USER INDUSTRY CLASSIFICATION

7/6/2021

PROJECTING ANALYTICAL VIEWS FOR ENERGY TRANSITION PRODUCTS TO DETERMINE PROPER CLASSIFICATION OF COMPANY INDUSTRY TO CORRECT 'OTHER' CATEGORY

USER INDUSTRY CLASSIFICATION

PROJECTING ANALYTICAL VIEWS FOR ENERGY TRANSITION PRODUCTS TO DETERMINE PROPER CLASSIFICATION OF COMPANY INDUSTRY TO CORRECT 'OTHER' CATEGORY

PURPOSE

Energy Transition has been a central topic in the news for the last 6 months and many companies are beginning to see the change in culture and policies and preparing for the position themselves properly. A certain energy data company would like to know how they can improve their data collection process and retroactively correct thousands of data points for companies. These data points classify a company's industry as "other" rather than their actual industry. The goal of this project will be to use the company views to cluster the company in the proper industry type. This will allow for improved analytics and future targeting of company to solution products.

For better reference, please see Figure 1 that shows what percentage of the collected data is mislabeled as 'other' for industry.

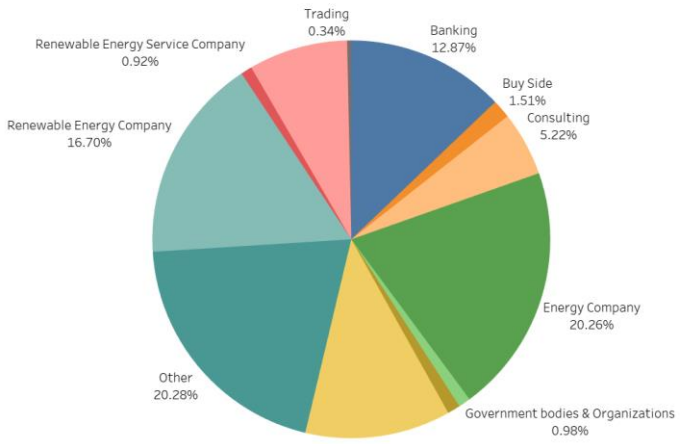

Figure 1 Analytical Views by Industry (Data created in Tableau)

Data Wrangling

The user usage data was queried from a user utility database and exported into a CSV file. It encompasses the last 3 years. Analytical views were utilized to both show the need for the determination of these industries (such as in Figure 1) as well as for the predictive model itself. The data itself was clean but needed a bit of massaging. The render time column was converted to seconds from milliseconds. The standard time column was converted to a python friendly timestamp. The document names column also contained erroneous names for documents that had "??" characters populating it. These were removed and allowed the data to be ready for analysis.

Company	Industry	Document Name	Analytics Views
1832 Asset Management	Buy Side	Africa to see a huge renewable energy boom dri...	2
1832 Asset Management	Buy Side	Apache and Total hit the jackpot for a third t...	1
1832 Asset Management	Buy Side	Are major US shale gas players well-hedged to ...	2
1832 Asset Management	Buy Side	Asia's offshore wind market to nearly match Eu...	1
1832 Asset Management	Buy Side	Biden's $2 trillion energy plan paves way for ...	2

Figure 2 Top 5 rows of cleaned data set

EDA

A deeper dive into the data was performed to get a better understanding of the data. The distribution of analytical views showed something interesting that needed further investigation. The data looks uniform with almost all view totals being under 5 views. However, the data is extremely right skewed. Figure 3 represents the distribution of the analytical views.

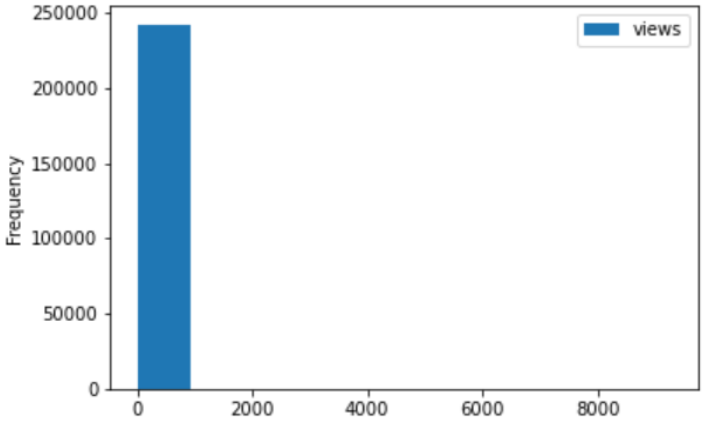

Figure 3 Distribution of analytical views

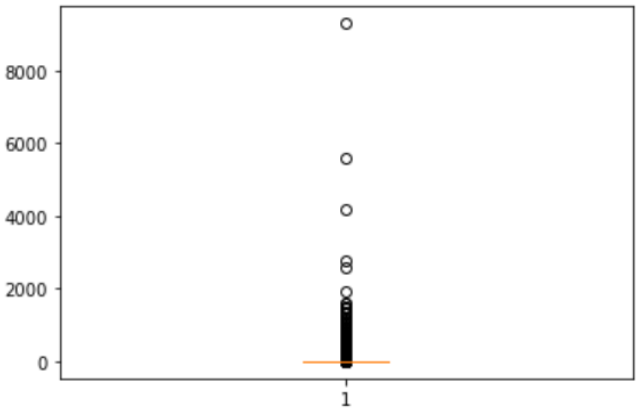

Figure 4 Boxplot of analytical views

A closer look at the boxplot in Figure 4, confirms the right tail nature of the distribution. There are some extremely high outliers. These outliers proved to be actual values and were not errors so they could not be removed.

Preprocessing, Training, Modelling

Preprocessing of the data was done by first utilizing the encoding the data for the categorical columns. The only column not encoded is the analytical views. The data was then removed of the 'other' selections from the industry column. This would then be reused to predict later if needed. The rest of the data was split into a train/test set with the test being 30% of the data set. Random Forest and KNN were then used to cluster the data to determine if the industry could be separated by the document name, company name, and view count. The accuracy was then determined for each model. With an 82% accuracy score, the Random Forest model was chosen to move forward with in order to determine the actual industry of the 'other' or 'unknown' industry.

Model	ACC
Random Forest	0.82
KNN	0.57