**Title: IMDb Score Prediction Project Documentation**

**Table of Contents**

1. Introduction

2. Problem Statement

3. Design Thinking Process

4. Phases of Development

5. Dataset Description

6. Data Preprocessing Steps

7. Model Training Process

8. Choice of Regression Algorithm

9. Evaluation Metrics

10. Conclusion

**1. Introduction**

This document provides a comprehensive overview of the work conducted in the IMDb Score Prediction project. The project aimed to develop a machine learning model to predict IMDb scores for Netflix original films, documentaries, and specials. The dataset used in this project consists of all Netflix original films released as of June 1st, 2021, and includes their corresponding IMDb scores obtained through web scraping.

**2. Problem Statement**

The primary objective of the project was to build a predictive model that accurately estimates IMDb scores based on various features of Netflix original films, documentaries, and specials. The model would assist in evaluating the potential success or quality of a movie based on its characteristics.

**3. Design Thinking Process**

The development of the IMDb Score Prediction project followed the design thinking process, which comprises the following stages:

- Empathize: Understanding the problem domain and the needs of stakeholders.

- Define: Clearly defining the problem statement, goals, and objectives of the project.

- Ideate: Generating potential solutions and exploring different approaches to movie score prediction.

- Prototype: Building initial versions of the predictive model and iterating on it.

- Test: Evaluating and refining the model through experimentation and validation.

- Implement: Deploying the final model and integrating it into the desired application or system.

**4. Phases of Development**

The project was divided into two key phases:

**Phase 1: Data Acquisition and Preprocessing**

This phase involved the collection of data from the Wikipedia page, which included information about Netflix original films, documentaries, and specials. The IMDb scores were obtained by integrating this data with a dataset consisting of the corresponding IMDb scores. The phase included steps such as web scraping, data integration, exploratory data analysis (EDA), and data preprocessing.

**Phase 2: Model Training and Evaluation**

In this phase, a regression algorithm was selected and trained using the preprocessed dataset. The model was evaluated using appropriate evaluation metrics to assess its predictive performance. The phase included tasks such as model selection, data splitting, model training, and model evaluation.

**5. Dataset Description**

The dataset used in the project consisted of all Netflix original films released as of June 1st, 2021, along with Netflix documentaries and specials. The data was obtained by web scraping from a Wikipedia page. It was then integrated with a dataset that included the corresponding IMDb scores for the movies. IMDb scores are community-voted ratings, with the majority of films having 1,000+ reviews.

**6. Data Preprocessing Steps**

Before training the model, the dataset underwent several preprocessing steps. These steps included handling missing data, feature selection, feature encoding, and feature scaling. Missing values were imputed or removed, relevant features were selected, categorical variables were encoded, and numerical features were scaled to ensure consistent ranges.

**7. Model Training Process**

A regression algorithm, specifically a gradient boosting regressor, was chosen for training the model. The dataset was split into training and testing sets, and the model was initialized with suitable hyperparameters. The model was trained on the training set, and hyperparameter tuning techniques such as grid search or random search were applied. The trained model was evaluated on the testing set to assess its predictive performance.

**8. Choice of Regression Algorithm**

The gradient boosting regressor was selected as the regression algorithm for this project. This algorithm was chosen due to its ability to handle complex relationships between features and target variables, as well as its good performance in regression tasks. The advantages and suitability of this algorithm for predicting IMDb scores were considered, along with comparisons to other regression algorithms.

**9. Evaluation Metrics**

The performance of the trained model was evaluated using several metrics. Mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score were used as evaluation metrics. MSE measures the average squared difference between predicted and actual IMDb scores, while RMSE provides a more interpretable metric in the original scale of the scores. R2 score represents the proportion of variance in the target variable that can be explained by the model.

**10. Conclusion**

The IMDb Score Prediction project aimed to develop a machine learning model for predicting IMDb scores of Netflix original films, documentaries, and specials. The project followed the design thinking process and involved two key phases: data acquisition and preprocessing, and model training and evaluation. A regression algorithm, the gradient boosting regressor, was chosen for model training, and evaluation metrics such as MSE, RMSE, and R2 score were used to assess the model's performance. The project holds promise in assisting stakeholders in evaluating the potential success or quality of Netflix content.