Project Proposal: Movie Sales Forecasting

Objective:
The aim of this project is to forecast movie sales using historical data and provide recommendations for maximizing gross revenue. We will analyze the key factors that drive success, using feature importance techniques to support data-driven decision-making.

Project Workflow:

1. **Data Preprocessing (ETL)**
a. Remove Null Values: Identify and handle missing data (drop or impute).
b. Drop Irrelevant Columns: Remove unnecessary columns like movie_imdb_link and aspect_ratio.
c. Feature Engineering: Create new features like profit (gross revenue - budget).
d. Filter Data: Focus on U.S.-based movies and clean anomalies (e.g., incorrect dates, misspellings).
e. Convert Data Types: Ensure all columns are in the correct format.

---

2. Exploratory Data Analysis (EDA) and Statistical Tests

a. **Calculate Median and Quartiles**: Compute the **median** and **quartiles** for key metrics such as gross revenue, budget, profit, and movie duration. This will provide a more robust measure of central tendency and data spread, especially given the wide variation in monetary values.
b. **Visualize Distributions**: Use box plots to analyze the distribution of key features like gross revenue, budget, and duration. Box plots will help identify outliers, skewness, and the overall distribution of the data, including the interquartile range (IQR).
c. **Correlation Analysis**: Explore relationships between key features like budget, genre, and gross revenue using correlation metrics. This will help identify potential dependencies and patterns within the data.
d. **Distribution and Statistical Analysis**:
- **Distribution Analysis**: Examine the distribution of key features for skewness or kurtosis, which may influence the modeling process.
- **Statistical Testing**: Apply statistical tests such as **T-tests** or **ANOVA** to evaluate differences between groups (e.g., movie genres or budget categories). Check if the data meets the assumptions required for these tests (e.g., normality, homogeneity of variances) to derive meaningful insights, particularly when analyzing multiple genres combined.

---

3. Machine Learning Models

a. **Linear Regression Model**: Build a baseline model to predict gross revenue and evaluate its performance using metrics like R² and MSE.
b. **Neural Network Development**: Develop a Sequential model with **Keras**, experimenting with different layers and hyperparameters. **Optimize Neural Network (Hyperparameter Tuning)**:
- Adjust hyperparameters:
- Number of hidden layers and neurons.
- Activation functions (e.g., ReLU, sigmoid).
- Learning rates and batch sizes.
- Number of epochs.

c. **Avoid Overfitting**
- Implement regularization techniques if necessary.

- Use dropout layers to prevent overfitting, if warranted.

d. Consider implementing **regularization techniques like dropout layers** to avoid overfitting in the neural network.

e. **Compare both models against naive baseline** to show improvement as you use increasingly more complex models.

---

4. **Feature Importance Analysis**

a. Use **Permutation Importance**, **SHAP Values** or **Partial Dependence Plots** to determine the most significant features impacting movie revenue (e.g., actor popularity, genre).
b. Apply the selected method to the Neural Network model.
c. Interpret the results to understand feature impacts.

---

5. **Visualization and Results**

a. **Key Visualizations** (may be adjusted based on the identification of the most important features):
   - **Box office revenue by genre**, with a median gross revenue line in visualizations for comparison.
   - **Top 10 actors by office revenue**, with a median gross revenue line in visualizations for comparison.
   - **Trends in box office sales over time**, with a median gross revenue line in visualizations for comparison.

b. **Feature Importance Visualization**: Highlight the top factors affecting movie revenue.
c. **Accuracy rates of liner regression and deep learning models**.
d. **Distributions for gross revenue, profit, and budget**

---

6. Discuss results and conclusion:

a. **Compile Results**
   - Summarize findings from EDA and modeling.
   - Highlight key insights and trends.

b. **Draft Conclusion**
   - Discuss how the results address the project's purpose.
   - Provide recommendations to movie producers based on findings, if applicable.

---

7. Presentation and Final Deliverables:

a. Create a presentation.
b. **Assign Presentation Roles**
c. **Decide which slides to present and for how long to fit in the time limits.**
   - Ensure each team member presents at least one slide.
   - Assign sections based on individual contributions.

d. **Engagement Strategy:**
   - Show a 10 second movie scene and ask audience engaging questions (selecting ideas).

e. **Rehearsal**

- Practice the presentation to ensure smooth delivery and timing.
- Refine slides as needed.

---

8. **Code and Notebook Review**
   - Ensure all code is well-documented and notebooks are clean and readable.
   - I need to be able to run your entire notebook.
   - Submit **only one notebook for evaluation**. Our development/exploratory notebooks will be saved inside “Notebooks” folder.

---

9. **Finalize Reports**

a. Write a comprehensive report summarizing the methodology, results, and conclusions in README and presentation.

---

10. **Repository Organization**

a. We plan to verify if all files are properly organized in the repository.
b. Update the README with instructions on how to navigate the project, and/or with any other relevant details that may have changed since the last time update.

---

11. **Final Deliverables**:

- Comprehensive report summarizing findings and methodologies.
- Well-documented code for reproducibility.
- Clear, impactful visualizations to present key insights to stakeholders.
- Presentation with audience engaging tactics

Data Source:
Kaggle: this dataset includes historical movie data, including budget, gross revenue, actor popularity, and more. We will address potential limitations, such as monetary values not being adjusted for inflation and how the gross revenue was recorded.