**Discovering the Possibility of Improving Classification Modeling Results Using Explainable Machine Learning in Prediction Tasks**

**Helena Hu** xh2626**, Wendy Wang** cw3539**, Yuqi He** yh3653**, Donglizhen Shi** ds4224

**Abstract**

In this project, we aim to discover if SHAP and LIME and similar explainable ML tools can help us improve prediction in classification models. The first step is to find one or more metrics that help us benchmark SHAP and LIME’s performance. The second goal is to figure out if SHAP and LIME’s inferences can help us improve the models based on these metrics. We implemented classification models on 9 different datasets and tried improving them. We find that the existing conventional metrics are suitable for such experiments and SHAP and LIME, although unable to improve the model by a significant amount, can provide useful inferences that help make further decisions that are potentially improving the models to some extent.

1. **Introduction**

**1.1. Research Goals**

Explainable ML is getting significantly more popular each year for its ability to help researchers understand black-box models, at least to some extent. One question raised commonly in the academic community is “Can we somehow reverse engineer the process?”, which is really a fair question to ask since if we understand the model better, we should be able to improve it. With this question in mind, there are two important follow ups: 1. What metric is a good metric to measure “improvement” of models? 2. Are we able to “improve” models under some general setting with Explainable ML tools? These are the research questions of our project.

**1.2. General Setting**

Although it would be nice if we have the time and capacity to test our assumptions on all types of models, it is simply too much work for a 3 month project. Therefore we decided to start small, starting from a setting where we think can yield interesting results without too much hassle. We decided to restrict all our prediction tasks to classification and we decided to test out different types of existing metrics ( namely, what metric to decide our model to be better) and see if they might make a difference instead of creating our very own metric. It is also pretty important to select simpler explainable ML tools in terms of difficulty of implementation. There are many other ways to improve explainability and many of them are extremely powerful and sexy, but those can be too time consuming to implement. So SHAP and LIME are selected for this project, due to its simplicity in explaining as well as debugging, although some limitations apply.

Hence, our final research questions should be: **1. Are existing metrics good metrics to measure “improvement” of classification models? 2. Are we able to “improve'' classification tasks with SHAP and LIME?**

**1.3. Exlplainable ML with SHAP and LIME**

LIME is able to provide explainability to the decision-making of a model locally (for a given instance). Linear SHAP approximates feature importance for regression, which we utilized in our initial research. We will be testing both LIME and SHAP for their potential effectiveness on improving classification tasks.

**1.4. Deliverables and Overall Results**

The experiments are designed to be simple: we take 9 datasets that are very commonly used as examples for classification tasks, fit them in different commonly used classification models, then use different metrics to evaluate the performances. Once this step is done we then use LIME and SHAP to explain them. Depending on the explanation results, we eliminate a number of variables depending on the distribution of the Shapley values, then fit the model again, recording the new model’s results. We then compare these new results with the old model, and see if they are actually different from the original results.

The datasets are not necessarily synthesized, but close. An example is the iris dataset, it is one of the most commonly used for examples of classification predictions. The data is also very clean, so very little cleaning work is required. It is also important to note that if Explainable ML methods are unable to pass tests in these datasets, then they most likely won’t be effective on more complex datasets and tasks.

Due to time constraints, we focused on only the accuracy, F1-score and its variations( macro F1, micro F1, and weighted F1), and ROC/AUC outputs. These will be the metrics used.

The results are pretty much what we expected: in some datasets using LIME or SHAP (especially SHAP) can be useful to benefit the feature selection process. In other datasets, very little or no improvements are present, thus they need to be further examined, perhaps with a different explainable ML method. The metrics are in a trickier situation to summarize. The idea of our metric selection process is that if most metrics point us in one direction and one does not, we decide to trust the majority more. It is important to note that we are not here to pick one perfect metric to benchmark these experiments, rather to weed out some potentially subpar metrics. In fact, our results show that all metrics we choose are impacting our decisions roughly the same way, making them all good potential metrics for future experiments.

1. **Literature Review**

**2.1. LIME [1]**

From multiple previous work on LIME, we chose one that utilizes LIME on explaining classification tasks since our focus is also on classification tasks.

The paper [3] explains how to explain a machine learning model using LIME which is a novel technique that interprets a prediction locally. LIME is used by users to interpret and assess how trustworthy a model is.

To prove explanations vital, there is a text classification task which has a SVM model with high accuracy but the explanation of the prediction shows serious issues with it. It is tempting to see high accuracy predictions and use the model, but if the reasons for accurate predictions are not clear then an explanation is necessary. An image classifier using deep networks classified a guitar and the explanation of the prediction backed it up (The conclusion the model made was reasonable). The outcomes show that LIME is useful in aiding users to affirm trust in predictions. It is harder to trust a whole model instead of just predictions from a model.

LIME is crucial to our attempt on local explanations for different classification models. The paper provides ways to even do some of the most complicated classification tasks when LIME is present.

**2.2. SHAP [2]**

Compared to LIME and other traditional Shapley methods, SHAP is a new method. The original paper [4] and codes were analyzed, and many of the findings are helpful to our ultimate goal of explaining models.

In the paper, the authors aim to find a way to understand the explainable machine learning methods and propose a unified method to interpret model predictions. Previous explainable machine learning methods sometimes lack identification of a new class of feature importance measures and theoretical results showing unique solutions. How they are related or if one outperforms another in some situations has not yet been addressed.

Hence, the authors raised the concept of additive feature attribution methods, giving it three principles and proposed SHAP, a unified approach that satisfies all three principles. Building SHAP into FLAML will help us to identify the class of additive feature attributions methods and find the unique solution in the class that adheres to desirable properties.

SHAP uses the unification of six methods to show improved computational performance and consistency with human intuition. These six methods include LIME (produce local linear interpretation for models), DeepLIFT (recursively explains deep learning predictions), Layer-wise Relevance Propagation (Similar to DeepLIFT, but with reference activations of neurons set at 0), Shapley regression values (assigns importance values to features), Shapley sampling values (explains a model through sampling approximations and approximates effect of removing a variable), and Quantitative Input Influence (proposes a sampling approximation). These methods were closely examined by their adherence to the three properties of additive feature attributions and their relatedness.

Different merges of different methods (Kernel SHAP, linear SHAP, max SHAP, Deep SHAP, and low-order SHAP) were created with different SHAP value calculations and purposes. The methods were then compared in different tasks. The outcome shows that SHAP increases local accuracy and consistency compared to previous methods which use additive feature attributions.

1. **Methodology**

**3.1. Datasets**

As aforementioned in the introduction, we choose a total of 9 datasets that are not synthesized and clean for experimentation. Each dataset has a response variable where the output is categorical and later engineered to be nominal (with classes 0, 1, 2, for example). A typical example of our dataset is the *Abalone* dataset. *Abalone* has a response variable indicating the sex of abalone (0 for female, 1 for intersex, 2 for male), then 8 different variables indicating their size/weight. Figure 1 is a simple visualization of the features, excluding the response.

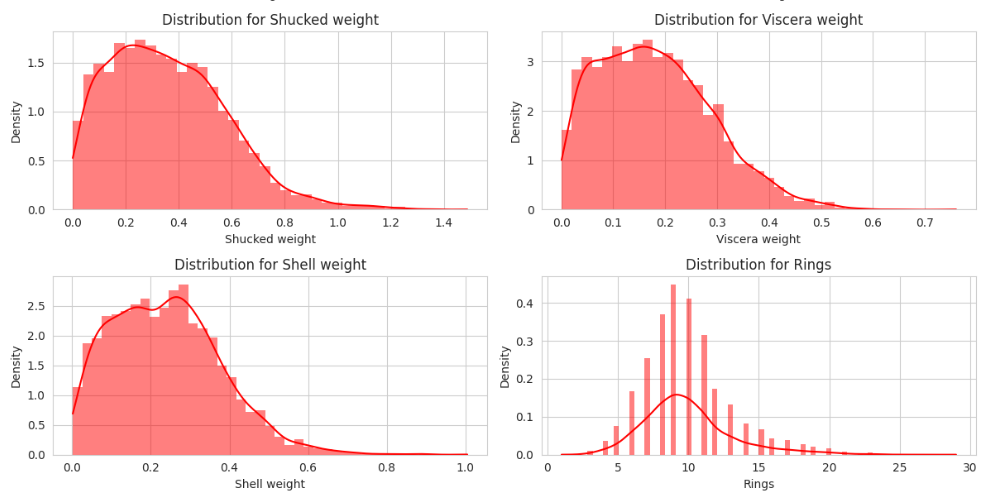
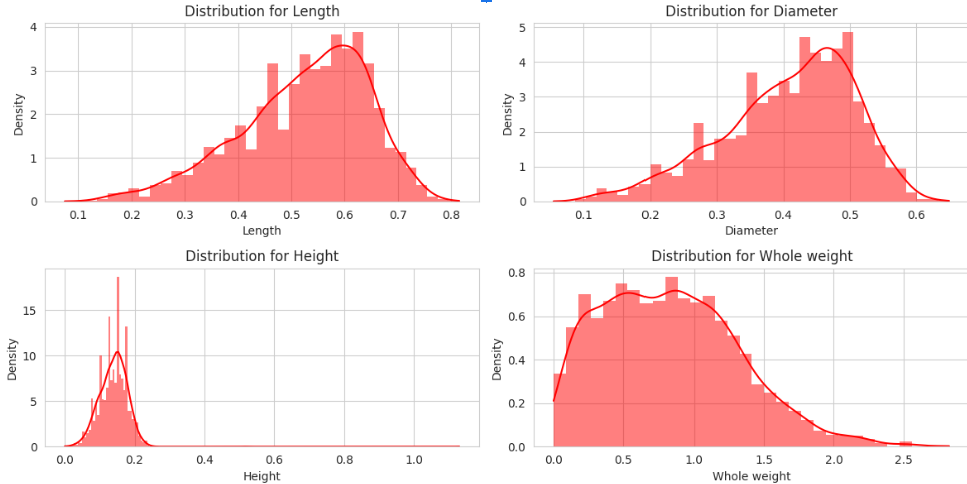


Figure 1: Abalone features

Other datasets are in similar nature: 1 response followed by numerous features.

**3.2. Modeling**

The models we fit are simple. We have chosen datasets that are suitable for classification predictions to start with and predictions using classification is exactly what we did. We chose 8 different methods of classification: Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Adaboost, XGboost (XGB), and Neural Network (NN, fit by the neural\_model defined by the simplest form in Keras).

We then fit these models to our data, using a training set and a testing set, of course (75-25 ratio). We conduct a preliminary model selection by looking at the metrics we choose collectively. For example, on the Iris dataset, we would look at results shown in Figure 2 and make a decision to stick with one particular model for benchmarking SHAP and LIME. Here we can see KNN is no doubt our final selection of the model we are improving through SHAP and LIME.

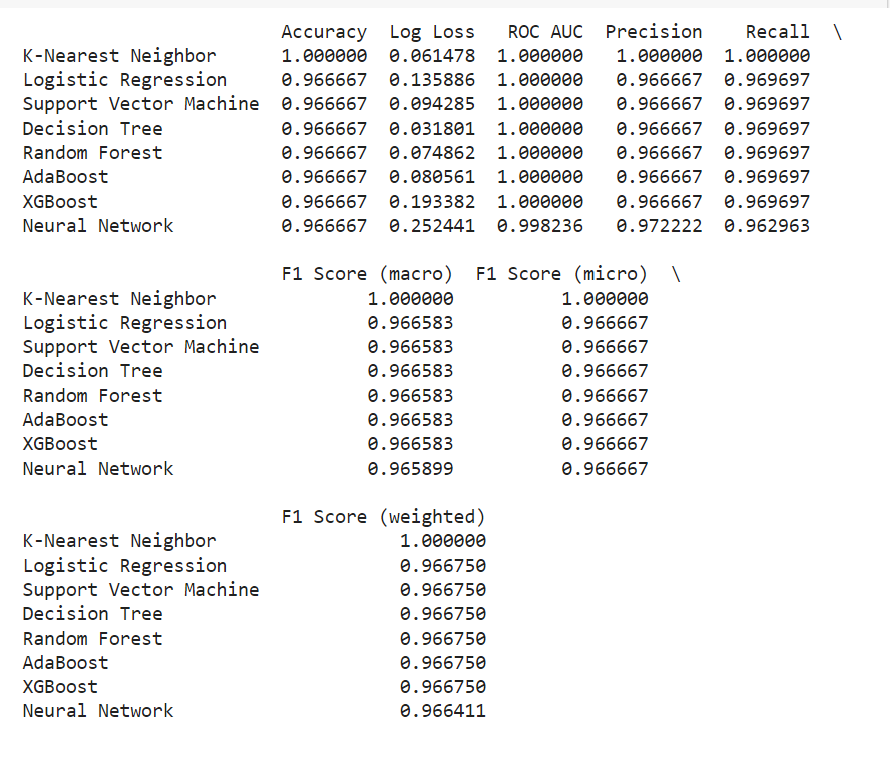


Figure 2: model metrics after fitting all methods to Iris

We choose to use Accuracy, Log loss, ROC AUC, Precision, Recall, and different variations of F1 scores as metrics. Under most circumstances, they would ( and should) give similar inferences. If one of the metrics show dramatically different inferences than the majority for a particular model on a particular dataset, we would stop considering it a decent metric.

So the final model in this stage is the best model based on our collective metrics, while all features are used. From this point on, nothing is changed in the models except for features. ROC/AUC plots and other inferences are created to help us make decisions in the next section, which we included some examples in the Appendix. More of those can be found in the Project Final Plots file.

**3.3. Explaining**

The next step is to obtain SHAP and LIME output for further inference. In our best model, SHAP and LIME were used to find the impact of each feature to the model. In all models, the complexity of features provide some favor to SHAP and LIME – the amount of features are usually a little overwhelming for the model to not overfit. Figure 3 and 4 shows two types of ways to visualize SHAP for each feature, and figure 5 shows how LIME interprets the models.

Figure 3 shows a local interpretation (Force plot) of our model fitted on dataset Abalone. There are a few features here that we would suspect giving very little impact to the predictions: Height, Diameter, Whole Weight, etc.

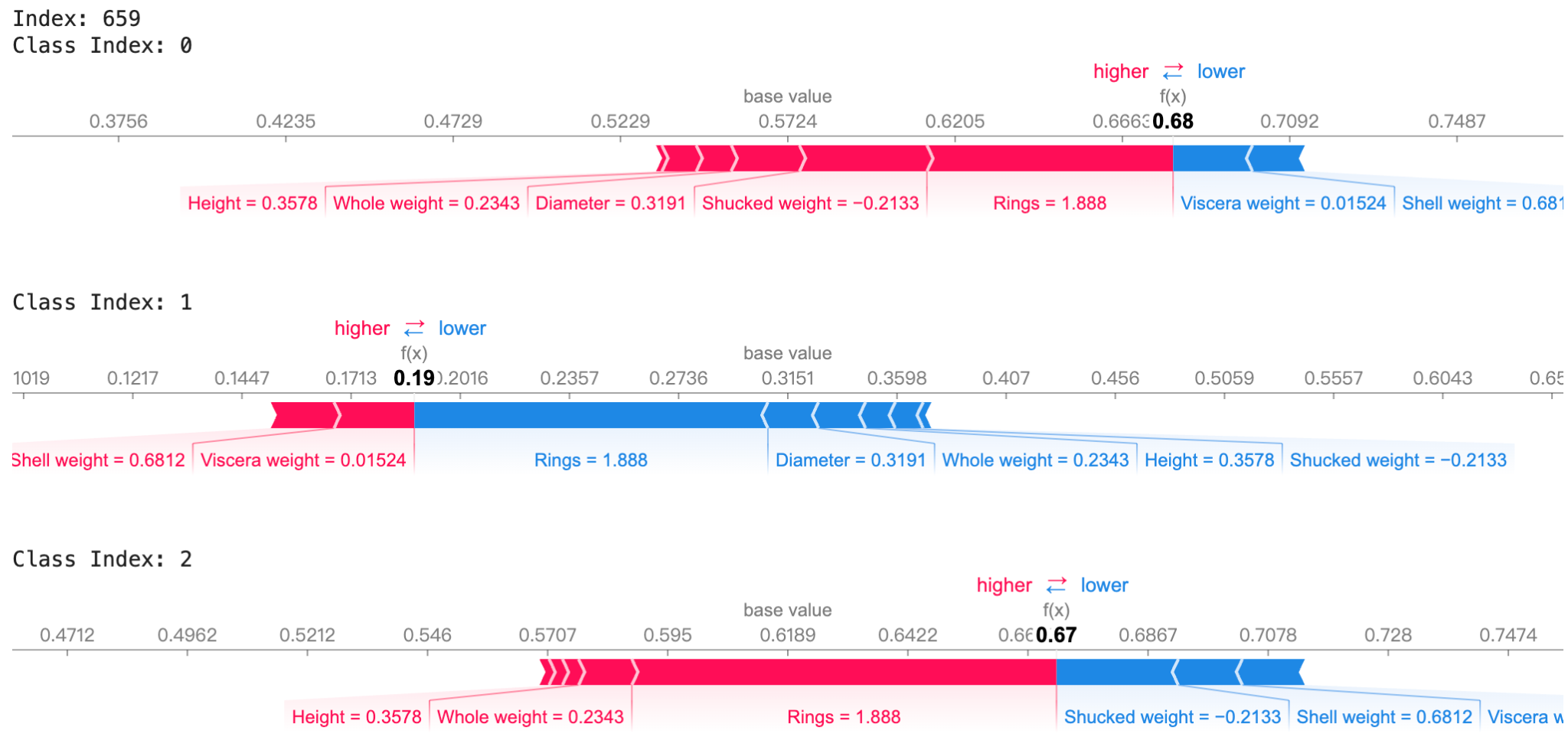


Figure 3: Local interpretation from SHAP on Abalone

Figure 4 shows a feature importance visualization created based on Shapley values of our model on Abalone. It pretty much aligns with our summary above, with a little more certainty. We have many reasons to drop Diameter or Height as features.

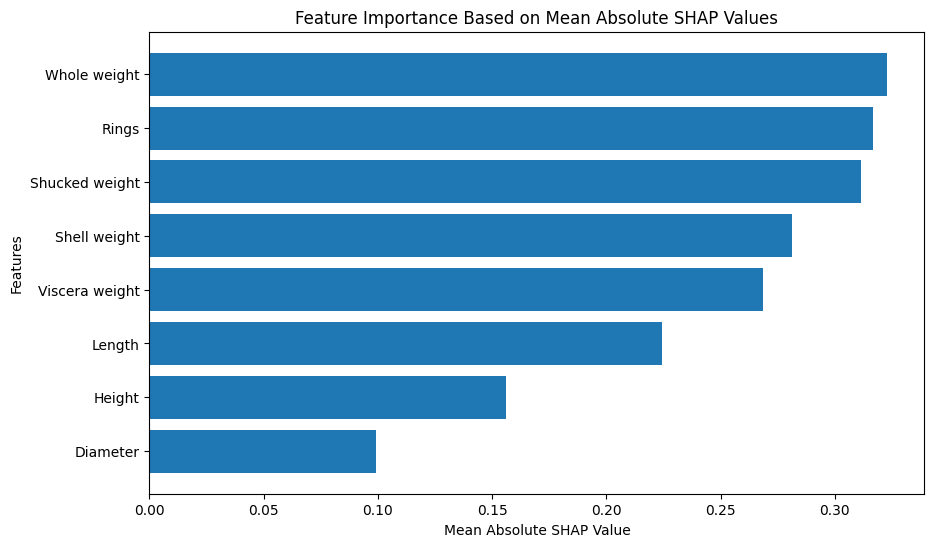


Figure 4: Global interpretation from SHAP on Abalone

Figure 5 shows a LIME interpretation of the model on Abalone. Notice that in the two instances we are looking at, Height is again contributing very little to the prediction. Although notice that in both cases, Length bet against the prediction outcome, these can all be taken into account. So height and length will be considered to be dropped based on the LIME output.

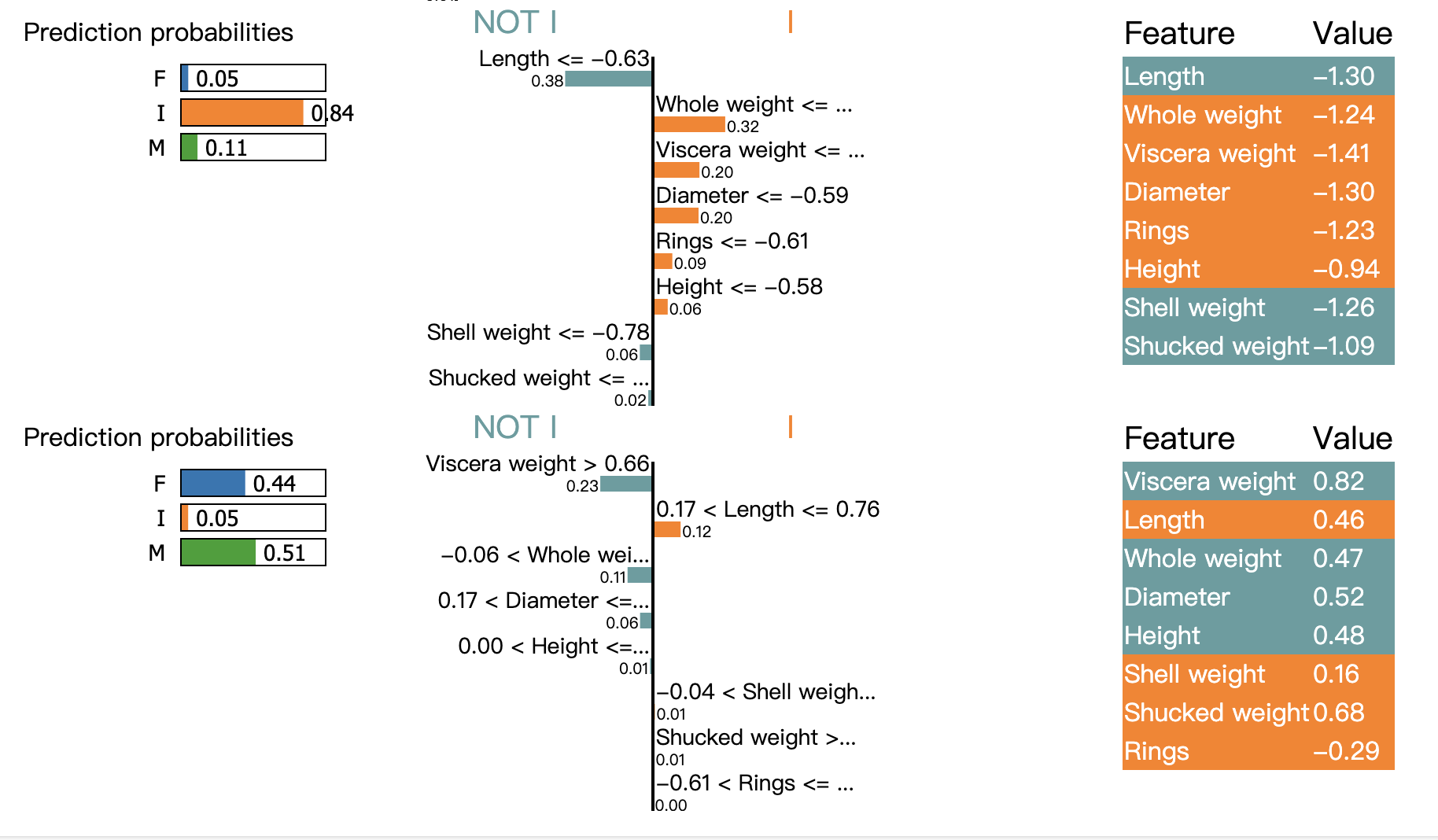


Figure 5: LIME output of classification model on Abalone

**3.4. Improving**

Based on observations in the previous part, we fit new models where everything stays the same except the features. We would remove undesirable features based on SHAP and LIME outputs, then evaluate the new model.

The criteria to remove features is a tricky one. We decided to keep this simple: Remove the feature with the lowest importance in SHAP, or with the least contribution in LIME, then fit a new model with the remaining features. If we see an improvement in any of our selected metrics, we will continue doing so until dropping variables starts drastically lowering all of our metrics.

The result for each model is that we would have one full model ( described in part 3.2), one refined model using SHAP, and one improved model using LIME. Their Accuracy, ROC/AUC, and F1 scores are recorded in a new dataset called Results. Figure 6 shows the head of Results. The reason to not use recall and precision is that they are essentially the same as accuracy and F1 scores. The reason to not use log loss here is that they are in the opposite direction when optimizing ( we want lower log loss and higher other metrics).

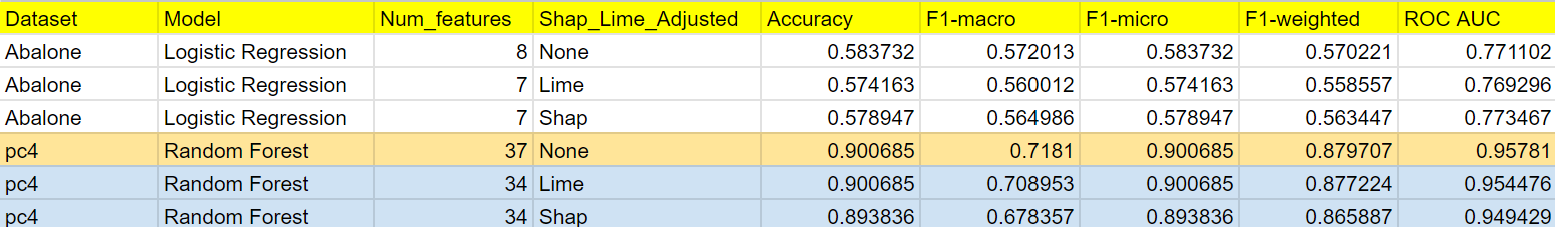


Figure 6: Head of Results dataset

**3.5. Testing**

There are two layers of testing our results. We want to know which of the metrics are useful in such an experiment, so there are a few scenarios to think about: if all metrics point us in the same direction, then no further analysis is needed, we can conclude that all metrics we chose for this experiment behave the same way. If there is one metric not pointing in the same direction as others, we will tend to conclude that this particular metric is not useful. If it is a pretty even split, further testing is needed.

On the other hand, we want to know, based on our metrics, if we are improving the models with SHAP and LIME. So here comes two tests: the first one is a formal Wilcoxon sign-rank test. We simply want to know if there is significant difference between the values of the metrics before and after applying SHAP and LIME. Another important thing is the eye test. There might not be significance in the difference, but if SHAP and LIME improve models by a tiny bit in every case, then there could be some effectiveness to it.

1. **Results**

**4.1. Metrics**

As aforementioned a Wilcoxon sign rank test was performed. We find that for all metrics, no statistically significant difference was found between before and after SHAP/LIME. This indicates that the metrics, at least under our circumstance, are all valid for this type of experiment.

**4.2. Improvement**

The Wilcoxon sign rank test answered our second research question by giving us extremely high p-values (all p-values from all the tests turn out to be >0.95, many of them are just 1). We fail to reject all null hypotheses and there are no significant differences between our metrics before improving the model and after improving the model.

However, there are two important notes to make. The first is that in most cases we improved our metrics by dropping one or more features. Although the improvement is deemed trivial by the Wilcoxon sign rank test, we are still using less features and getting consistently the same if not better results. The bottom line is that using SHAP/LIME does not hurt the modeling process, and could even give useful inferences in improving the model. Another important point is that we did limited experiments on limited models. We also use very limited features of SHAP/LIME. This is due to the complexity of considering all parties involved when fitting a model – we obviously want to keep as many things unchanged as possible to see if there is significance, but interactions between hyperparameters and features are not linear, these thoughts should be taken into account in the future.

1. **Conclusion**

In Summary, the metrics accuracy, ROC/AUC, and different variations of F1-scores are useful and competent metrics to decide whether a classification model is improved. SHAP and LIME does not significantly improve the classification models with its most basic features, but gives some useful inference to improving the model to some extent, depending on the nature of the data.

1. **Recommendations**

We believe we are up for a good start. A lot of things can be tried with experiments like this. For one, SHAP and LIME might work better on improving more complex models. So fitting a different type of model like regression or GNN is a good start in the future. The selection of datasets are important too. This time we only used real data, but synthesized data can show more obvious results for us to gauge the progress of SHAP and LIME. Obviously, the metrics we choose are a little too conventional. A different but reasonable metric can be used to gauge the improvements as well.

1. **Citations**

*[1]Ribeiro, Marco Tulio, et al. “‘why Should I Trust You?’: Explaining the Predictions of Any Classifier.” arXiv.Org, 9 Aug. 2016,* [*arxiv.org/abs/1602.04938*](http://arxiv.org/abs/1602.04938)*.*

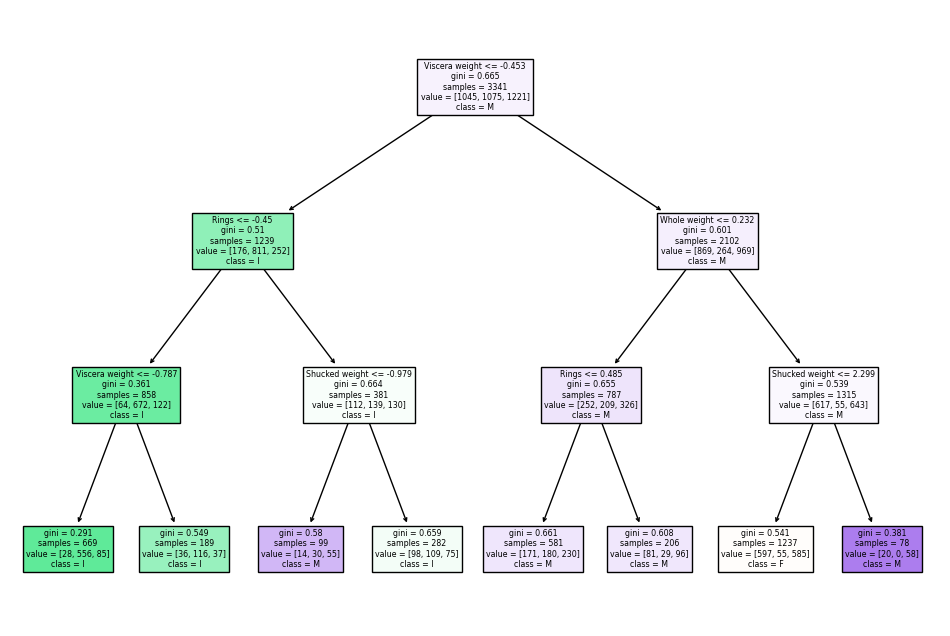
*[2]Lundberg, Scott, and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” arXiv.Org, 25 Nov. 2017, arxiv.org/abs/1705.07874.*

*[3]Burkart, Nadia, and Marco F. Huber. “A Survey on the Explainability of Supervised Machine Learning.” arXiv.Org, 16 Nov. 2020, arxiv.org/abs/2011.07876.*

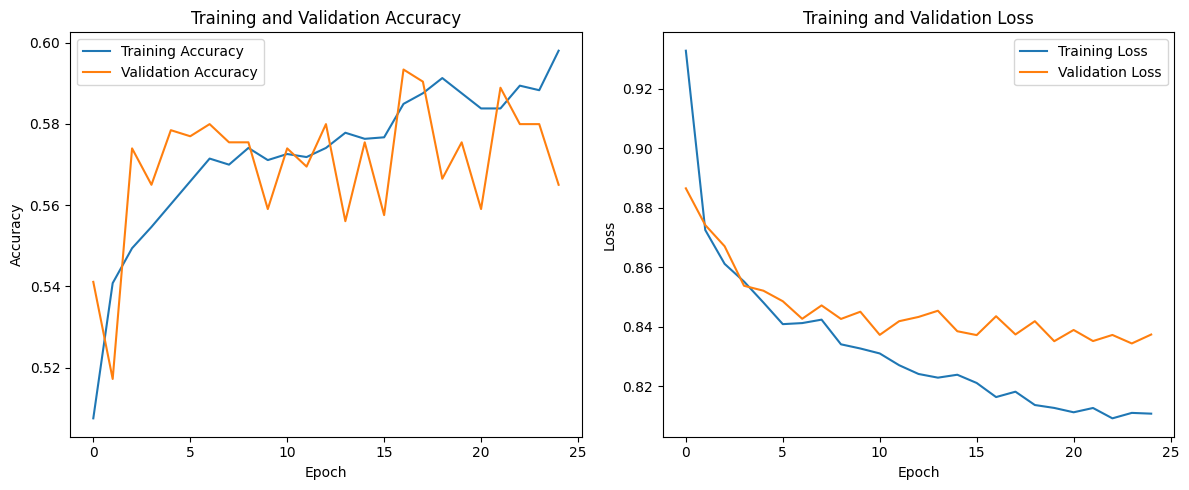
*[4]Weber, Patrick, et al. “Applications of Explainable Artificial Intelligence in Finance-A Systematic Review of Finance, Information Systems, and Computer Science Literature - Management Review Quarterly.” SpringerLink, Springer International Publishing, 28 Feb. 2023, link.springer.com/article/10.1007/s11301-023-00320-0.*

**Appendix**

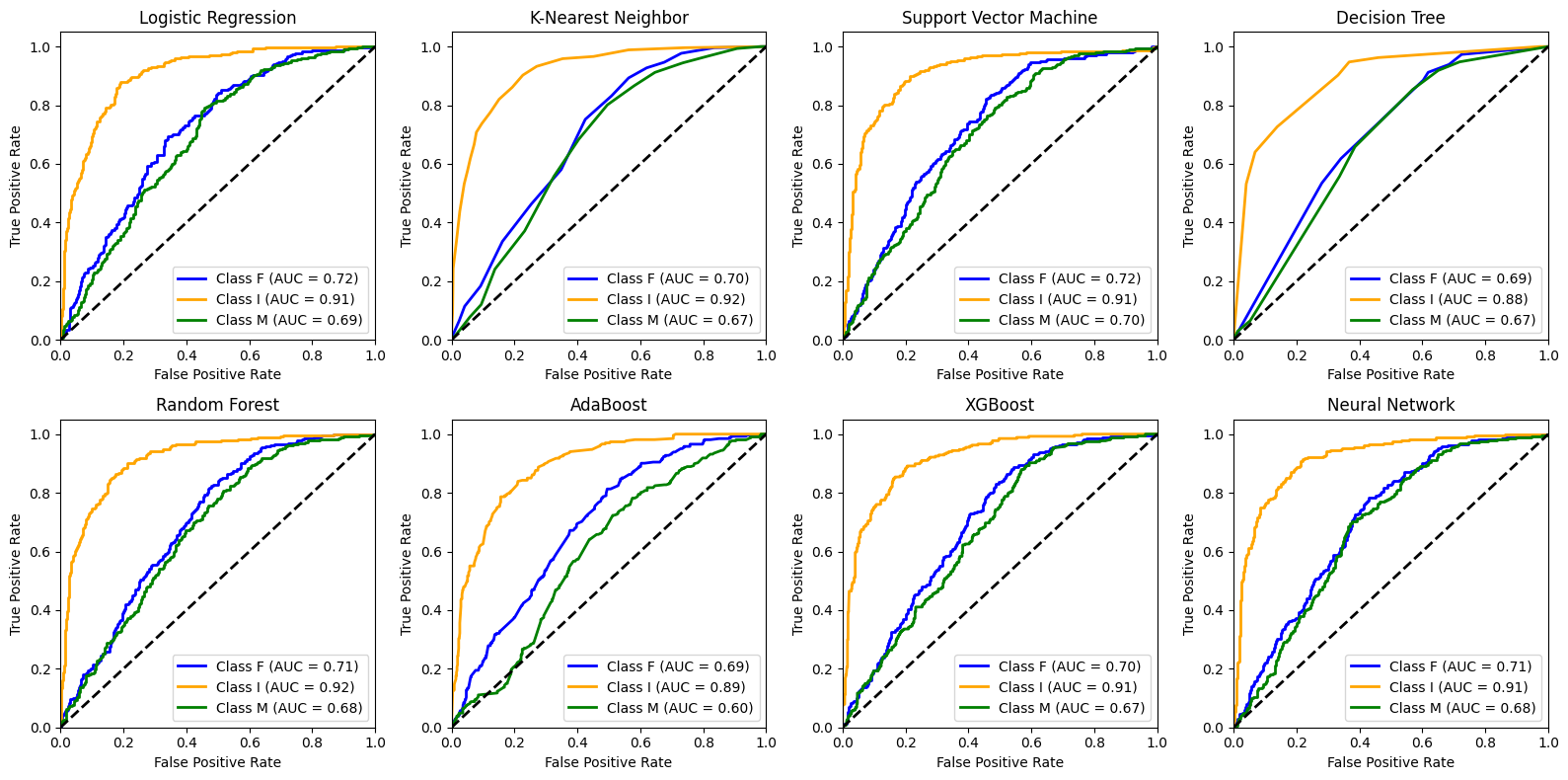
*Decision Tree*



*Neural Network Training Accuracy & Loss*



*ROC Curve*



Feature importance by LIME

