Jovani Benavides
Matthew Alvarez
Ira Dionisio
CSCI-164
20 April 2025

The Examination of Machine Learning Algorithms Across Different Data Sets

In this study, we evaluated the performance of various supervised machine learning algorithms across two distinct datasets: one related to movie ratings and the other focused on vehicle sales. The movie dataset, sourced from IMDb, was used to predict film ratings based on structured metadata using Linear Regression, Ridge Regression, and Multi-Layer Perceptron (MLP) models. For the car dataset, the objective was to predict vehicle selling prices using Linear Regression, Random Forest, and K-Nearest Neighbors (KNN). The analysis began with comprehensive data preprocessing for both datasets to ensure data quality, normalize feature distributions, and prepare the inputs for effective model training and evaluation.

For the car dataset, we used regression models to predict the selling price of a car given its feature set. For the preprocessing, the numerical features were standardized, and the categorical features were encoded using one-hot encoding. Using 80% training and 20% testing data, to ensure that the model was accurate. Using Mean Squared Error (MSE) and R-squared value, the regression models were tested for accuracy. The models included linear regression, best random forest, and K-nearest neighbors (KNN) regressor. Prior to tuning the hyperparameters, the MSE and R-squared values, respectively, for the linear regression model were 9.23 and 0.60, for the random forest regressor 0.77 and 0.97, and for the KNN regressor 1.28 and 0.94. Linear regression doesn’t benefit from hyperparameter tuning and was therefore skipped. The best random forest regressor model had found these to be the best hyperparameters

from a parameter grid: {'regressor__max_depth': 10, 'regressor__min_samples_leaf': 1, 'regressor__min_samples_split': 2, 'regressor__n_estimators': 100}. KNN Nearest neighbors found these to be the best hyperparameters from a parameter grid: {'regressor__n_neighbors': 3, 'regressor__p': 2, 'regressor__weights': 'distance'}. The new MSE and R-squared values, respectively, for best random forest are 0.75 and 0.97 and for KNN regressor are 1.09 and 0.95. While KNN had the most improvement from the hyperparameter tuning, the best random forest still had the best values for accuracy. That can be seen in the graphs given on GitHub.

For the movie dataset we utilized supervised learning to predict the IMBDb movie rating. Within the movie data set there were 5000 films and each film had several key features that help predict the score of a film. The goal of utilizing this database is to evaluate the performance of multiple regression models. Compare their predictive effectiveness, and benchmark the results against prior published studies. In order to ensure quality and consistency, the dataset went through preprocessing. Records that were missing were removed from the dataset and some features were log-transformed to avoid skewness and impact of extreme values. For our model to have the best performance we used features such as duration, log budget, log gross, number of critics reviewing, total facebook likes of the cast, the directors facebooks likes, and the content rating. All these features will be used to help our models predict the actual IMBd score for films. After running the 3 supervised regression models, all the models were evaluated using MSE and R^2. After analyzing the models performance we saw that MLP outperformed the other two models, achieving the lowest error MSE=0.7703 and the highest $R^2 = 0.3379$. This shows that the neural network was better able to capture the non-linear patterns present in the data.

As part of the project, we were tasked with comparing our work to prior research using the same dataset. For this comparison, we selected the study “Predicting IMDb Movie Ratings

Using Supervised Machine Learning" by Joe Cowell. In his analysis, Cowell used tree-based models, specifically Random Forest and Gradient Boosting, on a filtered subset of the IMDb dataset containing movies from 2000 to 2020. He began by modeling using only runtime as a feature, achieving an R^2 of 0.2687. As he incorporated additional features, his model's performance improved, reaching an R^2 of 0.432. In contrast, our study used the entire IMDb 5000 Movie Dataset (~5,000 entries), while Cowell used a smaller subset (~2,000 entries). This difference in dataset size and composition may partially explain the performance gap between our models. Additionally, while both studies used the same source data, the feature sets and preprocessing strategies differed significantly. Our approach focused on seven structured features, applying log transformations, categorical encoding, and feature scaling. Cowell used a broader feature set, including text-based fields and performed NLP-style preprocessing, genre extraction, and engineered new popularity indicators. These additional features likely provided his models with more predictive context, especially since genre and user interaction metrics are closely tied to public reception. Therefore, although Cowell's model outperformed ours in terms of R^2, the difference is largely attributable to feature richness, rather than fundamental modeling choices. It is also important to acknowledge that IMDb score prediction is inherently challenging due to the subjective nature of movie ratings. A film might be critically acclaimed yet poorly received by the public or vice versa. This leads to noise in the target variable that no model can easily resolve. Ultimately, our comparison suggests that extending our feature set to include genre data and audience engagement metrics would likely enhance performance. Furthermore, our results indicate that while MLP performed best among our models and tree-based models as used by Cowell may be more effective for this type of structured data.

This study demonstrated the practical application of supervised machine learning techniques to two distinct real-world prediction tasks: car price estimation and IMDb movie rating prediction. Through structured preprocessing, feature selection, and algorithmic experimentation, we achieved strong predictive performance on both datasets. The Random Forest model proved highly effective in predicting car prices, achieving an R^2 value of 0.97 after hyperparameter tuning, while the MLP model yielded the best results for predicting IMDb scores, albeit with more modest performance due to the subjective nature of the target variable. The comparative analysis with prior work highlighted the importance of feature diversity and engineering. Although our models did not outperform those in external studies, our methodology focused on clarity, reproducibility, and rigorous evaluation. Overall, this project not only reinforced fundamental machine learning concepts such as data preprocessing, model tuning, and evaluation, but also emphasized the limitations and challenges of predictive modeling in domains influenced by human preferences. Future improvements may include incorporating more advanced features, such as natural language elements and ensemble learning strategies, to further enhance performance and generalizability.

Work Cited

Cowell, J. (2020, April 17). Predicting IMDb Movie Ratings Using Supervised Machine Learning. Towards Data Science. https://towardsdatascience.com/predicting-imdb-movie-ratings-using-supervised-machine-learning-f3b126ab2ddb

Zhang, C. (2018). *IMDb 5000 Movie Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset

Mohaiminul. (2021). *Car Price Prediction* [Code and Data]. Kaggle. https://www.kaggle.com/code/mohaiminul101/car-price-prediction/input

UCI Machine Learning Repository. (n.d.). IMDb Dataset. University of California, Irvine. https://archive.ics.uci.edu/dataset/132/movie