DATA 621 - Final Project
Google Play Store

Using Linear Regression to Predict App Ratings

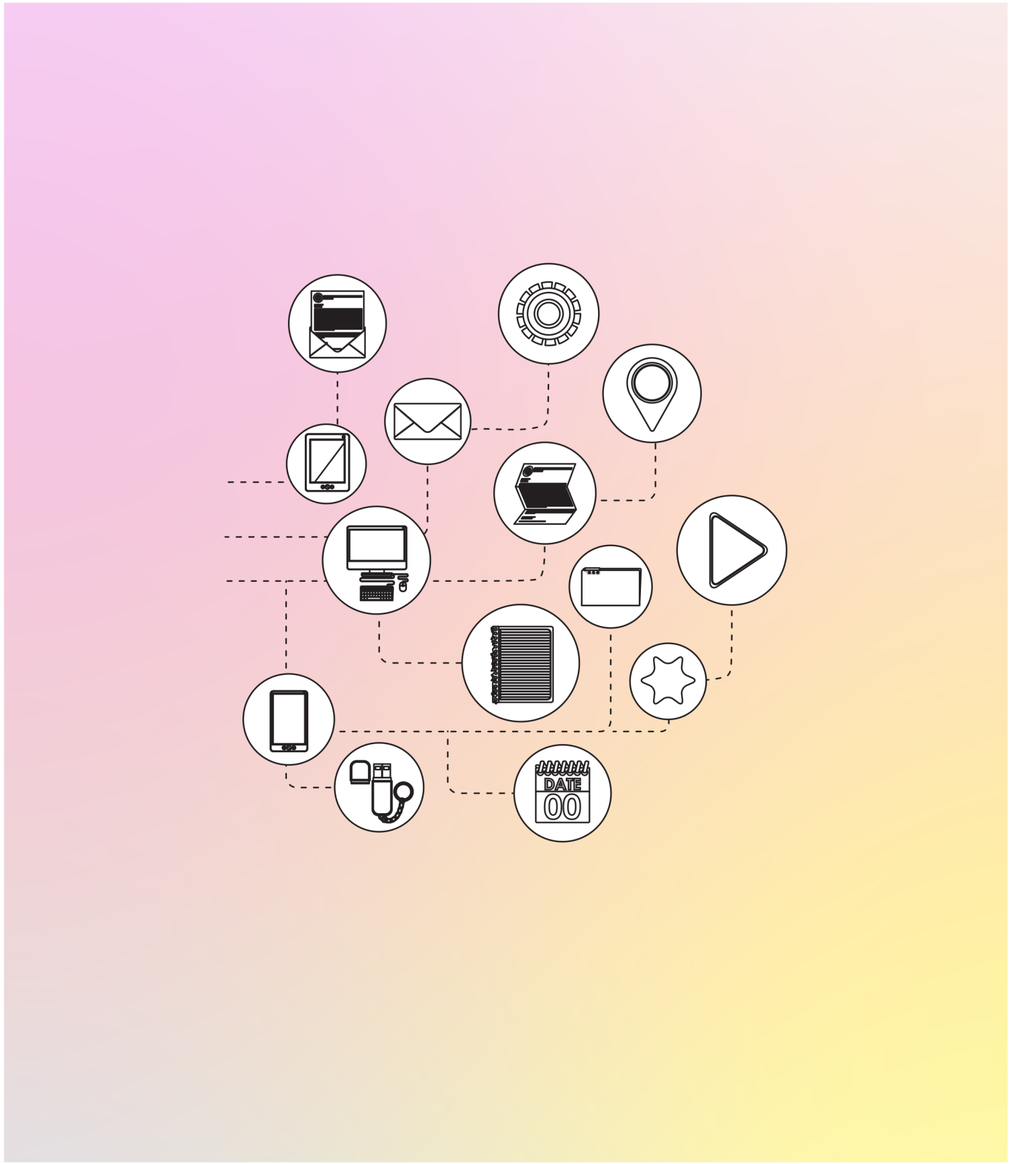
BY CHRISTIAN URIOSTEGUI, VYANNA HILL,
JOSE RODRIGUEZ

FALL 2023

Introduction

Finding apps that capture the interests of users in the form of reviews and volume of downloads, can secure key financial partners for developers. Highly reviewed apps also receive favorable promotions and prime positioning within app stores.

Our team, with the use of Google play store data we found on Kaggle, will create a best-fit regression model that can predict an app’s rating in the store based on features of an app. This predictive tool will assist in formulating strategies for the development of successful apps.



1. Data Exploration

About the Dataset

Some of the information contained in the dataset found include information on the app name, the category they fall under, the app's price, size and whether the app is free. Due to the large dataset (2.3 million+ rows) we will be using a much smaller sample for our analysis (~28,000 rows).

Variable Name	Definition
Ad Supported	Ad support in app
App ID	Package name
App Name	Name of the app
Category	App category (Adventure, Arcade, Social , etc)
Content Rating	Maturity level of app
Currency	App currency
Developer Email	Email of developer
Developer ID	Developer ID in Google playstore
Developer Website	Website of the developer
Editor Choice	Whether rate as editor choice
Free	Whether app is free or paid
In app purchases	Are there In-app purchases in app
Installs	Approximate install count
Last Updated	Last app update
Maximum Installs	Approximate minimum app install count
Minimum Android	Minimum android version supported
Price	App price
Privacy Policy	Privacy policy from developer
Rating	Average rating (1-5)
Rating Count	Number of ratings
Released	App launch date on Google playstore
Size	Size of application package

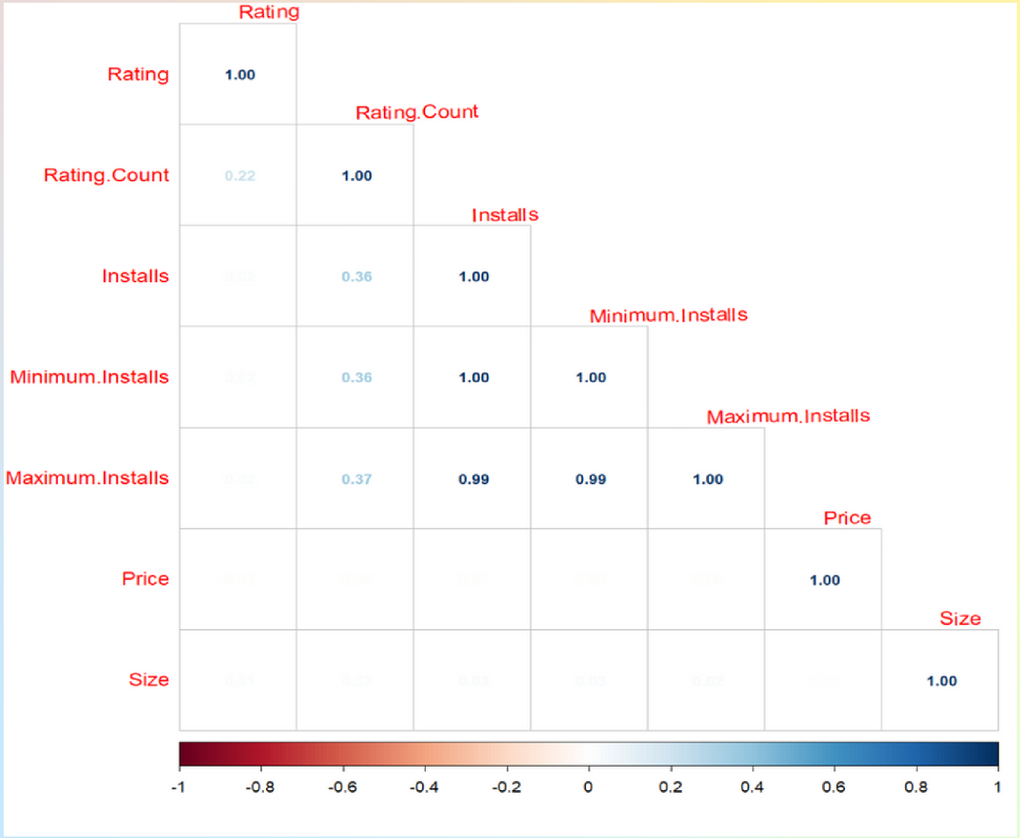
2. Data Preparation

ï..App.Name	App.Id	Category	Rating
0	0	0	270
Rating.Count	Installs	Minimum.Installs	Maximum.Installs
270	0	1	0
Free	Price	Currency	Size
0	0	0	0
Minimum.Android	Developer.Id	Developer.Website	Developer.Email
0	0	0	0
Released	Last.Updated	Content.Rating	Privacy.Policy
0	0	0	0
Ad.Supported	In.App.Purchases	Editors.Choice	Scraped.Time
0	0	0	0

Handling Missing Data

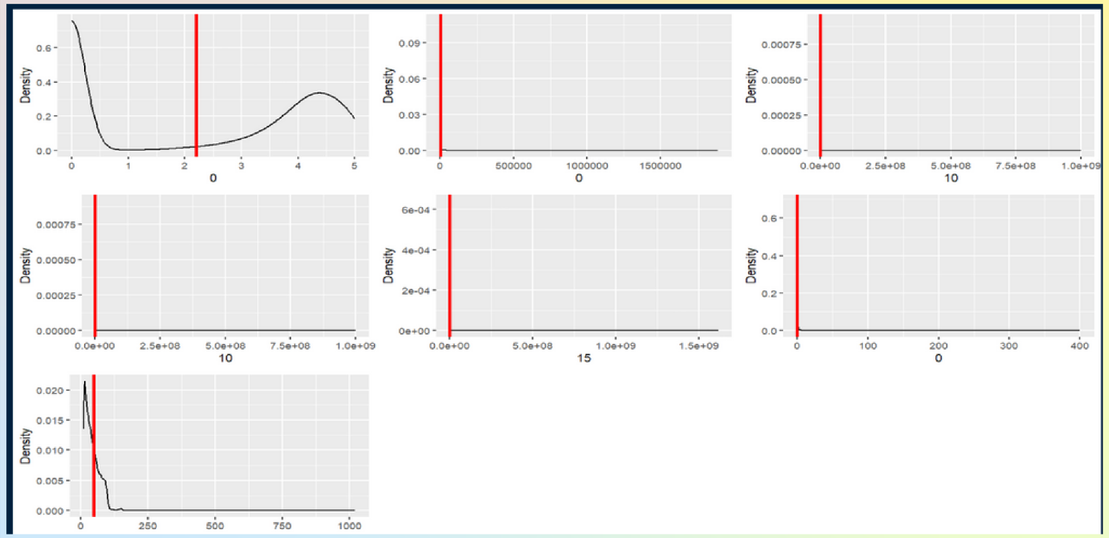
Missing data was handled by removing rows with NA values (Listwise Deletion). Two reasons prompted the omission of missing values: 1) removing missing values only removed about 5% of observations, 2) It is not known if the missing data points are missing completely at random (MCAR), hence imputation could introduce bias to the model.

Handling missing data is a crucial step in building linear regression models. Not addressing this can lead to model underperformance and significant loss in accuracy.



Checking for Multi-Colinearity

Prior to creating our models, we need to check for Multi-colinearity or the search for highly correlated variables. They can skew the data and as a result have to be removed. 'Maximum.Installs' and 'Minimum.Installs' were found to be highly correlated to 'Installs'.



Transformations

All numerical variables were found to be highly skewed. Transformations such as box-cox, logarithmic and square-root were explored. However, they did not normalize the data. Consequently, several assumptions are violated: 1) Normality, 2) linearity. Ultimately, this leads to assuming that the data exhibits complex relationships where a Nonparametric Regression could be the best choice.

3. Model Building

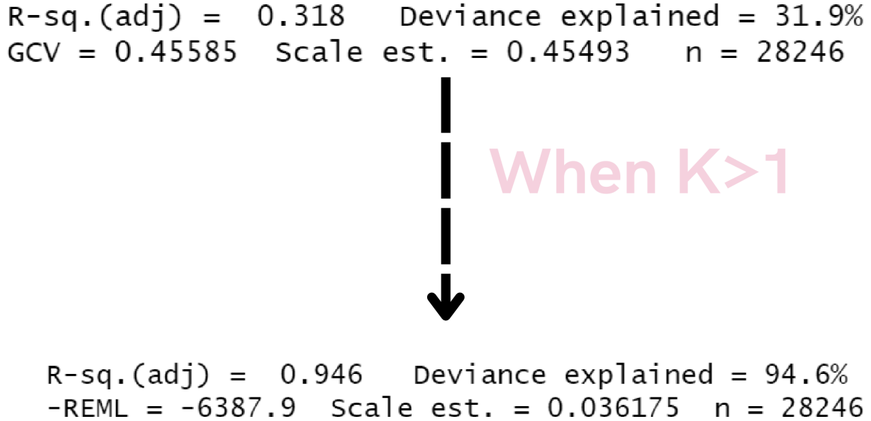
Choosing Our Model - GAM (Generalized Additive Models)

Because our variables weren’t found to be linear or normal, this violates the linearity and normality assumptions for traditional regression models like MLR/GLS/Robost regression. The only regression technique that can be modeled without parametric assumptions is (GAM General Additive Model). Thus, making it our choice when looking at different models.

Model Buidling with GAM

- The base GAM model was created with two features to gauge the R^2
  - The ‘Rating Count’ and ‘Category’ variables which saw an R^2 of 31%
- Models 1 through 4 added more features based on literature review
  - Models 1 & 4 work with additional features in the function
  - Models 2 & 3 introduced increased knots in its smoothed parameters

- Performance doubled from model one to two from the introduction of knots in the cubic splines of rating count and stars
  - Knots: the joining points of the segments in the segment lines
- There’s a safety range in the amount of K
  - When K is too large: no improvements in R^2 and process time triples
  - When K is too small: deviance and AIC/BIC doubles



4. Model Selection

Model Selection

model.build <chr>	df <dbl>	logLik <dbl>	AIC <dbl>	BIC <dbl>	deviance <dbl>	df.residual <dbl>	nobs <int>
gam1	67.44676	-16172.288	32481.90	33048.26	5197.4995	28178.55	28246
gam2	137.15845	6869.018	-13461.21	-12319.50	1016.8441	28108.84	28246
gam3	161.54426	12770.952	-25215.76	-23870.61	669.5219	28084.46	28246
gam4	70.30065	-23714.476	47571.99	48161.94	8865.9092	28175.70	28246

4 rows

Our updated model 3 performs the best. It has the lowest AIC, BIC and also has the highest Log-Likelihood. It also has the lowest deviance. Given the high performance of model 3, we will use this to perform our predictions.

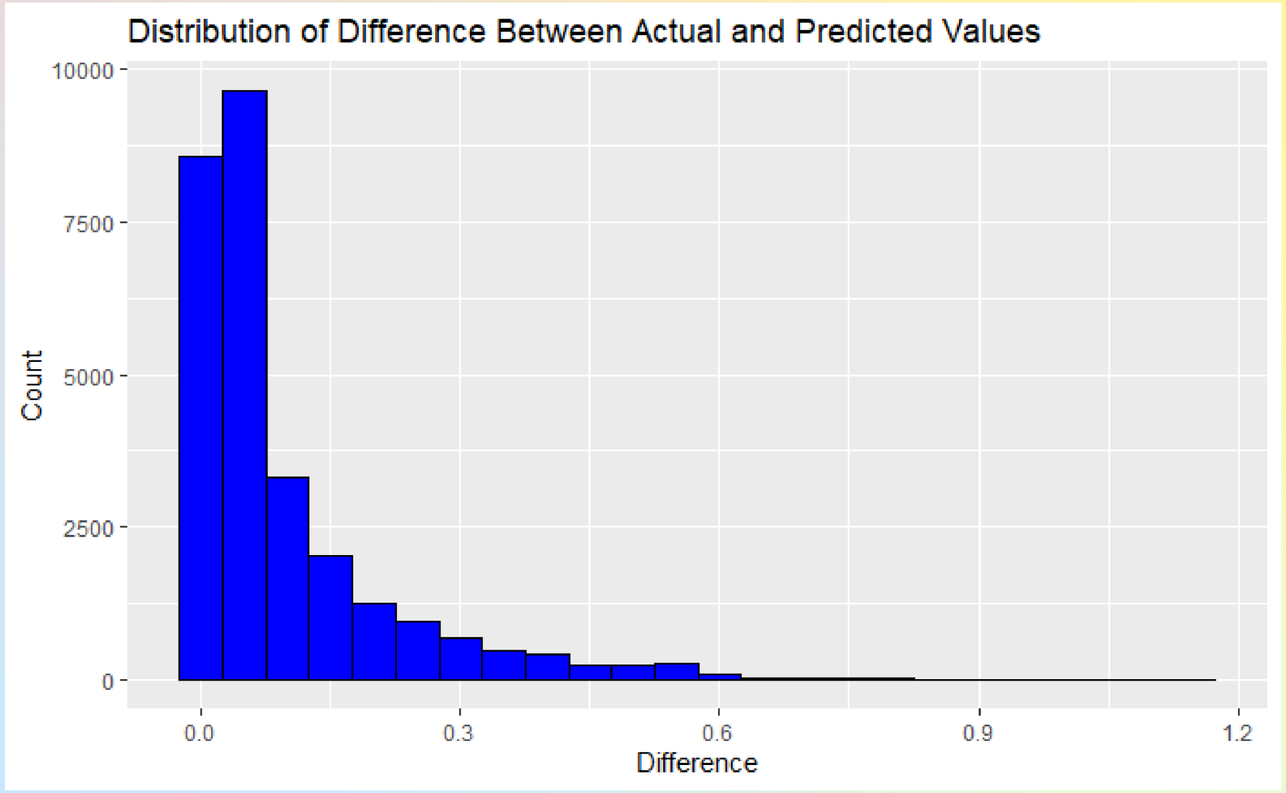
Feature Importance

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.8378909	0.0087042	96.263	< 2e-16	***
CategoryAdventure	0.0143087	0.0125410	1.141	0.25390	
CategoryArcade	0.0290436	0.0105846	2.744	0.00607	**
CategoryArt & Design	0.0038759	0.0135418	0.286	0.77471	
CategoryAuto & Vehicles	0.0094248	0.0136696	0.689	0.49053	
CategoryBeauty	0.0293035	0.0156618	1.871	0.06135	.
CategoryBoard	-0.0069876	0.0157647	-0.443	0.65760	
CategoryBooks & Reference	0.0433708	0.0094400	4.594	4.36e-06	***
CategoryBusiness	0.0250091	0.0094191	2.655	0.00793	**
CategoryCard	-0.0055891	0.0188815	-0.296	0.76722	
CategoryCasino	-0.0292344	0.0228092	-1.282	0.19996	
CategoryCasual	0.0067884	0.0104955	0.647	0.51777	
CategoryComics	0.0035115	0.0290545	0.121	0.90380	
CategoryCommunication	0.0268396	0.0108117	2.482	0.01305	*
CategoryDating	-0.0171642	0.0198066	-0.867	0.38617	
CategoryEducation	0.0365254	0.0090252	4.047	5.20e-05	***
CategoryEducational	0.0222155	0.0130591	1.701	0.08893	.

From the preferred model, the team reviewed the current collection of features and their coefficients in the regression model. Ad.SupportedTRUE, CategoryArcade, CategoryBooks & Reference, and CategoryBusiness were identified as statistically significant in the regression model.

The feature of ad supported is not a surprise as Wondwesen (2023) saw ad supported ads were more likely to have lower rating compared to the paid version in the app store.



Predictions

The team examined the predicted values of the predictions versus the actual values in the data set. The team noticed that none of the predicted values fall outside the confidence interval.

Final Thoughts

- Generalized Additive Modeling supports non-normal data through its support of smoothing the parameters of the model
  - The regression model improved its fit on the dataset to 96% (R^2) with the use of cubic splines on highly skewed features
- The use of splines in highly skewed data provides a closer fit as the smoothed parameters are no longer have the assumption of a relationship to the dataset
- For future applications
  - An expansion of the type of smoothing parameters on highly skewed features
  - An review on regression models with non parametric approaches

References

- Tafesse, Wondwesen. (2023). The differential effects of developers’ app store strategy on the performance of niche and popular mobile apps. Journal of Marketing Analytics. 11. 1-14. 10.1057/s41270-023-00216-8.1
- Lee, Gunwoong & Santanam, Raghu. (2014). Determinants of Mobile Apps' Success: Evidence from the App Store Market. Journal of Management Information Systems. 31. 133-170. 10.2753/MIS0742-1222310206.
- Kapoor, Anuj & Vij, Madhu. (2020). How to Boost your App Store Rating? An Empirical Assessment of Ratings for Mobile Banking Apps. Journal of theoretical and applied electronic commerce research. 15. 10.4067/S0718-18762020000100108.
- Wu, Huayao & Deng, Wenjun & Niu, Xintao & Nie, Changhai. (2021). Identifying Key Features from App User Reviews. 922-932. 10.1109/ICSE43902.2021.00088.