ENSF 612

ANALYSIS OF STACK OVERFLOW DATA

PROJECT REPORT

15th DECEMBER 2023

AMEY BRAHME

TEJPREET BAL

Table of Contents

List of Tables

List of Figures

1. Introduction and Motivation

In today's rapidly evolving technological landscape, the generation and accumulation of data have reached unprecedented levels. The sheer volume and complexity of this data presents both challenges and opportunities for organizations across various industries. Harnessing the power of big data has become imperative as it enables enterprises to gain valuable insights, make informed decisions, and enhance the overall efficiency of their operations. The ability to analyze vast datasets not only allows for a deeper understanding of customer behavior but also facilitates the identification of patterns and trends that can be leveraged for innovation and improvement. Big data projects, therefore, play a pivotal role in unlocking the latent potential within these massive datasets, paving the way for data-driven decision-making and sustainable growth.

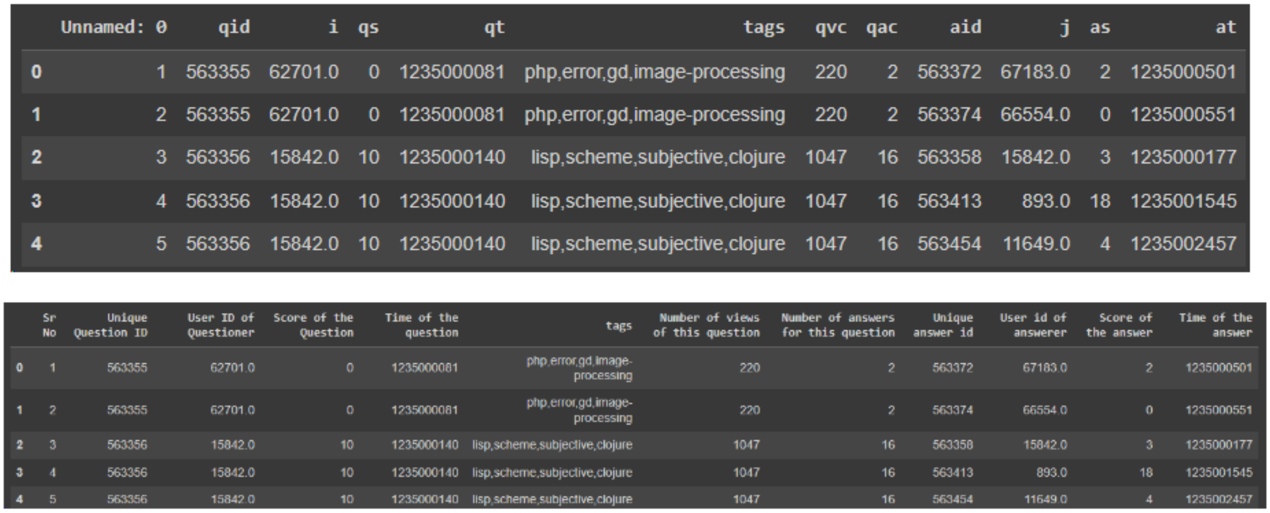
Motivated by the profound impact of big data analysis, one compelling area of focus is the examination of Stack Overflow data to predict the time it takes to receive a response to users' questions. Stack Overflow, a widely used platform for programming-related queries, serves as an invaluable repository of knowledge and expertise. By delving into this dataset, one can uncover patterns in user interactions, identify factors influencing response times, and ultimately enhance the user experience. Understanding the dynamics of question-response cycles on Stack Overflow not only contributes to the optimization of the platform but also sheds light on broader trends in collaborative problem-solving within the programming community. This targeted analysis exemplifies how big data projects can be tailored to address specific domains, offering actionable insights that hold the potential to streamline processes and elevate the overall quality of user engagement.

2. Data Collection

The data collection process for this analysis commenced by exploring various potential sources of Stack Overflow data. Initially, alternative options such as the latest data dumps from GitHub or torrent repositories were considered to ensure a comprehensive dataset. Techniques like web scraping were also contemplated as a means of obtaining the required information. However, due to time constraints and the availability of a substantial dataset from the University of California, Irvine (UCI) website, a decision was made to prioritize efficiency and reliability. The primary source of data for this project was the Stack Overflow dataset hosted at https://ics.uci.edu/~duboisc/stackoverflow/. This dataset comprised 263,541 rows and 12 columns, focusing on answers posted between February 18, 2009, and June 7, 2009. The information was contributed by 15,098 unique users, with a notable majority having answered fewer than 50 questions. This approach ensured a rich but manageable dataset for the subsequent analysis, striking a balance between the scope of the project and the practical constraints of data collection.

3. Data Inspection and Validation

Given that the dataset was released by the owners of Stack Overflow, a platform known for its commitment to data integrity and reliability, the need for extensive validation was alleviated. The inherent credibility of the source played a significant role in ensuring the reliability and accuracy of the data. While traditional validation steps, such as cross-referencing with external sources or conducting extensive data cleansing, might be essential in some cases, it was not necessary in this case. The Stack Overflow dataset was available in both R data format and CSV. A decision was made to utilize the CSV format. This choice was driven by practical considerations, such as the team's familiarity with CSV and its compatibility with the course scope, facilitating a smoother integration into the analysis pipeline. The dataset initially presented attributes with unintuitive titles, prompting a crucial step in the data inspection process—renaming the attributes with more descriptive headers to enhance clarity and comprehension.

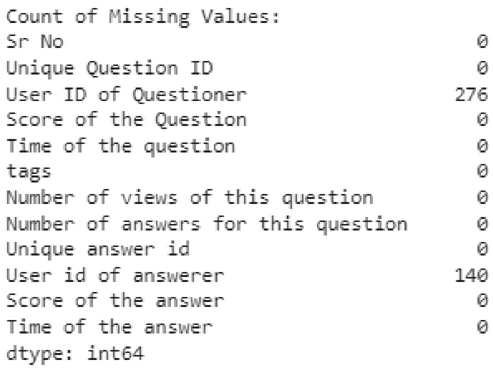


	Unnamed: 0	qid	i	qs	qt	tags	qvc	qac	aid	j	as	at
0	1	563355	62701.0	0	1235000081	php,error,gd,image-processing	220	2	563372	67183.0	2	1235000501
1	2	563355	62701.0	0	1235000081	php,error,gd,image-processing	220	2	563374	66554.0	0	1235000551
2	3	563356	15842.0	10	1235000140	lisp,scheme,subjective,clojure	1047	16	563358	15842.0	3	1235000177
3	4	563356	15842.0	10	1235000140	lisp,scheme,subjective,clojure	1047	16	563413	893.0	18	1235001545
4	5	563356	15842.0	10	1235000140	lisp,scheme,subjective,clojure	1047	16	563454	11649.0	4	1235002457

	Sr No	Unique Question ID	User ID of Questioner	Score of the Question	Time of the question	tags	Number of views of this question	Number of answers for this question	Unique answer id	User id of answerer	Score of the answer	Time of the answer
0	1	563355	62701.0	0	1235000081	php,error,gd,image-processing	220	2	563372	67183.0	2	1235000501
1	2	563355	62701.0	0	1235000081	php,error,gd,image-processing	220	2	563374	66554.0	0	1235000551
2	3	563356	15842.0	10	1235000140	lisp,scheme,subjective,clojure	1047	16	563358	15842.0	3	1235000177
3	4	563356	15842.0	10	1235000140	lisp,scheme,subjective,clojure	1047	16	563413	893.0	18	1235001545
4	5	563356	15842.0	10	1235000140	lisp,scheme,subjective,clojure	1047	16	563454	11649.0	4	1235002457

Figure 1. Attribute Titles before and after re-naming.

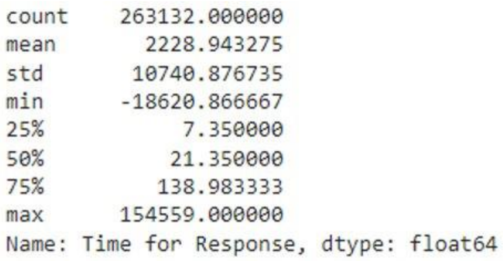
Subsequently, null values were identified within the dataset. Given that these null values were confined to the userID column, which was deemed irrelevant to the specific analysis goals, a decision was made to drop the column to streamline the dataset. To facilitate further analysis, the data was then loaded into a Pandas DataFrame, a widely used data manipulation library in Python. Basic exploratory steps, such as examining the shape of the DataFrame, were performed to gain an initial understanding of the dataset's dimensions.



```
Count of Missing Values:
Sr No                                  0
Unique Question ID                     0
User ID of Questioner                276
Score of the Question                  0
Time of the question                   0
tags                                   0
Number of views of this question       0
Number of answers for this question    0
Unique answer id                       0
User id of answerer                  140
Score of the answer                    0
Time of the answer                     0
dtype: int64
```

Figure 2. Pre-Processing of Data – Identifying Null Values

Basic statistical analysis was employed to gain an initial understanding of the dataset's characteristics. Measures such as the maximum, minimum, and average values were calculated to provide a broad sense of the data's distribution. However, recognizing the limitations of these summary statistics in capturing the nuances of the dataset, additional measures such as the first quartile (25th percentile), second quartile (50th percentile or median), and third quartile (75th percentile) were considered. The identification of quartiles was particularly valuable in highlighting the presence of outliers and the dataset's wide variability. As illustrated in Figure 3, the examination of response times reveals that 75% of the dataset's instances exhibited response durations below 139 minutes (about 2 and a half hours). However, the max and mean values present a contracting portrayal. This recognition underlines the need for data processing techniques to handle outliers and ensure a more robust and accurate analysis as described in subsequent sections of this report.



```
count    263132.000000
mean       2228.943275
std       10740.876735
min      -18620.866667
25%           7.350000
50%          21.350000
75%         138.983333
max      154559.000000
Name: Time for Response, dtype: float64
```

Figure 3. Rudimentary Statistical Analysis

4. Data Filtering

In the data filtering phase, the analysis extended beyond mere plotting of response times, incorporating comprehensive visualizations, such as box plots, to identify outliers and establish a cutoff criterion. Recognizing that an outlier post might exhibit multiple indications of deviance, additional attributes such as scores and the number of answers were integrated into the visual analysis. Negative values, inconsistent with the nature of response times, were expunged from the dataset. Negative values, which are incongruent with the inherent characteristics of response times, were systematically eliminated from the dataset. Additionally, responses exhibiting a score of 150 or greater were identified as outliers and subsequently excluded. The dataset underwent further refinement by excluding negative question scores and scores exceeding 150. The initial responses were preserved in the final dataset and all subsequent responses were removed.

In the effort to further streamline the dataset for a more focused analysis, several attributes were excluded to refine the dataset. Building on the earlier decision to omit userIDs, the dataset underwent further pruning by removing the serial number, Unique Question ID, Unique answer ID, Time of the Question, and Time of the answer. By dropping these attributes, the dataset is now aligned with the objectives of uncovering and predicting response time patterns on Stack Overflow, facilitating a more streamlined and purposeful exploration.

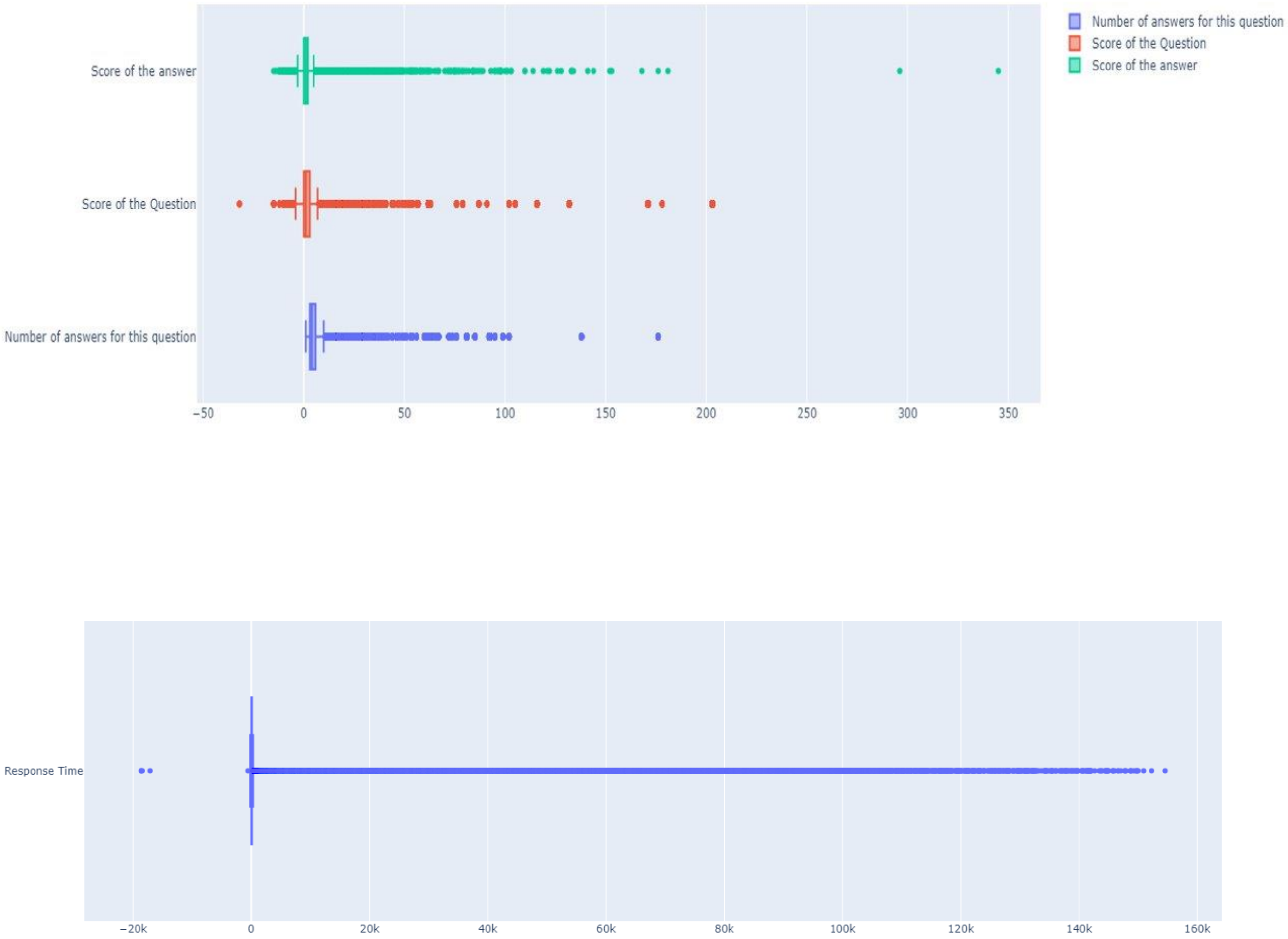


Figure 4. Box Plots for Raw Data

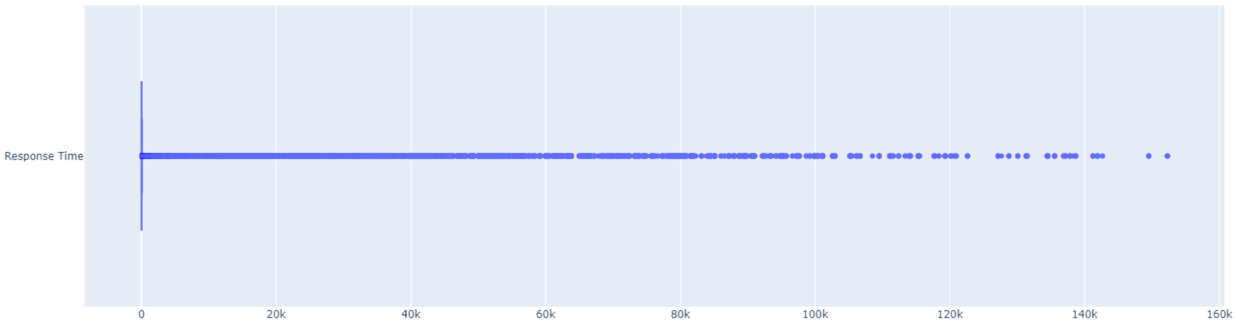
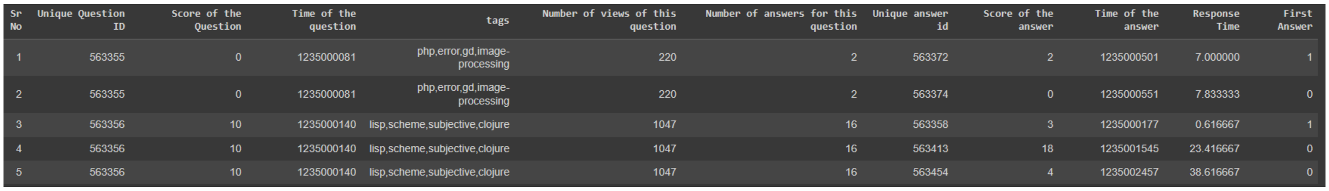


Figure 5. Box Plot of Response Times for trimmed Dataset

5. Data Transformations

In the data transformation phase, the timestamp information, originally presented in epoch time, underwent a transformation to derive more meaningful insights. A new column was introduced, representing the time taken for a response in minutes. This was achieved by calculating the difference between the epoch time of the question creation and the epoch time of the corresponding answer.

Given that questions on Stack Overflow can receive multiple answers, the analysis focused on isolating the response time for the first answer. This prioritization of the initial response time is significant as it provides insights into the quickest response a user receives. To facilitate this distinction, a new "First Answer" column was incorporated into the dataset. This column serves the purpose of identifying and marking the instances where the corresponding answer is the first response to a given question. This targeted transformation not only refines the dataset for the specific analysis goal but also sets the stage for subsequent modeling and evaluation of response time patterns on Stack Overflow.

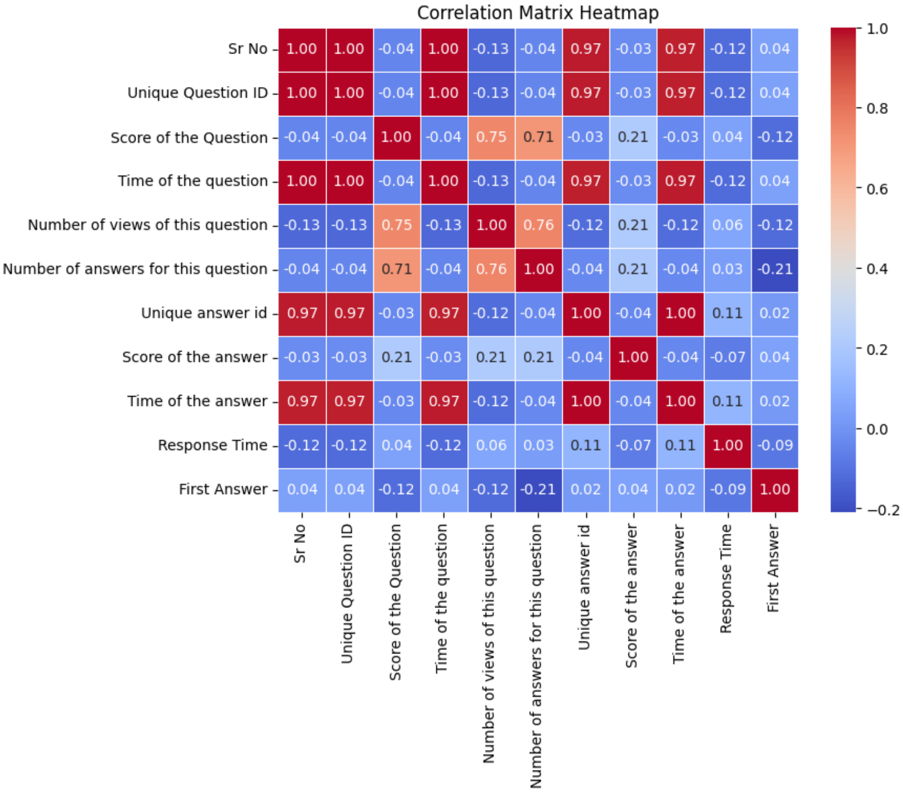


Sr No	Unique Question ID	Score of the Question	Time of the question	tags	Number of views of this question	Number of answers for this question	Unique answer id	Score of the answer	Time of the answer	Response Time	First Answer
1	563355	0	1235000081	php,error,gd,image-processing	220	2	563372	2	1235000501	7.000000	1
2	563355	0	1235000081	php,error,gd,image-processing	220	2	563374	0	1235000551	7.833333	0
3	563356	10	1235000140	lisp,scheme,subjective,clojure	1047	16	563358	3	1235000177	0.616667	1
4	563356	10	1235000140	lisp,scheme,subjective,clojure	1047	16	563413	18	1235001545	23.416667	0
5	563356	10	1235000140	lisp,scheme,subjective,clojure	1047	16	563454	4	1235002457	38.616667	0

Figure 6. Transformed Dataset Snapshot

6. Exploratory Data Analysis

In the exploration of data attributes, a correlation matrix was meticulously generated to visually represent the interrelationships among various aspects. Among the notable observations derived from this analysis, a salient finding pertains to the correlation between the score of a question and both the number of views and the number of answers. This correlation aligns with intuitive expectations, as questions attracting multiple answers are more likely to accrue a higher score, reflecting a greater likelihood of users finding satisfactory solutions when multiple perspectives are offered. Moreover, the inherent connection between the time a question is posed and the time at which an answer is provided emerges as another discernible pattern. The strong correlation observed in this context underscores the practical notion that timely responses contribute significantly to user satisfaction, with answers ideally being proffered promptly after the question is posed. In contrast, the remaining attributes exhibit a lack of significant correlation with each other, suggesting that their variations do not exhibit a discernible pattern of mutual influence. This analytical insight contributes to a comprehensive understanding of the dynamics inherent in the examined dataset, serving as a valuable foundation for further investigation and strategic decision-making.



	Sr No	Unique Question ID	Score of the Question	Time of the question	Number of views of this question	Number of answers for this question	Unique answer id	Score of the answer	Time of the answer	Response Time	First Answer
Sr No	1.00	1.00	-0.04	1.00	-0.13	-0.04	0.97	-0.03	0.97	-0.12	0.04
Unique Question ID	1.00	1.00	-0.04	1.00	-0.13	-0.04	0.97	-0.03	0.97	-0.12	0.04
Score of the Question	-0.04	-0.04	1.00	-0.04	0.75	0.71	-0.03	0.21	-0.03	0.04	-0.12
Time of the question	1.00	1.00	-0.04	1.00	-0.13	-0.04	0.97	-0.03	0.97	-0.12	0.04
Number of views of this question	-0.13	-0.13	0.75	-0.13	1.00	0.76	-0.12	0.21	-0.12	0.06	-0.12
Number of answers for this question	-0.04	-0.04	0.71	-0.04	0.76	1.00	-0.04	0.21	-0.04	0.03	-0.21
Unique answer id	0.97	0.97	-0.03	0.97	-0.12	-0.04	1.00	-0.04	1.00	0.11	0.02
Score of the answer	-0.03	-0.03	0.21	-0.03	0.21	0.21	-0.04	1.00	-0.04	-0.07	0.04
Time of the answer	0.97	0.97	-0.03	0.97	-0.12	-0.04	1.00	-0.04	1.00	0.11	0.02
Response Time	-0.12	-0.12	0.04	-0.12	0.06	0.03	0.11	-0.07	0.11	1.00	-0.09
First Answer	0.04	0.04	-0.12	0.04	-0.12	-0.21	0.02	0.04	0.02	-0.09	1.00

Figure 7. Correlation Matrix

Although some preparation was already completed as part of Sections 4 and 5, some further fine tuning was required before a pipeline could be used for fitting the data. The process initiated with the creation of a Spark context, and the computational workload was distributed across four workers, harnessing the parallel processing power of Spark. The initial dataset, represented as a Pandas DataFrame, was converted into a PySpark DataFrame. The DataFrame was then filtered to remove entries with negative response times and to exclude responses that were not the initial response.

The "tags" column, which contained multiple tags per entry, was exploded to individual tags to facilitate further analysis. To address the inherent challenges associated with machine learning algorithms and string data, the team employed StringIndexer to assign unique index values to the tags column, thereby augmenting interpretability. Taking an additional step, the team applied One-Hot Encoding to the indexed tags column, effectively circumventing limitations related to raw string-encoded categorical variables.

7. Model Building and Results

In this section, we describe the process of building and training machine learning models using Apache Spark. The analysis was conducted within the Google Colab environment, leveraging the power of Spark for big data processing.

7.1 Feature Engineering

As a crucial step in the model-building process, relevant features were selected and engineered to enhance the predictive capabilities of the machine learning models. The initial feature selection involved a curation of five columns: "Score of the Question," "Number of Views of This Question," "Number of Answers for This Question," "Score of the Answer," and "Tags Encoded." These columns were chosen from the original twelve, reflecting an initial attempt to capture essential aspects of the dataset for analysis. However, as elucidated in the subsequent section, the outcomes derived from this feature set proved to be suboptimal, prompting a reevaluation of the feature selection strategy. Consequently, a second feature selection was undertaken, focusing solely on two columns: "Tags Encoded" and "Response Time." This refined set of features aimed to distill the essential elements contributing to model performance, as revealed in the subsequent analysis.

7.2 Model Selection

PySpark pipelines were employed for Linear Regression and Random Forest Regressor algorithms for both sets of selected features described in Section 7.1. Furthermore, a decision Tree algorithm was employed only on the trimmed features. The utilization of pipelines ensured a systematic and reproducible approach to model building.

To refine the performance of the linear regression model, an exploration of hyperparameters was undertaken. This involved testing L1 regularization, L2 regularization, and a combination of both regularization techniques. The goal was to uncover the optimal configuration that would enhance the model's predictive capabilities for the given regression task.

Similarly, for the random forest regressor, a meticulous optimization process was conducted. The hyperparameters under scrutiny included varying the ‘maxDepth’ parameter within the range of 3 to 5 to mitigate the risk of overfitting. Additionally, the number of trees (numTrees) was explored in a range from 10 to 20, aiming to strike a balance for improved model generalization.

In addition to linear regression and random forest, the analysis extended to incorporate the evaluation of a decision tree model. This approach was pursued with the anticipation that the decision tree model might yield superior results, contributing further to the diverse set of models under consideration. Similar to the Random Forest, parameter optimization was performed for Decision Tree with maxDepth between 3 to 5 and maxBins from 32 to 64. The following sections provide a detailed account of the results obtained and insights derived from these comprehensive model evaluations.

7.3 Model Evaluation

The best parameters obtained through optimization are shown in Table 1. For linear regression it was observed that Lasso regression was best suited with a balanced regularization intensity of 0.5 whereas for decision tree and random forest, depth was limited to 3 to prevent overfitting.

Table 1. Optimized Hyperparameters for Linear and Random Forest Regression.

Model	ElasticNetParam (L1/L2 Regularization)	RegParam (Regularization Strength)	MaxDepth	NumTrees/ MaxBins
Linear Regression	1 (Lasso)	0.5	-	-
Random Forest Regression	-	-	3	10
Decision Tree	-	-	3	32

The measures used to evaluate the performance of predictions of various models were Root Mean Squared Error (RMSE), Mean Squared Error (MSE) and Mean Absolute Error (MAE) as highlighted in Table 2. All kinds of errors were significantly higher than the expected performance.

The linear and random forest models performed slightly better with the "Trimmed Selection" feature set, as evidenced by the slightly lower MAE. This suggests that the model may have achieved better predictive accuracy with a more concise set of features. The lower MAE suggests that the simplified feature set led to a more accurate model. However, it is important to note that a slight decrease of 6% for MAE with no real decrease in RMSE or MSE does not reveal any notable trend.

The comparison between linear regression, random forest and Decision Tree regression indicates a slight advantage for the linear model, demonstrating a performance improvement of 1.7% over the random forest and Decision Tree counterparts. The default linear model with Trimmed Feature set showed the best prediction results amongst all the tested models. Nevertheless, given the scale of errors under consideration, this 1.7% reduction is deemed inconsequential, and it is concluded that none of the models were able to do a good job in predicting response times. Choice between them might depend on other factors like interpretability, computational efficiency, or the specific characteristics of the dataset.

Table 2. Summary of Erros from various Models

Model	Feature Set	RSME	MSE	MAE
Linear Regression	Initial Selection, 5 Columns included	5,431	29,496,364	1,347
Linear Regression	Trimmed Selection, 1 Column	5,447	29,674,876	1,269
Linear Regression (Optimized)	Trimmed Selection, 1 Column	5,984	35,809,372	1,382
Random Forest Regression	Initial Selection, 5 Columns included	5,587	31,211,324	1,291
Random Forest Regression	Trimmed Selection, 1 Column	5,778	33,385,551	1,430
Random Forest (Optimized)	Trimmed Selection, 1 Column	5,786	33,477,181	1,421
Decision Tree Regression (Optimized)	Trimmed Selection, 1 column	5,822	33,892,728	1,428

8. Conclusion

This study aimed to forecast response times for questions posted on the widely utilized programming-oriented platform, Stack Overflow. The University of California (UCI) website served as a robust source for the dataset. The data underwent a systematic preprocessing phase, including attribute title standardization, null value identification and removal, as well as rudimentary statistical analysis. Outliers were pruned, and extraneous features like userID were eliminated to enhance data quality. Additional columns were introduced to capture the first answer to a question and represent time more intuitively. A correlation matrix was instrumental in unraveling interactions among dataset attributes.

Three regression models—Linear Regression, Decision Tree Regression, and Random Forest Regression—were meticulously fitted with two distinct feature sets, optimizing hyperparameters. Results, however, unveiled significant prediction errors, with RME ranging in the mid to high 5000s and MAE in the low to mid 1000s. While the linear model demonstrated superior performance, the marginal error differentials and elevated absolute error values prompted the conclusion that none of the models provided accurate response time predictions. The absence of substantial correlation between response times and other dataset attributes was identified as a potential contributor to this discrepancy.

To overcome these challenges, future steps involve exploring alternative datasets and employing different models such as Gradient Boosted Tree, Isotonic Regression, and Factorization Machine Regression. Despite the inherent complexities encountered, this project has furnished valuable insights into the intricacies of the data, establishing a solid foundation for subsequent explorations and analyses.