Domain

I will collect data from kaggle’s dataset Mobile App Statistics (Apple iOS app store). I would like to build up a model to predict users’ rating on new version of an app based on features including its size, price, number of ratings, etc. This model could help developer identify whether users will like this new version or not and what improvement could be made. Because there are many level of ratings, so I decide to stratified it into three levels: excellent(3.5 – 5), fair(2 – 3.5), poor(0-2).

Data

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| size\_millionbytes | float | Storage the app occupies |
| price | float | The amount of charges if the user wants to download the app |
| Rating\_count\_tot | int | Total number of ratings the app has. |
| Rating\_count\_ver | int | Total number of ratings the current version of app has. |
| User\_rating | float | Average user rating value(for all version) |
| User\_rating\_ver | float | Average user rating value for current version |
| Sup\_devices.num | int | Number of supporting devices |
| ipadSc\_urls.num | int | Number of screenshots showed for display |
| Lang.num | int | Number of supported languages |
| Prime\_genre | str | Primary Genre |

Known Unknowns

* I might change my predicted variable when I figure out something more reasonable.
* I might change my dataset when I figure out something more reasonable.