GENERAL MOTORS

BlacktopBuild @ UCal Berkeley

Part 1: Customer Feedback Verbatim Classification Competition

Overview

Complexity surrounding the holistic nature of customer experience has made measuring and understanding customer perceptions of automotive infotainment experiences extremely challenging. At the same time, the increasing volumes of unstructured textual data generated from customer feedback makes it even more difficult for GM business groups to analyze and interpret this information. The ability to rapidly assess customer experience feedback is critical for successful product deployment.

Text mining, a method enabling automatic extraction of information from textual data, is becoming more and more popular. Today, interpreting customer feedback requires significant engineering resources – as user-facing software in vehicles has grown exponentially in complexity, the teams and effort required to parse the feedback has become large. A flexible and scalable model is necessary to adapt to changing customer terminology and sentiment about the wide array of features offered in vehicles today.

In this competition, you are challenged to build a machine learning model that's capable of classifying customer feedbacks into pre-defined categories – you will pull on not only structured data sets that we provide, but also be challenged to find meaningful customer feedback from anywhere on the internet. The model will be verified against the pre-determined set of data to measure accuracy.

Data

You are provided with three sets of sampled customer feedbacks, which have been categorized by engineers into problem **areas** and **sub areas**.

You are free to scrape and use any customer feedback you find on the internet – just be sure to assess how representative it is of general customer body sentiment.

You must categorize and predict each customer feedback into these pre-defined area and sub areas using the model created, and finally come up with combined top 10 customer complaints.

File descriptions (3 sets)

1. **Train.csv**
   The training set contains customer verbatims, other structured data fields which may be used as inputs, and target area/sub area each one belongs. The variable to model for this exercise is labeled "**area"** and **"subarea**". These 3 files will have a prefix "source_1_", etc.

2. **Test.csv**
   The test dataset will be used to verify the model accuracy.
   These datasets will be provided late in the BlacktopBuild to prevent any data leaks, and after program submission.

Supplemental Data

ontology.csv

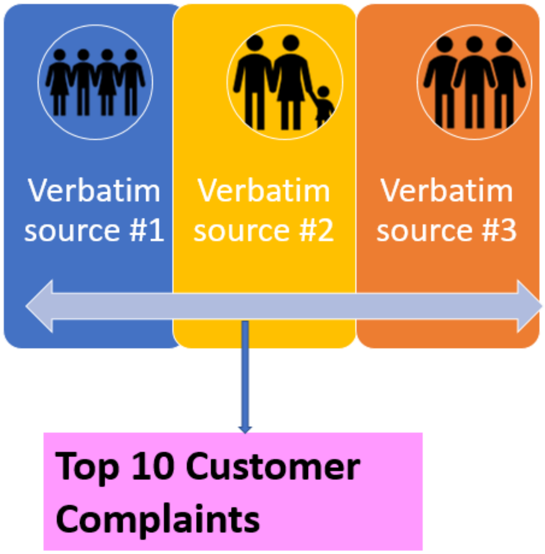
In addition to data sources online, we are providing an ontology, to help normalize the text data into more broad categories. This can be used to group test together into more standard language, which could aid in more accurate machine learning / deep learning models. The ontology consists of two levels, S1, and SY.

You are encouraged to come up with your own ontology, and apply to your model to increase model accuracy.

Submission Material

1. PowerPoint presentation to explain your approach to the problem and solution, including robustness, scalability, and flexibility.
2. Predictions against test datasets. Submission will include all fields in the test dataset plus prediction Area and Sub Area, and must include the original ID from the test.csv files.
3. All code used to create models. Code must include all libraries clearly listed at top of script.
4. Final trained model object files used to generate submission predictions.
5. **Prototype user interfaces (UI) using applications such as Bokeh, Dash, Django, RShiny, etc. that can ingest new text data and classify it into 'Area' and 'SubArea' automatically will be an added benefit when judging team performance.**

** All predictions will have similar weight in final grading. Source #1 has the most data volume, so will have the most weight.



Part 2: Discover relationship between customer complaints & defects

Overview

With the growth of complexity in modern automotive infotainment system, infotainment software has become very sophisticated. This design complexity leads to various challenges in software testing and identifying defects. Some of these are truly defects in that the requirements were not properly implemented; some are caused by changes made to other related systems; still others are requests for enhancement – improvements that would improve overall user experience. These defects are generally stored in a database and are resolved in a series of incrementally delivered updates.

Due to enormous volume of raised defects, it is not feasible to fix all the concerns within the confines of launching a product quickly, and indeed some issues may never be addressed. Today, defect prioritization is done by highly skilled engineers with significant experience in the field – however, this process is inefficient and leads to an overlap between true customer sentiment and personal experiences. This tribal knowledge is difficult to transfer to new engineers as more experienced engineers progress in their careers. It can take years for a new engineer to become proficient at the defect prioritization process.

In this competition, you are challenged to build a model that's capable of discovering engineering defects and the identifying top customer complaints. **Please note that not all defects are related to these complaints.** The model should indicate how they are related and to what degree are they related.

Data

You are provided with some sample engineering defects and the list of customer complaints. Definition of these defects will be provided. You are to discover the relationship between these customer complaints and engineering defects.

File descriptions

1. **Customer Complaints (10)** (Customer Complaints Definitions and Examples.docx)
   Each customer complaint has a detailed description, sample customer verbatim, and engineering defects that match the complaint.

2. **Engineering Defects (~ 1k record)** (BlackTop-Defects-Student.xlsx)
   Sample engineering defects that are potentially related to customer complaints.

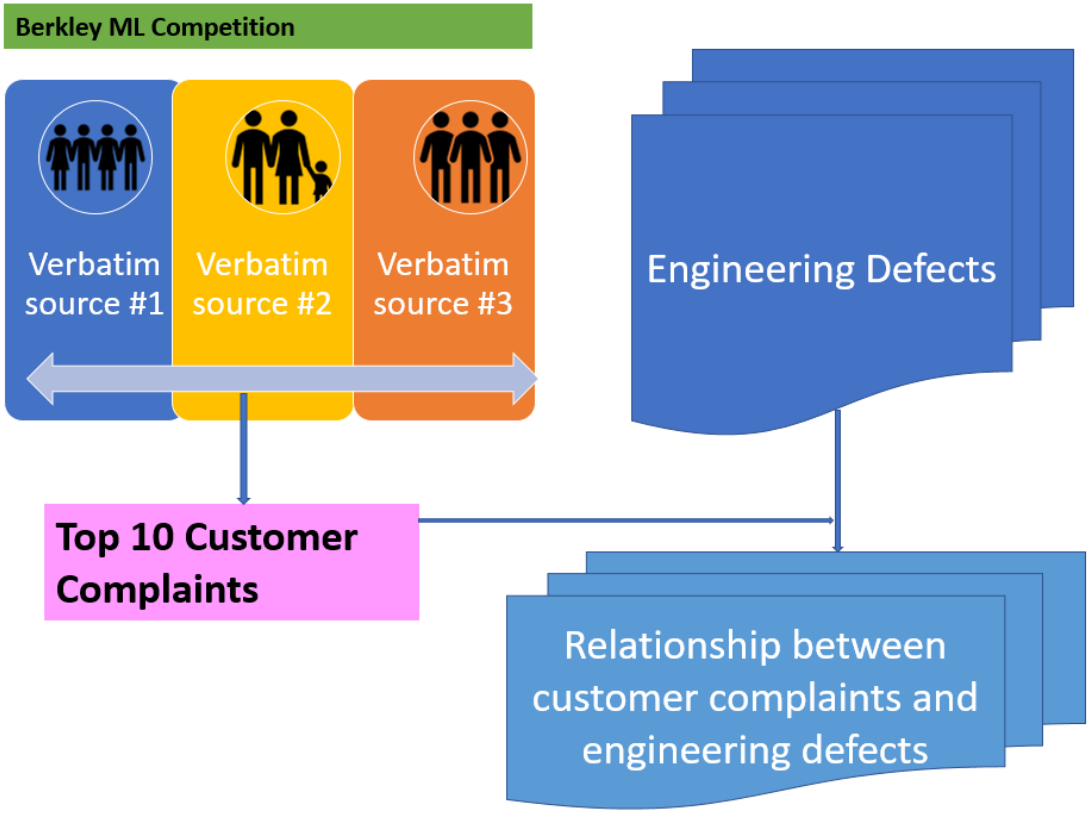
Submission Material

1. PowerPoint presentation to explain your solution, including robustness, scalability, and flexibility.
2. Prediction against test dataset. Submission will include all fields in the test dataset plus predicted area of correlation with customer complaints.
3. All code used to create models. Code must include all libraries clearly listed at top of script.
4. Final trained model object files used to generate submission predictions.

Assessment of Submission Material

In Part 1, Data source 1 prediction accuracy, assessed using F1 score, is the most important among the three data sources. Exact weightings for evaluation criteria will not be disclosed to student teams.

Overall, Part 2 will be weighted much more heavily than Part 1 when assessing team performance in the competition.


Berkley ML Competition
Verbatim source #1
Verbatim source #2
Verbatim source #3
Engineering Defects
Top 10 Customer Complaints
Relationship between customer complaints and engineering defects