MGT 6203 Group Project Progress Report

TEAM INFORMATION

Team #: 1

Team Members:

1. Team Member 1 Name; EdX username (MM)
   Ho Wei Sin; EdX username: howeisin
2. Team Member 2 Name; EdX username (MM)
   Kevin Fong Ka Chun, EdX username: kevin_chun
3. Team Member 3 Name; EdX username (MM)
   Quek Zhu Hui Joel, EdX username: joelquek
4. Team Member 4 Name; EdX username (MM)
   Soo Wen Jun; EdX username: SOOWENJUN

BACKGROUND INFORMATION

Project Title: Movie Success Factors

Background Information:

The film industry is a fiercely competitive sector, where achieving profitability and recognition through awards are crucial measures of success. Film production companies can greatly benefit from understanding the key factors that influence a movie's performance, allowing them to make informed decisions and maximize their return on investment.

Primary Research Question:

What are the primary factors that contribute to a **movie's profitability** and **likelihood of winning awards**?

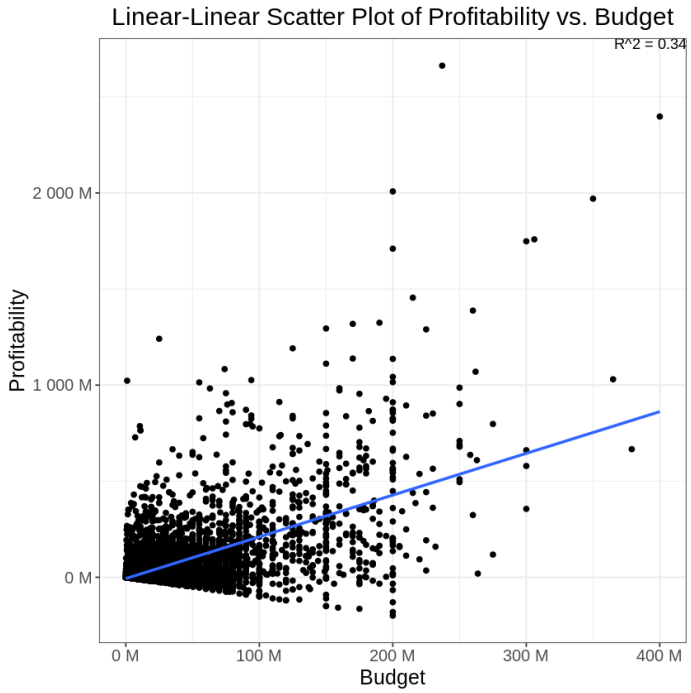
Supporting Research Questions:

1. What is the correlation between a movie's genre and its success?
2. Which factors play a role in determining whether a movie wins an Oscar?
3. Do movies released during holiday seasons, such as summer and winter, tend to perform better?
4. Are there common factors, such as production budget, MPAA rating, and IMDB rating that significantly impact a film's success when compared to other movies?

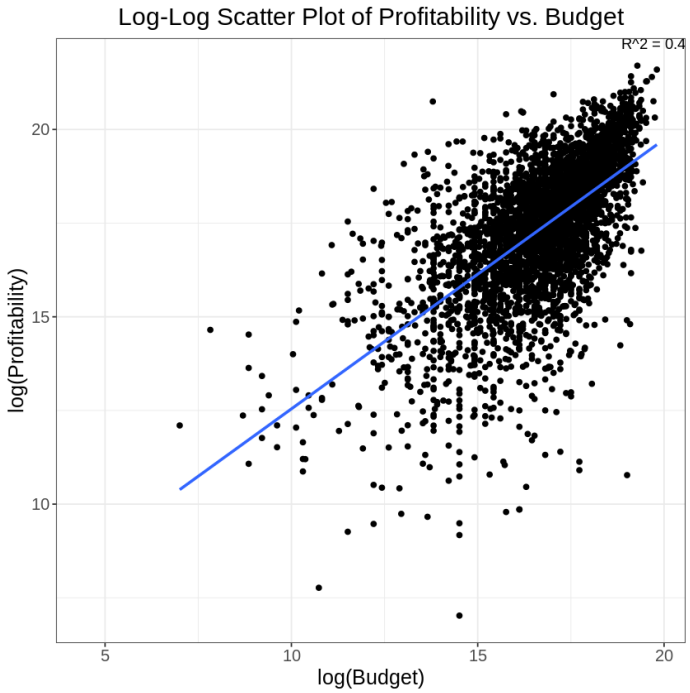
OVERVIEW OF PROBLEM AND PLANNED APPROACH

Our initial approach will involve conducting exploratory data analysis to obtain a better understanding of the distribution and associations among the variables. This analysis will include the utilization of visualizations, summary statistics, and correlation analysis. Through these techniques, we aim to uncover valuable insights about the data.

EDA: Profit (Unadjusted) Vs Budget (Unadjusted) Scatter Plot



EDA: Profit (Adjusted) Vs Budget (Adjusted) Scatter Plot

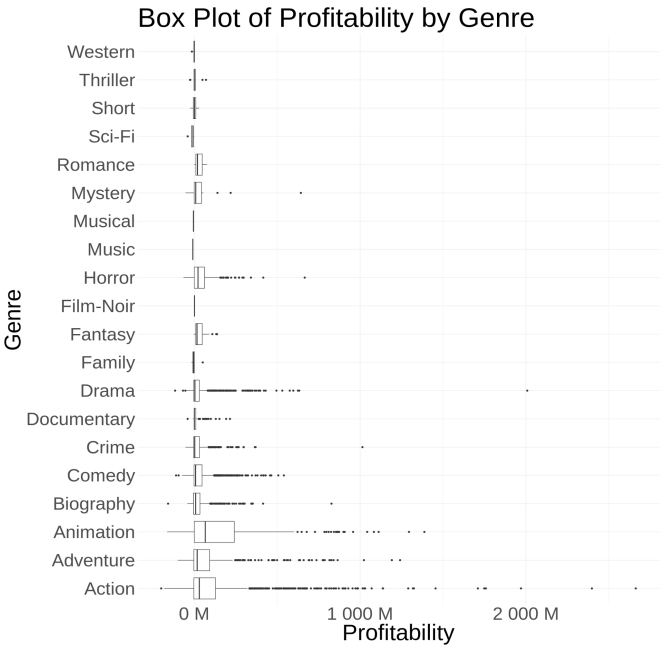


Observation 1

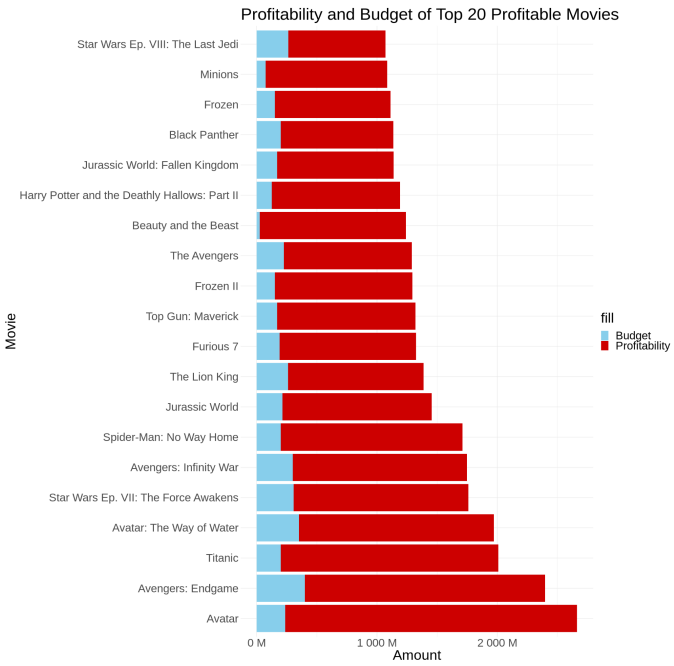
Linear -Linear plot shows heteroscedasticity in the data, thus a non-linear plot would be advised. The group adjusted the Profitability and Budget values using the year-on-year CPI values, and plotted all possible logarithmic plots and found the Log-Log plot to have the highest R-squared value. The plot and the summary of model quality can be found below.

Plot	R-Squared
Linear-Linear	0.34
Linear-Log	0.13
Log-Linear	0.28
Log-Log	0.4

EDA: Boxplots of Budget and Worldwide Gross aggregated by Genres



EDA: Top 20 Movies Profit-Budget Analysis

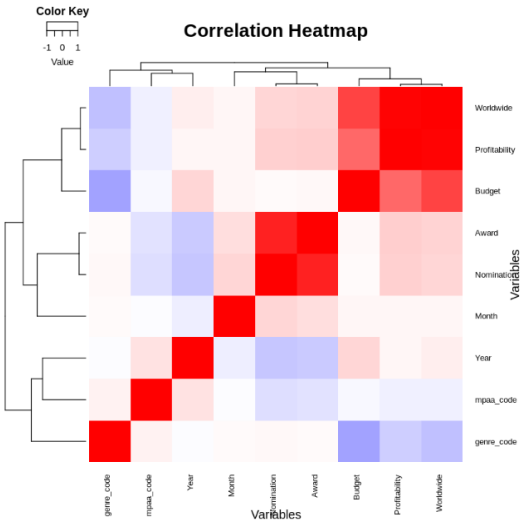


Observation 2

When comparing profit across genres, animation movies appear to do better based on the five number summaries followed by Action and Adventure. This observation is based on the median and max values of the boxplots.

It appears that the top 20 profitable movies have at least a 10x profit margin. The highest being Beauty and the Beast.

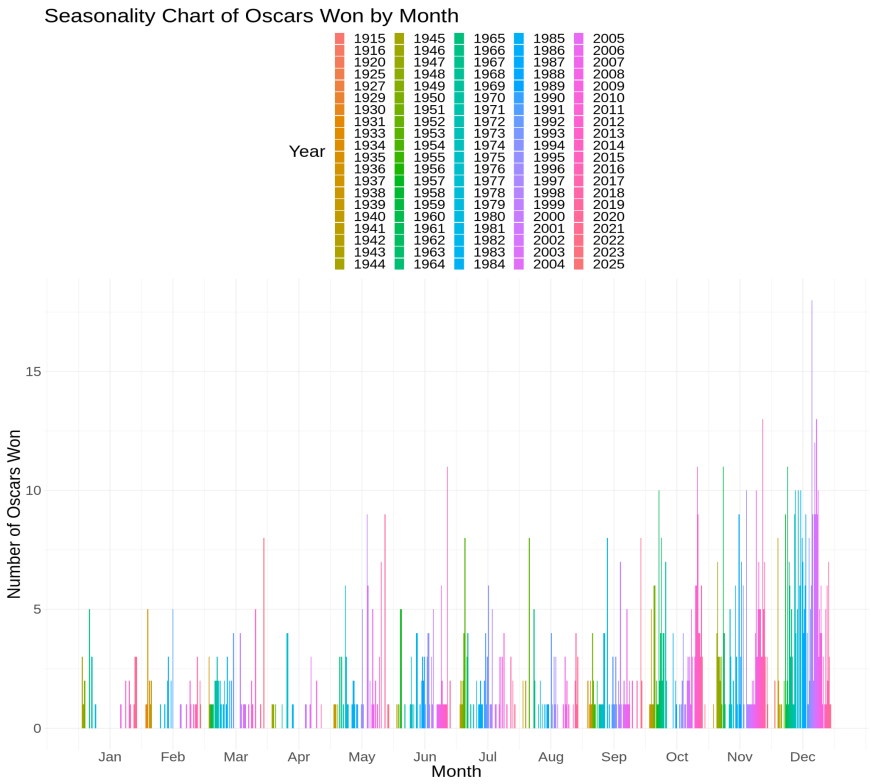
EDA: Correlation Heatmap



Observation 3

According to the updated correlation heatmap, **Budget** has a correlation of 0.59 with **Profitability**, which corroborates with the idea that a higher investment into a movie will lead to higher profitability. Other predictors demonstrate a weaker correlation with Profitability.

EDA: Seasonality Effect on Profitability and Likelihood of Winning An Oscar



Observation 4

From the chart, there seems to be a **higher chance of winning an Oscar** if the movie is released **in December** than compared to January.

HYPOTHESES/INSIGHTS

1. From the initial analysis, it can be observed that movies released in summer are less likely to win the Oscars.
2. Other considerations such as a larger budgetary investment and favorable critical reception (higher IMDB rating) contribute significantly to the success of a movie. This success can be measured by factors such as worldwide box office earnings and the achievement of prestigious awards like the Oscars.

PROJECT APPROACH AND MODELS

In carrying out the project, these are the major steps:

1. **Collecting and Preparing data**
   During the data preparation phase, both source datasets (movie titles with sales and ratings, movies with Oscar information) were merged into a unified dataset. The joins were performed based on the same movie title. If there is more than one movie of the same title, the older record was removed as there could be remakes of the same movie.

2. Feature Engineering

During the feature engineering phase, data imputation, data transformation, feature scaling and feature selection took place.

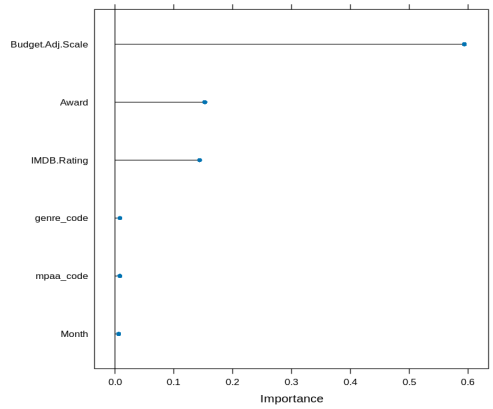
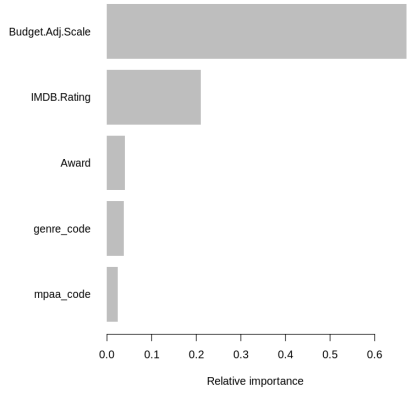
During initial analysis, some of the records had null Budget and Release dates, so data imputation was carried out using sources from the Internet. During the web scraping process, it was difficult to access the required data due to the unstructured data format of some websites. Another challenge encountered was the inclusion of unwanted characters in the Budget field and multiple Release dates. These required special handling.

Label encoding was used to encode categorical features such as IMDB Rating and MPAA into numeric value. CPI formula was obtained from the web and used to adjust monetary figures such as Profitability, Worldwide and Budget for inflation. These steps would have provided more accurate figures to predict monetary success and probability of winning awards of movies.

When fitting a regression model to the data, feature scaling was omitted at first, which gave rise to very high initial RMSE values. After feature scaling on monetary predictors, there was a significant improvement in the RMSE values and marginal improvements in R-Squared values.

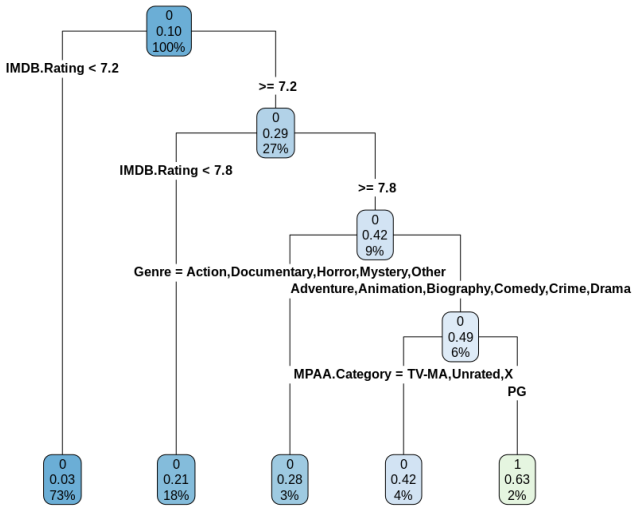
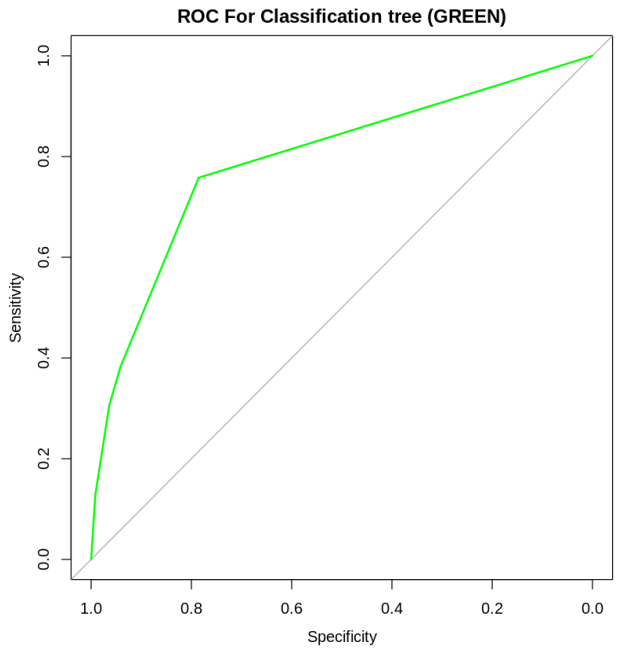
In ranking top 3 features for the model, both Elastic Net Regression and XGBoost models showed that Adjusted Budget, Award and IMDB Rating were significant in predicting profitability though in different order.

Feature Importance Graph

Elastic Net Regression	XGBoost
	

3. Model Selection and Evaluation

Choosing the models	The models chosen to answer the business questions included multilinear regression, decision trees and gradient boosting algorithms.
Training the models	Feature selection methods like regularization can be used to reduce the number of input variables prior to modeling. This helps in minimizing model training time and addressing overfitting concerns. Cross-validation techniques were employed during model training to prevent overfitting and reduce selection bias.
Evaluation of models	Depending on the type of problem (classification/regression), the performance of the models will be evaluated using respective metrics. For classification models, we use accuracy, confusion matrix, and AUC-ROC metrics. For regression models, we use Root Mean Squared Error (RMSE) and Adjusted R-Squared.

Classification Models	
Classification Model 1: Classification Tree for Likelihood of Winning an Oscar	Assessment of Classification Tree Model Quality using ROC/AUC
	

The following parameters would give the highest likelihood of winning an Oscars (63%).

IMDB Rating	> 7.8
Genre	adventure, animation, biography, comedy, crime or drama
MPAA Rating	PG

This model had an accuracy of 0.904 and an AUC of 0.79

	Actual No Oscar	Actual Win Oscar
Predict No Oscar	5213	44
Predict Win Oscar	519	76

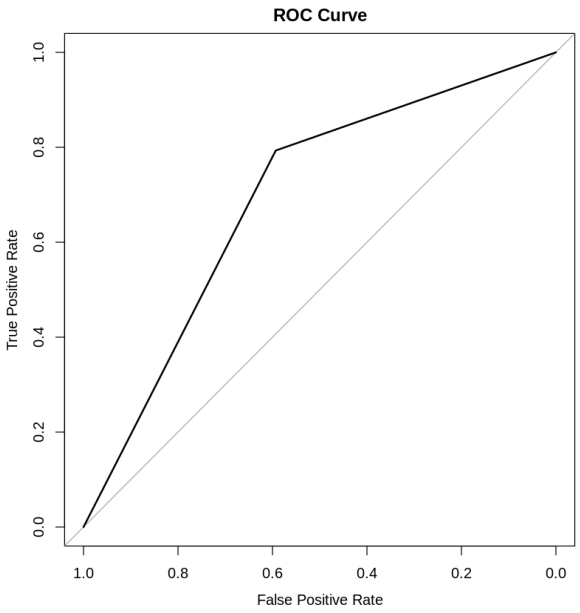
Classification Model 2: Logistic Regression for Likelihood of Winning an Oscar

Assess Logistic Regression Model Quality using ROC/AUC

Oscar_Winner ~ Budget + Worldwide + Duration + Month + Profitability

```
                 Estimate
(Intercept)   -6.444e+00
Budget        -1.836e-08
Worldwide      3.964e-09
Duration       3.028e-02
Month          1.190e-01
Profitability         NA
```

All coefficients are significant based on their p-values



Intercept	Represents estimated log-odds of response variable when all predictors are set to zero
Budget	A one-unit increase is associated with a decrease of -1.836e-08 in the log-odds of winning an Oscar, holding other variables constant. This means

	that higher values of Budget are associated with slightly lower odds of winning an Oscar.
Worldwide	A one-unit increase is associated with an increase of 3.964e-09 in the log-odds of winning an Oscar, holding other variables constant. This suggests that higher values of Worldwide are associated with slightly higher odds of winning an Oscar.
Duration	A one-unit increase is associated with an increase of 3.028e-02 in the log-odds of winning an Oscar, holding other variables constant. This means that longer durations are associated with higher odds of winning an Oscar.
Month	A one-unit increase is associated with an increase of 1.190e-01 in the log-odds of winning an Oscar, holding other variables constant. This suggests that later months (higher numerical values) are associated with higher odds of winning an Oscar.
Profitability	Without the specific coefficient value for Profitability, it is not possible to provide a precise interpretation.

Accuracy of the model is 0.613 and AUC is 0.6931

	Actual No Oscar	**Actual Win Oscar**
Predict No Oscar	3118	2140
Predict Win Oscar	123	472

Based solely on the coefficients, Duration and Month appear to be the most important predictors.

Regression Models	
Regression Model 1 Multilinear Regression on Profitability	Assess Regression Model Quality using RMSE
Profitability.Adjusted ~ Budget.Adjusted +	RMSE 9.715838e-15

IMDB.Rating + Month + Award + genre_code + mpaa_code

R-Squared 0.3982

```
Call:
lm(formula = Profitability.Adj.Scale ~ Budget.Adj.Scale + IMDB.Rating +
    Month + Award, data = oscar_df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8362 -0.2717 -0.0374  0.1650 12.8201

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.907512   0.065758 -13.801  < 2e-16 ***
Budget.Adj.Scale  0.579058   0.010184  56.857  < 2e-16 ***
IMDB.Rating       0.144269   0.010027  14.388  < 2e-16 ***
Month            -0.008347   0.003027  -2.758  0.00583 **
Award             0.162262   0.012585  12.893  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7769 on 5825 degrees of freedom
  (22 observations deleted due to missingness)
Multiple R-squared:  0.3986,    Adjusted R-squared:  0.3982
F-statistic: 965.3 on 4 and 5825 DF,  p-value: < 2.2e-16
```

Budget, IMDB.Rating and Award have p-values greater than 0.05 hence they are significant predictors for Profitability.

Intercept	when all predictors are zero, the expected profitability is approximately -0.908.
Budget.Adjusted	one unit increase is associated with an increase of 0.579 in the expected profitability, assuming all other variables remain constant.
IMDB.Rating	one unit increase is associated with an increase of 0.144 in the expected profitability, assuming all other variables remain constant.
Month	one unit increase is associated with a decrease of 0.008 in the expected profitability, assuming all other variables remain constant.
Award	one unit increase is associated with an increase of 0.162 in the expected profitability, assuming all other variables remain constant.

Gradient Boosting to Improve Regression Model	RMSE 0.8299 R-Squared 0.503

Feature	Gain	Cover	Frequency
<chr>	<dbl>	<dbl>	<dbl>
Budget.Adj.Scale	0.66687053	0.45553519	0.43548387
IMDB.Rating	0.20496256	0.15752831	0.19713262
Award	0.03730410	0.12385003	0.08781362
Month	0.03302834	0.07161175	0.09856631
mpaa_code	0.03083485	0.08458836	0.08960573
genre_code	0.02699961	0.10688635	0.09139785

```
#Calculate the root mean square error (RMSE) for test set
residuals = test_y - pred
RMSE = sqrt(mean(residuals^2))
print(paste0("RMSE = ", round(RMSE,4)))
```

```
[1] "RMSE = 0.8299"
```

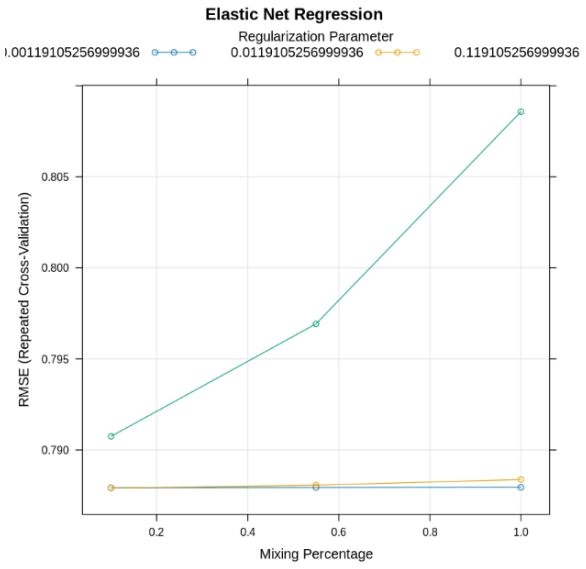
```
#Calculate R-squared for test set
y_test_mean = mean(test_y)

#Calculate total sum of squares
TSS = sum((test_y - y_test_mean)*(test_y - y_test_mean))

#Calculate residual sum of squares
RSS = sum(residuals^2)

#Calculate R-squared
R_squared = 1 - (RSS/TSS)
print(paste0("R-squared = ", round(R_squared,3)))
```

```
[1] "R-squared = 0.503"
```

Regression Model 2

Elastic Net Regression on Profitability

Assess Elastic Net Regression Model Quality using RMSE



glmnet variable importance Overall Budget.Adj.Scale 0.593758 Award 0.152654 IMDB.Rating 0.143879 genre_code 0.008385 mpaa_code 0.008354 Month 0.006284	RMSE 0.792947245137725 R_Squared 0.393605445992576

Conclusions

We used four models - Classification Trees, Logistic Regression, Multilinear Regression (Gradient Boosting) and Elastic-Net Regression. The reason for the model choices is that our team wanted two models for predicting Oscar likelihood, and two models for predicting profitability.

Out of all our models, the Classification Tree performed the best.

From our model outputs we can conclude the following

The following parameters would give the highest likelihood of winning an Oscars (63%).

IMDB Rating	> 7.8
Genre	adventure, animation, biography, comedy, crime or drama
MPAA Rating	PG
Duration	Longer
Month	Later part of the year

The following parameters would help predict profitability the best
Budget and IMDB.Rating
All our models affirm our intuition that a higher budget will lead to better Return On Investment (ROI). Also, PG film would mean more accessibility and hence a higher viewership. Also confirmed in our EDA, the later month will contribute to more movie success.

Both models also show that the top 3 important features that affect profitability of a movie are:

1. Budget.Adjusted
2. IMDB Rating
3. Award

Comparing RMSE values of both XGBoost gradient boosting and elastic net models, the XGBoost model has a lower RMSE value, indicating that it is a better fitting model.

GITHUB SOURCE CODE

The progress of our project can be viewed on the Team 1 Github page under the “Progress Report” Folder

https://github.com/MGT-6203-Summer-2023-Edx/Team-1/tree/main/Final%20Report

LITERATURE REFERENCES

1. Kaggle. (n.d.). Predicting movie success. Retrieved from https://www.kaggle.com/code/harshadeepvattikunta/predicting-movie-success

2. Folaron, D. (2019). Predicting movie box office success: Determining the influence of critical reception, genre, and budget. University of Tennessee, Knoxville. Retrieved from https://trace.tennessee.edu/cgi/viewcontent.cgi?article=3282&context=utk_chanhonoproj#:~:text=With%20information%20like%20the%20budget,more%20or%20less%20than%20expected.