# COVID-19's Impact on Small Businesses

Vir Handa[1], Vonny Jap[1], Kael Polkow[1], Om Sachdev[1], and Joseph Wibowo[1]

[1]Georgia Institute of Technology - Team 68

Fall 2022

## 1 Introduction

COVID-19 has made a significant impact on society and in particular, the economy. Studies have been performed to measure the impact of forced closures on small businesses. The conclusions from these studies were mainly drawn from survey and census data which are inherently limited in scope. Additionally, while these studies provided high-level context to the situation, they did not do more complex analysis.

We believe a model to predict business closures can better inform governments on how to formulate policies to minimize the negative impact on local businesses. In this paper, we develop a machine learning model trained on Yelp's academic dataset and mobility data from various vendors to predict closures of local brick-and-mortar businesses during the pandemic. Success of this project will be measured by the accuracy of the model produced.

## 2 Literature Survey

### 2.1 Why do Stores Close

Businesses close for many different reasons. In his study, Bates identifies demographic factors, such as education and race, that differentiate the discontinuation of a successful vs unsuccessful business [Bat05]. A group in Prague used geolocation data and k-means clustering to determine if location of a store has an impact on its sales [FS22]. Lastly, a study of businesses affected by Hurricane Katrina pointed out that disasters tend to intensify pre-existing business conditions [Mar+15]. These attributes are all potential features for a predictive model.

### 2.2 Methods to Predict Store Closure

There are existing studies that have tried to predict closures of certain small businesses using a model. Using online consumer reviews, Tao and Zhou performed a sentiment analysis to classify the reviews and use the results in a deep learning model to predict the closure of restaurants [TZ20]. In a similar study, a group at USC also used Yelp business data to train five different machine learning models that predict the closure of Las Vegas restaurants [Zha+]. Lastly, a study was done that used Foursquare and taxi mobility data to determine features that are predictive of retail business closures and also develop a model that can predict retail business closures with 80% accuracy [DSi+18].

### 2.3 Pandemic's Effect on Small Business

Policymakers knew stay-at-home mandates and social distancing would have a negative economic impact on businesses, but they could not quantify how much. In a study from Macau University, loss of customer flow was considered the most significant economic impact on local businesses when interviewing a panel of business owners [Alv+20]. Studies have consistently shown that the pandemic has led to one of history's largest drop in active business owners, but there was also a large surge in new businesses [Fai20][DH22]. These studies ultimately confirm the negative impact of the pandemic.

### 2.4 Data Availability

Data used for business closure prediction can be difficult to find. In a study from Drexel, Homebase provided a mobility dataset that, when combined with Google and Facebook data, was able to model a probability of business openings and closures during the pandemic [KLT22]. A team from GeoDS created a time-series dataset of visitor flows between two destinations to overcome limitations in existing mobility datasets [Kan+20]. For business closures, a team in India used the Yelp academic dataset to determine if a business is open or closed in their model [TD18]. Finally, to tie mobility and business data together, a researcher used SafeGraph's datasets to perform descriptive statistics on restaurant closures during COVID [Sed22]. These datasets and methodologies on how they are used will be used as guidance for our goal.

### 2.5 Mobility Data

Mobility data is a fairly new concept but has already shown a lot of potential. A team from Tel Aviv University used mobility data to try and predict Starbucks closures with a variety of machine learning algorithms [SZR21]. Pinar from Wharton was able to use mobility data from retail stores to show how much traffic they lost in 2020 [Yil21].

## 3 Method

In order to measure the pandemic's impact to small businesses, our approach is to use machine learning models to predict if a restaurant will close based on the business characteristics and mobility data trends.

### 3.1 Data collection and processing

For this project we used the academic dataset available from Yelp. The dataset is a snapshot of all businesses on Yelp up to early 2022 for major cities in America, so the results of the pandemic should be reflected in the dataset. The data has a boolean value for if the store is open or closed which will be the predicted value for our model.

However, there is no open or close date provided. To overcome this limitation, our team utilized the first and last review, tips, and check-ins date of each business to approximate the business open and closure dates. The business dataset comes with various features that can potentially be good candidates to predict the possibility of business closure. The main features were chosen based on their relevancy to restaurants and least non-null values. They include stars, review counts, parking, price range, offers delivery, offers takeout, and accepts credit cards.

The dataset is filtered for businesses that were closed before Jan 1, 2020 because we can assume the pandemic is not the reason for their closure. We also focused our study efforts on businesses related to food and restaurants as the dataset consists mostly of these businesses.

We are interested to learn if mobility data is predictive of business closures. The mobility dataset chosen is from Google who posted community mobility reports based on Google Maps data during the peak of COVID. It provides percent changes in movement to specific place categories over time. The benchmarks are aggregated to a central geolocation of an area. This dataset is joined to the yelp dataset by calculating the distance between a business and the closest location Google used to calculate their benchmarks.

The mobility dataset is a time-series so we generated features on declining visits by aggregating the data by location. From this, we can generate a feature that indicates if a location experienced a significant decline in movement to retail and restaurant stores. We generate the following features from this dataset: mean pct change, median pct change, whether the county mean is below the state's mean, trend, and pct of days that are negative.

Lastly, since we are using supervised learning models, we will split the dataset into training, validation, and test datasets (60/10/30).

### 3.2 Developing the Model

As stated, we selected features to train our model on. Since we were unaware of which features are most useful, we analyzed which features might be useful using filter methods where features will be evaluated based on their correlation with the outcome variable. To get the correlation statistics, we used chi-squared tests for categorical variables and Pearson's correlation for continuous variables.

The target feature of this dataset is whether or not a business closed during the pandemic. This can be modeled as a boolean value of 1 and 0. In other words, this is a binary classification problem. It is important to note that the dataset is quite imbalanced as there are many more open stores than closed stores so certain methods will work better than others. This paper will look to explore three models to see which one performs best. They are logistic regression, complement naive bayes, and random forests. All these models will be implemented using the `scikit-learn` python library.

Model evaluation will be done using a k-fold cross validation. Afterwards, the most performant model will be tuned to try and increase the model's accuracy using a grid search.

### 3.3 Visualization

In order to visualize the results of our experiment, we have decided to use a map showing the stores open and closed due to the pandemic and the results of our machine learning model predictions indicated by the dots on the map.

## 4 Experiments

For the setup, we created a training, validation, and test dataset using sklearn's `test_train_split` function. The data was stratified on the target feature to make sure each dataset has a similar class imbalance as the original dataset. The split percentage of each dataset is 60/10/30 respectively.

### 4.1 Feature Selection Experiments

For all feature selection experiments, the dataset used was the training dataset to prevent bias. To determine the best categorical features, a chi-squared test was used to determine which features had a relationship with our target feature. Continuous variables were grouped into bins to be used as categorical features. The `chi2` package in sklearn was used to perform the test on the dataset and calculate the p-values. The p-values are shown in Figure 1. `restaurants_takeout` and `business_accepts_cc` were found to have no relationship with our target variable.

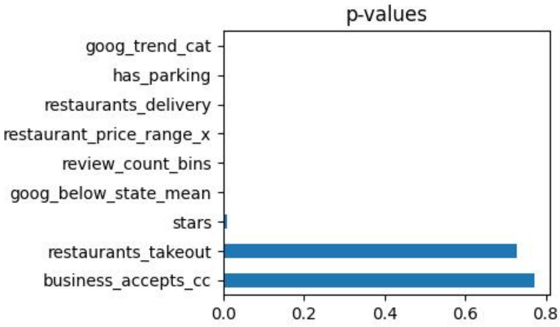

Figure 1: p-values from chi2 test

For continuous variables, Pearson's coefficient and mutual information were calculated to determine which features had a stronger relationship with the target variable. The `corr` function in pandas and `mutual_info_regression` package were used to calculate the values for each feature using the dataset. The results are shown in Figure 2 and 3.

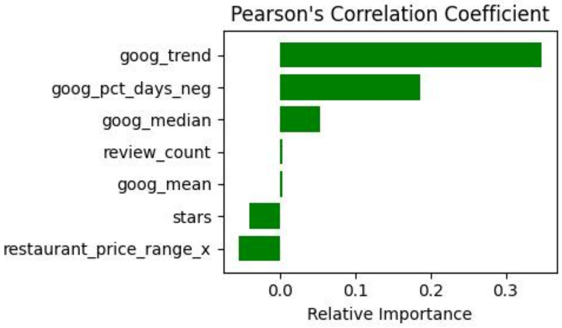

Figure 2: Correlation coefficients for each feature

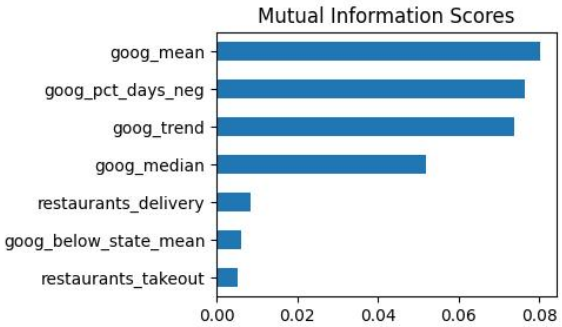

Figure 3: Mutual information scores for each feature

Our original hypothesis is that mobility data from the pandemic can be used to predict the impact on small businesses and based on these results, the mobility data seems much more correlated to the closure of a business than the business features.

### 4.2 Model Selection

For the model selection, we used a k-fold cross validation to determine which models performed best on the training dataset. `StratifiedKFold` from sklearn was used to create the folds to make sure each fold best represents the imbalances in the dataset. The hyperparameters of the models were set to the default values in sklearn as they are pretty reliable to get an idea on how effective each model is. The experiment used 5 folds.

When evaluating the models, we do not use accuracy as the evaluation metric. Since the dataset is quite imbalanced, the precision, recall and F1-score of the 0 label (closed business) were used to determine the effectiveness of the model since it takes into account this imbalance. We use this same f1-score definition for all figures. The mean of each of the metrics was then calculated to give the final result. The results of the experiment are in Figure 4.

| K-fold Validation Results | | | |
|---|---|---|---|
| Model | Precision | Recall | F1-score |
| Logistic Regression | 0.628 | 0.115 | 0.195 |
| Naive Bayes | 0.389 | 0.336 | 0.216 |
| Random Forest | 0.518 | 0.384 | 0.341 |

Figure 4: Table of model results

Since the random forest performed significantly better than any other model, we tried to improve the score by tuning the hyperparameters. The following parameters were used in sklearn's `GridSearchCV` class: `n_estimators`, `max_depth`, `max_features`, `min_samples_leaf`, `criterion`. Please refer to sklearn's documentation for a definition of each parameter. `StratifiedKFold` was used again as the splitting strategy for the cross validation to account for imbalances. The model is fitted on the training dataset and tested on the validation dataset. A split of 4 folds was used. The results are in Figure 5.

| Random Forest Hyperparams | | |
|---|---|---|
| Params | Mean F1 | Std F1 |
| ’criterion’: ’entropy’,<br>’max_depth’: None,<br>’max_features’: 9,<br>’min_samples_leaf’: 1,<br>’n_estimators’: 300 | 0.578 | 0.052 |
| ’criterion’: ’entropy’,<br>’max_depth’: None,<br>’max_features’: 9,<br>’min_samples_leaf’: 1,<br>’n_estimators’: 100 | 0.578 | 0.049 |
| ’criterion’: ’entropy’,<br>’max_depth’: 15,<br>’max_features’: 9,<br>’min_samples_leaf’: 5,<br>’n_estimators’: 100 | 0.575 | 0.044 |
| ’criterion’: ’entropy’,<br>’max_depth’: None,<br>’max_features’: 9,<br>’min_samples_leaf’: 1,<br>’n_estimators’: 300 | 0.570 | 0.044 |
| ’criterion’: ’gini’,<br>’max_depth’: None,<br>’max_features’: 9,<br>’min_samples_leaf’: 1,<br>’n_estimators’: 100 | 0.567 | 0.045 |

Figure 5: Table for random forest hyperparams results

Since the results of the grid search showed consistent scores across folds, overfitting doesn't seem to be an issue. Using the best parameters from the grid search, the model was then used to predict the test dataset. The result is shown in figure 6.

| Random Forest Test Results | |
|---|---|
| Metric | Score |
| Accuracy | 0.950 |
| Precision | 0.960 |
| Recall | 0.986 |
| F1-score | 0.657 |

Figure 6: Table for final model validation and test results

## 5 Conclusion

The results generally supported our hypothesis that mobility features are predictive of a restaurant's closure during COVID. While the random forest model f1 score is not high enough to indicate a very reliable model, the results show high potential that the model can be further tuned to become a useful model for future pandemics. Further investigation into other business features and mobility datasets is needed as well as keeping in mind the class imbalances in the data.

***All members contributed a similar amount of effort.***

## References

[Bat05] Timothy Bates. "Analysis of young, small firms that have closed: delineating successful from unsuccessful closures". In: *Journal of Business Venturing* 20.3 (2005), pp. 343–358. ISSN: 0883-9026. DOI: https://doi.org/10.1016/j.jbusvent.2004.01.003. URL: https://www.sciencedirect.com/science/article/pii/S0883902604000308.

[Mar+15] Maria I. Marshall et al. "Predicting small business demise after a natural disaster: an analysis of pre-existing conditions". In: *Natural Hazards* 79.1 (Oct. 2015), pp. 331–354. ISSN: 1573-0840. DOI: 10.1007/s11069-015-1845-0. URL: https://doi.org/10.1007/s11069-015-1845-0.

[DSi+18] Krittika D'Silva et al. "The role of urban mobility in retail business survival". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.3 (2018), pp. 1–22.

[TD18] Sharun S Thazhackal and V. Susheela Devi. "A Hybrid Deep Learning Model to Predict Business Closure from Reviews and User Attributes Using Sentiment Aligned Topic Model". In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2018, pp. 397–404. DOI: 10.1109/SSCI.2018.8628823.

[Alv+20] Jose C Alves et al. "Crisis management for small business during the COVID-19 outbreak: Survival, resilience and renewal strategies of firms in Macau". In: (2020).

[Fai20] Robert Fairlie. "The impact of COVID-19 on small business owners: Evidence from the first threemonths after widespread social-distancing restrictions". In: *Journal of Economics & Management Strategy* 29.4 (2020), pp. 727–740. DOI: https://doi.org/10.1111/jems.12400. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jems.12400. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jems.12400.

[Kan+20] Yuhao Kang et al. "Multiscale dynamic human mobility flow dataset in the US during the COVID-19 epidemic". In: *Scientific data* 7.1 (2020), pp. 1–13.

[TZ20] Jie Tao and Lina Zhou. "Can Online Consumer Reviews Signal Restaurant Closure: A Deep Learning-Based Time-Series Analysis". In: *IEEE Transactions on Engineering Management* (2020), pp. 1–15. DOI: 10.1109/TEM.2020.3016329.

[SZR21] Tal Shoshani, Peter Pal Zubcsek, and Shachar Reichman. "Predicting Store Closures Using Urban Mobility Data and Network Analysis". In: (2021).

[Yil21] Pinar Yildirim. "The Short Term Impact of COVID-19 on Brick-and-Mortar Retailers: Evidence from Retailtech". In: *Available at SSRN 3777740* (2021).

[DH22] Ryan A Decker and John Haltiwanger. "Business entry and exit in the COVID-19 pandemic: A preliminary look at official data". In: (2022).

[FS22] Tomáš Formánek and Ondřej Sokol. “Location effects: Geo-spatial and socio-demographic determinants of sales dynamics in brick-and-mortar retail stores”. In: *Journal of Retailing and Consumer Services* 66 (2022), p. 102902. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2021.102902. URL: https://www.sciencedirect.com/science/article/pii/S0969698921004689.

[KLT22] Andre Kurmann, Etienne Lalé, and Lien Ta. “Measuring Small Business Dynamics and Employment with Private-Sector Real-Time Data”. In: (2022).

[Sed22] Dmitry Sedov. “Restaurant closures during the COVID-19 pandemic: A descriptive analysis”. In: *Economics Letters* 213 (2022), p. 110380. ISSN: 0165-1765. DOI: https://doi.org/10.1016/j.econlet.2022.110380. URL: https://www.sciencedirect.com/science/article/pii/S0165176522000593.

[Zha+] Fan Zhang et al. “Las Vegas Business Closure Prediction Model”. In: ().