A blend of Supervised and UnSupervised Learning from Imbalanced Datasets

February 24, 2022

Registration number:	2101194
Project:	Imbalanced datasets
Link to GitHub:	https://github.com/ahmadraza346/CE888/tree/main/Learning%20From%20Imbalanced%

Executive summary (max. 250 words)	232
Introduction (max. 600 words)	509
Data (max. 500 words/dataset)	450, 429, 465
Methodology (max. 600 words)	599
Conclusions (max. 500 words)	263
Total word count	2947

Contents

1 Introduction 2

2 Data 2
- 2.1 Mushroom Data Set . . . 3
- 2.2 Gender Data Set . . . 5
- 2.3 Heart Data Set . . . 7

3 Methodology 9
- 3.1 Data Sets and Surrogates . . . 9
  - 3.1.1 Surrogates for Mushroom Data Set . . . 10
  - 3.1.2 Surrogates for Gender Data Set . . . 11
  - 3.1.3 Surrogates for Heart Data Set . . . 11
- 3.2 Proposed Model . . . 12
  - 3.2.1 Baseline Model . . . 12
  - 3.2.2 Bins Creation . . . 13
  - 3.2.3 K Mean . . . 13
  - 3.2.4 Random Forest Classifier . . . 13
  - 3.2.5 Model Testing . . . 13
  - 3.2.6 Model Results . . . 13

4 Conclusions 14

Abstract

Its very common to come across imbalanced data sets while working on classification models. Due to unequal representation of classes in these data sets, the learning of our models is compromised as they overlook the minority class. Among many existing solutions to this problem, re-sampling of the data sets is the most simple but not ideal approach where either majority class is down sampled or minority class is up sampled.In this project, we are proposing a new method of learning from imbalanced data sets based on a blend of supervised and unsupervised learning. For that purpose, we have created surrogates with low, medium and high ratio of imbalances from three chosen data sets that are publicly available and have been pre-processed and cleaned. As a baseline, we are going to perform stratified cross-validation on each data set and its surrogates using Random Forest Classifier. After receiving baseline results, we are again going to perform stratified cross-validation but this time using K-mean clustering and Random Forest Classifier both, where K-Mean clustering is going to aid in identifying the clusters having samples from more than one classes and Random Forest Classifier is used to train those clusters. Finally, to compare the results of base model and proposed model for data sets and their surrogates, we will draw box plots and perform permutation tests for the accuracy, F1, precision and recall scores obtained by cross-validation.

1 Introduction

With recent developments in the field of Science and Technology, bulks of data coming from a wide range of sources is being used extensively in decision making processes and real world applications. But, with the amount of data increasing exponentially, problems such as learning from imbalanced data sets have arised.[4] Data set is considered to be imbalanced when at least one class is having very few instances of the data while other class(es) contain majority of the instances from the data. Machine learning algorithms work best when the data sets are balanced otherwise they produce results that are biased towards the majority class.[3] One reason for the existence of imbalanced data sets is the rare instances of the minority class in real world. For example, in bank data set, its understood that number of fraudlant transactions is going to be very rare as compared to the normal transactions.

In case of imbalanced data sets, accuracy of the classifier on the majority class is exceptionally good while the accuracy for the minority class is very poor.[2] Most of the classification algorithms have poor performance on imbalanced data sets because they are accuracy driven and try to minimize the overall error assuming that all the classes are balanced and the errors produced by each class costs same, but in reality its not true.[7] In real world scenario, obtaining incorrect results from classification on imbalanced data sets can cost a lot. For example in diagnosis of a cancer, where class 1 represents cancer diagnosed and class 0 represents cancer not diagnosed. If class 1 is mis-classified(a person having cancer classified as not having cancer), it is going to cost a lot more than class 0 being mis-predicted(a person not having cancer but diagnosed as having cancer). In case of wrong prediction for class 1, it may cost the life of a person.[2]

Many works have been carried out to deal with data imbalances among which one is to modify the existing classifiers. Another one is to modify the data. Data modification technique is also termed as data re-sampling. In data re-sampling, either the majority class is under-sampled or the minority class is over-sampled. During under-sampling, the instances of a majority class are removed to create a balance while in over-sampling new instances are created in the minority class to achieve a balance. Both techniques have their pros and cons. During under-sampling, there is a possibility of losing the valuable information because of the removal of rows whereas over-sampling can result in over-fitting the model and increase the learning time.[5]

In this project, we are going to use a combination of supervised and unsupervised learning. The motivation for this project is to develop a new blended method that deals with the problem of imbalanced data without manipulating the data sets. We are going to perform stratified cross-validation using K-mean clustering and Random Forest Classifier and then compare the results with the baseline model trained on Random Forest Classifier. Evaluation metrics to be used are accuracy, F1, precision and recall scores.

2 Data

For this project, we are using three publicly available data sets. The reason for choosing these data sets is because they are balanced and perfectly fit the initial requirement of our project. I.e. To choose balanced data sets and create Imbalanced surrogates from them.

2.1 Mushroom Data Set

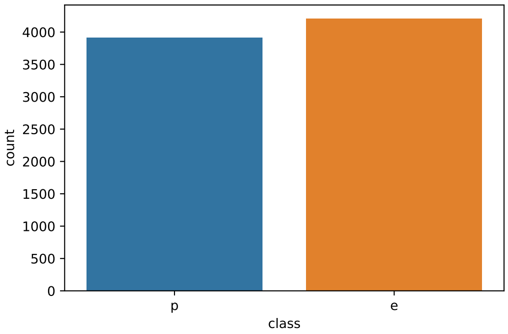

Figure 1: Data Distribution in Mushroom Data Set.

The first data set we have chosen for our project is a mushroom data set publicly available on kaggle. Its a binary classification problem where mushrooms are being classified as poisonous(p) or edible(e). The data set has 8124 total instances and 22 features where 3916(48.2 percent) instances belong to class poisonous(p) and 4208(51.8 percent) belong to class edible(e) as shown in Figure 1. Therefore, initially the data set is almost balanced with examples from both classes. As seen in Table 1 ,all features as well as target variable belonging to the Mushroom data set are categorical in nature. Furthermore, data set contains no null values or duplicate row.

class	cap-shape	cap-surface	cap-color	bruises	population	habitat
p	x	s	n	t	s	u
e	x	s	y	t	n	g
e	b	s	w	t	n	m
p	x	y	w	t	s	u
e	x	s	g	f	a	g

Table 1: Mushroom Data Set.

During data exploration, we have plotted the categorical features against their percentages in different classes and have gathered some insights from the data.

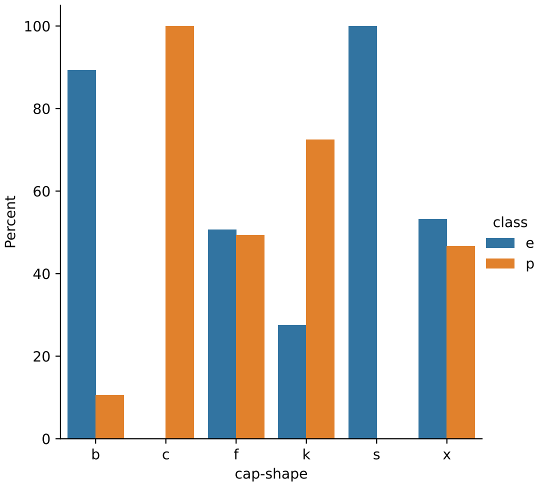

Figure 2: Percentages of cap-shapes belonging to class poisonous and edible.

- Conical cap-shaped(c) mushrooms are almost always poisonous while sunken(s) shaped mushrooms are always edible. Figure 2
- Grooves(g) cap-surfaced mushrooms are almost always poisonous.
- Green(r) and purple(u) mushrooms are almost always edible mushrooms.

- Almond(a) and Anise(l) odourded mushrooms are almost always edible while cresosote(c), foul(f), musty(m), pungent(p), spicy(s) and fishy(y) odourded mushrooms are almost always poisonous.

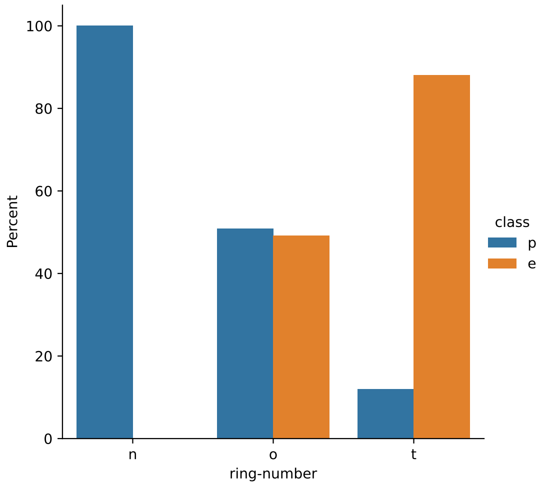

Figure 3: Percentages of ring numbers belonging to class poisonous and edible.

- Rooted(r) mushrooms are always edible.
- Mushrooms having stalk-color-above-ring buff(b), cinnamon(c) and yellow(y) are always poisonous while mushrooms having stalk-color-above-ring red(e), grey(g) and orange(o) are always edible.
- Mushrooms having stalk-color-below-ring buff(b), cinnamon(c) and yellow(y) are always poisonous while mushrooms having stalk-color-below-ring red(e), grey(g) and orange(o) are always edible.
- Mushrooms having veil color brown(n) and orange(o) are always edible while mushrooms with yellow(y)-colored veils are always poisonous.

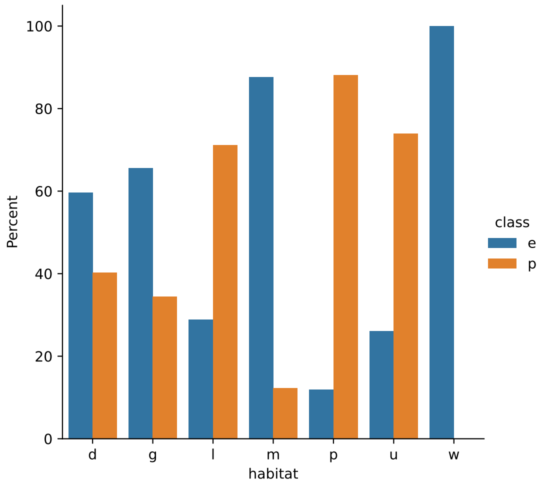

Figure 4: Percentages of habitats belonging to class poisonous and edible.

- Mushrooms without rings(n) are always poisonous. Figure 3
- Large(l) and none(n) ring-typed mushrooms are always poisonous while rings with flaring(f) type are edibles.
- Green(r) spore-printed mushrooms are always poisonous while buff(b), orange(o), purple(u), and yellow(y) are edibles.

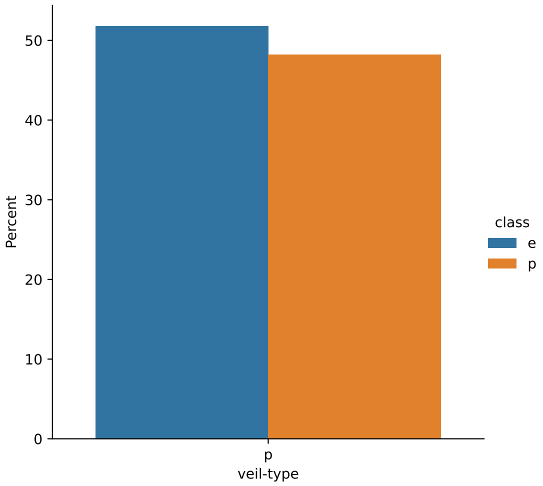

Figure 5: Percentages of veil-types belonging to class poisonous and edible.

- Abundantly(a) and numerously(n) populated mushrooms are always edible mushrooms.
- Mushrooms that exist in waste(w) are always edible. Figure 4
- Veil type has only one unique value p which is predicting edible and poisonous almost same number of times, so this feature is not much informative and we can drop it. Figure 5

In data pre-processing part, one hot encoding is performed on the categorical features. The reason for one hot encoding is that the features contain nominal values. i.e. they cannot be ranked so we cannot perform ordinal encoding. We have performed label encoding on the target column to convert the values to binary form. After performing one hot encoding on features, our feature vector size has increased from 22 to 117. We are not performing PCA at this stage because we are implementing Random Forest Classifier as our baseline model. But later on, in the second stage of the project where we are going to use a blend of supervised and unsupervised learning, we will be performing PCA to reduce our vector space.

2.2 Gender Data Set

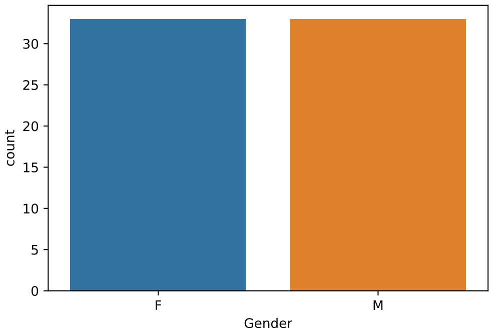

Figure 6: Data Distribution in Gender Dataset.

The second data set we have chosen for our project is a Gender Classification data set. The data is publicly available and was collected in Fall 2015 from university students of 21 nationalities studying various majors in various countries. Its a binary classification problem with a small data set to get an idea whether a person's gender can be predicted on their personal preferences. The data set has only 66 rows and 4 features divided equally among both class. I.e. Both classes contain 33 instances of the data set as shown in Figure 6. Therefore, initially the data set is perfectly balanced with examples from both classes. As seen in Table 2, all features as well as target variable belonging to the Gender data set are categorical in nature. Furthermore, data set contains no null values, but it does contain 4 duplicate rows.

Favorite Color	Favorite Music Genre	Favorite Beverage	Favorite Soft Drink	Gender
Cool	Rock	Vodka	7UP/Sprite	F
Neutral	Hip hop	Vodka	Coca Cola/Pepsi	F
Warm	Rock	Wine	Coca Cola/Pepsi	F
Warm	Folk/Traditional	Whiskey	Fanta	F
Cool	Rock	Vodka	Coca Cola/Pepsi	F

Table 2: Gender Data Set.

During data exploration, we have plotted the categorical features against their percentages in different classes and have gathered some insights from the data.

- Percentage of males who like cool and neutral colors is more than females while females prefer warm colors. Figure 7

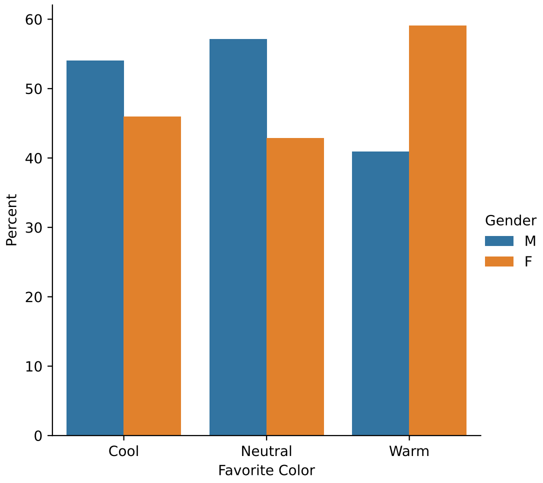

Figure 7: Percentages of Favorite Colors of Males and Females.

- Percentage of male and female liking the music genre folk/traditional is equal. In case of genre rock its also almost equal, while in case of other genres, percentage differs significantly.
- Percentage of Males who prefer to drink beer and vodka is more than females while females prefer wine, whiskey and other beverages. Also, percentage of females who drink is more than males. Figure 8
- 7up/Sprite and Coca Cola/Pepsi are more popular drinks in females while males prefer fanta and other soft drinks.

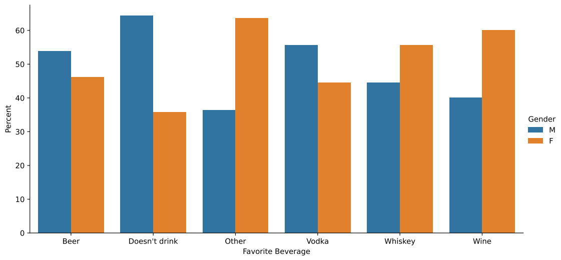

Figure 8: Percentages of Favorite Beverages of Males and Females.

During the pre-processing of the data, 4 duplicate rows were removed and one hot encoding was performed on the categorical features. The reason for one hot encoding is that the features contain nominal values.We also performed label encoding on the target column to convert the values to numerical form. After performing one hot encoding on features, our feature vector size has increased from 4 to 20. Afterwards, we dropped one column from each feature because while creating dummy variables at least two columns are co related. I.e. One variable can predict the values of other variables. This problem is known as dummy variable trap. Now we have a feature space of 16 for 62 rows but we are not performing PCA at this stage because we are implementing Random Forest Classifier as our baseline model. But later on, in the second stage of the project where we are going to use a blend of supervised and unsupervised learning, we will be performing PCA to reduce our feature space further.

2.3 Heart Data Set

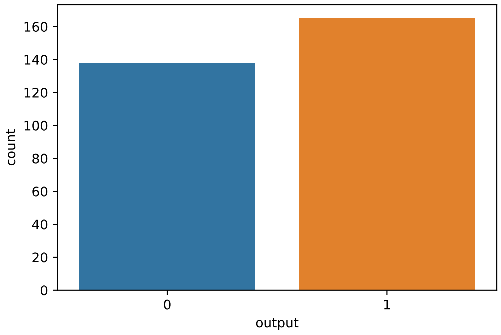

Figure 9: Data Distribution in Heart Data Set.

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Table 3: Heart Attack Data Set.

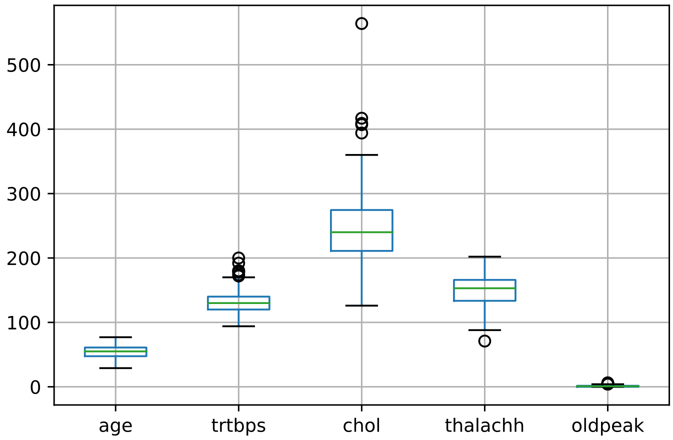

Figure 10: Box plot for numerical features in heart data set.

The third data set we have chosen for our project is a Heart Attack Classification data set. Its a binary classification problem with a medium sized data set used to predict the chances of a heart attack in person based on different features such as age, gender, cholesterol, etc of the person. The data set has 303 rows and 13 features. 165(54.5 percent) instances belong to class 1(more chances of heart attack) while 138(45.5 percent) belong to class 0(less chances of heart attack) as shown in Figure 9. Therefore, due to the negligible percentage of imbalance, data set is considered as

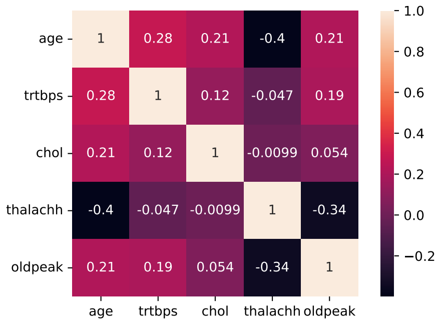

Figure 11: Correlations among numerical features for heart data set.

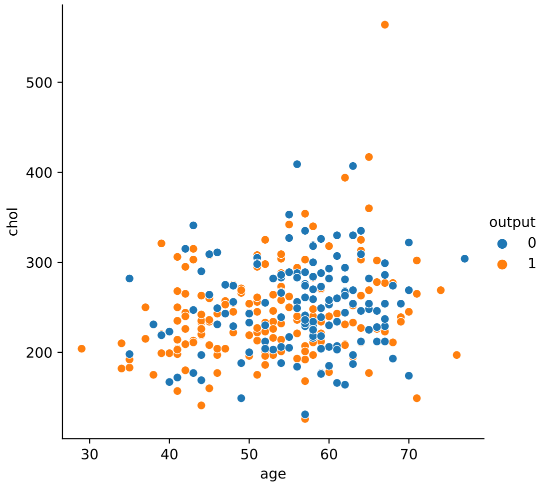

Figure 12: Relational plot among age and cholesterol for heart dataset.

balanced. As seen in Table 3, all features as well as target variable belonging to the Heart data set are appearantly numerical in nature. Furthermore, data set contains no null values, but it does contain 1 duplicate row.

To further understand the distribution of data among numerical features, we plotted a histogram. On observing the histogram in Figure 14, it can be seen that a few features such as: sex, cp, fbs, restecg, exng, slp, caa and thall that appeared to be numerical features initially are actually categorical features. Other features that are numerical in nature seems to have data that is skewed so there is a possibility of outliers present in them.

Boxplots are also drawn to get some insights about the outliers and quartiles of numerical features. From figure 10, it can be seen that resting blood pressure(trtbps), cholesterol, thalachh(maximum heart rate achieved and old peak have outliers present in them but we are not going to remove them as those values are realistic values. For example, cholesterol reaching above 400 is very rare but it happens under different medical circumstances. Correlation are also checked among these features with the help of plot. Thalachh(max heart rate achieved) has a weak negative correlation with all other features while oldpeak, cholesterol, trtbps(resting bloodpressure) and age have weak positive correlations among each other (Figure 11) .We also verified this through relational plots. For example, in Figure 12 the relation plot between age and cholesterol gives no significant insights about the relation between them as they are weakly correlated.

Like numerical features, categorical features are also plotted against their percentages in different classes.

- Chances of heart attack in females(0) is higher than males(1). Figure 13
- Chances of heart attack are high with chest pains (atypical angina(1), non anginal pain(2) and asymptomatic pain(3)).
- Surprisingly, with Exercise Induced Angina, chances of heartattack are low. Figure **??**

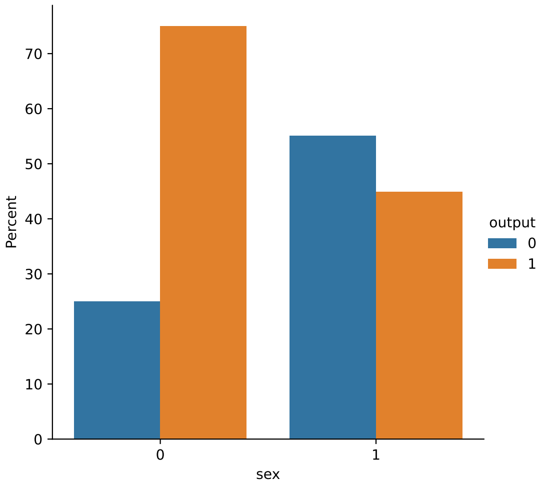

Figure 13: Chances of heart attack based on sex of a person.

- With Resting electro-cardiographic results of type1, chances of heart attack are high, with type 0 they are moderate while with type 2 they are really low.

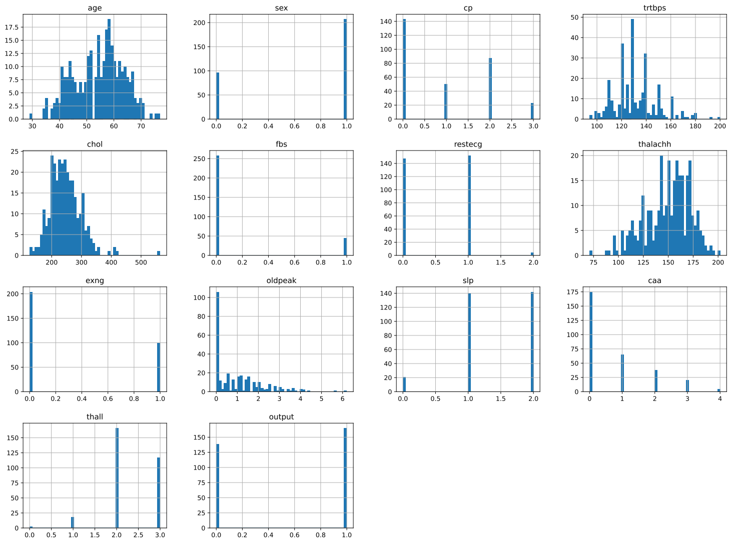

Figure 14: Histogram for heart data set.

At this stage of the project, we are not performing PCA or feature scaling on heart data set as we are using this data set for only Random Forest Classifier initially. All these steps will be performed after implementing the baseline model.

3 Methodology

3.1 Data Sets and Surrogates

To understand the effect of imbalance on data sets, we have chosen three data sets that are initially balanced and performed necessary data exploration and pre-processing steps on these data sets. Furthermore, three surrogates for each data set with low(65 percent) imbalance, medium(75 percent) imbalance and high(90 percent) imbalance are created. To create this imbalance, technique of random sub-sampling is applied to the class we intend to make the minority class. As a result we have now 9 surrogates and three original data sets to proceed with.

3.1.1 Surrogates for Mushroom Data Set

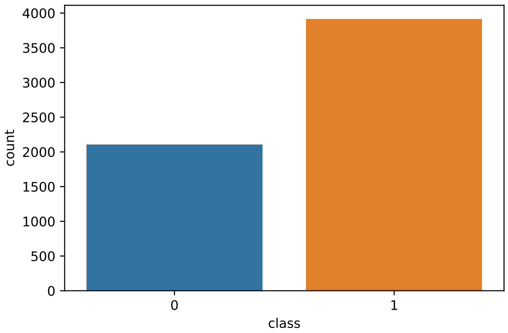

Figure 15: Low imbalance Surrogate for Mushroom Data Set.

- Surrogate for Mushroom Data set with low imbalance has 3916 instances for majority class(p/1) and 2109 instances for minority class(e/0). Figure15

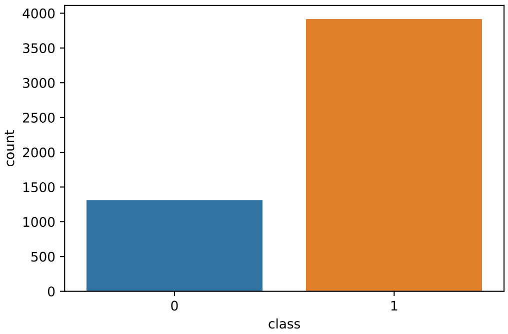

Figure 16: Medium imbalance Surrogate for Mushroom Data Set.

- Surrogate for Mushroom Data set with medium imbalance has 3916 instances for majority class(p/1) and 1307 instances for minority class(e/0). Figure 16

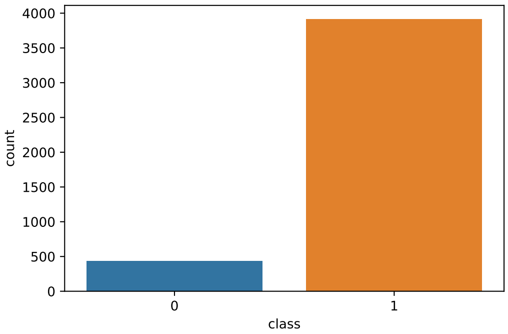

Figure 17: High imbalance surrogate for Mushroom Data Set.

- Surrogate for Mushroom Data set with high imbalance has 3916 instances for majority class(p/1) and 436 instances for minority class(e/0). Figure 17

3.1.2 Surrogates for Gender Data Set

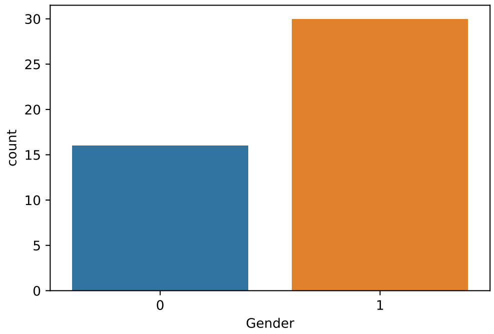

Figure 18: Low imbalance Surrogate for Gender Data Set.

- Surrogate for Gender Data set with low imbalance has 30 instances for majority class(Male/1) and 16 instances for minority class(Female/0). Figure 18

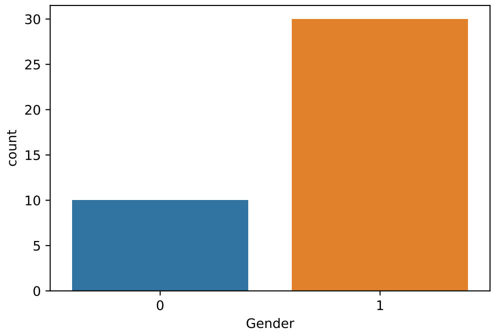

Figure 19: Medium imbalance Surrogate for Gender Data Set.

- Surrogate for Gender Data set with medium imbalance has 30 instances for majority class(Male/1) and 10 instances for minority class(Female/0). Figure 19

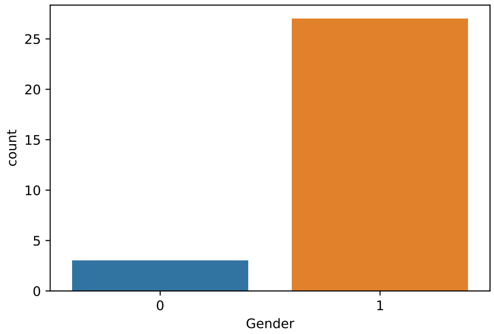

Figure 20: High imbalance surrogate for Gender Data Set.

- Surrogate for Gender Data set with high imbalance has 27 instances for majority class(Male/1) and 3 instances for minority class(Female/0). Figure 20

3.1.3 Surrogates for Heart Data Set

- Surrogate for Heart Data set with low imbalance has 164 instances for majority class 1 and 88 instances for minority class 0. Figure 21

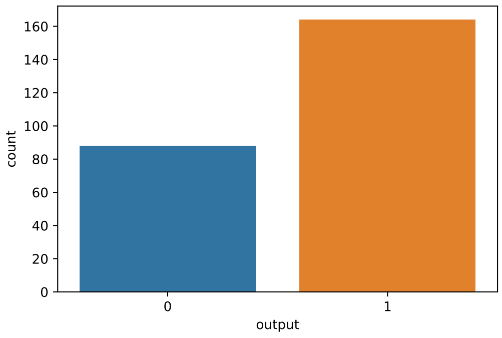

Figure 21: Low imbalance Surrogate for Heart Data Set.

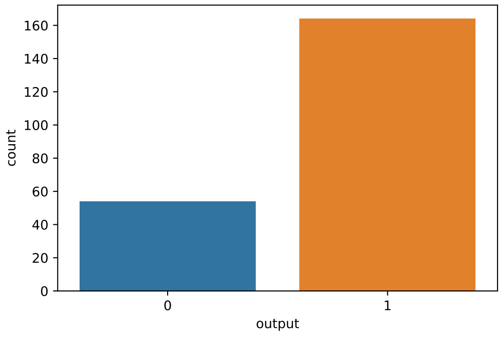

Figure 22: Medium imbalance Surrogate for Heart Data Set.

- Surrogate for Heart Data set with medium imbalance has 164 instances for majority class 1 and 54 instances for minority class 0. Figure 22

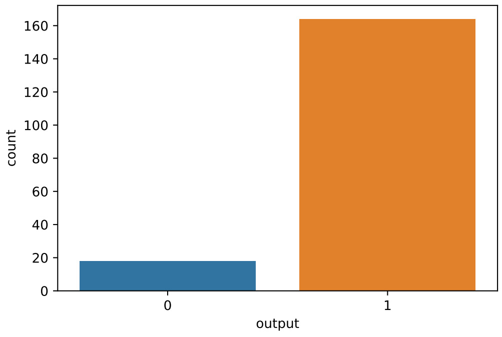

Figure 23: High imbalance surrogate for Heart Data Set.

- Surrogate for Heart Data set with high imbalance has 164 instances for majority class 1 and 18 instances for minority class 0. Figure 23

3.2 Proposed Model

3.2.1 Baseline Model

As a baseline model, we will perform stratified cross-validation on all data sets and their surrogates and train a random forest on them. The reason for stratification is that the ratio of imbalance remains same in every fold formed during cross validation. Metrics used to report the results of our baseline model will be accuracy, precision, recall and F1 scores. These results are going to be compared with the results from our proposed model to check its performance.

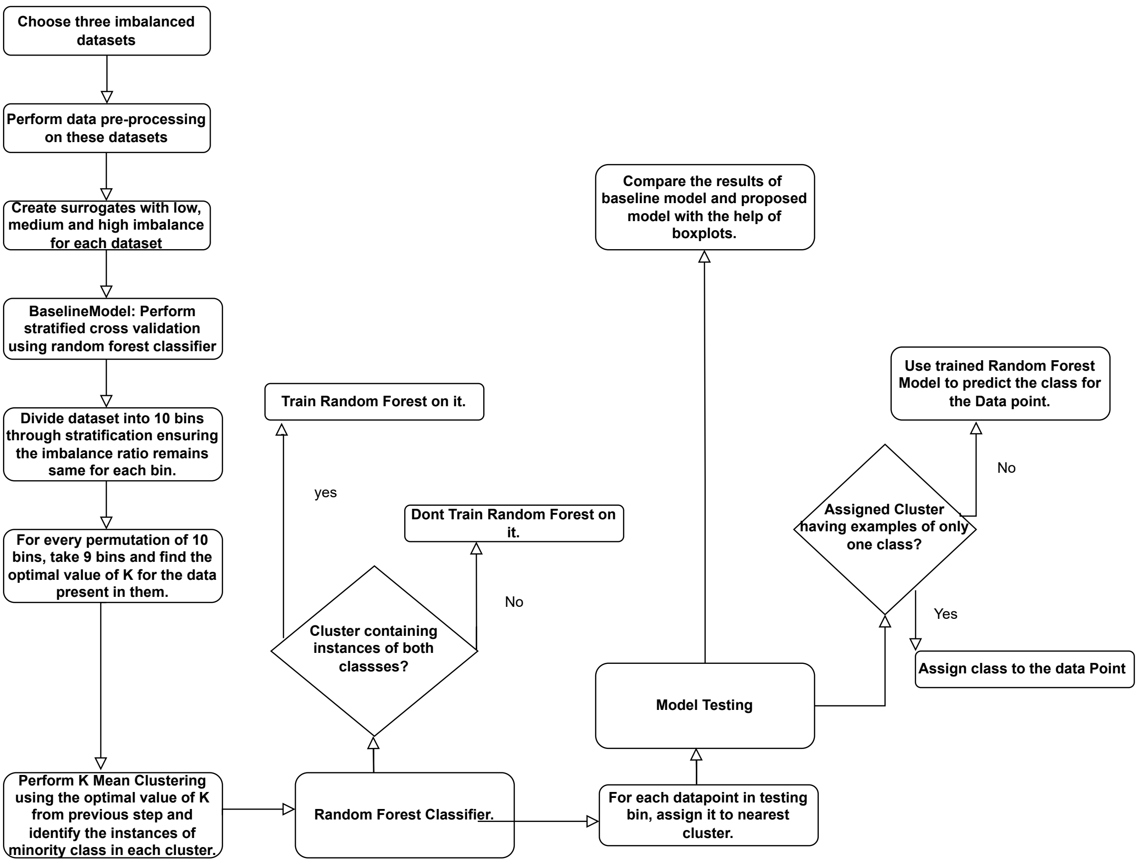
Choose three imbalanced datasets
Perform data pre-processing on these datasets
Create surrogates with low, medium and high imbalance for each dataset
BaselineModel: Perform stratified cross validation using random forest classifier
Divide dataset into 10 bins through stratification ensuring the imbalance ratio remains same for each bin.
For every permutation of 10 bins, take 9 bins and find the optimal value of K for the data present in them.
Perform K Mean Clustering using the optimal value of K from previous step and identify the instances of minority class in each cluster.
Random Forest Classifier.
Cluster containing instances of both classses?
yes
Train Random Forest on it.
No
Dont Train Random Forest on it.
For each datapoint in testing bin, assign it to nearest cluster.
Model Testing
Compare the results of baseline model and proposed model with the help of boxplots.
Assigned Cluster having examples of only one class?
No
Use trained Random Forest Model to predict the class for the Data point.
Yes
Assign class to the data Point

Figure 24: Project Flow.

3.2.2 Bins Creation

After obtaining the results from the baseline model, data for each data set and respective surrogates is divided into 10 parts through stratification to ensure that the imbalance ratio is not disturbed during the creation of these folds.

3.2.3 K Mean

For each permutation of the 10 bins created, we keep one bin for the testing and for the remaining 9 bins, K mean algorithm is run. Optimal number of clusters(k) to be used in K Mean algorithm is identified by applying two different methods. I.e. Elbow method and the Silhouette method. Details about Elbow and Silhouette method are available in [1], [6]. For each cluster, its centroid is calculated and instances of minority class in a cluster are identified and saved. The identification of samples in minority class is going to help in next step.

3.2.4 Random Forest Classifier

Once the samples of minority class in a cluster are identified, Random Forest Classifier is trained on all the clusters except those having zero samples from the minority class.

3.2.5 Model Testing

For every data point in the unseen testing bin, we assign it to the nearest cluster. If the cluster has instances of only one class, we assign that class to the data point otherwise the Random Forest Classifier trained on the data of remaining 9 bins is used to predict the class of data point.

3.2.6 Model Results

After having results for all the permutations, we find the average accuracy, F1, precision and recall scores for our model and compare it with the scores from base model. We will be using box plots to determine up to which extent results for our models differ.

4 Conclusions

Our proposed method is better than the existing methods in a way that it does not manipulate with the data or existing algorithms unlike data re-sampling and ensemble methods. As a result, chances of losing important data are ruled out and the problem of over fitting is avoided. It simply combines the techniques of supervised and un-supervised learning to learn from the imbalanced data sets. The advantage of using this technique is that we are dividing our data into clusters through unsupervised learning and training our supervised model only on the clusters containing data from both majority and minority classes. This aids the model to avoid overlooking the minority class during learning and produce results biased towards the majority class. Also, we are using multiple data sets of different sizes and having different types of features, therefore it is going to produce more realistic results.The advantage of using different data sets is that if our model is not performing well on a specific data set and performing good on other data sets with same amount of imbalance, it will be easy to determine whether the performance of our model is affected by the data imbalance or other factors such as features dimensionality and size of the data sets. Comparisons with baseline model is going to help us to evaluate the performance of our model. Evaluation metrics used are going to help us evaluate whether our model suffered from the data imbalance or not. Using surrogates with different ratio of imbalances is going to find out how much imbalance our proposed model can handle.

References

[1] P. Bholowalia and A. Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

[2] V. Ganganwar. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4):42–47, 2012.

[3] R. Ghorbani and R. Ghousi. Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access*, 8:67899–67911, 2020.

[4] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[5] A. Y.-c. Liu. *The effect of oversampling and undersampling on classifying imbalanced text datasets.* PhD thesis, Citeseer, 2004.

[6] K. Matsushima. The silhouette method. In *Introduction to Computer Holography*, pages 281–308. Springer, 2020.

[7] C. Veni. On the classification of imbalanced data sets. 07 2018.