Machine Learning Project Documentation

Model Refinement

Model Refinement

1. Overview

In this phase, I aimed to refine the initial models used for classifying `irrigation_frequency` by improving the overall accuracy through feature engineering, augmented data, and modifying the target classes. Initial efforts, which included experimenting with various machine learning algorithms and tuning hyperparameters, did not yield satisfactory results.

2. Model Evaluation

The initial models (Random Forest, XGBoost, etc.) returned accuracy scores below 0.2. Despite several iterations and efforts, including feature engineering and dataset manipulation, the accuracy remained consistently low, signaling that the features and data might not be suitable for the expected outcomes.

3. Refinement Techniques

To overcome the challenges faced, I performed feature engineering by adding and dropping features, but these changes didn't significantly impact model performance. As a more aggressive approach, I augmented the dataset and reduced the number of target classes. These changes proved effective, raising the accuracy to 0.80 with a Random Forest classifier and 0.79 with XGBoost.

4. Hyperparameter Tuning

After reducing the target classes and working with augmented data, I fine-tuned the hyperparameters of the Random Forest and XGBoost models. The tuning involved adjusting parameters such as `n_estimators, max_depth`. These refinements contributed to improved model stability and slight boosts in performance

5. Feature Selection

I experimented with various feature selection methods to improve the model's performance. This included feature importance using random forrestbased selection. However, this technique did not significantly impact the accuracy, indicating that the initial feature set may not have been the key limiting factor in the model's performance.

6. Code Implementation

Conclusion

In conclusion, the model refinement phase was crucial in improving performance, particularly through data augmentation and reducing the target classes. Despite initial challenges, these adjustments allowed the Random Forest model to achieve an accuracy of 0.80, while XGBoost closely followed at 0.79. The primary difficulty encountered was aligning the dataset's features with the desired outcomes, which will need further exploration in future work.

References

https://www.youtube.com/@RyanAndMattDataScience