Google Play 2018 Apps – 'Ratings' Regression Model

Data Analysis

Data Analysis changes

- There existed an anomalous 1.9 category which consisted of a single record and was deleted – interestingly this record had all its columns shifted to the left.
- The Reviews column was converted from string to integer as it was incorrectly a string due to the above record.
- Duplicates were dropped based on if the app name was the same.
- The Installs column could be better used as an integer and so was converted by dropping the "+" and casting as int.
- About 15% of the Ratings were null, so were removed as they were our target. Otherwise we'd also contaminate our data if we tried to fill it in somehow.
- Sizes were in Megabyte, Kilobyte and significant amount were labelled as "Varies with device". Decided to convert to integer and average the size based on Category using a separate dataframe before applying to the main dataframe.
- The Price column was in dollars as a string and needed to be changed to a float.
- Last Updated column had too many datapoints to use effectively, so extracted only the year.
- Removed the Genres column as it had too many datapoints as is, and was also already represented better by the Category column.
- Removed the Current Ver column as the values were relative to each app; there would be no adequate data to compare this with.

Insights

Rating, Reviews and Installs

Reviews and Installs don't seem to have any influence on Rating whatsoever, although, the amount of Reviews and Installs themselves do have a decent correlation of 0.62 on average.

Category: HOUSE_AND_HOME

Even though still low, the correlations between Rating and Reviews/Installs were significantly higher in the HOUSE_AND_HOME category compared to others.

- This could mean that encouraging higher app participation/engagement in this category may result in an increase in Rating.

Category: EVENTS

Interestingly, EVENTS' apps' Reviews and Installs correlated negatively with Rating, even if not a particularly strong negative correlation.

This could mean that EVENTS apps are currently subpar in some way when it comes to mass user experience and investment into this may improve the Rating.

- Or that the nature of the content of EVENTS apps are more likely to upset an individual for some reason, as more people use it – perhaps something like the availability of tickets/places for an EVENT decreasing significantly once more users are active on the app, therefore causing discontent?

- Despite this, the average rating of EVENTS apps is at 4.4, with the app average at a 4.2.

Reviews and Installs Columns Correlation

The average correlation of Installs and Reviews amongst all Categories is 0.62, but the following categories have a very strong correlation, ranging from 0.97 to 0.9, in descending order: EDUCATION, MAPS_AND_NAVIGATION, EVENTS, WEATHER, and BEAUTY.

- This strong relationship shows high user engagement – that these individuals are willing to give feedback. This makes them a great source of information to improve the app – these companies could actively reach out to these individuals for a more qualitative review of their app.

- But why is the engagement so high? For each Category, I looked at some of the most installed apps and their reviews.

-- EDUCATION: Consisted of learning apps, including language learning such as DuoLingo. Most reviews in this field were positive reviews about how well the individual had learnt with the app.

-- MAPS_AND_NAVIGATION: GPS apps such as Waze. A lot of the reviews are about either; liking/lacking features, functionality issues, or problems with updates.

-- EVENTS: Ticket apps such as Ticketmaster. Frustrations about app functionality (selecting tickets, seats, inputting credit card details, GPS issues, tickets not showing, search options, etc.) seem to be the most prevalent. This also explains why the number of Reviews slightly correlates with a negative Rating as lot of people are reviewing to complain.

-- WEATHER: Forecasting apps such as AccuWeather. This seems to be the most balanced thus far; some people praising the app and its features while others are having problems with the functionality, particularly with updates. It seems that there aren't many glaring issues like with EVENTS apps, but small frustrating ones. Something like this maybe due to Android update versions, or how the app interfaces with different types of phones.

-- BEAUTY: Consists of virtual makeup filters and selfie filters like Filters for Selfie. Not only do the reviews mainly report back how good these types of apps are, but there also seems to be a high amount of engagement within the review section itself – shown by the amount of likes across each review.

Type and Price Columns

Although the ratio of Paid to Free apps was around 1:12, we see that the Ratings do not vary much between the two – there seems to be no real correlation.

On the other hand, when an app is paid for, there seems to exist a slight positive correlation between its Price and its Rating. To confirm this more data points would be needed.

- If the above is true then it is likely that either paid apps have higher quality (because of a higher development budget), or something more psychological is happening with the individual (a sort of monetised confirmation bias).

Content Rating Column

Everyone, Teen, and Everyone 10+ categories remained quite similar in terms of the Rating value – this may show that, for the most part, age does not matter when it comes to determining a Rating when it comes to an age universal type of app.

Although, the Mature 17+ Rating seemed to be lower on average – potentially meaning that Mature apps are critiqued more harshly than universal apps, or that their quality is worse.

Unfortunately, there weren't enough data points for Adult and Unrated categories.

Last Updated Column

There exists a slight correlation between the Rating of the app and when it was Last Updated, this could mean that it is necessary to keep apps up to date to achieve a higher Rating.

More recently updated apps seem to have a higher install rates than outdated ones – this indicates that apps must be kept up to date, otherwise install rates will suffer as time goes on.

Also, the app Size is higher for apps which have been updated more recently, which is something to take into consideration when updating your app frequently – developers must make sure to remove unneeded bloat files within the app as time goes on.

Android Version Column

The apps compatible with only higher version Androids seem to reliably score a higher rating, whilst those apps which are compatible with even the lowest version of Android also score higher in general.

It seems that the apps which are compatible with some androids, but not others, have a much larger range of Rating. This hints at the idea of compatibility playing a significant part in low ratings – which the updating of the app or the phone may well influence.

Neural Network

Pre-processing changes

- Dropped the App name column
- One-hot encoded the following: Category, Type, Content Rating, and Android Ver
- Dropped the original dummy columns

The NN was created using Dense layers. Dropout layers and the EarlyStopping callback was used to reduce overfitting.

Multiple configurations were tried, including changing the number of neurons in layers, changing the number of layers, trying a different Dropout %, and changing the optimiser.

The results of the MAE and RMSE differed slightly, around 0.365 for MAE and 0.52 for RMSE. In this case using the MAE would be more appropriate as we're working with a tighter, contained Regression problem – between 0 and 5. Whereas RMSE would be more appropriate for punished higher values found in more open-ended Regression scenarios (eg. Estimating house prices).

Therefore, our Error of 0.365 is adequate for predicting the Rating of an app. Ultimately this translates to a 7.3% deviation in Rating.

RandomForest with RandomizedSearchCV

The RandomForestRegressor model performed slightly better than my custom tensorflow model, giving results of MAE 0.34 and RMSE of 0.491

Testing with 100 then 300 trees barely improved the model to MAE 0.338 and RMSE 0.49.

Initialised a random search grid at 100 iterations and passed through the RandomForest, the best_params_ for ONLY the randomised grid had an MAE of 0.334 and RMSE of 0.48 – but this gave an idea for what to analyse with the GridSearchCV

GridSearchCV Model

```
{'n_estimators': 200,
 'min_samples_split': 10,
 'min_samples_leaf': 1,
 'max_features': 'sqrt',
 'max_depth': None,
 'bootstrap': True}
```

Was the result from the random searched random forest.

The final best grid was:

```
{'bootstrap': True,
 'max_depth': None,
 'max_features': 'sqrt',
 'min_samples_leaf': 1,
 'min_samples_split': 12,
 'n_estimators': 300}
```

Which produced the same result of MAE 0.334 and RMSE 0.48.

From the base model we managed to attain an 8% refinement in MAE and a 7.7% refinement in RMSE, which is a tangible and adequate enhancement on the model itself. The combination between the random grid and grid search allowed us to hone in on which parameters affected this model the most, and then fine tune them. This allowed us to use the time more efficiently whilst also producing a better model to be used.

Ultimately, this final model translates into a 6.7% deviation in rating, which is better than the original model's 7.3% deviation.