AUTOMATIC TICKET CLASSIFICATION CASE STUDY

Customer Complaint Classification System

Problem Statement

In the dynamic landscape of customer support, efficient handling of customer complaints is crucial for maintaining customer satisfaction and loyalty. Our company has amassed a substantial amount of customer complaint data in JSON format. Unfortunately, this data is currently unlabelled, making it challenging to promptly route tickets to the appropriate departments for resolution. To address this issue, we propose the implementation of a robust classification system using Natural Language Processing (NLP) techniques.

Objective :The goal is to create a model capable of automatically categorizing customer complaints into distinct clusters based on the products or services mentioned in the tickets. This classification will allow for streamlined ticket management, ensuring that each issue is directed to the relevant department for swift and effective resolution.

classify tickets into the following five clusters based on their products/services:

Credit card / Prepaid card

Bank account services

Theft/Dispute reporting

Mortgages/loans

Others

In [1]:

```
# Importing required library
import json
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

# Importing libraries for text preprocessing and analysis
import re, nltk, spacy, string
nlp = spacy.load('en_core_web_sm')
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, TfidfTran
from sklearn.decomposition import NMF
from nltk.stem import WordNetLemmatizer
from textblob import TextBlob
from wordcloud import WordCloud, STOPWORDS
```

```
# Importing libraries for model evaluation metrics
from sklearn.model_selection import train_test_split, GridSearchCV, RandomizedSearchCV
from sklearn.metrics import f1_score, confusion_matrix, classification_report

# Libraries for avoiding warnings
import warnings
warnings.filterwarnings('ignore')

# Row/Column display limit
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', None)
```

```
C:\Users\asus\anaconda3\lib\site-packages\scipy\__init__.py:146: UserWarning: A NumPy ve
rsion >=1.16.5 and <1.23.0 is required for this version of SciPy (detected version 1.26.
2
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

The libraries are imported for basic numerical computation, data manipulation, data visualization, text preprocessing and analysis and model evaluation. The library spacy is used for production purpose. Known for speed and efficiency. Key features are tokenization, POS tag, named entity recogniton, lemmatization. NMF is non negative matrix factorization used for dimensionality reduction for topic modelling and feature extraction. Textblob is NLP lib built on top of NLTK provides cosistent API for NLP tasks and easy to use for text processing and analysis. Wordcloud is a data visualization tech. that displays most frequent words in the text. Generates visually appearing wordclouds, excludes stopwords from display.

In [2]:

```
# Loading the data
f = open('C:\\Users\\asus\\Desktop\\Automatic Ticket Classification\\Dataset\\complaint
data = json.load(f)
df = pd.json_normalize(data)
df.head()
```

Out[2]:

	_index	_type	_id	_score	_source.tags	_source.zip_code	_source.complaint_id	_source.
0	complaint-public-v2	complaint	3211475	0.0	None	90301	3211475	Attem collect not
1	complaint-public-v2	complaint	3229299	0.0	Servicemember	319XX	3229299	W notifi about

	_index	_type	_id	_score	_source.tags	_source.zip_code	_source.complaint_id	_source.
2	complaint-public-v2	complaint	3199379	0.0	None	77069	3199379	fea ter prol
3	complaint-public-v2	complaint	2673060	0.0	None	48066	2673060	Tr c pay pr
4	complaint-public-v2	complaint	3203545	0.0	None	10473	3203545	Fe in

The above code loads the json data and convert it to pandas dataframe. The opened json data is parsed to python data structures. The json normalize flattens the data such that nested dictionaries are expanded into seperate tables. The resulting data from json.load is dictionary/list or combination.

In [3]:
```
# Checking the shape of data
df.shape
```

Out[3]:
```
(78313, 22)
```

No of rows: 78313

No of columns: 22

In [4]:
```
# Checking the info about datatypes
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78313 entries, 0 to 78312
Data columns (total 22 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   _index                        78313 non-null  object 
 1   _type                         78313 non-null  object 
 2   _id                           78313 non-null  object 
 3   _score                        78313 non-null  float64
```

```
 4   _source.tags                    10900 non-null  object
 5   _source.zip_code                71556 non-null  object
 6   _source.complaint_id            78313 non-null  object
 7   _source.issue                   78313 non-null  object
 8   _source.date_received           78313 non-null  object
 9   _source.state                   76322 non-null  object
 10  _source.consumer_disputed       78313 non-null  object
 11  _source.product                 78313 non-null  object
 12  _source.company_response        78313 non-null  object
 13  _source.company                 78313 non-null  object
 14  _source.submitted_via           78313 non-null  object
 15  _source.date_sent_to_company    78313 non-null  object
 16  _source.company_public_response  4 non-null      object
 17  _source.sub_product             67742 non-null  object
 18  _source.timely                  78313 non-null  object
 19  _source.complaint_what_happened  78313 non-null  object
 20  _source.sub_issue               32016 non-null  object
 21  _source.consumer_consent_provided  77305 non-null  object
dtypes: float64(1), object(21)
memory usage: 13.1+ MB
```

In [5]:

```
# Checking the statistical distribution of data
df.describe().T
```

Out[5]:

	count	mean	std	min	25%	50%	75%	max
_score	78313.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

In [6]:

```
# Checking the null values present in data
(df.isnull().sum()/df.shape[0])*100
```

Out[6]:

```
_index                          0.000000
_type                           0.000000
_id                             0.000000
_score                          0.000000
_source.tags                   86.081493
_source.zip_code                8.628197
_source.complaint_id            0.000000
_source.issue                   0.000000
_source.date_received           0.000000
_source.state                   2.542362
_source.consumer_disputed       0.000000
_source.product                 0.000000
_source.company_response        0.000000
_source.company                 0.000000
_source.submitted_via           0.000000
_source.date_sent_to_company    0.000000
_source.company_public_response 99.994892
_source.sub_product            13.498397
_source.timely                  0.000000
_source.complaint_what_happened 0.000000
_source.sub_issue              59.117899
_source.consumer_consent_provided  1.287143
dtype: float64
```

In [7]:

```
df.head(20)
```

Out[7]:

	_index	_type	_id	_score	_source.tags	_source.zip_code	_source.complaint_id	_sourc
0	complaint-public-v2	complaint	3211475	0.0	None	90301	3211475	Atter colle no
1	complaint-public-v2	complaint	3229299	0.0	Servicemember	319XX	3229299	notif abou
2	complaint-public-v2	complaint	3199379	0.0	None	77069	3199379	fe te pr
3	complaint-public-v2	complaint	2673060	0.0	None	48066	2673060	pa p
4	complaint-public-v2	complaint	3203545	0.0	None	10473	3203545	i
5	complaint-public-v2	complaint	3275312	0.0	Older American	48227	3275312	Manag a
6	complaint-public-v2	complaint	3238804	0.0	None	76262	3238804	Manag a
7	complaint-public-v2	complaint	3249272	0.0	None	07753	3249272	pa p
8	complaint-public-v2	complaint	3351653	0.0	None	60621	3351653	Clos a
9	complaint-public-v2	complaint	3273612	0.0	None	99354	3273612	Manag a
10	complaint-public-v2	complaint	3233499	0.0	None	104XX	3233499	In infor

	_index	_type	_id	_score	_source.tags	_source.zip_code	_source.complaint_id	_sourc
								c
11	complaint-public-v2	complaint	3180294	0.0	None	750XX	3180294	In infor c
12	complaint-public-v2	complaint	3362102	0.0	None	08701	3362102	Clo a
13	complaint-public-v2	complaint	3319930	0.0	None	420XX	3319930	In infor c

	_index	_type	_id	_score	_source.tags	_source.zip_code	_source.complaint_id	_sourc
14	complaint-public-v2	complaint	3224980	0.0	None	920XX	3224980	Manag a
15	complaint-public-v2	complaint	3209411	0.0	None	None	3209411	Impro
16	complaint-public-v2	complaint	3311133	0.0	None	78748	3311133	Manag a
17	complaint-public-v2	complaint	3331023	0.0	None	770XX	3331023	Clos a

	_index	_type	_id	_score	_source.tags	_source.zip_code	_source.complaint_id	_sourc
18	complaint-public-v2	complaint	2647668	0.0	None	47331	2647668	Strugg mc
19	complaint-public-v2	complaint	3300211	0.0	None	32796	3300211	\ notif abou

The column _source.complaint_what_happened has multiple blanck values. Need to replace these empty spaces with nan.

In [8]:
```
# Replacing empty spaces of _source.complaint_what_happened column with nan
df['_source.complaint_what_happened'].replace('', np.nan, inplace=True)

# Checking the null count again
(df.isnull().sum()/df.shape[0])*100
```

Out[8]:
```
_index                          0.000000
_type                           0.000000
_id                             0.000000
_score                          0.000000
```

```
_source.tags                       86.081493
_source.zip_code                    8.628197
_source.complaint_id                0.000000
_source.issue                       0.000000
_source.date_received               0.000000
_source.state                       2.542362
_source.consumer_disputed           0.000000
_source.product                     0.000000
_source.company_response            0.000000
_source.company                     0.000000
_source.submitted_via               0.000000
_source.date_sent_to_company        0.000000
_source.company_public_response    99.994892
_source.sub_product                13.498397
_source.timely                      0.000000
_source.complaint_what_happened    73.092590
_source.sub_issue                  59.117899
_source.consumer_consent_provided   1.287143
dtype: float64
```

In [9]:

```
# Dropping nulls from '_source.complaint_what_happened' column
df.dropna(subset=['_source.complaint_what_happened'], inplace=True)

# Checking the shape of data
df.shape
```

Out[9]:

```
(21072, 22)
```

In [10]:

```
# Removing _ and source from columns
df.columns = [re.sub(r'^_', '', col) for col in df.columns]
df.columns = [re.sub(r'^source\.', '', col) for col in df.columns]

list(df.columns)
```

Out[10]:

```
['index',
 'type',
 'id',
 'score',
 'tags',
 'zip_code',
 'complaint_id',
 'issue',
 'date_received',
 'state',
 'consumer_disputed',
 'product',
 'company_response',
 'company',
 'submitted_via',
 'date_sent_to_company',
 'company_public_response',
 'sub_product',
 'timely',
 'complaint_what_happened',
 'sub_issue',
 'consumer_consent_provided']
```

In [11]:
```
# Define a function that gives text input and returns cleaned text
def cleaned_text(text):
    text = text.lower()  # Making lowercase chr
    text = re.sub(r'\[.*?\]', '', text)  # Remove shortest possible text enclosed in sq
    text = re.sub(r'[%s]'%re.escape(string.punctuation), '', text)  # Remove punctuatio
    text = re.sub(r'\w*\d\w*', '', text)  # Remove alphanumeric chrs
    return text
```

In [12]:
```
# Lets clean the text from '_source.complaint_what_happened' column
df['complaint_what_happened'] = df['complaint_what_happened'].apply(lambda x: cleaned_t
```

In [13]:
```
# Define a function that returns lemmatized text
def lemmatize_text(text):
    lemma_list = []
    document = nlp(text)
    for word in document:     # Extract lemmas (base word) for text and add it to list
        lemma_list.append(word.lemma_)
    return ' '.join(lemma_list)
```

In [14]:
```
# Apply the above function to complaint_what_happened column and add new column to df '
df['lemmatized_complaints'] = df.apply(lambda x: lemmatize_text(x['complaint_what_happe

df.head()
```

Out[14]:

	index	type	id	score	tags	zip_code	complaint_id	issue	date_recei
1	complaint-public-v2	complaint	3229299	0.0	Servicemember	319XX	3229299	Written notification about debt	2019-01T12:00 0!
2	complaint-public-v2	complaint	3199379	0.0	None	77069	3199379	Other features, terms, or problems	2019-02T12:00 0!

	index	type	id	score	tags	zip_code	complaint_id	issue	date_recei
10	complaint-public-v2	complaint	3233499	0.0	None	104XX	3233499	Incorrect information on your report	2019-06T12:00 0
11	complaint-public-v2	complaint	3180294	0.0	None	750XX	3180294	Incorrect information on your report	2019-14T12:00 0

	index	type	id	score	tags	zip_code	complaint_id	issue	date_recei
14	complaint-public-v2	complaint	3224980	0.0	None	920XX	3224980	Managing an account	2019-27T12:00 0!

In [15]:
```
# Create a new df 'cleaned_df' which have only 2 columns complaint_what_happened and le
cleaned_df = df[['complaint_what_happened', 'lemmatized_complaints']]
cleaned_df.head()
```

Out[15]:

	complaint_what_happened	lemmatized_complaints
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 \nin 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 \ni have a right to know this information as a consumer \n\nchase	good morning my name be xxxx xxxx and I appreciate it if you could help I put a stop to chase bank cardmember service \n in I write to chase ask for debt verification and what they send I a statement which be not acceptable I be ask the bank to validate the debt instead I been receive mail every month from they attempt to collect a debt \n I have a right to know this information as a

	complaint_what_happened	lemmatized_complaints
	account xxxx xxxx xxxx xxxx thanks in advance for your help	consumer \n\n chase account xxxx xxxx xxxx xxxx thank in advance for your help
2	i upgraded my xxxx xxxx card in 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 xxxxxxxx to xxxxxxxx without my consent xxxx has the recording of the agent who was misled me	I upgrade my xxxx xxxx card in and be tell by the agent who do the upgrade my anniversary date would not change it turn the agent be give I the wrong information in order to upgrade the account xxxx change my anniversary date from xxxxxxxx to xxxxxxxx without my consent xxxx have the recording of the agent who be mislead I
10	chase card was reported on however fraudulent application have been submitted my identity without my consent to fraudulently obtain services do not extend credit without verifying the identity of the applicant	chase card be report on however fraudulent application have be submit my identity without my consent to fraudulently obtain service do not extend credit without verify the identity of the applicant
11	on while trying to book a xxxx xxxx ticket i came across an offer for to be applied towards the ticket if i applied for a rewards card i put in my information for the offer and within less than a minute was notified via the screen that a decision could not be made i immediately contacted xxxx and was referred to chase bank i then immediately contacted chase bank within no more than of getting the notification on the screen and i was told by the chase representative i spoke with that my application was denied but she could not state why i asked for more information about the xxxx offer and she explained that even if i had been approved the credit offer only gets applied after the first account statement and could not be used to purchase the ticket i then explicitly told her i was glad i got denied and i was absolutely no longer interested in the account i asked that the application be withdrawn and the representative obliged this all happened no later than after putting in the application on notwithstanding my explicit request not to proceed with the application and contrary to what i was told by the chase representative chase did in fact go ahead to open a credit account in my name on this is now being reported in my credit report and chase has refused to correct this information on my credit report even though they went ahead to process an application which i did not consent to and out of their error	on while try to book a xxxx xxxx ticket I come across an offer for to be apply towards the ticket if I apply for a reward card I put in my information for the offer and within less than a minute be notify via the screen that a decision could not be make I immediately contact xxxx and be refer to chase bank I then immediately contact chase bank within no more than of get the notification on the screen and I be tell by the chase representative I speak with that my application be deny but she could not state why I ask for more information about the xxxx offer and she explain that even if I have be approve the credit offer only get apply after the first account statement and could not be use to purchase the ticket I then explicitly tell she I be glad I get deny and I be absolutely no long interested in the account I ask that the application be withdraw and the representative oblige this all happen no later than after put in the application on notwithstanding my explicit request not to proceed with the application and contrary to what I be tell by the chase representative chase do in fact go ahead to open a credit account in my name on this be now be report in my credit report and chase have refuse to correct this information on my credit report even though they go ahead to process an application which I do not consent to and out of their error
14	my grand son give me check for i deposit it into my chase account after fund clear my chase bank closed my account never paid me my money they said they need to speek with my grand son check was clear money was taking by my chase bank refuse to pay me my money my grand son called chase times they told him i should call not him to verify the check owner he is out the country most the time date happen check number xxxx claim number is xxxx with chase	my grand son give I check for I deposit it into my chase account after fund clear my chase bank close my account never pay I my money they say they need to speek with my grand son check be clear money be take by my chase bank refuse to pay I my money my grand son call chase time they tell he I should call not he to verify the check owner he be out the country most the time date happen check number xxxx claim number be xxxx with chase

In [16]:
```
# Extracting POS tag
def singular_noun(text):
```

```
    text_blob = TextBlob(text)
    return ' '.join([word for (word,tag) in text_blob.tags if tag=='NN'])
```

import nltk nltk.download('averaged_perceptron_tagger')

The avg perceptron tagger is a POS tagger used for assigning POS tags to each word of sentence. It uses perceptron ML algorithm to predict POS tags for words. It provides info about gramatical structure of sentence.

import nltk nltk.download('punkt')

The punkt is pretrained unsupervised ML model for sentence tokenize.

In [85]:
```
# Adding new column 'complaints_POS_removed' to cleaned_df after removing POS tag from
cleaned_df['complaints_POS_removed'] = df.apply(lambda x: singular_noun(x['lemmatized_c

cleaned_df.head()
```

Out[85]:

	complaint_what_happened	lemmatized_complaints	complaints_POS_removed	complaints_clean
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 \nin 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 \ni have a right to know this information as a consumer \n\nchase account xxxx xxxx xxxx xxxx thanks in advance for your help	good morning my name be xxxx xxxx and I appreciate it if you could help I put a stop to chase bank cardmember service \n in I write to chase ask for debt verification and what they send I a statement which be not acceptable I be ask the bank to validate the debt instead I been receive mail every month from they attempt to collect a debt \n I have a right to know this information as a consumer \n\n chase account xxxx xxxx xxxx xxxx thank in advance for your help	morning name stop bank cardmember service ask debt verification statement bank debt mail month debt right information consumer chase account thank advance help	morning name stop bank cardmember service ask debt verification statement bank debt mail month debt right information consumer chase account thank advance help
2	i upgraded my xxxx xxxx card in 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 xxxxxxxx to xxxxxxxx without my consent xxxx has the recording of the agent who was misled me	I upgrade my xxxx xxxx card in and be tell by the agent who do the upgrade my anniversary date would not change it turn the agent be give I the wrong information in order to upgrade the account xxxx change my anniversary date from xxxxxxxx to xxxxxxxx without my consent xxxx have the recording of the agent who be mislead I	card agent upgrade date agent information order account change date xxxxxxxx consent xxxx recording agent	card agent upgrade date agent information order account change date xxxxxxxx consent xxxx recording agent

	complaint_what_happened	lemmatized_complaints	complaints_POS_removed	complaints_clean
10	chase card was reported on however fraudulent application have been submitted my identity without my consent to fraudulently obtain services do not extend credit without verifying the identity of the applicant	chase card be report on however fraudulent application have be submit my identity without my consent to fraudulently obtain service do not extend credit without verify the identity of the applicant	card report application identity consent service credit identity applicant	card report application identity consent service credit identity applicant
11	on while trying to book a xxxx xxxx ticket i came across an offer for to be applied towards the ticket if i applied for a rewards card i put in my information for the offer and within less than a minute was notified via the screen that a decision could not be made i immediately contacted xxxx and was referred to chase bank i then immediately contacted chase bank within no more than of getting the notification on the screen and i was told by the chase representative i spoke with that my application was denied but she could not state why i asked for more information about the xxxx offer and she explained that even if i had been approved the credit offer only gets applied after the first account statement and could not be used to purchase the ticket i then explicitly told her i was glad i got denied and i was absolutely no longer interested in the account i asked that the application be withdrawn and the representative obliged this all happened no later than after putting in the application on notwithstanding my explicit request not to proceed with the application and contrary to what i was told by the chase representative chase did in fact go ahead to open a credit account in my name on this is now being reported in my credit report and chase has refused to correct this information on my credit report even though they went ahead to process an	on while try to book a xxxx xxxx ticket I come across an offer for to be apply towards the ticket if I apply for a reward card I put in my information for the offer and within less than a minute be notify via the screen that a decision could not be make I immediately contact xxxx and be refer to chase bank I then immediately contact chase bank within no more than of get the notification on the screen and I be tell by the chase representative I speak with that my application be deny but she could not state why I ask for more information about the xxxx offer and she explain that even if I have be approve the credit offer only get apply after the first account statement and could not be use to purchase the ticket I then explicitly tell she I be glad I get deny and I be absolutely no long interested in the account I ask that the application be withdraw and the representative oblige this all happen no later than after put in the application on notwithstanding my explicit request not to proceed with the application and contrary to what I be tell by the chase representative chase do in fact go ahead to open a credit account in my name on this be	try book xxxx ticket offer ticket card information offer minute screen decision bank chase bank notification screen chase representative application state information xxxx offer credit offer account statement use ticket account application representative oblige put application explicit request application chase chase fact credit account name report credit report chase information credit report application error	try book xxxx ticket offer ticket card information offer minute screen decision bank chase bank notification screen chase representative application state information xxxx offer credit offer account statement use ticket account application representative oblige put application explicit request application chase chase fact credit account name report credit report chase information credit report application error

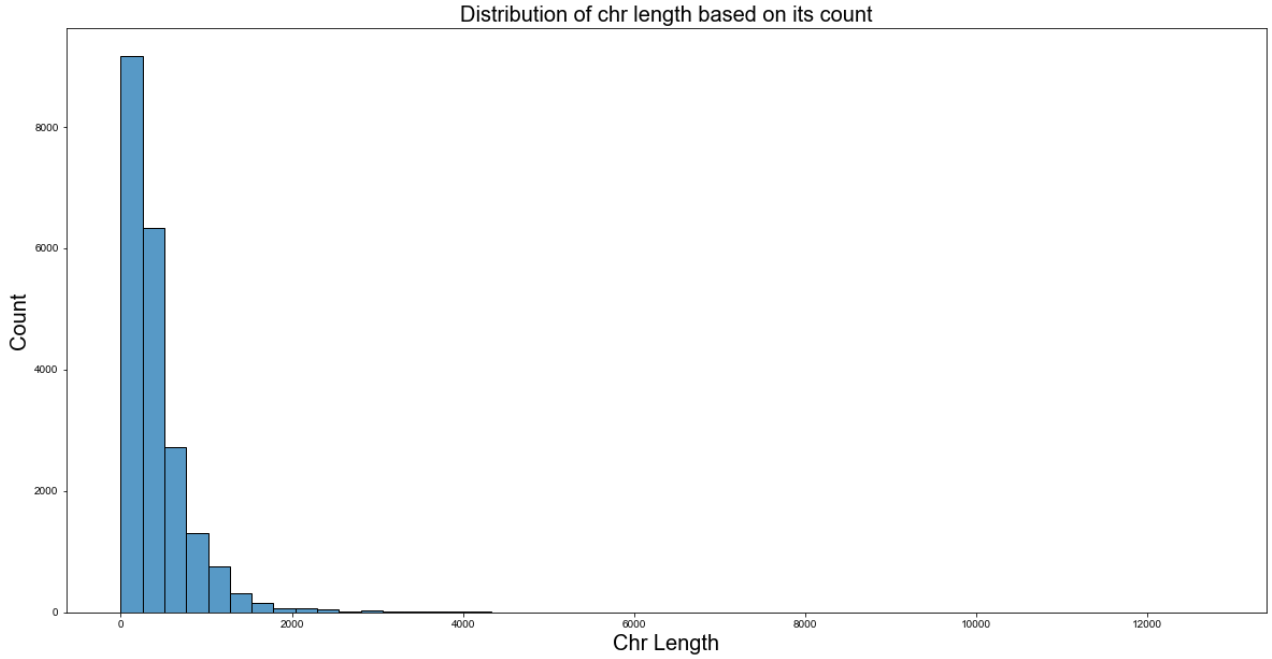
	complaint_what_happened	lemmatized_complaints	complaints_POS_removed	complaints_clean
	application which i did not consent to and out of their error	now be report in my credit report and chase have refuse to correct this information on my credit report even though they go ahead to process an application which I do not consent to and out of their error		
14	my grand son give me check for i deposit it into my chase account after fund clear my chase bank closed my account never paid me my money they said they need to speek with my grand son check was clear money was taking by my chase bank refuse to pay me my money my grand son called chase times they told him i should call not him to verify the check owner he is out the country most the time date happen check number xxxx claim number is xxxx with chase	my grand son give I check for I deposit it into my chase account after fund clear my chase bank close my account never pay I my money they say they need to speek with my grand son check be clear money be take by my chase bank refuse to pay I my money my grand son call chase time they tell he I should call not he to verify the check owner he be out the country most the time date happen check number xxxx claim number be xxxx with chase	son chase account fund bank account pay money son check money bank refuse money son call chase time check owner country time date check number claim number chase	son chase account fund bank account pay money son check money bank refuse money son call chase time check owner country time date check number claim number chase

Exploratory Data Analysis

In [86]:
```
# Visualize the data from complaints_POS_removed column based on character length
chr_len = [len(x) for x in cleaned_df['complaints_POS_removed']]
chr_len[:10]
```

Out[86]:
```
[159, 105, 74, 414, 161, 7, 629, 605, 1186, 51]
```

In [87]:
```
# Plot the histogram based on chr length
plt.figure(figsize=[20,10])
sns.histplot(data=chr_len, bins=50)
plt.title('Distribution of chr length based on its count', fontdict={'fontsize':20})
plt.xlabel('Chr Length', fontdict={'fontsize':20})
plt.ylabel('Count', fontdict={'fontsize':20})
plt.show()
```



In [88]:
```
# Removing pronouns from text corpus complaints_POS_removed
cleaned_df['complaints_clean'] = cleaned_df['complaints_POS_removed'].str.replace('-PRO
```

In [89]:
```
# Find top 30 unigrams along with their frequency in 'complaints_POS_removed' corpus
def top30unigrams(text, n=30):
    vector = CountVectorizer(stop_words='english').fit(text)
    BOW_model = vector.transform(text)
    BOW_model_sum = BOW_model.sum(axis=0)
    word_freq = [(word, BOW_model_sum[0, idx]) for word, idx in vector.vocabulary_.item
    word_freq = sorted(word_freq, key= lambda x: x[1], reverse=True)
    return word_freq[:n]
```

The function top30unigrams learns the vocabulary from input text data and then transform this data into matrix form where each row represents documents in text data and column represents a unique word in vocabulary. stopwords are removed during learning process. The function returns top 30 single words along with their frequency.

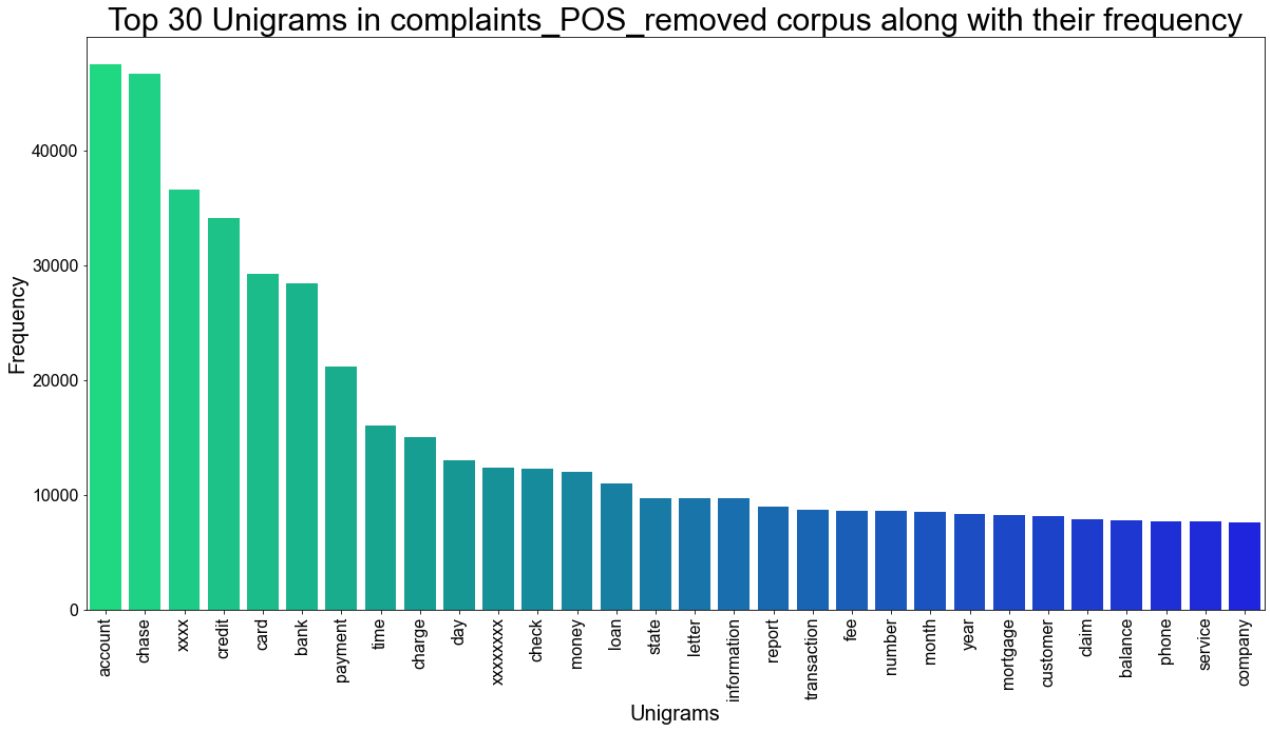
In [90]:
```
# Top 30 unigrams in complaints_POS_removed
top30uni = top30unigrams(cleaned_df['complaints_POS_removed'].values.astype('U'))
top30uni_df = pd.DataFrame(top30uni, columns=['unigrams', 'counts'])
top30uni_df.head()
```

Out[90]:

	unigrams	counts
0	account	47514
1	chase	46699
2	xxxx	36563
3	credit	34148

	unigrams	counts
4	card	29278

In [91]:
```
# Lets plot the barplot of top 30 unigrams
plt.figure(figsize=[20,10])
sns.barplot(x=top30uni_df['unigrams'], y=top30uni_df['counts'], palette='winter_r')
plt.xticks(rotation=90, fontsize=16)
plt.yticks(fontsize=16)
plt.xlabel('Unigrams', fontdict={'fontsize':20})
plt.ylabel('Frequency', fontdict={'fontsize':20})
plt.title('Top 30 Unigrams in complaints_POS_removed corpus along with their frequency'
plt.show()
```



In [92]:
```
# Find top 30 bigrams along with their frequency in 'complaints_POS_removed' corpus
def top30bigrams(text, n=30):
    vector = CountVectorizer(ngram_range=(2,2), stop_words='english').fit(text)
    BOW_model = vector.transform(text)
    BOW_model_sum = BOW_model.sum(axis=0)
    word_freq = [(word, BOW_model_sum[0, idx]) for word, idx in vector.vocabulary_.item
    word_freq = sorted(word_freq, key= lambda x: x[1], reverse=True)
    return word_freq[:n]
```

The count vectorizer is initialized with ngram range as (2,2) means it finds the bigrams from input text data. The first parameter 2 is min range and second one is max.

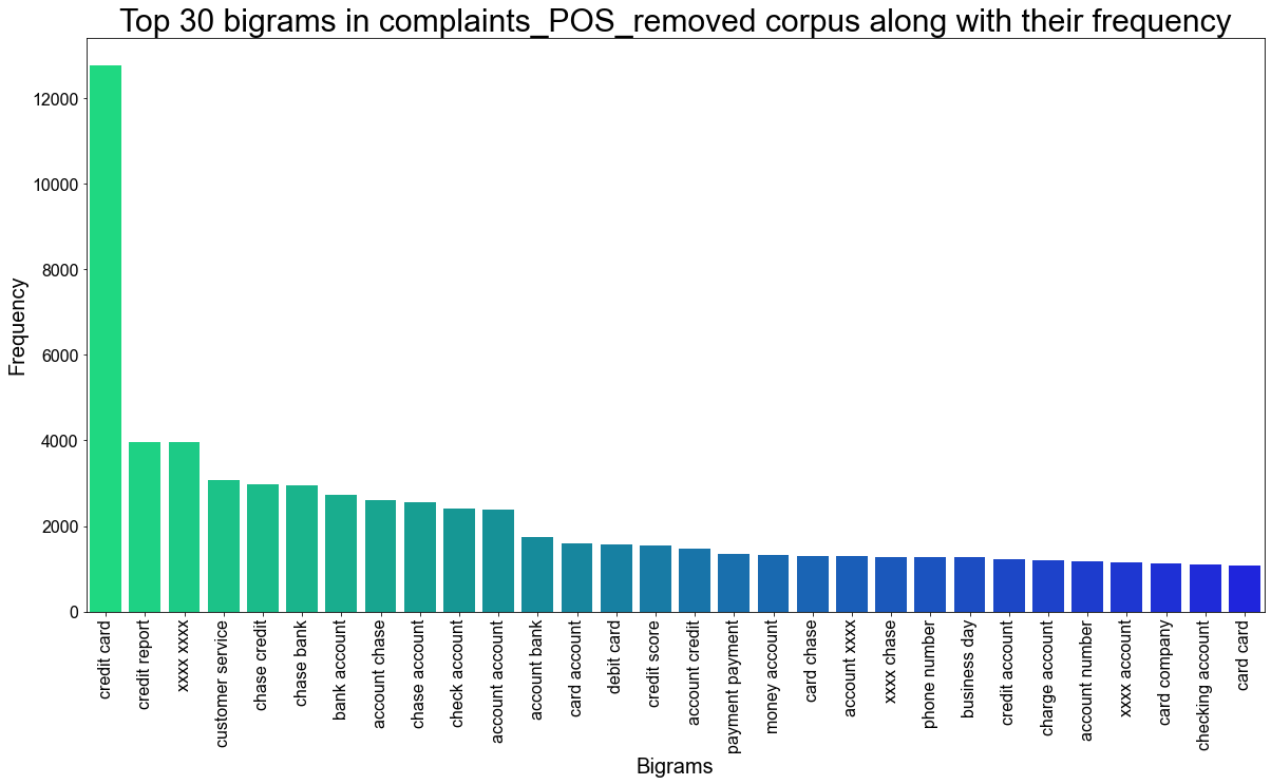
In [93]:
```
# Top 30 bigrams in complaints_POS_removed
top30bi = top30bigrams(cleaned_df['complaints_POS_removed'].values.astype('U'))
top30bi_df = pd.DataFrame(top30bi, columns=['Bigrams', 'counts'])
top30bi_df.head()
```

Out[93]:

	Bigrams	counts
0	credit card	12778
1	credit report	3955
2	xxxx xxxx	3953
3	customer service	3081
4	chase credit	2966

In [94]:

```
# Lets plot the barplot of top 30 bigrams
plt.figure(figsize=[20,10])
sns.barplot(x=top30bi_df['Bigrams'], y=top30bi_df['counts'], palette='winter_r')
plt.xticks(rotation=90, fontsize=16)
plt.yticks(fontsize=16)
plt.xlabel('Bigrams', fontdict={'fontsize':20})
plt.ylabel('Frequency', fontdict={'fontsize':20})
plt.title('Top 30 bigrams in complaints_POS_removed corpus along with their frequency',
plt.show()
```



In [95]:

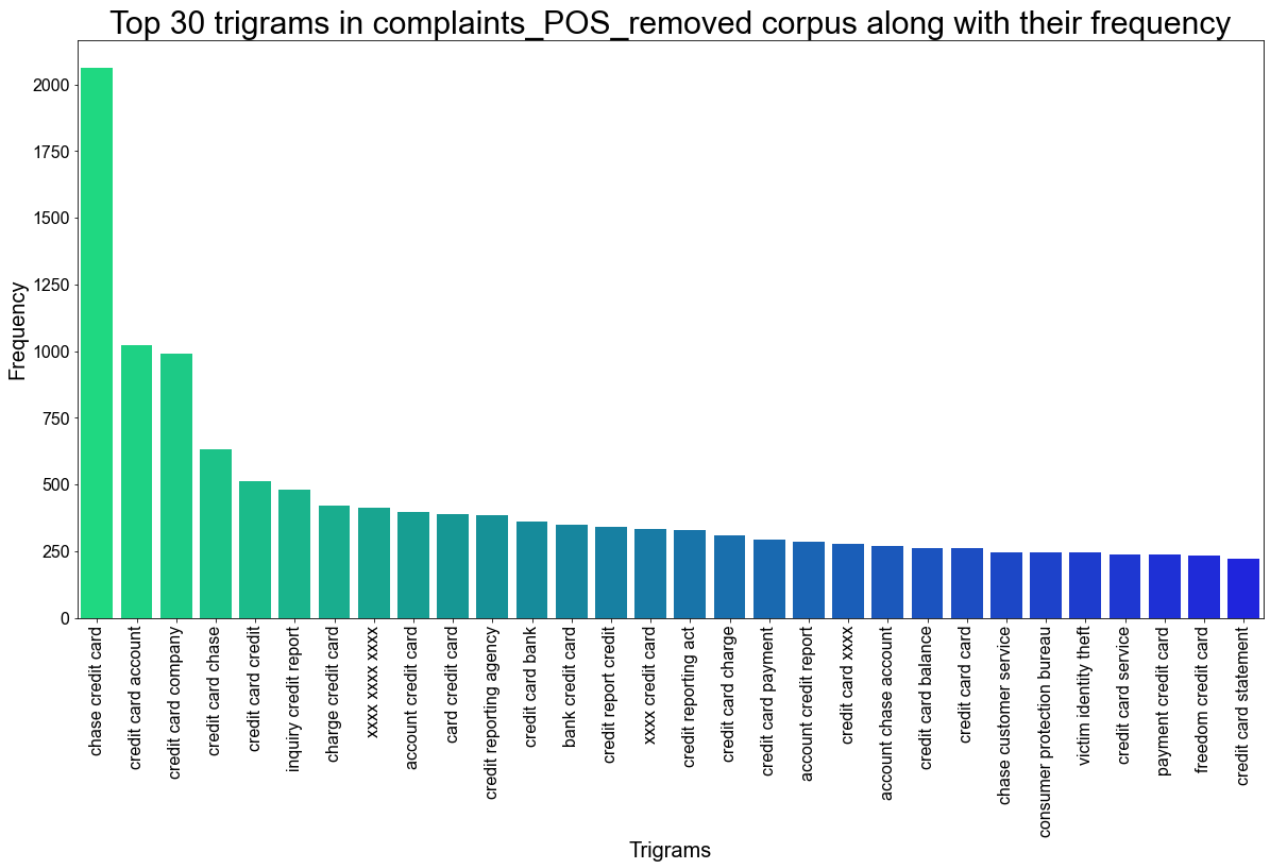
```
# Find top 30 trigrams along with their frequency in 'complaints_POS_removed' corpus
def top30trigrams(text, n=30):
    vector = CountVectorizer(ngram_range=(3,3), stop_words='english').fit(text)
    BOW_model = vector.transform(text)
    BOW_model_sum = BOW_model.sum(axis=0)
    word_freq = [(word, BOW_model_sum[0, idx]) for word, idx in vector.vocabulary_.item
    word_freq = sorted(word_freq, key= lambda x: x[1], reverse=True)
    return word_freq[:n]
```

In [96]:
```
# Top 30 trigrams in complaints_POS_removed
top30tri = top30trigrams(cleaned_df['complaints_POS_removed'].values.astype('U'))
top30tri_df = pd.DataFrame(top30tri, columns=['Trigrams', 'counts'])
top30tri_df.head()
```

Out[96]:

	Trigrams	counts
0	chase credit card	2063
1	credit card account	1022
2	credit card company	991
3	credit card chase	633
4	credit card credit	513

In [97]:
```
# Lets plot the barplot of top 30 trigrams
plt.figure(figsize=[20,10])
sns.barplot(x=top30tri_df['Trigrams'], y=top30tri_df['counts'], palette='winter_r')
plt.xticks(rotation=90, fontsize=16)
plt.yticks(fontsize=16)
plt.xlabel('Trigrams', fontdict={'fontsize':20})
plt.ylabel('Frequency', fontdict={'fontsize':20})
plt.title('Top 30 trigrams in complaints_POS_removed corpus along with their frequency'
plt.show()
```



In [98]:
```
# Remove the personnel info of the customer masked as xxxx in complaints_POS_removed co
cleaned_df['complaints_clean'] = cleaned_df['complaints_clean'].str.replace('xxxx', '')
```

```
cleaned_df.shape
```

Out[98]: (21072, 4)

In [99]:
```
cleaned_df.head()
```

Out[99]:

	complaint_what_happened	lemmatized_complaints	complaints_POS_removed	complaints_clean
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 \nin 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 \ni have a right to know this information as a consumer \n\nchase account xxxx xxxx xxxx xxxx thanks in advance for your help	good morning my name be xxxx xxxx and I appreciate it if you could help I put a stop to chase bank cardmember service \n in I write to chase ask for debt verification and what they send I a statement which be not acceptable I be ask the bank to validate the debt instead I been receive mail every month from they attempt to collect a debt \n I have a right to know this information as a consumer \n\n chase account xxxx xxxx xxxx xxxx thank in advance for your help	morning name stop bank cardmember service ask debt verification statement bank debt mail month debt right information consumer chase account thank advance help	morning name stop bank cardmember service ask debt verification statement bank debt mail month debt right information consumer chase account thank advance help
2	i upgraded my xxxx xxxx card in 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 xxxxxxxx to xxxxxxxx without my consent xxxx has the recording of the agent who was misled me	I upgrade my xxxx xxxx card in and be tell by the agent who do the upgrade my anniversary date would not change it turn the agent be give I the wrong information in order to upgrade the account xxxx change my anniversary date from xxxxxxxx to xxxxxxxx without my consent xxxx have the recording of the agent who be mislead I	card agent upgrade date agent information order account change date xxxxxxxx consent xxxx recording agent	card agent upgrade date agent information order account change date consent recording agent
10	chase card was reported on however fraudulent application have been submitted my identity without my consent to fraudulently obtain services do not extend credit without verifying the identity of the applicant	chase card be report on however fraudulent application have be submit my identity without my consent to fraudulently obtain service do not extend credit without verify the identity of the applicant	card report application identity consent service credit identity applicant	card report application identity consent service credit identity applicant
11	on while trying to book a xxxx xxxx ticket i came across an offer for to be applied towards the ticket if i applied for a rewards card i put in my information for the offer and	on while try to book a xxxx xxxx ticket I come across an offer for to be apply towards the ticket if I apply for a reward card I put in my information for	try book xxxx ticket offer ticket card information offer minute screen decision bank chase bank notification screen chase representative application	try book ticket offer ticket card information offer minute screen decision bank chase bank

	complaint_what_happened	lemmatized_complaints	complaints_POS_removed	complaints_clean
	within less than a minute was notified via the screen that a decision could not be made i immediately contacted xxxx and was referred to chase bank i then immediately contacted chase bank within no more than of getting the notification on the screen and i was told by the chase representative i spoke with that my application was denied but she could not state why i asked for more information about the xxxx offer and she explained that even if i had been approved the credit offer only gets applied after the first account statement and could not be used to purchase the ticket i then explicitly told her i was glad i got denied and i was absolutely no longer interested in the account i asked that the application be withdrawn and the representative obliged this all happened no later than after putting in the application on notwithstanding my explicit request not to proceed with the application and contrary to what i was told by the chase representative chase did in fact go ahead to open a credit account in my name on this is now being reported in my credit report and chase has refused to correct this information on my credit report even though they went ahead to process an application which i did not consent to and out of their error	the offer and within less than a minute be notify via the screen that a decision could not be make I immediately contact xxxx and be refer to chase bank I then immediately contact chase bank within no more than of get the notification on the screen and I be tell by the chase representative I speak with that my application be deny but she could not state why I ask for more information about the xxxx offer and she explain that even if I have be approve the credit offer only get apply after the first account statement and could not be use to purchase the ticket I then explicitly tell she I be glad I get deny and I be absolutely no long interested in the account I ask that the application be withdraw and the representative oblige this all happen no later than after put in the application on notwithstanding my explicit request not to proceed with the application and contrary to what I be tell by the chase representative chase do in fact go ahead to open a credit account in my name on this be now be report in my credit report and chase have refuse to correct this information on my credit report even though they go ahead to process an application which I do not consent to and out of their error	state information xxxx offer credit offer account statement use ticket account application representative oblige put application explicit request application chase chase fact credit account name report credit report chase information credit report application error	notification screen chase representative application state information offer credit offer account statement use ticket account application representative oblige put application explicit request application chase chase fact credit account name report credit report chase information credit report application error
14	my grand son give me check for i deposit it into my chase account after fund clear my chase bank closed my account never paid me my money they said they need	my grand son give I check for I deposit it into my chase account after fund clear my chase bank close my account never pay I my money they say they	son chase account fund bank account pay money son check money bank refuse money son call chase time check owner country time date check	son chase account fund bank account pay money son check money bank refuse money son

	complaint_what_happened	lemmatized_complaints	complaints_POS_removed	complaints_clean
	to speek with my grand son check was clear money was taking by my chase bank refuse to pay me my money my grand son called chase times they told him i should call not him to verify the check owner he is out the country most the time date happen check number xxxx claim number is xxxx with chase	need to speek with my grand son check be clear money be take by my chase bank refuse to pay I my money my grand son call chase time they tell he I should call not he to verify the check owner he be out the country most the time date happen check number xxxx claim number be xxxx with chase	number claim number chase	call chase time check owner country time date check number claim number chase

In [100…

```
# Feature Extraction using TfIdf Vectorizer
tfidf = TfidfVectorizer(max_df=0.95, min_df=2, stop_words='english')
DTM = tfidf.fit_transform(cleaned_df['complaints_POS_removed'])
```

TfIdf model is used for info retrieval and text mining to represent how important the word is to document or whole corpus. The max doc freq is set to 0.95 means the words occur more than 95% in documents ie too common are excluded and the words that are too rare occur less than 2 doc are also excluded. This controls the size and relevance of vocabulary.

In [101…

```
# Initialize the NMF model with no of topics 5
nmf_model = NMF(n_components=5, random_state=42)
nmf_model.fit(DTM)
```

Out[101…

```
NMF(n_components=5, random_state=42)
```

In [105…

```
# Checking the features obtained from tfidf model
len(tfidf.get_feature_names_out())
```

Out[105… 7364

In [107…

```
# Lets find the top 10 words with highest weights from nmf model and sort it in ascendi
first_topic = nmf_model.components_[0]
top10words_idx = first_topic.argsort()[-10:]
for index in top10words_idx:
    print(tfidf.get_feature_names_out()[index])
```

```
day
branch
xxxx
deposit
chase
fund
```

```
money
bank
check
account
```

The above words are the indicative of most imp terms associated with perticular topic.

In [110…

```
# Lets print the top 15 words from each topic
for index, topic in enumerate(nmf_model.components_):
    print(f'Top 15 words from Topic #{index}')
    print([tfidf.get_feature_names_out()[i] for i in topic.argsort()[-15:]])
    print('\n')
```

```
Top 15 words from Topic #0
['customer', 'transfer', 'transaction', 'business', 'number', 'day', 'branch', 'xxxx',
'deposit', 'chase', 'fund', 'money', 'bank', 'check', 'account']


Top 15 words from Topic #1
['letter', 'year', 'balance', 'application', 'debt', 'information', 'limit', 'company',
'score', 'account', 'chase', 'inquiry', 'report', 'card', 'credit']


Top 15 words from Topic #2
['foreclosure', 'house', 'bank', 'document', 'time', 'rate', 'letter', 'year', 'propert
y', 'chase', 'modification', 'home', 'xxxx', 'mortgage', 'loan']


Top 15 words from Topic #3
['refund', 'time', 'service', 'xxxxxxxx', 'purchase', 'fraud', 'claim', 'merchant', 'xxx
x', 'fee', 'dispute', 'chase', 'transaction', 'card', 'charge']


Top 15 words from Topic #4
['chase', 'account', 'credit', 'xxxx', 'pay', 'date', 'auto', 'time', 'xxxxxxxx', 'day',
'statement', 'fee', 'month', 'balance', 'payment']
```

These are the top 15 words present in each topic having highest weights.

In [111…

```
# Create best topic for each of the complaint in terms of int 0,1,2,3,4
best_topic = nmf_model.transform(DTM) # Extracting the topic wts for each complaint
best_topic[0].round(2)
best_topic[0].argmax()  # Dominant topic for first complaint
best_topic.argmax(axis=1) # Domnant topic for all complaints
```

Out[111…

```
array([0, 3, 1, ..., 3, 4, 4], dtype=int64)
```

The DTM is transformed using fitted nmf model resulting a matrix where each row consist of complaints(words) and each column consist of weight of that complaint in each topic to find the best topic. The first complaint of best topic is rounded to 2 decimal

places and finds the index of the topic having highest wts for first complaint. The index of the topic having highest wts of each complaint s then found. It operates over rows and provides an array of ints representing the dominant topic for corrosponding complaints.

In [112…

```
# Assign the best topic for each of the complaint in cleaned df
cleaned_df['Best Topic'] = best_topic.argmax(axis=1)
cleaned_df.head()
```

Out[112…

	complaint_what_happened	lemmatized_complaints	complaints_POS_removed	complaints_clean	Best Topic
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 \nin 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 \ni have a right to know this information as a consumer \n\nchase account xxxx xxxx xxxx xxxx thanks in advance for your help	good morning my name be xxxx xxxx and I appreciate it if you could help I put a stop to chase bank cardmember service \n in I write to chase ask for debt verification and what they send I a statement which be not acceptable I be ask the bank to validate the debt instead I been receive mail every month from they attempt to collect a debt \n I have a right to know this information as a consumer \n\n chase account xxxx xxxx xxxx xxxx thank in advance for your help	morning name stop bank cardmember service ask debt verification statement bank debt mail month debt right information consumer chase account thank advance help	morning name stop bank cardmember service ask debt verification statement bank debt mail month debt right information consumer chase account thank advance help	0
2	i upgraded my xxxx xxxx card in 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 xxxxxxxx to xxxxxxxx without my consent xxxx has the recording of the agent who was misled me	I upgrade my xxxx xxxx card in and be tell by the agent who do the upgrade my anniversary date would not change it turn the agent be give I the wrong information in order to upgrade the account xxxx change my anniversary date from xxxxxxxx to xxxxxxxx without my consent xxxx have the recording of the agent who be mislead I	card agent upgrade date agent information order account change date xxxxxxxx consent xxxx recording agent	card agent upgrade date agent information order account change date consent recording agent	3
10	chase card was reported on however fraudulent application have been submitted my identity without my consent to fraudulently obtain services do not extend credit	chase card be report on however fraudulent application have be submit my identity without my consent to fraudulently obtain service do not extend	card report application identity consent service credit identity applicant	card report application identity consent service credit identity applicant	1

	complaint_what_happened	lemmatized_complaints	complaints_POS_removed	complaints_clean	Best Topic
	without verifying the identity of the applicant	credit without verify the identity of the applicant			
11	on while trying to book a xxxx xxxx ticket i came across an offer for to be applied towards the ticket if i applied for a rewards card i put in my information for the offer and within less than a minute was notified via the screen that a decision could not be made i immediately contacted xxxx and was referred to chase bank i then immediately contacted chase bank within no more than of getting the notification on the screen and i was told by the chase representative i spoke with that my application was denied but she could not state why i asked for more information about the xxxx offer and she explained that even if i had been approved the credit offer only gets applied after the first account statement and could not be used to purchase the ticket i then explicitly told her i was glad i got denied and i was absolutely no longer interested in the account i asked that the application be withdrawn and the representative obliged this all happened no later than after putting in the application on notwithstanding my explicit request not to proceed with the application and contrary to what i was told by the chase representative chase did in fact go ahead to open a credit account in my name on this is now being reported in my credit report and chase has refused to correct this information on my credit report even though they went ahead to process an application	on while try to book a xxxx xxxx ticket I come across an offer for to be apply towards the ticket if I apply for a reward card I put in my information for the offer and within less than a minute be notify via the screen that a decision could not be make I immediately contact xxxx and be refer to chase bank I then immediately contact chase bank within no more than of get the notification on the screen and I be tell by the chase representative I speak with that my application be deny but she could not state why I ask for more information about the xxxx offer and she explain that even if I have be approve the credit offer only get apply after the first account statement and could not be use to purchase the ticket I then explicitly tell she I be glad I get deny and I be absolutely no long interested in the account I ask that the application be withdraw and the representative oblige this all happen no later than after put in the application on notwithstanding my explicit request not to proceed with the application and contrary to what I be tell by the chase representative chase do in fact go ahead to open a credit account in my name on this be now be report in my credit report and chase have refuse to correct this information	try book xxxx ticket offer ticket card information offer minute screen decision bank chase bank notification screen chase representative application state information xxxx offer credit offer account statement use ticket account application representative oblige put application explicit request application chase chase fact credit account name report credit report chase information credit report application error	try book ticket offer ticket card information offer minute screen decision bank chase bank notification screen chase representative application state information offer credit offer account statement use ticket account application representative oblige put application explicit request application chase chase fact credit account name report credit report chase information credit report application error	1

	complaint_what_happened	lemmatized_complaints	complaints_POS_removed	complaints_clean	Best Topic
	which i did not consent to and out of their error	on my credit report even though they go ahead to process an application which I do not consent to and out of their error			
14	my grand son give me check for i deposit it into my chase account after fund clear my chase bank closed my account never paid me my money they said they need to speek with my grand son check was clear money was taking by my chase bank refuse to pay me my money my grand son called chase times they told him i should call not him to verify the check owner he is out the country most the time date happen check number xxxx claim number is xxxx with chase	my grand son give I check for I deposit it into my chase account after fund clear my chase bank close my account never pay I my money they say they need to speek with my grand son check be clear money be take by my chase bank refuse to pay I my money my grand son call chase time they tell he I should call not he to verify the check owner he be out the country most the time date happen check number xxxx claim number be xxxx with chase	son chase account fund bank account pay money son check money bank refuse money son call chase time check owner country time date check number claim number chase	son chase account fund bank account pay money son check money bank refuse money son call chase time check owner country time date check number claim number chase	0

In [114…

```
# Now we will use the index of topics to classify any new complaint. Lets create traini
training_df = cleaned_df[['complaint_what_happened', 'Best Topic']]
training_df.head()
```

Out[114…

	complaint_what_happened	Best Topic
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 \nin 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 \ni have a right to know this information as a consumer \n\nchase account xxxx xxxx xxxx xxxx thanks in advance for your help	0
2	i upgraded my xxxx xxxx card in 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 xxxxxxxx to xxxxxxxx without my consent xxxx has the recording of the agent who was misled me	3
10	chase card was reported on however fraudulent application have been submitted my identity without my consent to fraudulently obtain services do not extend credit without verifying the identity of the applicant	1
11	on while trying to book a xxxx xxxx ticket i came across an offer for to be applied towards the ticket if i applied for a rewards card i put in my information for the offer and within less than a minute was notified via the screen that a decision could not be made i immediately contacted xxxx and was referred to chase bank i then immediately contacted chase bank within no more than of getting the notification on the screen and i was told by the chase representative i spoke with that my application was denied but she could not state why i asked for more information about the xxxx offer and she explained that even if i had been approved the credit offer only gets applied after the first account	1

	complaint_what_happened	Best Topic
	statement and could not be used to purchase the ticket i then explicitly told her i was glad i got denied and i was absolutely no longer interested in the account i asked that the application be withdrawn and the representative obliged this all happened no later than after putting in the application on notwithstanding my explicit request not to proceed with the application and contrary to what i was told by the chase representative chase did in fact go ahead to open a credit account in my name on this is now being reported in my credit report and chase has refused to correct this information on my credit report even though they went ahead to process an application which i did not consent to and out of their error	
14	my grand son give me check for i deposit it into my chase account after fund clear my chase bank closed my account never paid me my money they said they need to speek with my grand son check was clear money was taking by my chase bank refuse to pay me my money my grand son called chase times they told him i should call not him to verify the check owner he is out the country most the time date happen check number xxxx claim number is xxxx with chase	0

In [115…

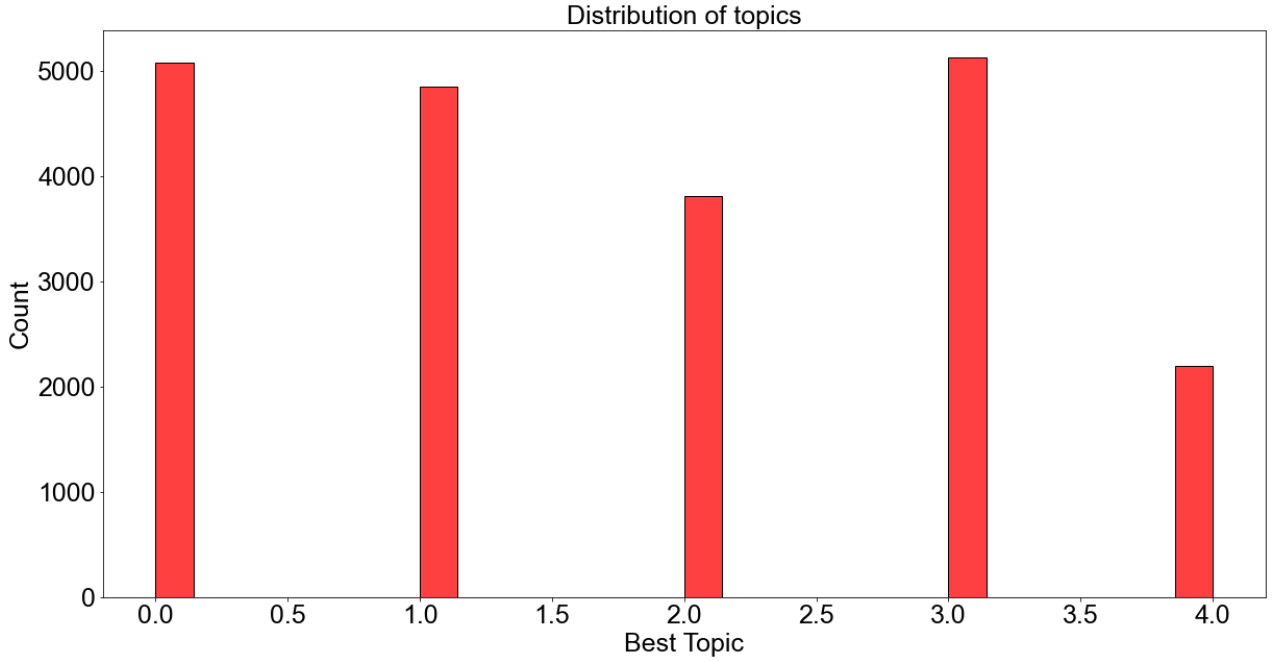
```
# Checking the value counts of topics
training_df['Best Topic'].value_counts()
```

Out[115…

```
3    5128
0    5084
1    4856
2    3807
4    2197
Name: Best Topic, dtype: int64
```

In [118…

```
# Visualize the distribution of topics using histplot
plt.figure(figsize=[20,10])
sns.histplot(data=training_df, x='Best Topic', color='r')
plt.title('Distribution of topics', fontsize=25)
plt.xlabel('Best Topic', fontsize=25)
plt.ylabel('Count', fontsize=25)
plt.xticks(fontsize=25)
plt.yticks(fontsize=25)
plt.show()3
```



Lets initialize the BOW model to get the matrix representation of input text based on token counts. The BOW model considers only the unique word (feature/vocabulary) from all the docs present in text data. It will not consider the order of words in docs and importance of word wrt whole corpus. It gives same importance to all words. Therefore we have to further use tfidf representation to get the frequency info from BOW model with imporatnce info. It takes into account not only frequency of word in the docs but also how unique that word is across all docs. This helps in giving more weights to words that are imp for spacific doc but not too common across entire corpus.

In [119…

```
# BOW model
BOW_model = CountVectorizer()
X_train_BOW = BOW_model.fit_transform(training_df['complaint_what_happened'])
```

In [120…

```
# TFIDF model
tfidf_model = TfidfTransformer()
X_train_tfidf = tfidf_model.fit_transform(X_train_BOW)
```

In [186…

```
# Lets do the train test split
X_train, X_test, y_train, y_test = train_test_split(X_train_tfidf, training_df['Best To

print(f'X_train shape: {X_train.shape}')
print(f'X_test shape: {X_test.shape}')
print(f'y_train shape: {y_train.shape}')
print(f'y_test shape: {y_test.shape}')
```

```
X_train shape: (15804, 33599)
X_test shape: (5268, 33599)
y_train shape: (15804,)
y_test shape: (5268,)
```

In [187…

```
# Lets create the model evaluation function for choosing the best model
def model_evaluation(y_test, y_pred, model_name):
    print(f'Classification Report {model_name}\n')
    print(classification_report(y_test, y_pred, target_names=["Bank Account services",
                                                              "Others", "Theft/Dispute

    plt.figure(figsize=[20,10])
    plt.title(f'Confusion matrix {model_name}\n', fontsize=20)
    plt.xticks(fontsize=20)
    plt.yticks(fontsize=20)
    conf_matrix = confusion_matrix(y_test, y_pred)
    sns.heatmap(conf_matrix, annot=True, cmap='Greens', xticklabels=["Bank Account serv
                                                              "Others", "Theft/Dispute
                yticklabels=["Bank Account services", "Credit card or prepaid card",
                                                              "Others", "Theft/Dispute
                annot_kws={"size": 20}, fmt='d')

    plt.show()
```

The function will print the classification report along with the heatmap of confusion matrix.

1) Naive Bayes

In [188…

```
# Importing the NB alogorithm from sklearn
from sklearn.naive_bayes import MultinomialNB

# Initializing NB algorithm, fitting the training data and predicting on test data
model_name = 'NB'
NB = MultinomialNB()
NB.fit(X_train, y_train)
y_pred = NB.predict(X_test)
```

In [189…

```
# Calculating the f1 score of model using weighted avg method
F1_score_NB = f1_score(y_test, y_pred, average='weighted')
F1_score_NB
```

Out[189…

```
0.6844028768108159
```

The descent f1 score is achieved without hyperparameters. Lets train hyperparameteres using grid search CV.

In [190…

```
# Hyperparameter training
params_NB = {'alpha': (1, 0.1, 0.01, 0.001, 0.0001, 0.00001),
             'fit_prior': [True, False]}

gridCV_NB = GridSearchCV(estimator=NB, param_grid=params_NB, scoring='f1_weighted', ver
gridCV_NB.fit(X_train, y_train)
gridCV_NB.best_params_
```

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
```

Out[190…
```
{'alpha': 0.1, 'fit_prior': False}
```

In [191…
```
# Creating model with best hyperparams
model_name = 'NB'
NB_tuned = MultinomialNB(alpha=0.1, fit_prior=False)
NB_tuned.fit(X_train, y_train)
y_pred_NB = NB_tuned.predict(X_test)
```

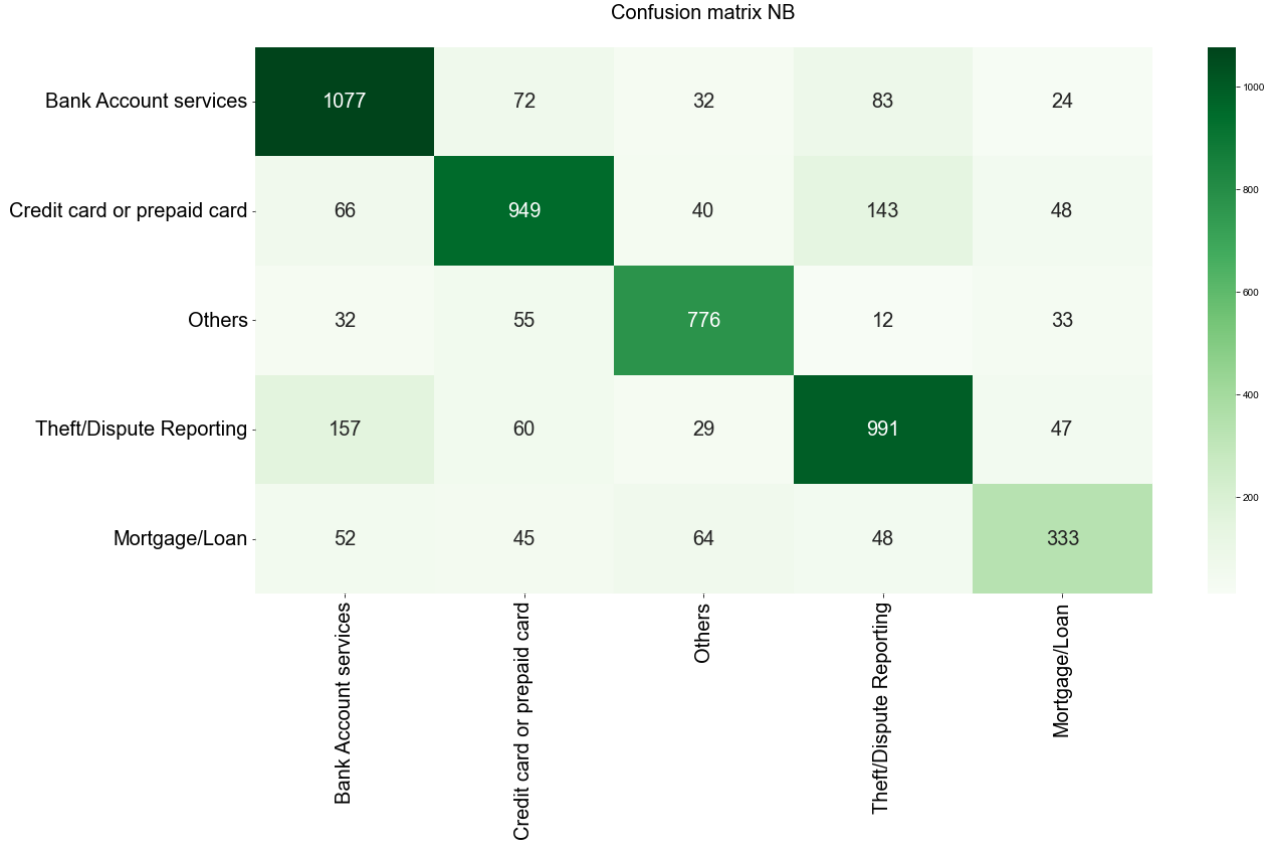
In [192…
```
# Calculating the f1 score of model using weighted avg method
F1_score_NB_tuned = f1_score(y_test, y_pred_NB, average='weighted')
F1_score_NB_tuned
```

Out[192…
```
0.7820923664188754
```

In [193…
```
# Lets evaluate the NB classifier
model_evaluation(y_test, y_pred_NB, model_name)
```

```
Classification Report NB
```

	precision	recall	f1-score	support
Bank Account services	0.78	0.84	0.81	1288
Credit card or prepaid card	0.80	0.76	0.78	1246
Others	0.82	0.85	0.84	908
Theft/Dispute Reporting	0.78	0.77	0.77	1284
Mortgage/Loan	0.69	0.61	0.65	542
accuracy			0.78	5268
macro avg	0.77	0.77	0.77	5268
weighted avg	0.78	0.78	0.78	5268



In [196…

```
# Creating df to store f1 scores of all models
f1_summary = pd.DataFrame([{'Model': 'Naive Bayes', 'f1 score': round(F1_score_NB_tuned
f1_summary
```

Out[196…

	Model	f1 score
0	Naive Bayes	0.78

2) Logistic Regression

In [197…

```
# Importing the LR alogorithm from sklearn
from sklearn.linear_model import LogisticRegression

# Initializing LR algorithm, fitting the training data and predicting on test data
model_name = 'Logistic Regression'
LR = LogisticRegression()
LR.fit(X_train, y_train)
y_pred = LR.predict(X_test)
```

In [198…

```
# Calculating the f1 score of model using weighted avg method
F1_score_LR = f1_score(y_test, y_pred, average='weighted')
F1_score_LR
```

Out[198…

```
0.9217255235034453
```

In [199…

```
# Hyperparameter training
params_LR = {'penalty': ['l1', 'l2'],
             'C': [0.001, 0.01, 0.1, 1, 10, 100],
             'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']}

gridCV_LR = GridSearchCV(estimator=LR, param_grid=params_LR, scoring='f1_weighted', ver
gridCV_LR.fit(X_train, y_train)
gridCV_LR.best_params_
```

```
Fitting 5 folds for each of 60 candidates, totalling 300 fits
```

Out[199…

```
{'C': 1, 'penalty': 'l1', 'solver': 'saga'}
```

In [201…

```
# Creating model with best hyperparams
model_name = 'Logistic Regression'
LR_tuned = LogisticRegression(C=1, penalty='l1', solver='saga')
LR_tuned.fit(X_train, y_train)
y_pred_LR = LR_tuned.predict(X_test)
```

In [202…

```
# Calculating the f1 score of model using weighted avg method
F1_score_LR_tuned = f1_score(y_test, y_pred_LR, average='weighted')
F1_score_LR_tuned
```

Out[202…

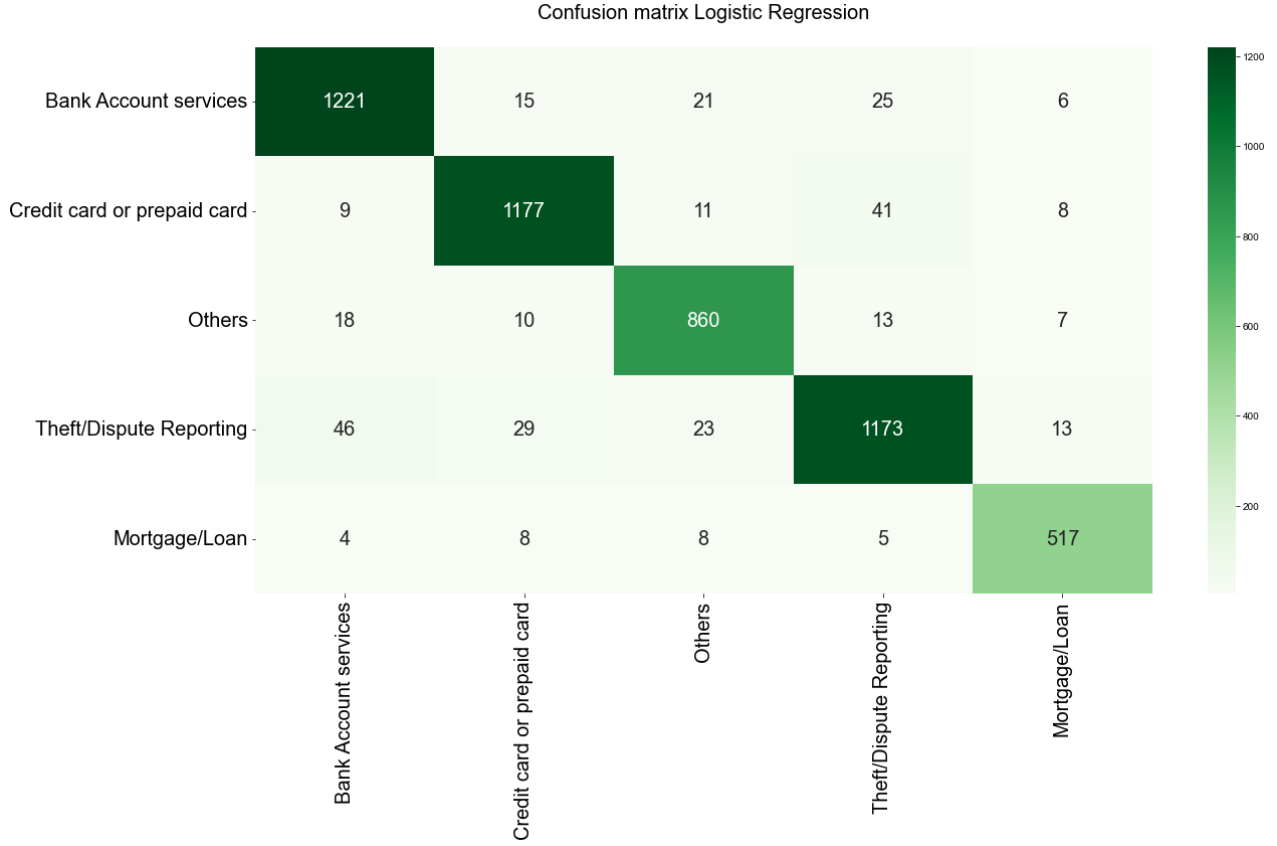
```
0.9392094693981582
```

In [203…

```
# Lets evaluate the LR classifier
model_evaluation(y_test, y_pred_LR, model_name)
```

```
Classification Report Logistic Regression
```

	precision	recall	f1-score	support
Bank Account services	0.94	0.95	0.94	1288
Credit card or prepaid card	0.95	0.94	0.95	1246
Others	0.93	0.95	0.94	908
Theft/Dispute Reporting	0.93	0.91	0.92	1284
Mortgage/Loan	0.94	0.95	0.95	542
accuracy			0.94	5268
macro avg	0.94	0.94	0.94	5268
weighted avg	0.94	0.94	0.94	5268



In [204…

```
# Update the summary df
f1_summary.loc[len(f1_summary.index)] = ['Logistic Regression', round(F1_score_LR_tuned
f1_summary
```

Out[204…

	Model	f1 score
0	Naive Bayes	0.78
1	Logistic Regression	0.94

3) Decision Tree

In [205…

```
# Importing the DT alogorithm from sklearn
from sklearn.tree import DecisionTreeClassifier

# Initializing DT algorithm, fitting the training data and predicting on test data
model_name = 'Decision Tree'
DT = DecisionTreeClassifier()
DT.fit(X_train, y_train)
y_pred = DT.predict(X_test)
```

In [206…

```
# Calculating the f1 score of model using weighted avg method
F1_score_DT = f1_score(y_test, y_pred, average='weighted')
F1_score_DT
```

Out[206…

```
0.7740651332777396
```

In [208…

```
# Hyperparameter training
params_DT = {'criterion': ['gini', 'entropy'],
             'max_depth': [5, 10, 15, 20, 25, 30],
             'min_samples_leaf': [1, 5, 10, 15, 20, 25]}

gridCV_DT = GridSearchCV(estimator=DT, param_grid=params_DT, scoring='f1_weighted', ver
gridCV_DT.fit(X_train, y_train)
gridCV_DT.best_params_
```

```
Fitting 5 folds for each of 72 candidates, totalling 360 fits
```

Out[208…

```
{'criterion': 'gini', 'max_depth': 30, 'min_samples_leaf': 20}
```

In [209…

```
# Creating model with best hyperparams
model_name = 'Decision Tree'
DT_tuned = DecisionTreeClassifier(criterion='gini', max_depth=30, min_samples_leaf=20)
DT_tuned.fit(X_train, y_train)
y_pred_DT = DT_tuned.predict(X_test)
```

In [210…

```
# Calculating the f1 score of model using weighted avg method
F1_score_DT_tuned = f1_score(y_test, y_pred_DT, average='weighted')
F1_score_DT_tuned
```

Out[210…

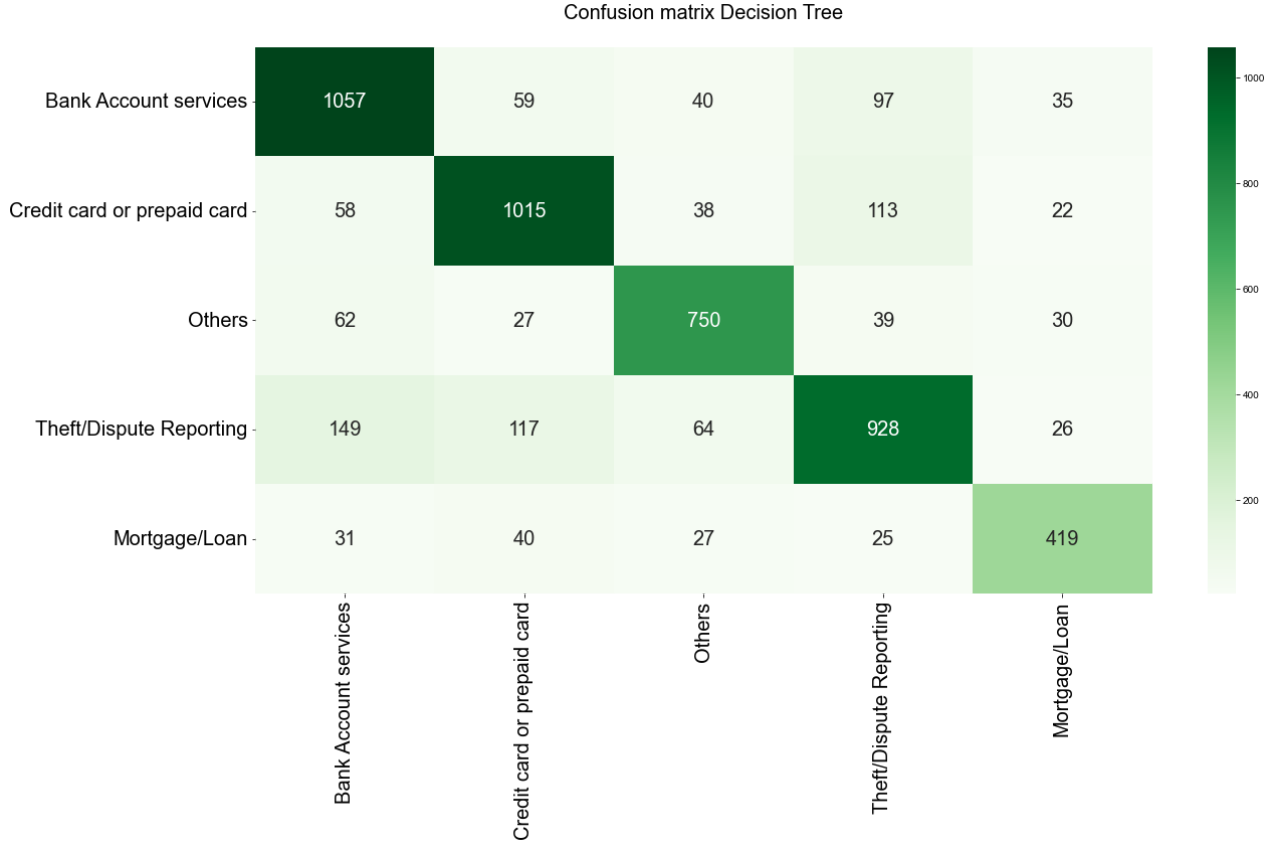
```
0.7909182590855947
```

In [211…

```
# Lets evaluate the DT classifier
model_evaluation(y_test, y_pred_DT, model_name)
```

```
Classification Report Decision Tree
```

	precision	recall	f1-score	support
Bank Account services	0.78	0.82	0.80	1288
Credit card or prepaid card	0.81	0.81	0.81	1246
Others	0.82	0.83	0.82	908
Theft/Dispute Reporting	0.77	0.72	0.75	1284
Mortgage/Loan	0.79	0.77	0.78	542
accuracy			0.79	5268
macro avg	0.79	0.79	0.79	5268
weighted avg	0.79	0.79	0.79	5268



In [212…

```
# Update the summary df
f1_summary.loc[len(f1_summary.index)] = ['Decision Tree', round(F1_score_DT_tuned, 2)]
f1_summary
```

Out[212…

	Model	f1 score
0	Naive Bayes	0.78
1	Logistic Regression	0.94
2	Decision Tree	0.79

4) Random Forest

In [232…

```
# Importing the RF alogorithm from sklearn
from sklearn.ensemble import RandomForestClassifier

# Initializing RF algorithm, fitting the training data and predicting on test data
model_name = 'Random Forest'
RF = RandomForestClassifier(max_depth=35)
RF.fit(X_train, y_train)
y_pred_RF = RF.predict(X_test)
```

In [233…

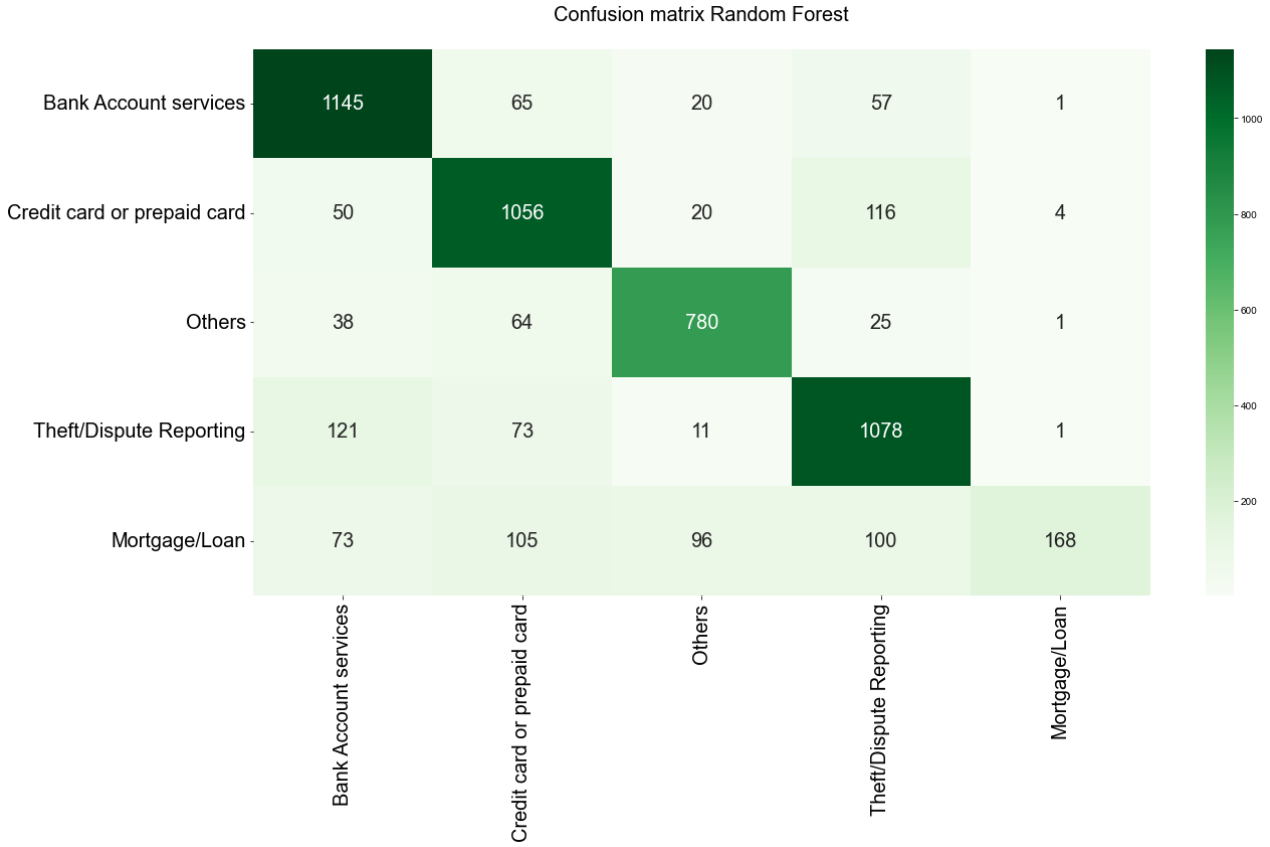
```
# Calculating the f1 score of model using weighted avg method
F1_score_RF = f1_score(y_test, y_pred_RF, average='weighted')
F1_score_RF
```

Out[233…
```
0.7899875523385225
```

In [234…
```
# Lets evaluate the RF classifier
model_evaluation(y_test, y_pred_RF, model_name)
```

```
Classification Report Random Forest
```

	precision	recall	f1-score	support
Bank Account services	0.80	0.89	0.84	1288
Credit card or prepaid card	0.77	0.85	0.81	1246
Others	0.84	0.86	0.85	908
Theft/Dispute Reporting	0.78	0.84	0.81	1284
Mortgage/Loan	0.96	0.31	0.47	542
accuracy			0.80	5268
macro avg	0.83	0.75	0.76	5268
weighted avg	0.81	0.80	0.79	5268



In [235…
```
# Update the summary df
f1_summary.loc[len(f1_summary.index)] = ['Random Forest', round(F1_score_RF, 2)]
f1_summary
```

Out[235…

	Model	f1 score
0	Naive Bayes	0.78
1	Logistic Regression	0.94
2	Decision Tree	0.79

	Model	f1 score
3	Random Forest	0.79

Conclusion: Logistic Regression model is predicting well with f1 score 0.94

In [ ]: