Project IV Monte Carlo Simulation - Spotify Dataset

Ipsita Bisht
Professor Alfaro-Córdoba
STAT 155
23 May 2025

1. Question

In this simulation, I would like to look at how the presence of hit songs affect the model's performance in calculating a hit or not-hit song. It is difficult to find a hit song within the dataset after running the ANN model on it, giving a low precision and recall score compared to non-hit songs. Due to this limitation, I wanted to design this experiment to assess how the model can perform alongside the RandomForest model when given different levels or proportions of hit songs within the dataset. Will their performance improve with a higher presence of hit songs? Or will it remain unchanged? For this simulation, we will be using the ANN model from Project 3 as well as the Random Forest model and use data generation techniques to mock up datasets with certain levels of hit songs.

2. Data

In order to test the models against these different levels of hits, a dataset needs to be generated for each of the three levels specified. The data-generating model that is used in this simulation takes in parameters as follows:

- n (number of samples)
- hit_proportion (the level of hit songs we want in the generated dataset)
- numeric_features (features to be simulated for dataset)
- genres (genres to be simulated which are already pre-selected)

In the generate_data() function, we take these parameters and create a new dataframe which contains the genres and numeric features and assigns songs to be hits based on the hit_proportion value. Using the np.random function, we generate the values for the numeric features based on the scale already defined (0.0 - 1.0) using a uniform distribution. Same goes for the genre features and we use the np.random.choice() function to simulate this for the n observations we want to generate.

3. Estimates

Due to the imbalances in this dataset, looking at the model's accuracy in predictions is misleading so in place of that, my simulation will estimate the precision, recall, and f1-score across 20 iterations. These metrics provide more insight and a higher quality evaluation of the model's predictions. Precision measures the proportion of true positives among all positive predictions. Recall helps measure the proportion of true positives among all actual positive predictions. The f1 score provides a harmonic mean of both precision and recall.

$$Precision = True\ Positive / True\ Positive + False\ Positive$$
$$Recall = True\ Positive / True\ Positive + False\ Negative$$
$$F1 = 2 * Precision * Recall / Precision + Recall$$

4. Methods

This simulation evaluates how the Artificial Neural Network and Random forest models perform in classifying hits versus non-hit songs under varying proportions of hit songs within the datasets generated.

- Artificial neural network: This model was chosen because of its ability to detect non linear relationships between multiple features however it is sensitive to class imbalances making it's performance not the best for identifying rare classes
- Random Forest: This model was chosen for its ability to handle imbalanced data and is not sensitive to feature scaling

These models will be assessed on the metrics mentioned earlier (Precision, Recall, F1 score) as we change the proportion of hit songs within a dataset. To ensure a proper evaluation, the models will be trained on datasets with the following levels of hit-rates:

- 1%
- 5%
- 10%

I hypothesized that with more hit songs in the dataset, the model's performance will improve across all metrics. With more hit songs, the model should have a better chance at learning the patterns needed to differentiate a hit from a non-hit.

5. Performance Criteria

The models' performance will be measured across different criteria across all iterations and levels of hit-rates. This is done through calculating the mean and standard deviation of:

- Precision
- Recall
- F1-Score

Accuracy will not be assessed as it is misleading in binary classification problems such as this one. Based on these performance criterias, the ideal hit rate ratio can be identified.
By aggregating results across 20 iterations, we will get an estimate of expected performance and its variability under each hit proportion scenario.

6. Simulation Plan

In order to run this experiment, there are two key parts that will be carried out. The first part is to generate the datasets needed based on the different levels assigned earlier. The next part is to iterate through the different levels and to pass it to the different models while aggregating the results to later display in a table

To simulate the dataset, there will be 20 iterations with 5000 samples in each dataset that will be curated. The parameters for these models will be fixed throughout the iterations to ensure a consistent environment.

For each simulation step:

1. Generate a new data set
2. Train both the ANN and RandomForest models
3. Test on the train-test split datasets
4. Calculate the precision, recall, and f1-score on the hit class
5. Record metrics

7. Anticipated Challenges or Limitations

The main challenge was figuring out how to balance the dataset given that the target class is the minority. This required the use of oversampling methods in the preprocessing steps to account for those minority classes. However, there are still issues with the sampling of the minority class and thus further methods will need to be researched and incorporated for the final product