CKME 136- Play Store User Rating Prediction – Initial Results

Anthony Le

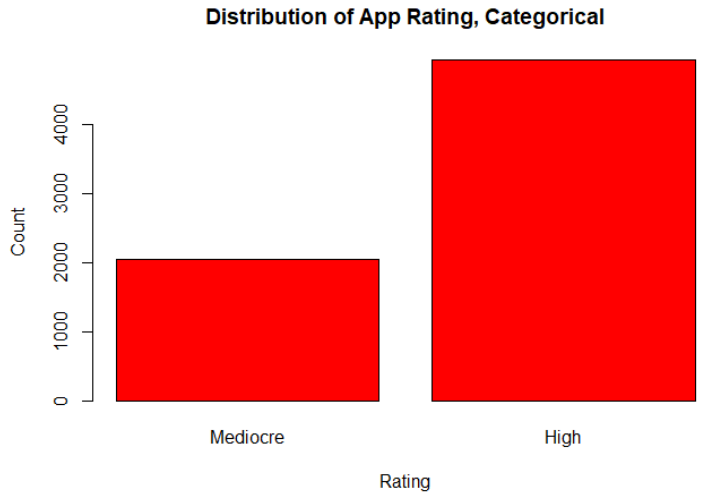
Github link: https://github.com/anthony-le-136/ckme136

Dataset Cleaning

My initial step was identifying attributes that can be easily removed. The App name, Current.Ver, and Last.Updated attributes are unique to each app and therefore not useful for classification. I also removed Genres as it is directly correlated to the Category attribute. After omitting NA and NaN values, the cleaned dataset, labeled **playstore**, consists of 6982 observations of 9 attributes.

Dataset Balancing and Splitting

The distribution of the predictor attribute, Rating, is skewed towards user ratings above 4.0. For the sake of simpler prediction models, I split the continuous data into two categories.



After taking 70% of the data to create a training set, I used the SMOTE function from the DMwR package to oversample the minority class in this set.

Model Building

For the initial model building, I've selected Random Forest and Naïve Bayes to evaluate different binary classifiers. The Mediocre class is taken to be the "positive" class. The following confusion matrices were created after inputting the 30% test set into each model.

Random Forest

Prediction	Mediocre	High
Mediocre	42	62
High	572	1418

Recall = 42/(42 + 62) = 0.40385

Precision = 1418/(1418 + 572) = 0.71256

Naïve Bayes

Prediction	Mediocre	High
Mediocre	278	610
High	336	870

Recall = 278/(278 + 610) = 0.3131

Precision = 870/(870+336) = 0.7214

It is clear the initial recall and precision metrics are subpar. The possible solutions may include using a validation dataset to fine tune the models, or try different classification algorithms.