PREDICTIVE ANALYSIS ON PLAY STORE AND APP STORE

Ritika S[1], Mahashruthi KB[2]
[1]- Dr. Velvadivu. P, Faculty, Department of Computing, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India
[2]- Dr. Sathya. C, Faculty, Department of Computing, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India

---***---

ABSTRACT: *In this study, we have analysed the two prominent app stores - Google Play Store and iOS App Store. The main aim of our project is to analyse the data on apps in both app/play store and provide meaningful insights. We have tried to predict the genre of app to the accuracy of App Store apps. To cluster the Play Store apps based on user reviews. Next, we have done Exploratory Data Analysis on Price Distribution for both. Further, we have predicted the price of free apps (App/Play stores) using Regression Model. The final result of this predictive analysis would tell us about how factors that we took into consideration have significant influence on the apps.*

KEYWORDS: Google Play store, iOS App store, Clustering, Classification, Linear Regression, Visualisation, Prediction, Compare, Predictive Analysis, Exploratory Data Analysis.

1. INTRODUCTION:

Google play store [9] and iOS app store are engulfed with a few thousands of similar and new applications. There has always been a debate on which market is best for users and why. App developers also come to a dilemma as to which market should be chosen to release their products. Also, users while downloading the app look for the previous user's review and ratings to know if its useful or not. The success of an app is generally determined by the number of installs and the user ratings. With the already available data having different attributes, we performed analysis and try to predict insights out of it.

2. LITERATURE REVIEW:

There has been a constant growth in the public and private information stored within the internet. They can hence provide important information for product and services refinement. Several studies have been done on this topic using various factors but there were very few when it came to comparing them. But when noticed there were none comparing both the app store and play store data. Either they are focussing on app store completely for analysis or completely on play store. Thus, we compare both the app store and play store data in order make price comparisons.

3. RESEARCH METHODOLOGY:

The data for the analysis was taken from Kaggle for both play store [1] and app store [2]. Data pre-processing was done using various methods. Since the attributes for the both datasets were named different, we first started by filtering the attributes to match both datasets. Since there were many columns that were not required for our analysis, we dropped them and formed new data frame. We started by performing basic descriptive statistics to understand the data. Then classification was done on the iOS App store data using Linear SVC algorithm to train a model in order to predict the genre of an app based on its corresponding description. We also performed clustering analysis on Google play store to form the homogeneous groups based on the user reviews. Then EDA was performed to understand the price distribution on both the platform. With the help of the above results, prediction of price of free apps was done using the Linear Regression model separately and compared to interpret the results.

4. DATA ANALYSIS AND INTERPRETATIONS:

4.1 Classification

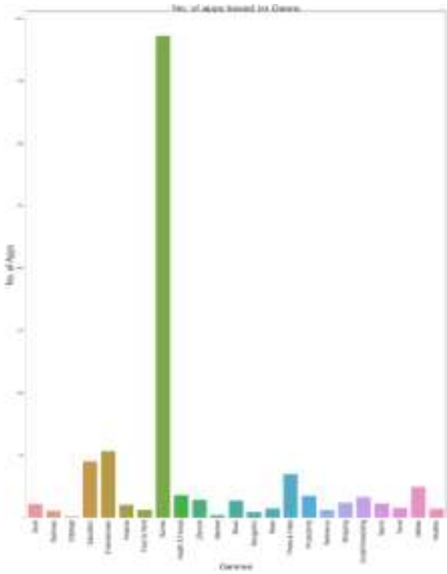

***Chart-**1 Genre distribution plot for App store*

It is clear that gaming genre has contributed significantly higher than almost any other genres. The rest of the genres excluding 'Games' seem to almost lie in a similar range.

We use Linear Support Vector Classifier algorithm [3] for this classification problem.

Next to train the model, as text cannot be used to classify, we find TF-IDF [4] of all the descriptions. The model obtained an accuracy of 81.5%. For testing the model, we take app descriptions from google store and check if the model can predict the genre of app correctly to the accuracy. When done so we found that 6 out of 10 times the model predicted the genre correctly.

```
APP NAME: Instant big profile Dp
CATEGORY: Social Networking
PREDICTED CATEGORY:

array(['Social Networking'], dtype=object)
```

***Fig** .1 Model predicting genre correctly.*

```
APP NAME: Jio Tv Live Cricket Game
CATEGORY: Entertainment
PREDICTED CATEGORY:

array(['Games'], dtype=object)
```

***Fig .2** Model predicting genre incorrectly.*

4.2 Clustering

App	Rating	Installs	Sentiment_Polarity	Sentiment_Subjectivity
7	4.0	500000.0	0.465906	0.493254
16	3.8	10000000.0	0.181294	0.443957
20	4.7	1000000.0	0.318145	0.591098
21	4.6	10000000.0	0.196290	0.557315
30	4.2	10000000.0	0.423659	0.512356

***Table-** .1. Table representing the attributes used for clustering.*

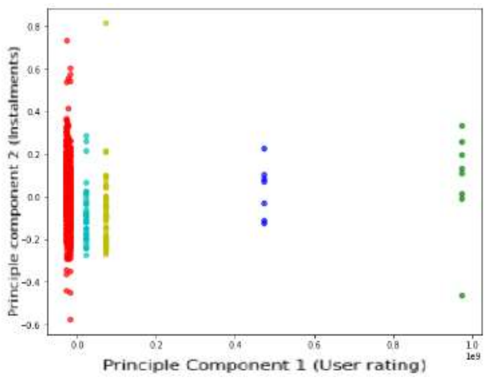

***Chart-**2. Scatter plot representing the optimum number of clusters needed.*

Principal component analysis [5] was used to reduce the dataset size and use to find the number of clusters. Here, we can see that 5 clusters are formed.

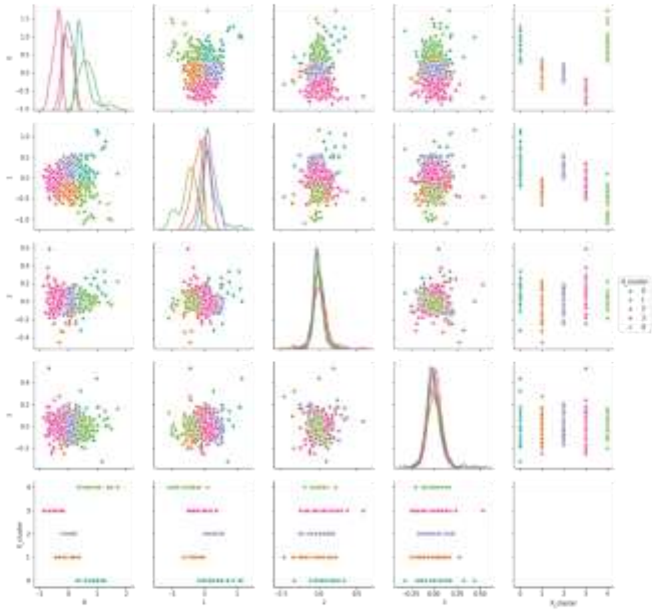

***Chart-**3. Pair plot representing the cluster formation for the attributes taken under four different projections.*

We can infer that the performance on the pipeline was pretty good. The clusters formed were only slightly overlapped and the cluster assignments were much better than random.

4.3 Visualizations on Price Distributions

In order to compare price distribution, from each dataset alike attributes are selected to obtain a accurate comparison.

```
1. Free apps are 4056
2. Counting (outliers) super expensive apps 7
 -  which is around 0.09726274836737529 % of the total Apps
 Thus we will dropping the following apps from Apple Store
```

	track_name	price	prime_genre	user_rating
115	Proloquo2Go - Symbol-based AAC	249.99	Education	4.0
162	NAVIGON Europe	74.99	Navigation	3.5
1136	Articulation Station Pro	59.99	Education	4.5
1479	LAMP Words For Life	299.99	Education	4.0
2181	Articulation Test Center Pro	59.99	Education	4.5
2568	KNFB Reader	99.99	Productivity	4.5
3238	FineScanner Pro - PDF Document Scanner App + OCR	59.99	Business	4.0

***Table -2**. Table representing iOS app store outliers*

```
1. Free apps are 8719
2. Counting (outliers) super expensive apps 15
 -  which is around 0.1601537475976938 % of the total Apps
 Thus we will dropping the following apps from Google Store
```

	App	Price	Category	Rating
4197	most expensive app (H)	399.99	Games	4.3
4362	💎 I'm rich	399.99	Lifestyle	3.8
4367	I'm Rich - Trump Edition	400.00	Lifestyle	3.6
5351	I am rich	399.99	Lifestyle	3.8
5354	I am Rich Plus	399.99	Games	4.0
5355	I am rich VIP	299.99	Lifestyle	3.8
5356	I Am Rich Premium	399.99	Finance	4.1
5357	I am extremely Rich	379.99	Lifestyle	2.9
5358	I am Rich!	399.99	Finance	3.8

***Table -**3 Table representing Google play store.*

Table 2 and Table 3 represents the outliers that are removed for getting an accurate model.

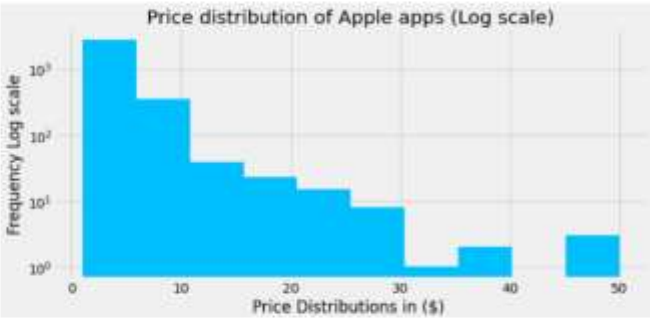

***Chart -4** Histogram representing the price distribution of iOS app store in dollars.*

***Chart-** 5 Histogram representing the price distribution of Google play store in dollars.*

From chart 4 and 5 we can say that the count of paid apps exponentially decreases as the price increases. Very few apps have been priced above $30 for Apple and $20 for Google.

	free	paid	total	paid_percent	free_percent
Education	132	321	453	70.860927	29.139073
Entertainment	334	201	535	37.570093	62.429907
Games	2257	1605	3862	41.558778	58.441222
Others	1166	832	1998	41.641642	58.358358
Photo & Video	167	182	349	52.148997	47.851003

***Table -4** iOS App store*

	free	paid	total	paid_percent	free_percent
Games	2605	239	2844	8.403657	91.596343
Lifestyle	821	89	910	9.780220	90.219780
Others	3983	216	4199	5.144082	94.855918
Productivity	639	40	679	5.891016	94.108984
Utilities	671	63	734	8.583106	91.416894

***Table -**5 Google Play store*

Table 4 and 5 show the summary of the entire reduced dataset.

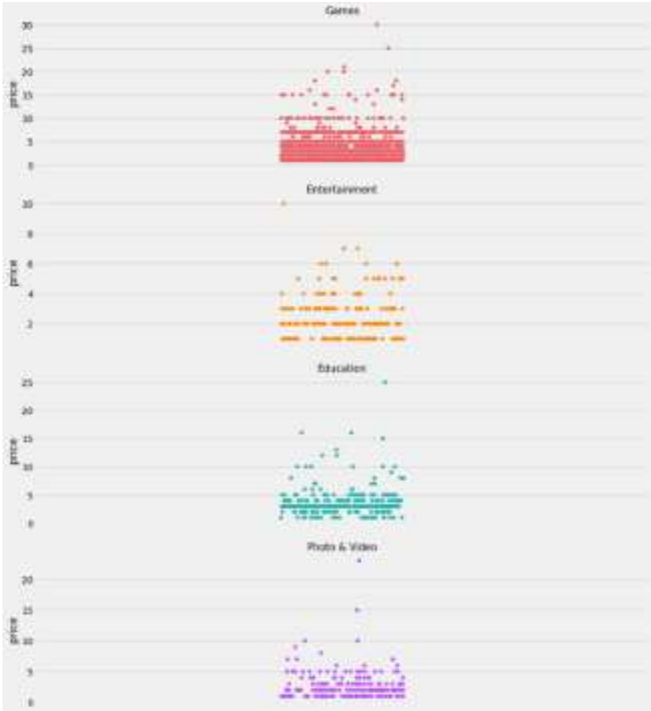

***Chart- 6**. Strip plot representing the price distribution affected by category for App store.*

We can see that the paid gaming apps are highly priced and distribution extends till $20 and paid Photo & Video apps have a lower price range.

Chart-*7 Strip plot representing the price distribution affected by category for Play store.*

We can see that the paid gaming apps are highly priced and distribution extends till $25 and paid Entertainment apps have aa lower price range.

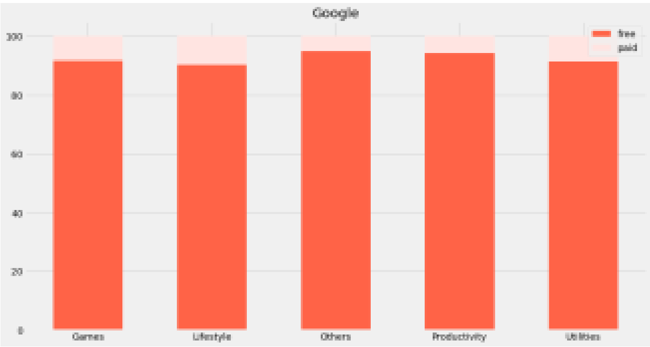

Chart-*8 Paid and Free app distribution in Play store.*

We can infer that the Games category has a significant % of paid apps. On the contrary, Lifestyle category hosts high % of free apps.

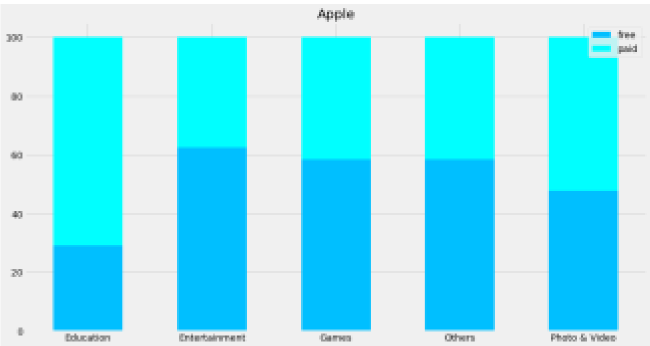

Chart- *9 Paid and Free app distribution in App store.*

We can infer that the Education category has significant % of paid apps. On the contrary, Entertainment and Games category hosts high % of free apps.

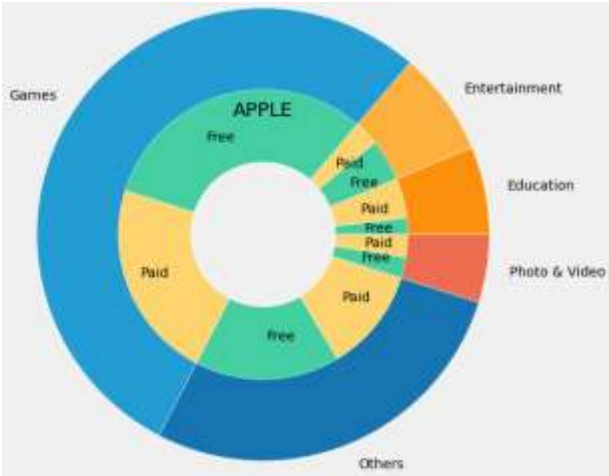

Chart-10 *Pie chart representing entire iOS app distribution.*

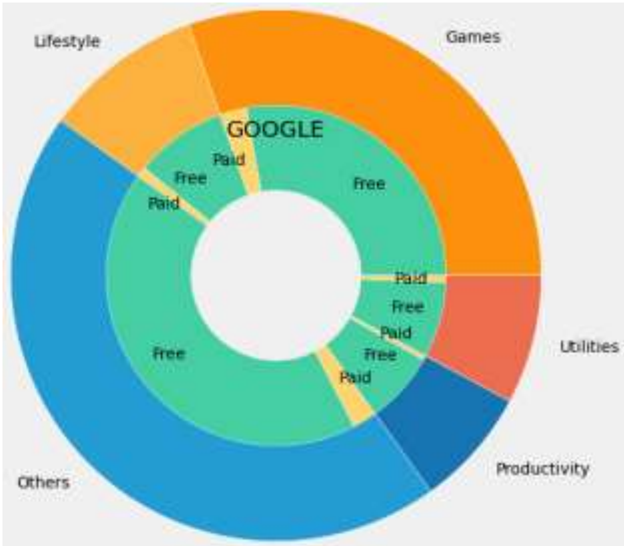

Chart-11 *Pie chart representing entire Google app distribution.*

It is clearly visible that the number of paid apps in Apple store is more than in Google store.

4.4 Linear Regression

We train a Linear Regression model by training the prices of paid apps and test the model by predicting the price of free apps for both the datasets.
After splitting the paid and free apps in the dataset, paid apps are used for training which are then tested with free app data. Predictions of the prices of free apps are obtained as shown in fig 3 and fig 4.

```
Predictions:  [3.21114487 2.67751486 1.42156336 ... 3.86647551 2.88439457 4.14470355]
```

fig- 3 *Price Prediction of free apps in App store*

```
Predictions for Price test :  [29.56945321 26.08833163 18.38885538 ... 10.1176386З 21.33755967
  9.35827372]
```

fig- 4 Price Prediction of free apps in Play store

We obtain a regression score[6] of 0 for both datasets (Regression score of 0 tells us that the predictions made are accurate while that of 1 tells us that the predictions made are perfect).

In order check validation of Regression Model made, we use the residual plot [7].

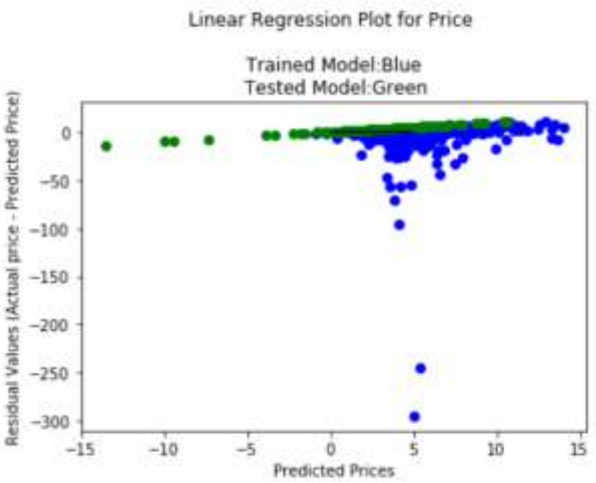

***Chart-** 12 Residual Plot for App store.*

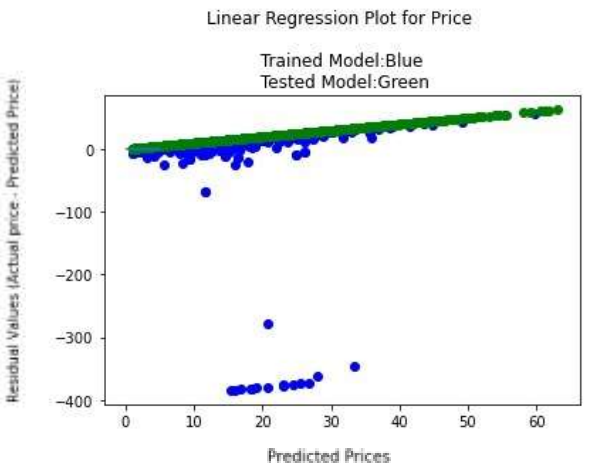

***Chart-** 13 Residual Plot for Play store.*

In plot chart 12 and 14 as we can see both the graphs have high density of points close to the origin and low density of points away from origin. Thus, we can conclude that it is a good residual plot, which in turn concludes the model to be a desirable one.

5. CONCLUSION:

- Successfully trained and tested classifiers based on genre in the app.
- Grouped apps based on user reviews using clusters.
- Built Regression models for price prediction

REFERENCES:

[1] Apple App store:
https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps#AppleStore.csv

[2] Google Play store:
https://www.kaggle.com/lava18/google-play-store-apps#googleplaystore.csv

[3] The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations. Let's get started.
https://pythonprogramming.net/linear-svc-example-scikit-learn-svm-python/#:~:text=The%20objective%20of%20a%20Linear,the%20%22predicted%22%20class%20is

[4] TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

[5] Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[6] How to evaluate regression models? -by Vimarsh Karbhari
https://medium.com/acing-ai/how-to-evaluate-regression-models-d183b4f5853d

[7]https://www.statology.org/residual-plot-python/#:~:text=A%20residual%20plot%20is%20a,check%20for%20heteroscedasticity%20of%20residuals

[8]https://www.irjet.net/archives/V7/i12/IRJET-V7I1248.pdf

[9]http://www.ru.ac.bd/stat/wp-content/uploads/sites/25/2020/03/ICDSSDG_COR-2019_paper_98F.pdf

[10]https://www.researchgate.net/publication/305728429_App_Store_Analysis_Using_Regression_Model_for_App_Downloads_Prediction