**Footfalls (FF)**

Footfalls are arguably the most important and posssibly the most unbiased metric to measure the performance of a film. We now look at the likely causal relationships of the different features to the footfalls, specifically India footfalls, attracted by a film. To avoid clutter we draw simpler DAGs that model the likely causal relationship between FF 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 FF. To ascertain the causal effect of W on FF we set up a regression model, regressing FF over B and W.

The p-value of the regression coefficient for W is 0.062 and the regression coefficient for W is 0.0486. We therefore conclude that there is likely **no direct causal path from W to FF**.

**Genre**

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

For all genre categories, the p-value of the regression coefficient is significantly lower than 0.05 and the regression coefficients are very large and we conclude that there is a **direct causal path from G to FF**.

**Run Length**

Y, B, S, R and F are likely confounders to the effect of L on FF. To condition on the confounders we set up regression model, regressing FF on Y, B, S, R, F and L. We ascertain the regression coefficient and the p-value for L to determine the likely causal path from L to FF.

The p-value of the regression coefficient for L is less than 0.05 and the regression coefficient for L is 0.3257. We therefore conclude that there is likely a **direct causal path from L to FF**.

**First Week Revenue**

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

The p-value of the regression coefficient for F is less than 0.05 and the regression coefficient for F is 0.5029. We therefore conclude that there is likely a **direct causal path from F to FF**.

**Release Screens**

Y and B are likely confounders and F and L are mediators to the effect of S on FF. 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 regression model, regressing FF on Y, B, S, F and L. We ascertain the regression coefficient and the p-value for S to determine the likely causal path from S to FF.

The p-value of the regression coefficient for S is less than 0.05 and the regression coefficient for S is 0.2321. We therefore conclude that there is likely a **direct causal path from S to FF**.

**Runtime**

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

The p-value of the regression coefficient for R is less than 0.05 and the regression coefficient for R is 0.1810. We therefore conclude that there is likely a **direct causal path from R to FF**.

**Budget**

Y is likely confounders and R, S, F and L are mediators to the effect of B on FF. To condition on all the confounders and mediators we set up regression model, regressing FF on Y, B, R, S, F and L. We ascertain the regression coefficient and the p-value for B to determine the likely causal path from B to FF.

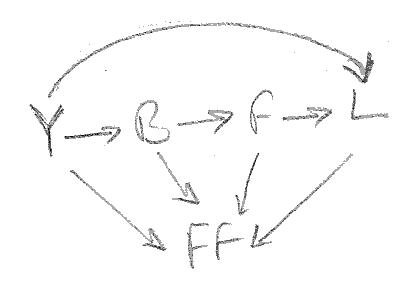
The p-value of the regression coefficient for B is greater than 0.05 and the regression coefficient for B is -0.0474. We therefore conclude that there is likely **no direct causal path from B to FF**.

**Release Year**

G, R, S, F and L are mediators to the effect of Y on FF. To condition on all the mediators we set up regression model, regressing FF on Y, G, R, S, F and L. We ascertain the regression coefficient and the p-value for Y to determine the likely causal path from Y to FF.

The p-value of the regression coefficient for Y is less than 0.05 and the regression coefficient for Y is -0.2014 We therefore conclude that there is likely a **direct causal path from Y to FF**.

**Summary**

In summary, we conclude that the **total footfalls of a film is influenced by:**

* **year of release,**
* **genre,**
* **runtime**
* **number of screens released to,**
* **first week revenue, and**
* **run length**.

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 |

**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: 71.69
* Percentage of estimates for test set that are off by less than 35% from true value: 85.54
* Percentage of estimates for test set that are off by less than 45% from true value: 88.62
* Percentage of estimates for test set that are off by less than 55% from true value: 92.00

**Test Set Performance**

* Percentage of estimates for test set that are off by less than 25% from true value: 61.23
* Percentage of estimates for test set that are off by less than 35% from true value: 77.54
* Percentage of estimates for test set that are off by less than 45% from true value: 84.31
* Percentage of estimates for test set that are off by less than 55% from true value: 89.85