Machine Learning to Guess the Genre of a Movie

Dylan Schafer

I want to create a program that can predict the genre of a movie based on runtime, English name, original name, rating, number of votes, and release year. I think I can get this to be 80% accurate. Some movies have multiple genres, and to help increase my accuracy if the program can guess one of the main genres that counts as guessing the genre of the movie.

The first parameter that I had to modify was the chunk size of the data that I was loading into my memory. This helped with the memory allocation, so that the computer continued running smoothly and was able to process the data in more manageable pieces instead of all of it at the same time. The next big parameter that was modified us the Random Forest classifiers. These parameters define the range of values that were explored during hyperparameter tuning. The class weight parameter was added to adjust the model to handle class imalances, which I thought would be helpful since there was great imbalances in data related to some genres and others. After running some tests, the balanced mode ended up being the most accurate, so I didn't end up implementing this idea further. There is also some precaution coding at the beginning, to confirm that there aren't any movies that have missing data. This wasn't an issue since the data we were using was pretty clean to begin with.

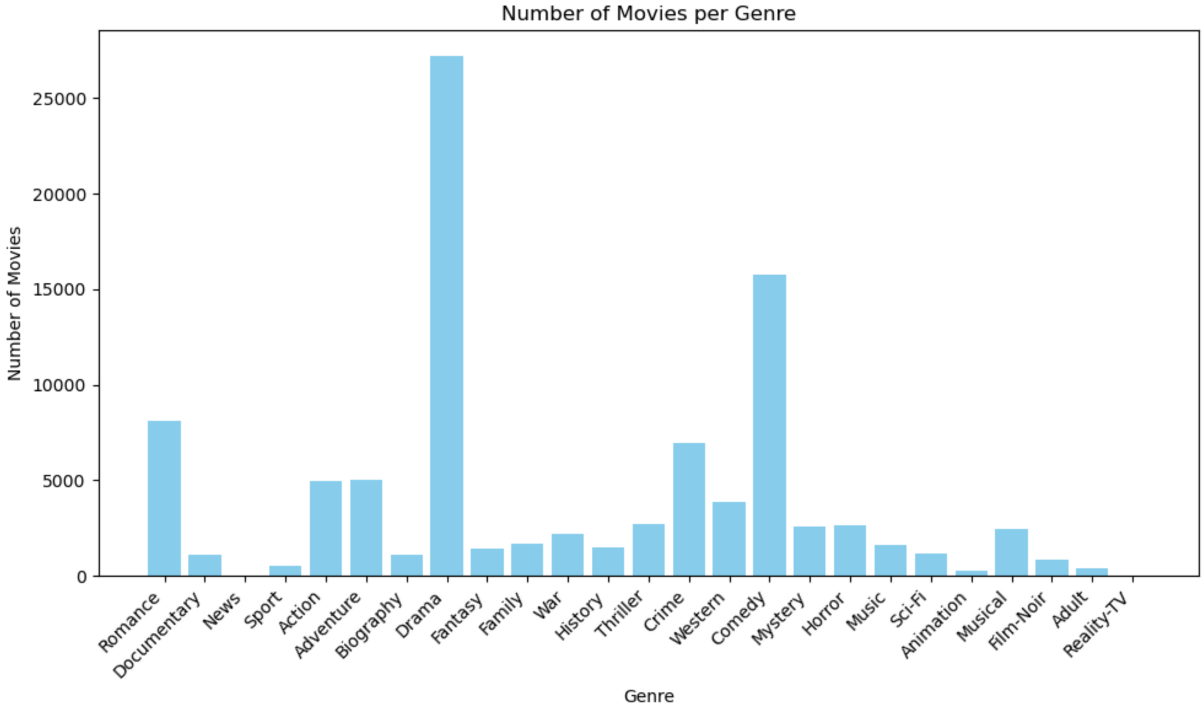
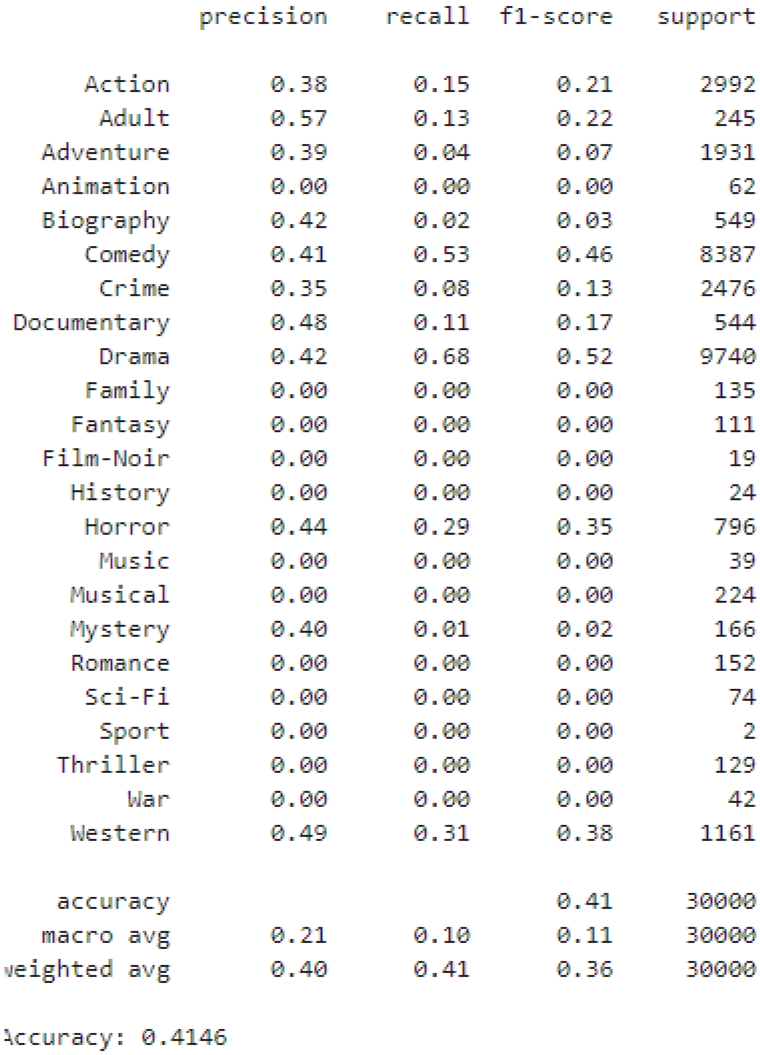


Fig 1. A bar graph of the number of movies per genre

During coding, some of the challenges were faced. The data was already in pretty good shape, but there were some problematic parts that I had to deal with. Some of the movies had multiple genres listed. I had to write some code to split the genres within one block of data. This didn't take too much difficulty, about one line of code was written to deal with this. After generating a list of all the genres, I created a bar graph to get some insight into what I was dealing with. I noticed that many of the genres didn't have very many movies associated with it.



```
              precision    recall  f1-score   support

      Action       0.38      0.15      0.21      2992
       Adult       0.57      0.13      0.22       245
   Adventure       0.39      0.04      0.07      1931
   Animation       0.00      0.00      0.00        62
   Biography       0.42      0.02      0.03       549
      Comedy       0.41      0.53      0.46      8387
       Crime       0.35      0.08      0.13      2476
 Documentary       0.48      0.11      0.17       544
       Drama       0.42      0.68      0.52      9740
      Family       0.00      0.00      0.00       135
     Fantasy       0.00      0.00      0.00       111
   Film-Noir       0.00      0.00      0.00        19
     History       0.00      0.00      0.00        24
      Horror       0.44      0.29      0.35       796
       Music       0.00      0.00      0.00        39
     Musical       0.00      0.00      0.00       224
     Mystery       0.40      0.01      0.02       166
     Romance       0.00      0.00      0.00       152
      Sci-Fi       0.00      0.00      0.00        74
       Sport       0.00      0.00      0.00         2
    Thriller       0.00      0.00      0.00       129
         War       0.00      0.00      0.00        42
     Western       0.49      0.31      0.38      1161

    accuracy                           0.41     30000
   macro avg       0.21      0.10      0.11     30000
weighted avg       0.40      0.41      0.36     30000

Accuracy: 0.4146
```

Fig 2. The Accuracy of the program per genre

This lack of data for certain genres was a large challenge for the code, since the program wasn't able to generate a lot of connections between the other data and these smaller genres. For these genres, the program scored very low. This tanked the overall prediction average. For fun, I removed all these points from the data and achieved an 80 percent accuracy. I decided not to include this in the final code, because there were only a handful of genres left, which made the prediction so much easier.

After creating the program and running it for 30 minutes, I was only able to achieve. 40%. I overestimated my abilities with python, and ended up struggling with the coding. I also underestimated how difficult it would be to predict the genre of a movie. I thought that there would be a heavy correlation between the ratings and time period a movie came out, and the genre of said movie. There were also many genres of movies that only had a couple of movies associated with it, and these movies were extremely difficult to predict since the genre in general had less data associated with it. The most predictable genre ended up being drama, which makes sense since this was the largest genre.