart: Recursive Partitioning and Regression Trees.* https://cran.r-project.org/package=rpart.

Touvier, Mathilde, Philippine Fassier, Mathilde His, Teresa Norat, Doris S.M. Chan, Jacques Blacher, Serge Hercberg, Pilar Galan, Nathalie Druesne-Pecollo, and Paule Latino-Martel. 2015. "Cholesterol and breast cancer risk: A systematic review and meta-analysis of prospective studies." doi:10.1017/S000711451500183X.

World Cancer Research Fund/American Institute for Cancer Research. 2018. "Diet, nutrition, physical activity and breast cancer." *Continuous Update Project Expert Report 2018*, 50. doi:10.1007/s12082-007-0105-4.

World Health Organization|International Agency for Research on Cancer-Global Cancer Observatory. 2018. "Age standardized (World) incidence rates, breast, all ages.” Vol. 876. http://gco.iarc.fr/today.

Xia, Hui, Shushu Ma, Shaokang Wang, and Guiju Sun. 2015. “Meta-analysis of saturated fatty acid intake and breast cancer risk.” *Medicine (United States)* 94 (52): 1–10. doi:10.1097/MD.0000000000002391.

Zang, Jiajie, Meihua Shen, Sufa Du, Tianwen Chen, and Shurong Zou. 2015. “Breast Cancer The Association between Dairy Intake and Breast Cancer in Western and Asian Populations : A Systematic Review and Meta-Analysis” 18 (4): 313–22.

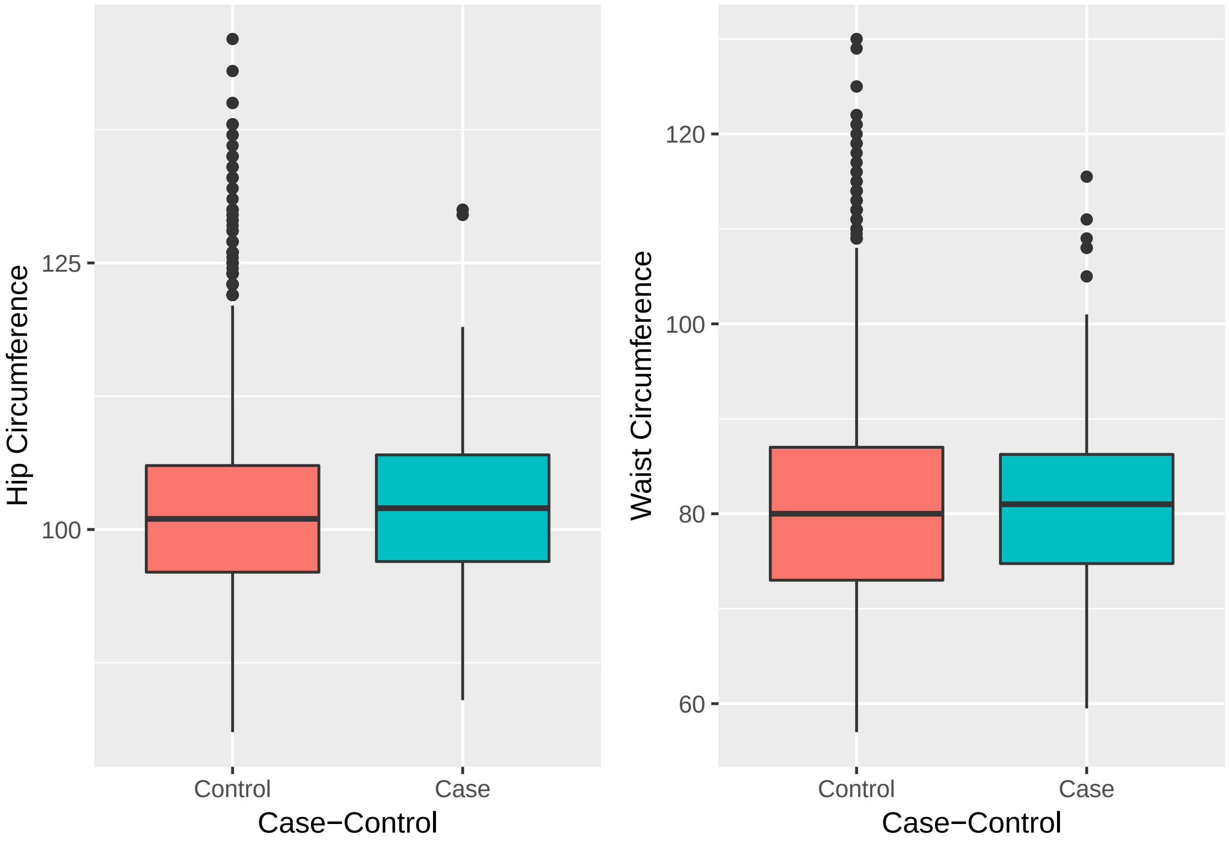
Ziegler, Regina G. 2018. “Anthropometry and Breast Cancer.” *The Journal of Nutrition* 127 (5): 924S–928S. doi:10.1093/jn/127.5.924s.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and variable selection via the elastic net.” *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67 (5): 768. doi:10.1111/j.1467-9868.2005.00527.x.

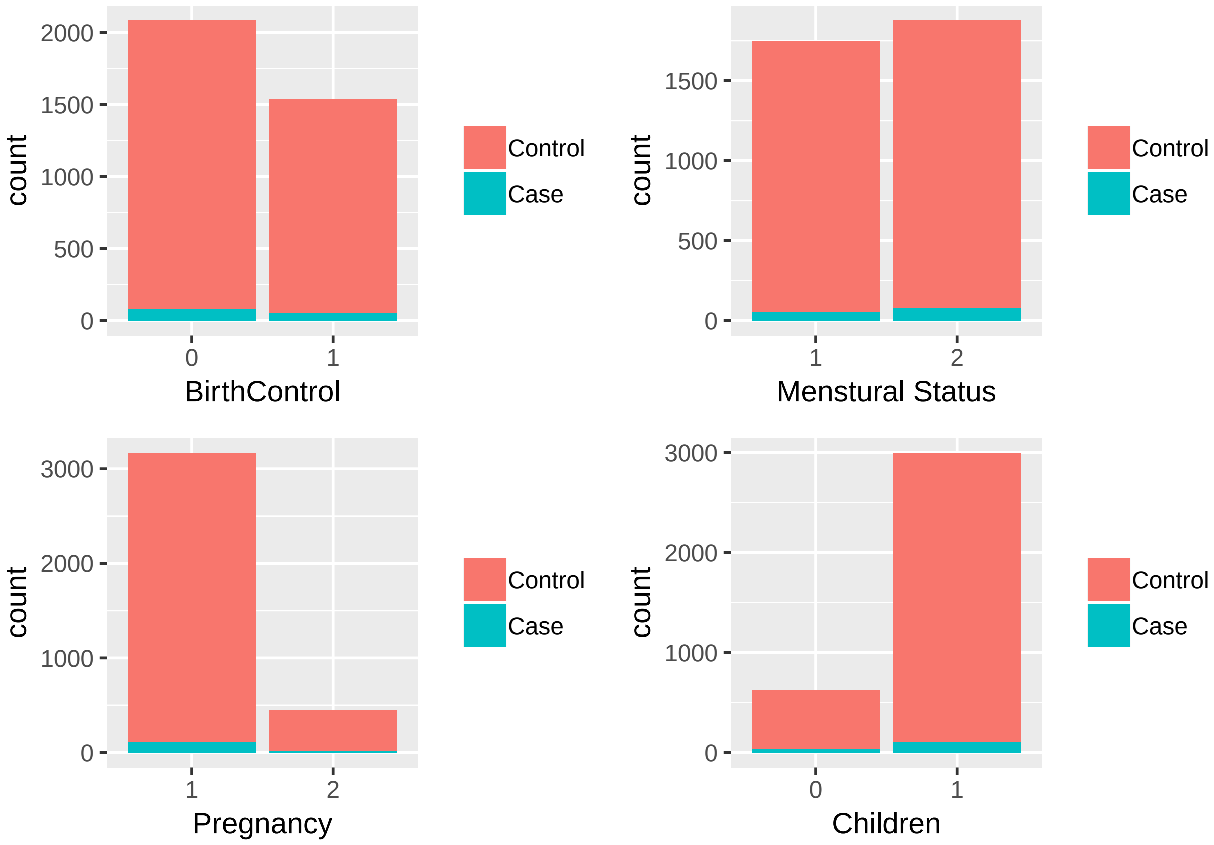
# 11 Appendix

## 11.1 Exploratory

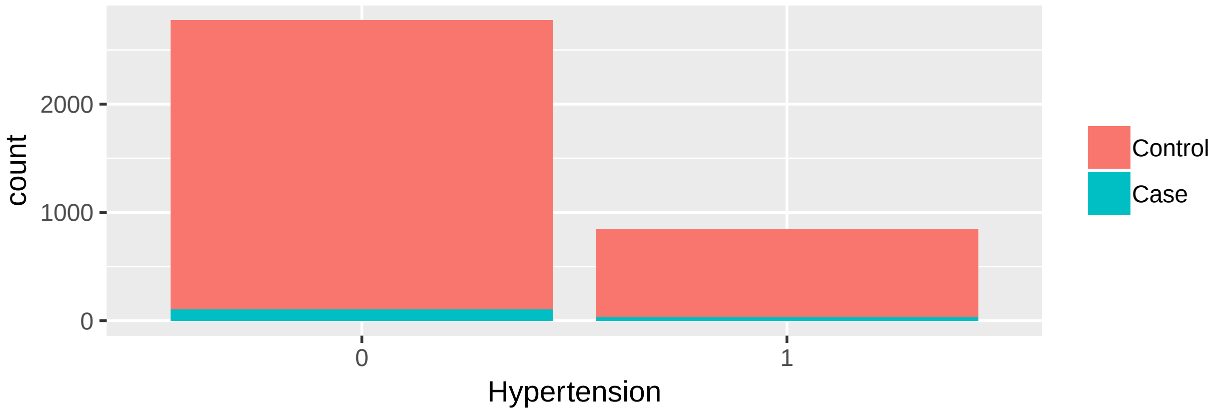
### 11.1.1 Anthropometric measurement plots

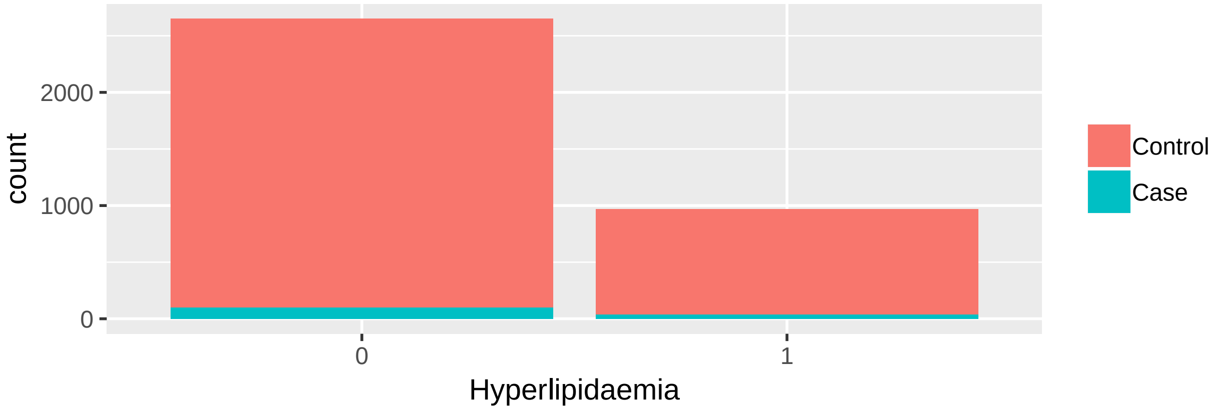


### 11.1.2 Reproductive factor plots

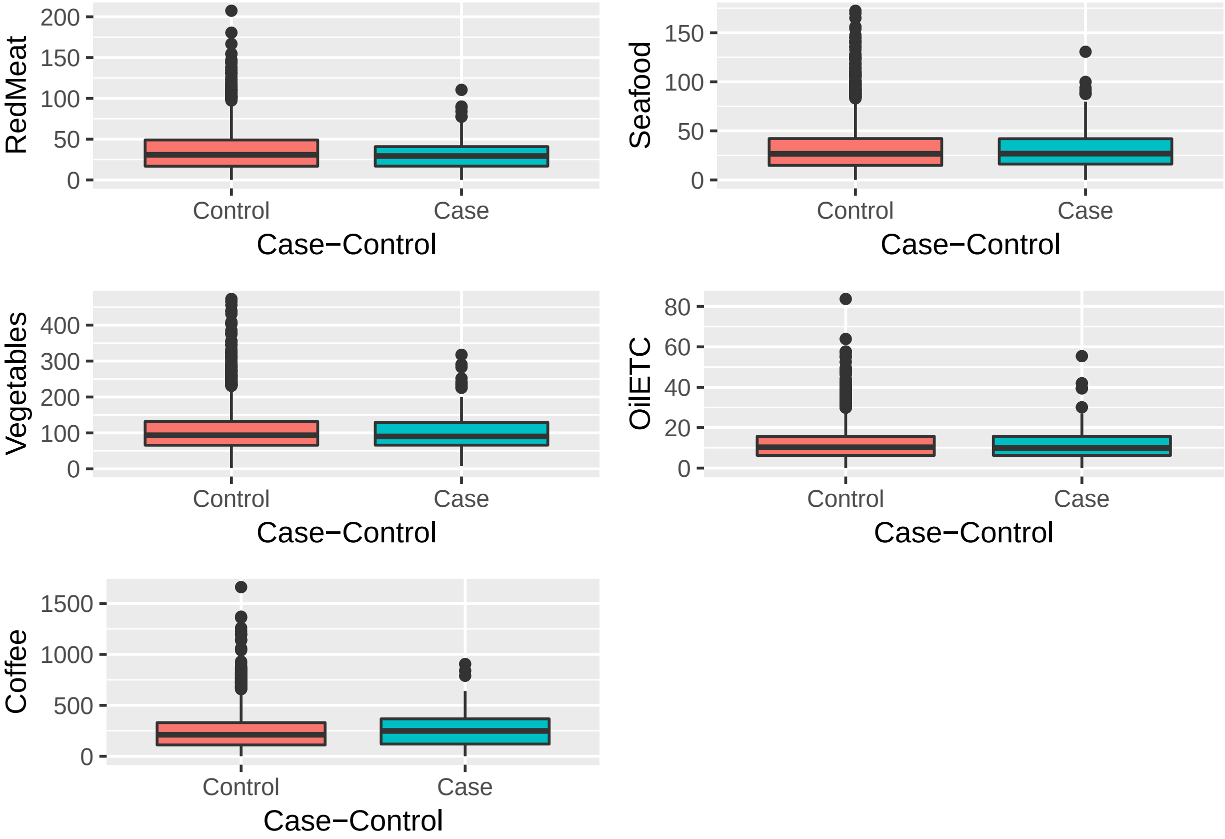


### 11.1.3 Lifestyle choices plots





### 11.1.4 Nutritional and Dietry Information plots



## 11.2 Initial Test Results

```
knitr::kable(Exploratory[-which(Exploratory$p.value <= 0.05),],
             digits = 2)%>%kable_styling(latex_options =
                                           "HOLD_position", font_size = 14)
```

| | test.stat | p.value | p.adjust |
|---|---:|---:|---:|
| Height | 229138.50 | 0.54 | 1 |
| HipCircumference | 206293.50 | 0.15 | 1 |
| WaistCircumference | 204526.00 | 0.11 | 1 |
| WHR | 207035.50 | 0.18 | 1 |
| Alcohol | 257450.00 | 0.09 | 1 |
| RedMeat | 248972.50 | 0.31 | 1 |
| Seafood | 232397.50 | 0.71 | 1 |
| CerealandGrains | 252271.50 | 0.20 | 1 |
| Vegetables | 236294.00 | 0.96 | 1 |
| OilETC | 237469.50 | 0.96 | 1 |
| Dairy | 249038.50 | 0.31 | 1 |
| Fruits | 229881.50 | 0.56 | 1 |
| Coffee | 218769.50 | 0.13 | 1 |
| BirthControl | 0.16 | 0.69 | 1 |
| Menstruating | 3.07 | 0.08 | 1 |
| Pregnant | 0.50 | 0.48 | 1 |
| AgeAtFirstBirth | 6.92 | 0.14 | 1 |
| Children | 2.80 | 0.09 | 1 |
| Complication | 0.84 | 0.66 | 1 |
| Phy | 2.53 | 0.28 | 1 |
| Cigarettes | 2.38 | 0.12 | 1 |
| Hypertension | 0.02 | 0.88 | 1 |
| Hyperlipidaemia | 0.04 | 0.83 | 1 |

### 11.3 Logistic Model Output

```
#Deciding the link function
model1 <- glm(casecontrol~.,data = train,
              family = binomial(link = "logit"))
model2 <- glm(casecontrol~.,data = train,
              family = binomial(link = "probit"))
model3 <- glm(casecontrol~.,data = train,
              family = binomial(link = "cloglog"))
kable(AIC(model1, model2,model3))
```

| | df | AIC |
|---|---|---|
| model1 | 35 | 951.1912 |
| model2 | 35 | 949.6383 |
| model3 | 35 | 951.4067 |

#### 11.3.1 Model Analysis

```
#Assessing the model
GenralizedRSqr <- function(dev, nulldev,n){
  Rsqr <- 1-exp((dev-nulldev)/n)
  return(Rsqr)
}
GenralizedRSqr(model1$deviance,model1$null.deviance,nrow(train))
```

```
[1] 0.01633503
```

```
summary(model1)
```

```

Call:
glm(formula = casecontrol ~ ., family = binomial(link = "logit"), 
    data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.8210  -0.3073  -0.2445  -0.1919   3.0987  

Coefficients: (4 not defined because of singularities)
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     1.862e+00  1.906e+01   0.098   0.9222
Height                          1.807e-02  9.618e-02   0.188   0.8510
Weight                          1.564e-02  1.147e-01   0.136   0.8915
HipCircumference               -9.164e-02  1.244e-01  -0.737   0.4614
```

```
WaistCircumference              7.154e-02  1.517e-01   0.472   0.6372
Age                             1.549e-02  2.285e-02   0.678   0.4978
BirthControl1                   2.501e-01  2.299e-01   1.088   0.2766
Menstruating2                   4.343e-01  3.543e-01   1.226   0.2202
AgeAtMeno less than 45 years   -4.721e-01  3.920e-01  -1.204   0.2284
AgeAtMeno45 to 49 years        -8.243e-01  3.315e-01  -2.487   0.0129 *
AgeAtMenogreater than 49 years         NA         NA      NA       NA
Pregnant2                      -3.680e-01  4.137e-01  -0.889   0.3738
AgeAtFirstBirth1               -1.220e+00  4.947e-01  -2.467   0.0136 *
AgeAtFirstBirth2               -9.650e-01  4.453e-01  -2.167   0.0302 *
AgeAtFirstBirth3               -8.516e-01  4.363e-01  -1.952   0.0510 .
AgeAtFirstBirth4               -1.006e+00  4.385e-01  -2.294   0.0218 *
Children1                              NA         NA      NA       NA
Complication1                   2.067e-01  2.441e-01   0.847   0.3971
Complication2                          NA         NA      NA       NA
BreastFeedNever Breastfed      -3.669e-01  3.046e-01  -1.204   0.2284
BreastFeedBreastfed                    NA         NA      NA       NA
Phy1                           -1.755e-01  2.569e-01  -0.683   0.4945
Phy2                            5.390e-02  2.382e-01   0.226   0.8210
Cigarettes1                    -3.421e-01  2.151e-01  -1.590   0.1117
```

```
Hypertension1                -7.920e-02  2.461e-01  -0.322   0.7476

Hyperlipidaemia1             -2.999e-01  2.431e-01  -1.234   0.2174

Alcohol                      -1.310e-04  7.781e-04  -0.168   0.8663

WhiteMeat                     6.239e-03  3.393e-03   1.839   0.0659 .

RedMeat                      -4.151e-03  4.589e-03  -0.905   0.3657

Seafood                      -2.922e-03  4.268e-03  -0.685   0.4936

CerealandGrains              -1.035e-03  1.196e-03  -0.865   0.3871

Vegetables                    7.755e-04  2.168e-03   0.358   0.7205

OilETC                       -5.792e-03  1.475e-02  -0.393   0.6946

Dairy                        -2.317e-03  1.595e-03  -1.452   0.1465

Fruits                        9.142e-05  5.763e-04   0.159   0.8740

Coffee                        7.990e-04  5.353e-04   1.493   0.1356

Sweets                       -6.875e-03  3.539e-03  -1.942   0.0521 .

WHR                          -8.229e+00  1.540e+01  -0.534   0.5931

BMI                           6.109e-02  2.864e-01   0.213   0.8311

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 928.90  on 2896  degrees of freedom
Residual deviance: 881.19  on 2862  degrees of freedom
AIC: 951.19

Number of Fisher Scoring iterations: 6
```

```
Anova(model1)
```

```
Analysis of Deviance Table (Type II tests)

Response: casecontrol
                   LR Chisq Df Pr(>Chisq)
Height               0.0352  1    0.85113
Weight               0.0186  1    0.89143
HipCircumference     0.5628  1    0.45313
WaistCircumference   0.2277  1    0.63324
Age                  0.4603  1    0.49747
BirthControl         1.1740  1    0.27859
Menstruating                 0
AgeAtMeno            7.1023  2    0.02869 *
Pregnant                     0
```

```
AgeAtFirstBirth      0.9758  3    0.80711
Children                     0
Complication         0.7287  1    0.39329
BreastFeed           1.5507  1    0.21303
Phy                  0.7427  2    0.68982
Cigarettes           2.5927  1    0.10736
Hypertension         0.1044  1    0.74656
Hyperlipidaemia      1.5793  1    0.20886
Alcohol              0.0286  1    0.86565
WhiteMeat            3.1828  1    0.07442 .
RedMeat              0.8445  1    0.35811
Seafood              0.4839  1    0.48667
CerealandGrains      0.7669  1    0.38117
Vegetables           0.1266  1    0.72201
OilETC               0.1569  1    0.69200
Dairy                2.2282  1    0.13551
Fruits               0.0249  1    0.87456
Coffee               2.1004  1    0.14726
Sweets               4.4496  1    0.03491 *
WHR                  0.2949  1    0.58707
```

```
BMI                  0.0454  1    0.83133
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 11.3.2 Variable Selection

## Method 2

```
#Performing Variable Selection
stepmodel1 <- step(model1, direction = "both", trace = 0)
GenralizedRSqr(stepmodel1$deviance,
               stepmodel1$null.deviance,nrow(train))
```

```
[1] 0.01117587
```

```
summary(stepmodel1)
```

```


Call:
glm(formula = casecontrol ~ Height + HipCircumference + AgeAtMeno +
    BreastFeed + Cigarettes + Sweets + BMI, family = binomial(link = "logit"
    data = train)


Deviance Residuals:
    Min       1Q   Median       3Q      Max

-0.6843  -0.3059  -0.2519  -0.2098   2.9584


Coefficients:

                                Estimate Std. Error z value Pr(>|z|)

(Intercept)                    -6.707482   2.758680  -2.431  0.01504 *

Height                          0.032383   0.017873   1.812  0.07000 .

HipCircumference               -0.030420   0.020425  -1.489  0.13640

AgeAtMeno less than 45 years   -0.056085   0.373871  -0.150  0.88076

AgeAtMeno45 to 49 years        -0.319302   0.322478  -0.990  0.32210

AgeAtMenogreater than 49 years  0.475148   0.225368   2.108  0.03500 *

BreastFeedNever Breastfed      -0.952770   0.335219  -2.842  0.00448 **

BreastFeedBreastfed            -0.579914   0.235093  -2.467  0.01363 *

Cigarettes1                    -0.347230   0.208722  -1.664  0.09619 .

Sweets                         -0.007395   0.003444  -2.147  0.03178 *

BMI                             0.088902   0.039388   2.257  0.02400 *

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 928.90  on 2896  degrees of freedom
Residual deviance: 896.35  on 2886  degrees of freedom
AIC: 918.35

Number of Fisher Scoring iterations: 6
```

```
Anova(stepmodel1)
```

```
Analysis of Deviance Table (Type II tests)

Response: casecontrol
                 LR Chisq Df Pr(>Chisq)
Height             3.2532  1    0.07128 .
HipCircumference   2.1505  1    0.14252
AgeAtMeno          7.9689  3    0.04666 *
BreastFeed         9.2324  2    0.00989 **
Cigarettes         2.8423  1    0.09181 .
Sweets             5.5146  1    0.01886 *
BMI                4.6397  1    0.03124 *
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#The dredge function was used for updating the model
updated <- update(stepmodel1,.~.-Height-HipCircumference)
GenralizedRSqr(updated$deviance,updated$null.deviance,nrow(train))
```

```
[1] 0.009872884
```

### Method 2

```
# For variable selection interaction  enet
xmat <- model.matrix(model1)[,-1]
enetstep<- glmnet(xmat, train$casecontrol, family="binomial",
                  alpha=0.5)
set.seed(18008373)
cvenetstep<- cv.glmnet(xmat, train$casecontrol,
                       family="binomial",type ="auc",alpha=0.5,
                       nfolds=5)
nullD <-enetstep$nulldev
pickme <- which(enetstep$lambda == cvenetstep$lambda.min)
coef <- coef(enetstep)[, pickme]
GenralizedRSqr(deviance(enetstep)[pickme],nullD, nrow(train))
```

```
[1] 0.01145797
```

```r
#Fitting it without penalization
model4<- glm(casecontrol~ Weight+AgeAtMeno+AgeAtFirstBirth+
               Children+BirthControl+BreastFeed+Phy+Cigarettes+
               WhiteMeat+Dairy+Coffee+Sweets+RedMeat+
               CerealandGrains +Hyperlipidaemia,data = train,
             family = binomial(link = "logit"))
summary(model4)
```

```
Call:
glm(formula = casecontrol ~ Weight + AgeAtMeno + AgeAtFirstBirth +
    Children + BirthControl + BreastFeed + Phy + Cigarettes +
    WhiteMeat + Dairy + Coffee + Sweets + RedMeat + CerealandGrains +
    Hyperlipidaemia, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6913  -0.3087  -0.2463  -0.1967   3.1756

Coefficients: (2 not defined because of singularities)
```

```
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.3482831  0.6831195  -4.901 9.51e-07 ***
Weight                      0.0157050  0.0086125   1.824   0.0682 .
AgeAtMeno less than 45 years  0.0405282  0.3822710   0.106   0.9156
AgeAtMeno45 to 49 years    -0.2974765  0.3331516  -0.893   0.3719
AgeAtMenogreater than 49 years  0.5840030  0.2458545   2.375   0.0175 *
AgeAtFirstBirth1           -0.8938460  0.3852227  -2.320   0.0203 *
AgeAtFirstBirth2           -0.6224796  0.3087711  -2.016   0.0438 *
AgeAtFirstBirth3           -0.4691397  0.2889478  -1.624   0.1045
AgeAtFirstBirth4           -0.6319459  0.2999614  -2.107   0.0351 *
Children1                          NA         NA      NA       NA
BirthControl1               0.2219294  0.2185821   1.015   0.3100
BreastFeedNever Breastfed  -0.3566525  0.3033021  -1.176   0.2396
BreastFeedBreastfed                NA         NA      NA       NA
Phy1                       -0.1378078  0.2527410  -0.545   0.5856
Phy2                        0.0656306  0.2344864   0.280   0.7796
Cigarettes1                -0.3529640  0.2117789  -1.667   0.0956 .
WhiteMeat                   0.0058276  0.0032979   1.767   0.0772 .
Dairy                      -0.0023239  0.0015605  -1.489   0.1364
Coffee                      0.0008151  0.0005298   1.538   0.1239
```

```
Sweets                         -0.0067458  0.0034869  -1.935   0.0530 .
RedMeat                        -0.0047886  0.0045436  -1.054   0.2919
CerealandGrains                -0.0011193  0.0011456  -0.977   0.3285
Hyperlipidaemia1               -0.2933442  0.2380473  -1.232   0.2178
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 928.90  on 2896  degrees of freedom
Residual deviance: 886.45  on 2876  degrees of freedom
AIC: 928.45

Number of Fisher Scoring iterations: 6
```

```r
GenralizedRSqr(model4$deviance,model4$null.deviance,nrow(train))
```

```
[1] 0.01454697
```

```r
#Stepping
stepmodel4 <- step(model4, direction = "both", trace = 0)
GenralizedRSqr(stepmodel4$deviance,stepmodel4$null.deviance,
               nrow(train))
```

```
[1] 0.01036001
```

```
#All possible subset selection was performed again for comparison
updatedmodel4 <- glm(casecontrol~ Weight+AgeAtMeno+Children+
                       Cigarettes+Sweets,data = train,
                     family = binomial(link = "logit"))
kable(AIC(stepmodel1,updated, updatedmodel4))
```

| | df | AIC |
|---|---|---|
| stepmodel1 | 11 | 918.3458 |
| updated | 9 | 918.1607 |
| updatedmodel4 | 8 | 916.3366 |

### 11.3.3 Model Dignostics

```
#Assessing the fit
probstest <- predict(stepmodel4 , newdata = test, type = "response")


#Predictions - threshold chosen from training in main report
predstest <- as.factor(ifelse(probstest>0.038,"X1","X0"))
levels(predstest) <- levels(y)
```

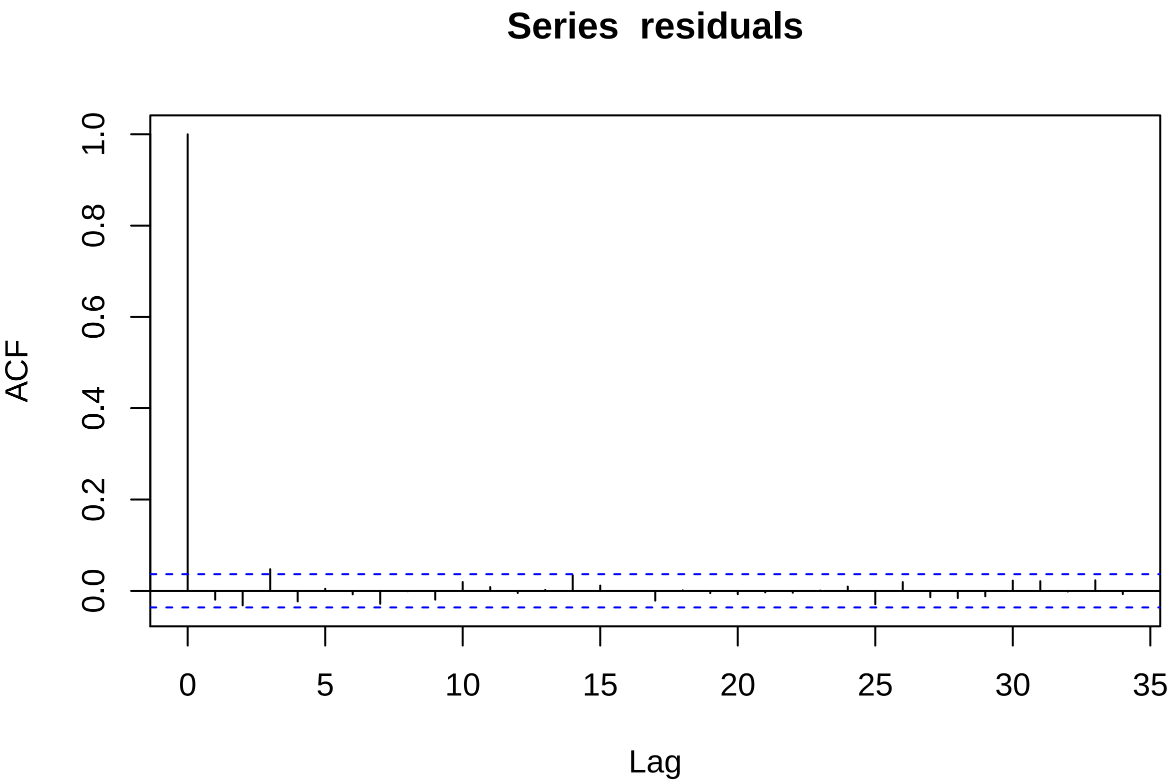
```r
kable(modelmeasures(predstest,test$casecontrol))
```

| Percision | Recall | F_Measure | Sensitivity | Specificity | Balanced_Accuracy |
|---:|---:|---:|---:|---:|---:|
| 0.0397112 | 0.4074074 | 0.0723684 | 0.4074074 | 0.6178161 | 0.5126117 |

```r
#Check for correlation - none

residuals <- residuals(stepmodel4 , type = "pearson")

runs.test(residuals)
```

```

    Runs Test - Two sided


data:  residuals

Standardized Runs Statistic = -0.018576, p-value = 0.9852
```

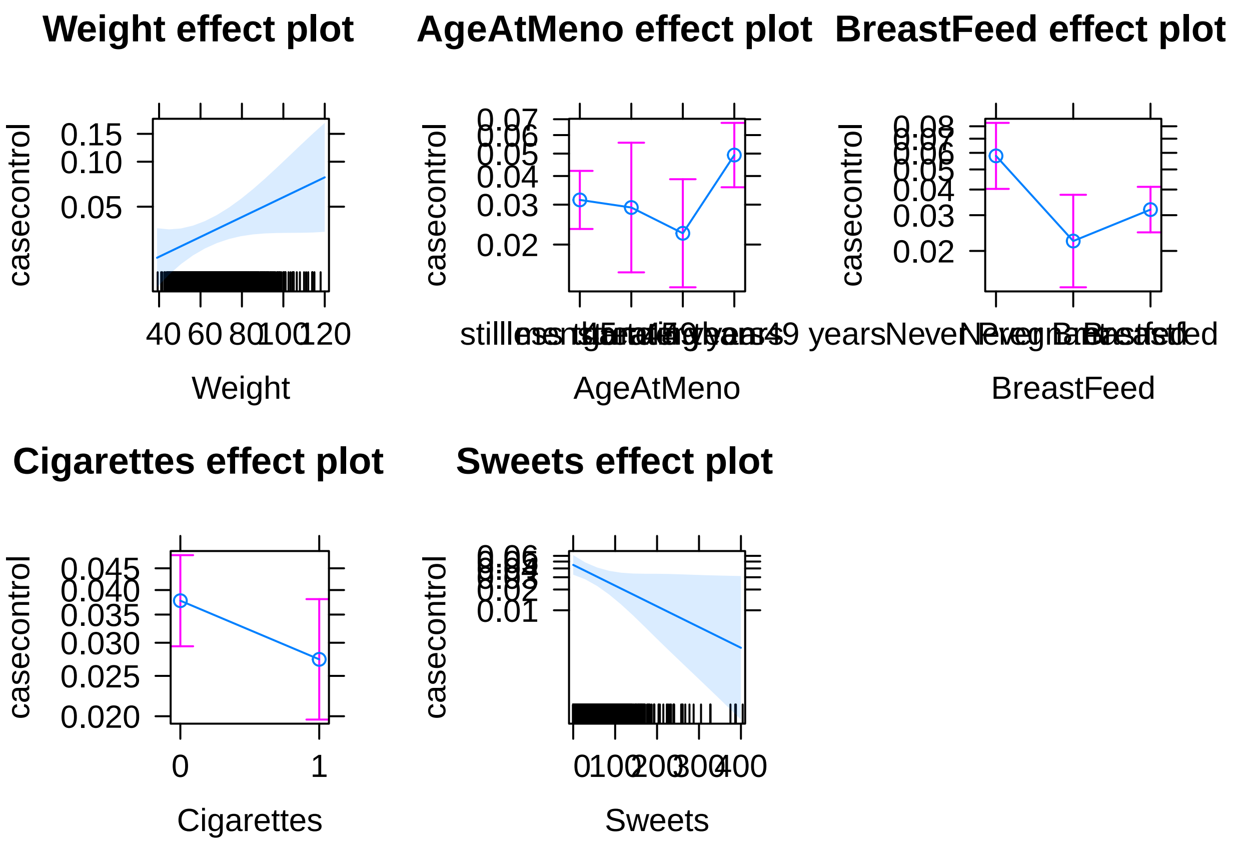
```r
acf(residuals)
```



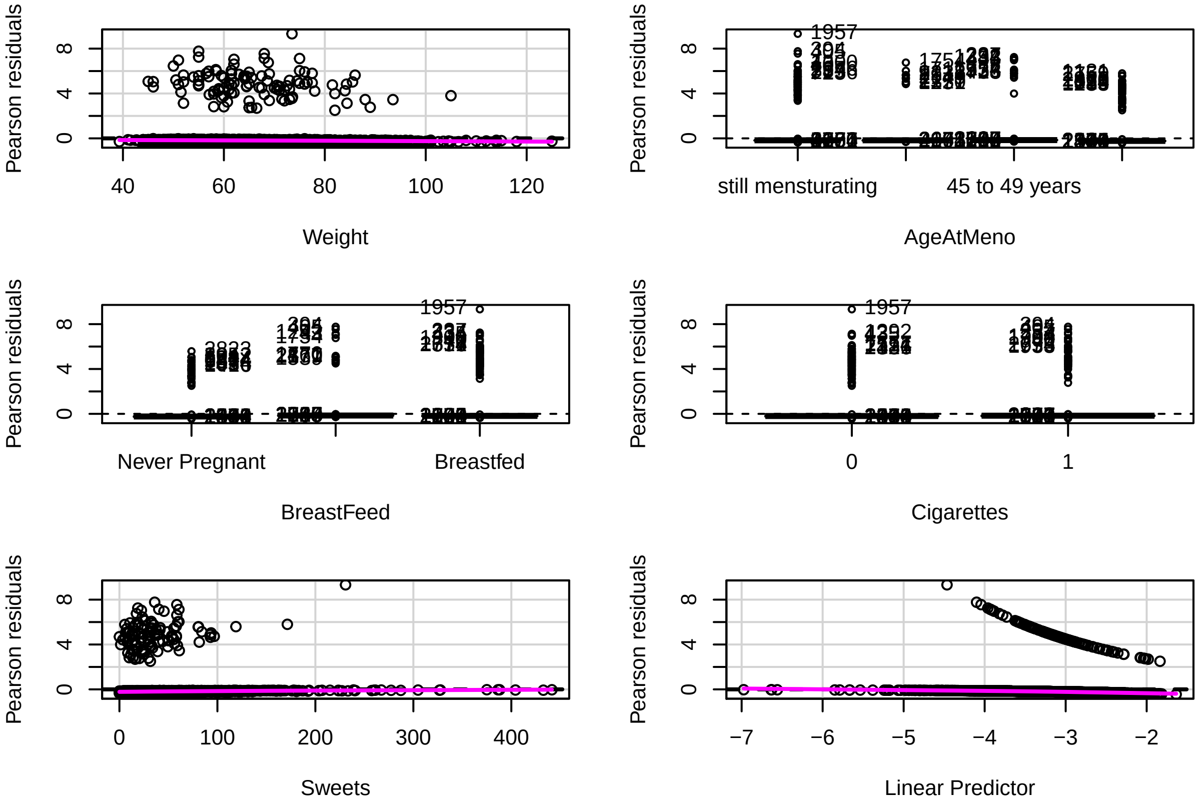
```
#Checking for collinearity

kable(vif(stepmodel4))
```

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---:|---:|---:|
| Weight | 1.017240 | 1 | 1.008583 |
| AgeAtMeno | 1.033382 | 3 | 1.005488 |
| BreastFeed | 1.027060 | 2 | 1.006697 |
| Cigarettes | 1.024659 | 1 | 1.012255 |
| Sweets | 1.014927 | 1 | 1.007436 |

```
#Checking Linearity

plot(allEffects(stepmodel4))
```



```
#Mean-Variance

residualPlots(stepmodel4)
```



```
            Test stat Pr(>|Test stat|)
Weight         2.0448           0.1527
AgeAtMeno
BreastFeed
Cigarettes
Sweets         0.7299           0.3929
```

### 11.4 Model without imputation

```r
stepmodelraw <- step(glm(casecontrol~.,data = na.omit(dat2),
                         family = binomial(link = "logit")),
                     trace = 0)
summary(stepmodelraw)
```

```

Call:
glm(formula = casecontrol ~ AgeAtMeno + BreastFeed + Cigarettes +
    WhiteMeat + CerealandGrains + Sweets, family = binomial(link = "logit")
    data = na.omit(dat2))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5583  -0.3035  -0.2606  -0.2204   2.9431

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.688652   0.334005  -8.050  8.3e-16 ***
AgeAtMeno1       0.018867   0.338692   0.056  0.95558
AgeAtMeno2      -0.141966   0.278980  -0.509  0.61084
```

```
AgeAtMeno3       0.444986   0.211819   2.101  0.03566 *
BreastFeed1     -0.932412   0.324721  -2.871  0.00409 **
BreastFeed2     -0.337400   0.223715  -1.508  0.13151
Cigarettes1     -0.282112   0.190349  -1.482  0.13832
WhiteMeat        0.005638   0.002912   1.936  0.05287 .
CerealandGrains -0.001739   0.001060  -1.641  0.10073
Sweets          -0.003701   0.002759  -1.341  0.17982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1084.6  on 3404  degrees of freedom
Residual deviance: 1059.2  on 3395  degrees of freedom
AIC: 1079.2

Number of Fisher Scoring iterations: 6
```

```
stepraw2 <- step(glm(casecontrol~ Weight+AgeAtMeno+AgeAtFirstBirth+
                       Children+BirthControl+BreastFeed+Phy+
                       Cigarettes+WhiteMeat+Dairy+Coffee+Sweets+
```

```r
                       RedMeat+ CerealandGrains +Hyperlipidaemia,
                     data = na.omit(dat2),
                     family = binomial(link = "logit")),trace = 0)
summary(stepraw2)
```

```
Call:
glm(formula = casecontrol ~ AgeAtMeno + BreastFeed + Cigarettes +
    WhiteMeat + Sweets + CerealandGrains, family = binomial(link = "logit")
    data = na.omit(dat2))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5583  -0.3035  -0.2606  -0.2204   2.9431

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.688652   0.334005  -8.050  8.3e-16 ***
AgeAtMeno1       0.018867   0.338692   0.056  0.95558
AgeAtMeno2      -0.141966   0.278980  -0.509  0.61084
```

```
AgeAtMeno3       0.444986   0.211819   2.101  0.03566 *
BreastFeed1     -0.932412   0.324721  -2.871  0.00409 **
BreastFeed2     -0.337400   0.223715  -1.508  0.13151
Cigarettes1     -0.282112   0.190349  -1.482  0.13832
WhiteMeat        0.005638   0.002912   1.936  0.05287 .
Sweets          -0.003701   0.002759  -1.341  0.17982
CerealandGrains -0.001739   0.001060  -1.641  0.10073
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1084.6  on 3404  degrees of freedom
Residual deviance: 1059.2  on 3395  degrees of freedom
AIC: 1079.2

Number of Fisher Scoring iterations: 6
```

### 11.5 Important interactions

#### 11.5.1 Menopausal interactions

modelmeno <- glm(casecontrol~(Height+Weight+HipCircumference+ WaistCircumference+ BMI+Age+BirthControl+ AgeAtMeno+Pregnant+Children + AgeAtFirstBirth+Complication+BreastFeed+ Cigarettes+ Hypertension +Hyperlipidaemia+Phy+ Alcohol+WhiteMeat+RedMeat+Seafood+Vegetables+ OilETC+CerealandGrains+ Dairy+Fruits+Coffee+ Sweets)*Menstruating,data = train, family = binomial(link = “logit”))

steppedmodelmeno <-step(modelmeno, direction = “both”,trace = 0)

GenralizedRSqr(deviance(steppedmodelmeno),nullD, nrow(train))

**Penalised**

xmatrixmeno<- model.matrix(steppedmodelmeno)[,-1]

enetmeno<- glmnet(xmatrixmeno, train$casecontrol, family=“binomial”,alpha=0.5)

set.seed(18008373)

cvenetmeno<- cv.glmnet(xmatrixmeno, train$casecontrol, family=“binomial”, type =“auc”, alpha=0.5, nfolds=3)

pickmemeno <- which(enetmeno*lambda* == *cvenetmeno*lambda.min)

coefmeno <- coef(enetmeno)[, pickmemeno]

|  | x |
|---|---|
| (Intercept) | 0.0463173 |
| Height | 0.9907619 |
| Age | 1.0474001 |
| AgeAtMeno1 | 0.5049985 |
| AgeAtMeno2 | 0.3895854 |
| AgeAtMeno3 | 0.9389478 |
| BreastFeed1 | 0.1243222 |
| BreastFeed2 | 0.4086184 |
| Cigarettes1 | 0.9804555 |
| WhiteMeat | 1.0055985 |
| Dairy | 0.9978516 |
| Coffee | 1.0008283 |
| Sweets | 0.9926276 |
| Menstruating2 | 0.0352370 |
| Height:Menstruating2 | 1.0431755 |
| Age:Menstruating2 | 0.9385740 |
| BreastFeed1:Menstruating2 | 5.3204401 |
| BreastFeed2:Menstruating2 | 1.5287924 |
| Cigarettes1:Menstruating2 | 0.5488463 |

GenralizedRSqr(deviance(enetmeno)[pickmemeno],nullD, nrow(train))

#### 11.5.2 Pregnant interactions

modelpreg <-glm(casecontrol~(Height+Weight+HipCircumference+ WaistCircumference+ BMI+Age+BirthControl+ AgeAtMeno+Children + Menstruating+ AgeAtFirstBirth+Complication+BreastFeed+ Cigarettes+ Hypertension +Hyperlipidaemia+Phy+ Alcohol+WhiteMeat+RedMeat+Seafood+Vegetables+ OilETC+CerealandGrains+ Dairy+Fruits+Coffee+ Sweets)*Pregnant,data =

train, family = binomial(link = “logit”))

steppedmodelpreg <- step(modelpreg,direction = “both”,trace = 0)

**Penalised**

xmatrixpreg<- model.matrix(steppedmodelpreg)[,-1]

enetpreg<- glmnet(xmatrixpreg, train$casecontrol, family=“binomial”,alpha=0.5)

set.seed(18008373)

cvenetpreg<- cv.glmnet(xmatrixpreg, train$casecontrol, family=“binomial”, type =“auc”, alpha=0.5, nfolds=3)

pickmepreg <- which(enetpreg*lambda* == *cvenetpreg*lambda.min)

coefpreg <- coef(enetpreg)[, pickmepreg]

| | x |
|---|---|
| (Intercept) | 0.2929296 |
| Weight | 1.0339234 |
| HipCircumference | 0.9711652 |
| AgeAtMeno1 | 0.9853866 |
| AgeAtMeno2 | 0.7447192 |
| AgeAtMeno3 | 1.7298098 |
| BreastFeed1 | 0.2886969 |
| BreastFeed2 | 0.4283689 |
| Cigarettes1 | 0.6971042 |
| Hypertension1 | 0.7004623 |
| Hyperlipidaemia1 | 0.8808113 |
| WhiteMeat | 1.0073212 |
| Seafood | 0.9897901 |
| Sweets | 0.9926294 |
| Pregnant2 | 0.4767641 |
| Hypertension1:Pregnant2 | 5.5125254 |
| Hyperlipidaemia1:Pregnant2 | 0.2909256 |
| WhiteMeat:Pregnant2 | 0.9875607 |
| Seafood:Pregnant2 | 1.0213758 |

```
GenralizedRSqr(deviance(enetpreg)[pickmepreg],nullD, nrow(train))
```

### 11.5.3 Age interactions

```
modelage <- glm(casecontrol~(Height+Weight+HipCircumference+ WaistCircumference+ BMI+Children+BirthControl+ AgeAtMeno+Pregnant + Menstruating+ AgeAtFirstBirth+Complication+BreastFeed+ Cigarettes+ Hypertension +Hyperlipidaemia+Phy+ Alcohol+WhiteMeat+RedMeat+Seafood+Vegetables+ OilETC+CerealandGrains+ Dairy+Fruits+Coffee+ Sweets)*Age,data = train,
family = binomial(link = “logit”))

steppedmodelage <- step(modelage, direction = “both”,trace = 0)

GenralizedRSqr(deviance(steppedmodelage),nullD, nrow(train))
```

**Penalised**

```
xmatrixage<- model.matrix(steppedmodelage)[,-1]

enetage<- glmnet(xmatrixage, train$casecontrol, family=“binomial”,alpha=0.5)

set.seed(18008373)

cvenetage<- cv.glmnet(xmatrixage, train$casecontrol, family=“binomial”, type =“auc”, alpha=0.5, nfolds=3)

pickmeage <- which(enetagelambda == cvenetagelambda.min)

coefage <- coef(enetage)[, pickmeage]
```

|  | x |
|---|---:|
| (Intercept) | 2.481691e+14 |
| Height | 7.999965e-01 |
| AgeAtMeno1 | 8.768746e-01 |
| AgeAtMeno2 | 6.606924e-01 |
| AgeAtMeno3 | 1.508002e+00 |
| BreastFeed1 | 3.608539e-01 |
| BreastFeed2 | 5.423958e-01 |
| Cigarettes1 | 7.236508e-01 |
| Alcohol | 1.009659e+00 |
| WhiteMeat | 1.005362e+00 |
| Vegetables | 1.033019e+00 |
| OilETC | 8.329087e-01 |
| CerealandGrains | 9.839379e-01 |
| Sweets | 9.932298e-01 |
| Age | 4.704350e-01 |
| Height:Age | 1.004763e+00 |
| Alcohol:Age | 9.998002e-01 |
| Vegetables:Age | 9.993447e-01 |
| OilETC:Age | 1.003547e+00 |
| CerealandGrains:Age | 1.000299e+00 |

```
GenralizedRSqr(deviance(enetage)[pickmeage],nullD, nrow(train))
```

### 11.5.4 Age at first birth interactions

```
modelageat1 <- glm(casecontrol~(Height+Weight+HipCircumference+ WaistCircumference+ BMI+Children+BirthControl+ AgeAtMeno+Pregnant + Menstruating+ Age+Complication+BreastFeed+ Cigarettes+ Hypertension +Hyperlipidaemia+Phy+ Alcohol+WhiteMeat+RedMeat+Seafood+Vegetables+
```

OilETC+CerealandGrains+ Dairy+Fruits+Coffee+ Sweets)*AgeAtFirstBirth,data = train, family = binomial(link = “logit”))

steppedmodelageat1 <- step(modelageat1, direction = “both”, trace = 0)

GenralizedRSqr(deviance(steppedmodelageat1),nullD, nrow(train))

**Penalised**

xmatrixageat1<- model.matrix(steppedmodelageat1)[,-1]

enetageat1<- glmnet(xmatrixageat1, train$casecontrol, family=“binomial”, alpha=0.5)

set.seed(18008373)

cvenetageat1<- cv.glmnet(xmatrixageat1, train$casecontrol, family=“binomial”, type =“auc”, alpha=0.5, nfolds=3)

pickmeageat1 <- which(enetageat1$lambda == cvenetageat$1lambda.min)

coefageat1 <- coef(enetageat1)[, pickmeageat1]

|  | x |
|---|---:|
| (Intercept) | 0.1055161 |
| Weight | 1.0388225 |
| HipCircumference | 0.9678309 |
| AgeAtMeno1 | 0.0767038 |
| AgeAtMeno2 | 0.4350359 |
| AgeAtMeno3 | 1.5271547 |
| BreastFeed1 | 0.7816471 |
| BreastFeed2 | 1.2023899 |
| Cigarettes1 | 0.6887196 |
| Hypertension1 | 2.8651163 |
| Hyperlipidaemia1 | 0.5477813 |
| Phy1 | 2.6634608 |
| Phy2 | 2.8240330 |
| Coffee | 1.0008050 |
| Sweets | 0.9917694 |
| AgeAtFirstBirth1 | 0.5975028 |
| AgeAtFirstBirth2 | 0.8856993 |
| AgeAtFirstBirth3 | 1.1581561 |
| AgeAtFirstBirth4 | 1.0456099 |
| AgeAtMeno1:AgeAtFirstBirth1 | 113.4837756 |
| AgeAtMeno2:AgeAtFirstBirth1 | 2.2757381 |
| AgeAtMeno3:AgeAtFirstBirth1 | 0.9356771 |
| AgeAtMeno1:AgeAtFirstBirth2 | 0.4077901 |
| AgeAtMeno2:AgeAtFirstBirth2 | 1.0000000 |
| AgeAtMeno3:AgeAtFirstBirth2 | 2.6133730 |
| AgeAtMeno1:AgeAtFirstBirth3 | 13.9858001 |
| AgeAtMeno2:AgeAtFirstBirth3 | 2.0515890 |
| AgeAtMeno3:AgeAtFirstBirth3 | 0.8258435 |
| AgeAtMeno1:AgeAtFirstBirth4 | 16.7325399 |
| AgeAtMeno2:AgeAtFirstBirth4 | 2.9085848 |
| AgeAtMeno3:AgeAtFirstBirth4 | 1.0000000 |
| Hypertension1:AgeAtFirstBirth1 | 0.2316755 |
| Hypertension1:AgeAtFirstBirth2 | 0.1940246 |
| Hypertension1:AgeAtFirstBirth3 | 0.1383044 |
| Hypertension1:AgeAtFirstBirth4 | 0.3670123 |
| Hyperlipidaemia1:AgeAtFirstBirth1 | 0.0338467 |
| Hyperlipidaemia1:AgeAtFirstBirth2 | 0.6272337 |
| Hyperlipidaemia1:AgeAtFirstBirth3 | 3.5712387 |

```
GenralizedRSqr(deviance(enetageat1)[pickmeageat1],nullD, nrow(train))
```

#### 11.5.5 Penalised all interactions

### Elastic Net

```
xmatrix<- model.matrix(modelinterc)[,-1]

enet<- glmnet(xmatrix, train$casecontrol, family=“binomial”,alpha=0.5)

set.seed(18008373)

cvenet<- cv.glmnet(xmatrix, train$casecontrol, family=“binomial”,type =“auc”,
alpha=0.5, nfolds=3)

pickme <- which(enetlambda == cvenetlambda.min)

coef <- coef(enet)[, pickme]
```

|  | x |
|---|---|
| (Intercept) | 0.0384555 |
| Height | 1.0000000 |
| Weight | 1.0000000 |
| HipCircumference | 1.0000000 |
| WaistCircumference | 1.0000000 |
| BMI | 1.0000000 |
| Age | 1.0000000 |
| BirthControl1 | 1.0000000 |
| Menstruating2 | 1.0000000 |
| Pregnant2 | 1.0000000 |
| Children1 | 0.9496763 |
| AgeAtMeno1 | 1.0000000 |
| AgeAtMeno2 | 1.0000000 |
| AgeAtMeno3 | 1.0000000 |
| AgeAtFirstBirth1 | 1.0000000 |
| AgeAtFirstBirth2 | 1.0000000 |
| AgeAtFirstBirth3 | 1.0000000 |
| AgeAtFirstBirth4 | 1.0000000 |
| Complication1 | 1.0000000 |
| Complication2 | 1.0000000 |
| BreastFeed1 | 1.0000000 |

GenralizedRSqr(deviance(enet)[pickme],nullD, nrow(train))

**Glm with the combinations from the elastic net**

model5 <- glm(casecontrol~ Age*Weight+ Pregnant*AgeAtMeno+ Pregnant*Seafood+ AgeAtMeno*AgeAtFirstBirth + AgeAtMeno*Complication+ AgeAtMeno*Coffee+ AgeAtFirstBirth*Phy+ BreastFeed*Sweets+ CerealandGrains*Dairy+ Children*Seafood+ BirthControl*AgeAtMeno+ AgeAtMeno*WhiteMeat+ AgeAtFirstBirth*Hyperlipidaemia+ Complication*Dairy+ Cigarettes*Sweets,data = train, family = binomial(link = “logit”))

GenralizedRSqr(model4*deviance, model*5null.deviance,nrow(train))

stepmodel5 <- step(model4, direction = “both”, trace = 0)

GenralizedRSqr(stepmodel5*deviance, stepmodel*5null.deviance,nrow(train))

### 11.5.6 Final model

```
#Penalising
xmat2 <- model.matrix(stepmodel4)[,-1]
enet2<- glmnet(xmat2, train$casecontrol, family="binomial",alpha=0.5)
set.seed(18008373)
cvenet2<- cv.glmnet(xmat2, train$casecontrol, family="binomial",
                    type ="auc",alpha=0.5, nfolds=3)
```

```r
pickme2 <- which(enet2$lambda == cvenet2$lambda.min)

GenralizedRSqr(deviance(enet2)[pickme2],nullD, nrow(train))
```

```
[1] 0.01033503
```

```r
coef2 <- exp(coef(enet2)[, pickme2])

coefenet<- data.frame(Odds = coef2)

kable(coefenet)
```

| | Odds |
|---|---:|
| (Intercept) | 0.0319397 |
| Weight | 1.0159264 |
| AgeAtMeno less than 45 years | 0.9510995 |
| AgeAtMeno45 to 49 years | 0.7266894 |
| AgeAtMenogreater than 49 years | 1.5920459 |
| BreastFeedNever Breastfed | 0.3933234 |
| BreastFeedBreastfed | 0.5566876 |
| Cigarettes1 | 0.7314924 |
| Sweets | 0.9933692 |

```r
GenralizedRSqr(deviance(enet2)[pickme2],nullD, nrow(train))
```

```
[1] 0.01033503
```

## 11.6 Data Mining methods

*The output is not shown here but the entire code has been given*

### 11.6.1 Tree

```
fitControl <- trainControl(method = “repeatedcv”, number = 5, repeats = 5, classProbs = TRUE, summaryFunction = twoClassSummary)

set.seed(123)

ctree <- train(x = x,y = y, method = “rpart”, trControl = fitControl, parms = list(split = “gini”), metric = “ROC”,tuneLength = 10)
```

#### 11.6.1.1 Weighted tree with class imbalance

```
loss <- matrix(c(0,100,5,0), ncol=2)

set.seed(12345)

ctreeloss <- train(x,y, method = “rpart”, trControl = fitControl, parms = list(split = “gini”, loss = loss), tuneLength = 10, metric = “ROC”)
```

#### 11.6.1.2 Tree adjusting probilites to maximise ROC

```
set.seed(123456)
```

```
ctreeprob <- train(x,y, method = “rpart”, parms = list(split = “gini”, prior = c(.04,.96)), metric = “ROC”, tuneLength = 10, trControl = fitControl)

ctreeprob1 <- train(x1,y1, method = “rpart”, parms = list(split = “gini”, prior = c(.04,.96)), metric = “ROC”, tuneLength = 10, trControl = fitControl)
```

#### 11.6.1.3 Downsampled

```
set.seed(54321)

downsample <- downSample(train,train$casecontrol)

downsample <- downsample%>%select(-Class)

set.seed(123456)

ctreedown <- train(x= downsample[,c(1:28,30,31)],y= downsample$casecontrol, method = “rpart”, trControl = fitControl, parms = list(split = “gini”), tuneLength = 10, metric = “ROC”)
```

#### 11.6.1.4 Upsampled

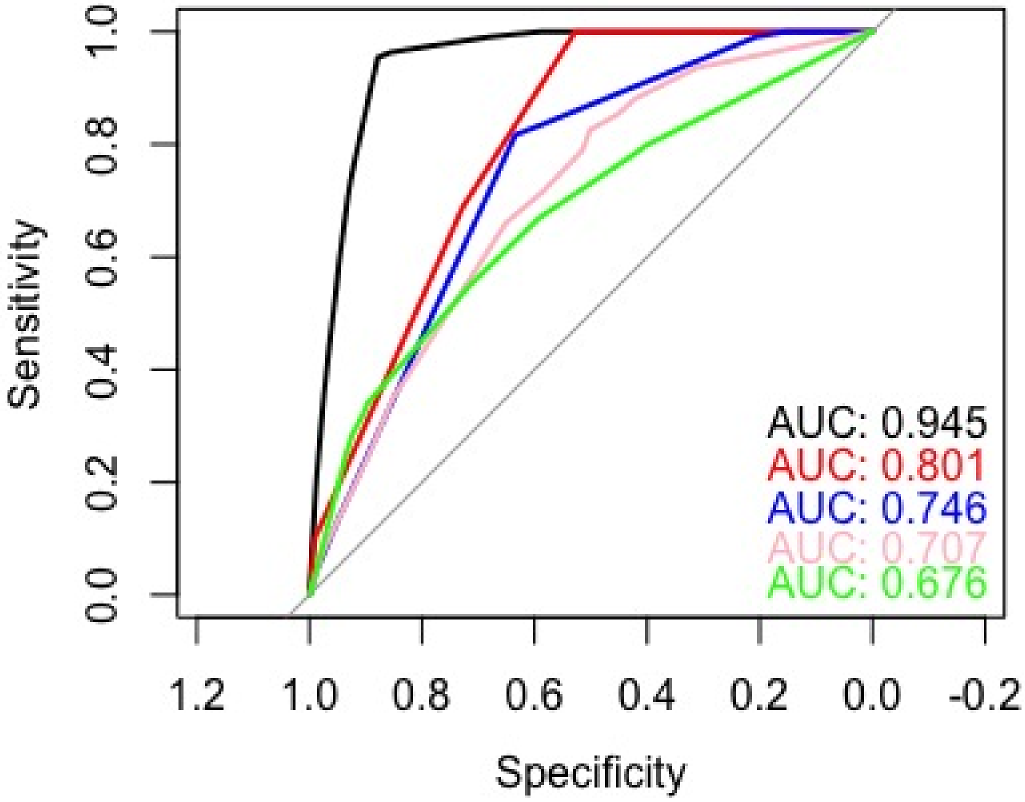
```
set.seed(54321)

upsample <- upSample(train,train$casecontrol)

upsample <- upsample%>%select(-Class)

set.seed(123456)
```

```
ctreeup <- train(x= upsample[,c(1:28,30,31)],y= upsample$casecontrol,
method = “rpart”, trControl = fitControl, parms = list(split = “gini”),
tuneLength = 10, metric = “ROC”)
```

#### 11.6.1.5 SMOTE

```
library(DMwR)

set.seed(54321)

smotesample <- SMOTE(casecontrol˜., data = train)

set.seed(123456)

ctreesmote <- train(x= smotesample[,c(1:28,30,31)],y= smotesample$casecontrol,
method = “rpart”, trControl = fitControl, parms = list(split = “gini”),metric
= “ROC”,tuneLength = 10)
```

### Comparison



#### 11.6.2 Random Forests

##### 11.6.2.1 Balanced Random forest

min <- min(table(train$casecontrol))

*500 trees*

set.seed(18008373)

rforestBal <- train(x,y, method = "rf", metric = "ROC", trControl = fitControl,tuneGrid=tunegrid, importance = TRUE, strata = train$casecontrol, *1000 trees*

set.seed(18008373)

rforestBal2 <- train(x,y, method = “rf”, metric = “ROC”, trControl = fitControl,tuneGrid=tunegrid, importance = TRUE, strata = train$casecontrol, sampsize = c(min,min), ntree = 1000)

#### 11.6.2.2 SMOTE

*500 trees*

set.seed(18008373)

rforestSMOTE<- train(x= smotesample[,c(1:28,30,31)],y= smotesample$casecontrol, method = “rf”,metric = “ROC”, trControl = fitControl, tuneGrid=tunegrid)

*1000 trees*

set.seed(18008373)

rforestSMOTE2<- train(x= smotesample[,c(1:28,30,31)],y= smotesample$casecontrol, method = “rf”,metric = “ROC”, trControl = fitControl, tuneGrid=tunegrid, ntree = 1000)