Sparsely Connected Layers for Financial Tabular Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

While neural networks are the norm for unstructured data such as images and texts, their performance often lags behind more traditional machine learning models such as gradient boosted trees for tabular data, evidenced by academic studies, industrial practices, and Kaggle competitions. In particular, there is no easy way to increase the number of layers for neural networks applied on tabular data, which has been a key reason in their success on unstructured data. Deeper fully connected networks also suffer from the vanishing gradient problem, and convolutional layers and transformers are not directly applicable to tabular data in general. Special constructs such as skip layers and attentions have been adapted to tabular data models with limited success. In this paper, we show that for the consumer financial tabular data, while standard 2-layer neural networks fall behind gradient boosted trees as usual, sparsely connected layers can increase the network depth and reliably outperform gradient boosted trees. The superior performance seems to come from the ability of the sparse layers to reduce correlations from the input data, which is a common problem in high-dimensional tabular data. Therefore we are hopeful the method could be applicable to other domains with similar challenges.

1 Introduction

The past decade has witnessed the rapid progress of neural networks on unstructured data such as images and text, much of which are due to special network constructs that enable very deep architectures without suffering from the vanishing gradient problem. In particular, convolutional neural networks have dominated almost all vision tasks [15], and more recently transformers have dominated the language models [23].

On the other hand, while tabular data is arguably much simpler and easier to process, there is no generic deep neural architecture available for it. Standard 2-layer neural nets often fall behind modern gradient boosted tree models (GBDT) such as XGBoost [5], LightGBM [13], and CatBoost [18]. For example, the 2022 Kaggle competitions have been dominated by GBDTs, winning 7 out of the 10 challenges [7]. Another survey paper using academic datasets concluded that decision tree models are still state of the art, winning on 7 out of the 9 datasets [4].

Motivated by the performance gap, there is a significant amount of research on deep neural networks for tabular data, which can be largely divided into three categories. First, we can combine neural networks and tree-based methods. For example, [16] proposes a network structure that imitates a binary decision tree, while [20] argues that an ensemble of the two approaches works better. Second, the wide and deep network architecture [6] can be applicable for some datasets. Taking a step further, DeepFM [10] introduces a learned Factorization Machine [19] into a deep network. Finally, there is recent interest in using the attention mechanism to learn higher order interaction between features. TabNet [2] uses sequential attention for implicit feature selection, and each layer places importance on only a few features. Another method performs both inter-sample and inter-feature attention [21].

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

While working with our industrial scale financial tabular data, we also discover that XGBoost significantly outperforms the standard 2-layer neural networks. Adding additional fully connected layers hurts the performance even more. Other more sophisticated network architectures have difficulty meeting the baseline performance in our problem setting. However, a simple sparsely connected layer that groups the input data in meaningful ways surprisingly beats XGBoost by a reliable margin. Additional analysis shows that our high dimensional financial data, not surprisingly, has highly correlated features. The sparsely connected layer seems to be especially effective in eliminating the correlation automatically. Thus, we believe that this method could be applicable in other domains with high-dimensional tabular data.

2 Our Method

2.1 Financial tabular data

Machine learning is widely applied in the consumer finance domain, for example, propensity models to acquire customers and risk models for underwriting. A person's financial history is without doubt very useful in predicting their future financial behaviors. Therefore there are many data providers that featurize and standardize personal financial data for model-building purposes [9]. Example attributes of these datasets include how many inquiries in the past six months, how many credit cards are open, the maximal balance of the past three months, etc.

Because these datasets are designed for general purposes, there are many thousands of attributes that cover many financial industries, a range of time periods, and a variety of behaviors. For example, loans are grouped into auto, personal, mortgage, student, retail, etc, and inquiry attributes include the number of inquiries in 14-day, 1-month, 3-month, 6-month, etc. Derogatory behaviors include delinquency, charge-off, collection, bankruptcy, etc.

2.2 The data correlation problem

These thousands of financial attributes are usually correlated. For example, the 3-month inquiry count is a subset of the 6-month count, and charge-off accounts are often preceded by delinquency.

It is well known that correlated features can dramatically impair a model's performance, more so in neural networks than GBDTs because correlations incorrectly amplify the gradients while decision trees can ignore redundant features altogether. This is likely a big reason why GBDTs often outperform neural networks, as data correlation is very common in high-dimensional tabular data.

However, reducing data correlation is nontrivial. First of all, with large scale high-dimensional data, the correlation among attributes almost always has a smooth distribution so there is no clear cutoff threshold for feature removal, meaning correlation will always exist no matter how many features we remove. Second, with sparse labels, any subtle difference between two correlated features can have non-negligible impact on model performance. Finally, as business objectives or macro conditions change, feature reduction analysis has to be repeated, which is tedious and error prone. Ideally we would like the model to select features and eliminate correlations automatically.

2.3 Our model architecture

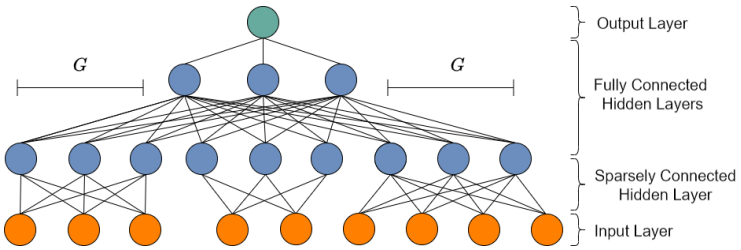

Figure 1: Our proposed model architecture.

Given information about correlated features, we can put those features beside each other in the neural network's input layer. The first layer of this model consists of small MLPs for each input feature

group whose output size is G, a hyperparameter we tune. We call this a sparse layer since only the group's neurons are connected in it. Figure 1 shows a visualization of our proposed model, with the special case that we only have 2 hidden layers. In practice, any number of fully connected hidden layers is possible, as this model becomes a regular MLP after the input layer.

3 Experiments

We evaluate all models on a proprietary dataset, and attempt to model consumer propensity. The dataset consists of tens of millions of samples and thousands of features, with a positive sample ratio in the single digits. For all neural networks, numerical features were scaled to 0 mean and 1 variance and categorical features were one-hot encoded. Missing values were mean imputed. Error bars were not reported due to the computational complexity of retraining models with datasets of this size. Experiments ran on a 30 GB RAM AWS ECS g2.2xlarge instance. Training time took an average of 2 hours per model.

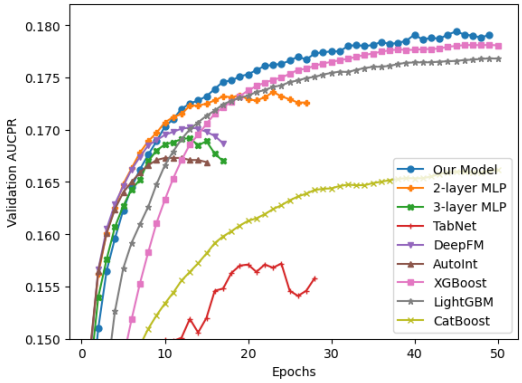

Figure 2: Various models' validation AUCPR on proprietary classification dataset.

3.1 Results

To evaluate our model's goodness, we compare our model to many existing models that are commonly used in tabular data applications, such as GBDT methods (XGBoost, LightGBM, and CatBoost) [5, 13, 18], and specialized neural networks (TabNet, DeepFM, and AutoInt) [2, 10, 22]. Open source implementations were used for all of these models, and can be found here [8, 24, 5, 17, 25].

We can immediately see from Figure 2 that our approach, beats all the other methods compared. It is worth noting that both XGBoost and LightGBM perform quite well when compared to the other models, which continues to demonstrate the strength of GBDT models on tabular data.

We conduct ablation study to investigate the effect of increasing the number of layers of the model. Figure 2 shows the performance of our model compared to a (fully connected) 2- and 3-layer MLP. In the spirit of fairness, our model's second hidden layer has the same size as the first and second hidden layers in the 2- and 3-layer MLPs. The first hidden layer of the 3-layer MLP has the same number of neurons as the sparse layer in our model. Simply increasing the number of hidden layers clearly lowers the model's performance, yet our specialized architecture, even though it has the same number of neurons as the 3-layer MLP, is the clear winner here too. This demonstrates that our model's superior performance doesn't only come from widening/deepening the neural network.

3.2 The effect of the sparse layer

Figure 3 shows how the sparsely connected layer in our model removes the correlation in the input data. The left figure shows one input group, where there are lots of highly correlated feature clusters. The right figure shows that after the sparse layer, there is no visible correlation. The middle figure shows that there is still some leftover correlation when using fully connected layers, demonstrating that regular neural networks cannot remove the correlation as well as our model can.

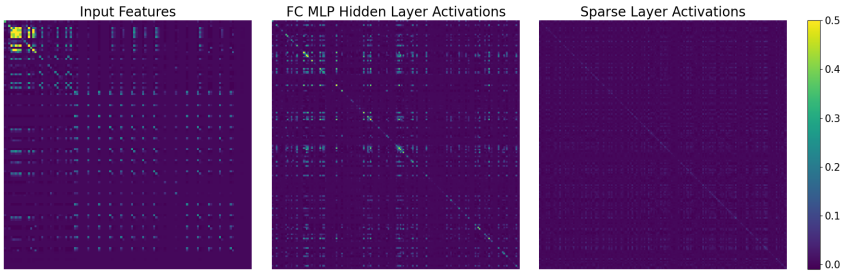

Figure 3: The covariance matrices of the input features, output of a 3-layer MLP first hidden layer, and the output of the sparse layer in our model. Note activation here refers to the value computed after applying the activation function.

This also shows that grouping highly correlated features together is important, which is easy in our case as the financial features are naturally grouped by financial industries, e.g., automobile, mortgage, personal loan, etc. We also tried to group features in other reasonable fashions, and got similar results. Grouping features randomly drops the performance significantly, from 0.1794 to 0.1730.

The sparsely connected layer acts in a similar fashion as the embedding layer, which is also why we choose a fixed number G of output for each group, regardless of the group input size. Similarly to embedding, our sparsely connected layer can significantly reduce network complexity and the number of model parameters.

4 Related Works

Tabular data presents unique challenges for deep learning, as traditional neural networks have struggled to outperform GBDT methods like XGBoost, LightGBM, and CatBoost [5, 13, 18]. These tree-based methods excel at handling tabular data due to their ability to capture complex feature interactions and manage missing values. Despite their success, tree models do not naturally incorporate domain knowledge, which can be a limitation in scenarios where feature relationships are known but not explicitly provided to the model.

Recent efforts have explored adapting neural networks to work more effectively with tabular data. TabNet [2] leverages attention mechanisms to dynamically focus on relevant features. DeepFM [10] and AutoInt [22] attempt to model feature interactions explicitly, with the former combining factorization machines with deep neural networks to capture both first- and second-order interactions, and the latter using self-attention mechanisms to model higher-order feature interactions without manual engineering. While Group-Connected Multilayer Perceptron Networks (GMLP) [12] also focuses on exploiting feature groupings, they do so by dynamically learning groups during the training phase using techniques like temperature annealing and entropy loss to encourage sparsity. These models, however, rely on the network to discover feature groupings and relationships during training, rather than encoding domain-specific knowledge directly into the model.

Additionally, Sparsely-Connected Neural Networks (SCNN) [1] reduce the number of connections in neural networks to save memory and energy consumption while maintaining competitive accuracy. SCNN uses stochastic binary masks to generate sparsity. Our model shares the goal of introducing sparsity but differs in that its sparsity is driven by the natural structure of the data — specifically, by grouping related features — rather than relying on random mask generation or pruning.

5 Conclusion

Sparsely connected layers represent a new approach to dealing with high-dimensional tabular data by explicitly incorporating domain-driven feature groupings into the neural network architecture. This positions it as a simple yet effective solution compared to more complicated neural network architectures for tabular data. Since large-scale high-dimensional tabular data often exhibits similar behaviours as our consumer financial data, we hope to see more use cases of our proposed solution.

References

[1] Arash Ardakani, Carlo Condo, and Warren J. Gross. Sparsely-connected neural networks: Towards efficient vlsi implementation of deep neural networks, 2017. URL https://arxiv.org/abs/1611.01427.

[2] Sercan O. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

[4] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, June 2024. ISSN 2162-2388. doi: 10.1109/tnnls.2022.3229161. URL http://dx.doi.org/10.1109/TNNLS.2022.3229161.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 11 of *KDD '16*, page 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

[6] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

[7] ML Contests. Tabular Data Competitions - Winning Strategies. https://mlcontests.com/tabular-data/, March 2023. Accessed 2 September, 2024.

[8] Dreamquark-AI. Tabnet, 2023. URL https://github.com/dreamquark-ai/tabnet.

[9] Experian. Premier Attributes. https://www.experian.com/business/products/premier-attributes, 2024. Accessed 2 September, 2024.

[10] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: A factorization-machine based neural network for ctr prediction, 2017. URL https://arxiv.org/abs/1703.04247.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. URL https://arxiv.org/abs/1502.03167.

[12] Mohammad Kachuee, Sajad Darabi, Shayan Fazeli, and Majid Sarrafzadeh. Group-connected multilayer perceptron networks, 2020. URL https://arxiv.org/abs/1912.09600.

[13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[14] Guolin Ke, Jia Zhang, Zhenhui Xu, Jiang Bian, and Tie-Yan Liu. TabNN: A universal neural network solution for tabular data, 2019. URL https://openreview.net/forum?id=r1eJssCqY7.

[15] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4): 541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

[16] Haoran Luo, Fan Cheng, Heng Yu, and Yuqi Yi. Sdtr: Soft decision tree regressor for tabular data. *IEEE Access*, 9:55999–56011, 2021.

[17] Microsoft. Lightgbm, 2024. URL https://github.com/microsoft/LightGBM.

[18] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features, 2019. URL https://arxiv.org/abs/1706.09516.

[19] Steffen Rendle. Factorization machines. In *2010 IEEE International conference on data mining*, pages 995–1000. IEEE, 2010.

[20] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[21] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[22] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19. ACM, November 2019. doi: 10.1145/3357384.3357925. URL http://dx.doi.org/10.1145/3357384.3357925.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[24] Shen Weichen, Zan Shuxun, Wang Ze, Zhang Wutong, Zhang Yuefeng, Huo Junyi, Zeng Kai, Chen K, Cheng Weiyu, and Tang. Deepctr-torch, 2022. URL https://github.com/shenweichen/DeepCTR-Torch.

[25] Yandex. Catboost, 2024. URL https://github.com/catboost/catboost.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction cover the main points that will be discussed in the data - trying to beat the performance of GBDT methods on financial tabular data using neural networks.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We mention the reason why error bars weren't included (we are comparing 9 models on a dataset of millions of rows and thousands of columns - it is an expensive task). We also make it clear that our results have only been demonstrated on financial data, and leave our method's performance on other domains as an open question.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: There are no theoretical proofs in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: Because our data is proprietary and models may be used for business use-cases, we cannot disclose our hyperparameters. We did however take steps to explain the content of the dataset and describe all the model architectures in depth.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Because our data is proprietary and models may be used for business use-cases, we cannot disclose our hyperparameters. We did however take steps to explain the content of the dataset and describe all the model architectures in depth.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification: Because our data is proprietary and models may be used for business use-cases, we cannot disclose our hyperparameters. We did however take steps to explain the content of the dataset and describe all the model architectures in depth.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We mention the reason why error bars weren't included (we are comparing 9 models on a dataset of millions of rows and thousands of columns - it is an expensive task). We also make it clear that our results have only been demonstrated on financial data, and leave our method's performance on other domains as an open question.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Information was provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We followed everything to the best of our knowledge.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: No societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All open-source code was cited appropriately.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Nothing has been released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.