MGT 6203 Project Final Report

Hal Zhang, Usha Sharma, Brandon Ritchie, Carlos Moncada, Ngoc Le

GitHub Repository: https://github.gatech.edu/MGT-6203-Spring-2024-Canvas/Team-23

Dataset: https://www.kaggle.com/datasets/kamaumunyori/income-prediction-dataset-us-20th-century-data/data

Introduction

In an era where data-driven decisions are paramount, understanding the nuances behind employment and income levels has never been more critical. This project delves into the intricate dynamics of labor statistics to identify factors influencing income levels, particularly focusing on the dichotomy between those earning above and below $50,000 annually. The importance of this study is underscored by the analysis provided by Autor, Katz, & Kearney [1], which highlights the growing wage inequality in the United States and its implications on socioeconomic mobility. Our approach brings a fresh perspective by employing advanced data cleaning techniques, feature selection, and a comparative analysis of various predictive models, addressing the urgent need to understand the forces behind this growing disparity.

Through meticulous data cleaning and preparation, we address challenges such as skewed distributions and inconsistent missing data encoding, ensuring our analysis rests on a solid foundation. By evaluating a suite of models including Logistic Regression, Random Forest, K Nearest Neighbor, and XG Boosting we aim to uncover the most significant predictors of income levels in addition to finding the best model for classifying income. We chose these specific four models because we know that logistic regression and random forest will help us with variable selection while XG Boosting could lead to better accuracy. K Nearest Neighbor serves as a non-parametric model that can also classify income. Our project stands out by not only identifying significant predictors but also by interpreting their implications in the context of current labor market trends. This endeavor is not just an academic exercise; it provides actionable insights for policymakers, economists, and social scientists seeking to understand and mitigate income disparities. Actions into improving income disparity can lead to lasting impacts in people’s lives especially children [2].

Our initial hypothesis is that there will be two different types of factors that contribute to income disparity, uncontrollable and controllable with respect to the person. We predict that less education and less hours worked are factors that a person can control and can lead to lower wages. While we predict that lower age, female gender, being of non-white race, and not being a citizen are factors that a person cannot control that can lead to lower wages. It is important to distinguish between the two, as they require different policy recommendations for change. We also hypothesize that random forest will perform the best on our unbalanced dataset as the model inherently performs bootstrapping which can help balance datasets.

Data Cleaning

Before applying any analytical models, we needed to do preliminary analysis and cleaning of the data. The full dataset had 191,000 observations. We initially analyzed the response variables that we wanted to use, “wage_per_hour”, a quantitative variable expressing hourly salary and “income_above_limit”, a categorical variable expressing whether a person makes a salary greater or less than 50,000 dollars a year. The values of wage per hour were right skewed ranging from 0 to 9999 with a mean of 1398.5. Due to the dataset not providing the units for wage per hour and the abnormal spread, we decided to not use wage per hour for our response variable; however, we are able to use income_above_limit and converted that column to binary values with 1 being people who made above 50,000 dollars. As a result of using income_above_limit as our only response variable, we must discard the test dataset that was provided and split up our training dataset to do training, validation, and testing.

We then cleaned the predictor variables. Our dataset has 43 columns. We cleaned the columns by first looking at which columns have missing values. Unfortunately, this dataset had different values for missing values in different columns, so we changed all the missing values to “NA” to keep everything consistent. Then we looked up all the columns that had more than 20 percent of NA values and eliminated them as we believed the data would be too sparse to interpolate. We were left with 27 predictor variables after taking out those columns.

For the 27 predictor variables left many of them were categorical columns. We went through each and calculated the percent above income threshold to determine which factors were most similar. Factors with similar classifications and like percentages were combined to avoid multi-collinearity in the modelling step. For example, we combined all the categories for education that were under high school (1st grade, 2nd grade, etc.) into one unifying category of “under high school”. To make these categorical variables easy for the models to use, we “one hot encoded” them. Finally, we filtered the cleaned data into two separate data frames of employed and unemployed and will only be using the employed dataset for our analysis as we want to focus on what factors lead to differences in income.

After having a cleaned dataset, we performed more data manipulation for preparation of our analysis. We split the cleaned dataset into training/cross validation and testing datasets. Then we undersampled the majority class of the response variable (income_above_limit of 0) to create a new dataset. In addition, we also oversampled the minority class of the response variable (income_above_limit of 1) which produced another dataset. This resulted in three datasets that we can run our models on. The distribution of the predictor class for each dataset is shown below in Figure 1.

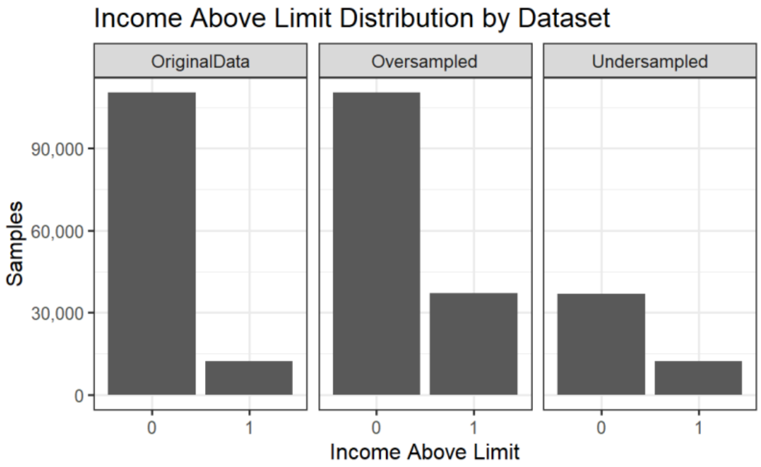


Figure 1: Composition of response variable for the three datasets

Data Analysis

Overview of Modeling

For the data analysis, we evaluated the data on logistic regression, random forest, gradient boosting, and k nearest neighbor to see which ones would produce the best sensitivity, specificity, and balanced accuracy. In addition, these models can be used for variable selection. Logistic regression model provides good interpretability of the results, which can be used for increasing income. The choice of employing a Random Forest model, in particular, was informed by Breiman's foundational work on this methodology [3], which has been noted for its robustness in handling complex datasets like ours. Random Forest's ability to deal with unbalanced data and its importance in variable selection made it an invaluable tool in our analysis, aiming to uncover the most significant predictors of income levels. Gradient boosting was evaluated because of its versatility, high performance, great insights into feature importance and capturing complex relationship between features and response variable. KNN was chosen for evaluation because of its simplicity. KNN does not make any assumptions about data distributions and is robust to noisy data, but has limitations like computational inefficiency. These four models cover a wide range of possibilities for our dataset and provide comparison points which help in analyzing the data.

For all four of the models we used, we ran them on three types of training datasets: original, oversampled and undersampled mentioned in the data cleaning section using income_above_limit as the response variable and the rest of the variables as the predictors. These models were then tested with the same test dataset for comparison.

For each of the models, we used the same comparison points. The details for each are described below.

Logistic Regression

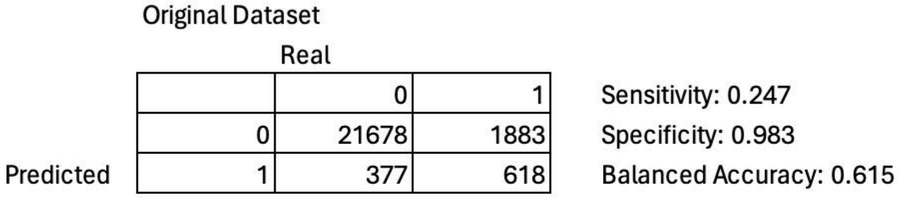
We initially created logistic regression models on the three datasets and compared their sensitivity, specificity, and balanced accuracy.

Figure 2 shows significant variables and their coefficients.

```
##                                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                           -3.330183   0.779920  -4.270 1.96e-05 ***
## age                                    0.035093   0.001486  23.616  < 2e-16 ***
## week_working_hours_above_50            1.431819   0.044161  32.422  < 2e-16 ***
## head_of_house                          0.439572   0.036865  11.924  < 2e-16 ***
## us_citizen                             0.424746   0.094228   4.508 6.55e-06 ***
## educationAssociates                   -2.306073   0.084400 -27.323  < 2e-16 ***
## educationBachelors                    -1.333401   0.073638 -18.108  < 2e-16 ***
## educationHS_Grad                      -3.100372   0.074835 -41.429  < 2e-16 ***
## educationLess_than_HS                 -3.892045   0.096968 -40.138  < 2e-16 ***
## educationMasters                      -0.717032   0.081162  -8.835  < 2e-16 ***
## educationSome_College                 -2.520259   0.075976 -33.172  < 2e-16 ***
## genderFemale                          -1.178819   0.037162 -31.721  < 2e-16 ***
## raceBlack                             -0.380718   0.064597  -5.894 3.78e-09 ***
## raceHispanic                          -0.420334   0.068435  -6.142 8.14e-10 ***
## tax_statusNonfiler                    -2.447870   0.294626  -8.308  < 2e-16 ***
```

Figure 2: Statistically Significant Variables of Logistic Regression

The following shows the confusion matrix for all three datasets and performance parameters.

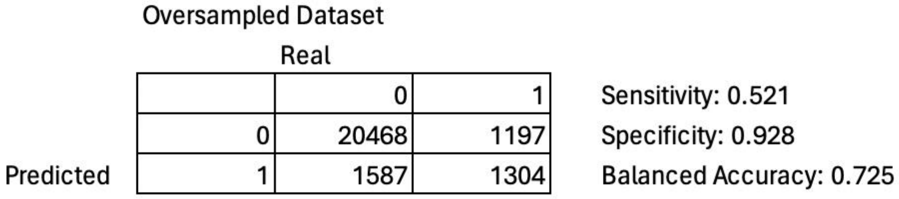


Original Dataset

Real

		0	1
	0	21678	1883
Predicted	1	377	618

Sensitivity: 0.247
Specificity: 0.983
Balanced Accuracy: 0.615

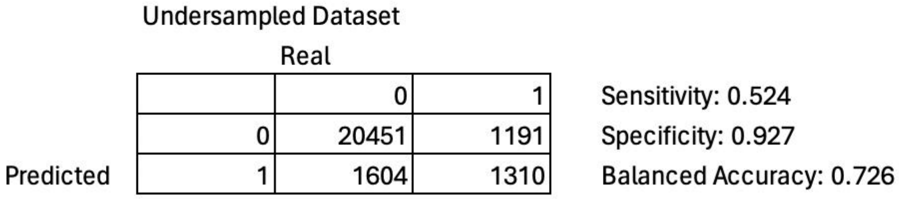


Oversampled Dataset

Real

		0	1
	0	20468	1197
Predicted	1	1587	1304

Sensitivity: 0.521
Specificity: 0.928
Balanced Accuracy: 0.725



Undersampled Dataset

Real

		0	1
	0	20451	1191
Predicted	1	1604	1310

Sensitivity: 0.524
Specificity: 0.927
Balanced Accuracy: 0.726

Conclusion

From the generated models, the outcome is mostly as we hypothesized. Parameters which were found to be most significant are higher education, with the coefficients indicating that higher income is associated with higher level of education. Race and gender are also significant. Females have lower income as compared to males. Non-white races also have lower income as compared to white. Some outcomes are interesting, for example being born in the United States is not significant, even though being a U.S. citizen is significant and it impacts income positively.

Random Forest

We then created random forest models, initially doing hyperparameter tuning with cross validation and found that the best number of variables to select for branching for each decision tree in the random forest (mtry) was 28 (all of our predictor variables). The random forest model was run on the original, oversampled, and undersampled datasets and the confusion matrices for each of the datasets are listed below along with sensitivity, specificity, and balanced accuracy.

Original Dataset

		Real	
		0	1
Predicted	0	21398	1787
	1	657	714

Sensitivity: 0.285
Specificity: 0.970
Balanced Accuracy: 0.628

Oversampled Dataset

		Real	
		0	1
Predicted	0	20420	1087
	1	1635	1414

Sensitivity: 0.564
Specificity: 0.926
Balanced Accuracy: 0.745

Undersampled Dataset

		Real	
		0	1
Predicted	0	20160	829
	1	1895	1672

Sensitivity: 0.669
Specificity: 0.914
Balanced Accuracy: 0.791

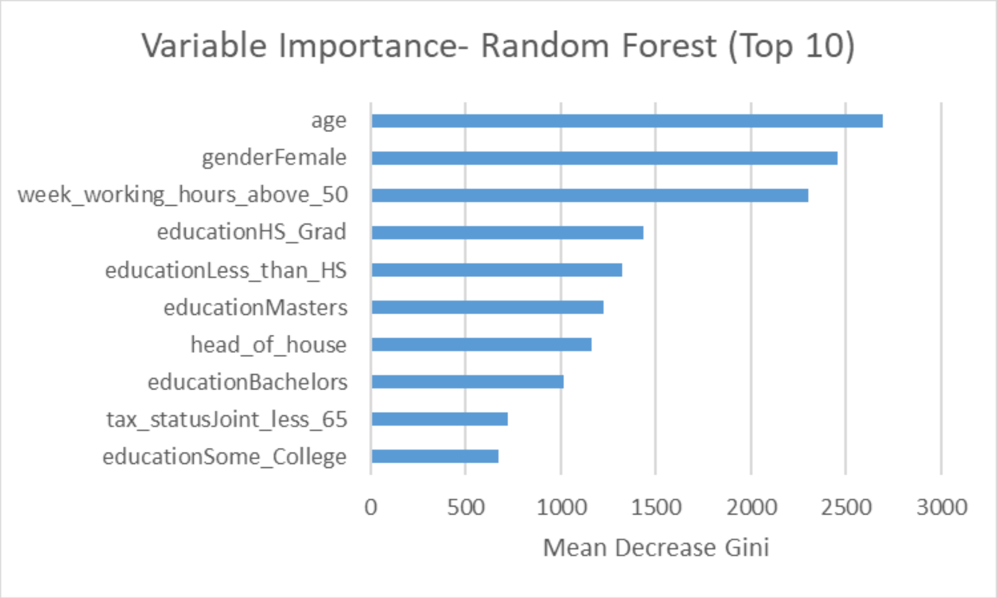


Figure 3: Top 10 Most Important Variables for Random Forest Model

Conclusion:

Even though random forest bootstraps data, it can still perform poorly on imbalanced datasets as can be seen by the poor sensitivity and balanced accuracy on our original dataset. However, with both under and oversampling, not only did the sensitivity increase with a minor decrease in specificity but the balanced accuracy increased. The variable importance shown by the random forest model in Figure 3 were what we hypothesized; the age, gender, number of hours worked, and education play a significant role in the classification of income

K Nearest Neighbors

The K-Nearest Neighbors (KNN) model is a non-parametric method used for classification and regression. It was employed as part of a suite of predictive models to analyze income prediction data from the US 20th-century dataset. This section of the report focuses on the implementation, results, and evaluation of the KNN model, specifically examining its performance on original, oversampled, and undersampled datasets.

Parameter Tuning and Model Configuration

- Cross-Validation Setup: We configured the model to utilize 10-fold cross-validation, enabling us to rigorously evaluate its performance and ensure the reliability of our predictions.
- K Values Tuning: We explored a range of neighborhood sizes (k values) from 3 to 11. This approach allowed us to determine the optimal number of neighbors that provides the best balance between sensitivity and specificity.
- Training Control: The model was set up to save predictions and compute class probabilities, facilitating detailed performance analysis, particularly using ROC metrics as the benchmark.

Model Performance

Original Dataset

		Real	
		0	1
Predicted	0	21636	1930
	1	419	571

Sensitivity: 0.228
Specificity: 0.981
Balanced Accuracy: 0.605

Oversampled Dataset

		Real	
		0	1
Predicted	0	19940	824
	1	2115	1677

Sensitivity: 0.671
Specificity: 0.904
Balanced Accuracy: 0.787

Undersampled Dataset

		Real	
		0	1
Predicted	0	20088	1138
	1	1967	1363

Sensitivity: 0.545
Specificity: 0.911
Balanced Accuracy: 0.728

Analysis

The KNN model performed variably across different data preparations:

- The original dataset showed poor sensitivity, indicating a difficulty in correctly identifying individuals earning above $50,000, likely due to the imbalanced nature of the data.
- The oversampled dataset demonstrated significantly improved sensitivity and balanced accuracy, suggesting that increasing the representation of the minority class helped the model learn better to identify higher earners.
- The undersampled dataset also showed improved performance over the original, although it was slightly less effective than the oversampled in terms of balanced accuracy.

Conclusion

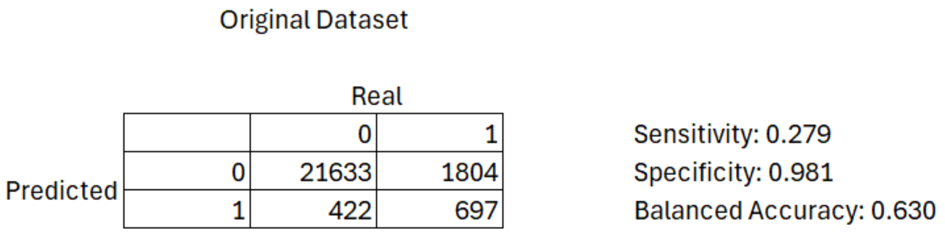
The varying performance of the KNN model across different datasets highlights the challenges of working with imbalanced data. Oversampling the minority class proved to be the most effective method in improving model sensitivity and balanced accuracy, indicating that this approach may be preferable for future model training sessions to enhance predictive performance. While KNN provided valuable insights, its performance also suggests a need for continued exploration of feature selection and potentially combining it with other model types to enhance predictive accuracy.

XG Boosting

The XG boost algorithm is a highly efficient and scalable boosting algorithm. It utilizes L1 (Lasso) and L2 (Ridge) regularization to help prevent overfitting. As mentioned above the model was trained and tuned on three training sets (full training data, over sampled training data, and under sampled training data) using cross validation. Time was spent tuning the hyperparameters on the full dataset. Because of the significant amount of time tuning, these same parameters were applied when training the other datasets.

The tuned hyperparameters include **nrounds** (number of boosting iterations), **max_depth** (maximum depth of each decision tree in the ensemble), **eta** (learning rate for gradient descent), **gamma** (minimum loss reduction required for further partitioning of leaves), **colsample_bytree** (Fraction of features to consider when building each tree), **min_child_weight** (minimum sum of the weights for a child), and **subsample** (subsample ratio of the training instances).

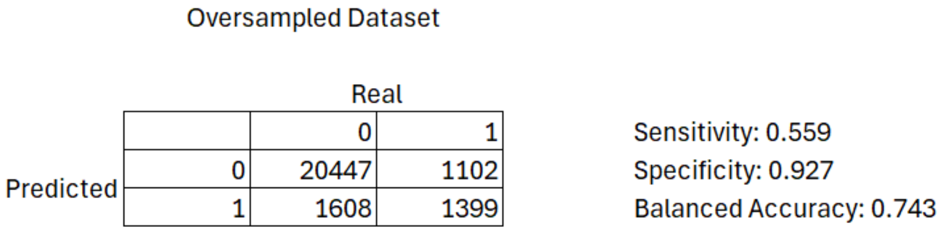
After each model was trained, each model was fitted to the full training dataset (no sampling techniques) and performance metrics were computed as shown below.



Original Dataset

		Real	
		0	1
Predicted	0	21633	1804
	1	422	697

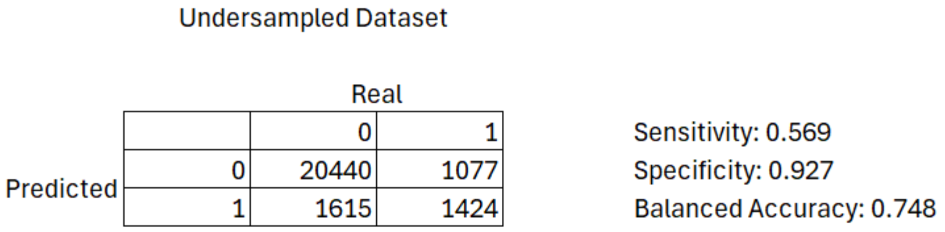
Sensitivity: 0.279
Specificity: 0.981
Balanced Accuracy: 0.630



Oversampled Dataset

		Real	
		0	1
Predicted	0	20447	1102
	1	1608	1399

Sensitivity: 0.559
Specificity: 0.927
Balanced Accuracy: 0.743



Undersampled Dataset

		Real	
		0	1
Predicted	0	20440	1077
	1	1615	1424

Sensitivity: 0.569
Specificity: 0.927
Balanced Accuracy: 0.748

As we can see, the under sampled and over sampled models performed significantly better than the full data model. Specifically, the sensitivity is approximately 70% better using the models trained on under/over sampled data. This means that these models do better at predicting the minority class and would be preferable when identifying the minority occurrence is important.

Discussion

The results of the logistic regression and random forest as seen above showed what we hypothesized, being younger in age, being a female, having less education, working less hours, being a non-citizen, and being of non-white race all led to decreased incomes. This was consistent in both models which is a good sign that these predictor variables are significant. It is surprising how much having less education affects salary (logistic regression coefficient of –3.89) in comparison to being female (-1.18), being black (-0.38). While being a non-tax filer also leads to much lower salary (-2.45); this might be due to the fact that the primary earner for a household usually files the taxes.

Knowing these factors lead to lower wages, it is important for us to think about what we can recommend to policymakers and businesses to bridge this wage gap. For controllable factors, it is important to incentivize education by either paying for the cost of education or guaranteeing a pay raise if a person were to obtain more education. A policy to encourage people to work more hours could be more difficult. Providing better child or pet care could indirectly lead to people with families to be able to work more hours.

The uncontrollable factors, however, will require specific interventions by policymakers and businesses to ensure there is no discrimination. Sometimes, controllable factors can have indirect effects on uncontrollable factors. For example, college students of age 18-22, have less time to work and therefore leads to decreased salaries. These correlations are something that we can study in more detail in the future with more data provided.

While it is crucial to know what factors lead to wage imbalances, it is also important to have an accurate model to be able to classify a person's salary based on their characteristics. To compare the overall performance of all of our models, we decided to use balanced accuracy as it is a good measure to compare models for classifying unbalanced datasets. The concern for classifying unbalanced datasets is that the model can just predict the same result (the majority class) and obtain a very high accuracy however the model would do a poor job predicting the minority class. In our case, the sensitivity (the true positive rate) would be low given the increased number of false negatives. The balanced accuracy averages specificity and sensitivity, thereby considering the number of false negatives and false positives. The balanced accuracies for all of our models were around 0.6 for the original dataset; the specificities were really high but sensitivities were low. To deal with that, we oversampled the minority class and undersampled the majority class producing two additional datasets. As shown in our results, this significantly improved the sensitivity and balanced accuracy without sacrificing too much of the specificity. Out of the four models we performed on three separate datasets, the random forest on the undersampled data performed the best with a balanced accuracy of 0.791. Even though the random forest was the best model in terms of prediction, the logistic model is the most interpretable and can produce coefficients that show us the magnitude of difference a variable makes. XG Boosting and K Nearest Neighbor also did an excellent job in making predictions but like the random forest model, they are difficult to interpret.

Conclusion

We started off the project with the goal of determining which factors, controllable and uncontrollable, led to income inequality and which data analytic model would be able to classify our income dataset the best. Along the way, we learned about how unbalanced datasets could skew the accuracy of models and overcame this by using balanced accuracy to compare our models. We also under and oversampled our datasets, which improved the balanced accuracy of our models. Finally, we concluded that as we hypothesized, less education, being female, being of non-white race, being a non-citizen, and being younger in age all led to lower salaries. It was surprising to us how much having less education affected salary more so than the other variables. After obtaining this information, we can recommend to policymakers and businesses to have incentives for obtaining more education and making sure there is no salary discrimination against females, non-white races, and non-citizens. Businesses and policymakers can use the random forest or logistic regression model to classify people and identify who needs help from intervention the most. Studies we could do in the future that can lead to improved recommendations is to use other datasets in order to find if any controllable factors such as education, number of hours worked, and other variables have correlations with uncontrollable factors such as age, race, citizenship status as this could allow us to make even more specific recommendations.

Works Cited

1. Autor D, Katz L, Kearney M. Trends in U.S. Wage Inequality: Revising the Revisionists. Rev Econ Stat. Available from: https://scholar.harvard.edu/lkatz/publications/trends-us-wage-inequality-revising-revisionists. Accessed March 13, 2024.
2. Ratcliffe C, Kalish E. Escaping Poverty: Predictors of Persistently Poor Children's Economic Success. Urban Inst; May 2017. Available from: https://urban.org.
3. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
4. Birla S, Kohli K, Dutta A. Machine Learning on imbalanced data in Credit Risk. In: Proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON); 2016 Oct; Vancouver, BC, Canada. IEEE; 2016. p. 1-6.
5. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5:221–232. Available from: https://doi.org/10.1007/s13748-016-0094-0.