Homework: Twitter Analytics

Background

You will act as a data analyst/scientist in this HW. A social media data analytics project involves several steps or phases: business understanding, data collection & understanding, data (text) preprocessing (feature engineering), descriptive analytics, model building (e.g., classification, topic modeling), and storytelling. You have learned what these steps are and what activities you need to be involved during each step or phase. Now, it is your turn to apply your knowledge/skills into this HW.

This HW is designed to help you master the concepts and techniques for Twitter Analytics, a set of computational techniques to extract intelligence from Twitter data. You're reminded that Twitter data, among social media data, is most popularly used in industry applications (e.g., marketing, supply chain, HR, business strategy, journalism, emergency planning, homeland security).

In this HW, you will (1) collect Twitter data ("big data") using movie hashtags, (2) extract valuable information[1] (e.g., # of tweets, # of retweets, # of users) from Twitter data, and (3) build a multiple regression model to predict movie sales (or revenues).

This HW is semi-open ended. This is an individual HW. As a data analyst/scientist, you need to be creative and eager to find something valuable from Twitter data and develop a predictive (regression) model to predict future sales.

I have developed this HW based on the idea presented in this article (https://www.fastcompany.com/1604125/twitter-predicts-box-office-sales-better-prediction-market-updated) This article explains what is needed to build a predictive model (regression model in this case because you're trying to predict sales) to predict movie sales with Twitter data. The articles is noting:

"*Then they took two different approaches, dealing with two very different performance metrics: the first weekend performance, which is largely built on buzz and the second weekend performance, which is largely built whether people actually like the movie.*

To predict first weekend performance, they built a computer model, which factored in two variables: the rate of tweets around the release date and the number of theaters its released in. Lo and behold, that model was 97.3% accurate in predicting opening weekend box office. By

[1] More variables (e.g., # of tweets per movie, # of users per movie) are better for regression model, meaning you should extract as many variables as possible from Twitter data.

contrast, the Hollywood Stock Exchange, which has been the gold standard for opening box-office predictions, had a 96.5% accuracy.

Meanwhile, to predict second-weekend performance, the authors created a ratio of positive tweets to negative ones. Then they blended that with the Tweet rate metric in another prediction algorithm. This time, the method was 94% accurate."

This HW has two parts:

Part I - you need to predict first week performance (or sales) only: domestic box office sales. This first week performance prediction is the first part of this HW.

Part II - The second part of this HW is predicting second-week performance.

The following information is useful for your HW:

1. You can get the list of movies coming soon from the following sites (there could be other websites from which you can find this information):
    a. https://www.rottentomatoes.com/browse/upcoming/
    b. http://www.imdb.com/movies-coming-soon/
    c. https://www.movieinsider.com/movies/october/2017

2. You can get movie sales data from the following site (there could be other websites from which you can find movie sales data):
    a. http://www.the-numbers.com/box-office-chart/weekend/2017/10/06

3. To my knowledge (I may be wrong on this!), new movies are released on every Friday (e.g., Oct 6, 2017). For example, "Blade Runner 2049" was released on Oct 6 2017 and the first-day performance data can be found from http://www.the-numbers.com/movie/Blade-Runner-2049#tab=box-office Then, to predict first week performance, you would start data collection from about a week earlier (Sept 29, 2017).

4. When you collect movie tweets, I would use movie hashtags. For example, #bladerunner2048 is the hashtag for "Blade Runner 2049". I think there could be more than one hashtag for a movie used by moviegoers. To build a regression model, you need a minimum of 30 samples (or Twitter data about 30 movies or more).

5. You would use English tweets only in the analysis.

6. For this data collection, you should use your Ubuntu server (t2.micro). You don't need a fast server, but your server needs a large volume (I would say 30GB) to store Twitter data. In the AWS module, we learned how to run Python script on AWS for a long period

of time.

I suggest the following timeline for data collection:

For HW Part I

- First, find out the movies to be released on Oct 20, 2017.
- Second, start Twitter data collection from Oct 13, 2017 until Oct 19, 2017. After data collection, you need to extract information, which can be used as dependent variables (or X variables) in your regression model.
- Third, collect the sales data for Oct 20, 2017. This sales data will be the dependent variable (or y value) in your regression model.
- HW due date: Oct 26

For HW Part II

- Start the second round of Twitter data collection from Oct 20, 2017 until Oct 26, 2017 using the same hashtags you used for the first round of data collection. After data collection, you need to extract information, which can be used as dependent variables (or X variables) in your regression model. This list of dependent variables must be same as those used in your regression model for HW Part I.
- Collect the sales data for Oct 27, 2017. This sales data will be the dependent variable (or y value) that your regression model needs to predict.
- HW due date: Nov 3

Regression model:

- I expect you have a good understanding of regression analysis and how to build a regression model.
- For this HW, you need to use Python (scikit-learn and/or statsmodels python package for machine learning) to build regression model. If you're new to regression analysis using Pytion, you can watch my video lecture and review jupyter notebooks on this topic.
- Part I
    - You build (or train) a multiple regression model. Hopefully, a good one!
- Part II
    - You will apply your regression model to predict second-week movie sales.

Requirements

1. Create the Ipython notebook (named as "yourlastname_firstinitial_ HW_TwitterAnalytics_PredictingMovieSales.ipynb") and complete the **tasks** below. **Each cell must be properly numbered and formatted using Markdown**.
2. For any questions regarding Python coding, you should consult lecture notes (Jupyter Notebooks) and https://stackoverflow.com/.

3. Your Jupyter Notebook needs to be properly formatted using Markdown and comment.

Required procedures & task

You are required to build a regression model to predict future movie sales using Twitter data.

A. HW Part I

1. First, import your dataset(s) and provide descriptive statistics of your Twitter data. You need to show all your works in the jupyter notebook to answer the below questions.
    a. Entire dataset
        i. How many movies are in the dataset?
        ii. What hashtags are used in data collection?
        iii. How many hashtags are used in data collection?
        iv. How many tweets in your dataset?
        v. How many unique users in your dataset?
        vi. What is the ratio of original tweets and retweets?
    b. Per movie
        i. How many tweets per movie?
        ii. How many unique users per movie?
        iii. What are top five movies in terms of # of tweets?
        iv. What are top five movies in term of # of unique users?
        v. What are top five movies in terms of movie sales?
2. Second, you need to demonstrate your skills in regression analysis on Twitter data and movie sales information.
    a. I strongly encourage you to build multiple regression models using different regression algorithms (e.g., ordinary least square, lasso, svm, randomforestregressor)
    b. Report the performance of your regression models in terms of variance score (R-squared) and mean square error (MSE)
    c. Explain your best model in terms of variance score and mean square error.
    d. What variables appear to be important in terms of predicting movie sales?
3. Third, demonstrate your skills in "storytelling". To me (and your client), this is the most important part of this HW.
    a. Explain your best model in terms of variance score and mean square error.
    b. What variables appear to be important in terms of predicting movie sales?
    c. How confident are you about your regression model?
    d. Any suggestions to improve the model quality?
    e. Any potential areas (industry applications) Twitter data can be used to predict future?

B. HW Part II

1. First, import your datand provide descriptive statistics of your Twitter data. You need to show all your works in the jupyter notebook to answer the below questions.
    a. Entire dataset
        i. How many movies are in the dataset?
        ii. How many tweets in your dataset?
        iii. How many unique users in your dataset?
        iv. What is the ratio of original tweets and retweets?
    b. Per movie
        i. How many tweets per movie?
        ii. How many unique users per movie?
        iii. What are top five movies in terms of # of tweets?
        iv. What are top five movies in term of # of unique users?
        v. What are top five movies in terms of movie sales?
    c. Compare first-week tweets (# of tweets) and second-week tweets (# of tweets) per movie
        i. Report the growth (or decline) of tweets in percentage per movie
        ii. Report top five movies in terms of growth
2. Second, you need to predict second-week performance (sales) per movie. This stage is called "model deployment" in data analytics, which is the final stage of a typical data analytics project.
    a. Apply your best regression model and predict second-week sales per each movie
    b. Report the results (predicted sales from your regression model)
    a. Report the actual second-week sales per movie (this information can be acquired from http://www.the-numbers.com/box-office-chart/weekend/2017/10/06
    c. Report mean square error of your prediction. Easy to calculate this value:
    MSE = | actual second-week sales – predictive second-week sales | / # of movies

    This value (MSE) will be used to evaluate the overall performance of your regression model and select the best regression model in this class !!!

How to Format Your Jupyter Notebook:

- Start with K-State Honor Code "**On my honor, as a student, I have neither given nor received unauthorized aid on this academic work**."
- Include the questions (tasks) and the question numbers, using Markdown, prior to each cell containing python codes.
- Must be professional and neat

Submission

HW Part I

- Complete Ipython notebook in **HTML** version (yourlastname_firstinitial_ HW_TwitterAnalytics_PartI.html).
- If you wish, you can submit two HTML files (one for descriptive analytics and the other one for regression). If this is the case, name them "yourlastname_firstinitial_ HW_TwitterAnalytics_PartI_descriptiveanalytics.html" and "yourlastname_firstinitial_ HW_TwitterAnalytics_PartI_regressionmodel_and_storytelling.html" respectively.

HW Part II

- Complete Ipython notebook in **HTML** version (yourlastname_firstinitial_ HW_TwitterAnalytics_PartII_ModelDeployment.html).