Credit Card Fraud Detection Under Extreme Class Imbalance

Monisha Vodnala[1], Fairy Dmonte[2], Harshitha Sanikommu[3]
INFO 6105 Data Science Engineering Methods and Tools
Spring 2019 2
Professor Nik Bear Brown

***Abstract*—Credit card is one of Indians fastest growing trend. Online payment doesn't require physical card. Anyone having details of the card can access it. This has compelled the financial organizations to implement and continuously improve their fraud detection system. However, credit card fraud dataset is highly imbalanced and different types of misclassification errors may have different costs and it is essential to control them, to a certain degree, to compromise those errors. Classification techniques are the promising solutions to detect the fraud and non-fraud transactions. Unfortunately, in certain condition, such as imbalance data, classification techniques do not perform well when it comes to huge numbers of differences in minority and majority cases. In this study, we focus on resampling methods, Random Under Sampling, Random Over Sampling, Synthetic Minority Oversampling Technique, Hybrid Sampling were applied to overcome the rare events in the dataset. Then, the three resampled datasets were classified using classification techniques. The performances are measured by their accuracy, precision, Recall, ROC-AUC curve (AUC). The findings disclosed that by resampling the dataset, the models were more practicable, gave better performance and were statistically better.**

Keywords—Credit card; imbalanced dataset; misclassification error; resampling methods; random under sampling; random oversampling; synthetic minority oversampling technique

1. Introduction

Businesses migrating and evolving to online business and money is managed electronically in an ever-growing cashless banking economy and cash usage was gradually replaced by credit card. [1]. Credit payment has been the most popular mode of payment ever since. Based on [2], credit payment purchases are classified into two: i) physical card purchase and ii) virtual card purchase. For payments on online purchases categorized as virtual card purchases needed information such as card numbers, expiration data, and secure codes. As the number of credit card user's increase, fraudulent transactions have been increased too. In the article [3] stated that, identifying and locating the fraudsters is difficult since they are hidden behind the internet. People who face frauds, bear all the cost including various fee like card issuer fee, charges and administrative charge. [4]. Therefore, it is necessary we have an effective credit card fraud detection system.

Prior to proceeding with fraud detection system, we need to understand that the credit card dataset is huge. Usually, the number of frauds in real dataset is very rare as compared to genuine transaction. With such imbalanced data, the performance degrades in terms of accuracy. If ill legitimate transaction is classified as legitimate or visa-versa it will affect customers which will indirectly affect the financial institution. [8, 9, 10]. Maes (2002) have provided some capacity that a fraud detection system should have in order to perform a good result [11]. The system should be able to: i) handle imbalance class, ii) handle noise, iii) avoid the overlapping data iv) adapt themselves to new kinds of frauds, v) evaluate the classifier using good metrics, and vi) detect the behaviour of the frauds. Recent research in [12] stated the major challenges to build a fraud detection System i) the data distribution evolves over time because of seasonality and new attack strategies ii) fraudulent transactions represent only a very small fraction of all the daily transactions and iii) the fraud detection problem is intrinsically a sequential classification task.

In 2017, Haixian et al. stated that identifying a rare event is difficult due to its less frequency and misclassification results in heavy cost. The 3 main solutions identified by review paper are resampling, cost-sensitive learning, and ensemble methods [13]. The most popular one is Resampling. It's resampling method balances the dataset to alleviate the skewed dataset problem. The second one i.e. cost-sensitive Learning can be incorporated to data level and algorithmic level. The third method i.e. ensemble method improves the performance of a single classifier that outperforms. The review paper target two approaches that should be performed to deal with imbalance data i) Using under-sampling, over-sampling or combination of both i.e. Hybrid method to balance the data. ii) Modifying the classification methods or optimizing the performance of learning algorithm. Both the papers give us some insight on which methods are more efficient and are commonly used in imbalance data problems

Burez (2009) applied several methods i.e. evaluation metrics, ii) cost-sensitive learning, ii) resampling methods and iv) boosting to handle the imbalanced class. ROC being his evaluation metrics for analysis and stochastic gradient boosting learner for boosting. For cost-sensitive learning, he used random forest and random under sampling as for resampling methods [15]. The study in [16] proposed an efficient resampling method and obtained comparable classification results between random under sampling and random over sampling. The experiments were carried out using four large imbalanced Bioinformatics datasets. They have recommended 100%- under(0.75)-over method for obtaining comparable classification results to the over sampling results. In 2002,

[17] has proposed Wrapper-based Random Oversampling (WRO) to handle class imbalanced problem. Wrapper is a pre-processing method that incorporates the classifier output to guide pre-processing. The method oversampled the minority class data randomly and the classifier is optimised. They evaluated the WRO with real dataset that they obtained from UCI repository. WRO has better results in most experiment compared to Synthetic Minority Over Sampling Techniques (SMOTE) and random over sampling. Research in [18] investigated the resampling methods specifically on data from Spotify users. They used the most common oversampling methods: random oversampling and SMOTE, and the most common under sampling method: random under sampling. Yan and Han (2018) proposed RE-sample and Cost-Sensitive Stacked Generalisation (RECSG) based on 2-layer learning models to solve the imbalanced problem in 18 benchmark datasets [19]. The experimental results and statistical tests showed that the RECSG approach improved the classification performance.

In reviewing the literature, resampling methods is the main focus of this study due to its simplicity and compatibility with existing classification models to handle the rarity event in massive credit card dataset. There is no research yet were found on the association between credit card fraud and resampling methods. Therefore, the aim of this study is to investigate the classification models' ability to classify the fraud and non-fraud transactions, and to examine if the different resampling methods could improve the performances of the models. The research methodology of the study is conducted in Section 2. Thereafter in Section 3, the experimental setup is described. Next, the results and discussions are presented in Section 4. This study ends with conclusion remarks and future works in Section 5.

2. Research Methodology

This section gives brief description of the methodology of this study. In addition, this section also discusses each step of the methodology. Fig. 1 displays the framework of research methodology of this study.

Step 1: We identified the imbalance credit card dataset which was publicly available on Kaggle for our research.

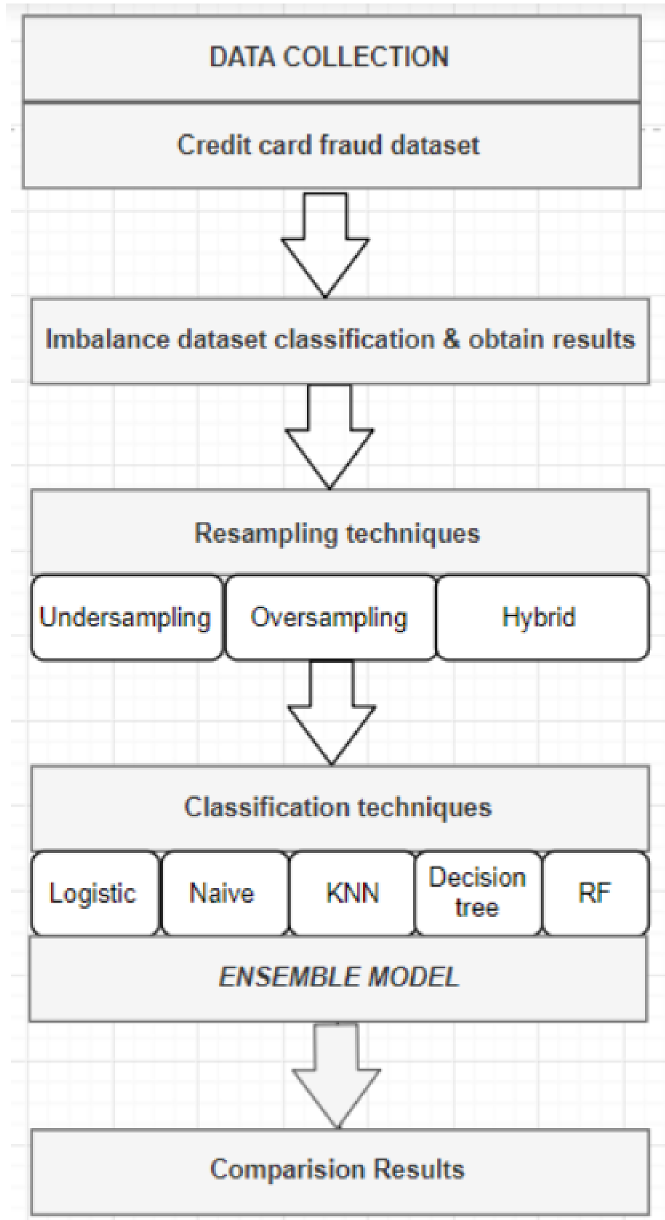
Step 2: We applied classification techniques and checked the results on the Imbalanced dataset.

Step 3: We applied under sampling techniques and checked the results.

Step 4: We repeated the process for oversampling, Hybrid techniques and Ensemble model and compared the results.

Step 5: We compared the overall results and mentioned the conclusions respectively

For this research we have divided the dataset into test and train by 33% and 66% respectively. Future scope we can also change the sample percentages of division of data and analyze the results.



A. *Data Collection*

Getting a real-life data for fraud detection is a real challenge due to data sensitivity and privacy issue. We are using a publicly available dataset downloaded from Kaggle which transaction is basically made in September 2013 by European Cardholders. It has a total of 284807 transactions and 492 fraud transactions. Fraud detection accounts for 0.1792% of all transaction which is very low. This clearly states how imbalanced our data is.

B. Classification Techniques

This dataset falls under the category of binary classification. Either the transaction is classified as genuine (0) or fraud (1). We first resample the data , then train our models using different classifiers to evaluate various methods. In this study, six different classification techniques were explored: Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), Decision Tree(DT), KNN and Ensemble Learning. A summary of the strength and limitations of the classifiers used in this study is given in Table II.

TABLE I. COMPARISON OF RESAMPLING METHODS

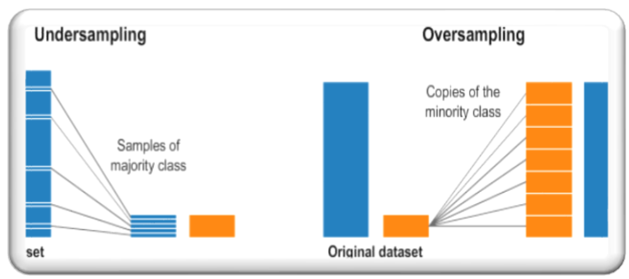
	RUS	**ROS**	**Hybrid Sampling**
Process	Randomly discards the data to shrink the majority class.	Duplicates the data randomly to expand minority class.	Uses combination of ROS and RUS to create a balance dataset.
Strength	Less convergenc e times	Doesn't lose any information and can produce better results.	Effectively improves the classification accuracy of minority class.
Limitation	Discarding data randomly can result into data loss.	Multiple tied instances can result in overfitting.	Can reduced the performance of classification due to data synthetic

TABLE II. STRENGTH AND LIMITATION OF THE CLASSIFICATION TECHNIQUES

Classifier	**Strength**	**Limitation**
Naïve Bayes	Doesn't need a lot of training data and it's insensitive to irrelevant features.	Makes strong impression on the shape od the data distribution.
Logistic Regression	It is a classification model, and also gives us probabilities.	Interpretation is difficult since interpretation of the weights is multiplicative and not additive.
Random Forest	Have high variance and maintains accuracy even when large portion of the data is missing.	It is complex, hard and time-consuming to construct all the trees
KNN	The K-Nearest Neighbor (KNN) Classifier is a very simple classifier that works well on basic recognition problems.	k-NN doesn't perform well on imbalanced data. If we consider two classes, A and B, and the majority of the training data is labeled as A, then the model will ultimately give a lot of preference to A. This might result in getting the less common class B wrongly classified.
Decision Tress	A major decision tree analysis advantages is its ability to assign specific values to problem, decisions, and outcomes of each decision.	Provide less information on the relationship between the predictors and the response. Biased toward predictors with more variance or levels.
Ensembled Model	Ensemble of learners typically outperforms any one learner	Difficult to measure correlation between classifiers from different types of learners. Learning time and memory constraints

C. Resampling Methods

Three widely-used methods for resampling in this study are Under Sampling, Over Sampling and Hybrid sampling. For under sampling, RUS, Cluster Centroid and AllKNN is chosen. For over sampling ROS, SMOTE and ADASYN were chosen as oversampling methods because of its widely usage. Furthermore, ROS is an intuitive way of balancing data, whereas SMOTE is more complex creating synthetic samples using K-Nearest Neighbor (KNN). For Hybrid sampling we used the Techniques SMOTEEN and SMOTE Tomek.



I. UNDERSAMPLING METHODS

This method reduces the majority class by randomly selecting observations from the predominant class or informative where an algorithm processes the data to extract observations without losing valuable information. Advantage- Works better on a big data set by removing some samples which helps with the run time and storage problems. Disadvantage-can result into valuable information lose.

1. Random Under Sampler(RUS):

Randomly remove samples from the majority class, with or without replacement. This is one of the earliest techniques used to alleviate imbalance in the dataset, however, it may increase the variance of the classifier and may potentially discard useful or important samples.

2. Cluster Centroid:

Cluster Centroids uses the concept of finding cluster centroid (clusters are created encircling data-points belonging to the majority class). The cluster centroid is found by obtaining the average feature vectors for all the features, over the data points belonging to the majority class in feature space.

3. AllKNN:

Class to perform under sampling based on KNN methods.

4. Tomek links:

TOMEK - Tomek links can be used as an under-sampling method or as a data cleaning method. Tomek links to the over-sampled training set as a data cleaning method. Thus, instead of removing only the majority class examples that form Tomek links, examples from both classes are removed.

5. RESULTS AFTER UNDERSAMPLING TECHNIQUES

When we used Random Under sampling, the recall is high for decision tree, but the precision is very low i.e., 1.35%. So, we are ignoring it. Both logistic regression and random forest have recall 87.16% but the precision is 7.94% and 6.04% respectively. So, we can clearly say that random forest is the best metric for random under sampling.

When we used Cluster Centroids, the recall is very high i.e., 100% but the precision is very low with values in the range 0.23% and 1.19%. So, we can clearly say that this resampled method is not the best method.

TOMEK LINKS AND AIIKNN resampled methods are not giving good results while balancing the datasets. So, we can conclude that these resampled methods are not good for our project.

II. OVERSAMPLING METHODS

This method works by increasing the number of minority class. It either duplicates the entries or manufacturing the data that is the same of what we have. Advantage- there is no loss of data. Disadvantage- multiple instances can lead to overfitting.

1. Random Over Sampler(RUS):

ROS(Random Oversampling Method)- Random Oversampling involves creating the sample sets of minority classes and adding it to the training data thus balancing the classes. With the random oversampling method, we can verify the true power of ensemble models. Our ensemble model combined the best part of the simple models and could get a very high AUC (0.967), high precision (85.19%) and high recall (77.7%). Looking at the confusion matrix we can see the cases for false negative and false positive as 33 and 20, respectively. The logistic regression was the model with the lowest number of false negatives (only 18) and the random forest was the model with the lowest number of false positives (an astonishing result of only 3!). So, if we try to tweak the weights for each model in the ensemble model, we can achieve an even better result.

2. SMOTE:

The analysis for SMOTE method is very similar to the random oversampling method. Again, we can see that the ensemble model could get the best parts of all simple models, resulting in a good performance. Once again, tweaking the weights for each simple model might be a way to enhance the result.

3. ADASYN

ADASYN is a improve version of SMOTE. We add random weights to the samples, making it more realistic. This increases the variance in them instead of being linearly correlated to the parent sample. Our Random Forest produces much better results under ADASYN with Accuracy-99.94%, Precision-85.82%, Recall-77.70% and ROCAUC of 0.934.

4.RESULTS AFTER OVERSAMPLING TECHNIQUES

With the random oversampling method, we can verify the true power of ensemble models. Our ensemble model combined the best part of the simple models and could get a very high AUC (0.967), high precision (85.19%) and high recall (77.7%). Looking at the confusion matrix we can see the cases for false negative and false positive as 33 and 20, respectively. The logistic regression was the model with the lowest number of false negatives (only 18) and the random forest was the model with the lowest number of false positives (an astonishing result of only 3!). So, if we try to tweak the weights for each model in the ensemble model, we can achieve an even better result.

The analysis for SMOTE method is very similar to the random oversampling method. Again, we can see that the ensemble model could get the best parts of all simple models, resulting in a good performance. Once again, tweaking the weights for each simple model might be a way to enhance the result.

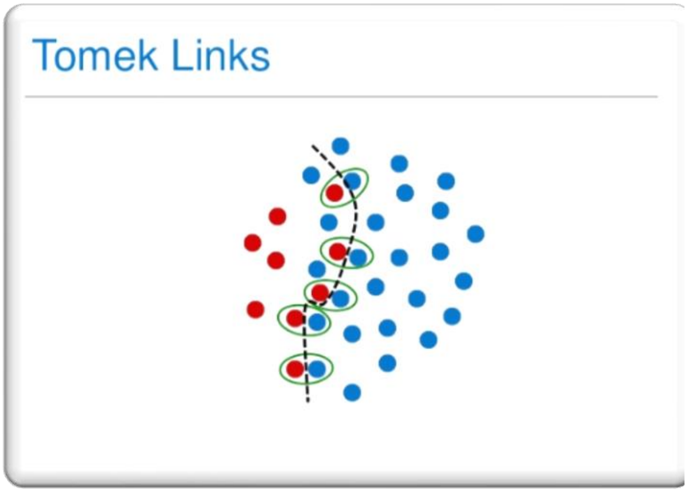
An interesting result is that in all cases of resampling, the highest AUC score among the simple models was using logistic regression. In our opinion, binary classification problems should have at least this model as a benchmark when the objective is to maximize the AUC.

III. HYBRID METHODS

A hybrid sampling combines Random under sampling techniques and Random over sampling techniques to give optimal results. The idea behind this is to combine the strengths of the individual techniques while lessening the drawbacks. Two of the most famous methods are:

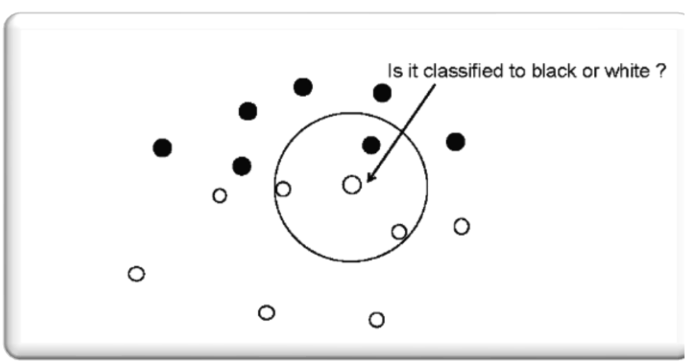
1.SMOTE-TOMEK:

We implemented SMOTE Tomek technique. Tomek links can be used as an under-sampling method or as a data cleaning method. Tomek links to the over-sampled training set as a data cleaning method. Thus, instead of removing only the majority class examples that form Tomek links, examples from both classes are removed. Implementing this technique with Random forest gives us the Accuracy of 99.93%, Precision of 83.82%, Recall of 77.03% and ROCAUC of 0.93077.



2.SMOTE-EEN:

SMOTEEEN Edited Nearest Neighbor removes any example whose class label differs from the class of at least two of its three nearest neighbors. The ENN method removes the instances of the majority class whose prediction made by KNN method is different from the majority class. ENN method can remove both the noisy examples as borderline examples, providing a smoother decision surface. ENN tends to remove more examples than the Tomek links does, so it is expected that it will provide a more in-depth data cleaning. In this technique, Random Forest gives an accuracy or 99.93%, Precision of 80.41%, Recall of 80.41% and ROCAUC of 0.93077.



3.RESULTS AFTER HYBRID TECHNIQUES

Clearly SMOTE EEN, Random forest gives the best result with a Precision of 80.41% which is our True Positive on Actual Results and Recall of 80.41% which is our True Positive on Predicted Results.

As we saw that we were not getting good results with Random Oversampling and Random Under sampling methods, we are building hydride methods by combining both oversampling and under sampling methods. Hybrid methods take one method from oversampling and one method from under sampling, combines them and performs balancing.

With SMOTE TOMEK, we can see that recall is high for logistic regression and naive Bayes but the precision for both is very low. The recall and precision for random forest is 80.41%. So, we can clearly say that random forest is the best metric for this SMOTE TOMEK.

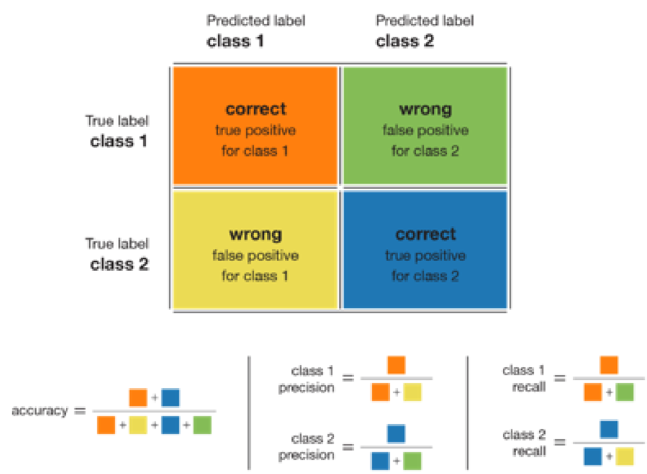
With SMOTE EEN, we can see that recall is high for logistic regression and naive Bayes but the precision for both is very low. The recall and precision for random forest is 77.03% and 83.82% respectively. So, we can clearly say that random forest is the best metric for this SMOTE EEN.

D. Performance Evaluation

In this study, performance evaluations were conducted to assess the performance of the classification methods for each resampling technique. The correlation between these are presented in a confusion matrix below. Performance of all classifiers were compared in terms Accuracy, Precision, Recall and ROCAUC Curve (AUC). These metrics are calculated using the confusion matrix as shown below.

E. Confusion matrix of credit card dataset

	Classified as Fraud	**Classified as Non-Fraud**
Fraud	True Positive (TP)	False Negative (FN)
Non-Fraud	False Positive (FP)	True Negative (TN)



- high recall + high precision : the class is perfectly handled by the model
- low recall + high precision : the model can't detect the class well but is highly trustable when it does
- high recall + low precision: the class is well detected but the model also includes points of other classes in it
- low recall + low precision : the class is poorly handled by the model
- Accuracy- ration of number of correct predictions
- Precision- Percentage of results that are relevant i.e. True Positive/Actual Results
- Recall- Total results correctly classified by the algorithm i.e. True Positive/Predicted results.

- ROCAUC-ROC is probability and AUC represent degree or measure of Separability.
-

Table: 1- Results table on Imbalaced Data

Methods	Model accuracy	Precision	Recall	ROC_AUC
Logistic Regression	99.9%	0.88	0.63	0.98
RandomForest Classifier	99.9%	0.91	0.76	0.94
Naive Baiye Classifier	99.9%	0.80	0.67	0.97

These are the results we obtained when we tried to classify Imbalance data. As expected accuracy is >99.9% But the accuracy is really impressive which is not the right metric.Our model accuracy is> 99.9%.It is a poor measure of performance on imbalanced datasets. And as expected, recall is very low in all the classifiers. That proves that our system is directly failing to classify the data.

Table: 2- Results table using Oversampling techniques

TABLE VI. COMPARISON RESULTS OF Random Over-Sampling BY RATIO 33:66

Resampling Methods: Random Over Sampling				
Techniques	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	97.84%	6.65%	87.84%	0.968
Naive Bayes	97.41%	5.32%	83.11%	0.955
KNN	99.90%	68.55%	73.65%	0.884
Decision tree	99.91%	75.37%	68.24%	0.841
Random forest	99.95%	97.25%	71.62%	0.915
Ensemble model	99.94%	85.19%	77.70%	0.966
Resampling Methods: SMOTE				
Techniques	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	99.14%	14.95%	85.14%	0.966
Naive Bayes	97.69%	5.91%	82.43%	0.954
KNN	99.82%	48.13%	78.38%	0.898
Decision tree	99.78%	41.60%	73.65%	0.867
Random forest	99.94%	84.06%	78.38%	0.93
Ensemble model	99.91%	72.02%	81.76%	0.968
Resampling Methods: ADASYN				
Techniques	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	99.07%	13.98%	85.14%	0.965
Naive Bayes	97.55%	5.62%	83.11%	0.955
KNN	99.82%	48.33%	78.38%	0.904
Decision tree	99.78%	42.57%	71.62%	0.857
Random forest	99.94%	85.82%	77.70%	0.934
Ensemble model	99.91%	70.35%	81.76%	0.97

Table: 3- Results table using under sampling and hybrid techniques

TABLE VI. COMPARISON RESULTS OF Random Under-sampling Techniques BY RATIO 33:66

Resampling Methods: Random Under-sampling Method				
Techniques	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	98.23%	7.94%	87.16%	0.9721
Naive Bayes	96.55%	4.07%	83.78%	0.9537
KNN	97.16%	4.90%	83.78%	0.9479
Decision tree	88.67%	1.35%	89.19%	0.8892
Random forest	97.64%	6.07%	87.16%	0.9645
Ensemble model	98.09%	7.36%	86.49%	0.9671
Resampling Methods: Using Classification Centroid				
Techniques	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	85.96%	1.13%	92.57%	0.9658
Naive Bayes	91.82%	1.76%	84.46%	0.9238
KNN	74.37%	0.62%	91.89%	0.9149
Decision tree	25.77%	0.23%	100.00%	0.6282
Random forest	10.56%	0.19%	100.00%	0.8039
Ensemble model	70.26%	0.56%	95.95%	0.9717

Resampling Methods: AllKNN				
Techniques	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	99.92%	86.54%	60.81%	0.9614
Naive Bayes	97.76%	5.93%	80.41%	0.9554
KNN	99.96%	96.67%	78.38%	0.9999
Decision tree	100.00%	100.00%	100.00%	1
Random forest	99.99%	100.00%	93.92%	0.9999
Ensemble model	85.96%	1.13%	92.57%	0.9658

TABLE VI. COMPARISON RESULTS OF Hybrid Techniques BY RATIO 33:66

Resampling Methods: SMOTEEN				
Techniques	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	99.13%	14.89%	85.14%	0.9678
Naive Bayes	97.71%	5.95%	82.43%	0.9544
KNN	99.80%	45.35%	79.05%	0.8981
Decision tree	99.74%	37.58%	75.68%	0.8772
Random forest	99.93%	80.41%	80.41%	0.93077
Ensemble model	99.90%	68.54%	82.43%	0.9691
Resampling Methods: SMOTE TOMEK				
Techniques	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression	99.14%	15.09%	85.14%	0.9667
Naive Bayes	97.70%	5.91%	82.43%	0.9546
KNN	99.82%	48.13%	78.38%	0.8981
Decision tree	99.78%	41.92%	73.65%	0.8673
Random forest	99.93%	83.82%	77.03%	0.9307
Ensemble model	99.92%	73.94%	82.43%	0.9683

IV. Conclusions and Future Work

Our main aim in this research was to investigate the ability of classification model to classify as fraud and non-fraud transactions and to research more on how resampling methods can improve the performance of our model. In all the six classifier, Random Forest proved to be the best classifier. RF succeeded to get higher accuracy compared to NB, LR, KNN, DT and EM for the resampling methods

- The accuracy score of this algorithm gave an accuracy of 99% which seems impressive, but is it really? As expected Recall for all the methods i.e. Logistic Regression, Random Forest Classifier, Naive Bayes is 63%, 76% and 67% respectively. Recall is the metric for our fraudalent observations. If this low that means we didn't do a good job in classifying fraudalent data.

- Hybrid techniques tend to perform well compared to Undersampling and Oversampling techniques
- SMOTE-ENN is best among all the methods because the recall and precision both are around 80.41% compared to other methods in which one of them is less.
- Undersampling methods are usually not that effective and our results also show the same. Moreover there might be loss of data, we can reject these models.
- Out of oversampling techniques SMOTE and ADASYN performs decently. But can combine with undersampling methods like TOMEK, its more effective.
- Every technique gives different results when run on algorithms, like SMOTE gives some results on Random forest which are very different from results when run on Naive Bayees or Logistc
- Tree based classifiers such as Random forest works better when Resampling techniques are applied.
- Ensemble models need not perform better always, our table results show that.

V. FUTURE SCOPE

The entire research has been carried out by splitting the data into 67% train and 33% test.We can also change the test and train sampling percentage and check the results , which will be different.We can also do neural network analysis and check if there's any change. Predictive models can also be built to improve efficiency of our fraud system.

VI. REFERENCES

[1] Ki, Y., & Yoon, J. W. (2018, January). PD-FDS: Purchase Density based Online Credit Card Fraud Detection System. In KDD 2017 Workshop on Anomaly Detection in Finance (pp. 76-84).

[2] Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. International Journal of System Assurance Engineering and Management, 8(2), 937-953.

[3] Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. IEEE ACCESS, 6, 14277-14284

[4] Quah, J. T., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert systems with applications, 35(4), 1721-1732.

[5] Reddy, M. S., Indraja, S., & Nikhil, L. (2017). Implementation Of Neural Network For Cashless Transactions In Credit Card Transactions. International Journal of Recent Trends in Engineering & Research.

[6] Dal Pozzolo, A., Caelen, O., & Bontempi, G. (2015, September). When is undersampling effective in unbalanced classification tasks?. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 200-215). Springer, Cham.

[7] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In Computational Intelligence, 2015 IEEE Symposium Series on (pp. 159-166). IEEE.

[8] Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. IEEE Intelligent Systems and Their Applications, 14(6), 67-74.

[9] Chan, P. K., & Stolfo, S. J. (1998, August). Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In KDD (Vol. 98, pp. 164-168).

[10] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

[11] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002, January). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).

[12] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), 3784-3797.

[13] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing,
G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220-239.

[14] Santoso, B., Wijayanto, H., Notodiputro, K. A., & Sartono, B. (2017, March). Synthetic Over Sampling Methods for Handling Class Imbalanced Problems: A Review. In IOP Conference Series: Earth and Environmental Science (Vol. 58, No. 1, p. 012031). IOP Publishing.

[15] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

[16] Batuwita, R., & Palade, V. (2010, July). Efficient resampling methods for training support vector machines with imbalanced datasets. In Neural Networks (IJCNN), The 2010 International Joint Conference on (pp. 1- 8). IEEE.

[17] Ghazikhani, A., Yazdi, H. S., & Monsefi, R. (2012, May). Class imbalance handling using wrapper-based random oversampling. In Electrical Engineering (ICEE), 2012 20th Iranian Conference on (pp. 611-616). IEEE.

[18] Jagelid, M., & Movin, M. (2017). A Comparison of Resampling Techniques to Handle the Class Imbalance Problem in Machine Learning: Conversion prediction of Spotify Users-A Case Study.

[19] Yan, J., & Han, S. (2018). Classifying Imbalanced Data Sets by a Novel RE-Sample and Cost-Sensitive Stacked Generalization Method. Mathematical Problems in Engineering, 2018.

[20] Credit Card Fraud Detection. Accessed: June. 7, 2018. [Online]. Available: https://www.kaggle.com/dalpozz/creditcardfraud

[21] https://thesai.org/Downloads/Volume9No11/Paper_55-Handling_Class_Imbalance_in_Credit_Card_Fraud.pdf

[22] https://arxiv.org/pdf/1608.06048.pdf