Learning to work with imbalanced datasets

Adetya Aggarwal

An imbalanced classification problem is a problem where the ratio of target classes is highly skewed or biased. this poses a problem as most of the predictive classification problems were designed with the assumption that there is equal distribution among classes with maybe slight variances. When using these algorithms with highly imbalanced data, the resulting models perform poorly specially with classes which are most under-represented in the dataset.

In this project I have tried to learn about some of the most popular data manipulation techniques used in the industry to rectify the problem.

**Data**

I tried to model credit card fraud system where the model would be able to classify transactions as fraud or legal. The dataset is highly imbalanced as the number of fraud transactions in the dataset are only about .001% of the total transactions.

**Techniques**

*Under Sampling* – Under Sampling is a procedure where the size of the most represented class (also called the majority is class) is reduced while the size of the minority class is kept the same and equally sized classes are created.

*Over Sampling* – In over sampling, the number of events of minority class are randomly duplicated and added to the training dataset.

*Over-Under Sampling* – In over-under sampling, both over and under sampling is used by reducing the size of the majority class and oversampling the minority class, creating a balanced dataset.

*SMOTE -* Synthetic Minority Oversampling Technique creates synthetic examples of the minority class in the dataset from existing minority class. Examples are selected that are close in the feature space.

**Algorithms**

Different algorithms were used with the augmented data using the techniques above. I selected algorithms based on popularity, simplicity, and complicity. The algorithms used were -

* Naïve Bayes
* Decision Tree
* Random Forest
* Neural Network

**Results**

For this analyses, accuracy wasn’t chosen as a parameter because as the dataset is highly imbalanced, accuracy would be high as most of the majority class would be predicted correctly. However, as we are more interested the model’s capability to predict minority class hence Precision, Recall and F1-Score were calculated.



**Conclusion**

The results are very subjective in nature. If the goal is precision, that is, how much of the fraud predictions were correct then, random forest with original untouched dataset gives the best results. However, if the goal is recall, that is, of the total fraud transactions, how many were caught, then a Neural Network with SMOTE Sampling provides the best results.

Overall, Random Forest used with untouched original data provides the best F1 score. This speaks volumes as to how good random forest algorithm is and hence is extensively used in the industry.

Graphs

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated