**Credit Card Fraud Detection using Stacked Ensemble**

**Project Report**

|  |  |
| --- | --- |
| Submitted by: Keyrun Adhikari | Submitted to: Dr. Keith Dillon |

**Introduction**

Credit Card Fraud Detection using Stacked Ensemble uses multiple Machine Learning supervised classifiers as base models and uses predictions from them as features along with the original features to predict credit card fraud transactions using ensemble stacked model.

**Dataset**

Dataset used in this project is Credit Card Fraud Detection dataset collected and analyzed during a research collaboration of Worldline and the Machine Learning Group ([http://mlg.ulb.ac.be](http://mlg.ulb.ac.be/)) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection [1]. It can be found on Kaggle [here](https://www.kaggle.com/mlg-ulb/creditcardfraud).

The dataset class is labeled as 1 for fraudulent and 0 for genuine, and is highly imbalanced, that is the positives (fraud) amount for 0.172% of 284,807 transactions. There are 30 features and 1 target class column in the dataset among which 28 features contain numerical values obtained after PCA transformation. The original feature names are changed to V1, V2, V3, ...., V28 due to confidentiality issues. The only features that are not transformed using PCA are Amount and Time [2]. The 67% of the dataset is used for training and remaining as test set.

**Approach Used**

4 supervised classifiers were stacked to original features to train an ensemble model in a 2-level architecture.

**Level 1**

The level 1 consisted of following base models.

* KNeighborsClassifier
* RandomForestClassifier
* XGBClassifier
* LogisticRegression

All the base models were tuned using RandomizedSearchCV or GridSearchCV methods with 5-fold cross-validation. Both training and test datasets were trained with best models after hyperparameter tuning to get predictions for stacked model using 5-fold cross-validation.

**Level 2:**

In level 2, XGBClassifier was trained using the stacked predictions along with the original features.



Figure 1: Stacked Ensemble Model

**Cross-validation and Hyperparameter Tuning**

5-fold cross-validation was used for validation.

Two methods RandomizedSearchCV or GridSearchCV were used for hyperparameter tuning. Both of these methods are passed with model to tune, parameters grid, scoring metric, and cross-validation fold. Then, the best set of hyperparameters from the parameters grid for all the models based on F1-score using 5-fold cross-validation were found using best\_estimator\_ attribute of the method.

The difference between two methods is GridSearchCV runs the model with all the combinations of the parameters from parameters grid and can be used for models where we only have one or few hyperparameters to tune as well as are faster to train, while RandomizedSearchCV runs the model with random combinations of the parameters and can be used for models where we have more hyperparameters to tune as well as are slower to train.

**Results**

Following scores were achieved with base models and ensemble stacked model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifiers**  **Scores** | **KNN** | **Random Forest** | **XGBoost** | **Logistic Regression** | **Ensemble XGBoost** |
| **Precision** | 0.4838 | 0.9426 | 0.9531 | 0.8761 | 0.9461 |
| **Recall** | 0.2013 | 0.7718 | 0.8187 | 0.6174 | 0.8255 |
| **F1** | 0.2843 | 0.8487 | 0.8808 | 0.7244 | 0.8817 |
| **TN** | 93806 | 93831 | 93832 | 93825 | 93831 |
| **FP** | 32 | 7 | 6 | 13 | 7 |
| **FN** | 119 | 34 | 27 | 57 | 26 |
| **TP** | 30 | 115 | 122 | 92 | 123 |

Table 1: Scores

**Conclusion**

This project gave me idea on how to deal with imbalanced datasets as well use hyperparameter tuning methods. Also, I was able to learn how stacked ensemble models work and how to implement them. Both hyperparameter tuning methods GridSearchCV and RandomizedSearchCV took significant time to run for models with multiple hyperparameters to tune. Among the base models, XGBoost performed better than other models, however stacked ensemble XGBoost performed slightly better than base XGBoost classifier. Also, the ensemble XGBoost detected the highest number of true positives, that is most number of fraud transactions from test set among all the models. F1-score gives better evaluation for imbalanced datasets as it is the harmonic average of precision and recall.

**References**

1. Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015
2. Machine Learning Group, ULB. Credit Card Fraud Detection. https://www.kaggle.com/mlg-ulb/creditcardfraud/home (accessed 13, November 2018)