Tree-based mixed-effects

# Tree-Structured Methods for Longitudinal Data

Segal

[Https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475220?Journalcode=uasa20](https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475220?journalCode=uasa20)

* The regression tree methodology is extended to repeated measures and longitudinal data by modifying the split function so as to accommodate multiple responses
* Note that these trees cannot be used for the prediction of future periods for the same objects.
* The first using as node impurity a function of the likelihood of an autoregressive or compound symmetry model
* The earliest effort to extend regression tree methodology to longitudinal and [clustered data](https://www.sciencedirect.com/topics/computer-science/clustered-data" \o "Learn more about clustered data from ScienceDirect's AI-generated Topic Pages)
* Does not allow for splitting on time-varying [covariates](https://www.sciencedirect.com/topics/mathematics/covariate" \o "Learn more about covariates from ScienceDirect's AI-generated Topic Pages).
  + Consequently, (1) no random or subject-specific effect of these covariates is allowed, and (2) all repeated observations from a given subject cannot be split across different nodes
  + Note that these trees cannot be used for the prediction of future periods for the same objects.

# RE-EM Trees: A New Data Mining Approach for Longitudinal Data

Rebecca J. Sela · Jeffrey S. Simonoff

[Https://archive.nyu.edu/bitstream/2451/28094/5/Trees\_with\_Random\_Effects%203-28-11.pdf](https://archive.nyu.edu/bitstream/2451/28094/5/Trees_with_Random_Effects%203-28-11.pdf)

* This paper presents a methodology that combines the structure of mixed effects models for longitudinal data with the flexibility of tree-based estimation methods.
  + The RE-EM tree accounts for the structure of longitudinal data while allowing for unbalanced panels and prediction of future time periods
* MERT and RE-EM trees are designed for Gaussian response data.
* The main idea is to fit a tree after removing the random effects part of the model, update the estimates (or predictions) of the random effect and cycle until convergence.
* Includes the ability to use time-varying attributes in the construction of a flexible representation for the underlying relationship between the response and the attributes; indeed, by including time as a potential attribute, it is possible to fit completely different tree structures for different time periods if the tree splits on time.

# Mixed effects random forest for clustered data

Ahlem Hajjem, François Bellavance & Denis Larocque

[Https://www.sciencedirect.com/science/article/pii/S0167715210003433?Casa\_token=frmhmkflliaaaaaa:1yqwd5vb41bjtpiykrqs\_Zx3cj2nn8rdqicq0BhgX2hgL5LdARdCs\_avlwqgyibkfzntda2glapm](https://www.sciencedirect.com/science/article/pii/S0167715210003433?casa_token=FRmhMkfLliAAAAAA:1yqwd5vb41bjtpIYKRQs_Zx3cj2nn8rdqicq0BhgX2hgL5LdARdCs_aVLWQGyIbkFzNtDa2GlapM)

* Proposed a [mixed effects regression](https://www.sciencedirect.com/topics/mathematics/mixed-effect-regression" \o "Learn more about mixed effects regression from ScienceDirect's AI-generated Topic Pages) tree (MERT) method. In contrast to Segal, MERT can appropriately deal with the possible random effects of observation-level covariates and can split observations within clusters since observation-level covariates are candidates in the splitting process.
* The main idea is to fit a tree after removing the random effects part of the model, update the estimates (or predictions) of the random effect and cycle until convergence.
* MERT and RE-EM trees are designed for Gaussian response data.

# Generalized mixed effects regression trees

[Ahlemhajjem](https://www.sciencedirect.com/science/article/pii/S0167715217300895?casa_token=E6GVB61nKDsAAAAA:MjHjDFnRzFHMkF59RkXwf4QAW9-uvotLghTtplAMA_hMoCr0lASgyfmoH-Yy5yjMrV_iFQsgz5dC" \l "!)[a](https://www.sciencedirect.com/science/article/pii/S0167715217300895?casa_token=E6GVB61nKDsAAAAA:MjHjDFnRzFHMkF59RkXwf4QAW9-uvotLghTtplAMA_hMoCr0lASgyfmoH-Yy5yjMrV_iFQsgz5dC" \l "!)[denislarocque](https://www.sciencedirect.com/science/article/pii/S0167715217300895?casa_token=E6GVB61nKDsAAAAA:MjHjDFnRzFHMkF59RkXwf4QAW9-uvotLghTtplAMA_hMoCr0lASgyfmoH-Yy5yjMrV_iFQsgz5dC" \l "!)[b](https://www.sciencedirect.com/science/article/pii/S0167715217300895?casa_token=E6GVB61nKDsAAAAA:MjHjDFnRzFHMkF59RkXwf4QAW9-uvotLghTtplAMA_hMoCr0lASgyfmoH-Yy5yjMrV_iFQsgz5dC" \l "!)[françoisbellavance](https://www.sciencedirect.com/science/article/pii/S0167715217300895?casa_token=E6GVB61nKDsAAAAA:MjHjDFnRzFHMkF59RkXwf4QAW9-uvotLghTtplAMA_hMoCr0lASgyfmoH-Yy5yjMrV_iFQsgz5dC" \l "!)[b](https://www.sciencedirect.com/science/article/pii/S0167715217300895?casa_token=E6GVB61nKDsAAAAA:MjHjDFnRzFHMkF59RkXwf4QAW9-uvotLghTtplAMA_hMoCr0lASgyfmoH-Yy5yjMrV_iFQsgz5dC" \l "!)

[Https://www.sciencedirect.com/science/article/pii/S0167715217300895?Casa\_token=E6GVB61nKDsAAAAA:mjhjdfnrzfhmkf59rkxwf4qaw9-uvotlghttplama\_hmocr0lasgyfmoh-Yy5yjMrV\_ifqsgz5dc#b5](https://www.sciencedirect.com/science/article/pii/S0167715217300895?casa_token=E6GVB61nKDsAAAAA:MjHjDFnRzFHMkF59RkXwf4QAW9-uvotLghTtplAMA_hMoCr0lASgyfmoH-Yy5yjMrV_iFQsgz5dC#b5)

* Following the steps of the [generalized linear mixed models](https://www.sciencedirect.com/topics/mathematics/generalized-linear-mixed-model" \o "Learn more about generalized linear mixed models from ScienceDirect's AI-generated Topic Pages) (glmms), we propose a tree based method, named “generalized mixed effects regression tree” (GMERT), which is suitable for non-gaussian data (e.g., binary outcomes and count data). The proposed GMERT method can handle unbalanced clusters, and can incorporate observation-level covariates and their potential random effects.
* This extension uses the penalized quasi-likelihood (PQL) method for the estimation and the expectation-maximization (EM) algorithm for the computation

# Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin a1c

* Ngufor, formalized the problem of longitudinal/clustered supervised machine learning, as that of learning the two components of a non-linear mixed-effects model separately through an iterative expectation maximization-like algorithm, in which we alternatively estimate the fixed-effect component using machine learning methods and the random-effect component using GLMM.

Boosting

# Regularization for Generalized Additive Mixed Models by Likelihood-Based Boosting

[Andreas Groll](https://pubmed.ncbi.nlm.nih.gov/?term=Groll%20A%5BAuthor%5D)\* and  [Gerhard Tutz](https://pubmed.ncbi.nlm.nih.gov/?term=Tutz%20G%5BAuthor%5D)#

# A Gradient Boosting Machine for Hierarchically Clustered Data

[Patrick J. Miller](https://pubmed.ncbi.nlm.nih.gov/?term=Miller%20PJ%5BAuthor%5D), [Daniel B. Mcartor](https://pubmed.ncbi.nlm.nih.gov/?term=McArtor%20DB%5BAuthor%5D), and  [Gitta H. Lubke](https://pubmed.ncbi.nlm.nih.gov/?term=Lubke%20GH%5BAuthor%5D)

# Gaussian Process Boosting

Fabio Sigrist

# Generalized Linear Mixed Models Based on Boosting

Gerhard Tutz and Andreas Groll

# Gradient boosting for linear mixed models

Colin griesbach, benjamin sa ̈ fken and elisabeth waldmann

# Model-based Boosting in R A Hands-on Tutorial Using the R Package mboost

Benjamin Hofner∗† Andreas Mayr† Nikolay Robinzonov‡ Matthias Schmid†

# Addressing imbalanced insurance data through zero-inflated poisson regression with boosting

[Simon C.K. Lee](https://www.cambridge.org/core/search?filters%5BauthorTerms%5D=Simon%20C.K.%20Lee&eventCode=SE-AU)

Xgboost in healthcare

# Prediction Model of Dementia Risk Based on xgboost Using Derived Variable Extraction and Hyper Parameter Optimization

Seong-eun ryu 1, dong-hoon shin 2, and kyungyong chung 1

<https://sci-hub.se/10.1109/access.2020.3025553>

* “ID variable is removed from independent variables, since it does not influence the dependent variable CDR at all.” This research used the open source OASIS-1 and OASIS-2 data\footnote{Combining these data sets is adviced against by OASIS, so this analysis might contain more flaws on the data processing aspect, than discussed here.} which is a combination of cross-sectional and longitdudinal data. Through deleting the ID variable, and ignoring the correlated structure inherent to the data, this research fails to address are the problems that arise with using machine learning methods using longitudinal data. Therefore, their conclusions are unreliable.

# Risk prediction for repeated measures health outcomes: A divide and recombine framework

l[rafiqul I.chowdhury](https://www.sciencedirect.com/science/article/pii/S235291482200003X" \l "!)[a](https://www.sciencedirect.com/science/article/pii/S235291482200003X" \l "!)[jabed H.Tomalb](https://www.sciencedirect.com/science/article/pii/S235291482200003X#!)

* In statistical parlance, the research questions are: (i) calculating the risk of health conditions for a patient at a time point given the risk factors and the history of health conditions (i.e., calculating conditional probability), (ii) estimating the risk of health conditions over a sequence of time points given the risk factors and the history of health conditions observed over time (i.e., the joint probability), (iii) predicting the health condition for the next time point given the risk factors and the history of health conditions (i.e., predicting future states of the disease)
* Then from the second onward follow-ups, previous outcomes are added to incorporate temporal dependency according to the proposed framework.
  + In order to share information using models between time points, we propose to include previous outcomes as covariates starting second time point and beyond
* The proposed framework divides the complex multivariate problem into several univariate problems by fitting a model to the data specific to a single subset.
* The proposed framework of this paper divides the data into multiple subsets using observed time points. Then fits a marginal model to the first subset and a conditional model to the subsequent subsets and recombines the models to obtain joint probabilities, which are used to predict the sequence of responses for all the time points and trajectories. The chain rule of probability theory is used to recombine the marginal, conditional and joint models.
* This framework employed several statistical and machine learning models to obtain the marginal and conditional models as the base learner. The divide and recombine framework is used on top of the statistical and machine learning models as an umbrella method.
* We calculate marginal and conditional probabilities using statistical and machine learning models to predict trajectory risks using the proposed framework. Using the framework, and by augmenting the covariates using previous responses, we extended the following machine learning algorithms originally developed for cross-sectional data to predict the risk trajectories for repeated responses.
* It may be noted that all the predicted probabilities are for individual patients.
* However, to keep the manuscript short, we did not try some learning algorithms recently gained popularity and attention, such as XGBoost. In the proposed framework, one can easily employ this algorithm. The goal is to check if these models along with the divide and recombine framework are useful to answer the research questions posed earlier.

Xgboost, a Machine Learning Method, Predicts Neurological Recovery in Patients with Cervical Spinal Cord Injury

Tomoo Inoue, Daisuke Ichikawa, Taro Ueno, Maxwell Cheong, Takashi Inoue, William D. Whetstone, Toshiki Endo, Kuniyasu Nizuma, and Teiji Tominaga

* Niet longitudinal

Interpretable classifiers for prediction of disability trajectories using a nationwide longitudinal database

Yafei Wu, Chaoyi Xiang, Maoni Jia & Ya Fang

* We used growth mixture model (GMM) to identify the heterogeneous disability trajectories of the targeted population from 2002 to 2018. GMM is able to establish several latent category groups considering individual and population heterogeneity, and the individuals in each category group enjoy the same or similar average growth trajectory (the same intercept and slope), which is used to describe the changes of individuals in category groups over time
* In this study, five ML methods were included to carry out three-class and two-class outcome predictions on the whole data set and the feature selection data set screened by LASSO.
* In addition, we discovered that the ensemble learning method (RF and XGBoost) had the best performance among the five models, which is attributed to the fact that it form a stronger classifier by combining multiple weak learners [37], even if some weak ones get wrong predictions, the others can also correct the error to varying degrees [36].
* The mixed growth model was used to identify disability trajectories, and five machine learning models were further established to predict disability trajectories using epidemiological variables.
* SHAP

page1image493644928

Xgboost, a novel explainable AI technique, in the prediction of myocardial infarction, a UK Biobank cohort study

Alexander Moore1, Max Bell 2,3

* We compared two machine learning methods, XGBoost and logistic regression in predicting risk of MI
* Wel panel data
* Did nothing to account for intertemporal effects

# A comparative analysis of machine learning approaches to predict C. Difficile infection in hospitalized patients

Saarang Panchavati BS 1, Nicole S. Zelin MD 1, Anurag Garikipati MS, Emily Pellegrini meng, Zohora Iqbal phd \*, Gina Barnes MPH, Jana Hoffman phd, Jacob Calvert msc, Qingqing Mao phd, Ritankar Das msc

* For continuously measured features, such as vital sign and laboratory measurements, XGBoost used the last measured value, as well as summary statistics (mean, standard deviation) of all values measured during the first 6 hours of hospital care.
* Risk stratification of individual patients early in a hospital admis- sion may assist in the prevention of CDI in high risk patients.25,26 Using real-world EHR data, we developed 3 different MLAs to predict CDI at any point during an inpatient stay based on only 6 hours of data. In clinical practice, use of a machine learning-based CDI predic- tion tool may enable patients to benefit from increased monitoring and treatment earlier in their disease course and facilitate timely implementation of appropriate infection control practices.
* Our results demonstrate that MLAs can predict CDI with excellent discrimination (AUROC > 0.8).27 Many of the important features used by the models to predict CDI were similar across MLAs, and have pre- viously been identified as risk factors for CDI.11,12 The highest per- forming MLA in terms of AUROC values was the XGBoost model, while the neural networks achieved higher sensitivities at the opti- mized operating points.
* We have demonstrated that MLAs using just the first 6 hours of hospitalization data can predict CDI with high discrimination. We have also shown that XGBoost can achieve comparable predictive performance to the more complex neural networks, and that differ- ent training techniques to account for the low prevalence of CDI in training data are optimal for different MLA architectures.

%% Tree-based mixed-effects

\cite{segal1992tree}

- The regression tree methodology is extended to repeated measures and longitudinal data by modifying the split function so as to accommodate multiple responses

- Note that these trees cannot be used for the prediction of future periods for the same objects.

- The first using a function of the likelihood of an autoregressive or compound symmetry model as node impurity

- The earliest effort to extend regression tree methodology to longitudinal and clustered data

- Does not allow for splitting on time-varying covariates.

o Consequently, (1) no random or subject-specific effect of these covariates is allowed, and (2) all repeated observations from a given subject cannot be split across different nodes

o Note that these trees cannot be used for the prediction of future periods for the same objects.

\cite{sela2009re}

- This paper presents a methodology that combines the structure of mixed effects models for longitudinal data with the flexibility of tree-based estimation methods.

o The RE-EM tree accounts for the structure of longitudinal data while allowing for unbalanced panels and prediction of future time periods

- MERT and RE-EM trees are designed for Gaussian response data.

- The main idea is to fit a tree after removing the random effects part of the model, update the estimates (or predictions) of the random effect and cycle until convergence.

- Includes the ability to use time-varying attributes in the construction of a flexible representation for the underlying relationship between the response and the attributes; indeed, by including time as a potential attribute, it is possible to fit completely different tree structures for different time periods if the tree splits on time.

\cite{HAJJEM2011451mert}

- Proposed a mixed effects regression tree (MERT) method. In contrast to Segal, MERT can appropriately deal with the possible random effects of observation-level covariates and can split observations within clusters since observation-level covariates are candidates in the splitting process.

- The main idea is to fit a tree after removing the random effects part of the model, update the estimates (or predictions) of the random effect and cycle until convergence.

- MERT and RE-EM trees are designed for Gaussian response data.

\cite{HAJJEM2017114gemert}

- Following the steps of the generalized linear mixed models (glmms), we propose a tree based method, named “generalized mixed effects regression tree” (GMERT), which is suitable for non-gaussian data (e.g., binary outcomes and count data). The proposed GMERT method can handle unbalanced clusters, and can incorporate observation-level covariates and their potential random effects.

- This extension uses the penalized quasi-likelihood (PQL) method for the estimation and the expectation-maximization (EM) algorithm for the computation

\cite{ngufor2019mixed}

- Ngufor, formalized the problem of longitudinal/clustered supervised machine learning, as that of learning the two components of a non-linear mixed-effects model separately through an iterative expectation maximization-like algorithm, in which we alternatively estimate the fixed-effect component using machine learning methods and the random-effect component using GLMM.

**xxx**

%% Xgboost in healthcare

\cite{ryu2020prediction}

- “ID variable is removed from independent variables, since it does not influence the dependent variable CDR at all.” This research used the open source OASIS-1 and OASIS-2 data\footnote{Combining these data sets is advised against by OASIS, so this analysis might contain more flaws on the data processing aspect, than discussed here.} which is a combination of cross-sectional and longitudinal data. Through deleting the ID variable, and ignoring the correlated structure inherent to the data, this research fails to address are the problems that arise with using machine learning methods using longitudinal data. Therefore, their conclusions are unreliable.

\cite{moore2022xgboost}

- We compared two machine learning methods, XGBoost and logistic regression in predicting risk of MI

- Wel panel data

- Did nothing to account for intertemporal effects

\cite{panchavati2022comparative}

- For continuously measured features, such as vital sign and laboratory measurements, XGBoost used the last measured value, as well as summary statistics (mean, standard deviation) of all values measured during the first 6 hours of hospital care.

- Risk stratification of individual patients early in a hospital admis- sion may assist in the prevention of CDI in high risk patients.25,26 Using real-world EHR data, we developed 3 different MLAs to predict CDI at any point during an inpatient stay based on only 6 hours of data. In clinical practice, use of a machine learning-based CDI predic- tion tool may enable patients to benefit from increased monitoring and treatment earlier in their disease course and facilitate timely implementation of appropriate infection control practices.

- Our results demonstrate that MLAs can predict CDI with excellent discrimination (AUROC > 0.8).27 Many of the important features used by the models to predict CDI were similar across MLAs, and have pre- viously been identified as risk factors for CDI.11,12 The highest per- forming MLA in terms of AUROC values was the XGBoost model, while the neural networks achieved higher sensitivities at the opti- mized operating points.

- We have demonstrated that MLAs using just the first 6 hours of hospitalization data can predict CDI with high discrimination. We have also shown that XGBoost can achieve comparable predictive performance to the more complex neural networks, and that differ- ent training techniques to account for the low prevalence of CDI in training data are optimal for different MLA architectures.

\cite{inoue2020xgboost}

- Niet longitudinal

\cite{wu2022interpretable}

- We used growth mixture model (GMM) to identify the heterogeneous disability trajectories of the targeted population from 2002 to 2018. GMM is able to establish several latent category groups considering individual and population heterogeneity, and the individuals in each category group enjoy the same or similar average growth trajectory (the same intercept and slope), which is used to describe the changes of individuals in category groups over time

- In this study, five ML methods were included to carry out three-class and two-class outcome predictions on the whole data set and the feature selection data set screened by LASSO.

- In addition, we discovered that the ensemble learning method (RF and XGBoost) had the best performance among the five models, which is attributed to the fact that it form a stronger classifier by combining multiple weak learners [37], even if some weak ones get wrong predictions, the others can also correct the error to varying degrees [36].

- The mixed growth model was used to identify disability trajectories, and five machine learning models were further established to predict disability trajectories using epidemiological variables.

- SHAP

\cite{chowdhury2022risk}

- In statistical parlance, the research questions are: (i) calculating the risk of health conditions for a patient at a time point given the risk factors and the history of health conditions (i.e., calculating conditional probability), (ii) estimating the risk of health conditions over a sequence of time points given the risk factors and the history of health conditions observed over time (i.e., the joint probability), (iii) predicting the health condition for the next time point given the risk factors and the history of health conditions (i.e., predicting future states of the disease)

- Then from the second onward follow-ups, previous outcomes are added to incorporate temporal dependency according to the proposed framework.

o In order to share information using models between time points, we propose to include previous outcomes as covariates starting second time point and beyond

- The proposed framework divides the complex multivariate problem into several univariate problems by fitting a model to the data specific to a single subset.

- The proposed framework of this paper divides the data into multiple subsets using observed time points. Then fits a marginal model to the first subset and a conditional model to the subsequent subsets and recombines the models to obtain joint probabilities, which are used to predict the sequence of responses for all the time points and trajectories. The chain rule of probability theory is used to recombine the marginal, conditional and joint models.

- This framework employed several statistical and machine learning models to obtain the marginal and conditional models as the base learner. The divide and recombine framework is used on top of the statistical and machine learning models as an umbrella method.

- We calculate marginal and conditional probabilities using statistical and machine learning models to predict trajectory risks using the proposed framework. Using the framework, and by augmenting the covariates using previous responses, we extended the following machine learning algorithms originally developed for cross-sectional data to predict the risk trajectories for repeated responses.

- It may be noted that all the predicted probabilities are for individual patients.

- However, to keep the manuscript short, we did not try some learning algorithms recently gained popularity and attention, such as XGBoost. In the proposed framework, one can easily employ this algorithm. The goal is to check if these models along with the divide and recombine framework are useful to answer the research questions posed earlier.