Yelp Elite Member classification

Rajat Jain

Yelp Datasets

1. Yelp has introduced comprehensive dataset for conducting research and analysis
2. This dataset contains seven CSV files. Ranging across different domains.
    - Users
    - Business
    - Tip
    - Reviews
    - Checking
    - Business_Attributes
    - Business_Hours

Data Statistics

1. 5,200,000 user reviews
2. Information on 174,000 businesses
3. The data spans 11 metropolitan areas

EDA

User Dataset

1. Major of the users are yelping between year 3 to 7.
2. There are total 75 people who are most oldest yelpers. They are yelping from last 10 years.
3. There are three users Shila , Kim, Victor with review count over 9000. They must be food/restaurant reviewers.
4. Most likely maximum amount of cool, funny or useful reviews a yelper will write is from 0 to 10.
5. Categories of reviews writing does not depend on the how old you are with the yelp.
6. Old or young no difference between them while giving the stars.

Business Dataset

Found the Recipe for being the successful business

1. For head start I considered only five star rated businesses.
2. Las Vegas and Phoenix seems to be safe cities for opening the business.
3. Home Services, Beauty and Spas are the major categories of the five star business.
4. Home Services is the not only most reviewed category but also people review it most positively
5. Westside region of the Las Vegas has most successful 5 star business.

Recipe - Open a Home Service business in the west side region of the Vegas

NOTE- Perform sentiment analysis on the reviews of the people about Home Services in Vegas. Find the gist of the problems the consumers are facing from present and built business around it.

Yelp Elite Program

1. Yelp is well known as the most popular tool used by people to discover new places through reviews.
2. For engagement they added the social aspect to their business. Eg. “Friends” feature, a “Fans” feature, votes and likes on reviews and much more.
3. In this problem I am interested in a unique social feature Elite use feature.
4. They are basically handpicked by the yelp who are most active and produce exceptional reviews
5. With more and more users who want to become part of the club has created the tough competition for user to become elite.
6. So yelp on daily basis receives lot of applications especially from the young yelpers
7. I want to automate this process with building a model which help selecting the members to elite.

Problem Statement

1. Out of 1326100 total users, 1265282 are NON-elite, 60818 are elite.
2. Mean review count for elite members is 234 whereas for Non-elite are 60
3. Generally people with large review counts are selected as elite.
4. Interestingly there are 25 members whose review counts are less than 10 are selected as Elite.
5. I want to build a classifier to classify users as elite or not when review count is less then 10.

Imbalanced Dataset

1. There are total 8916906 users with review count less than or equal to 10.
2. Out of which 8916881 are non-elite and 25 are elite.
3. Clearly the case of imbalanced dataset

First Intuition

1. Normal Classifier accuracy metric will not work for this solution .
2. Classifier will be biased towards the majority class '0'
3. Classifier will wrongly classify minority class ‘1’ as ‘0’
4. One solution is to collect more data for review count for class 1 which is not possible in our case.
5. Another solution is to perform undersampling or oversampling .
6. OverSampling adds copies of the minority class.
7. UnderSampling deletes the instances from the majority class.

Business Constraints

1. Non-elite should not be classified as elite at any cost.
2. Biased the classifier to reduce false positives.

Experiments Approach

1. For hypothesis generation only considered undersampling.
2. Three undersampling techniques (NearMiss, ClusterCentroid, RandomUnderSampler).
3. Reason for choosing them solely on behalf of the computation nature of them.
4. Used 5 different ratios for majority class 1, 2,3,5, 10.
    a. Ratio 10 means - Majority class(0) 250 and Minority class(1) is 25
5. Experimented with three classifiers
    a. Logistic Regression
    b. Random Forest
    c. SVM
6. Recall used as a evaluation metric to score the classifier.
7. Performed the error analysis extensively using outlier detection techniques.
8. Performed feature engineering using different methods to extract important features.

Results

1. Random UnderSampling has performed the best, shown least variance in the ratios
2. Logistic Regression and Random Forest both performed better than the SVM
3. Ratio 5 is a interesting case. Recall score increased at this sampling after the dip

Error Analysis

1. Three false positives are detected (Class 0 is classified as Class1)
2. Performed Statistical based Extreme Value Analysis.
    a. 2 out of 3 found out to be extreme value for review counts.
3. Used two further Methods, Angle Based Outlier Detection and KNN for further outlier analysis.
    a. Both Models failed to predict any outliers

Feature Engineering

1. Performed Feature Engineering using three methods
   a. Correlation
   b. Recursive feature elimination (RFE)
   c. Ensemble Method - ExtraTreesClassifier
2. Used changed in the recall score from baseline as the evaluation metric
3. All the three methods failed to improved the recall score.

Method	Change_in_score_from_baseline
Correlation	-0.11
RFE	-0.24
ExtraTreeClassifier	-0.16

Future Work

1. This problem can be derived further from here into either as a Anomaly detection or skew Classification problem.
2. Need to dig further into finding the interesting pattern of those 25 elite users.
3. We can also segregate the problem on behalf of the different categories and derive the category based features.
4. Other case studies in same domain can also be researched
    a. Microsoft faces the same problem of unbalanced dataset Click Prediction in Bing Search Ads.
    b. Facebook have many billions of observations of users daily. To present ads they downsampled negative examples at different rates.

References

Microsoft Case Study
https://www.microsoft.com/en-us/research/wp-content/uploads/2017/04/main-1.pdf

Google Ad Prediction Case Study

https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41159.pdf