Eluvio Data Science Challenge

Fumika Isono

Date: Sunday, April 25, 2021

Objective

We set the goal of the project to be: **Predicting the number of upvotes based on the given variables.**

Summary

We built a model to estimate up votes of the subreddit worldnews in Reddit.

- We used **google colab** to code and load the provided large dataset. Although it did not save significant time, we made use of the access to **GPU** for training the model.
- We made use of various python time/calendar modules and **NLP** modules to convert the variable 'time_created' and 'title' to features useful for predicting the up votes.
- We used **Gradient Boosted Trees** as a model. Hyper-parameters were chosen based on the grid search and **cross-validation**.
- We correctly estimated the upvotes within +/- 50 votes with **87.2% accuracy with gradient boosted tree model**, whereas **with the global mean(baseline), only 4.3%** fall within +/- 50 votes of the true upvotes.

Methodology

1. We first analyzed the data, and made sure that the appropriate data types were correctly loaded. We then made sure we understood the dataset correctly, which turned out to be the data of posts from Reddit, worldnews subreddit. Because the dataset does not have too many features, we decided to analyze the features one by one and see how we could engineer them to derive meaningful features.This provided us more intuition to understand the data set and how important the features are in estimating upvotes.

2. We then engineered relevant features. -e.g., time features in 4 different spans (day, week, year, all-time), holidays, in the US central Time zone. We converted the time variable to the US central Time zone, based on the Reddit users and the population distribution in each time zone. We made use of Sentiment Intensity Analyser, readability scores, document vectorization to perform NLP on 'title' variables.

3. We used mean absolute value as a loss function since the distribution of upvotes is heavily skewed.

4. We split the dataset to 80% training set, 10% validation set, and 10% test set. We never used the test set for training. We used cross validation when tuning hyper parameters with the training set, then used validation set for the final training.

5. We used two models for predicting the up votes:
    a. Global mean, as a baseline
    b. Gradient Boosted Trees using the MAE loss function.

6. They yielded the following results

	Training MAE	Validation MAE	Test MAE
Model 1: Global Mean	186.16	189.03	172.36
Model 2: Gradient Boosted DT	**110.49**	**112.22**	**102.99**

The Gradient Boosted decision tree model successfully performed well, as expected.

7. We can also provide a more business centric view of the performance of our mode:

	Percentage of prediction within +/- 50 votes accuracy (test set)	Percentage of prediction within +/- 100 votes accuracy (test set)
Model 1: Global Mean	**87.23 %**	**90.56 %**
Model 2: Gradient Boosted DT	4.27 %	21.60 %

We see that we successfully increased the percentage of prediction within specific ranges of error.

8. A study of the importance of the features showed that the most important features are the authors, number of days elapsed, and the length of the title.

Many more details can be found in the google colab.