# Analyzing the Effect of Noise Injection on Gradient Boosting Machine Regression Models

Hamza Bachnak, *Student, Pennsylvania State University*, Jeremy J. Blum, *Professor, Pennsylvania State University*, Mahfuza Farooque, *Professor, Pennsylvania State University*

**Abstract**— **Decision trees are a widely used algorithm for regression tasks. Much like its name suggests, a decision tree resembles a tree, beginning from a single root node, splitting off into different branch nodes, and finally ending on leaf nodes. Typically, for a regression task, a single value in the leaf node is used for a prediction. The discrete nature of trees results in outputs that are discontinuous. In gradient boosting machines (GBMs), multiple decision trees are then bundled together to produce output. We hypothesize that one reason for the large number of trees in GBM models is the discontinuous nature of tree-based predictions, and therefore, if other means were used to smooth out GBM predictions, smaller models could produce similar results. This research investigates the ability of noise injected into input vectors during the prediction to allow for smaller GBM model, potentially reducing runtime and memory usage, without sacrificing model accuracy.**

I. INTRODUCTION

One of the more heavily utilized tools in machine learning is the decision tree algorithm. In following decision tree, we note the new cases of larynx cancer within different age groups of different sexes using data provided by Center of Disease Control [1]. The decision tree indicates the factor of age to display various odds of contracting larynx cancer.

A diagram of a number of cells

Description automatically generated with medium confidence

Rates of New Larynx Cancers By Age Reported By CDC [1]

To use the decision tree, shown above, we would look at the age group we are currently targeting. If the age of the individual is within the range of age group 1 (45 <= x <= 54), then we would proceed down the left branch. Otherwise, we would see if the age of the individual satisfies age group 2’s age range and proceed down the right branch, if so. This process is repeating till we reach a leaf node containing a probability value.

A graph of a number of people

Description automatically generated

Rates of New Larynx Cancers by Age Reported by CDC Represented in Bar Graph [1]

Looking closely at the prior graph, we can observe a large discrepancy of being diagnosed with larynx cancer between ages 54 and 55 with a drastic increase of nearly double. This drastic increase in probability leads to discontinuities when interpreting such data. These discontinuities within data can lead to more extreme gaps between variables with close values when producing prediction models and thus a less accurate model.

To help combat such gaps within data, different types decision tree-base models have been utilized within machine learning. Techniques such as Gradient Boosting Machines (GBM’s) and Random Forest both create vast forests of trees. Other techniques such as Linear Model apply linear regression models in the leaves of produced linear trees. Each of these approaches sacrifice computational runtime, the size and scale of the produced models, or both.

We hypothesize that one reason for the large number of trees in GBM models is the discontinuous nature of tree-based predictions, and therefore, if other means were used to smooth out GBM predictions, smaller models could produce similar results. This hypothesis is tested by train a baseline GBM using lightGBM followed by training a smaller GBM using fewer trees. When making predictions using the smaller GBM model, noise is added to the input vectors with a goal of smoothing out the predictions. The accuracy of the smoothed predictions is then compared to the baseline GBM. For this research, we will be analyzing datasets of used US cars, provided on Kaggle.

II. Literature Review

Efficiency and speed are two of the most prominent factors when it comes to performance in this age of technology. To facilitate tasks to reach peak performance, people create sets of instructions, formulas, or algorithms. The dictionary of Britannica diligently defines an algorithm as “a specific procedure for solving a well-defined computational problem ​[2]​.” Within the field of computer science, algorithms are necessary to help create working formulas and functions to conduct tasks, from artificial intelligence to databases, security programs, and so forth. They also help us understand why certain methods are able to overpower others, logically.

As technology has advanced in recent years, algorithms and complex systems have taken advantage of available resources. One field that has spawned by such technology and has grown exponentially is Machine Learning, as many different types of algorithms have been developed in this field. Over the years, different attempts have been made to develop more efficient algorithms to help produce models, to be applied to systems.

*A.* *Regression versus Classification*

Before delving into specific types of Machine Learning algorithms, an important detail to note is the type of tasks the algorithm is supposed to accomplish. Such tasks may be split into two groups: Regression and Classification. Classification algorithms are designed to map values, assigning them to a predefined label, thus categorizing them. An example of classification may be that of labeling produce as fruit or vegetable, based on variables such as seeds, roots, leaves, etc. Regression, on the other hand, is the process of mapping data to find a relationship between the variable to produce a model that may predict continuous outcomes. An example of regression may be that of predicting the average amount of rainfall to occur in a week based on prior data and current variables such as humidity and temperature. The major distinction between both classification and regression is that classification algorithms provide a discrete label, while regression algorithms provide an output of continuous values​ [3]​​ [4]​​ [5]​.

*B. Decision Trees*

Decision trees are one of the more utilized methods used in machine learning to help train new models using provided data sets. Much like its name infers, a decision tree resembles a tree, beginning from a single root node, splitting off into different branch nodes, and finally ending on leaf nodes. A simple example of a decision tree would be a question of: “Is it currently raining?” If the answer in the example is a yes, the decision tree would proceed to a leaf node that may provide a response of “bring an umbrella,” overwise, the tree will proceed to another leaf node with a conclusion of “no umbrella needed”. A decision tree does not have to be binary but such trees tend to be the most common in machine learning​ [6]​​ [7]​.

*C. Gradient Boosting Machines*

Ensemble learning is the combination of different models to produce a better performing model. The Gradient Boosting Model (GBM) is an algorithm that makes use of ensemble learning by training several models sequentially. GBM then predicts errors or outstanding values of initial models and sums up to find a starting model. Afterwards, one weak learner is added at a time to the model and previous learners are kept unchanged. GBM soon analyzes and weighs the patterns displayed and strengthens the model accordingly, based on the weak predictions. The modeling is stopped once the GBM cannot locate any pattern to the model ​ [7]​​ [8]​​ [9]​.

*D. Bagging, Random Forest, and Boosting algorithms*

Bagging is a type of ensemble learning method that combines results and takes the mean to produce a resulting model. The algorithm can be split into three key steps:

1. Randomly sampled datasets of original training data (bootstrapping).
2. Build and fit several classifiers to each of the diverse copies.
3. Take the average of all predictions to produce final result.

Random forest is another ensemble learning method in which multiple random decision trees from a set combine to produce a resulting, more generalized model. Finally, and the most relevant ensemble learning method to this project, is the boosting method. The basic concept behind boosting is correcting errors of previous models and using them to train newer models based off the more improved pattern ​[7]​​ [11]​​ [12]​.

*E. Data Gaps and Outliers*

Some issues that may arise before and after processing data to produce models are gaps, outliers, and overfitted data within both the training and test data. All of these reduce legibility and effectiveness of the models produced.  ​ [7]​​ [8]​​ [11]​​ [12]​

*F. Prior attempts*

Prior attempts to smoothen a curve, without overfitting data, using both greedy and non-greedy methods have been used ​[10]​. Most of these methods tend to focus on the processing phase of data while using known and working pre-processing techniques to help produce both a training set of data as well as test set of data ​[7]​​ [11]​​ [12]​.

III. METHODS

To test the hypothesis, we constructed a baseline GBM followed by training a smaller GBM, with fewer trees. Before testing both GBM’s, we shuffled the Dataset . After shuffling the dataset, we split apart the dataset into 3 distinct chunks: 20% test data, 20% validation data, 60% training data. Following the data split, we standardized individual chunks using z-score normalization in relationship to the test data. We used one-hot encoding to help improve upon the applicability of categorical data. Afterwards, the training of each individual GBM began.

1. *Baseline GBM*

During the training process of the baseline GBM, we used an early stopping technique to help determine the optimal number of trees necessary to find an efficient GBM model. We used this trained model to fit the dataset onto the validation set and determine the RMSE (Root-Mean-Square Error) value from it.

A diagram of a training process

Description automatically generated Flowchart Representation for Construction of Baseline GBM

1. *Smaller GBM*

After training and validating the baseline GBM model, we trained the smaller GBM model using a ratio of estimators used within the baseline model (concluding in a ratio of trees between the baseline GBM model and the smaller GBM model). After fitting the dataset using the trained model, we remedy the reduced quantity of trees, averaging iterations of Gaussian noise to each produced predication set.

We had determined multiple ratios of estimators to help display a trend within the effectiveness of the smaller GBM model. The Gaussian noise applied upon the validation set to produce the prediction set, was decided using a range of 0 to 0.1 applied on the valid data.

A diagram of a company

Description automatically generated

Flowchart Representation for Construction of Smaller GBM

1. *Testing*

After constructing both the baseline GBM and the smaller GBM, we applied each model to the test data. Data collected from each model consisted of RMSE, time, and memory. Finally, we record and compare such data collected by repeating this process, reiterating from shuffling the datasets.

IV. RESULTS

Currently, we have not produced substantial results to derive conclusions from but intend to continue researching this topic. Our goal is to publish the results of this extended research during the Penn state Fall 2023 Honors thesis exhibition.

V. CONCLUSION

In this research, we had prepared and presented our hypothesis of the potentiality of reducing the number of trees within a forest of GBM without sacrificing the smoothness or efficiency of the produced model. We trained a baseline GBM using lightGBM followed by training a smaller GBM using fewer trees to help test this thesis. Results on this research are inconclusive, but we intend on extending the research throughout the upcoming Fall 2023 semester and plan to present our findings during the Penn State Harrisburg Fall 2023 Honors thesis exhibition.

A future research discussion may involve research into a particular algorithm to help determine the most optimal ratio is a potential future research topic, as we had set ratios to be percentage-based and are potentially suitable for this dataset.

# REFERENCES

|  |  |
| --- | --- |
| [1] | Center of Disease Control and Prevention, "cdc.gov," Center of Disease Control and Prevention, 2023. [Online]. Available: https://gis.cdc.gov/Cancer/USCS/#/Demographics/. [Accessed 26 June 2023]. |
| [2] | I. Encyclopædia Britannica, "Britannica," Encyclopædia Britannica, Inc, 6 May 2021. [Online]. Available: https://www.britannica.com/science/computer-science/Algorithms-and-complexity. [Accessed 8 April 2022]. |
| [3] | J. Brownlee, "Difference Between Classification and Regression in Machine Learning," 11 December 2017. [Online]. Available: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/. [Accessed 22 5 2022]. |
| [4] | Y. Su, "Prediction of air quality based on Gradient Boosting Machine Method," in *2020 International Conference on Big Data and Informatization Education (ICBDIE)*, Zhangjiajie, China, 2020. |
| [5] | V. V. G. M. G. E. G. Spyros Theocharides, "Day-ahead Forecasting of Solar Power Output from Photovoltaic Systems Utilising Gradient Boosting Machines," in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC) (A Joint Conference of 45th IEEE PVSC, 28th PVSEC & 34th EU PVSEC)*, Waikoloa, HI, USA, 2018. |
| [6] | W.-Y. Loh, "Regression tree models for," in *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, Institute of Mathematical Statistics, 2006, pp. 210-228. |
| [7] | P. Cunningham, "Ensemble Techniques," Academia, 2007. |
| [8] | Y. Shi, J. Li and Z. Li, "Gradient Boosting with Piece-Wise Linear Regression Trees," in *Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, 2019. |
| [9] | E. A. P. A. I. A. A. J. J. M. H. H. A. N. C. A. Norshakirah Aziz, "A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems," in *2020 International Conference on Computational Intelligence (ICCI)*, Bandar Seri Iskandar, Malaysia, 2020. |
| [10] | J. Xie, V. Rojkova, S. Pal and S. Coggeshall, "A Combination of Boosting and Bagging for KDD Cup 2009," in *Journal of Machine Learning Research*, San Diego, 2009. |
| [11] | A. V. Konstantinov and L. V. Utkin, "A Generalized Stacking for Implementing," in *Cyber-Physical Systems*, Springer Cham, 2021, pp. 3-16. |
| [12] | V. O. Zhilinskiy, "GLONASS Satellite Pseudorange Errors Mitigation Using Gradient Boosting Machine," in *2021 XV International Scientific-Technical Conference on Actual Problems Of Electronic Instrument Engineering (APEIE)*, Novosibirsk, Russian Federation, 2021. |
| [13] | J. C. Xueping Zhu, "Risk Prediction of P2P Credit Loans Overdue Based on Gradient Boosting Machine Model," in *2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, Shenyang, China, 2021. |
| [14] | T. S. A Yovan Felix, "Flood Detection Using Gradient Boost Machine Learning Approach," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dubai, United Arab Emirates, 2019. |
| [15] | J. Raymaekers, P. J. Rousseeuw, T. Verdonck and R. Yao, "Fast Linear Model Trees by PILOT," 2023. |