ORIE 4741 Project Proposal

Impact of Accuracy on Feature Explanation Methods

Brian Liu

Christina Zhou

1 Objective:

The objective of this project is to research how popular feature explanation techniques such as LIME and SHAP perform when the accuracy of the underlying model is varied. Through this research, we hope to develop an algorithm that can robustly interpret black-box models that have varying accuracy levels.

2 Background:

Complicated models such as boosted trees, support vector machines, and neural networks are often used in data science to solve real-world problems. One major drawback of these black-box models is that the relationships between their feature inputs and final outputs have no intuitive interpretation. In contrast, simpler models such as linear regression have easy to understand feature interpretations. For example, the coefficient of a linear regression model, β_j, has the interpretation of being the mean change in the response per unit change in the predictor x_j. Model explanation methods such as LIME and SHAP were introduced to make black-box models interpretable. These methods often involve approximating black-box models locally with simpler explainer models to derive real world insight. However, these methods assume that the underlying black-box model fits the data very accurately. In practice, often times data scientists are required to extract features from models that are not accurate enough to be put into production. We hope to quantify how the performance of feature explanation methods vary with model accuracy.

3 Approach:

The first step of our project is to test out several feature explanation methods on a simple, well studied data set. We will fit a black-box model to the data, as well as linear/logistic regression and an interpretable tree model such as random forest. We can then compare the feature importance rankings between the regression and tree models with the feature importance rankings provided by using SHAP and LIME on the black-box model. By throttling how much data we train our models on, we can manipulate the model test accuracy. We can then reapply our feature explanation techniques to see how our feature explanations vary with model accuracy. From there, we can then work towards building a more robust feature explanation method.

4 Data:

We will test our experiment on datasets that have many different features and feature types, and that can be accurately modeled by a variety of different methods. We can then reduce the accuracy of these models to see the impact on feature explanations.

- For classification modeling, the famous Titanic dataset would work well for our experiment. Most classification algorithms work extremely well on this data; logistic regression, random forest, and SVM classifiers are all $> 80\%$ accurate when predicting passenger survivor .
- For regression modeling the Ames house pricing dataset would work well for our experiment. This dataset has over 80 features and is used in many data science classes to teach multiple linear regression.

5 References:

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2016, pp. 1135–1144.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 4768–4777. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

ORIE 4741 Project Proposal

Impact of Accuracy on Feature Explanation Methods

Brian Liu

Christina Zhou

1 Objective:

The objective of this project is to research how popular feature explanation techniques such as LIME and SHAP perform when the accuracy of the underlying model is varied. Through this research, we hope to develop an algorithm that can robustly interpret black-box models that have varying accuracy levels.

2 Background:

Complicated models such as boosted trees, support vector machines, and neural networks are often used in data science to solve real-world problems. One major drawback of these black-box models is that the relationships between their feature inputs and final outputs have no intuitive interpretation. In contrast, simpler models such as linear regression have easy to understand feature interpretations. For example, the coefficient of a linear regression model, β_j, has the interpretation of being the mean change in the response per unit change in the predictor x_j. Model explanation methods such as LIME and SHAP were introduced to make black-box models interpretable. These methods often involve approximating black-box models locally with simpler explainer models to derive real world insight. However, these methods assume that the underlying black-box model fits the data very accurately. In practice, often times data scientists are required to extract features from models that are not accurate enough to be put into production. We hope to quantify how the performance of feature explanation methods vary with model accuracy.

3 Approach:

The first step of our project is to test out several feature explanation methods on a simple, well studied data set. We will fit a black-box model to the data, as well as linear/logistic regression and an interpretable tree model such as random forest. We can then compare the feature importance rankings between the regression and tree models with the feature importance rankings provided by using SHAP and LIME on the black-box model. By throttling how much data we train our models on, we can manipulate the model test accuracy. We can then reapply our feature explanation techniques to see how our feature explanations vary with model accuracy. From there, we can then work towards building a more robust feature explanation method.

4 Data:

We will test our experiment on datasets that have many different features and feature types, and that can be accurately modeled by a variety of different methods. We can then reduce the accuracy of these models to see the impact on feature explanations.

- For classification modeling, the famous Titanic dataset would work well for our experiment. Most classification algorithms work extremely well on this data; logistic regression, random forest, and SVM classifiers are all $> 80\%$ accurate when predicting passenger survivor .
- For regression modeling the Ames house pricing dataset would work well for our experiment. This dataset has over 80 features and is used in many data science classes to teach multiple linear regression.

5 References:

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2016, pp. 1135–1144.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 4768–4777. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf