Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/59/20231091

Download Behavior Analysis Based on Google Play Store Data

Jinzi Zheng[1,a,*]

[1]*College of Mathematics, Shandong University of Science and Technology, Qingdao, China*
a. zhengjinzi2002@ldy.edu.rs
**corresponding author*

Abstract: As mobile apps continue to grow, app stores, as the main channel for users to download apps, are becoming increasingly important for developers and platforms. Understanding users 'download behavior and accurately predicting users' preferences and needs can effectively improve the effect of the recommendation system in the application mall, improve user experience and improve the download conversion rate. However, traditional rule-based recommender systems often face the problems of data sparsity and model complexity. Therefore, it is an urgent and valuable topic to analyze user download behavior combined with machine learning technology and to provide personalized recommendations and services. This study uses the download behavior information of Google Play Store users in 2018, and use three classic machine algorithms (linear regression, random forest and SVM) to model and predict the software rating, dig deep into various factors affecting the rating, and gain deep insight into users' preferences and behavior patterns. This will provide more accurate recommendation results for the application mall, improve the application quality and popularity, and improve the user satisfaction and loyalty, and provide an important reference for optimizing the recommendation system and personalized service of the application mall.

Keywords: Google Play Store, behavioral analysis, random forest, linear regression, SVM.

1. Introduction

According to research network in 2022 research report, the application store as application distribution platform, the application download channels according to the user source ranked first, the highest proportion, and for developers, app store users with high natural trust (64.1%), users download demand precision, high conversion rate (55.3%), is an important user source channels [1]. Most developers want to have good analysis templates or tools to help them retain customers and grab the market. There are many applications that require in-depth research, with iOS consumers offering 22 ratings a day on one device, while Facebook can receive more than 4,000 feedback per day [2]. Google launched its own app store, Android Market, in October 2008. Android Market Is based on the open source code operating system Android, it is not apple harsh audit system, publishing applications in Google app store is very simple, is positioned as a open content distribution system, it can help use the Android operating system mobile end users to find, buy, download and install all kinds of content [3]. This made Google Play Store quickly hot, with millions of users registering through Google to get personal information and download and use the apps on their own electronic devices. Thousands of developers have uploaded their works to the Google Play store, and the software has quickly reached a million downloads [4]. Incentive installation services can manipulate app store metrics, including installed numbers, and while the Android market does not encourage this, the findings suggest that these manually entered tags can effectively improve the app store data and help mobile app developers attract money from venture capitalists [5].

McIlroy et al. looked at the frequency of updates of apps in the Google Play store and found that nearly half of these frequently updated apps did not publish information about the reasons and principles of the updates. Research shows that developers frequently update software is not be refused by users, but the frequency depends on the category of the store. Developers don't need to care too much about the details of the new update, because users sometimes don't pay attention to the information [6]. Yang's team studied the distribution model through online surveys and analyzing 500,000 user reviews over three years (2016-2019). They found that apps that accompanied longer release notes tended to get higher ratings from users. In addition, improving the speed and frequency of software updates can also get higher ratings [7]. Noei et al. studied mall reviews and found that men submitted more reviews than women that could infer gender, and that developers responded more to men than women. This results in women's needs that cannot be addressed by developers, so developers should consider gender in responding to comments to help mitigate bias from the feedback loops [8]. With the development of Google App Mall, related analytics software has also emerged. Provider By.U is the first digital provider in Indonesia, extracting provider-related information by performing sentiment analysis. One of these features is available through the By.U app comment feature on Google Play Store. After processing the data, it then uses a SVM and a TF-IDF to extract the feature points. The TF-IDF + SVM using the 5-fold validation had a fairly high accuracy, with an average accuracy of nearly 85% [9]. Finally, the relationship between technical and social functions and the prevalence was apps by Businge et al. Found that both technical and social factors play an important role in explaining the popularity of applications [10].

The rest part of the paper is organized as follows. The Sec. 2 first introduces the basic knowledge of machine learning, behind the paper adopts three classical machine learning algorithm (linear regression, SVM, decision tree and random forest), this paper introduces the basic principles of several classification learning algorithm, for the user in the application mall download behavior analysis and Rating prediction provide theoretical support. The Sec. 3 performed multiple regression analysis for each model, considering whether to retain dummy variables and whether to include categorical variables. Where, the regression not containing the type part is used only for incorrect comparisons. Then, one took the MSE as the index to compare the actual results with the results of the model predictions, so as to evaluate and verify the accuracy of the model. Finally, this study will explore the variables that affect the results and analyze their proportion. The Sec. 4 mainly talks about the shortcomings and limitations of the model, and puts forward some ideas and prospects for this. The Sec. 5 is the conclusions and prospects. First, this paper will make a simple summary before and after the user behavior analysis test, and then analyze the shortcomings in the experimental process, and put forward the user behavior analysis in the process in the future.

2. Data and Method

The Google Play Store information is scraped from the Kaggle. Each application contains information about category, rating, number of reviews, software size, etc. Play Store Application data can help developers understand user needs and behavior patterns, reveal the rules of market competition, find problems and challenges, and provide reference for marketing and advertising. The raw data section is shown in Table. 1.

Table 1: Raw data selection.

App	Rating	Reviews	Size	Installs	Content Rating	Genres	Current Ver	Android Ver
Photo Editor & Candy Camera & Grid & Scrap Book	4.1	159	19M	10,000+	Everyone	Art & Design	1.0.0	4.0.3 and up
Coloring book moana	3.9	967	14M	500,000 +	Everyone	Art& Design; Pretend Play	2.0.0	4.0.3 and up
U Launcher Lite-FREE Live Cool Themes, Hide...	4.7	87510	8.7M	5,000,00 0+	Everyone	Art & Design	1.2.4	4.0.3 and up
Sketch -Draw & Paint	4.5	215644	25M	50,000,0 00+	Teen	Art & Design	Varies with device	4.2 and up
Pixel Draw - Number Art Coloring Book	4.3	967	2.8M	100,000 +	Everyone	Art& Design;Cre ativity	1.1	4.4 and up

Random forest regression is an ensemble learning algorithm that uses a forest composed of multiple decision trees for regression prediction, and in the process of constructing each decision tree, for the segmentation of each node, the random forest selects only a subset of all features for consideration. This increases the diversity between decision trees and improves the generalization ability of the entire random forest. When a prediction needs to be made for a new sample, the sample is fed into each decision tree in the random forest to get the prediction result for each tree. For regression problems, the average or weighted average of all decision tree predictions is calculated as the final prediction for the entire random forest. The algorithm is suitable for regression problems and performs well when dealing with large amounts of feature and noise data. When it comes to random forest regression models, one has to talk about decision trees, which are nonparametric supervised learning methods. Essentially from the training data to summarize a set of decision rules, used to solve classification and regression problems, the rules are: root nodes, internal nodes, leaf nodes (labels) composed of a tree map to present, unlike the classification tree, the prediction of the regression tree is based on the average of the case reaching the leaf node as the output of the predicted value. This study plot sDecision Tree For Max depth =3.

Random forest can be said to be the plus version of Bagging algorithm, which is improved on the basis of Bagging algorithm. Bagging The calculation takes a random sampling method, i.e., one will collect the duplicate samples. In the Bagging algorithm, the same number of samples is the same as the M number of the training set. Therefore, the number of samples and the number of training sets are the same, but their content is different. If one does a random sampling of T times for the training set with m samples, the T sampling sets will vary due to the randomness.bagging, the process of the algorithm is very simple and is shown as follows for t=1,2..., T:

- The training set was randomly sampled the t time for m times to obtain the sampling set containing m Dt samples
- Train the t-th weak learner Gt (x) with the sampling set Dt

If the classification algorithm predicts, the category or category that T weak learners cast the most votes is the final category. For the regression problem, usually using the simple averaging method, one can arithmetic average the regression results obtained by T weak learners and then output the final model.

Linear regression is a common regression analysis method used to model the linear relationship between the input feature x and the output target y. Linear regression assumes a linear relationship between x, y, and finding the best parameters minimizes the error between the model's predicted value and the true value. It is simple and easy to interpret and is suitable for problems that predict continuous outputs. By learning the sample mapping relationship f (x) =y, the resulting prediction result y is a continuous value variable. The general expression of the prediction function of the multiple linear regression is:

$$\mathrm{f}(\mathrm{x_i}) = \theta_0 + \theta_1 \mathrm{x}_1 + \theta_2 \mathrm{x}_2 + \cdots + \theta_\mathrm{m} \mathrm{x}_\mathrm{m} \tag{1}$$

If $\mathrm{x}_0 = 1$, one adds into the Eq. (1) to obtain:

$$\mathrm{f}(\mathrm{x_i}) = \theta_0 \mathrm{x}_0 + \theta_1 \mathrm{x}_1 + \theta_2 \mathrm{x}_2 + \cdots + \theta_\mathrm{m} \mathrm{x}_\mathrm{m} = \sum_{\mathrm{i}=0}^{\mathrm{m}} \theta_\mathrm{i} \mathrm{x}_\mathrm{i} \tag{2}$$

changing Eq. (2) into a vector expression to:

$$\mathrm{f}(\mathrm{x}) = \theta^\mathrm{T} \mathrm{x} + \mathrm{b} \tag{3}$$

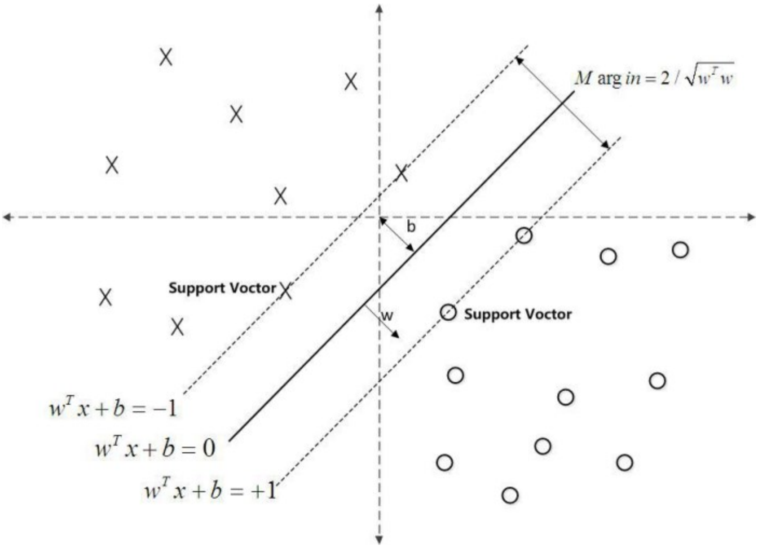

Figure 1: A sktech of the SVM.

Table 2: Data rename.

Rating	Reviews	Size	Installs	Content Rating	Category	Genres
4.1	159	19000000.0	10000	0	0	0
3.9	967	14000000.0	500000	0	0	1
4.7	87510	8700000.0	5000000	0	0	0
4.5	215644	25000000.0	50000000	1	0	0
4.3	967	2800000.0	100000	0	0	2

SVM are mostly used for binary classification and multiclassification problems, and the goal of SVM is to find an optimal hyperplane that separates as many different classes of samples as possible. In the search for optimal hyperplanes, SVM find the closest points to these planes, which are called support vectors. Support vectors are the key points that determine the position of hyperplanes, and they are located on the boundaries of different categories. During the solution process, SVM finally transforms the problem into a convex optimization problem. By introducing the Lagrange multiplier, the original problem is transformed into a dual problem. By solving the dual problem, the optimal separation hyperplane and the corresponding decision function can be obtained. For new sample points, their class is determined by calculating their value in the decision function. If the value of a decision function is greater than a certain threshold, it is classified as one class, otherwise it is another class. As shown in the Fig. 1, the separation hyperplane is $w^T x + b=0$.It can be shown that there is only one such hyperplane. The vectors parallel to the hyperplane with a certain functional distance of the two hyperplanes, which one defines as the support vector, as shown in the dashed line of the Fig. 1. As for the data, Rgus first created this dataset, which has integer encoding of categorical variables, defined as D1, and then, this study created another dataset that created virtual values specifically for each category instance in the dataset, defined as D2, and the part of D2 is as given in Table. 2, where one sets the rating of the continuous variable Rating to Y, while the parameter x contains the comment, size, installation, type, price, content, type, and category.

3. Results and Discussion

3.1. Correlation Analysis

There are 33 categories. Through the visual analysis of the data, one found the most types of software in Google Mall are games and family types. When it comes to rating, one found that software ratings in the app store are generally high, and even half of the apps have a score of more than 4.5, indicating that the software in the Google Play Store is of high quality. Among them, the categories of health, fitness, and books received the highest ratings. Correspondingly, dating apps perform poorly, most of them are lower than the average score of the mall, because such software is more likely to contain scam information, so it is undoubtedly a big challenge for developers to make a good dating software. In addition, some low-rated software has appeared in the Home & Finance category (seen from Fig. 2).

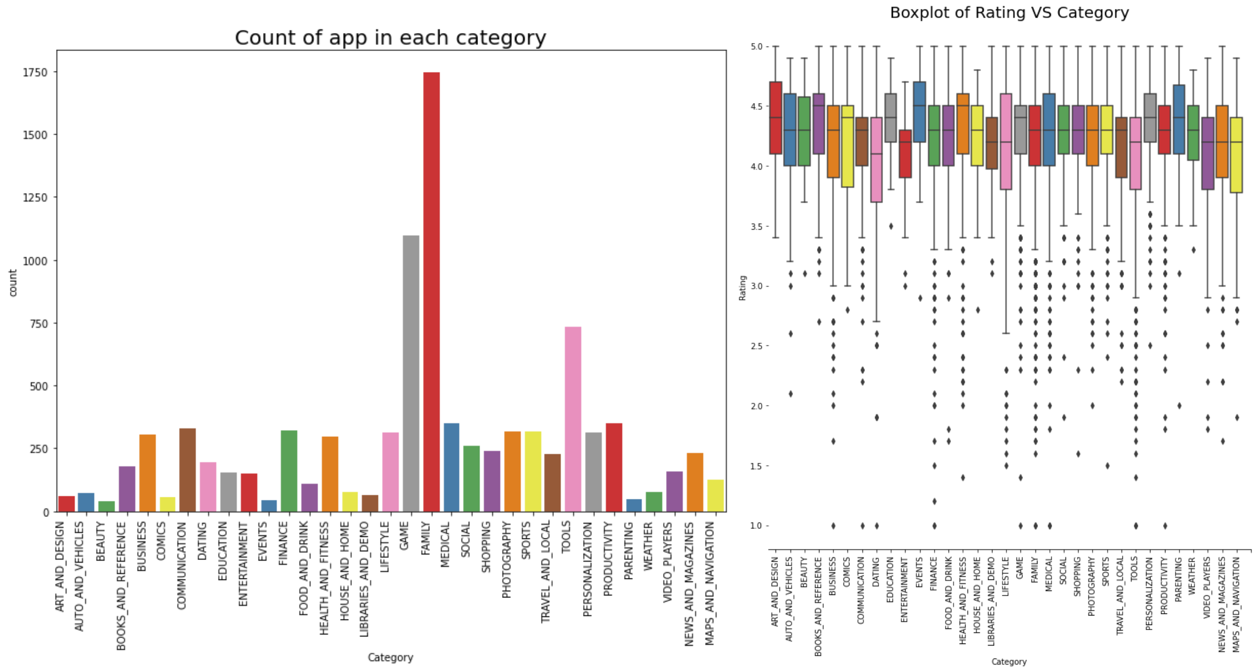

Figure 2: Counts analysis.

Then compare Downloads of Paid and Free apps. As can be seen from the Fig. 3, the majority of customers prefer to download free apps. Beside some apps cost around $400, the majority of the apps cost less than $100. So, one chooses a data-only paid app that costs less than $250. Then, one found that the app categories "Game", "Medical", "Family", and "Tools" have the most ratings. The costliest app category is "Medical", with a maximum price of $80. The majority of the apps cost within $20. Then, one filtered out the apps that needed to be paid for and made a scatter plot of rating and size. Looking at the scatter plot in Fig. 4, it is not difficult to find that among these money-spending software, the high ratings are software that occupies relatively little memory, because most of the good paid software is refined in a certain function, such as the more popular "Football Manager", which is a simulated management mobile game with exquisite football themes, with a score of 4.6, designed for football lovers. In the game, players will become a professional football manager, can build their own football club, there are many players to recruit, can train, improve the level of their team, there are a large number of matches to participate in, experience real football matches. Therefore, delicate and small software is easier to get the public's love, while paid programs with large memory are not favorable. Now, one has carried out correlation tests between variables.

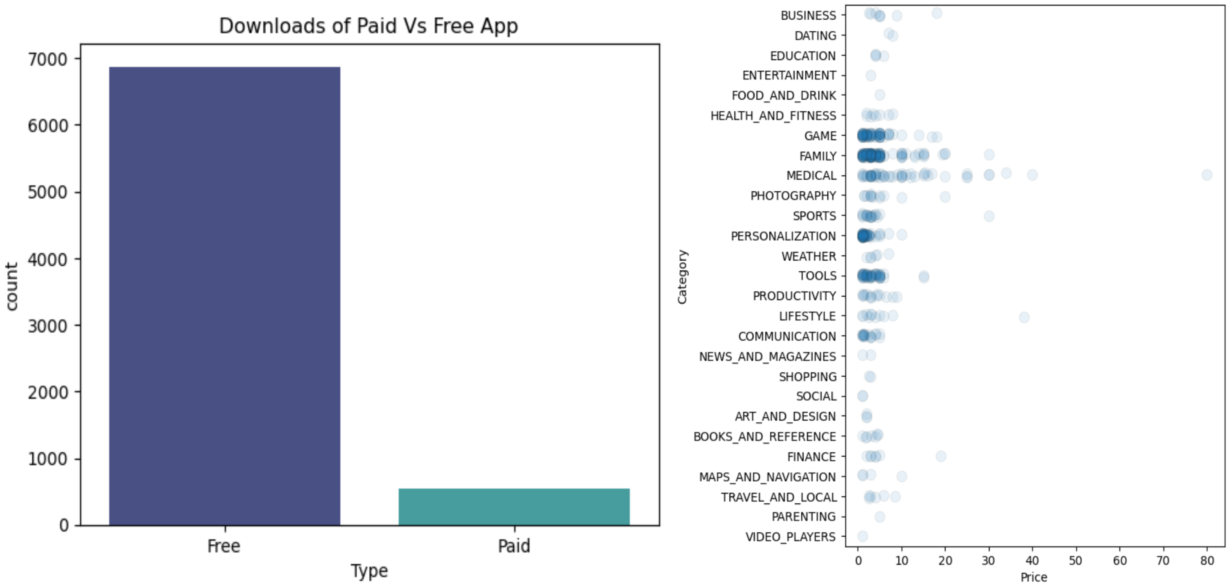

Figure 3: Download of paid and price distribution.

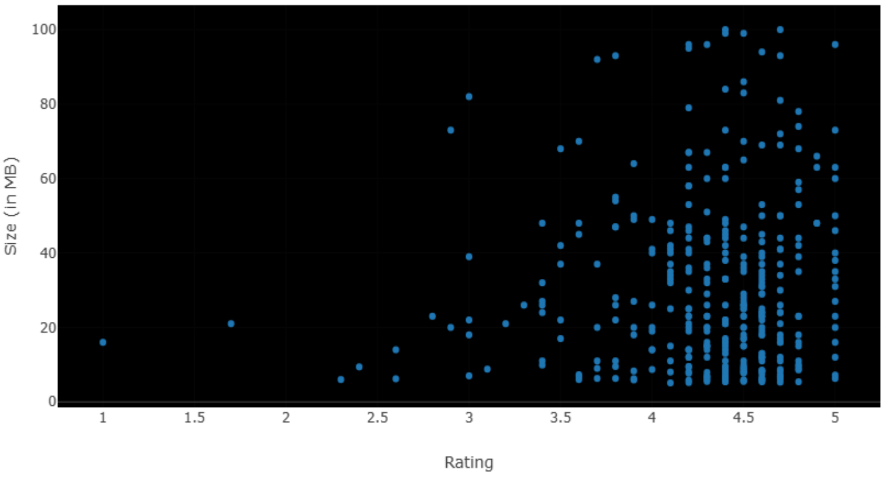

Figure 4: Size as a function of rating.

The correlation coefficient between the number of reviews and the number of downloads reached 0.626, showing an obvious positive correlation, which shows that people are more willing to download software with a large number of reviews and are willing to tend to the public's choice. At the same time, it can also be seen that there is no shortage of positive reviews in the app store, and many are willing to give feedback after downloading (seen from Table 3).

Table 3: Correlation analysis.

	Rating	Reviews	Price	Content	Size	Installs
Rating	1	0.08	-0.021	0.005	-0.019	0.053
Reviews	0.08	1	-0.01	0.082	0.037	0.626
Price	-0.021	-0.01	1	-0.017	0.018	-0.011
Content	0.005	0.082	-0.017	1	-0.015	0.05
Size	-0.019	0.037	0.018	-0.015	1	0.017
Installs	0.053	0.626	-0.011	0.05	0.017	1

3.2. Model Evaluation

The results given in Fig. 5 show that regardless of whether the model contains dummy variables, the predicted values obtained by them are not much different from the actual mean, and the standard deviation of the model with the dummy variables is higher than that of the model without the dummy variables. In the scatterplot, the slope of the regression line with dummy variables is smaller, which means that after adding dummy variables, the influence of independent variables on the rating of the software is weaker.

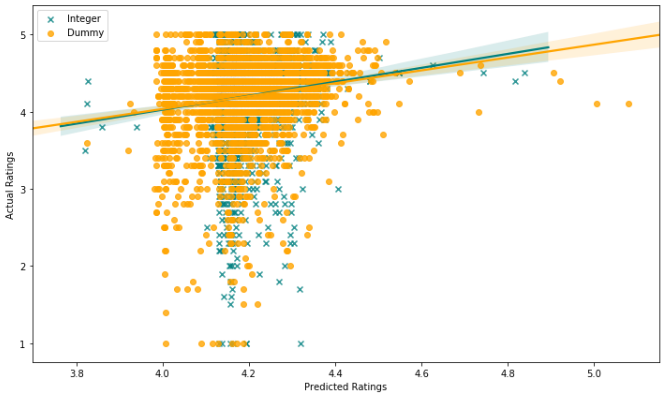

Figure 5: Linear model.

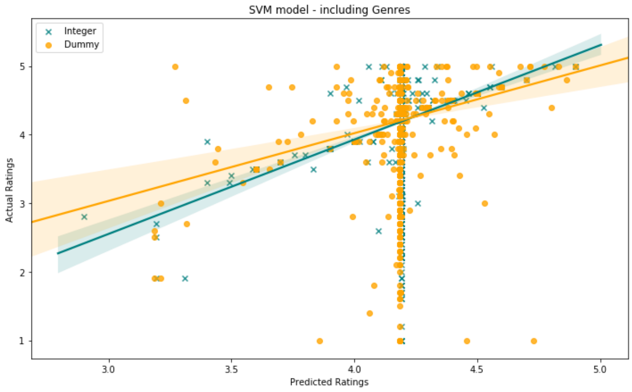

Figure 6: SVR model.

In the SVR model presented in Fig. 6, one sees that compared with the actual mean, its prediction accuracy is higher, although the standard deviation of virtual coding is still larger than that of integer coding, but compared with the linear regression model, it has been greatly improved, and the model is more stable. Random forest regression performs similarly in models with or without dummy variables, with similar means and standard deviations (seen from Fig. 7). Although he has some deviations in the predicted values compared to other models, he is the best at plotting graphs.

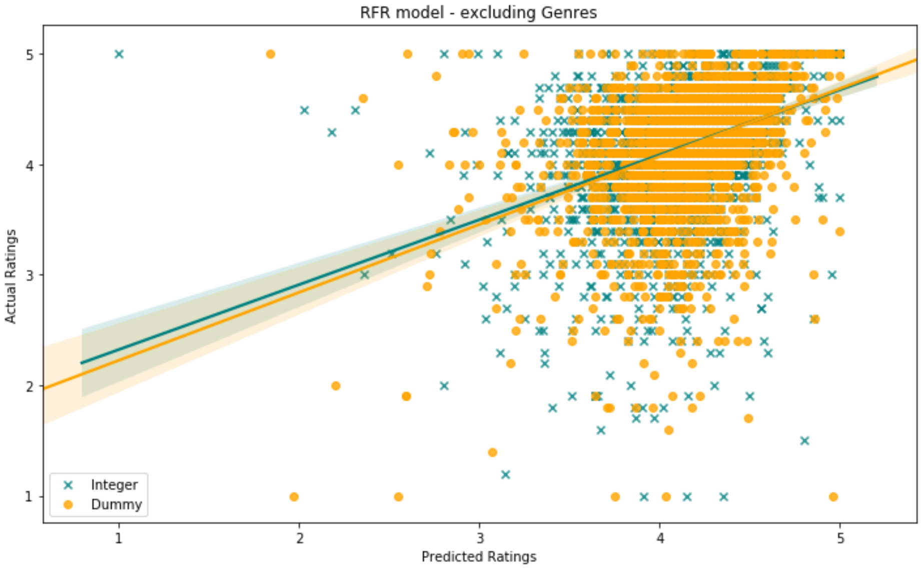

Figure 7: Random forest.

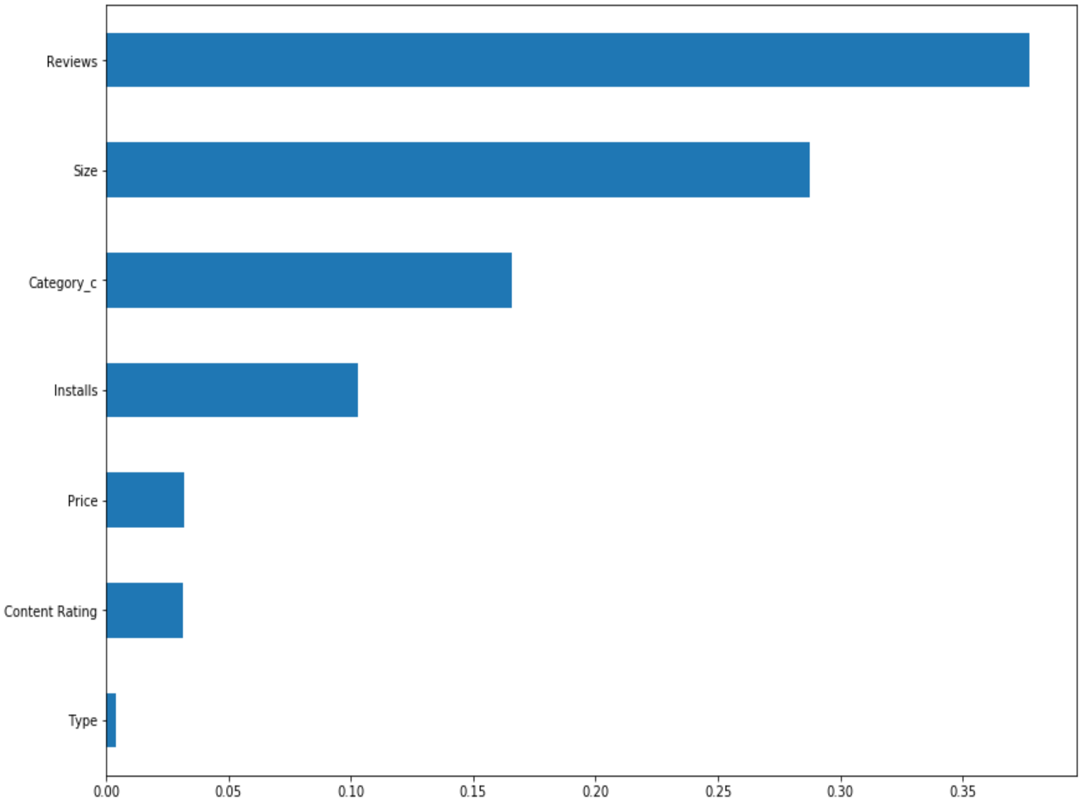

Figure 8: Variables.

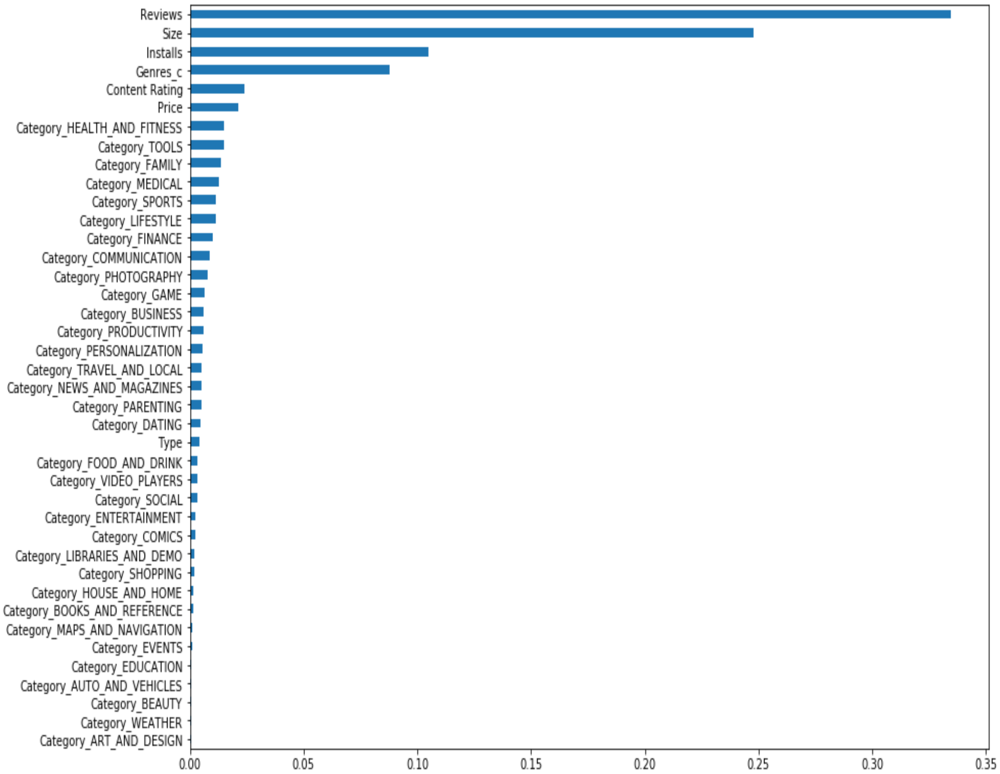

Figure 9: Variables for different types.

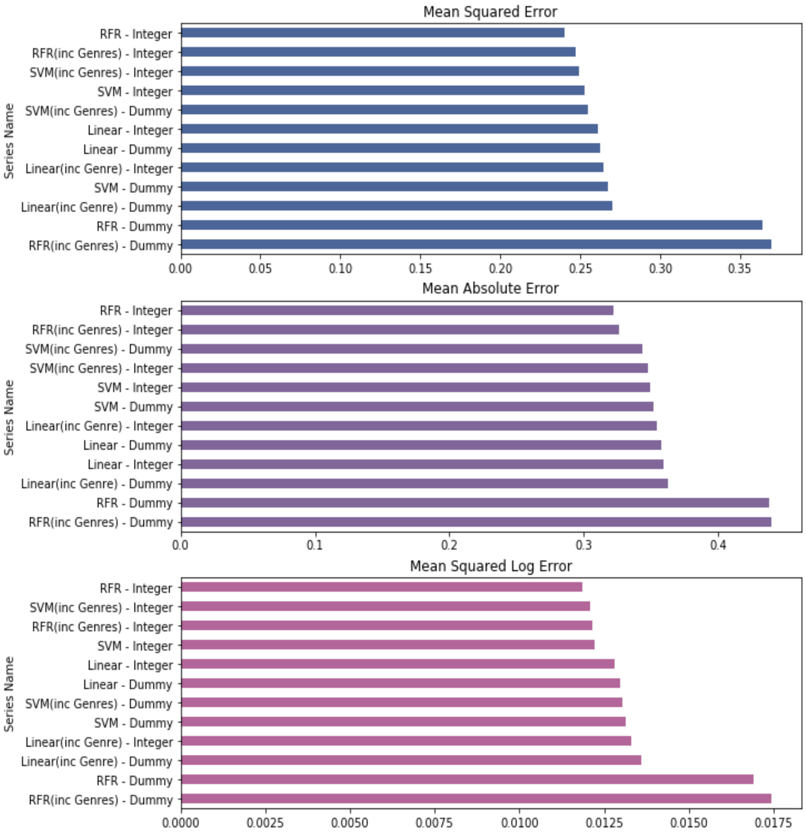

Figure 10: MSE, MAE and MSLE.

3.3. Explanation

One obtained the predicted values and their accuracy of each of the three models through regression, and analyzed the degree of influence of the variables in the regression model. Next, one looks at which independent variables have a large correlation with ratings, ranking according to the degree of impact, the top few are reviews, size, category, and number of installs. After further subdivision, variables such as reviews and size still play the biggest role in software ratings (seen from Fig. 8 and Fig. 9), but it is unexpected that the variables of the tool category also have such a high contribution, even higher than the categories such as family and life. In dummy coding, the SVM model with categorical variables had the lowest overall error rate, while in integer coding, the RFR model without categorical variables realized the best results(Fig. 10). In fact, the error values of these models are similar, so different data types are likely to change the choice of optimal model. What’s more, it is hard to believe that the error of the model with dummy variables in the RFR is exceptionally large, which one did not find from the fit plot above.

4. Limitations and Prospects

In terms of models, linear regression assumes a linear relationship between independent and dependent variables. For the problem of nonlinear relationships, linear regression may not get accurate predictions, especially when dealing with nonlinear problems. Moreover, linear regression is sensitive to outliers, which may affect the fitting effect and prediction results of the model. However, random forest models have high complexity and tend to overfit, especially when the number of trees is high. Moreover, random forests are not suitable for high-dimensional data: random forests do poorly process high-dimensional data, which may lead to overfitting or dimensional disaster.

The disadvantages and limitations of SVM regression are complex parameter selection, high computational complexity, and sensitivity to missing values. SVM regression, which requires some experience or parameter tuning skills. Longer training time, especially when the amount of data is large, increases the computational complexity. The SVM regression is sensitive to missing values, and requires additional processing if missing values are present in the data. In view of the limitations and shortcomings of the above models, one made data cleaning and feature selection, cleaned the data and outlier processing, and improved the stability and prediction ability of the model. Moreover, the most relevant features are selected by the feature selection technique, reducing the multicollinearity problem. It can also improve the prediction accuracy and robustness by integrating the results of multiple different models. Moreover, a better understanding and practical experience selects algorithms and models tailored to the problem with more domain-specific data.

With the continuous development of technologies, one looks forward to developing reinforcement learning and deep learning. With the continuous development of reinforcement learning and deep learning, these technologies will play a greater role in the field of prediction and modeling, providing more accurate and complex prediction models. It is expected to achieve automatic model selection and tuning, which can reduce the need for manual intervention and improve modeling efficiency and accuracy.

5. Conclusion

To sum up, this paper analyses the data of Google App Mall, and analyses the download and consumption behaviour of users in the app mall. This study performed regression analysis through linear regression, SVM and random forest model to reveal the relationships between different influencing factors for applied ratings, help to analyse and evaluate policies, reveal potential factors and predict future development. Although the model itself has a little lack of accuracy and overfitting problems, I believe that with the increasing application of multimodal data, more prediction models combining multiple data types will appear to improve the comprehensiveness and accuracy of the prediction results. It is hoped that this research can help developers understand users' needs and behaviour patterns, reveal the rules of market competition, find out problems and challenges, and provide reference for marketing and advertising. By deeply studying user download behaviour, one can better meet user needs and promote the development of the mobile app market.

References

[1] *Shanghai iResearch Market Consulting Co., LTD. (2022) Mobile Application Operations Growth Insights White Paper. IResearch series, 7, 17-260.*

[2] *Adnan, M. (2020). A Methodology for Comparison of User Reviews with Rating of Android Apps using Sentiment Analysi. Master's Thesis, Southwest University of Science and Technology.*

[3] *Qu, A. (2013). Research on user behavior of mobile application mall based on TAM and IDT models. Master's thesis, Beijing University of Posts and Telecommunications.*

[4] *Viennot, N., Garcia, E. and Nieh, J. (2014). A measurement study of google play. In The 2014 ACM international conference on Measurement and modeling of computer systems, 221-233.*

[5] *Farooqi, S., Feal, Á., Lauinger, T., McCoy, D., Shafiq, Z. and Vallina-Rodriguez, N. (2020). Understanding incentivized mobile app installs on google play store. In Proceedings of the ACM internet measurement conference 696-709.*

[6] *McIlroy, S., Ali, N. and Hassan, A.E. (2016). Fresh apps: an empirical study of frequently-updated mobile apps in the Google play store. Empirical Software Engineering, 21, 1346-1370.*

[7] *Yang, A.Z., Hassan, S., Zou, Y., and Hassan, A.E. (2022). An empirical study on release notes patterns of popular apps in the Google Play Store. Empirical Software Engineering, 27(2), 55.*

[8] *Noei, E. and Lyons, K. (2022). A study of gender in user reviews on the Google Play Store. Empirical Software Engineering, 27(2), 34.*

[9] *Fransiska, S., Rianto, R. and Gufroni, A.I. (2020). Sentiment Analysis Provider by. U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method. Scientific Journal of Informatics, 7(2), 203-212.*

[10] *Businge, J., Openja, M., Kavaler, K. et al. (2019). Studying android app popularity by cross-linking github and google play store. 2019 IEEE 26th international conference on software analysis, evolution and reengineering (SANER), 287-297.*