Impact of Noise on Machine Learning Algorithms Performance

How noise in datasets affect the performance of ML algorithms

Neha Tiwari  
 MSc Computer Science - Data Science  
 Trinity College Dublin  
 Dublin, Ireland  
 tiwarin@tcd.ie

Krishna Hariramani  
 MSc Computer Science - Data Science  
 Trinity College Dublin  
 Dublin, Ireland  
 hariramk@tcd.ie

Ankur Rangwala  
 MSc Computer Science - Data Science  
 Trinity College Dublin  
 Dublin, Ireland  
 rangwala@tcd.ie

INTRODUCTION

We never get perfect data in real-world, hence our data models and decisions based on corrupted (noisy) data tends to be inaccurate. Presence of noise can make a negative impact on system performance by increasing the complexity and training time and decreasing the efficiency of the resulting model. Majority of prevailing machine learning algorithms have already incorporated different tactics to handle noise, despite that, noise still can affect algorithms performance adversely. Hence, knowing the exact impact of noise in our dataset is an essential issue and should be evaluated to make an effective model to take a better decision.

RELATED WORK

Machine Learning, noise impacts, Gaussian noise, Machine Learning Algorithms, Naive Bayes Classifier, Logistic regression, Random Forest Classifier.

METHODOLOGY

Figure 1 shows the project workflow diagram at macro level.

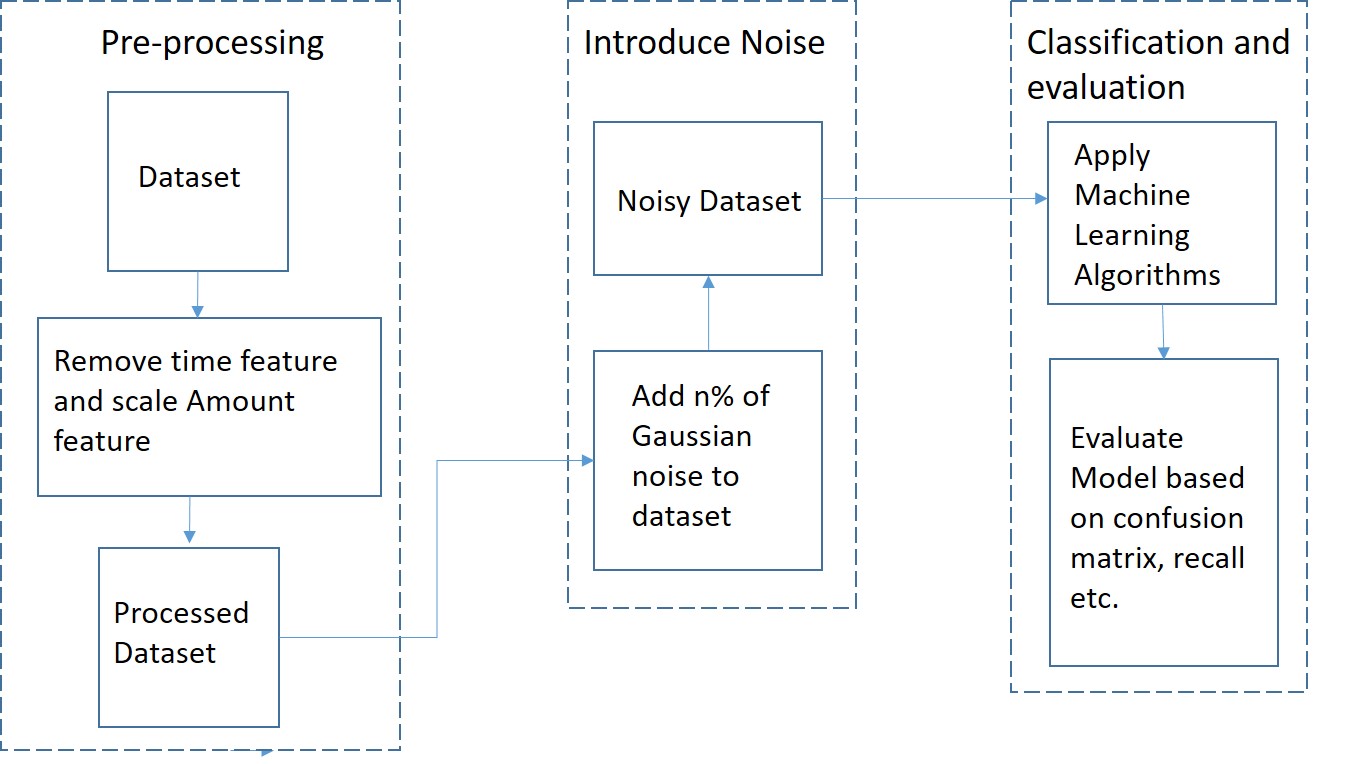


Figure 1: Project Workflow Diagram

We have divided our methodology into three sub sections mentioned below:

1. Data Analysis and Pre-Processing

For our project, we used credit card fraud dataset from Kaggle []. In the dataset there are 284807 records of transactions (made in September 2013 over two days) and 28 normalized numerical features which were obtained by using PCA on a set of original features, due to confidentiality issues the names of features are not revealed by author, other than these 28 features, two additional features ‘Time’ and ‘Amount’ are also provided for additional information. We drop the Time feature as it follows an increasing trend (from 0 to 172792) and does not give any information for labels in the format it is given, we normalize and scale the Amount feature as it can give important information like the amount of fraud and the value which can be saved by employing the fraud detection models. In the end, we get 29 scaled and normalized numerical features and one label column ‘Class’ containing 0's and 1's (1 denoting fraud transaction).

2. Introduce Noise to dataset

After pre-processing the dataset, we add noise to it and evaluate performance of three popular classification algorithms (Logistic Regression, Naïve Bayes Classifier and Random Bayes Classifier) at different noise levels.

For adding n% of noise to a given dataset, we use the algorithm shown in Figure 2[]. Figure 3 shows an example of how we have added noise to the dataset.

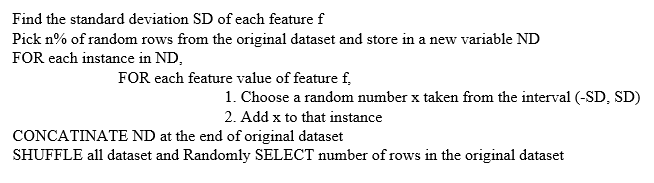


Figure 2: Algorithm for adding noise to a dataset

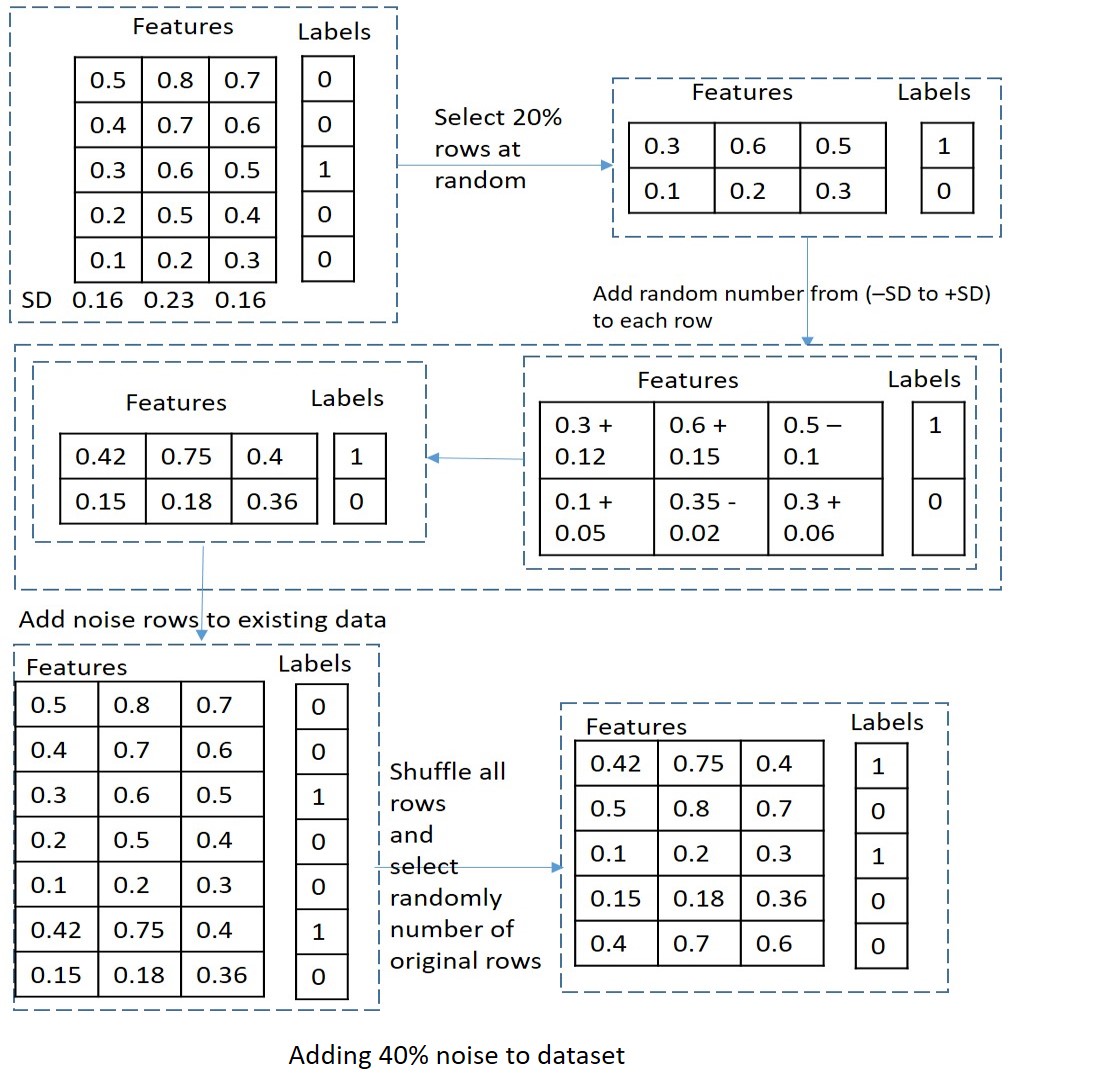


Figure 3: Example of addition of 40% noise to a dataset

3. Classification and Evaluation

After adding noise to the dataset, the dataset was split into train set and test set with a 70:30 split. We use Logistic Regression, Naïve Bayes Classifier and Random Forest Classifier with default hyper parameters for classification of frauds as we are monitoring the effect of noise on different ML algorithms. We will discuss hyper-parameter tuning more in future work section.

For evaluation of our models we have used a probability threshold of 0.5 for the classification of a transaction as a fraud. If the probability of a transaction being fraud is given to be more than 50% by an algorithm, then the transaction is classified as a fraud else regular. Also, we are monitoring the effect of noise on different ML algorithms so as far as we keep the probability threshold same for every algorithm it should not affect the impact of noise on the performance of an algorithm.

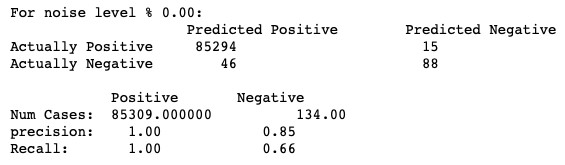
Since our dataset is highly unbalanced, the negative class (frauds) account for only 0.172% of all transactions evaluation metrics like accuracy and AUC under ROC should not be considered for evaluating the models. Since we are interested in correctly predicting the negative class (frauds) correctly, we have used confusion matrix and recall for negative class (TN/TN+ FN) as a metric to evaluate our model’s performance. We are also showing specificity (TN/TN+FP) at different noise levels, why we do that is discussed in more detail in results section.

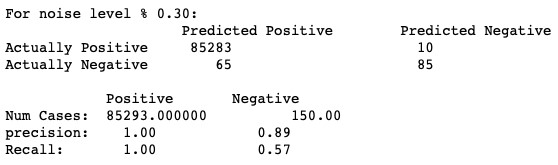
RESULTS & DISCUSSIONS

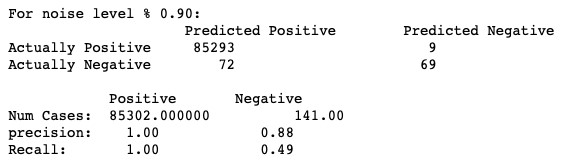
While performing tests with different ML algorithms, below are the results:

1. Logistic Regression

Figure 4 shows the confusion matrix at different noise levels, where we can observe that, as noise increases the ratio of True-Negative to False-Negative decreases.







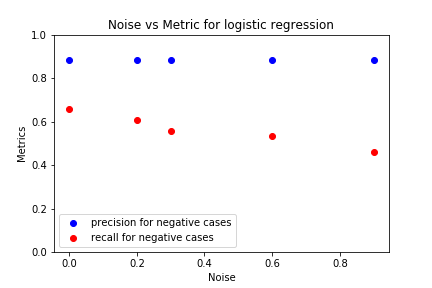
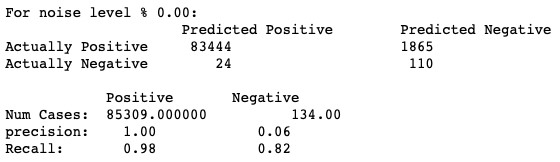
Figure 4: Confusion Matrix & Performance Evaluation at noise levels - 0%, 30%, 90%

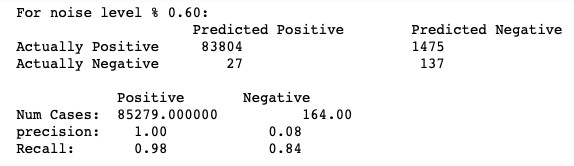
Figure 5: Noise Vs Metrics

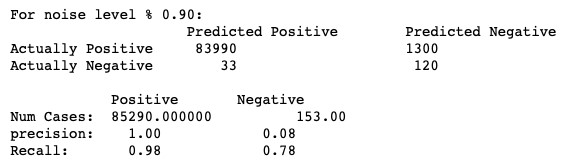
From the confusion matrix and the graph, as the ‘Noise’ level increases the recall for negative class (TN/TN+FN) value follows a decreasing trend generally as expected. Also, we can see that precision for negative class or specificity (TN/TN+FP) remains same due to class imbalance. The impact of noise data over logistic regression is that as we increase noise from 0% to 90%, the recall falls from 66% to 49%.

2. Naive Bayes classifier

Figure 6 shows the confusion matrix at different noise levels, where we can observe that, as noise increases the ratio of True-Negative to False-Negative do not follow any trend in general which means that classification for fraud is more robust to noise but this comes in contrast to precision since this model classifies many regular transactions as fraud i.e. number of False-positive are increasing. Same can be concluded from the graph.







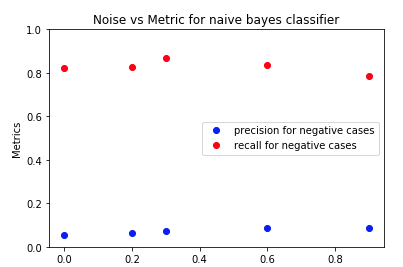
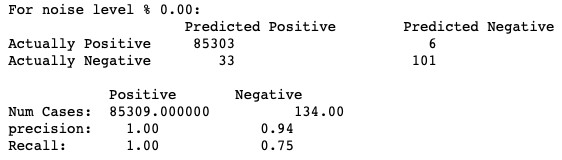
Figure 6: Confusion Matrix & Performance Evaluation at noise levels - 0%, 60%, 90%

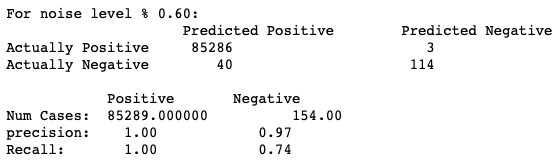
Figure 7: Noise Vs Metrics

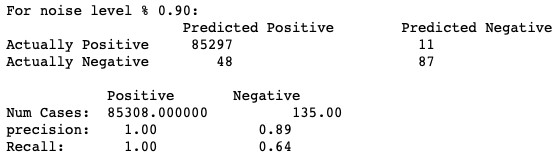
From the above metrices and graph, we can see the ‘Recall’ value is getting decreased as the ‘Noise’ level increases.

3. Random forest classifier

Figure 8 shows the confusion matrix at different noise levels, where we can observe that, as noise level increases there is no proper trend of True-Negative to False-Negative, i.e. number of False-positive are increasing. Same can be concluded from the graph.







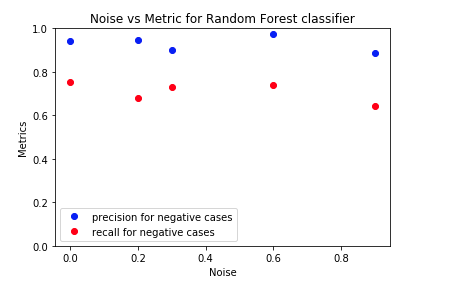
Figure 8: Confusion Matrix & Performance Evaluation at noise levels - 0%, 60%, 90%

Figure 9: Noise Vs Metrics

The above metrices and graph, shows that ‘Recall’ value is getting decreased as the ‘Noise’ level increases.

From the above experiments, we concluded that, ‘Naive Bayes classifier’ gives us considerably good result, but at the cost of precision.

LIMITATIONS & OUTLOOKS

Due to time constraints, currently, we applied only three popular machine learning algorithms (Naive Bayes classifier, Logistic Regression and Random Forest classifier) to evaluate the impact of noise on the above-mentioned algorithms. Also, we calculated the noise impact based on default hyper-parameters for each algorithm. However, we would attempt to investigate the impact after performing hyper-parameter tuning on the existing algorithms, in addition to experimenting with other algorithms like decision tree and k-nearest neighbors (KNN) as part of future work.

REFERENCES

[1] FNM Surname (2018). Article Title. Journal Title, 10(3), 1–10.

[2] F.N.M Surname, Article Title, https://www.acm.org/publications/proceedings-template.

[3] F.N.M Surname and F.N.M Surname, 2018 Article Title, The title of book two (2nd. ed.). Publisher Name, City, State, Country.

[4] Ian Editor (Ed.). 2018. *The title of book two* (2nd. ed.). University of XXX Press, City, Chapter 100. DOI: <http://dx.doi.org/10.1000/0-000-00000-0>.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00