Analysis of Machine Learning Models

**Summary**

In Module 17, we were given sample sets of code to walk thru each module and observe the machine learning behavior. These analyses were based on supervised data. We had a specific dataset that was either large or small, and wanted to determine an outcome based off the provided info.

For the challenge, we had the opportunity to evaluate the following machine learning modules and determine best model for looking at credit risks. Below is a table of the models and the results.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tests | Balanced Accuracy | Pre | Rec | Spe | F1 | Geo | iba |
| Random oversampling | 0.66 | 0.99 | 0.58 | 0.74 | 0.73 | 0.66 | 0.42 |
| SMOTE | 0.65 | 0.99 | 0.68 | 0.62 | 0.81 | 0.65 | 0.43 |
| Under sampling | 0.55 | 0.99 | 0.41 | 0.68 | 0.58 | 0.53 | 0.27 |
| Combo | 0.55 | 0.99 | 0.57 | 0.72 | 0.72 | 0.64 | 0.4 |
| Random Forest | 0.99 | 0.99 | 1 |  | 0.99 |  |  |
| Easy Ensemble | 1 | 0.99 | 1 |  | 0.99 |  |  |

The precision is very good on all of these tests. However, a deeper review of the data and information about the classification report reveals important information to factor in for our best model choice. Below is a brief description of the model and whether or not the model should be used.

Oversampling

In **random oversampling,**instances of the minority class are randomly selected and added to the training set until the majority and minority classes are balanced.  These model was the most balanced and the precision was the same as all the others.

SMOTE

The**synthetic minority oversampling technique (SMOTE)** is another oversampling approach to deal with unbalanced datasets. SMOTE did not perform as well as the random sample. Although the balance is close at .65, it did not perform as well in all of the other sections. It could be useful as a second option to the oversampling model.

Undersampling

In random **undersampling**, randomly selected instances from the majority class are removed until the size of the majority class is reduced, typically to that of the minority class. This model does not have an accuracy score or F1 that deems it useful.

Combo

**Combo** or SMOTEENN is applied, instead of SMOTE. As with SMOTE, the minority class is oversampled; however, an undersampling step is added, removing some of each class’s outliers from the dataset. The result is that the two classes are separated more cleanly. Although this process cleaned up the data significantly, the accuracy dropped significantly as well. This drop in accuracy deems the model unusable.

Random Forest

Instead of having a single, complex tree like the ones created by decision trees, a **random forest** algorithm will sample the data and build several smaller, simpler decision trees. Each tree is simpler because it is built from a random subset of features. This model does not appear to have enough high risk credit data to build a model that is useful.

Easy Ensemble

**Ensemble** learning builds on the idea that two is better than one. A single tree may be prone to errors, but many of them can be combined to form a stronger model. A random forest model, for example, combines many decision trees into a forest of trees. This model does not appear to have enough high risk credit data to build a model that is useful.

**Conclusion**

I would not use any of the aforementioned samples. Although each of the models serve their own purpose and attempt to deal with the data differences in a way that will lead us to a decision, it seems that we need a better balance of data.

However, the random sampling and SMOTE appear to factor in some of deficiencies of our original dataset and balances out the information a little better, making our model more realistic. It allowed us to see the more of the high risk information compared to the rest of the data. These models come the closest to the type of information we’d like to work with.