Extreme Gradient Boosting Based Question Similarity Predictor

Vrushabh Waman[1], Shreyank Patil[2], Ujwal Deshpande[3], Rupali Sathe[4]

[1,2,3]Department of Information Technology Pillai HOC College of Engineering and Technology Rasayani, India

[4]Professor, Department of Information Technology Pillai HOC College of Engineering and Technology Rasayani India

Abstract - *Question and Answer Websites are a very popular medium for people to share knowledge and experience. We aim to ease the searching of questions on these websites. In this paper we have compared multiple machine learning algorithms to predict similar questions using real world data set.*

Key Words**: Machine Learning, Linear Regression, Logistic Regression, Extreme Gradient Boosting, Log Loss, Quora, Question and Answer**

1.INTRODUCTION

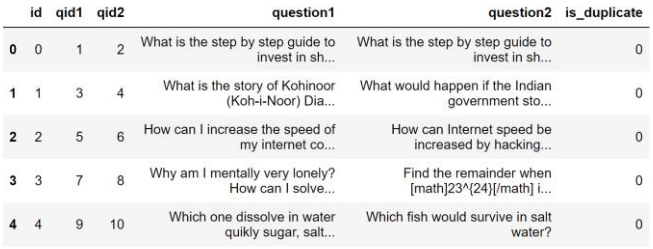
Nowadays question answer websites like Quora, stack overflow, Yahoo Answers, Google scholars have become a very popular medium to search questions for laymen as well as domain experts. Many people ask similar worded questions multiple times. Multiple questions with the same intent can cause users to spend more time seeking the best answer for their question. Similarly, writers should not feel the need to answer the same question multiple times. We have used a machine learning based approach for searching questions on the website. The main problem in these tasks is to identify that the questions asked are duplicates of questions that have already been asked. This could be useful to instantly provide answers to questions that have already been answered. We are tasked with predicting whether a pair of questions are duplicates or not.

2. LITERATURE SURVEY

We've Compared Multiple Models, but the Random Models is the worst-case scenario. The test Log-Loss using Random Model is 0.8865. whereas when the data is linearly separable i.e., while using Logistic Regression we encountered test Log-Loss of 0.4237. In linear support vector machines where the data is separated using 2 plains, we got a test Log-Loss of 0.5078.

3. DATA ACQUISITION

We have used the Quora data set for implementing machine learning algorithms. This data set has more than 4 lakh data points. Each data point consists of two questions and a binary value which states whether they are similar or not.



	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

Chart -1: Example Points from Dataset

4. EXPLORATORY DATA ANALYSIS

The data set is imbalanced, 63% of the data points are not similar i.e., more than 2.5 lakh question pairs are not similar.

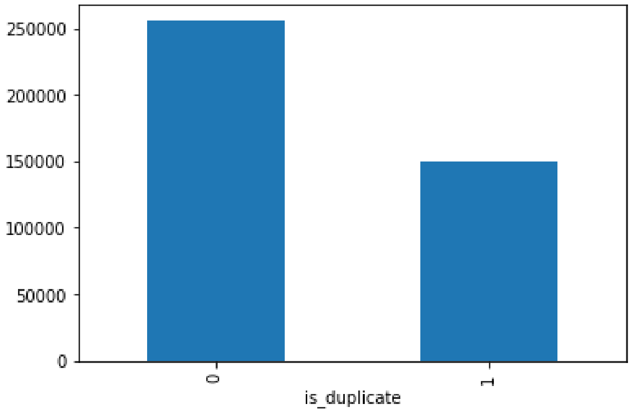


Chart -2: Imbalance Dataset

5. PRE-PROCESSING

5.1 Data Cleaning

In the process of cleaning of data, we have included steps like

- Removing of stop words
- Removing HTML tags
- Removing punctuations
- Expanding Contractions
- Performing Stemming
- Filling N.A. Values

5.2 Feature Extraction

We have constructed new features based on the length and frequency of the questions and number of words contained in the question. We also have appended features which are a derivative of above features. we have also created NLT and fuzzy features [1].

5.3 Featurization

Featurization is a way to change text data into numerical vectors. Questions are represented in TF idf weighted word2vec format. We have featurized each question into 300 vectors. Therefore, in total we have 600 features.

6. MODEL TRAINING, EVALUATION AND INTERPRETATION

In total we have 794 features and one binary output feature for training the model. We have done a random train test split of 70:30 where the training data has equal class distribution. The key performance indicator used is log-loss. lower is the loss, better is the performance of the model.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss

Chart -3: Formula to Calculate Log Loss

We have tuned the hyper parameters (alpha) in the range of 10^{-5} to 10^{1}.

This XGboost[5] model is trained with an objective as "binary logistic". whereas the maximum depth of the tree is set to 4 and maximum rounds to stop training is 400. the train and test Log-Loss after 400 rounds is 0.3511 and 0.3517 respectively.

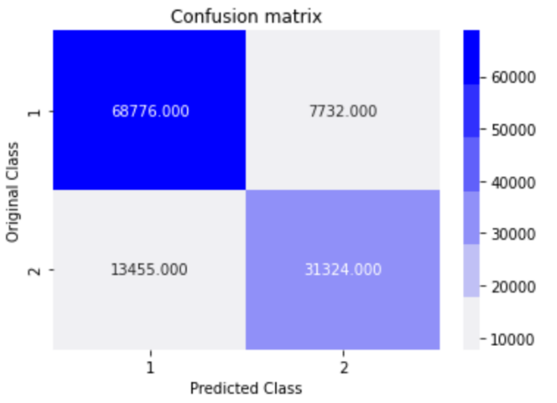


Chart -4: Confusion matrix for Extreme gradient boosting

7. VISUAL UNDERSTANDING OF SYSTEM

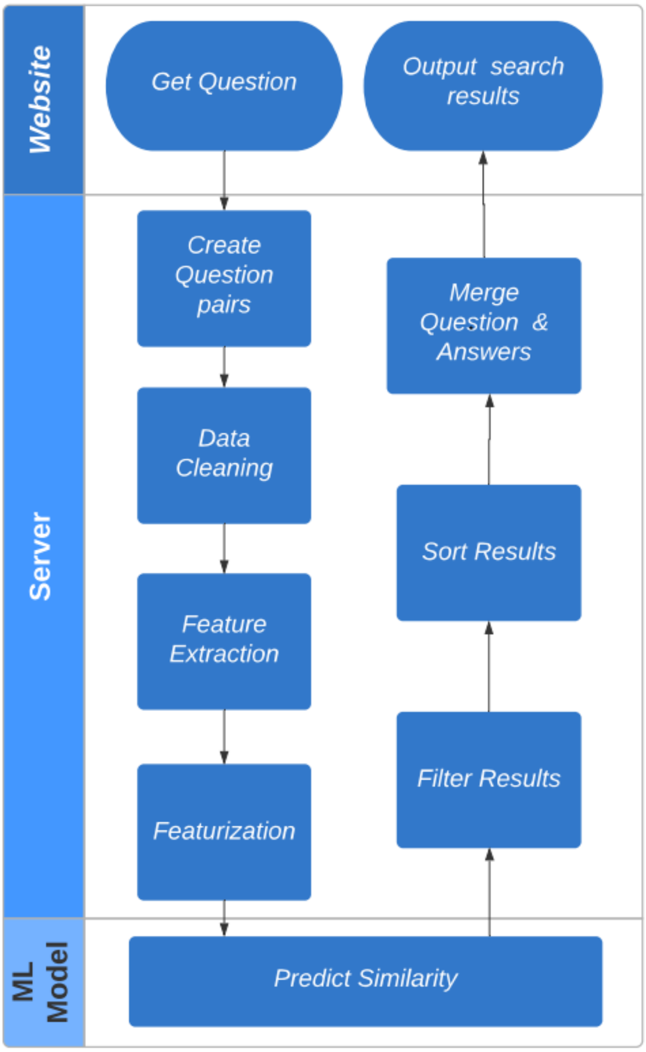


Chart -5: Swimlane Diagram of Implemented System

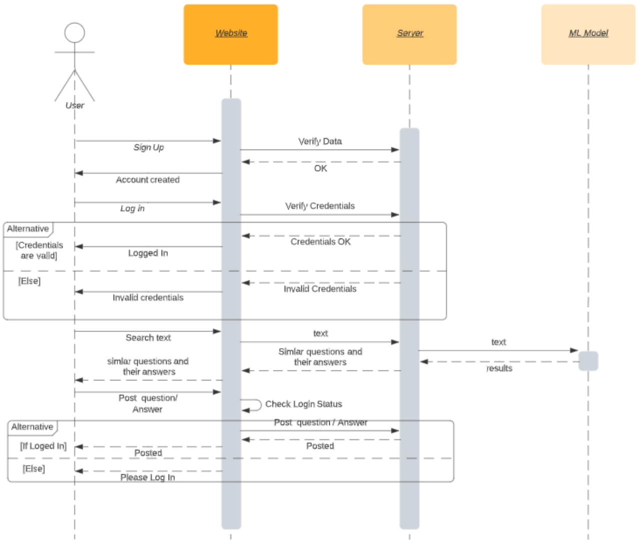
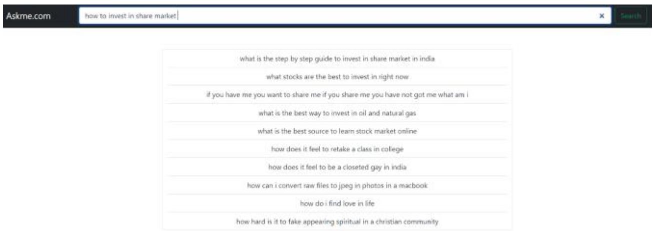


Chart -6: Sequence Diagram of Implemented System

8. RESULT AND ANALYSIS



how to invest in share market

what is the step by step guide to invest in share market in india
what stocks are the best to invest in right now
if you have me you want to share me if you share me you have not got me what am i
what is the best way to invest in oil and natural gas
what is the best source to learn stock market online
how does it feel to retake a class in college
how does it feel to be a closeted gay in india
how can i convert raw files to jpeg in photos in a macbook
how do i find love in life
how hard is it to fake appearing spiritual in a christian community

Chart -6: Screenshot of website with search results.

The implemented system performs superior to the model implemented with logistic regression or linear SVM as the log loss for the implemented system using Extreme Gradient Boosting is less.

9. CONCLUSIONS

Thus, we have implemented the above system on the website using Flask and we have achieved satisfactory results on the search question.

ACKNOWLEDGEMENT

We remain very committed to Prof. RUPALI SATHE for providing us with son and technical support and guidance. We would also like to thank our guide for providing us with his professional opinion and good suggestions for all stages of the project. We would like to take this opportunity to thank Professor Dr.J.E.NALAWADE Head of Information Technology for his motivation and invaluable support. This acknowledgment is incomplete without thanking academic and non-academic staff for their department for their kind support. We would also like to thank Dr.J.MATHEW, Principal of Pillai HOC College of Engineering and Technology, Rasayani for providing the necessary infrastructure and resources for the project.

REFERENCES

[1] M. Ramze Rezaee, B. Goedhart, B.P.F. Lelieveldt, J.H.C. Reiber , Fuzzy Feature Selection (1999).

[2] Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll, An Introduction To Logistic Regression Analysis and Reporting (2002).

[3] Lubor Ladicky, Philip Hilaire Torr, Linear Support Vector Machines (2011).

[4] Sonam Sonam, Ayushi Verma, Sangeeta Lal, Neetu Sardana, TagStack:Automated System for Predicting Tags in StackOverflow (2019).

[5] Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System (2016)