This report details a causal inference analysis that aims to investigate the hypothesized impact of movie budgets on IMDb scores within the film industry. The underlying premise is that **larger budgets** allow for higher production values, better cast and crew, more extensive marketing campaigns, and ultimately, a potentially **higher quality and more appealing movie** to audiences, which should be reflected in IMDb scores.

Data Preprocessing:

The dataset was processed to categorize the IMDb scores into binary groups of 'high score' and 'low score', with the threshold set at 7.4. This binary categorization simplifies the outcome variable, allowing the application of classification techniques to infer causal relationships. The movie budgets were similarly categorized into 'high budget' and 'low budget' groups based on a threshold set at the 75th percentile, introducing a clear binary treatment variable into the analysis.

Feature Engineering:

The release_year variable was transformed into categorical 'eras', reflecting significant periods in cinematic history. This transformation captures the potential shifts in audience preferences, technological advancements, and industry standards, providing a nuanced control over temporal variations that might influence the IMDb score independently of the budget.

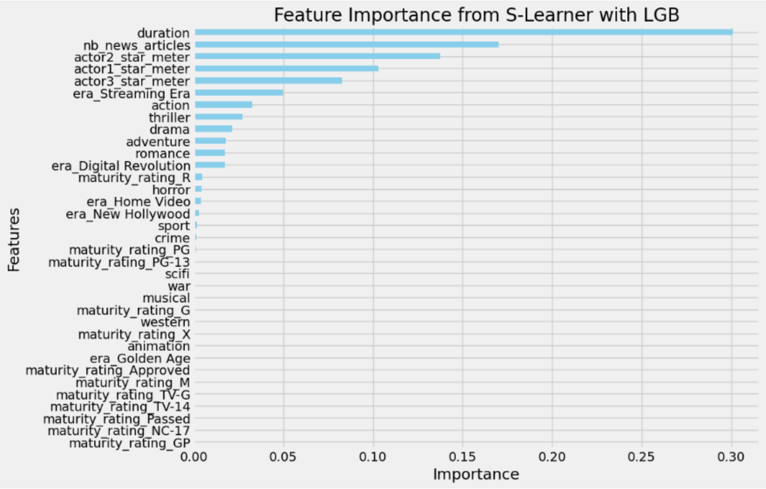
Covariate Selection:

A set of covariates was chosen based on their relevance to both the treatment (budget) and the outcome (IMDb score). These include movie characteristics such as duration and genre indicators, as well as measures of publicity and star power. These covariates were then standardized to ensure comparability and to aid in the interpretation of model coefficients.

Why S-learner:

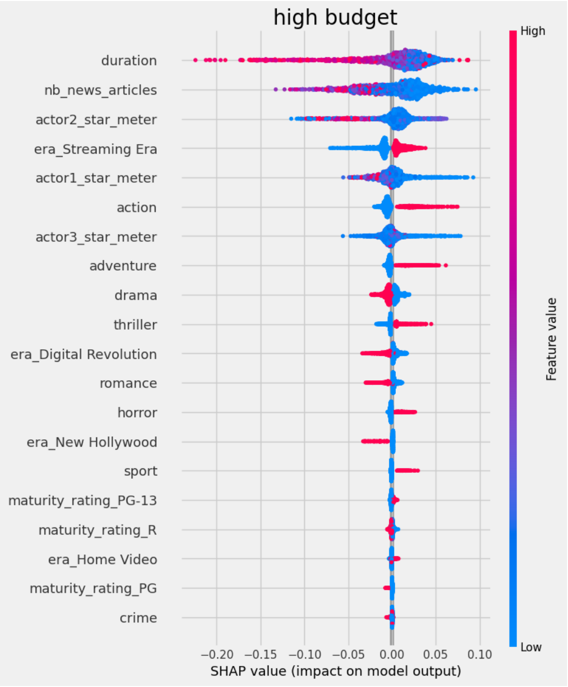
The S-learner was chosen despite the imbalances present in outcomes, covariates, and treatments due to its robustness and its ability to handle complex, non-linear relationships within the data effectively. Moreover, S-learner might be less sensitive to treatment group imbalance compared to the other meta learner since it models the treatment effect within the context of the entire dataset.

LGB Result:

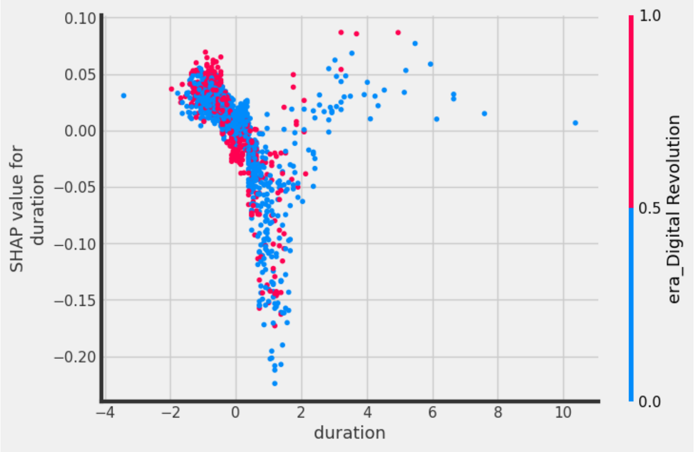


The feature importance indicates that 'duration' and 'number of news articles' are the most significant predictors, suggesting that longer movies and those with more media coverage tend to be associated with higher IMDb scores. The star meters for actors 2 and 1 follow, hinting that the presence of well-known actors contributes positively to a movie's IMDb score.

The influence of the 'Streaming Era' is notable, reflecting the changing landscape of movie consumption and possibly the increasing importance of digital distribution in reaching audiences. Genres such as action, thriller, and drama also appear to be important factors, indicating that these genres are perhaps more likely to attract higher budgets and, in turn, influence IMDb scores.



Higher values of 'duration' (as indicated by the red dots) tend to push the model towards predicting a 'high budget', which in the context of this model, suggests a positive relationship with higher IMDb scores.

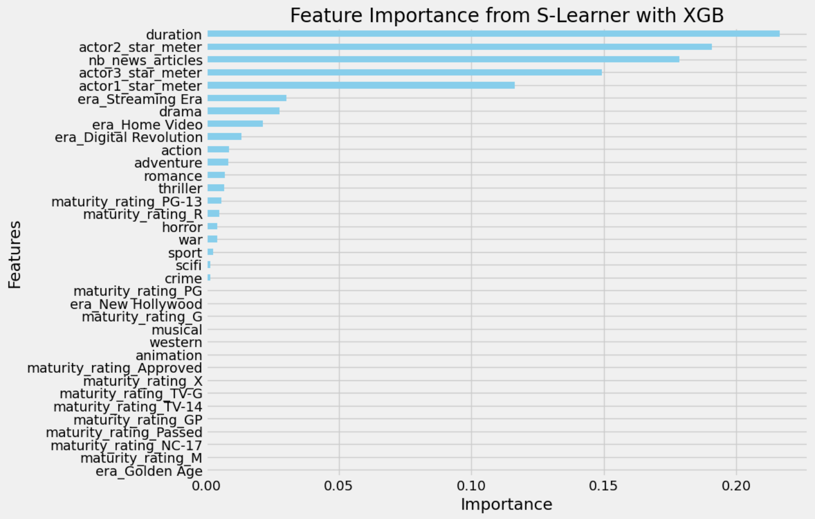


Movies of longer duration in the Digital Revolution era are likely to be associated with higher IMDb scores. The blue dots indicate a negative relationship for shorter movies, particularly outside of the Digital Revolution era.

ATE = -0.07484183173384468

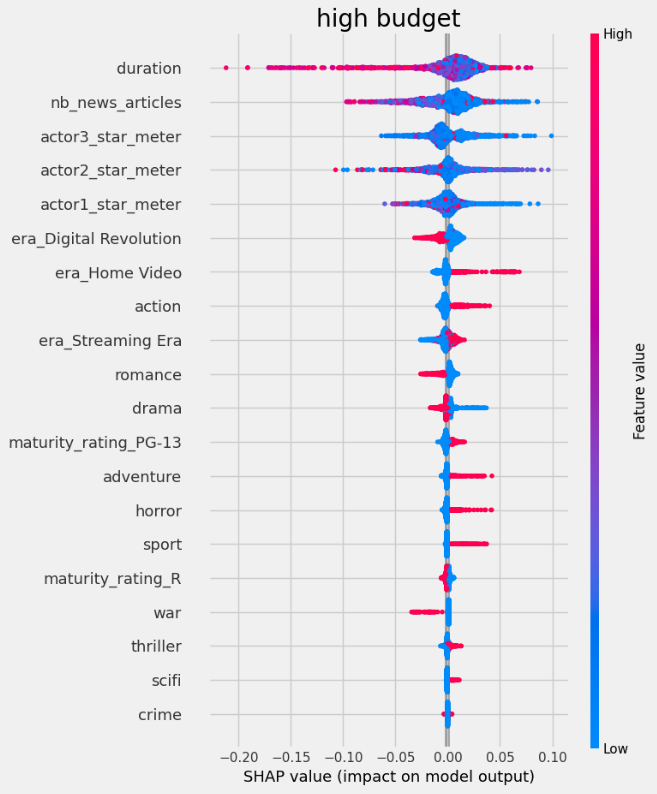
The ATE estimated by the S-learner with LightGBM is -0.07484183173384468, which suggests that being in the 'high budget' category is associated with a 0.0748 decrease in the likelihood of having a high IMDb score, on average, when all other factors are held constant. This result is counterintuitive, as the initial hypothesis was that higher budgets would correspond to higher IMDb scores.

XGB Result:

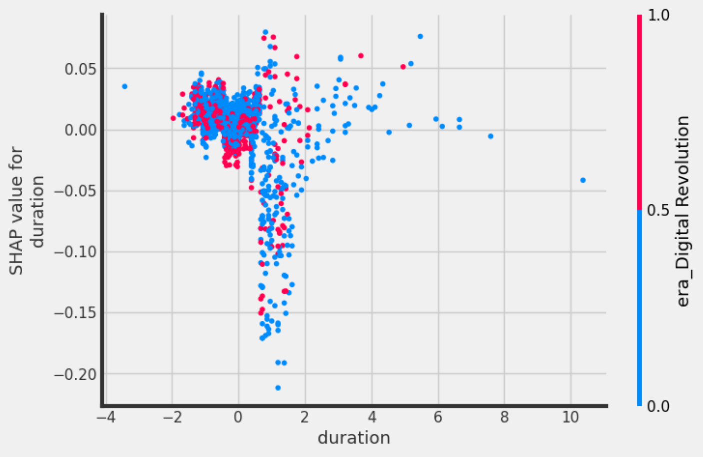


Similar to the LightGBM model, the XGBoost model indicates that 'duration' and 'actor2_star_meter' are important features. Notably, 'nb_news_articles' seems to have less relative importance in this model compared to LightGBM, suggesting that the way XGBoost handles the data and feature interactions might differ.

The 'actor3_star_meter' has increased in importance in this model, indicating that the third-billed actor's star power may play a more significant role in predicting IMDb scores when the XGBoost algorithm is used. The 'era_Streaming Era' and 'era_Digital Revolution' still appear as important features, reaffirming the influence of these time periods on a movie's reception.



This shows that higher values of 'duration' have a more uniform effect across the predictions, while the 'actor' star meters show a diverse impact, with some increasing and others decreasing the likelihood of a high IMDb score.



The SHAP dependence plot for 'duration' colored by 'era_Digital Revolution' again confirms the positive impact of longer movie durations on the predicted IMDb scores, particularly in the Digital Revolution era. The plot suggests a strong interaction effect between movie duration and the era in which a movie was released, influencing its budget and subsequent IMDb score.

ATE = -0.050369384819301664

This value is closer to zero than the ATE from the LightGBM model but still indicates a negative association between having a high budget and the likelihood of a high IMDb score. This suggests that even after controlling for various factors, higher budgets do not necessarily translate to higher IMDb scores according to the XGBoost model.

Interpretation of Negative ATEs:

When both the LightGBM and XGBoost S-learner models indicate a negative Average Treatment Effect for the relationship between high budget movies and IMDb scores, several interpretations and implications need to be considered:

1. **Expectations vs. Delivery**: High budget movies often come with high expectations. When such movies do not meet audience expectations, which can be elevated by the budget, it may result in lower IMDb scores. The negative ATE could reflect the gap between what is expected and what is delivered.
2. **Diminishing Returns**: There may be a point at which increasing the budget does not significantly enhance the perceived quality or enjoyment of a movie. This phenomenon of diminishing returns suggests that beyond a certain budget threshold, additional investment has a smaller impact on audience ratings.
3. **Audience Bias**: Movie-goers might have a bias against high budget movies, perceiving them as less artistic or original compared to independent, lower-budget films. This could contribute to a systematic undervaluation of high budget movies in IMDb scores.
4. **Overfitting to Blockbusters**: High budget films are often blockbuster types, with specific characteristics that may not align with factors that contribute positively to IMDb scores. For example, they might be more action-oriented and less focused on narrative depth, which could affect critical reception.

Further Steps:

- **Model Validation**: The consistency of the negative ATE across two models increases confidence in the robustness of the findings. However, it's crucial to validate these results further, possibly through different methodologies or data subsets, to rule out model-specific biases.
- **Covariate Influence**: Both models identify similar covariates as important, such as duration and actor star meters. These factors seem to have a more direct relationship with IMDb scores than the budget itself. This could mean that while the budget is important, other movie characteristics play a more decisive role in determining IMDb scores.

- **External Validity**: The findings should be compared with industry knowledge and other research to ensure external validity. For instance, if similar studies have found a positive relationship between budget and IMDb score, it would be important to understand why this analysis differs.