**Data Analysis and Machine Learning classification of customer complaint dataset**

1. **Objectives: The objectives fulfilled in the following task are as follows-**
   1. Identify the most representative sub-sample from the dataset
   2. Perform Data cleaning, pre-processing on the chosen sub-sample
   3. Visualisation of statistics and insights on the data
   4. Perform NLP tasks such as sentiment analysis, Emotion Detection, Topic Modeling
   5. Implement Machine Learning Model
   6. Describe the performance of the ML model based on performance metrics such as accuracy, precision, recall, f1-score
2. **Data Analysis:**

The data analysis has been performed on the given dataset - [link](https://catalog.data.gov/dataset/consumer-complaint-database) as follows-

* **Cleaning** - The data has been cleaned for the unnecessary index values by dropping the index. The data has also been checked for the Null values. Strings have replaced columns containing NaN values.
* **Sampling -** Data Sampling has been done on the dataset to find the most representative features for further analysis and classification. Only required 5 features have been considered, and the rest have been dropped.
* **Pre**-**processing** – The pre-processing for NLP tasks includes removing stop words, punctuation, special characters from the string. The text from the customer complaint narrative has been pre-processed using the required techniques. It has also been tokenised to handle it in further processing. The NaN values replaced by XYZ are then dropped to balance the data and increase the quality of the analysis.

Data pre-processing also includes multiple operations in terms to get the clean text to analyse the sentiments from it.

1. **Sentiment Analysis:**

* The text from the complaints has been then analysed to collect sentiments from it. The analysis has been performed using the TextBlob library.
* The dataframe containing polarity, subjectivity, and sentiments has been concatenated with the text.
* The text has been then categorised into sentiments based on their polarity score. The Positive, Negative, Neutral sentiments have been assigned to the text based on the score.
* Though the text contained customer complaints, positive sentiment has been analysed in the complaints surprisingly.
* The analysis has been visualised using the bar plot based on the counts of sentiments detected in the individual text.

1. **Emotion Detection**

* The 5 categories of emotions have been detected from the customer complaint narrative such as Fear, Sad, Surprise, Happy, Angry
* Text2emotion library has been utilised to detect the emotions from the text sequences.
* Fear is the dominant emotion among the complaints, which can be clearly seen in the visualisation
* Surprisingly the complaints showed significantly less anger in the text analysis, and the other emotions, sad and surprised, were more than anger.
* The unexpected emotion count can be seen for happy emotion, almost four times higher than angry emotion.

1. **Topic Modelling**

* Topic modelling has been performed on the text to understand the insights such as what are the most complaints and issues are about.
* It is an attempt to gather insights for operations, such as identifying product categories based on the issues mentioned by customers.
* The text from issue and sub-issue has been concatenated to get a specific sequence length, which is pre-processed and further used for topic modelling using the Latent Dirichlet Allocation (LDA) model.
* Text from complaints has also been used for topic modelling in terms to understand what category of product has the most complaints about.
* The Issues document Topic 7: shown 73.3250732200502 % where the topic contained ‘Topic 7’: ‘information belongs else someone incorrect report atm debit using card.’
* For Complaints document Topic 9 shown 69.69914888607647 %, where it included ‘Topic 9’: ‘credit report account inquiry identity theft information fraudulent remove victim’ vocabulary

1. **Applying Machine Learning Classifiers using stratified K-fold cross-validation**

* Multiple machine learning classifiers have been applied to classify the text from the customer complaint narrative based on its sentiment.
* The text was first cleaned and pre-processed and divided into train and test datasets using stratified K-folds.
* The Sentiments were taken as labels, and the customer complaints have been provided for the prediction.
* The train, test split with 10 stratified folds, has been provided to the tokenizer and then used in the classification models.
* Stochastic Gradient Descent Classifier, Decision Tree classifier, Random Forest Classifier, Ridge Classifier have been implemented, and their performance has been further compared to find the best classifier among them.
* The accuracy provided by these classifiers is the mean of the accuracies collected from each fold. Validation accuracy has also been compared to find a better performing ML classifier.

1. **Applying Machine learning classifier without stratified K-fold cross-validation**

* Data train-test split has been performed using Sklearn library and further used into the dedicated ML Classifiers with Tfidf vectorisation.
* Passive Aggressive Classifier, Naïve Bayes Classifier have given 82% and 58% accuracies, respectively.
* A for loop has been implemented to compare these two classifiers’ performances to train the models and extract performance metrics based on the classification reports.
* After the training and prediction using these two ML models, all the performance metrics, classification reports, and confusion matrix have been analysed.

1. **Machine Learning classifiers and their comparative analysis**

* Though the passive-aggressive classifier has given 82% accuracy, multiple other traditional (SOTA) and suitable ML classifiers have been implemented to perform a comparative analysis.
* These classifiers include Perceptron, Stochastic Gradient Classifier, Naive Bias Classifier, Decision Tree Classifier, Random Forest Classifier, Adaboost Classifier, Ridge Classifier, SVC Classifier, Voting Based Classifier.
* The methodology followed to train these models were the same, and we also evaluated precision, recall, f1-score, and accuracy.
* The comparative analysis is based on the results achieved by these models, which were further visualised against each other’s performance in performance metrics figures.
* The performance metrics achieved by all the classifiers can be seen in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier Name | Accuracy  % | Precision  % | Recall  % | F1-score  % |
| Passive Aggressive Classifier | 81.2 | 79 | 80 | 81 |
| Naïve Bayes Classifier | 58.39 | 77 | 43 | 39 |
| perceptron | 80 | 93 | 95 | 94 |
| Stochastic Gradient Descent Classifier | 79 | 95 | 95 | 95 |
| Naive Bias Classifier | 55 | 68 | 93 | 78 |
| Decision Tree Classifier | 70 | 92 | 94 | 93 |
| Random Forest Classifier | 74 | 94 | 96 | 95 |
| AdaBoost Classifier | 72 | 89 | 98 | 93 |
| Ridge Classifier | 79 | 95 | 94 | 95 |
| SVC Classifier | 67 | 96 | 89 | 93 |
| Voting Based Classifier | 80 | 96 | 94 | 95 |

* The table above describes the highest accuracy of the passive-aggressive classifier - 81.2%, whereas the other classifiers provide moderate accuracies, such as Stochastic Gradient Descent Classifier -79%, Random Forest Classifier – 74%, AdaBoost Classifier – 72, and Ridge Classifier achieves 79% accuracy.
* The highest f1-score is given by the random forest classifier and the other two classifiers, which is 95%.