Project Name

Movie Box Office Prediction – Machine Learning & Data Analysis

GitHub: https://github.com/Ashish-3007/Movie-Box-office-Prediction

Highlights

- Built a **machine learning pipeline** to predict worldwide box office revenue of movies using historical data (2000–2024).
- Engineered features from **domestic/foreign revenue split, audience votes, genres, and release season** to improve model performance.
- Compared multiple regression algorithms (**Linear Regression, Random Forest, XGBoost**) using metrics like **R^2, MAE, RMSE**.
- Achieved best performance with **XGBoost**, delivering robust revenue predictions on unseen movies.
- Designed a **sample prediction system** where new movie attributes (e.g., genre, release season) can be input to estimate revenue potential.

Technology Used

- **Data Source:** CSV dataset
- **Preprocessing & EDA:** Pandas, NumPy, Matplotlib, Seaborn
- **Feature Engineering:** One-hot encoding for genres & release seasons, normalization of revenue distribution
- **Machine Learning Models:** Linear Regression, Random Forest, XGBoost
- **Evaluation Metrics:** R^2 Score, MAE, RMSE

Conclusion

The project successfully built a machine learning model using Random Forest Regressor to predict box office revenue with high accuracy. It highlighted the significance of features like budget, genres, and release year in revenue estimation, showcasing data-driven insights for the film industry's financial forecasting.

Future Scope

The project can be expanded by integrating real-time social media sentiment, audience reviews, and streaming data to enhance prediction accuracy. Incorporating deep learning, ensemble models, and global market trends will allow broader generalization, making the system valuable for production houses, distributors, and investors in strategic decision-making.

Overall Summary

This project integrates **data preprocessing, feature engineering, and supervised machine learning** to solve a real-world problem in the film industry. It showcases strong analytical and technical skills in building predictive models and provides a business-facing application by offering a **data-driven approach to box office forecasting**.

Articulation of Learning

Through this project, I gained expertise in data preprocessing, feature engineering, and applying Random Forest for regression. I learned the importance of domain-specific data cleaning, model evaluation using metrics, and visualization for insight extraction. This experience strengthened my understanding of applying machine learning to real-world business and creative industries.