****

**PROPOSAL AND STATEMENT OF WORK**

**AIDI 1003 – CAPSTONE TERM 1**

**AMIT MARAJ**

AHNCH BALA 100424062

SONAKSHI KARKERA 100720763

SURBHI THAKUR 100732335

ARUN KALAESWARAN 100771700

**September 27, 2019**

## Executive Summary

The main goal of this project is to apply machine learning to predict a restaurant’s success based on the customer’s star rating and finding which restaurant features have the most impact on its average rating by performing sentiment analysis on the text of user reviews.

We’ll be using the dataset provided by Yelp as part of their Dataset Challenge 2019 to train and test our classification model. The dataset includes information about local businesses in 10 metropolitan areas across 2 countries.

Along with the main goal, we have two additional goals. The second goal is to help Yelp differentiate themselves from competitors. The review industry is a saturated market with many competitors vying for the top spot. Yelp, arguably one of the most popular review site in North America, would like to remain at the top by using the data they have in creative ways. Our model will help open a new revenue channel for Yelp as well as potentially increase revenue. Our third and final goal is to ultimately help local restaurants achieve success within their city. Our model will let them be aware of any imminent closure, so that they can take the appropriate steps to rectify their business strategy.

## Rationale

Restaurant owners do not know about the closure of their restaurant early enough to prevent it from happening. We have proposed a solution that can warn business owners of closure early enough so that they can take preventative measures. As stated before, our goal out of this project is to look out for features that better predict the future of a restaurant.

Our prototype will be able to predict this based on the customer’s star rating and finding which restaurant features have the most impact on restaurant’s average rating by performing sentiment analysis on user reviews. Our key metric for model evaluation will be precision and accuracy. Though accuracy will indicate how well our model is performing, we’ll be targeting on getting high precision value for “closed” class.

There will be 3 beneficiaries of our prototype. Yelp, our main beneficiary, will get to open a new revenue stream through consulting.  Secondly, clients/restaurants get to receive expert advice from Yelp in order to prevent closure and financial loss. Finally, consumers will receive a more fine-tuned and catered food experience.

## Why is this problem important to the organization?

This problem is important to Yelp, as they want to retain as many businesses on their platform as possible. More businesses on the platform would generate more advertising revenues since it would create more existing web pages to advertise on. Having more businesses on Yelp would make them the industry leader in food recommendations, which would create more trust in their platform and make restaurants want to be on Yelp or have consumers want to search on Yelp.

## Problem Statements

**1.** When restaurants are already active on Yelp based on their Yelp profile we will be able to determine if they will close or not.

**2.** Using this model, Yelp can differentiate themselves from their competitors like Google Reviews, OpenTable, etc. Yelp would be providing a service that can help their clients make business decisions instead of just being a platform that consumers would use for making food decisions.

**3.** Existing restaurants will be able to increase their customer experience using our recommendation. By determining whether they will close or not would give clients an opportunity to improve their user experience and try and prevent their business from closing.

## Data Requirements

We will be using a data set that has been provided by Yelp that they created for the Yelp Dataset Challenge. The Yelp Dataset Challenge wants students to use Yelp’s data for academic or teaching purposes in innovative ways. The dataset contains 6,685,900 reviews, 192,609 businesses, 200,000 pictures and 10 metropolitan areas. It is a large json file that is located on their website at this link; https://www.yelp.com/dataset/download.

We require our data set to have the following requirements: *Business Name, City, Stars, Review Count, Is Open, Categories, Attributes, Hours of Operation, Review Date, Review Text, Review Star*. Within these requirements we will constrain certain features like city to specifically 3-5 metropolitan areas. We also plan to segregate the ratings from 0-3 and then 4-5. ‘Is Open’, is the most important feature as it will tell us whether a restaurant is open or not as of January 19, 2019. We may have to make assumptions on the Categories feature, as some restaurants may be fusion but can still be labelled as Thai food or Indian food. Attributes will play a key role in our data as well, as there are many attributes that can affect the longevity of a business like whether it’s a take-out restaurant and if they have street or garage parking. We will have to sort through the many attributes and select the ones that are most important. The amount of attributes may essentially lead our dataset into having 20-30 features. This would mean we would have to carefully filter the most important features and apply those to our model. We are assuming Yelp’s reviews provided aren’t spam reviews and are actual reviews by individuals. It is an ongoing problem of reviews being created by bots, however we will make the assumption all the reviews are intact. We also assume that the restaurants that are open or closed as of January 19, 2019 are still open or closed as of today.

## Data Collection

An open source dataset was provided on Yelp’s website that Yelp regularly updates for its Dataset Challenge. We retrieved the datasets of different release year, namely 2018 and 2019 respectively. Each dataset contains 6 JSON files: business, check-in, tip, photo, the user. For the scope of this project, we’ll only be using the business and review JSON files. Since these are authentic and detailed datasets provided by Yelp, the file sizes are quite high (8GB each). Due to this, we may need to switch to cloud options for training our model.

## Data Exploration

Our aim, when it comes to data exploration, is to use the data provided by Yelp in 2018 to predict the closure of a restaurant in 2019. Essentially, this is a binary classification problem in which we’ll have to utilize a variety of classification algorithms as well as perform sentiment analysis on restaurant reviews using NLP to improve the accuracy of our prediction.

In order to achieve this, we will be looking for features (statistical data from the business.json file) that best describes a restaurant in a more objective way, along with analyzing text features from their reviews. The business.json file contains business from select cities and each business has 15 attributes. The review.json file has 9 attributes. After determining our scope, we’ve filtered and decided on the features that would have the greatest impact on our prediction:

|  |  |
| --- | --- |
| *Independent Variables* | *Dependent Variables* |
| business\_id | is\_open |
| name |  |
| city |  |
| state |  |
| stars |  |
| review\_count |  |
| categories |  |
| review\_id |  |
| review\_stars |  |
| date |  |
| text |  |

Since this is a binary classification problem, the variable i*s\_open* is mapped as 1 for an open restaurant and for those that were closed, it was mapped as 0. It is also worth mentioning that in order to have all the reviews for recorded business in the year 2018, we have to merge the two datasets on business\_id.

# Data Preparation and Cleaning

In order to prepare the data, we’ll first filter the Yelp data for businesses only as that is our main focus. To do this, we select businesses with they keyword “restaurant” within the “category” column and “Toronto” as “city”. This will give us data on only restaurants within Toronto. *Note*: We’ll be building a model for Toronto initially and then duplicating it for different cities.

As additional steps for preparing data we will be dropping additional features and handling missing values.

There are around 5256 open and 2711 closed restaurants that are shown in the dataset. The number of observations belonging to the open class is significantly higher than those belonging to the closed class. Hence, we handle an imbalance in classes using appropriate sampling techniques.

The following will also be done in order to clean our data:

* Normalize the review text by converting it to lower–case and removing special characters
* Tokenize text by converting unigrams, bigrams.
* Consider performing [lemmatization](https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html).

## Data Engineering

For categorical features we will be clubbing the most common word pairs across domain features as one variable. We will keep building on it and will come up with several other features. (Eg: Location, ambience, service etc)

As for numerical features, we have review count, stars, review\_stars, location and chain\_restaurant-if a restaurant appears more than 3 times in a dataset it will be marked as 1 (chain restaurant) and 0 otherwise.

## Training and Testing Dataset

We’ll train our model based on the features we have selected and evaluate the model using 60/20/20 train, validation and test dataset. We’ll train our model using the training dataset, validate the performance of various models using the validation dataset and predict using the test dataset.

## Methodology and Algorithms

The problem at hand is a binary classification problem, so using Logistic regression would be an ideal classifier in this situation. We will also be using complex models like SVM and Naïve Bayes to compare it with a simple model like logistic regression and choose the most appropriate one based on its performance. The evaluation metric we will be using to measure the performance of the model is Precision and to measure the performance of a classifier we will be using Recall.

In our context, consider (X, Y) to be a data point where X is the set of features and Y is the binary value. Y ⊆ L, where L = {Closed, Open}.

Let h be a classifier.

Let z = h(x) by h for the datapoint (x, Y). Then,

Precision = Out of the categories predicted, how many of them are true category.

Recall = Out of the total true categories, how many of them were predicted to “close”.

Another metric is the error rate to give an overview of how good the predictor is. In our case, even though a prediction is wrong, we want to know how deviated the prediction is: for a restaurant with a 4.5-star rating, classifying it as a 1- star restaurant is a totally different story from classifying it as a 4-star restaurant, even though both predictions are wrong. Therefore, here we need the help of the confusion matrix. For the limit of space, we will only include the confusion matrix for our” best” predictor - logistic regression.

## Software

For software. we will be using Jupyter Notebook to create the model.

Project Plan

**A screenshot of a cell phone

Description automatically generated**