Group_Project: Helping Yelp!

Dataset Description

We’re working with the Yelp dataset on Kaggle(link), which contains information about businesses across 11 metropolitan areas in four countries. The dataset is split into 5 tables: (see image 1)

Our goal is to predict if a business will be closed based on its business attributes.

Importance of the problem

Yelp very recently refreshed its board of directors and is currently transforming its business model from CPM to CPC. (impression-based internet advertising to cost-per-click performance-based model) With the new management team onboard, **Yelp is now committed to mid-teen CAGRs**(Compound Annual Growth Rate) from 2019 - 2023.

In Yelp’s most recent earnings presentations, they state that revenues from subscription software and new partnerships are growing at 30%. However, these account for only 5% of Yelp’s total business. This means that for the short term, Yelp will drive its growth rate by pushing its sales team to growth their advertising accounts in a consistent way, while at the same time cutting operating budgets aggressively. We anticipate that the following 1-2 years will be a painful period for Yelp’s current sales representatives, and we checked Yelp’s glassdoor reviews for confirmation (see Image 2).

We hope our model can help Yelp’s sales representatives in the following ways:

- **Filter out dying local businesses** so they don’t have to call everyone on their cold call number lists
- Utilize the time they free up to building relationships and improve **client retention** & explore multi-location and national accounts
- Decide whether to **onboard a business** on its platform looking at its attributes and predicting if they’ll survive or not

Exploratory analysis

Classifying the businesses by categories, we observed that **Restaurants and Shopping are the top categories with most businesses** (see Image 3). We have considered all cuisines of restaurants under the same category.

When classifying by location, it was surprising to note that New York was not in the list of top 10 cities with restaurants being a leading category. We checked the dataset and found that it did not have entries for New York. There was also a **positive correlation between the number of business and review count for the same location** (see Image 4).

We then looked at the distribution of stars and review counts and found that they are skewed. So,we used the median values of them as threshold to segment the businesses into four segments (see Image 5) - "I: Going great, II: Increase footfall, IV: Increase service and III: God save them". We expected to see higher number of businesses open in the I and IV categories and it was what the data showed as well. Almost all the businesses in the GS segment was closed. This shows that **stars and review count have an effect on the open/close status of the businesses.**

We wanted to look at **business attributes that affect the open/close status of a business**. Our understanding was that the users might be looking for specific attributes when they visit businesses and the absence of it would have an adverse effect on businesses. Considering restaurants being the top category, we intuitively selected coefficients like Parking Garage, Alcohol, Accepts credit cards, Live music, Open 24 hours and dogs allowed. When analyzed, the **parking garage, alcohol and allowing dogs** has a **positive correlation** with the target while having **valet parking** and **being open for 24 hours** has **no significant effect** on the target (see Image 7).

Solution and insights

With our exploratory analysis above, we see that most of the local businesses are either in the "Restaurants" or "Shopping" category. Now our goal is to fit a model to each business category respectively. We will try **logistic regression, KNN, and random forest classifier**s, and compare them with a **baseline where the model assumes that all businesses are open**.

We did grid search cross-validation with the following parameters for KNN and Random Forest :{*KNN:* number_of_neighbors = [3,7,11,19,29],
Random Forest: number_of_estimators = [3,7,11,19,29]}

Restaurants

	Model			
	Baseline	Logistic regression	KNN (n=29)	Random Forest (n=29)
Accuracy score	70.39%	70.96%	71.22%	70.96%

Shopping

	Model			
	Baseline	Logistic regression	KNN (n=11)	Random Forest (n=11)
Accuracy score	83.13%	82.93%	83.41%	83.33%

Insights

1. **Top features of shopping malls are highly correlated than restaurants** (it is easier to come up with a winning formula to make money in retail). However, it also means that **non-parametric models are better suited for predicting shopping mall success**.
2. Yelp's overall strategy shifting their focus away from local restaurants seems to be a step in the right direction, because it is **harder to build long-lasting relationships with customers in an industry where business life cycles are shorter**
3. Even though **KNN** works best for our data, it **doesn't give us information about the direction and magnitude of each feature**, so we'll explain **variable importance with regression coefficients** instead:

a) Restaurants:

BYOB - The restaurants that do not allow BYOB are worse off, because the decreased revenue due to lost customers are more than the additional income from selling beverages. Giving customers the freedom of choice (Yes with cork fee) gives us the best performance.

Delivery - Working with delivery options seems to be a double-edged sword, offering this service creates additional workload on the cooks and servers without additional tips. For most restaurants this seems to negatively affect the service and food quality that dine-in customers are receiving.

Dogs Allowed- Allowing dogs greatly increases the chances of a restaurant remaining open in the near future according to our model. However, we feel that this may be a proxy of how well a restaurant's kitchen layout and outdoor seating areas are designed, due to the additional inspections that the restaurant has to go through during their permit application process.

b) Shopping:

Happy hour- This one is self-explanatory because setting up special discounts during special holidays both attracts customers and entices them to buy more stuff than normal.

Wheelchair access- What's worth noting here is that shopping malls that have wheelchair access as NA outperform both malls with and without wheelchair access, by a large margin too. Our hypothesis is that the former malls are in buildings or areas that are so well designed that disabled people could move around at will, and the idea of searching for handicapped-accessible pathways doesn't even cross their minds.

Sources:

Yelp investor relations - https://www.yelp-ir.com/events-and-presentations/default.aspx

SQN investors - https://www.sqnletters.com/

Simplywall.st Services - https://simplywall.st/stocks/us/media/nyse-yelp/yelp#

Yelp - https://www.yelp.com

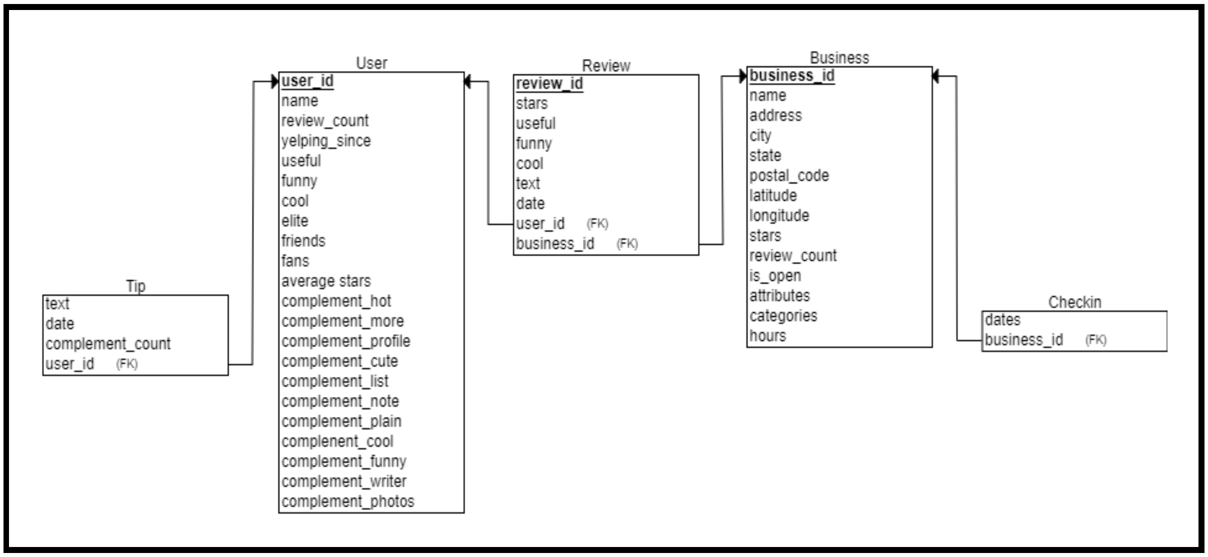
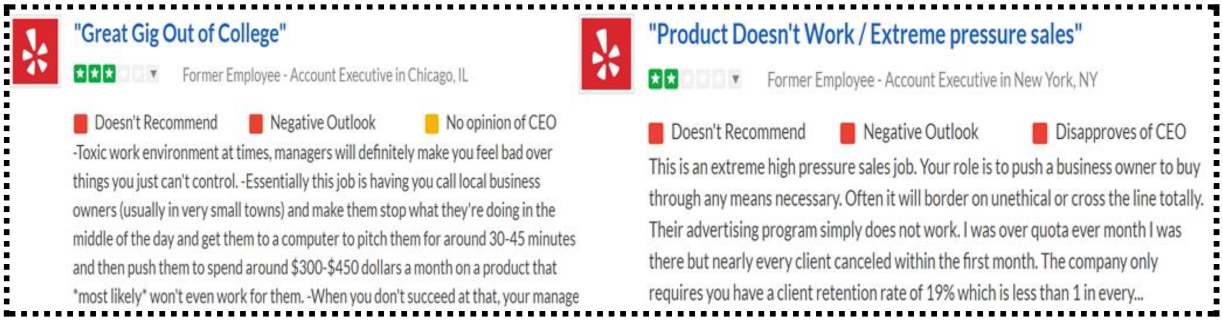


Image 1: Tables in the dataset



"Great Gig Out of College"

★★★☆☆ Former Employee - Account Executive in Chicago, IL

🟥 Doesn't Recommend 🟥 Negative Outlook 🟨 No opinion of CEO

-Toxic work environment at times, managers will definitely make you feel bad over things you just can't control. -Essentially this job is having you call local business owners (usually in very small towns) and make them stop what they're doing in the middle of the day and get them to a computer to pitch them for around 30-45 minutes and then push them to spend around $300-$450 dollars a month on a product that *most likely* won't even work for them. -When you don't succeed at that, your manage

"Product Doesn't Work / Extreme pressure sales"

★★☆☆☆ Former Employee - Account Executive in New York, NY

🟥 Doesn't Recommend 🟥 Negative Outlook 🟥 Disapproves of CEO

This is an extreme high pressure sales job. Your role is to push a business owner to buy through any means necessary. Often it will border on unethical or cross the line totally. Their advertising program simply does not work. I was over quota ever month I was there but nearly every client canceled within the first month. The company only requires you have a client retention rate of 19% which is less than 1 in every...

Image 2: Glass door reviews of Yelp

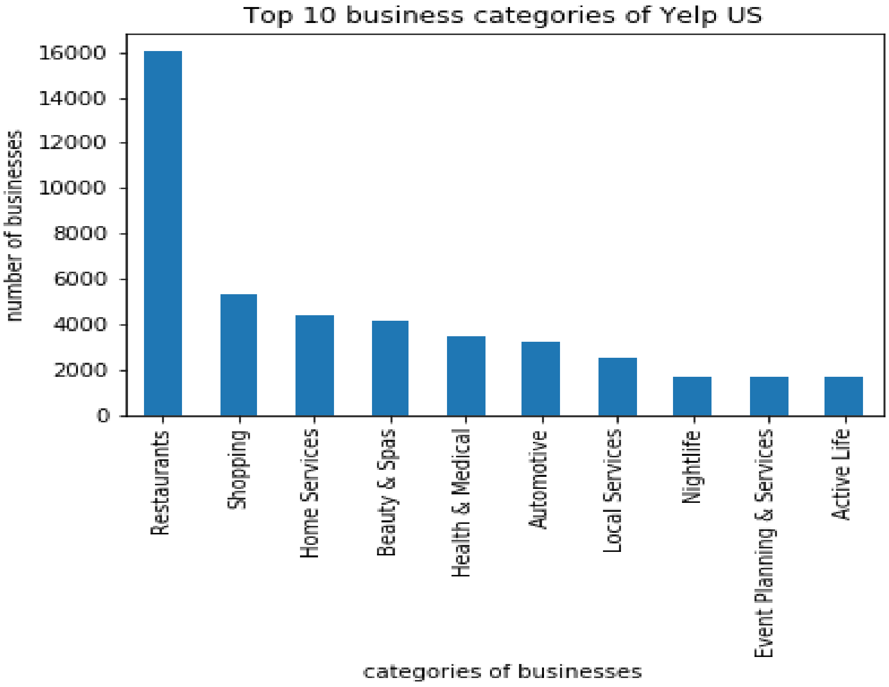


Image 3 : Top business categories in Yelp

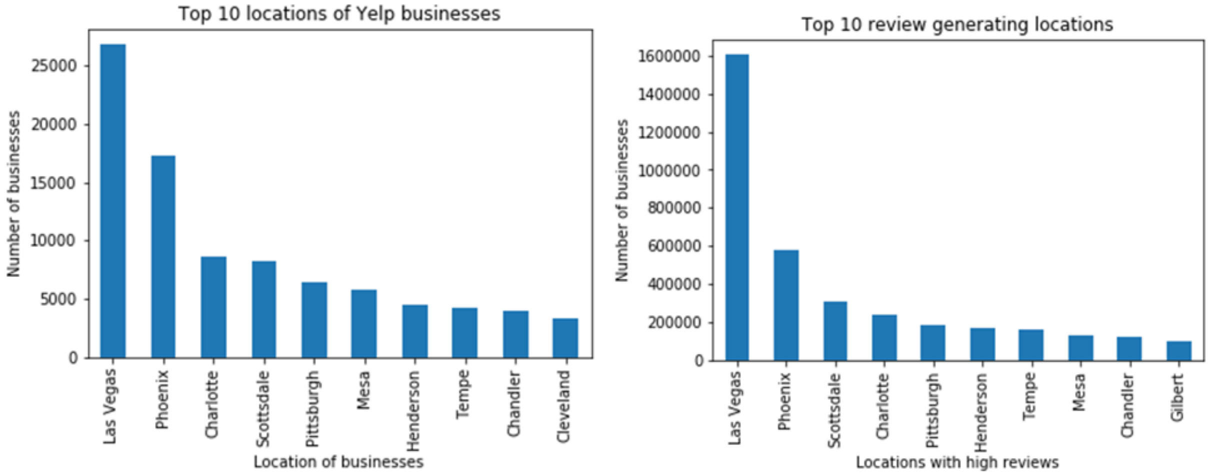
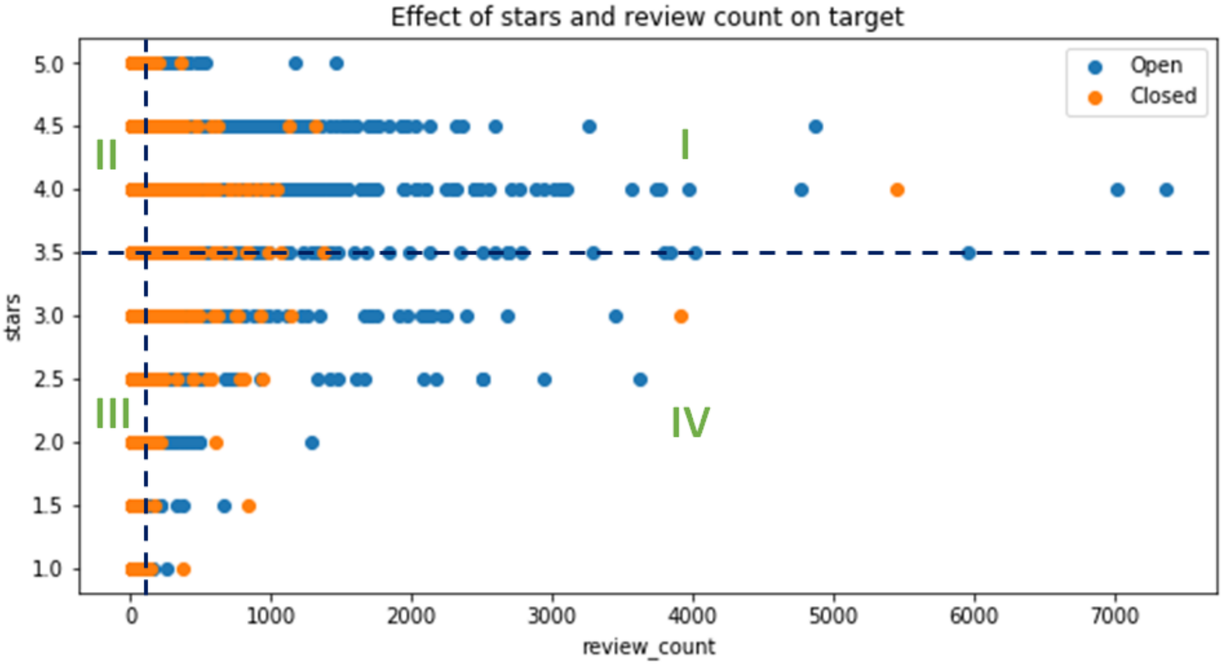


Image 4 : Top locations of Yelp businesses and top review generating locations



I : Going Great II : Increase footfall III : God save them IV : Improve service

Image 5: Effect of stars and review count on Target

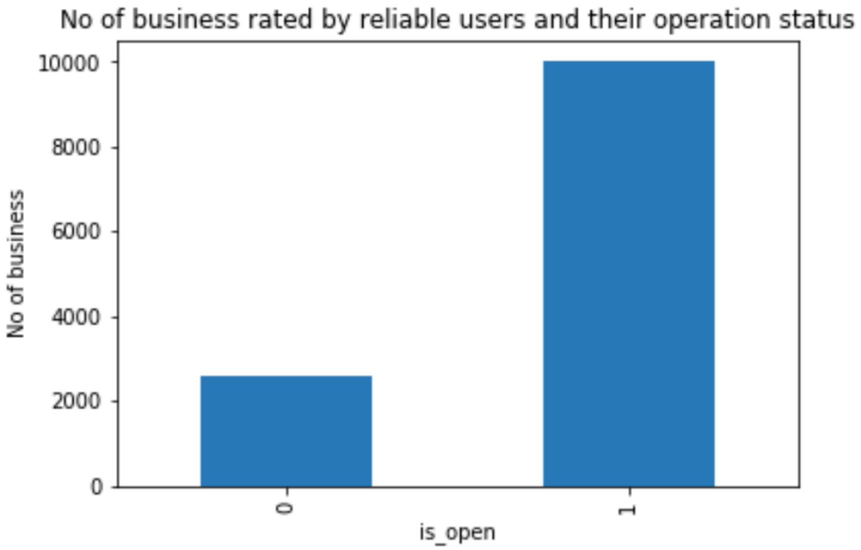
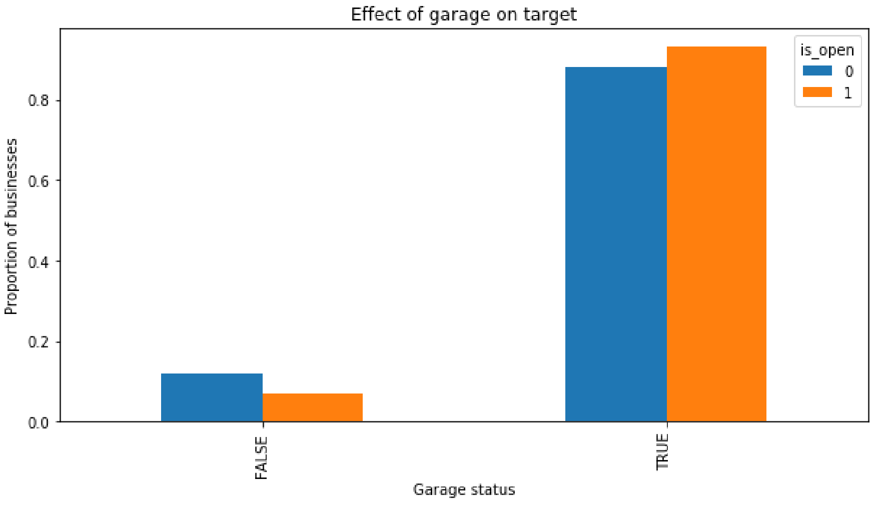


Image 6: No of business rated by reliable reviewers and their operation status



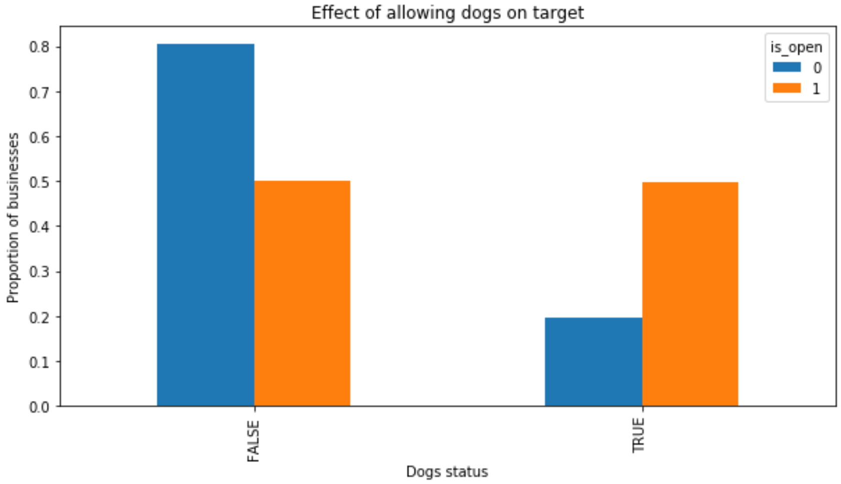


Image 7 (i): Effect of parking Garage on Target (ii) : Effect of Dogs Allowed on Target