**Data Analytics Capstone Executive Summary**

**Capstone Project Name:** “Siskel and Ebert” style Regression ML Model that predicts popular movie ratings as ‘high’ or ‘low.’

**Project Topic:** This capstone project utilizes two popular cinema movie datasets sourced from Kaggle.com. These datasets encompass a wide range of information about 10,000 movies, including IMDb ratings, vote counts, and box office earnings. Through this dataset, we gain valuable insights into the key metrics contributing to a movie's success and its performance relative to other films in the cinema industry.

**Research Question**: What is the anticipated IMDb (Internet Movie Database) rating a movie can achieve by leveraging a set of explanatory variables extracted from a popular movie dataset?

IMDb.com is the largest internet movie database, featuring popular user ratings, reviews for movies and shows, and more. Ranked as the 52nd most visited website in 2019, this site began as a fan-operated movie database and has been owned and operated by Amazon since 1998. The ‘rating’ variable in this dataset serves as an indicator of a movie's popularity among both fans and critics. Identifying the variables that exhibit strong correlations with higher ratings can enhance predictive modeling for the cinema industry.

The established null and alternative hypotheses for this analysis project is:

**Null hypothesis**: The independent variables have no statistically significant effect on the dependent variable, average IMDb movie rating of high or low from the Internet Movie Database.   
**Alternate Hypothesis**: The independent variables have a statistically significant effect on the dependent variable, average IMDb movie rating of high or low from the Internet Movie Database.

**Data Analysis Process:** The steps involved in this analysis and building the model process include data collection, data wrangling, identification and treatment of duplicates, missing or null values, and outliers in the data. Encoding categorical variables, conducting data exploration, and performing feature selection are the final steps in the data preparation process before splitting and training the machine learning model.

Interesting trends and insights started showing up using the genre’s variables based on ‘high’ and ‘low’ ratings as well as the count of movies with top 25 actors. Due to one of the online datasets out of date information, these two variables weren’t used in the regression model.

A graph of different colored squares

Description automatically generated

A blue and white bar graph

Description automatically generated

Correlation chart and heat map was used as the statistical technique for feature selection by the analyst. These techniques showed correlated variables to the ranking variable to include in the model: ‘metascore’, ‘runtime’, and ‘vote\_average’. These three variables were used to train the model to get the highest accuracy for this project.

A screenshot of a cell phone

Description automatically generated

**Findings:** The final logistic regression prediction model, created in Python, had an accuracy of 86%. The number of observations for this model was 6,358. The dependent variable, which is the binary outcome variable used to predict the ranking of a movie’s success among fans. The independent variables, or feature variables, used was ‘runtime’, ‘metascore’, ‘vote\_average.’ The pseudo r-squared score indicates a good fit model at 0.4493.

The standard error statistic measured the uncertainty associated with the coefficient estimates. The smaller the errors are indicating more precise estimates. Those scores and variables are runtime: 0.002, metascore: 0.002, and vote\_average: 0.079.

P-value score needs to be less than 0.05. This score proves the variable is statistically significant in predicting the outcome. In this case the p-value for all the dependent variables is less than 0.05, indicating there is strong evidence to reject the null hypothesis, indicating the dependent variables do have an impact on the prediction ranking outcome.

**Limitations:** The regression model achieved an accuracy of 86%. However, there were certain limitations that needed to be addressed to achieve this level of performance. The most significant limitation pertained to the availability of an adequate number of data points required to construct a statistically robust machine learning model. To accurately predict whether a new movie will receive a high or low rating, numerous factors come into play. The pandemic undeniably altered how fan’s view, rate, and watch movies. This project enabled the training of a machine learning model to forecast human behavior based on previous human opinions, thus allowing for predictions of future ratings. The larger movie dataset contained data that hadn't been updated beyond 2017. If this data were updated to 2023, it's possible that the model would have incorporated different variables or outcomes.

**Next Steps:** There are several ways to advance this research project. A data analyst could construct a comparative linear regression model to predict a movie's precise numeric rating, using this model as a base. Additionally, an adding a web scraping project to acquire additional movie data, such as budgets of films, to gain deeper insights into the factors influencing a movie's ratings. Further analysis could be performed to formulate a composite target variable, encompassing factors like votes cast, gross profit, budget, etc., instead of relying solely on an already established rating.

**Expected Benefits**: In terms of implementing this rating model system for predicting ratings in the cinema industry, it's important to consider the historical context. The cinema rating duo, Siskel and Ebert, set the standard until 1999. Their ratings and opinions were grounded in their personal experiences with the movies they reviewed. Their success stemmed from their ability to provide contrasting perspectives and critiques based on their individual life experiences and opinions. This machine learning prediction model is designed to replicate the tasks performed by human critics. It aims to demonstrate to the cinema industry which insights and indicators have the most significant impact on ratings.