Implications of Feature Engineering on Predicting Foundation Settlements

Joachim M. Kirstein

January 19, 2025

1 Introduction

This article intends to evaluate the implications of feature engineering on prediction accuracy of two machine learning models when predicting foundation settlements in plane strain conditions. Performance of linear regression and random forest have been evaluated.

The prediction accuracy is plotted in relation to settlement values and the plots constitutes the basis for discussion. The aim of this article is to highlight the implications of feature engineering.

2 Dataset

The dataset are features and outputs of around 5600 simple FE models. The models have been created by a script generating random models within a range of parameters. A train test split has been applied with a ratio of 90 to 10 %.

The models are made with a stiff plate element, exposed to a vertical force with varying eccentricity. The subgrade consists of random layers per 0.5 m. The subgrade layers are modeled with a Mohr-Coulomb soil model with a high failure criterion.

- Stiff plate element, width varies between 1-4 m.
- Load applied is an out-of-plane lineload equivalent to 100 kN/m/m · foundation width.
- Load is placed with a random eccentricity between 0 and 0.3 · foundation width.
- Bottom boundary conditions are equivalent to 2·foundation width.
- Total model width is 4·foundation width. The foundation is placed at model center.
- Subgrade layers are created with E-modulus randomly between 10 and 100 MPa. Poissons ratio of 0.3 is applied. Angle of friction of 40° and c'=300 kPa have been applied.

3 Results

In the following, three sets of results are presented. The first is without any feature engineering applied to the dataset. The second contains some feature engineering on the subgrade E-modulus, and the third contains feature engineering on the relationship between layer influence compared to foundation width. The implication of the modified features are discussed in each result subsection.

3.1 No Feature Engineering

Figure 1 shows the prediction error of linear regression and random forest model with no feature engineering applied.

Generally, the linear regression predictions are +/- 30 % of correct values. with best fit around 4-8 mm settlements. The largest deviations are observed at very small settlements, but also at large settlements consistent errors are observed.

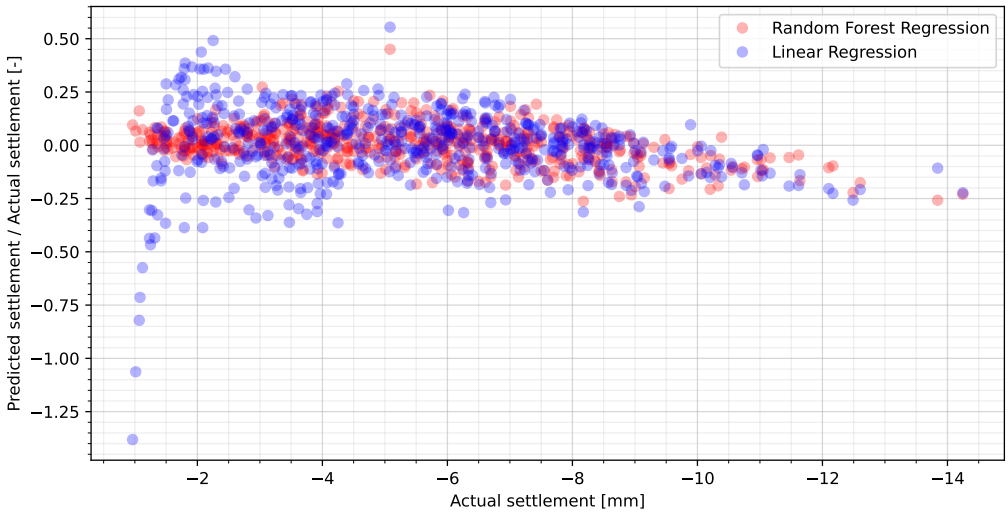

Figure 1: Predictions with no feature engineering applied

The random forest are predicting better than the linear regression, and much better at low settlements, however a discrepancy of +/- 20 % are observed.

3.2 Inverse Subgrade E-modulus

The settlements to E-modulus are an inverse linear correlation; with an increased E-modulus, the settlements are expected to decrease, whereas with an increased E-modulus, the settlements are expected to go towards infinity.

Therefore, applying the inverse E-modulus instead of the E-modulus as features to the calculations, the linear regression model is expected to capture the inverse-linear behavior of the settlement to E-modulus correlation.

Figure 2 shows the results, and compared to the results from figure 1, the predictions by the linear regression model has significantly improved. Now, the settlements between 3 and 10 mm are quite accurate. However, a large discrepancy is still observed for settlements below 2 mm.

This improvement demonstrates how incorporating an inverse E-modulus feature aligns the model input with the expected geotechnical behavior, enabling even simple linear models to better represent the underlying soil-structure relationship.

The random forest model remains unchanged, which correlates well with the decision tree model that it applies. No influence of inverting E-modulus is expected.

3.3 Subgrade depth to foundation width ratio

For smaller foundation widths, the upper layers become significantly more important compared to the deeper layers, as the stress propagation spreads out causing almost no strain in the deeper layers. The increment of this effect decreases per foundation width, as the foundation size increases. Hence, the difference between a 1-2 m foundation width is significantly larger than a 4-5 m foundation width. When the linear regression model fits to all model results and they are weighted evenly, it is evident, that feature coefficient for the deeper layers are overestimated for small foundations, and that the upper layers are significantly underestimated for smaller foundations.

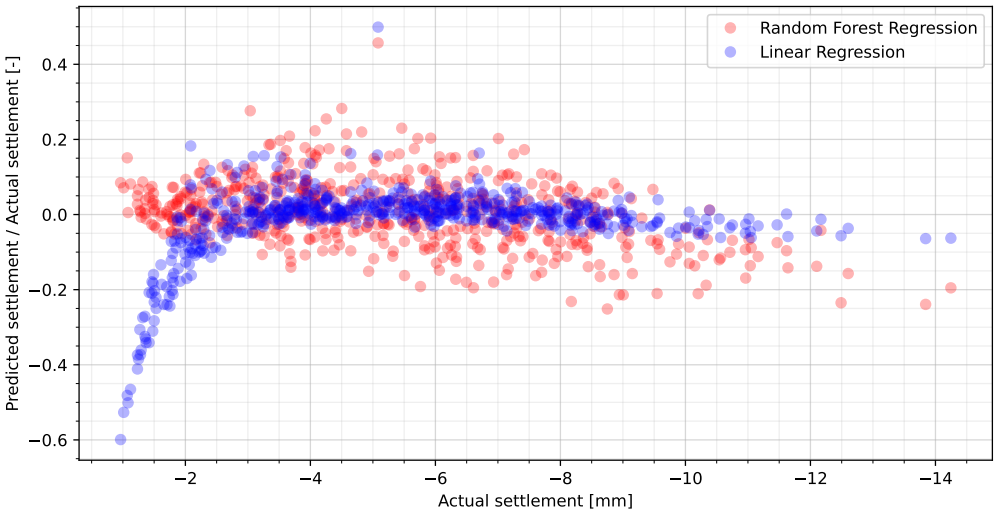

Figure 2: Predictions with inversed stiffness modulus.

The effect can to some degree be captured in the linear regression model by adding a new feature per soil layer, which is the soil layers' inverse stiffness multiplied by foundation size. This provides a parameter that allows the foundation size' influence on the soil layer (to a first degree) to be captured. That feature coefficient per soil layer along with the soil layer coefficient itself, allows for a non-linear soil layer influence in relation to foundation size, and allow a larger foundation to weigh the deeper soil layer coefficients higher.

The results are shown in figure 3 and it is evident that the linear regression model performs significant better, especially for the lower settlement values.

Again the random forest model performs similarly to previous results, as expected.

The feature coefficients are shown in figure 4. The figure contains 16 width-times-inverse-stiffness (one for each layer), along with 16 inverse stiffness values. The last two shows eccentricity and foundation width. From the graph it is interesting how each of the soil layers are values evenly except for the very top ones, which are weighed very high, whereas the very deepest are weighed low. This allow the small foundation sizes to show significant influence of the top layers, while not impacting the larger foundation sizes significantly. The larger foundation sizes can capture deeper layers by the width-times-inverse-stiffness, where the top layers are actually lower than the layers just below the surface. The influence drops again for the deeper layers. The bottom layers are having lower coefficients which is to be expected.

4 Conclusion

This study highlights how feature engineering grounded in geotechnical principles can significantly enhance the performance of simple machine learning models, particularly linear regression. By leveraging knowledge of foundation settlement behavior, such as inverse relationships with stiffness and the varying influence of soil layers, can create features that capture complex interactions. More advanced models like random forests are not affected by such feature engineering.

This exercise demonstrates the value of domain knowledge in improving performance and understanding of machine learning models.

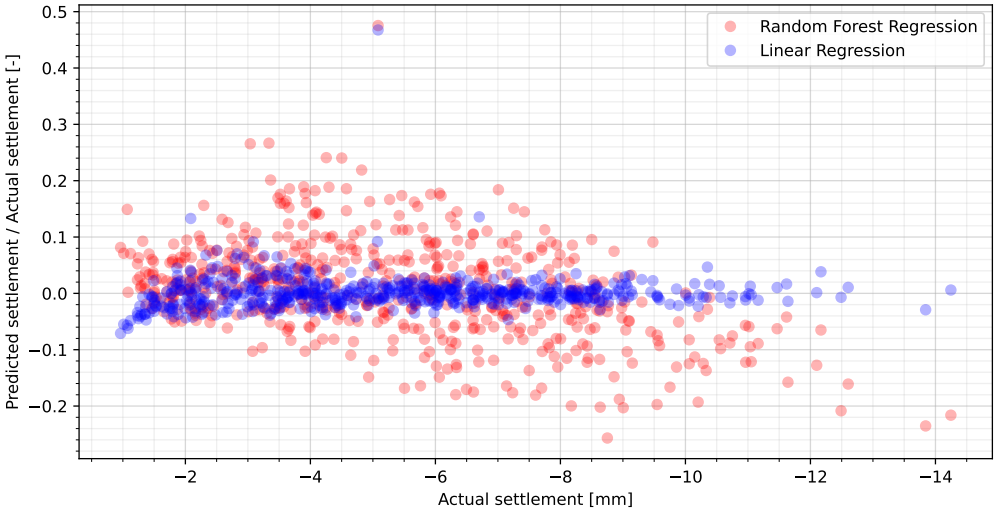

Figure 3: Predictions with inverse stiffness and width-depth relation.

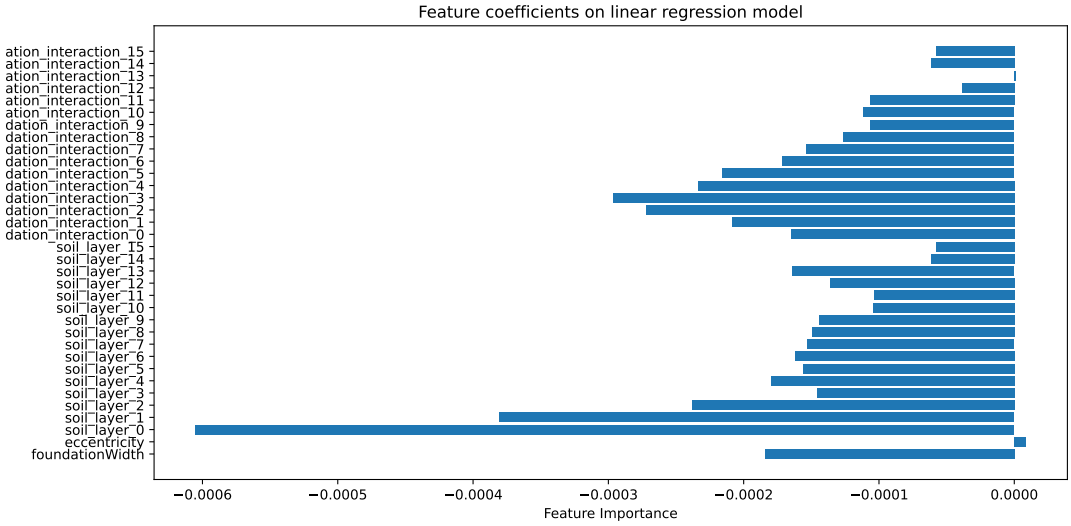

Figure 4: Feature coefficients on linear regression model