IBM Data Science Professional Certificate Capstone - Predicting Venue Like Count Using Natural Language Processing of Venue "Tip" Text

Andrew Barber
github.com/acbarber

Abstract

With such a large quantity of textual data on the internet today, there is an enormous opportunity for exploration and analysis. However, much of this data lacks the necessary context to conduct such research, making "labelled data", or data with an external measure of validation, extremely valuable. In this paper we attempt to leverage labelled textual data extracted from the Foursquare API (https://developer.foursquare.com/), making a model to predict venue like counts based on existing venue "tips". We found that our model successfully outperformed the baseline average like count metric, recording a mean squared error (MSE) of 10.99 compared to the baseline's MSE of 12.57.

1 Introduction

In many cases data is limited in scope and format. This is especially true now, with an abundance of unstructured data in the form of natural language - it can be hard to transform and extract meaning from the data in a valuable way. Labelled natural language processing (NLP) data presents a unique opportunity to extract meaning from natural language and hopefully extrapolate these findings to a wider category of unlabelled data. In the case of this paper, we have examined venue data from the Foursquare API including venue "tips" (reviews of the venue) and venue like count.

Our data was taken from venues scattered around the 40 major neighborhoods of Manhattan where we attempted to predict venue like counts based on the tips left for that venue. This NLP prediction task is used for user recommendation systems and may be useful for predicting the popularity of a venue based on unlabelled comments where another measure of popularity does not exist. It may also be used as a supporting method for venue popularity evaluation when there are other methods of popularity such as like count or rating but the quantity of NLP data is far more extensive. If my model can successfully predict like counts in Manhattan venues, it could then be applied to popularity evaluation applications such as listed above or represent a viable base for future research.

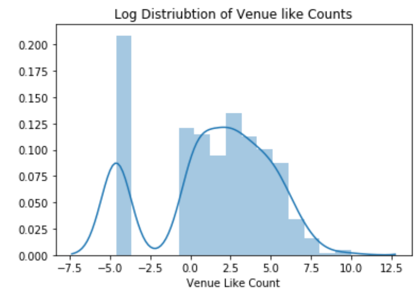

Figure 1: Log Distribution of Venue Like Counts

2 Problem Definition and Data

As mentioned above, we used the Foursquare API for this paper. Specifically we looked at "tips" for various venues scattered across the 40 neighborhoods of Manhattan. Our problem addressed how to leverage this "tips" text in a way that could accurately predict a venue's like count. In order to retrieve a good mix of venues, we pulled the twelve nearest venues to each of the Manhattan neighborhoods. We then extracted all "tips" from the venues and concatenated them together to form the tip text for each venue. Rows with "N/A" values in either the "Venue Like Count" or "Venue Tip Text" fields were dropped. The final number of records was 459.

The challenge with this dataset came with the large number of zeros for like counts. In order to produce a model that wouldn't automatically predict 0 for all venues, putting the target field on

Figure 2: Word Cloud for Venue Tip Text Vocabulary

a log-scale was necessary. In order to avoid log values of "-infinity" when transforming the like counts, we also implemented a smoothing technique where a a value of 0.01 was added to all like counts. The resulting bimodal log-normal distribution can be seen in Figure 1 above where one mode represents the ln(0) and the other at around ln(12).

When looking at the textual features of the dataset, the most common words in the venue "tips" can be seen in figure 2 below. These words seem predominantly positive and represent an interesting finding regarding this dataset and perhaps others like it: if the features of the text reflect similar sentiments, then how can we best leverage the text to differentiate between venues that would receive a large number of likes and venues that would not? These features make sentiment analysis especially difficult when examining such an approach as a future direction for this research.

For the purposes of this paper, all textual data was stripped of punctuation and made lowercase. We also utilized NLTK's word-tokenizer function (https://www.nltk.org/api/nltk.tokenize.html) in order to separate our text into tokens. This allowed us to transform our text into features which we then used in our model. We will go into greater detail regarding our methodology in the next section.

3 Methodology

In order to accurately represent the textual information in our dataset, we decided to use the bag-of-words (BOW) model. In this method, sentences are represented by an array of length n, where n is the length of a vocabulary (in this case all unique tokens for the venue tips in the dataset). The count of individual words in the sentence are then placed at the words corresponding index in the array. In this way, sentences can be represented by a vector in an n-dimensional space.

Given that our problem is a regression problem, the simplest and most accessible model to implement would be a linear regression model. We decided that this would be a good starting point for future research on this, and similar datasets. To evaluate this model we determined the average like counts across all venues in the dataset. A model that outperforms this baseline will then be more accurate at predicting venue like counts than simply taking a typical number of likes for a given venue.

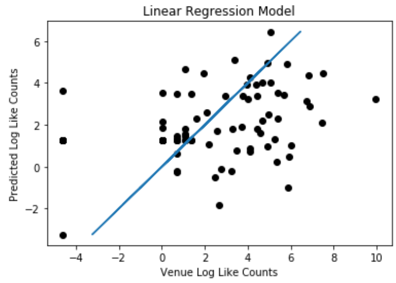

Figure 3: Linear Regression Predicted Values Versus True Values

4 Evaluation and Results

When evaluating our results, we determined the MSE between the true venue log like counts and the predicted log like counts for both of our models. Referring to Table 1 below, you will see that our model outperformed the baseline using this metric. Figure 3 above represents a more in-depth representation of our results, showing how our model's predictions compared to the true values. This result indicates that the linear regression model produces less error when predicting the venue log like counts than the baseline model.

5 Discussion

Our successful results represent a significant proof-of-concept regarding this dataset. Given a simple textual feature-extraction method (BOW) and a simple linear regression model, we can predict our target field of venue log like counts with less mean squared error than simply looking at the average of the venue log like counts. Future research can then use our model as a baseline to test different regression models.

One of the main limitations of this research, however, comes from the dataset itself. With a final record count of only 459 venues, statistical analysis and model evaluation is not as powerful as with larger datasets. Another limitation lies in

Model	MSE
Baseline - Mean Venue Log Like Count	12.53
Linear Regression (BOW)	11.00

Table 1: Evalutation of models

the narrow scope of the data. Looking at Manhattan venues exclusively limits the researcher's ability to make generalizations regarding the results of their research. Expanding the scope of the data will then further increase the impact of the results.

Future models created for this task could vary in numerous ways. For one, different text-representation techniques could be used such as term-frequency independent-document-frequency (TF-IDF), word embeddings or pre-trained word embeddings. These NLP methods may produce better results than our BOW model. Different regression models may also be used to increase performance. Specifically, neural networks could be leveraged in the form of recurrent neural networks (RNNs) such as long-short-term models (LSTMs) or convolutional neural networks (CNNs).

6 Conclusion

Overall, we found that our research represents a promising jumping-off point for future discoveries in the context of our problem. After utilizing the "tip" text of venues scattered throughout the 40 neighborhoods of Manhattan we found that we could more accurately predict the like count these venues received than simply assuming the like count for each venue was the average of the like counts for all venues. This discovery can now be used as a proof-of-concept for popularity-prediction in a variety of other datasets.