**Capstone 2 Project Ideas**

1. **Elo Merchant Category Recommendation**

Help understand customer loyalty and build a recommendation engine with discount from credit card provider.

Data Link - <https://www.kaggle.com/c/elo-merchant-category-recommendation#description>

train.csv and test.csv files - These contain the card\_ids that will be used for training and prediction.

The historical\_transactions.csv and new\_merchant\_transactions.csv files contain information about each card's transactions.historical\_transactions.csv contains up to 3 months' worth of transactions for every card at any of the provided merchant\_ids.new\_merchant\_transactions.csv contains the transactions at *new* merchants (merchant\_ids that this particular card\_id has not yet visited) over a period of two months.

merchants.csv contains aggregate information for each merchant\_id represented in the data set.

**The data is formatted as follows:**

train.csv and test.csv contain card\_ids and information about the card itself - the first month the card was active, etc. train.csv also contains the target.

historical\_transactions.csv and new\_merchant\_transactions.csv are designed to be joined with train.csv, test.csv, and merchants.csv. They contain information about transactions for each card, as described above.

merchants can be joined with the transaction sets to provide additional merchant-level information.

**File descriptions**

* train.csv - the training set
* test.csv - the test set
* sample\_submission.csv - a sample submission file in the correct format - contains all card\_ids you are expected to predict for.
* historical\_transactions.csv - up to 3 months' worth of historical transactions for each card\_id
* merchants.csv - additional information about all merchants / merchant\_ids in the dataset.
* new\_merchant\_transactions.csv - two months' worth of data for each card\_id containing ALL purchases that card\_id made at merchant\_ids that were *not visited in the historical data*.

**Data fields**

Data field descriptions are provided in Data Dictionary.xlsx.

1. **Quora Insincere questions classification**

**Link -** [**https://www.kaggle.com/c/quora-insincere-questions-classification/data**](https://www.kaggle.com/c/quora-insincere-questions-classification/data)

An existential problem for any major website today is how to handle toxic and divisive content. Quora wants to tackle this problem head-on to keep their platform a place where users can feel safe sharing their knowledge with the world.

An insincere question is defined as a question intended to make a statement rather than look for helpful answers. Some characteristics that can signify that a question is insincere:

* Has a non-neutral tone
  + Has an exaggerated tone to underscore a point about a group of people
  + Is rhetorical and meant to imply a statement about a group of people
* Is disparaging or inflammatory
  + Suggests a discriminatory idea against a protected class of people, or seeks confirmation of a stereotype
  + Makes disparaging attacks/insults against a specific person or group of people
  + Based on an outlandish premise about a group of people
  + Disparages against a characteristic that is not fixable and not measurable
* Isn't grounded in reality
  + Based on false information, or contains absurd assumptions
* Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

The training data includes the question that was asked, and whether it was identified as insincere (target = 1). The ground-truth labels contain some amount of noise: they are not guaranteed to be perfect.

Note that the distribution of questions in the dataset should not be taken to be representative of the distribution of questions asked on Quora. This is, in part, because of the combination of sampling procedures and sanitization measures that have been applied to the final dataset.

**File descriptions**

* **train.csv** - the training set
* **test.csv** - the test set
* **sample\_submission.csv** - A sample submission in the correct format
* **enbeddings/** - (see below)

**Data fields**

* **qid** - unique question identifier
* **question\_text** - Quora question text
* **target** - a question labeled "insincere" has a value of 1, otherwise 0

This is a Kernels-only competition. The files in this Data section are downloadable for reference in Stage 1. Stage 2 files will only be available in Kernels and not available for download.

## What will be available in the 2nd stage of the competition?

In the second stage of the competition, we will re-run your selected Kernels. The following files will be swapped with new data:

* test.csv - This will be swapped with the complete public and private test dataset. This file will have ~56k rows in stage 1 and ~376k rows in stage 2. The public leaderboard data remains the same for both versions. The file name will be the same (both test.csv) to ensure that your code will run.
* sample\_submission.csv - similar to test.csv, this will be changed from ~56k in stage 1 to ~376k rows in stage 2. The file name will remain the same.

**Embeddings**

External data sources are not allowed for this competition. We are, though, providing a number of word embeddings along with the dataset that can be used in the models. These are as follows:

* **GoogleNews-vectors-negative300** - <https://code.google.com/archive/p/word2vec/>
* **glove.840B.300d** - <https://nlp.stanford.edu/projects/glove/>
* **paragram\_300\_sl999** - <https://cogcomp.org/page/resource_view/106>
* **wiki-news-300d-1M** - <https://fasttext.cc/docs/en/english-vectors.html>

1. **Uber data**

<https://data.world/login?next=%2Frmiller107%2Ftravel-time-uber-movement%2Fworkspace%2Fproject-summary>

The data contains the average travel time calculations between two ‘zones’ by time and date variables. Build a model to make predictions about travel time and high congestion zones by leveraging some of the attributes in the data such as the pickup location, destination, time of day, and average travel time.