**Project Report: To mail or not to mail**

**Introduction:**

A details customer product usage and socio-demographics dataset is provided and the target is to build a prediction model for column V86 (mobile home policies) .

The training data consists of 5422 observations and 86 columns as mentioned below:

V1 int64

V2 int64

V3 int64

V4 int64

V5 int64

V6 int64

V7 int64

V8 int64

V9 int64

V10 int64

V11 int64

V12 int64

V13 int64

V14 int64

V15 int64

V16 int64

V17 int64

V18 int64

V19 int64

V20 int64

V21 int64

V22 int64

V23 int64

V24 int64

V25 int64

V26 int64

V27 int64

V28 int64

V29 int64

V30 int64

V31 int64

V32 int64

V33 int64

V34 int64

V35 int64

V36 int64

V37 int64

V38 int64

V39 int64

V40 int64

V41 int64

V42 int64

V43 int64

V44 int64

V45 int64

V46 int64

V47 int64

V48 int64

V49 int64

V50 int64

V51 int64

V52 int64

V53 int64

V54 int64

V55 int64

V56 int64

V57 int64

V58 int64

V59 int64

V60 int64

V61 int64

V62 int64

V63 int64

V64 int64

V65 int64

V66 int64

V67 int64

V68 int64

V69 int64

V70 int64

V71 int64

V72 int64

V73 int64

V74 int64

V75 int64

V76 int64

V77 int64

V78 int64

V79 int64

V80 int64

V81 int64

V82 int64

V83 int64

V84 int64

V85 int64

V86 int64

All columns are numerical datatypes

**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