Big Data Continual Assessment

The prediction of application popularity based on features identified in App Store data

Shannen Dowlinga

aIT Tralee, Co. Kerry, Ireland

Abstract

App Stores possess a variety of information, detailing application features and user feedback. Information about application features, price, number of downloads, reviews and genres are some of the insights to be found. Much of this data can be invaluable in foreseeing the popularity of an application. Correlations between popularity and reviews are often reported to be strong, however price and review correlations do not follow the same pattern. This paper has examined this, along with establishing the extent to which these features may be beneficial in predicting an application’s popularity. Following investigation, it was determined some features are more capable of being utilised for this task. Previous research has determined the usefulness of app ratings for prediction and this paper has further proved the case. The exploration of using the number of app downloads feature, to perform the same task, was carried out and found to need more investigation for a conclusive result.

*Keywords:* mobile apps; App Store; Android; iOS; app features; popularity; downloads; ratings; prediction;

1. Background

The following sections introduce the research undertaken, as well as detailing the methodology, findings and conclusions reached upon completion.

* 1. Introduction

The motivation for this project originates from research conducted on App Store analysis. Several papers were investigated on the topic, including conference proceedings and reports, on predicting app success, app ratings and determining relationships between app store and application features. Informed by this research, the paper intends to continue on the theme of predicting an application’s popularity. Using data analysis and machine learning methods, the ability to predict the number of application installations or downloads, has been explored. Main methods of enquiry followed, during the implementation phase of the project, have applied Decision Tree and Support Vector Machine methodologies to the prediction task. Initial phases of the project intended to compare findings from Android and iOS data, however this has not been achievable due to the limitations of iOS data available.

* 1. Literature Review

Mobile applications have seen an exponential growth in their development and availability in recent years. App Stores on portable devices have been instrumental in this growth, providing a single source of distribution. With this source, comes a large volume of data. App Stores provide information on apps in terms of size, price, rating, versions, number of downloads and updates and much more. Along with this, customer information can also be obtained. This provides a resource to developers to understand the features that entice users to choose their application. Realising the potential of an application and developing new apps around this, is extremely achievable (Harman, et al., 2012). Due to this growth it can be said, that there has been a saturation in terms of the number of applications available today. This has influenced the lifespan of many applications. App features can be utilised to predict the sustainability of an app. Some of these features include the price, number of updates, category and rating of an app. This can be a negative or positive effect, where a free and regularly updated application, has a greater lifespan (Lee & Raghu, 2014).

These features lend themselves to generate a thorough description of an application, while some can influence other features. Correlations between app features, has been the topic of much research. Some findings have determined that an app’s rating has a direct effect on the number of downloads it records. In contrast to that, it has been found that the pricing of an application, does not affect either of these attributes (Harman, et al., 2012). Research has previously been undertaken by University College London, to utilise App Store information, in predicting application ratings. This research has proven that an application’s rating can be determined, solely from the features it offers. Similarly, ratings can be predicted based on a developer’s description of an app, they have created. Correlations between numeric features, i.e. price, rating and downloads, have also been found in this research, which has been reoccurring throughout the review of literature studied (Sarro, et al., 2018).

Some important points to note, are that much of this data, cannot be proven to be completely correct and accurate. Customers may leave incorrect ratings or reviews, pricing may not portray the true cost of the application, i.e. a subscription service or *“in app purchases”* may be encountered later, and developers may also have a biased opinion on the application they have created. Similarly, a highly rated application may not be as popular as expected. It may only have one positive review or good rating which is also a biased view (Finkelstein, et al., 2014).

Following a review of these findings, this paper aims to further confirm, or disprove, these discoveries and attempt to anticipate the popularity of the applications.

1. Implementation
   1. Methodology

The Knowledge Discovery in Databases process, (KDD), was followed in development of this project. The beginnings of the project involved identifying a problem question and area of research. This was done based on the review of papers, mentioned in the previous section. Following the choice of this research question, the required resources, datasets and additional research, were gathered as part of the data requirement stage.

The next phases of the project, involved pre-processing the data, performing analysis, gathering descriptive statistics and creating visualisations of the data. The primary task in this phase was the pre-processing of the acquired dataset. The data was cleaned, removing any noisy and inconsistent data, missing values were replaced where appropriate, by calculating the mean value of the remainder of the dataset. Following some basic pre-processing tasks, statistics and visualisations were employed to learn more about the data. This included displaying descriptive statistics for the app features, such as mean and median values, plotting a number of graphs for visualisation, along with presenting the correlations between the attributes in the dataset.

Table 1. Sample of dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| App Name | Reviews | Price | Type | Category | ContentRating | Rating | Installs |
|  |  |  |  |  |  |  |  |
| Google | 4828372 | 0 | Free | SOCIAL | Teen | 4.2 | 1000000000 |
| Candy Crush Saga | 22419455 | 0 | Free | FAMILY | Everyone | 4.4 | 500000000 |
| Weather Live | 76593 | 5.99 | Paid | WEATHER | Everyone | 4.5 | 500000 |

The final stage of the project development entailed applying machine learning algorithms to the data. The classification algorithms employed were the *k-*Nearest Neighbor, Decision Tree and Support Vector Machine classifiers. The models developed using these algorithms, were created with the intention of being used for predictive modelling. Initial results, returned for the prediction of ‘Installs’, were poor, meaning another approach was considered. Based on research, the option of prediction of application popularity, from the Rating attribute, was considered a possibility. The findings below clearly demonstrate the vast difference in the accuracy of results for both approaches.

* 1. Findings

The results, following experimentation, clearly show the difference in the ability to predict the Installs and Rating features. Accuracy values are almost double in contrast, of both experiments. It could be said that it is much more feasible to utilise an application’s rating in predicting app popularity. This approach was considered during development, following the acknowledgment that the preliminary results were not reliable.

Table 2. Algorithm accuracy results.

|  |  |  |  |
| --- | --- | --- | --- |
| Predicted Attribute | *k*NN \* | CART \* | SVM \* |
|  |  |  |  |
| Installs | 0.42 | 0.41 | 0.31 |
| Rating | 0.73 | 0.67 | 0.77 |

\*Results rounded to two decimal places.

* 1. Discussion

During the data pre-processing phase of the project, the skew of the data was discovered, using the class distribution and skew methods from Python. As the dataset was large, with 10,000 entries, the data was imbalanced and had the potential of overfitting. Aiming to reduce the affect this may have, some more entries were removed, in an attempt to balance this further. The ‘Installs’ feature was highly skewed, and as it was the focus of the investigation, emphasis was put on reducing this imbalance, however, poor results were still obtained in later stages.

There are some suggestions as to why the results were poor for the Installs attribute. One possibility may be due to the imbalance of data. As outlined in the accompanying notebook, some ranges of Installs had thousands of entries, as opposed to tens or hundreds, for many more. Attempts were made to reduce this contrast, however, it was difficult, as removal or replacements led to an imbalance in other attributes, i.e. Rating, Reviews etc. Another suggestion is that the possible values of the Installs feature, was not specific enough. The original dataset had values such as 10,000+, 500,000+, 1,000,000+ and so on. In cleaning the dataset, the ‘+’ symbol was removed, leaving 10,000, 500,000 etc. This is not an accurate measure of installations, merely a range in which the actual value could be a number of quantities. The range of downloads is also extremely large, fluctuating from 0 and 1 to 1,000,000,000. It was thought these values may be sufficient but that does not appear to be the case on viewing the findings.

In keeping with the literature reviewed as part of the project, some correlations were found between features. One correlation of note involved the relation between Installs and Reviews. This suggests an app with more downloads, has more reviews, which is understandable. There was an obvious correlation between Price and Type (free/paid), however it was not as strong as might be expected. Many correlations relating to the price feature were negative. This may be due to less installations of paid apps, meaning less reviews, ratings etc.

* 1. Conclusion

Following the findings of this paper, and the initial recording of poor results, it could be determined that more investigation is needed to accurately predict application installs or downloads. First step for improvement in achieving this goal, would be to utilise a more consistent dataset while further data pre-processing and cleaning is also required to the current dataset, if possible. The paper did conclude and confirm the accuracy of the approach to use ratings as a determinant of application success. This feature was accurately predicted and proves the benefit of the previous research into this topic.

1. Illustrations

The following illustrations outline the comparison between the three algorithms, employed in the prediction model. The image demonstrates the poor accuracy described in the findings above, as well as, the improvement on change of prediction attribute.

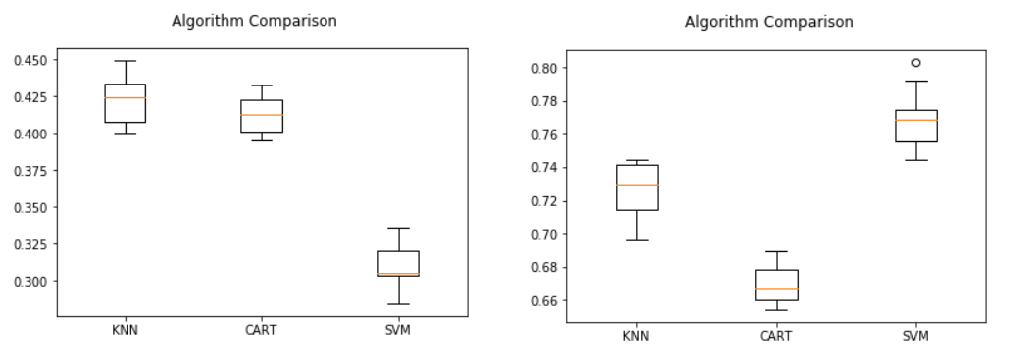
****

Fig. 1. Comparison of *k-*Nearest Neighbor, Decision Tree and Support Vector Machine classifiers. (a) ‘Installs’ prediction model; (b) ‘Rating’ prediction model.

1. Glossary
   1. Glossary of terms used in the document

* KDD – Knowledge Discovery in Databases
* *k*NN – *k-*Nearest Neighbor
* CART – Decision Tree
* SVM – Support Vector Machine

References

Finkelstein, A. et al., 2014. *App Store Analysis: Mining App Stores for Relationships between Customer, Business and Technical Characteristics ,* London: University College London.

Harman, M., Jia, Y. & Zhang, Y., 2012. *App Store Mining and Analysis: MSR for App Stores,* London: University College London.

Lee, G. & Raghu, T. S., 2014. *Determinants of Mobile Apps Success: Evidence from App Store Market,* Tempe, Arizona: Arizona State University.

Sarro, F., Harman, M., Jia, Y. & Zhang, Y., 2018. *Customer Rating Reactions Can Be Predicted Purely Using App Features,* London: University College London.