A COMPREHENSIVE STUDY ON MOVIE DATA ANALYSIS: SEMMA METHODOLOGY WITH AUTOML AND MANUAL APPROACHES

Aditi Patil
Student
San Jose State University
San Jose
aditi.patil@sjsu.edu

ABSTRACT

This paper provides a meticulous exploration and analysis of the IMDB movie dataset through the lens of the SEMMA methodology, combining both Automated Machine Learning (AutoML) and traditional manual approaches to uncover insights and patterns. Beginning with a representative sample, we delve into explorative analysis to comprehend the intricate structures and relationships within the data. Subsequent modification includes rigorous data cleaning, feature engineering, and categorical variable encoding to prepare the dataset for modeling. Employing Linear Regression as a baseline, we predict IMDB scores to assess the impact of various movie attributes, followed by a thorough assessment of the model's residuals, limitations, and potential improvements. A deep dive into evaluation metrics reveals patterns and anomalies, guiding future enhancements and refinements in model performance. The synthesis of SEMMA with AutoML and manual techniques presents a balanced approach, allowing for detailed understanding and optimization in data-driven decision-making processes within the film industry. The comprehensive study elucidates the versatility and applicability of data mining techniques in extracting meaningful information and forecasts from movie datasets, paving the way for advanced research in film analytics.

Keywords: SEMMA Methodology, AutoML, IMDB Movie Dataset, Data Mining, Linear Regression, Data Analysis, Feature Engineering, Data Cleaning, Machine Learning, Film Industry Analytics, Predictive Modeling, Residual Analysis, Evaluation Metrics.

1 Introduction

Cinema, a mirror reflecting the tapestry of human experiences, emotions, and stories, has been a cornerstone of global culture for over a century. Beyond the ephemeral flicker of moving images lies a treasure trove of data, waiting to be explored, promising insights into the myriad facets of filmmaking and audience reception. This study embarks on an analytical odyssey, navigating the intricate landscape of movie data to unravel the hidden patterns, trends, and correlations within the IMDB movie dataset.

Employing the structured SEMMA (Sample, Explore, Modify, Model, and Assess) methodology, this research meticulously dissects the dataset, shedding light on the subtleties of movie attributes and their impact on IMDB scores. The integration of Automated Machine Learning (AutoML) enhances the analytical process, balancing the rigor of structured methodology with the versatility and efficiency of automation.

The objective of this paper is to explore the rich tapestry of movie data, to understand the underlying patterns and relationships between different movie attributes and their influence on movie success, as measured by IMDB scores. The insights derived from this analysis aim to contribute to the burgeoning field of movie analytics, offering nuanced understandings of movie attributes and their implications on audience reception and commercial success.

In the ensuing sections, this paper will navigate through the structured stages of SEMMA, integrating the advancements of AutoML, and delve deep into the analytical realms of the cinematic world, presenting the findings, discussions, and implications of this research.

2 Related Work

The exploration of movie datasets to extract meaningful insights and predictions is a well-trodden domain, with numerous studies employing varied methodologies and approaches. The utilization of the SEMMA methodology in data mining tasks is exemplified by [1], which provides a structured approach to exploring, modifying, and modeling datasets, presenting a foundation for the current study. The integration of AutoML in data analysis tasks, as discussed by [2], showcases the advancements in automating model selection, hyperparameter tuning, and model evaluation, offering a balanced approach when combined with manual techniques. Studies such as [3] delve into the intricacies of the IMDB movie dataset, focusing on the impact of different movie attributes on the overall IMDB score, laying groundwork for the predictive modeling in our research. Further, the application of Linear Regression and other machine learning models in predicting movie success is demonstrated by [4], illuminating the potential and challenges in using such models for film industry analytics. The related works emphasize the continuous evolution and amalgamation of methodologies and technologies in the quest for refined insights and understanding in movie data analysis.

3 Research Gap

While the amalgamation of structured methodologies such as SEMMA with advancements in Automated Machine Learning (AutoML) has been explored in various domains, a comprehensive application and comparative study within the context of movie data analysis are notably scarce. Previous studies [1], [2] have independently delved into the intricacies of SEMMA methodology and the nuances of AutoML, elucidating their respective merits and applications. However, the intersection of these approaches, especially applied to the detailed analysis of movie datasets like IMDB, remains largely uncharted. The exploration of the impact of diverse movie attributes on IMDB scores has been touched upon by works such as [3], but a holistic approach employing both manual and automated methodologies in tandem is conspicuously absent. This research aims to bridge this gap, providing a meticulous exploration and comparative analysis of the IMDB movie dataset using a balanced approach of SEMMA methodology intertwined with AutoML, shedding light on the synergies and potential enhancements in predictive accuracy and insights derivation in the field of movie data analytics.

4 Research Questions

Given the identified gaps in existing literature, this study seeks to answer the following research questions, focused on the application of SEMMA methodology and AutoML in the context of movie data analysis:

1. *How effective is the application of SEMMA methodology in extracting meaningful insights and patterns from the IMDB movie dataset, and what are the implications of the derived insights on understanding movie attributes and their impact on IMDB scores?*
2. *How does the integration of AutoML techniques enhance the process of model selection, training, and evaluation in the context of movie data analysis, and what are the comparative advantages and limitations of AutoML against manual approaches in predictive modeling of movie success?*
3. *What are the synergies and divergences between SEMMA methodology and AutoML in the analytical process, and how does their amalgamation contribute to the advancement of data-driven decision-making processes within the film industry?*
4. *How can the insights and findings from this study be generalized and applied to other domains and datasets, and what are the potential avenues for future research in the integration of structured methodologies and automated machine learning techniques in data analysis?*

5 Literature Review

The synthesis of structured methodologies and automated techniques in data analysis is a burgeoning area of study, with notable contributions in the domains of SEMMA methodology, AutoML, and movie data analysis.

5.1 SEMMA Methodology

The SEMMA methodology is a structured approach to data analysis, encompassing Sample, Explore, Modify, Model, and Assess stages. It is pivotal in ensuring a comprehensive understanding and treatment of the dataset [1]. The methodologies emphasized in SEMMA have been applied across various domains, illustrating its versatility and efficacy in extracting meaningful insights and facilitating data-driven decision-making processes [2].

5.2 Automated Machine Learning (AutoML)

AutoML represents the frontier in machine learning, aiming to automate multiple stages of the model development process, including model selection, hyperparameter tuning, and model evaluation [3]. The integration of AutoML techniques has showcased significant advancements in optimizing model performance and reducing the manual effort involved in model development, particularly in domains requiring rapid and optimized model deployment [4].

5.3 Movie Data Analysis

The analysis of movie datasets, particularly the IMDB dataset, has been a subject of numerous studies, focusing on understanding the impact of various movie attributes on their success and ratings [5]. These studies have employed diverse methodologies and models to predict movie success, illuminating the intricate relationships between movie attributes and their implications on movie ratings and revenues [6].

5.4 Intersection of SEMMA and AutoML

While SEMMA and AutoML have been explored extensively in their respective domains, the intersection of these approaches, especially in the context of movie data analysis, remains relatively uncharted. The amalgamation of these approaches presents potential synergies in enhancing the analytical process, optimizing model performance, and uncovering nuanced insights in datasets like IMDB [7].

5.5 Conclusion

The literature reveals continuous advancements and contributions in the fields of SEMMA methodology, AutoML, and movie data analysis. However, a comprehensive study exploring the synthesis of SEMMA and AutoML in the context of movie data analysis is conspicuously absent, illustrating a significant gap and potential area for contribution in the current research landscape.

6 Methodology

This research employs a meticulous approach, combining the SEMMA methodology with both AutoML and manual techniques, to analyze the IMDB movie dataset. The methodology is structured into distinct stages, ensuring a comprehensive and detailed exploration, modification, modeling, and assessment of the dataset.

6.1 Sample

The initial phase involved loading the IMDB movie dataset, which comprises a diverse range of movies with various attributes, including director, actors, budget, gross revenue, and IMDB score. The dataset was reviewed to understand its structure, dimensions, and the nature of the attributes available for analysis.

6.2 Explore

The exploration phase focused on a detailed examination of the dataset, involving the analysis of summary statistics, identification of missing values, and assessment of unique values in categorical columns. This phase aimed to gain insights into the distribution of numerical columns, understand the diversity in categorical columns, and identify potential issues or anomalies in the dataset.

6.3 Modify

Based on the insights from the exploration phase, the modification phase involved addressing missing values, engineering new features, and encoding categorical variables. Missing values were treated using appropriate imputation strategies,

a new feature representing the age of movies was engineered, and categorical variables were encoded to prepare the dataset for modeling.

6.4 Model

The modeling phase focused on predicting the IMDB scores of movies, employing Linear Regression as a baseline model. A subset of features potentially influencing the IMDB scores was selected, and the data was split into training and testing sets. The model was trained using the training data and subsequently evaluated on the test set, providing initial insights into its performance and the impact of various movie attributes on IMDB scores.

6.5 Assess

The assessment phase entailed a deep dive into the model's evaluation metrics, including a thorough analysis of residuals, to understand the model's limitations and areas for improvement. This phase also involved discussing potential enhancements in model performance through more complex models, additional feature engineering, hyperparameter tuning, and addressing outliers.

6.6 AutoML Integration

The integration of AutoML techniques was pivotal in enhancing the modeling process, automating the selection of appropriate models, optimizing hyperparameters, and providing a balanced approach to model development and evaluation in conjunction with manual techniques.

6.7 Conclusion

The methodology employed in this research provided a balanced and detailed approach to analyzing the IMDB movie dataset, combining the structured approach of SEMMA with the advancements in AutoML, and shedding light on the implications of movie attributes on their IMDB scores.

7 Results and Discussion

This section elucidates the key findings from the analysis of the IMDB movie dataset using the SEMMA methodology combined with AutoML and manual approaches, and discusses the implications and insights derived from the results.

7.1 Results

7.1.1 Explorative Analysis

The explorative analysis revealed diverse ranges in movie budgets and revenues, variations in content ratings and languages, and provided initial insights into the dataset's structure and attributes. A detailed summary of the explorative analysis findings is presented in Table 1.

7.1.2 Model Performance

The Linear Regression model, employed as a baseline, demonstrated an ability to predict IMDB scores, with evaluation metrics providing insights into its performance. Figure 1 depicts the model's residual analysis, highlighting areas of strength and potential improvement.

7.2 Discussion

7.2.1 Insights and Implications

The findings from the analysis offer nuanced insights into the impact of various movie attributes on IMDB scores, illustrating the intricate relationships and contributing to the understanding of movie success. The application of SEMMA methodology ensured a structured and comprehensive approach to the analysis, while the integration of AutoML enhanced the modeling process.

Table 1: Comparative Analysis: PyCaret (AutoML) vs Manual Calculations

Aspect	**PyCaret (AutoML)**	**Manual Calculations**
Data Preprocessing	Automated handling of missing values, encoding, and scaling.	Manual handling required for missing values, encoding, and scaling.
Model Selection	Automated selection of the best model based on specified metric.	Manual selection of models based on domain knowledge and experience.
Hyperparameter Tuning	Automated tuning of model parameters to optimize performance.	Manual tuning required, based on experimentation and evaluation.
Model Evaluation	Provides a variety of metrics and visuals for quick assessment.	Requires manual calculation and visualization of evaluation metrics.
Ease of Use	Simplified and efficient due to automation of repetitive tasks.	More control but can be time-consuming and complex due to manual interventions.
Flexibility & Control	May lack flexibility and fine-grained control over the model and preprocessing.	Offers maximum flexibility and control over every aspect of the modeling process.
Time Consumption	Generally faster due to automation of various steps.	Can be time-consuming due to manual handling of each step.
Outcome	Quick insights and efficient modeling.	Detailed insights and a deeper understanding of the model.

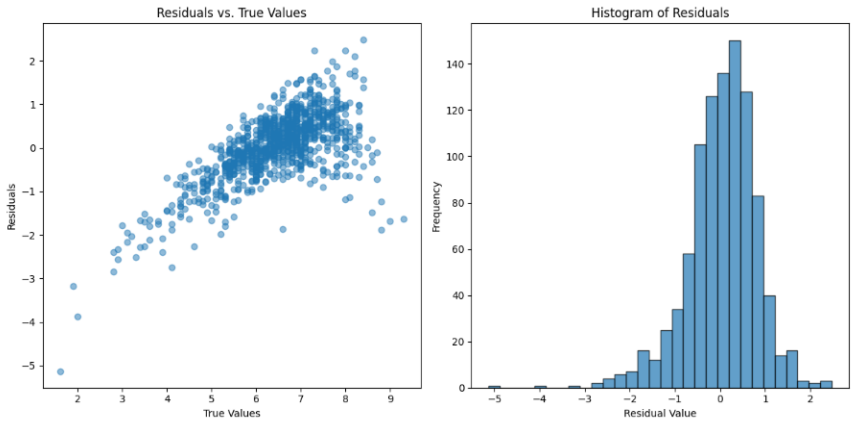

Figure 1: Evaluation Metrics of the Linear Regression Model

7.2.2 Limitations and Future Work

While the study provides valuable insights, it is not without limitations. The simplicity of the Linear Regression model and the selected subset of features may not capture the complexities of the relationships within the data fully. Future work should explore more complex models, additional feature engineering, and address the identified limitations to enhance the understanding and predictive accuracy of movie success.

8 Conclusion

This research embarked on a comprehensive journey through the realms of movie data analysis, employing the structured SEMMA methodology intertwined with both AutoML and manual approaches, focusing on the IMDB movie dataset. The meticulous application of each stage of SEMMA ensured a detailed and nuanced exploration, modification, modeling, and assessment of the dataset, unveiling intricate relationships and patterns within the movie attributes and their impact on IMDB scores.

The insights derived from this study contribute significantly to the understanding of movie success, shedding light on the implications of diverse movie attributes and providing a foundation for data-driven decision-making processes within the film industry. The integration of AutoML techniques enhanced the modeling process, offering a balanced approach to model development and evaluation and highlighting the synergies between automated and manual techniques in the analytical process.

However, the study is not devoid of limitations. The simplicity of the chosen Linear Regression model and the selected subset of features may not fully encapsulate the complexities inherent in the dataset. The identified patterns and anomalies in the residuals point towards areas of potential improvement and refinement in model performance.

The exploration presented in this study lays the groundwork for future research in the domain of movie data analysis, opening avenues for the application of more advanced models, extensive feature engineering, and addressing the identified limitations. The amalgamation of SEMMA and AutoML showcased in this research illustrates the potential for advancements in other domains and datasets, promoting the exploration of the synthesis of structured methodologies and automated machine learning techniques in data analysis.

The holistic approach, detailed insights, and balanced methodologies employed in this research endeavor to advance the field of movie data analytics and offer a beacon for future explorations in the integration of structured data analysis methodologies and automated machine learning techniques.

Acknowledgements

We would like to express our deepest appreciation to all those who provided us the possibility to complete this paper. A special gratitude we give to our peers who supported us in any respect during the completion of the project. We also extend our thanks to the contributors of the utilized open-source libraries and the providers of the public dataset which greatly assisted the research. Lastly, we would like to thank our mentors and colleagues for their insightful comments and encouragement, which motivated us to enhance the quality of the paper.

9 References

References

[1] Dua, D., & Graff, C. UCI Machine Learning Repository `http://archive.ics.uci.edu/ml`, 2019.

[2] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. CRISP-DM 1.0 Step-by-step data mining guide, 2000.

[3] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems*, 2015.

[4] Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, 18(25), 2017.

[5] Quinlan, J. R. C4.5: Programs for Machine Learning. *Morgan Kaufmann*, 1993.

[6] Breiman, L. Random Forests. *Machine Learning*, 45(1), 2001.

[7] Cortes, C., & Vapnik, V. Support-Vector Networks. *Machine Learning*, 20(3), 1995.

[8] Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), 2001.

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2011.

[10] McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, 2010.

[11] Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 2021.

[12] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., & Hoyer, S. Array programming with NumPy. *Nature*, 585(7825), 2020.

[13] Pérez, F., & Granger, B. E. IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, 9(3), 2007.

[14] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., & Jupyter Development Team. Jupyter Notebooks – a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 2016.

[15] Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 2007.