CPSC 66 Final Report: Examples and Requirements

Student 1 Name JBOOKER1@SWARTHMORE.EDU
Zixi Gao ZGAO1@SWARTHMORE.EDU
Student 3 Name NLOWE1@SWARTHMORE.EDU
Student 4 Name OOGUNDE1@SWARTHMORE.EDU

Abstract

A one- or two-paragraph abstract that outlines the central goal and results of the project. This is your 30-second elevator pitch where you sell a reader on reading your paper. It should be 200 words maximum.

1. Introduction

What you attempted to do and what was the motivation for your work. You should provide some context about the problem including any relevant background about the task and related work.

2. Methods

This paper aimed to predict film revenue based on several film features, such as budget, runtime, etc. We developed a pipeline for data analysis:

1. Data preprocessing to clean up the dataset.

2. Exploratory analysis to evaluate potential relationships between revenue and individual features using scatter plots and simple linear regression.

3. Advanced machine learning models application to capture nonlinear interactions among features.

We developed a pipeline to classify film revenue by features, illustrated in Figure 1.

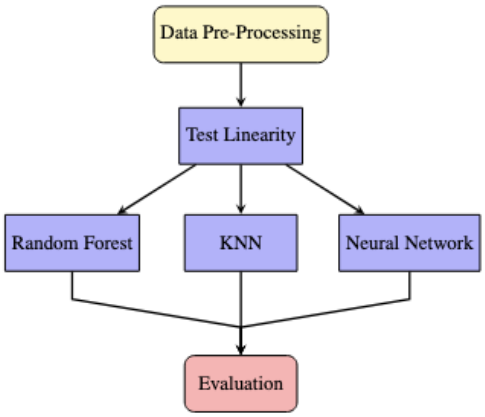

Figure 1. Workflow for film revenue classification. Starting with data pre-processing, followed by assessing linear relationships between features and revenue. Three non-linear machine learning models, Random Forest, K-Nearest Neighbors (KNN), and Neural Network, are applied. The models are then evaluated to determine their effectiveness in predicting film revenue.

2.1. Linearity Test

To evaluate the initial relationship between individual features and revenue, we performed scatter plot analysis and simple linear regression for interpretable visual and statistical insights. Scatter plots had revenue plotted on the y-axis and each feature on the x-axis to visualize data distribution. Simple linear regression was applied, and the coefficient of determination (R^2) was calculated to measure the percentage of variance in revenue explained by each feature.

The predicted revenue ($\hat{y}$) was computed using the formula:

$$\hat{y} = m \cdot x + c$$

where:

- m (slope) represents the rate of change in revenue for

CPSC 66 Machine Learning Proceedings, Swarthmore College, Spring 2021.

a unit change in the feature:

$$m = \frac{\sum (x - \text{mean}(x)) \cdot (y - \text{mean}(y))}{\sum (x - \text{mean}(x))^2}$$

- c (intercept) is the predicted revenue when the feature value (x) is zero:

$$c = \text{mean}(y) - m \cdot \text{mean}(x)$$

The coefficient of determination (R^2) was calculated as:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \text{mean}(y))^2}$$

An example of pseudocode for generating the plots and R^2 is as follows:

Algorithm 1 Linearity Validation for Feature and Revenue

Input: Dataset with numerical features and revenue
Output: Combined scatter plots with R^2 values
Initialize a combined visualization graph
for *each feature* x_i *in features* **do**
  Extract feature x_i and revenue yi from the dataset
  Create a scatter plot of all (x_i, yi) pairs
  Annotate the plot with calculated R^2
Save all annotated plots as a combined visualization graph

Our analysis of the scatter plots and R^2 values informed the selection of features for more advanced modeling.

2.2. Models Application

Building on the results of the linear regression analysis, we implemented the following machine learning models to capture non-linear relationships and interactions among features:

2.2.1. Neural Network

A neural network is a non-linear machine learning model inspired by the human brain that consists of interconnected layers of nodes to learn patterns through training (Hardesty, 2017). Their suitability for revenue prediction lies in two key aspects:

1. The ability to extract abstract representations in hidden layers, enabling them to handle the complexity of multi-class classification.
2. The capability to handle the large dataset, which supports robust training and enhances the model's capacity to generalize effectively, allowing accurate classification of films into three revenue ranges.

The neural network's performance is evaluated using accuracy and loss plots, while hyperparameters are tuned iteratively to optimize these metrics.

3. Experiments and Results

3.1. Linearity Test

3.1.1. Results

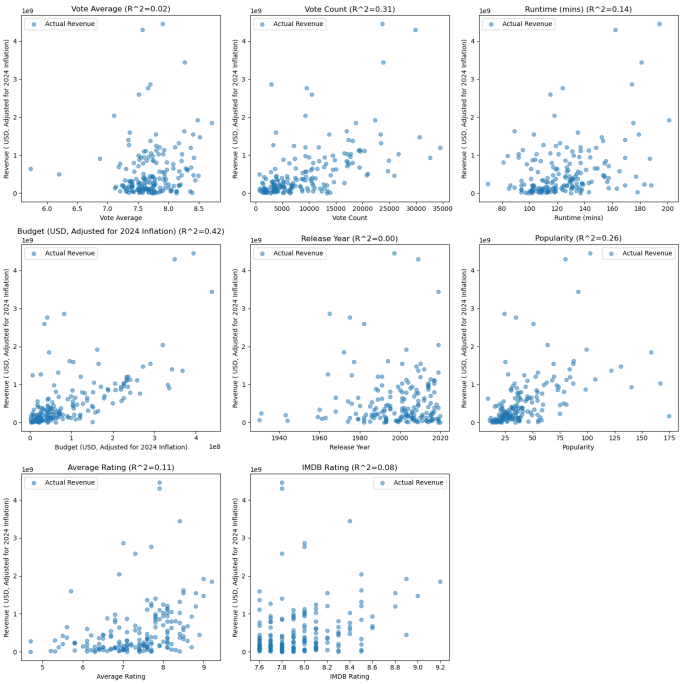

Figure 2. Scatter plots illustrating the relationship between individual features on the x-axis and revenue on the y-axis. The R^2 values indicate the strength of linearity. R^2 values closer to zero indicate weak linear relationships, while R^2 values closer to one indicate stronger linear relationships. The results demonstrate weak to moderate linear correlations between individual features and revenue.

The linearity analysis revealed that no individual feature demonstrates a strong linear relationship with revenue (Fig 1.). Among all features, **Vote Count** exhibits the strongest linear relationship with revenue from the scatter plot. However, the $R^2 = 0.31$ is relatively low, indicating that only a small portion of revenue variance is explained by Vote Count. **Budget** shows moderate correlation with revenue ($R^2 = 0.42$); however, the scatter suggests significant variability for low-budget films. **Popularity** ($R^2 = 0.26$) exhibits non-linear patterns for high popularity scores. **Vote Average** ($R^2 = 0.02$), **Runtime** ($R^2 = 0.14$), **Release Year** ($R^2 = 0.00$), **Average Rating** ($R^2 = 0.11$), and **IMDb Rating** ($R^2 = 0.08$) showed weak or negligible correlations, with scatter plots indicating non-linear trends (Figure 2).

3.1.2. Discussion

The results suggest that no individual feature exhibits a strong linear relationship with revenue. This finding aligns with previous research highlighting the limitations of linear models for revenue classification, given the non-linear interactions among features (Sen Sharma, et al., 2021). Consequently, we concluded that non-linear models are better suited to capture the complex interactions among features and improve revenue classification.

3.2. Neural Network

To predict film revenue, a neural network was implemented and trained on the dataset containing over 8,000 samples.

3.2.1. Experimental Methodology

Data Collection and Preprocessing Data preprocessing was conducted as detailed in the Method section. All numerical features were normalized and standardized using the `StandardScaler` to ensure uniform feature scaling. Revenue values were adjusted for 2024 inflation and categorized into **two** or **three** classes for classification tasks. For the binary classification task, revenue was divided into:

- **Low**: $\leq 50,000,000$
- **High**: $> 50,000,000$

For multi-class classification, revenue was categorized into:

- **Low**: $\leq 25,000,000$
- **Medium**: $25,000,000 <$ revenue $\leq 120,000,000$
- **High**: $> 120,000,000$

The thresholds for revenue classification were manually selected to ensure a balanced distribution of films across the different classes, thereby enhancing the model's performance on both binary and multi-class classification tasks.

Model Architecture and Training A feed-forward neural network with three hidden layers was designed, employing ReLU activation functions, dropout for regularization, and batch normalization to stabilize training. The model was compiled with a stochastic gradient descent (SGD) optimizer and a categorical cross-entropy loss function to handle the multi-class target. A 70/30 train-test split was used for evaluation, and the model was trained for 300 epochs with a batch size of 64.

Evaluation Training and validation accuracy were tracked over epochs to monitor convergence and assess the model's generalization. The final model's accuracy was evaluated on the validation set, and the classification performance was summarized.

3.2.2. Result

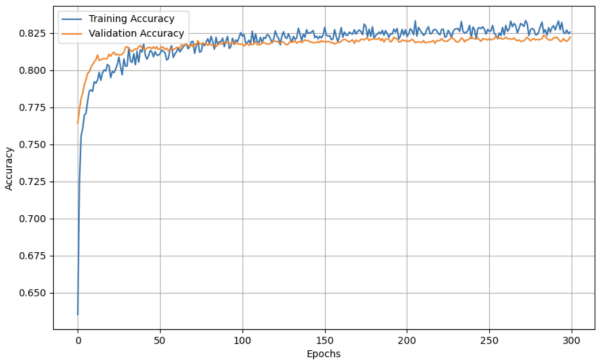

Figure 3. Training and validation accuracy of the neural network over 300 epochs with 2 revenue classes. The x-axis represents training epochs, and the y-axis represents accuracy. The curves demonstrate consistent convergence, with validation accuracy stabilizing at 82%, indicating effective generalization.

Table 1. Classification Report for Neural Network on Film Revenue Classification (Two Classes).

Class	Precision	Recall	F1-Score
Low Revenue (0)	0.81	0.83	0.82
High Revenue (1)	0.83	0.82	0.82
Accuracy		0.82	
Macro Avg	0.82	0.82	0.82
Weighted Avg	0.82	0.82	0.82

We observed the following during the training and evaluation of the neural network:

1. Both the training and validation accuracy curves showed rapid improvement during the initial epochs, stabilizing around 250 epochs (Figure 3).
2. The training accuracy stabilized at approximately 82%, and the validation accuracy remained close throughout, indicating minimal overfitting (Figure 3).
3. The final validation accuracy of 82% suggests that the neural network effectively captures the non-linear relationships among features (Figure 3).
4. The classification report further supports the model's performance. The precision, recall, and F1-scores

for all classes are consistent (Precision: 0.81, Recall: 0.83 for Class 0; Precision: 0.83, Recall: 0.82 for Class 1), and the macro-average and weighted-average metrics confirm that the model performs well across all classes, without bias toward any specific revenue range (Table 1).

Table 2. Classification Report for Neural Network on Film Revenue Classification (Three Classes).

Class	Precision	Recall	F1-Score
Low Revenue (0)	0.80	0.75	0.77
Medium Revenue (1)	0.62	0.65	0.63
High Revenue (2)	0.80	0.81	0.80
Accuracy		0.73	
Macro Avg	0.74	0.73	0.74
Weighted Avg	0.74	0.73	0.73

We also ran the neural network for three revenue classes and received lower accuracy and higher variability. Neural network achieves higher overall accuracy (82%) in the binary classification compared to 73% in three-class classification. The three-class classification also exhibits moderate precision and recall variability compared to binary class classification. In Table 2., the middle revenue range (Class 1) achieves Precision 0.62 and Recall 0.65, which are notably lower compared to the high (Class 2: Precision 0.80, Recall 0.81) and the low revenue range (Class 0: Precision 0.80, Recall 0.75) (Table 2). These results highlight the model failed to classify the middle revenue class compared to the other classes.

3.2.3. Discussion

The results demonstrate that the neural network effectively models the complex interactions among features, achieving **82% accuracy** and balanced performance for two revenue classes, outperforming linear models. This highlights the model's robustness in handling varying data distributions. The close alignment of training and validation curves further underscores the model's ability to generalize effectively.

For the three-class revenue classification, the neural network achieved a lower overall accuracy of **73%**, with greater variability. This variability is expected due to the middle revenue range often sharing characteristics with low- and high-revenue films. The additional complexity also makes it challenging for the neural network to accurately model the boundaries between classes.

Overall, the neural network demonstrates strong performance, with better accuracy and more consistent results in the two-class scenario due to the more distinct separation of classes. However, the specific features contributing most to the classification remain unclear, as the neural network operates as a black-box model.

4. Conclusions

Lessons learned. Wrap up the paper with a restatement of the initial hypothesis and your findings. Discuss unanswered questions/possible future work to further the study of this central question.

5. Social impact

Lessons learned. Wrap up the paper with a restatement of the initial hypothesis and your findings. Discuss unanswered questions/possible future work to further the study of this central question.

Acknowledgments

Place acknowledgements in an unnumbered section at the end of the paper. Typically, this will include to colleagues who contributed to the ideas, individuals who reviewed your submission, or external sources who helped acquire data.

Hardesty, L. Explained: Neural Networks. *MIT News Office*. Archived on 18 March 2024. Retrieved on 2 June 2022, 14 April 2017.

Sen Sharma, A., Roy, T., Rifat, S. A., & Mridul, M. A. Presenting a Larger Up-to-date Movie Dataset and Investigating the Effects of Pre-released Attributes on Gross Revenue. *arXiv preprint*, arXiv:2110.07039 [cs.IR], 13 October 2021. https://arxiv.org/abs/2110.07039