Learning from Imbalanced Datasets (Supervised and Unsupervised Learning)

Harshada Raut, MSc in Data Science, University of Essex, U.K

Abstract—The main aim of the paper is to test a new approach to deal with Imbalanced Data sets using Supervised and Unsupervised learning.In real life incidence, imbalanced classes appear in many domains and at the time of dealing with it we neglect how to tackle classification problem significantly.Hence in this assignment we will perform cross validation and apply k-means clustering on three of such databases to conclude which method is best fit for imbalance datasets and how it affects the results.However, we all know Road accident is a serious issue these days which results in a serious death and some serious disabilities all around the world.Whereas Titanic brings out the history of a mysterious incident as there is no definite evidence of passengers who died in doomed situation. Moreover, the Diabetes diseases, not only lowers quality of life and greatly increases medical expenses, but also significantly increases disease-related deaths.Therefore I considered that it is necessary to perform analysis on above imbalance data sets and to make better prediction in order to avoid some such situations in near future.

Index Terms—Imbalanced Data, classification, Clustering, supervised learning,unsupervised learning.

1 Introduction

Imbalanced classes appear in many domains such as Fraud detection, Spam filtering, Diseases Screening,Advertising Click-through etc.It always has a binary classification variable known as target which plays an important role through out the analysis. A data set is known as an Imabalanced dataset when instances of one of the two classes is higher than the other, in another way, the number of observations is not the same for all the classes in a classification dataset. Suppose you are working in a company and asked to create a model that predicts whether a product is defective or not.You decide to use classifier and train it on the data achieving 96.2% percent accuracy.but a week later you find out that around 3.8% of the products are defective and your model just always answers "NOT DEFECTIVE" leading to 96.2% accuracy.Many times we don't realize that we are working with an imbalance dataset and hence couldn't tackle classification problems that comes with it.

The sole purpose of this assignment is to generate as many information and insights about the data as possible.Not only we dig more but we also can find any problem that might exist in the dataset. Imbalance data set is a scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes.The main question faced during data analysis is – How to get a balanced dataset by getting a decent number of samples for these anomalies given the rare occurrence for some them?

However, Domain accuracy is not an appropriate measure to evaluate model performance while working with imbalance data set. We can deal significantly with an imbalance dataset by improving classification algorithms. Another method is to balance classes in the training data before providing the data as input to the alogorithm.The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This is done in order to obtain approximately the same number of instances for both the classes.Thus we will be using random under-sampling to balance distribution by randomly eliminating majority classes.This process will be carried out until the majority and minority class instance are balanced out.The advantage of using random forest is it

can help improve run time and storage problems by reducing the number of training data samples when the training data set is huge.[1] The Pima- Indians Diabetes has less diabetic patients than that of diabetic which resulted imbalance at around 65% which is an answer to very crucial issue of diabetes deceases.It has now led to become one of the major diseases that result in deaths.
On the other hand the road accident is another major problem rapidly increasing around the globe. hence I chose to perform analysis on Oman Data set which has 89% of imbalance ratio. However the Titanic data set is has this ratio of 68% of dead people resulted after an investigation of how many people survived from an Infamous incident.I have performed cross validation and random forest to make it balanced to perform k-means clustering effectively.

2 Background

Different machine learning algorithms have been developed recently to tackle this problem, which mostly have been based on sample techniques, cost sensitive learning and ensemble methods.As the matter of fact we know, accuracy is not the best criteria to be considered while dealing with an imbalance dataset. We can handle this problem by using Logistic Regression algorithm, KNN, SMOTE and accuracy metric from Scikit-learn.Whereas many machine learning algorithms are designed to maximize overall accuracy by default. Thus another way of is to use Up-sampling minority classes.It is the process of random duplication of observations from the minority class in order to reinforce its signal.First we will separate observations from each class into different DataFrames.further, we will resample the minority class "With replacement", setting the number of samples to match that of the majority class. Finally, we will combine the up-sampled minority class DataFrame with the original majority class DataFrame and thus we can get a model which will give meaningful accuracy.
Moreover, Down-Sampling is another technique which can be used to deal with large datasets.It randomly removes observations from the majority class to prevent domination of learning algorithm. It follows similar process as up-sampling only the difference is we will resample the class "Without replacement" and will match to that of minority class instead of majority.
So far, we've looked at two ways of addressing imbalanced classes by resampling the dataset. Now finally we will consider use of Decision Trees on these datasets because their hierarchical structure allows them to learn signals from both the classes.Thus, we will train the dataset using Random Forest on the original imbalance dataset. [2]

3 Methodology

We will began our analysis by plotting some basic plots and generate correlation matrix for variables to identify relation among them.But before we proceed we must run test for null values and if there are any we better drop those columns from the dataset which will make it more suitable for clustering.

3.1 Cross-validation on dataset using a decision tree and Random Forest

Further we will check the Imbalance ratio of our dataset by adding a new variable named Z-score in the database. It will give the ratio of Imbalance in percentage. Moving on to the next step we will perform cross - validation on our datset using Random Forest and Decision tree. We will use Labelencoder to make sure that all the variable will fit in the transformation matrices.We will perform 10 k-fold Cross validation by spliting data set into test and training.We will also produce the ROC curve for the same and get the Accuracy mean of the dataset.

3.2 Kmeans clustering with Elbow and Silhouette Method

In step 3 we will partition the data set into 10 bins with leave one out approach keeping the imbalance ratio from original dataset.Our main

aim to perform Elbow and Silhouette method on datasets using k-Means clustering.The elbow method helps to choose the optimum value of 'k' (number of clusters) by fitting the model with a range of values of 'k'. Here we would be using a 2-dimensional data set but the elbow method holds for any multivariate data set.[3][4] And it has seen that the value of 'Epsilon' decreases with iteration number for each dataset. Moreover the Silhouette analysis is used to study the separation distance between the clusters.This method shows effectively that how close each point in one cluster is to other points in the neighboring clusters.Therefore it also provides a way to access parameters like number of clusters visually as shown in above figure.

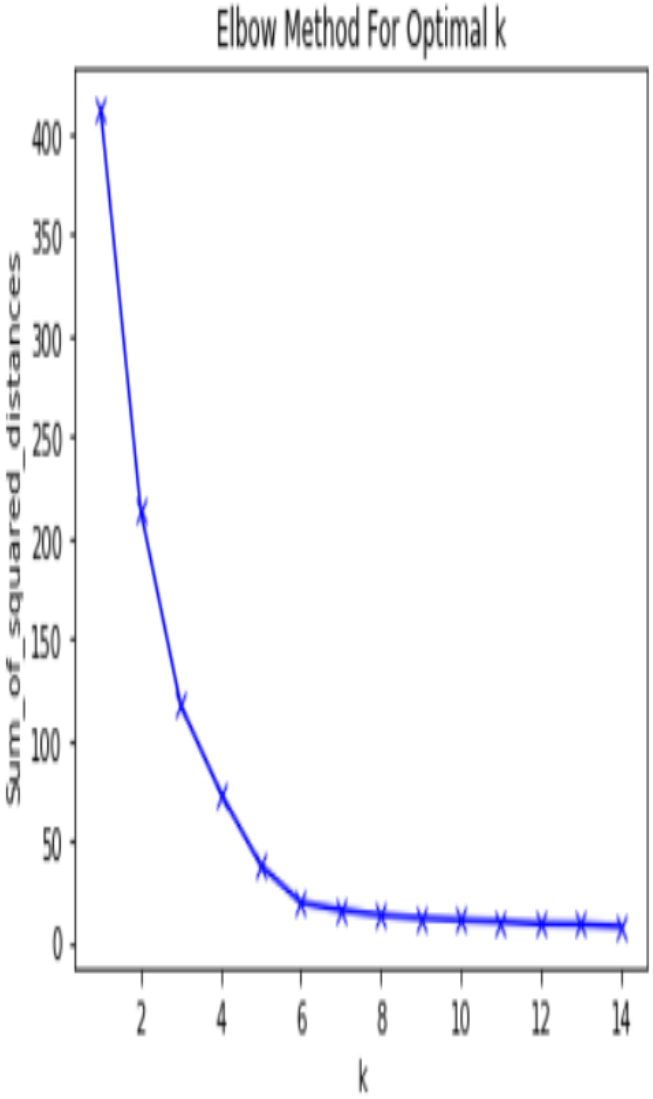

Fig. 1: Elbow Methodology

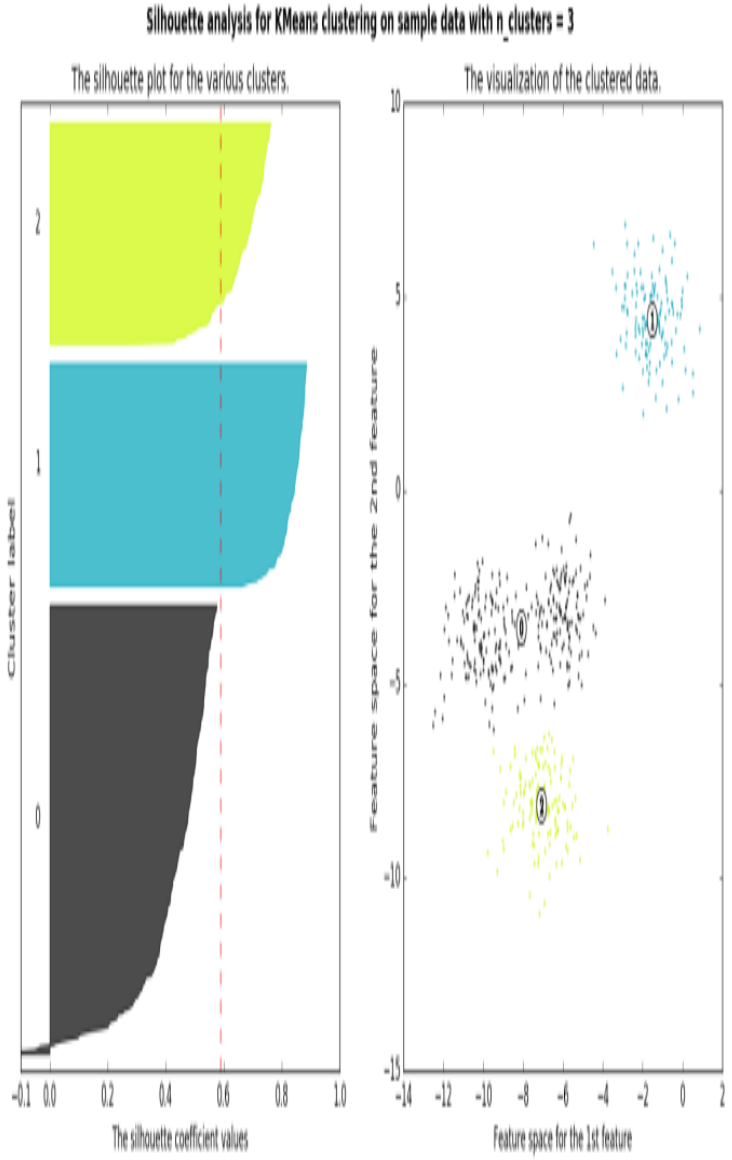

Fig. 2: Silhouette Methodology

3.3 Random Forest algorithm with two class data

Now for each cluster we will identify its centroid and the number of samples of the minority class in that cluster as shown below.Numbers of minority and majority class have been calculated which helped to separate two class clustered bin data from whole bin data for hold out bin. Further we will Train a random forest for each of the clusters that

contains samples from more than one class. The next task is to assign x to its closest cluster by using unseen k-fold technique.

3.4 Random forest and decision tree with with class data

Cluster assignment performed for rest nine bins and two class data for all 10 bins separated from dataset.Random forest and decision tree trained with two class data with leave one out approach and evaluation metrics saved. this results for both algorithms are better than the cross-validation with random forest and decision tree results.

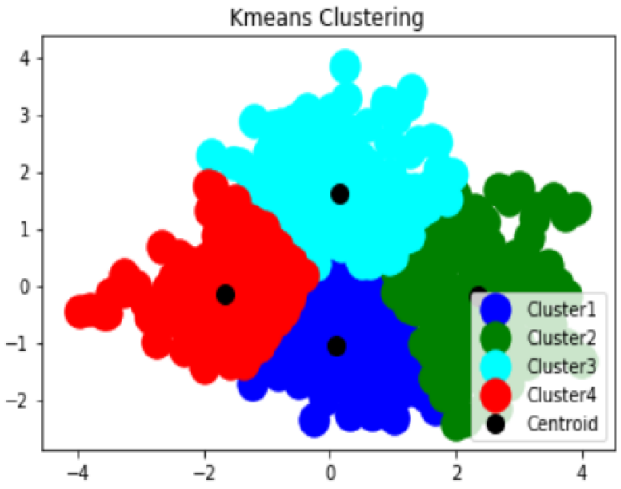

Fig. 3: Kmeans clustering

At last we will obtain boxplots of the cross-validation result for each method.We will compare new methods to that of old one to check which method the results significant accuracy.

4 Experiment

To achieve the aim of the given task, Three datasets(Binary Classification)with specified imbalance ratio have been taken from Kaggle to run test. To began with, The Pima Indians Diabetes Database s originally from the National Institute of Diabetes and Digestive and Kidney Diseases.The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.The datasets consists of several medical predictor variables and one target variable Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.It has 768 instances and 10 features with imbalance ratio of 65 percent. As only 268 patient has diabetes since 500 of them don't have it.

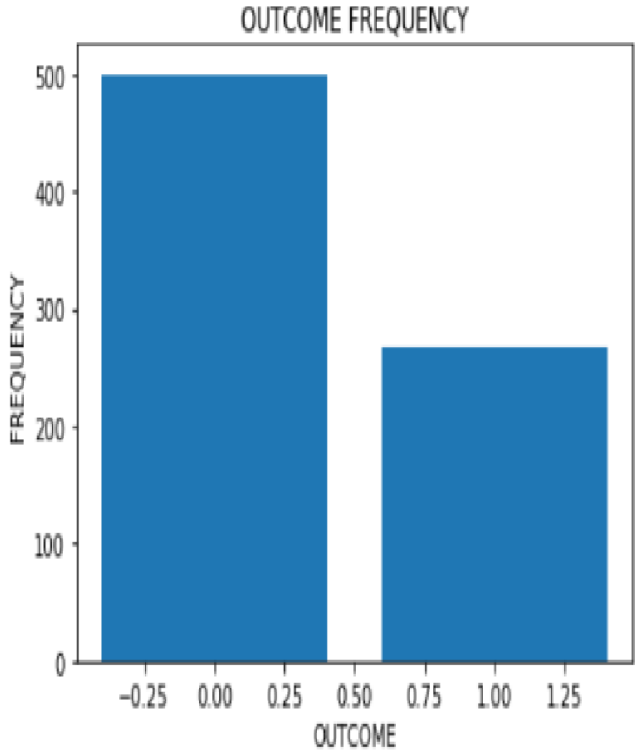

Fig. 4: Frequency chart of Diabetes dataset

Moving on to the next, The Oman dataset has 24192 instances with 16 features.The data set has one target variable(deaths) and rest are predictors. The aim of this challenge is to find out the cause of many deaths due to road accidents occurred in Oman country. It has proven that over speeding is the main cause behind the road accidents in Oman, followed by bad behavior and also the other factor such as negligence played an important

role in road accident.There are 21647 values with 0 deaths so up sampling method was used increase size of minority class 1 equal to 21647. After the balancing the total data-set became equal to 43294. Now the dimensions of data set are (43294,11). The data set is converted into x and y variables, in which x = all input variable and y = target variable 'deaths'. The balancing is made because its a serious problem with some data sets, keeping same level of ratio (imbalanced class) can result in models with increased bias towards the majority class(minority-class instances predicted as being in the majority class)

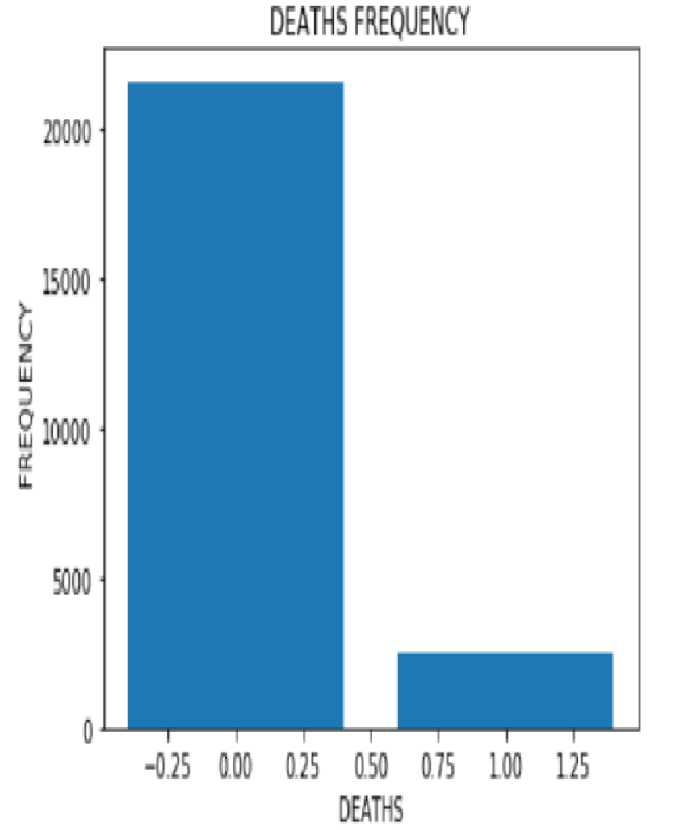

Fig. 5: Frequency chart of Oman dataset

Last but not the least, The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone on board, resulting in the death of 1502 out of 2224 passengers and crew which resulted imbalance ratio over 67 percent. The target variable(Survived) and predictor variables are taken into consideration to identify what sort of people died in this tragedy.To analyse these datasets we will be Using the Elbow method and the Silhouette method, to identify the number of clusters and Later on will run k-means clustering.This testing will be carried out to select a final clustering that with the lowest output criteria.

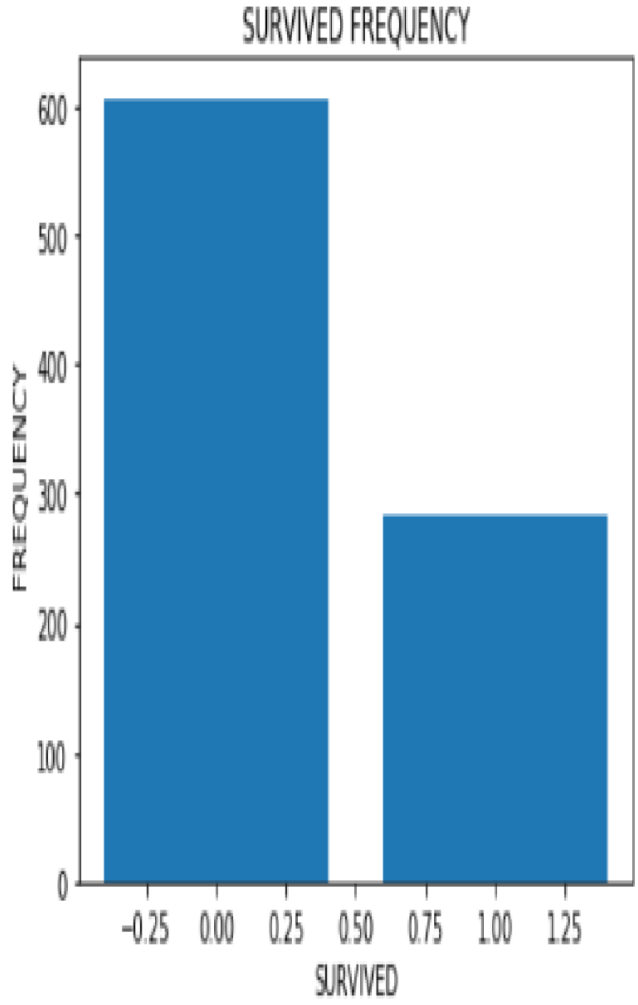

Fig. 6: Frequency chart of Titanic dataset

5 Discussion

In this point it is crucial to consider how the evaluation of the results will take place. Classi-

fication accuracy is not a good metric in applications with class imbalance issues, because it places more weight on the frequent classes than on uncommon classes. Therefore, a classifier cannot perform well on the rare classes.
The most frequent metric, though, is ROC analysis and the associated use of the area under the ROC curve (AUC) to assess overall classification performance. AUC it is not biased against the minority class and shows how correctly the model separates the positive and negative examples and ranks them.Specifically ROC curve is a two-dimensional chart in which True Positive rate (percentage of the correctly classified positive observations) is plotted on the y-axis and False Positive rate (percentage of the misclassified negative observations) is plotted on the x-axis. Furthermore, another well known evaluation metric is the F value, which combines precision and recall, and is high when both of them are high.
Ultimately, even if ROC is a popular and strong measure to evaluate performance of binary classifiers, the precision-recall (PRC) plot is highly recommended as the most explanatory tool for visual analysis.

6 Conclusion

Handling Imbalance Dataset is really a difficult task if one do not know how to tackle classifier problems. Such as in Oman datasets the Imbalance ratio of binary classifier deaths was higher at around 89%. If not handled significantly it will show the accuracy rate approximately 1 In spite the fact that are many people got injured and several died because of over speeding in Oman. If it wasn't for training dataset with accurate classifiers this situation would have been easily neglected by the society.
When diabetes dataset taken into consideration it was clear that over 40% of people are suffering through this decease.If the situation will not be controlled it can become the leading cause to the death. Moreover more than 65% passangers of Titanic died that night and many of them were found missing mysteriously. All the above data sets contains binary classifier hence predicting outcomes was a challenge.
After creating box plot for cross-validation of the datasets it was clear that old methods such as stratify cross validation, Decision tree, Random forest and k-means clustering are better than that of new methods where we assign x to each cluster to 10 bins with leave one out aproach.However, Imbalance dataset is quite difficult to analyse as if variables are not partitioned well the Decision tree will always throw an error for binary classification.The real task was to partition data set to variables. But once you get an Idea how to deal with imbalance data set by using appropriate classifiers you realize it is an interesting problem.

References

[1] A. Vidhya, "Imbalanced data : How to handle imbalanced classification problems." [Online]. Available: https://www.analyticsvidhya.com/blog/2017/03/i data-classification/
[2] unknown, "How to handle imbalance dataset in machine learning," 2017. [Online]. Available: https://elitedatascience.com/imbalanced-classes
[3] S. J. Franklin, "Elbow method of k-means clustering using python," Analytics Vidhya, 2019. [Online]. Available: https://medium.com/analytics-vidhya/elbow-method-of-k-means-clustering-algorithm-a0c916adc540
[4] W. Fu and N. Y. U. Patrick O. Perry Stern School of Business, "Estimating the number of clusters using cross-validation," 2017. [Online]. Available: https://arxiv.org/pdf/1702.02658.pdf