Manish Patel

Automated Customer Complaint Classification

Data Science Intensive Capstone Project, August 2023 Cohort

Problem_Identification

- The company is seeking to develop a model that can accurately categorize customer complaints according to the products and services it provides. The complaints are in the form of json text format, and the company aims to automate the process of evaluating and assigning each complaint to the appropriate department.
- The company plans to speed up the process of resolving customer issues by segregating the clusters. Efficient handling of complaints can result in increased customer satisfaction and stronger loyalty.
- The solution should enable the company to streamline its customer support ticket system, improve service quality, and enhance customer experience.

Data Information

Data acquisition

https://www.kaggle.com/datasets/venkatasubramanian/automatic-ticket-classification

No of Records : 78313

No of fields : 22

Data Acquired for the Period : 2011-2021

Project Objective

- To reach our desired goal, it is crucial to extract relevant information from the 'complaint_what_happened' column. This column is likely to contain feedback or responses provided by users and customers. Through analysis of this data, we can gain valuable insights and reach informed conclusions that can greatly influence our decision-making process. Therefore, it is of utmost importance to efficiently acquire and interpret the data from this column.

- Develop a model that can accurately categorize customer complaints into one of the following five clusters:

  1. Credit card / Prepaid card
  2. Bank account services
  3. Theft/Dispute reporting
  4. Mortgages/loans
  5. Others

Text Preprocessing Pipeline

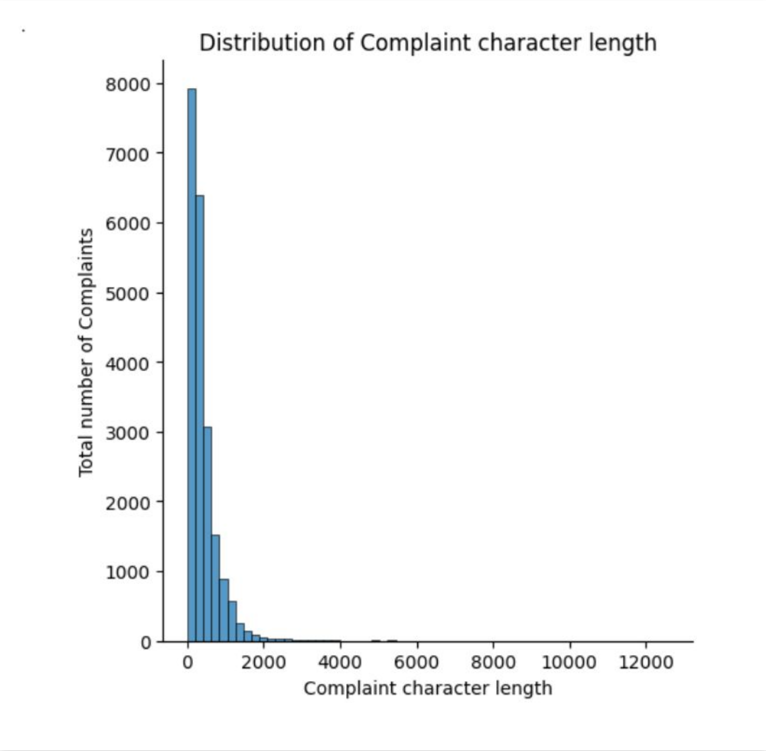
After filtering out empty complaints, our next steps involve:

- Lowercasing all text entries.
- Eliminating content enclosed in square brackets.
- Stripping away punctuation marks.
- Omitting words containing numerical digits.

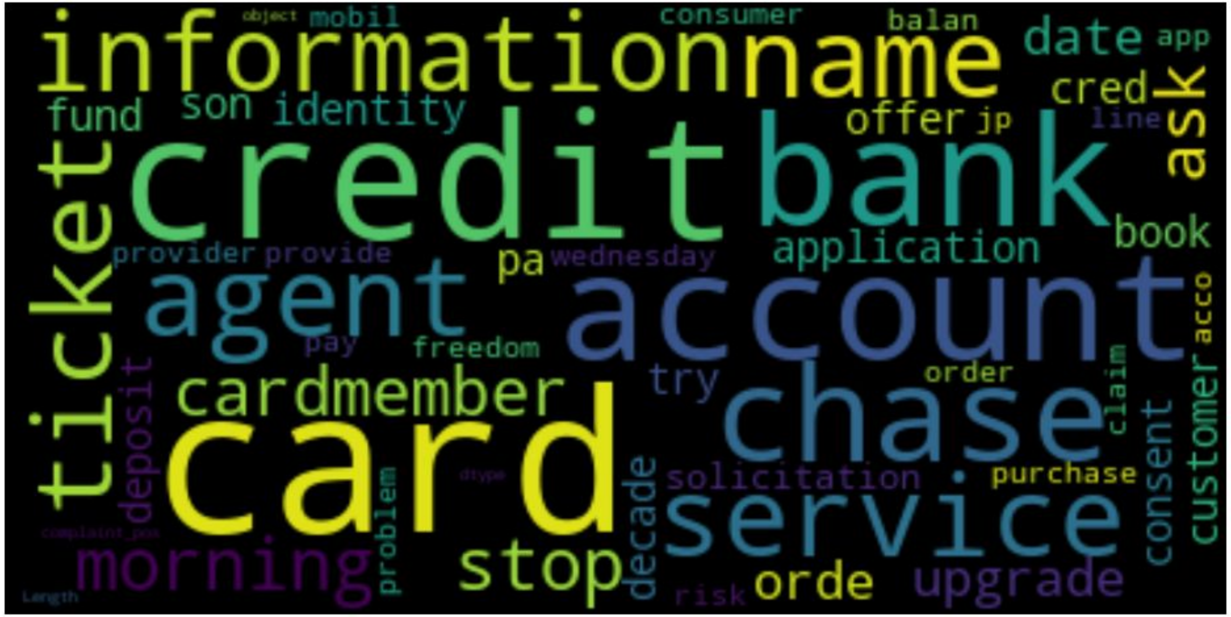
Following these cleaning procedures, we proceed to:

- Lemmatize the text, transforming words into their base or dictionary forms (e.g., 'running' to 'run', 'better' to 'good').
- Employing POS tags to extract relevant words from the texts."

"The following graph displays the length of each complaint in terms of character count "

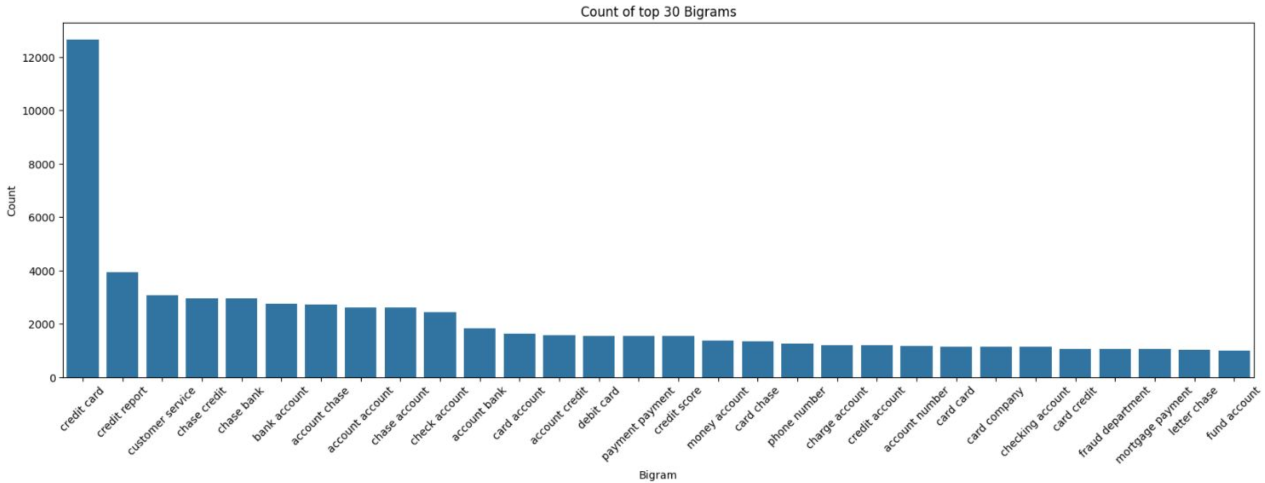


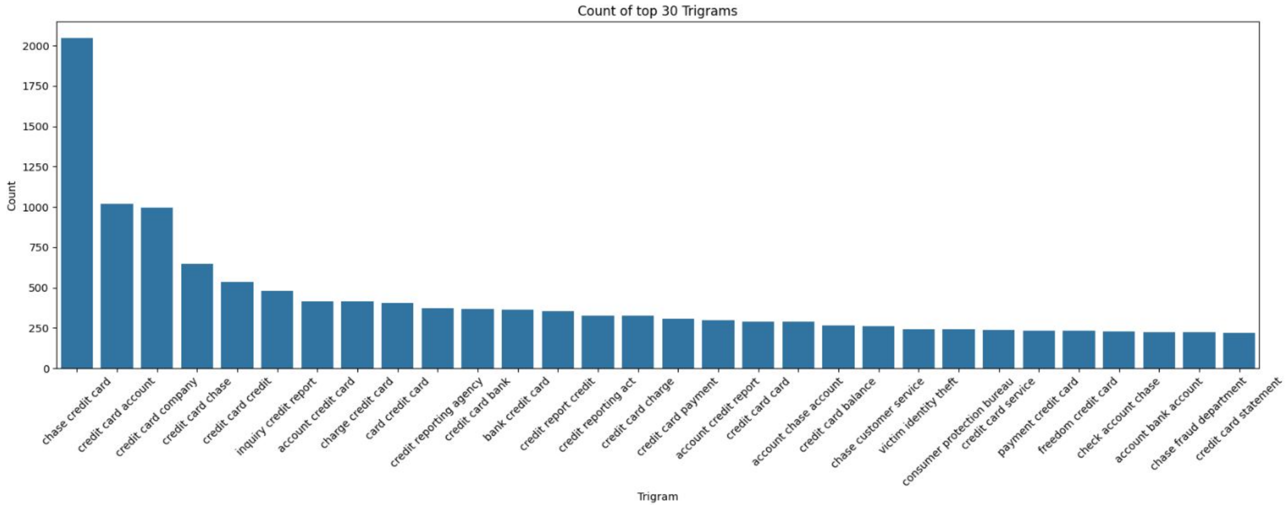
"After processing the text, a word cloud was generated to highlight the top 60 words by frequency among all the articles."



Bigram Trigram

Our analysis utilized bigram and trigram analysis techniques to thoroughly explore the dataset. Through this approach, we uncovered the frequency of word combinations and their associations. These findings are encapsulated in bar plot, offering a succinct and informative overview of our research outcomes.





Topic Modeling

We're working through a series of tasks and processes to complete our project successfully. Our goal is to deliver a high-quality outcome that exceeds expectations.

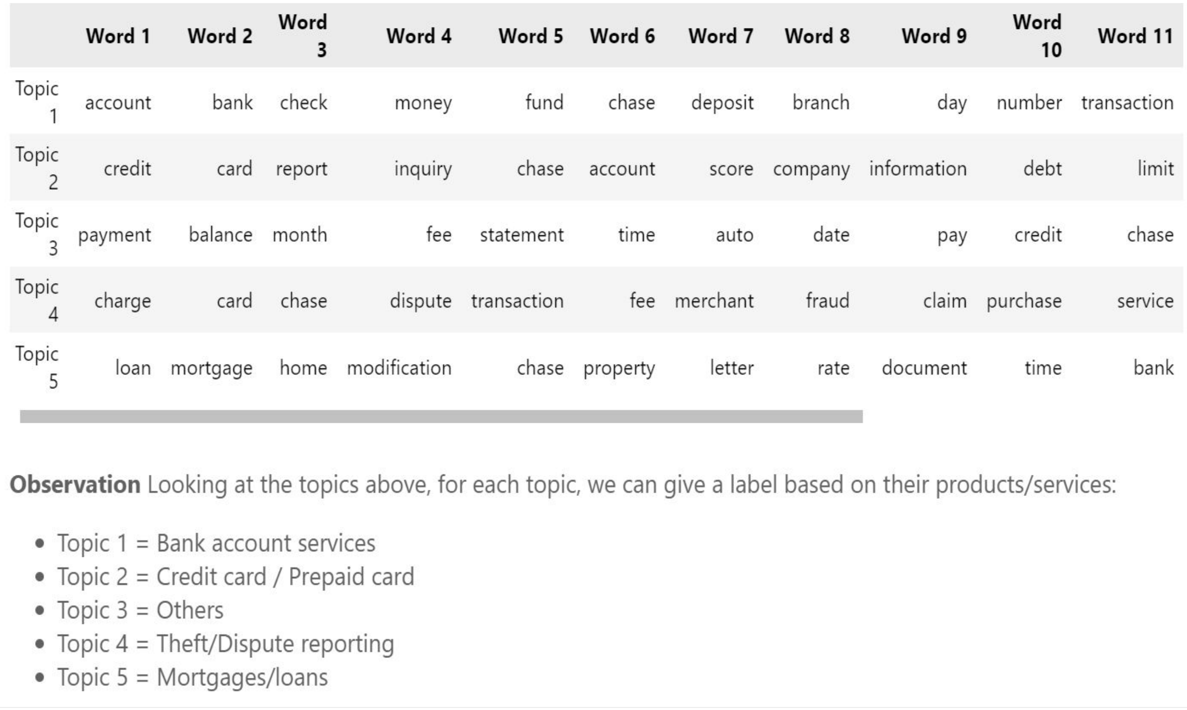
TFIDF:- We used TFIDF, a technique in NLP, to determine the importance of a word in a document. It creates a weight for each word based on its importance within the document and rarity across all documents in the corpus.

DTM:- The Document-Term Matrix (DTM) is a tool used in natural language processing that represents the frequency of terms or words in each document within a corpus. It's a matrix where rows represent documents and columns represent terms.

Non-Negative Matrix Factorization (NMF): NMF is a technique to uncover latent topics within text data. It breaks down data into smaller matrices and identifies patterns. This helps to understand underlying themes present within a large corpus of text data. It has various applications such as improving search algorithms and developing effective marketing strategies

Topic Modeling

Following the application of three different techniques, namely TFIDF, DTM, and NMF to our dataset, we were able to obtain a comprehensive data frame that showcases five distinct topics that were discovered from the data.



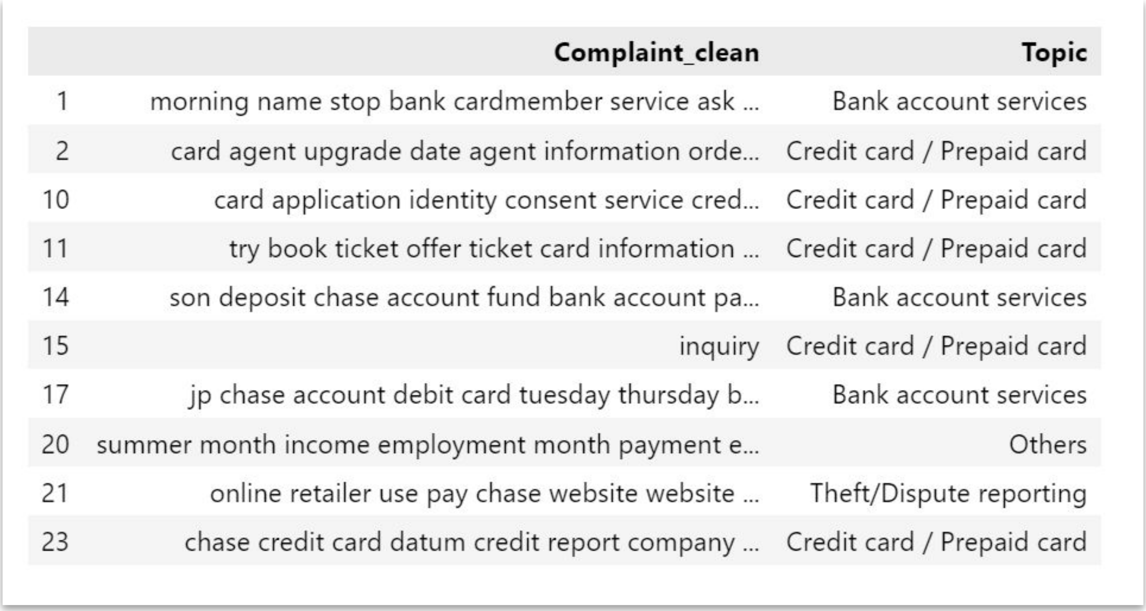
	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11
Topic 1	account	bank	check	money	fund	chase	deposit	branch	day	number	transaction
Topic 2	credit	card	report	inquiry	chase	account	score	company	information	debt	limit
Topic 3	payment	balance	month	fee	statement	time	auto	date	pay	credit	chase
Topic 4	charge	card	chase	dispute	transaction	fee	merchant	fraud	claim	purchase	service
Topic 5	loan	mortgage	home	modification	chase	property	letter	rate	document	time	bank

Observation Looking at the topics above, for each topic, we can give a label based on their products/services:

- Topic 1 = Bank account services
- Topic 2 = Credit card / Prepaid card
- Topic 3 = Others
- Topic 4 = Theft/Dispute reporting
- Topic 5 = Mortgages/loans

Topic Modeling

We found multiple topics in our customer complaint dataset using topic modeling. To aid in further analysis, we created a new variable named 'Topic' to capture the primary customer issues. The aim is to enhance the accuracy and efficiency of our models in resolving customer complaints.



	Complaint_clean	Topic
1	morning name stop bank cardmember service ask ...	Bank account services
2	card agent upgrade date agent information orde...	Credit card / Prepaid card
10	card application identity consent service cred...	Credit card / Prepaid card
11	try book ticket offer ticket card information ...	Credit card / Prepaid card
14	son deposit chase account fund bank account pa...	Bank account services
15	inquiry	Credit card / Prepaid card
17	jp chase account debit card tuesday thursday b...	Bank account services
20	summer month income employment month payment e...	Others
21	online retailer use pay chase website website ...	Theft/Dispute reporting
23	chase credit card datum credit report company ...	Credit card / Prepaid card

Modeling

CountVectorizer:- We used CountVectorizer, a tool from the sci-kit-learn library in Python, to convert our text data into a numerical representation. This helped us identify the frequency of each word and prepare the dataset for machine learning model development.

TfidfTransformer:- TfidfTransformer is a technique used in NLP and information retrieval to extract the most relevant features or keywords from text. It reduces the weight of frequently occurring words and scales down their impact to highlight less common but more informative words. The tf-idf score for a term is calculated by multiplying its frequency in a document by the inverse frequency of the term in the corpus.

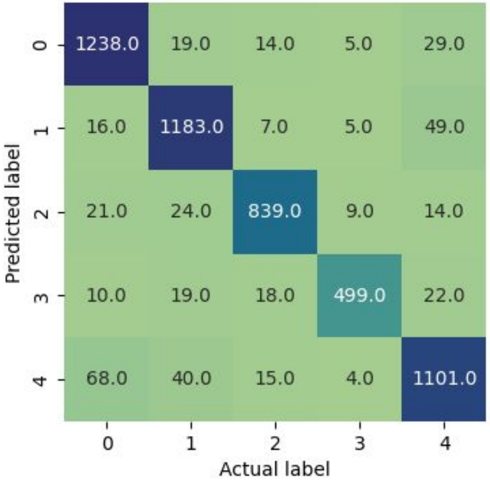
Training and Testing:- Our dataset is divided into two parts- 75% for training and 25% for testing to ensure accurate results.

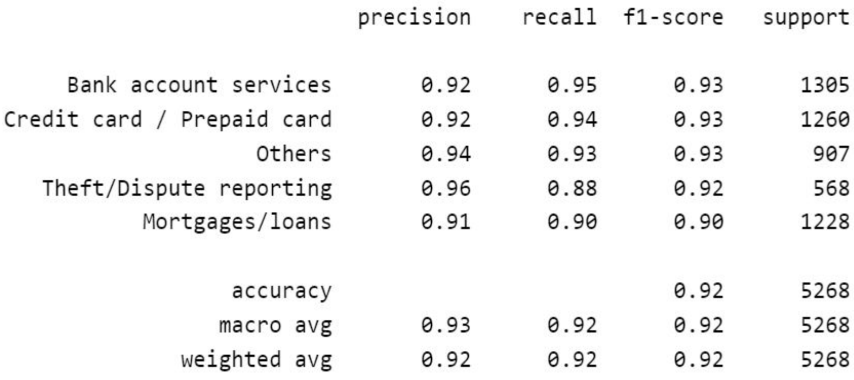
Modeling

Models we used for modeling:

Model 1 - Logistic Regression
Model 2 - Decision Tree Classifier
Model 3 - Random Forest Classifier

Model 1 - Logistic Regression:- We improved the logistic regression accuracy score from 91.91% to 92.25% after tuning the hyperparameters. We also achieved good precision, recall, and f1-scores for all topics. Check figure for the results.





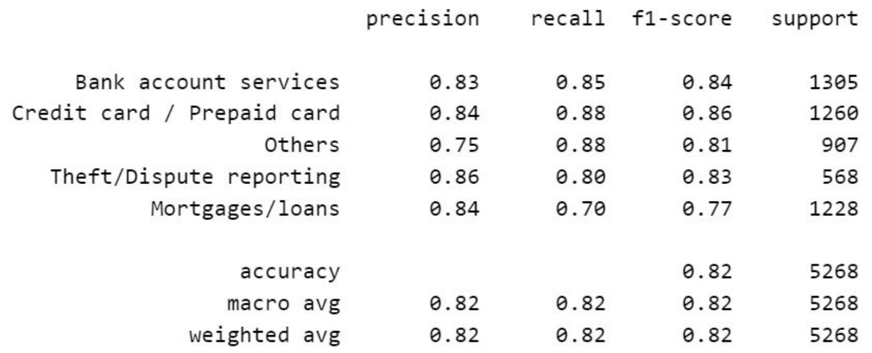
```
                          precision    recall  f1-score   support

   Bank account services       0.92      0.95      0.93      1305
Credit card / Prepaid card       0.92      0.94      0.93      1260
                  Others       0.94      0.93      0.93       907
Theft/Dispute reporting       0.96      0.88      0.92       568
         Mortgages/loans       0.91      0.90      0.90      1228

                accuracy                           0.92      5268
               macro avg       0.93      0.92      0.92      5268
            weighted avg       0.92      0.92      0.92      5268
```

Modeling

Model 2 -Logistic Regression:- We applied Decision Tree Classifier to our training and testing data and achieved an accuracy score of 77%. Despite tuning the hyperparameters, we were unable to improve the score. Unfortunately, we did not achieve satisfactory precision scores, recall, and f1-scores for all topics. The result is displayed in figure.

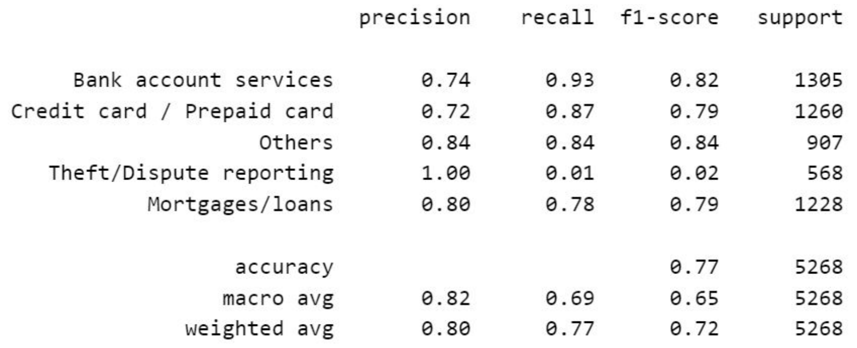


```
                          precision    recall  f1-score   support

    Bank account services       0.83      0.85      0.84      1305
  Credit card / Prepaid card       0.84      0.88      0.86      1260
                   Others       0.75      0.88      0.81       907
  Theft/Dispute reporting       0.86      0.80      0.83       568
          Mortgages/loans       0.84      0.70      0.77      1228

                 accuracy                           0.82      5268
                macro avg       0.82      0.82      0.82      5268
             weighted avg       0.82      0.82      0.82      5268
```

Random Forest Classifier:- We achieved an 77.02% accuracy score using logistic regression on our training and testing data. Despite our best efforts in tuning the hyperparameters, we were unable to improve the score. Unfortunately, we encountered challenges in achieving satisfactory precision scores, recall, and f1-scores for all topics. A visual representation of our results can be found in figure.



```
                          precision    recall  f1-score   support

    Bank account services       0.74      0.93      0.82      1305
  Credit card / Prepaid card       0.72      0.87      0.79      1260
                   Others       0.84      0.84      0.84       907
  Theft/Dispute reporting       1.00      0.01      0.02       568
          Mortgages/loans       0.80      0.78      0.79      1228

                 accuracy                           0.77      5268
                macro avg       0.82      0.69      0.65      5268
             weighted avg       0.80      0.77      0.72      5268
```

Concluding the modeling process:

After thorough testing, the 'LogisticRegression' model emerged as the clear winner among the three models evaluated. With an impressive accuracy score of 92% and superior precision, recall, and f1-score across all topics, it outperformed both the 'DecisionTreeClassifier' and 'RandomForestClassifier'. This confirms the 'LogisticRegression' model's effectiveness and suitability for the dataset. Hence, we conclude that 'LogisticRegression' is the optimal choice for modeling the given dataset.

Thank you

Manish

Email:-pmanish790@yahoo.in