A Comparative Study of Logistic Regression and Neural Networks Across Varying Sample Sizes on the HIGGS Dataset

Muhammad Azeem Bhatti(2504437)
Department of Creative Technologies
Air University
Email: 2504437@students.au.edu.pk

***Abstract*—Large-scale machine learning datasets require careful model selection to balance predictive performance and computational efficiency. Classical models such as Logistic Regression (LR) offer simplicity and speed but are limited to linear decision boundaries. Neural Networks (NNs) can capture non-linear interactions but impose substantially higher computational costs. In this research, we conduct a detailed empirical comparison of LR and a Multilayer Perceptron Neural Network (MLP) on the 11-million-sample HIGGS dataset. We evaluate performance across varying fractions of the training dataset (1%–100%) using accuracy, precision, recall, F1-score, ROC AUC, PR AUC, and training time. Our results show that LR saturates early, achieving an F1-score of 0.686 and ROC AUC of 0.684, while the NN reaches an F1-score of 0.785, ROC AUC of 0.852, and PR AUC of 0.865. However, NN training time grows to over 3200 seconds at full data, compared to 86 seconds for LR. We conclude with practical recommendations for model selection on large tabular datasets.**

I. Introduction

Modern machine learning applications increasingly rely on large-scale datasets with millions of samples and complex feature interactions. Choosing an appropriate model in such settings requires balancing predictive performance, interpretability, and computational cost. Logistic Regression (LR) remains a widely used baseline for classification tasks due to its simplicity and efficiency. However, LR is fundamentally limited by its linear decision boundary. Neural Networks (NNs), in contrast, offer powerful non-linear modeling capabilities but require significantly greater computational resources.

In this paper, we compare LR and NNs on the HIGGS dataset, a large tabular dataset with 11 million simulated particle collision events. This dataset presents a challenging testbed due to its size and non-linear feature relationships. We train both models on increasing fractions of the training data to evaluate how model performance and computational cost scale with dataset size.

Our contributions include:

- A systematic comparison of LR and NN performance across 1%–100% training sample sizes.
- A detailed analysis of ROC and PR curves for both models.
- Insights into computational scalability for large datasets.
- Clear recommendations on when to prefer LR or NN in real-world machine learning pipelines.

II. Related Work

Logistic Regression remains a standard baseline for classification tasks, especially in domains requiring interpretability such as healthcare and finance [3]. Neural Networks have demonstrated superior performance on complex datasets due to their non-linear modeling capacity [2]. The HIGGS dataset was introduced in [1], where deep neural networks were shown to outperform traditional machine learning models.

III. Dataset

The HIGGS dataset contains 11 million samples with 28 real-valued features. The target variable indicates whether an event corresponds to a Higgs boson signal or background process. The dataset is approximately balanced with 53% positives and 47% negatives.

The dataset consists of:

- Low-level kinematic features
- High-level engineered features

A detailed statistical summary was generated during preprocessing.

IV. Methodology

A. Preprocessing

No missing values were present in the dataset. Duplicate rows were detected but retained. Features were standardized using z-score normalization. An 80/20 stratified train-test split was applied.

B. Models

1) Logistic Regression: We used scikit-learn's implementation with:

```
solver="saga", max_iter=1000, n_jobs=-1
```

2) Neural Network: A feed-forward MLP with:

```
hidden_layer_sizes=(128,64),
activation="relu",
solver="adam",
max_iter=50
```

C. Training Fractions

Models were trained on:

$1\%, 2\%, 3\%, 4\%, 5\%, 10\%, 15\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90$

D. Metrics

We evaluated:

- Accuracy
- Precision
- Recall
- F1-score
- ROC AUC
- PR AUC
- Training time

V. Results

A. Logistic Regression

LR performance saturates after approximately 5% of data. Table I summarizes LR metrics at 100% training data.

TABLE I
LOGISTIC REGRESSION PERFORMANCE AT 100% DATA

Metric	Value
Accuracy	0.641
Precision	0.639
Recall	0.742
F1-score	0.687
ROC AUC	0.684
PR AUC	0.683
Training Time (s)	86.75

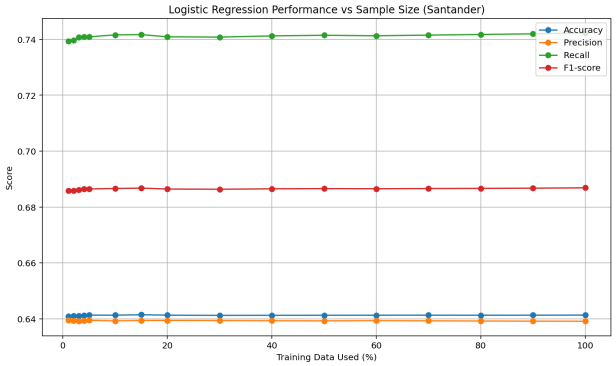

Fig. 1. LR Performance vs Training Size

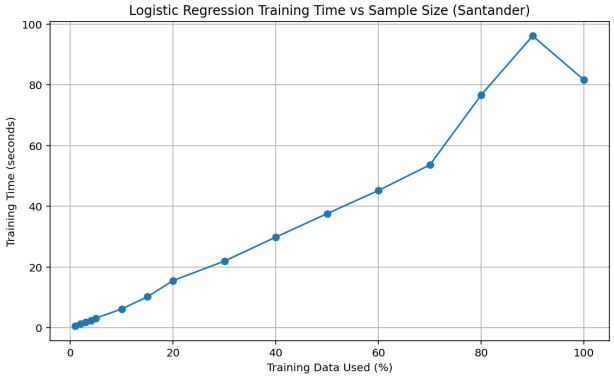

Fig. 2. LR Training Time vs Training Size

B. Neural Network

NN consistently outperformed LR across all metrics. Table II summarizes full-data performance.

TABLE II
NEURAL NETWORK PERFORMANCE AT 100% DATA

Metric	Value
Accuracy	0.768
Precision	0.772
Recall	0.797
F1-score	0.785
ROC AUC	0.852
PR AUC	0.865
Training Time (s)	3203.9

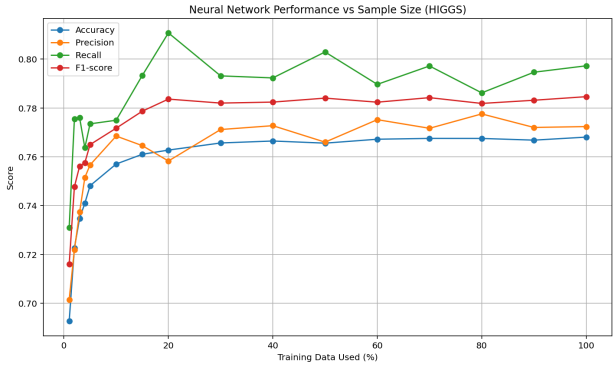

Fig. 3. NN Performance vs Training Size

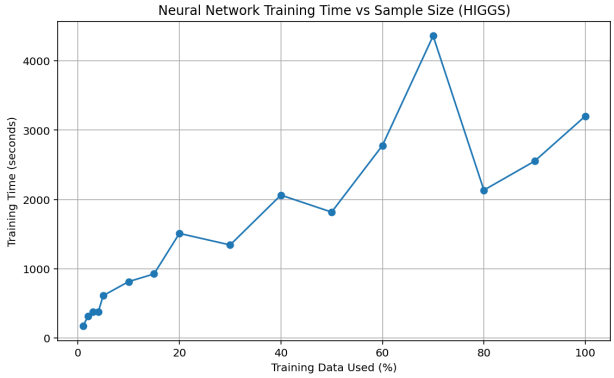

Fig. 4. NN Training Time vs Training Size

C. ROC & PR Curve Comparison

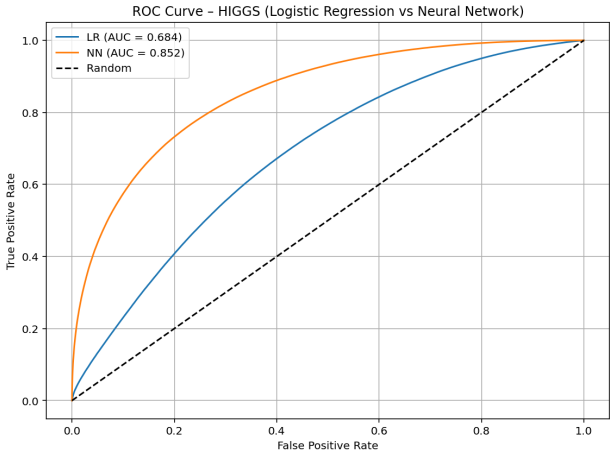

Fig. 5. ROC Curve: LR vs NN

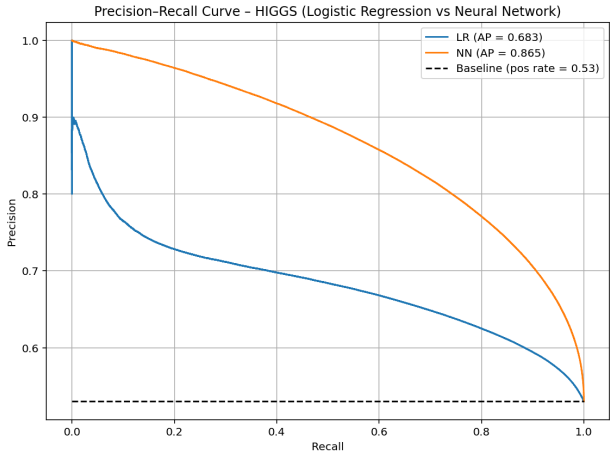

Fig. 6. PR Curve: LR vs NN

VI. Discussion

The experimental results highlight clear behavioural patterns:

- LR saturates extremely early and cannot exploit additional data.
- NN benefits significantly from large datasets and non-linear structure.
- NN requires substantially more training time (40–50x that of LR).
- PR and ROC curves demonstrate NN's superior discriminative ability.

These findings suggest that LR is preferred when:

- computational resources are limited,
- interpretability is needed,
- or the dataset is mostly linear.

NN is preferred when:

- the dataset is large,
- feature interactions are complex,
- and accuracy is the primary goal.

VII. Conclusion

We presented a detailed comparison of Logistic Regression and Neural Networks on the massive HIGGS dataset. The Neural Network decisively outperformed Logistic Regression across all performance metrics, though at the cost of significantly higher training time. For large-scale, non-linear datasets, Neural Networks should be the preferred choice. Logistic Regression remains a valuable baseline due to its speed and simplicity but is insufficient for complex tasks such as HIGGS classification.

References

[1] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature Communications*, 2014.
[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
[4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011.