# Mid-Program Project

## PG Certification in Data Science – Electronics & ICT Academy IITG



**TABLE OF CONTENTS**

## BACKGROUND: CONSUMER COMPLAINT RESOLUTION

When consumers are not happy with some aspect of a business, they choose to reach out to the customer service and might raise a complaint. Businesses try their best to resolve the complaints that they receive. However, it might not always be possible to appease every customer.

Unhappy consumers might raise follow-up questions/complaints about the resolutions provided, and this is detrimental to the business as it points to systemic failures in the Customer Support division and could lead to poor brand image. Disputed complaints which are being/have been resolved could be a critical dataset to derive essential learnings for any business.

Predicting whether a complaint resolution will be accepted or rejected by a consumer can enable a business to proactively look at complaints which might be disputed and hence save unnecessary escalation as well as their reputation. Systemic issues can be identified by noticing which complaints have a higher potential to be disputed, and customer support agents can be trained to pay more attention or enhance the quality of communication for certain types of complaints.

The **Consumer Financial Protection Bureau (CFPB) in the United States** receives several consumers' complaints about the dealings of financial companies. It sends these complaints about their products and services to them for eliciting a response. The CFPB makes sure that these complaints are published here soon after the company responds or after 15 days since sending the complaint to the company.

## DATASET

You have been provided with a dataset containing the following columns –

- **Date received**: Date when the complaint was received
- **Product**: Type of product identified in the complaint, e.g., "Student loan"
- **Sub-product**: Type of sub-product identified in the complain
- **Issue**: The issue raised in the complaint, e.g., "Struggling to repay your loan."
- **Sub-issue**: E.g., "Problem lowering your monthly payments."
- **Consumer complaint narrative**: This is a consumer-submitted description of "what happened". Reasonable steps have been taken to remove personal information that could be used to identify the consumer
- **Company public response**: The response to a consumer's complaint. It can be from a pre-set list of options, e.g., "Company believes the complaint is the result of an isolated error"

- **Company**: For which the complaint is about
- **State**: Derived from the consumer's mailing address
- **ZIP Code**: Derived from the consumer's mailing address
- **Consumer consent provided:** Flag to specify whether the consumer allowed the publishing of their complaint description
- **Submitted via**: E.g., "Web" or "Phone."
- **Date sent to the company**
- **Company response to consumer**
- **Timely response**: Flag specifying if the response was timely
- **Consumer disputed**: Flag specifying if the consumer disputed the resolution
- **Complaint ID**: Identifier for each complaint

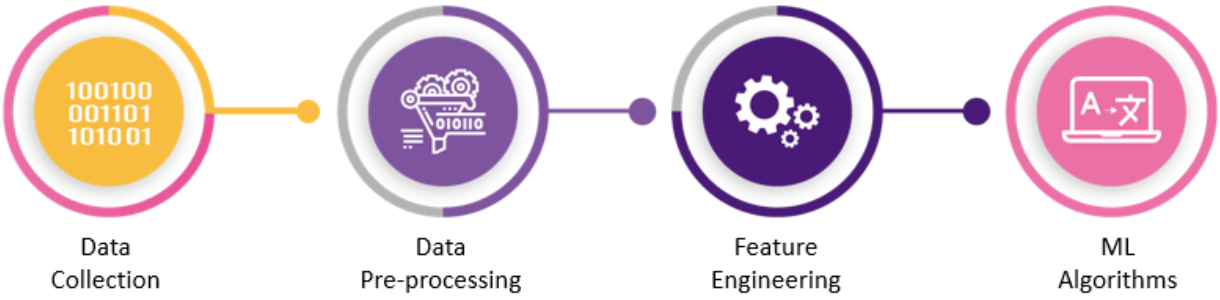
Two files have been provided.

- **Training Data**: Edureka_Consumer_Complaints_train.csv
- **Test Data**: Consumer_Complaints_test.csv [Does not have the Consumer Disputed column]



## TASKS

**Process of Building an ML Model**




Data Pre-processing and Exploratory Data Analysis

1. You should be able to answer questions like - "What types of issues have a higher chance of being disputed?", "Do timely responses lead to a lower chance of a

complainer not being happy with the conclusion?", "Is there a geographical skew in the complaints that are received and conclusions that are disputed?" and more.

2. Once you've had a deeper understanding of the provided dataset, you should be able to build a model to identify the type of product for which a consumer has raised a complaint using the narrative provided. Here, you would need to use natural language processing techniques that you've learnt during the course and build features from the text columns.

Feature Engineering and ML Model Building with Algorithms

3. Lastly, you should use the numerical, categorical and text columns (Consumer complaint narrative, Company public response etc..) to build features and then train a few predictive models to figure out whether you can identify if the resolution for a consumer complaint will be disputed. You should try and find the best hyperparameters by searching through a bunch of combinations to get better prediction metrics.



## EVALUATION CRITERION

### PART 1 - DATA EXPLORATION

These are guidelines for you to explore the given dataset and for understand the underlying distributions. You should explore the data by building interesting Visualizations. Try answering the following question with regards to the training dataset provided to you.

- Analyze the missing values in the columns
- What is the number of unique values and most frequently occurring categories in the Categorical Columns?
- Can you identify the top issues raised by consumers in this dataset?
- Are there some products which receive a higher number of complaints?
- Do all the companies receive the same number of complaints?
- How are the complaints submitted - through which medium?
- What is the geographical distribution of the complaints?
- Do the complaints rise in any specific month or day of the week?

- How do companies respond to the complaints? What are the most common responses received?
- Does responding to complaints in a timely manner alter the number of consumers that disputed the company response and those that did not?

**PART 2 - TEXT BASED MODELLING**

Use the consumer complaint narrative to predict the type of product the consumer identified in the complaint. For example, "Checking or savings account" or "Student loan."

- You would need to remove NAs from the Product and the Consumer Complaint Narrative column before attempting to build a Text Classifier
- You can handle columns the text in the Consumer Complaint Narrative column creatively. See if you can create some good features from this column using CountVectorizer or TF-IDF



**PART 3 - CLASSIFICATION MODELS AND FEATURE ENGINEERING**

Build a model to predict whether the consumer will dispute the resolution of the complaint or not.

- **Feature Engineering** - These are some general rules about feature engineering or the art of choosing the right features
    - Do not use date columns as is; you can use them to create other features. For example, to extract which month of the year, the complaint was filed. Was it the first week or last week of the month? What was the gap between the filing of complaints and the data being sent to the company? These are just ideas, feel free to make any other features from these
    - You can convert strings/object type columns to date_time data using pd.to_datetime
    - It does not make sense to use Consumer ID as a predictor
    - Before removing NAs from data, do check if there are columns which have too many NaN. See whether you need to impute those values or need to drop that column altogether before you start removing NA observations from the entire data
    - It does not make sense to use ZIP CODES as a numeric variable

- Consider making features for presence of NaNs itself

- **Model Training and Evaluation** - Once you have extracted the relevant features for your experiment, you can try grid search or random search available in 'sklearn' to tune hyperparameters in the model that you choose
  - Break your train data into two parts and use one to build a model and test its performance on the other. This way, you will have some idea on your approximate score you might get on your test data
  - This way, you can evaluate whether you are doing well or not, or whether your solution needs improvement or not. You can explore cross-validation to give you a more robust validation
  - If you are creating any new features on your training data or modifying features in the train; you will have to do that for test data also. This is needed so that the model which was built using the training data can be used for the test data to make predictions
  - It is a large dataset, might take a lot of time to run



## SUBMISSION GUIDELINES

Submission should be in the following formats –

- A code file with the relevant steps for each task in Python
- A CSV file with the "predictions" column only. Also, the number of rows in the submission CSV should be the same as test data.