**Semester Project**

**Data Mining Steps/Strategies**

Team: T-10

Team members: Tianran Huang, Jialu Liu, Lisha Xia,

Shuyan Xia, Xiaochen Yan, Shiqin Zeng, Zhongyan Zheng

Project Title: Amazon QA Conversation Analysis

[**Task 1: Explore Data Characteristics 3**](#_cwko0st2875j)

[1.1 Exploratory Variables 3](#_qm4yqoa3ofbj)

[1.1.1 Reviews Distribution Based on Date 4](#_crjza416diei)

[1.1.2 Word Characteristics 5](#_geq56jk687hu)

[(a) Sentiment Polarity Distribution 5](#_gxszv04weguy)

[(b) Average Sentiment Polarity per Day 5](#_p1z5h6rkpjaj)

[(c) Topic Modeling Using Review Texts 5](#_b3on5151ffk)

[1.1.3 Numerical Data Characteristics 6](#_fy5rg3zaby4n)

[1.2 Response Variables 6](#_ay5pg5ublu4u)

[**Task 2: Develop investigation steps using background knowledge from the literature review 6**](#_szqlnn828jqq)

[2.1 Topic modeling 7](#_4froznk2wp3b)

[2.2 Sentiment Analysis 8](#_gahgkl4gtavn)

[2.3 Regression/Classification 10](#_2jn3g97032cr)

[2.3.1 Feature Engineering: 10](#_938g6ihhzsz2)

[2.3.2 Model Selection: 10](#_5vr5ko92nbc8)

[2.3.3 Model Training and Evaluation: 11](#_tc6nncuxsvx)

[2.3.4 Evaluate metrics: 11](#_ef17o53w0aop)

[**Task 3: Make plans of how and what to study the “properties” of the developed procedure 11**](#_6lh41pisi26c)

[3.1 Random Forest Classification 12](#_pr6kmwlulo6o)

[3.1.1 Confusion Matrix 12](#_phgbzymuhbjm)

[3.1.2 Accuracy 12](#_xq0k39siggi2)

[3.1.3 Precision 13](#_aro223z68109)

[3.1.4 Recall (Sensitivity) 13](#_c2iv9evltmn)

[3.1.5 F1 Score 13](#_1ns5jfdd8r6b)

[3.2 Regression Metrics 13](#_m5jiaic5tehd)

[3.2.1 Mean Squared Error(MSE) 13](#_ayzw85y05agp)

[3.2.2 Mean Absolute Percentage Error(MAPE) 14](#_ezlixvgbirdk)

[3.3 Cross Validation 14](#_4qj5iar8geci)

# **Task 1: Explore Data Characteristics**

One product was explored to explain the data characteristics as shown in Figure 1:

# 

Figure 1: Product instance in Amazon

## 

## 1.1 Exploratory Variables

Review Section: Using an electronic scale product as an example, one entry from the raw data is presented as follows:

{

"reviewerID": "AND3GQSC4A06R",

"asin": "B001KXZ808",

"reviewerName": "NaN",

"helpful": [0, 0],

"reviewText": "We have had this scale for almost a year and a half; it is working really well overall. It is considerably more consistent and our favorite out of all of the scales that we have owned. Sometimes it does like to fluctuate the poundage on you and you have to weigh a few times to get a consistent reading. (Reason for the 4 not 5 stars.)My wife also says that one of her main downsides is that she feels like it is a little delicate with it being glass with our kids. But in saying that, with our kids, this scale gets rough-housed and it is holding up very well.",

"overall": 4.0,

"summary": "Overall Great Scale",

"unixReviewTime": 1405987200,

"reviewTime": "2014-07-22"

}

### 1.1.1 Reviews Distribution Based on Date

We sorted and selected the most recent 1,000 entries from the dataset. The data distribution based on date is described in Figure 2, and sentiment analysis is shown in Figure 3.

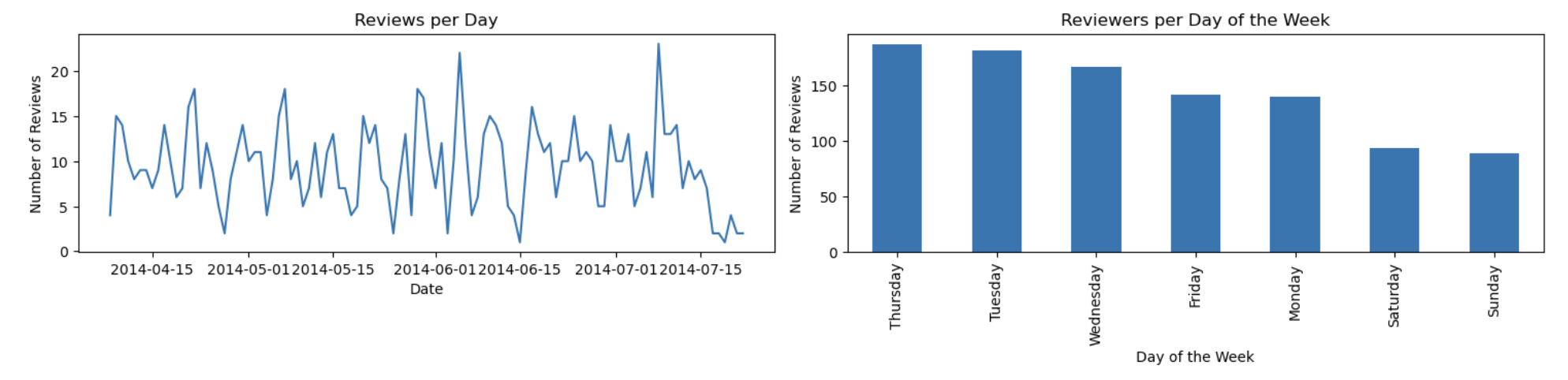


Figure 2: Data distribution based on date

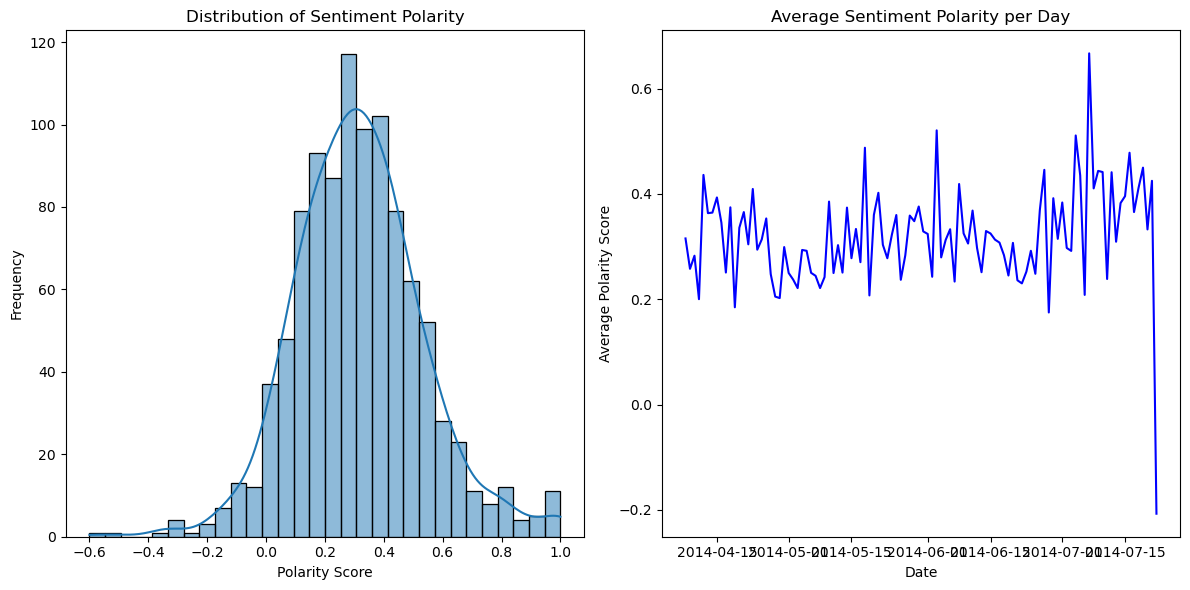


Figure 3: Sentiment analysis distribution

### 

### 1.1.2 Word Characteristics

#### (a) Sentiment Polarity Distribution

The histogram illustrates the distribution of sentiment polarity scores found in all the reviews. These scores vary from -1, representing a highly negative sentiment, to +1, denoting a highly positive sentiment. Scores close to 0 signify a neutral sentiment. Analyzing the shape of this distribution helps in understanding the general sentiment towards the product, whether it leans more toward positive or negative.

#### (b) Average Sentiment Polarity per Day

This line graph represents the daily average sentiment polarity. Variations in the graph reflect daily shifts in the product's overall reviews. High points signal days marked by predominantly positive discussions, whereas low points reveal days with more negative sentiments. Such visualizations are crucial for grasping the product's review trends over time. They become especially revealing when linked to particular issues or subjects addressed in the QA section, providing deep insights into the dynamics of product feedback.

#### (c) Topic Modeling Using Review Texts

For each review, the topic model provides a distribution over topics (Figure 4), indicating the presence and proportion of each topic within the review. Transform these topic distributions into a vector of topic weights for each review, and each element in this vector represents the weight or contribution of a topic to the review, effectively summarizing its thematic content. Finally, we can extract the reviews-topic distribution as a feature set.

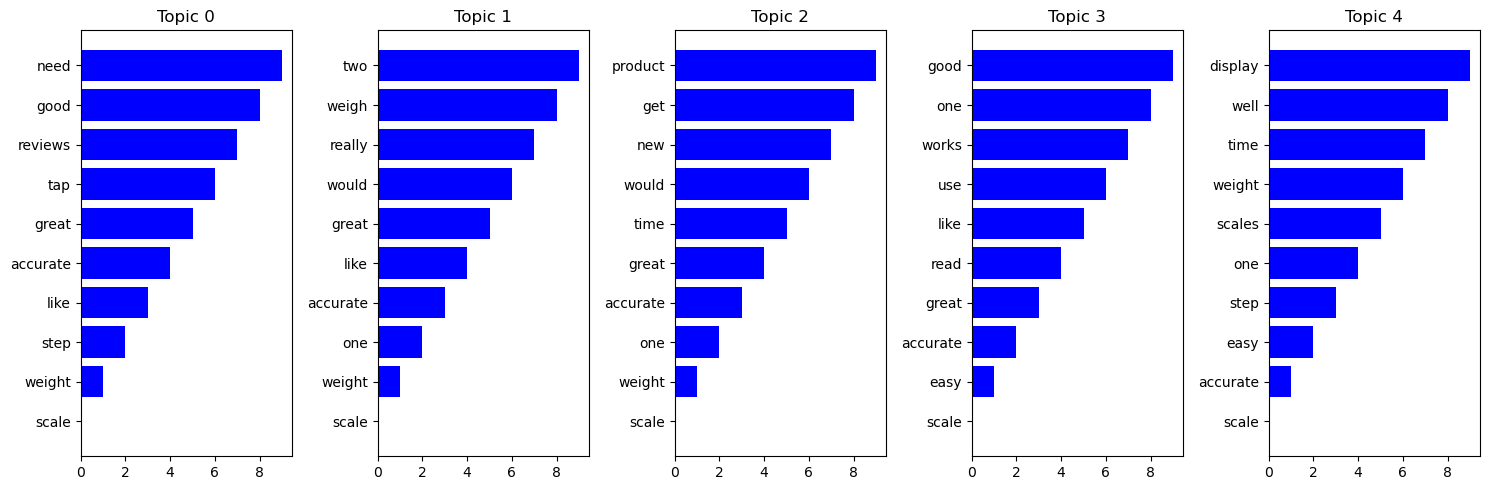


Figure 4: Distribution of topics based on topic modeling

## 

### 1.1.3 Numerical Data Characteristics

Review Section: Helpful: [a,b] transfer the list to the total number of the responses (b) and a ratio to present the helpful ratio (a/b) respectively.

Product price: a numerical number of the price.

Review date: transfer the date to the year, month, day, and week of the day.

## 1.2 Response Variables

Overall score: [1.0, 2.0, 3.0, 4.0, 5.0]

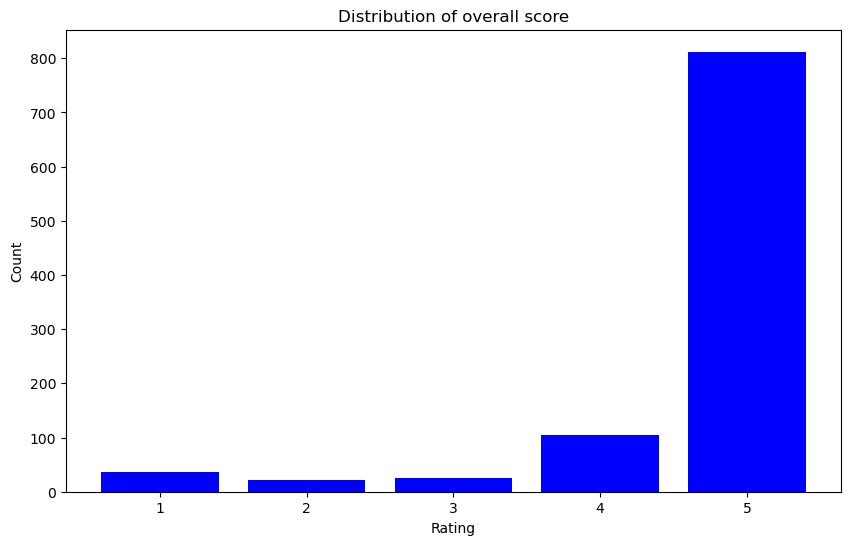


Figure 5: Distribution of overall score

The result (Figure 5) shows there is a significant disparity in overall rating distribution. The most defining characteristic is the large disparity in the number of instances among different rating scores. Over 90% of the data belongs to the rating 5 class. Several resampling methods will be considered to the original dataset, or use ensemble learning techniques that combine multiple models to improve the prediction of minority class examples.

# **Task 2: Develop investigation steps using background knowledge from the literature review**

Given that the majority of the data consists of text, the data mining effort concentrated on converting this qualitative data into a quantitative form for incorporation into the model. Various researchers have investigated multiple methods, such as extracting topics from reviews and analyzing the sentiments expressed in them. In this section, we will offer a summary of these techniques employed in our data mining study.

## 2.1 Topic modeling

This process begins by preparing the text data through several steps: data cleaning, lowercasing, tokenization, stopwords removal, lemmatization, and stemming. It then extracts the text features using the Bag Of Word (BOW) method and the Term Frequency-Inverse Document Frequency (TF-IDF) and constructs a dictionary and a corpus essential for employing Latent Dirichlet Allocation (LDA) for topic modeling. The LDA model is executed to unearth the underlying topics within the review data. It highlights the keywords tied to each topic. The outcome is a collection of topics, where each is delineated by words that hold the most relevance to it. This approach offers insights into the primary themes discussed for each product. Here are the detailed investigation steps:

1. Data Pre-Processing
   1. Data Cleaning
      1. For missing values and duplicate reviews, we will remove the entire row of the product review.
      2. Remove punctuation and special characters.
   2. Lowercasing
      1. Convert all text to lowercase ensures consistency and reduces the complexity of the data.
   3. Tokenization
      1. Break down the reviews into individual words or tokens.
   4. Stopwords Removal
      1. Remove common stopwords like "and", "the", "is" and words with less than 3 characters.
   5. Lemmatization
      1. Change words to their base or root form (e.g., "walking" to "walk").
   6. Stemming
      1. Removes prefixes and suffixes to reduce words to their stems (e.g., "walking" to "walk").
2. Feature Extraction
   1. Use the Bag Of Word (BOW) method to represent text data.
   2. Use the Term Frequency-Inverse Document Frequency (TF-IDF) to evaluate the relevance of a word within a document in a collection of documents.
3. Modeling
   1. Use Latent Dirichlet allocation (LDA) to find the dominant topic with two different inputs, one from BOW, and another one from TF-IDF.
   2. After getting the dominant topic, merge the keyword and the dominant keyword as the feature data, with the category as the y data.
   3. Divide the data with 0.1 as the test size and 0.9 as our train size.
   4. Classify the reviews based on X to the right categories via algorithms such as logistic Regression and Multi-Layer Perceptron.

## 2.2 Sentiment Analysis

Sentiment analysis can help understand customer satisfaction, product strengths and weaknesses, and consumer trends. There are primarily three types of sentiment analysis:

1. Lexicon-based (unsupervised): Relies on a predefined lexicon (or dictionary) of words that have been assigned positive, negative, or neutral sentiments. This method calculates sentiment scores based on the presence and combination of these words in the text.
2. Machine Learning-based (supervised): Involves training a model on a dataset where the correct sentiments are known to predict sentiments on new, unseen data.
3. Deep Learning: Uses neural networks to capture deeper linguistic patterns and context. This approach often results in higher accuracy but requires more computational resources and a larger dataset for training.

Given that the scores corresponding to the reviews are known and considering the nuances and complexities of human language, we choose to conduct the Machine Learning-based approach for our analysis.

Steps to Conduct Sentiment Analysis using Machine Learning:

1. Data Preprocessing:
2. Cleaning: Remove special characters and extra white spaces.
3. Tokenization: Break down each text piece into individual words or tokens.
4. Normalization: Convert all words to lowercase to ensure consistency.
5. Removing Stopwords: Remove words such as "and" and "the" that do not contribute meaning to the sentiment.
6. Stemming/Lemmatization: Reduce words to their base or root form.
7. Feature Extraction:
8. Convert text data into numerical features that machine learning models can process. Common methods include BOW, TF-IDF as in the Topic Modeling session above, and word embeddings (e.g., Word2Vec, GloVe).
9. Model Selection and Training:
   1. Choose a machine learning model suitable for text classification or regression. Popular choices include Logistic Regression, Naive Bayes, Support Vector Machines (SVM), and deep learning models like LSTM (Long Short-Term Memory) networks.
   2. Split the data into training and testing sets.
   3. Train the selected model on the training set and tune it using cross-validation.
10. Evaluation:
    1. Evaluate the model's performance on the testing set using metrics such as accuracy, precision, recall, or F1 score.
    2. Perform error analysis to understand where the model is making mistakes and why.
11. Prediction and Insight Generation:
    1. Apply the trained model to new product reviews to predict sentiment.
    2. Analyze the sentiment results to extract actionable insights and interpretations.
    3. Identify common themes in positive and negative reviews to help inform product improvements, marketing strategies, and customer service approaches.

## 2.3 Regression/Classification

Given the nature of the data and the goals outlined in the Literature Review, Theme Modeling, and Sentiment Analysis sections, we first approached the problem using a classification approach. By categorizing reviews into different satisfaction or quality ratings, it is possible to tell us indirectly about product feedback scores.

Therefore, our strategy turns to the use of classification models that classify product reviews based on features extracted from topic modeling and sentiment analysis. This approach not only allows us to predict the sentiment (positive, negative, neutral) associated with each review, but also to categorize them based on the themes revealed by the LDA model. Here is how we will proceed:

### 2.3.1 Feature Engineering:

From Theme Modeling: Use the dominant themes of each comment and their distribution as features. These themes, represented as categorical variables, indicate the dominant themes discussed in the comments.

From Sentiment Analysis: Converting sentiment scores into categorical labels (positive, negative, neutral). In addition, the frequency of positive and negative keywords identified through text mining will be used as features indicating the tone and sentiment of the comments.

### 2.3.2 Model Selection:

Considering that our goal is to categorize reviews into predefined categories (topics) and sentiments, we chose to use classification models such as the Random Forest classifier, which are well suited to deal with the categorical nature of the features and predict discrete results. Random Forest is an integrated learning approach that improves overall prediction accuracy and robustness by combining predictions from multiple decision trees. The reasons for choosing Random Forest as a classification model include:

1. Ability to handle high-dimensional data: random forests are able to handle situations where the number of features far exceeds the number of samples, which is common in text data analysis. By constructing multiple trees and randomly selecting a subset of features at each decision point, it reduces the risk of overfitting the model.
2. Resistance to overfitting: Random Forest mitigates the overfitting problem that can occur with a single tree by integrating multiple decision trees. The randomness of each tree ensures the overall robustness of the model.
3. Feature Importance Assessment: Random Forest provides an intuitive understanding of which features have the greatest impact on predictive outcomes, which is valuable for interpreting the model's predictions and understanding key factors in our textual data.

### 2.3.3 Model Training and Evaluation:

Data Segmentation: First, divide the data into a training set and a test set. Common ratios include 70% training data and 30% test data, or use cross-validation to further improve model stability and accuracy.

Training the model: Use the training set data to train the random forest model. This includes choosing the right number of trees, tree depth, and other hyperparameters, which can be optimized by grid search and cross-validation.

### 2.3.4 Evaluate metrics:

We will evaluate the models using classification metrics such as accuracy, precision, recall, and F1 scores to ensure that they effectively categorize comments.

By analyzing the categorization results, we can understand which features (topics, sentiment scores, keyword frequency) are most predictive of review categorization and will highlight product aspects that influence customer feedback.

# **Task 3: Make plans of how and what to study the “properties” of the developed procedure**

Based on the description of the response variable, the prediction could be a classification or a regression.

## 3.1 Random Forest Classification

### 3.1.1 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model and shows the counts of true positives, false positives, true negatives, and false negatives. It is important to understand the types of errors the model is making. It is used to identify a detailed breakdown of the model's predictions and provides insights into the specific types of errors (e.g., false positives, false negatives) made by the model, helping in targeted improvements.

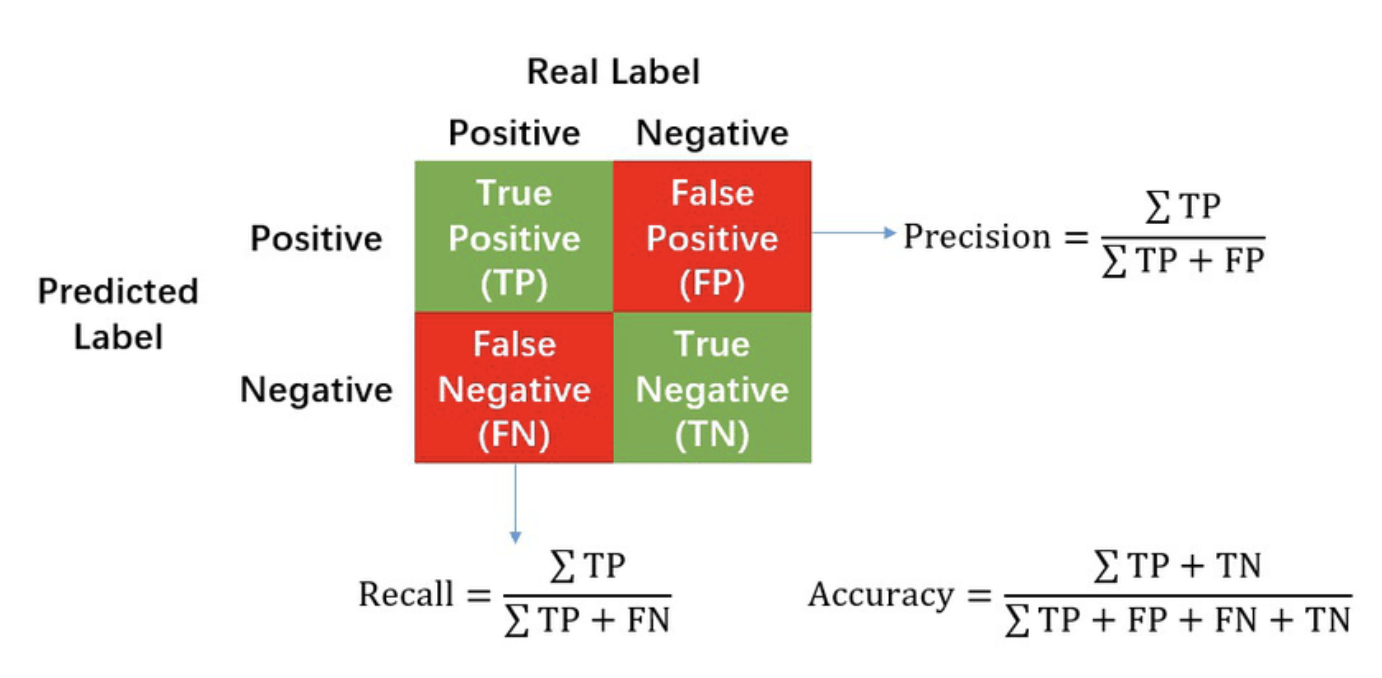


Figure 2. Calculation of Precision, Recall and Accuracy in the confusion matrix.

### 3.1.2 Accuracy

Accuracy is the ratio of correctly predicted instances to the total instances in the dataset, providing a basic measure of the model's correctness. It is important for a general assessment of the model's performance, especially when classes are balanced and the cost of false positives and false negatives is similar. Accuracy is easy to understand and calculate, providing a high-level view of model performance compared to other metrics.

### 3.1.3 Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, measuring the model's exactness, especially when the cost of false positives is high. It ensures that positive predictions made by the model are accurate, making it crucial in scenarios where false positives are costly.

### 3.1.4 Recall (Sensitivity)

Recall is the ratio of correctly predicted positive observations to all observations in the actual class, measuring the model's ability to capture all positive instances. F1 score is important when there is an imbalance between the classes or when both false positives and false negatives are costly, providing a balanced measure of precision and recall.

### 3.1.5 F1 Score

The F1 score is the weighted average of Precision and Recall. It provides a balance between precision and recall, making it useful when there is an uneven class distribution or when both false positives and false negatives are costly. Thus it is useful in imbalanced datasets, providing a single metric that balances precision and recall.

## 3.2 Regression Metrics

### 3.2.1 Mean Squared Error(MSE)

MSE is the average of the squared differences between predicted and actual values. Larger values indicate larger errors.

MSE=

MSE is sensitive to large errors due to squaring the differences, commonly used and easy to interpret. Also it is useful for optimization algorithms as it provides a smooth, differentiable function.

### 3.2.2 Mean Absolute Percentage Error(MAPE)

MPE measures the percentage difference between predicted and actual values. MAPE takes the absolute value of it and provides a percentage measure of the average absolute error. MAPE is easy to understand and interpret, providing a straightforward percentage measure of accuracy.

MAPE=

## 3.3 Cross Validation

Use k-fold cross-validation to assess how well the model generalizes to new data. Train and evaluate the model on different subsets of the data to ensure robust performance.

* Cross-validation provides a more robust evaluation of a model's performance compared to a single train-test split.
* Cross-validation simulates the model's performance on different subsets of the data, giving insights into how well it generalizes to new, unseen samples.
* Cross-validation helps prevent data leakage, where information from the test set unintentionally influences the training process. Each fold is treated as an independent evaluation, reducing the risk of biased performance estimates.
* It provides a more comprehensive assessment of a model, which helps identify whether the model's performance is consistent or varies significantly based on the data it encounters.