**Creating a Regression Model to Predict Future Box Office Earnings of Movie Franchise Films**

Maira Asmat

**Abstract**

The goal of this project was to scrape data from a public website and use that data to build a regression model and use that model in a prediction. I worked with data related to movies that are part of franchises to predict future box office totals, using feature engineering to creating and refining a ridge regression model and creating predictions based on that model.

**Design**

I based the idea of this project around understanding if future movie sequels can still be profitable, and even if they can be more profitable than their predecessors and other (standalone) films. Ideally, this model could be used by movie directors and producers when deciding what to create.

**Data**

Data was webscraped from the Box Office Mojo franchise page, which included the features of rating, release date, distributor, and runtime. This dataframe was merged with data scraped from website The Numbers to obtain budget and domestic gross, as well as IMDb to obtain genres. With all three dataframes merged together, there were 365 rows/data points with 23 columns/features. (In the final model, this was reduced to 4 features.)

**Algorithms**

*Feature Engineering*

1. Converting categorial features such as genre to binary dummy variables
2. Using EDA to determine which features could be useful to the final model

*Models*

A simple linear regression, a ridge regression, and a lasso regression were all tested before a ridge regression was chosen due to its slightly higher r-squared value and its slightly smaller margin of error (when compared to that of the linear regression).

*Model Evaluation and Selection*

The data underwent a rigorous validation method: cross-validation and testing. Thus, 80% of the data was used for a cross-validated training process and 20% was used for final testing evaluation. This method was chosen due to there being a small amount of data points, so the most efficient method was needed.

**Tools**

* BeautifulSoup to webscrape
* Numpy and Pandas for exploratory data analysis and data manipulation
* Scikit-learn for modeling
* Matplotlib and Seaborn for data visualizations/plotting

**Communication**

All visuals are embedded within the slides, which are attached in the GitHub repository and will be presented.