Lazyapp customer review prediction

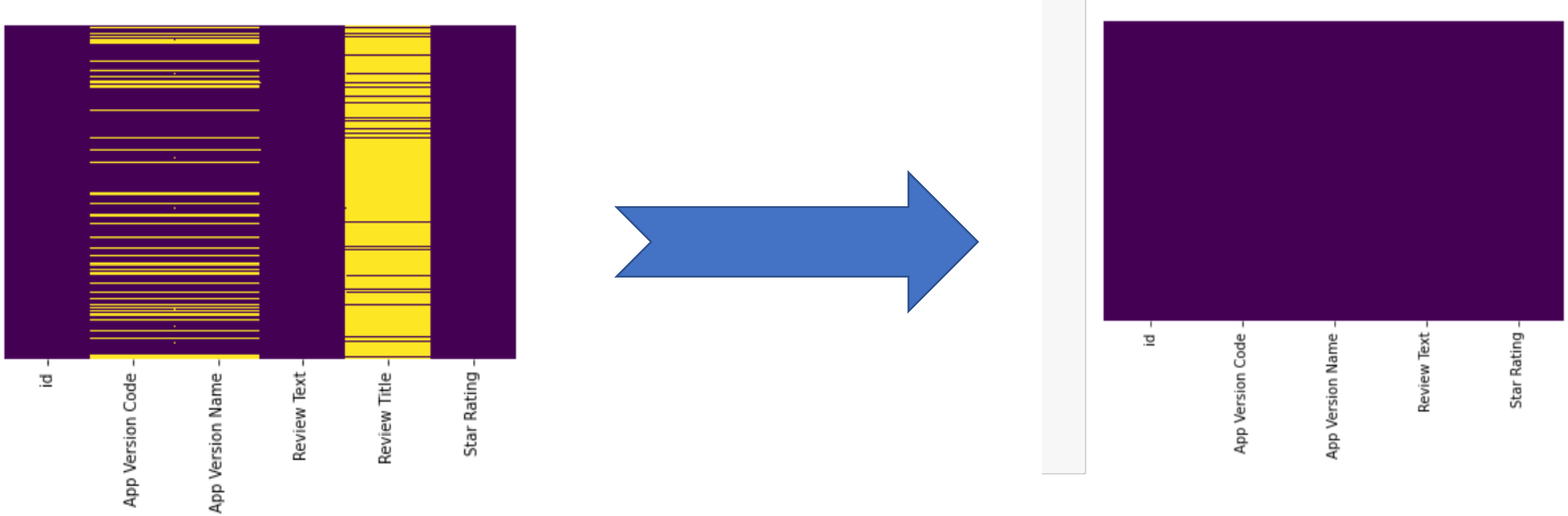
ABHISHEK THOMAS

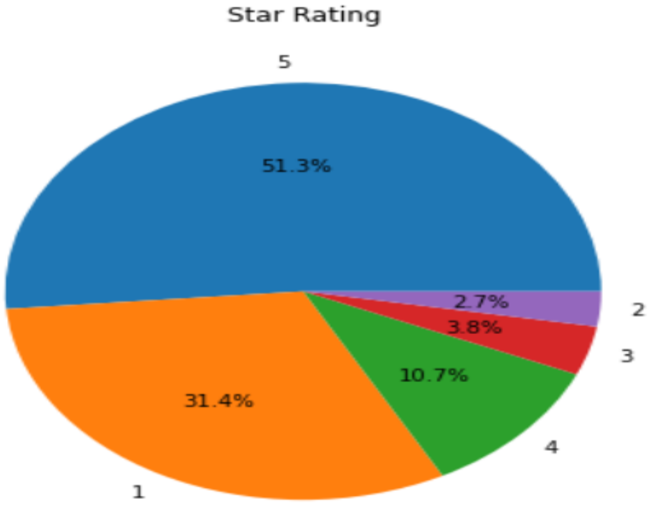
Dataset

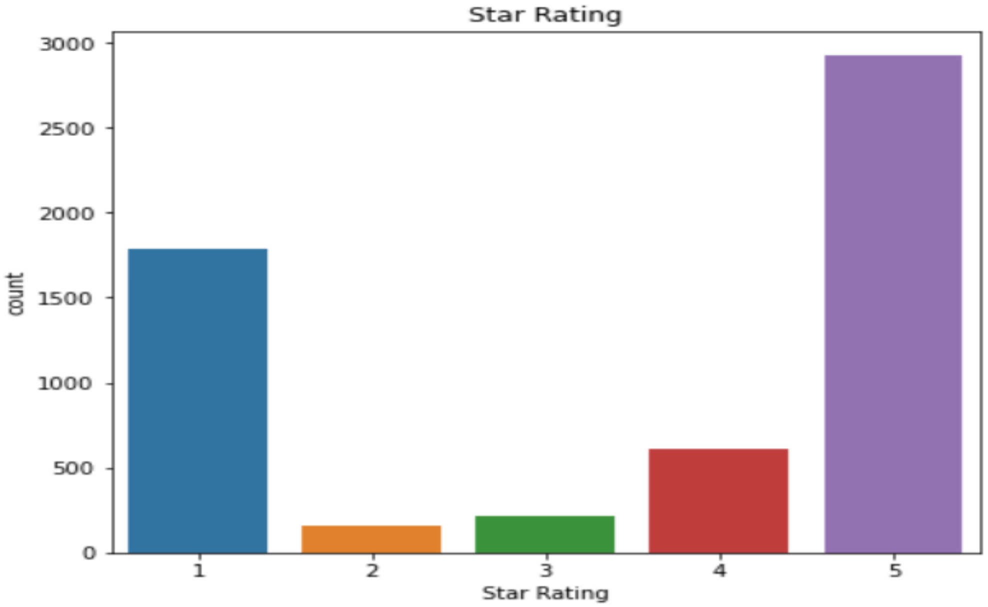
- id
- App version code
- App version type
- Review text
- Review title
- Rating

EDA

- We find there are more than 80% missing values in review titles and app version code and app version type occur are categorical values that occur multiple times for each id , so they are actually related to star reviews
- We find that 51% of star rating are 5 star and only 2.7% have 2 star
- There are review text which are mostly in Hindi-English


id
App Version Code
App Version Name
Review Text
Review Title
Star Rating
id
App Version Code
App Version Name
Review Text
Star Rating


Star Rating
5
51.3%
2
2.7%
3
3.8%
4
10.7%
1
31.4%


Star Rating
count
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
Star Rating

Feature engineering

- We find that that id is unique value and review title has more than 80% null values so we drop review title column
- We see that app version code and app version type occur multiple types rows and it is related to the reviews, so we try to make a new column merging the two values and finding dummy columns for this
- The only other column left is review text which and we can see that most of the reviews are given in hindi, so nlp methods like lemitizaton and stemming wont do any good. However we will remove the tokenize, remove stopwords , remove special characters , tfidf and then use document similarty
- Finally we would merege the dummy columns and document similarity matrix to find there

Final features

- Review text in which the stopwords , special characters , tfidf and document similarity is used
- App_code which was dummy variable and made by merging app version code and app version type

Models used

- KNN – we use this model and also use different parameters to find the best knn model .(ACCURACY – 66.2%)
- Random forest – we use this model and also use differn parameters to find the best random forest model (ACCURACY - 68.8%)

With all the best parameters combined we find that RANDOM FOREST WAS MORE ACCURATE, however we need to deep dive more and see metrics like precision and recall , changing the optimal threshold and reducing the features to OPTIMIZE OUR MODEL AND GIVE BETTER CLASSIFICATION

- SIMILARY WE HAVE TO FEATURE ENGINEER OUR TEST DATASET AND THEN USE RANDOM FOREST TO PREDICT THE MODELS