Is That A Duplicate Quora Question?

3rd June'18

Team:
Srishti Sawla
Amrinder Singh Bedi
Ashish Jha
Shravani Ghatnekar
Rahul Deora

Agenda

1. Problem Statement
2. EDA
3. Cleaning of Data
4. Feature Engineering
5. Model Building
6. Model Evaluation and Selection

What is Quora?

- Quora is a platform to gain and share knowledge
- Best place to ask questions and connect with people to get quality answers
- Over 100 million people visit Quora every month, and many people ask similarly worded questions.

Problem Background

- 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. Hence the challenge is to classify whether question pairs are duplicates or not. Doing so will make it easier to find high quality answers to questions resulting in an improved experience for Quora writers, seekers, and readers.

Problem Statement

- Given a pair of questions, predict if the questions have same meaning or intent

Data Overview

	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

Where,

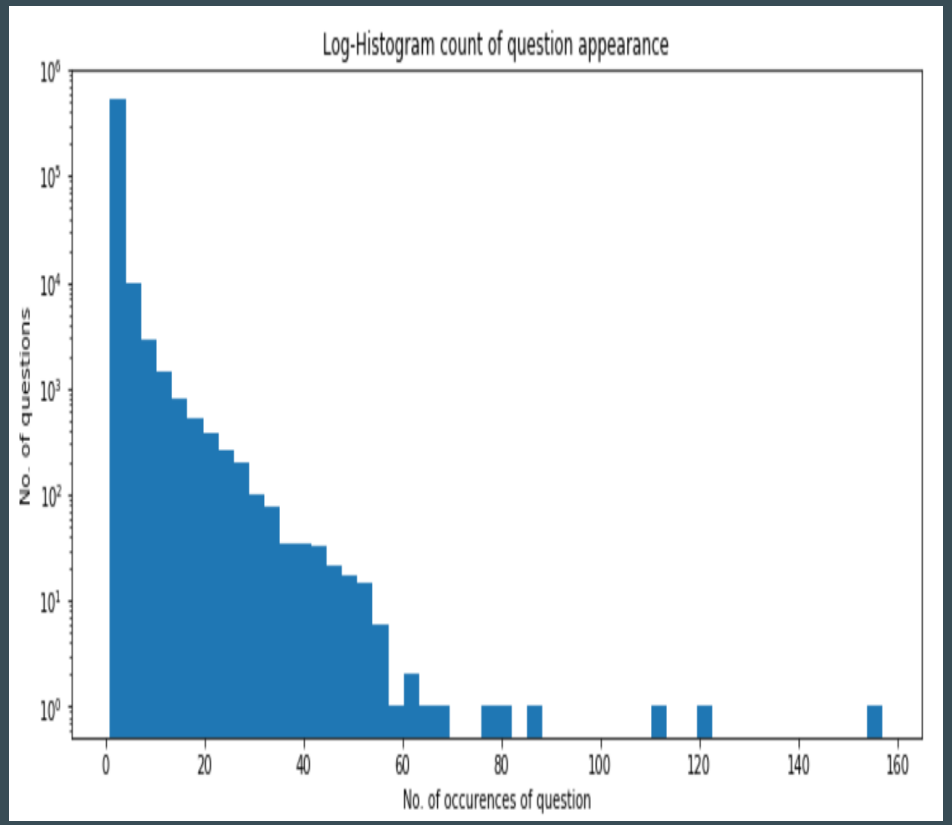
1. id: ID of the item
2. qid1: ID of question1
3. qid2: ID of question2
4. question1: Full text of question 1
5. question2: Full text of question 2
6. is_duplicate: 1 if the questions having same meaning
   0 if the questions not having same meaning

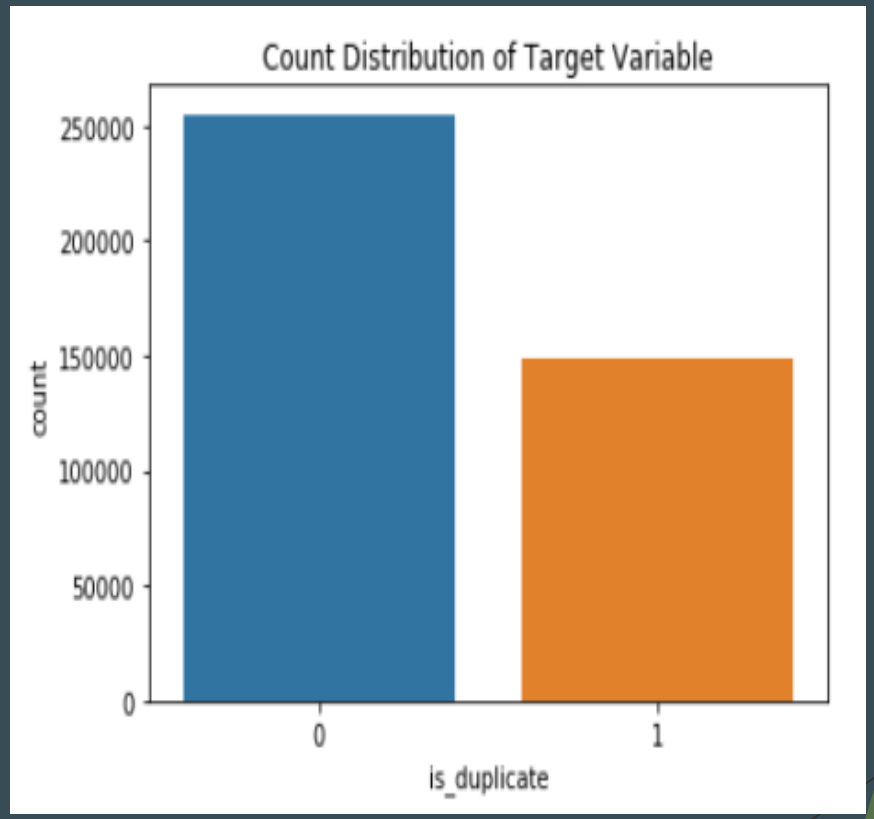
BASIC EDA

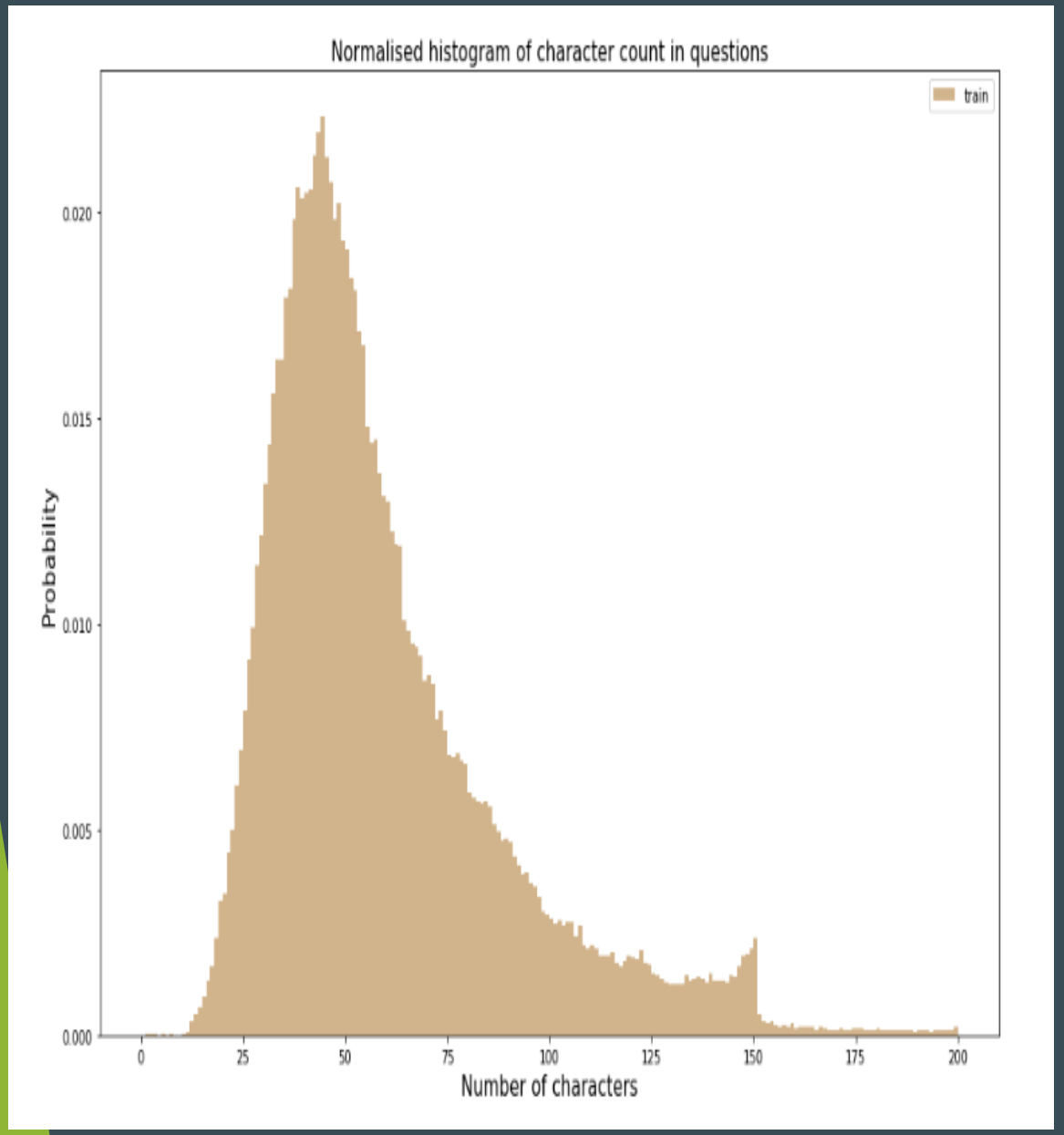
Data Overview

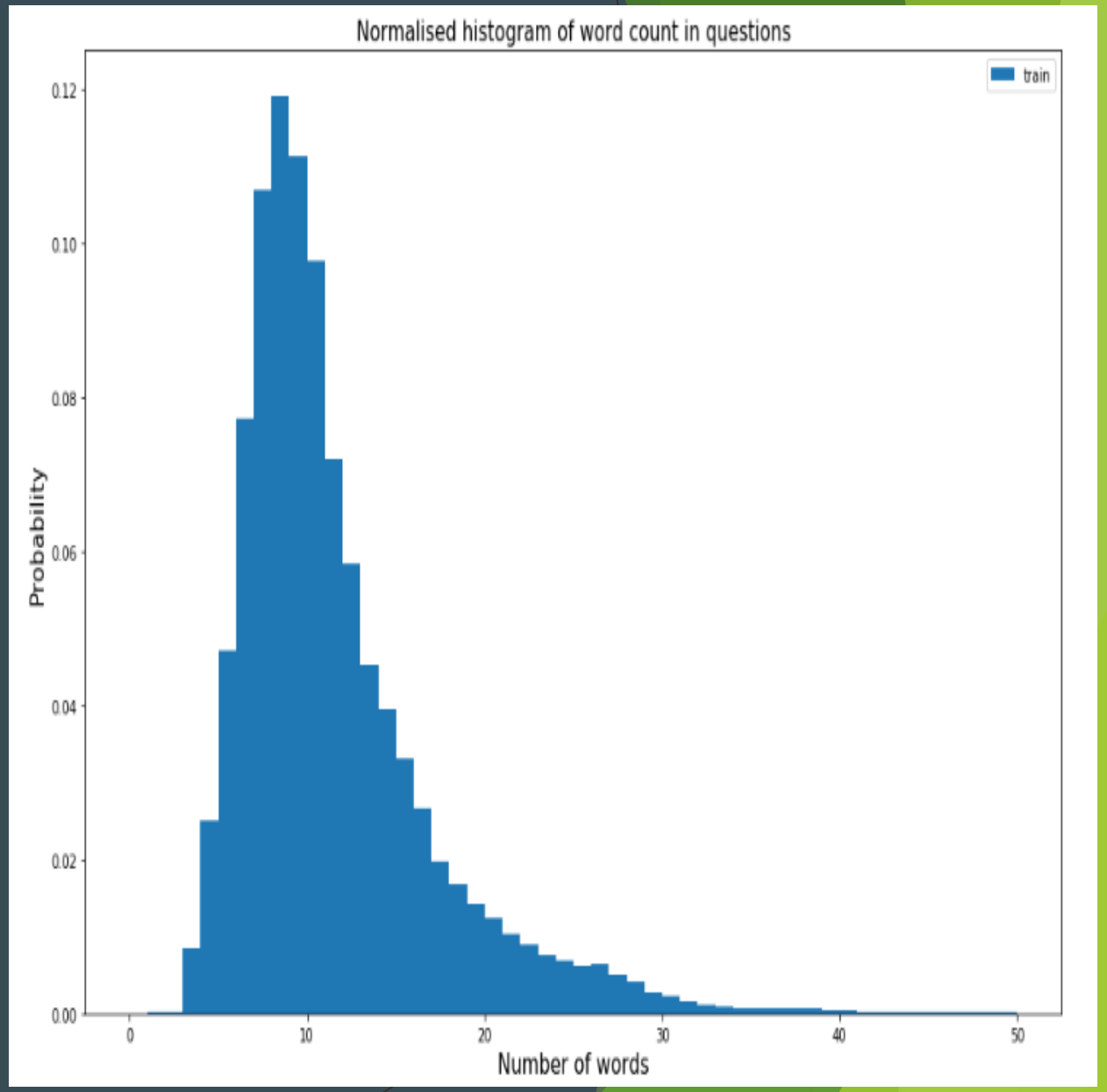
- No. of question pairs: 404290
- Unique Question in Dataset:537933
- Questions appearing multiple times:111780
- Duplicate Proportion:36.92
- Questions with question marks: 99.88%
- Questions with [math] tags: 0.12%
- Questions with full stops: 6.31%
- Questions with capitalised first letters: 99.81%
- Questions with capital letters: 99.95%
- Questions with numbers: 11.83%

Histogram for Occurrence of Questions






Normalised histogram of character count in questions
train
Probability
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
125
150
175
200
Number of characters


Normalised histogram of word count in questions
train
Probability
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
Number of words

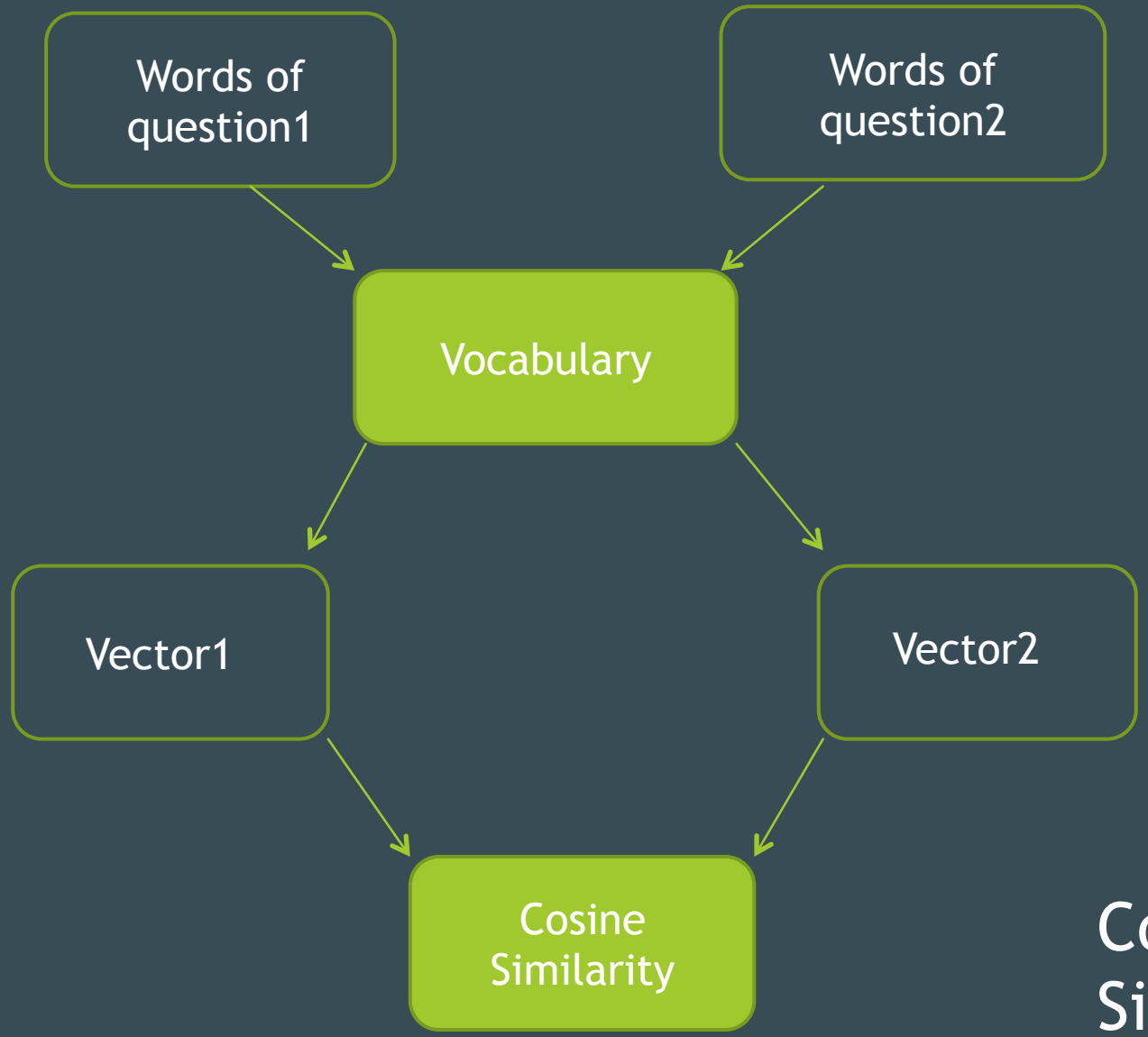
- Two word clouds generated from question 1 and question 2 show that there are a few duplicate most common words in them.

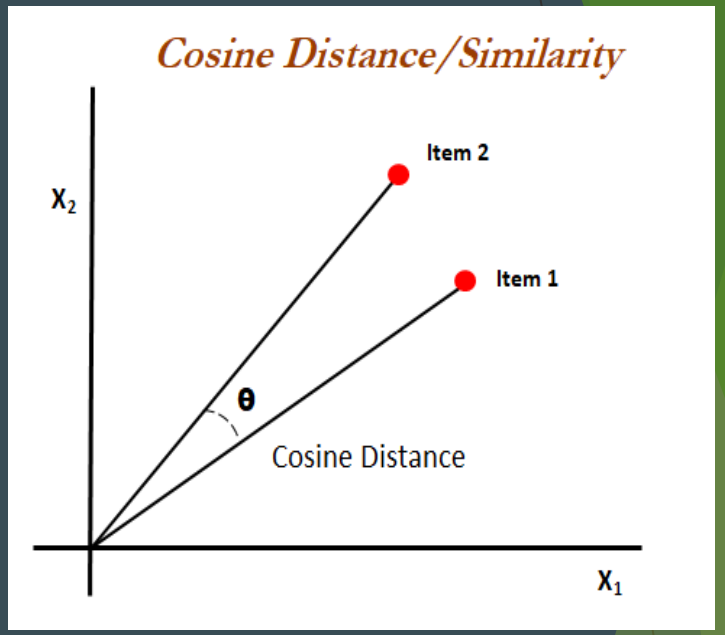
Data Cleaning

- Removed Punctuation marks
- Removed selective StopWord(punctuation marks)
- Lemmatisation

Feature Engineering

- Cosine Similarity
- Length of questions
- Length Difference
- First and Last Word Comparison
- Polarity

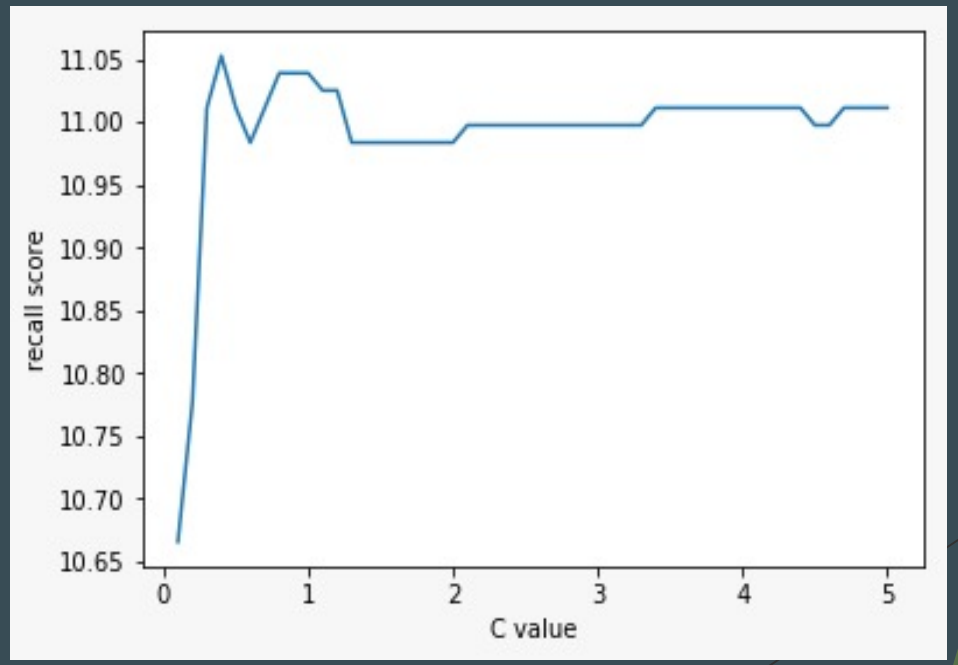




Cosine Similarity

Data Modelling

- Logistic Regression as baseline model
- Naïve Bayes
- Random Forest



Model Evaluation

- Evaluation Metric used: Logloss
- Final Model Selected: Random Forest

  Logloss for RF: 3.12

THANK YOU