

Inspection score prediction of Restaurants in New York City using Machine Learning

Jashwanth Reddy Goraka
Kumbam Nithin Goud

Problem Statement:

- This project aims to develop a machine learning model that can accurately predict the inspection scores of restaurants in New York City based on various factors such as Cuisine type, location, violation type, etc.
- By predicting inspection scores, this model can help the New York City Department of Health and Mental Hygiene (DOHMH) identify high-risk restaurants and prioritize inspections, which can lead to improved food safety and reduced health risks for consumers.
- This project will involve cleaning and preprocessing the DOHMH New York City Restaurant Inspection Results dataset, feature engineering to extract relevant information, and testing various machine learning algorithms to identify the best-performing model for the task of inspection score prediction.

Introduction:

The restaurant industry plays a vital role in New York City's vibrant culinary landscape, catering to a diverse population of residents and tourists. Ensuring the safety and quality of food establishments is of utmost importance to protect public health and maintain high standards. To assist regulatory authorities in efficiently monitoring and prioritizing inspections, machine learning algorithms can be leveraged to predict inspection scores for restaurants.

This document aims to explore the problem of inspection score prediction for restaurants in New York City using machine learning techniques. By analyzing various features and historical inspection data, we can develop models that estimate the likelihood of a restaurant receiving a specific inspection score.

Predicting inspection scores has several practical implications. First and foremost, it can assist regulatory bodies, such as the New York City Department of Health and Mental Hygiene (DOHMH), in optimizing their inspection schedules and resource allocation. By identifying high-risk establishments in advance, they can prioritize inspections and interventions, thereby improving overall food safety.

Restaurant owners and operators can benefit from inspection score predictions. By understanding the factors that contribute to a higher or lower inspection score, they can take proactive measures to maintain compliance with health regulations and improve their operations. This can lead to increased customer satisfaction, better brand reputation, and potentially higher revenue.

To tackle this problem, we employ a data-driven approach using machine learning algorithms. By utilizing historical data on inspections, violations, and restaurant attributes, we can build predictive models that learn patterns and relationships between various factors and inspection scores. These models can then be used to generate predictions for future inspections based on new or updated information.

Overview:

The Data Pipeline as follows:

1. Data Acquisition: The data is used from this [link](#). It has around 206K records.
2. Data Preprocessing: This step involves cleaning the data, handling missing values, and transforming it into a suitable format for machine learning algorithms.
3. Feature Selection and Engineering: We will identify relevant features from the available data and explore the creation of new meaningful features that can enhance prediction performance.
4. Model Selection: We will evaluate and compare different machine learning algorithms suitable for this problem, considering their strengths, limitations, and interpretability.
5. Model Training and Evaluation: The selected models will be trained on the prepared data and evaluated using appropriate metrics to assess their predictive performance.

Modeling:

To analyze and predict outcomes based on data, various machine learning algorithms are employed. Here's an overview of the mentioned algorithms, highlighting their usage and characteristics:

Linear Regression:

Linear regression is a supervised learning algorithm used for predicting continuous numeric values. It establishes a linear relationship between the independent variables (features) and the dependent variable (target) by fitting the best-fitting line through the data points. The algorithm calculates the coefficients that minimize the sum of squared residuals, allowing us to estimate the target variable based on the input features. Linear regression assumes a linear relationship and is widely used for tasks such as sales forecasting, price prediction, and trend analysis.

Logistic Regression:

Logistic regression is a classification algorithm used to predict categorical outcomes. It estimates the probability of an instance belonging to a particular class by fitting a logistic function to the input features. The logistic function transforms the output into a probability between 0 and 1. Logistic regression is commonly employed for binary classification problems, where there are two classes, such as spam detection or disease diagnosis. It can also be extended to handle multi-class classification tasks using techniques like one-vs-rest or softmax regression.

GradientBoostingRegressor:

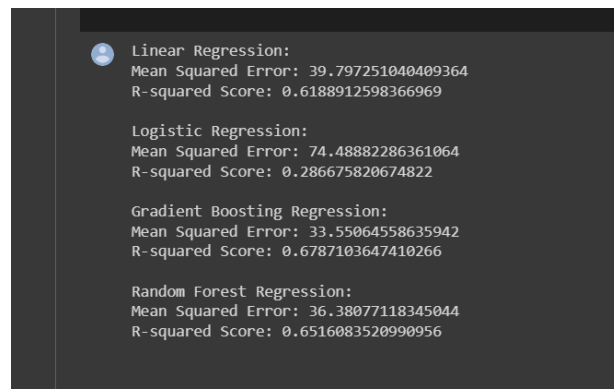
GradientBoostingRegressor is a boosting algorithm that combines multiple weak predictive models, typically decision trees, to create a powerful predictive model. It works by iteratively fitting new models to the residuals of the previous models, with each subsequent model focusing on the previously mispredicted instances. GradientBoostingRegressor is used for regression tasks, where the target

variable is continuous. It is known for its ability to handle complex datasets and produce highly accurate predictions. This algorithm is commonly used in various domains such as finance, healthcare, and online advertising.

RandomForestRegressor:

RandomForestRegressor is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions to make the final prediction. Each decision tree is built on a random subset of the data, with random feature selection at each node. RandomForestRegressor is used for regression tasks and is known for its ability to handle high-dimensional data, capture complex relationships, and provide robust predictions. It is widely used in applications such as stock market prediction, real estate valuation, and customer demand forecasting.

After running all these model, these are our accuracy scores for each model:



```
Linear Regression:
Mean Squared Error: 39.797251040409364
R-squared Score: 0.6188912598366969

Logistic Regression:
Mean Squared Error: 74.48882286361064
R-squared Score: 0.286675820674822

Gradient Boosting Regression:
Mean Squared Error: 33.55064558635942
R-squared Score: 0.6787103647410266

Random Forest Regression:
Mean Squared Error: 36.38077118345044
R-squared Score: 0.6516083520990956
```

Performance Optimization:

We wanted to improve our model accuracy by using few optimization techniques on Gradient Boosting Regression since it was the best model based on our evaluation metrics.

Hyperparameter tuning: Hyperparameter tuning is the process of selecting the optimal configuration of hyperparameters for a machine learning model. It involves systematically exploring different hyperparameter values and evaluating their impact on the model's performance. The goal is to find the combination of hyperparameters that maximizes the model's predictive accuracy or minimizes the error.

After this, we got an improved R2 score of 0.695.

Following hyperparameter tuning, we did cross-validation and early stopping to further improve our model.

Cross-validation: Cross-validation is a technique used to assess the performance of a machine learning model by dividing the data into multiple subsets or "folds." The model is trained and evaluated multiple times, with each fold serving as both a training and validation set. This helps to obtain a more reliable estimation of the model's performance.

Early stopping: Early stopping is a regularization technique used during the training of a machine learning model. It involves monitoring a validation metric, such as loss or accuracy, and stopping the training process when the performance on the validation set starts to degrade. This prevents overfitting and helps to find an optimal point where the model generalizes well on unseen data.

Both these methods did not improve our score which suggests that the model may have reached its performance limit with the given data and architecture.

Conclusion:

Through our comprehensive analysis and modeling, we have gained valuable insights into the factors influencing restaurant inspection scores in New York City. Our findings reveal a significant impact of correlated factors on inspection scores, indicating the relevance of considering multiple variables in assessing restaurant performance.

As we move forward, there remains ample opportunity to enhance our analysis. Integrating additional data sources such as Yelp reviews and weather data holds promise for obtaining deeper insights into restaurant performance and safety, enabling a more holistic evaluation.

This project underscores the immense potential of machine learning and data analysis in elevating food safety and sanitation standards in restaurants. By leveraging these powerful tools to detect potential issues and identify areas for improvement, we can strive towards fostering a safer and healthier dining experience for all individuals.