Consumer Complaints Resolution

Consumer complaint resolution is important to any business. In this particular case we have been given detailed consumer complaints along with whether consumer disputed with the conclusion. If we are able to predict this, consumer who is more likely to dispute a conclusion can be given more attention as to how the complaints are handled as well as how persuasively the final conlusions are conveyed to them.

Your target here is to build prediction model for column "Consumer disputed".

Data Files

Training Data = Consumer_Complaints_train.csv

Test Data = Consumer_Complaints_test.csv

Formal Problem Statement ¶

Your target here is to build prediction model so that you are able to predict which consumer is more likely to dispute the resolution of a complaint i.e. you need to make predictions for the "Consumer disputed" column.

All the column names are self explanatory. You need to build your model on train data. Test data does not have the response column "Consumer disputed", you need to predict those values and submit it in a csv format.

Evaluation Criterion

Part 1:

You will first attempt Part 1 of this project which is a quiz. You can access it through LMS. This quiz needs to be answered based on exploration of the dataset given and some generic questions about algorithms discussed in the course. Consider only the training dataset for data cleaning and exploration to answer the quiz questions. There will be 10 questions of which you need to get at least 7 correct in order to pass the project.

Part 2:

Here you work on creating the machine learning models and choosing the one which gives the best performance. You can refer to the Project Process Guides provided in LMS to understand how to approach and work on a project.

In order to get a passing grade in this project you need to have the AUC score of at least 0.54 for your test data predictions.

Submission:

Submission CSV should resemble the file provided in the LMS:

Sample Submission = 'sample_submission.csv'

Column names and value types should match. Also number of rows in the submission csv should be exactly the same as test data. If this is not taken care of, your submission will not be graded.

General Guidelines for the project:

- 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 first week or last week of the month. What was the gap between between 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.
- You can handle columns Issue & Consumer Complaint Narrative, creatively. See if you can create some good feature from this column containing text data. [tfidf ?]
- It doesnt make sense to use Consumer ID as predictor.
- Break your train data into two parts and use one to build model and test its performance on the other. This way you will have some idea on your approximate score you might get on your submission. Otherwise you'll have to wait for few days post submission without knowing whether you are going to do well or not; or whether your solution needs improvement or not.
- 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 obs from the entire data.
- 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 doesnt make sense to use ZIP CODES as a numeric variable.
- It is a large dataset, might take a lot of time to run.
- You can discuss anything on QA forum. Although threads which explicitly disclose too much information will be removed from the forum.
- Consider making features for presence of NaNs itself.

We have also uploaded a benchmark script, it gives you auc score on test data, slightly less than what is required to pass the course. You can include your ideas to make better predictions and make submissions. You can make as many submissions you want if you want. [We might ask you to submit the script which was used to generate the submission at any time].

Once you have passed the required criterion for csv submission; for recieving certification from us; you'll need to submit a project report [soft copy] not exceeding 4 pages.

In order to clear this project, you are required to clear both, Part 1 as well as Part 2 of this assignment.

Wish you all the best!