Predicting App Ratings on the Google Play Store

Shelley Jones

Agenda

- Context
- Business Question
- Data Dictionary
- Exploratory Data Analysis
- Data Pre-Processing and Visualisations
- Unsupervised Learning
- Supervised Learning
- Model Evaluation
- Conclusions
- Questions

Context

- App ratings on the Google Play Store play a crucial role in the success of mobile applications:
  - User trust and credibility
  - Visibility and discovery
  - Impact on download rates
  - User feedback and improvement
  - Influence on monetisation
  - Competitive advantage
- App ratings are essential for building trust, enhancing visibility, driving downloads, gathering user feedback, and maximizing revenue potential
- For businesses, maintaining high ratings is not just about attracting downloads; it’s about fostering a positive relationship with users, which ultimately contributes to long-term success.

Business Question

"Using data that is already publicly available, can we develop a machine learning framework that utilizes both supervised and unsupervised models to predict the Google Play Store rating of a mobile app based on features such as user reviews, download count, app updates, and category, achieving an accuracy of at least 85% within the next four weeks?"

- **Specific**: Clearly defines the goal (predicting app ratings) and mentions the use of both supervised and unsupervised models, along with the features considered
- **Measurable**: Establishes a clear success criterion (achieving at least 85% accuracy) for evaluating the model's performance
- **Achievable**: Assumes access to the necessary data and resources, making the goal realistic
- **Relevant**: Aligns with business objectives related to app performance and user satisfaction
- **Time-bound**: Specifies a timeframe (within the next four weeks) for developing the framework.

Data Dictionary

The dataset contained 13 features in which 'Rating' was the target feature.

- App: Name of the application
- Category: Category the app belongs to e.g. ART_AND_DESIGN, FINANCE, COMICS, BEAUTY
- Rating: The current average rating (out of 5) of the app on Google Play
- Reviews: Number of user reviews given on the app
- Size: Size of the app in MB (megabytes) or KB (kilobytes)
- Installs: Number of times the app was downloaded from Google Play
- Type: Whether the app is paid or free
- Price: Price of the app in US$
- Content Rating: The age group suitable for the app
- Last Updated: Date on which the app was last updated
- Genres: The genres associated with the app.
- Android ver: The minimum Android OS version required to run the app
- Current ver: The current version of the app

Exploratory Data Analysis

[12]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
7869	POCKET ATLAS OF CT HEAD	MEDICAL	4.3	22	18M	10,000+	Free	0	Everyone	Medical	March 16, 2018	2.0	4.0.3 and up
7422	CJ'S TIRE AND AUTO INC.	PRODUCTIVITY	5.0	5	11M	100+	Free	0	Everyone	Productivity	May 30, 2018	1.0.1	4.1 and up
5753	AW	FAMILY	NaN	0	Varies with device	5+	Free	0	Everyone 10+	Strategy	August 28, 2015	Varies with device	Varies with device
3516	File Browser by Astro (File Manager)	PRODUCTIVITY	4.3	609182	9.2M	50,000,000+	Free	0	Everyone	Productivity	July 2, 2018	6.4.0	4.0 and up
7445	CJ Infinity	FOOD_AND_DRINK	NaN	0	16M	10+	Free	0	Everyone	Food & Drink	January 5, 2018	1.1	4.1 and up
4270	Guess the song of J Balvin	GAME	NaN	28	8.9M	1,000+	Free	0	Everyone	Trivia	December 24, 2017	1.1	4.1 and up
473	Talkray - Free Calls & Texts	COMMUNICATION	4.2	244863	Varies with device	10,000,000+	Free	0	Everyone	Communication	May 29, 2018	Varies with device	Varies with device
290	TurboScan: scan documents and receipts in PDF	BUSINESS	4.7	11442	6.8M	100,000+	Paid	$4.99	Everyone	Business	March 25, 2018	1.5.2	4.0 and up
9923	CALIOPE EU: Air Quality	HEALTH_AND_FITNESS	3.9	21	2.2M	1,000+	Free	0	Everyone	Health & Fitness	October 30, 2015	1.1.2	4.0 and up
5022	Ae Allah na Dai (Rasa)	BOOKS_AND_REFERENCE	4.7	263	9.1M	10,000+	Free	0	Everyone	Books & Reference	January 1, 2015	2.0	2.1 and up

- 10841 rows, 13 columns
- Lots to think about!

Exploratory Data Analysis

	count	mean	std	min	25%	50%	75%	max
Rating	9367.0	4.193338	0.537431	1.0	4.0	4.3	4.5	19.0

- A rating of 19.0 should not be possible
- All but one column are 'object' types and will need some sort of cleaning / encoding (categorical data)
- Null values for Rating

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             10841 non-null  object
 1   Category        10841 non-null  object
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object
 4   Size            10841 non-null  object
 5   Installs        10841 non-null  object
 6   Type            10840 non-null  object
 7   Price           10841 non-null  object
 8   Content Rating  10840 non-null  object
 9   Genres          10841 non-null  object
 10  Last Updated    10841 non-null  object
 11  Current Ver     10833 non-null  object
 12  Android Ver     10838 non-null  object
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

Data Pre-Processing and Visualisations

Rating:

- What about the value of 19.0?

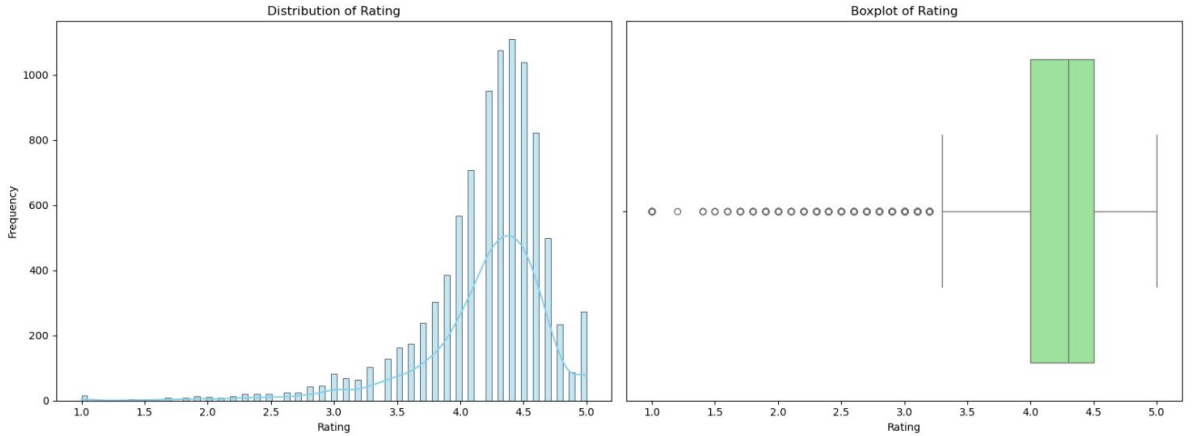
	App	Category	Rating	Reviews	\
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	

	Size	Installs	Type	Price	Content Rating	Genres	...	\
10472	1,000+	Free	0	Everyone	NaN	February 11, 2018	...	

- Data in this row was misaligned
- Dropped the row

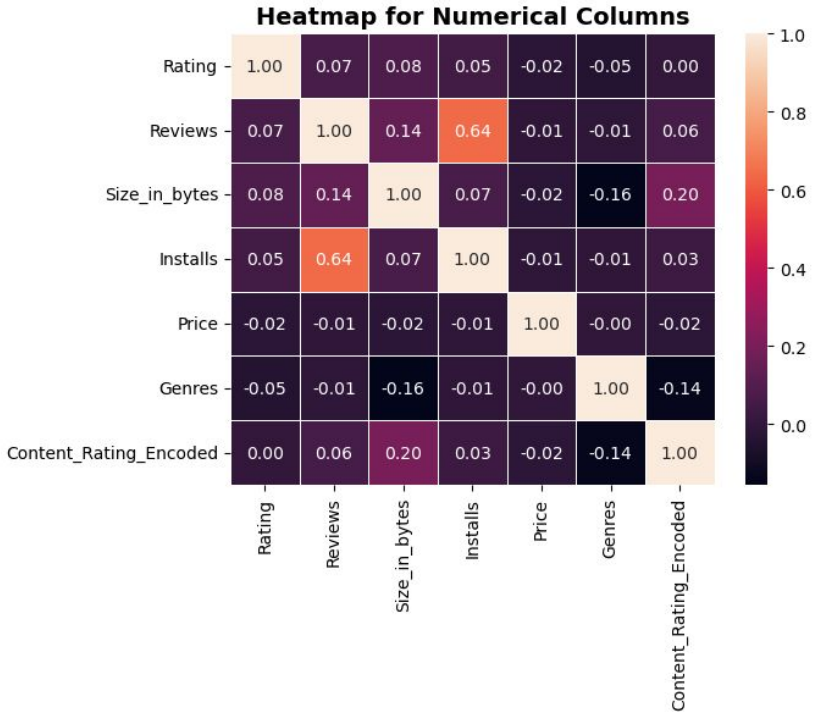
Data Pre-Processing and Visualisations

- Skewed left; majority are higher ratings (little unexpected)
- 1474 missing values
- Concern over data leakage, can’t impute from another feature
- Opted to keep data clean and deleted rows with missing values



Data Pre-Processing and Visualisations

- Continued cleaning and converting data
- Correlation between Reviews and Installs
- Rating not correlated with anything



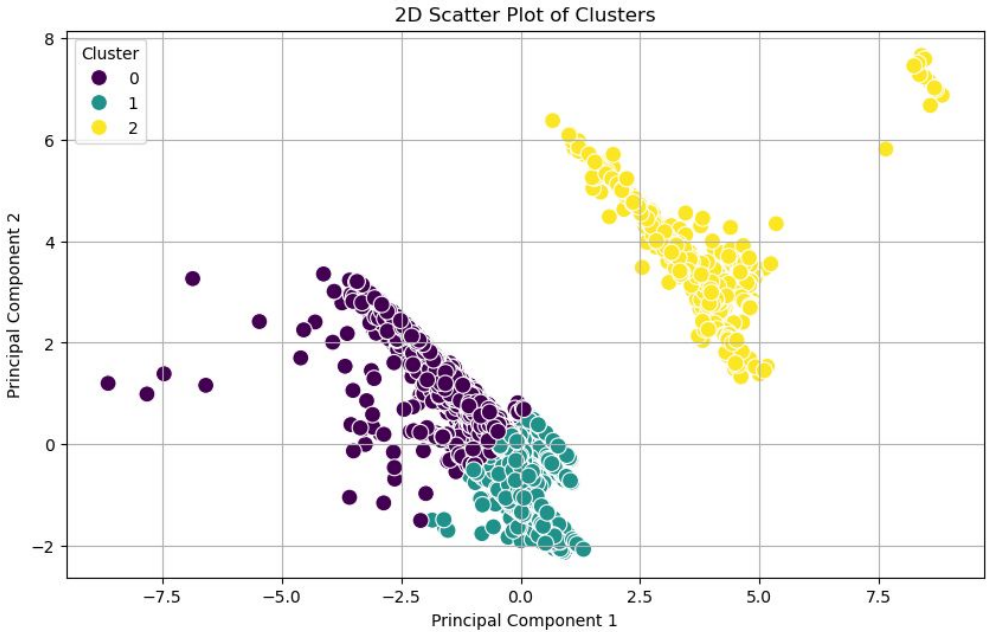
Unsupervised Learning

KMeans

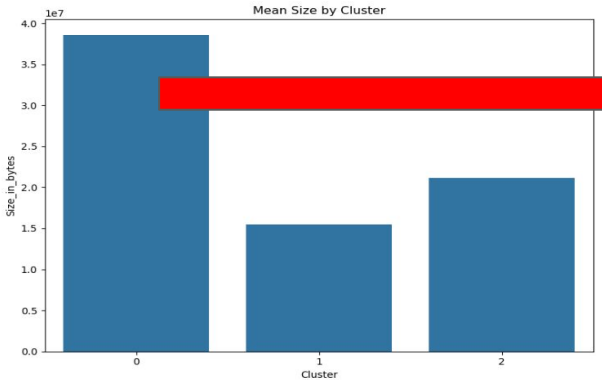
- Dropped categorical columns (e.g. App, Last Updated, Current Ver, Android Ver)
- Defined the target and predictor features:
  - Target = Rating
  - Predictor features: All other columns
- Performed Train-Test split
- Scaled the data
- Dimensionality reduction with PCA (2 components)
- Used Elbow Method for Optimal K
- Fit the model and predicted cluster labels

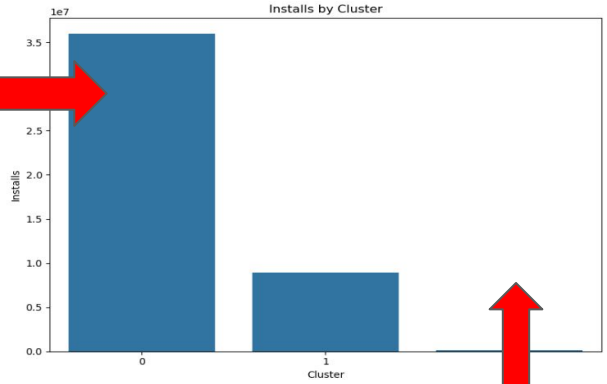
Cluster Visualisation

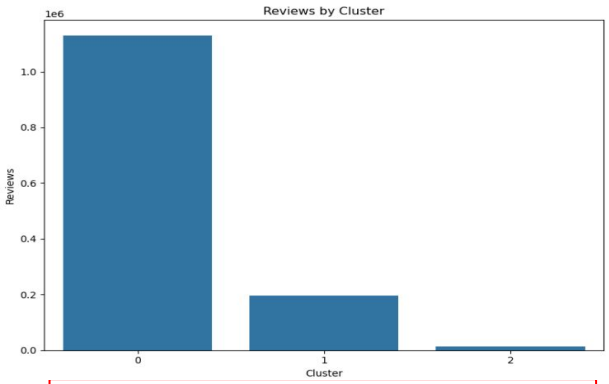
- Fairly separated clusters
- Distinct groups in data?

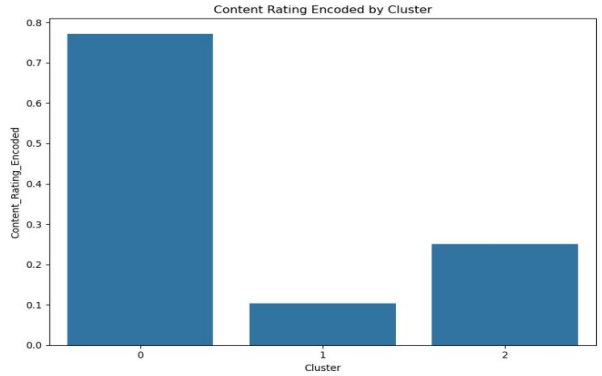


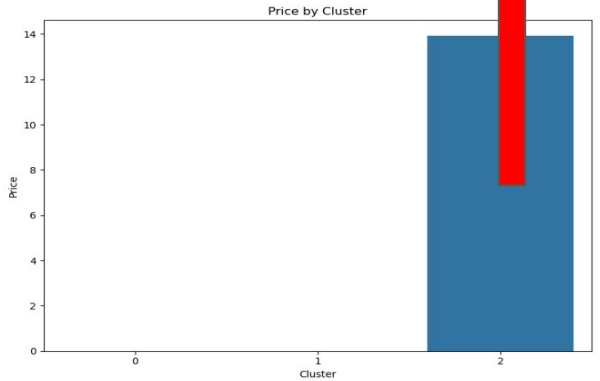
Cluster Comparison

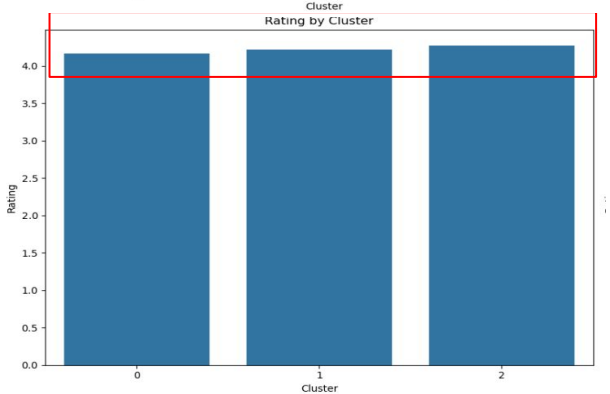


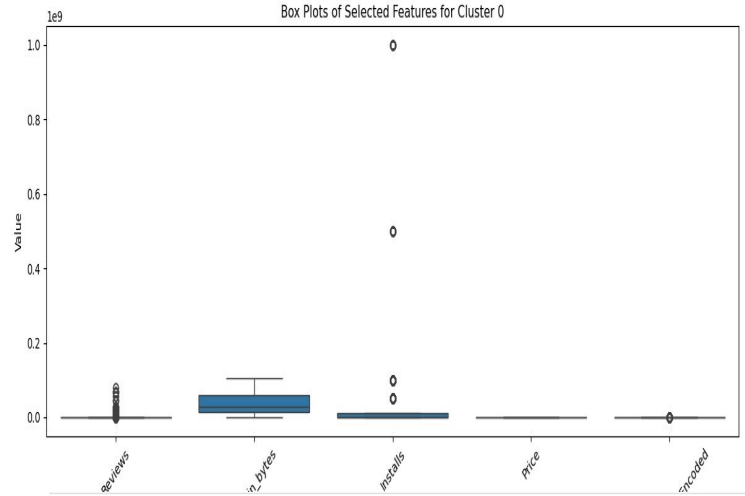


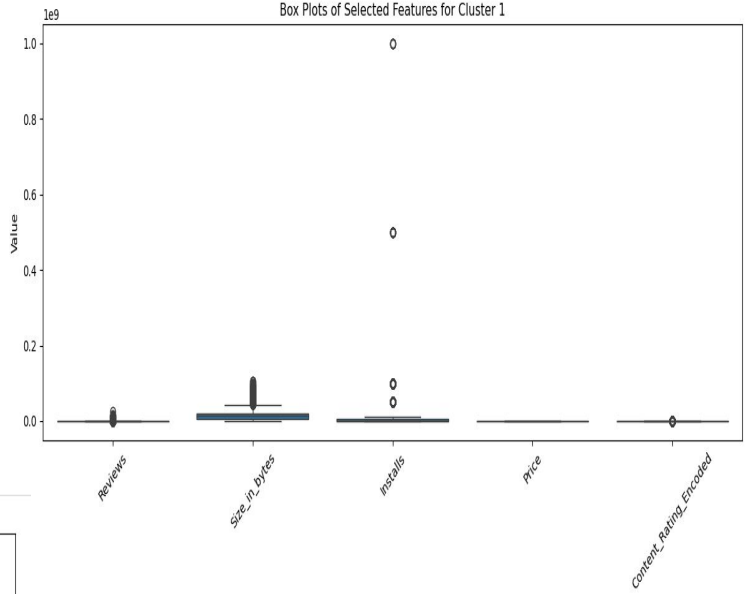


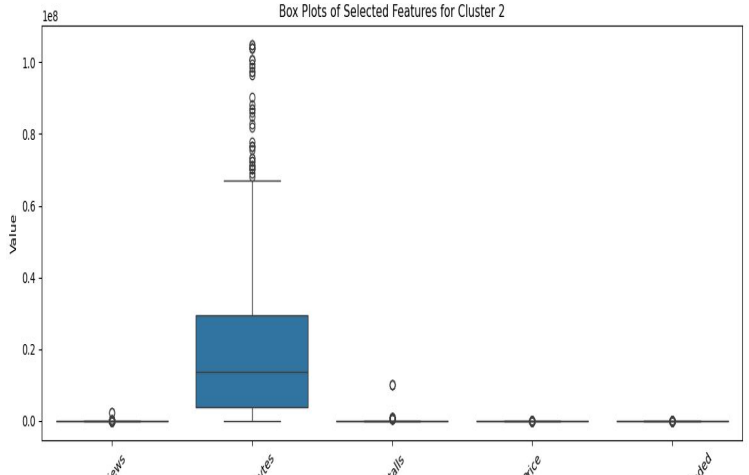




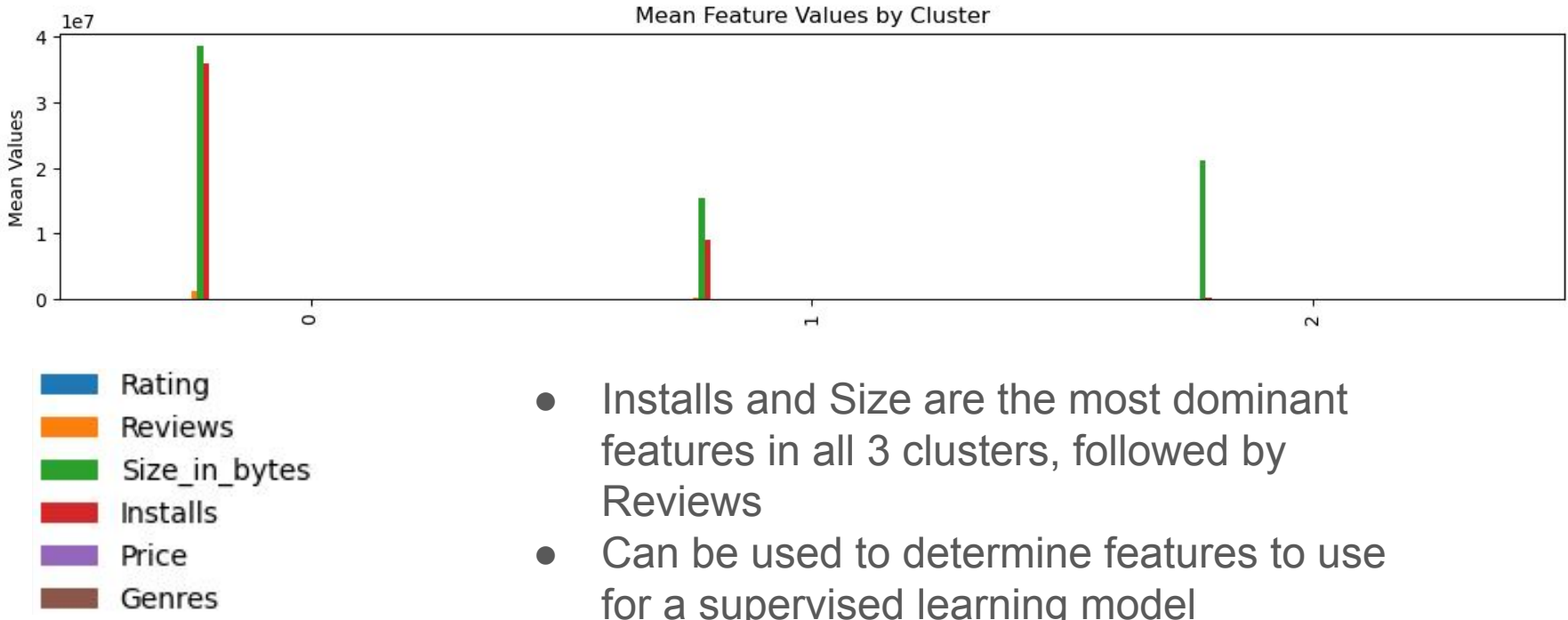



Box Plots of Selected Features for Cluster 0
1e9
1.0
0.8
0.6
0.4
0.2
0.0
Value
Reviews
Size_in_bytes
Installs
Price
Content_Rating_Encoded


Box Plots of Selected Features for Cluster 1
1e9
1.0
0.8
0.6
0.4
0.2
0.0
Value
Reviews
Size_in_bytes
Installs
Price
Content_Rating_Encoded


Box Plots of Selected Features for Cluster 2
1e8
1.0
0.8
0.6
0.4
0.2
0.0
Value

Dominant Features



- Installs and Size are the most dominant features in all 3 clusters, followed by Reviews
- Can be used to determine features to use for a supervised learning model

Supervised Learning

- Cluster labels were appended to the original dataset
- Defined the target and predictor features:
  - Target = Rating
  - Predictor features: All other columns
- Performed Train-Test split
- Scaled the data
- Performed modelling using a number of regression techniques
- Evaluated each model

Comparisons

Method	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R-Squared
Linear Regression	Train: 0.35 Test: 0.36	Train: 0.26 Test: 0.24	Train: 0.51 Test: 0.49	Train: 0.04: Test: 0.02
KNN	Train: 0.30 Test:0.37	Train: 0.19 Test: 0.28	Train: 0.44 Test: 0.53	Train: 0.28 Test: -0.13
Random Forest	Train: 0.11 Test: 0.37	Train: 0.03 Test: 0.28	Train: 0.18 Test: 0.53	Train: 0.88 Test: -0.13
Gradient Boosting	Train: 0.27 Test: 0.31	Train: 0.15 Test: 0.21	Train: 0.39 Test: 0.46	Train: 0.44 Test: 0.15
Bagging	Train: 0.16 Test: 0.31	Train: 0.06 Test: 0.21	Train: 0.24 Test: 0.46	Train: 0.79 Test: 0.14
Stacking	Train: 0.22 Test: 0.31	Train: 0.10 Test: 0.21	Train: 0.31 Test: 0.45	Train: 0.64 Test: 0.16

Conclusions

Headlines:

- None of the models captured the underlying patterns in the data correctly. Points to consider include:
    - Data Quality: Noisy data / outliers. I did not remove any potential outliers
    - Feature Selection: From the heatmap, no features had a strong relationship with Rating
    - Model Complexity: I used a range of model complexities to combat underfitting
    - Data Leakage: Probably not an issue as my scores were low
    - Rating Distribution: Ratings were skewed

Other Points to Consider:

- Stratify the Ratings across Train & Test datasets (should have done!)
- Investigate the data quality further especially for noisy data
- Search for additional relevant datasets (there is one which captures review comments)
- Convert Rating to categorical data (e.g. low, medium, high) and use Classification methods
- Look into Ordinal Logistic Regression (Rating has a natural order i.e. 1 to 5)

Finally:

- None of the models produced results that a business could rely upon; further work would be required.

References

https://www.kaggle.com/datasets/lava18/google-play-store-apps

Questions?