Characteristics Classification of Apple Store Dataset Using Clustering

Debasish Rath
MTECH (PG)
DSML
Pes University
Bangalore, Karnataka, India
debasishrath301@gmail.com

Medha B
MTECH (PG)
DSML
Pes University
Bangalore, Karnataka, India
medhab14126@gmail.com

Subham Budigas
MTECH (PG)
DSML
Pes University
Bangalore, Karnataka, India
subhamd1205@gmail.com

Supriya S K
MTECH (PG)
DSML
Pes University
Bangalore, Karnataka, India
supriyask021@gmail.com

***Abstract*—This research is interested in the user ratings of Apps on Apple Stores. The purpose of this research is to have a better understanding of some characteristics of the good Apps on Apple Store so Apps makers can potentially focus on these traits to maximize their profit. The data for this research is collected from kaggle.com, Four different attributes contribute directly toward an App's user rating: rating_count_tot, rating_count_ver, user_rating and user_rating_ver. The relationship between Apps receiving higher ratings and Apps receiving lower ratings is analyzed using Exploratory Data Analysis and Data Science technique "clustering" on their numerical attributes. Apps, which are represented as a data point, with similar characteristics in rating are classified as belonging to the same cluster, while common characteristics of all Apps in the same clusters are the determining traits of Apps for that cluster. Both techniques are achieved using Jupyter Notebook and libraries including pandas, numpy, seaborn, and matplotlib. The data reveals direct correlation from number of devices supported and languages supported to user rating.**

Keywords—Characteristics Classification of Apple Store Dataset Using Clustering, Machine Learning, Decision Tree, Pca, Recommendation System, Ratings.

I. Introduction

As smart phones entering people's life, Apps for different operating systems for smart phones create brand new markets for Apps developers. Gradually, mobile Apps become profitable and grow faster than ever as new technologies and features are added to mobile devices. Mobile Apps Stores like Apple Store allow users to rate their experience with Apps, and users usually use ratings from other users to determine whether to download an App or not. To maximize a mobile App company's profit, it is important to understand what the users think a good App is like. This research aims to have a better understanding on what traits are users for mobile Apps are looking for when using them.

Previously, the same data set has been used with a focus on popularity of different Genres, and the relationship between how willing users are to pay for Apps within different Genre and the Genres themselves. For the purpose of this project, clustering is used to deal with numerical attributes of this data.

For this project, the importance of Genre is increased, to understand multiple numerical attributes and provide an easy visual representation of the result, clustering is chosen. This research helps to find some meaningful correlations between one or some of the attributes and user rating.

II. Literature Survey

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.

Principal Component Analysis (PCA) is basically a statistical procedure to convert a set of observation of possibly correlated variables into a set of values of linearly uncorrelated variables. Each of the principal components is chosen in such a way so that it would describe most of the still available variance and all these principal components are orthogonal to each other. In all principal components first principal component has maximum variance. PCA basically search a linear combination of variables so that we can extract maximum variance from the variables. Once this process completes it removes it and search for another linear combination which gives an explanation about the maximum proportion of remaining variance which basically leads to orthogonal factors. In this method, we analyze total variance.

The explosive growth in the amount of available digital information and the number of visitors to the Internet have created a potential challenge of information overload which hinders timely access to items of interest on the Internet. Information retrieval systems, such as Google, Devil Finder and Altavista have partially solved this problem but prioritization and personalization (where a system maps available content to user's interests and preferences) of information were absent. This has increased the demand for

recommender systems more than ever before. Recommender systems are information filtering systems that deal with the problem of information overload by filtering vital information fragment out of large amount of dynamically generated information according to user's preferences, interest, or observed behavior about item]. Recommender system has the ability to predict whether a particular user would prefer an item or not based on the user's profile.

Recommender systems are beneficial to both service providers and users . They reduce transaction costs of finding and selecting items in an online shopping environment . Recommendation systems have also proved to improve decision making process and quality . In e-commerce setting, recommender systems enhance revenues, for the fact that they are effective means of selling more products . In scientific libraries, recommender systems support users by allowing them to move beyond catalog searches. Therefore, the need to use efficient and accurate recommendation techniques within a system that will provide relevant and dependable recommendations for users cannot be over-emphasized.

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

III. WORKING PRINCIPLE

In this research, we use Clustering to find out patterns in the dataset. Furthermore, they also intended to find the best accuracy of the models used. In this project by using K_Means algorithm to get accurate results. We used different base learning algorithms, such as Naïve Bayes, Logistic Regression, kNN, Decision Tree. A block diagram is given below to show the proposed model of Apple Store Dataset

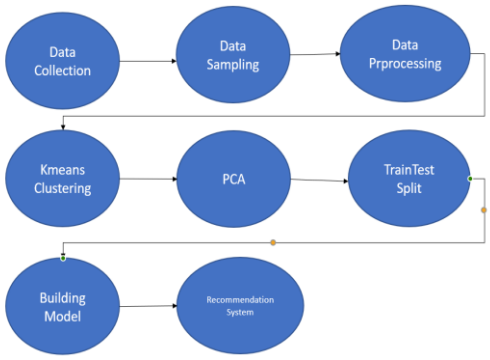

Fig. 1. Block diagram of the proposed model.

A. *Data*

Dataset is one of the most important factor in any kind of machine learning and predictions. In the research of Apple Store Dataset, the most technical and one of the toughest tasks is to find out the proper dataset applicable to the models. As the aim is to find out the proper dataset, we used a dataset that is collected from kaggle. There were seven thousand one hundred and ninety seven rows of data in the dataset. The dataset had seventeen columns, which were 'id','track_name','size_bytes','currency','price','rating_count_tot','rating_count_ver','user_rating','user_rating_ver','ver','count_rating','prime_genre','sup_devices.num','ipadSc_urls.num','lang.num','vpp_lic',

Where 'track_name','ver','count_rating''prime_gener' and 'currency are categorical columns' and rest are all numerical columns.

B. *Data Sampling*

Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined. It enables data scientists, predictive modelers and other data analysts to work with a small, manageable amount of data about a statistical population to build and run analytical models more quickly, while still producing accurate findings Sampling can be particularly useful with data sets that are too large to efficiently analyze in full -- for example, in big data analytics applications or surveys. Identifying and analyzing a representative sample is more efficient and cost-effective than surveying the entirety of the data or population.
An important consideration, though, is the size of the required data sample and the possibility of introducing a sampling error. In some cases, a small sample can reveal the most important information about a data set. In others, using a larger sample can increase the likelihood of accurately representing the data as a whole, even though the increased size of the sample may impede ease of manipulation and interpretation.

C. *Data Preprocessing*

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be

directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. There are several steps of data pre-processing, which are,

i. Getting the data
ii. Importing libraries
iii. Importing datasets
iv. Finding Missing Data
v. Splitting dataset into train and test split

To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset.

In order to perform data pre-processing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data pre-processing which are: Numpy, Matplotlib, Pandas.

We need to import the datasets which we have collected for our machine learning project and the next step of data pre-processing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Feature scaling is the final step of data pre-processing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominate the other variable.

D. K-Means Clustering

Clustering is the task of grouping together a set of objects in a way that objects in the same cluster are more similar to each other than to objects in other clusters. Similarity is a metric that reflects the strength of relationship between two data objects. Clustering is mainly used for exploratory data mining. It has manifold usage in many fields such as machine learning, pattern recognition, image analysis, information retrieval, bio-informatics, data compression, and computer graphics.

K-Means falls under the category of centroid-based clustering. A centroid is a data point at the center of a cluster. In centroid-based clustering, clusters are represented by a central vector or a centroid. This centroid might not necessarily be a member of the dataset. Centroid-based clustering is an iterative algorithm in which the notion of similarity is derived by how close a data point is to the centroid of the cluster.

E. Principle Components Analysis

Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving an eigenvalue/eigenvector problem, and the new variables are defined by the dataset at hand, not a priori, hence making PCA an adaptive data analysis technique. It is adaptive in another sense too, since variants of the technique have been developed that are tailored to various different data types and structures. This article will begin by introducing the basic ideas of PCA, discussing what it can and cannot do. It will then describe some variants of PCA and their application.

F. Train-Test Split

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.

G. Recommendation Systems

Recommender systems are beneficial to both service providers and users . They reduce transaction costs of finding and selecting items in an online shopping environment . Recommendation systems have also proved to improve decision making process and quality . In e-commerce setting, recommender systems enhance revenues, for the fact that they are effective means of selling more products . In scientific libraries, recommender systems support users by allowing them to move beyond catalog searches. Therefore, the need to use efficient and accurate recommendation techniques within a system that will provide relevant and dependable recommendations for users cannot be over-emphasized.

This collects relevant information of users to generate a user profile or model for the prediction tasks including user's attribute, behaviours or content of the resources the user accesses. A recommendation agent cannot function accurately until the user profile/model has been well constructed. The system needs to know as much as possible from the user in order to provide reasonable recommendation right from the onset. Recommender systems rely on different types of input such as the most convenient high quality explicit feedback, which includes explicit input by users regarding their interest in item or implicit feedback by inferring user preferences indirectly through observing user behavior. Hybrid feedback can also be obtained through the combination of both explicit and implicit feedback. In E-learning platform, a user profile is a collection of personal information associated with a specific user. This information

includes cognitive skills, intellectual abilities, learning styles, interest, preferences and interaction with the system. The user profile is normally used to retrieve the needed information to build up a model of the user. Thus, a user profile describes a simple user model. The success of any recommendation system depends largely on its ability to represent user's current interests

H. Building a Model

Model building is a hobby that involves the creation of physical models either from kits or from materials and components acquired by the builder. The kits contain several pieces that need to be assembled in order to make a final model. Most model-building categories have a range of common scales that make them manageable for the average person both to complete and display. A model is generally considered physical representations of an object and maintains accurate relationships between all of its aspects.

The model building kits can be classified according to skill levels that represent the degree of difficulty for the hobbyist. These include skill level 1 with snap-together pieces that do not require glue or paint; skill level 2, which requires glue and paint; and, skill level 3 kits that include smaller and more detailed parts. Advanced skill levels 4 and 5 kits ship with components that have extra-fine details. Particularly, level 5 requires expert-level skills.

Model building is not exclusively a hobbyist pursuit. The complexity of assembling representations of actual objects has become a career for several people. There are, for instance, those who build models to commemorate historic events, employed to construct models using past events as a basis to predict future events of high commercial interest.

Below is the base models result before pca with their accuracy score and the best models classification report is shown

```
Model name : LogisticRegression()
accuracy score  0.7548611111111111
Model name : KNeighborsClassifier()
accuracy score  0.7159722222222222
Model name : DecisionTreeClassifier()
accuracy score  1.0
Model name : GaussianNB()
accuracy score  0.36736111111111114
```

Fig. 2. Accuracy score of the base models before pca.

```
Train Set Accuracy:74.90012159110648
Test Set Accuracy:75.48611111111111

Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         0
           1       1.00      0.75      0.86      1440

    accuracy                           0.75      1440
   macro avg       0.50      0.38      0.43      1440
weighted avg       1.00      0.75      0.86      1440
```

Fig. 3. Classification report of Logistic Regression.

Below is the base models result after pca with their accuracy score and the best models classification report is shown

```
Model name : LogisticRegression()
accuracy score  0.9916666666666667
Model name : KNeighborsClassifier()
accuracy score  1.0
Model name : DecisionTreeClassifier()
accuracy score  0.9979166666666667
Model name : GaussianNB()
accuracy score  0.9493055555555555
```

Fig. 4. Accuracy score of the base models after pca.

```
Train Set Accuracy:99.27045336112559
Test Set Accuracy:99.16666666666667

Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98       341
           1       1.00      0.99      0.99      1099

    accuracy                           0.99      1440
   macro avg       0.98      0.99      0.99      1440
weighted avg       0.99      0.99      0.99      1440
```

Fig. 5. Classification report of Logistic Regression.

IV. Result and Discussion

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. Data points are clustered based on feature similarity. Here we use the elbow plot to find the optimal value of k which is shown below in the figure. Here three is the optimal value of the clusters.

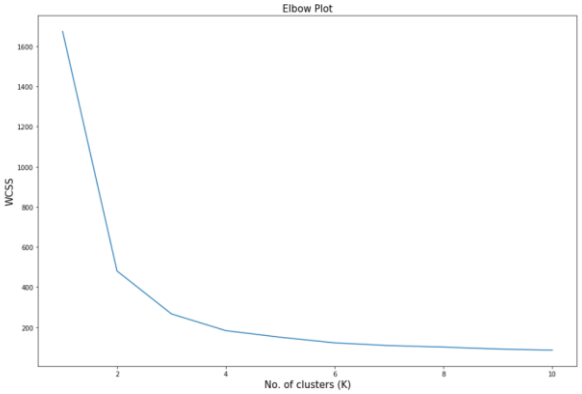

Fig. 6. Elbow plot for optimal number of clusters.

PCA is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. We have used PCA for our dataset and below is the figure of explained variance ratio of the features. Here for one features that explains 99 percent of the data.

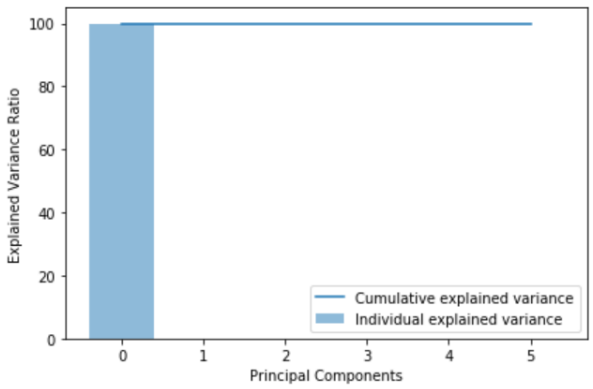

Fig. 7. PCA explained variance ratio.

Popularity based Recommendation system is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. They are primarily used in. As the name suggests Popularity based recommendation system works with the trend. It basically uses the items which are in trend right now. Below figure shows the popular prime_genre with corresponding user_rating and rating_count_tot.

prime_genre	user_rating	rating_count_tot
Photo & Video	3.800860	349
Games	3.685008	3862
Education	3.376380	453
Entertainment	3.246729	535

Fig. 8. Popularity based Recommendation System.

A content based recommender works with data that the user provides, either explicitly (rating) or implicitly. Based on that data, a user profile is generated, which is then used to make suggestions to the user. As the user provides more inputs or takes actions on the recommendations, the engine becomes more and more accurate. Below figure shows top five recommendations for the apps who used the application.

- Top 5 Recommendations for the Apps based on content
- WeChat
- SCRABBLE Premium for iPad
- Google Chrome – The Fast and Secure Web Browser
- Fox News
- Mint: Personal Finance, Budget, Bills & Money

Fig. 9. Content based Recommendation System.

Collaborative filtering (CF) is a technique used by recommender systems. Collaborative filtering has two senses, a narrow one and a more general one. In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person. For example, a collaborative filtering recommendation system for preferences in television programming could make predictions about which television show a user should like given a partial list of that user's tastes (likes or dislikes). these predictions are specific to the user, but use information gleaned from many users. Below figure shows the estimated ratings for the prime_genre.

	id	App_name	est_rating
0	567141387	Jetpack Joyride	4.500000
1	567141387	你我贷理财-P2P理财管家	4.500000
2	567141387	Out There Chronicles - Ep. 1	4.500000
3	567141387	PS Deals+ - Games Price Alerts for PS4, PS3, Vita	4.500000
4	567141387	ESPN Fantasy Football Baseball Basketball Hockey	4.500000

Fig. 10. Collaborative based Recommendation System.

Most recommender systems now use a hybrid approach, combining collaborative filtering, content-based filtering, and other approaches. There is no reason why several different techniques of the same type could not be hybridized. Hybrid approaches can be implemented in several ways: by making content-based and collaborative-based predictions separately and then combining them; by adding content-based capabilities to a collaborative-based approach (and vice versa); or by unifying the approaches into one model for a complete review of recommender systems). Several studies that empirically compare the performance of the hybrid with the pure collaborative and content-based methods and demonstrated that the hybrid methods can provide more accurate recommendations than pure approaches. These methods can also be used to overcome some of the common problems in recommender systems such as cold start and the sparsity problem, as well as the knowledge engineering bottleneck in knowledge-based approaches. Below figure shows the top 5 prime_genres having maximum correlation with Business.

```
track_name
Master For Minecraft Pocket Edition - Ultimate Guide For PE   0.000634
我想有个家                                                    0.000604
Helper for Pokemon Go                                         0.000587
Psych                                                         0.000583
The Krustashians                                              0.000565
Name: "Burn your fat with me!!", dtype: float64
```

Fig. 11. Hybrid Recommendation System.

V. Conclusion

In today's world, uses of machine learning algorithms can play a significant role to understand people's opinion automatically and use that to improve certain product or service efficiently. In this research clustering (k-means clustering), dimensionality reduction (Pca) and recommendation system (popularity based recommendation system, content based recommendation system, collaborative recommendation system, hybrid recommendation system)

has been implemented so it is beneficial for the apps and the users.

We want to improve the algorithms that are used in this research. In the future, the goal is to apply unsupervised machine learning in the model, such as the neural network. We are looking forward to making the decision-making process easier when it comes to selecting apps for the users. We believe that this model will be user-friendly and accurate enough for users to find the best apps which is highly recommended.

REFERENCES

[1] https://ieeexplore.ieee.org/document/9034568
[2] https://ieeexplore.ieee.org/document/8612897
[3] https://ieeexplore.ieee.org/document/9321202
[4] https://ieeexplore.ieee.org/document/5966542
[5] https://ieeexplore.ieee.org/document/8929775
[6] Building Recommender Systems with Book by Frank Kane.
[7] Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries
[8] Machine Learning: Make Your Own Recommender System
[9] Recommender Systems Handbook
[10] https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps