ECON 424 Project Proposal - Group 3

Submitted by: Raphael Shawn Gozali (20805288) & Jason Tedjosoesilo (20801292)

Questions

For this project, these are the questions that we will be answering:

1. Without any interactions, which covariate(s) affect revenue the most?
2. Are there any certain genres, actors, or keywords that affect revenue the most?
3. If we do interactions between genres, actors, and keywords, are there any particular combination(s) that affect revenue the most?
4. Do people's preferences on movie genres and actors change over the years/decades?

Methods

To answer the questions above, we will be doing these methods:

1. Do preprocessing on data by creating necessary columns for regression, i.e. adjusted revenue (removing inflation factor)
2. Use parameter reduction and regression via LASSO to get a certain number of covariates and see which covariates impact the most to the revenue.
3. Get the genres, actors, and keywords from their respective columns (each separated by "-"), filter these words by number of appearance more than a certain number, and then use these words as covariates in regression.
4. Get the words from step 3 and create interactions between them. Include these interactions as covariates in the regression.
5. Separate the data by year/decade published and do the modelling using LASSO on these subgroups and see whether the model changes over time.

Data

In this project, we will be using the Movies dataset from Kaggle: https://www.kaggle.com/datasets/akshaypawar7/millions-of-movies?resource=download.
This dataset consists of around 723 thousand rows of movies, with their descriptions and details separated into columns. The response variable that we are focusing on from this dataset is the "revenue" column, which shows us the revenue of the movie. The covariates that might be useful from this dataset includes: date published, budget, genres, casts, keywords, and runtime.

Results

From this project, we hope to get a good model based on this data where revenue is based only on a small number of covariates, i.e a small number of elasticities. We also hope to get interesting keywords and interactions between genres and casts that can affect revenue of the movie. Changes in people's preferences of the movies over the years/decades are also expected here, since the trends between generations are likely to be different.