BEST SPORTS STORES OPENING

Analysis of Yelp reviews for Sporting Goods Stores

Team Members:

Chris Lee, Juan Tordoya, Junyue Deng, Supriya Bhatia, and William Salas

ABSTRACT

This report presents our analysis for Best Sports for stores opening target locations, stores' features and main subjects of customer dissatisfaction for other brands.

MSA 8050 Scalable Data for Analytics

OBJECTIVE AND BACKGROUND

Best Sports Companies is looking to penetrate the market with a new brand of stores. They have commissioned us to identify the geographic areas where they can start businesses, features that other stores are lacking, and most areas of improvement of existing Sporting Goods stores. Our team will find the most beneficial metropolitan areas, and the top complaints from existing stores so that they can improve on launching their brand. Main objectives include data preprocessing, sentiment analysis, topic modeling and so on.

We have selected the public Yelp reviews database to conduct our analysis. Our team will process several million records form the database, to extract the relevant data for our analysis.

HYPOTESIS

Hypothesis 1: Geographic area in USA with top number of stores will be included in the list of options to open the store. We expect to find metropolitan areas with high population density, and high disposable income

Hypothesis 2: After analyzing the review text and review stars of companies in sporting categories, we can determine areas of improving after running logistic regression prediction and LDA classification. We expect to find topics that are highly relevant for similar stores in both positive and negative reviews, and determine which are the areas of improvement in new stores

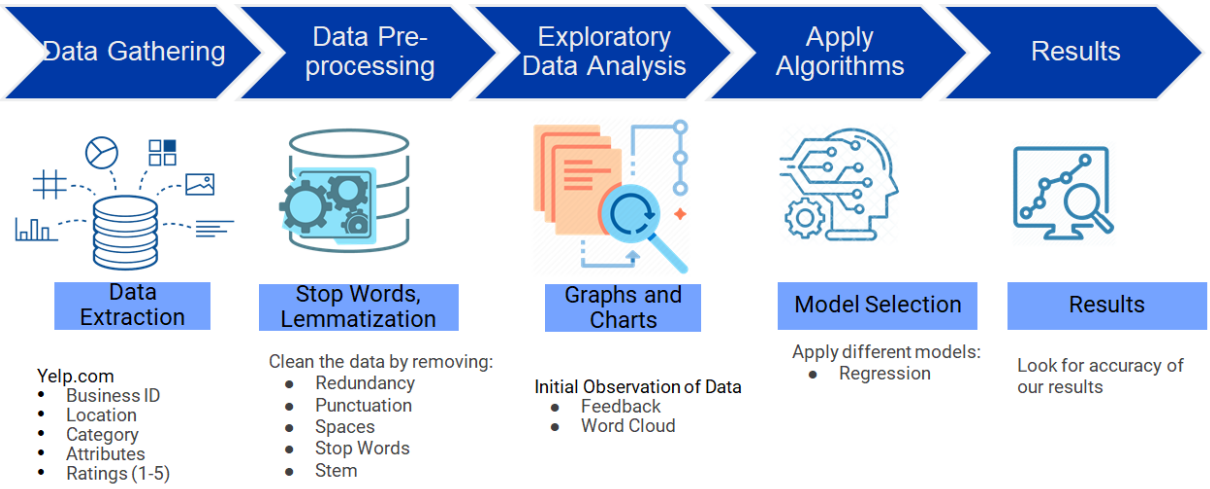
Hypothesis 3: Parking attributes will also be an important factor in evaluating the geographical location. We will explore the importance of different business attributes by checking the correlation between attributes and rating. This will give more insights on location and improvement such as product offerings and so on.

METHODOLOGY

There is plenty of data in the data set, but after we took a look at all the data, the data that we decided that would be most useful to us are: Business (ID, Name, Location, Category), Business Attributes (Parking attributes), Business Hours and amount of Check-ins, Reviews and Tips (Reviews, text, stars, etc.), User (yelping_since and review count for threshold).

We have solved the problem in PySpark utilizing Natural Language Processing (NLP), Sentiment Analysis, and Aggregation techniques. The techniques and algorithms can change throughout our assignment as we do not know what problems we may face.

We have used following steps to arrive at the conclusions:



EXPLORATORY DATA ANALYSIS

We extracted information of all businesses in the Sporting Goods category, and then filter only the reviews for the selected businesses using joins. Our selection yielded 1872 companies and 24340 reviews. We also filtered the companies' attributes entity to find companies in the Sporting Good category that have identified attributes, we found 1782 companies with reported attributes

Additionally, after further inspection of the reviews we found 74% positive sentiment as compare to only 12% negative

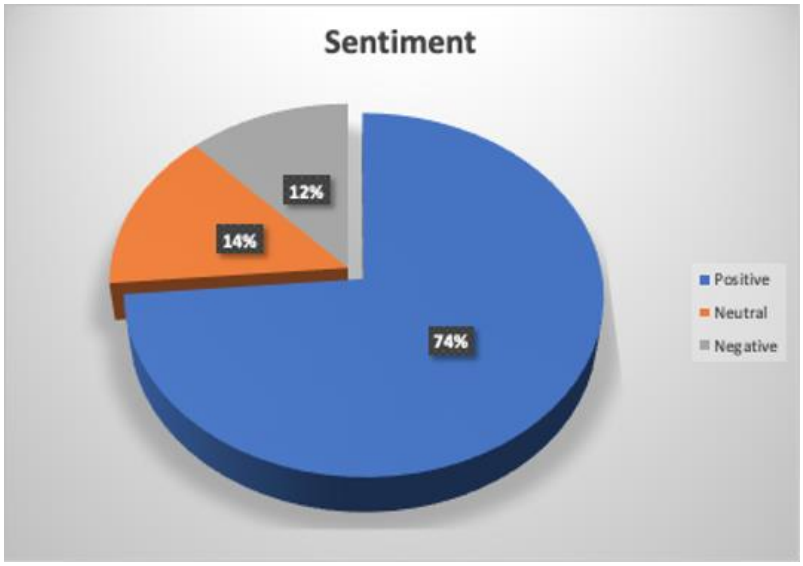


Figure: Positive, Negative and Neutral sentiment distribution

In the business table we identified 1872 business considered into 'sporting goods' category, located in 15 different states and 160 metropolitan areas. The most relevant metropolitan areas, in the US, in terms of frequency are: 11.5% in Las Vegas, 5.9% in Scottsdale and 4.5% in Charlotte, the only cities with more than 4% of the business.

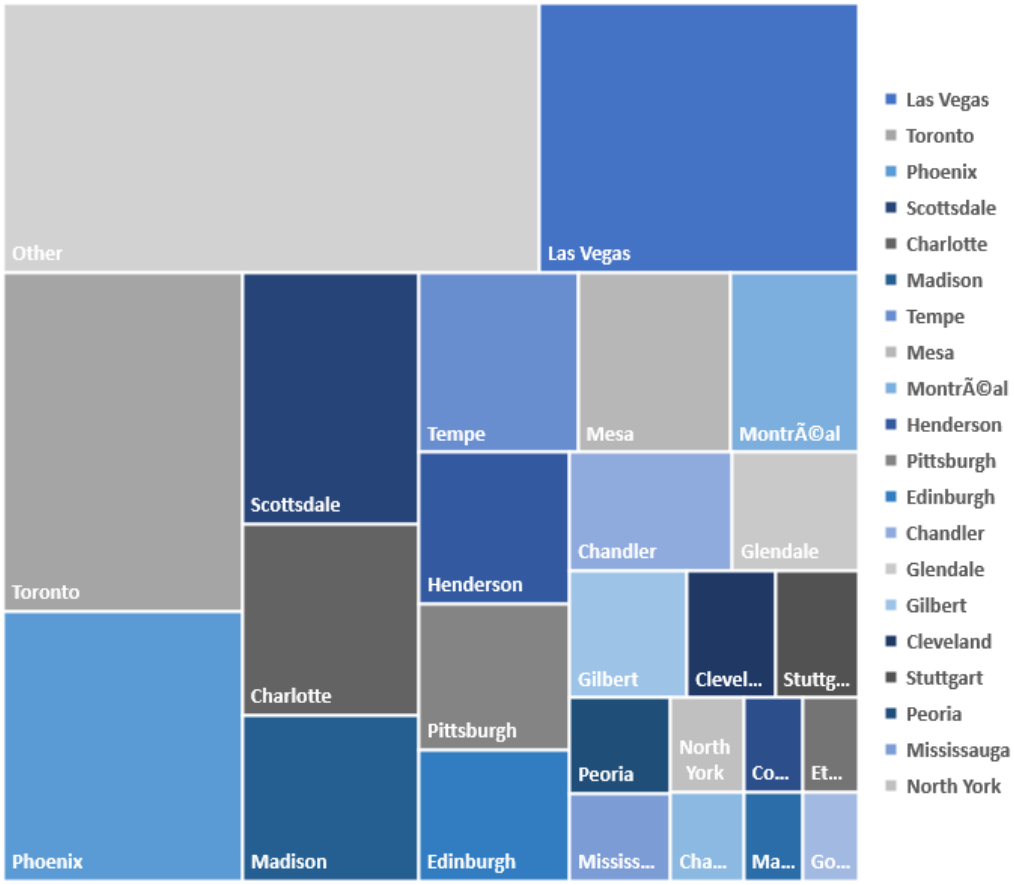


Figure: Distribution of business by city

In addition to the business information, which provided us the geographical distribution of the business we are working with, we have 24240 reviews for those 1872 businesses, being the 10% of the total number of reviews in the database, and, with average rating of 3.97. We also have access to the attribute information, in which the business may have, true, false or non-mark, indicating if any service or facility is provided for the customers.

We tried to visually identify any correlation between the business attributes, and the customers rating, as a first look into what we expect to infer the customers value the most, and the result is shown as follows in the graph.

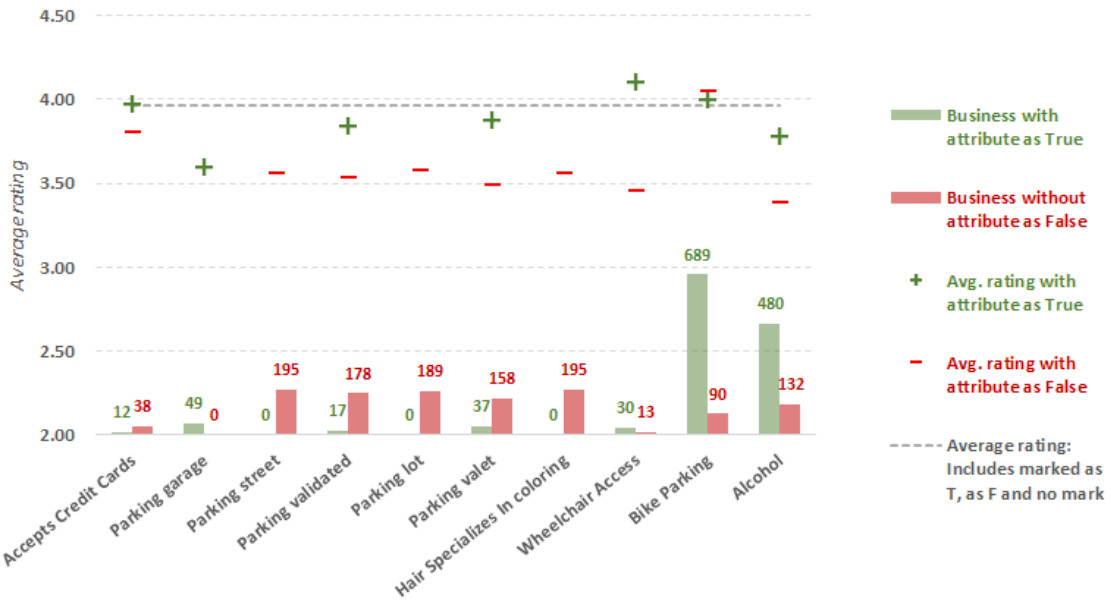


Figure: Average rating by attribute

As a result, there are some attributes, like *Wheelchair Access*, which determine a meaningful impact in the average rating, or parking related attributes, that apparently have high value for customers.

PRE-PROCESSING

Pre-Processing is an important part of both sentiment analysis and topic modeling. Both which utilizes Natural Language Process in which the text needs to be normalized. The general process of which we implemented goes as follows:

1. Removing punctuations from the text.
2. Converting all the text to lowercase.
3. Tokenizing the text.
4. Removing stop words and numbers from the text.
5. Stemming/lemmatizing the text to obtain the root word.
6. Removing shortwords, words with minute amounts of characters.

With the pre-processed data we begin preparing the training and testing data set. We first split the data into two categories of satisfaction: Low satisfaction and High Satisfaction. In splitting the satisfactions, we did not include the reviews with a star rating of 3 because we consider it to be neutral.

1. Low Satisfaction = Stars < 3, label = 0
2. High Satisfaction = Stars >= 4, Label = 1

With the newly defined low and high satisfactions, we created our training and testing data set with a random split of 70%/30% respectively.

MODEL SELECTION

After having the data pre-processed, we transformed the data into a sparse matrix using the CountVectorizer wrapper. We will use a logistic regression model to optimize a regularization value, in which will be used on the test data on the final model.

Using the idea of logistic regression, we created a pipeline and the model for logistic regression with a Cross-Validation K-Fold of 10. With the model, we did an LDA Classification to classify the positive and negative comments to discover the topmost talked about topics in each sentiment.

RESULTS

Our regression analysis produced an 89% accuracy, which indicates that our positive and negative topics will be highly accurate. After doing an LDA classification of negative and positive sentiments, we come up with the following topics:

On the positive comments we found topics related to store service and staff's knowledge of the products. We also found that it is very important for the customers the variety of accessories, and selection of specific items like bikes and shoes

On the negative reviews, we found that customers complain mostly on employees and prices. Some interesting topic that customers are dissatisfied is the selection of magazines, and golf items. There are also some negative comments about the stores' customer services and purchasing experience

Figure: Positive and Negative topics

Some of the negative words that customers use in their negative comments are words like store, bike and service, which may indicate a bad customer experience in the bike and bike accessories department. Another interesting finding is that customers also complain about shoes selection, which is surprising since we are analyzing sporting goods stores. We would suggest looking into this further to find the root cause.

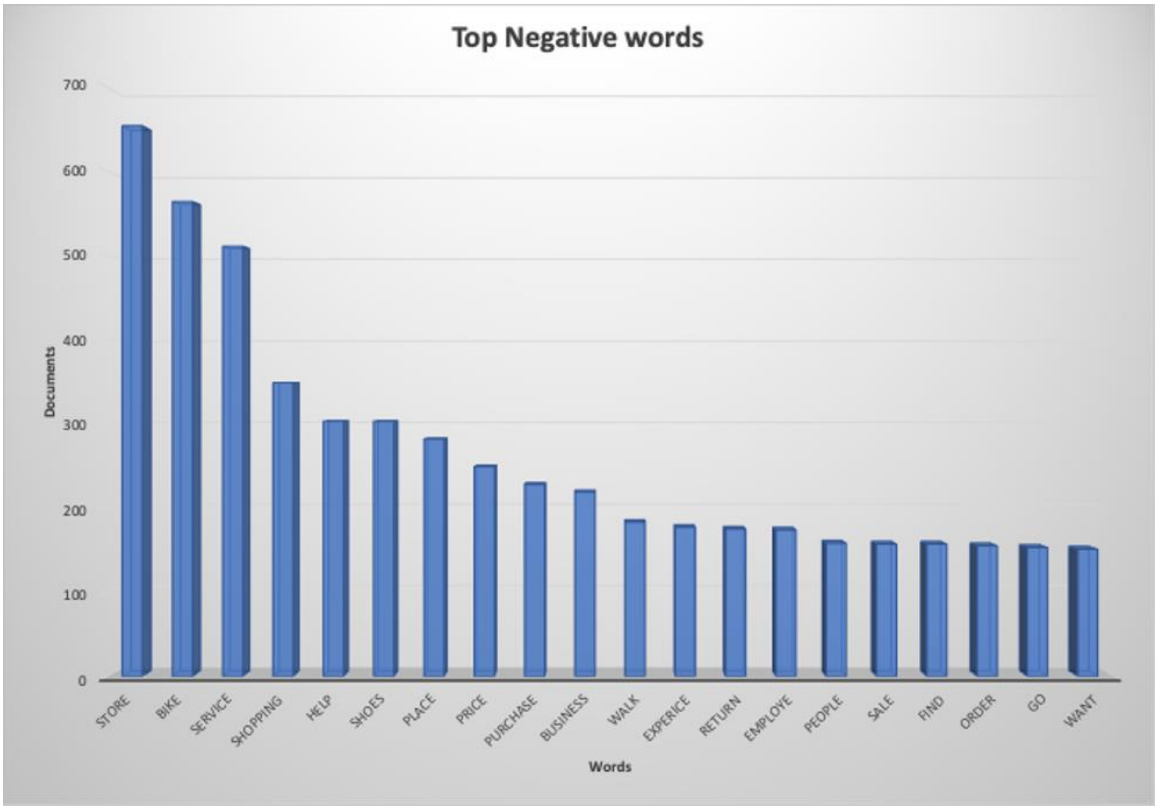


Figure: Topmost negative words

In general, we can say that customers are satisfied with the stores and stores' services, but it seems like there is a good amount of complaints about prices, purchase experience and selection of items related to biking and shoes.

CONCLUSION AND RECOMMENDATION

Store Location:

Best Sports companies should look to open their new stores in three regions, West in Las Vegas and Scottsdale, Center in Peoria, and East in Charlotte and North NY.

Based on the data analysis, the stores are located in highly concentrated metropolitan areas with a high-income level. Due to the nature of the sporting goods business, these areas are the most effective and profitable areas in the country

Positive sentiment topics:

Clients are satisfied in general about the store location and buying experience.

As mentioned above, the location of the sporting goods store is important for the consumer, as well as the buying experience. The clients are mostly satisfied with the buying experience that most of these stores provide

Negative sentiment topics:

Clients are looking for more variety in bikes and biking accessories. Also, they are looking for shoes and more help from the store staff.

Based on the analysis, the clients are looking for more variety of articles, they are very interested in bikes and bike's articles. We recommend Best Sports to focus its attention to this type of articles

Most talked about words in a negative way:

On the negative side, the most talked about words that the company should focus their advertising are related to store, staff and purchasing features like returns and client service experience.

Stores attributes:

Best Sports Companies should focus on offering the following attribute for their stores: some parking related facilities, wheelchair easy access or to accept credit card, which are the most valued for the customers.