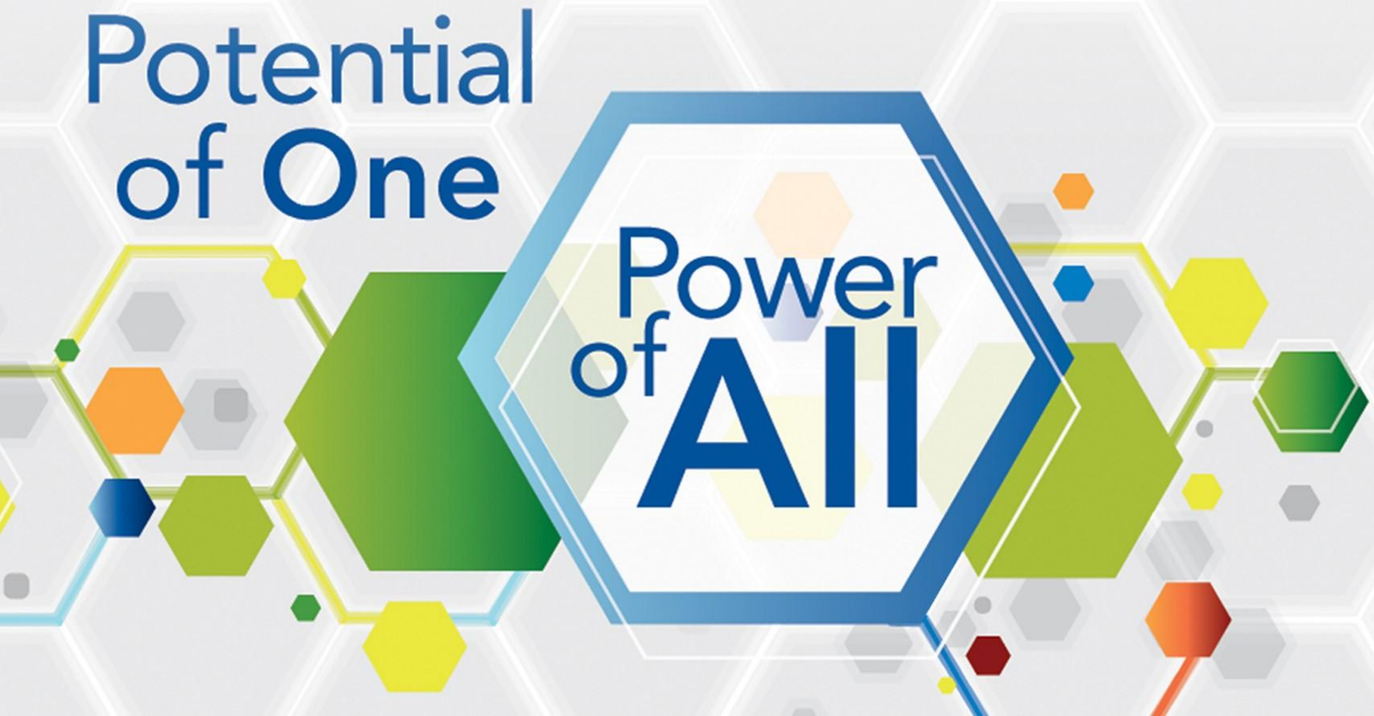
SAS GLOBAL FORUM

Text Analytics: Predicting the success of Newly Released Free Android Apps Using SAS® Enterprise Miner™ and SAS® Sentiment Analysis Studio



Text Analytics: Predicting the success of Newly Released Free Android Apps using SAS® Enterprise Miner™ & Sentiment Analysis Studio

Author(s) - Reddy, Vandana & Dugar, Chinmay
Faculty Advisor- Dr. Goutam Chakraborty, Professor (Marketing)
Department of MIS and SAS and OSU data mining certificate program, Oklahoma State University, Stillwater, Oklahoma 74075

Abstract

With smart-phone and mobile apps market developing so rapidly, the expectations about effectiveness of mobile applications is high. Marketers and app developers need to analyze huge data available much before the app release, not only to better market the app, but also to avoid costly mistakes. The purpose of this poster is to build models to predict the success rate of an app to be released in a particular category. The poster summarizes the success trends across various factors and also introduces a new SAS® macro %getappdata, which we developed for web crawling and text parsing. A Linear Regression model with least Average Squared Error is selected as the best model, and number of installations, app maturity content are considered as significant model variables. App category, user reviews, and average customer sentiment score are also considered as important variables in deciding the success of an app.

Objective

- Data has to be collected for 270 android apps under the "Top free newly released apps" category from https://play.google.com/store. SAS® web-crawling macro will be used to collect data from the website.
- SAS® Sentiment Analysis Studio will be used to calculate the average customer sentiment score for each app.
- Model building has to be done in SAS® Enterprise Miner™ to predict the rank of an app by considering average rating, number of installations, total number of reviews, number of 1-5 star ratings, app size, category, content rating, and average customer sentiment score as independent variables.

Method

- Data Preparation
  Data has been collected for 270 android apps under "Top free newly released apps" category from https://play.google.com/store. We have developed %macro getAppData to crawl Google play website for data collection.

Explanatory Data Analysis

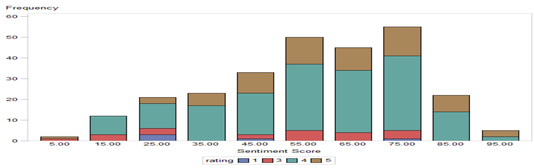


Figure 1. This shows the frequency distribution of Sentiment Scores and the distribution of average rating.

Figure 2. This shows the frequency distribution of app genre and the distribution of average rating in each genre.

- Model Building
  Linear Regression, Neural and Auto-Neural network models have been built to predict the rank of an app by considering average rating, number of installations, total number of reviews, number of 1-5 star ratings, app size, category, Average Sentiment Score and content rating as independent variables. Linear Regression model with least Average Squared Error is selected as best model and number of installations, app maturity content are considered as significant model variables. App category and user reviews are also considered as important variables in deciding the success of an app. In this study the success rate is analyzed within each app category, content rating level, average rating level and range of number of installations.

Results

Figure 3: Modeling

Figure 4: Sentiment Analysis Results

Model	ASE
Linear Regression	0.04288
Neural Network	0.053233
Auto Neural Network	0.054532

Figure 5: Model Accuracy Results

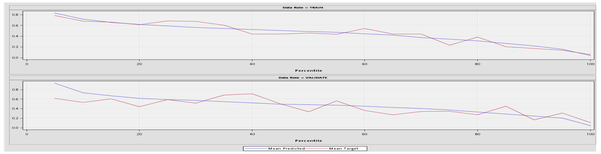


Figure 6: Score rankings overlay plot for Linear regression model

Conclusions

- Average Squared Error for all the models is captured and from the score rankings overlay plot (mean predicted) for validation data of best model, it is clear that the model is accurate in predicting the rank with slightly high ASE in few percentile ranges.
- Apps with content rating level 'High' are observed to be more successful when compared with apps in other content rating levels.
- Apps with high average rating are observed to have more number of installs. Average rating and number of installations are the primary reference factors for an customer to decide whether to install an app or not.
- Apps with high average rating are observed to have high positive sentiment score.
- Future work includes considering the customer text reviews into account for calculating the sentiment scores for each app to have a better understanding on customer feedback.

References

- Roehl, William G. (2011), *Using SAS® to Help Fight Crime: Scraping and Reporting Inmate Data* [PDF file]. Available from http://support.sas.com/resources/papers/proceedings11/140-2011.pdf.
- http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/e
- %getappsdata macro is available from authors on request.

SAS® GLOBAL FORUM

Washington, D.C.

March 23–26, 2014