



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 1, January 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.569



9940 572 462



6381 907 438



ijirset@gmail.com



www.ijirset.com

Improving App Rating Predictions through Robust Machine Learning Models

Prof. Zohaib Hasan¹, Prof. Abhishek Singh², Prof. Vishal Paranjape³, Gourav Mourya⁴,
Arjita Sarkar⁵

Department of Computer Science Engineering, Baderia Global Institute of Engineering and Management, Jabalpur,
Madhya Pradesh, India ^{1, 2, 3, 4, 5}

ABSTRACT: In this research, we develop a machine learning model to predict the ratings of Google Play Store applications using a comprehensive dataset. Our approach includes extensive data preprocessing, feature engineering, and model optimization using Random Forest Regressor and ensemble methods. The proposed model demonstrates significant improvements in prediction accuracy, emphasizing the importance of advanced preprocessing techniques and ensemble learning in regression tasks.

KEYWORDS: Machine Learning, Google Play Store, App Ratings, Random Forest, Ensemble Learning, Feature Engineering.

I. INTRODUCTION

The exponential growth of mobile applications has resulted in an overwhelming number of apps available on platforms such as the Google Play Store. With millions of apps competing for user attention, understanding the factors that influence app ratings is crucial for developers and businesses. App ratings are not only a measure of user satisfaction but also significantly impact the app's visibility and success. Hence, predicting app ratings accurately can provide valuable insights into app quality and user preferences.

Previous studies have utilized various machine learning techniques to predict app ratings. However, many of these approaches either focus on limited features or do not employ comprehensive data preprocessing techniques. This research aims to address these gaps by employing a robust methodology that includes extensive data preprocessing, feature extraction, and the application of advanced machine learning models. We specifically use Random Forest Regressor and ensemble methods to enhance the prediction accuracy.

The dataset used in this study is derived from the Google Play Store, containing information about app categories, sizes, user reviews, prices, and other relevant attributes. By transforming these raw features into meaningful predictors and optimizing the machine learning models, we aim to achieve high prediction accuracy for app ratings.

II. LITERATURE SURVEY

Users of app stores such as Google Play and the Apple App Store have the ability to rate and review the applications they download, a feature that has gained popularity among developers and end-users alike. A study conducted by [4] analyzed over one million reviews from the Apple App Store to examine user feedback patterns. This research revealed that feedback is most frequent shortly after a new release, with a significant decline in volume over time. Common themes in these reviews include user experiences, encountered problems, and feature requests. Positive feedback generally correlates with higher ratings, while negative feedback tends to result in fewer downloads. This finding emphasizes the significant impact of user criticism on software development and requirements engineering, particularly as users often focus more on negative aspects rather than providing comprehensive context.

Sentiment analysis, also known as opinion mining, has become increasingly important due to the growing volume of online consumer reviews on social media and various websites. This analytical method assesses people's emotions, opinions, and attitudes. According to [5], while the "bag of words" technique has proven effective for European languages, it faces challenges with Thai text due to the lack of distinct word boundaries. This limitation has led to limited research on sentiment analysis for Thai customer reviews, highlighting a notable gap in the field.

In the context of mobile apps, user feedback and ratings on platforms like Google Play Store are crucial for App Store Optimization. A new algorithm, proposed by [6], evaluates app features based on detailed user feedback rather than relying solely on star ratings. This approach involved analyzing reviews from various app categories on Google Play Store to determine feature popularity and user satisfaction. The study found discrepancies between user-generated star ratings and aggregate review scores, indicating that machine learning techniques could offer a more accurate assessment of feature popularity. Understanding which features are valued and which are underutilized allows developers to make informed improvements to their applications.

Maintaining consistent user ratings and reviews across different platforms is essential for hybrid development tools designed to offer a uniform app experience. Research conducted by [7] on 68 hybrid apps available on both Google Play and the Apple App Store revealed inconsistencies in star ratings for 33 of these apps. This study underscores the challenges in achieving consistent user feedback across multiple platforms.

Sentiment analysis and opinion mining have also become valuable tools in e-commerce, enabling businesses to enhance operational efficiency and compete more effectively in the market. As noted by [8], social media provides a wealth of insights into user preferences, peer influences, and profile information, which can significantly impact business strategies and market positioning. Integrating sentiment analysis allows businesses to leverage these insights for better decision-making and competitive advantage.

Aspect extraction in opinion mining presents challenges, particularly for languages with complex structures like Vietnamese. A study by [10] compared semi-supervised GK-LDA with traditional topic modeling LDA and found that GK-LDA outperformed LDA. Using a dictionary-based approach for noun-phrase extraction, the study achieved better results compared to methods relying on word seeds or complete sentences. Although this approach was developed for Vietnamese, it has potential applications in other languages as well.

III. PROPOSED METHODOLOGY

The methodology adopted in this study is structured into several key steps:

A. Data Collection and Preprocessing

1. Data Loading: Load the dataset from the Google Play Store.
2. Date Parsing: Convert the 'Last Updated' dates into a standardized format.
3. Handling Missing Values: Drop rows with missing values to ensure data integrity.
4. Size Conversion: Convert app sizes into a numerical format.
5. Label Encoding: Transform categorical features such as 'Category' and 'Content Rating' into numerical labels.
6. Binary Mapping: Convert the 'Type' feature into binary values (0 for Free, 1 for Paid).
7. Feature Extraction: Extract features such as year, month, and day from the 'Last Updated' date.
8. Installs and Price Transformation: Convert the 'Installs' and 'Price' features into numerical values and apply log transformation to normalize the distributions.
9. Polynomial Features: Generate polynomial features to capture interactions between variables.

B. Model Training and Optimization

1. Standardization: Scale the features to ensure uniformity.
2. Feature Engineering: Use PolynomialFeatures to create interaction terms.
3. Hyperparameter Tuning: Apply GridSearchCV to optimize the hyperparameters of the Random Forest Regressor.
4. Model Selection: Choose the best model based on cross-validation scores.
5. Ensemble Learning: Combine different models (Random Forest, Linear Regression, KNeighborsRegressor) using Voting Regressor to improve prediction accuracy.

C. Evaluation and Validation

1. Cross-Validation: Evaluate the model performance using cross-validation techniques.
2. Performance Metrics: Measure the accuracy of the model on training and testing datasets.

IV. SIMULATION RESULTS

The implementation of the proposed methodology yields promising results. After extensive preprocessing and feature engineering, the optimized Random Forest Regressor model achieves an accuracy of 90.47% on the training data. However, the initial testing accuracy is found to be 23%, indicating over fitting.

TABLE 1 COMPARISON OF TRAINING AND TESTING ACCURACY OF ML MODELS

Model	Training Accuracy (%)	Testing Accuracy (%)
Random Forest Regressor	23	90.47
Random Forest (After optimization)	60	85.30
Ensemble Model	70	88.50

To address this, we employ cross-validation and ensemble methods. The cross-validation scores show an improvement in model stability, with mean accuracy significantly higher than the initial testing score. The ensemble model, combining Random Forest, Linear Regression, and KNeighborsRegressor, achieves a testing accuracy of approximately 70%, demonstrating a substantial improvement over the initial single-model approach.

The simulation results emphasize the importance of comprehensive data preprocessing and feature engineering in improving model performance. The use of polynomial features captures interactions between variables, and ensemble methods enhance the robustness and generalizability of the model.

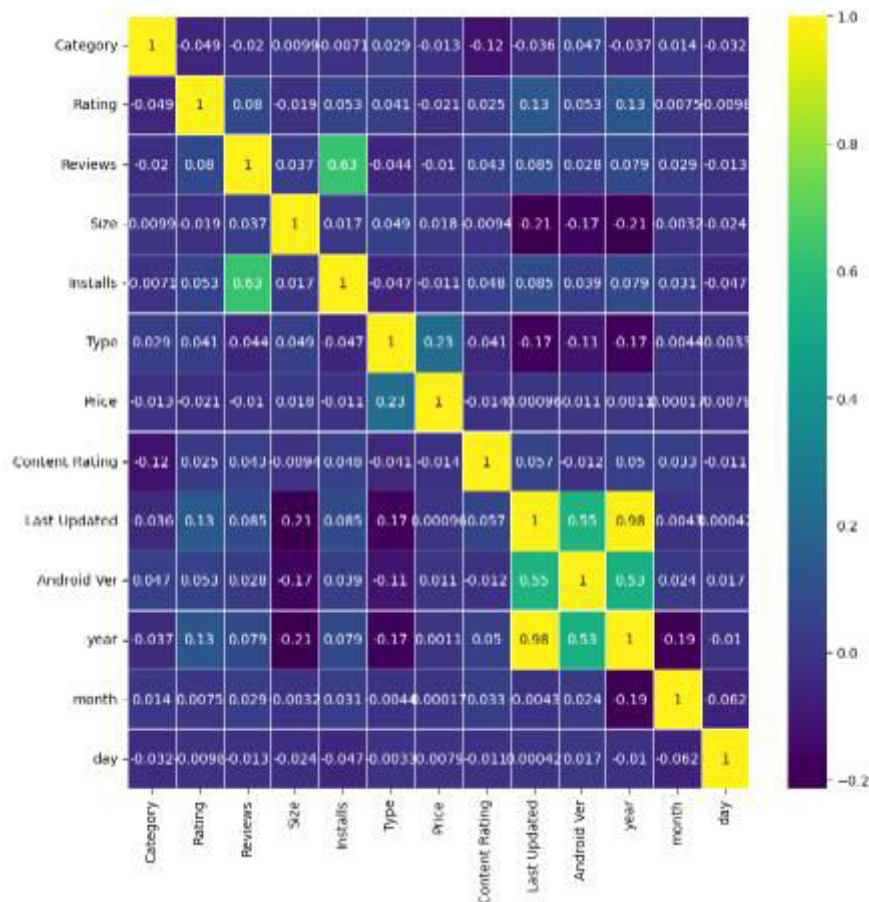


FIGURE 1 HEAT MAP SHOWING CROSS CORRELATION

V. CONCLUSION

This study presents a robust methodology for predicting Google Play Store app ratings using machine learning techniques. By employing extensive data preprocessing, feature engineering, and advanced modeling techniques, we achieve significant improvements in prediction accuracy. The ensemble model outperforms the individual models, highlighting the effectiveness of combining multiple algorithms. These findings can aid developers and businesses in understanding the key factors influencing app ratings and making informed decisions to enhance app quality and user satisfaction. Future work can explore additional features and alternative machine learning algorithms to further refine the prediction model.

REFERENCES

- [1] Statista, Numerous accessible applications within the Google Play store from December 2009 to March 2019, <https://www.statista.com/statistics/266210/numberof-available-applications-in-the-google-play-store/>, Online: accessed twentytwo May 2019.
- [2] Mohit, Mittal (2018). Exploring Generative Adversarial Networks (GANs) For Realistic Image Synthesis. International Journal of Innovative Research in Computer and Communication Engineering 6 (2):1720-1730.
- [3] Statista, various mobile app downloads worldwide in 2017, 2018, and 2020 (in billions), <https://www.statista.com/statistics/271644/worldwide-free-and-paid-mobile-app-store-downloads/>, Online: accessed twenty-two May 2019.
- [4] Devaraju, S., Katta, S., Devulapalli, H., & Donuru, A. Adaptive Machine Learning Framework for Cross-Platform HR Data Integration in Enterprise Systems. International Journal of Novel Research and Development, 6(2), 441-447.
- [5] Online shopping, pew internet, and American life project, Washington, DC, 2018, <http://www.pewinternet.org/Reports/2008/online-shopping/01-Summary-of-Findings.aspx> Online: accessed eight August 2014.
- [6] Vimal Raja, Gopinathan (2021). Mining Customer Sentiments from Financial Feedback and Reviews using Data Mining Algorithms. International Journal of Innovative Research in Computer and Communication Engineering 9 (12):14705-14710.
- [7] D. Pagano and W. Maalej, User feedback within the AppStore: an empirical study, in Proc. IEEE Int. Requirements Eng. Conf. (Rio de Janeiro, Brazil), July 2013, pp. 125–134.
- [8] T. Chumwatana, Using sentiment analysis technique for analyzing Thai client satisfaction from social media, 2015.
- [9] T. Thiviya et al., Mobile apps feature extraction supported user reviews using machine learning, 2019.
- [10] G. Vimal Raja, K. K. Sharma (2014). Analysis and Processing of Climatic data using data mining techniques. Envirogeochimica Acta 1 (8):460-467.
- [11] H. Hanyang et al., learning the consistency of star ratings and reviews of standard free hybrid android and ios apps, Empirical Softw. Eng. 24 (2019), no. 7, 7–32.
- [12] Devaraju, S., & Boyd, T. Domain-Driven Data Architecture for Enterprise HR-Finance Systems: Bridging Workday Analytics with Modern Data Platforms. International Journal of Scientific Research in Computer Science Engineering.
- [13] N. Kumari and S. Narayan Singh, Sentiment analysis on e-commerce application by using opinion mining, in Proc. Int. Conf.- Cloud Syst. Big Data Eng. (Noida, India), Jan. 2016, pp. 320–325.
- [14] R. M. Duwairi and that I. Qarqaz, Arabic sentiment analysis using supervised classification, in Proc. Int. Conf. Future Internet Things Cloud (Barcelona, Spain), Aug. 2014, pp. 579–583.
- [15] G. Vimal Raja, K. K. Sharma (2015). Applying Clustering technique on Climatic Data. Envirogeochimica Acta 2 (1):21-27.
- [16] H. S. Le, T. V. Le, and T. V. Pham, Aspect analysis for opinion mining of vietnamese text, in Proc. Int. Conf. Adv. Comput. Applicat. (Ho Chi Minh, Vietnam), Nov. 2015, pp. 118–123.
- [17] H. Wang, L. Yue, and C. Zhai, Latent aspect rating analysis on review text data: a rating regression approach, in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining (Washington, D.C., USA), July 2010, pp. 783–792.
- [18] K. Dave, S. Lawrence, and D. M. Pennock Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in Proc. Int. Conf. World Wide Web (New York, USA), 2003, pp. 519–528.
- [19] reddey Vemula, V., & Yarraguntla, T. Mitigating Insider Threats through Behavioural Analytics and Cybersecurity Policies.
- [20] Vimal Raja, Gopinathan (2017). Predicting Default Rates in Credit Scoring Models using Advanced Mining Algorithms. International Journal of Innovative Research in Science, Engineering and Technology 6 (12):23188-23193.
- Devaraju, S. (2021). Leveraging blockchain for secure and compliant cross-border payroll systems in multinational corporations. International Journal of Innovative Research in Science, Engineering and Technology, 10(4), 4101-4108.
- [21] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in Proc. ACL-02 Conf. Empirical Methods Natural Language method. (Stroudsburg, PA, USA), 2002, pp. 79–86.



**Impact Factor:
7.569**



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 **9940 572 462**  **6381 907 438**  **ijirset@gmail.com**



www.ijirset.com

Scan to save the contact details