Stack Overflow Questions Quality Classification

Group 11: Parv Thakkar

Introduction

- Stack Overflow is a website that is widely used by the developers to come together and solve common problems and help each other.
- There are many platforms like stack overflow but among them stack overflow stands out and has the greatest number of visits and user base.
- It is important for a question-and-answer platform to have good quality of questions as this makes easier for the person asking for help. This also reduces the response time as a good quality question will be answered fast compared to a question which requires edits.
- Therefore, this project is built to categorize the quality of the question into 3 categories namely High Quality (HQ), Low Quality requiring Edits (LQ_EDIT), and Low Quality Closed (LQ_CLOSE).

Dataset

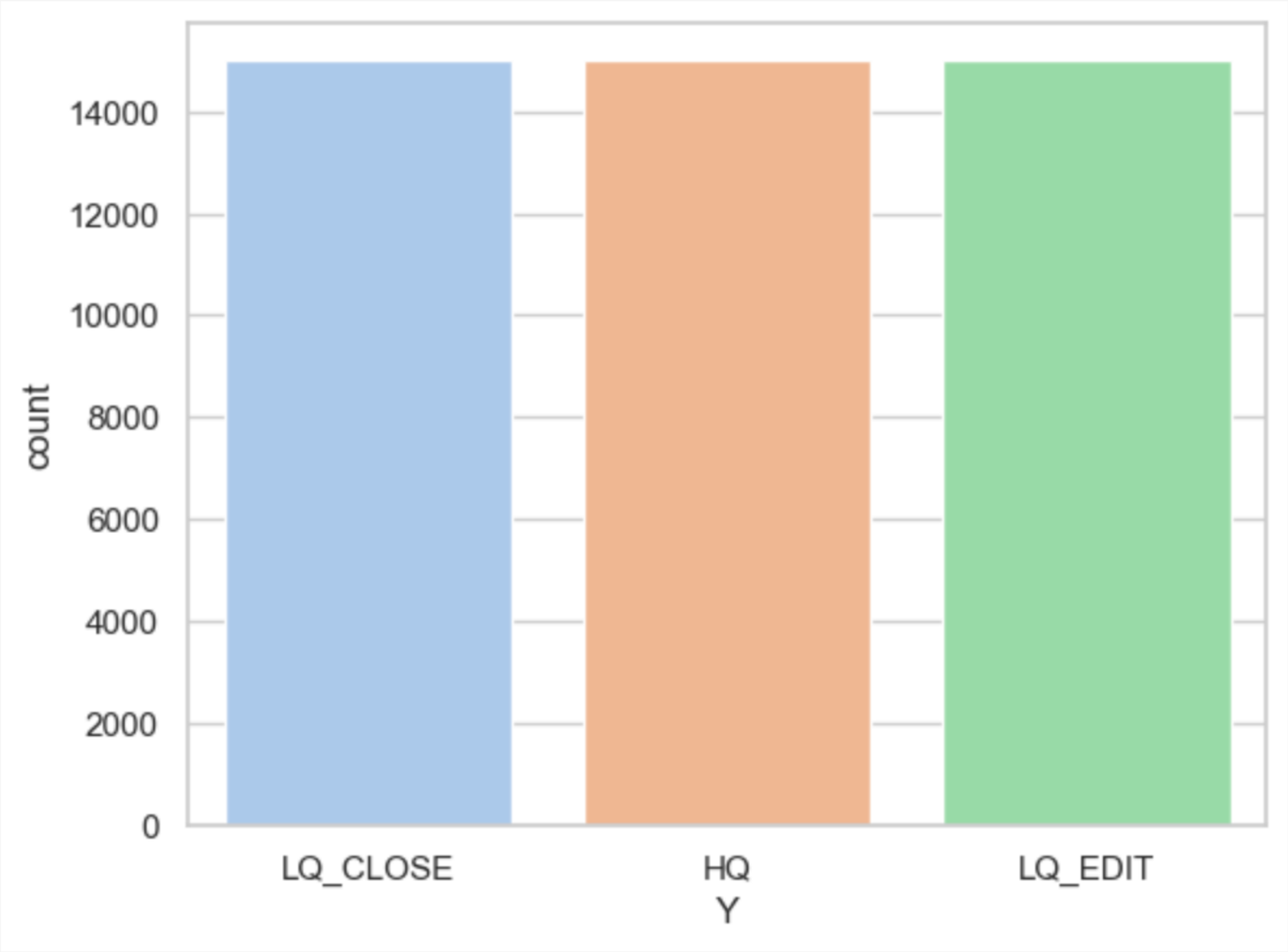
- Datasets consists of 2 CSV files – Train.csv and Valid.csv
- 6 columns in each file – ID, Title, Body, Tags, CreationDate, Y
- Train.csv consists of 45,000 stack overflow questions:
  - 33% are high quality questions.
  - 33% are low quality questions with edits.
  - 33% are low quality questions which are closed.
- Test.csv consists of 15,000 stack overflow questions.

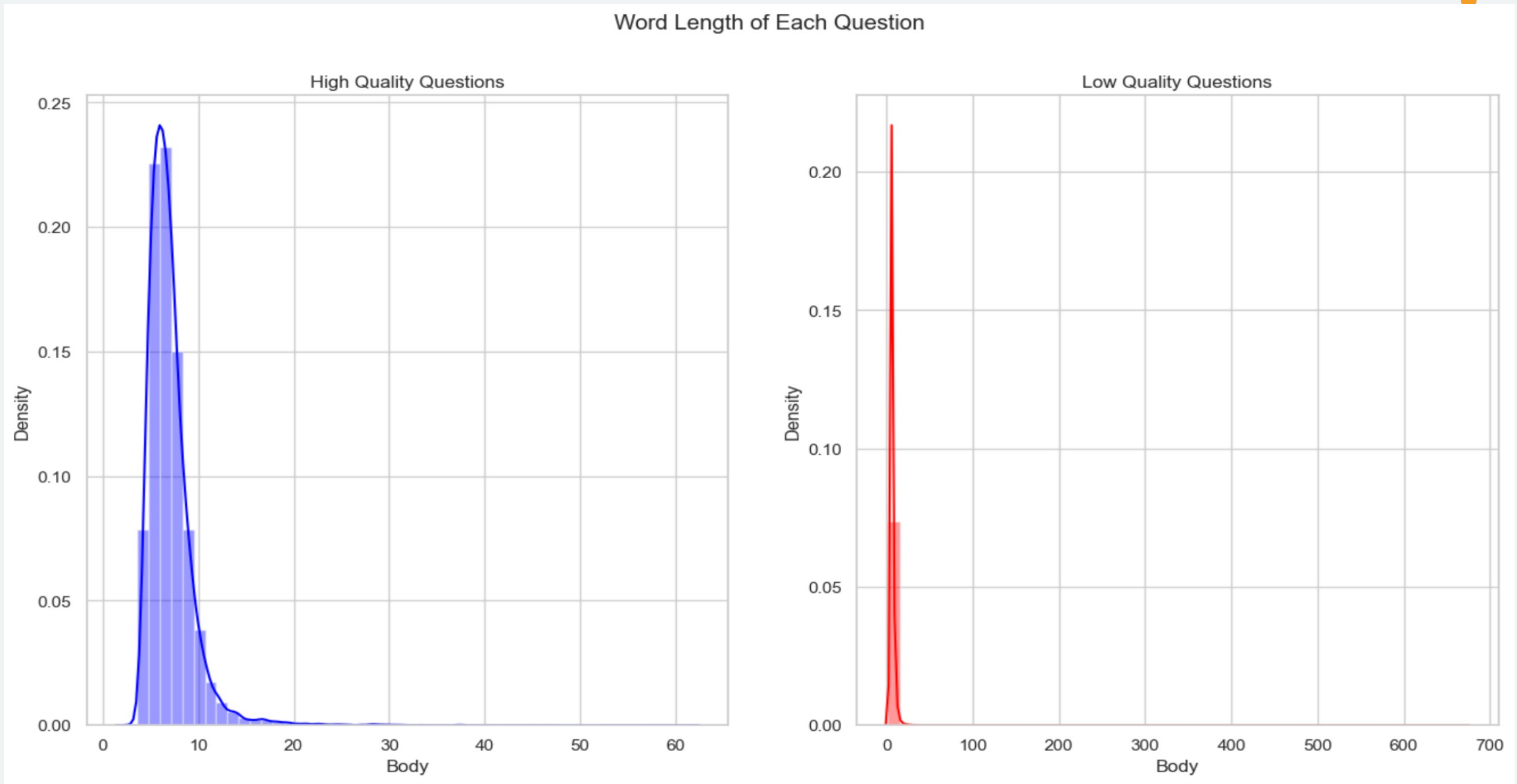
Data Pre-Processing

- This data pre-processing step includes the following tasks:
  - Dropping unnecessary columns.
  - Label Encoding the Target Variable.
  - Combining Title and Body into one column.
  - Removing HTML tags from the text.
  - Removing newlines and tabs/whitespaces/links/special characters/stopwords.
  - Removing punctuations.
  - Correcting misspelt words.
  - Lemmatization/Stemming.

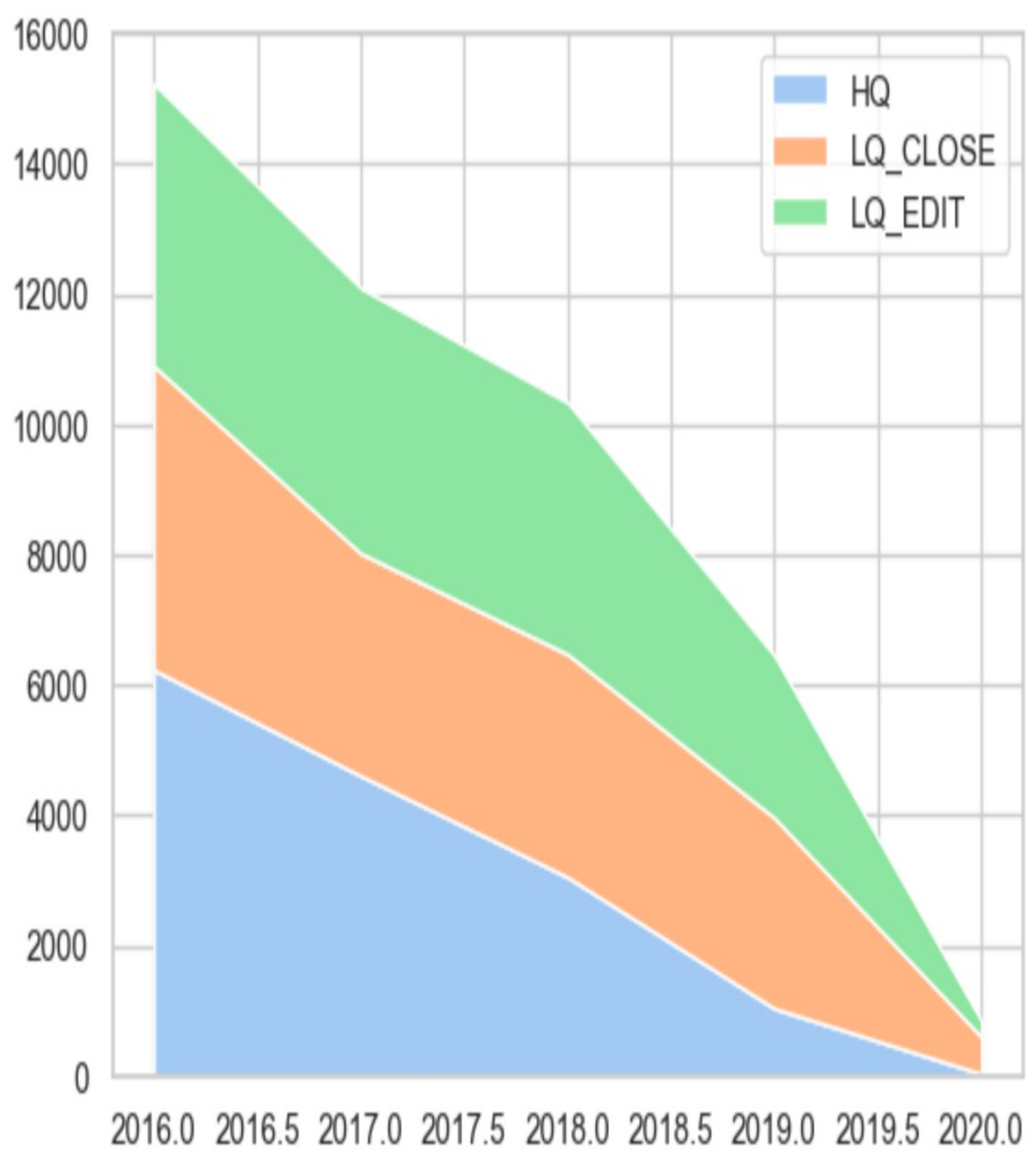
Exploratory Data Analysis

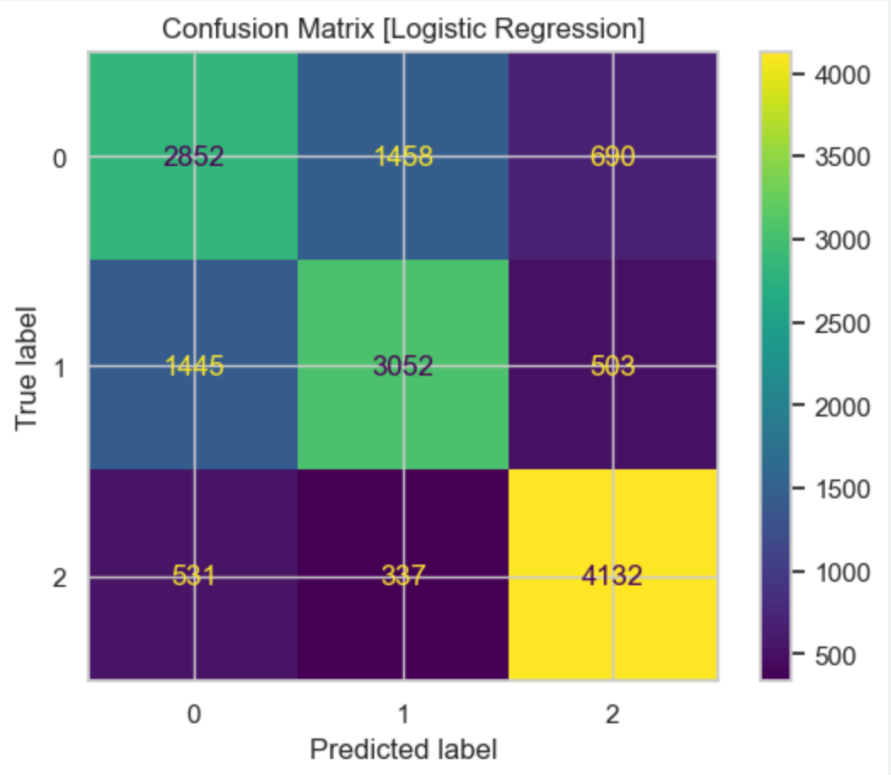
Target Variable Distribution

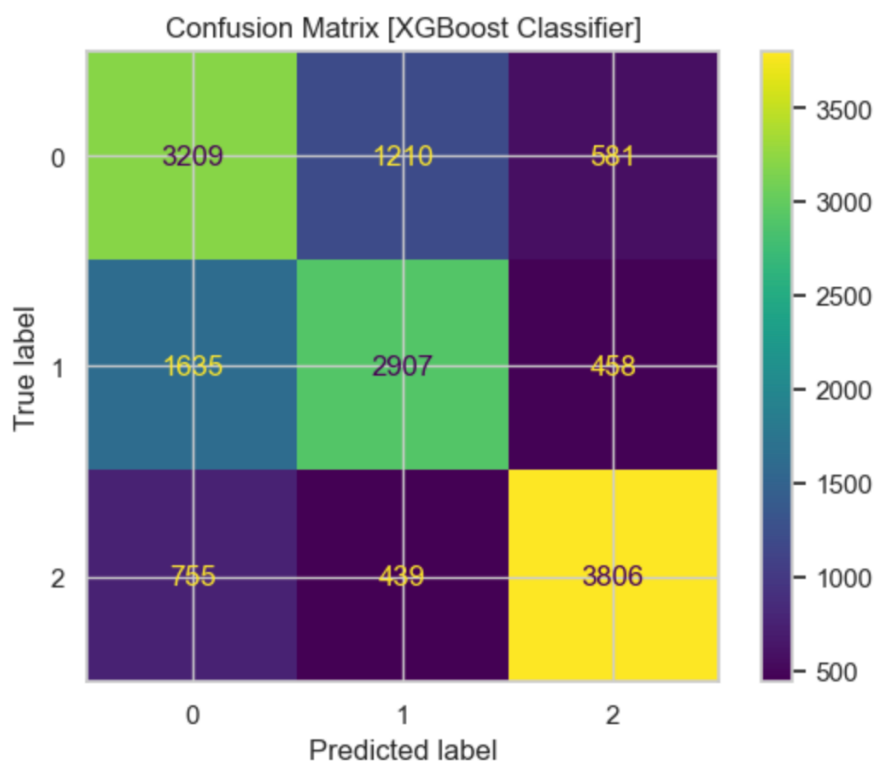


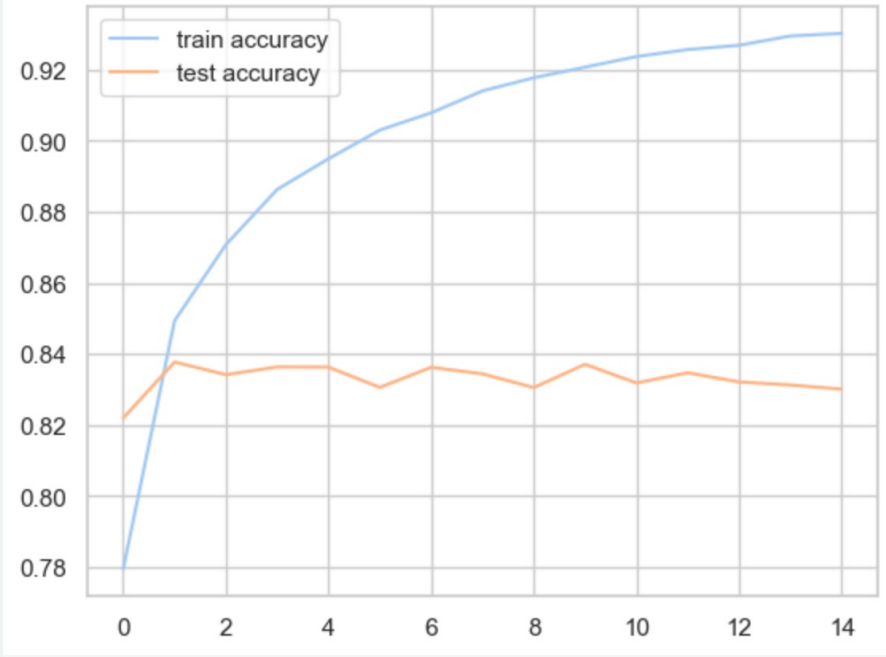

Word Length of Each Question
High Quality Questions
Density
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
60
Body
Low Quality Questions
Density
0.20
0.15
0.10
0.05
0.00
0
100
200
300
400
500
600
700
Body


Popular Technologies
javascript
ios
Name
dtype
android
python
Java
object
Length
azure
swift


16000
14000
12000
10000
8000
6000
4000
2000
0
2016.0
2016.5
2017.0
2017.5
2018.0
2018.5
2019.0
2019.5
2020.0
HQ
LQ_CLOSE
LQ_EDIT


Confusion Matrix [Logistic Regression]
True label
Predicted label
0
1
2
2852
1458
690
1445
3052
503
531
337
4132
4000
3500
3000
2500
2000
1500
1000
500


Confusion Matrix [XGBoost Classifier]
True label
Predicted label
0
1
2
3209
1210
581
1635
2907
458
755
439
3806
3500
3000
2500
2000
1500
1000
500


train accuracy
test accuracy
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0
2
4
6
8
10
12
14

Machine Learning Models

- **Logistic Regression:** It is a popular linear classification model that works well with large datasets.
- **Multinomial Naïve Bayes:** It is a probabilistic classifier that is commonly used for text classification tasks.
- **Random Forest Classifier:** It is an ensemble learning method that combines multiple decision trees to improve the accuracy of the model.
- **Decision Tree Classifier:** Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions.
- **K Nearest Neighbor Classifier:** It is a non-parametric classification model that determines the class of a new data point by finding the most similar training examples.
- **XGBoost:** It is a powerful gradient-boosting algorithm that is widely used for supervised learning tasks.

Deep Learning Model

- The model architecture consists of an input layer that takes in the tokenized sequences of questions with variable lengths.
- The next layer is an embedding layer with 2.5 million parameters which maps the input sequences to dense vectors of fixed size.
- The output of the embedding layer is fed into a bidirectional LSTM layer with 98,816 parameters.
- The bidirectional LSTM layer processes the input sequences in both forward and backward directions to capture the contextual information.
- The output of the first LSTM layer is then fed into a second bidirectional LSTM layer with the same number of parameters to further refine the contextual information.
- A dense layer with 32 units and a dropout layer is added to prevent overfitting.
- The final output layer has 3 units for the 3 classes: High Quality, Low Quality with Edits, and Low-Quality Closed.

Hyper parameter Tuning

Hyper Parameters	Training Data AUC	Test Data AUC
EPOCH = 15, BATCH SIZE = 32	93%	83%
EPOCH = 5, BATCH SIZE = 64	95%	82%
EPOCH = 10, BATCH SIZE = 64	96%	81%

Comparing Results

Model	Accuracy
Logistic Regression	67%
Multinomial Naïve Bayes	64%
Random Forest Classifier	64%
Decision Tree Classifier	51%
K Nearest Neighbors Classifier	54%
XGBoost Classifier	66%
LSTM	83%

Conclusion

- The project aim was to classify the stack overflow questions in categories based on the quality of the question.
- A dataset with 60,000 questions was taken from the Kaggle website was used to train and test the model.
- Performed Exploratory Data Analysis on the dataset to extract valuable insights.
- Built various machine learning models like Logistic Regression, Multinomial Naïve Bayes, Random Forest Classifier, Decision Tree Classifier, K Nearest Neighbors Classifier and XGBoost and compared their accuracies.
- Built a deep learning model using LSTM and achieved the highest Test and Train accuracy of 83% and 96% respectively.

Future Work

- Collect and work with a larger dataset with more recent Stack Overflow questions.
- Improve or Create new deep learning models to avoid overfitting.
- Try hyper parameter tuning for general machine learning models.
- Explore words or n-grams in questions that have a rating of low quality.

Thank You!