I. Data Preprocessing and Exploratory Data Analysis (EDA):

From my preliminary inspection of the given dataset, I found challenges concerning data quality and class distribution. Key findings include:

Missing Values:

The existence of missing values in the sub_category and crimeaditionalinfo columns required a deliberate imputation and removal process. My approach to missing subcategories was based on the self-relationship in category and sub_category, and chosen for the best data-oriented approach. Where a direct mapping was unavailable, I employed fuzzy matching techniques complemented by manual corrections guided by domain expertise. This combined strategy aimed to maximize data retention while ensuring logical consistency. For the crimeaditionalinfo column, given its centrality to the NLP analysis, rows with missing descriptions were removed to maintain the integrity of the textual analysis, acknowledging the trade-off between data quantity and analytical quality.

Feature Engineering:

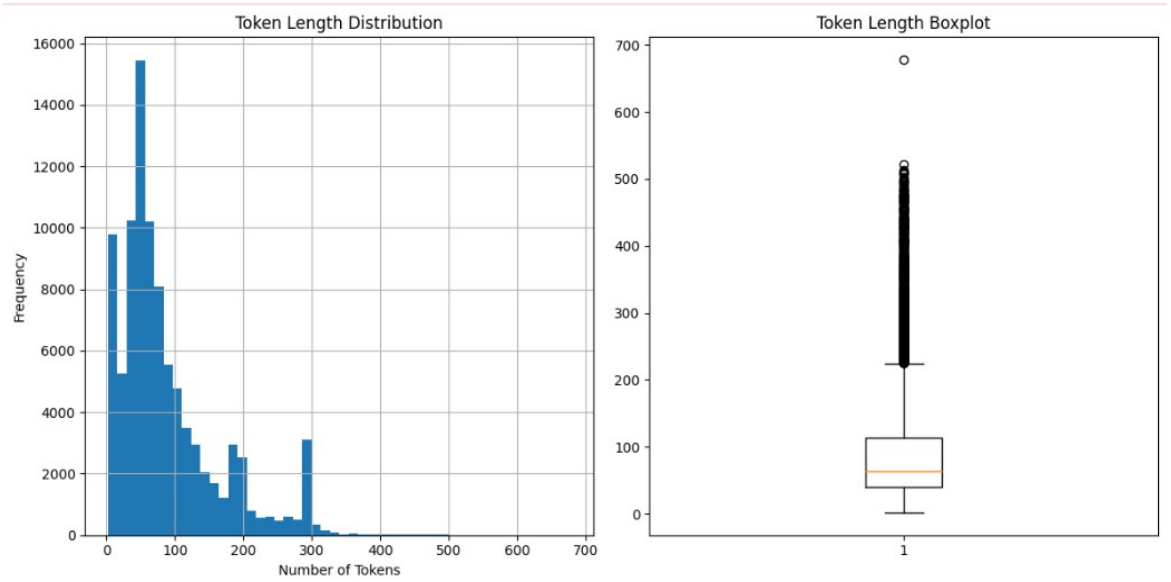
Based on the potential importance of text variables, I created new derived variables named text_length and word_count from the column called crimeaditionalinfo. These features were to feed more quantity information to the model in terms of descriptions.

Class Imbalance:

One of the key observations that I got from the EDA was that "Online Financial Fraud" accounted for most of the dataset. This severely imbalanced distribution may be a threat to the model in terms of training, since it could lead to a bias towards the majority class. I presumed to face issues with class weighting or similar mitigation techniques during model training.

Textual Variability:

To my surprise, the EDA showed a significant variability in the length of the crime descriptions within different categories. This finding hinted that maybe the level of detail or complexity of crimes varies by type and therefore affects the model's ability to learn effective representations.




```
Token Length Analysis:
Total texts: 93665
Mean length: 87.81
Median length: 64.00
Max length: 678
Min length: 2
10th percentile: 14.00
25th percentile: 40.00
50th percentile: 64.00
75th percentile: 114.00
90th percentile: 194.00
Using max sequence length: 267
```

II. Natural Language Processing (NLP):

The heart of my analysis was to extract meaningful insights from the textual crime descriptions. My approach included the following steps :

Text Preprocessing To analyze text data, I applied a standard preprocessing pipeline: tokenization, lowercasing, and stop word removal. Tokenization This step separated the descriptions into individual words using NLTK's word_tokenize function. Lowercasing ensures that words from different cases are represented uniformly, such that semantic meaning is not lost with regard to capitalization issues. Removing stop words Using NLTK's stopwords corpus, it filters out common English words like "the," "a," and "an," most of which don't hold great semantic meaning in this context.

Word Frequency Analysis:

On this step, I proceeded to word frequency analysis to identify the most frequent terms in the cleaned crime descriptions. This was presented graphically via a bar chart showing the top 20 most frequent words. The results yielded recurring patterns

such as financial crimes, where words like "account," "bank," "money," and "fraud" appeared prominently.

Word Cloud Visualization:

I created a word cloud to offer a more intuitive way to represent word frequencies. This visualization obviously brought the dominance of terms related to financial crime to light, thus providing a rapid and intuitive view of the main topics within the dataset.

III. Category and Subcategory Analysis:

In addition to the textual analysis, I investigated the distribution of and the relationships between crime categories and subcategories:

Category Distribution:

The bar chart visualization for the most prominent crime categories proved "Online Financial Fraud" to be the largest category, thus again making the training of the model with class imbalance a matter of serious consideration.

Train

Text Statistics:
---

	text_length	word_count
count	93665.000000	93665.000000
mean	398.365772	70.199231
std	329.303728	58.360233
min	1.000000	0.000000
25%	199.000000	33.000000
50%	284.000000	52.000000
75%	494.000000	90.000000
max	1517.000000	370.000000

Analyzing Text Length Distribution...

Key Insights:
1. Most common crime category: Online Financial Fraud (61.3% of all cases)
2. Average description length: 398.4 characters
3. Category with most detailed descriptions: RapeGang Rape RGRSexually Abusive Content

Total number of records: 93,665
Number of categories: 14

Number of sub-categories: 37

Test

Total number of records: 31,222
Number of categories: 14
Number of sub-categories: 37

Missing Values:
---
category 0
sub_category 0
crimeaditionalinfo 0
text_length 0
word_count 0
dtype: int64

Text Statistics:
---

	text_length	word_count
count	31222.000000	31222.000000
mean	399.151848	70.394465
std	326.924879	58.032038
min	1.000000	0.000000
25%	199.000000	33.000000
50%	288.000000	52.000000
75%	496.000000	91.000000
max	1521.000000	393.000000

Key Insights:
1. Most common crime category: Online Financial Fraud (60.5% of all cases)
2. Average description length: 399.2 characters
3. Category with most detailed descriptions: RapeGang Rape RGRSexually Abusive Content

Distribution among Subcategories:

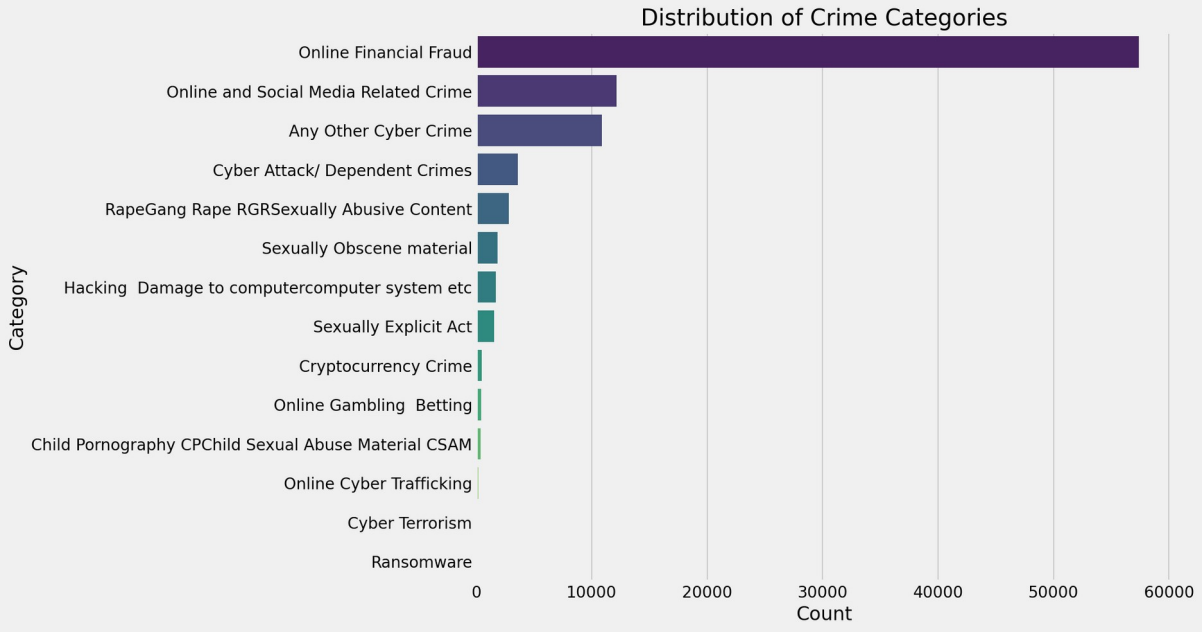
To understand crime types at a more granular level, I visualized the top 15 most prominent subcategories. In doing so, I got an idea about the occurrence of some cybercrime types based on this dataset.

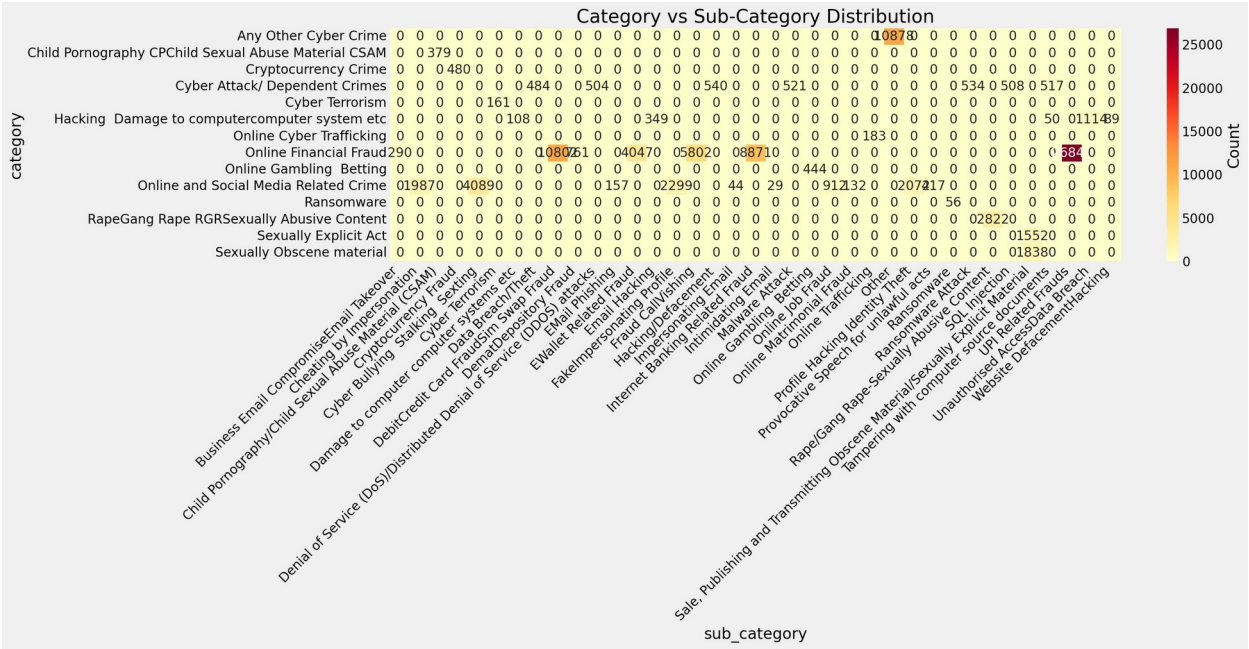
Category-Subcategory Inter-relationship:

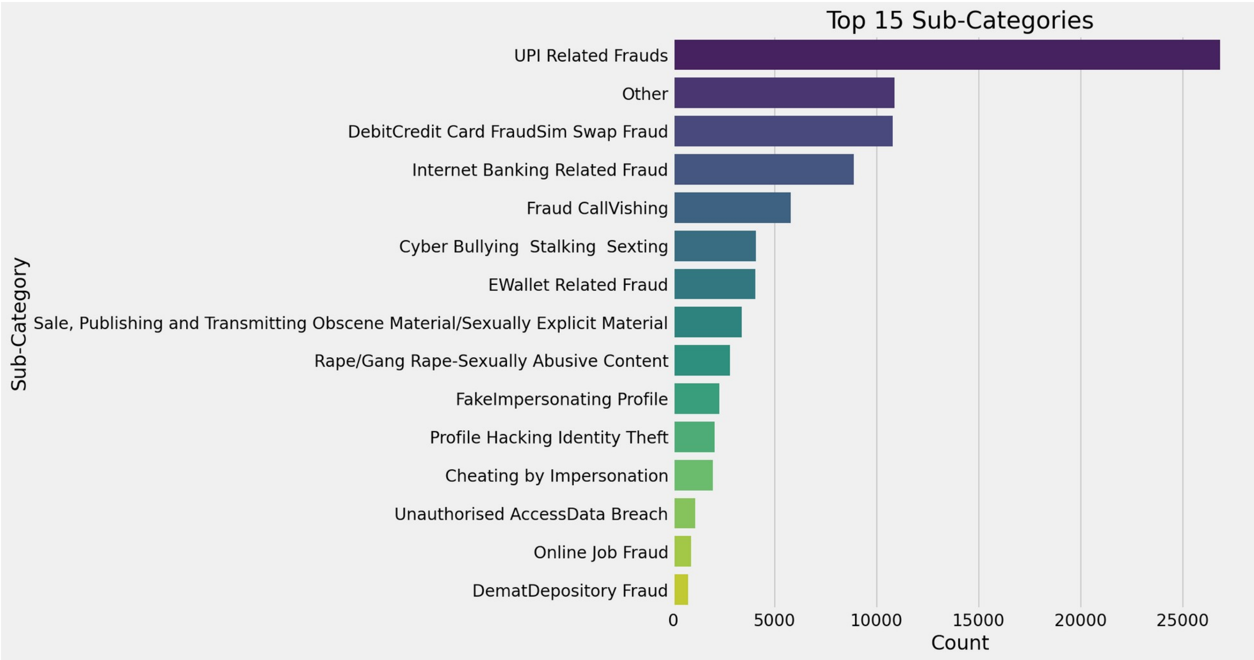
I used a heatmap of the cross-tabulation between category and subcategory variables to investigate how categories and subcategories co-vary. This helped me to get a

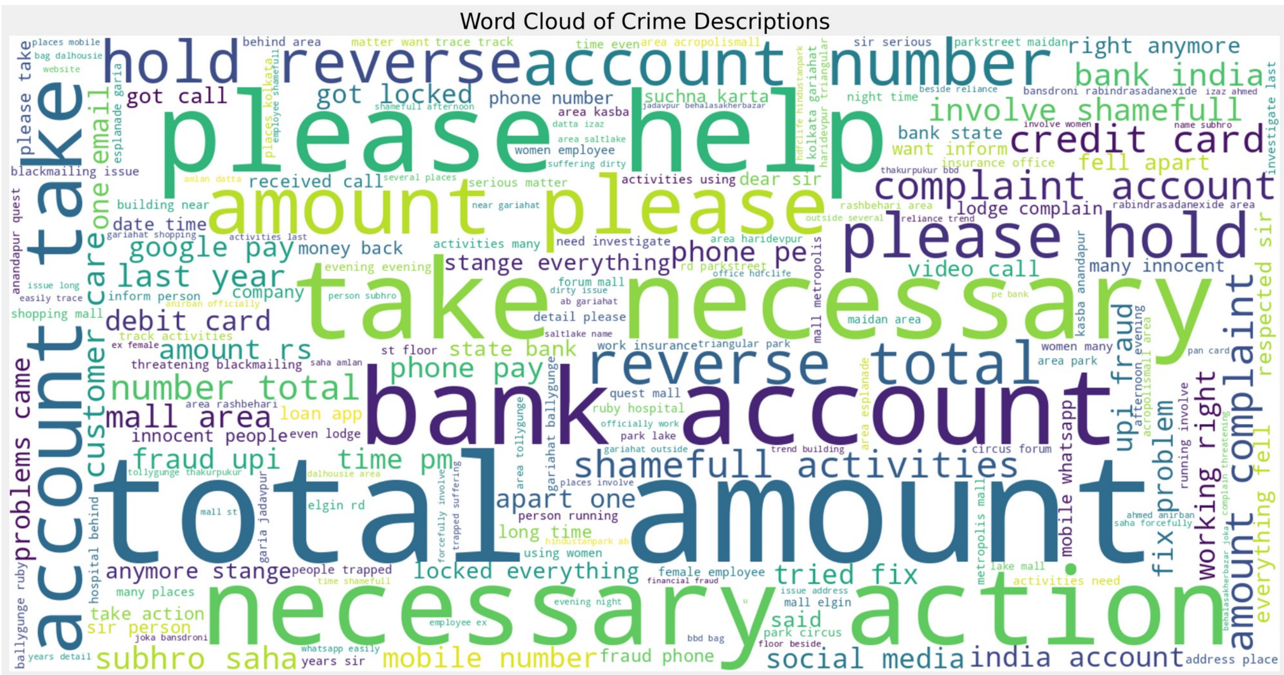
clearer understanding of a hierarchical structure created by the crime categories system, indicating which subcategories co-occur with each main category. Such insight into model predictions may be useful in making meaningful interpretations of the relationships between various types of crime.

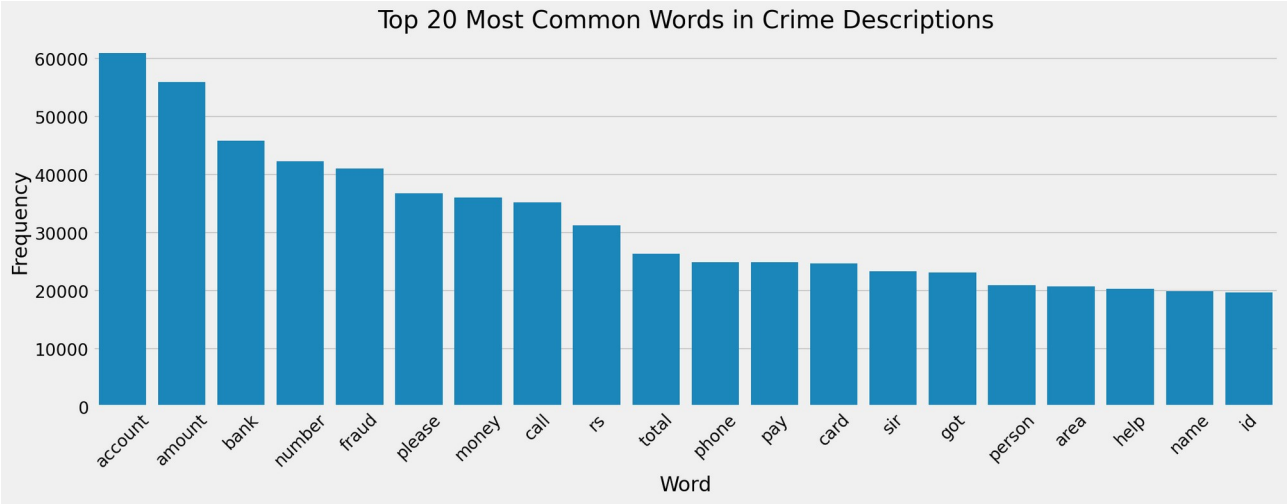
Train

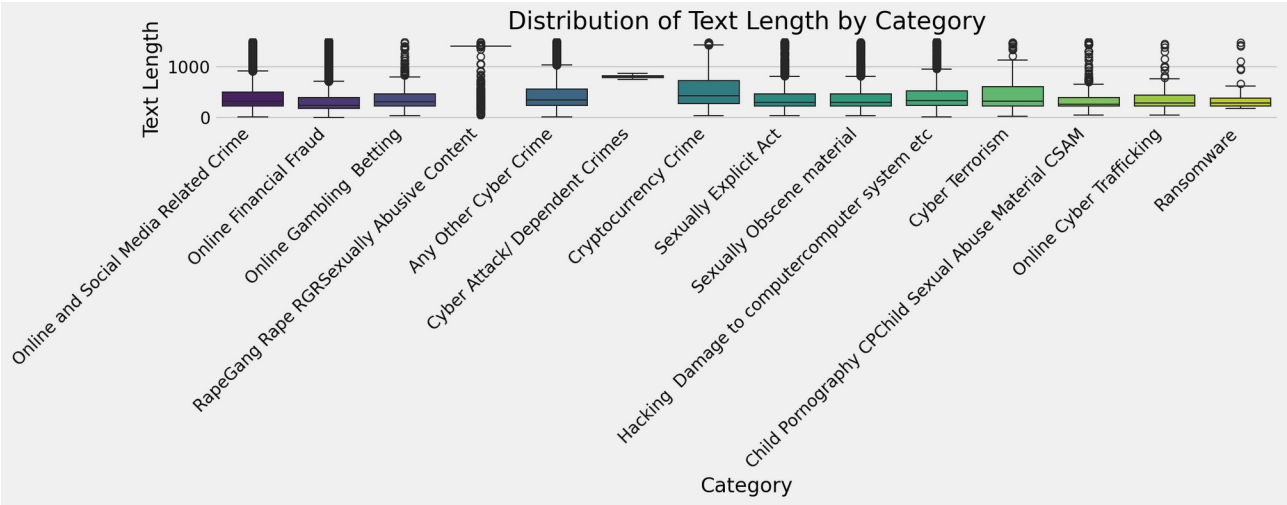




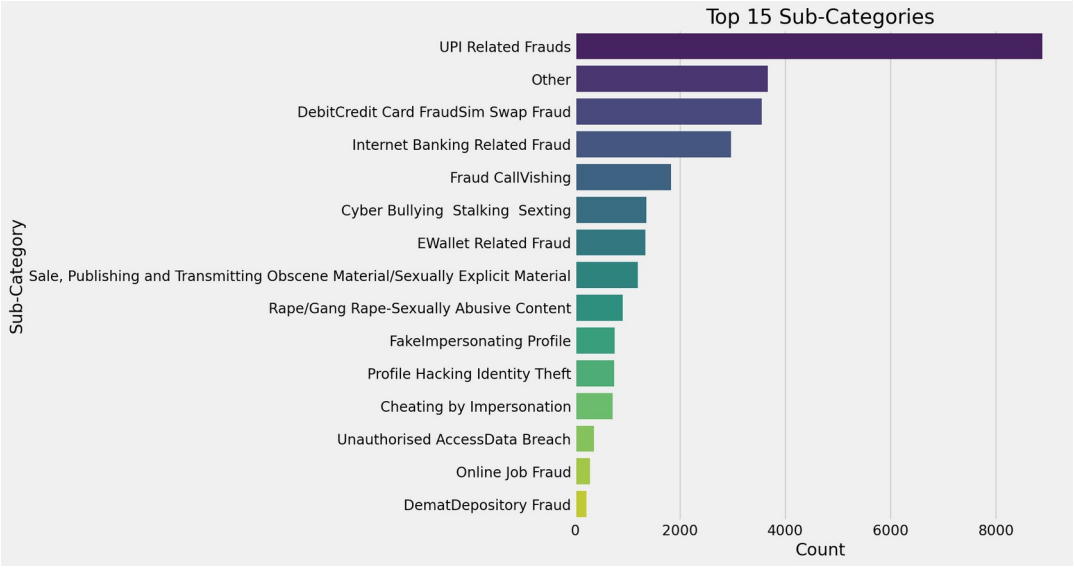

Top 15 Sub-Categories
UPI Related Frauds
Other
DebitCredit Card FraudSim Swap Fraud
Internet Banking Related Fraud
Fraud CallVishing
Cyber Bullying Stalking Sexting
EWallet Related Fraud
Sale, Publishing and Transmitting Obscene Material/Sexually Explicit Material
Rape/Gang Rape-Sexually Abusive Content
FakeImpersonating Profile
Profile Hacking Identity Theft
Cheating by Impersonation
Unauthorised AccessData Breach
Online Job Fraud
DematDepository Fraud
Sub-Category
0
5000
10000
15000
20000
25000
Count

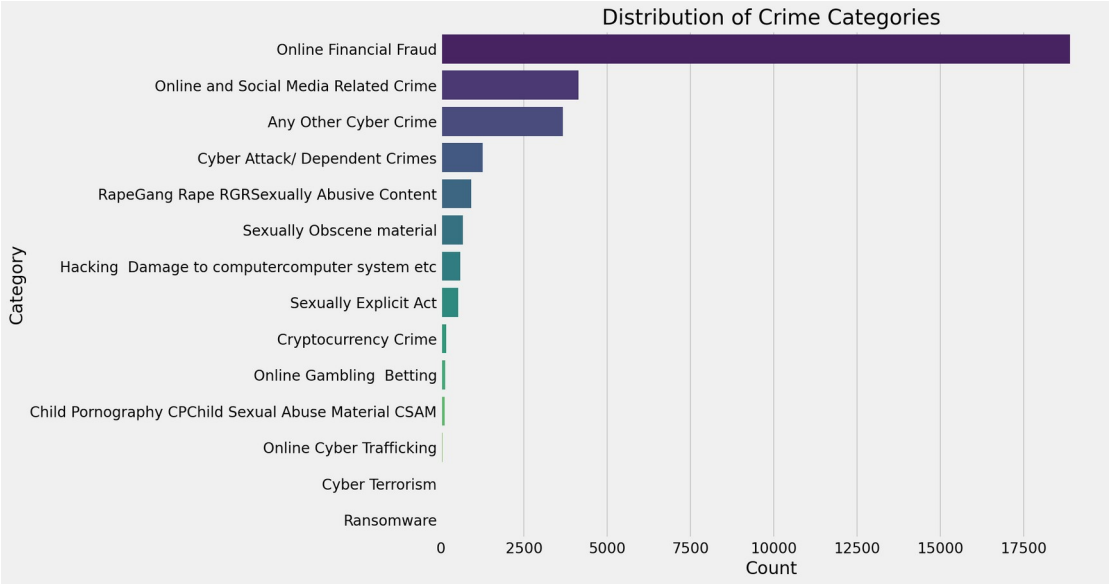

Word Cloud of Crime Descriptions
total amount
bank account
take necessary
necessary action
please help
amount please
please hold
account take
hold reverse
account number
reverse total
complaint account
credit card
shamefull activities
bank india
involve shameful
debit card
google pay
phone pay
phone pe
last year
number total
mall area
fraud upi
time pm
got locked
got call
video call
social media
india account
mobile number
subhro saha
tried fix
locked everything
stange everything
customer care
problems came
amount complaint
working right
upi fraud
right anymore
fell apart
received call
money back
state bank
apart one
innocent people
take action
sir person
dear sir
lodge complain
bank state
loan app

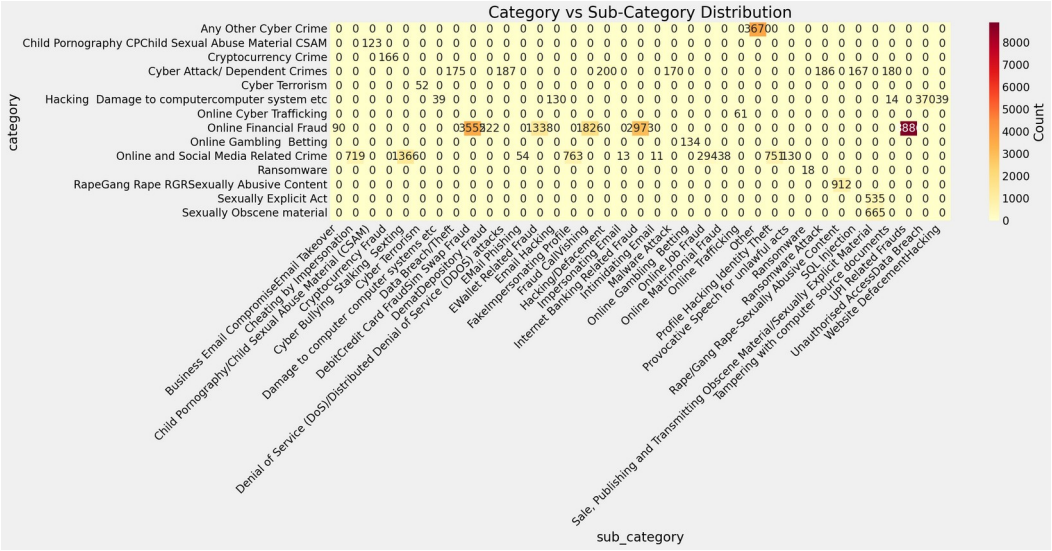

Top 20 Most Common Words in Crime Descriptions
Frequency
60000
50000
40000
30000
20000
10000
0
account
amount
bank
number
fraud
please
money
call
rs
total
phone
pay
card
sir
got
person
area
help
name
id
Word

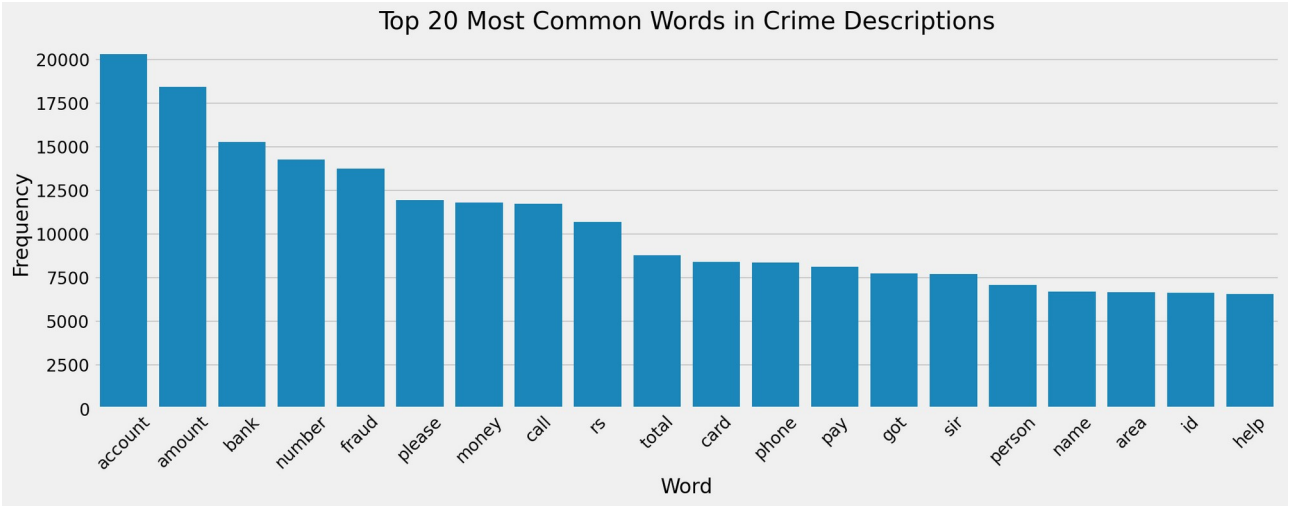

Distribution of Text Length by Category
Text Length
1000
0
Online and Social Media Related Crime
Online Financial Fraud
Online Gambling Betting
RapeGang Rape RGRSexually Abusive Content
Any Other Cyber Crime
Cyber Attack/ Dependent Crimes
Cryptocurrency Crime
Sexually Explicit Act
Sexually Obscene material
Hacking Damage to computercomputer system etc
Cyber Terrorism
Child Pornography CPChild Sexual Abuse Material CSAM
Online Cyber Trafficking
Ransomware
Category

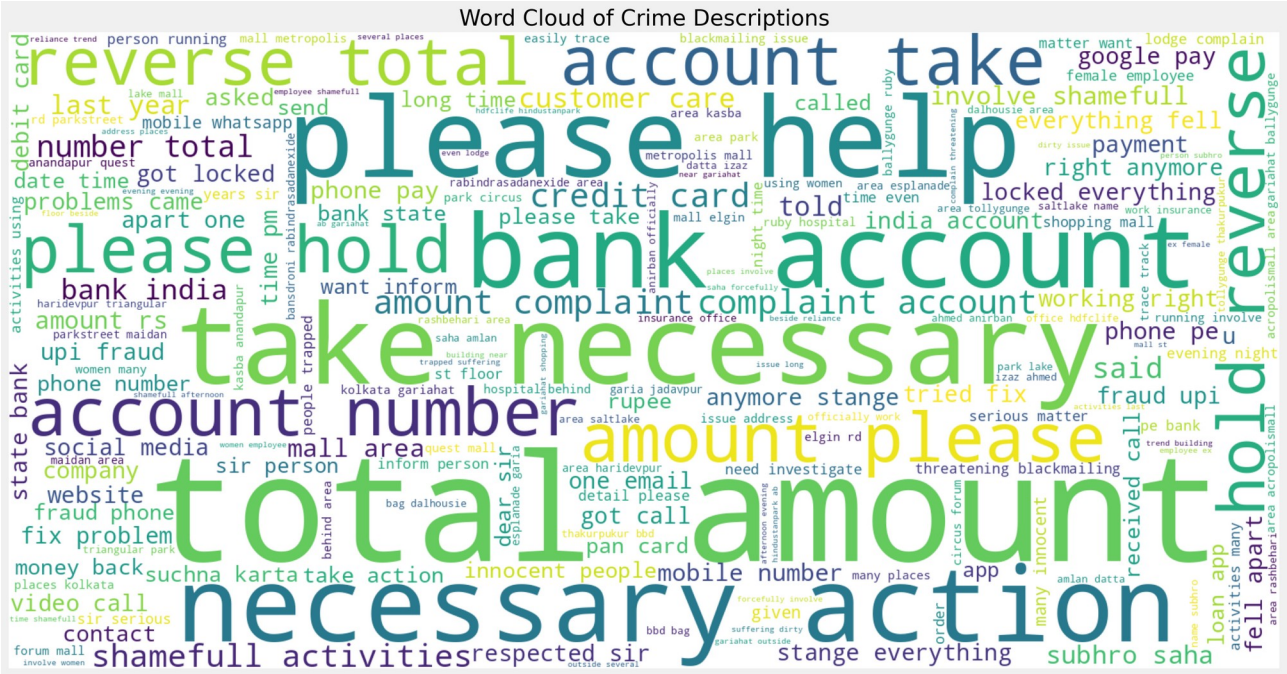
Test


Top 15 Sub-Categories
Sub-Category
UPI Related Frauds
Other
DebitCredit Card FraudSim Swap Fraud
Internet Banking Related Fraud
Fraud CallVishing
Cyber Bullying Stalking Sexting
EWallet Related Fraud
Sale, Publishing and Transmitting Obscene Material/Sexually Explicit Material
Rape/Gang Rape-Sexually Abusive Content
FakeImpersonating Profile
Profile Hacking Identity Theft
Cheating by Impersonation
Unauthorised AccessData Breach
Online Job Fraud
DematDepository Fraud
0
2000
4000
6000
8000
Count


Distribution of Crime Categories
Online Financial Fraud
Online and Social Media Related Crime
Any Other Cyber Crime
Cyber Attack/ Dependent Crimes
RapeGang Rape RGRSexually Abusive Content
Sexually Obscene material
Hacking Damage to computercomputer system etc
Sexually Explicit Act
Cryptocurrency Crime
Online Gambling Betting
Child Pornography CPChild Sexual Abuse Material CSAM
Online Cyber Trafficking
Cyber Terrorism
Ransomware
Category
0
2500
5000
7500
10000
12500
15000
17500
Count


Category vs Sub-Category Distribution
category
sub_category
Count
Any Other Cyber Crime
Child Pornography CPChild Sexual Abuse Material CSAM
Cryptocurrency Crime
Cyber Attack/ Dependent Crimes
Cyber Terrorism
Hacking Damage to computercomputer system etc
Online Cyber Trafficking
Online Financial Fraud
Online Gambling Betting
Online and Social Media Related Crime
Ransomware
RapeGang Rape RGRSexually Abusive Content
Sexually Explicit Act
Sexually Obscene material
Business Email CompromiseEmail Takeover
Cheating by Impersonation
Child Pornography/Child Sexual Abuse Material (CSAM)
Cryptocurrency Fraud
Cyber Bullying Stalking Sexting
Cyber Terrorism
Damage to computer computer systems etc
Data Breach/Theft
DebitCredit Card FraudSim Swap Fraud
Denial of Service (DoS)/Distributed Denial of Service (DDOS) attacks
DematDepository Fraud
EMail Phishing
EWallet Related Fraud
Email Hacking
FakeImpersonating Profile
Fraud CallVishing
Hacking/Defacement
Impersonating Email
Internet Banking Related Fraud
Intimidating Email
Malware Attack
Online Gambling Betting
Online Job Fraud
Online Matrimonial Fraud
Online Trafficking
Other
Profile Hacking Identity Theft
Provocative Speech for unlawful acts
Ransomware
Rape/Gang Rape-Sexually Abusive Content
Ransomware Attack
Sale, Publishing and Transmitting Obscene Material/Sexually Explicit Material
SQL Injection
Tampering with computer source documents
Unauthorised Access/Data Breach
UPI Related Frauds
Website DefacementHacking


Top 20 Most Common Words in Crime Descriptions
Frequency
Word
account
amount
bank
number
fraud
please
money
call
rs
total
card
phone
pay
got
sir
person
name
area
id
help



V. Analysis of Length of Text

There is a likelihood of text length affecting the model performance. Thus, I did an analysis of the distribution of text length across categories. Box plots are used to clearly visualize the variation in description lengths. This indicated that certain categories like "RapeGang Rape RGRSexually Abusive Content" tend to have longer and potentially more detailed descriptions as compared to others. It would inform feature engineering that makes features tailored to the characteristics of certain categories.

V. XGBoost Model and Evaluation:

In this experiment, I made use of the XGBoost model, which is a prominent gradient boosting algorithm suited ideally for classification operations. My implementation has taken into account the following important considerations:

TF-IDF Features :

Text descriptions were converted into numeric features through TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. TF-IDF captures the importance of words in a document, relatively to a larger corpus and hence serves as robust representation of textual data for our model.

Class Weighting:

I utilized balanced class weighting during training of the model. That is, it gives higher weights to underrepresented classes. This allows one to penalize misclassifications over an underrepresented class much more severely than the rest and accordingly teach the model towards more well-balanced representations.

Model Parameters:

The XGBoost model was initialized with a set of hyperparameters aimed at balancing model complexity and performance. These parameters included objective, eval_metric, tree_method, learning_rate, max_depth, and others. These parameters govern various aspects of the model's training process, influencing its ability to learn from the data and generalize to unseen examples.

Evaluation Metrics:

The performance of the model was evaluated using a whole set of metrics, including precision, recall, F1-score, and support for each category and subcategory. Because of class imbalance, weighted F1-score was selected as the performance metric that gave a balanced measure for the overall effectiveness of the model. A classification report with detailed metrics across the individual classes was further generated to understand how the model was performing on each class separately.

Category prediction is fairly good, with the weighted F1-score equaling 0.6681, whereas subcategory prediction has a lower score with a weighted F1-score at 0.4926. This may be due to increased complexity and granularity of the subcategory classification task with additional inherent class imbalance.

==========

Total training completed at 2024-11-22 22:55:12

Total training time: 166.3 minutes
==========

==========
Evaluating model performance
==========

==========
Starting predictions at 2024-11-22 22:55:12
==========

Transforming texts to TF-IDF features...
Feature matrix shape: (31222, 5000)

Making category predictions...

Making subcategory predictions...

Predictions completed in 0.1 minutes

Category Classification Report:

	precision	recall	f1-score	support
0	0.32	0.45	0.37	3670
1	0.18	0.41	0.25	123
2	0.21	0.84	0.34	166
3	1.00	1.00	1.00	1265
4	0.00	0.00	0.00	52
5	0.20	0.66	0.30	592
6	0.00	0.02	0.00	61
7	0.94	0.69	0.80	18890
8	0.04	0.33	0.06	134
9	0.57	0.34	0.42	4139
10	0.06	0.17	0.09	18
11	1.00	0.91	0.95	912
12	0.13	0.24	0.17	535
13	0.21	0.37	0.27	665
accuracy			0.62	31222
macro avg	0.35	0.46	0.36	31222
weighted avg	0.76	0.62	0.67	31222

Category F1 Score (Weighted): 0.6681

Subcategory Classification Report:

	precision	recall	f1-score	support
0	0.05	0.21	0.08	90
1	0.14	0.21	0.17	719
2	0.23	0.40	0.30	123
3	0.29	0.77	0.43	166
4	0.47	0.50	0.48	1366
5	0.00	0.00	0.00	52
6	0.07	0.08	0.07	39
7	0.17	0.19	0.18	175
8	0.67	0.68	0.67	3555
9	0.06	0.21	0.09	222
10	0.15	0.17	0.16	187
11	0.10	0.20	0.13	54
12	0.43	0.49	0.46	1338
13	0.27	0.44	0.33	130
14	0.38	0.45	0.41	763
15	0.24	0.34	0.28	1826
16	0.15	0.12	0.14	200
17	0.00	0.00	0.00	13
18	0.59	0.59	0.59	2973
19	0.00	0.00	0.00	11
20	0.10	0.11	0.10	170
21	0.05	0.20	0.08	134
22	0.20	0.66	0.31	294
23	0.04	0.11	0.06	38
24	0.00	0.00	0.00	61
25	0.47	0.17	0.25	3670
26	0.45	0.42	0.43	751
27	0.05	0.42	0.09	130
28	0.15	0.22	0.18	18
29	0.12	0.11	0.12	186
30	0.99	0.91	0.95	912
31	0.12	0.13	0.12	167
32	0.41	0.28	0.33	1200
33	0.15	0.11	0.13	194
34	0.83	0.56	0.67	8886
35	0.17	0.51	0.26	370
36	0.04	0.05	0.04	39
accuracy			0.47	31222
macro avg	0.24	0.30	0.25	31222
weighted avg	0.56	0.47	0.49	31222

Subcategory F1 Score (Weighted): 0.4926

VI. Implementation Plan (Moving Forward):

Following the results and model evaluation, I propose the following steps for further development and improvement:

Add Augmentation:

To address the imbalance class, I would suggest exploring data augmentation techniques like SMOTE or back-translation. These methods synthetically generate samples for under-represented categories to enhance the model's ability to learn a better representation of the classes.

Advanced Feature Engineering:

Beyond TF-IDF, I would propose exploring more advanced text-embedding techniques, such as Word2Vec or BERT, to capture the semantic relationships among words. Another approach might be to seek data from diverse contexts to gain further contextual information.

Model tuning and exploration - Further hyperparameter fine-tuning of the XGBoost model, using for example, techniques in grid search or Bayesian optimization, is suggested. In addition, exploration of additional architectures of the model, deep learning models are known to perform well when having complex dependencies between data elements.

In-depth Error Analysis: A thorough error analysis is crucial for understanding the specific types of misclassifications the model makes. This analysis can reveal biases, feature limitations, or data quality issues that require attention. Manual inspection of misclassified examples can provide valuable insights.

Deployment and Monitoring:

Once the model reaches satisfactory performance, it is deployed to a production environment for real-time classification. The deployed model needs to be continuously monitored, and new data is fed into it at regular intervals for retraining it to maintain accuracy and update according to changing cybercrime trends.

Libraries Used:

- pandas
- numpy
- matplotlib
- seaborn

- wordcloud
- nltk
- sklearn
- xgboost
- difflib
- re