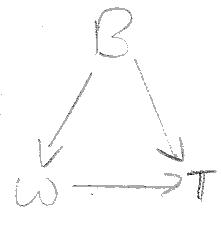
**Total Nett Gross Revenue (T)**

We want to build a prediction model that allows industry participants to predict the likely total nett gross revenue from a movie. To that end, we now look at the likely causal relationships of the different features to the total nett gross revenue earned by a film. To avoid clutter we draw simpler DAGs that model the likely causal relationship between T 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 T. To ascertain the causal effect of W on T we set up a regression model, regressing T over B and W.

The p-value of the regression coefficient for W is 0.857 and we conclude that there is likely **no direct causal path from W to T**.

**Genre**

Y is a confounder and FF is the mediator for the effect of G on T. To condition on Y and FF, we set up a regression model, regressing T 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 L.

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

**Footfalls**

Y, G, R, S, F and L are likely confounders to the effect of FF on T. To condiction on the confounders we set up regresion model for each year, regressing T on FF, Y, G, R, S, F and L and then evaluate the correlation between T 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 0.2054 indicating that FF has a significant effect on T. We conclude that there is likely a **direct causal path from FF to T**.

**Run Length**

Y, B, S, R and F are likely confounders and FF is a mediator to the effect of L on T. 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 T 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 0.0823 indicating that L has a weak causal effect on T. We conclude that there is likely **no** **direct causal path from L to T**.

**First Week Revenue**

Y, I, B and S are likely confounders and FF a mediator to the effect of F on T. To condition on all the confounders and the mediator we set up regression model, regressing T 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 T.

The p-value for the coefficient of F is less than 0.05 and the regression coefficient is 0.9576 implying that F has a significant association with T and there is a **direct causal path from F to T**.

**Release Screens**

Y and B are likely confounders and F and FF are mediators for the effect of S on T. 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 T on Y, B, 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 T.

The p-value for the coefficient of S is less than 0.05 and the regression coefficient is -0.1942 implying that S has significant association with T and we conclude and there is a **direct causal path from S to T**.

**Runtime**

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

The p-value for the coefficient of R is less than 0.05 but the regression coefficient is -0.0581 implying that R does not have significant causal association with T and we conclude and there is **no** **direct causal path from R to T**.

**Budget**

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

The p-value for the coefficient of B is less than 0.05 but the regression coefficient is -0.0388 implying that B does not have significant causal association with T and we conclude and there is **no** **direct causal path from B to T**.

**Inflation**

Y is a likely confounders and F a mediator to the effect of I on T. To condition on all the confounders and mediators, we set up a linear regression model, regressing T 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 T.

The p-value for the coefficient of I is less than 0.05 but the regression coefficient is -0.0952 implying that I does not have significant causal association with T and we conclude and there is **no** **direct causal path from I to T**.

**Release Year**

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

The p-value for the coefficient of Y is less than 0.05 but the regression coefficient is 0.1369 implying that Y has a weak causal association with T and we conclude and there is likely a **direct causal path from I to T**.

**Summary**

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

* **release year,**
* **the number of screens the film releases to,**
* **first week revenue, and**
* **footfalls**

The DAG for this model is presented alongside.

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 |

**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: 92.31
* Percentage of estimates for test set that are off by less than 35% from true value: 97.85
* Percentage of estimates for test set that are off by less than 45% from true value: 99.69
* Percentage of estimates for test set that are off by less than 55% from true value: 99.69

**Test Set Performance**

* Percentage of estimates for test set that are off by less than 25% from true value: 69.54
* Percentage of estimates for test set that are off by less than 35% from true value: 83.08
* Percentage of estimates for test set that are off by less than 45% from true value: 89.54
* Percentage of estimates for test set that are off by less than 55% from true value: 94.15