Exercise Machine Learning

Consider a dataset where each line represents an actor or actress for a specific movie. The dataset includes the following information:

- **Unnamed: 0:** Index or identifier
- **tconst:** Movie ID
- **ordering:** Position of the actor in the movie
- **nconst:** Actor ID
- **category:** Actor or Actress
- **characters:** Character played by the actor (nconst) in the movie (tconst)
- **primaryTitle:** Primary title of the movie
- **originalTitle:** Original title of the movie
- **isAdult:** 0 for not an adult film, 1 for an adult film
- **startYear:** Start year of the movie
- **runtimeMinutes:** Movie runtime in minutes
- **genres:** Movie genres
- **primaryName:** Actor or Actress primary name
- **birthYear:** Actor or Actress birth year
- **deathYear:** Actor or Actress death year
- **primaryProfession:** Primary profession (may be other than actor)
- **averageRating:** Movie average rating
- **numVotes:** Number of votes for the movie

1. Business Understanding:

Question 1: Can we identify any trends in the popularity of genres over time by considering the number of votes?

Question 2: What is the distribution of movie ratings for movies featuring top-rated actors, and how does it compare to movies with lesser-known actors?

Question 3: Can we predict the success of a movie (measured by the number of votes) based on the genres it belongs to?

Question 4: How has the average rating of movies changed over time, and can we identify any specific years with a significant increase or decrease in average ratings?

Question 5: What is the relationship between an actor's birth year and the genres they tend to participate in, and how has this evolved over the years?

2. Data Understanding:

Question 6: What is the distribution of movie runtimes for different genres, and can we identify genres with consistently longer or shorter movies?

Question 7: How does the distribution of average ratings vary for actors with different primary professions?

Question 8: Are there correlations between an actor's birth year and the genres of movies they participate in?

Question 9: Can we identify any outliers in the distribution of the number of votes for movies?
Question 10: How do movies with a higher number of votes typically perform in terms of average rating, and is there a correlation between these two metrics?

Question 11: What are the most common pairs of genres that co-occur in movies, and how does their prevalence vary over the years?

Question 12: Can we identify any patterns in the distribution of runtime for movies with different content ratings (adult vs. non-adult)?

3. Data Preparation:

Question 13: How can we handle categorical variables like genres in a machine learning model?

Question 14: How can we handle imbalanced classes when building a model to predict whether a movie is an adult film?

Question 15: Can we create a feature that represents the average rating of movies an actor or actress has participated in?

4. Modeling:

Question 16: Can we apply advanced feature engineering techniques to improve the performance of the regression model?

Question 17: How does the performance of the regression model change when using a different algorithm, such as Gradient Boosting?

Question 18: How does the performance of a Support Vector Machine (SVM) classifier compare to the Random Forest classifier in predicting whether a movie is an adult film?

Question 19: Can we build an ensemble model that combines predictions from multiple classifiers for better classification performance?

5. Evaluation:

Question 20: How well does the classification model perform in predicting whether a movie is an adult film, and what are the key factors contributing to this prediction?

Question 21: What is the impact of different hyperparameter values on the performance of the regression model?

Question 22: What additional metrics, such as precision, recall, and F1-score, can provide a more detailed evaluation of the classification model?

Question 23: How does the performance of the regression model change when considering only movies released after a certain year?

Question 24: Can we implement cross-validation to get a more robust estimate of the model's performance?

Question 25: Can we implement cross-validation to get a more robust estimate of the model's performance?

6. Deployment:

Question 26: How can we deploy the classification model to provide real-time predictions or flag potentially sensitive content?

Question 27: What challenges might be encountered when deploying a machine learning model in a production environment, and how can they be mitigated?

Question 28: Can we implement model explainability techniques to enhance the interpretability of the classification model's decisions?