Pattern Recognition and Machine Learning

Indian Institute of Technology, Jodhpur

Minor Project
Report

Harshil Kaneria (B21CS033)
Amisha Kumari (B21CS007)

Department of Computer Science, IIT Jodhpur
March 31, 2023

The Credit Card Fraud dataset is highly imbalanced. The problem this dataset could lead to is that it can bias machine learning models towards the majority class, resulting in poor performance on the minority class.

Techniques used to address the problem are:

- **Undersampling**: A technique to balance class distribution by removing samples from the majority class in the dataset.
- **Oversampling**: A technique to balance class distribution by duplicating or generating new samples for the minority class in the dataset.
- **Combination of Undersampling and Oversampling**: A technique that combines undersampling and oversampling to balance class distribution in the dataset.
- **Cost Sensitive Learning**: A technique that assigns different costs to different classes to improve the performance of a classifier on imbalanced datasets by balancing misclassification cost.
- **Ensemble Methods**: A technique that combines multiple models to improve the overall performance of a classifier, including methods like bagging, boosting, and stacking.

Different Machine learning Algorithms implemented:

- **Logistic Regression:** Logistic regression is a supervised learning algorithm used for binary classification that models the relationship between a dependent binary variable and one or more independent variables using a logistic function, which outputs a probability value representing the likelihood of a sample belonging to a particular class.
- **Random Forest Classification:** A random forest model is a supervised learning algorithm that creates multiple decision trees and combines their outputs to make predictions, making it a powerful classifier that can handle complex datasets.
- **Extreme Gradient Boosting Classification:** An extreme gradient boosting (XGBoost) model is a supervised learning algorithm that uses a gradient boosting framework to iteratively improve the performance of decision trees, resulting in a highly accurate and robust classifier.
- **Support Vector Machine Classification:** A support vector machine (SVM) model is a supervised learning algorithm that finds the best hyperplane to separate the data into classes, making it an effective classifier for both linearly and nonlinearly separable datasets.
- **K Nearest Neighbor Classification:** A K-nearest neighbor (KNN) model is a supervised learning algorithm that classifies new data points based on the

class of the K-nearest training data points, making it a simple yet effective classifier for small datasets.

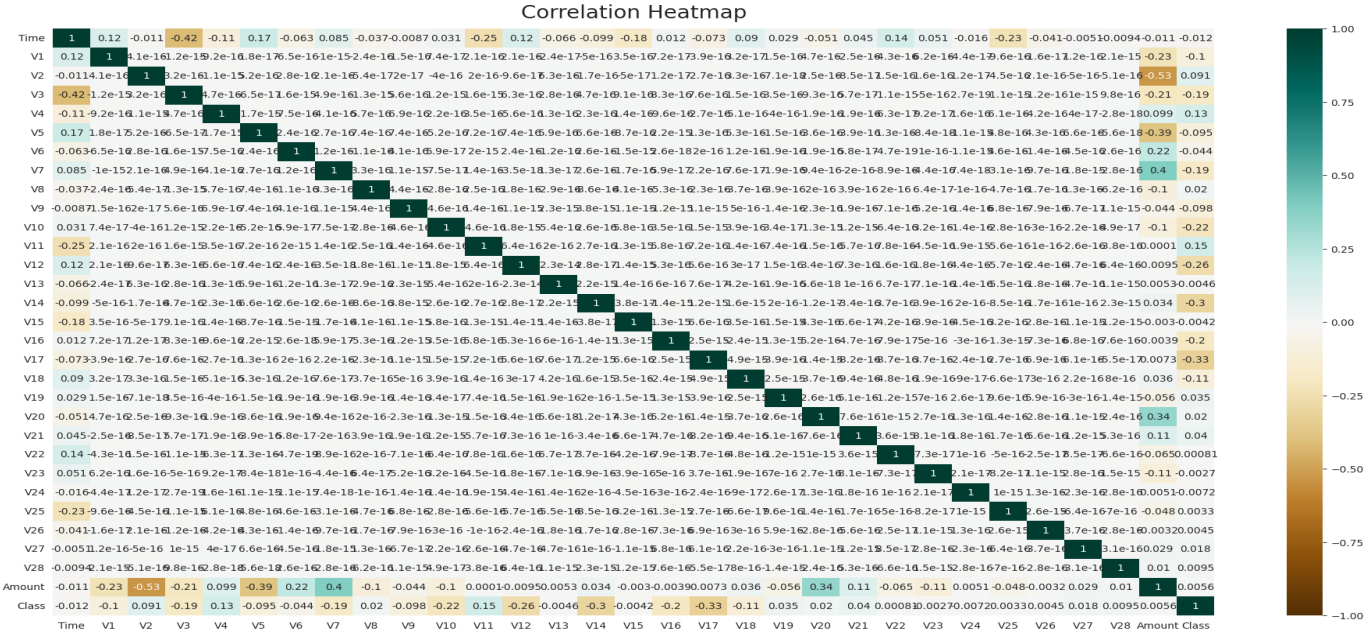
The metric which will be appropriate to evaluate this kind of imbalanced dataset is **recall**. We would consider class 1 (Fraudulent class) as a positive class. This implies 100 % recall would mean that all the actual fraudulent transactions are being classified correctly as fraudulent transactions. So higher the recall the better is the model.

Data Visualisation and Inferences:

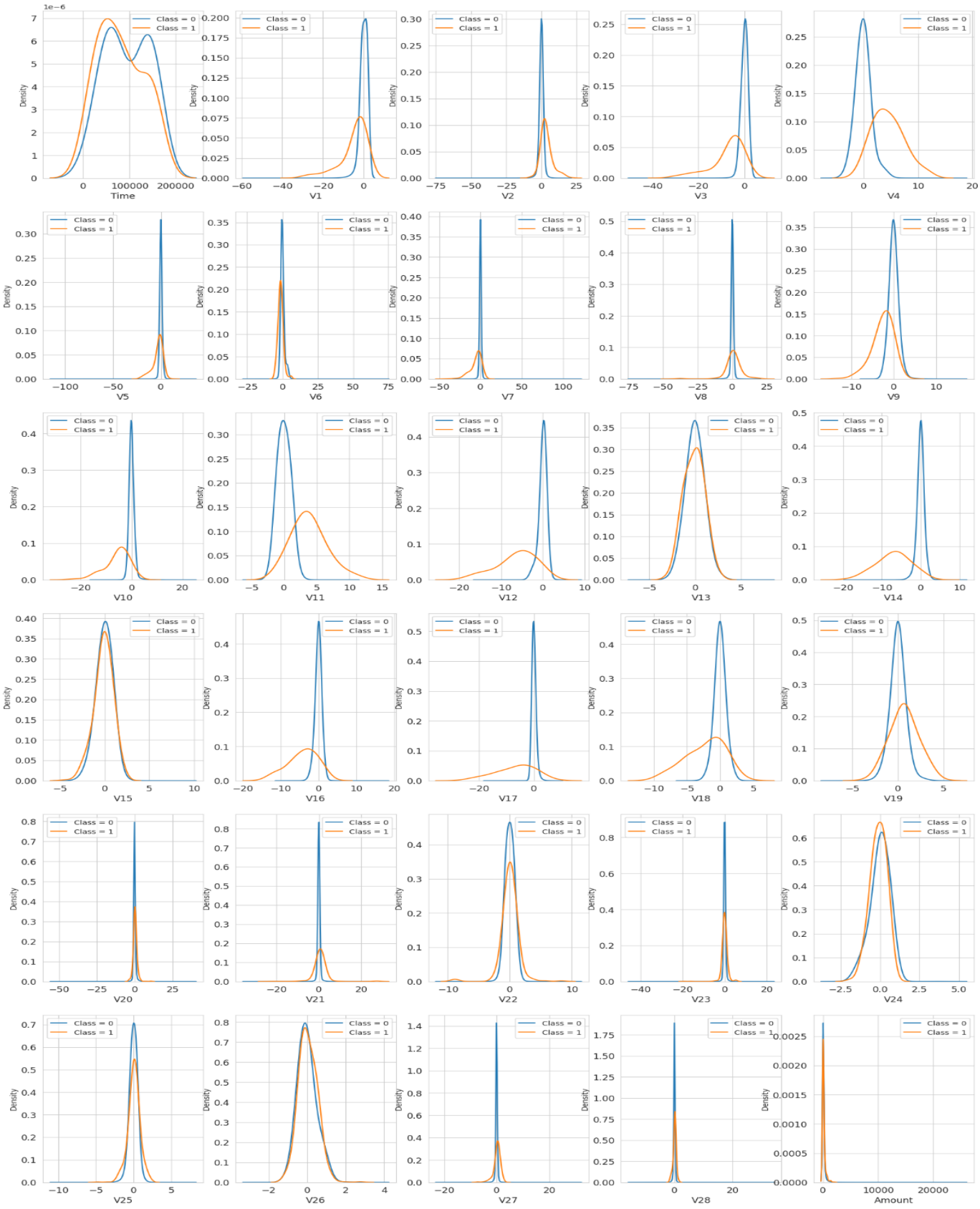
From the pearson correlation heatmap shown below we can infer the following points:

1)V2 V4 V11 and V19 are more positively correlated with class. So, the higher the value, the more likely it is to belong to a Fraudulent class.

2)V10 V12 V14 and V17 are more negatively correlated with class. So, the lower the value , the more likely it is to belong to a Fraudulent class.



From the density graph given below we can observe that all the features follow gaussian distribution and the density of class 0 for each feature is higher than class 1 density around the mean. We can also see that the mean of class 0 is greater than the mean of class 1 in most of the features.


Class = 0
Class = 1
Density
Time
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
Amount

Observations:

Under-Sampling:

Advantages: Reduced computational complexity , Reduced bias towards the majority class , Improved interpretability , Improved generalization.
Disadvantages: potential loss of information , risk of increasing the variance of the model due to the smaller sample size.

- **Random Under Sampling:**

Sampling Techniques	Model	Accuracy	Precision	Recall	ROC-AUC score
Random Under Sampling	Logistic Regression	97.09 %	4.84 %	92.92 %	0.98
	Random Forest Classifier	96.92 %	4.51 %	91.15 %	0.98
	XGB Classifier	96.57 %	4.13 %	92.92 %	0.99
	SVM Classifier	98.22 %	7.46 %	89.38 %	0.99

It is observed that best recall is given by the Logistic Regression model and XGB classifier. So for Random Undersampling technique XGB classifier and Logistic Regression model will work best.

- **Edited Nearest Neighbours (ENN):**

Sampling Techniques	Model	Accuracy	Precision	Recall	ROC-AUC score
ENN - Edited Nearest Neighbor	Logistic Regression	99.93 %	84.27 %	66.37 %	0.97
	Random Forest Classifier	99.94 %	79.31 %	79.31 %	0.96
	XGB Classifier	99.94 %	81.74 %	83.19 %	0.98
	SVM Classifier	99.93 %	81.0 %	71.68 %	0.96

It is observed that best recall is given by the XGB classifier. XGB is a boosting method. So for ENN technique XGBclassifier will work best.

Over-Sampling:

Advantages: Increased representation of minority class , Reduced bias towards the majority class , Improved generalization , Reduced risk of information loss
Disadvantages: potential for overfitting , risk of introducing noise or bias into the dataset

- **Random Over Sampling:**

Sampling Techniques	Model	Accuracy	Precision	Recall	ROC-AUC score
Random Over Sampling	Logistic Regression	97.57 %	5.69 %	92.04 %	0.98
	Random Forest Classifier	99.96 %	93.68 %	93.68 %	0.96
	XGB Classifier	99.96 %	90.29 %	82.3 %	0.98
	SVM Classifier	98.74 %	9.87 %	84.96 %	0.97

It is observed that best recall is given by the Random Forest Classifier. It's an Ensemble method. So for Random Oversampling technique, Random Forest Classifier will work best.

- **SMOTE (Synthetic Minority Over Sampling Technique):**

Sampling Techniques	Model	Accuracy	Precision	Recall	ROC-AUC score
SMOTE	Logistic Regression	97.42 %	5.43 %	92.92 %	0.98
	Random Forest Classifier	99.95 %	82.91 %	82.91 %	0.98
	XGB Classifier	99.93 %	72.93 %	85.84 %	0.99
	SVM Classifier	98.34 %	7.83 %	87.61 %	0.97

It is observed that best recall is given by the Logistic Regression model. So for SMOTE technique, the Logistic Regression model will work best.

Combining Under and Over Sampling:

- **SMOTEENN (Synthetic Minority Over-sampling Technique Edited Nearest Neighbors):**

Sampling Techniques	Model	Accuracy	Precision	Recall	ROC-AUC score
SMOTEENN	Logistic Regression	97.27 %	5.15 %	92.92 %	0.98
	Random Forest Classifier	99.94 %	76.98 %	76.98 %	0.97
	XGB Classifier	99.9 %	64.05 %	86.73 %	0.99
	SVM Classifier	98.31 %	7.73 %	88.5 %	0.97

It is observed that best recall is given by the Logistic Regression model. So for SMOTE ENN technique, the Logistic Regression model will work best.

Cost Sensitive Learning:

Advantages: Improved accuracy , Customized trade-off between precision and recall, Better decision making , More efficient resource allocation

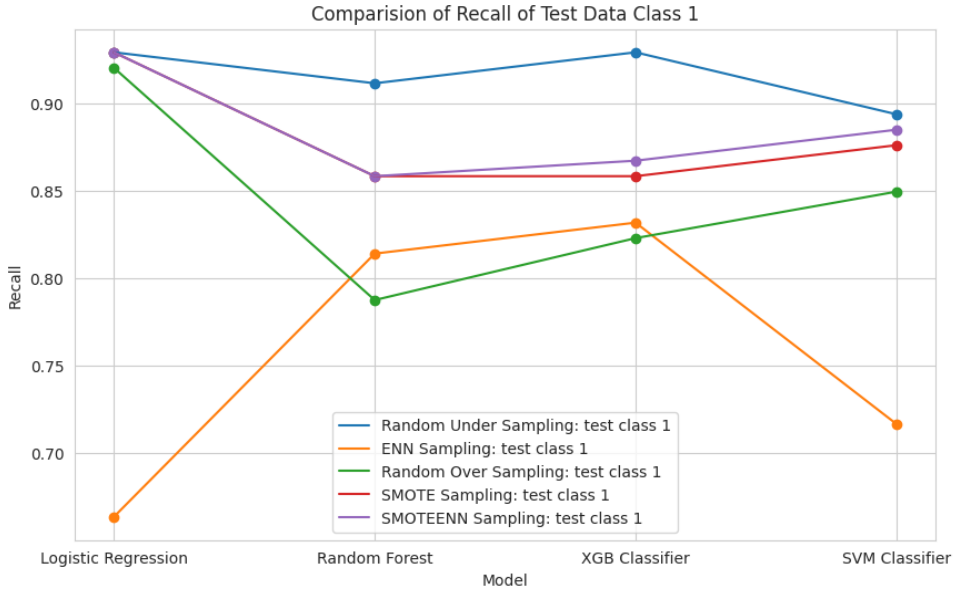
Disadvantages: Increased complexity , There is need for careful selection of cost matrices or weighting schemes

Technique	Model	Accuracy	Precision	Recall	ROC-AUC score
Cost Sensitive Learning	Logistic Regression	97.58 %	5.71 %	92.04 %	0.98
	Random Forest Classifier	99.96 %	94.57 %	94.57 %	0.95
	XGB Classifier	99.96 %	89.42 %	82.3 %	0.99
	KNN Classifier	99.95 %	90.32 %	74.34 %	0.93

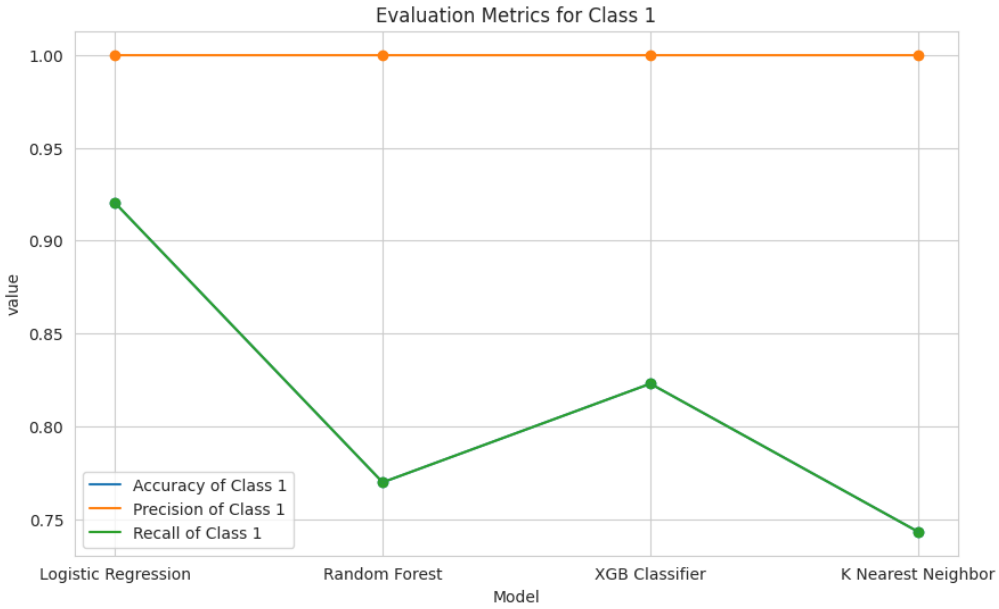
It is observed that best recall is given by the Random Forest Classifier. So for the Cost Sensitive Learning technique, the Logistic Regression model will work best.

The graph given below shows the comparison of Recall scores for different sampling techniques with different classifiers on class 1 of test data. From the graph we can observe

that Random Under-sampling has the highest recall of all the variants. This may be due to high variance as the dataset is small. We have Recall score close to 90% for SVM classifier for SMOTEENN technique. The accuracy and recall score of class 1 would be the same as class 1 is positive class and we don't have any negative class i.e. class 0 in the dataset used for calculating the following graph.



The graph below shows the Accuracy, Precision and Recall scores of class 1 of test data for Cost Sensitive Learning technique. Accuracy and Precision of Class 1 on test data for Cost Sensitive Learning is 100 % while recall varies for different classifiers. The best recall is given by the Logistic Regression Model. As this is computed after training on full dataset and then recall being above 90% is a good fit for fraud detection. The Logistic Regression has performed well for the dataset overall.



Conclusion:

We can conclude that for an imbalance dataset like this combining under and over sampling would work well. Also Cost sensitive learning works well for the same. Eventually we can also conclude that the Logistic Regression has performed well in all the sampling techniques and cost sensitive learning techniques. It has given a very good recall and accuracy for SMOTE and SMOTEENN sampling techniques that correspond to over-sampling and combining over and under-sampling techniques respectively.