Describe event

The project name is Film Analytics. The primary objective of this TMDB movie data analysis project is to predict movie ratings by analyzing various factors. The film industry holds significant financial importance, with global movie sales and streaming platform revenues exceeding $100 billion. Numerous factors influence a movie's success, and understanding these factors is crucial for production companies and individual agents when deciding which projects to participate in. The project's data source is The Movie Database (TMDB), which includes various variables for 10,760 movies released between 1961 and 2015. The dataset includes information on movie budgets, revenues, box office, genres, release years, and more. The core of this research is to analyze the trend of movie ratings over time and identify which factors affect these ratings.

The key steps of the project include data cleaning, feature engineering, model building, and evaluation. Initially, the dataset needs to be explored to handle missing values and outliers. Then, categorical variables will be converted to dummy variables, and numerical variables will be standardized. During the modeling process, I will attempt linear regression and random forest models, and evaluate the models' performance through cross-validation.

Evaluation

Through this project, I learned how to clearly structure a project directory and efficiently store data within various directories. I also gained experience in transferring local data to remote repositories like GitHub. The Random Forest model performed excellently in predicting movie ratings, explaining about 81.2% of the variation in ratings. Furthermore, cross-validation ensured the model's robustness and generalization capability. However, the initial linear regression model did not perform well, indicating the need to consider more nonlinear relationships when selecting models.

Bring out emotion

In the early stages of the project, I felt anxious about handling large amounts of data and various variables. However, as the project progressed, I gradually adapted to this complexity and enjoyed the problem-solving process. I felt particularly excited and satisfied when I discovered the excellent performance of the random forest model. This sense of accomplishment has motivated me to continue exploring more possibilities in the field of data science.

Review in light of previous experience

In conducting the movie data analysis project, I reflected on my previous experiences with data analysis projects and how they applied to this project. In past projects, I often encountered issues with data quality, such as numerous missing values and outliers. Confirming the types of missing values sometimes required repeatedly checking the original data tables, which was very time-consuming. While handling the TMDB data, I found similar issues. However, by using filtering commands in R, I could quickly identify the types of missing values, such as spaces or the # symbol.

In previous projects, I typically used a single model for predictions. In this project, I tried multiple models (linear regression and random forest) and evaluated their performance through cross-validation. This helped me better understand the importance of model selection and parameter tuning, as well as how to use cross-validation to assess a model's generalization capability.

Identify lessons learned

Through this TMDB data analysis project, I learned several important lessons:

1. **Model Selection and Evaluation**: When choosing a model, one should not rely solely on a single model but should try multiple models and evaluate their performance using cross-validation. Different models may perform differently on various datasets and tasks, and comparing them can help find the most suitable model.

2. **Parameter Tuning**: Model parameters have a significant impact on model performance. In this project, tuning the parameters of the random forest model (such as mtry) significantly improved its predictive ability. This made me realize the importance of parameter tuning in data analysis projects.

3. **Documentation and Version Control**: In this project, I used Git for version control, regularly updating and committing code. This not only facilitated collaboration but also helped quickly backtrack and resolve issues when they arose.

Establish follow up actions

In future projects, I plan to continue deepening my knowledge of big data processing methods and complex model construction techniques. Additionally, I will make greater use of methods such as cross-validation and parameter tuning to ensure the robustness and accuracy of the models.

Feedback on those actions

After the project concludes, I will improve my project implementation methods based on the lessons learned and summarized in my reflection logs. By adopting more systematic processes and technical approaches, I aim to enhance project efficiency and the quality of the results.