AppoMania

—Dhwani
—Naresh
—Sandip
—Sirisha

INTRODUCTION

- Analysis of Google Play Store and iOS App Store datasets

- Comparison of apps between these two markets

Datasets

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

https://www.kaggle.com/lava18/google-play-store-apps#googleplaystore.csv

Datasets

Google

- App
- Category
- Rating
- Reviews
- Size
- Installs
- Type
- Price
- Content Rating
- Genres
- Last Updated
- Current Ver
- Android Ver
- Translated Review
- Sentiment
- Sentiment Polarity
- Sentiment Subjectivity

Apple

- id
- track_name
- size_bytes
- currency
- price
- rating_count_tot
- rating_count_ver
- user_rating
- user_rating_ver
- ver
- cont_rating
- prime_genre
- sup_devices.num
- ipadSc_urls.num
- lang.num
- vpp_lic
- track_name
- app_desc

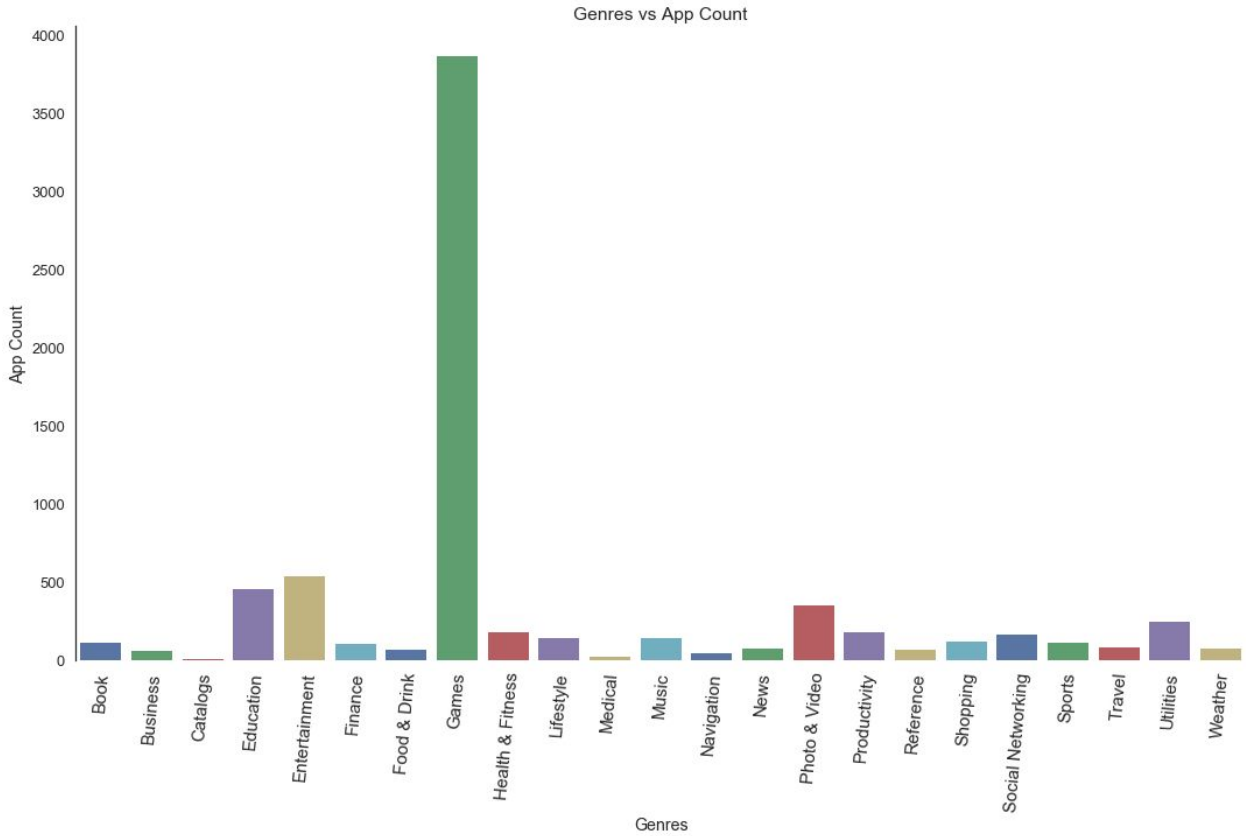
Dataset Infographics

Classifying apps based on the description - Apple

- Purpose
  - Aid the developers and App Stores to correctly classify the Apps based on the description.
  - Also to recommend a genre for a new App.

- Columns Used
  - App Description - An elaborate description of the app given by the App developers.
  - Category - Actual category/genre of the app.

Classification(Genre) - Distribution



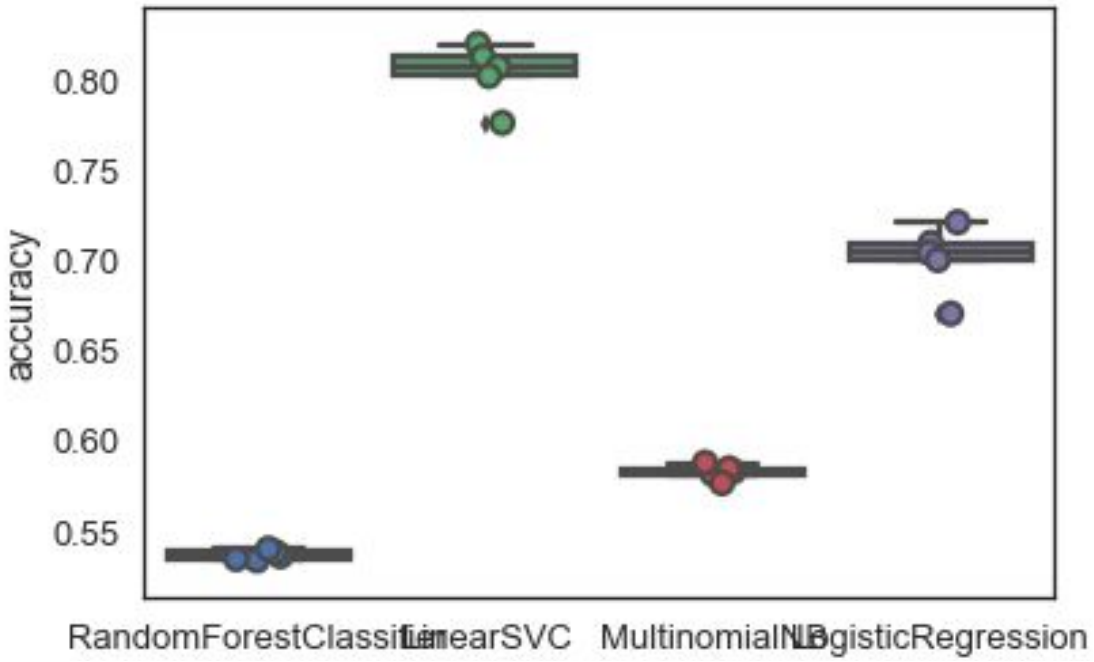
Classification(Genre) - Feature Engineering

- Description
  - It is a text column.
  - Need a way to weigh each words in the description appropriately.

- tf-idf Vectorization
  - term frequency–inverse document frequency
  - It is a measure used to evaluate the importance of a word in a text/document.
  - Converts strings to features based on its importance.
  - Used in Text mining and Searching. (Also in EECS 767)

Classification(Genre) - Result

Model Name	Prediction Rate
Linear SVC	80.3460
Logistic Regression	70.1049
Multinomial NB	58.2739
Random Forest Classifier	53.6624



Prediction Rate using Cross Validation

Classification(Genre) - Result Cntd

- Generated Data Set
  - Randomly selected 31 recently released apps from Google Store.
  - Retrieved its app description and corresponding genre.
  - Predicted the genre of the apps using the trained Linear SVC model.
  - Compared the results between the actual vs predicted.
- Results
  - Prediction rate was 80.64%.
  - Model correctly predicted 25 out of 31.

Classification(Genre) - Result cntd

Incorrectly Classified

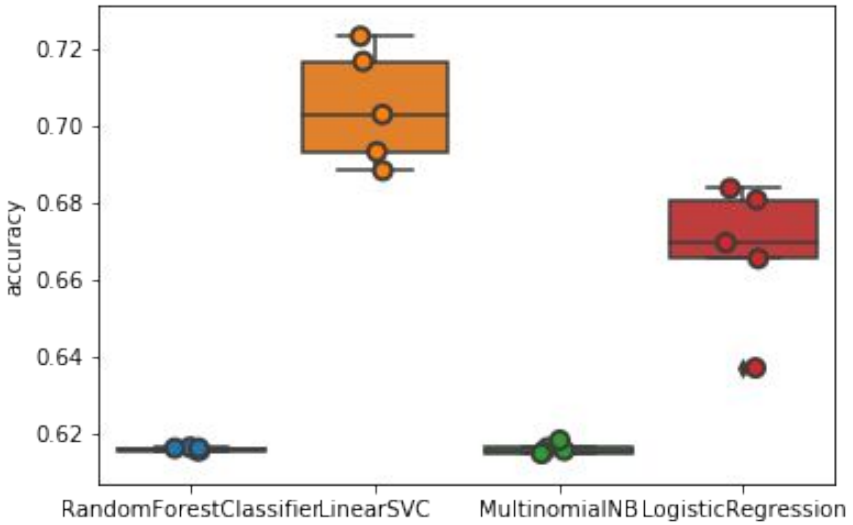
App name	App Description	Actual Category	Predicted Category
City Coach Bus Simulator 2019	City Coach Bus Simulator 2019 is interesting, ...	Travel	Games
Video Status and Quotes (videostatus.net)	Short 30 Seconds videos that you can download ...	Entertainment	Photo & Video
Premium Stickers For WhatsApp	Do you also want to share some cool stickers w...	Lifestyle	Social Networking
Hey Duggee: The Squirrel Club	The Squirrel Club app enables fans to create t...	Entertainment	Education
Jio Tv Live Cricket Game	Jio TV Live Finger Cricket Game- IPL is just a...	Entertainment	Games
iTranslator - Smart Translator - Voice & Tex	iTranslator - Voice & Text Smart Translate is ...	Productivity	Utilities

Classification(Content Rating)

- Purpose
  - Aid the developers and App Stores to correctly label the Apps for Content Rating.
  - Also to recommend Content Rating for a new App.
- Columns Used
  - App Description - An elaborate description of the app given by the App developers.
  - Category - Category/genre of the app.
  - Content Rating- Content Rating of the App. For Ex: 4+, 12+,19+
- Feature Engineering
  - Tf-idf Vectorization and Label Encoding

Classification(Content Rating) - Result

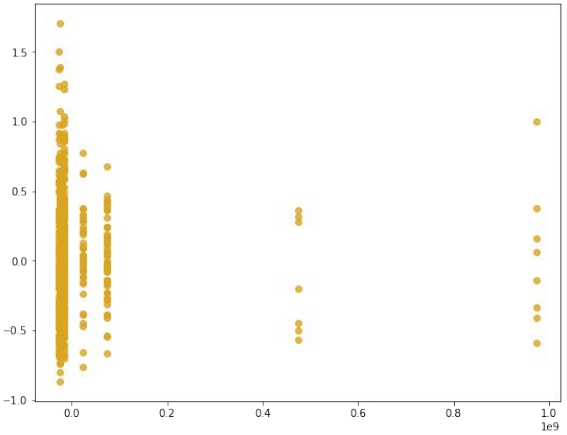
Model Name	Prediction Rate
Linear SVC	70.5015
Logistic Regression	66.7371
Multinomial NB	61.5951
Random Forest Classifier	61.5951

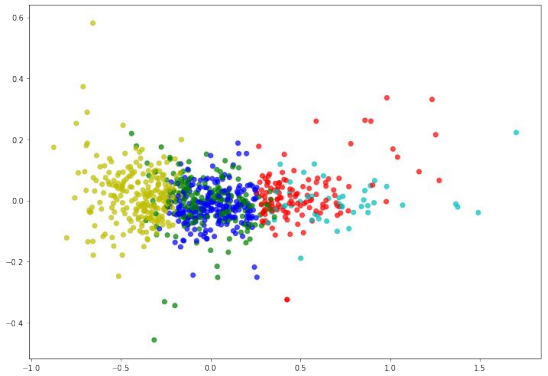


Prediction Rate using Cross Validation

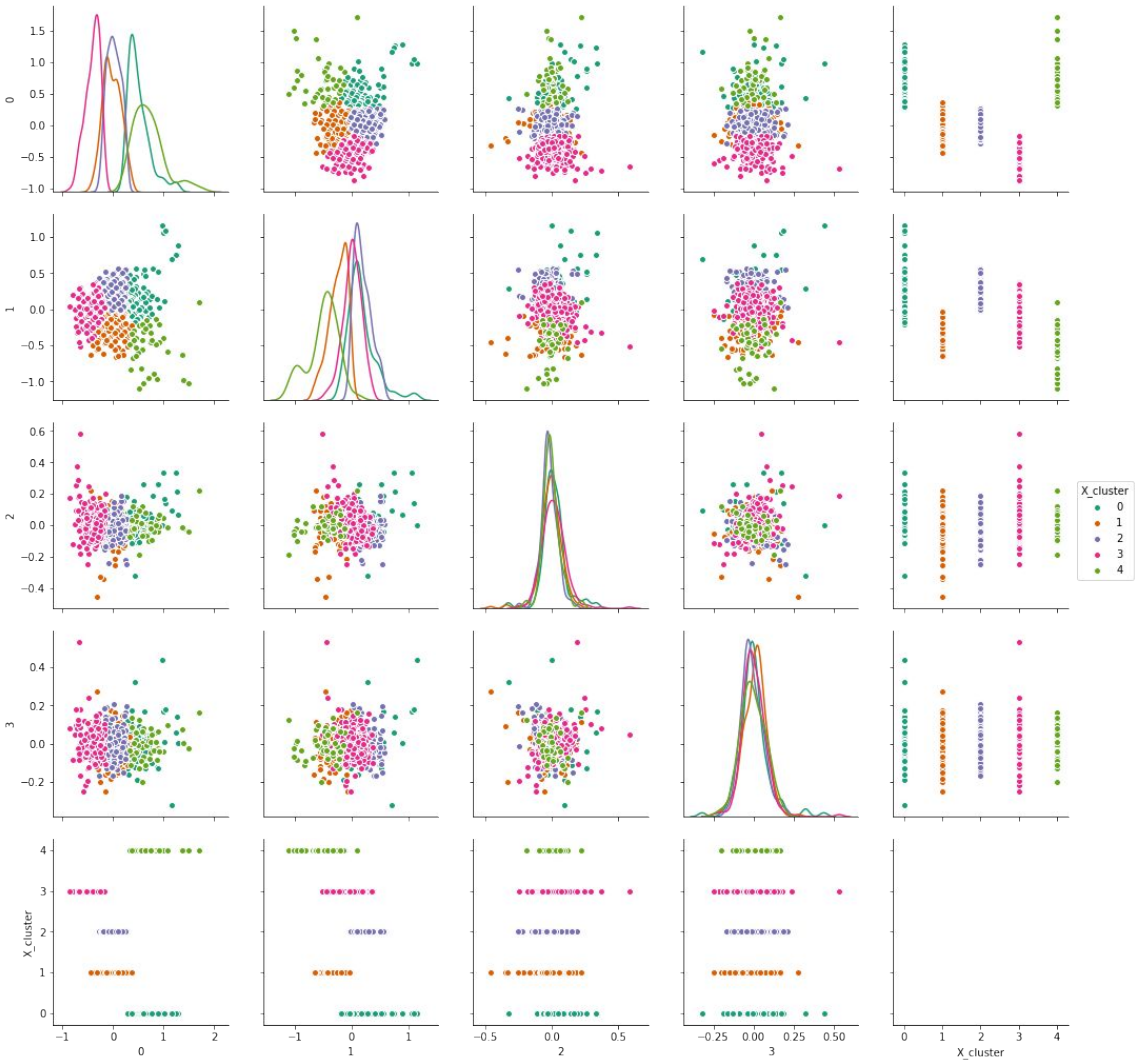
Clustering Based On User Reviews - google

- Purpose
  - Cluster the apps based on User Reviews.
  - Rank the apps in each cluster.
- Columns Used
  - Sentiment - Sentiment for each comments. Ex. Positive/Negative.
  - Sentiment Polarity - Denotes Sentiment Orientation.
  - Sentiment Subjectivity - Denotes peoples feeling.
  - Ratings - App Rating
- Feature Engineering
  - PCA - To Extract features

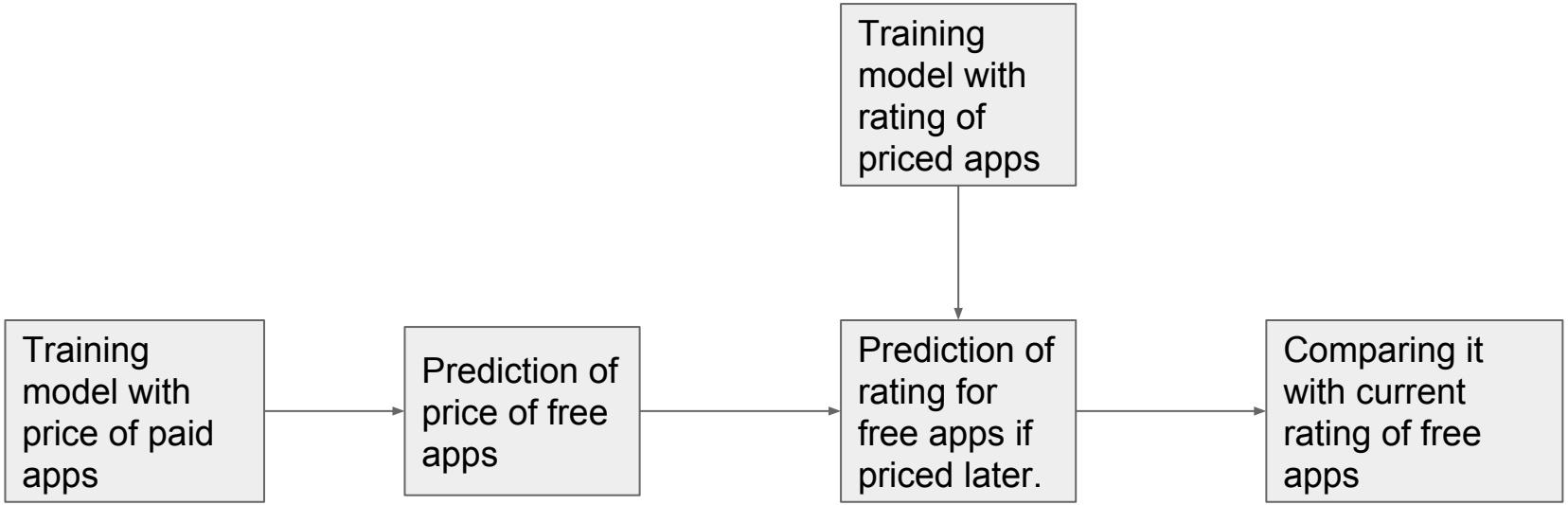




Pair Plot - Variations between 4 Principal Component



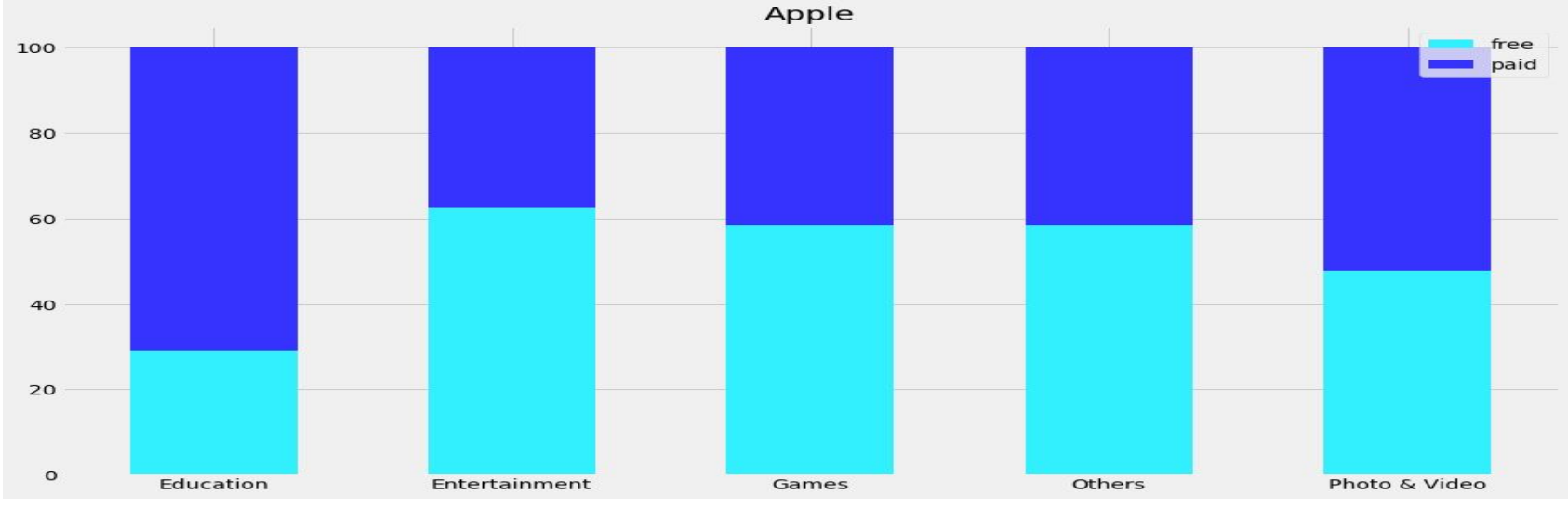
Price and rating prediction

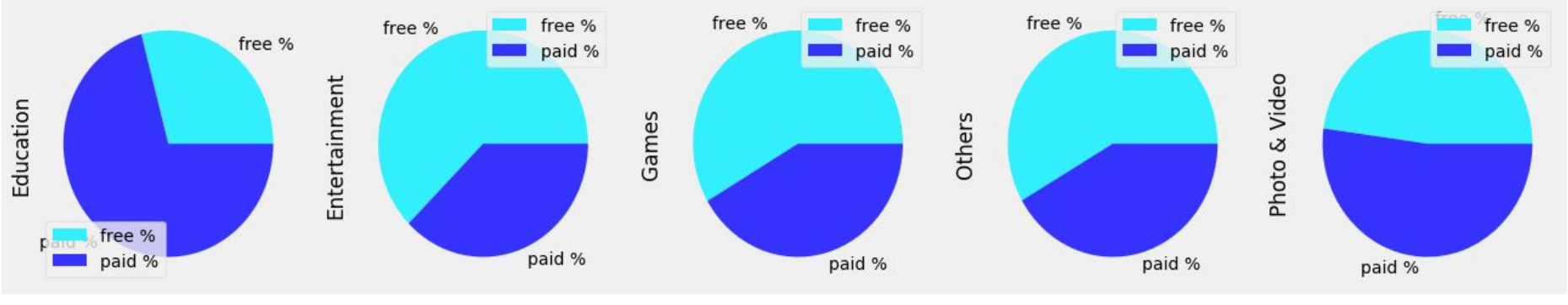


Implementation model of prediction system

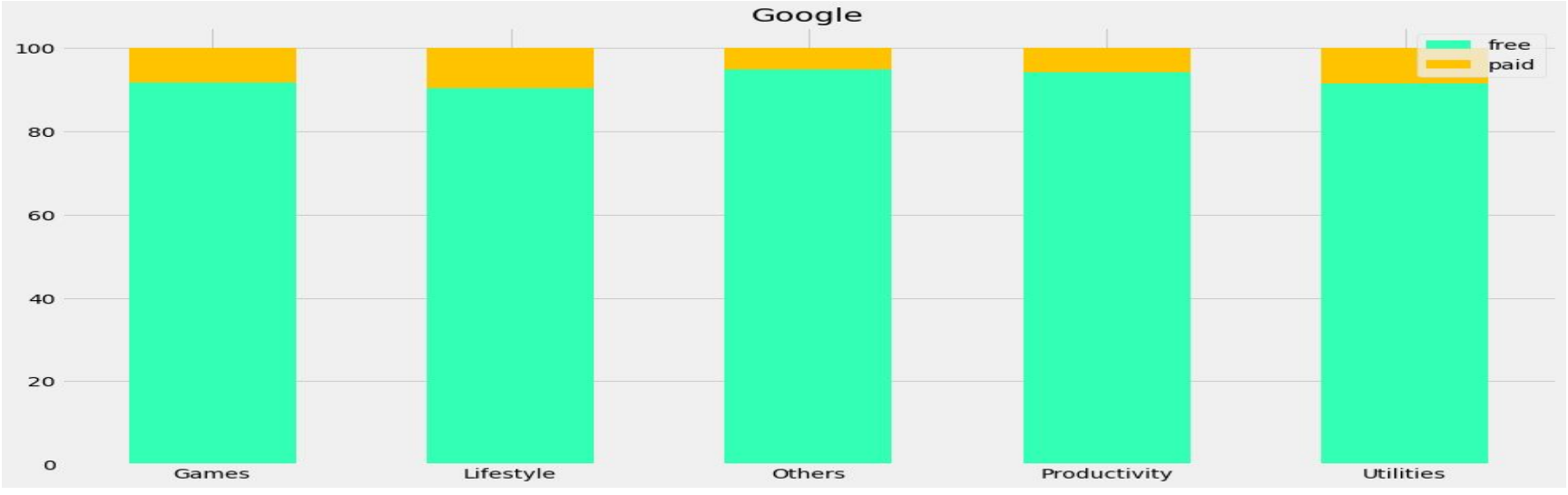
PRICE DISTRIBUTION OF PAID APPS

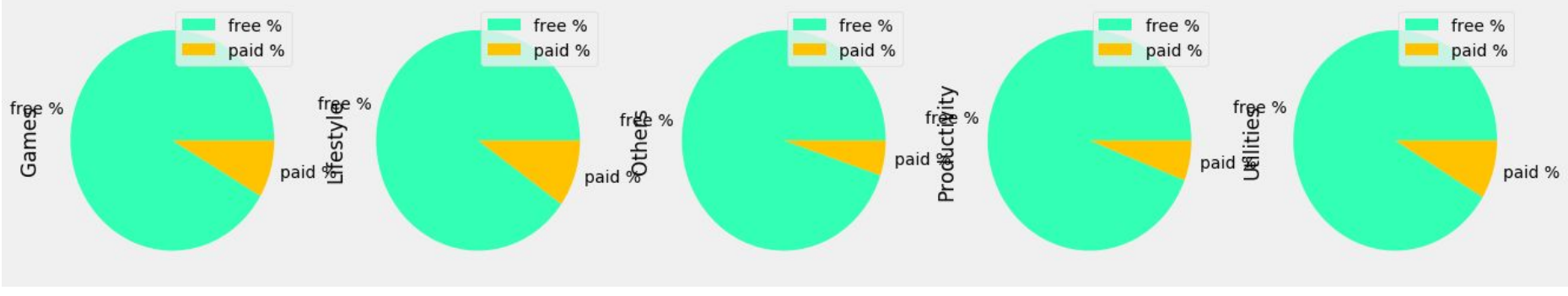
paid apps Vs Free apps - Apple



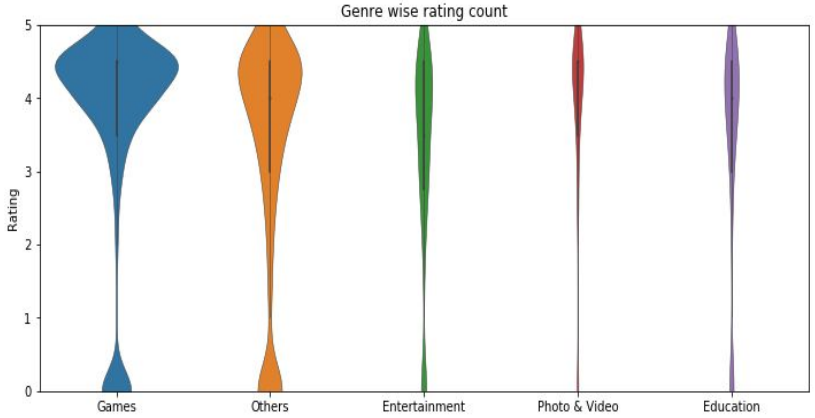


PAID APPS VS FREE APPS - GOOGLE



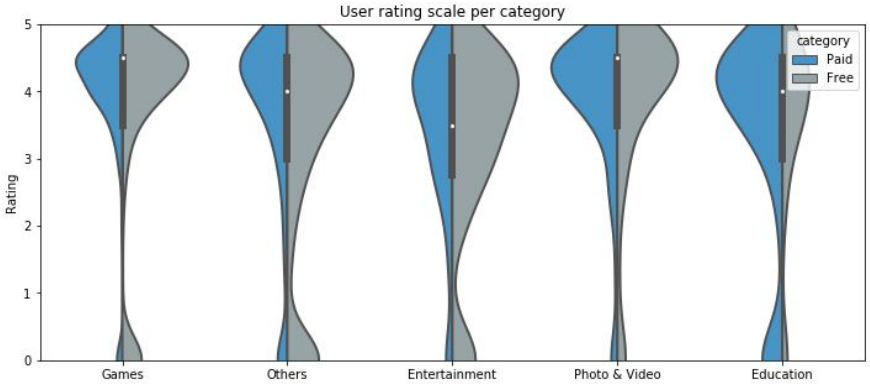


Predicting popularity if free app is priced - apple

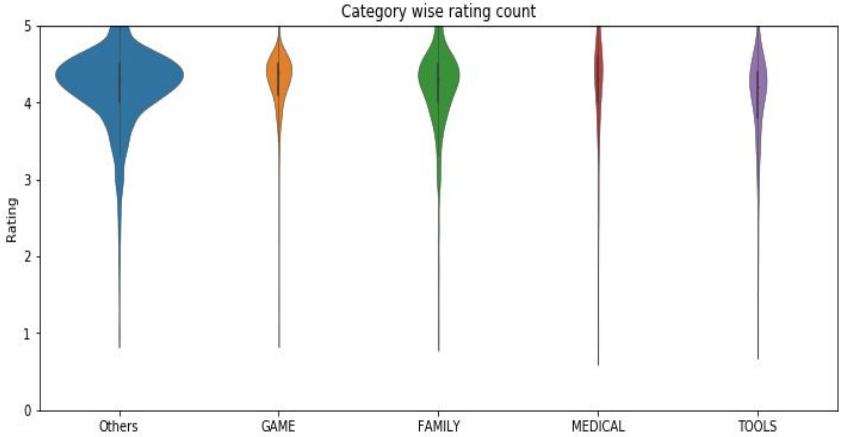


- In most of the categories, free apps are more popular

- More number of apps in rating range 4 - 5.

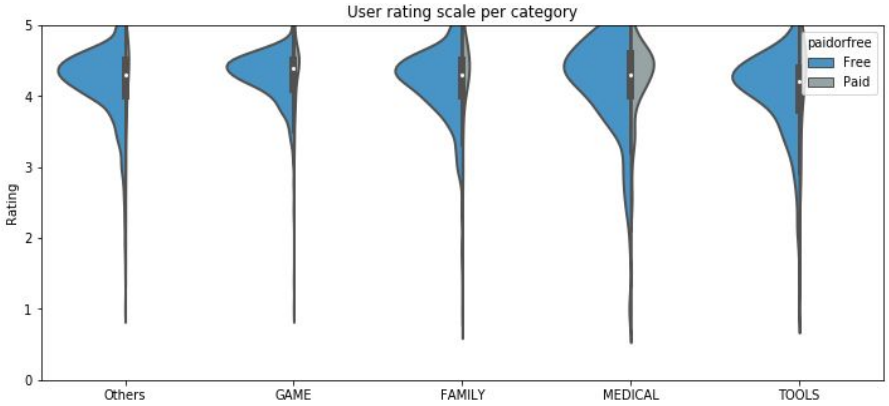


Predicting popularity if free app is priced - google



- Almost all free apps are more popular than priced apps.

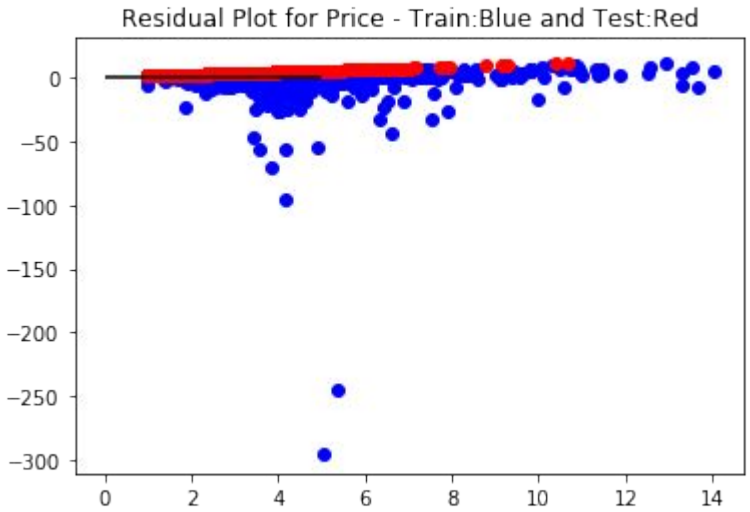
- More number of apps in rating range 4 - 5.



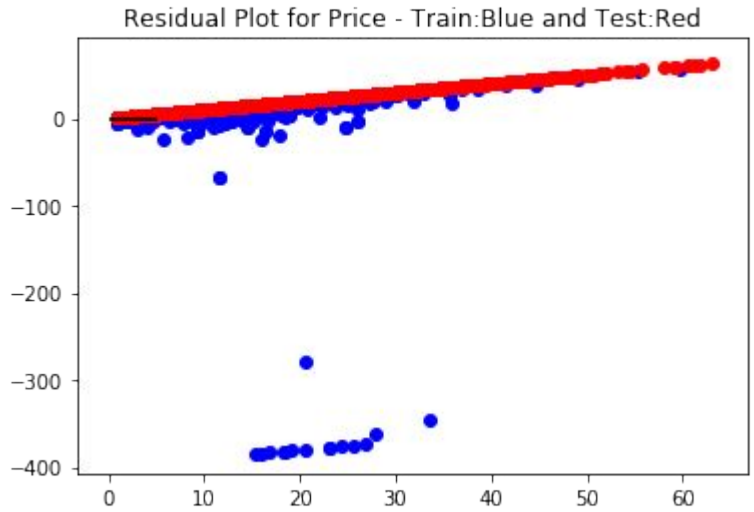
Predicting prices for free apps

- Using **Linear Regression Model** we tried to predict the price of free apps of both android and apple stores by training the model with the price of paid apps

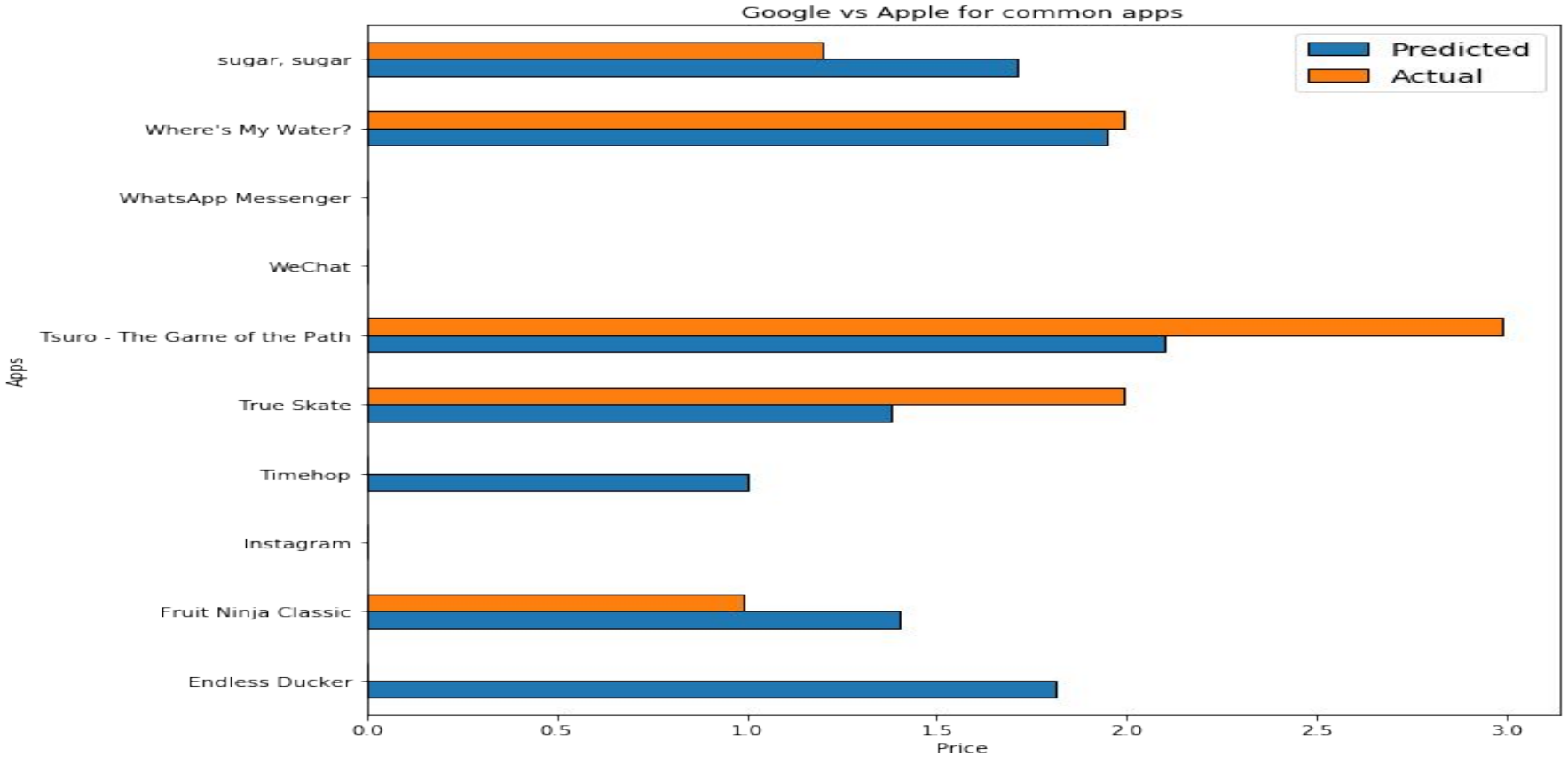
APPLE



Google



Comparing the Apple Store and Play Store by prices

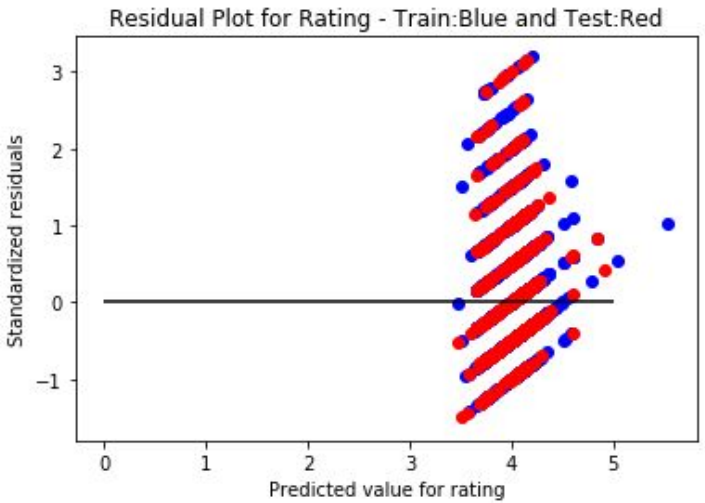


- Apple store is 26% more expensive than the Google Play store.

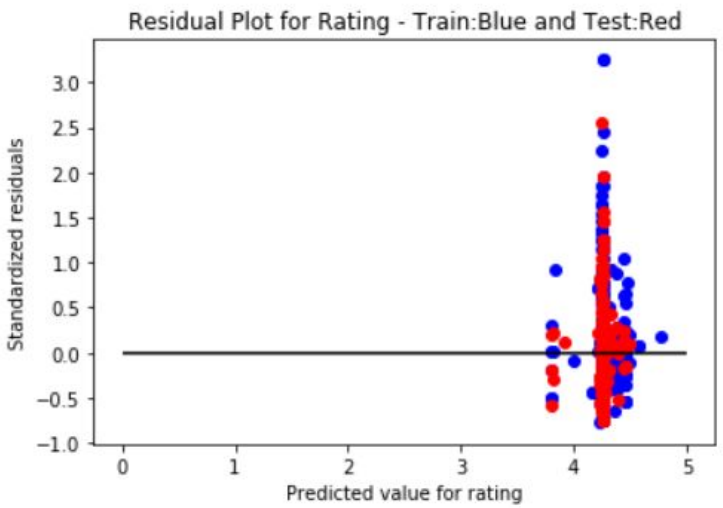
Predicting popularity if free app is priced

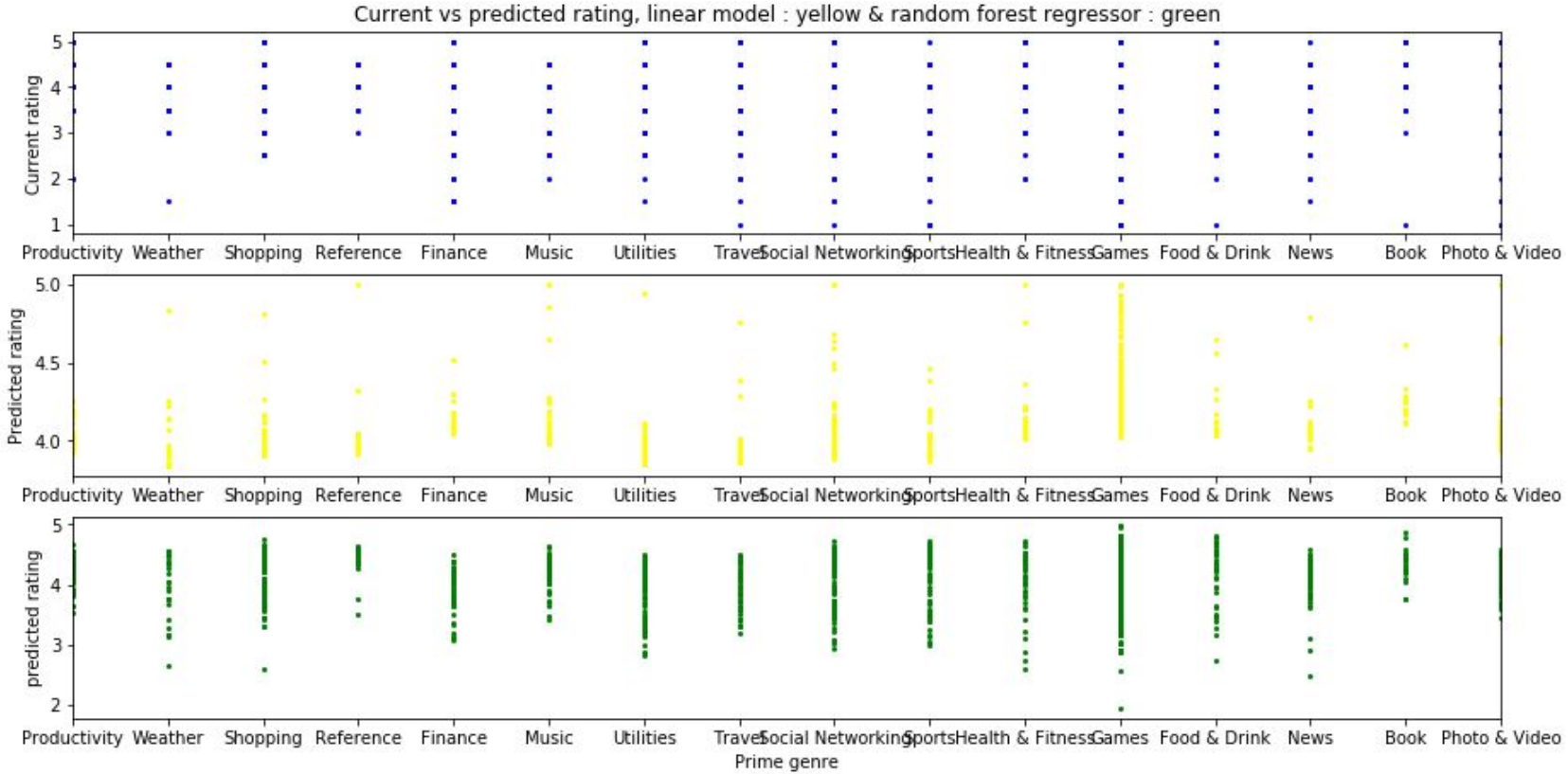
- Model used : Linear regression model and Random forest regressor.
- Trained model with the ratings of paid apps.

Apple

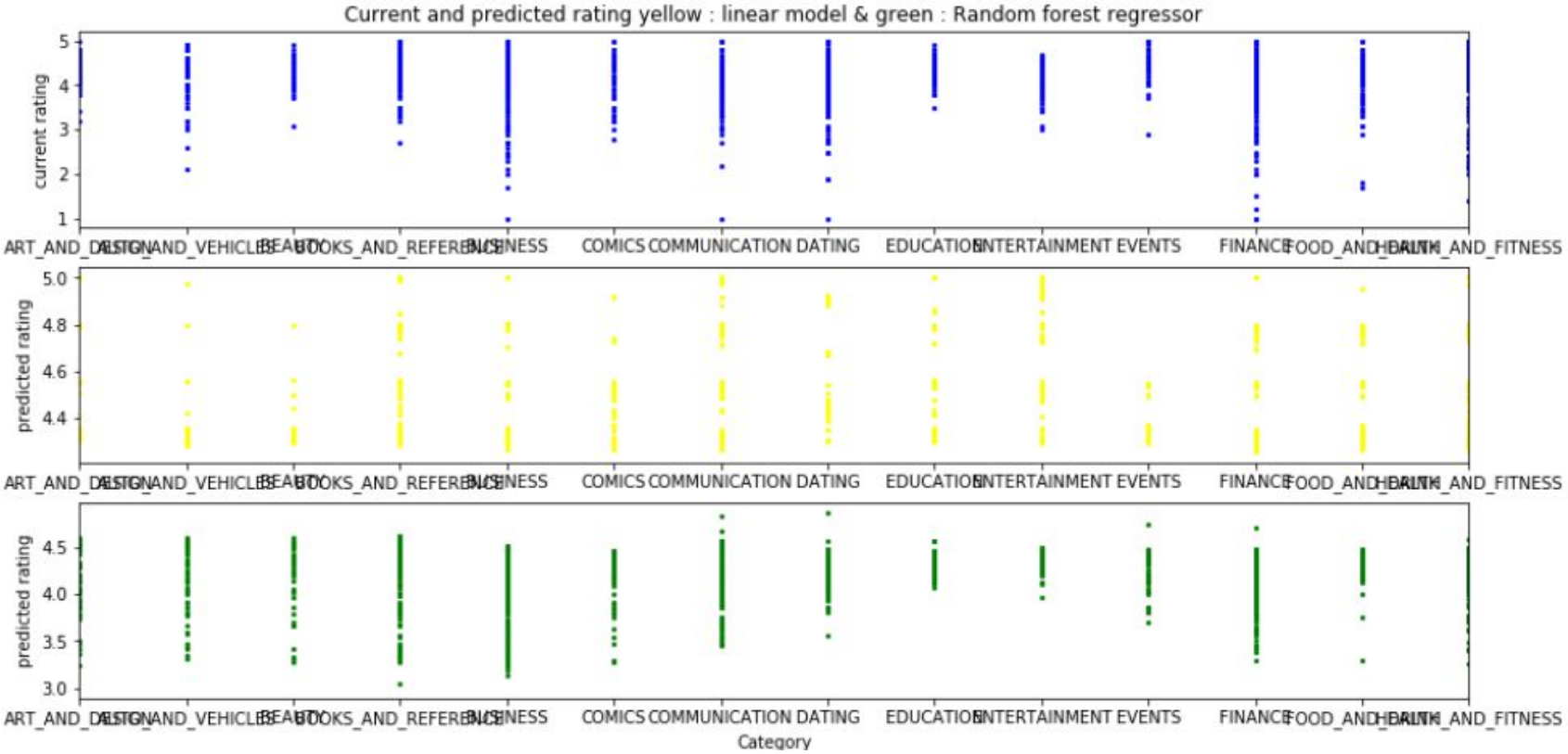


Google





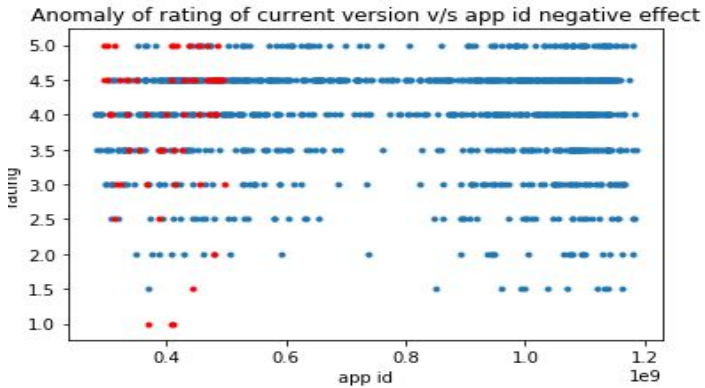
Rating prediction for free apps in apple store



Rating prediction for free apps in Google play store

Rating anomaly

- Outlier detection: User rating compared to previous version rating for apps of Applestore.
- Model used: Isolation forest



id	track_name	size_bytes	currency	price	rating_count_tot	rating_count_ver	user_rating	user_rating_ver
429851711	Flashlight !	14336000	USD	0.0	35769	21	4.5	2.0

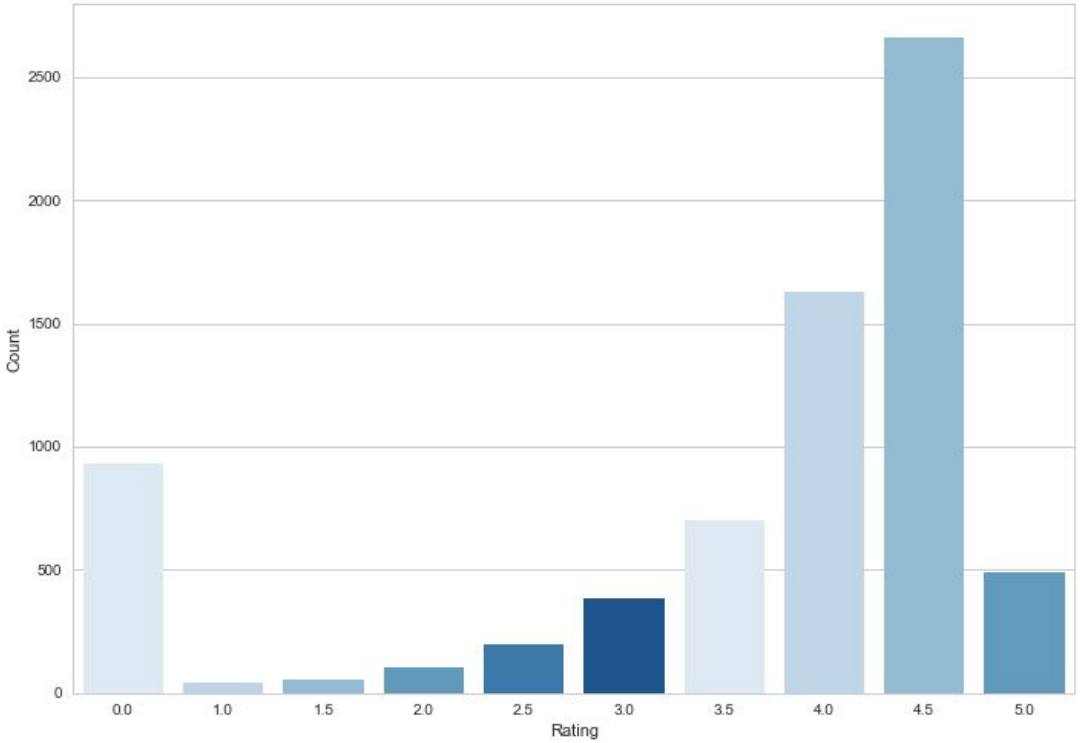
Recommender Systems

- Two different types of recommender systems built for Google and Apple App stores
  - **Simple Recommender**
  - **Content-Based Recommender**
  - **Collaborative-Filtering Based Recommender**

Simple Recommender

Returns top 15 apps based on weighted metrics

- Weighted rating computed for each app based on average rating per app and number of reviews
  - **Pros** - Easy to implement
  - **Cons** - Only provides generalized recommendations



	id	track_name	user_rating	rating_count_tot	score
1104	487119327	Head Soccer	5.0	481564	4.985514
302	350642635	Plants vs. Zombies	5.0	426463	4.983664
3086	930574573	Sniper 3D Assassin: Shoot to Kill Gun Game	5.0	386521	4.981996
2177	698255242	Geometry Dash Lite	5.0	370370	4.981221
498	387428400	Infinity Blade	5.0	326482	4.978733
1877	625334537	Geometry Dash	5.0	266440	4.974025
810	436491861	Domino's Pizza USA	5.0	258624	4.973254
2731	887947640	CSR Racing 2	5.0	257100	4.973099
1802	600674056	Pictoword: Fun 2 Pics Guess What's the Word Tr...	5.0	186089	4.963090
350	363282253	Plants vs. Zombies HD	5.0	163598	4.958160
1484	552039496	The Room	5.0	143908	4.952620
1817	606190854	Iron Force	5.0	141634	4.951884
1918	635573390	Sniper Shooter: Gun Shooting Games	5.0	134080	4.949267
467	381471023	Flashlight ◎	5.0	130450	4.947905
884	448639966	Pic Collage - Picture Editor & Photo Collage M...	5.0	123433	4.945054

Fig: Table showing the top 15 Play store apps based on weighted score

Content Based Recommender

Recommends apps similar to other apps

- Computes pairwise similarity scores based on app description
- Calculate TF-IDF vectors for each app
- Different measures of similarity applied- cosine similarity, Euclidean
- Applied a popularity filter to improve recommendations - based on ratings
    - **Pros** - Does a good job recommending apps that belong to the similar categories
    - **Cons** - Does not take user input(rating per app etc) into consideration

```
In [56]: get_recommendations('Evernote - stay organized')

Out[56]: 4354              FastEver 2 - Quick memo app for Evernote
         535                                                Noteshelf
         2318      Notepad+: Take Notes, Annotate and Write on PDF
         3931      Whink - Note taking, Annotate & Record Lectures
         654                                      Microsoft OneNote
         344                                            Notability
         3359               Carbo - Handwriting in the Digital Age
         228                     ScanBizCards Business Card Reader
         4019                                                  Bear
         508             CamScanner +| PDF Document Scanner and OCR
         1213    Paper by FiftyThree - Sketch, Diagram, Take Notes
         426                                            Notes Plus
         188                                          Epson iPrint
         1522    MetaMoJi Note - note taking and PDF annotation...
         5481                                               Jottit
         639                                SecurityCam for iPhone
         3607     Scanner For Me - PDF Scan with OCR for Documents
         4962             WatchNotes - Display notes on watch face
         1173                    ALON Dictaphone - Voice Recorder
         1637       InstaLogo Logo Creator - Graphic design maker
         220       Scanner Pro - PDF document scanner app with OCR
         399               AudioNote - Notepad and Voice Recorder
         184                                    My Measures PRO
         2522                          Scanbot - Scanner App & Fax
         820     TextGrabber — image to text: OCR & translate p...
         Name: track_name, dtype: object
```

Fig: Recommendations for Play store using cosine similarity

```
improved_recommendations('Evernote - stay organized')
```

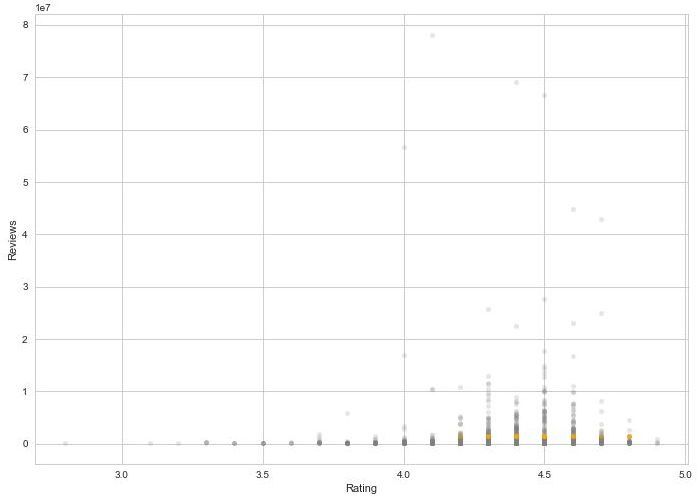
	track_name	user_rating	rating_count_tot	score
654	Microsoft OneNote	4	39638	3.949069
220	Scanner Pro - PDF document scanner app with OCR	4	31912	3.938346
1213	Paper by FiftyThree - Sketch, Diagram, Take Notes	4	18219	3.901642
344	Notability	4	17594	3.898895
535	Noteshelf	4	7562	3.816731
2318	Notepad+: Take Notes, Annotate and Write on PDF	4	6288	3.795641
1637	InstaLogo Logo Creator - Graphic design maker	4	6263	3.795178
426	Notes Plus	4	6257	3.795067
508	CamScanner +| PDF Document Scanner and OCR	4	5482	3.779594
2522	Scanbot - Scanner App & Fax	4	3936	3.740511

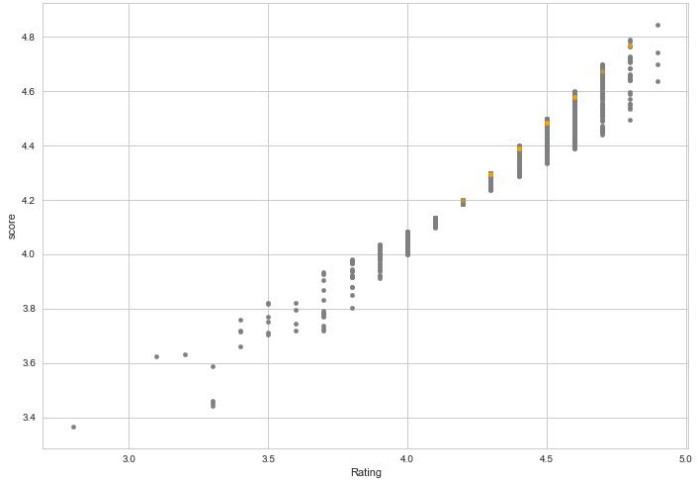
Fig: Improved recommendations using popularity filter

Collaborative-Filtering

- **User-based filtering**
  - Recommendations based on similar interests with other users

- **Item-based filtering**
  - Similar to content-based
  - Recommend based on past ratings from users





CONCLUSION

- Successfully trained and tested classifiers based on multiple criterion
- Built regression models for price/rating prediction
- Grouped apps based on user reviews using clusters
- Developed recommender systems to suggest most suitable apps to any user

Thank you !

Questions?