MAJOR PROJECT

PREDICT DUPLICATE QUESTION ON QUORA

BY : - Shwetank Chaudhary
(07811503120)

Mentor : Dr. Surinder Kaur

Introduction

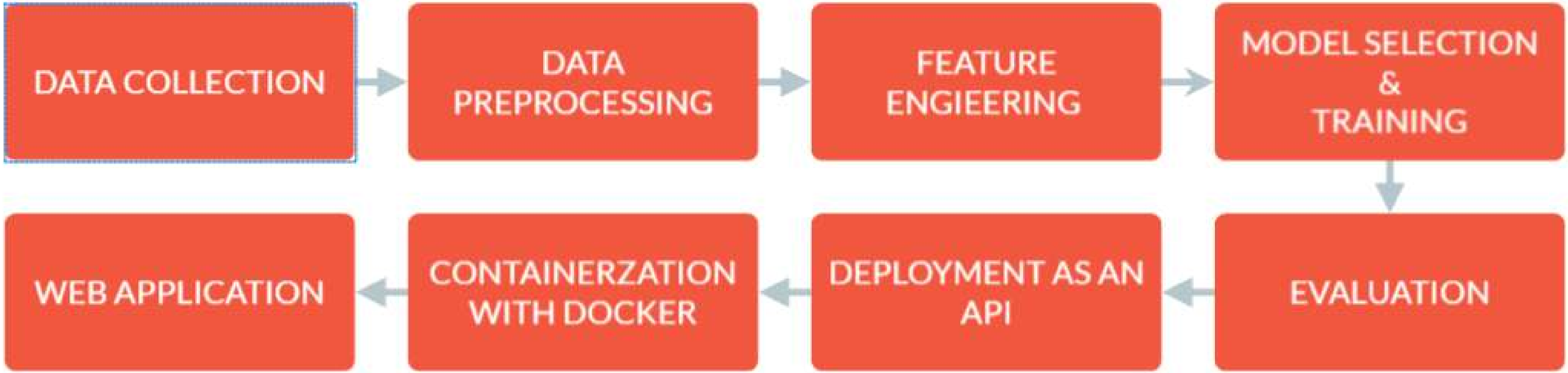
PROBLEM STATEMENT

In the age of information abundance, users often encounter multiple versions of the **same question** across online platforms like Quora. However, distinguishing between genuine inquiries and duplicate questions can be challenging, leading to redundancy and clutter within the platform. This redundancy not only diminishes user experience but also hampers the effectiveness of search and recommendation systems.

OBJECTIVE

Goal is to build a classifier that predicts
whether or not a question is a repeated redundant of already question,
along with the suggestive method to improvise the question.

HOW DOES IT WORK?



DATA COLLECTION

DATA PREPROCESSING

FEATURE ENGIEERING

MODEL SELECTION & TRAINING

EVALUATION

DEPLOYMENT AS AN API

CONTAINERZATION WITH DOCKER

WEB APPLICATION

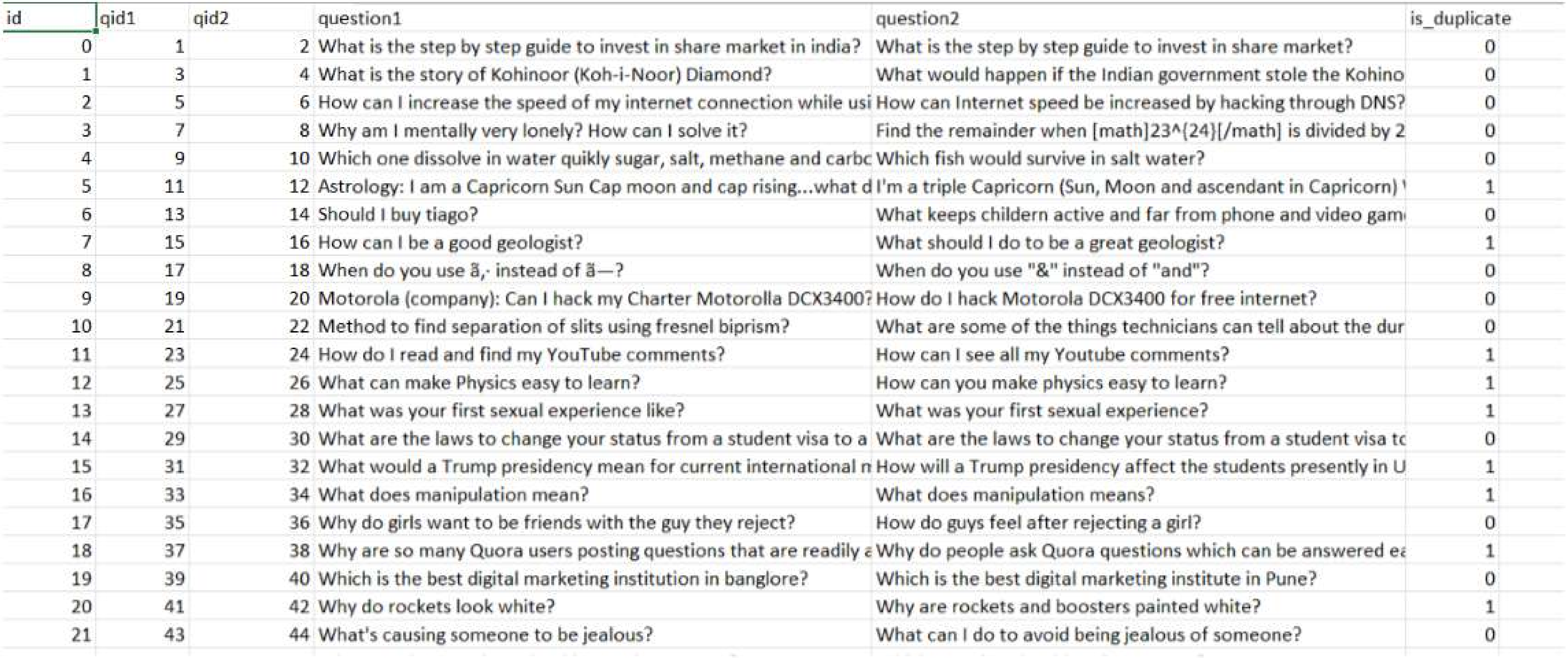
PREREQUISITES FOR PROJECT DEVELOPMENT

- **Python** – You should be familiar with Python programming with its syntax and indentations.
- **Pandas** – Data analysis is essential before building a model. Pandas is a Python library that helps to analyze a high volume of data with straightforward functions and methods.
- **Matplotlib** – Graphs help us to understand the data in a better way, so if you hold knowledge of Python visualizing libraries, then it will help to get the solution quickly.
- **Sklearn** – You should be familiar with machine learning and feature engineering because we will use them to extract different features and train machine learning.

DATASET DESCRIPTION

The dataset we will use is a very popular dataset that Quora hosted in one of the Kaggle's.

The dataset contains only 5 columns, of which two columns contain 2 different questions, 2 column contains the respective question id, and the last column indicates the target variable whose value is in binary format (1 means duplicate, and 0 means non-duplicate)



id	qid1	qid2	question1	question2	is_duplicate
0	1	2	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
1	3	4	What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohino	0
2	5	6	How can I increase the speed of my internet connection while usi	How can Internet speed be increased by hacking through DNS?	0
3	7	8	Why am I mentally very lonely? How can I solve it?	Find the remainder when [math]23^{24}[/math] is divided by 2	0
4	9	10	Which one dissolve in water quikly sugar, salt, methane and carbc	Which fish would survive in salt water?	0
5	11	12	Astrology: I am a Capricorn Sun Cap moon and cap rising...what d	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) '	1
6	13	14	Should I buy tiago?	What keeps childern active and far from phone and video gam	0
7	15	16	How can I be a good geologist?	What should I do to be a great geologist?	1
8	17	18	When do you use ã,· instead of ã—?	When do you use "&" instead of "and"?	0
9	19	20	Motorola (company): Can I hack my Charter Motorolla DCX3400?	How do I hack Motorola DCX3400 for free internet?	0
10	21	22	Method to find separation of slits using fresnel biprism?	What are some of the things technicians can tell about the dur	0
11	23	24	How do I read and find my YouTube comments?	How can I see all my Youtube comments?	1
12	25	26	What can make Physics easy to learn?	How can you make physics easy to learn?	1
13	27	28	What was your first sexual experience like?	What was your first sexual experience?	1
14	29	30	What are the laws to change your status from a student visa to a	What are the laws to change your status from a student visa tc	0
15	31	32	What would a Trump presidency mean for current international n	How will a Trump presidency affect the students presently in U	1
16	33	34	What does manipulation mean?	What does manipulation means?	1
17	35	36	Why do girls want to be friends with the guy they reject?	How do guys feel after rejecting a girl?	0
18	37	38	Why are so many Quora users posting questions that are readily a	Why do people ask Quora questions which can be answered ea	1
19	39	40	Which is the best digital marketing institution in banglore?	Which is the best digital marketing institute in Pune?	0
20	41	42	Why do rockets look white?	Why are rockets and boosters painted white?	1
21	43	44	What's causing someone to be jealous?	What can I do to avoid being jealous of someone?	0

05

PROJECT DEVELOPMENT OVERFLOW

It is good practice to clear the mindset and general project flow steps. so below are the simple steps we will follow to complete the project.

- Basic Data Analysis
- Feature Engineering
- Model Development
- Optimize the model to increase performance
- Web application creation
- Deployment over cloud

FEATURE ENGINEERING

Feature engineering is a classic way of adding new features to the data that dominates to predict output variables and improve the model's accuracy. A crucial feature creates a direct impact on the model. Feature engineering consists of transformation, scaling, feature extraction, feature encoding, etc.

We will add 7 more features to our existing dataset. The bag of words model for questions 1 and 2 questions 2 will produce different features that will be passed to the Machine learning model after analysis.

ADVANCE FEATURES :

1. **Token Features**

cwc_min: This is the ratio of the number of common words to the length of the smaller question
cwc_max: This is the ratio of the number of common words to the length of the larger question
csc_min: This is the ratio of the number of common stop words to the smaller stop word count among the two questions
csc_max: This is the ratio of the number of common stop words to the larger stop word count among the two questions
ctc_min: This is the ratio of the number of common tokens to the smaller token count among the two questions
ctc_max: This is the ratio of the number of common tokens to the larger token count among the two questions
last_word_eq: 1 if the last word in the two questions is same, 0 otherwise
first_word_eq: 1 if the first word in the two questions is same, 0 otherwise

2. Length Based Features

mean_len: Mean of the length of the two questions (number of words)

abs_len_diff : Absolute difference between the length of the two questions (number of words)

longest_substr_ratio: Ratio of the length of the longest substring among the two questions to the length of the smaller question

3. Fuzzy Features

fuzz_ratio: fuzz_ratio score from fuzzywuzzy

fuzz_partial_ratio: fuzz_partial_ratio from fuzzywuzzy

token_sort_ratio: token_sort_ratio from fuzzywuzzy

ALGORITHM USED :

- Random Forest algorithm
- XGBoost Algorithm

MACHINE LEARNING MODELING PART

The data is now ready, and you must repeat the steps above to train the Random Forest and XGboost models for NLP Project. The random forest accuracy is approximately **78.7**, and XGboost gives **79.2** percent. So by doing this much optimization, we could boost the performance by 2 to 2.5 percent.

ACCURACY SCORE :

Model	Initial accuracy	7 features	15 Features
Random Forest	74 %	77%	78.4
XGBoost	73%	76%	79.4

CONFUSION MATRIX:

Actual\Predicted	0	1
0	✓	✗
1	✗	✓

THANK YOU

So much!