Chapter 21
Using Machine Learning to Predict Business Survival in COVID-19

Dhruv Gargi and Riddha Mathur

Abstract The global recession due to the pandemic has knocked the business landscape and brought the world to its knees. There were a number of renowned companies that made the headlines for being the top industry hard hits. Nonetheless, there were businesses that survived this pandemic and navigated the COVID complexities so effectively that it tipped the scales in their favor. We attempt to study the factors that helped these businesses masterfully work their way through the conundrums of coronavirus pandemic. We first build a dataset that entailed information pertinent to businesses and relevant COVID-related information that was sourced from Yelp and other platforms. We used a variety of classifiers to make predictions about the survival of these businesses followed by that after assessing their performance through varied methods. The model efficiency was classified based on several rating techniques to evaluate both underperforming and profitable businesses.

Keywords Global recession · Coronavirus pandemic · Business survival · Feature engineering · Machine learning · Feature importance

21.1 Introduction

The success of a business is dependent on the axiom that it stays open. The COVID-19 pandemic poses an existential threat to small businesses, with more than 400,000 lost since the crisis began [7]. The outbreak of Novel Coronavirus disease is a grave menace to the entire world affecting millions of people [11]. Technology adoption has a critical role for business survival during the COVID-19 crises especially with small businesses [1]. As compared to large firms, new startups and small firms show high flexibility in their reactions to the crisis, partly due to the low level of bureaucracy

D. Gargi (✉)
Manipal University, Manipal, India
e-mail: dslisanoob@gmail.com

R. Mathur
Goa Institute of Management, Sanquelim, Goa, India
e-mail: riddha15mathur@gmail.com

S. Bhattacharyya et al. (eds.), *Human-Centric Smart Computing*, Smart Innovation, Systems and Technologies 316, https://doi.org/10.1007/978-981-19-5403-0_21

and limited social responsibility compliance [4]. However, small businesses are more susceptible to cash flow problems created by the COVID-19 pandemic putting them in jeopardy of survival [6]. The recent and still scarce literature in this field seeks to provide suitable solutions to prevent irreparable disruption and help strengthen business, but does not apply advanced statistical methods to that end [5]. In this project, we develop several machine learning models to predict and analyze whether a business remains open or permanently closes during times of COVID-19.

We conduct research on a dataset containing records of 153,843 American businesses shown in Yelp comprising of multiple segments including COVID-19 features. The records comprise of information such as location, reviews, number of stars, business attributes, business categories, and COVID-19 features like delivery or takeout, virtual services offered, COVID Banner, etc. We then employ the technique of feature engineering to apply appropriate transformations to create a parsed dataset that can be utilized for training and evaluation. Essentially, feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be instrumental in machine learning. In this process, we also add data from external sources namely median income and population for the postcode the corresponding business is located in.

It is important to note that COVID-19 also indirectly affects businesses through some of the other features not cited as COVID features. For instance, the postcode median income feature. With COVID-19, increased job losses will very likely translate to lower median incomes in a locality and if median income relative to postcode is an important feature, it will have a significant impact on business survival odds.

We devise a pipeline to effectively upsample (SMOTE-NC) the minority class (closed businesses, which takes up around 20% of original dataset), train, validate, and test the performance of our classifiers, including K-Nearest Neighbors, Logistic Regression, Random Forest, and Neural Network. The results for Naive Bayes and Support Vector Machine are shown in the Appendix.

We assess feature significance using the Permutation Importance algorithm provided by `eli5` library to determine the 10 most important features in this classification task.

We believe that this paper elevates general understanding of the impact of COVID-19 on American businesses and the economy as whole. The results of our paper highlight the effect that this pandemic will have on businesses across the country. Moreover, this paper can serve as insight to people making business decisions in the time of COVID-19.

21.2 Related/Similar Work

After reviewing the literature, we derived information from a variety of research papers in enhancing our research. The works reviewed and their implementation in this research can be described as follows:

21.2.1 COVID-19 Impact on Businesses

The outbreak of the Coronavirus disease 2019 (COVID-19) has had significant ramifications for businesses of all sizes. Businesses are required to navigate through the financial and operational challenges or else face the prospect of imminent closure. For instance, an Accenture study details how businesses are responding to this unique challenge. Another study conducted a survey on 5,800 small businesses to study the effect of COVID.

With the onset of the coronavirus pandemic, hotel occupancy dropped significantly and Hotel chains were forced to try new approaches to make money and maintain the cash flow. Lau, Arthur, et al. [10] applied DeLone and McLean's Information System Success Model to examine the adopted digital technologies. Katare, Bhagyashree, et al. [8] state that drivers of income loss are not necessarily associated with time to recovery and businesses that are undercapitalized are more likely to suffer higher income loss, longer time to recovery, and less likely to be resilient.

A study by Adam, Abdalla, Nawal, et al. [2] shows that the structural equation modeling results that the innovation practices SMEs adopt to face the repercussions of COVID-19 has a positive impact on the performance and likelihood of business survival.

Adejare, Bimbo Onaolapo et al. [3] examine COVID-19 pandemic and business survival as a mediation on the performance of firms in the Fast moving consumer goods (FMCG) sector: insight for the future of business operation.

21.2.2 Machine Learning-Based Classification Models

The problem of whether a business will survive is a typical classification problem. Various machine learning models have been proposed to address the general classification problem, such as K-Nearest Neighbors, Naive Bayes, Logistic Regression, Decision Tree, and Random Forest. Among them, Neural Networks, are recently proposed and demonstrated to be a powerful model in terms of the prediction accuracy. However, it is known to lack interpretability, *i.e.*. a black box, which means that it is hard to get the confidence interval or the lower bound for the final results. Although some works proposed to interpret the inner mechanism of neural networks, the uncertainty estimation of deep neural networks is still an open problem.

21.3 Dataset

We make use of the Yelp Dataset[1] to solve our question. We draw features from the following segments of the dataset:

[1] https://www.yelp.com/dataset.

Table 21.1 Features and their representation

Feature	Representation
COVID-19 features	Binary variables
Location	Continuous variables
Stars	Continuous variables
Review Count	Continuous variables
Chain	Binary variables
Business attributes	Ordinal categorical variables and binary variables
Business categories	Binary variables
Income postcode	Ordinal categorical variables
Population postcode	Continuous variables

Data including label, location, number of stars, review count, business attributes, and business categories. COVID data relevant to a business such as virtual services offered, COVID Banner.

We make use of the features in the datasets to perform essentially a join operation to obtain joint unique records based on business id.

In this process, we also perform feature engineering on raw datasets to obtain a final parsed dataset comprising of appropriate features that can be used to develop our classifiers.

21.3.1 Feature Engineering

In essence, the notion that we can represent predictors in several ways in a model and that some of these representations are better than others, leads to the idea of feature engineering. It is basically the process of creating representations of data that increase the effectiveness of a model [9]. In this paper, we utilize the features in the above mentioned Yelp dataset to perform a join operation and do feature engineering to procure a final parsed dataset (Table 21.1).

- **COVID-19 Features**: We directly use the features' highlights, delivery or takeout, Grubhub enabled, Call To Action enabled, Request a Quote Enabled, COVID Banner, Temporarily Closed, and Virtual Services Offered from the COVID features dataset. These features are represented as binary variables.
- **Location**: The features latitude and longitude are used to represent a business location. These two features are represented as continuous variables.
- **Stars**: It refers to the rating of a business. Since float values are accepted, this feature is represented as a continuous variable.
- **Review Count**: It denotes the number of Yelp reviews for each business. This feature is represented as a continuous variable.

- **Chain**: We use a simple Natural Language Processing trick to determine whether a business (uniquely identified by its id) is a chain or not. We use the business name feature and count the number of times each business name appears in the dataset. If it only appears one time, the corresponding business (with respect to its id) is not a chain. If it appears more than once, the corresponding business is a chain. This feature is represented as a binary variable.
- **Business Attributes**: We use the following attributes as features in our models: Business Accepts Credit Cards, Bike Parking, WiFi, Business Parking, Offers Alcohol, Has TV, Noise Level, Price Range, and Outdoor Seating. Noise Level and Price Range are ordinal categorical variables. The remaining features are represented as binary variables. When records do not have the above attributes, features represented by binary values are set to 0 (false). For Noise Level and Price Range (both range from 14), the value 2 is assigned.
- **Business Categories**: The Yelp dataset has a feature called "categories". In spite of this name, relative to the official Yelp category list (link below), this feature provides a list of sub-categories for each business. Using the above Yelp category list, we map each sub-category to its main category in order to reduce model dimensionality and sparsity. Based on the mapping, each business is assigned categories from the following category list: Active Life, Arts and Entertainment, Automotive, Beauty and Spas, Education, Event Planning and Services, Financial Services, Food, Health and Medical, Home Services, Hotels and Travel, Local Flavor, Local Services, Mass Media, Nightlife, Pets, Professional Services, Public Services and Government, Religious Organizations, Restaurants, and Shopping.
  These categories are features represented as binary variables.
  For example, if a business in the Yelp dataset has `Aquariums and Museum` as its feature value for "categories", we utilize the Yelp category list to map Aquariums to the main category of Active Life and Museum to the main category of Arts & Entertainment. We set the corresponding binary variables to 1 and the remaining ones to 0.
- **External Sources** to boost business data:
  In the Yelp Dataset, each business has a corresponding American postal code. We make use of this to add the features of median income and population relative to the business' postcode.

(1) **Income Postcode**: The median income for a postcode is obtained from the Individual Income Statistics released by the IRS. The IRS releases income data as brackets and not exact figures. This feature is an ordinal categorical variable. For instance, a value of 1 is indicative of income between $1 and $25,000, 2 is between $25,000 and $50,000, 3 is between $50,000 and $75,000, 4 is between $75,000 and $100,000, 5 is between $100,000 and $200,000, and 6 is income greater than $200,000.

(2) **Population Postcode**: The population for a postcode is obtained from the Zip Code Database. This feature is represented as a continuous variable.

Attributes of final dataset that we will use to develop and evaluate our classifiers:

- N: 153843 American businesses,
- D: 45 features
- y: is_Open (1: business is open, 0: business is closed)
- **x**: all the features mentioned above

21.4 Approach

21.4.1 Preprocessing

In Sect. 21.3.1, we address the issue of preprocessing to obtain our final, parsed dataset that we can use to train and evaluate our models. We employ some more data preparation techniques like SMOTE and Min-Max normalization in our Training & Evaluation Pipeline.

21.4.2 SMOTE

In the Yelp Dataset, only around 20 percent of American Businesses are listed as closed. To address this class imbalance, we make use of SMOTE to generate synthetic samples to obtain a balanced class distribution while training, which helps boost our models' generalization ability.

SMOTE-NC stands for Synthetic Minority Over-sampling Technique for Nominal and Continuous features, which is a class balancing technique based on nearest neighbors judged by Euclidean Distance between data points in feature space. SMOTE-NC is taken from the `imbalanced-learn` library in `sci-kit learn`, which helps to create synthetic data for categorical as well as quantitative features in the dataset.

Specifically, we apply SMOTE-NC to the training set only and not the validation or test set. This is done to prevent the bleeding of information. The SMOTE algorithm creates synthetic data points by utilizing nearest neighbors of samples. Thus, if the minority class' (closed businesses) nearest neighbors end up in the validation or test set, the synthetic data points in the training set partially capture their information. Therefore, if doing so, we would get over-optimistic values for the tenfold cross validation and test set accuracy.

21.4.3 Min-Max Normalization

We acquire insights from an article[2] to apply Min-Max normalization appropriately to avoid data leakage in the training and evaluation pipeline for the classifiers that

[2] https://machinelearningmastery.com/data-preparation-without-data-leakage/.

employ data normalization (KNN, Logistic Regression, and Neural Nets). Since our KNN model uses distance to compare feature values, it is important for features to be scaled to the same range to avoid unfairly over-weighting or under-weighting features. Normalization is applied for logistic regression to speed up solver convergence. It is also applied for Neural Networks to make the training process more stable and final classification accuracy higher. Normalization is not applied for Random Forest. The application of Min-Max normalization in the Training & Evaluation Pipeline is explained further in Sect. 21.4.4.

21.4.4 Training and Evaluation Pipeline

We adopt the following approach to train and evaluate our classifiers:

- The final parsed dataset is divided into the train and test set using the classic 80:20 split.
- The train set is then used to perform tenfold cross validation using balanced accuracy metric (explained in Sect. 21.4.6) to inform model selection.
- In each fold of cross validation, the train set is divided into the train fold set and the validation fold set. SMOTE-NC is applied on the train fold set to upsample the minority class (closed businesses). If required, Min-Max normalization is fit to the upsampled training fold set and the transformation is applied to the upsampled training fold set and the validation fold set. The classifier is trained on the upsampled training fold set and then evaluated on the validation fold set.
- Wherever possible and appropriate, cross validation is used to guide model parameter selection.
- After having selected model parameters through tenfold cross validation, we then upsample the entire train set using SMOTE. If required, Min-Max normalization is fit to the upsampled train set and the transformation is applied to the train and test sets. We then train the classifier. This classifier is then evaluated on the unseen testing data.

21.4.5 Machine Learning Models

21.4.5.1 KNN

Library: `sklearn.neighbors.KNeighborsClassifier`
KNN variant: Distance Weighted Nearest Neighbors
K value: Given the large dataset, we use the most commonly used value for K, which is the $\sqrt{N}$, where N is the number of samples in the train set (the upsampled train set in this case). N is 200110. Thus, K is set to 450.

Distance Metric: We use Euclidean based on tenfold cross validation results. During our experiments, we get better results with Euclidean distance than with Manhattan distance.

21.4.5.2 Logistic Regression

Library: `sklearn.linearmodel.LogisticRegression`
Convergence Algorithm: Stochastic Average Gradient (SAGA)

We selected this convergence algorithm because it performs better than other solvers during our assessment with tenfold cross validation.

21.4.5.3 Random Forest

Library: `sklearn.ensemble.RandomForestClassifier`

Number of Decision Trees: 100

This algorithm draws 80% of the input training set and uses bootstrap sampling to construct decision tree.

Splitting Criterion: Information Gain

Stopping Criteria: (1) Less than or equal to 500 samples at node (decided based on tenfold cross validation). (2) No more splits left. (3) All samples at node have same label.

21.4.5.4 Neural Network

Library: `PyTorch`

We use the Multi Layer Perceptron Artificial Neural Network as our model, which is a combination of 5 fully connected layers and activation functions called ReLU. Here we choose cross entropy as the loss function, and utilize the learning rate of 1×10^{-3}, batch size 1024, 20 training epochs and Adam optimizer to conduct the experiments.

21.4.6 Metrics

Here we show some important definitions:
True Positive Rate (Sensitivity): Proportion of businesses classified correctly as open from the set of open businesses.
False Positive Rate: Proportion of businesses classified incorrectly as open from the set of closed businesses.
True Negative Rate (Specificity): Proportion of businesses correctly classified as closed from the set of closed businesses.

Table 21.2 Model evaluation

Classifier	Tenfold CV balanced-accuracy (%)	Test balanced-accuracy (%)	Test Vanilla accuracy (%)
KNN	71.20	70.82	77.45
Logistic Regression	70.34	70.36	76.87
Random Forest	70.41	70.56	82.65
Neural Network	71.32	71.72	79.31

False Negative Rate: Proportion of businesses classified incorrectly as closed from the set of open businesses.
Balanced Accuracy: This is defined as the average recall obtained on each class (average of sensitivity and specificity). We use this metric across both the tenfold cross validation and test data because the validation and test sets are imbalanced. The goal is to find a classifier that performs well across both classes (open and closed businesses) and this metric helps us in achieving this.
Vanilla Accuracy: The common accuracy metric is used on the unseen test data as well to add further perspective to the results.
Class-specific metrics: We also compute precision, recall, and F1 score on the unseen test data to assess how our classifiers perform with respect to each of the two classes.

21.4.7 Library Usage

We use libraries for our classifiers instead of using the code from class for the following reasons:

- Code from libraries is cleaner and easier to organize given that only function calls are involved.
- Using libraries allows model parameters to be easily updated.
- In the case of logistic regression, the sci-kit learn model converges faster than the code used for class.

21.5 Results

21.5.1 Model Evaluation

From the ML models that we use to check for business survival including KNN, Logistic Regression, Random Forest, and Neural Net, we get the following results:

Table 21.3 Model evaluation on each class

	Class 0 (Closed businesses)			Class 1 (Open businesses)		
Model type	Precision (%)	Recall (%)	F	Precision (%)	Recall (%)	$F1$
KNN	41.84	60.38	49.43	90.19	81.26	85.49
Logistic regression	40.90	60.10	48.67	90.05	80.62	85.07
Random forest	52.50	51.51	52.00	89.22	89.60	89.41
Neural network	44.97	59.77	51.33	90.31	83.67	86.86

From Table 21.2 we can see that all models have relatively similar performance on the test balanced-accuracy metric with the Neural Net performing slightly better than the rest. The Random Forest classifier has the best performance according to the vanilla accuracy metric.

21.5.2 Class-Specific Metrics

Across the board, the models are better at predicting open businesses than closed ones. All models have higher precision, recall and $F1$ scores for Open Businesses (Table 21.3).

Random Forest is the best according to the sensitivity metric, KNN is the best according to the specificity metric, and as mentioned above, Neural Network is the best when balancing the sensitivity and specificity metrics (balanced accuracy).

21.5.3 Feature Importance

Algorithm: Permutation Importance, Library:
This method determines feature importance by assessing how much the score for the respective metric (we use the balanced accuracy metric) decreases when the values for a feature are randomly shuffled.

We use this method because it is model agnostic and allows us to determine the important features for various models (did not use neural networks and KNN since they take too long). In this paper, we present the important features for the Random Forest model, which achieves the third highest balanced accuracy on the test set. The following are the 10 most important features in determining business survival:

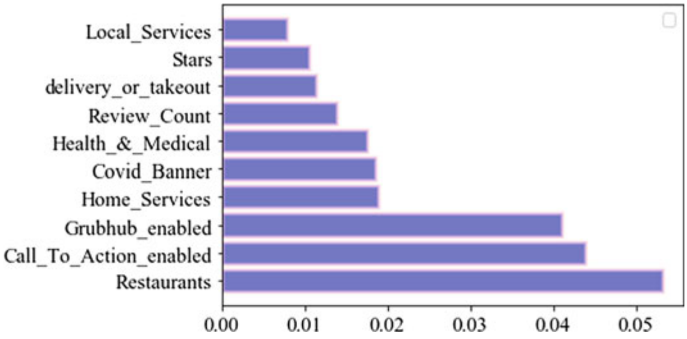

Fig. 21.1 The Top 10 most important features. For example, a weight of 0.05 means balanced-accuracy decreases by 5% without the corresponding feature

(1) Restaurants

(2) Call_To_Action_enabled

(3) Grubhub_enabled

(4) Home_Services

(5) Covid_Banner

(6) Health_&_Medical

(7) Review_Count

(8) delivery_or_takeout

(9) Stars

(10) Local_Services

Based on Fig. 21.1, COVID features like `Call_To_Action_enabled`, `Grubhub_enabled`, `Covid_Banner`, and `delivery_or_takeout` are important in this classification task. It is also worth noting that Health_&_Medical is an important feature.

21.6 Conclusions and Future Work

21.6.1 Conclusions

In this paper, we build a comprehensive business dataset with rich features, collected from diverse information sources. Then we formulate the problem of predicting business survival and employed several machine learning methods to solve it. In this process, we develop an exhaustive training and evaluation pipeline combining techniques like SMOTE, cross validation and make use of multiple performance metrics. Results show that all models are significantly better at classifying open businesses than closed ones. Finally, we determine the 10 most important features that influence business survival. The presence of four COVID features suggests that a business' survival prospects are reliant on providing services that ensure customer health and safety. For, e.g., restaurants having Grubhub enabled. We hope that our work can help save the economy by informing and guiding business owners, investors and other people in power to make wise business decisions.

21.6.2 Future Work

The Yelp dataset has a file that contains reviews written by users for businesses. Natural Language Processing can be leveraged and this data can be used to perform sentiment analysis. This will give us a sense of what customers think about the business and this sentiment can be modeled as a feature to help us possibly obtain better classifiers. In addition to this we would also like to gather a larger dataset with more countries in order to build models that generalize better and aren't limited to specific countries.

References

1. Abed, S.S.: A literature review exploring the role of technology in business survival during the covid-19 lockdowns. Int. J. Org. Anal. (2021)
2. Adam, N.A., Alarifi, G.: Innovation practices for survival of small and medium enterprises (SMES) in the COVID-19 times: the role of external support. J. Innov. Entrepreneurship **10**(1), 1–22 (2021)
3. Adejare, B.O., Olaore, G.O., Udofia, E.E., Adenigba, O.A.: Covid-19 pandemic and business survival as mediation on the performance of firms in the FMCG-sector
4. Alves, J.C., Lok, T.C., Luo, Y., Hao, W.: Crisis management for small business during the covid-19 outbreak: survival, resilience and renewal strategies of firms in Macau (2020)
5. Carracedo, P., Puertas, R., Marti, L.: Research lines on the impact of the covid-19 pandemic on business. a text mining analysis. J. Bus. Res. **132**, 586–593 (2021)
6. Giunipero, L.C., Denslow, D., Rynarzewska, A.I.: Small business survival and covid-19-an exploratory analysis of carriers. Res. Transp. Econ. 101087 (2021)

7. Hamilton, S.: From survival to revival: how to help small businesses through the COVID-19 crisis. The Hamilton Project Policy Proposal pp. 2020–14 (2020)
8. Katare, B., Marshall, M.I., Valdivia, C.B.: Bend or break? Small business survival and strategies during the COVID-19 shock. Int. J. Disaster Risk Reduction **61**, 102332 (2021)
9. Kuhn, M., Johnson, K.: Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press (2019)
10. Lau, A.: New technologies used in COVID-19 for business survival: insights from the hotel sector in china. Inform. Technol. Tour. **22**(4), 497–504 (2020)
11. Rakshit, D., Paul, A.: Impact of COVID-19 on sectors of Indian economy and business survival strategies. Available at SSRN 3620727 (2020)