Task 6. Data Science example.

Perform a Text Classification on consumer complaint dataset

(https://catalog.data.gov/dataset/consumer-complaint-database) into following categories.

0 Credit reporting, repair, or other

1 Debt collection

2 Consumer Loan

3 Mortgage

Steps to be followed -

1. Explanatory Data Analysis and Feature Engineering

2. Text Pre-Processing

3. Selection of Multi Classification model

4. Comparison of model performance

5. Model Evaluation

6. Prediction

Text classification on consumer complaint datasets categorizes complaints into predefined categories (like debt collection, credit reporting, mortgage and more), improving customer service efficiency. This model can evolve into an automated system, reducing manual sorting and addressing. Machine learning and AI can classify new complaints with high accuracy, routing them to appropriate teams for resolution, enhancing customer satisfaction and revenue.

I trained a text classifier on a large dataset of 4,055,791 customer complaints submitted to the CFPB by US financial institutions, using Jupyter Notebook (Anaconda3) to streamline the process and avoid repetitive dataset uploads.

The project involves labelling complaints to a specific product, a supervised problem statement. Different machine learning algorithms will be tested to determine the best one. The model will classify new complaints and assess its performance, using natural language processing for text pre-processing.

Several actions must be taken to complete a text categorization work on the Consumer Complaint Database, as I outline in the following step-by-step guide:

**At the very beginning** we need to import everything what we are going to use in our project execution

**Step 1: Explanatory Data Analysis (EDA) and Feature Engineering:**

* Load the dataset and explore it to understand its structure, columns, and distribution.
* Examine the target variable ('Product') to see the distribution of categories.
* Handle missing data.
* Perform necessary feature engineering, like TF-IDF features and important feature selection.
* We also use replace function to reduce the product complain category.
* Also creating new features from the text, like text length, word count for better understanding.

**Step 2: Text Pre-Processing:**

* Tokenization: Split the text into words or tokens.
* Removing Stop Words: Remove common words that do not carry much information.
* In place of Encoding Labels on product category I create a separated column category\_id to do the same working.
* Here we do use TF-IDF and chi-Square for feature engineering

**Step 3: Selection of Multi-Classification Model:**

* Split the dataset into training and testing sets.(20% test and 80% train data)
* Choose a suitable text classification algorithm such as:
  + Multinomial Naive Bayes
  + Linear Support Vector Machines (SVM)
  + Random Forest Classifier
  + Logistic Regression

**Step 4: Comparison of Model Performance:**

* Train multiple classification models with the pre-processed text data.
* Evaluate the models using metrics mean accuracy.
* Perform cross-validation to assess model robustness.

**Step 5: Model Evaluation:**

* Select the best-performing model based on evaluation metric and also evaluate the other matrices like accuracy, precision, recall, F1-score.
* Visualize the results using confusion matrices.

**Step 6: Prediction:**

* Use the chosen model to make predictions on new or unseen data.
* Convert the numerical predictions back to their corresponding category labels.