Patient's age prediction based on medical diagnosis of teeth maturity *

De La Torre Camilo *Universite Toulouse 1 Capitole*
Campan Alexis *Universite Toulouse 1 Capitole*

Our study involves patient's age prediction based on teeth maturity that is assesed by doctors. Teeth are characterised by several levels of maturity (going from A to H), data is often incomplete and for some patients, teeth can't be evaluated because the patient does not have them yet. After several attempts to mitigate the incompleteness of our records, we obtained a very low Mean Absolute Error of 0.9 years in predicting unseen data. The model that best generalized to unseen records was a tuned GBRT (Gradient Boosting Regression Tree). Lastly, we also analyse the importance of different teeth maturities and interestingly discovered that not all teeth have the same predictive power on patient's age.

Keywords: Teeth, Age, Regression, Machine Learning, GBRT

Introduction

The data we study explores teeth maturity of several individuals, specifically, we analysed 2847 records. By teeth maturity we mean that each tooth's condition is analysed by a doctor and classed on one of the possible maturity levels (going from A to H, A being the least mature, and H being the most mature). A total of 8 teeth are analysed by doctors, these involve : two incisors teeth, one canine tooth, two premolar teeth, finally, three molar teeth. In addition to teeth maturity indices, we also know the patient's sex, age and medical ID.

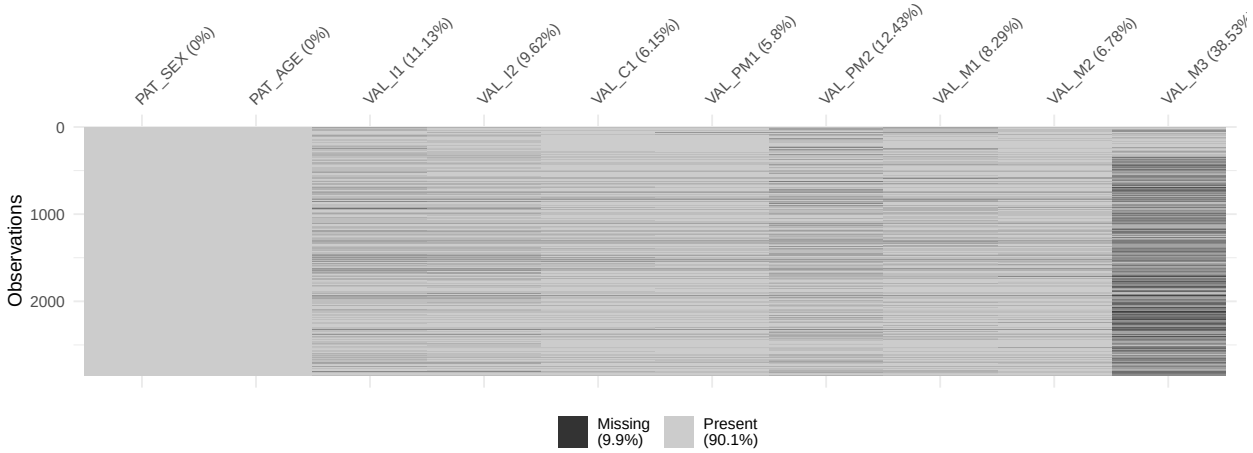
Incompleteness of the data

The data at hand is not complete: out of our 2847 records, for 1436 records at least a tooth score is missing (50% of the data). Moreover, not all teeth are equally misrepresented, as we might expect, molar teeth scores are less likely to be present in our records because several patients do not have them yet. In addition, because of multiple reasons, is it possible that a patient of a given age lost or had removed one or several teeth, this, we think, complicates missing values interpretation because several causes, unrelated to age, can affect it.

Figure X shows missing values per column in our data, moreover, it shows the interaction between missing values across columns.

We observe that our dependent variable, patient's age, is always complete. We also note that the last molar has a very high percentage of missing values. Besides the last molar (M3), another two teeth have over 10% of missing values (I1 and PM2).

*We would like to thanks the doctors who delivered their experienced assesment of teeth maturity as well as our teacher who provided us the data to work with.



Nonetheless, because of the first reason of incompleteness we presented (that teeth grow at different stages in life) we think that imputing missing values makes sense, our approach is explained when we discuss about our methodology.

Methodology

As predictive power is our main goal, we are interested in finding the model that minimizes an error score on unseen data. The score we chose is the Mean Absolute Error defined as $mae = (\frac{1}{n}) \sum_{i=1}^{n} |y_i - x_i|$. ## Ordinal encoding

Handling missing values

dependent Kontopantelis et al. [1]

Outlier removal

Model developement

Baseline results on several algorithms

Results on best tuned GBRT

Feature importances

Conclusion

References

[1] Evangelos Kontopantelis et al. “Outcome-sensitive multiple imputation: a simulation study”. In: *BMC Medical Research Methodology* 17.1 (Jan. 2017), p. 2. ISSN: 1471-2288. DOI: 10.1186/s12874-016-0281-5. URL: https://doi.org/10.1186/s12874-016-0281-5.