**Exhibitor Share (E)**

We want to build a prediction model that allows Exhibitors to predict the likely revenue from a movie. To that end, we now look at the likely causal relationships of the different features to the exhibitor share of total nett gross earned by a film. To avoid clutter we draw simpler DAGs that model the likely causal relationship between E and a feature and will include other features that were seen to have causal relationship, in the analysis earlier, with the feature under investigation.

**Release Week**

B is a confounder for the effect of W on E. To ascertain the causal effect of W on E we set up a regression model, regressing E over B and W.

The p-value of the regression coefficient for W is 0.917 and we conclude that there is likely **no direct causal path from W to E**.

**Genre**

Y is a confounder and FF is the mediator for the effect of G on E. To condition on Y and FF, we set up a regression model, regressing E over FF, G and Y. Since G is a categorical variable we use dummy variables to one-hot encode G. All variables in the regression model are standardized. We ascertain the regression coefficient and the p-value for G to determine the likely causal path from G to E.

For all but two genre categories, the p-value of the regression coefficient is significantly greater than 0.05 and for the two genre categories with small p-values the absolute value of the corresponding regression coefficient is 0.06 or lower. We can conclude that there is likely **no direct causal path from G to E**.

**Total Nett Gross Revenue**

Y, S, F and FF are likely confounders to the effect of T on E. To condition on the confounders we set up a regression model, regressing E over Y, S, F, FF and T. All variables in the regression model are standardized. We ascertain the regression coefficient and the p-value for T to determine the likely causal path from T to E.

The p-value of the regression coefficient of T is less than 0.05 and the regression coefficient itself is 1.3861 and we conclude that there is a **direct causal path from T to E.**

**Footfalls**

Y, G, R, S, F and L are likely confounders to the effect of FF on E. To condiction on the confounders we set up regresion model for each year, regressing E on FF, Y, G, R, S, F and L and then evaluate the correlation between E and FF by examiming the p-values of the regression coefficients of FF.

The p-values for the regression coeffiecient of FF is below 0.05 and the regression coefficient itself is 0.1542 indicating that FF has a significant effect on E. We conclude that there is likely a **direct causal path from FF to E**.

**Run Length**

Y, B, S, R and F are likely confounders and FF is a mediator to the effect of L on E. To condition on the confounders and the mediator we set up regression model for each year, regressing FF on Y, B, S, R, F, FF and L and then evaluate the correlation between E and L by examining the p-values of the regression coefficients of L.

The p-values for the regression coeffiecient of FF is below 0.05 but the regression coefficient itself is 0.1412 indicating that L has a causal effect on E. We conclude that there is a **direct causal path from L to E**.

**First Week Revenue**

Y, I, B and S are likely confounders and FF a mediator to the effect of F on E. To condition on all the confounders and the mediator we set up regression model, regressing E on Y, I, B, S, FF and F. We ascertain the coefficient value and its p-value to determine the existense of a causal path from F to E.

The p-value for the coefficient of F is less than 0.05 and the regression coefficient is 0.9259 implying that F has a significant association with E and there is a **direct causal path from F to E**.

**Release Screens**

Y and B are likely confounders and T, F and FF are mediators for the effect of S on E. Admittedly, S here is the number of screens the film opens to in its first week. We however make the assumption that the number of screens that a film gets through its run is likely a function of the number of screens that it opens to and hence we use S as a proxy variable for the number of screens that the film plays to during its run.

To condition on all the confounders and mediators, we set up a linear regression model, regressing E on Y, B, T, F, FF and S. We ascertain the values of the coeffecient value and its p-value to determine the likely causal path from S to E.

The p-value for the coefficient of S is less than 0.05 and the regression coefficient is 0.0924 implying that S has at best a weak causal effect on E and we conclude and there is likely **no** **direct causal path from S to E**.

**Runtime**

B and Y are confounders and FF and L are mediators to the effect of R on E. To condition on all the confounders and mediators, we set up a linear regression model, regressing E on Y, B, FF, L and R. We ascertain the values of the coeffecient value and its p-value to determine the likely causal path from R to E.

The p-value for the coefficient of R is less than 0.05 but the regression coefficient is -0.0775 implying that R does not have significant causal association with E and we conclude and there is **no** **direct causal path from R to E**.

**Budget**

Y and I are likely confounders and L and F are mediators to the effect of B on E. To condition on all the confounders and mediators, we set up a linear regression model, regressing E on Y, I, B, L and F. We ascertain the values of the coeffecient value and its p-value to determine the likely causal path from B to E.

The p-value for the coefficient of B is less than 0.05 but the regression coefficient is -0.0445 implying that B does not have significant causal association with E and we conclude and there is **no** **direct causal path from B to E**.

**Inflation**

Y is a likely confounders and F a mediator to the effect of I on E. To condition on all the confounders and mediators, we set up a linear regression model, regressing E on Y, I and F. We ascertain the values of the coeffecient value and its p-value to determine the likely causal path from I to E.

The p-value for the coefficient of I is greater than 0.05 and the regression coefficient is -0.0078 implying that I does not have significant causal association with E and we conclude and there is **no** **direct causal path from I to E**.

**Release Year**

T, L, F and FF are mediators to the effect of Y on E. To condition on all mediators, we set up a regression model, regressing E over Y, T, L, F and FF. We ascertain the values of the coeffecient value and its p-value to determine the likely causal path from Y to E.

The p-value for the coefficient of Y is less than 0.05 but the regression coefficient is 0.0239 and we conclude and there is likely **no** **direct causal path from I to E**.

**Summary**

In summary, we conclude that the **Total Nett Gross Revenue of a film is influenced by:**

* **first week revenue,**
* **run length,**
* **footfalls, and**
* **total nett revenue**

We update the summary table of causation:

|  |  |
| --- | --- |
| **Feature** | **Influenced/Affected By** |
| Release Year |  |
| Inflation | Release Year |
| Genre | Release Year |
| Budget | Release Year, Inflation |
| Runtime | Release Year, Budget |
| Release Week | Budget |
| Release Screens | Release Year, Budget |
| First Week Revenue | Release Year, Inflation, Budget, Release Screens |
| Run Length | Release Year, Budget, Release Screens, Runtime, First Week Revenue |
| Footfalls | Release Year, Genre, Release Screens, Runtime, First Week Revenue, Run Length |
| Total Nett Revenue | Release Year, Release Screens, First Week Revenue and Footfalls |
| Exhibitor Share | First Week Revenue, Run Length, Footfalls and Total Nett Revenue |

**Prediction Model Summary**

We fitted a Gradient Boosted Ensemble Regression Model to a training set and evaluated its performance against a test set, both sets drawn from the data available. The performance of the model:

* Percentage of estimates for test set that are off by less than 25% from true value: 97.54
* Percentage of estimates for test set that are off by less than 35% from true value: 98.15
* Percentage of estimates for test set that are off by less than 45% from true value: 99.38
* Percentage of estimates for test set that are off by less than 55% from true value: 99.38

**Test Set Performance**

* Percentage of estimates for test set that are off by less than 25% from true value: 68.31
* Percentage of estimates for test set that are off by less than 35% from true value: 82.15
* Percentage of estimates for test set that are off by less than 45% from true value: 90.77
* Percentage of estimates for test set that are off by less than 55% from true value: 94.46