Yelp Predictor

Yelp Data Analysis and Prediction

Darshan Swami, Chandra Harsha, Sidhanth Subhash Jain, Tanishq Tapiawala

INTRODUCTION

When you invest in a new firm, you want it to be profitable for you as the investor. However, there are many unknowns in the market for new businesses, such as where to invest, where to position it, what category it should go under, and how to identify the customer base or target market. There are merely just few ways to find out or answer to these questions. One of the most prominent is Yelp, where you can find reviews and ratings for multiple businesses from any location.

Yelp is a global online review network, including restaurants, shopping malls, hotels, and tourism businesses around the world. Users can rate merchants, write reviews, and exchange service experiences on it. The Yelp open-source dataset is the official dataset for scientific research and machine learning competitions produced by Yelp Inc. 908,915 tips by 1,987,897 users. Over 1.2 million business attributes like hours, parking, availability, and ambience.

Objective 1: The Business Part. Precisely describe the business problem Yelp Predictor aims to solve, discuss why this problem is important, and explore how our data-science techniques can make a difference.

Objective 2: The Data Part. Formulate the business problem as a data problem and develop a data science solution for implementation.

Objective 3: Sentimental Analysis. Develop a prototype of Yelp Predictor by acquiring a sample data set and implementing the previously designed math solution.

OBJECTIVE 1: The Business Part

The business problem to solve.

Yelp Predictor is a predictive analytics tool aimed at investors, customers, and business owners. A business review rating can have a significant impact on revenue. Online product and service reviews frequently have a significant impact on future purchases. Many analysts concluded that "a one-star increase in Yelp rating leads to a 5-9 percent increase in revenue" using the Yelp dataset. As a result, an algorithm that can predict the quality of a potential business would be extremely useful to potential owners. The ratings and reviews can help us find better restaurants and services.

Why this problem is important to solve.

According to the most recent Bureau of Labor Statistics data, nearly one in every five businesses in the United States fail within the first year (BLS). With 32.5 million small businesses in the United States, some will undoubtedly fail, whether small or large. Financial constraints, workforce issues, and owner burnout can all cause businesses to fail. Furthermore, the percentage of businesses that fail varies greatly depending on the state or industry. Having an algorithm that tells investors which business categories to target and assisting business owners in understanding the reviews about their business can have a significant impact on the number of failed businesses in the United States. Yelp Predictor solves all of these issues using a completely automated machine learning approach.

The ideas behind Yelp Predictor to solve this problem.

When you make an investment in a new business, you want it to be profitable for you. However, there are numerous unknowns in the market for new businesses, such as where to invest, where to position it, what category it should fall into, and how to identify the customer base or target market. There are only a few options for finding out or answering these questions. One of the most well-known is Yelp, which allows you to find reviews and ratings for businesses in any location. Data Pirates attempted to analyze Yelp data and train machine learning models to determine the best company, location, and category to enter for your business's success.

Differences that can be made via our data science approach.

Utilizing sentiment-based analytics offers a number of advantages as a data science approach. As '*data is the new oil'* every business decision is fuelled by data to shoot to the moon. We can leverage the power of data analysis and conject unseen insights. The key to a successful business is customer satisfaction and with the help of our technology we can obtain real life sentiments from the reviews and ratings posted by the customers, which can help the owners or investors to make precise business decisions aptly. Data may disappoint but it never lies and from the data we could see that the restaurant business is the one that is affected more by the reviews. We have focused our analysis to restaurant reviews, to identify the likelihood of a review being positive or negative and how it's rating can affect the business.

This idea deserves investment from the Sharks.

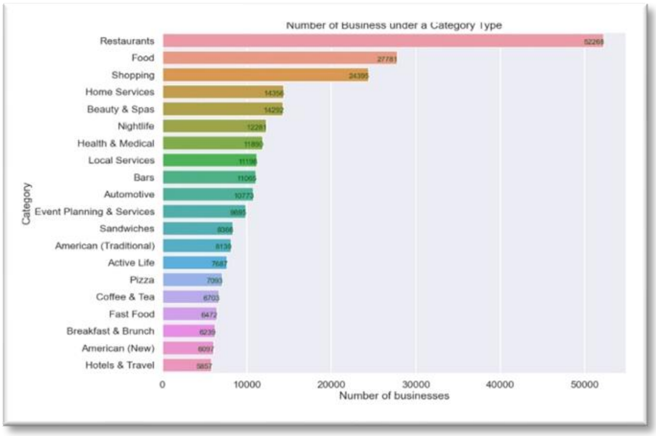
Yelp Predictor is Software-as-a-Service targeted towards investors who want to start a business but don't have any idea of where to invest and what business to invest in. We provide cutting edge solutions based on the evaluation of the data from the country's best review rating website 'Yelp' where you have millions of reviews on thousands of businesses. Yelp Predictor's unique methods offer competitive advantages using higher computational exploratory data analysis. The most unique feature of our company is to provide real life sentimental analysis of reviews to understand the actual sentiments of the customers for a business. This allows investors to use our services with confidence that their investment is not at risk.

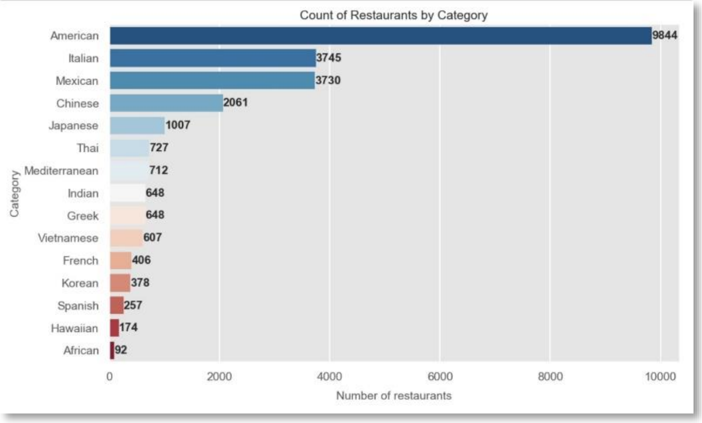
We are seeking a $100,000 investment from the Sharks, in return for a 10% equity stake in our company. We believe this valuation is reasonable given the need (and future need) which our product fulfils, the large market size, and the high potential for future growth.

OBJECTIVE 2: The Data Part

Conjecture 1: Restaurants are the most common business category.

The conjecture is true, the number of restaurants in the business sector is the highest at 53,000. We may therefore presume that opening a restaurant is a smart choice if one wishes to launch a new business



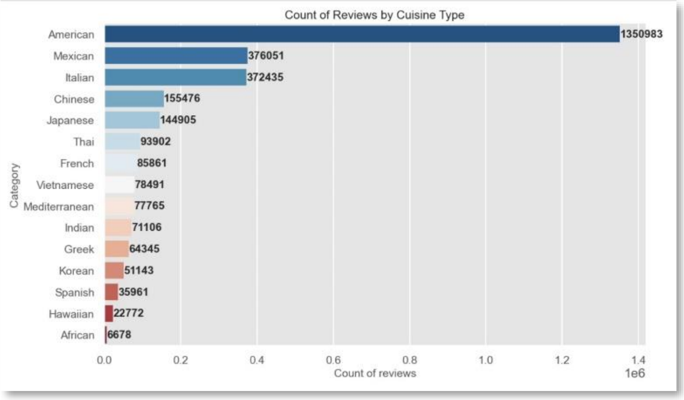


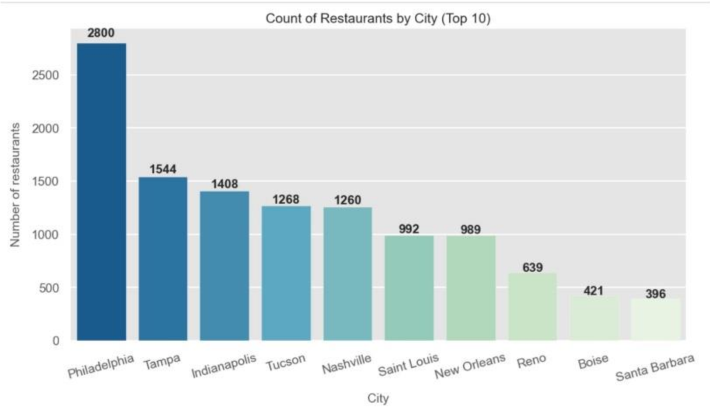
Conjecture 2: The Highest count of restaurants are of Italian origin.

The conjecture is false. The most eateries are found in the categories with dark blue backgrounds. On the other hand, the categories with dark red backgrounds have the fewest eateries. Most eateries are of American descent. Therefore, we may assume that individuals prefer American cuisine, and we can use this information as guidance when beginning a new business.

Conjecture 3: The most reviewed cuisine is French.

The restaurants with the most reviews are of American cuisine. So, when selecting what Category of cuisine eatery, we should start and we can take this information into account while opening a new restaurant business.



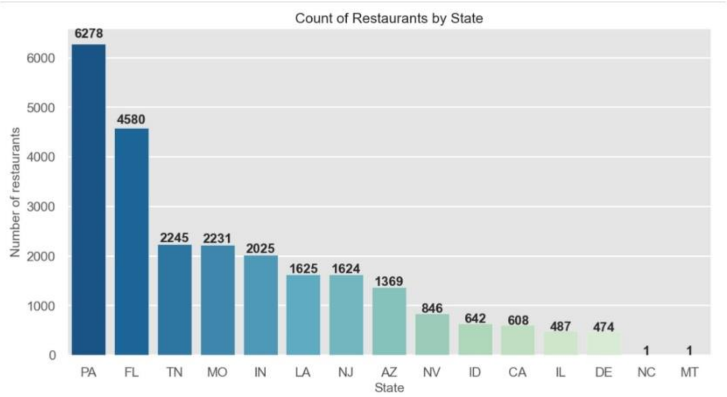


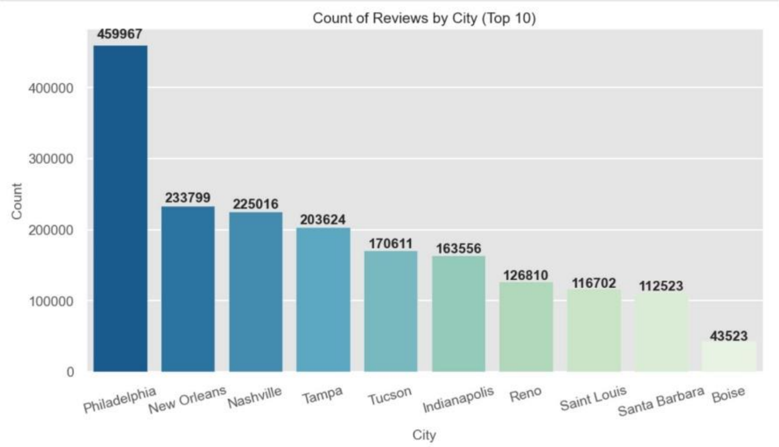
Conjecture 4: Chances of facing competition are more in Philadelphia city.

The restaurants with the most reviews are of American cuisine. So, when selecting what Category of cuisine eatery, we should start and we can take this information into account while opening a new restaurant business.

Conjecture 5: Pennsylvania has the greatest number of restaurants.

This conjecture is true as we saw earlier that Philadelphia has the highest number of restaurants as well as highest number of reviews so we can start an American cuisine eatery in Philadelphia city of Pennsylvania state. As the customer base is large so we have plenty of opportunity.



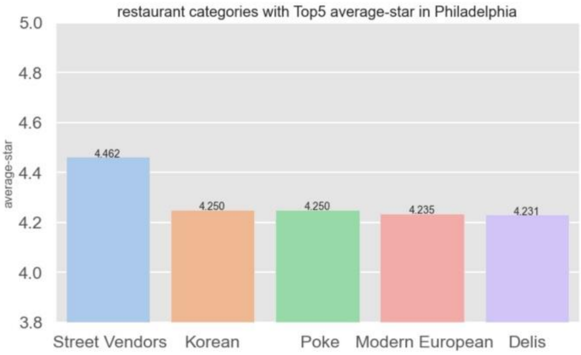


Conjecture 6: Philadelphia city has the greatest number of reviews.

This conjecture is true Philadelphia is the city with the most count of reviews and Boise has the least number of reviewers.

Conjecture 7: Street vendors have the most average star rating in the city of Philadelphia.

This conjecture is true, and street vendors have the most average star rating in Philadelphia. We can see that people prefer street food in Philadelphia So if we want to start a new business, one can think about investing in street food.



OBJECTIVE 3: Sentimental Analysis Part

Sentiment analysis is a highly effective tool for a business to not only take a look at the overall brand perception, but also evaluate customer attitudes and emotions towards a specific product line or service. This data-driven approach can help the business better understand the customers and detect subtle shifts in their opinions in order to meet changing demand.

Step 1: Using Count Vectorizer.

CountVectorizer means breaking down a sentence or any text into words by performing pre-processing tasks like converting all words to lowercase, thus removing special characters. In NLP models can't understand textual data they only accept numbers, so this textual data needs to be vectorized.

Step 2: Removing Stop words Porter stemming.

The Porter stemming algorithm is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems.

Step 3: Applying Logistic Regression.

Applied logistic regression to find the probability of the reviews being positive or negative. The accuracy obtained after performing logistic regression was 95.9% and F1 score for the model was 0.929

Step 4: Random Under Sampling and Random Over Sampling.

Random under sampling involves randomly selecting examples from the majority class and deleting them from the training dataset. In the random under-sampling, the majority class instances are discarded at random until a more balanced distribution is reached.

Random oversampling duplicates examples from the minority class in the training dataset and can result in overfitting for some models.

Step 5: Using TFIDF.

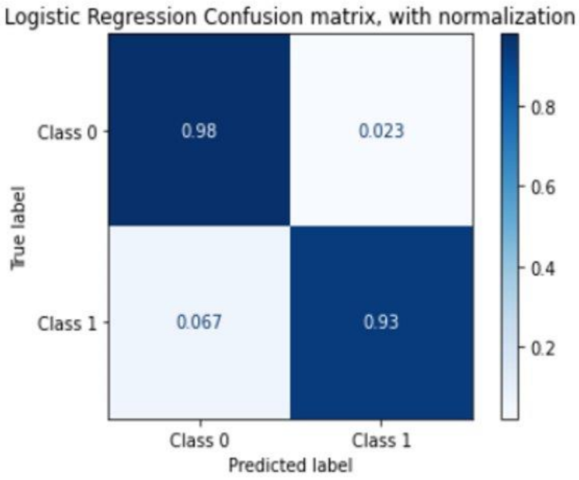
Transforms text to feature vectors that can be used as input to estimator. vocabulary_ Is a dictionary that converts each token (word) to feature index in the matrix, each unique token gets a feature index. We used TFIDF to compare the results with Count Vectorizer to determine better vectorization method for sentimental analysis. The accuracy of the logistic regression model after using TFIDF was 97.5% and F1 score 0.957 which shows an increase in accuracy of the model compared to Count Vectorizer.

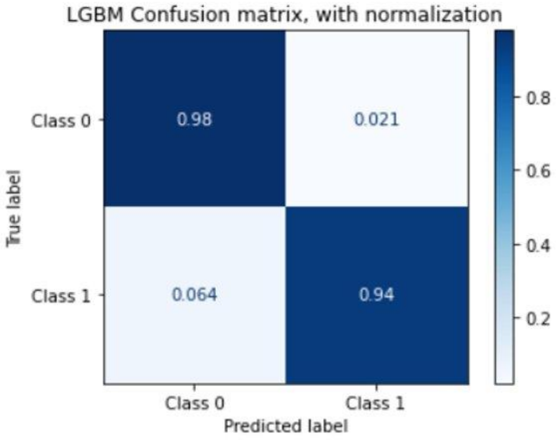
Step 6: Model Comparison.

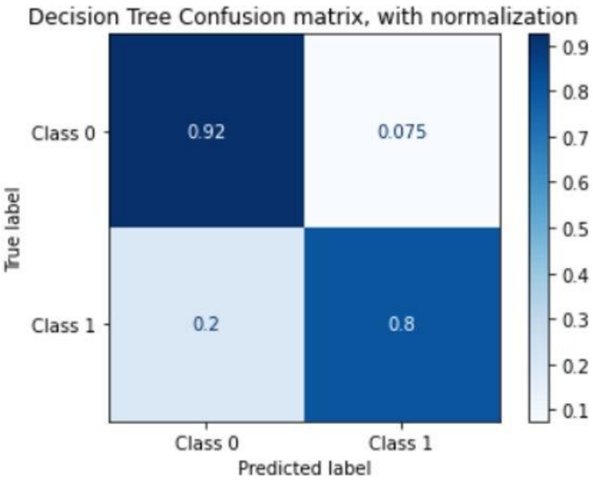
In order to find the best model to find the probability of the sentimental analysis we applied two more models to the dataset viz. Light GBM and Decision Tree and compared the results.

	Accuracy	F1	RMSE
Logistic Regression	**96.64**	**0.938**	**0.036**
Light GBM	**96.66**	**0.942**	**0.034**
Decision Tree	**89**	**0.81**	**0.11**

From the above table we can confirm that Light GBM gives the best accuracy amongst all the models.







CONCLUSIONS

We began investigating the difficult-yet-important problem of sentimental analysis using only the Yelp dataset's user reviews. We began by obtaining the Yelp Data set. It included three different json files containing information about Reviews, Businesses, and Users. We then did some preliminary research. We primarily focused on the restaurant reviews submitted by users. We discovered that 53% of users gave 5 stars (the highest rating) and 20% gave 1 star (lowest). We concentrated on the Restaurant category because it had the most businesses. We even came to the conclusion that American cuisine was the most popular. We can also see that the most reviews are for American cuisine. We then investigated the same using demographic data. We were surprised to learn that the state of Pennsylvania has the most restaurants. In addition, we can confirm that Philadelphia has the most restaurants. **So, as a business idea, we can correctly state that an investor can invest in a street vendor in Philadelphia to earn a profit.**

Then we moved on to the most crucial part of the Case Study, Sentiment Analysis. Sentiment analysis is a powerful tool for businesses that allows them to assess not only overall brand perception, but also customer attitudes and emotions toward a specific product line or service. In order to meet changing demand, this data-driven approach can help businesses better understand their customers and detect subtle shifts in their opinions. We used vectorization methods such as CountVectorizer and TF-IDF. Porter Stemmer assisted us with data pre-processing. Then we used various models such as Logistic Regression, LightGBM, and Decision Tree to determine which model would produce the best results for determining the sentiment of any given review. **We discovered that the LightGBM model was the best, with an accuracy of 96.66%, an F1 score of 0.942, and an RMSE of 0.034.**

APPENDIX 1: 90 SECOND “SHARK TANK” PITCH

Hello Sharks, we are Yelp Predictor. We are seeking a $100,000 investment in return for 10% Equity in our company. Yelp Predictor is a Software-as-a-Service company for our clients to help them invest smartly.

Sharks do you know Nearly 1 in 5 U.S. businesses fail within the first year, according to the latest data from the U.S. Bureau of Labor Statistics (BLS). With 32.5 million small businesses across the nation, some are undoubtedly bound to fail, whether small or large. Businesses can falter for various reasons, including financial constraints, workforce issues and owner burnout. Plus, the percentage of businesses that fail can vary widely based on the state or industry.

Yelp Predictor is Software-as-a-Service targeted towards investors who want to start a business but don’t have any idea of where to invest and what business to invest in. We provide cutting edge solutions based on the evaluation of the data from the country’s best review rating website ‘Yelp’ where you have millions of reviews on thousands of businesses. Yelp Predictor’s unique methods offer competitive advantages using higher computational exploratory data analysis. The most unique feature of our company is to provide real life sentimental analysis of reviews to understand the actual sentiments of the customers for a business. This allows investors to use our services with confidence that their investment is not at risk.

Thank you for your time.