MATH 5472 Course Project

Can Yang, Department of Mathematics, HKUST

October 30, 2025

Project Title: What If Without the XXX Method

Objective:
The goal of this project is to critically analyze a well-established method in statistical machine learning by exploring its necessity, approach, benefits, and competitors. This will help you develop a deeper understanding of the method's significance in the field.

Instructions:

1. **Choose a Method:**

   Select a well-established method in statistical machine learning to focus on for your essay. Examples include methods like Support Vector Machines, Decision Trees, Neural Networks, etc.

2. **Essay Structure:**
   Your essay should be organized into four main sections, following the NABC framework:
   - **Need:**
     Discuss the problem or need that the XXX method addresses. Why was this method developed? What gaps in knowledge or practice does it fill?
   - **Approach:**
     Describe how the XXX method works. What are its main principles and techniques? Provide a clear overview of its operational framework.
   - **Benefits:**
     Highlight the advantages of using the XXX method. What are its strengths compared to other methods? How has it contributed to advancements in statistical machine learning?
   - **Competitors:**
     Identify other methods that serve as alternatives or competitors to the XXX method. Compare their approaches, benefits, and limitations in relation to the XXX method.
3. **Essay Requirements:**
   - Length and Format: **the NeurIPS paper format**.
   - References: Include at least 5 scholarly references to support your analysis.
4. **Submission Guidelines:**
   - Submit your essay as a PDF document by **December 10, 2025**.
   - Ensure that your name, student ID, and course title are included on the title page.
5. **Evaluation Criteria:**
   Your essay will be evaluated based on the following criteria:
   - Clarity and coherence of ideas
   - Depth of analysis in each section (Need, Approach, Benefits, Competitors)

- Quality of writing and adherence to academic standards
- Proper use of references and citations

Hint:

- Do not simply use words but conduct experiments (simulations or real-world evidence are fine; but real-world evidence is preferred). **Eassy without experimental results can earn at most 5 points**.
- You should setup your Github website such that your results are accessible and reproducible. Please include this website in your essay.
- An example of the essay given by DeepSeek is provided at the end of this document. The total points for this project are 40, while the DeepSeek version can only earn less than 5 points. When you prepare your essay, please think carefully about how your essay can add value compared to general AI tools.
- You may consider methods listed as "Ten Statistical Ideas that Changed the World" at website https://ledaliang.github.io/journalclub/. Of course, you can choose some other methods which are not in the list but your favourite.

Deadline:
Please submit your completed essay by **December 10, 2025**.

Feel free to reach out if you have any questions or need further clarification on the project.
Good luck!

----------A DeepSeek Version----------

What if Without Gradient Boosted Trees

In the sprawling landscape of modern data science, few algorithms have achieved the near-ubiquity and practical impact of gradient boosted trees (GBTs). From predicting customer churn to powering financial risk models and winning data science competitions, frameworks like XGBoost, LightGBM, and CatBoost have become the default weapons of choice for structured data. It is a tool so deeply woven into the fabric of the field that its absence is almost unthinkable. Yet, by engaging in this thought experiment—what if gradient boosted trees had never been developed?—we can gain a profound appreciation for their role and envision a profoundly different, and perhaps more diverse, technological present.

The most immediate and visible impact would be felt in the competitive arena of data science. Platforms like Kaggle would be unrecognizable. Since the advent of XGBoost in 2016, its dominance on leaderboards has been so absolute that a common piece of advice for entrants is "just use XGBoost." Without it, the competitive landscape would be a far more eclectic and fragmented battlefield. Ensembles of diverse algorithms—from sophisticated support vector machines and elastic nets to meticulously tuned random forests—would vie for supremacy. The "wisdom of the crowd" in model stacking would be even more critical, fostering a culture of creative ensemble building rather than a frequent reliance on a single, dominant workhorse.

Innovation might have been channeled into perfecting neural networks for tabular data earlier or into reviving and refining other statistical learning methods.

Beyond competitions, the commercial and industrial world would operate with a tangible lag in predictive precision. Countless applications that silently rely on GBTs would be less efficient and less accurate. Recommendation systems would be slightly more off-mark, leading to lost revenue and poorer user experience. Fraud detection systems would have a higher false-negative rate, allowing more malicious transactions to slip through. Credit scoring models would be less nuanced, potentially leading to either riskier loans or the unjust exclusion of worthy borrowers. The entire "personalization" engine that drives modern digital commerce—from targeted advertising to dynamic pricing—would be a cruder instrument. This would not mean a complete halt to progress, but rather a cumulative "drag" on efficiency across global industries, amounting to billions in lost value and countless suboptimal automated decisions.

In this alternate reality, the vacuum left by GBTs would not be empty. It would likely be filled by two primary contenders. The first is the random forest algorithm. As a robust and parallelizable ensemble method, random forests would undoubtedly be the go-to for many tabular data problems. They are powerful and resilient, but they generally plateau at a lower predictive accuracy than a well-tuned GBT. The second contender would be deep learning. Without the overwhelming success of GBTs on structured data, the immense research and development momentum behind neural networks might have spilled over into conquering tabular data sooner. We might have seen the development of specialized architectures for tabular datasets a decade earlier, making deep learning a more common and mature solution for problems that are currently the exclusive domain of boosting.

This shift in the technological center of gravity would have had a cascading effect on the culture and skillset of data scientists. The field might be less homogenized. Instead of a common path that often leads to "import XGBoost," practitioners would need a deeper, more intimate understanding of the strengths and weaknesses of a wider array of models. The role would lean more towards that of a "machine learning mechanic," skilled in assembling and fine-tuning a diverse toolkit, rather than a "power user" of a single, supremely effective tool. This could have fostered a more profound theoretical understanding at the expense of the rapid prototyping and deployment that GBTs enable.

In conclusion, a world without gradient boosted trees would be a world of diminished precision, slower industrial automation, and a more fragmented technical landscape. While innovation would have undoubtedly found other paths, the unique combination of predictive power, computational efficiency, and handling of messy, real-world data offered by GBTs has provided a singular accelerant to the field of applied AI. Their absence would not have halted progress, but it would have altered its trajectory, making it more pluralistic and perhaps more challenging. Ultimately, this thought experiment serves as a powerful reminder that technological progress is not always a smooth, inevitable curve. It is often punctuated by breakthrough innovations that, like gradient boosted trees, so effectively solve core problems that they reshape the world in their image, making it difficult to remember what came before.