CHAPTER I
INTRODUCTION

This chapter discusses the background, research questions, research objectives, research significance, and research delimitations, which serve as the foundation for developing the movie popularity prediction model in Indonesia. The significant growth of the Indonesian film industry is not always accompanied by certainty regarding a movie's success in attracting audiences; therefore, a data-driven approach is required to understand the factors influencing movie popularity. Consequently, this study proposes the utilization of the gradient boosting algorithm to construct a prediction model for movie admissions based on pre-release metadata.

1.1 Background

The Indonesian film industry has exhibited significant growth in recent years. A surge in movie admissions, the increasing productivity of local production houses, and the expanding diversity of offered genres demonstrate a growing enthusiasm among the Indonesian public toward domestic cinema. This upward trend is further supported by audience behavior studies, which indicate that Indonesian films possess a robust domestic market potential, particularly when production factors are successfully aligned with audience preferences [1]. Nevertheless, not all Indonesian films achieve high movie admissions, indicating that a film's success remains highly volatile and cannot be reliably determined through intuition alone.

This performance uncertainty is evident in the phenomenon where low-budget productions can unexpectedly achieve box office success, whereas highly promoted films do not always secure adequate audience reception. The case study of the film '*Uang Panai'* exemplifies how cultural proximity, local representation, and actor appeal can significantly enhance audience acceptance [2]. This example illustrates that audience decisions are influenced by factors established prior to a film's release, which are directly correlated with movie metadata such as genre, actors, director, and other production attributes. This pre-release information shapes

initial audience perceptions and serves as the baseline for preliminary evaluation before individuals decide to watch a film.

Prior research demonstrates that film attributes, such as genre, actors, and directors, can be leveraged to identify audience preference patterns. Furqon et al. found that combinations of these attributes are associated with the audience's propensity to select specific films [3]. Meanwhile, Memon and Hussain demonstrated that pre-release features can be utilized to predict box office revenue using regression-based models [4]. Another study reinforces that a film's success is influenced by various production features, such as genre, budget, and rating, which can be deployed to predict the level of movie popularity [5]. Therefore, pre-release metadata exhibits an empirical correlation with film performance, indicating that data-driven analysis can facilitate a more objective prediction of movie popularity.

Within the context of the Indonesian film industry, the most widely utilized indicator of popularity is the number of movie admissions. Movie admissions represent the actual consumption by the audience and serve as the foundational metric for assessing a film's success across various industry reports and academic research [6]. Therefore, this study utilizes movie admissions as the primary prediction target.

Technological advancements enable movie metadata to be harvested from online databases such as The Movie Database (TMDb). Metadata including genre, release date, actors, and other attributes are publicly available and widely utilized in research. Various studies have leveraged the TMDb dataset to predict movie ratings, box office revenue, and popularity levels using machine learning approaches [7]. The availability of comprehensive and structured metadata renders it a highly relevant data source for research based on predictive analytics.

Aligning with the improved availability of data, the application of machine learning algorithms in predicting film performance continues to evolve. Numerous studies indicate that gradient boosting-based models are capable of delivering superior performance compared to traditional statistical methods. This is supported by research from Memon and Hussain, which demonstrates that pre-release film features can be utilized to predict a movie's success prior to its theatrical release [4].

Within the domain of modern ensemble learning, gradient boosting frameworks, specifically XGBoost, LightGBM, and CatBoost are widely recognized for their superior predictive performance on tabular datasets, such as movie metadata. A comparative analysis by Kumar and Kumar highlights that these three algorithms offer distinct advantages in terms of optimization accuracy, computational efficiency, and their native capacity to handle categorical features [8]. XGBoost incorporates robust regularization mechanisms to mitigate overfitting, whereas LightGBM achieves high computational efficiency through its leaf-wise tree growth strategy. Complementing these approaches, CatBoost is engineered specifically to process categorical features effectively, eliminating the need for complex pre-processing pipelines [8], [9].

Furthermore, contemporary research highlights the critical role of hyperparameter tuning in optimizing the performance of boosting models. Tang demonstrated that parameter optimization within XGBoost can yield significant improvements in both the stability and accuracy of box office predictions [10]. These findings align with Wu et al., who reported that machine learning-based models, including XGBoost, are capable of generating higher predictive accuracy compared to conventional regression models [11]. However, systematic hyperparameter tuning remains rarely applied to movie admission predictions within the Indonesian cinema context, thereby presenting a critical research opportunity.

Driven by these gaps, this study moves beyond a single-algorithm approach to construct and compare three distinct predictive models: the XGBoost Regressor, CatBoost Regressor, and LightGBM Regressor. These models are trained exclusively on pre-release metadata such as genre, director, writer, producer, and lead cast aligning with the methodologies recommended in feature-based pre-release literature [4]. Previous studies have shown that machine learning approaches can improve the accuracy of movie performance predictions, and that using pre-release metadata is effective for predicting a film's success before its release. However, most of these studies still focus on predicting box office revenue in a global context, and they have not yet fully explored the comparative use of modern gradient boosting models with structured hyperparameter optimization.

Additionally, research that specifically uses movie admissions as the popularity indicator within the Indonesian film industry is still relatively limited.

These limitations highlight the need to develop a predictive approach that is not only accurate but also relevant to the characteristics of the Indonesian film industry. Therefore, this study proposes a comparative approach by implementing three modern gradient boosting algorithms and performing hyperparameter optimization on each model to achieve optimal performance. Using movie admissions as the prediction target is expected to provide a more concrete representation of audience acceptance, so that the results of this study can offer practical benefits in supporting production and distribution decisions. Additionally, this research analyzes the contribution of each pre-release metadata feature to the predictions, aiming to improve model interpretability and understand the factors that most influence movie admissions.

Therefore, this study is based on the hypothesis that utilizing pre-release metadata with an optimized gradient boosting machine learning approach can improve the prediction accuracy of movie admissions compared to conventional approaches. Additionally, it is hypothesized that there are performance differences among the XGBoost, LightGBM, and CatBoost algorithms in modeling the relationship between movie metadata and admissions, meaning that a comparison of these three models can identify the most effective one for the Indonesian film industry.

1.2 Research Question

In a study, the problem formulation is necessary to provide clear boundaries regarding the aspects to be examined and to guide the development of the proposed solution. The research questions are formulated based on the initial phenomena identified in the Indonesian film industry and the need to predict movie performance prior to its release. Based on this background, this study focuses on the following research questions :

1. How can predictive models for movie admissions in Indonesia be developed using the XGBoost, CatBoost, and LightGBM algorithms?

2. Which algorithm among XGBoost, CatBoost, and LightGBM delivers the best performance in predicting movie admissions in Indonesia?
3. Which pre-release metadata features have the greatest influence on movie admission predictions based on the best-performing model?

1.3 Research Objectives

The research objectives are formulated to address the previously stated research questions. Each objective describes the outcomes to be achieved through the series of analysis, modeling, and evaluation processes conducted in this study. In general, the objectives of this study are as follows :

1. To develop and compare the performance of predictive models based on XGBoost, CatBoost, and LightGBM in predicting movie admissions in Indonesia.
2. To determine the best-performing model based on the evaluation of regression metrics.
3. To identify the metadata features that have the greatest influence on the predictions through feature importance and SHAP analysis on the best-performing model.

1.4 Research Significance

This study is expected to contribute to both academic and industrial practice. The significance of this research is outlined to demonstrate the utility of the obtained results, both for the advancement of science and for stakeholders in the film industry. The benefits of this study cover two main aspects: academic significance and practical significance, as follows :

1.4.1 Academic Significance

Academically, this study helps expand research on predictive modeling in the film domain using machine learning approaches. The contribution of this research is also expected to enrich the literature on the use of modern boosting algorithms, particularly in the context of predictions based on pre-release metadata. The expected academic benefits include :

1. Contributing to the development of research on movie predictions based on pre-release metadata.
2. Contributing to the scientific literature on the application of modern boosting algorithms in predicting movie popularity.

1.4.2 Practical Significance

Beyond its theoretical contributions, this study offers practical value that can be utilized by film industry professionals. The movie admission predictions generated by the models are expected to assist stakeholders in making strategic decisions during both the production and distribution phases. The specific practical benefits include :

1. Helping producers, distributors, and investors predict potential movie admissions prior to release.
2. Serving as a basis for determining marketing strategies, budgeting, and movie release scheduling.

1.5 Research Limitations

To ensure the study remains focused and well-directed, research scope boundaries are necessary to define the limits of the object, data, and methods used. These boundaries are also important to avoid expanding the scope into areas irrelevant to the primary objective. The scope of this study is bounded as follows:

1. The research objects are limited to movies released in Indonesia between 2007 and 2025, based on data available on the Cinepoint website.
2. The selected movies must have a duration of more than 60 minutes and exclude short films.
3. The pre-release metadata used is limited to genre, director, writer, producer, and lead actors.
4. The indicator of movie popularity used is the number of movie admissions.
5. The algorithms used are strictly limited to XGBoost, CatBoost, and LightGBM, without comparison to any models outside of these three algorithms.