Comp 4449 Data Science Capstone October 13, 2021

Paul O'Leary Midterm Project

Credit Card Fraud Detection

Purpose: More and more of us have received notifications from our credit card companies asking if we purchased gas at a station in Appleton Wisconsin at 3:30 AM, while we were safely in our beds a thousand miles away. According to the Federal Trade Commission, 390,000 reports of credit card fraud were made in 2020, accounting for a third of Identity Theft cases that year. The median cost to the consumer affected by fraud was $311 (FTC), with total losses in the United States of around $10 Billion (Nilson Report). Because credit card companies generally do not make consumers responsible for fraudulent charges, they work hard to prevent fraud from happening at all, and when it does happen, to catch it quickly. This project will attempt to identify fraudulent credit card charges from sample data representing charges made during two days in 2013.

Data Description: The dataset was collected during a research collaboration of Worldline and the Machine Learning Group of Université Libre de Bruxelles. (For full citation please see "Acknowledgements") The data is available via Kaggle at this link: https://www.kaggle.com/mlg-ulb/creditcardfraud, and consists of 284,807 credit card charges made over two days by European cardholders in September of 2013. Due to confidentiality issues, the data has been transformed by Principal Component Analysis, and consists of strictly numeric values. The following fields are present in the data:

"Time":	The elapsed time since the beginning of that day.
"V1" through "V28":	The PCA transformed and scaled data.
"Amount":	The amount of the charge in some monetary unit
"Class":	The target variable, with "0" indicating a non-fraudulent charge and "1" indicating a fraudulent charge.

The dataset is severely imbalanced – only 492 records of the 284,807 total represent fraudulent charges, which is 0.173% of the records in the dataset. Many classification techniques executed with the data in its present state will simply choose the majority class, earn an Accuracy Score of 99+%, and consider the job well done. Alternate methods to prepare the data and create the classification model will be investigated.

Data Preparation: Standard examination of the data was performed, with basic statistical results and histograms for all the fields. Correlation was examined, but because the data has been PCA transformed, all fields were kept throughout the analysis. The decision was made to drop the “Time” column from further analysis. The “Amount” field was scaled using Sklearn’s Standard Scaler, to level that value with the scaled versions of fields “V1” through “V28”.

The majority of the data preparation involved methods to try to balance the data – most classification algorithms expect somewhat balanced data to produce acceptable models. For this project, the following balancing techniques were investigated:

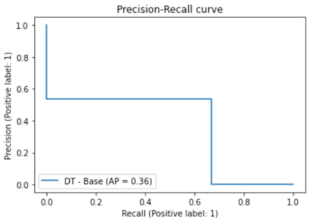
- Generic Under-sampling – records were sampled randomly from the majority class to match the number of records in the minority class.
- Generic Over-sampling – records were repeatedly sampled randomly from the minority class to match the number of records in the majority class.
- SMOTE – Synthetic Minority Oversampling TEchnique. SMOTE synthesizes new examples of the minority class by using a form of ‘nearest neighbor’. SMOTE chooses a member of the minority class, finds the nearest neighbor, and creates a new data point on the line between the two chosen points. This new point is labeled as a member of the minority class.
- SMOTE with Tomek Links – After SMOTE creates new members of the minority class, Tomek Links samples members of the majority class and removes those points if the nearest neighbor is a member of the minority class.

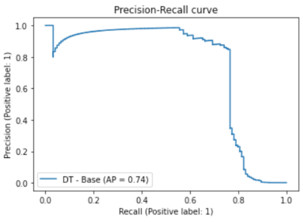
<u>Results:</u> Full results are available in the accompanying Jupyter Notebook.

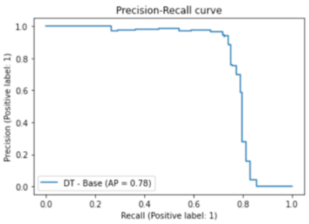
The various levels of unbalanced and balanced data were classified using a Decision Tree classifier, a Support Vector Classifier and a Random Forest Classifier.

Classification with the Original Files:

	Decision Tree	SVC	Random Forest
Accuracy	0.999	0.999	0.9995
F1	0.595	0.678	0.790
AUPRC	0.359	0.516	0.646



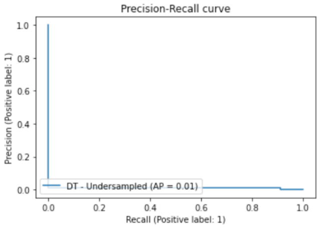


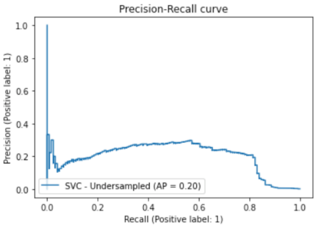


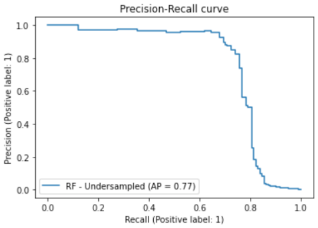
The original files classified surprisingly well, especially by the Random Forest Classifier, with 82/124 fraudulent charges classified correctly, but 3 were falsely called fraud, while 42 were fraudulent but labeled as non-fraudulent.

Generic Under-sampled data (majority class records matched to the number of minority class records):

	Decision Tree	SVC	Random Forest
Accuracy	0.870	0.949	0.983
F1	0.018	0.043	0.114
AUPRC	0.008	0.020	0.053



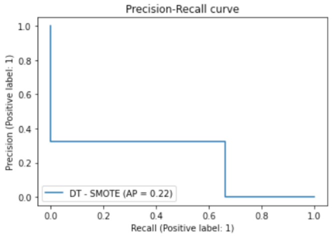


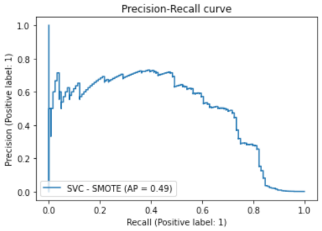


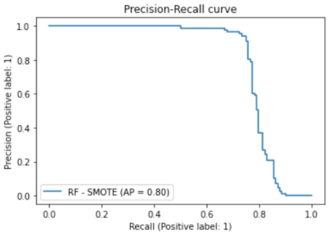
As expected, the reduced number of records in the majority class make this the weakest performing classification attempt.

For SMOTE balanced data:

	Decision Tree	SVC	Random Forest
Accuracy	0.998	0.999	0.9996
F1	0.437	0.578	0.834
AUPRC	0.216	0.338	0.705







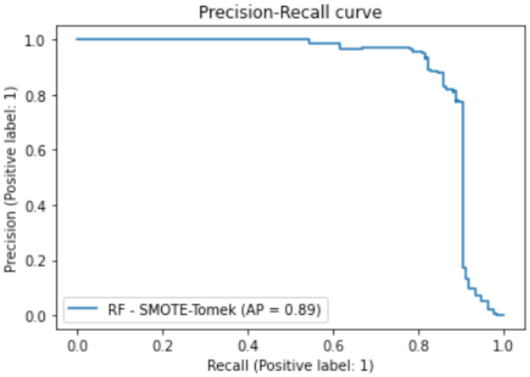
SMOTE improves on all previous attempts at classification, with 93/124 fraudulent records classified correctly, with 6 falsely called fraud, and 31 falsely labeled as non-fraudulent.

For the final full classification attempt, SMOTE with Tomek Links was run on a train set of 70% of the data, with the remaining 30% in the test set. The classification was done by a tuned Random Forest Classifier:

	Random Forest
Accuracy	0.9995
F1	0.854
AUPRC	0.730

Confusion Matrix:

```
        [[85286  21]
         [ 19    117]]
```



The tuned Random Forest Classifier, with SMOTE – Tomek Links balanced data performed the best, with 117/136 fraudulent charges correctly predicted, with 21 falsely called fraud, and 19 falsely labeled as non-fraudulent.

Conclusions.

Each step taken with more sophisticated methods to balance the data incrementally improved the classification algorithm's ability to create a predictive model. As expected, the generic under-sampling approach loses too much information about the majority class, while generic over-sampling does not enhance the information about the minority class. The SMOTE and SMOTE with Tomek Links performed best, with the high cost of processing power and time.

On all the created datasets, the Random Forest Classifier created the best models, with less processing power and time required, than the Support Vector Classifier specifically. The final tuned model earned an AUPRC Score of 73%, correctly labeling 117 out of 136 fraudulent charges, with 40 mislabeled.

While the results are promising, other methods of detecting fraudulent credit card charges should be pursued. Initial testing of Outlier Detection algorithms using unsupervised learning methods, such as OneClassSVM and Isolation Forests was begun, but more work is needed.

Notes: The processing power required to run some of the balancing and classification tasks undertaken in this project heavily taxed this researcher's Macbook Pro, and greatly increased the time necessary for multiple iterations. The PCA transformed data, with-non descriptive labels, removed all context from initial data analysis, and made it, frankly, quite dull.

Acknowledgements

https://www.kaggle.com/mlg-ulb/creditcardfraud

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on https://www.researchgate.net/project/Fraud-detection-5 and the page of the DefeatFraud project

Please cite the following works:

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

Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon

Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE

Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)

Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier

Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing

Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi Information Sciences, 2019

Yann-Aël Le Borgne, Gianluca Bontempi Machine Learning for Credit Card Fraud Detection - Practical Handbook