Applying Logistic Regression Analysis in Modeling Settlement Analysis with Ground Improvement

Seok Hyeon Chai(✉), Tara Stojimirovic, and Thamer Yacoub

Rocscience, Toronto, ON, Canada
steve.chai@rocscience.com

Abstract. Settlement analysis plays an essential role in providing a safe measure for the stability of the foundation and other structural components. The settlement analysis with numerical methods such as Finite Element method (FEM), Finite Difference Method, or other sophisticated constitutive models are continuing to grow with the advancement of technology around the world. Machine Learning (ML) has been gaining a high interest in recent years in many different fields of engineering for its ability to predict an outcome given input parameters with certain patterns. Although the theories in ML existed before, there has been a significant growth in recent years due to overflowing information that is available online. There are numerous papers that introduce ML application in attempt to connect these two complex models together to predict a soil property and the behavior by comparing the results with the existing numerical analysis methods or field results. Also, ML algorithm has been used in research for predicting fragility curve analysis with given set of training data to determine the risk factor of structural buildings subjected to earthquake loads. There are many areas of focus on achieving accuracy with some complex models. Also, there are many resources available which discuss the modelling aspect of the analysis that can incorporate the ML algorithms in simple models. The paper introduces the modelling aspect of analyzing settlement with selective hyper parameters to achieve the target settlement. The paper introduces a logistic regression modelling using ground improvement result as a target variable and settlement and soil properties as the parameters to train. The study provides a model using ML algorithms to predict whether the soil model has ground improvement or not, with the selected hyper-parameters using settlement analysis software, Settle3. The paper provides the prediction accuracy with different methods of ML algorithms including logistic regression, K-nearest neighbor (KNN), stochastic gradient decent, Support Vector Machines (SVM) etc. The accuracy shows very high confidence with given parameters and more details on the future studies of the paper is discussed.

Keywords: Settlement analysis · Ground improvement · Machine Learning · Logistic regression · KNN · SVM · Sequence modelling

R. E. Hammah et al. (Eds.): RIC 2023, AHE 19, pp. 287–297, 2023.
https://doi.org/10.2991/978-94-6463-258-3_30

1 Introduction

Machine learning algorithms are being applied in various places across engineering disciplines and daily lives. The application ranges from scientific advancements in research to day-to-day activities such as waking up routine to driving to work. Due to the recent advancement of ML in recent years, the focus on expanding the application of the data science field to other disciplines has been sharper than before. More attention to the use of ML for civil engineering applications has been no exception. The application of Neuron Network in estimating spatial variability of soil profile using CPT boreholes, to even estimating stress strain behavior of steel plates using non-conventional Hooke's laws equation [1] have fascinating findings which utilizes ML concepts to civil engineering applications.

One notable research involving application of Machine Learning algorithm in geotechnical engineering is related to determining the characteristics of soil based on CPT (Cone penetration testing) borehole logs. CPT points are in-situ tests that can be used to identify the soil type. Based on characteristics of soil type, the CPT field test can be used for stratigraphic profiling [4]. The challenge remains as to the realistic soil profile representation of the location of interest as there are limited number of CPT boreholes that can be collected in the field. Many researchers have been investigating different ways to predict the soil classification with spatial variability of soil profile under the ground. Ching and Phoon (2017) implemented sparse Bayesian Learning algorithm to analyze site-specific data of CPT data for predicting site characterization. Then, Ching et al. [2] developed an improved method to address non lattice data in three-dimensional probabilistic site characterization by determining the missing data points of CPT data based on the surrounding CPT data. Below shows a diagram of the lattice and non-lattice data sets with respect to the CPT holes.

Based on this algorithm, the full 3D surface can then be characterized based on the surrounding known nodes of CPT data points. Figure 1 shows the 3D graphical representation of Ic values which are soil characteristic values that can be used to categorize soil types (Roberton 1990, 2010). Similarly, Fig. 2 shows the Bayesian Learning algorithm in showing the 2D data points with location of the CPT boreholes and its corresponding Ic values (entry number Z). On the left are the original data points and the right shows filled data points.

Then, Ic values that is used for characterization of soil types in the SBT (Soil Behavior Type) chart.

The example above was provided to demonstrate different applications of ML algorithms that can range from determining a complex soil profile creation. However, ML algorithm can also be used to recognize simple patterns of analysis to obtain the results. Logistic regression is a simplistic ML algorithm that classifies a binary output (1 or 0) based on multivariable parameters. There are many optimization algorithms for minimizing the cost function between the predicted equation based on training sets versus the test set. The motivation for this paper is derived from the variety of areas where ML theories are used across different disciplines to predict the pattern of next behavior given the inputs from the user. This is based on the previous inputs provided by the users which then the ML algorithm will apply the appropriate steps to train the dataset to predict the behavior with user-selected target parameters. There are many different methods in

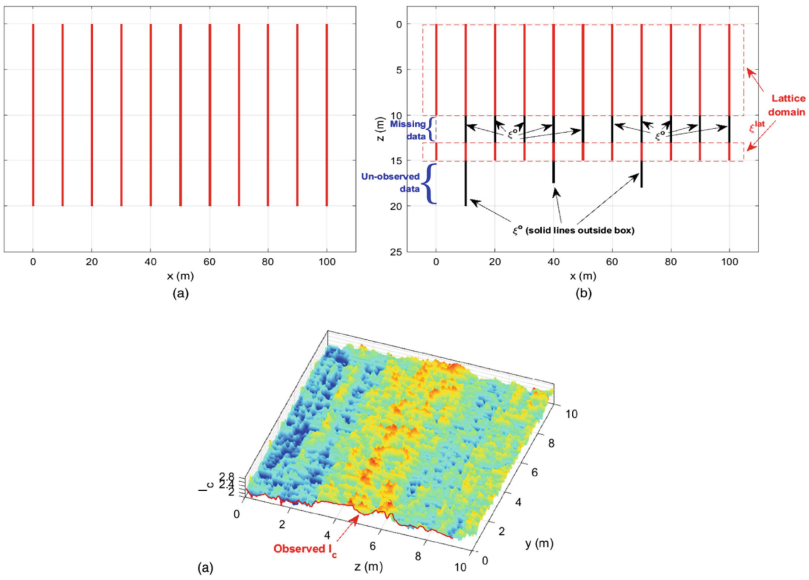

Fig. 1. Diagram showing the lattice (left) to non-lattice (right) data, then 3D view of the site characterization [2]

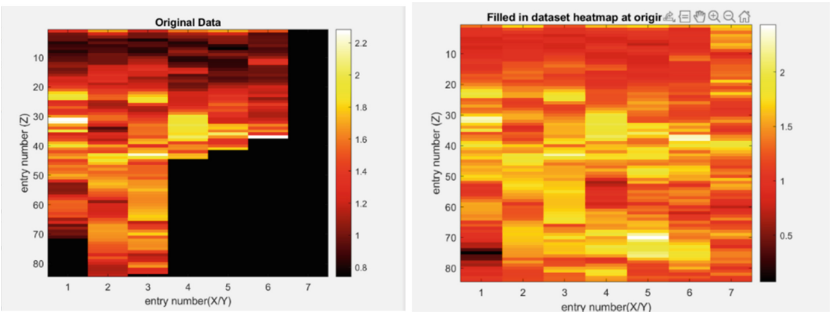

Fig. 2. Original data and filled data using the sparse Bayesian Learning algorithm

machine learning, specifically related to logistic regression. Gradient descent, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), etc. are some of the most popular topics. They all have similar overall structure, yet each provides a unique attribute in achieving the results.

The paper investigates each of the ML methods, then goes to explain the model description along with hyper parameter selection, to data collection which is provided in the model description of the paper. In the results discussion, different case studies provide which ML logistic regression method is the most viable option for predicting the application of ground improvement with respect to the parameters from settlement analysis results.

The paper provides a framework and the feasibility for applying ML algorithm to predict a ground improvement feature in the model with settlement analysis results. It provides a preliminary design and simple framework for applying a sequence model of settlement analysis which can expand to many other modelling applications.

2 Model Description

Settle3 models have been created with loads and ground improvement. In order to keep the models simple to track the performance of the ML algorithm, only the following parameters were collected for training and testing the ML algorithm (Fig. 3).

The collected parameters are the following:

Es – elastic young's modulus of soil.
Load – magnitude of the load defined.
Grd_improv = ground improvement (0 for not-applied, 1 for applied in the mode).
Elastic_mod = elastic modulus of ground improvement.
Bot_elev = bottom elevation of ground improvement.
Diameter = diameter of the ground improvement.
Settlement = total settlement of the model.
Piezo = water table (0 for no-water table, 1 for water-table in the model).
Pwp = pore water pressure.
Immediate = immediate settlement.

Batch compute in Settle3 allows users to compute multiple files with loads and ground improvement sections. Then, it will read the output the results for the settlement analysis with the selected parameters (Fig. 4).

Above is an example of batch compute in developer menu of Settle3. 'Get ML Info' will only collect the file information with respect to the parameters chosen from above.

Below are the results that is obtained once the files are computed and the parameters are retrieved from the model (Table 1):

Some parameter model assumptions made are provided below:

- Loads are applied in the same stage as stone column stage.
- Single query point results are obtained for each file. Meaning this first query point determines the total settlement of the file.

```
# train data_trainset:
data_train = pd.read_csv('ML_largedata_train.csv', header=0)
data_test = pd.read_csv('ML_largedata_test.csv', header=0)
# data_train = data_train.dropna()
print(data_train.shape)
print((data_train.columns))
```

```
(288, 10)
Index(['Es', 'load', 'grd_improv', 'elastic_mod', 'bot_elev', 'diameter',
       'settlement', 'piezo', 'pwp', 'immediate'],
      dtype='object')
```

Fig. 3. ML dataset of parameters selected.

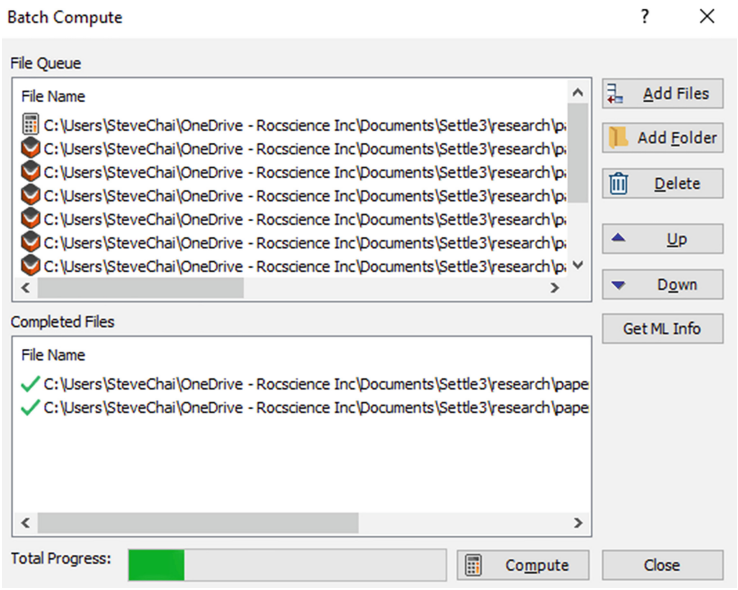

Fig. 4. An example of how the data was collected using batch compute option in Settle3.

Table 1. Table showing some examples of the train dataset

Es	load	grd_improv	elastic_mod	bot_elev	diameter	settlement
500	0	0	0	0	0	0.100327
500	0	0	0	0	0	0.100327
500	0	1	80000	−52	0.6	0.0578818
500	0	1	80000	−25	0.6	0.0378451
500	0	0	0	0	0	0.100327
500	0	0	0	0	0	0.100327

– Flexible foundation load type is considered, so the strain contour profile underneath the load is non-uniform.
– For simplicity of collecting the dataset, no embankment loads are considered. Only the loads are used.

In total, 200 models have been used to train the ML algorithm, while 102 models are used for testing the trained set.

Some of the models used for the data collection are shown in Fig. 5.

These are some of the examples of the models used for data collection. The models range from simplistic load and stone column applications to more complex geometry with combination of the loads and stone columns. All the models have a load applied at a specific region followed by either stone columns applied or not applied. Note that the

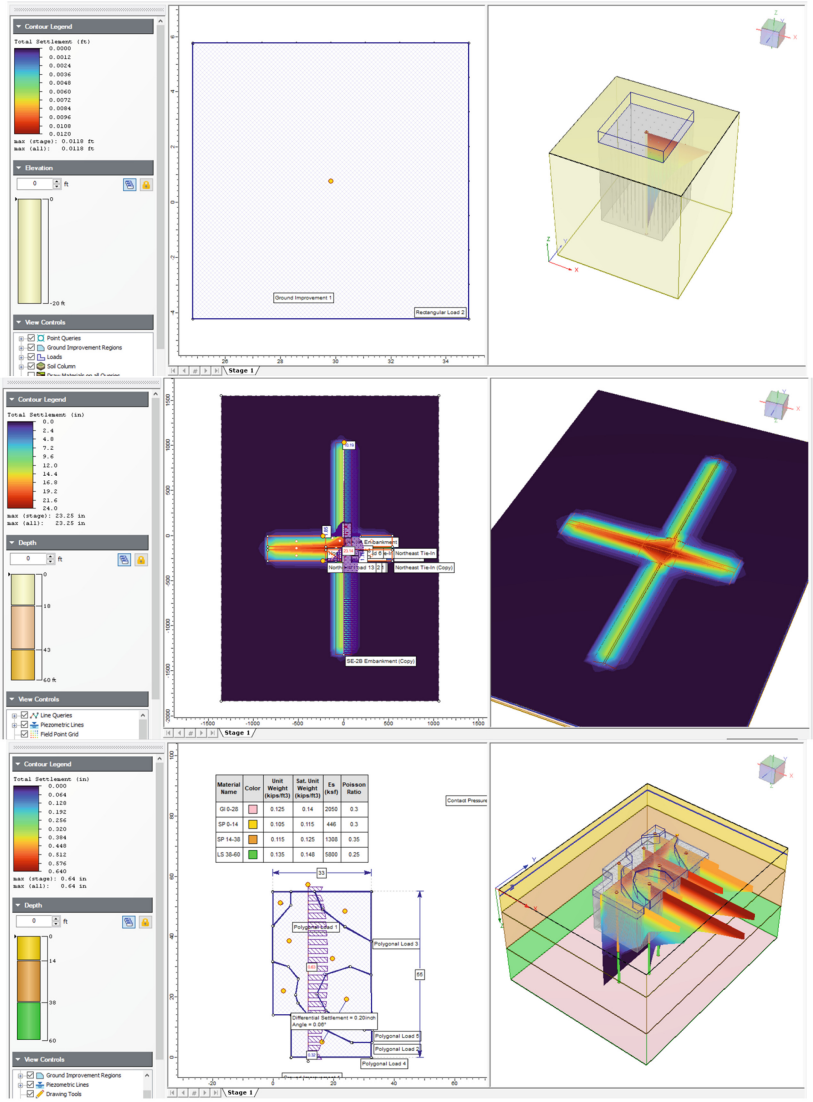

Fig. 5. Ground improvement models in Settle3 used for training dataset.

parameters collected using the models are based on the query point at a specific location of the model. This means that only the results computed from the loading stress at query point with respect to the soil properties underneath the loads are collected. Only single query point immediate displacements are taken from each model.

3 Results and Discussions

Jupyter notebook is used for running the ML algorithms. Full python code can be provided upon request. The study was first carried out without these parameters: Elastic modulus of stone columns, bottom elevation, and diameter of the ground improvement, and no piezometric lines and immediate settlement.

As mentioned previously, 200 models have been used for training the hyperparameters in the mode, while 102 models were used for testing. Below shows the test set up for logistic regression function in python (Fig. 6).

By introducing the f1-score and precision and recall values as shown below, we see that the accuracy of 88% has been achieved using logistic regression.

Figure 7 shows the ROC curve with red line as indication of the ratio of false positive rate to true positive rate. Normally, the model is described as performing well when the ML model is further away from the red line towards the true positive rate. The graph shows relatively good results, but better results can be achieved. Other ML algorithms have been used in this paper to test the performance of the predictive model with the given dataset. Below is the summary of the results with other models and their score (0 to 100%) (Fig. 8).

Note that there are some methods with 100% accuracy, such as random forest model or decision tree model. This does not mean that these algorithms are the perfect solution. Rather, there may be overfitting of parameters. The more complex the models become with more parameters introduced, the more ML algorithm will end up overfitting the curves with given parameters, which ends up perfectly predicting the test set with many parameters introduced. This was the observed behavior with other algorithms as well when the study was initially carried out with all the parameters mentioned in the model description.

Taking a look at KNN, better precision to recall results are shown in Fig. 9 and Fig. 10.

Therefore, the paper provides a relatively simple approach in developing predictive model whether the settlement analysis model has ground improvement or not based on the soil properties parameters and its resulting settlement. Note that the study had limited

```
# newX = []
X_train = []
X_test = []
y_train = [] # 2nd col of data_trainset (survived)
y_test = [] # 2nd col of data_trainset (survived)
data_train = data_train.dropna()

X = data_train.loc[:, data_train.columns != 'grd_improv']
y = data_train.loc[:, data_train.columns == 'grd_improv']

print(X.shape)
print(y.shape)

X_train ,X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3,random_st
logmodel = LogisticRegression()
logmodel.fit(X_train,y_train)
```

```
(199, 3)
(199, 1)
```

Fig. 6. Feeding the training data set in logistic regression using Python script.

```
predictions = logmodel.predict(X_test)
acc_log = round(logmodel.score(X_train,y_train)*100,2)
print(classification_report(y_test,predictions))
confusion_matrix(y_test,predictions)
```

```
              precision    recall  f1-score   support

         0.0       0.90      0.98      0.94        54
         1.0       0.00      0.00      0.00         6

    accuracy                           0.88        60
   macro avg       0.45      0.49      0.47        60
weighted avg       0.81      0.88      0.84        60
```

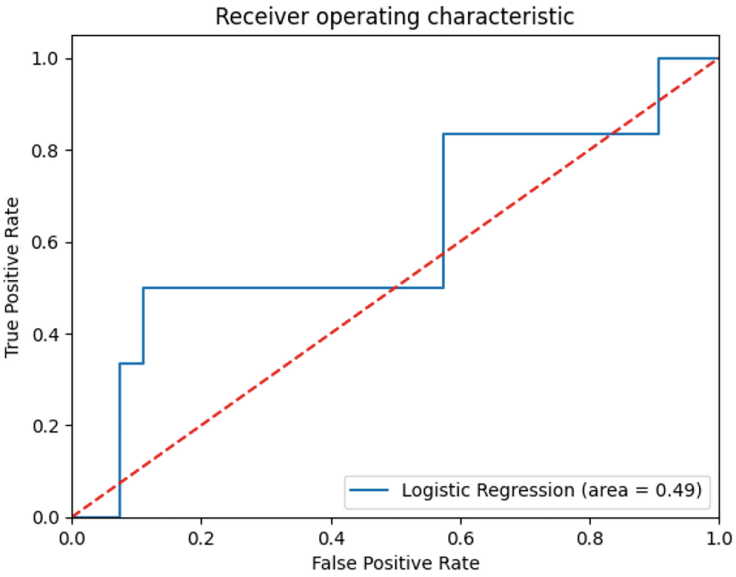

Fig. 7. ROC curve for logistic regression.

number of parameters as introducing more parameters resulted in reaching almost perfect score for different models which proved that the parameters may have been overfitted for these selected parameters. Therefore, further studies for applying regularization of the overfitted parameters, or introducing different parameters that does not rely on other parameters such as ground improvement depth with its stiffness may further provide more accurate prediction model, which is out of the scope for this paper.

As the objective of this paper is to introduce logistic regression analysis with different methods with settlement analysis models, the hyper parameter selection and the target prediction parameter was selected for the purpose of simplicity. This also provides a good starting point for expanding the application to eventually allow the ML algorithms to provide suggestions with sequence modeling approach. The goal in this paper was to predict whether the model has ground improvement or not. This means that the performance of ML algorithm to detect whether the model has ground improvement or not will increase with more dataset.

Score	Model
100.00	Random Forest
100.00	Decision Tree
91.37	KNN
82.01	Logistic Regression
79.86	Perceptron
79.86	Stochastic Gradient Decent
57.55	Support Vector Machines
46.04	Naive Bayes

Fig. 8. Prediction model score with different ML algorithms used.

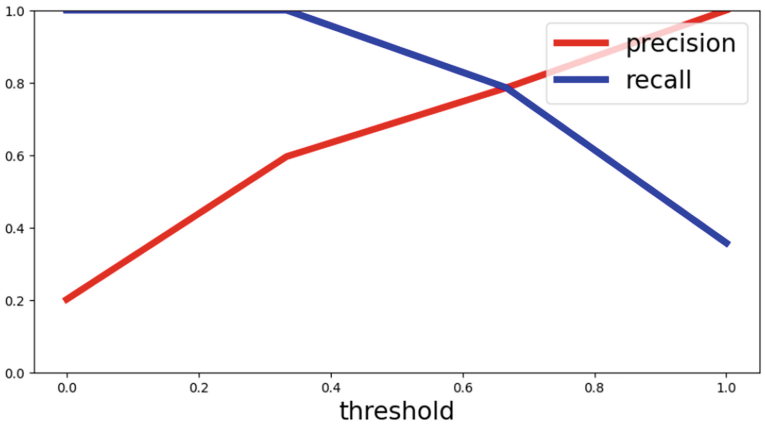

Fig. 9. Threshold of precision to recall curve with KNN model.

4 Future Studies

As mentioned previously, regularization of parameters to overcome overfitting of data or dealing with the biased dataset, to the selection of the parameters to improve the ML algorithm performance are always the challenge. More study will be required to investigate the performance of different ML applications for settlement analysis and its various features in Settle3.

The motivation for this study initially came from Recurrent Neuron Network (RNN) where it allows each layer of the network to contribute to the overall analysis of the results by predicting the most likelihood of the combination of each step to the predicted result. Also, this algorithm has Long Short-Term Memory (LSTM) which remembers a state of the layer which may affect the other layers within the overall steps of the analysis.

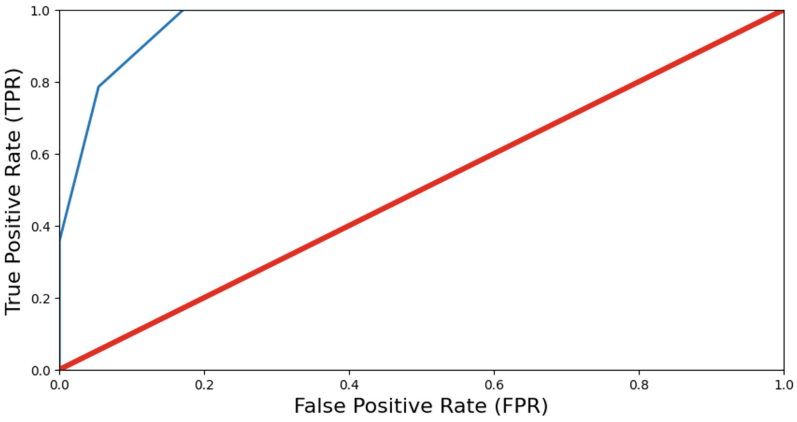

Fig. 10. ROC curve for KNN model.

This also resolves the issue with vanishing gradient descent in the training process. The paper study was implemented as the beginning platform for building ML algorithms with simple dataset. In the future, more thorough analysis of the data, as well as different ML applications will be investigated to allow engineers and practitioners to fully utilize the concepts to geotechnical software analysis applications related to sequence modelling.

References

Ackerman, Daniel (2021) "New AI tool calculates materials' stress and strain based on photos". MIT News Office. Link: https://news.mit.edu/2021/ai-materials-stress-strain-0422.

Ching, Jianye & Zhiyong, Yang & Phoon, Kok-Kwang. (2021). Dealing with Nonlattice Data in Three-Dimensional Probabilistic Site Characterization. Journal of Engineering Mechanics. 147. 06021003. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001907.

Phoon, Kok-Kwang & Ching, Jianye & Wang, Yu. (2019). Managing Risk in Geotechnical Engineering – From Data to Digitalization. 13–34. https://doi.org/10.3850/978-981-11-2725-0-SL-cd.

Robertson, P.K. (1990) "Soil classification using the cone penetration test" Department of Civil Engieergin. The university of Alberta, Edmonton, Alta., Canada. Canada Geotech J 27. 151–158.

Niklas Donges (2018) "Predicting the survival of Titanic Passengers". *Towards Data Science*. Link: https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8