Jongwook Choe:
Professor Tao He
Math 448 Tuesday-Thursday
Due 4/23/2024

MATH 448 Project Report: Understanding Public Data and Predict Probability of Success of Newly Opening Business.

Abstract

This report represents the results of a predictive modeling project aimed at forecasting the case count for a given neighborhood based on historical data and geographical features. The project involves data preprocessing, model training using Random Forest Regression, evaluation, and prediction.

Introduction:

Haven't you worried about stepping on poop while walking down the road? Well, I did. "San Francisco's public poo problem hits record level. They call it the Bay Area brownout. San Francisco authorities have recorded more than 125,000 cases of human feces found on public streets since 2020." (Washington Times) According to the article, just the cases reported in 2020 are more than those reported from 2011 to 2019. I believe it started after the covid hit. The city is starting to get better but we want clean shoes. I thought if we opened a shoe cleaning business in a neighborhood where there is a higher chance people shoes it could be a big hit! Throughout the analysis, the goal is to predict where will be the optimized opening place for the new business. The data will be trained by merging public data and Yelp API. I will create a new equation to estimate the success rate based on review counts and counts of surrounding human waste.

Data Description:

This dataset includes cases generally associated with a place or thing (for example parks, streets, or buildings) and created July 1, 2008 or later. Cases generally logged by a user regarding their own needs (for example, property or business tax questions, parking permit requests) are not included.

The first data set consists of 8,183 rows and 15 columns. The following table briefly describes the variables/columns of the data set.

Column Name	Description	Type
CaseID	The unique ID of the service request was created.	Number
Opened	The date and time when the service request was made.	Date & Time
Closed	The date and time when the service request was closed.	Date & Time
Updated	The date and time when the	Date & Time

	service request was last modified.	
Status	A single-word indicator of the current state of the service request. (Note: GeoReprot V2 only permits ‘open’ and ‘closed’)	Plain Text
Responsible Agency	The agency responsible for fulfilling or otherwise addressing the service request.	Plain Text
Category	The human-readable name of the service request type.	Plain Text
Request Type	The human-readable name of the service request subtype.	Plain Text
Request Details	The human-readable name of the service request details.	Plain Text
Address	Human readable address or description of location.	Plain Text
Supervisors District	San Francisco Supervisor Distrity as defined in ‘Supervisor Districts as of 2022’	Number
Neighborhood	San Francisco Neighborhood as defined in ‘SF Find Neighborhoods’	Plain Text
Point	Combination of Latitude and Longitude for Sorcrata native maps.	Location
Source	Mechanism or path by which the service request was created typically ‘Phone’, ‘Text/SMS’, ‘Website’, ‘Moblie App’, ‘Twitter’, etc but terms may vary by system.	Plain Text
Media URL	A URL to media associated with the request, e.g. an image.	Website URL

This dataset provides comprehensive information about various service requests in San Francisco, including their status, categories, locations, and timestamps. Analyzing this data can provide insights into service request patterns, response times, and resource allocation for different agencies and neighborhoods in the city.

To effectively manage and analyze the service request dataset from San Francisco, it's imperative to perform data cleaning and reduce the number of columns to enhance efficiency and accuracy in subsequent analyses.

	Year	Neighborhood	Case_Count	Latitude	Longitude
0	2008	Alamo Square	88.0	37.775396	-122.435890
1	2008	Anza Vista	20.0	37.781189	-122.441219
2	2008	Apparel City	31.0	37.737862	-122.404304
3	2008	Aquatic Park / Ft. Mason	35.0	37.805496	-122.422005
4	2008	Ashbury Heights	34.0	37.763199	-122.449364
...	...	...	...	...	...
2735	2024	Fruitvale	NaN	37.784940	-122.233930
2736	2024	Lorin	NaN	37.852550	-122.265850
2737	2024	Cleveland Heights	NaN	37.802227	-122.240967
2738	2024	Eastlake	NaN	37.797787	-122.256973
2739	2024	Cleveland Heights	NaN	37.801851	-122.244714

2740 rows × 5 columns

Model Training:

The dataset consists of two main components: train data and test data. The train data contains historical records of case counts, neighborhood information, and geographical coordinates. The test data, on the other hand, is from yelp api used for prediction purposes.

Before model training, the train data underwent preprocessing steps, including renaming columns and concatenating train and test data for feature engineering. Categorical variables, such as 'Neighborhood', were one-hot encoded, while numeric features like 'Year', 'Latitude', and 'Longitude' were passed through without preprocessing.

Model Evaluation:

The model's performance was evaluated using Mean Absolute Error (MAE) as the evaluation metric. MAE measures the average absolute difference between the predicted and actual case counts. A lower MAE indicates better predictive accuracy. The validation MAE was calculated to assess the model's performance on unseen data. Liner Regression, Random Forest, and Lasso Regression these models will be implemented and see which one best predict the data.

- Linear Regression:

Linear Regression is a simple and interpretable model that assumes a linear relationship between the input features and the target variable. Linear Regression predicts the case count based on the linear combination of neighborhood information and geographical coordinates. It provides a baseline model for comparison and is easy to interpret, but it may not capture complex relationships in the data.

- Random Forest Regression:

Random Forest Regression is an ensemble learning technique based on decision trees, where multiple decision trees are trained on different subsets of the data and their predictions are averaged to make the final prediction. Random Forest Regression captures nonlinear relationships between the input features and the target variable. It is robust to overfitting, handles high-dimensional data well, and generally provides accurate predictions. However, it may be less interpretable compared to linear models.

- Lasso Regression:

Lasso Regression is a linear model that performs both variable selection and regularization by adding a penalty term to the loss function, which encourages sparsity in the coefficients. Lasso Regression may automatically select relevant features (neighborhood information and geographical coordinates) and shrink the coefficients of less important features. It can be useful for feature selection and dealing with high-dimensional data, but it may not perform well if there are many irrelevant features or if the relationships are highly nonlinear.

```
Validation MAE for Linear Regression: 946.3355867593406
Validation MAE for Random Forest Regression: 461.75196473551637
Validation MAE for Lasso Regression: 930.9391709337834
```

Based on the validation Mean Absolute Error (MAE) values, the Random Forest Regression model has the lowest MAE, followed by the Lasso Regression model, and then the Linear Regression model. Lower MAE indicates better prediction accuracy. Therefore, the Random Forest Regression model performs the best among the three models on the validation set. However, it's important to consider other factors such as model interpretability, computational complexity, and the specific requirements of your application when choosing the best option. If you prioritize model interpretability and simplicity, you might prefer the Linear Regression model despite its higher MAE. On the other hand, if prediction accuracy is paramount and you can handle the computational complexity, the Random Forest Regression model would be the preferred choice.

Therefore, a Random Forest Regression model was chosen for its ability to handle complex relationships and provide robust predictions. The model was trained using the preprocessed train data, where the target variable was the case count, and features included neighborhood information and geographical coordinates.

Prediction:

After model training and evaluation, predictions were made on the test data to forecast the case counts for different neighborhoods. The predicted case counts were added as a new column in the test data. The test data was then sorted based on the predicted case counts in descending order.

To further analyze the predictions, **the top 10 businesses with the highest predicted case counts** were identified. This was achieved by merging the sorted test data with existed Yelp business data using latitude and longitude coordinates. The merged data provided insights into the predicted case counts alongside relevant business and neighborhood information.

	Name	Neighborhood	Predicted_Case_Count
0	Carlos Shoe Service	South of Market	11718.58
1	Museum Parc Cleaners	South of Market	11718.58
2	Bay Breeze Cleaners	Lower Nob Hill	11256.20
3	Busy B Coin	Lower Nob Hill	11255.89
4	Busy B Coin	Lower Nob Hill	11255.89
5	Today's Laundromat	Lower Nob Hill	11255.89
6	Today's Laundromat	Lower Nob Hill	11255.89
7	Busy B Coin	Lower Nob Hill	11255.89
8	Busy B Coin	Lower Nob Hill	11255.89
9	Today's Laundromat	Lower Nob Hill	11255.89
10	Today's Laundromat	Lower Nob Hill	11255.89
11	Cleanly	Lower Nob Hill	11252.35
12	Cleanly	Lower Nob Hill	11252.35
13	Cleanly	Lower Nob Hill	11252.35
14	Cleanly	Lower Nob Hill	11252.35
15	Laundry Locker by Mulberrys	Lower Nob Hill	11252.05
16	Laundry Locker by Mulberrys	Lower Nob Hill	11252.05

The predictive modeling approach presented in this report demonstrates the potential for forecasting case counts based on historical data and geographical features. The Random Forest Regression model, coupled with appropriate preprocessing and evaluation techniques, offers a robust framework for predicting case counts in different neighborhoods. Further refinement and validation of the model could enhance its accuracy and applicability in real-world scenarios.