



# Predict Business Closure With Yelp Data



An Exercise In Supervised Learning By: Emy Parparita



# Objective

Predict business closure using Yelp data.

## 1. Diablo's Cantina - CLOSED

☆☆☆☆☆ 1135 reviews

\$\$ · Mexican, Bars

The Strip

3770 Las Vegas Blvd S  
Las Vegas, NV 89109  
(702) 730-7979



Although this location is now closed, it's still worthy of a review! D = Directly on the strip, attached to the Monte Carlo! Easy to access ... [read more](#)

## 1. Roy's Restaurant - CLOSED

☆☆☆☆☆ 457 reviews

\$\$\$ · Sushi Bars, Seafood, Steakhouses

Eastside

620 E Flamingo Rd  
Las Vegas, NV 89119  
(702) 691-2053



Overall, this was a great experience! We went for a birthday, and the staff goes above and beyond to make your whole party feel special. I'm not sure... [read more](#)

## 1. Zeffirino - CLOSED

☆☆☆☆☆ 406 reviews

\$\$\$ · Italian, Seafood, Beer Bar

The Strip

3377 Las Vegas Blvd S  
Las Vegas, NV 89109  
(702) 414-3500



Food: The food was excellent. Rating: Four stars Service: Very attentive while being efficient and friendly. Rating: Five stars Atmosphere: Very... [read more](#)

## 1. So Good Cafe - CLOSED

☆☆☆☆☆ 82 reviews

\$ · Vietnamese, Cafes, French

Chinatown

5600 Spring Mountain Rd  
Las Vegas, NV 89146  
(702) 818-5955



Wow! So tasty. I walked in and just told them to bring me something good. So I got the traditional Vietnamese coffee (that'll give you a caffeine... [read more](#)

# Data Source

---

- The Yelp [Dataset](#), as a SQL dump (MySQL): “... is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes”
- Relevant Tables:
  - attribute
  - business
  - category
  - review

# Schema

## attribute

Field	Type
id	int(11)
business_id	varchar(22)
name	varchar(255)
value	mediumtext

## category

Field	Type
id	int(11)
business_id	varchar(22)
category	varchar(255)

## business

Field	Type
id	varchar(22)
name	varchar(255)
neighborhood	varchar(255)
address	varchar(255)
city	varchar(255)
state	varchar(255)
postal_code	varchar(255)
latitude	float
longitude	float
stars	float
review_count	int(11)
is_open	tinyint(1)

## review

Field	Type
id	varchar(22)
business_id	varchar(22)
user_id	varchar(22)
stars	int(11)
date	datetime
text	mediumtext
useful	int(11)
funny	int(11)
cool	int(11)

# Data Selection

## Cities with most businesses

city	business_cnt
Las Vegas	26809
Phoenix	17213
Toronto	17211
Charlotte	8554
Scottsdale	8228
Pittsburgh	6355
Montréal	5973

## Cities with most reviews

city	review_cnt
Las Vegas	1605343
Phoenix	576709
Toronto	430985
Scottsdale	308529
Charlotte	237118
Pittsburgh	179471
Henderson	166884

Will use **Las Vegas, Restaurants** category (most likely to be reviewed), reviews after **2015/01/01**.

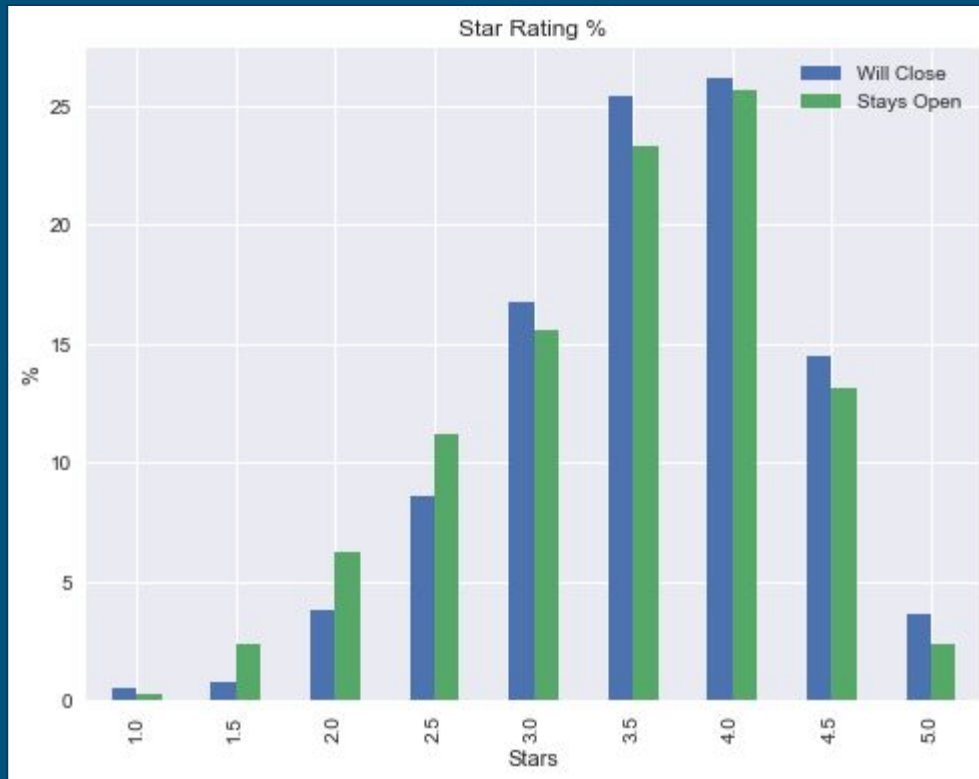
Number of selected businesses: **4085**, out of which **795** are closed

Train/Test Split: **80/20**, indexed by **business\_id**

# Stars Relevance

Is the average star rating a good predictor?

It doesn't look that way.



# Feature Selection

Feature	Desc	Indicator	Engineered	Keep
bptr_cnt	Number of similar businesses/neighborhood	✗	✓	✓
cat_...	Category	✓	✗	✗
nbr_...	Neighborhood	✓	✗	✗
rating	Global -1/0/1 based on stars binning	✗	✓	✓
review_count	Total number of reviews	✗	✗	✓
review_sentiment	Most recent N/% reviews, converted to -1/0/1 sentiment and summed up	✗	✓	✓
same_name_cnt	Number of businesses with the same name (chain?)	✗	✓	✓
stars	Global number of stars	✗	✗	✗
zip_	Postal code	✓	✗	✗

# Target and Score Selection

---

**Target:** `will_close`, 0/1 derived as  $(1 - \text{is\_open})$

**Scores:**

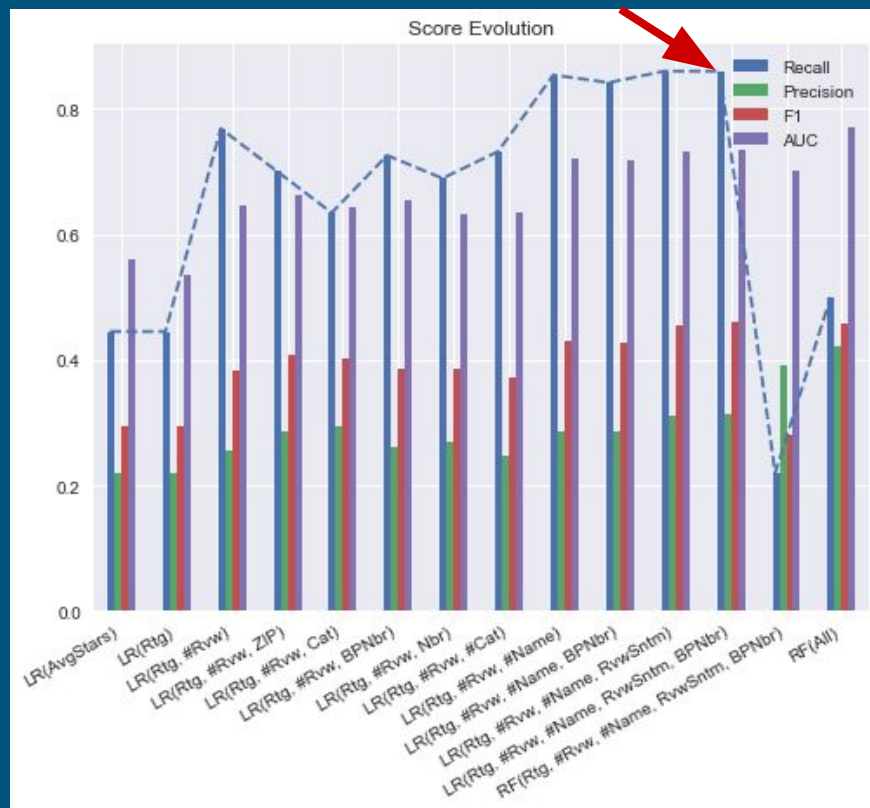
- **Recall** (the most important score, false positives are less risky than false negatives)
- Precision
- F1
- AUC



# Feature Trials

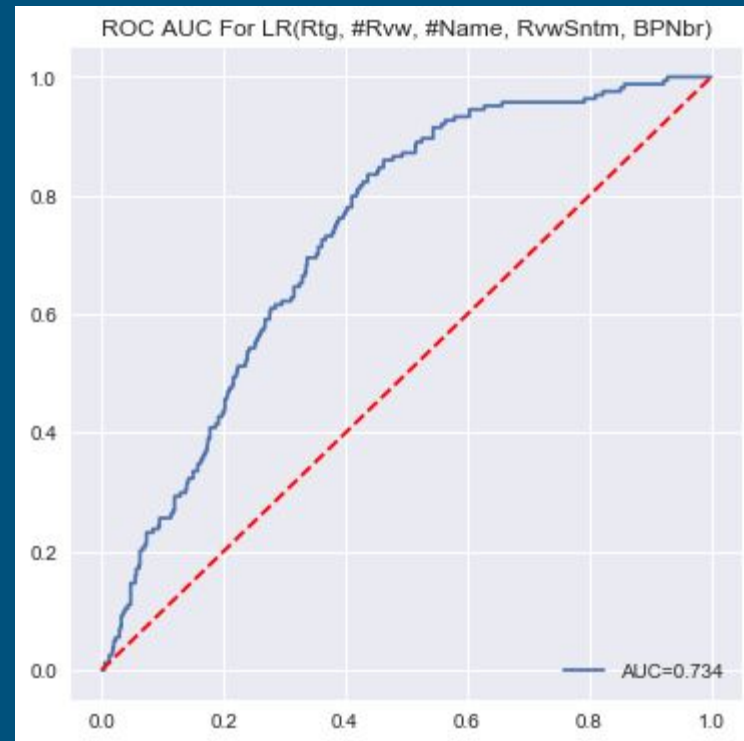
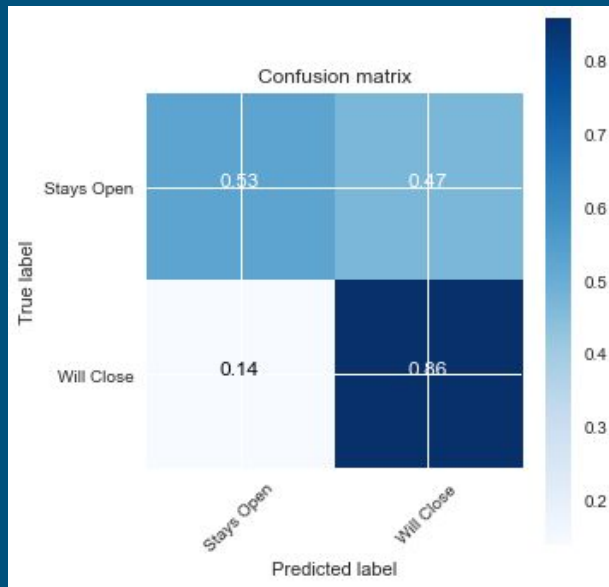
Different sets of features and models were tried:

- Location (ZIP, neighborhood) and categories do not help
- Logistic Regression has better recall than RandomForest
- Precision is low



# Final Test Score

Recall	0.860
Precision	0.314
F1	0.460
AUC	0.734



# Conclusions

---

- The model could be deployed as a Web App to be used by owners to check if their business is at risk of closing
- Unfortunately the current implementation has very low precision, “cry wolf” syndrome would lead to dismissal
- Possible refinements:
  - NLP for review and tips parsing
  - Better feature engineering
  - Include economic data, e.g. changes in lease or energy costs
  - Include demographics data

# Fun Fact

---

Feature importance as reported by RandomForest running with all features:

zip_93013	0.152091	
bpmr_cnt	0.118295	
zip_89199	0.102510	
zip_89183	0.038803	
cat_brazilian	0.031475	
cat_seafood	0.029128	
zip_89179	0.025255	
cat_hot_dogs	0.011923	
<b>cat_hookah_bars</b>		<b>0.011857</b>
zip_89121	0.010710	
<b>rating</b>		<b>0.010450</b>