Sina Negarandeh | 300395579
Bahar Emami Afshar | 300377401
Machine Learning - CSI5155
Winter 2024

Semi-Supervised Learning, Label Scarcity, and Class Imbalance
Final Project Report

Data Preprocessing:

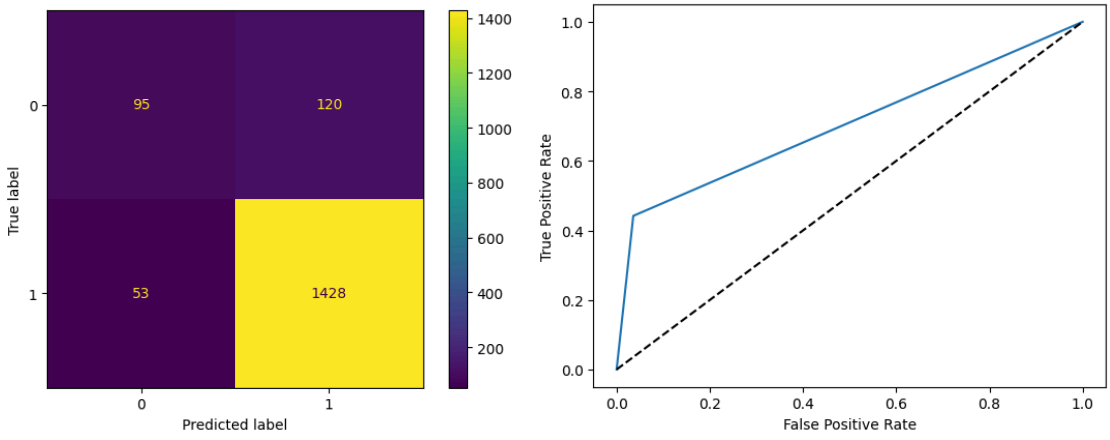
For the preprocessing phase, we conducted a thorough analysis of the dataset, including examining the data structure, feature formats, and distributions, and identifying any missing values. Following this, a series of preprocessing techniques were applied, including dropping and imputing missing values, encoding categorical features via hashing, binarization of the target column, and normalizing numerical values. Subsequently, using the importance values calculated by XGBoost, we conducted feature selection, eliminating those features that had negligible impact on the predictions.

Before algorithm training, we divided the data into labeled training, unlabeled pool, and testing sets for evaluation purposes. Our framework offers flexibility, allowing us to adjust the ratio of labeled to unlabeled data as requested in the requirements of the project description.

Supervised Learning:

In the supervised learning phase, we initially partitioned the data into a 75% training set and a 25% test set. Following the application of the mentioned preprocessing techniques, we further balanced the training set using the BorderlineSMOTE oversampling algorithm. Finally, we applied XGBoost and SVM classifiers with a 5-fold cross-validation strategy, optimizing its hyperparameters through Bayesian search against the balanced preprocessed dataset.

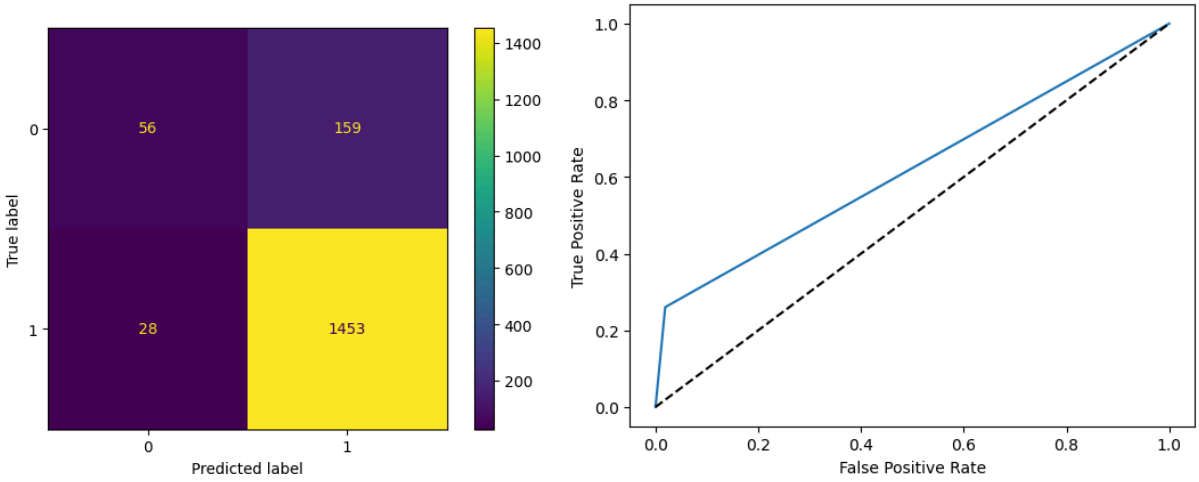
For evaluation, we visualized the ROC curves and computed performance metrics. The outcomes of our supervised learning phase are summarized below.



XGBoost Supervised Learning - Balanced Data a) Confusion Matrix b) ROC curve

```
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.96      0.94      1481
           1       0.64      0.44      0.52       215

    accuracy                           0.90      1696
   macro avg       0.78      0.70      0.73      1696
weighted avg       0.89      0.90      0.89      1696


Accuracy: 0.90
Micro F1 score: 0.90
Macro F1 score: 0.73
```

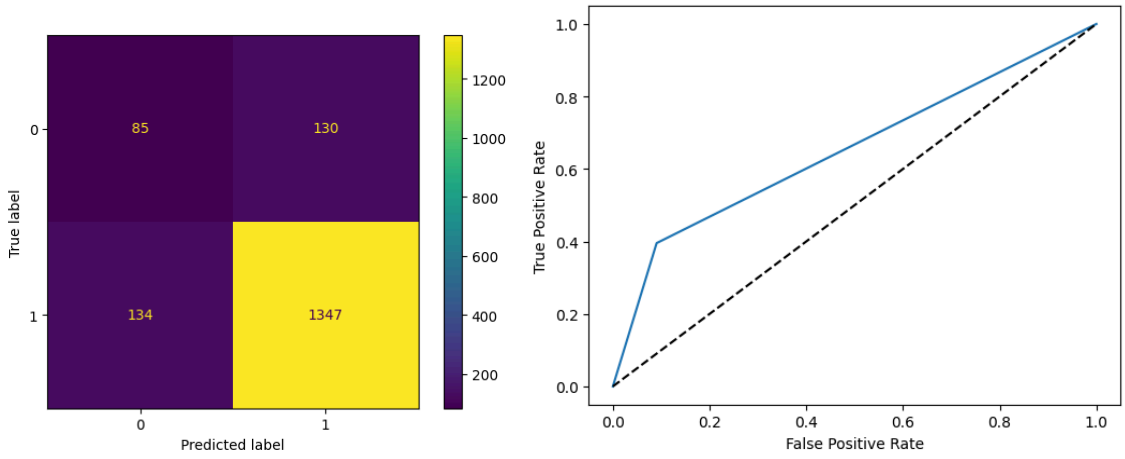
XGBoost Supervised Learning - Balanced Data Classification Report



XGBoost Supervised Learning - Original Imbalanced Data a) Confusion Matrix b) ROC curve

```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.98      0.94      1481
           1       0.67      0.26      0.37       215

    accuracy                           0.89      1696
   macro avg       0.78      0.62      0.66      1696
weighted avg       0.87      0.89      0.87      1696


Accuracy: 0.89
Micro F1 score: 0.89
Macro F1 score: 0.66
```

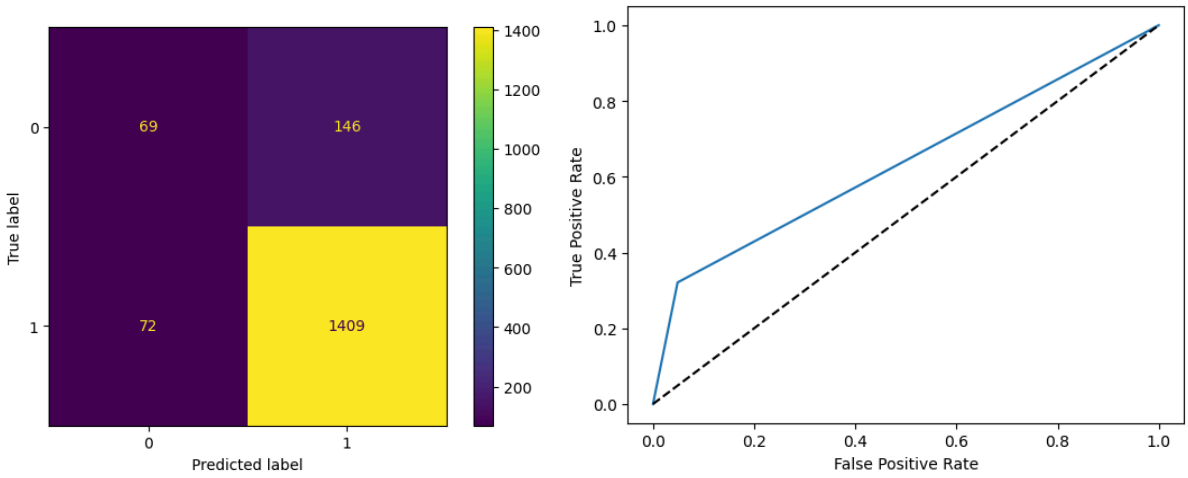
XGBoost Supervised Learning - Original Imbalanced Data Classification Report



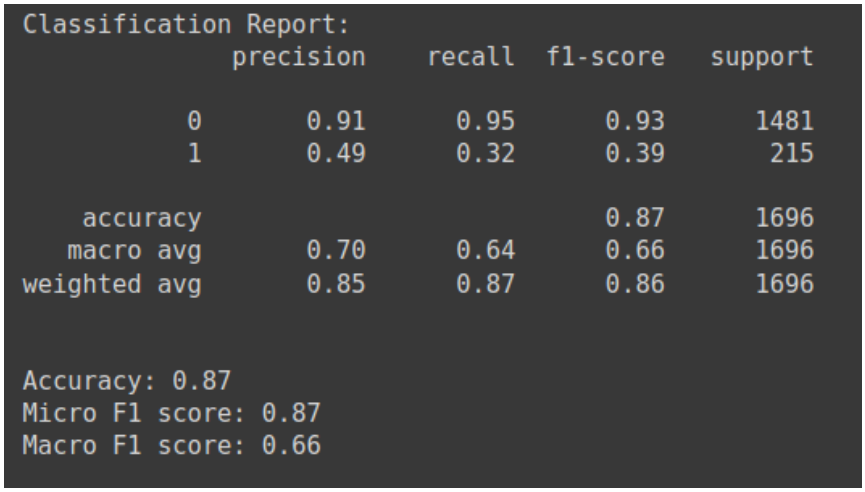
SVM Supervised Learning - Balanced Data a) Confusion Matrix b) ROC curve

```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.91      0.91      1481
           1       0.39      0.40      0.39       215

    accuracy                           0.84      1696
   macro avg       0.65      0.65      0.65      1696
weighted avg       0.85      0.84      0.84      1696


Accuracy: 0.84
Micro F1 score: 0.84
Macro F1 score: 0.65
```

SVM Supervised Learning - Balanced Data Classification Report



SVM Supervised Learning - Original Imbalanced Data a) Confusion Matrix b) ROC curve



```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.95      0.93      1481
           1       0.49      0.32      0.39       215

    accuracy                           0.87      1696
   macro avg       0.70      0.64      0.66      1696
weighted avg       0.85      0.87      0.86      1696


Accuracy: 0.87
Micro F1 score: 0.87
Macro F1 score: 0.66
```

SVM Supervised Learning - Original Imbalanced Data Classification Report

Semi-Supervised Learning:

In this project, we implemented different algorithms from Inductive Semi-Supervised Classification methods, such as Self-Training, Co-Training, SemiBoost (from Wrapper methods), and Label Propagation (from Manifold Learning). Wrapper methods involve taking a Supervised Learning model and utilizing it to learn from data that already has labels. Subsequently, the trained model is employed to predict labels for unlabeled data, resulting in pseudo labels. A selection criterion, such as highest confidence, is then applied to incorporate some pseudo labels into the labeled data. Afterward, the model is retrained using both labeled data and pseudo labels. For the implementation of Self-Training, we utilized the Self-training classifier from the scikit-learn library, with XGBoost employed as the Supervised Learning model. For Co-Training and SemiBoost, we employed the LAMDA-SSL library, which offers a comprehensive and user-friendly toolkit containing 30 Semi-Supervised Learning algorithms. Co-Training involves the collaboration of two Supervised learners to enhance each other's training. In our implementation, we utilized XGBoost and SVM as the Supervised learners. Ensemble learning involves combining multiple learners, and in this project, we utilized SemiBoost as the Semi-Supervised Learning Ensemble Algorithm, with SVM serving as the base learner. Finally, within the scikit-learn library, there are two methods for Manifold Learning: Label Propagation and Label Spreading. We implemented Label Propagation in our project.

Evaluation and Results:

For the evaluation of these algorithms, we considered both the level of unlabelled data and class imbalance. Each algorithm is tested with both balanced and imbalanced data. Additionally, each Semi-Supervised model is tested with 5 levels of unlabelled data: 50%, 75%, 90%, 95%, and 99%. For each test, the classification report is displayed, which includes accuracy and F1-measure, along with the classification matrix, the runtime for training, and the ROC curve. After gathering all of these test results, we employ the Friedman test and the Nemenyi post-hoc test to determine whether there are any statistical differences between the performances of the algorithms. The evaluations of the Supervised Learning models, XGBoost and SVM, are provided in the Supervised Learning section.

Balanced Data						
	Supervised Learning (XGBoost)	Supervised Learning (SVM)	Semi-Supervised Learning Self-Training Algorithm (XGBoost)	Semi-Supervised Learning Co-Training Algorithm (XGBoost and SVM)	Semi-Supervised Learning Ensemble Algorithm (SemiBoost)	Intrinsically Semi-Supervised Learning Algorithm (Label Propagation)
0% Unlabelled Data	Accuracy: 0.90 Macro F1 score: 0.73 Training runtime: 1.00 seconds	Accuracy: 0.84 Macro F1 score: 0.65 Training runtime: 18.31 seconds	-	-	-	-
50% Unlabelled Data	-	-	Accuracy: 0.88 Macro F1 score: 0.67 Training runtime: 11.22 seconds	Accuracy: 0.82 Macro F1 score: 0.68 Training runtime: 1134.76 seconds	Accuracy: 0.82 Macro F1 score: 0.66 Training runtime: 1878.64 seconds	Accuracy: 0.84 Macro F1 score: 0.65 Training runtime: 1420.60 seconds
75% Unlabelled Data	-	-	Accuracy: 0.90 Macro F1 score: 0.62 Training runtime: 8.62 seconds	Accuracy: 0.62 Macro F1 score: 0.52 Training runtime: 951.38 seconds	Accuracy: 0.84 Macro F1 score: 0.63 Training runtime: 1240.38 seconds	Accuracy: 0.85 Macro F1 score: 0.64 Training runtime: 307.92 seconds
90% Unlabelled Data	-	-	Accuracy: 0.88 Macro F1 score: 0.61 Training runtime: 4.44 seconds	Accuracy: 0.26 Macro F1 score: 0.26 Training runtime: 329.12 seconds	Accuracy: 0.85 Macro F1 score: 0.59 And Training runtime: 768.09 seconds	Accuracy: 0.83 Macro F1 score: 0.62 Training runtime: 244.02 seconds
95% Unlabelled Data	-	-	Accuracy: 0.88 Macro F1 score: 0.57 Training runtime: 3.99 seconds	Accuracy: 0.18 Macro F1 score: 0.17 Training runtime: 212.00 seconds	Accuracy: 0.87 Macro F1 score: 0.55 Training runtime: 312.17 seconds	Accuracy: 0.82 Macro F1 score: 0.60 Training runtime: 224.76 seconds
99% Unlabelled Data	-	-	Accuracy: 0.84 Macro F1 score: 0.58 Training runtime: 4.41 seconds	Accuracy: 0.15 Macro F1 score: 0.14 Training runtime: 128.56 seconds	Accuracy: 0.87 Macro F1 score: 0.52 Training runtime: 244.84 seconds	Accuracy: 0.76 Macro F1 score: 0.57 Training runtime: 195.44 seconds
Performance Comparison of Semi-Supervised Learning Algorithms Friedman Test: statistic: 4.92 P-value: 0.18 There are no statistical differences between the performances of the Semi-Supervised Learning algorithms						

The table shown above summarizes the performance of each algorithm trained on balanced data with different levels of unlabeled data. The summarization includes accuracy, Macro F1 Score, and training runtime for each algorithm, as well as the result of the Friedman test and the Nemenyi post-hoc test for comparing all algorithms. However, the detailed results of the evaluation for each algorithm are included in the notebook, which contains information such as training runtime, confusion matrix, classification report, and ROC curve. Additionally, the table above only shows the results of evaluating algorithms trained on balanced data. In the notebook, you can find the results of evaluating algorithms trained on imbalanced data as well.

Significant Observations:

- As the percentage of unlabeled data increases, the performance metrics such as F1 score and Accuracy decrease, which is expected.

- Between all of the semi-supervised learning approaches, with 50% of unlabeled data Co-training outperforms other algorithms with an F1 Score of 68%. However as the imbalance ratio increases, Co-training performance significantly drops and achieves the lowest performance among others. This can be improved by using different classifiers and adjusting multiple views of the dataset to the algorithms.
- Label Propagation (Manifold) approach showed the most consistent performance across various imbalance ratios.
- Among all semi-supervised learning approaches, Self-training proved to be more applicable to real-world problems, as other approaches are intensively time-consuming. Self-training gains a similar level of accuracy in a notably shorter time.

Lessons Learned:

- In conclusion, Semi-Supervised learning is a great approach when dealing with sparse labeled data, we can still achieve desirable performance even with limited labeled data. This can be useful in many real-world problems when the process of labeling is time-consuming and resource-intensive. While acquiring data may be straightforward and cost-free in many instances, the expertise required for labeling is often difficult to find. This is where Semi-Supervised approaches offer a crucial advantage.
- We gained valuable insights and a deeper understanding of algorithms by comparing and implementing various Semi-Supervised approaches. We learned about the strengths and weaknesses of each of the algorithms, as well as ways to improve them. This knowledge helps us in selecting the most suitable semi-supervised approach for similar future problems.
- Addressing imbalanced data while implementing Semi-Supervised approaches with limited labeled data posed another challenge, one which we effectively navigated using the practices learned from Assignment 1.
- This assignment offers great insights into the interplay between imbalanced data, varying levels of unlabelled data, and diverse semi-supervised algorithms.