CSI5155 - Fall 2024

Assignment 2 - Explainable AI (XAI)

<Vrishab Davey><300438343>

In explaining the performance of both Random Forest and Decision Tree models, SHAP did a great job when applied to Chocolate and Mushroom datasets. This was great insight into how the decisions were made in understanding why certain predictions were done, why others were not, when the models succeeded or failed, and how we might improve the future predictions.

Chocolate Dataset

Best model:

model parameters for Random Forest Learner: {'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 20, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 20, 'min_weight_fraction_leaf': 0.0, 'monotonic_cst': None, 'n_estimators': 50, 'n_jobs': None, 'oob_score': False, 'random_state': 42, 'verbose': 0, 'warm_start': False}

Accuracy for Test Set (Random Forest Model): 0.7294429708222812

Worst model:

model parameters for Decision Tree Learner: {'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'entropy', 'max_depth': 3, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 6, 'min_samples_split': 10, 'min_weight_fraction_leaf': 0.0, 'monotonic_cst': None, 'random_state': 42, 'splitter': 'best'}

Accuracy for Test Set (Decision Tree Model): 0.7241379310344828

Mushroom Dataset:

Best Model:

model parameters for Random Forest Learner: {'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 5, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 5, 'min_weight_fraction_leaf': 0.0, 'monotonic_cst': None, 'n_estimators': 100, 'n_jobs': None, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}

Accuracy (Best Model): 0.7379095163806553

Worst Model:

Worst model parameters for Decision Tree Learner: {'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'entropy', 'max_depth': 10, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 20, 'min_weight_fraction_leaf': 0.0, 'monotonic_cst': None, 'random_state': None, 'splitter': 'best'}

Accuracy (worst Model): 0.6833073322932918

How and why the algorithms are making certain decisions?

The SHAP analysis portrayed the contributions of features for each of the models. For example, on the Chocolate dataset, Ethnicity, Ascore, Cscore, and Escore were currently the most governing in random forest and decision tree models. The Random Forest model was complex with an accuracy of 72.9% on the test and had a wide range of influential features. It uses 50 estimators and a depth of 20. This contrasts with the model Decision Tree, which, on the far end, relies more on Ethnicity and Ascore in a model that had attained an accuracy of 72.4%, although its max depth is 3; hence, it cannot catch most of the feature interactions given how shallow this depth is. The best model was the Random Forest model obtained in the mushroom dataset which had an accuracy score of 73.8%. The most influential features this model showed were Country, Age, and SS. SHAP values demonstrated that these variables tend to have diverse impacts and show different trends across the samples. This Decision Tree model had a general accuracy of 68.3% and was the worst model. While it reflected a very similar trend that the most crucial features were Countr yand Age but failed to capture complex interactions, overfitting within some categories, this simpler structure in this model made it less adaptive to the dataset's complexity, which can explain its lower accuracy.

Why the Algorithms Didn't Make Other Decisions?

Looking at the SHAP values also revealed why some features had little influence. For instance, Gender and Country(Decision Plot) in the Mushroom Dataset's Decision Tree model had little to no SHAP influence.This means that they contributed less to the decisions the model was making on classification. With that insight comes knowledge of why features, although influential in the Chocolate Dataset Random Forest model, the model did not give them greater prioritization in its predictions. This contrast helps usunderstand how different models and datasets require unique feature considerations for effective prediction.

When the Algorithms Worked/ When They Failed?

The Random Forest model seemed to work quite well on the Chocolate dataset, with high accuracy and a very consistent SHAP value for impactful features indicating that they were able to assess the hidden features pretty well. In contrast, the slightly lower performances of the Decision Tree model were still able to capture some more subtle relationships in the data. The problems were different with the Mushroom dataset, in that these models returned pretty bad accuracies, with a Random Forest at 73% and a DecisionTree at 68%. Further SHAP analysis will show how features such as Country and SS are just not good enough to capture the complex interactons in a way that predicts well. Thus indicating the reality that these models do not generalize across all the patterns in this dataset.

Model Trustworthiness and Room for Improvement

The SHAP plots helped to assess the Trustworthiness of each model. Since it was clear which features were responsible for which predictions and how much, the Random Forest models were more reliable in that aspect as the plots were self-explanatory as to how they were able to capture the intrinsic details. Models of Decision Trees were more interpretable but, in return, limited in reliability due to them being superficial and not being able to capture such intrinsic details. To make the accuracy better of the Decision Tree models, it would be useful to increase the complexity of the trees probably with feature engineering. On the other hand, Random Forests probably can be further tuned w.r.t a number of their parameters over estimators or max depth, for example, to optimally balance interpretability and accuracy.