CM3015 Machine Learning and Neural Networks

Mid-Term Report

Authors: Teo Qi Xuan

Table of Contents

1. Abstract

The primary aim of this project is to develop and evaluate machine learning models to predict IMDb scores. Leveraging a dataset which contains relevant information such as runtime, release year, IMDb votes etc, the goal is to create an accurate predictive model that can effectively predict and estimate IMDb scores. The project aims to explore the relationships between the feature and target variables, providing insights into the factors that significantly contribute to the perceived quality of the movies/shows. Additionally, an assessment of different machine learning models will be conducted to optimise the predictive performance and enhance the overall understanding of the predictive modelling process.

2. Introduction

In the realm of machine learning, the ability to predict and understand movie ratings would benefit movie creators as it would be able to provide valuable insights into audience preferences and cinematic trends. The primary aim of this project is to leverage machine learning models to accurately predict IMDb scores for movies, contributing to the ongoing exploration of predictive modelling in the context of film evaluation. IMDb scores serve as a pivotal metric in gauging audience perceptions and influencing viewer choices, making the development of robust predictive models crucial for both movie creators and movie enthusiasts.

The investigation resolves around a comprehensive dataset comprising essential features such as release year, runtime, IMDb votes, and age_certification, offering a rich variety of information integral to the movie-watching experience. The unique characteristics and challenges embedded within this dataset propel the quest to uncover nuanced relationships between the features and the IMDb score, thereby enhancing the understanding of the intricate dynamics shaping audiences' judgments.

Despite the prolific research in the broader machine learning landscape, there exists a notable gap in the literature concerning the specific nuances of predicting IMDb scores. The project aims to address this gap by delving into the complexities of movie rating prediction, presenting an opportunity to contribute novel insights and methodologies to the ever-evolving field of predictive modelling within the realm of filmmaking. Through the use of machine learning algorithms, this project aspires to offer a comprehensive understanding of the intricate interplay between movie attributes and audience-perceived quality, paving the way for advancements in both cinematic analytics and machine learning applications.

3. Background

The project harnesses the power of machine learning algorithms to predict IMDb scores, relying on a foundation of well-established principles and methodologies within the field. The primary algorithm employed is linear regression, a fundamental and widely used supervised learning technique. It models the relationship between the dependent variables (IMDb scores) and the independent variables (release year, runtime, age certification, IMDb votes). The algorithm aims to fit a linear equation to the observed data, aiming to capture the underlying patterns and associations within the chosen features.

In addition to linear regression, polynomial regression was also used to capture non-linear dependencies present in the data. The principle of polynomial regression involves transforming the input features into higher-degree polynomials, introducing curvature and enabling the model to better fit the training data. This increased flexibility is crucial when dealing with complex relationships, such as those found in movie rating prediction, where factors influencing audience preferences are often multifaceted and non-linear.

The choice to incorporate polynomial regression is motivated by the need to unravel hidden patterns and capture the inherent complexity in the dataset, ultimately enhancing the model's predictive capabilities. The combination of linear and polynomial regression enables a more nuanced understanding of the relationships between movie features and IMDb scores.

The project aligns with established concepts in the machine learning literature, where the interplay between linear and polynomial regression is explored in diverse applications. Concepts such as model complexity, regularisation, and the impact of polynomial degrees on model performance are integral to the project's methodology, reflecting a comprehensive understanding of the chosen algorithms. The utilisation of the different machine learning models offers a holistic approach to capturing the richness of factors influencing movie ratings.

4. Methodology

The project took a systematic approach to prepare and analyse the dataset for IMDb score prediction.

The preprocessing steps taken are:

1. Handling Missing Values: Any instances of missing values were removed to ensure the integrity of the dataset.
2. Column Selection: Columns such as index, id, and imdb_id were dropped as they were irrelevant to the predictive modelling.
3. Data Type Conversion: The IMDb vote column was converted to integer format.

The use of k-fold cross-validation was motivated by the need to assess the generalisation performance of the models. By partitioning the dataset into k subsets and systematically using each as a validation set while k-1 remaining subset from the training set. This technique provides a robust evaluation metric that helps mitigate issues related to overfitting or model instability. The reported average mean squared error across folds serves as a reliable indicator of the model's overall predictive capability.

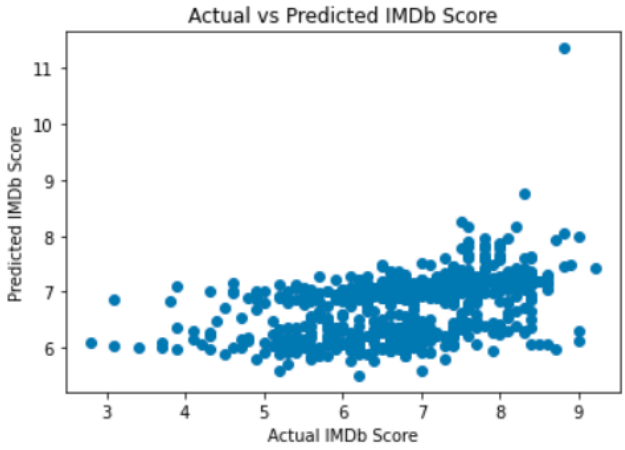
Decisions throughout the project were guided by the principle of achieving a balance between model complexity and generalisation. The preprocessing steps taken were aimed to streamline the dataset without compromising the quality of information. The choice of polynomial regression was driven by the hypothesis that the relationship between features and IMDb scores might exhibit nonlinear patterns. Cross-validation was essential to avoid overfitting and ensure the model's performance on diverse subsets of the dataset. This methodology allowed for a comprehensive exploration of the dataset, consideration of nonlinear relationships, and robust evaluation of model performance through cross-validation.

5. Results

5.1. Linear Regression

The linear regression model underwent training on a dataset featuring both categorical and numerical data. Categorical variables such as type and age certification were encoded using the OneHotEncoder. Evaluation on a test set yielded the subsequent performance metrics:

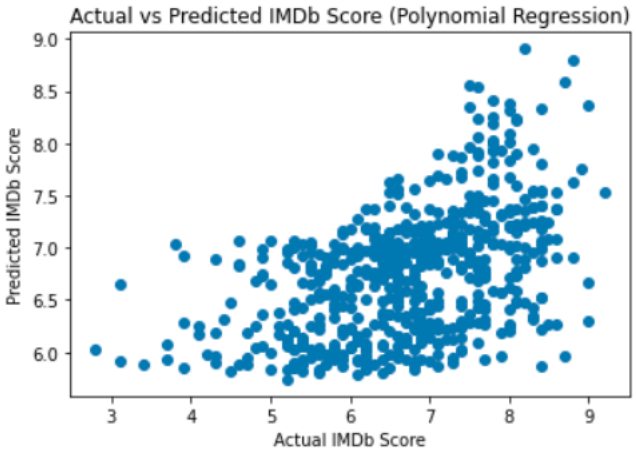
```
Mean Squared Error: 0.9829865907557113
Model Score: 0.1974577249716778
```



5.2. Polynomial Regression

In the case of polynomial regression model with a degree of 2, a pipeline incorporating Polynomial Features and Linear Regression was employed. The model's performance on the test set is as follows:

```
Polynomial Regression Mean Squared Error: 0.9375811544408333
Polynomial Regression Model Score: 0.23452820233269767
```



5.3. Cross-Validation

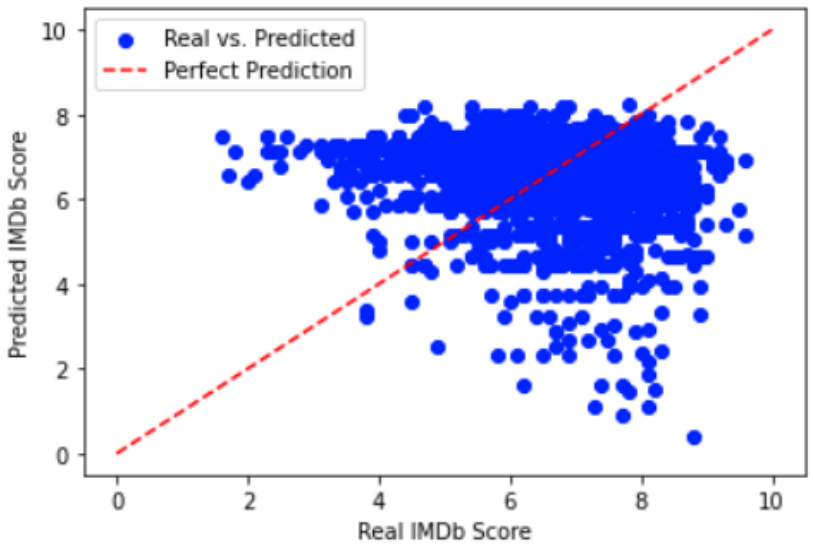
To evaluate the model's robustness, K-Fold cross-validation was executed with k = 5. The mean squared error for each fold and the average mean squared error across all folds are

```
Fold 1, Mean Squared Error: 14.56543689640645
Fold 2, Mean Squared Error: 1.107524883449692
Fold 3, Mean Squared Error: 1.2220036944184478
Fold 4, Mean Squared Error: 1.0210765806469742
Fold 5, Mean Squared Error: 1.3880145152569632
Average Mean Squared Error: 3.860811314035705
```

5.4. Custom Gradient Descent/Linear Regression

A custom gradient descent method was implemented for linear regression. The model was trained using scaled features and underwent 1000 iterations, resulting in the following insights:

```
Iteration 0, Mean Squared Error: 46.58591898225645
Iteration 100, Mean Squared Error: 4.108791369062303
Iteration 200, Mean Squared Error: 2.398924822650133
Iteration 300, Mean Squared Error: 2.326607807359628
Iteration 400, Mean Squared Error: 2.320253543820202
Iteration 500, Mean Squared Error: 2.3167197311918897
Iteration 600, Mean Squared Error: 2.3134502960279377
Iteration 700, Mean Squared Error: 2.3103249372749914
Iteration 800, Mean Squared Error: 2.3073233374444504
Iteration 900, Mean Squared Error: 2.3044310866390356
```



6. Evaluation

The evaluation of this project encompasses a critical examination of its strengths and weaknesses, keeping in mind the overarching aim to predict IMDb scores based on selected features. The success of the project is not solely based on achieving the set objectives but also hinges on the meticulousness and depth of the undertaken processes.

6.1. Strengths

1. Diverse Model Exploration: The project explored multiple machine learning models including linear regression and polynomial regression. This diversity allows for a comprehensive understanding of how varying models respond to the dataset.
2. Cross-Validation: By using K-Fold cross-validation it provides a robust assessment of the model's generalisation capabilities. It helps mitigate overfitting and provides a more realistic estimation of the model's performance.
3. Custom Gradient Descent/Linear Regression: The inclusion of the custom-made gradient descent method for linear regression demonstrates the understanding of the algorithms. The model was created only with numpy and matplotlib.

6.2. Weaknesses

1. Limited Feature Exploration: The choices of features used can be considered limited. The inclusion of additional relevant features could enhance the model's predictive capabilities.
2. Assumption of linearity: The project assumes that there is a linear relationship between the features and the target. This might oversimplify the underlying dynamics of the dataset.

6.3. Critical Awareness

The project acknowledges the challenges and complexities associated with predicting IMDb scores accurately. The aim to predict subjective ratings based on a set of features is inherently ambitious, and the project recognises the limitations within the scope of the dataset. It is crucial to underscore the project's aim while ambitious aligns with the objective of machine learning applications.

6.4. Insights and Understanding

The project has fostered a deeper and more comprehensive understanding of machine learning algorithms, model evaluation techniques, and the iterative nature of model refinement. The implementation of custom models demonstrates the understanding of the mechanics behind the models employed.

6.5. Summary

In summary, the project strikes a balance between ambition and pragmatism. While the aim of predicting the scores is challenging, the project contributes valuable insights into the strengths and weaknesses of various regression models. The critical awareness demonstrated throughout the project showcases a nuanced understanding of the complexities inherent in predictive modelling.

7. Conclusions

In conclusion, the project aims to predict IMDb scores. The findings revealed insights into the model's predictive capabilities and their alignment with the project's aim.

Linear Regression: The model demonstrated moderate performance, revealing a correlation between the features selected and the IMDb scores. However, its simplicity may limit capturing complex relationships.

Polynomial Regression: Introducing a non-linear relationship improved the predictive prowess, with a lower mean squared error and enhanced the model's score.

Custom Gradient Descent/Linear Regression: The implementation offered hands-on exploration and demonstrated a deep understanding of machine learning algorithms. The mean squared error gets better after each iteration. However, the method might not be scalable.

Cross-Validation: K-Fold cross-validation underscored the model's robustness, validating the generalisation capabilities.

While the results are promising, further feature exploration and advanced techniques could enhance the predictive accuracy. These findings provide a foundation for future improvements.

8. References

Netflix IMDB Scores. https://www.kaggle.com/datasets/thedevastator/netflix-imdb-scores. Accessed 19 Dec. 2023.