**Project Report: Consumer Complaints Resolution**

**Introduction:**

A details consumer complaints data is provided and the target is to build a prediction model for column Consumer Disputed.

The training data consists of 478421 observations and 18 columns as mentioned below:

Date received object

Product object

Sub-product object

Issue object

Sub-issue object

Consumer complaint narrative object

Company public response object

Company object

State object

ZIP code object

Tags object

Consumer consent provided? object

Submitted via object

Date sent to company object

Company response to consumer object

Timely response? object

Consumer disputed? object

Complaint ID int64

All columns are categorical/object datatypes except for the column “Complaint ID”

**Data Preparation:**

The column **complaint id** is just a serial number or a consumer identitty. Hence, it would have no impact of model building and we should remove this column.

The number of missing values in each column is mentioned below:

Date received 0

Product 0

Sub-product 138473

Issue 0

Sub-issue 292625

Consumer complaint narrative 403327

Company public response 388029

Company 0

State 3839

ZIP code 3848

Tags 411215

Consumer consent provided? 342934

Submitted via 0

Date sent to company 0

Company response to consumer 0

Timely response? 0

Consumer disputed? 0

Complaint ID 0

The percentage of missing values in column **Tags** is 411215/478421 = 85.953%. Hence, this column should not be considered in model building.

The percentage of missing values in column **Consumer complaint narrative** is 403327/478421 = 84.303 %. Out of 75094 non-missing values, number of unique values are 74019. Hence, we can not make direct dummies for this column. However, we can create features from the text data using tfidf vectorizer.

Note:- It a good idea to delete the column having more that 50% missing values provided it does not contain crucial information.

The column **Date received** and **Date sent to company** are actually date. We can derive different features from them such as difference in dates by converting them to date objects

The new features added from **Date received** and **Date sent to company** arethe following:

**day\_diff :** difference between **Date received** and **Date sent to company** in days

**complaint received month:**  complaint was received on which month of the year

**complaint received weekday:** complaint was received on which day of the week