Preventing Bar Closure

Group-15: Max Schleck, Shravan Kaul, Jiayang Wang

Introduction

- Yelp's quarterly economic status report labeled the restaurants as one of the top two failing service industries.
- Within the restaurant category, bars continually rank as the restaurant type with the highest rate of closure consistently in this post Covid landscape.
- Research reasons why bars are continuing to fail at a higher rate.

Data Cleaning

- First we preprocessed all the business and reviews json files to scrape the information.
- We chose Pennsylvania as it had the most number of businesses in both open and closed categories and wanted to stick to one demographic for best results to find reasons of bar closure in a region.

CLOSED

state	
PA	87904
FL	58308
MO	35461
LA	34526
TN	32557
IN	30183
CA	18659
NV	15342
AZ	14869
NJ	7779

OPEN

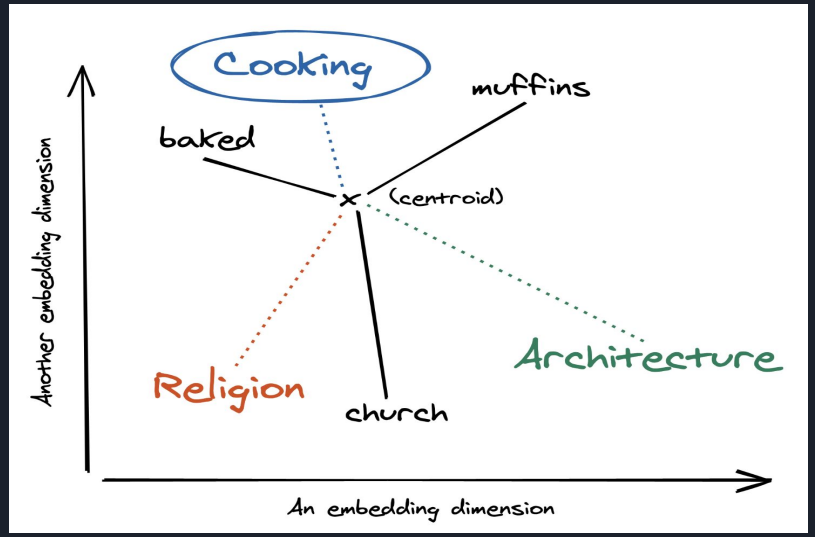
state	
PA	327963
FL	271844
LA	175574
TN	172216
MO	112721
IN	111094
AZ	80898
NV	80231
CA	74105
NJ	39575

Text Cleaning

- Tokenization, Lemmatization and Removal of Stop Words.
- Language detection to English
- Modified Stopwords list for added safety for text classification model to remove unnecessary words not related to bar aspects.
- Profanity filter words to avoid skewing of review data.

Language Processing and Sentiment Analysis

- Defining aspect labels after EDA : Ambience, Food, Drinks, Price and Service
  Creating word embedding from one of models like BERT, ELMO, Glove and Word2vec for reviews as well as aspect labels.
- Using centroid based distance calculation to classify reviews into aspects (NearestNeighnors)
- Choosing Sentiment analysis model between TextBlob, Vader & DistilBERT.



Resulting Hazards Model:

Surv(time,is_open)~Drink + Ambience + Price + strata(Service)

	Hazard Ratio:
Price	**1.19**
Ambience	**1.24**
Drinks	**1.13**

```
Likelihood ratio test= 40.85  on 3 df,   p=7e-09
Wald test            = 41.42  on 3 df,   p=5e-09
Score (logrank) test = 41.61  on 3 df,   p=5e-09
```

Assumption Checking

Proportional Hazards

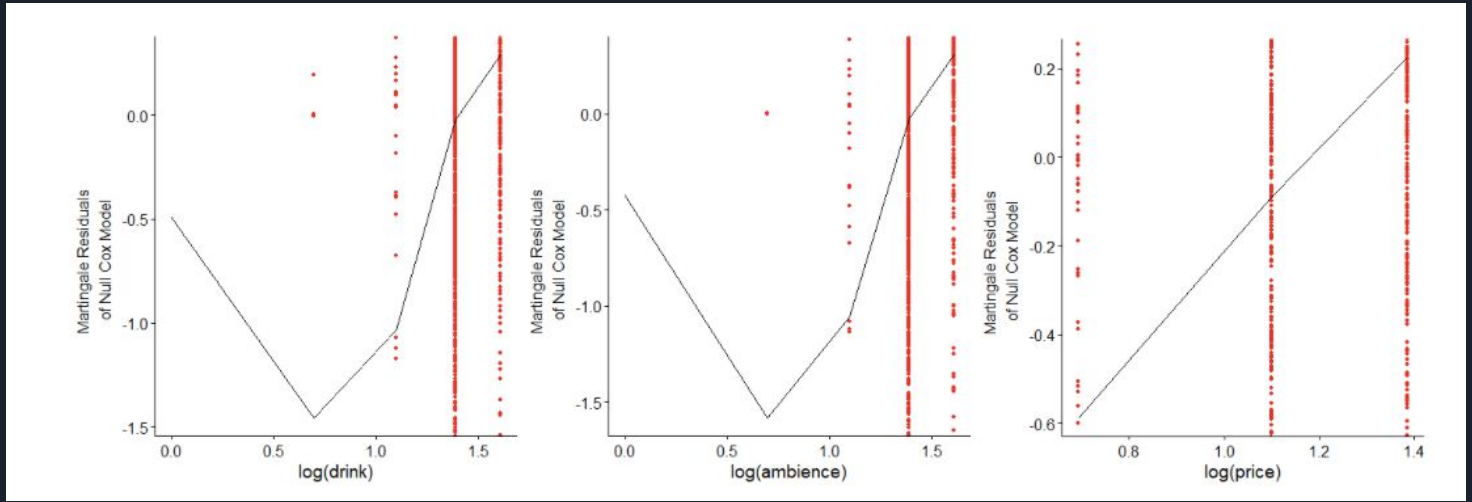
- The CPH model requires the relative hazard (or risk of closure) to be constant over time and as our other predictors vary.
- This assumption is checked by referencing the scaled Schoenfeld residuals correlation with time.
- Generating these correlation coefficients (p values in the table to the left) shows that none of our predictors are correlated at the 10% significance level.
- This assumption was met.
-

	chisq	df	p
food	2.3652	1	0.12
drink	0.0398	1	0.84
ambience	0.2869	1	0.59
price	0.0252	1	0.87
GLOBAL	4.6059	4	0.33

Assumption Checking

Linearity of log(Hazard) and Predictors

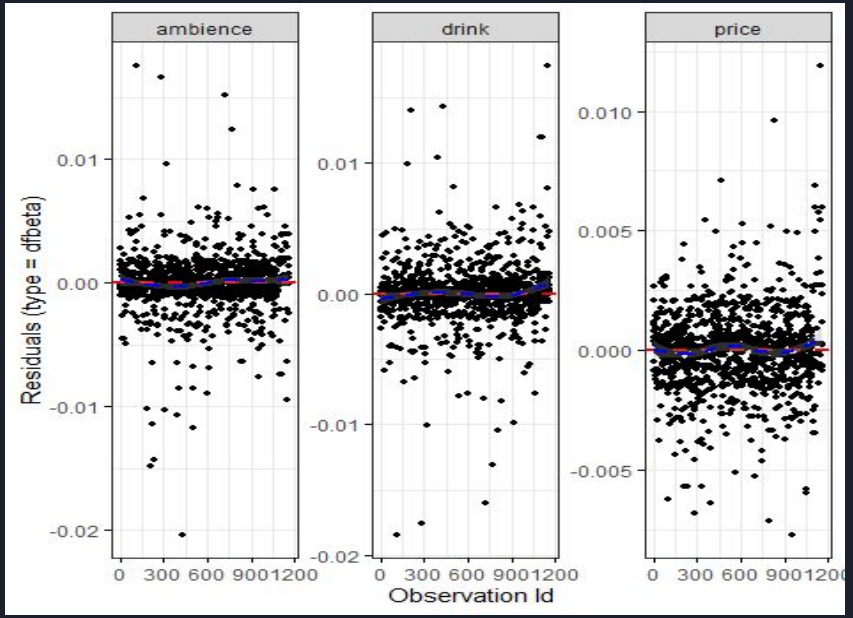
- The CPH model additionally requires there to be a roughly linear relationship between the log of our hazard and our predictors.



Assumption Checking

Influential Points

- we looked to remove potential outliers impacting our model.
- This is checked visually by reviewing a dfbeta plot. This plot represents the changes to the regression coefficients that occur when said point is removed.
- Looking at the dfBeta plots to the right we see that all points are within this threshold meaning we do not need to do any outlier removal at this point with our data.



Recommendations:

The Monday's Method!

Focus:

- Improving culture
- Happy hour/Drink Deal opportunities
- Seasonal Drink Menus

Ignore:

- Food

Future Improvements

Post Covid world calculations:

- Give the data 2 years to collect

Broaden state scope

Multiprocessing NLPs

Statestreet on a Saturday at 2pm during Isolation