AI-ML Assignment , Report

Content:

1. Problem Statement
2. Approach's to solve problem
3. Methods and implementation
4. Result and outcome discussion
5. Conclusion

Problem Statement:

Library called 'Kitaab' stores tons of books from long back and now they entered into the digital world to enhance the readers and they have launched the online reading platform, looking for some of interesting features for their platform which are possible by Data science methods

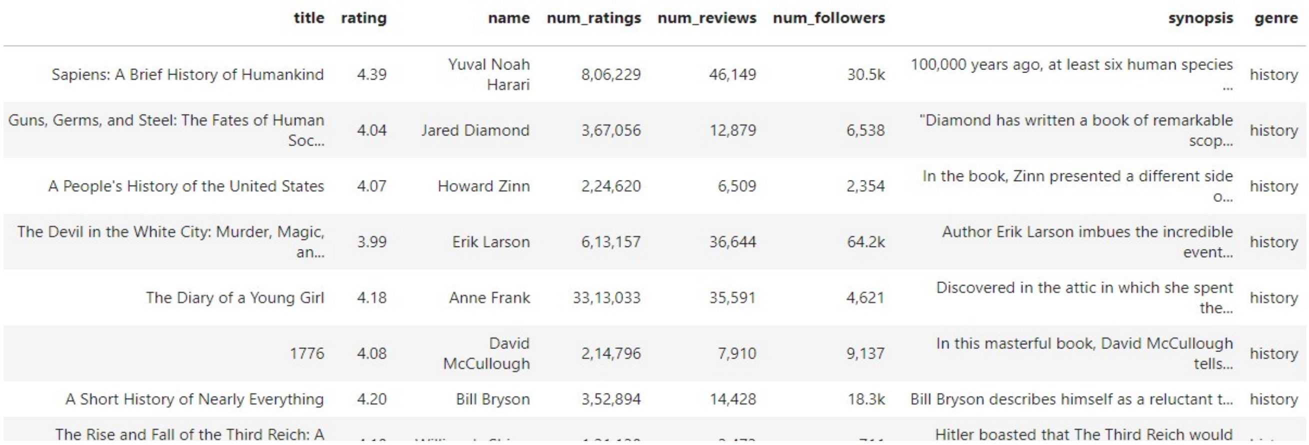
Tasks:

1. Develop ML model which predicts the Genre of the book ,which helps in automatically tag the genre to the uploaded books
2. ML model for book rating prediction which helps in sorting and making tread of books on platform

Tasks 1:

ML model for predicting the Genre of books:

Dataset understanding:Given dataset contains total 1540 data and 8 features includes information about title of the book , name of author,number of ratings,number of reviews ,number of followes,synopsis ,rating and genre



title	rating	name	num_ratings	num_reviews	num_followers	synopsis	genre
Sapiens: A Brief History of Humankind	4.39	Yuval Noah Harari	8,06,229	46,149	30.5k	100,000 years ago, at least six human species ...	history
Guns, Germs, and Steel: The Fates of Human Soc...	4.04	Jared Diamond	3,67,056	12,879	6,538	"Diamond has written a book of remarkable scop...	history
A People's History of the United States	4.07	Howard Zinn	2,24,620	6,509	2,354	In the book, Zinn presented a different side o...	history
The Devil in the White City: Murder, Magic, an...	3.99	Erik Larson	6,13,157	36,644	64.2k	Author Erik Larson imbues the incredible event...	history
The Diary of a Young Girl	4.18	Anne Frank	33,13,033	35,591	4,621	Discovered in the attic in which she spent the...	history
1776	4.08	David McCullough	2,14,796	7,910	9,137	In this masterful book, David McCullough tells...	history
A Short History of Nearly Everything	4.20	Bill Bryson	3,52,894	14,428	18.3k	Bill Bryson describes himself as a reluctant t...	history
The Rise and Fall of the Third Reich: A	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Hitler boasted that The Third Reich would	[illegible]

Fig1: Explains about given dataset

Approach's : Genre of the book mainly derived from the either the book title or synopsis [Which explain main abstract about the book] ,so for this task other than synopsis feature remaining features are dropped as these features were irrelevant to task. Synopsis predictor is a text dataset and we either use traditional encoding methods or NLP methods for text preprocessing.

Main disadvantage of using the encoding techniques like one-hot encoding will create the high dimension preprocessed results which restricts the development of generalized model for datasets and requires complex model also, lacks in extracting meaning full vectors from text dataset. NLP(natural language preprocessing) will overcome all this disadvantages and helps in extracting meaning full vectors with low dimension dense matrix.

Our target contains multiclass outputs and requires classifiers algorithms from classification

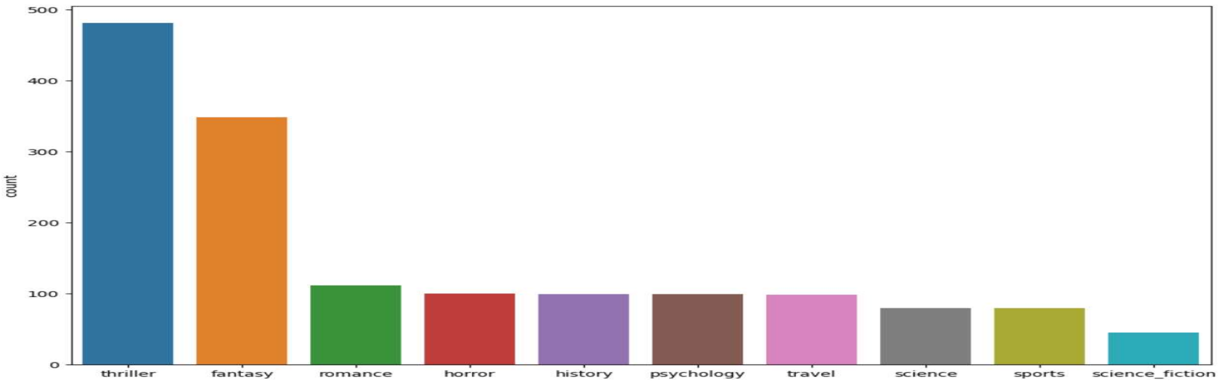
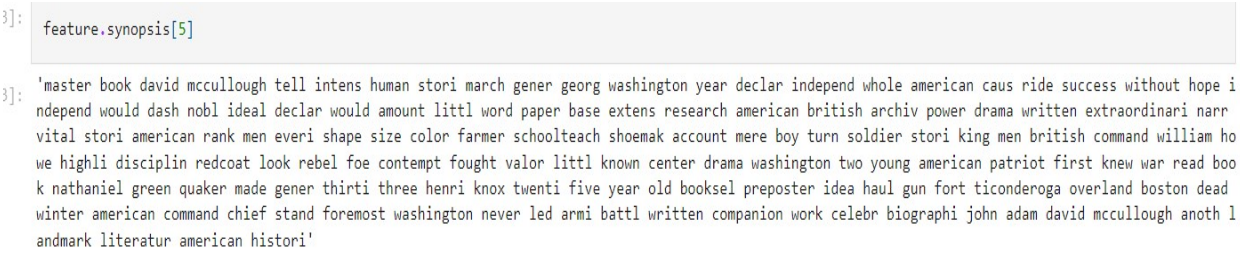


Fig2: Visualization on target dataset[Genre]

Methods and Implementation:

- Text cleaning : Removing the Non alphabetic characters, punctuations and lower the whole text sequence
- Text Preprocessing: normalize text and prepare words and documents for further processing in Machine Learning
  a. Lemmatization: Algorithms extracts more meaning full words from stemmed words
  b. Stemming: Stemming is process of reduce infected word to their word stem
  c. Removing the Stop words: Removing common stopwords like in,the,a,an…
- Encoding [Count vectorization, TFIDF.Word2Vec] :Preprocessed data are vectorized using TFIDF method
- Model training : After splitting the data[80/20] trained with some of base models like [SVM, multiNB and Randomforestclassifier] to high model score out of it
- Hyperparameter tuning and evaluation: Further SVM performed with best score out of other model and tunned the SVM using RandomsearchCV to get best parameters to built final model and evaluated the performance using classification report[confusion matrix]
- Sample testing: Tested the model on random sample from given dataset

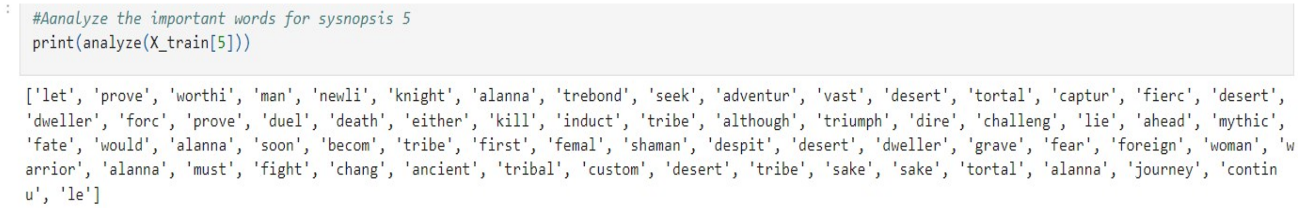
Results:



```
3]: feature.synopsis[5]

3]: 'master book david mccullough tell intens human stori march gener georg washington year declar independ whole american caus ride success without hope i
ndepend would dash nobl ideal declar would amount littl word paper base extens research american british archiv power drama written extraordinari narr
vital stori american rank men everi shape size color farmer schoolteach shoemak account mere boy turn soldier stori king men british command william ho
we highli disciplin redcoat look rebel foe contempt fought valor littl known center drama washington two young american patriot first knew war read boo
k nathaniel green quaker made gener thirti three henri knox twenti five year old booksel preposter idea haul gun fort ticonderoga overland boston dead
winter american command chief stand foremost washington never led armi battl written companion work celebr biographi john adam david mccullough anoth l
andmark literatur american histori'
```

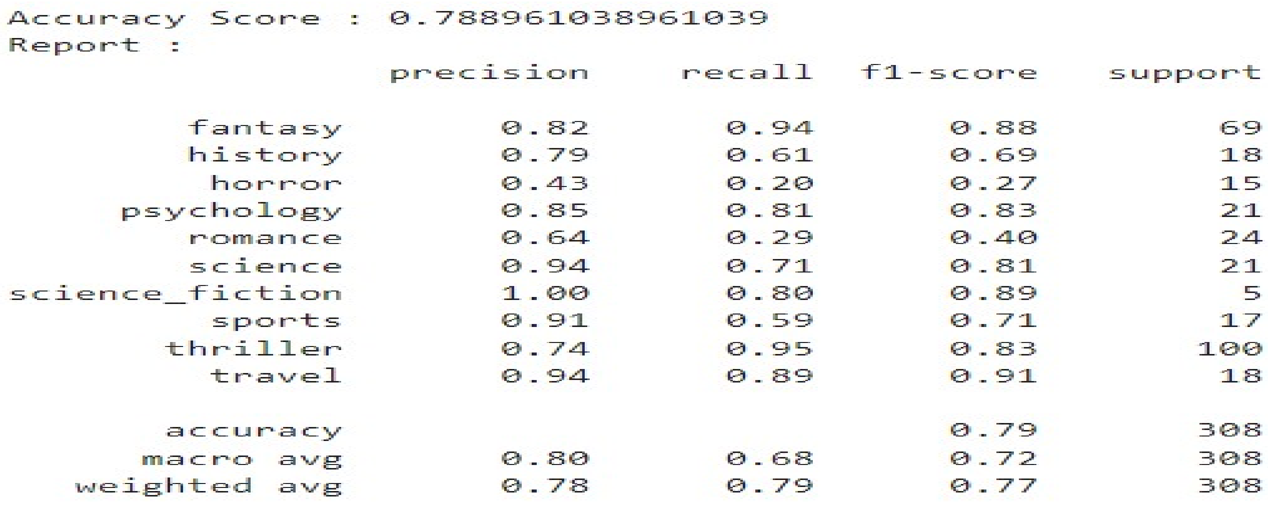
Fig3: Preprocessed text sequence



```
#Aanalyze the important words for sysnopsis 5
print(analyze(X_train[5]))

['let', 'prove', 'worthi', 'man', 'newli', 'knight', 'alanna', 'trebond', 'seek', 'adventur', 'vast', 'desert', 'tortal', 'captur', 'fierc', 'desert',
'dweller', 'forc', 'prove', 'duel', 'death', 'either', 'kill', 'induct', 'tribe', 'although', 'triumph', 'dire', 'challeng', 'lie', 'ahead', 'mythic',
'fate', 'would', 'alanna', 'soon', 'becom', 'tribe', 'first', 'femal', 'shaman', 'despit', 'desert', 'dweller', 'grave', 'fear', 'foreign', 'woman', 'w
arrior', 'alanna', 'must', 'fight', 'chang', 'ancient', 'tribal', 'custom', 'desert', 'tribe', 'sake', 'sake', 'tortal', 'alanna', 'journey', 'contin
u', 'le']
```

Fig4: Important words for vectorization



```
Accuracy Score : 0.788961038961039
Report :
                 precision    recall  f1-score   support

        fantasy       0.82      0.94      0.88        69
        history       0.79      0.61      0.69        18
         horror       0.43      0.20      0.27        15
     psychology       0.85      0.81      0.83        21
        romance       0.64      0.29      0.40        24
        science       0.94      0.71      0.81        21
science_fiction       1.00      0.80      0.89         5
         sports       0.91      0.59      0.71        17
       thriller       0.74      0.95      0.83       100
         travel       0.94      0.89      0.91        18

       accuracy                           0.79       308
      macro avg       0.80      0.68      0.72       308
   weighted avg       0.78      0.79      0.77       308
```

Fig5: Final model accuracy score for tunned parameters

Task2:

<u>Model to predict the Rating of the book</u>

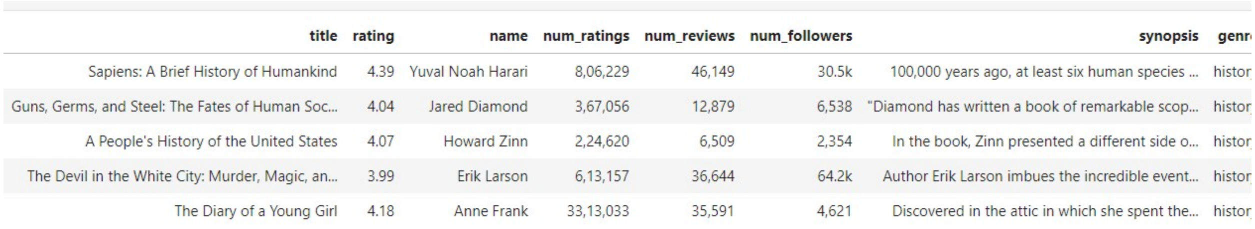
Approach: Ratings of the book were continues values in the range of 0 – 5 ,so regressor algorithms helps in predicting the ratings.Ratings mainly depends on number of ratings,type of genre,book,authors and synopsis feature is somewhat irrelevant to predict the ratings from other available features.So synopsis is dropped.

Methods and Implementation:

- Simple EDA on datasets to get statistical understanding

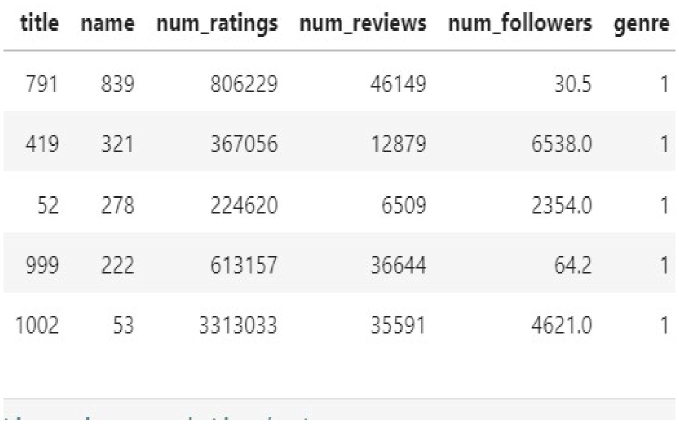
- Feature Engineering : Manipulation on NaN values ,converting objects types to numerical after cleaning delimetors,Encoding the categorical vlaues
- Feature selection : using correlation heatmap method
- Model training :base models [linear regressor,Rf regressor,Adaboost,XGBoost]
- Hyperparameter tunning and evaluation

Results:



title	rating	name	num_ratings	num_reviews	num_followers	synopsis	genr
Sapiens: A Brief History of Humankind	4.39	Yuval Noah Harari	8,06,229	46,149	30.5k	100,000 years ago, at least six human species ...	histor
Guns, Germs, and Steel: The Fates of Human Soc...	4.04	Jared Diamond	3,67,056	12,879	6,538	"Diamond has written a book of remarkable scop...	histor
A People's History of the United States	4.07	Howard Zinn	2,24,620	6,509	2,354	In the book, Zinn presented a different side o...	histor
The Devil in the White City: Murder, Magic, an...	3.99	Erik Larson	6,13,157	36,644	64.2k	Author Erik Larson imbues the incredible event...	histor
The Diary of a Young Girl	4.18	Anne Frank	33,13,033	35,591	4,621	Discovered in the attic in which she spent the...	histor

Fig5: Dataset Before Preprocessing



title	name	num_ratings	num_reviews	num_followers	genre
791	839	806229	46149	30.5	1
419	321	367056	12879	6538.0	1
52	278	224620	6509	2354.0	1
999	222	613157	36644	64.2	1
1002	53	3313033	35591	4621.0	1

Fig6:Dataset after preprocess

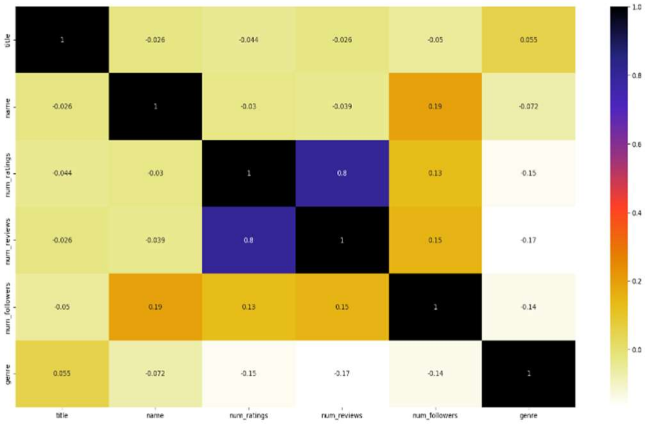


Fig7:Correlation heatmap for feature selection

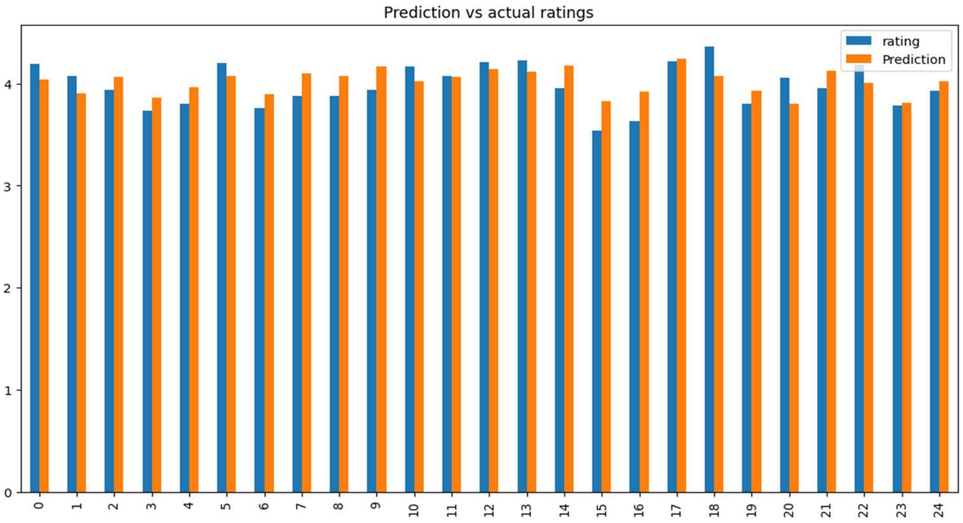


Fig8: Prediction and actual value visualization

Conclusion:

Task1 : In predicting Genre final model score was around 79% further it can be increased by adding some data into dataset and training the model or using transfer learning model for building complex model on text data

Task2: In predicting Rating XGB model results with good output. Performance can be increased by adding some more data and scaling techniques