CAPSTONE PROJECT – REPORT AUTOMATIC TICKET CLASSIFICATION - USING NLP

Presented by:

MADHU NUTHULA on 20/11/2023

STUDENT AT INSTITUTE OF DATA 2022-2023

Graduate Certificate in Data Science & Artificial Intelligence

Table of Contents:

S.No	Content	Page No
1	Problem Statement	2
2	Importance	2
3	Benefits of customer complaints	3
4	Project Key Process	4
5	About Dataset	4
6	Exploratory Data analysis	5
7	Text Processing	6,7
8	Clustering the complaint Categories	8, 9
9	Word Cloud Visuals per Category	10
10	Test Train Split	11
11	Vectorization	11,12
12	Scale and apply dimensionality reduction:	12
13	Modelling & Evaluation	13
14	Confusion Matrix -Best Model	14
15	Conclusion	14
16	Additional project work references	15
17	Libraries or Packages used in this project	15
18	References	16

Problem statement:

Cluster the complaints received by a financial institution in to 5 categories. Create a model that can classify the tickets automatically which will help route the complaint to appropriate department and resolve it faster.

Importance:

Complaints are vital part of the business to be addressed in timely manner. Most customers do not complain, they just leave the company and joins the competitor. It is an opportunity for a business to know the drawbacks and address them.

When a customer complains he/she is already disappointed and needs to be addressed sooner before it is too late. Otherwise it can lead to further damage; the disappointed customer spreads the bad experience in the community which will impact brand goodwill to the company and eventually leave the company. If the complaint is handled effectively, the customer will be rather happy and understands that the company values his complaint then in turn sharers the positive experience to others. This will improve the brand image and attracts new customers.

For Financial Institutions, a late response can lead to major financial loss. If the complaint is regarding non-compliance or any other serious matters. This can be taken to the regulatory bodies (organizations that enforce the financial institutions) and they can penalize huge compensation. This could be avoided by early detection of the complaint category and routing it to the right department where the responsible person allocated in the department can understand and relate to the issue and seriousness of it.

So, If the complaint is allocated to the right department, then there is high chance to solve the complaint immediately. Thus, we can:

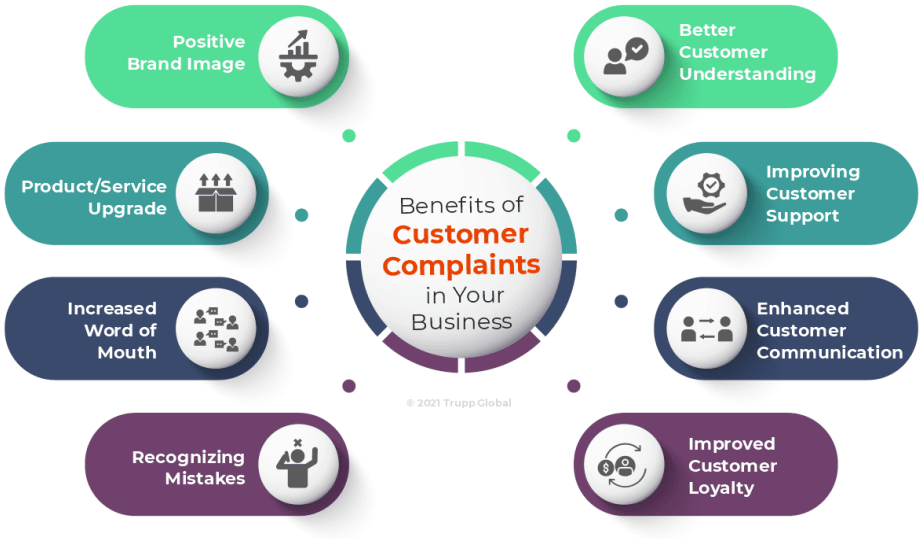
- Prevent a dissatisfied customer from churning
- Reveal customer pain points

- Build a positive brand reputation
- Avoid monetary loss in compensations
- Improve customer loyalty towards the brand

Every business should admit if there is **“NO CUSTOMER - NO BUSINESS”**

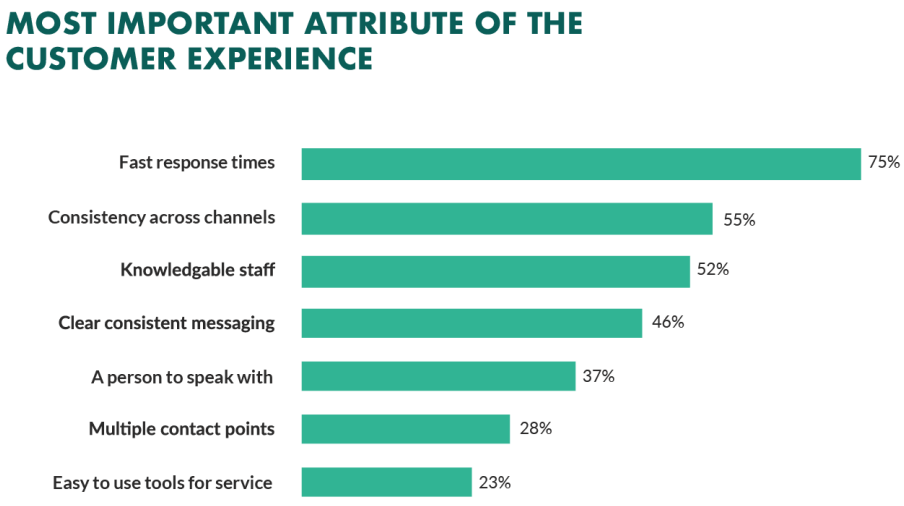
Benefits of Customer Complaints:

Below image shows the benefits of customer complaints:



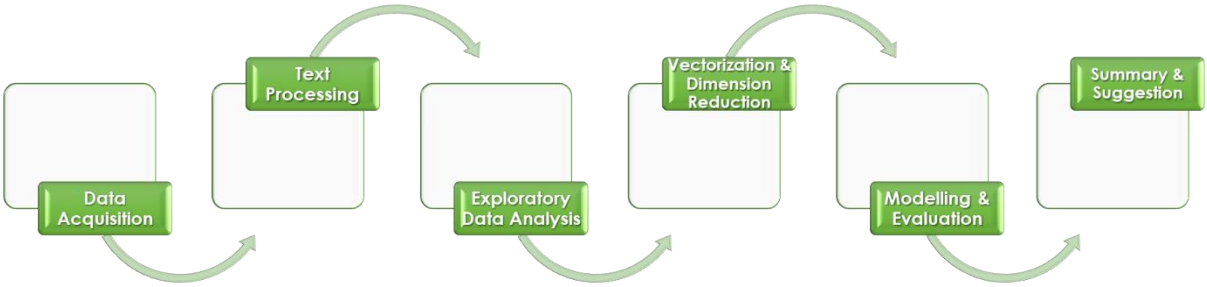
As per Super Office CRM, a fast response makes your customers feel important. Speed matters in customer service because your customers demand it!

CMO council found that the most important attribute of a good customer experience, according to the customers' themselves, is a fast response time. Below is



This makes it very obvious to have a solution in place to resolve the complaints faster to progress further in business, win customer loyalty, beat the competitors and lead the market.

Project Key Process:



About the Dataset:

The Dataset is sourced from Kaggle, which contains the complaints record of a financial institution from The Consumer Financial Protection Bureau (CFPB) a federal U.S. agency that acts as a mediator when disputes arise between financial institutions and consumers. Via a web form, consumers can send the agency a narrative of their dispute.

- Extracted required information from the JSON file, only extracted below mentioned 5 features as they are the relevant attributes to address the business problem.
- Initial extracted dataset was in shape (78313 rows, 5 columns)
- Columns:
  - Complaints - Actual complaint text
  - Product - Complaint Category chosen by customer
  - Sub_product - Complaint sub-category chosen by customer
  - Issue - Issue about the complaint written by customer
  - Sub_issue - Sub issue about the complaint written by customer

Combined columns “Complaints”, “Issue” and “sub_issue” as column “Complaints” as these three columns were written information by the customer about the complaint.

Combined the columns “Product”, “sub_products” to single column as “Category”

Dataset Source: https://www.kaggle.com/datasets/abhishek14398/automatic-ticket-classification-dataset

Exploratory Data analysis:

Once the required data is gathered we have performed some initial data analysis to understand the Dataset thoroughly. Below are some charts of EDA highlights noticed:

Screenshot of the Data frame:

	complaints	product	sub_product	sub_issue	issue	category
0	, Debt is not yours, Attempts to collect debt...	Debt collection	Credit card debt	Debt is not yours	Attempts to collect debt not owed	Debt collection, Credit card debt
1	Good morning my name is XXXX XXXX and I apprec...	Debt collection	Credit card debt	Didn't receive enough information to verify debt	Written notification about debt	Debt collection, Credit card debt
2	I upgraded my XXXX XXXX card in XX/XX/2018 and...	Credit card or prepaid card	General-purpose credit card or charge card	Problem with rewards from credit card	Other features, terms, or problems	Credit card or prepaid card, General-purpose c...
3	, , Trouble during payment process	Mortgage	Conventional home mortgage		Trouble during payment process	Mortgage, Conventional home mortgage
4	, Charged too much interest, Fees or interest	Credit card or prepaid card	General-purpose credit card or charge card	Charged too much interest	Fees or interest	Credit card or prepaid card, General-purpose c...
5	, Problem using a debit or ATM card, Managing...	Checking or savings account	Checking account	Problem using a debit or ATM card	Managing an account	Checking or savings account, Checking account

Screenshot of the Data frame: (after combining the columns)

	complaints	category
0	, Debt is not yours, Attempts to collect debt...	Debt collection, Credit card debt
1	Good morning my name is XXXX XXXX and I apprec...	Debt collection, Credit card debt
2	I upgraded my XXXX XXXX card in XX/XX/2018 and...	Credit card or prepaid card, General-purpose c...
3	, , Trouble during payment process	Mortgage, Conventional home mortgage
4	, Charged too much interest, Fees or interest	Credit card or prepaid card, General-purpose c...
5	, Problem using a debit or ATM card, Managing...	Checking or savings account, Checking account

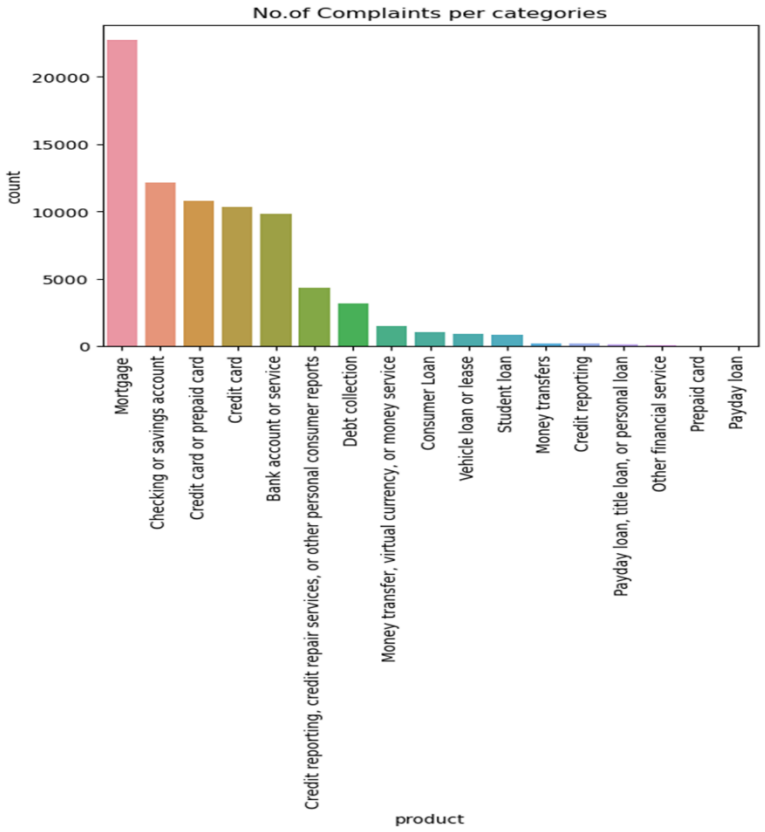
Dataset Shape: (78313, 5)

Category count: 17

Sub_Category count: 73

Issue count: 154

Sub_issue count: 206



Text Processing:

Text processing is the automated process of analysing and sorting unstructured text data to gain valuable insights. This allows machine learning models to get structured information about the text to use for analysis, manipulation of the text, or to generate new text.

Text normalization includes:

- converting all letters to lower or upper case
- converting numbers into words or removing numbers
- removing punctuations, accent marks and other diacritics
- removing white spaces
- expanding abbreviations
- removing stop words, sparse terms, and particular words
- text canonicalization - converting data that involves more than one representation into a standard approved format.

Data Preprocessing:

- Tokenization — convert sentences to words
- Removing unnecessary punctuation, tags
- Removing stop words — frequent words such as “the”, “is”, etc. that do not have specific semantic
- Stemming — words are reduced to a root by removing inflection through dropping unnecessary characters, usually a suffix.
- Lemmatization — Another approach to remove inflection by determining the part of speech and utilizing detailed database of the language.

In this project we used Spacy Library for processing text, created functions to get sentence count and word count for each complaint and added them as new features in to the Data Frame.

EDA highlights after text processing:

complaint with Maximum sentences: 265 sentences (screenshot of part of the complaint below)

```
14476    -- -- -- -- -- Forwarded message -- -- -- -- - From : XXXX XXXX XXXX Date : Wed, XX/XX/2019 at XXXX XXXX Subject : F
wd : Follow-up : XXXX XXXX Police Report Filing on Saturday XX/XX/2019 by XXXX XXXX Regarding Fraud To : XXXX Begin forwarded
message XXXX From : XXXX XXXX XXXX Date : XX/XX/2019 at XXXX XXXX PDT To : XXXX Subject : Fwd : Follow-up : XXXX XXXX Police
Report Filing on Saturday XX/XX/2019 by XXXX XXXX Regarding Fraud Hi there, I am writing to provide a written record of my ex
perience with XXXX. I will not sign a hold harmless letter at this time as it is my understanding that XXXX is fully liable f
or enabling their client 's scam behavior. \n\nI am on a call with one of XXXX 's representatives now who has informed me tha
t there is not a local email that I can sent my full report to, and thus I am emailing this email. \n\nAs a result, XXXX has
on file a full record of what transpired and this can be referenced by XXXX representatives moving forward. \n\nPlease let me
know what questions you have. \n\nBest, XXXX -- -- -- -- -- Forwarded message -- -- -- -- - From : XXXX XXXX XXXX Date : Sat,
XX/XX/2019 at XXXX XXXX Subject : Follow-up : XXXX XXXX Police Report Filing on Saturday XX/XX/2019 by XXXX XXXX Regarding Fr
aud To : XXXX Hi Officer XXXX, Thank you for taking your time to speak with me today and file a police report regarding the f
raud I experienced on XX/XX/2019. Please find enclosed a detailed account of what occurred and what I have provided to my ban
k ( JP Morgan Chase ) in order to move forward with a Fraud claim I have filed with JP Morgan Chase. \n\n-- Start Secure Mess
age To JP Morgan Chase -- I write to file a report in writing for a scam that was committed. \n\nOn XX/XX/2019, I reached out
to a property manager, XXXX XXXX, regarding a XXXX posting. XXXX 's email is XXXX and his phone number is XXXX. \n\nOn XX/XX/
2019, I completed a full rental application and was accepted. I was told to send {$3300.00} to the Landlord, XXXX XXXX, as ou
tlined in the lease agreement. This amount was to cover the first month of rent plus the security deposit. XXXX 's email is X
XXX. I do not know XXXX 's phone number. \n\nOn XX/XX/2019, I sent the requested {$3300.00} to : Account Name : XXXX XXXX Acc
ount Number : XXXX Routing Number : XXXX Swift Cose : XXXX Beneficiary Address : XXXX XXXX XXXX XXXX XXXX, CA XXXX. \n\nMore
```

Complaint with Maximum words: 5585 words (screenshot of part of the complaint below)

```
31952    Alleged Account # XXXX FINAL RESPONSE TO JPMORGAN CHASE BANK NA CO MORTGAGE DEPT Reference number XXXX, XXXX : NOTIC
E - MOST IMPORTANT - The documents for this alleged Mortgage Assignment is nothing but forgeries and robo signing at best. NO
ACTUAL FINDINGS - Within the request they the bank JPMorgan Chase NA Mortgage Co was requested to provide proof of said signa
tures that of the original signatures not forgeries. The signatures are proven not to be that of XXXX XXXX XXXX nor that of X
XXX XXXX XXXX the real persons in flesh. Whom do not and has never given any permission to JPMorgan Chase, nor XXXX Mortgage
XXXX XXXX XXXX XXXX Funding. WRONG IDENTITY - JPMorgan Chase Bank has referred to XXXX XXXX XXXX & XXXX XXXX XXXX as XXXX XXX
X XXXX and XXXX XXXX XXXX, as well as XXXX XXXX or as in XXXX XXXX using the same address within this claim. This is complete
ly and utterly identity theft, fraud no contract. There are no forms presented of an actual ledgers, copy of a check for paym
ents signed by the real persons nor an accurate date or timeline of events for this loan to have taken place. What has been p
resented to all seen in this complaint is nothing best of forgeries, identity fraud, fraud no contract, and fraud in the incl
usion! The attempts of communications with XXXX XXXX XXXX and XXXX XXXX XXXX are therefore cut off, and has never truly exist
ed with the Plaintiffs/real party ( s ) in interest on in the acts of harassment. The document dated XXXX XXXX, XXXX are all
forgeries as stated multiple times to JPMorgan Chase Bank to its employees, and agents as in CEOs XXXX, CFO and Representativ
es. Their response is internally created by former employees of XXXX Mortgage XXXX XXXX XXXX XXXX Funding that of XXXX XXXX a
nd XXXX XXXX XXXX whom may or many not be real persons. Proof of the existence of an account of the actual establishment of d
ebt account but the actual Sentient human XXXX XXXX XXXX and that of XXXX XXXX XXXX duly signed and written out by both parti
es and not any unilateral agreement. This would include but not limited to the actual agreement upon which the signature page
has direct reference to the entire agreement XXXX XXXX XXXX and XXXX XXXX XXXX is an artificial entity, a title, of the limit
ed liability fictitious corporation which is legal trade mark, which constitutes valuable legal interest of which all right,
```

Text Cleaning:

Used Regular Expressions to remove following from the complaints text

- reduce multiple spaces and newlines to only one
- remove double quotes
- remove
- remove urls
- remove characters"x" (which is used to hide confidential information)

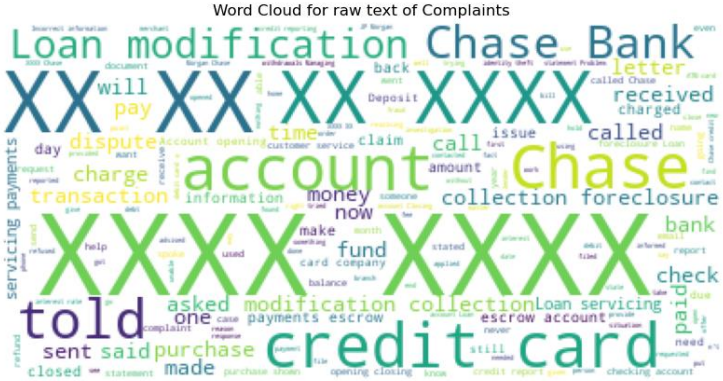
Used Spacy to tokenize the sentences, remove punctuations, remove stop words, lemmatize and join the lemmatized words together. Applied this to all the complaints and amended a new column in to the dataframe as “lemma_comp”.

Below is the screenshot of the dataframe after cleaning: (column “lemma_comp” is after cleaning)

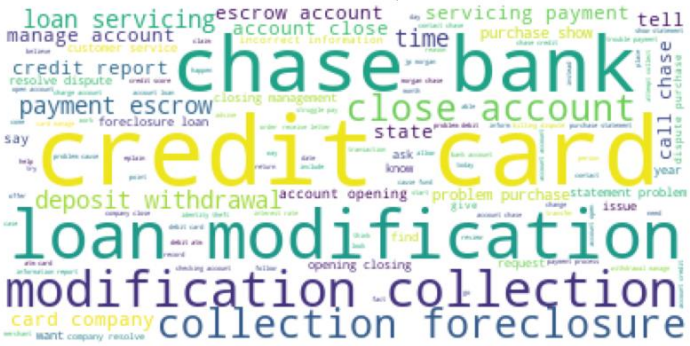
	complaints	category	sentence_length	word_count	lemma_comp
0	, Debt is not yours, Attempts to collect debt not owed	Debt collection, Credit card debt	1	11	debt attempt collect debt owe
1	Good morning my name is XXXX XXXX and I appreciate it if you could help me put a stop to Chase Bank cardmember services. In 2018 I wrote to Chase asking for debt verification and what they sent me a statement which is not acceptable. I am asking the bank to validate the debt. Instead I been receiving mail every month from them attempting to collect a debt. I have a right to know this information as a consumer. Chase account # XXXX XXXX XXXX XXXX Thanks in advance for your help., Didn't receive enough information to verify debt, Written notification about debt	Debt collection, Credit card debt	7	103	good morning appreciate help stop chase bank cardmember service write chase ask debt verification send statement acceptable ask bank validate debt instead receive mail month attempt collect debt right know information consumer chase account thank advance help receive information verify debt write notification debt
2	I upgraded my XXXX XXXX card in XX/XX/2018 and was told by the agent who did the upgrade my anniversary date would not change. It turned the agent was giving me the wrong information in order to upgrade the account. XXXX changed my anniversary date from XX/XX/XXXX to XX/XX/XXXX without my consent! XXXX has the recording of the agent who was misled me., Problem with rewards from credit card, Other features, terms, or problems	Credit card or prepaid card, General-purpose credit card or charge card	4	74	upgrade card /2018 tell agent upgrade anniversary date change turn agent give wrong information order upgrade account change anniversary date consent recording agent mislead problem reward credit card feature term problem
3	, , Trouble during payment process	Mortgage, Conventional home mortgage	1	6	trouble payment process
4	, Charged too much interest, Fees or interest	Credit card or prepaid card, General-purpose credit card or charge card	1	8	charge interest fee interest

Word Cloud Visuals for all the complaint text: (before cleaning & after cleaning)

Word Cloud for raw text of Complaints



Word Cloud for Complaints



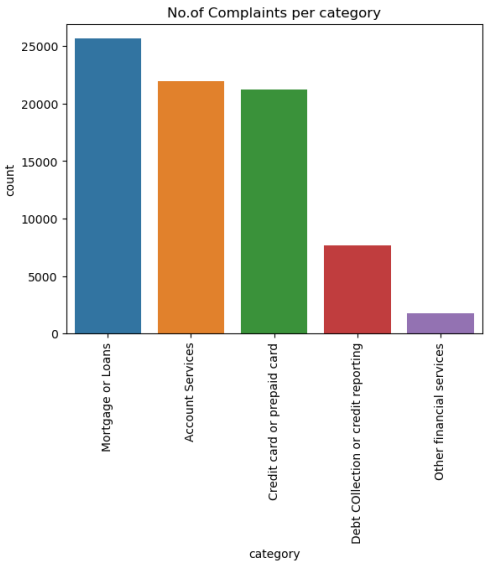
Clustering the complaint Categories:

Categorizing various categories in to 5 categories. Below is the list of categories and the categories merged in to them.

Mortgage or Loans: Encoded as: 1	Mortgage Consumer Loan Vehicle loan or lease Student loan Payday loan, title loan, or personal loan Payday loan
Credit card or prepaid card: Encoded as: 2	Credit card or prepaid card Credit card Prepaid card
Account Services: Encoded as: 3	Checking or savings account Bank account or service
Debt Collection or credit reporting: Encoded as: 4	Debt collection Credit reporting, credit repair services, or other personal consumer reports Credit reporting
Other financial services: Encoded as: 5	Money transfer, virtual currency, or money service Money transfers Other financial service

After Clustering the complaints below is the complaints count for each category.

Mortgage or Loans	25657
Account Services	21963
Credit card or prepaid card	21202
Debt Collection or credit reporting	7703
Other financial services	1788



At this point now we have our Target which is Category and feature which is cleaned and lemmatized text.

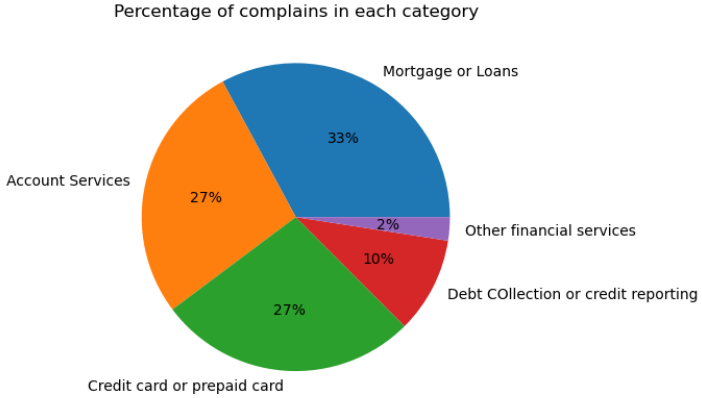
Feature = "lemma_comp" (cleaned and lemmatized complaint text)

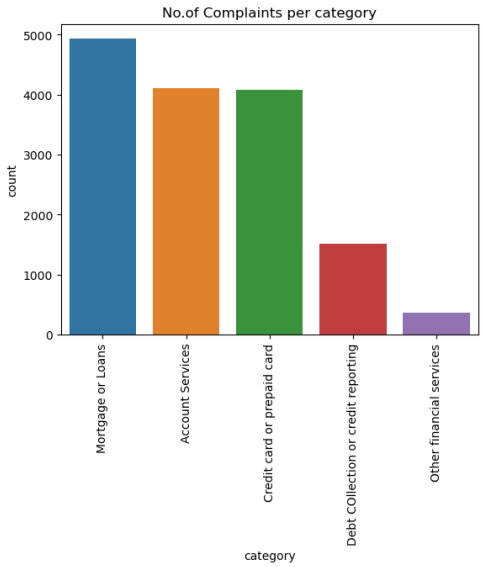
Target = "Category" (which is to be converted in to numerical format)

Numerical representation of catecories:	
Mortgage or Loans	1
Credit card or prepaid card	2
Account Services	3
Debt Collection or credit reporting	4
Other financial services	5

We have gone ahead to perform vectorization and modelling on the full dataset. But the computer resources were not adequate to process it. (16 GB RAM was not enough)

We have randomly sampled the data of 15000 records to help suit the computational requirement. The sampled dataset turned out to be an exact resemblance of the real dataset. Below are the charts to see the proportion of complaints in each category.





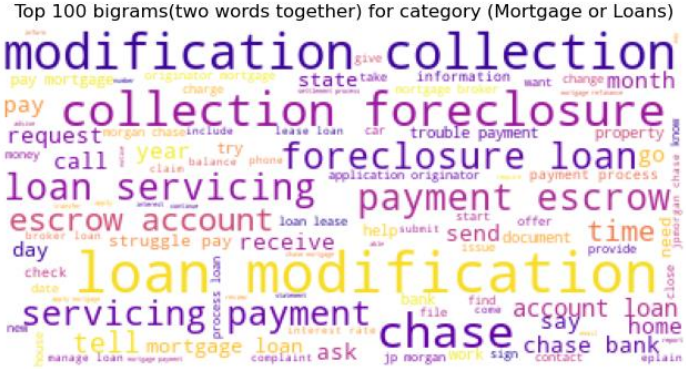
Word Cloud Visuals per Category:

To better understand if the clustering of categories is appropriate, we have performed some word cloud visuals to see the 100 most occurring bigrams (two words together).

After observing them it is clear that the clustering has been done appropriately. There are some irrelevant key words that do not match the categories but we need to accept the fact that it is the nature of text to vary based on person, issue and longer complaints that address other departments within the same complaint.

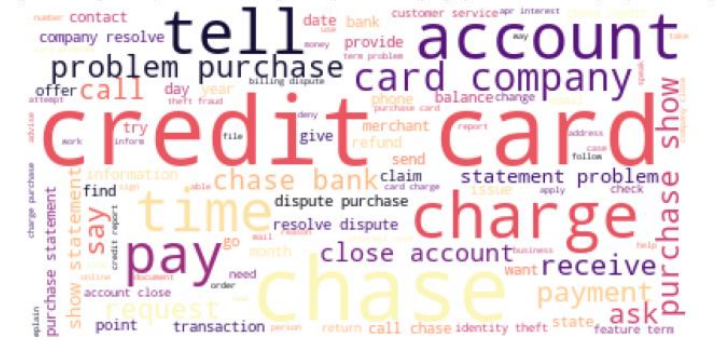
Here are the Word cloud visuals for each category:

Top 100 bigrams(two words together) for category (Mortgage or Loans)

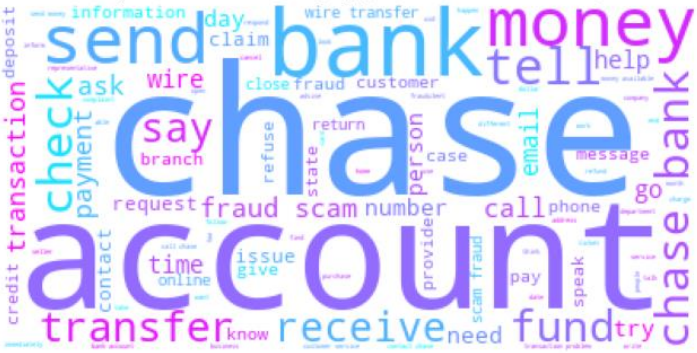


Top 100 bigrams(two words together) for category (Account Services)

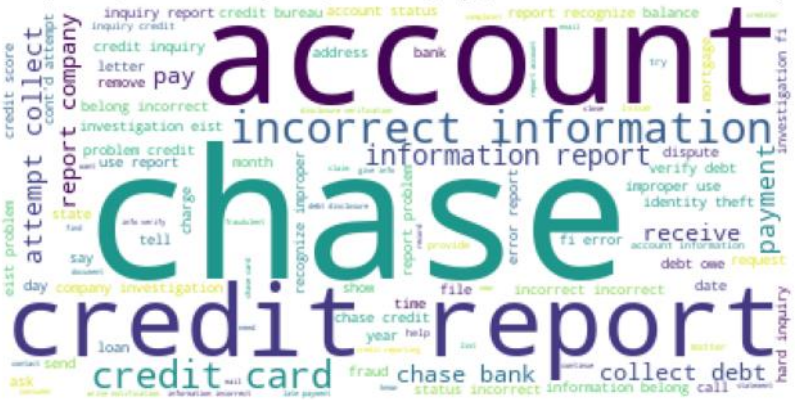
Top 100 bigrams(two words together) for category (Credit card or prepaid card)



Top 100 words for category (Other financial services)

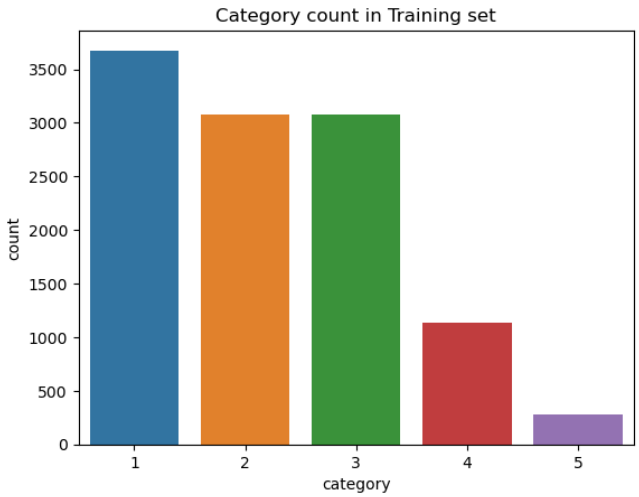


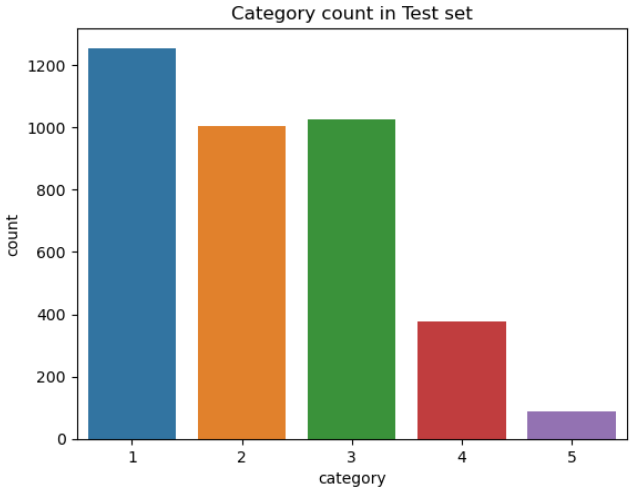
Top 100 bigrams(two words together) for category (Debt Collection or credit reporting)



Test Train Split:

Have split the data in to train and test sets with 80% and 20% proportion respectively. Training sets will be used to train the model and Test set will be used evaluate the trained model.





Vectorization:

Now is the real fun part, we need to convert the text in to numerical format (in to vectors). Each vector represents a word. We have chosen Count Vectorizer and tfidf Transformer to perform this.

What is Count Vectorizer: As per Scikit-Learn documentation it converts a collection of text documents to a matrix of token counts. This implementation produces a sparse representation of the counts using scipy.sparse.csr_matrix.

If you do not provide an a-priori dictionary and you do not use an analyzer that does some kind of feature selection then the number of features will be equal to the vocabulary size found by analyzing the data.

What is tf-idf Transformer: As per Scikit-Learn documentation it transform a count matrix to a normalized tf or tf-idf representation.

Tf means term-frequency while tf-idf means term-frequency times inverse document-frequency. This is a common term weighting scheme in information retrieval, that has also found good use in document classification.

The goal of using tf-idf instead of the raw frequencies of occurrence of a token in a given document is to scale down the impact of tokens that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus.

Now that we understand what count vectorizer does and what it-idf transformer does, we can relate to the implementation of these both techniques together. In our understanding this is the best way to identify the unique words of complaint that relate to the category. This will help machine learning models to understand the importance of word vectors in each category.

Procedure:

1) Fit the X_train data to count vectorizer
2) Use the count vectorizer(from 1st step) to Transform X_train
3) Fit the TF-IDF Transformer with (transformed X_train from 2nd step)
4) Use TF-IDF Transformer from step3 to transform the (transformed X_train from 2nd step)
5) Convert the transformed (X_train from 4th step) to an array
6) With X_test Transform using count vectorizer in step1 and use tfidf-transformer from step 3 to transform.

After vectorizing the shape of **X_train is (11250, 9494) and X_test is (3750, 9494).** That is very high number of dimensions for the machine learning algorithm to process.

Scale and apply dimensionality reduction:

Before applying we standardize the data using Standard Scaler. Standardization makes all variables contribute equally.

StandardScaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. StandardScaler can be influenced by outliers (if they exist in the dataset) since it involves the estimation of the empirical mean and standard deviation of each feature.

For dimensionality reduction we used PCA (Principal Component Analysis) from sklearn decomposition library. **Principal component analysis, or PCA,** is a statistical procedure that allows us to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analysed.

Results: After applying PCA we are able to reduce 9494 dimensions to 1781 dimensions with 90% explained variance (we only missed 10% of data) but this will help the models perform faster.

Modelling & Evaluation:

We have tried below models to train and evaluate against our train and test sets. Below are the summary of results:

Model	Results (Accuracy)
Logistic Regression	0.89
Support Vector Classifier	0.86
Decision Tree Classifier	0.82
AdaBoost Classifier	0.81
Gradient Boosting Classifier	0.83
Random Forest Classifier	0.85

Have chosen the models with best accuracy i.e. Logistic Regression and Random Forest Classifier for hyper parameter tuning. (Support Vector Classifier was performing very poorly on minority class, so avoided it)

Model (Hyper Parameter Tuned)	Results (Accuracy)
Logistic Regression	90%
Random Forest	86%

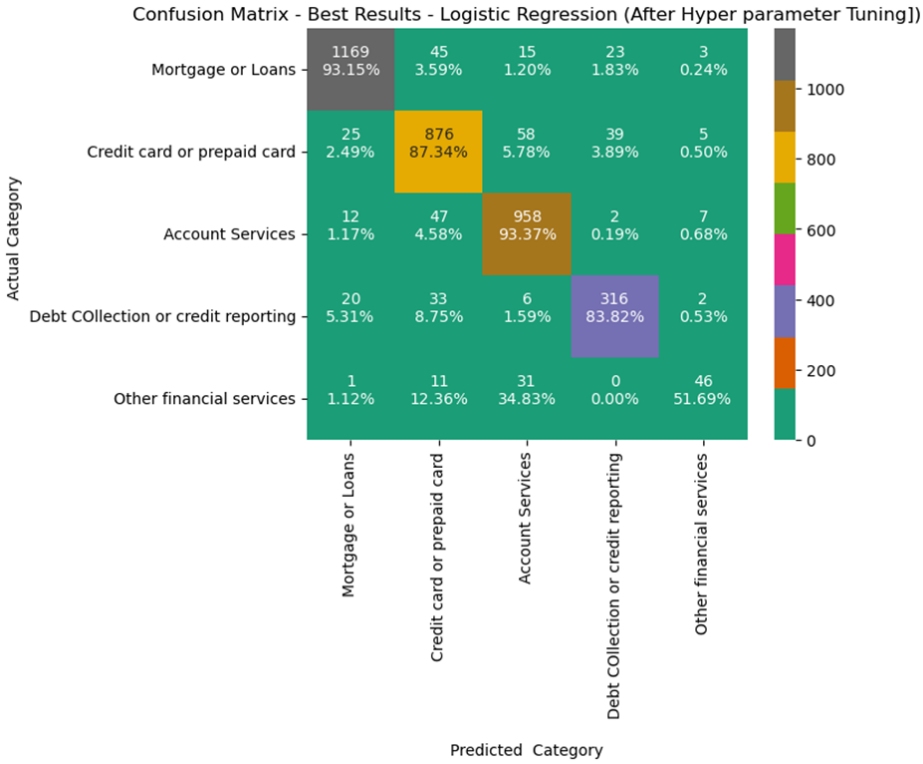
Considering the high level of accuracy and simplicity of the model we chose Logistic Regression as the best model for this project.

Best Model
Logistic Regression with Accuracy of 90%

Confusion Matrix -Best Model:

Below is the confusion matrix for Logistic Reggression:

Category “Other Financial Services” has 51% accuracy; due to the less no.of complaints in it. Can be improved by acquiring more data in for that category.



Conclusion:

Summary:

With the help of proposed model now we can categorize the complaints with 90% accuracy and this can be routed through the appropriate departments.
Financial Institutions can benefit:

- Faster complaint resolution
- Improved customer satisfaction
- Prevent financial loss
- Identify areas to improve
- Win the competitors and lead the Market.

Suggestion:

- Adequate computing resources for advanced modelling to gain better results
- Can be implemented in to any business sector
- Unsupervised learning using Topic modelling

Finally, we would say our model will help address the complaints faster and by the correct department with most accuracy which will improve customer satisfaction.

Ultimately,

Customer Satisfaction = Business Growth

If you value your Business, you must value your CUSTOMERS

Additional project work references:

To get deeper into the coding part please refer to the Jupyter notebooks presented;

Notebook Name	**Summary**
Automatic_Ticket_Classification_Capstone_Project	Data extraction, EDA, Data cleaning, Text Processing
EDA2 - Capstone	Extensive EDA on dataset with reduced records
Modelling_Capstone_3	Test-Train split, Vectorization, Scalling, dimension Reduction, Training & evaluating Machine Learning models, Conclusion

Also refer to presentation slides – **"Automatic Ticket Classification - Using NLP – Presentation"**

Libraries or Packages used in this project:

Below is the code used to import libraries (this will be helpful to understand where the pacages are imported from)

Note: pip install is required to import some libraries, this is not shown in the notebook as they have been already installed on the local computer.

```
import numpy as np
import pandas as pd
import json
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import nltk
import spacy
from spacy.tokenizer import Tokenizer
import en_core_web_md

from sklearn.model_selection import train_test_split
import re
from spacy import displacy
import string
from collections import Counter
from wordcloud import WordCloud
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.pipeline import Pipeline

from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import average_precision_score
from sklearn.metrics import roc_auc_score
from sklearn.metrics import precision_recall_curve
from sklearn.metrics import roc_curve
from sklearn.metrics import auc
```

```
from sklearn import metrics
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from sklearn.utils import shuffle
import warnings
warnings.filterwarnings('ignore')
```

References:

We thank the users from Kaggle who previous worked on this dataset, there code was an inspiration for us to progress further in the project. Below are the reference links:

Automatic Ticket Classification Dataset | Kaggle

consumer-finance-complaints · Datasets at Hugging Face

https://www.kaggle.com/code/vikram92/multiclass-complaints-classification-using-bi-lstm

https://www.kaggle.com/code/abhishek14398/automatic-ticket-classification-case-study

https://www2.deloitte.com/content/dam/Deloitte/us/Documents/regulatory/us-aers-the-power-of-complaints-042115.pdf

https://www.zendesk.com/blog/customer-complaints-10-tips-manage-better/

https://www.consumerfinance.gov/