**Project Report: Movie Rental Duration Prediction 🎬**

This report details a machine learning project aimed at **predicting the duration of movie rentals** based on historical data. The project evaluates and compares the performance of different regression models to identify the most effective approach for this prediction task.

**1. Project Goal 🎯**

The primary objective of this project was to **compare the predictive power of various regression techniques** (specifically Lasso + OLS and Random Forest) in estimating movie rental duration and to **identify the optimal model** for this application.

**2. Dataset 📊**

The analysis utilized the rental\_info.csv dataset. This dataset contains crucial information related to movie rental transactions.

**Key columns used from the dataset include:**

* rental\_date: The date when a movie was rented out.
* return\_date: The date when the movie was returned.
* special\_features: A text field detailing any special features included with the movie (e.g., "Deleted Scenes", "Behind the Scenes").

**3. Methodology 🛠️**

The project involved several key stages, from data preparation to model training and evaluation.

**3.1 Data Preparation and Feature Engineering**

* **Rental Duration Calculation**: A new target variable, rental\_length\_days, was engineered by calculating the difference in days between return\_date and rental\_date.
* **Dummy Variable Creation**: Binary (dummy) features, deleted\_scenes and behind\_the\_scenes, were created based on the presence of these keywords within the special\_features column.
* **Data Splitting**: The dataset was divided into training and testing sets (80% for training, 20% for testing) to ensure robust model evaluation.

**3.2 Lasso and OLS Regression**

* **Feature Selection with Lasso**: A **Lasso Regression** model (alpha=0.3) was initially employed. Lasso's regularization property helps in **identifying and selecting the most relevant features** by driving the coefficients of less important predictors to zero. Only features with positive coefficients after Lasso regularization were retained.
* **OLS Model Training**: An **Ordinary Least Squares (OLS) Regression** model was then trained using this reduced set of features.
* **Evaluation**: The performance of this combined approach was assessed using the **Mean Squared Error (MSE)**.

**3.3 Random Forest Regression and Hyperparameter Tuning**

* **Model Selection**: A **Random Forest Regressor**, an ensemble learning method, was chosen for its ability to handle non-linear relationships and interactions between features.
* **Hyperparameter Optimization**: To achieve optimal performance, **Randomized Search Cross-Validation (RandomizedSearchCV)** was used. This technique efficiently explores a defined hyperparameter space (n\_estimators and max\_depth) to find the best combination of parameters for the Random Forest model, cross-validating across 5 folds.
* **Evaluation**: The tuned Random Forest model's performance was also evaluated using **Mean Squared Error (MSE)** on the test set.

**4. Results 📈**

The **Mean Squared Error (MSE)** was the primary metric used to compare the models. A lower MSE indicates a more accurate model.

* The **Random Forest Regressor** typically demonstrated superior performance, yielding a **lower MSE** compared to the Lasso + OLS combined approach. This suggests that the ensemble nature of the Random Forest, combined with effective hyperparameter tuning, allowed it to capture the underlying patterns in the data more effectively.

**5. Conclusion ✅**

Based on the evaluation, the **Random Forest Regressor** proved to be the **most effective model** for predicting movie rental duration in this project. Its ability to handle complex relationships and its robust performance make it a suitable choice for this type of predictive task.

**6. Future Work 💡**

Several avenues can be explored to further enhance this project:

* **Expanded Feature Engineering**: Investigate the inclusion of additional domain-specific features, such as movie genre, audience ratings (e.g., IMDb scores), director and actor information, or release year, which could provide more predictive power.
* **Alternative Regression Models**: Experiment with other advanced regression algorithms like Gradient Boosting Machines (e.g., XGBoost, LightGBM) or Support Vector Regressors (SVR) to see if further performance gains can be achieved.
* **Enhanced Hyperparameter Tuning**: Implement more exhaustive hyperparameter tuning strategies, such as **GridSearchCV**, if computational resources permit, to explore the hyperparameter space more thoroughly.
* **Robust Cross-Validation**: Integrate more comprehensive cross-validation techniques across all models to ensure the generalizability and stability of the results.
* **Error Analysis and Visualization**: Conduct a deeper analysis of model residuals to understand where predictions deviate and visualize actual vs. predicted values to gain more granular insights into model strengths and weaknesses.