Restaurant Risk Prediction using Machine Learning Algorithms

Aman Patel, Kamal Panchal, Kuldip Patel, Shripal Shah

Abstract—Due to tough competition and various choices for consumers, opening a new restaurant has been challenging. Food safety is profoundly affected by the diversity of food and its raw materials. As the global economy has increasingly been liberalized, food imports have also increased, highlighting the importance of food risk management in protecting consumer health. Hence, we have developed different techniques to predict the risk category of restaurants. This technique improves training and test sets by involving various steps in preprocessing data, such as null values, noise, and outlier detection. After cleaning the data, the most important features are selected, and the data is divided into training and testing groups. Training data are used to consider and train machine learning algorithms from various families. Model accuracy is used to measure algorithm performance. With k-fold cross-validation, most models performed averagely on the test set. A comparison of algorithms is made using performance metrics to select the best model within each family. Computer metrics are used to compare and select the best machine learning model.

Index Terms—Machine Learning, Over-fitting, Accuracy, True positive, True negative, Precision, Recall, F1-score.

1 Introduction

TODAY, machine learning enhances every piece of technology used by users. Artificial intelligence includes the field of machine learning. Understanding data structure and incorporating it into the model, which can subsequently be used to predict values relying on the same type of data, are the last steps in applying machine learning. Machine learning ultimately seeks for connections and trends in the data. Machine learning can comprehend the link between various features of the data itself, as opposed to typical software programs, where commands are deliberately designed to carry out a certain activity. This experiment investigates several machine learning models' techniques for predicting the risk category for eateries.

Restaurant inspections are crucial to preventing foodborne infections. Machine learning may be used to foretell if a restaurant consistently exhibits a particular pattern of behavior or sometimes defaults. As a result, future risk classifications for restaurants may be predicted and high-risk establishments can undergo routine inspections.

The information gathered for every restaurant within San Francisco, California, during the examination is included in this report. After that, the data is cleaned and pre-processed to eliminate any noise. Different machine learning techniques will be modified after the data has been cleaned, and the models' propensity to correctly forecast the risk category will be evaluated. In order to choose the optimal model for this given dataset, multiple comparisons are conducted between several models.

2 Problem Statement

In the U. S., 844,000 commercial restaurants offer more than 54 billion meals annually . Restaurant meals account for 46% of all American food spending. In terms of food, the restaurant holds a market share of more than 40%. 44% of individuals in the United States dine at restaurants every day. Every year, the Centers for Disease Control and Prevention receive reports of an average of 550 outbreaks of foodborne disease. The outbreak increased by more than 40% between 1993 and 1997. Commercial food producers have been blamed for this rise. Public health and public health agencies have a crucial job to do: prevent foodborne outbreaks of illnesses brought on by eating places.

Local, state, or health officials frequently examine restaurants in the United States. Through inspection results, the primary goal of the inspection is to avoid foodborne diseases. The goal of inspection results is to increase food safety. Data collected by the inspector during the inspection is stored in a database. Public access to and regular publication of this data in regional media. We only used data from San Francisco, California in the United States for this investigation. Mostly on website of the Office of the Chief Data Officer, the City and County of San Francisco, this information is easily accessible. It is possible to anticipate variables such as inspection score, inspection type, risk category, etc. from this data, which is updated often.

In order to rate restaurants according to risk category-low, medium, and high-this post will attempt to examine numerous data elements and apply them efficiently. Ranking these factors enables restaurants classified as high-risk to be constantly monitored and enables appropriate action to be taken prior to an epidemic.

3 Literature Review

Opening a new restaurant has been a challenge in the current age of tough competition and various choices for consumers. As a result, the rating of a restaurant becomes an essential parameter for judging its quality. New customers are attracted to a good rating. Each new business wants to know if it will succeed in the future. In this paper, we validate predicted Ratings for New Restaurants. If necessary, one can compare various options for a particular characteristic of a planned Restaurant based on multiple factors that are characteristics of the planned restaurant. Opening a new restaurant can also help one make the right decisions. Thus, protecting them against investment losses, saving time, and making the whole process a calculated risk. Predictions are made using seven different regression models by analyzing factors that can be controlled easily before setting up a new restaurant. Finally, model metrics are compared to select the best regression model.

This paper proposes an early warning system for food. First, a gray prediction model is improved, and a quality index is constructed. Food quality is measured using the quality index model. Following that, a gray prediction model is applied to predict the future trend of food quality indexes. Finally, the food security risk is assessed by comparing predicted trends with standard limits proposed by experts. Taking measures to protect people's health from the risks associated with risky foods.

To improve border inspection methods for imported food, ensemble learning was used to develop risk prediction models. It was intended to increase the hit rate of non-conforming products so that border control of food products can be strengthened to protect public health. We developed models using five algorithms to assess each imported food batch's risk. As a way of evaluating the models, a confusion matrix was constructed to calculate predictive performance indicators, such as positive prediction value (PPV), recall, the harmonic mean of PPV and recall (F1 score), and the area under the curve. Compared to a single algorithm, ensemble learning achieved better and more stable prediction results. Furthermore, according to results of comparable data periods, the non-conformity hit rate increased significantly after the online implementation of ensemble learning models, indicating that ensemble learning effectively predicted risk. Additionally to improving the inspection hit rate of non-conforming food, the results of this study can be used to improve existing random inspection methods, thus strengthening food risk management capabilities.

4 Proposed Methodology

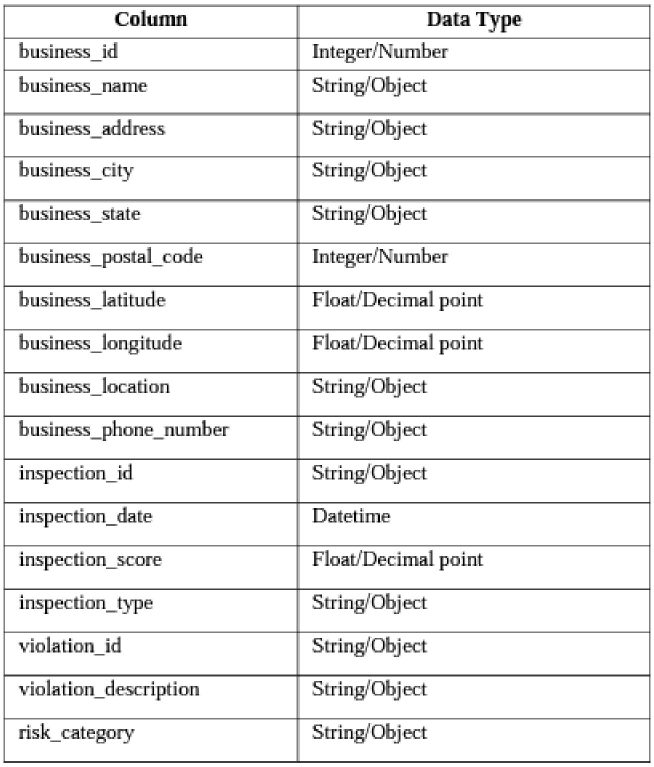
4.1 Dataset

A online site called DataSF is managed and maintained by the city of San Francisco. Everyone has access to this site. The goal of DataSF is to make it possible to use data to enhance municipal operations and services over time. All datasets accessible through this site are updated often and are also known as LIVE data. The publicly available data promotes widespread trust, openness, and accountability and is not false. It will be beneficial to conduct experiments using LIVE data and apply the developed trained model for the larger good.

4.1.1 Source and Description

Restaurant Scores - LIVES Standard is the name of the dataset that was utilized in this study. DataSF[3] is the source of the dataset. You can choose data from a variety of categories when visiting the online portal. The category for this dataset is ”Health and Social Services” [5]. This data represents actual values and is updated regularly. Since its creation on October 28, 2015, this dataset has been updated every day. The LIVES Flattened schema, which is based on LIVES version 2.0 listed on the Yelp website [5], is used to store LIVES San Francisco restaurant inspection data. 53.7000 rows and 17 columns make up the dataset. However, this study trains the model and evaluates several machine learning techniques using data collected from 2016 to February 22, 2019.

Below is the summary of dataset, with column name and their data type. Business in column name refers to restaurant.



Column	Data Type
business_id	Integer/Number
business_name	String/Object
business_address	String/Object
business_city	String/Object
business_state	String/Object
business_postal_code	Integer/Number
business_latitude	Float/Decimal point
business_longitude	Float/Decimal point
business_location	String/Object
business_phone_number	String/Object
inspection_id	String/Object
inspection_date	Datetime
inspection_score	Float/Decimal point
inspection_type	String/Object
violation_id	String/Object
violation_description	String/Object
risk_category	String/Object

Fig. 1. This is an image of the data description

4.1.2 Preprocessing

Every column in the data was thoroughly examined as part of the preprocessing. Preprocessing aims to change the data type of every column in a specific symmetry. Here, we attempt to convert each column to a numerical data type. A significant amount of noisy data was eliminated during the preparation stage. In order to improve model training, a lot of redundant data from particular columns was also deleted during feature engineering while unique elements were maintained. The complete data preprocessing procedure for each column with feature selection is shown below.

• business_name : Name of the company is listed in this column. However, business id may be used to specifically identify restaurants, thus it was removed.

• business_address: This functionality is not necessary because we can track a business's location using postal codes, latitude, and longitude information instead. This was thus abandoned.

• business_city: San Francisco was found to be the dataset's predominant business city. This column was removed because it was unable to make a significant contribution to data training. business_state:In the entire dataset, it was found that California was the business state. This column was removed because it cannot help with data training.

• business_phone_number:Combining longitude and latitude is this column. It was omitted since it causes duplication in the data. When determining the risk category for a restaurant, this doesn't seem to be a crucial element. The column has been removed.

• inspection_type:Data was transformed from this into categorical data. However, it was discovered that just one category was constant across all rows after eliminating null values. This column was so removed.

The misspelling was discovered during further investigation. All postal codes have the pattern 9XXXX, and the value of business postal code is something like ”64110”. This information has been altered, though. Additionally, there were several literal strings like ”CA,” ”Ca,” and ”941.” Because zip codes cannot be assessed based on other values, rows with these values have been eliminated. Some postcodes also had the 9XXXX-XXXX format, which preserved the first five digits in order to maintain symmetry but reduced the columns. Zip code and latitude were found to contain null values while looking for null and distinct values. Because they could have an impact on data training, these rows were eliminated from the data set. Additionally, check date was divided into three columns—year, month, and day—to aid in training the model.

4.1.3 Feature Section

Feature selection for training data is very important. If you don't choose the right attributes, your model may be underfit or overfit. This makes our model very unreliable. It is important to choose features that tell the story or pattern in your data. Our model learns patterns in the data and tries to make predictions based on them. Some of the features below have been redesigned and others have been kept.

• business_id: This column is very important because it uniquely identifies each restaurant. • business_post_code:This column can uniquely identify a restaurant's location and can be an important feature.

• business_latitude: This column lists the latitudes of available restaurants on a geographic map and estimates popular locations.

• business_longitude: As above, used in geographic maps to estimate popular locations.

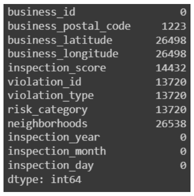
• inspection_score: This is a very important feature and has not changed.

• violation_id: This column uniquely identifies the description of the violation. So that was part of the training data.

• violation_type: This data has been converted to categorical numeric data because it has the last unique value. For example: sanitation:1, legality:2, non-compliance:3 and lack of infrastructure:4

• risk_category: This column is a very important part of the dataset. Predict risk categories from a data set. Therefore, the categories were converted to numerical data. That is, in the process of sorting the classes Low Risk: 1, Medium Risk: 2, High Risk: 3,

We also preferred to transform the data into categorical numerical values and eliminated most of the least estimated classes. After feature engineering, the dataset has a total of 22,273 rows and 8 columns. An overview of the dataset is shown in the below image.



```
business_id                 0
business_postal_code     1223
business_latitude       26498
business_longitude      26498
inspection_score        14432
violation_id            13720
violation_type          13720
risk_category           13720
neighborhoods           26538
inspection_year             0
inspection_month            0
inspection_day              0
dtype: int64
```

Fig. 2. This is an image from the preprocessed data after feature selection.

5 Data Visualization

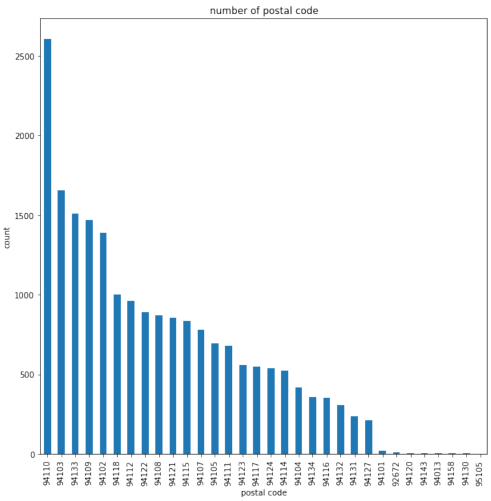


Fig. 3. Compared to other postal codes, only a few have very low densities and relative data.

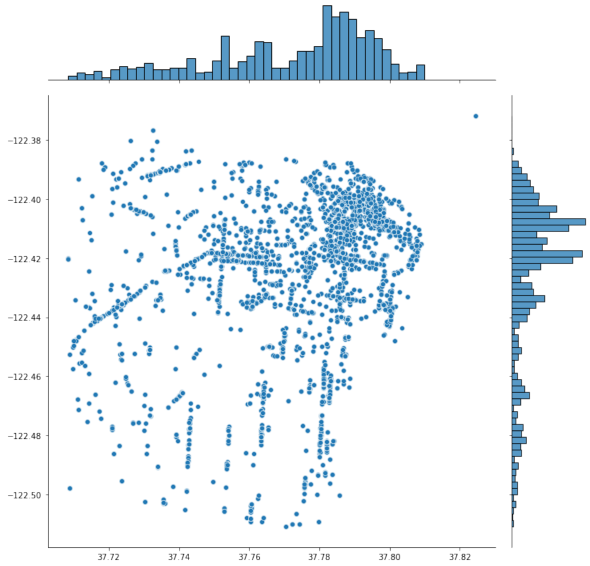


Fig. 4. Restaurants run their businesses in certain regions, which are hotspots. For this reason, it would be interesting to predict a restaurant's risk category based on its location

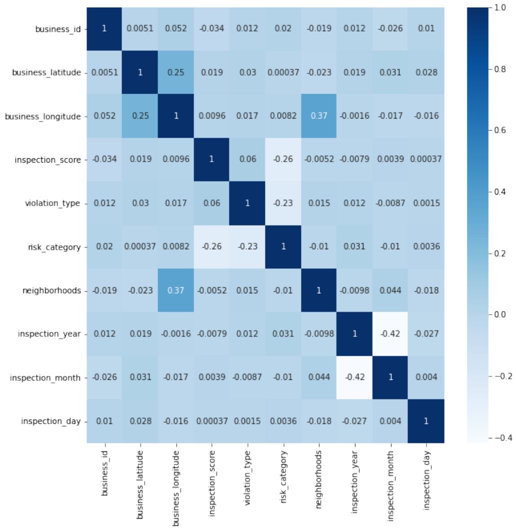
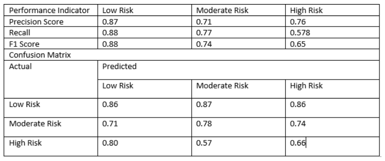


Fig. 5. This is an image of correlation between all the numerical features after the pre processing

6 Result Analysis

6.1 Linear models

When training logistic regression models with solvers like Newton-cg produced accuracy better than 73%, it was not possible to identify the decision limit that can be classified by risk category. It's possible that the data simply cannot be fit by a logistic model and will instead deliver random outcomes if the model is fitted with random data. Since it was unable to predict the medium and high risk categories, these categories had no values for its performance indicators.



Performance Indicator	Low Risk	Moderate Risk	High Risk
Precision Score	0.87	0.71	0.76
Recall	0.88	0.77	0.578
F1 Score	0.88	0.74	0.65
Confusion Matrix			
Actual	Predicted		
	Low Risk	Moderate Risk	High Risk
Low Risk	0.86	0.87	0.86
Moderate Risk	0.71	0.78	0.74
High Risk	0.80	0.57	0.66

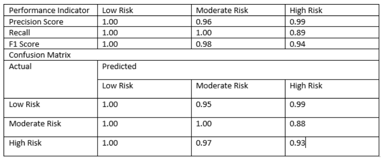
Fig. 6. This is an image of the data description

6.2 Geometric Model

K-Nearest Neighbors, which uses Manhattan and Euclidean distance, is the method used by the two trained models. The initial model that took Manhattan's distance into account produced some rather good findings. It was 80% accurate on average. The high risk category is where this model performs best, whereas the low risk category is where it performs best in terms of recall and accuracy. However, compared to the first model, the second model that uses Euclidean distance has a much lower average accuracy 65%. Take note that by forecasting two classes, these two models created confusion.

6.3 Probabilistic Model

In the probabilistic model, we trained Gaussian Naïve Bayes. With an average accuracy of 98%, this model did well. It might also signify an over-fitting of the model, though. K-fold cross validation with a number of tests was utilized to evaluate over-fitting. It is over-fitted because the accuracy after training and cross-validation stays at 98

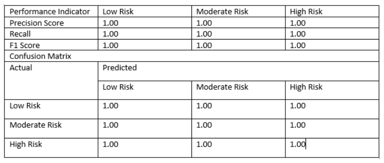


Performance Indicator	Low Risk	Moderate Risk	High Risk
Precision Score	1.00	0.96	0.99
Recall	1.00	1.00	0.89
F1 Score	1.00	0.98	0.94
Confusion Matrix			
Actual	Predicted		
	Low Risk	Moderate Risk	High Risk
Low Risk	1.00	0.95	0.99
Moderate Risk	1.00	1.00	0.88
High Risk	1.00	0.97	0.93

Fig. 7. This is an image of the data description

6.4 Logical Model

The logic model decision tree classifier was trained using the maximum depth of fault; the average accuracy of the tree ranges from 92% to 98%. The majority of genuine positive and true negative findings were accurately anticipated, achieving 100% accuracy, indicating that it is also overfitting.



Performance Indicator	Low Risk	Moderate Risk	High Risk
Precision Score	1.00	1.00	1.00
Recall	1.00	1.00	1.00
F1 Score	1.00	1.00	1.00
Confusion Matrix			
Actual	Predicted		
	Low Risk	Moderate Risk	High Risk
Low Risk	1.00	1.00	1.00
Moderate Risk	1.00	1.00	1.00
High Risk	1.00	1.00	1.00

Fig. 8. This is an image of the data description

7 Conclusion

Our approach made an effort to divide restaurants into three danger categories using information provided by the San Francisco authorities. The dataset includes of restaurant violations, inspection results, and location data. Daily updated, uncleaned data, on the other hand, has a lot of features that aren't crucial; the clutter serves as noise to the model during training. Preprocessing was done on the data to make it suitable for training on several model families.

In the feature selection process, inspection score obtained the highest rating, while day, month, and year received the lowest rating. A few years ago, certain year and month columns were missing from the training data. All types of models were taken into consideration and trained. Some models were found to overfit the data, while others were unable to determine the decision limit for the current categorization. However, for this data set, the geometric model performed flawlessly. The average accuracy was around 70%, despite some problems with some courses' interpretation.

References

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3323064/
[2] Olsen SJ, MacKinnon LC, Goulding JS, Bean NH, Slutsker L Surveillance for foodborne-disease outbreaks, United States, 19931997. MMWR CDC Surveill Summ. 2000;49:1–62
[3] https://datasf.org/
[4] https://datasf.org/about/
[5] https://data.sfgov.org/Health-and-Social-Services/Restaurant-ScoresLIVES-Standard/pyih-qa8i/
[6] https://machinelearningmastery.com/k-fold-cross-validation/

[7] https://datascience.stackexchange.com/questions/9167/what-doesrmse-points-about-performance-of-a-model-in-machine-learning

[8] https://blog.minitab.com/blog/adventures-in-statistics-2/regressionanalysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

[9] https://machinelearningmastery.com/confusion-matrix-machinelearning/

[10] https://towardsdatascience.com/metrics-to-evaluate-your-machinelearning-algorithm-f10ba6e38234

[11] https://developers.google.com/machine-learning/crashcourse/classification/accuracy

[12] https://machinelearningmastery.com/compare-machine-learningalgorithms-python-scikit-learn/

[13] https://blog.minitab.com/blog/adventures-in-statistics-2/regressionanalysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

[14] https://scikit-learn.org/stable/modules/linear_model.htmllinearmodel

[15] https://docs.aws.amazon.com/machine-learning/latest/dg/linearmodels.html

[16] https://scikitlearn.org/stable/auto_examples/linear_model/plot_bayesian_ridge.htmlsphx-glr-auto-examples-linear-model-plot-bayesian-ridge-py