



Exploratory Analysis on Effect of Poor Feature Selection on the Predictive Performance of Machine Learning Models

Ujjwal Pratap Singh

Student at Indian Institute of Technology, Kharagpur, West Bengal, India.

OPEN ACCESS

Article Citation:

Ujjwal Pratap Singh, "Exploratory Analysis on Effect of Poor Feature Selection on the Predictive Performance of Machine Learning Models", International Journal of Recent Trends in Multidisciplinary Research, July-August 2026, Vol 6(04), 165-167.



©2026 The Author(s). This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Published by 5th Dimension Research Publication

Abstract: Feature selection plays a crucial role in determining the predictive performance of machine learning models. While previous studies have primarily focused on developing effective feature selection techniques, the impact of poor feature selection has received comparatively less attention. This exploratory study investigates the effect of poor feature selection on the predictive performance of machine learning models. Experiments were conducted on multiple publicly available regression datasets using Linear Regression, Random Forest Regression, and Artificial Neural Networks. Model performance was evaluated using Root Mean Squared Error (RMSE), and a one-sample, one-tailed t-test was performed to assess the statistical significance of the results. The findings indicate that poor feature selection significantly degrades the predictive performance of all three models. These findings highlight the importance of feature selection in developing reliable and accurate machine learning models

Key Words: AI, Artificial Intelligence, Machine Learning, ML, Feature, Feature Selection, Prediction, Performance, Linear Regression, Random Forest, ANN.

1. Introduction and Literature Review

Machine learning (ML) is a field of artificial intelligence that enables computers to learn patterns from data and make predictions or decisions. Machine Learning models feed on data, and, in this case make predictions out of it. Model selection is one of the most crucial phases of the machine learning process, as a bad model choice can result in unnecessary training, noisy findings, misleading conclusions, or poor performance [1].

However, an ML application's performance may not necessarily depend on model. The decision of what constitutes a good model is often made by the developer and relies on a variety of factors [2]. Even complex models fail. The performance of models depends on many factors, one of which are features [3].

Datasets originating from different disciplines, such as biology, retail, finance, geography, economics, and social sciences may contain thousands of features for every record. Some of these features have little or no contribution (that is – redundant and irrelevant features) to machine learning models. For this reason, eliminating them from the learning process may help in increasing learning speed, reduce the risk of overfitting, decrease computational requirements, create simple models with high generalization performance, and facilitate the interpretability of results. Selecting relevant features is an indispensable task, and feature selection approaches are used to detect such features [4].

The relevancy of features can be classified as weakly relevant feature or strongly relevant feature [5]. A weakly relevant feature is that feature which have little to no effect on the model and vice versa for strongly relevant feature. Yu and Liu [6] pointed out that "an optimal feature subset should include all strongly relevant features, none of the irrelevant features, and a subset of weakly relevant features".

Existing studies primarily focus on feature selection methods rather than investigating the consequences of deliberately poor feature selection.

This study deals with exploring and analysing the effect of poor feature selection in different types of machine learning models.

Research Question: Is there any effect of poor feature selection on the predictive performance of ML models?

2. Research Methodology

The study was conducted using three machine learning models: Linear Regression, Random Forest Regression ($n_{\text{estimators}} = 100$), and Artificial Neural Networks (ANNs) (4 layers of network, learning rate = 0.01, iterations of training loop = 10000, ReLU activation function, Adam optimizer). Seven publicly available regression datasets were obtained from Kaggle to evaluate the impact of feature selection on model performance.

For each dataset, the data was divided into training and testing sets using the train-test split method (80:20). In addition, K-fold cross-validation (5-fold) was employed to obtain a more reliable estimate of model performance (5-fold cross validation was done on the training data). The Root Mean Squared Error (RMSE) was used as the evaluation metric for comparing the predictive performance of the models.

To investigate the effect of feature selection, two experimental scenarios were considered:

1. **Two Random Features (Poor Feature Selection):** All possible combinations of two features were generated from each dataset. Each combination was used to train and evaluate the models, and the corresponding RMSE was recorded. The average RMSE across all two-feature combinations was then computed.
2. **Four Selected Features:** Four features with the highest magnitude of correlation coefficients with the target variable were selected. These features were used to train each model, and the resulting RMSE was calculated.

Finally, a one-sample, one-tailed t-test ($\alpha=0.05$) was performed separately for each machine learning model to determine whether the mean percentage increase in RMSE was significantly greater than zero.

3. Discussion

	2 Random Features			4 Selected Features		
	Lin. Reg	Random Forest	ANN	Lin. Reg	Random Forest	ANN
1	5930	4194	6614	4058.50	1985	3780
2	13.4	14.7	13.6	8.70	6.73	8
3	0.6	0.6	0.8	0.53	0.58	0.53
4	9.2	10.5	9.7	6.92	7.34	7
5	0.43	0.48	0.43	0.37	0.38	0.36
6	1.5	1.6	1.5	0.95	0.78	0.72
7	0.97	0.53	5.64	0.50	0.24	5.61

Note: RMSE in 1st dataset is high because it deals with the prediction of car prices. Car prices are in lakhs and crores, so an RMSE of around 4000 is acceptable. Also, the aim of the study is not to make precision predictive models.

Across all seven datasets, the RMSE obtained using two randomly selected features was consistently higher than that obtained using four selected features. This indicates that selecting informative features plays an important role in improving predictive accuracy. Highest percentage increase of RMSE in Linear Regression was noted to be 94%, in Random Forest Regression it was 120.8% and highest percentage increase in ANN was 108.3%.

No decrease in RMSE was observed when using two randomly selected features compared with four selected features across the datasets considered. This observation supports the finding that poor feature selection negatively affects predictive performance.

4. Basic Statistical Analysis [One sample t-test]

S.no.	Linear Regression	Random Forest	ANN
	% age Increase (LR)	% age Increase (RF)	% age increase (ANN)
1	46.1	111	75
2	54	118	70
3	13.2	3.5	51
4	32.9	43.8	38.6
5	16.21	26.3	19.45
6	57.9	105	108.3
7	94	120.8	0.53

Ho (Null Hypothesis): $\mu_0 = 0\%$ (poor feature selection does not affect predictive performance of ML models)

H1 (Alternative Hypothesis): $\mu_0 > 0\%$ (poor feature selection affects predictive performance of ML models)

Exploratory Analysis on Effect of Poor Feature Selection on the Predictive Performance of Machine Learning Models

For Linear Regression:

$N=7$; $\mu = 44.9\%$; $\text{std} = 27.8\%$

Standard Error (SE) = $\text{std}/\sqrt{N} = 10.51$

t-value = $(\mu - \mu_0)/SE = 4.27$

degrees of freedom (df) = $N-1 = 6$

for one tailed t-test, $\alpha=0.05$, $df=6$, critical $t = 1.94$, $4.27 > 1.94$

Decision: Reject Null Hypothesis

The same procedure was repeated for the Random Forest Regression and Artificial Neural Network models. The resulting statistics are summarized in the table below.

Model	Mean %age Increase	STD	t-value	p-value	Decision
LR	44.9	27.8	4.27	0.0025	Reject Ho
RF	75.9	49.34	4.05	0.0035	Reject Ho
ANN	51.84	36.31	3.78	0.0047	Reject Ho

5. Result and Conclusion

As t-value exceeds the critical t-value in all the three models (Linear Regression, Random Forest and Artificial Neural Network), also, for all models, the corresponding p-value was <0.05 . Therefore, the null hypothesis was rejected for all three machine learning models. These findings suggest that poor feature selection negatively affects the predictive performance of the machine learning models evaluated in this study.

6. Limitations

This study was conducted using only seven regression datasets, which limits the generalizability of the findings. There was one outlier, each in, Random Forest and ANN table, which may have influenced the statistical result. Future studies should evaluate a larger number of datasets and additional machine learning algorithms.

Funding: This is an independent research and this research has received no external funding

References

1. Jie Ding, Vahid Tarokh, and Yuhong Yang. 2018. Model selection techniques: An overview. IEEE Signal Processing Magazine 35, 6 (2018), 16–34.
2. Yair Weiss and William T Freeman. 2007. What makes a good model of natural images? In 2007 IEEE conference on computer vision and pattern recognition. IEEE, 1–8
3. J. Miao and L. Niu, “A Survey on Feature Selection,” Procedia Comput. Sci., vol. 91, no. Itqm, pp. 919–926, 2016, doi: 10.1016/j.procs.2016.07.111
4. M. Büyükkeçeci and M. C. Okur, “A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning”, Gazi University Journal of Science, vol. 36, no. 4, pp. 1506–1520, Dec. 2023, doi: 10.35378/gujs.993763.
5. Kohavi, R., John, G.H., “Wrappers for feature subset selection”, Artificial Intelligence, 97(1-2):273-324, (1997).
6. Yu, L., Liu, H., “Efficient Feature Selection via Analysis of Relevance and Redundancy”, Journal of Machine Learning Research, 5: 1205-1224, (2004).