AI ML Challenges'W2023 - Paper Review

Name: Pushpak Vijay Katkhede **Date: 03/21/2023**

- ***Paper Title →***
  "LightGBM: A Highly Efficient Gradient Boosting Decision Tree"

- ***URL→***
  https://papers.nips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

- ***Citation →***

  G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Efficient and Accurate Gradient Boosting Machine for Large-Scale and High-Dimensional Data," in Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, Dec. 2017, pp. 6266-6276.

- ***Challenge Identified →***
  The authors observed below challenges when scaling Gradient Boosting Machine:
  1. The primary difficulty in applying GBMs to huge datasets is computational efficiency. Conventional GBM implementations can be computationally expensive, especially when dealing with huge datasets with many features. Long training timeframes and significant resource needs can make large-scale machine learning issues problematic, if not impossible.
  2. Memory efficiency is the second challenge in scaling GBMs to huge datasets. Conventional GBM implementations demand a lot of memory to hold the models and intermediate data structures, especially when dealing with huge datasets. This can lead to memory constraints that impede the training of larger models or the usage of more complicated features.
  3. Model correctness is the third hurdle in scaling GBMs to huge datasets. Because to overfitting or underfitting, traditional GBM implementations may be unable to achieve high accuracy on big and high-dimensional datasets. When a model becomes too complicated, it learns the noise in the data, whereas underfitting occurs when the model is too simple and fails to grasp the underlying patterns.

- ***Solution Proposed →***
  To address the challenges the authors proposed a technique called LightGBM. The feature binning is done by Light GBM. The LightGBM chooses the number of bins based on the distribution of the data, which results in more efficient use of memory, as opposed to the traditional method of using pre specified number of bins. This decreases the requirements for memory and computation. LightGBM uses a new approach called GOSS to address the problem of unbalanced datasets. Down sampling instances with lower gradients are included in the GOSS. This increases model accuracy by decreasing the impact of irrelevant data and focusing on the most useful occurrences. Data sampling techniques and parallelization are some of the features of LightGBM. For example, provides multi-threading to make use of the processing capabilities of several CPUs and data partitioning to train on subsets of data in parallel. Overall, LightGBM addresses the challenges of memory efficiency, model accuracy, and computational inefficiency by using techniques such as histogram-based feature binning, GOSS, leaf-wise growth, parallelization, and data sampling.