Male Fertility Analysis: Identifying Significant Variables and Predicting Fertility Outcomes

Abstract
Male infertility is a common disease that accounts for 40-50% of total infertility cases. This study aims to find significant predictors and the best model for predicting male fertility. UCI Machine Learning collected semen samples and participant information relating to 9 socio-demographic, environmental, lifestyle, and health predictors from 100 volunteers. We compared regression, Bayesian, and 4 tree-based methods (CART, bagging, random forest, and boosting) and found the boosting model to have the best predictive power. Our findings agreed with existing literature that (older) age, exercise, and (the presence of) medical conditions/injuries are significant predictors of male infertility.

1. Background & Significance

Infertility is a disease that many silently struggle with. As conversation around the disease increases and its stigma decreases, we wish to better understand the causes behind this disease in the hopes that it may be cured. This analysis hence seeks to identify important predictors of infertility so greater research can be done on these factors, and, if such factors are related to lifestyle, perhaps changes can be implemented in one's life for better fertility outcomes. Male infertility is of particular importance because it is responsible for 40-50% of total infertility cases [4]. The literature indicates that greater age is associated with greater medical issues, greater fragmentation in DNA in sperm, and a reduction in sperm quality and sexual function– all of which are associated with poor fertiltiy outcomes [3]. It has also been observed that sedentary men have greater infertility outcomes in addition to men who frequently drink alcohol and smoke cigarettes [1][2]. Given this introduction, we were interested in whether our results would agree with the established literature and/or discover new trends, which led us to three research questions:

1) *Which factors contribute best to predicting male fertility?*
2) *Are there differences in male fertility based on socio-demographic, environmental, health, and lifestyle factors?*
3) *Which of our models best predicts male fertility?*

2. Methodology

2a. Data Description

The dataset was collected from UCI Machine Learning in 2013. Semen samples and information relating to 9 socio-demographic, environmental, lifestyle, and health predictors were collected from 100 volunteers.

2b. Variables

The 9 predictors were Season in which the analysis was performed, Age, Childish diseases (yes/no), Accident or serious trauma (yes/no), Surgical intervention (yes/no), High fevers in the last year, Frequency of alcohol consumption, Smoking habit, and Number of hours spent sitting per day. The response was Diagnosis (Altered/Normal) and a dummy variable was created for the response as (1/0), which the boosting model required due to its Bernoulli distribution (as the response is binary).

2c. Analytic Methods

In checking assumptions before modeling, we note the predictor is categorical and there are no missing values, so there are no missingness or multicollinearity issues. We then examine boxplots of the two quantitative variables Age and Number of hours sitting per day for outliers and remove any errors (ie if greater than 24 was recorded for Number of hours sitting per day). Next, we divide 70% of our data into a training set and 30% into a validation set to use for all the models (note: this form of cross-validation suffers from randomness). We then were interested in comparing three classes of models: (1) Regression, (2) Bayesian, and (3) Tree-based methods. (1) For the regression model, we planned to fit a logistic regression model on the training set then conduct variable selection through stepwise regression with AIC and BIC criteria, after which we will refit the model with the significant predictors and conduct a chi-squared goodness of fit test to ensure the model fits the data adequately well. If it passes, we will choose a threshold via cross-validation that optimizes sensitivity (as we are interested in identifying infertile outcomes) and minimizes misclassification without sacrificing specificity by considering thresholds on a .01 increment from .05 to the maximum threshold such that we will obtain predictions for both outcomes, which is around the maximum posterior probability that an observation in the validation set is Altered. Finally, we will compute misclassification rate,

sensitivity rate, and AUC metrics from predictions on the validation set. (2) For the Bayesian method we will conduct multivariate discriminant analysis where we will first confirm the assumptions of multicollinearity, multivariate normality (where we will also confirm the two quantitative variables are significant through partitioning boxplots of each variable by fertility outcome and eliminating the variables whose distributions do not significantly differ between the outcomes), and equal variance with Box's M Test after which we will conduct analogous variable selection of the qualitative variables through barcharts. We will then fit the model on the training set with the significant predictors, choose a threshold via cross-validation and compute the same metrics. (3) We will lastly attempt 4 tree-based methods: CART, bagging, random forest, and boosting. For CART we will fit the model on the training set and obtain the optimal size through 10-fold cross-validation, then again choose a threshold via cross-validation and compute the same metrics. For the other tree-based methods we will fit the model on the training set, choose a threshold through cross-validation, and compute the same metrics. In addition we will identify the most significant predictors in the bagging and random forest models through variable importance plots, and for the boosting model we will use the var.monotone feature to identify whether each predictor has a positive, negative, or no relationship with the response.

3. Results

3.1 Logistic Regression

The stepwise regression with both AIC and BIC criteria resulted in no significant predictors. We conducted a goodness of fit test to find the model fit the data adequately well ((G(M,M0)=53.713,df=69,p=.91). Further analysis was not possible without predictors.

3.2 Bayesian Modeling

The assumption of multivariate normality was violated as the quantitative variables were eliminated since their distributions did not significantly differ between the two outcomes [Appendix, Figure 1 and Figure 2]; multivariate discriminant analysis could not be performed.

3.3 Tree-Based Methods

3.3a Classification and Regression Tree (CART)

A size of 2 was optimal [Appendix, Figure 3] resulting in our final tree [Appendix, Figure 4]. Any threshold between .05 and .3 performed equally well so a threshold of .3 was chosen to compute the metrics [Appendix, Figure 5].

3.3b Bagging

Thresholds between .29 and .43 equally maximized sensitivity and specificity and minimized misclassification so .3 was again chosen to compute the metrics [Appendix, Figure 6]. The variable importance plot found that Number of hours sitting per day was the most important [Appendix, Figure 7].

3.3c Random Forest

Thresholds between .3 and .41 equally maximized sensitivity and specificity and minimized misclassification so .3 was again chosen [Appendix, Figure 8]. The variable importance plot found that Number of hours sitting per day was the most important [Appendix, Figure 9].

3.3d Boosting

Thresholds between .38 and .5 equally maximized sensitivity and specificity and minimized misclassification so .5 was chosen [Appendix, Figure 10]. The var.monotone feature demonstrated that all the features have no relationship with the response.

3.3e Model Comparison

From Figure 11 all the models perform well with low misclassification and high AUC and sensitivity. However, the boosting model has the highest AUC (.885) and sensitivity (1) although a slightly higher misclassification of .069 compared to the bagging and random forest models which had misclassification of .034. The models besides boosting had the same sensitivity rate of .667 and CART had the highest misclassification of .172 which is characteristic of CART.

4. Discussion

4.1 Discussion

(1) In response to our first research question, none of our factors were significant predictors of male fertility, as shown by the stepwise regression with the logistic model and the var.monotone boosting result. However, from the bagging and random forest trees, Number of hours sitting per day is the most important predictor whereas CART identified Age and Accident or serious trauma as the most important predictors (Age was also ranked highly at second place in bagging and random forest but not Accident or serious trauma). Our findings agree with the literature that age, exercise, and medical issues are important factors in male fertility. (2) In response to our second research question, CART shows Age to be the most prominent factor showing differences in male fertility: A greater proportion of participants aged older than 29.5 years were infertile than those younger than 29.5 years. Additionally, if a participant was older than 29.5 years and had an accident or serious trauma there was an even greater proportion who were infertile (.3) than those in this age group who did not have an accident or serious trauma (.05). This agrees with the literature. (3) In response to our third research question, boosting is the most effective model in predicting male fertility since it has the highest AUC (.885) and sensitivity (1) at the expense of little more misclassification (.069) than the bagging and random forest models. Ultimately, we agreed with the literature that (older) age, exercise, and (the presence of) medical conditions/injuries are significant predictors of male infertility and we hope that our boosting model can preemptively predict male infertility so a solution can be found earlier on.

4.2 Model Considerations & Future Improvements

We were limited in our data, so we were unable to perform stepwise regression or multivariate data analysis. In the future, it would be helpful to find a larger dataset with additional factors that could be significant in predicting male fertility (i.e. tobacco use, overweight) especially quantitative ones so we will be able to conduct multivariate discriminant analysis. Additionally, our dataset was created in 2013, and the interpretability of the results may have changed since then due to medical advancements, and a more recent dataset would more accurately represent the current population. Finally, because our dataset has limited data, we recognize that our models are likely not externally valid so we do not recommend applying them to the general population.

4.3 Data Ethics

There is volunteer bias within our sample so there may be differences between those who volunteered to participate and those who did not; perhaps those who knew they were infertile were less likely to volunteer. Additionally, our dataset is not completely transparent: The data was originally collected by UCI, however, we obtained it from Kaggle, an open-source dataset website which allows users to make edits/contributions to datasets, of which there have been at least four in our dataset but it is unknown what they were.

Appendix

Figure 1: Boxplot of Age for Altered and Normal

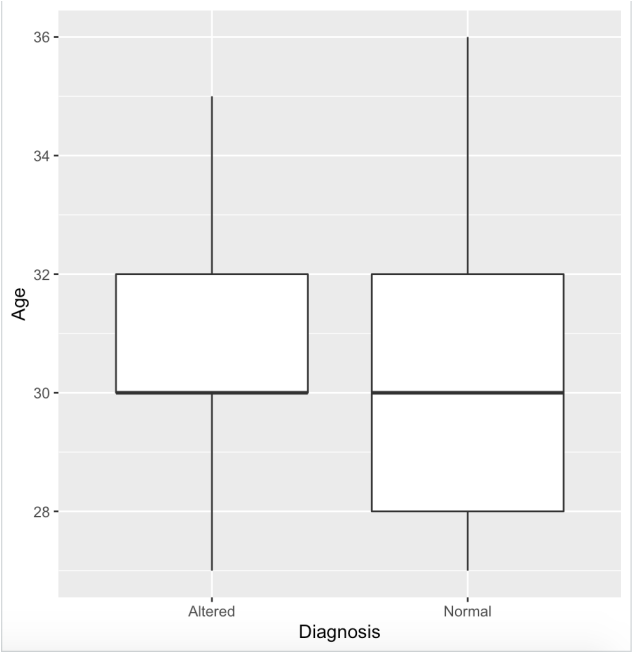

Figure 2: Boxplot of Number of hours sitting per day for Altered and Normal

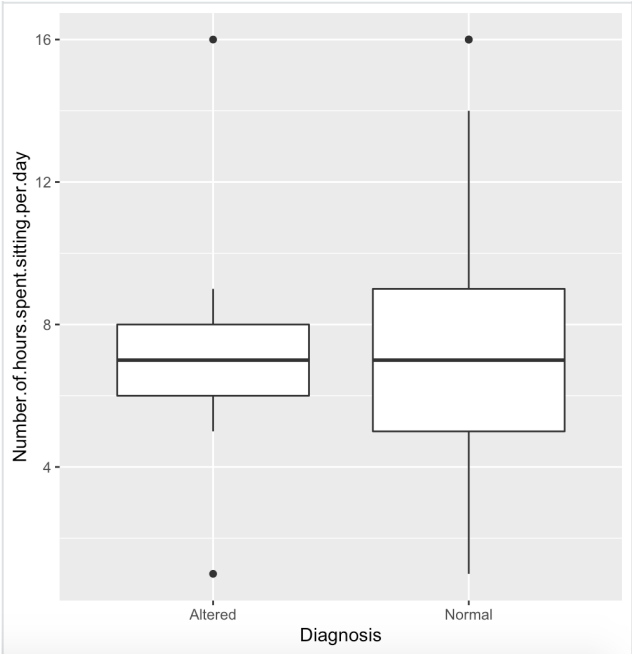

Figure 3: CART Terminal Node Size vs Cross-Validated Error

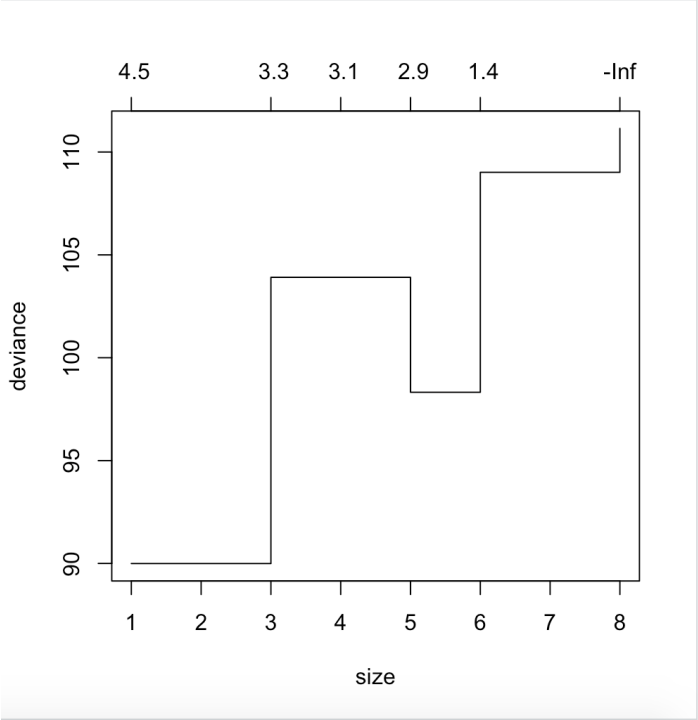

Figure 4: CART for Male Fertility

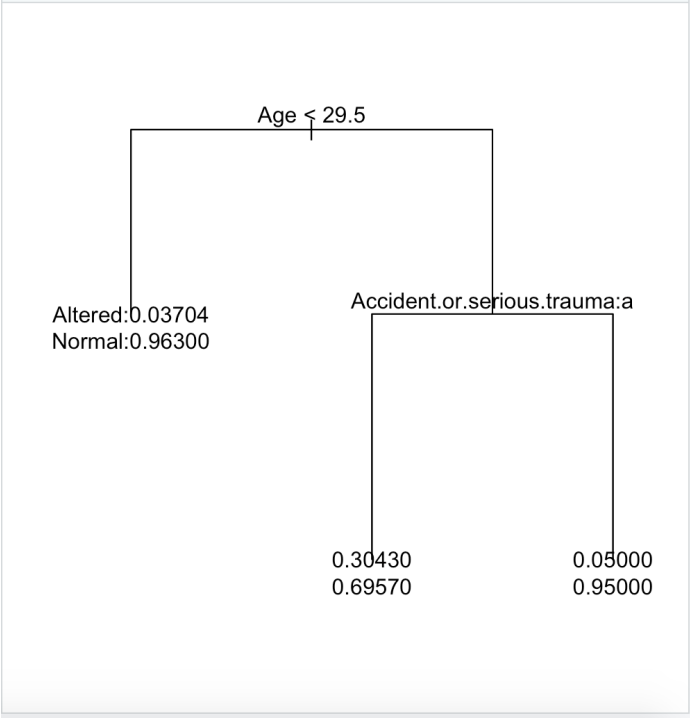

Figure 5: Determining Threshold for CART

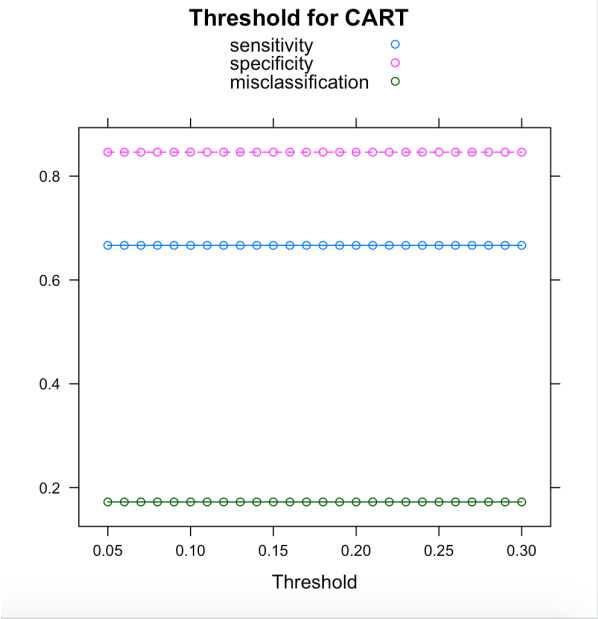

Figure 6: Determining Threshold for Bagging Model

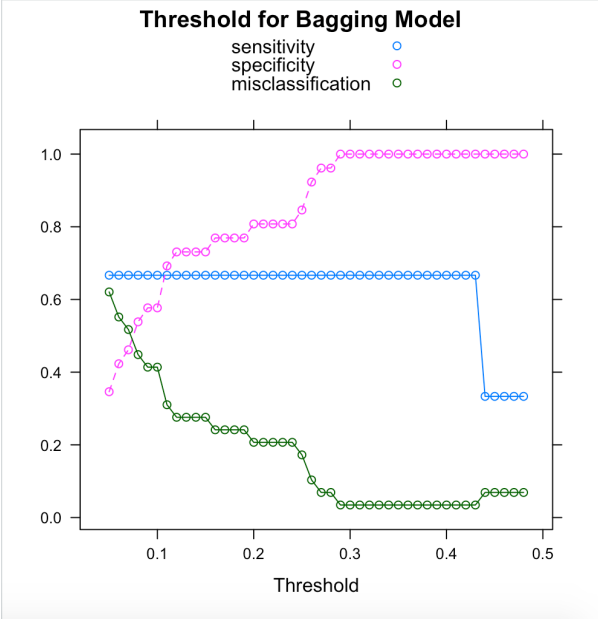

Figure 7: Bagging Model Variable Importance Plot

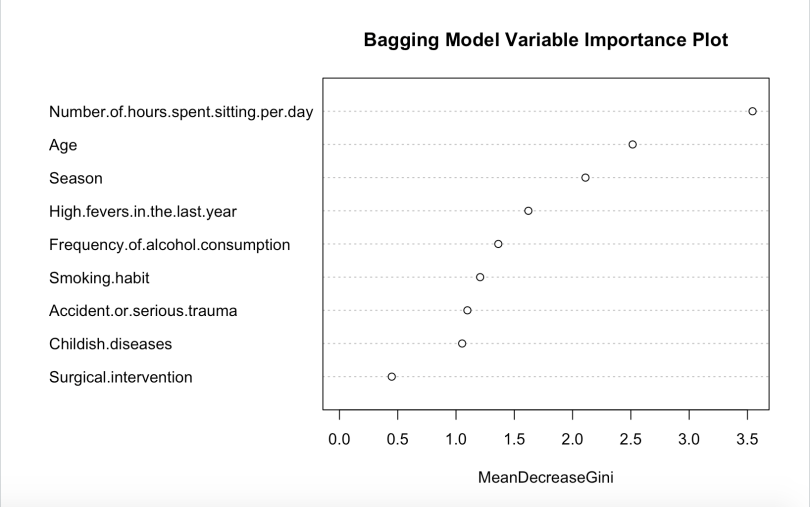

Figure 8: Determining Threshold for Random Forest Model

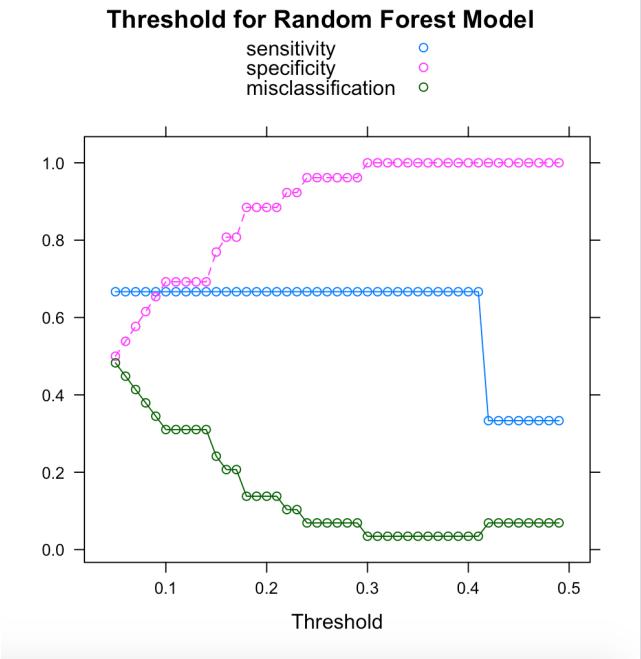

Figure 9: Random Forest Model Variable Importance Plot

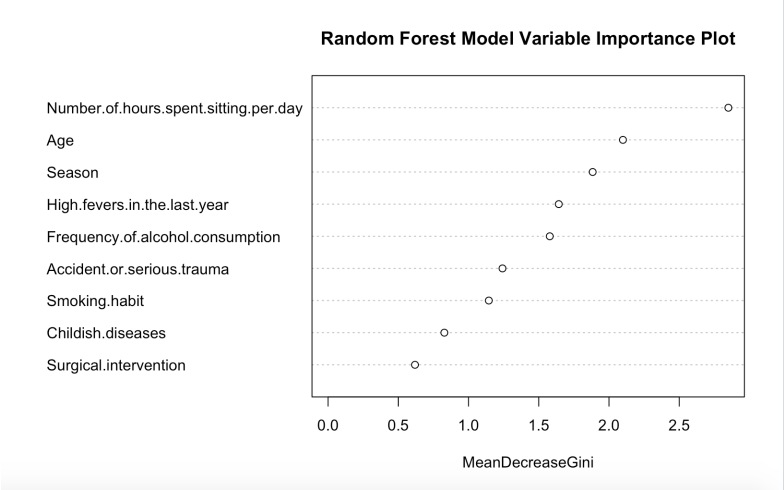

Figure 10: Determining Threshold for Boosting Model

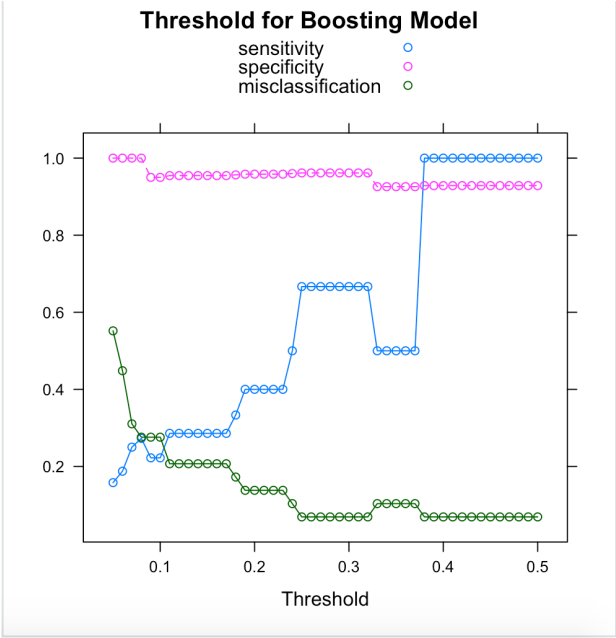

Figure 11: Model Comparison Metrics

Method	**Misclassification**	**Sensitivity**	**AUC**
CART	0.172	0.667	0.840
Bagging	0.034	0.667	0.737
Random Forest	0.034	0.667	0.795
Boosting	0.069	1	0.885

References

[1] Kumar, Naina, and Amit Kant Singh. “Trends of male factor infertility, an important cause of infertility: A review of literature.” Journal of human reproductive sciences vol. 8,4 (2015): 191-6. doi:10.4103/0974-1208.170370.

[2] Harris, Isiah D et al. “Fertility and the aging male.” Reviews in urology vol. 13,4 (2011): e184-90.

[3] Gaskins AJ, Mendiola J, Afeiche M, et al. Physical activity and television watching in relation to semen quality in young men. British Journal of Sports Medicine 2015;49:265-270.

[4] Gaur DS, Talekar MS, Pathak VP. Alcohol intake and cigarette smoking: impact of two major lifestyle factors on male fertility. Indian J Pathol Microbiol. 2010 Jan-Mar;53(1):35-40. doi: 10.4103/0377-4929.59180. PMID: 20090219.