Land Cover Classification

Leila Maritim

Multilabel classification task

Code: https://github.com/LeilaMemoi/AMINI/blob/main/Amini_Task1.ipynb

Observations

Given the nature of the data, the following key observations were made:

- The target labels have **no correlation**.
- The dataset contains features **lat, lon, mlat, mlon, x, and y**, with high correlation among them. To reduce redundancy, only **mlat and mlon** were retained.
- Each location instance can belong to **more than one class**, meaning the classes are **not mutually exclusive**.
    - Among the target labels: **"cropland"** and **"building"** are **binary classifications**.
    - **"wcover"** is a **multiclass classification**.
- These factors indicate that the problem is best approached as a **multi-label/task classification** task.

Decisions Made

1. Model Selection

The selected approach utilizes **Scikit-learn's MultiOutputClassifier**, which extends base estimators to support multi-label classification.

Approach

- Each target variable is modeled **independently**, meaning a separate classifier is trained for each column in Y_train.
- This avoids unnecessary dependencies between target labels, which is ideal since they are **uncorrelated**.
- The **Random Forest** and **XGBoost** classifiers were wrapped in a MultiOutputClassifier, and hyperparameter tuning was conducted to optimize performance.

Alternative Approach: Chain Models

- **Classifier Chains** extend base estimators like MultiOutputClassifier, but each model in the chain uses both X and the predictions from earlier models as inputs.
- This would capture class dependencies, but since the target labels are **not correlated**, it was **not the best choice** for this task.

2. Evaluation Metrics

Selecting a metric that effectively evaluates **multi-label classification** was crucial.

- The chosen metric is **Hamming Loss**, which measures the fraction of incorrectly predicted labels out of the total labels:
- **Why Hamming Loss?**
    - It accounts for **both** false positives and false negatives.
    - It is well-suited for **multi-label classification** where each instance can belong to multiple categories.

By leveraging MultiOutputClassifier and optimizing for **Hamming Loss**, the approach ensures a robust model for the classification task.

Findings and recommendation

Fine-Tuned Results

Through **hyperparameter tuning**, a slight improvement in the models' performance was observed, with **XGBoost** achieving a marginally lower Hamming Loss compared to **Random Forest**. Due to its slightly better performance, **XGBoost** was selected as the best model to predict the class probabilities in the test dataset.

Random Forest Hamming Loss: 0.2105

XGBoost Hamming Loss: 0.2035

Further Refinements

Although the current models are performing reasonably well, there are several avenues for further improvement:

1. **Other Evaluation Metrics**

While **Hamming Loss** is a useful metric for multi-label classification, exploring other metrics that are more sensitive to classification errors could provide additional insights.

- **Subset Accuracy** could be a more stringent metric for evaluating the performance, as it measures the proportion of instances where all labels are correctly predicted.
- **F1-Score** or **Precision-Recall** for each label can also be explored to assess the models' balance between precision and recall.

2. **Exploring Multi-Layer Perceptron (MLP)**

An alternative approach to consider is using a **Multi-Layer Perceptron (MLP)**, which has native capabilities for multi-label classification. MLPs can capture complex relationships between input features and target labels and could potentially offer better performance for this multi-label task.

Fine-tuning an MLP model could yield further improvements, especially if combined with regularization techniques to prevent overfitting.

Conclusion

Although **XGBoost** currently provides the best results in terms of **Hamming Loss**, exploring additional metrics and model architectures like MLP could help refine the classification performance even further. Incorporating these refinements could potentially enhance the overall accuracy and robustness of the multi-label classification system.