**GOOGLE PLAY STORE**



**Team:**

**Kshama Joshi**

**Menitha Magani**

**Ramya Shetty**

***CONTENTS***

1) ABSTRACT

* Problem
* Methods Used
* Results Obtained

2) MATERIALS AND METHODS

* Dataset Description
* Tools and Techniques

3) RESULTS AND DISCUSSIONS

4) CONCLUSION

***ABSTRACT***

PROBLEM:

Our aim is to analyze and predict the rating of an application in google play store.

METHOD USED :

* Multiple Linear Regression
* Stepwise Regression

In order to improve the performance of model we performed stepwise regression by removing Installs and Content Rating

* Random Forest

Based on performance and error, Random Forest is the best suitable model for our data set

RESULTS OBTAINED :

* Multiple Linear Regression

R square value – 0.9839

* Stepwise Regression

R square value – 0.984

* Random Forest

R square value – 0.9854

***MATERIALS AND METHODS***

DATASET DESCRIPTION :

* Source – Kaggle
* Data Dictionary

Size – 10841 rows and 13 columns

Variables:

- App (Application name)

- Category (Category to which the app belongs)

- Rating (Overall user rating of the app)

- Reviews (Number of user reviews for the app)

- Size (Size of the app)

- Installs (Number of user installs/downloads for the app)

- Type (If the app is paid or free)

- Price (Price of the app)

- Content Rating (Age group the app is targeted at )

- Genres (The genre to which the app belongs)

- Last Updated (Date when app was last updated on Play Store)

- Current Ver (Current version of app available on Play Store)

- Android Ver (Minimum required Android version)

TOOLS AND TECHNIQUES :

* Python
* Tableau

***RESULTS AND DISCUSSIONS***

We decided to apply regression on our data set so we checked for Linearity and Independence conditions. Both the assumptions were satisfied by the data so Multiple Linear Regression was applied. The R square value of the model was 0.9839.

We wanted to check if the model can improve by applying stepwise regression. At first, installs variable was removed based on its p-value and our model resulted with an R square value of 0.984. We went ahead and removed Content rating which returned the same R square value of 0.984. So, we stopped Stepwise Regression at this step.

Based on these outputs, the other two LINE assumptions, that is, Normality and Equal Variance were checked. We notice that the data was left skewed and not normally distributed. Hence, the LINE assumptions were violated. In order to normalize the data, we tried transforming it by applying log, square and cube transformations. The data remained skewed after trying all these transformations.

The data has to be normally distributed to apply regression. Therefore, we apply Random Forest on our data which results in an R square value of 0.9854 (98.54%).

Hence, Random Forest is the best suitable model for the Google Play Store dataset to predict the rating of an application.

***CONCLUSION***

* Our model has data that belongs majorly to only a few categories. So, our model is trained to perform well if the newly created application belongs to one of these categories. But what if someone would want to create an application for a category that has very few rows/very less information so there are chances our model might predict a false rating. So, the future work can be gathering more data in order to train our model to give accurate rating predictions.
* Business objective accomplished by applying Random Forest model on our data to predict the rating of an application on Google Play Store. It will benefit those people/organization who wish to create a new android application by giving them an insight as to whether or not it is going to be successful based on various factors.