Mini Project 2 – Consumer Complaint Resolution Analysis Using Python

Scenario: Product review is the most basic function/factor in resolving customer issues and increasing the sales growth of any product. We can understand their mindset toward our service without asking each customer.

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

So Here, we will analyze data, and with the help of different algorithms, we are finding the best classification of customer category so that we can predict our test data.

Objective: Use Python libraries such as Pandas for data operations, Seaborn and Matplotlib for data visualization and EDA tasks, Sklearn for model building and performance visualization, and based on the best model, make a prediction for the test file and save the output.

The main objective is to predict whether our customer is disputed or not with the help of given data.

Dataset description:

Customers faced some issues and tried to report their problems to customer care.

Dispute: This is our target variable based on train data; we have two groups, one with a dispute with the bank and another don't have any issue with the bank.

Date received: The day complaint was received.

Product: different products offered by the bank (credit cards, debit cards, different types of transaction methods, accounts, locker services, and money-related)

Sub-product: loan, insurance, other mortgage options

Issue: Complaint of customers

Company public response: Company's response to consumer complaint

Company: Company name

State: State where the customer lives (different state of USA)

ZIP code: Where the customer lives

Submitted via: Register complaints via different platforms (online web, phone, referral, fax, post mail)

Date sent to company: The day complaint was registered

Timely response?: Yes/no

Consumer disputed?: yes/no (target variable)

Complaint ID: unique to each consumer

Tasks to be performed:

The following tasks are to be performed:

Note: Please complete the given steps on both train and test data.

- Read the Data from the Given excel file.
- Check the data type for both data (test file and train file)
- Do missing value analysis and drop columns where more than 25% of data are missing
- Extracting Day, Month, and Year from Date Received Column and create new fields for a month, year, and day
- Calculate the Number of Days the Complaint was with the Company and create a new field as “Days held”
- Drop "Date Received","Date Sent to Company","ZIP Code", "Complaint ID" fields
- Imputing Null value in “State” by Mode
- with the help of the days we calculated above, create a new field 'Week_Received' where we calculate the week based on the day of receiving.
- store data of disputed people into the “disputed_cons” variable for future tasks
- Plot bar graph of the total no of disputes of consumers with the help of seaborn
- Plot bar graph of the total no of disputes products-wise with the help of seaborn
- Plot bar graph of the total no of disputes with Top Issues by Highest Disputes, with the help of seaborn
- Plot bar graph of the total no of disputes by State with Maximum Disputes
- Plot bar graph of the total no of disputes Submitted Via different source
- Plot bar graph of the total no of disputes where the Company's Response to the Complaints
- Plot bar graph of the total no of disputes where the Company's Response Leads to Disputes
- Plot bar graph of the total no of disputes. Whether there are Disputes Instead of Timely Response
- Plot bar graph of the total no of disputes over Year Wise Complaints
- Plot bar graph of the total no of disputes over Year Wise Disputes
- Plot bar graph of Top Companies with Highest Complaints
- Converte all negative days held to zero (it is the time taken by the authority that can't be negative)

- Drop Unnecessary Columns for the Model Building
  like:'Company', 'State', 'Year_Received', 'Days_held'
- Change Consumer Disputed Column to 0 and 1(yes to 1, and no to 0)
- Create Dummy Variables for categorical features and concat with the original data frame like: 'Product,' 'Submitted via,' 'Company response to consumer,' 'Timely response?'
- Scaling the Data Sets (note: discard dependent variable before doing standardization) and Make feature Selection with the help of PCA up to 80% of the information.

- Splitting the Data Sets Into X and Y by the dependent and independent variables (data selected by PCA)
- Build given models and measure their test and validation accuracy:
  - LogisticRegression
  - DecisionTreeClassifier
  - RandomForestClassifier
  - AdaBoostClassifier
  - GradientBoostingClassifier
  - KNeighborsClassifier
  - XGBClassifier
- Whoever gives the most accurate result uses it and predicts the outcome for the test file and fills its dispute column so the business team can take some action accordingly.

Note: You have been provided with an IPYNB file to perform the above tasks

Here, some of the commands have already been provided to you. These commands are related to text pre-processing which comes under Natural Language Processing that we will be covering in the upcomimg classes.