DSC 680 – MILESTONE 2: NIC CAGE MOVIE RATING PREDICTION

TRISH GIRMUS

6/1/2022

Business Problem

Academy Award winner Nicolas Cage has had a successful career in acting since his first movie role in 1982. Over the span of his career, the movie star has acted in 104 movies across multiple genres. Those movies have attained various ratings. Five new movies are slated to debut later this year through 2023, and two that are TBA. Based on the prior ratings of movies, the business problem is to predict the movie rating for these five new movies based on its title. Logistic regression modeling will be used to predict the movie ratings.

Background/History

Nicolas Kim Coppola was born on January 7, 1964 and is the nephew of famous director Francis Ford Coppola. He changed his name to Nicolas Cage so he would not be associated as a relative of the Coppola family. This occurred at the beginning of his acting career. His first role was a comedy television pilot titled *The Best of Times* in 1981. In 1982, Cage had a role in *Fast Times at Ridgemont High*, his first movie debut (this is the only movie where his last name is Coppola in the credits) (IMDb.com, Nicolas Cage). Cage earned a Golden Globe nomination for Best Actor – Motion Picture Musical or Comedy for *Moonstruck* in 1987. 1996 was a good year for Nic Cage as he won an Academy Award, Screen Actors Guild (SAG) Award, and Golden Globe Award for his performance in *Leaving Las Vegas* (wikipedia.org). In 1998, Cage was inducted into the Hollywood Walk of Fame. In addition to acting, he has also added the titles of director and producer to his credits (wikipedia.org).

Data Explanation

Secondary data for this project was collected through the Imdb movie database using the Cinemagoer package in Python. Nicolas Cage's person ID #0000115 was selected, and the key 'filmography' was chosen. From the main DataFrame, the movie ID was then appended. The DataFrame was then saved into a .csv file, and a backup copy was also created. Overall, the data contained 32 features and 109 rows. All data used for the project was categorical and numerical.

Below is a list of the 32 features and the definition of each.

Name [feature_name]:

- Title [title] – movie title
- Rating [rating] – moving rating
- Genres [genres] – movie genre
- Votes [votes] – total votes per movie
- Languages [languages] – language movie aired in
- Runtimes [runtimes] – how many minutes movie ran
- Countries [countries] – countries movie aired in
- Plot Outline [plot outline] – outline of movie plot
- Director [director] – movie director
- Year [year] – year of movie airing
- Writer [writer] – movie writer
- Producer [producer] – movie producer
- Composer [composer] – movie composer
- Cinematographer [cinematographer] – movie cinematographer
- Editor [editor] – movie editor
- Editorial Department [editorial department]
- Casting Director [casting director] – movie casting director
- Art Direction [art direction]
- Costume Designer [costume designer] – movie costume designer
- Make Up [make up] – people who did make up for movie
- Production Manager [production manager] – movie production manager
- Assistant Director [assistant director] – movie assistant director
- Art Department [art department]
- Sound Crew [sound crew] – movie sound crew
- Special Effects [special effects] – names of people who did special effects for movie
- Visual Effects [visual effects] – names of people who did visual effects for movie
- Stunt Performer [stunt performer] who performed stunts in movie
- Production Companies [production companies]
- Distributors [distributors]
- Special Effects Companies [special effects companies]
- Other Companies [other companies]
- Original Air Date [original air date] – original air date of movie

Methods

While analyzing the data, it was discovered that most of Nic Cage's movies were categorized by multiple genres in alphabetical order. A histogram was created of the individual genres. Those genres were then contained in a dictionary indicating the instances of movies per genre. As Figure 1 shows, there were 21 genres total. The top five genres with the most instances of movies were 'Thriller' at 33, followed by 'Drama' at 29, 'Action' at 27, 'Comedy' at 22, and 'Crime' at 18. A function was then created to assign the genre with the most instances from the group of multiple genres by replacing it as that only instance. Once this was completed, the next step was to find the top genres with the most instances for prediction. By using assumption and approximation, the decision was made to extract genres with 10 or more instances per movies only. This is because it was ideal for training and testing with more data to work with. The top genres with 10 or more instances per movies included: 'Comedy', 'Crime', 'Drama', 'Thriller', 'Adventure', 'Romance', 'Fantasy', 'Action', 'Horror', and 'Mystery'.

It was also discovered during data cleaning that the movies with premiere dates for 2022, 2023 and TBA were also missing other values that were necessary for prediction. Unfortunately, this data would not be sufficient to use for predicting ratings and as a result, it was removed from the main DataFrame. After removing, the dataset contained 33 features and 104 rows.

The next step of data cleaning entailed adding numeric fields for the columns [runtimes] and [year] (the added fields were created as [runtimes_n] and [year_n]). Then, the ratings needed to be reclassified. As the average across all movies was a 6.0 (this was concluded from the histogram), it was determined that this would be the threshold as to whether a movie was a success or a failure. A new column was created as [reclassified_rtg]. If the rating was equal to or greater than a 6.0, the function would return a 1. If the rating was below a 6.0, it would return a 0. The genre was then reclassified. As mentioned earlier, the multiple genres were listed in alphabetical order (versus an order of most genres listed first, next, last, etc.) in most of Cage's movies. The genre 'Thriller' had the most instances with 55, followed by 'Drama' with 28, then 'Comedy' with 11, 'Action' with 8, 'Fantasy' with 1 & 'Horror' with 1.

A list of six fields were then created for regression: [title], [reassigned_genre], [votes], [runtimes_n], [year_n], and [reclassified_rating]. The 'Drama' genre was determined as the genre to be used for prediction. This decision was based on the genre containing the second highest number of instances and using 'Thriller' might appear to be biased. Using any genre lower than 'Thriller' or 'Drama' would not work for modeling with a smaller number of instances. Since there was no longer any data for prediction (it had to be removed as it had too many missing values in the predicted fields), a subset of four movies was selected from the 'Drama' genre and removed from the main dataset. The selection of these four movies was a subjective decision. The movies that were chosen included: *The Runner, City of Angels, It Could Happen to You, and Deadfall*. After the four movies were removed, the main dataset contained 6 features and 100 rows. A numeric ID column was added which was then a new field for the

DataFrame index. The [reassigned_genre] column was dropped as it was determined that only one genre ('Drama') would be used. It also did not contribute anything to the prediction accuracy.

Ranges were then created for numeric features [votes], [runtimes], and [year] as these would be used as predictors. Percentiles from the histogram were used (minimum, 25, and 75). This was to ensure an objective method was used to reassign the numeric values to categorical variables (0, 1, 2). For the DataFrame of the four predicted movies, features [votes], [runtimes], and [years] were reclassified.

Analysis

Logistic Regression modeling was used for analysis. A train and test split set were created (for the genre 'Drama') to help prevent overfitting. Nine movies from the test set were randomly generated for prediction each time the model ran. For fields that were dropped, the [reclassified_rating] feature was added. The test data set size was 35%, and the training data was 65%. True values were already known in the test dataset for the rating. The model would then try and predict those. Using np.ravel (this is to avoid an sklearn error), the next step was to fit the model. From there, the model was used to predict the test data from the train data. A confusion matrix was then created for ratings in the test dataset from the predicted dataset. The accuracy of prediction showed whether the test dataset had a rating greater than the average of 6.0. The model ran 10 times, and as seen in the bar chart in Figure 2, the mean model accuracy was 72.5%, and the variance of model accuracy was 3.40%.

When the model was executed on the dataset to predict the test data, the backup copy of the original dataset was used to see the movie title. This was because [title] was dropped when the model was created. Another confusion matrix was created to determine the prediction accuracy. As seen in Figure 3, this model also ran 10 times, with a mean prediction accuracy of 71.40% and a variance of prediction accuracy of 1.13%.

Conclusion

The means for both model and prediction accuracies showed favorable results and conclusions can be drawn. While predicting nine out of 15 movies may seem like a stretch on a project like this, predictions in life may be made on something much less. It is important to note however, that a conclusion cannot be made running the model only one time. Using the subset of movies from the genre 'Drama' showed confidence that a predicted rating is higher than the average rating, which was 6.0 (when ran 10 times).

Assumptions

It was assumed when starting this project that a movie rating could be predicted based on missing values. While working through the project however, it was discovered that the business problem would have to be revised. Another assumption was the data itself. Once more analysis was conducted, it wasn't known that the genres, for example, were grouped together

by movie and in alphabetical order. This goes back to understanding the data and ensuring that decisions aren't made that could affect the outcome of modeling.

Limitations

Time was a limitation for this project as there were plans to do more with prediction modeling. Unfortunately, the time that was invested with the project focused on the prior business problem and once that was discovered, a lot of reworks was involved. Also, by changing the business problem I had to use existing movie ratings for the prediction. I would have liked to have used a different approach if I had more time to research.

Challenges

The topic of this project was enjoyable, and the data was easier to comprehend versus past projects. As I've usually worked with Linear Regression modeling on prior projects, learning a new type of regression modeling was time-consuming for me. It was also a challenge working with categorical data and converting it, as I haven't worked with that type of data in over a year. It was good to step outside of my comfort zone and try a new model, but there wasn't enough time to really do the analysis and modeling that I wanted to do. I know practice in future projects will expedite the time it normally takes by having more experience.

Future Uses/Additional Applications

Knowing that I have a good model for predicting movie ratings, a future use might be to try the model with a different actor and find additional features for prediction. Another idea with the existing dataset would be using the feature [votes] instead of [rating] to predict. Or, instead of reclassifying and assigning the genre with the most instances, use any movie regardless of genre.

Recommendations

It is recommended to iterate the existing model at least 100 times to see what other results occur. I would also recommend using additional genres and running the model with those. Another type of modeling might be beneficial if trying to predict other features, such as sentiment analysis. Or possibly, using a time series study to see how many movies Nic Cage films a year and look at that as a new business problem. There are a lot of options with this type of data, which makes projects fun to do.

Implementation Plan

The first step in the implementation plan is to run the model 100 times as suggested in the recommended section. The next step would be to add more genres and do modeling for those.

Ethical Assessment

This project was conducted for informative and entertainment purposes only, by sharing results of the predicted rating accuracy. Since this is not damaging to Nic Cage's career, I do not feel there are any ethical implications working on this based on those results. While a lot of assumptions were made with the data, I tried to think objectively when making decisions. As I continue to work more with data, I understand that those decisions can have an impact on the focus of the study, or those who participated in it.

APA References

IMDb.com. (n.d.). *Nicolas Cage*. IMDb. Retrieved from https://www.imdb.com/name/nm0000115/bio?ref_=nm_ov_bio_sm

Wikimedia Foundation. (2022, March 27). *List of awards and nominations received by Nicolas Cage*. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/List_of_awards_and_nominations_received_by_Nicolas_Cage

Wikimedia Foundation. (2022, May 26). *Nicolas Cage*. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Nicolas_Cage

Data References

Cinemagoer. PyPI. (n.d.). Retrieved from https://pypi.org/project/cinemagoer/

Used to retrieve data from IMDb database for Nicolas Cage person ID: 0000115

Appendix

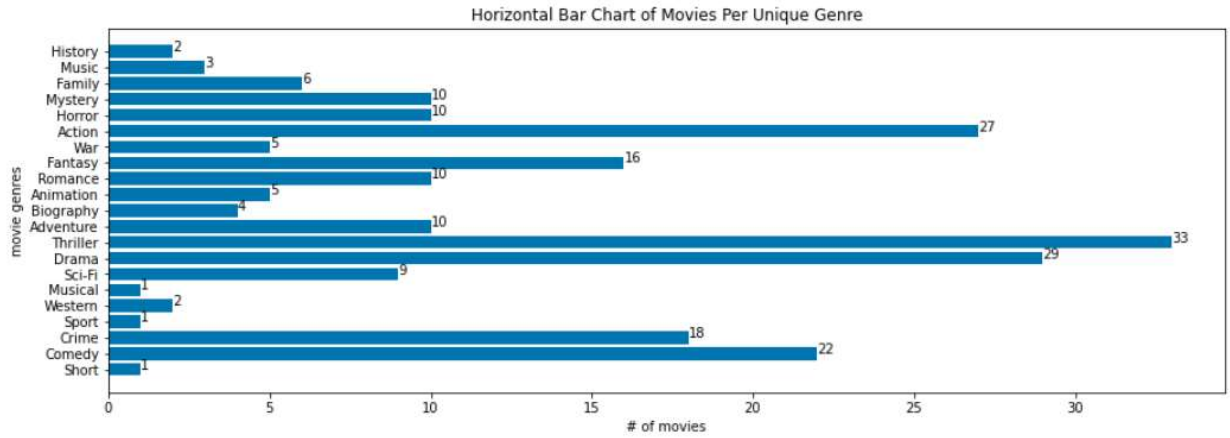
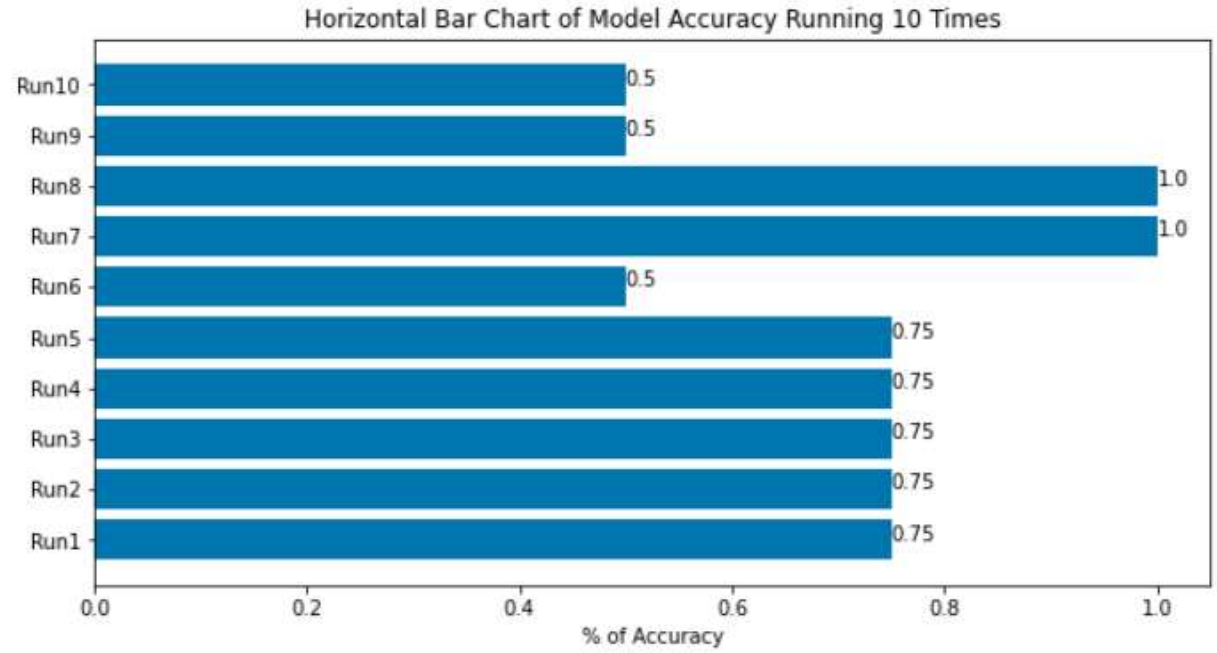
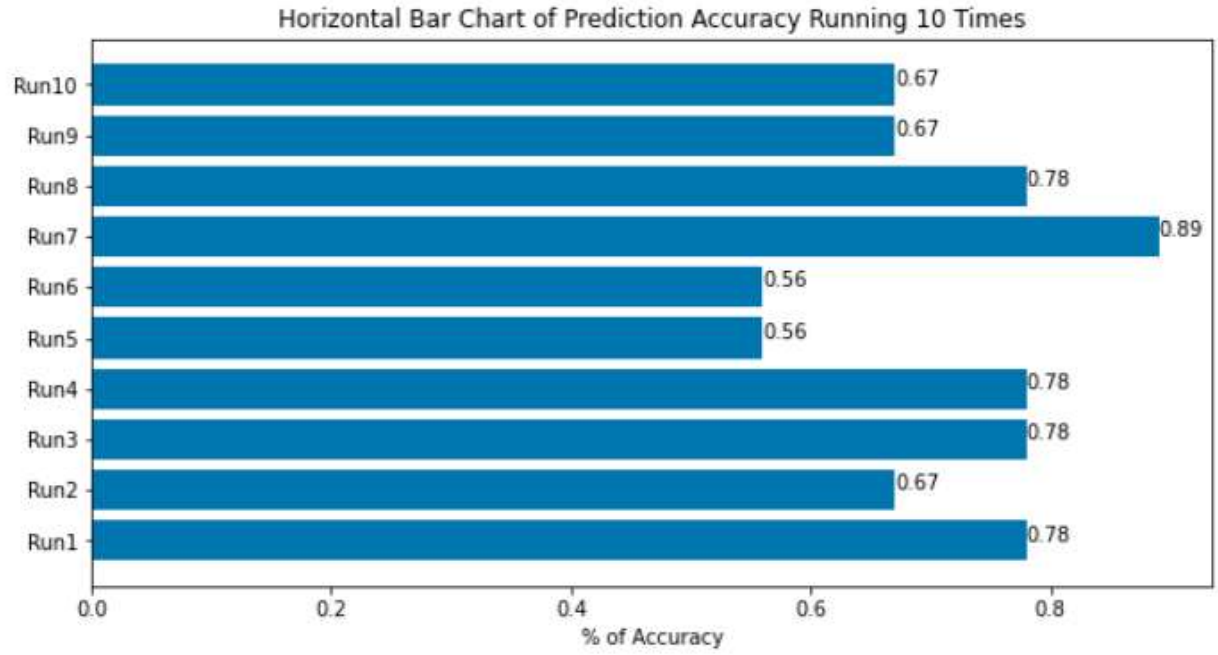


Figure 1- Bar Chart of How Many Movies Per Unique Genre




```
Mean model accuracy: 0.725
Variance of model accuracy: 0.034027777777777775
```

Figure 2- Bar Chart of Model Accuracy When Running Model 10 Times




```
Mean prediction accuracy: 0.7140000000000001
Variance of prediction accuracy: 0.01129333333333333
```

Figure 3- Bar Chart of Prediction Accuracy When Running Model 10 Times

10 Questions an Audience Would Ask You

1) Why did you choose Nic Cage for this project?

2) Will the accuracy always be the same?

3) How did you choose the attributes for modeling?

4) What did you do with the invalid data?

5) Do you feel the size of the data set was adequate for modeling?

6) Why did you select movie title as predictor for project's original idea?

7) What factors made you choose specific movies? Did you use all the movies or certain categories?

8) How did you reassign and trim the genres? What is the reason to limit the number of genres to that particular instance?

9) How did you reclassify numeric values?

10) How did you decide to pick the four movies?