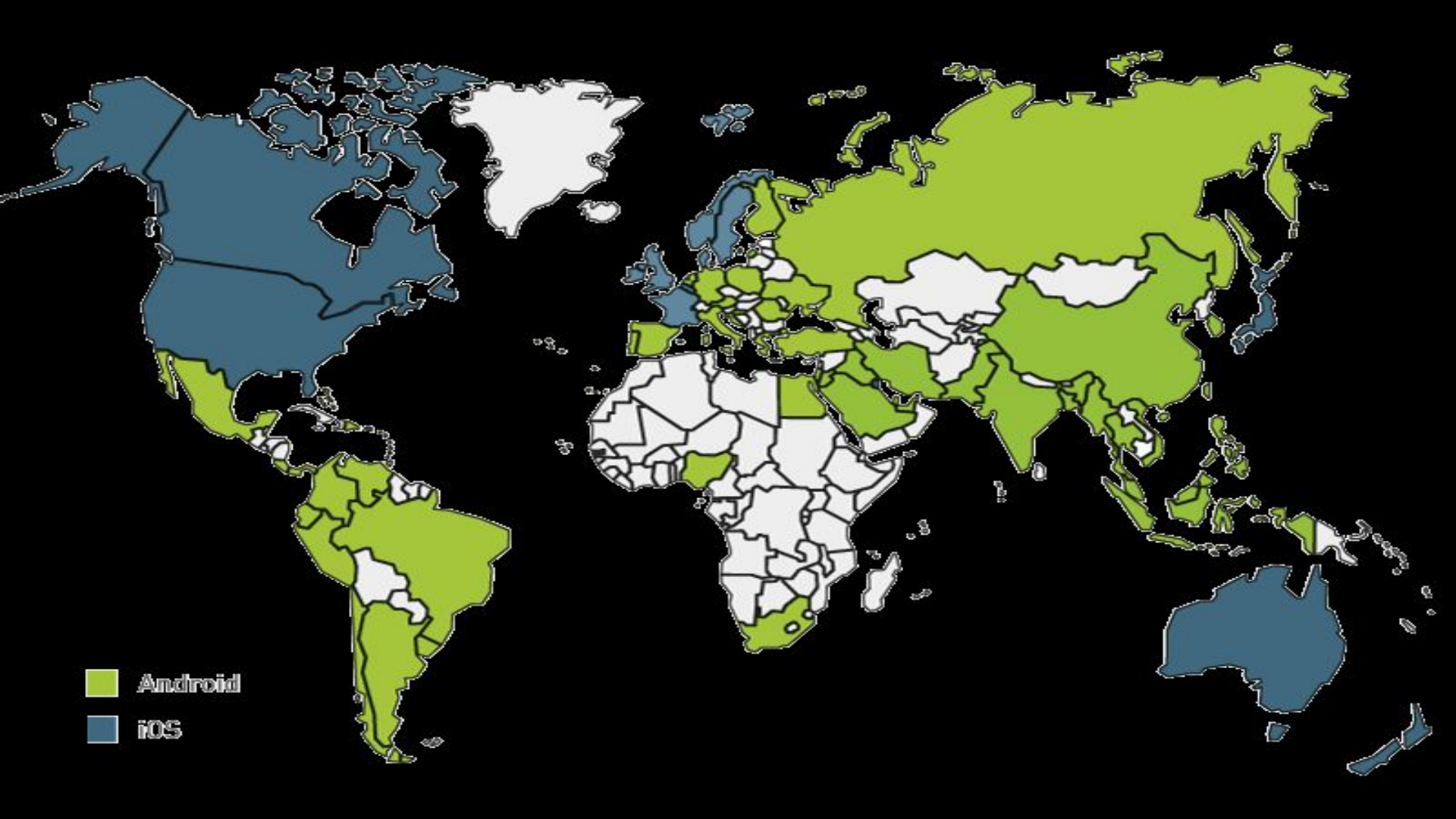
ANALYSIS OF APPLE APP STORE DATA

Dithya Sridharan

PROJECT OBJECTIVE

This data set contains more than 7000 Apple iOS mobile application details. The data was extracted from the iTunes Search API at the Apple Inc website. Android holds about 54.2% of the smartphone market, while iOS is 43% as of the year 2017. This data set contains key information to analyse top trending apps in iOS Apple Store. This project is determined to analyse the existing strategy to drive growth and retention of a future user.


Android
iOS

INSPIRATION FOR THE PROBLEM

Most Downloaded

Apple gives credits to apps with the highest number of downloads each year. We have assumed total rating is a rough indicator of number of downloads(as an app will be rated only after installing yet.)

Most Popular

Apple gives credits to apps with most user ratings. user_rating gives average user rating value for all versions.

Most Revenue

Apple gives credits to highest revenue generating apps on iOS. Revenue of each app is calculated using (no of downloads * price of app)

UNDERSTANDING THE DATA

1. "id" : App ID
2. "track_name": App Name
3. "size_bytes": Size (in Bytes)
4. "currency": Currency Type
5. "price": Price amount
6. "rating_count_tot": User Rating counts (for all version)
7. "rating_count_ver": User Rating counts (for current version)
8. "user_rating" : Average User Rating value (for all version)
9. "user_rating_ver": Average User Rating value (for current version)
10. "ver" : Latest version code
11. "cont_rating": Content Rating
12. "prime_genre": Primary Genre
13. "sup_devices.num": Number of supporting devices
14. "ipadSc_urls.num": Number of screenshots showed for display
15. "lang.num": Number of supported languages
16. "vpp_lic": Vpp Device Based Licensing Enabled

EDA TARGET ANALYSIS

EDA analysis is done by identifying correlation among the predictors. It identifies the relation between the target variables and the independent variables. The target variables :

1) Top 10 downloaded apps

2) Top 10 rated apps

3) Top 10 apps with most revenue

Data Preprocessing

Missing Values

The data set is found to have many observations with missing values. Usually missing values should be replaced by mean of the feature in order to not lose relevant info. But upon, display it was seen that the observations with missing values have missing values for all features indication that there is no information from those observations. So we have omitted those observations.

Data Transformation

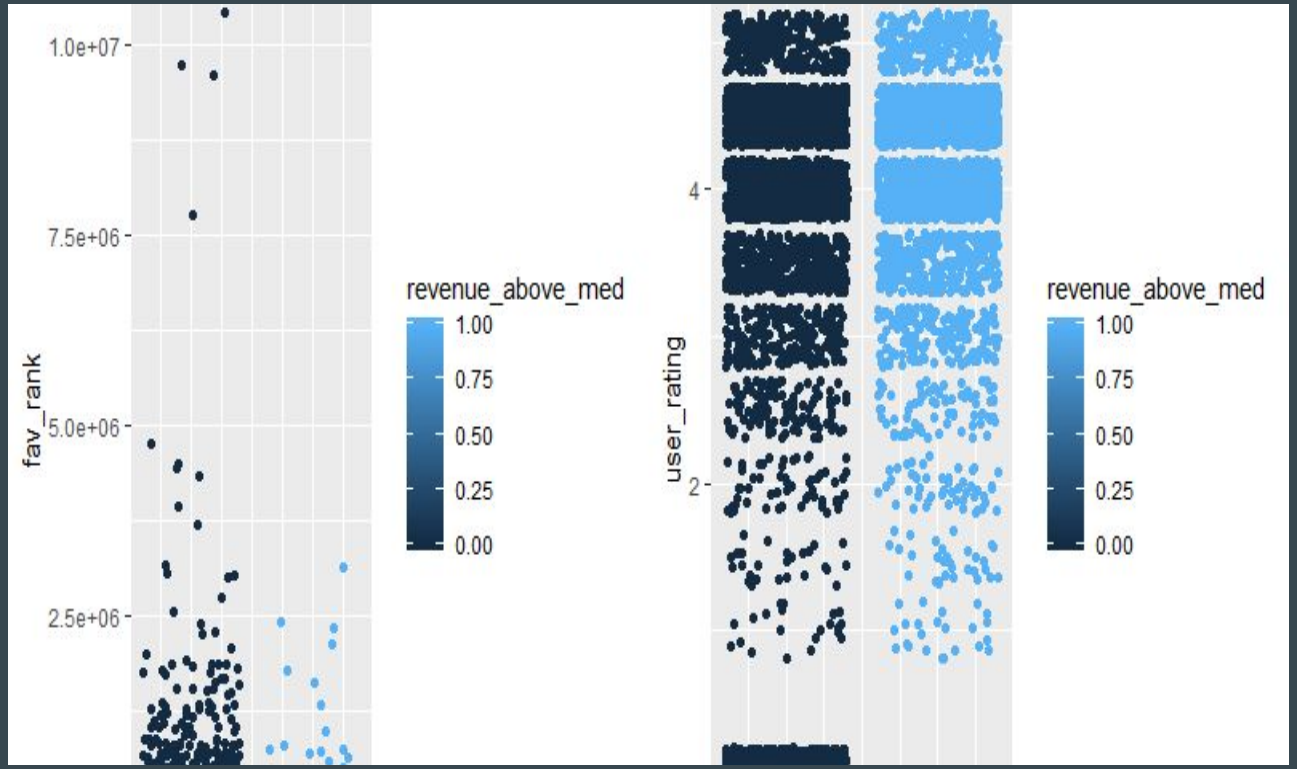
We have mutated two response variables -

1) Revenue - revenue generated by the apps(calculated as no of downloads * price)

2) Fav rank - Finding favourite rank of each app(calculated as no of downloads * user rating)

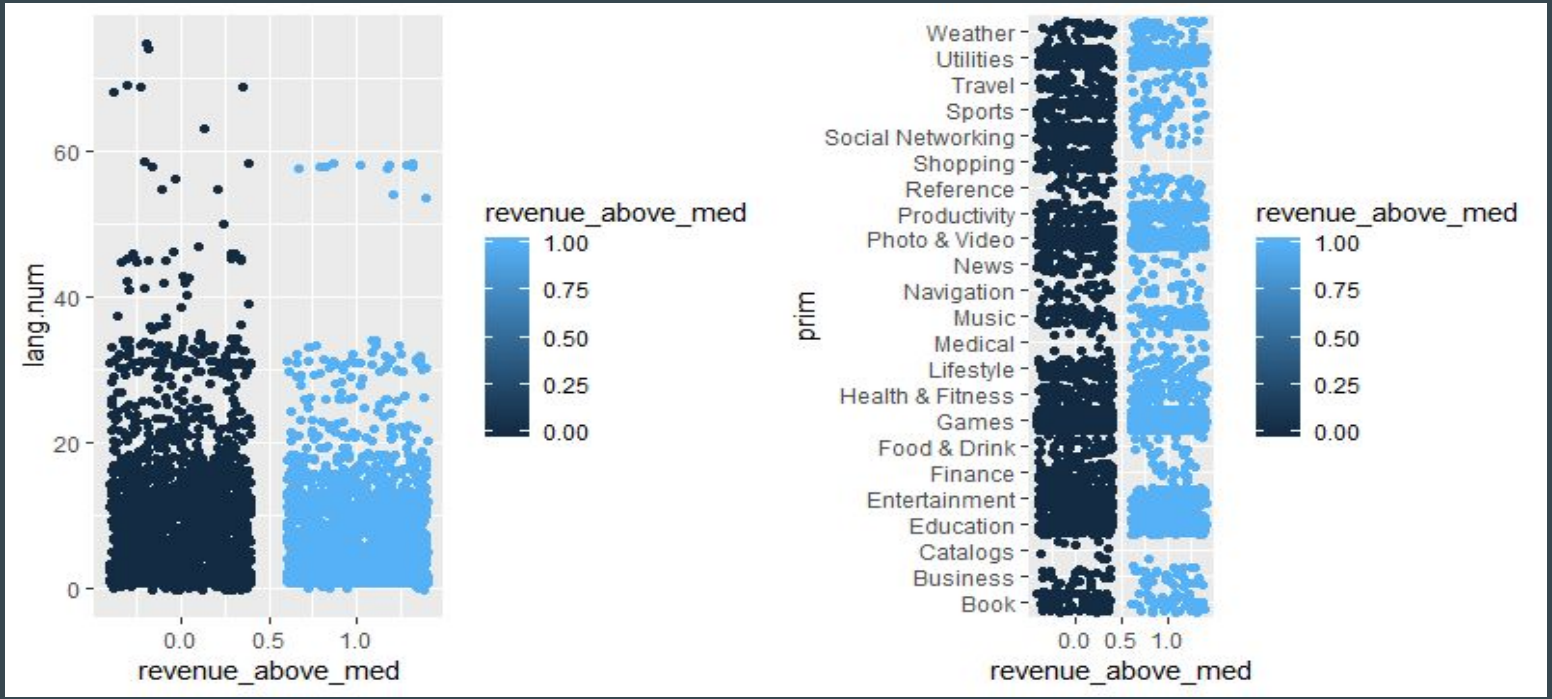
3) As we know that , we need categorical values for Boosting: Hence, we transformed the response variable(user_rating) to high and low category by mean.

EDA ANALYSIS



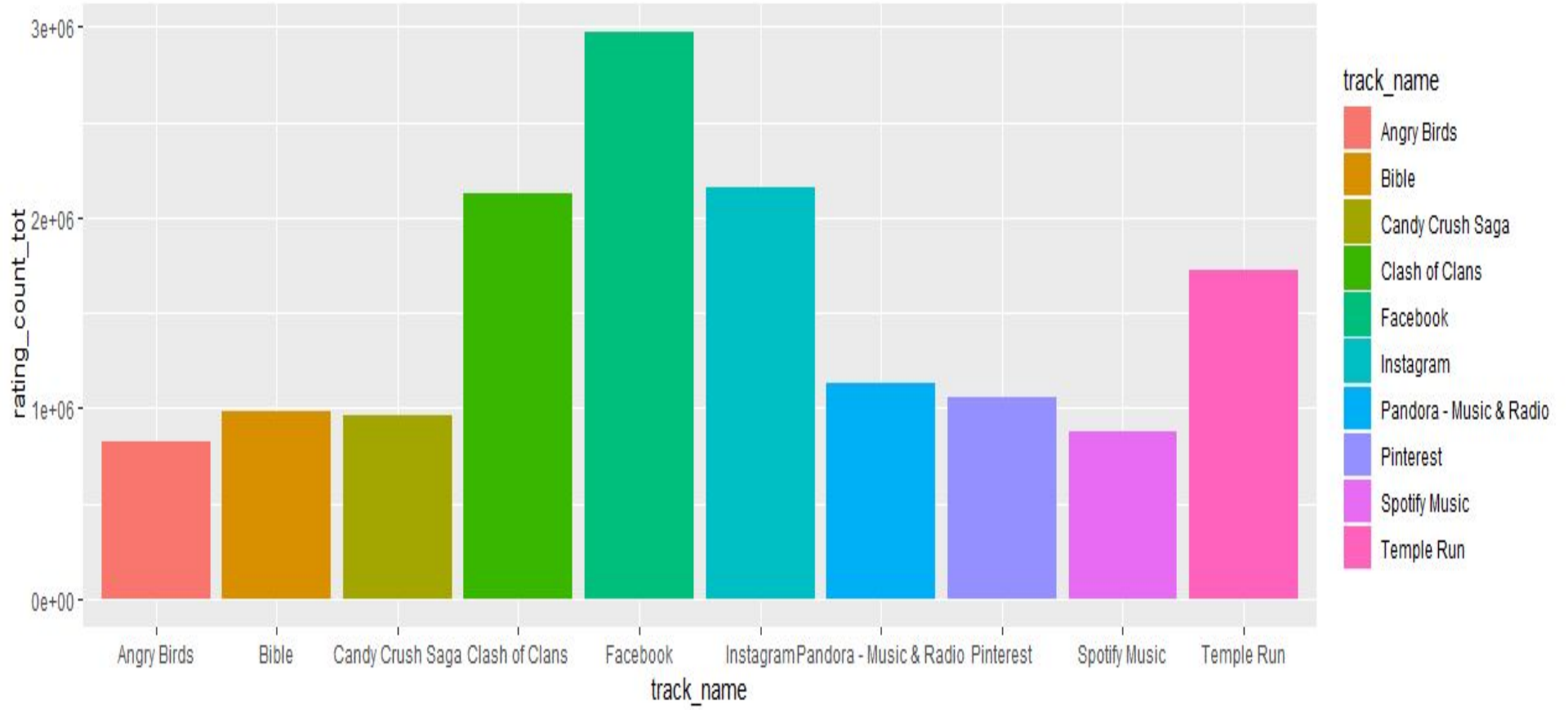
1 - In this graph, those that do good business(revenue>median) might not necessarily have a favourite rank.This is because some of the top ranked apps are free

2 - In this graph we can see that apps that make revenue and apps that do not have similar user ratings. So user rating does not necessarily depend on price of the app specified

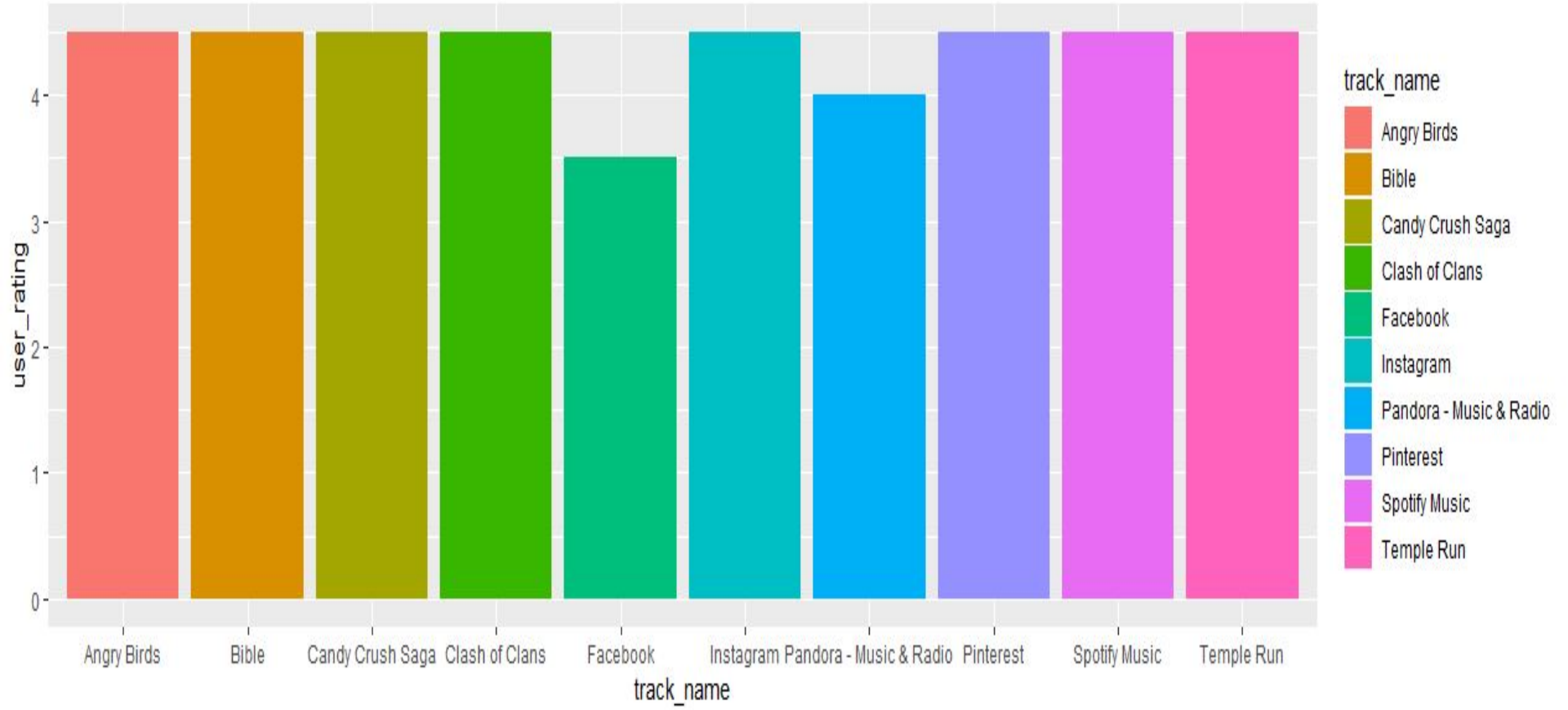


- In graph 1, we can see that apps that revenue made does not depend on the number of languages supported
- In graph 2, we can see that apps that made high revenue fall under education genre

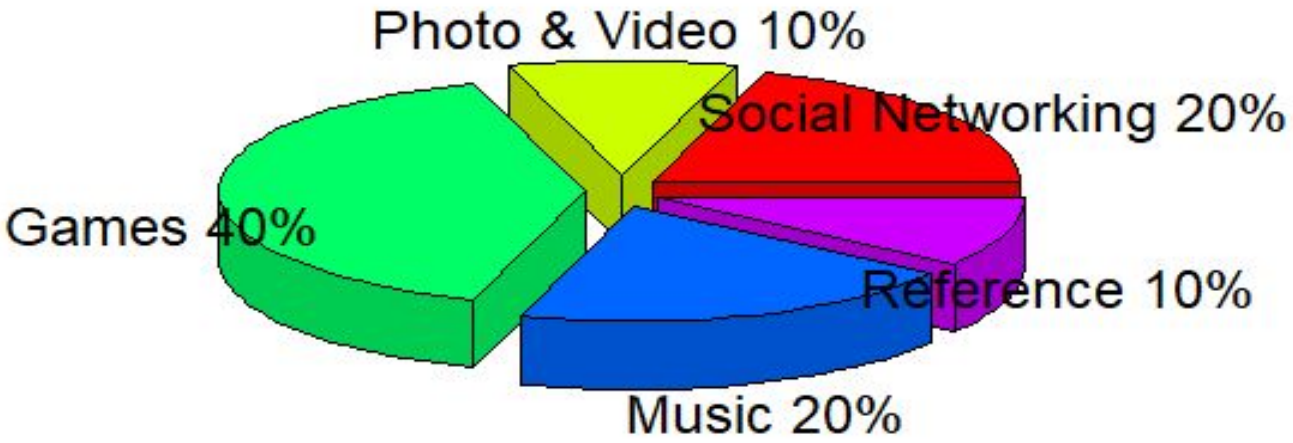
TEN MOST DOWNLOADED APPS FOR FURTHER ANALYSIS



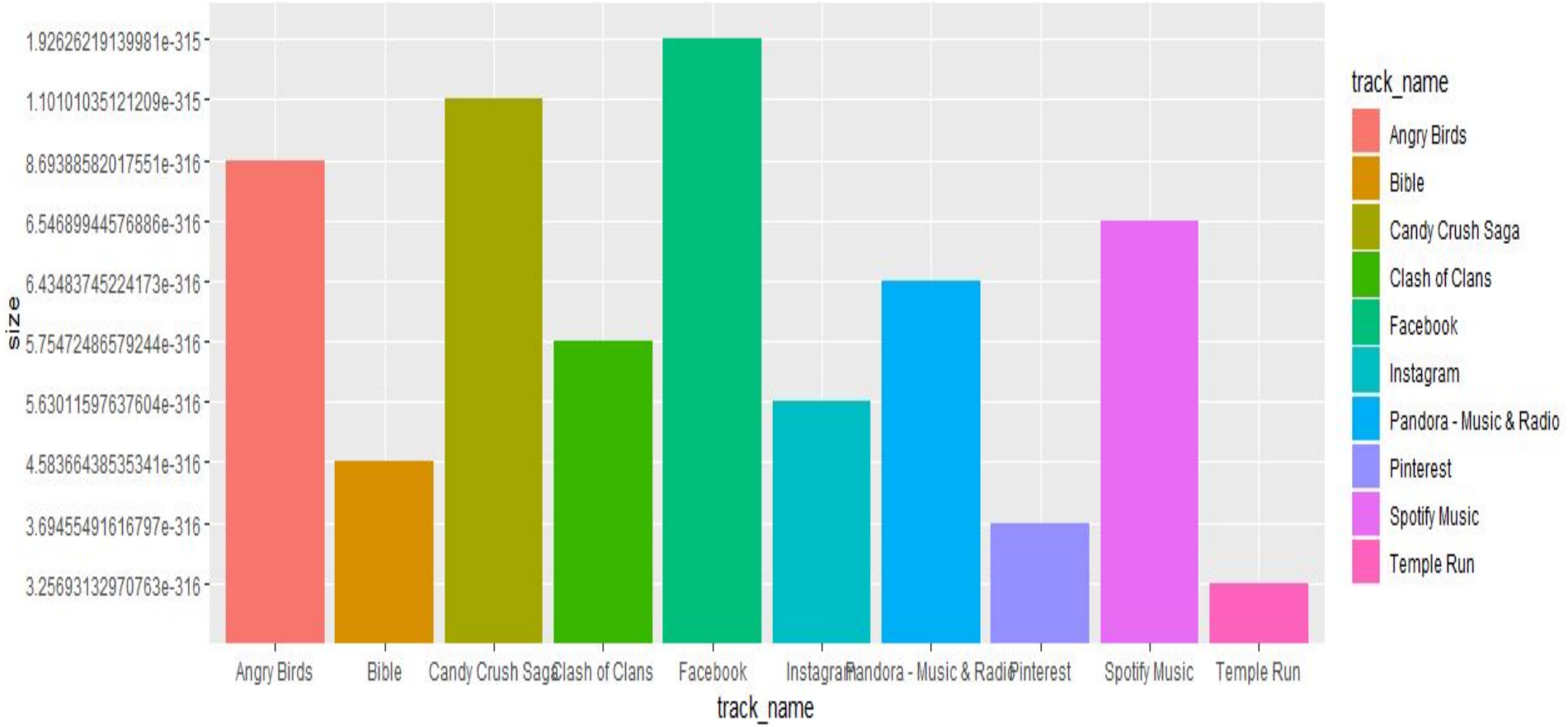
TEN BEST RATED APPS



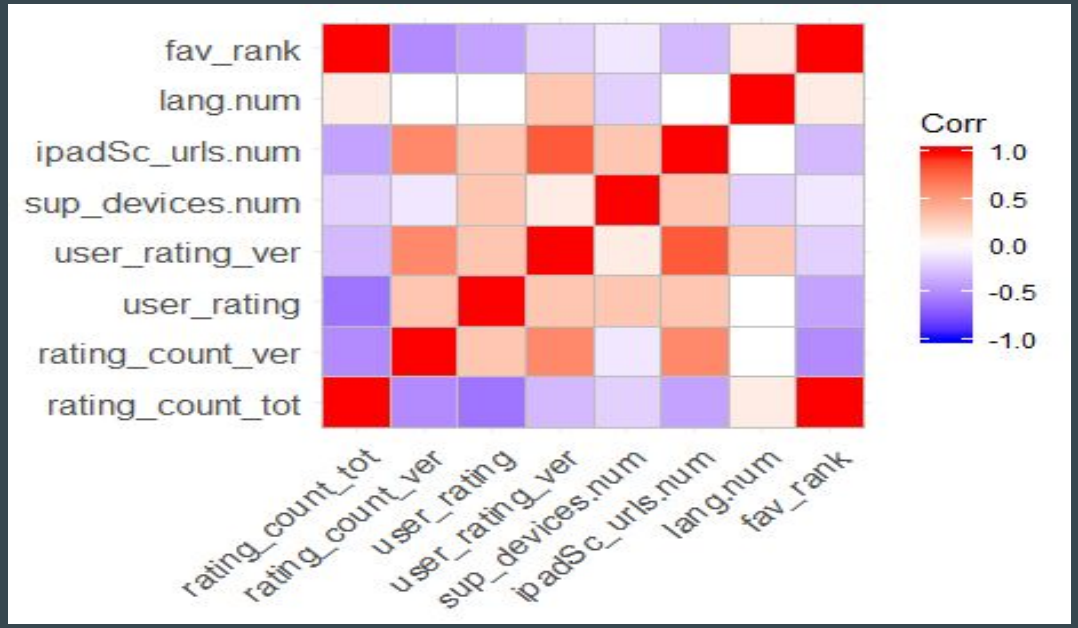
PIE CHART OF TOP TEN RATINGS AGAINST PRIMARY GENRE



SIZE COMPARISON FOR TEN BEST RATED APPS

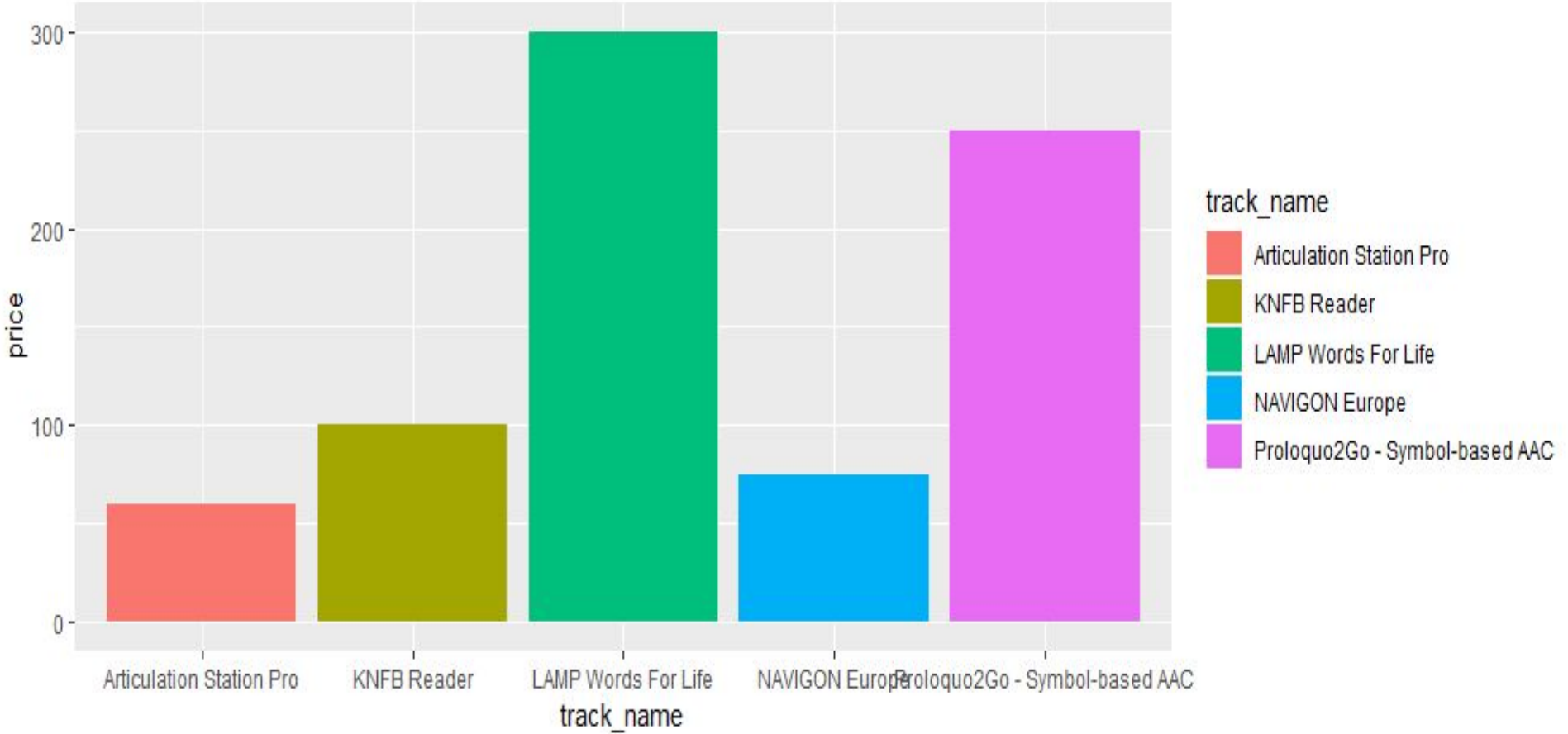


CORRELATION MATRIX FOR PREDICTORS FOR TEN BEST RATED APPS



This graph gives the correlation of feature of top 10 apps. We can see that the following features are correlated to user ratings:
1)No of screen shots available
2)No of supporting devices
3)Ratings of current version of app
4)Total count of current version downloaded
5)We see that no of languages app supports does not relate to user rating
6)Also, all the top apps downloaded are free

TOP 5 APPS THAT CARRY A PRICE



Prediction and Analysis

- For regression analysis, we have taken user rating as the response variable. This is based on the assumption that more the user rating, more number of downloads the app will incur in the future. This is based on the EDA Analysis.
- We will be doing the following :
- Feature selection using XgBoost and Random forests
- Model fitting and prediction using Bagging, Boosting, Logistic Regression and Neural Networks
- We will be highlighting the challenges faced and the future scope of our prediction.

FEATURE SELECTION

We had 17 attributes. Out of which we did not consider attributes likes app_id, app_name as they will not contribute much to the prediction.

For further feature selection, we used:

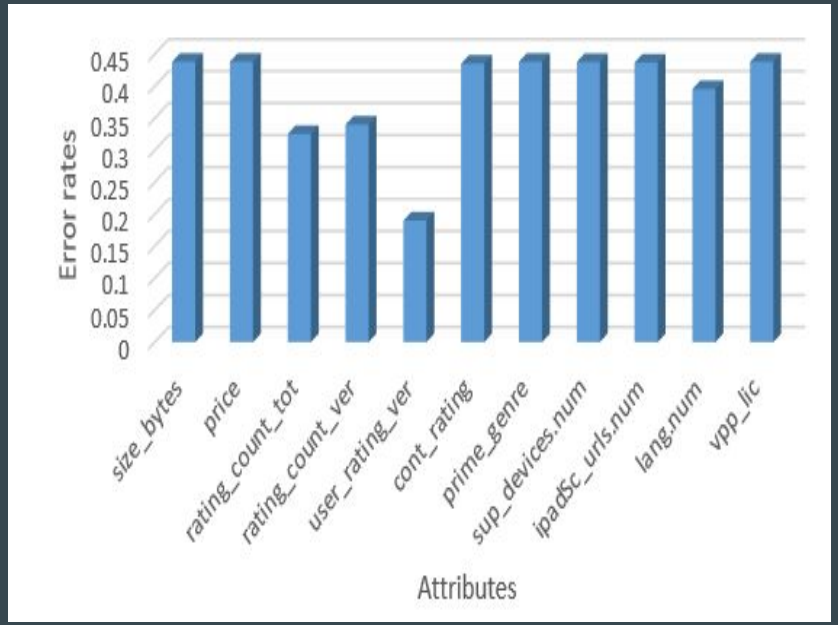
- Random Forest
- XGBoost

Why XGBoost for Feature Selection?

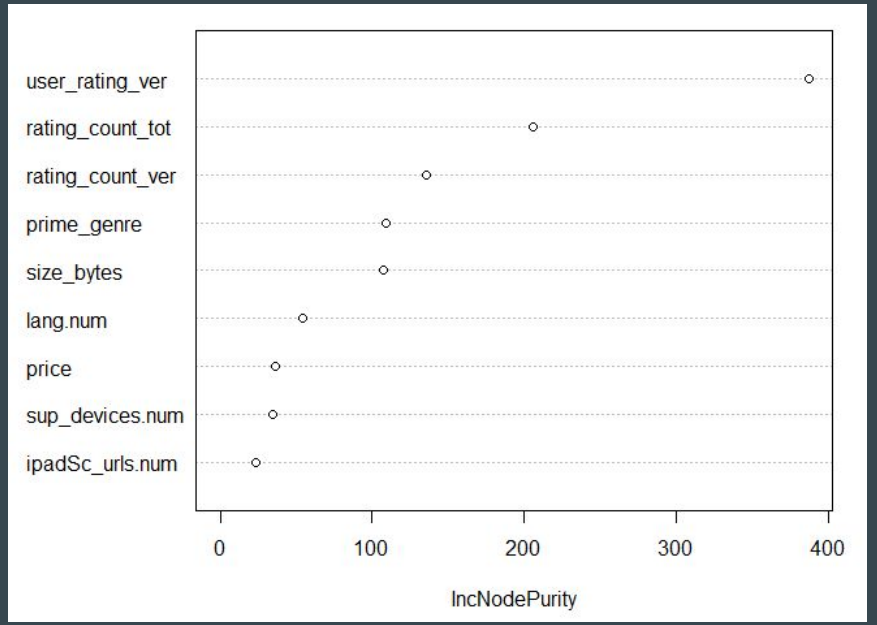
- XGBoost finds the correlation of each of the attribute with the response variable individually, and gives the error.
- The attribute with the minimum error with the response have higher importance.
- This importance is calculated explicitly for each attribute in the dataset, allowing attributes to be ranked and compared to each other.
- It can then use a threshold to decide which features to select.

Feature Selection

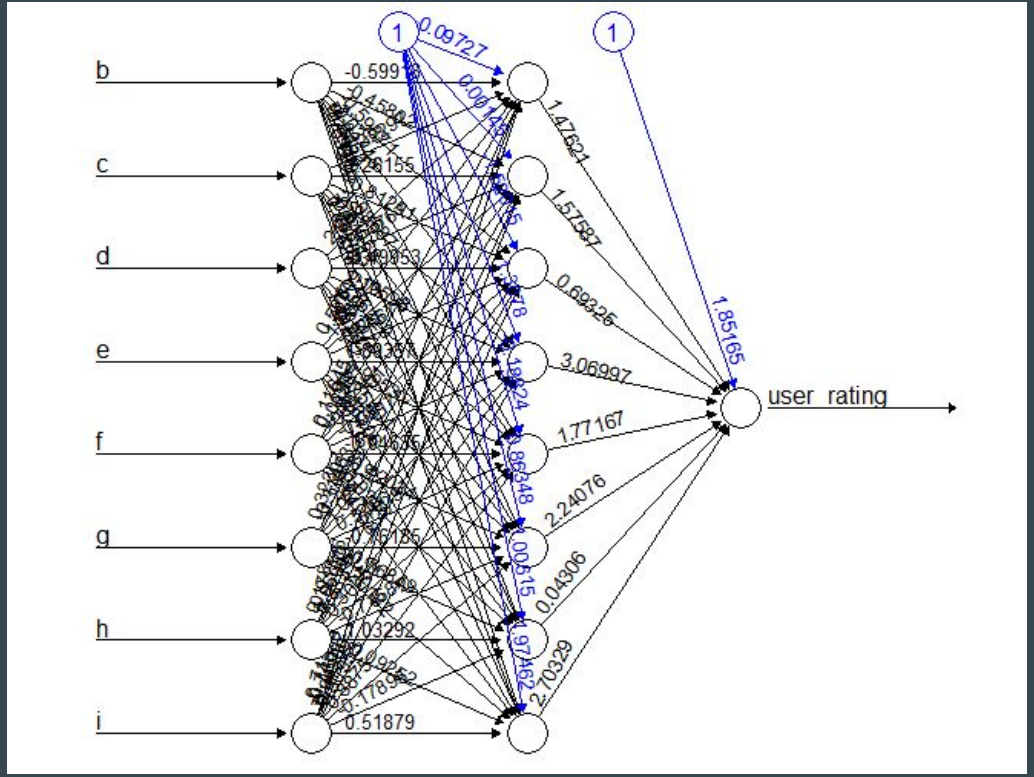
XGBoost



Random Forest



Neural Networks



Accuracy: 73.2%

Accuracy

Boosting	76.7%
Neural Networks	73.2%
Bagging	72%
Logistic Regression	61%

Conclusion

- EDA results and Prediction results agree.
- We concluded some important factors that influence Apple App store business :
  - Apps that do good business(revenue>median) might not necessarily have a favourite rank.This is because some of the top ranked apps are free
  - We can see that apps that make revenue do not have similar user ratings. So user rating does not necessarily depend on price of the app specified
  - We can see that apps that revenue made does not depend on the number of languages supported
  - We can see that apps that made high revenue fall under education genre
  - None of the top 10 downloaded apps have any revenue or price associated
  - Games apps are the most popular genre followed by social networking apps
  - Most of the gaming apps have a heavy size to download but are still most popularly downloaded apps
  - Overall, important features that Apple App store should concentrate on is : No of screenshots available for the app, No of supporting devices, ratings of the current version, size of the app and no of languages supported.