Movie Rating Score Prediction

Chengyang Miao
Boston University
Boston, MA
cymiao@bu.edu

Zhitao Gan
Boston University
Boston, MA
owengan@bu.edu

Abstract

The goal of this project is to predict the movie rating from Amazon Movie Reviews by using the available features. We first applied different models, such as **K-Nearest Neighbor**, **Decision tree** and **Multinomial Naive Bayes**, to numerical data and text data separately to test which model results in the best accuracy. Followed by that, we came up with two methods, **Voting system** and **text representation of number**, to make both numerical data and text data engage in our predictions.

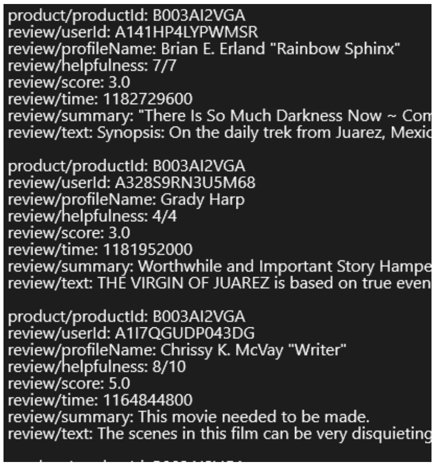
1 Data preparation

1.1 Data conversion

We download the data from

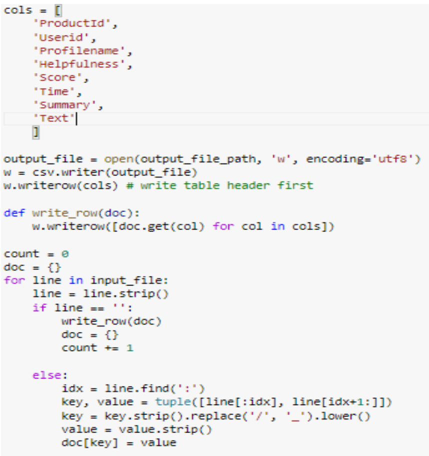
https://snap.stanford.edu/data/web-Movies.html

We started by transforming the text file to csv file. We notice that the format of the raw text data (**Figure 1**) is not supported by the built-in conversion function in pandas. We wrote our own function to handle it (**Figure 2**). **Figure 3** is the final outlook of the CSV file obtained by using our function. Luckily, the number of rows containing **NaN values** in our data set are small enough to be dropped, so we drop all the **NaN values**.



```
product/productId: B003AI2VGA
review/userId: A141HP4LYPWMSR
review/profileName: Brian E. Erland "Rainbow Sphinx"
review/helpfulness: 7/7
review/score: 3.0
review/time: 1182729600
review/summary: "There Is So Much Darkness Now ~ Con
review/text: Synopsis: On the daily trek from Juarez, Mexic

product/productId: B003AI2VGA
review/userId: A328S9RN3U5M68
review/profileName: Grady Harp
review/helpfulness: 4/4
review/score: 3.0
review/time: 1181952000
review/summary: Worthwhile and Important Story Hampe
review/text: THE VIRGIN OF JUAREZ is based on true even

product/productId: B003AI2VGA
review/userId: A1I7QGUDP043DG
review/profileName: Chrissy K. McVay "Writer"
review/helpfulness: 8/10
review/score: 5.0
review/time: 1164844800
review/summary: This movie needed to be made.
review/text: The scenes in this film can be very disquieting
```

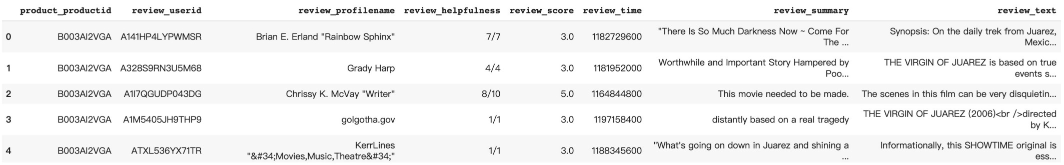
Figure 1: Raw text data



```
cols = [
    'ProductId',
    'Userid',
    'Profilename',
    'Helpfulness',
    'Score',
    'Time',
    'Summary',
    'Text'
    ]

output_file = open(output_file_path, 'w', encoding='utf8')
w = csv.writer(output_file)
w.writerow(cols) # write table header first

def write_row(doc):
    w.writerow([doc.get(col) for col in cols])

count = 0
doc = {}
for line in input_file:
    line = line.strip()
    if line == '':
        write_row(doc)
        doc = {}
        count += 1

    else:
        idx = line.find(':')
        key, value = tuple([line[:idx], line[idx+1:]])
        key = key.strip().replace('/', '_').lower()
        value = value.strip()
        doc[key] = value
```

Figure 2: Conversion function

NeurIPS 2021 AI for Science Workshop.



	product_productid	review_userid	review_profilename	review_helpfulness	review_score	review_time	review_summary	review_text
0	B003AI2VGA	A141HP4LYPWMSR	Brian E. Erland "Rainbow Sphinx"	7/7	3.0	1182729600	"There Is So Much Darkness Now ~ Come For The ...	Synopsis: On the daily trek from Juarez, Mexic...
1	B003AI2VGA	A328S9RN3U5M68	Grady Harp	4/4	3.0	1181952000	Worthwhile and Important Story Hampered by Poo...	THE VIRGIN OF JUAREZ is based on true events s...
2	B003AI2VGA	A1I7QGUDP043DG	Chrissy K. McVay "Writer"	8/10	5.0	1164844800	This movie needed to be made.	The scenes in this film can be very disquietin...
3	B003AI2VGA	A1M5405JH9THP9	golgotha.gov	1/1	3.0	1197158400	distantly based on a real tragedy	THE VIRGIN OF JUAREZ (2006)<br />directed by K...
4	B003AI2VGA	ATXL536YX71TR	KerrLines "&#34;Movies,Music,Theatre&#34;"	1/1	3.0	1188345600	"What's going on down in Juarez and shining a ...	Informationally, this SHOWTIME original is ess...

Figure 3: loaded cvs overview

1.2 feature extraction

There are 7 features in this data set:

ProductID - unique identifier for the product

UserID - unique identifier for the user

ProfileName - name of the user

Helpfulness -fraction of users who found the review helpful

Time - timestamp for the review

Summary - brief summary of the review

Text - text of the review

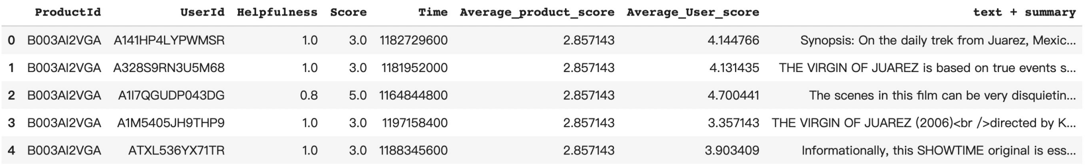
There are 3 useful features in the raw data frame. They are **Text**, **Summary**, **Helpfulness**. We decided to add two new features – **Average product scores** and **Average user scores**.

People's opinions on movie are subjective mainly depending on personal preferences and experiences. However,

1. logistically good movies will be appealing to the majority of audiences. This means that the average scores of good movies are likely to have higher scores whereas those of bad movies are likely to have lower scores. With respect to this, we added the first feature – **Average product scores** which represents **The average scores of the product**.

2. Users who have higher standard or have a special taste in movies will tend to give low scores which results in low average scores. On the contrary, those who enjoy watching different kinds of movies and have low standard will tend to give high scores which results in high average scores. With respect to this, we added our second feature – **Average user scores** which represents **The average scores given by the users**

In addition, We merge the **Text** column and **Summary**, because it is easier to process text data in one column. The final data set looks like **Figure 4**



	ProductId	UserId	Helpfulness	Score	Time	Average_product_score	Average_User_score	text + summary
0	B003AI2VGA	A141HP4LYPWMSR	1.0	3.0	1182729600	2.857143	4.144766	Synopsis: On the daily trek from Juarez, Mexic...
1	B003AI2VGA	A328S9RN3U5M68	1.0	3.0	1181952000	2.857143	4.131435	THE VIRGIN OF JUAREZ is based on true events s...
2	B003AI2VGA	A1I7QGUDP043DG	0.8	5.0	1164844800	2.857143	4.700441	The scenes in this film can be very disquietin...
3	B003AI2VGA	A1M5405JH9THP9	1.0	3.0	1197158400	2.857143	3.357143	THE VIRGIN OF JUAREZ (2006)<br />directed by K...
4	B003AI2VGA	ATXL536YX71TR	1.0	3.0	1188345600	2.857143	3.903409	Informationally, this SHOWTIME original is ess...

Figure 4: data

2 Analysis on numerical columns

We first tried to see if dimension reduction could be applied since it will help us reduce the size of the data. We applied **Standardscaler** method to normalize the data. Then we used **pca** technique to plot a graph (**Figure 5**) with $X - axis$ as $principlecomponents$ and $Y - axis$ as $explainedvariance$. Based on the graph, we could observe that most of the components have high variances which means that they all have significant information. Therefore, we concluded that dimension reduction could not be applied.

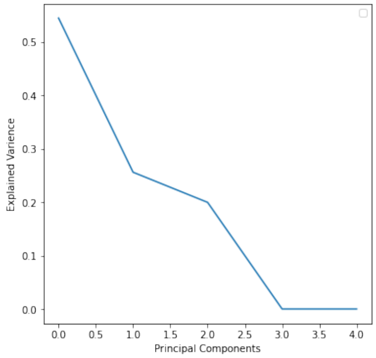


Figure 5: standard scalar.

2.1 KNN model

We first tried to find the best parameters of K value in this model. We used a **for loop** to obtain a list of accuracy by assigning k from 1 to 30. Then we plot a graph (**Figure 6**) to find out which k has the best accuracy. As shown on the graph, the accuracy starts to plateau at $k = 10$. Therefore, we built our KNN model with $k = 10$. We obtained our prediction results and plotted a graph as shown in **Figure 7**. We could see that KNN model does a good job on predicting the scores generating an accuracy of 0.7446. Through further observation based on the graph, we found out that it performed poorly on predicting Score 2 and Score 3 having an accuracy below 0.5.

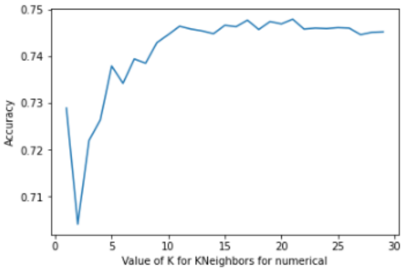
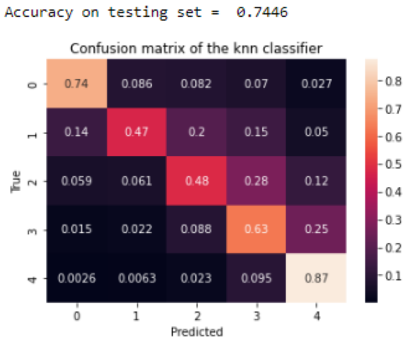


Figure 6: accuracy of value k from 1 to 30



```
Accuracy on testing set =  0.7446
```

Figure 7: accuracy of value k from 1 to 30

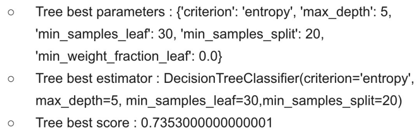
2.2 Decision tree model

The second model we chose was **decision tree model**. To obtain the best parameters for it, we used a method called **GridSearchCV** which is a common method to find the best parameters of model with multiple parameters. The outcome is shown in **Figure 8**. We then fit our train data into the model where we set the parameters according to GridSearchCV output. The result is shown in **Figure 9**. The accaracy of **decision tree model**is 0.7371 which is a bit lower than **knn model**. The similarity is that it also has bad performance on the prediction of Score 2 and Score 3.

3 Analysis on Text data

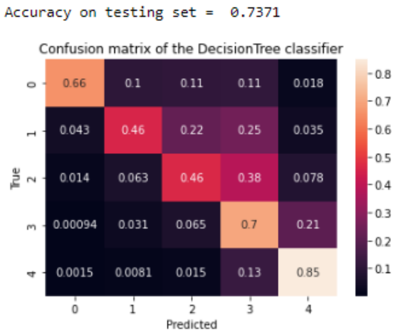
3.1 Preprocessing text

We first cleaned the text by using our own function **Figure 10**. There are three steps in our function:



- Tree best parameters : {'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 30, 'min_samples_split': 20, 'min_weight_fraction_leaf': 0.0}
- Tree best estimator : DecisionTreeClassifier(criterion='entropy', max_depth=5, min_samples_leaf=30,min_samples_split=20)
- Tree best score : 0.7353000000000001

Figure 8: GridSearchCV output



```
Accuracy on testing set =  0.7371
```

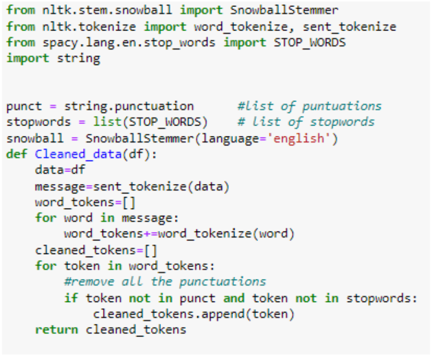
Confusion matrix of the DecisionTree classifier

True \ Predicted	0	1	2	3	4
0	0.66	0.1	0.11	0.11	0.018
1	0.043	0.46	0.22	0.25	0.035
2	0.014	0.063	0.46	0.38	0.078
3	0.00094	0.031	0.065	0.7	0.21
4	0.0015	0.0081	0.015	0.13	0.85

Figure 9: confusion matrix of Decision tree model

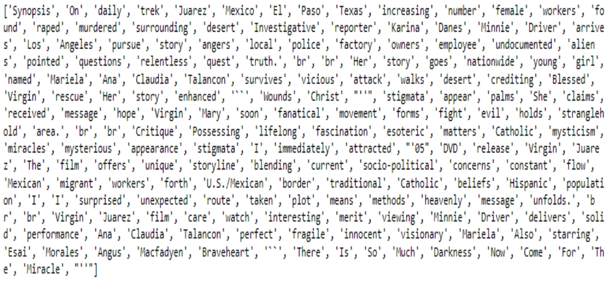
1. **tokenize the sentence**
2. **remove all the punctuation**
3. **remove all the stopwords**

The outcome of the text passing through the function will result in a list of tokens of words with stopwords removed as shown in **Figure 11**. Then we fit them into a method named **Tfidf** which stands for **term frequency**−**inverse document**. This is a common technique to transform text into a meaning representation of numbers (**Figure 12**).



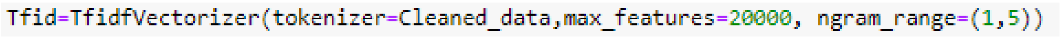
```python
from nltk.stem.snowball import SnowballStemmer
from nltk.tokenize import word_tokenize, sent_tokenize
from spacy.lang.en.stop_words import STOP_WORDS
import string


punct = string.punctuation      #list of puntuations
stopwords = list(STOP_WORDS)    # list of stopwords
snowball = SnowballStemmer(language='english')
def Cleaned_data(df):
    data=df
    message=sent_tokenize(data)
    word_tokens=[]
    for word in message:
        word_tokens+=word_tokenize(word)
    cleaned_tokens=[]
    for token in word_tokens:
        #remove all the punctuations
        if token not in punct and token not in stopwords:
            cleaned_tokens.append(token)
    return cleaned_tokens
```

Figure 10: function for cleaning text



```
['Synopsis', 'On', 'daily', 'trek', 'Juarez', 'Mexico', 'El', 'Paso', 'Texas', 'increasing', 'number', 'female', 'workers', 'fo
und', 'raped', 'murdered', 'surrounding', 'desert', 'Investigative', 'reporter', 'Karina', 'Danes', 'Minnie', 'Driver', 'arrive
s', 'Los', 'Angeles', 'pursue', 'story', 'angers', 'local', 'police', 'factory', 'owners', 'employee', 'undocumented', 'alien
s', 'pointed', 'questions', 'relentless', 'quest', 'truth.', 'br', 'br', 'Her', 'story', 'goes', 'nationwide', 'young', 'girl',
'named', 'Mariela', 'Ana', 'Claudia', 'Talancon', 'survives', 'vicious', 'attack', 'walks', 'desert', 'crediting', 'Blessed',
'Virgin', 'rescue', 'Her', 'story', 'enhanced', '``', 'Wounds', 'Christ', "''", 'stigmata', 'appear', 'palms', 'She', 'claims',
'received', 'message', 'hope', 'Virgin', 'Mary', 'soon', 'fanatical', 'movement', 'forms', 'fight', 'evil', 'holds', 'strangleh
old', 'area.', 'br', 'br', 'Critique', 'Possessing', 'lifelong', 'fascination', 'esoteric', 'matters', 'Catholic', 'mysticism',
'miracles', 'mysterious', 'appearance', 'stigmata', 'I', 'immediately', 'attracted', "'05", 'DVD', 'release', 'Virgin', 'Juare
z', 'The', 'film', 'offers', 'unique', 'storyline', 'blending', 'current', 'socio-political', 'concerns', 'constant', 'flow',
'Mexican', 'migrant', 'workers', 'forth', 'U.S./Mexican', 'border', 'traditional', 'Catholic', 'beliefs', 'Hispanic', 'populati
on', 'I', 'I', 'surprised', 'unexpected', 'route', 'taken', 'plot', 'means', 'methods', 'heavenly', 'message', 'unfolds.', 'b
r', 'br', 'Virgin', 'Juarez', 'film', 'care', 'watch', 'interesting', 'merit', 'viewing', 'Minnie', 'Driver', 'delivers', 'soli
d', 'performance', 'Ana', 'Claudia', 'Talancon', 'perfect', 'fragile', 'innocent', 'visionary', 'Mariela', 'Also', 'starring',
'Esai', 'Morales', 'Angus', 'Macfadyen', 'Braveheart', '``', 'There', 'Is', 'So', 'Much', 'Darkness', 'Now', 'Come', 'For', 'Th
e', 'Miracle', "''"]
```

Figure 11: outcome of text



```python
Tfid=TfidfVectorizer(tokenizer=Cleaned_data,max_features=20000, ngram_range=(1,5))
```

Figure 12: TfidfVectorizer

3.2 Analyze on text

After preprocessing the text data, we started to look for good models. The first two models that we tried were **LinearSVC** and **KNN**. However, the outcomes are surprising low as shown in **Figure 14 and Figure 15**. Therefore, we tried another model – **Multinomial Naive Bayes**. This model gives us an accuracy of 0.7371 which is a lot higher compared with **LinearSVC** and **KNN**(**Shown in figure 15**). However, the prediction of this model on text has the same problem which is poor performance on the class 2 and class 3.

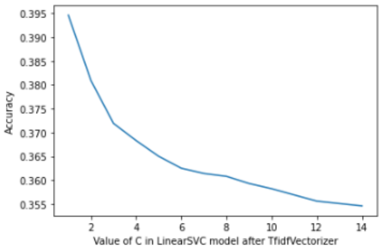


Figure 13: function for cleaning text

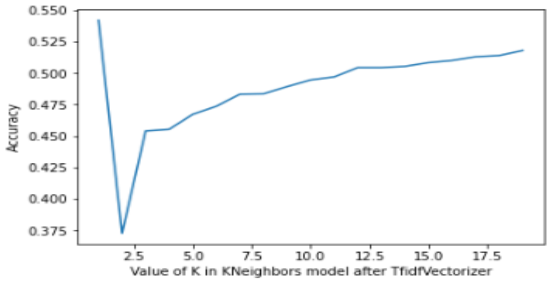
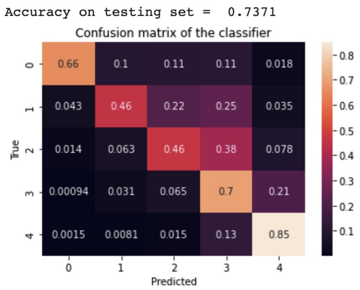


Figure 14: outcome of text



```
Accuracy on testing set =  0.7371
```

Figure 15: Multinomial NB model

4 Further approach on the combination of text and numerical data

4.1 Voting system

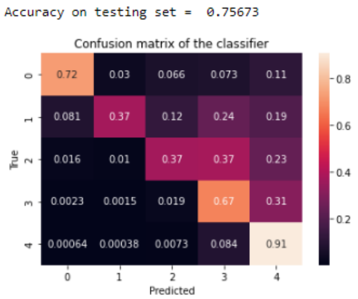
This approach works as following:

1. Make three sets of predictions based on three models

2. The score predicted by the most models (more than half) is the final prediction

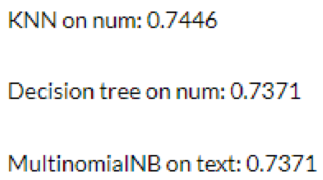
3. If predictions are all different, use the prediction of MultinomialNB

We thought this **Voting system** would improve the accuracy because of the idea that if one model makes a wrong prediction, the other two could overwrite it. We use **KNN** model and **decision tree** model on numerical data and **MultinomialNB** model on text data. Then we make the three predictions vote. The prediction that gets the most vote will be our final decision. The result is shown below. It does improve the accuracy exceeding all the previous model (**Shown in Figure 16 and 17**). However, by observing the confusion matrix, we found that it has higher general accuracy because it has higher accuracy predicting class 1, 4, and 5. It has worse performance on class 2 and 4 compared to other models.



```
Accuracy on testing set =  0.75673
```

Figure 16: Outcome of voting system

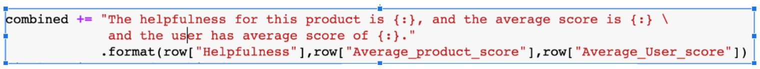


```
KNN on num: 0.7446

Decision tree on num: 0.7371

MultinomialNB on text: 0.7371
```

Figure 17: Existing result

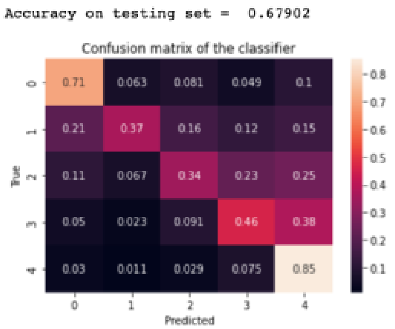
4.2 Representation of number in text form

We came up with this method where the numerical data is expressed in a text sentence and then added it to the beginning of the text (**Shown in Figure 18**). Then the predictions are made based on what we followed in text analysis. We were hoping that this adds the influence of numerical data into the text model. However, this results in a decrease in accuracy (**Figure 19 and 20**).



```
combined += "The helpfulness for this product is {:}, and the average score is {:} \
             and the user has average score of {:}."
            .format(row["Helpfulness"],row["Average_product_score"],row["Average_User_score"])
```

Figure 18: Result of combined data analyze

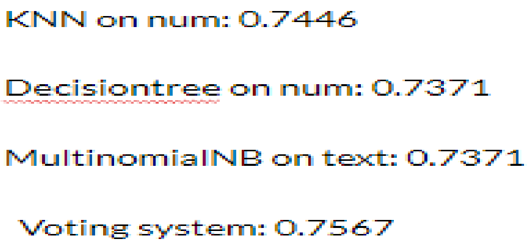


```
Accuracy on testing set =  0.67902
```

Confusion matrix of the classifier

True \ Predicted	0	1	2	3	4
0	0.71	0.063	0.081	0.049	0.1
1	0.21	0.37	0.16	0.12	0.15
2	0.11	0.067	0.34	0.23	0.25
3	0.05	0.023	0.091	0.46	0.38
4	0.03	0.011	0.029	0.075	0.85

Figure 19: Outcome of voting system



```
KNN on num: 0.7446

Decisiontree on num: 0.7371

MultinomialNB on text: 0.7371

 Voting system: 0.7567
```

Figure 20: Existing result

5 Summary

By applying different models to numerical data(**K-Nearest Neighbor**, **Decision tree**) and text data (**K-Nearest Neighbor**, **LinearSVC**, **MultinomialNB**), we concluded that model selection is really important, especially when it comes to text data. To find the best model for the data set, multiple attempts on different models are crucial and necessary. In addition, Innovation in your own techniques, such as voting system we came up with, based on your own knowledge and thorough observation of the data set can also make a huge difference!

References

[1] J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013.