CSC <535/635> Data Mining

Evaluation of various Classification Methods using a Blood Transfusion dataset

Submitted to:

Dr. Jamil Saquer

Author(s):

*Normandy Bryson*

**Abstract**

Classification methods are a primary tool used for data mining in an increasingly data centric world. While the theoretical learning of classification methods often present data as normally distributed or evenly split most data in real world applications of interest are not. In fact, most areas of interest utilizing a data mining technique suffer from skewed data or imbalanced data. The research paper aims to present a review of current research on differences between merely skewed data and imbalanced data by first defining each term and second by discussing in brief terms various methods available to solve such issues. The second half of the paper will involve an in-depth look at a dataset which involves skewed data and the results obtained from various classification methods in sklearn. These methods will include results which do not attempt to correct the skew and methods which do. Overall, four different classification methods will be compared.

**Introduction**

Data mining techniques, specifically classification methods, have been available for quite some time to researchers and industry leaders seeking to maximize their own information gain. However, the advent of a data centric society, where each year we are capturing more data than the year previous, provides both great opportunity and struggles primarily because real-world data captured may not follow a normal distribution or may have class outcomes where one class is heavily represented compared to another class.

Representativeness is an important characteristic in classification because the model is affected by skews and imbalances in the data. According to Google’s Machine Learning course, “a classification data set with skewed class proportions is called imbalanced” [1].The website goes on to explain a mild imbalance occurs when 20-40% of the data is the minority class, moderate imbalances occur when 1-20% of the data is the minority class, and extreme when <1% of the data is the minority class. The issue with such an explanation however comes from the problems big data bring when utilizing classification techniques.

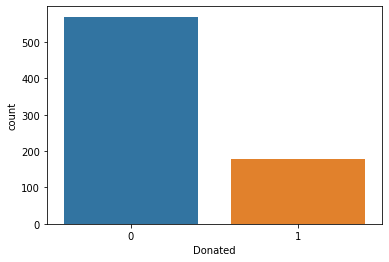
The primary argument during this portion of the paper is that while all imbalanced datasets are also skewed not all skewed datasets are imbalanced. The implication here, is that from the above explanation when talking about imbalanced and skewed instead of one problem, what we really have is two problems. The fundamental issue with combing these problems as one is that as big data progresses and class imbalances get larger newer concepts of tackling these imbalances will have to emerge.

As, [2] note “starting as a problem of skewed distributions of binary tasks, this topic evolved way beyond this conception.” One issue in attempting to make a distinction between skewed and imbalanced is no formal definition has been provided by researchers. For instance, in 2 papers we are provided with 4 different measures of imbalance. In, [3] researchers first claim a “high-class imbalance (i.e., a majority-to-minority class ratio between 100:1 and 10,000:1” and then later “any class imbalance (e.g., 50:1) level that makes modeling and prediction of the minority class a complex and challenging task can be considered high-class imbalance by the domain experts” [3]. In the next paper by, [2], the first definition is provided as “class imbalance concentrate on imbalance ratios ranging from 1:4 up to 1:100” though later the authors claim the focus is really “with problems on imbalance ratio ranging from 1:1000 up to 1:5000” [2]. Compared to the Google definition from above none of these works seem to provide a clear resolution to what is or isn’t imbalanced.

Yet imbalanced can be easily differentiated from skew in that with skew “even if the disproportion is high, but both classes are well represented and come from non-overlapping distributions we may obtain good classification rates” [2]. However, as those definitions from above make clear, on imbalanced ratios, there is a lack of representativeness in the data such that rarity existence. This rarity problem also known as the class imbalance problem occurs “when the class of interest is relatively rare as compared with other class(es)” [4]. There is also an extremeness with class imbalance problems such that, especially in big data, the “mal-effects can be felt much more severely due to the extreme degrees of class imbalance” [3]. The primary take-away from such information is imbalance problems involve rarity and extremeness while skewness may not.

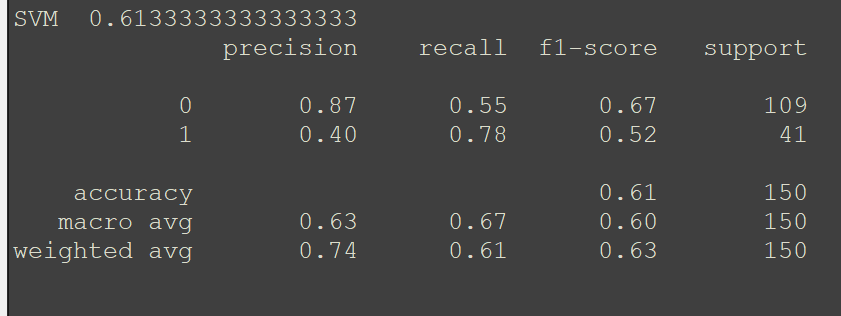
Current literature proposes for both skewed and imbalanced data three possible methods for obtaining better results. Both skewed data and imbalanced may use “over- or under- sampling a process to take the different cost between classes into account” [5]. Both problems may also use algorithm-leveling methods and/or hybrid methods. Hybrid methods involve the combination of data-leveling methods with algorithm-leveling methods while algorithm-leveling methods such as cost-sensitive approaches apply a higher value or cost to the underrepresented class so that “we boost its importance during the learning process (which should aim at minimizing the global cost associated with mistakes” [2]. Unfortunately, greater in-depth discussion of methods applied to imbalance problems, problems current researchers face with imbalance problems, and future directions for solving imbalance problems is more than this paper is designed to speak toward. The purpose of providing such an in-depth discussion over imbalanced versus skewed shows current terminology may not be completely solidified. Along with lack of clear terminology it is likely higher imbalance ratios will create a need for newer methods. While current methods for approaching both imbalance and skew problems are the same there is no guarantee this will be so in the future. To provide greater insight into classification methods the rest of this paper will highlight the evaluation of classification methods using a skewed dataset.

**Experimental Setup and Results**

The current experiment utilized the Blood Transfusion Service Center Data Set found at [6]. The data consists of 4 features R (Recency - months since last donation), F (Frequency - total number of donation), M (Monetary - total blood donated in c.c.), and T (Time - months since first donation) all of which are considered real valued attributes. The outcome variable was a binary classification consisting of 1 for donating blood and 0 for not donating blood during March 2007. No pre-processing was required during the initial run of this experiment and an 80/20 train-test split was used.

For the initial evaluation of four classification methods no skewed correcting methods were used. As indicated from the output class the skew was more heavily toward 0 (not donating) at 76% versus 1 (donating) at 24%. Accuracy results for each of the 4 classification methods shows SVM (74%), Logistic Regression (73%), Naïve-Bayes (75%), and KNN (75%). However, using an accuracy measure is not appropriate since in a “highly skewed data distribution, the overall accuracy metric is no longer sufficient” [4]. Since accuracy is no appropriate for discovering how well our model provides insight let’s look at other measures commonly used such as precision, recall, and f1.

Looking at Support Vector Machines first:

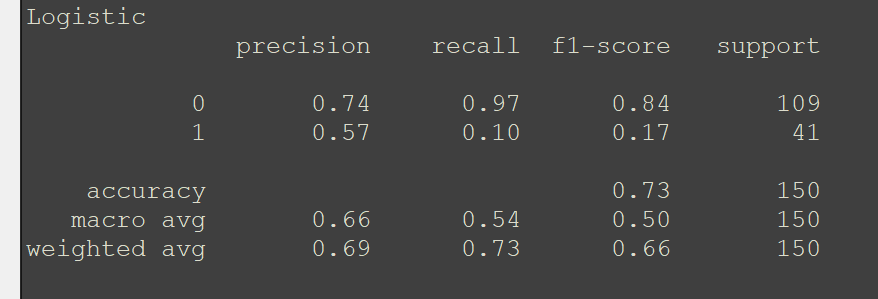


When we examine SVM through the classification report provided by pandas we find for precision SVM is adapt at correctly classifying 0, the dominant class, when it is considered the positive class. However, when SVM is required to correctly classify 1, the minority class, the model takes a nosedive and only reaches 40%. This is not surprising as precision is affected by skew.

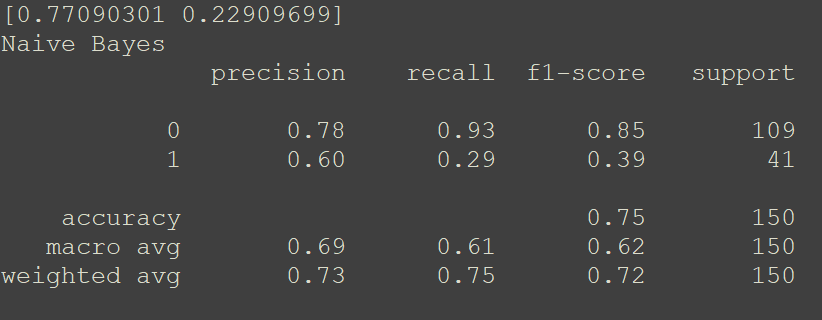
Examining recall, which corresponds to the percentage of positive samples correctly identified, we see a flip-flop, it is important to note here other algorithms discussed later will not show the same flip-flop behavior as SVM, with 0, the dominant class, obtaining a 55% recall and 1, the minority class, obtaining a 78% recall. The reasons for this flip-flop are not clear but it is likely such behavior affected the f1 score outcome.

Generally, speaking when relying on precision and/or recall we are often better off looking at f1 score as well. The f1 score provides greater clarity on the ability of the model as compared to precision/recall because precision/recall are only capable of capturing either FP or FN but not both whereas f1 will capture both FP and FN. The f1 score can also be viewed as a more realistic measure, or maybe more pessimistic measure, since the value of f1 will often be closer to the smaller value between precision/recall. We can see for 0, dominant class, precision was 87%, recall 55%, resulting in a f1 score of 67%. For class label 1, the minority class, precision was 40%, recall 78%, resulting in a f1 score of 52%. Not surprising the model performed better in terms of f1 score when classification required a label 0 versus a 1.

Next looking at Logistic Regression:

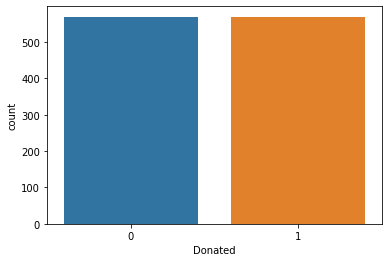


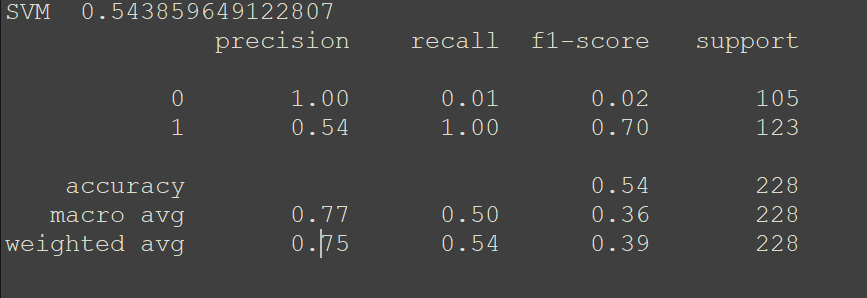
For our Logistic Regression model, regarding precision when the positive class was 0, the dominant class, precision rate was 74%. However, when the positive class was 1, the minority class, precision rate drops to 57%. In terms of recall we find for 0 we obtain a recall rate of 97%, extremely high especially since recall is not affected by skew. When examining recall for 1 though we see an extreme drop down to 10% meaning our model using Logistic Regression was only able to correctly classify samples with a class label of 1 ten percent of the time. Regarding the f1 score we see for 0, the dominant class, a result of 84% and for 1, the minority class, a result of 17%. Here again our model appears to have overfitted to the 0 label rather than learning a model for which to classify unknown samples.

Looking at Naïve Bayes:

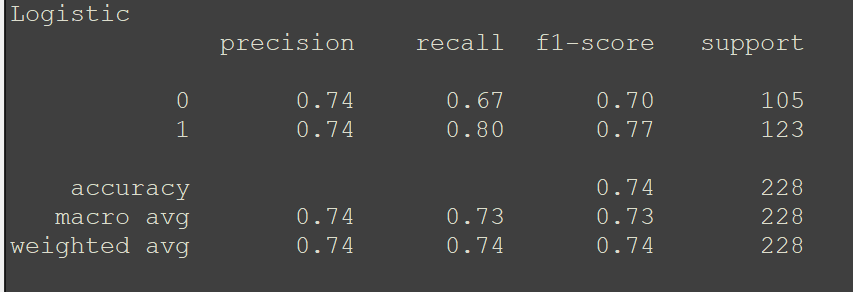
For Naïve Bayes we find much of the same regarding the previous two models. An overfitting to class label 0 resulting in a model which is unlikely to classify unknown samples with any sort of usability. An important aspect shown in the picture is the a-prior sample percentage shown in the top-left. We can clearly see during the process of running Naïve Bayes it is sampling with a distribution rate similar to what is found in our sample population.

The previous results provide a glimpse into utilizing a model which has been built from a skewed distribution. The result is we end up with a model not usable for any real meaningful classification. If this is truly the problem, then what does our results look like when we sample in such a way as to create equal distribution. Pandas provides the option of up-sample or down-sample. Both will be down but only up-sampling will be discussed as down-sampling had similar results as up-sampling. Up-sampling is a method of taking the minority sample and through sampling with replacement create two equal class label outcomes with the number of samples based on the dominant class. After performing up-sampling we can see our distributions are equal:

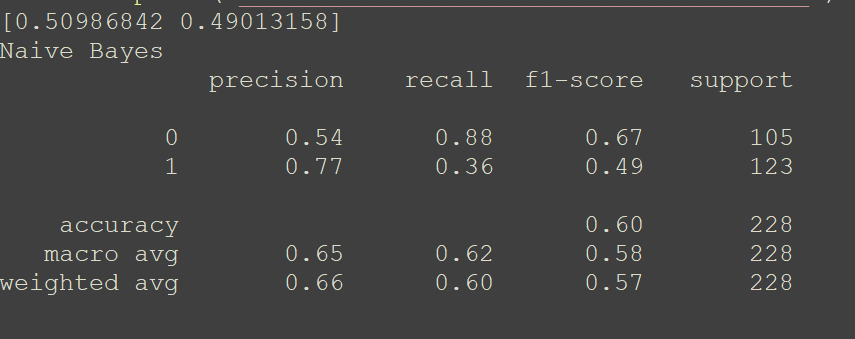
How will equal class distributions impact our various evaluation measures regarding classification models. Unfortunately, for accuracy little has changed as SVM (54%) results in a reduction in total accuracy when compared to the skewed distribution, Logistic Regression (74%) results in no change from one distribution to another, Naïve Bayes (60%) is another drop in accuracy when compared to the skewed distribution, and KNN (72%) slight reduction in accuracy compared to the skewed distribution. Yet when we examine precision, recall, and f1 score for our various evaluation model we will see a substantial difference when compared to the skewed distribution.

When looking at SVM:

For SVM we find when the positive class is 0 a precision rate of 100%, a clear sign something is likely wrong with our model, especially when we examine the recall for class label 0 is 1% and the f1 score is 2%. When the positive class is 1 a precision rate of 54% is obtained which is better than the skewed distribution regarding SVM, we find a recall rate of 100%, again another clear sign something is off with this model, and a f1 score of 70%.

When looking at Logistic Regression:

For Logistic Regression we find a much more usable model, or at least more realistic in terms of the behavior the model is espousing. When the positive class is 0 a precision rate of 74% is obtained which is the same as with our skewed distribution, a recall rate of 67% a substantial reduction compared to the skewed distribution, and a f1 score of 70% which is again a reduction when compared to the skewed distribution. When the positive class is 1, however, we find a better model when compared to the skewed distribution such that for precision we have a 74%, recall 80%, and f1 score of 77%. In every class when we up-sampled the once minority class performed substantially better.

When looking at Naïve Bayes:

The Naïve Bayes model performed more balanced compared to the skewed distribution model, but results did not provide a usable model. One interesting aspect for Naïve Bayes is it kept the class distribution similar to what our balanced distribution looks like as found in the top left.

As noted above, down-sampling was also conducted but results obtained were similar to what was found using up-sampling. These results provided better performance on the various classification measures of precision, recall, and f1 score for the minority class but did not result in an overall better model when compared to skewed distribution or up-sampling distribution.

**Conclusion**

All this information taken together points to a couple important aspects regarding data mining. One, while the initial distribution of data was skewed no amount of up-sampling or down-sampling provided a greater model for which to classify unknown samples. This is likely because the current features do not provide great enough insight such that a function can be made to map input into correct output label. It may have been more beneficial to engage in up-sampling or down-sampling at the same time as removing certain features but due to the scope of this paper such a combination was viewed as not beneficial.

Two, results for the various classification methods tend to cluster together not only in a generic way but also in terms of the real number values obtained. In terms of generic clustering I mean results which were bad tend to stay bad even when using other methods and in terms of real numbers each method tended to reflect this generic clustering by providing real number values substantially less than what is typically desired for a usable model.

Last, definitional ambiguity regarding imbalanced and skewed is a huge problem which shows no signs of being solved any time soon. I see little reason for not classifying the current research data set as being imbalanced other than some arbitrary cutoff of 4:1 provided by a previous paper. With an increasingly data-based society growing most problems of interest to researchers and industry involve the minority class and with imbalance problems likely to keep growing this problem needs newer solutions.

References

[1] Imbalanced Data. (n.d.). Retrieved from <https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data>.

[2] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. doi: 10.1007/s13748-016-0094-0

[3] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, *5*(1). doi: 10.1186/s40537-018-0151-6

[4] Tang, Y., Zhang, Y.-Q., Chawla, N., & Krasser, S. (2009). SVMs Modeling for Highly Imbalanced Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(1), 281–288. doi: 10.1109/tsmcb.2008.2002909

[5] Gadi, Manoel & Lago, Alair & Mehnen, Jorn. (2010). Data Mining with Skewed Data. 173-188. doi: 10.5772/9382.

[6] Yeh, I-Cheng, Yang, King-Jang, and Ting, Tao-Ming, "Knowledge discovery on RFM model using Bernoulli sequence, "Expert Systems with Applications, 2008, <https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center>.