Alma Mater Studiorum
University of Bologna

Duplicate Question Pair

Course: Machine Learning Mini Project

Presented by Priti Kumari Gupta

Contents

1 Introduction

In today's digital age, where information is readily available at our fingertips, question-and answer platforms have become increasingly popular. Quora is one of the popular platforms where users can ask questions and receive answers from a diverse community. Over 100 million people visit Quora every month, so it's no surprise that people ask similar-worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. Quora uploaded a competition on Kaggle where users were challenged to tackle the problem of duplicate questions having different threads.

This project aims to detect duplicate question pairs on Quora by utilizing the power of machine learning techniques. The main aim of the project is to develop a model capable of accurately identifying and classifying duplicate question pairs, automating the process, and improving the overall user experience on Quora. At present, Quora utilizes a Random Forest model for detecting duplicate questions. Following a similar approach, we initially implemented a Random Forest model to evaluate its performance on the dataset. Later implement xgboost, Logistic Regression, Naive Bayes to check the performance.

2 Data

After Importing the Library and dataset, we performed a preliminary Exploratory Data Analysis to gather insights and understand its characteristics. Quora provided two datasets, the train and test datasets. The training dataset contained approximately 400,000 pairs of questions, while the test dataset comprised a set of 1 million question pairs. The training dataset contains real-time data whereas the test data contains computer-generated data. The primary focus has been on the train set as those are the real-time values. Quora, as mentioned on Kaggle, has acknowledged that the labeling of the 'is_duplicate' column is performed by humans, resulting in a certain degree of inconsistency. It is important to consider, the ground truth labels on this dataset as being knowledgeable but not entirely accurate, as they may contain instances of incorrect labeling. More over, it was also indicated that the labels, on the whole, represent reasonable agreement, this may frequently do not hold true on a case-by-case basis for individual items in the dataset. The training dataset has five columns, including an ID column, two columns representing the ID of the question set 1 and question set 2, two columns containing the text of question 1 and

question 2 respectively, and 2 fields indicating duplicates. The dataset consists of the following columns:

- **Id**: This column represents a unique identifier for each question pair.
- **qid1**: This column contains the ID of the first question in the pair.
- **qid2**: This column contains the ID of the second question in the pair.
- **question1,question2**: This column includes the text of the first question in the pair. question2: This column includes the text of the second question in the pair.
- **is_duplicate:**: This column indicates whether the question pair is classified as a duplicate or not. It typically contains binary values (0 or 1) where 1 denotes a duplicate pair and 0 denotes a non-duplicate pair.

```
[4] df.head()
```

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0

Figure 1: Distribution

The main objective of the project is to identify and assess the level of duplication between the two below-given questions. From the above analysis, it can be concluded that a binary classification problem is present in which we will be presented with two questions and our objective is to determine whether they are duplicates or not. After analyzing the dataset, it was found that there are few null values. One value was missing in the question 1 column whereas two more values were missing in the question 2 column.

The null values in the dataset were considered negligible in comparison to its overall dataset size, hence we dropped the rows completely. further,it was also checked if there were more than one row exactly similar to each other, and found that there were none like this. Additionally, a few more analyses were conducted to check how

```
#to check the null value
df.isnull().sum()
```

```
id              0
qid1            0
qid2            0
question1       1
question2       2
is_duplicate    0
dtype: int64
```

Figure 2: Null value

many question pairs were duplicates and how many were not, As a result, it was found that out of the total question pairs, 255,024 were identified as non-duplicates, while 149,263 were identified as duplicates. The percentage of non-duplicates and duplicates are 63.07 and 36.92 respectively.

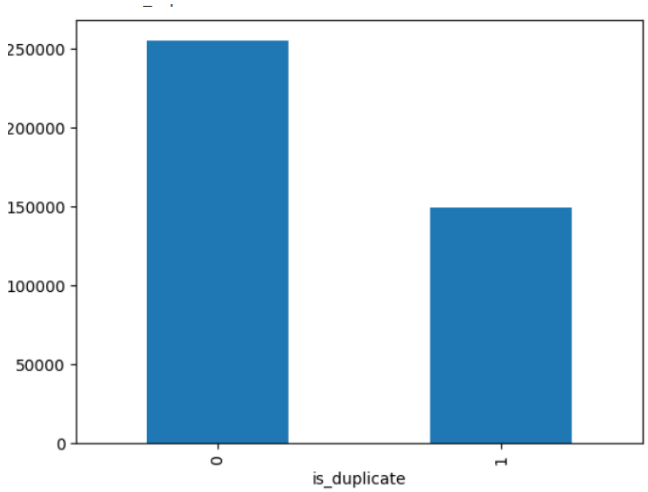

Figure 3: is_null distribution

The next target is to find the total number of unique and repeated questions out of 800K questions and found that 537929 were unique and 111778 were repeated. Additionally, to gain clarity on the frequency of repetition for the repeated questions few more analysis was performed. From the above distribution graph, we can conclude that there is one question that is repeated 160 times, one got repeated

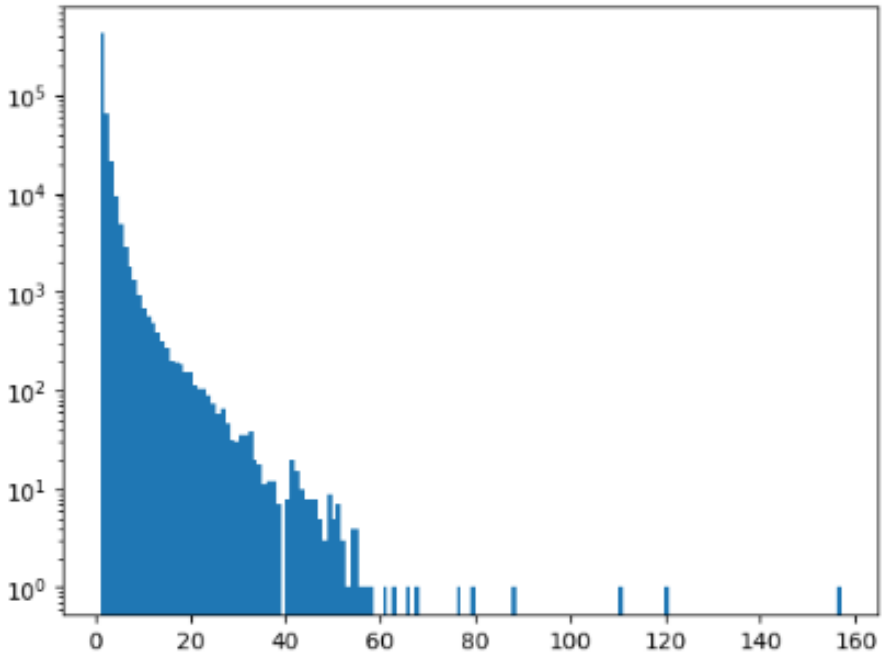

Figure 4: Distribution

120 times, and so on whereas the unique questions have the highest number. From Kaggle It was found the other people working on the project have identified the occurrence of the word/pattern 'MATH' approximately 900 times.

3 System Description

we have implemented random forest on our dataset to understand the performance. To implement the Random Forest model, we conducted feature engineering on the dataset. This involved extracting new features from the existing data. The features added during the feature engineering process are mentioned below with details:

- **q1_len**: char length of question1.

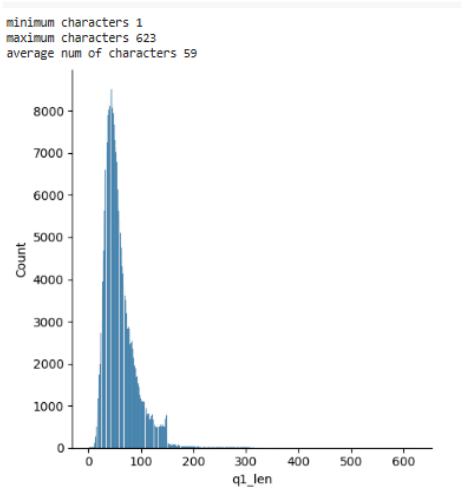

Figure 5: q1_length

- **q2_len**: char length of question2.

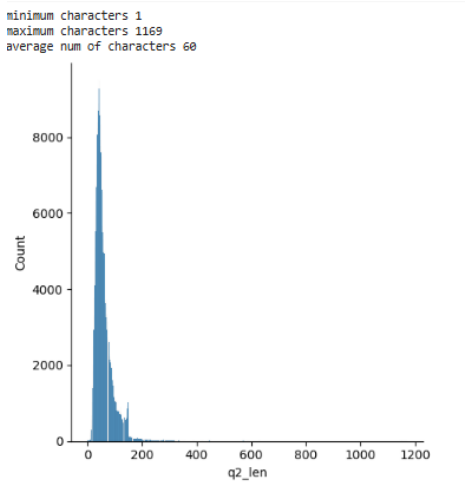

Figure 6: q2_length

- **q1_num_words**: Number of words in question1 obtained by splitting the questions in the question1 column using ” ”.

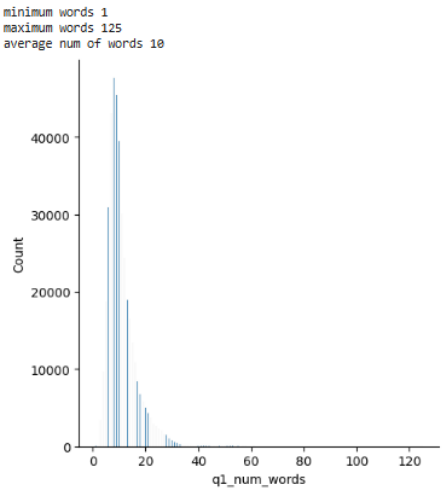

Figure 7: q1_num_word

- **q2_num_words**: Number of words in question2 obtained by splitting the questions in the question2 column using ” ”.

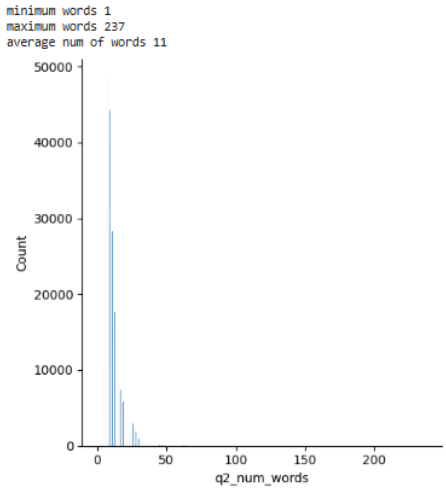

Figure 8: q2_num_word

- **word_common**: number of common unique words obtained by taking the intersection of the set of collection of all words in question1 and question2

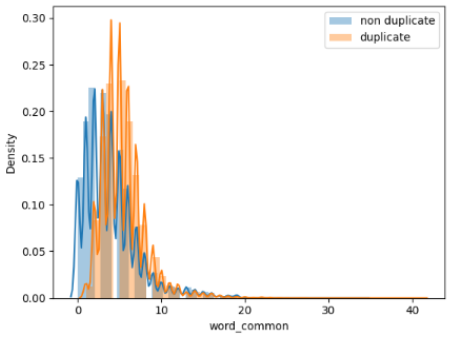

Figure 9: word_common

- **word_total**: sum of the total number of words in question1 and question2 obtained by taking the sum of a set of collections of all words in question1 and question2.

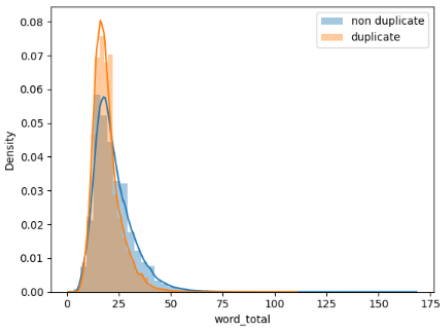

Figure 10: word_total

- **word_share**: word_common/word_total

The two columns question1 and question2 were dropped to evaluate the performance of the two models xgboost and random forest. Later the performance of the models was noted on the few features of the question pair set.

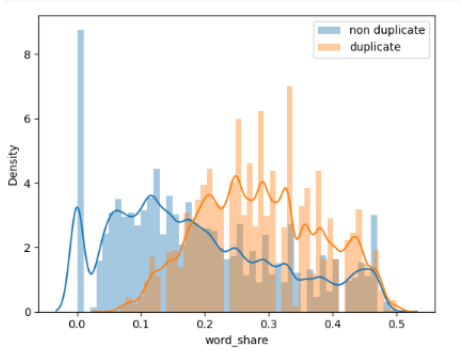

Figure 11: word_share

4 Models

4.1 RandomForest

After we applied feature engineering to extract some features from the data. We implemented Random Forest as this is the model Quora uses currently and we wanted to check its performance with our approach. The Random Forest gives an accuracy of 77.35 and F1 score of 67.9 respectively.

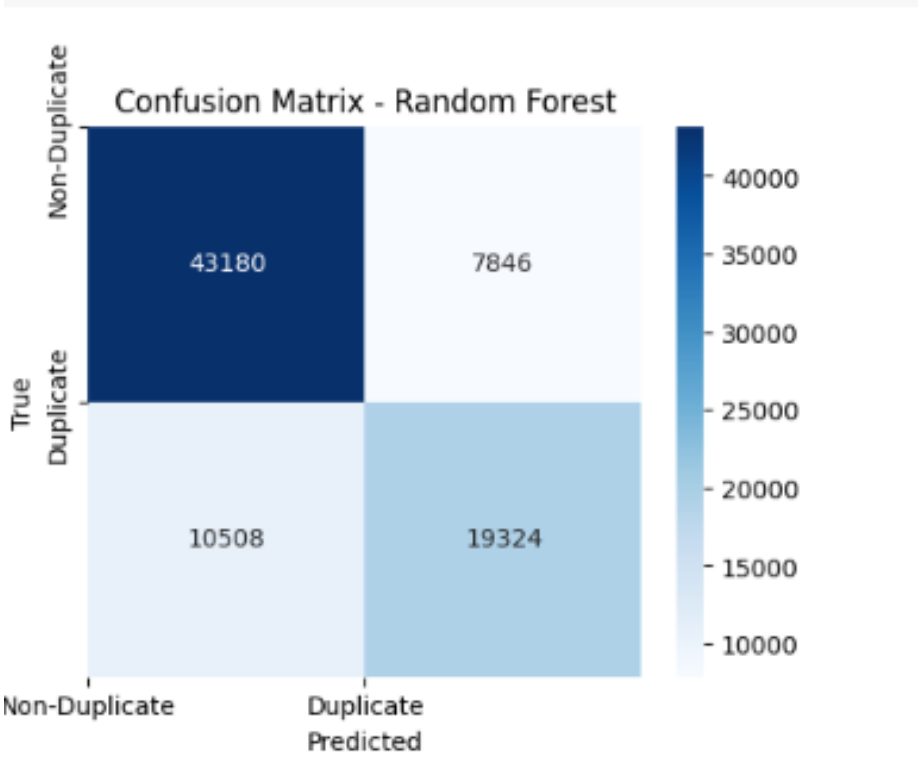

Figure 12: RandomForest Confusion Matrix

4.2 XGBoost

After the RandomForest We chose XGBoost to implement on the model as it is capable of handling complex, non-linear relationships between features and can effectively capture interactions between the question features and identify important patterns that contribute to question similarity. It is also known for being able to handle a large number of features and automatically learn feature interactions, making it suitable for text-based classification tasks. We implemented it to the same dataset and as expected we got slightly better results(accuracy 78.37 — f1score- 69.1) than RandomForest.

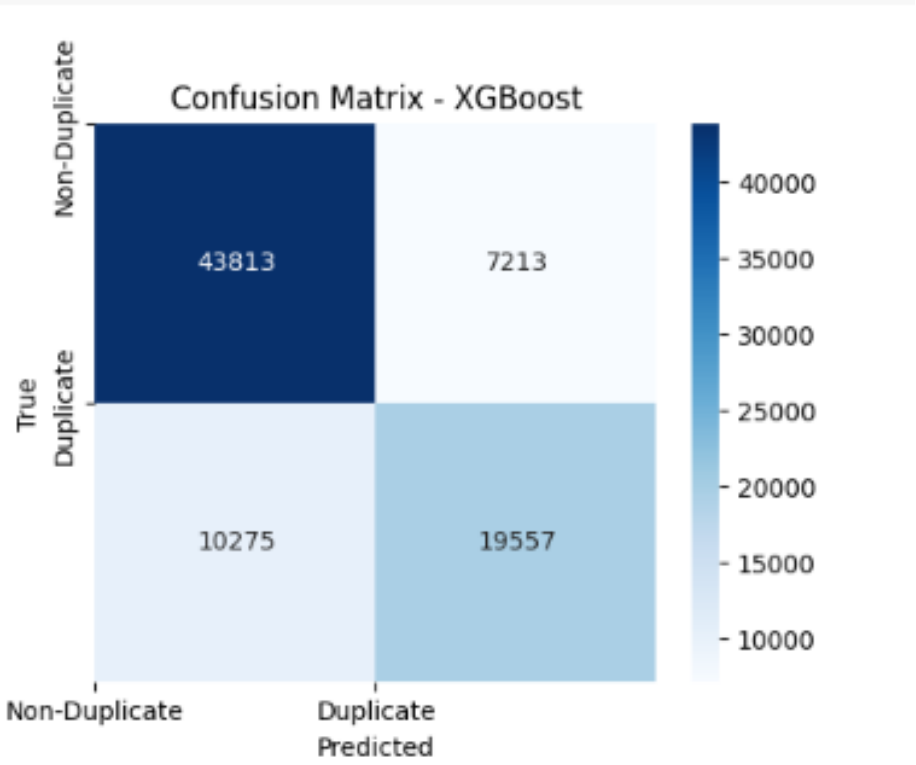

Figure 13: XGBoost Confusion Matrix

4.3 Logistic Regression and Naive Bayes

After checking with our two ML models,I have implemented two more model i.e Naive Bayes and Logistic regression but got wrost result as compared to RandomForest and XGBoost. The Accuracy and F1 score of Logistic Regression are 70.5 and,45.4 whereas the Accuracy and F1 score of Naive Bayes are 68.45 and 52.6 respectively.

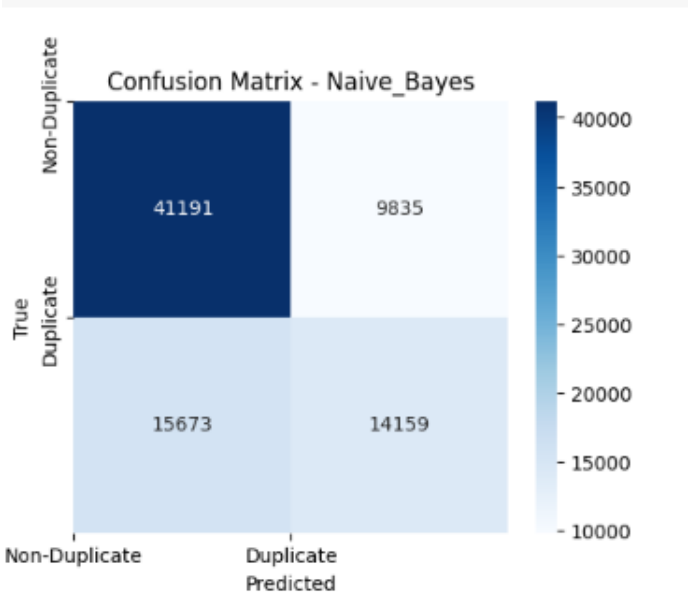

Figure 14: Naive Bayes Confusion Matrix

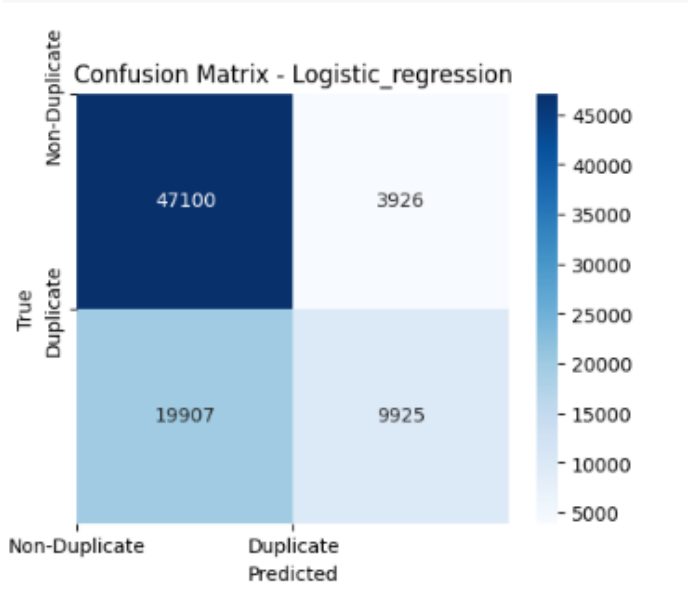

Figure 15: Logistic regression Confusion Matrix

5 Model Comparision and Conclusion

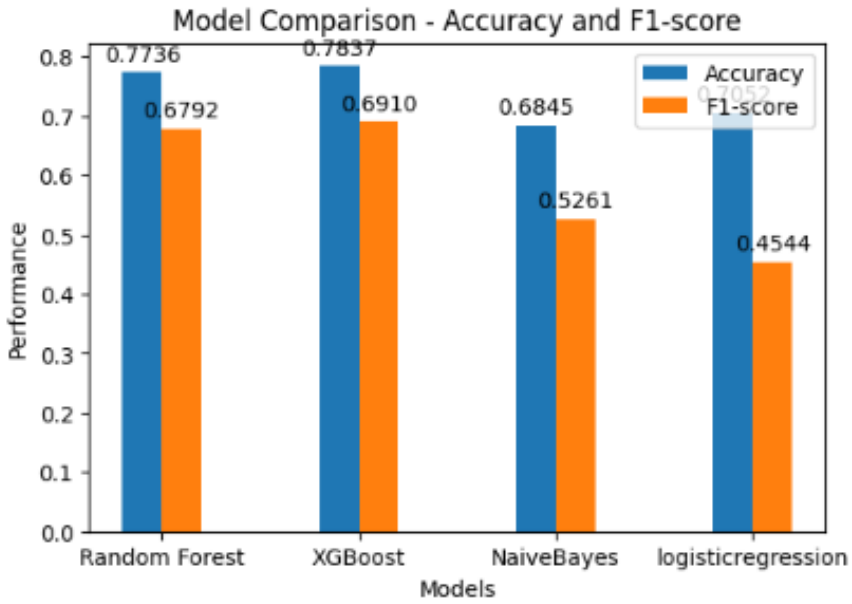

Figure 16: Models Comparision

From the above Model results it is clear that our two best models are XGboost with an accuracy of 78.37 and RandomForest with an accuracy of 77.35. According to accuracy, the XGBoost model gives a slightly better result than RandomForest but when we see the confusion Matrix of both the models, for the column of False Positive where models have predicted a Duplicated Question as Non Duplicate is more in XGBoost as compared to RandomForest.