UNIVERSITY
OF MIAMI

Survival Analysis for Small Business: Evidence from Yelp

Alan Huang, Xiru Pan

Introduction

Restaurant Survival Prediction

- Restaurant sector plays a significant role by contributing over **$833 billion** to the economy.
- The turnover rate for restaurants in their first year can reach **26%**.

Build Machine Learning Model to Predict the Survival of Restaurants

Introduction

Online review platform is important in shaping customer purchase decision.

- **90%** of consumers of small businesses used the internet to find a local business near them
- As many as **93%** of people say that online review platform affects their buying decision.
- Prior research shows that online review positively relate to revenue and customer satisfaction, but few studies have investigated how online review affect business survival

Use Online Review + Business Attributes Information to Predict Restaurant Survival

Related Works

Business survival/failure prediction

- Previous research use ratio analysis (Beaver, 1966), Altman Z-scores (Altman, 1967), and Bayesian models (Sarkar and Sriram, 2001) to predict bankruptcy
- Rely on financial metric, and suitable for medium to large corporation
- These models focus on finance management but offer limited operational strategies
- Business's success or failure also heavily depends on its industry context

Survival/failure in the restaurant industry

- Prior research shows that food quality, consistency, franchising, adaptability, and marketing strategies (Lee 1987; Nizam 2017; Wang & Kim 2021)
- Our objective is to utilize customer-generated reviews and ratings, combined with various business attributes, to develop a machine learning model that forecasts restaurant survival

Data

- We collect the dataset from *Yelp.com*.
- Yelp has become a major force in local commerce, with **186 million** people using it each month to post nearly **150 million** business reviews

- The dataset includes 150,346 restaurants across 1,416 cities in 27 states.
- 6.9 million reviews, 908,915 tips, 1.9 million users

Data Preprocessing

- Merge datasets (business data, tip data, and review data)
- Binary variables were converted to integers
- Inconsistent data entries were standardized
- Replace missing value with NaN.

Variable Description

Variable	Description
is_open	integer, 0 or 1 for closed or open, respectively
stars	integer, star rating
review_count	Integer, total review received
ByAppointmentOnly	Categorical, whether it is by appointment only
BusinessAcceptsCreditCards	Categorical, whether accept credit card,
BikeParking	Categorical, whether bike parking is available
WiFi	Categorical, whether it provide free, paid or no wifi
HasTV	Categorical, whether it has TV or not

Machine Learning Models

Generalized Linear Models

- **Logistic Regression**

It models the probability that a given input belongs to a particular category (e.g., a restaurant will survive or fail) by using the logistic function to ensure that the output lies between 0 and 1.

$$log\left(\frac{p}{1-p}\right) = X\beta$$

- **Probit Regression**

The probit model is similar to logistic regression but uses the cumulative distribution function of the normal distribution to link the linear predictors to the probability of the binary outcome.

$$\Phi^{-1}(p) = X\beta$$

Machine Learning Models

Support Vector Machines

•Support Vector Machines (SVM) are a powerful and versatile class of supervised machine learning algorithms used for classification.

•SVM can handle both linear and non-linear boundary by using different kernel

Linear Discriminant Analysis/Quadratic Discriminant Analysis

•LDA/QDA are classification models as well as a dimensionality reduction technique

•LDA assumes that the probability distribution of the input features is Gaussian, that each class has the same covariance matrix

•QDA allows each class to have its own covariance matrix

Decision Tree

•It is a tree-structured model where internal nodes represent the features of a dataset, branches represent decision rules, and each leaf node represents an outcome.

Results Comparison

Model	Accuracy
Logistic Model	0.8035
Logistic Model with Quadratic Terms	0.8037
Probit Model	0.8025
Probit Model with Quadratic Terms	0.8029
SVM with Linear Kernel	0.7954
SVM with Radial Basis Kernel	0.8007
LDA Model	0.8008
QDA Model	0.7391
Decision Tree Model	0.7954

- Accuracy evaluated based on 5-fold cross validation
- Logistic model with quadratic term have higher accuracy
- Non-linearity relationships between the features and the response variable

Future Work

Expansion of Feature Set

- Customer sentiment analysis from reviews, geographical factors, competitive landscape analysis, and temporal features such as seasonality

Utilization of Diverse Performance Metrics

- Precision, Recall, and F1

Incorporation of Additional Machine Learning Models