Problem 1:

Can we predict the probability of issue resolution when photographic evidence is submitted, while considering the impact of the source and location of the report?

Approach : Usage of logistic regression on the dataset wherein the photos of the issues were submitted.

Flagging: Cases that were closed within the target date were flagged as 1, issue not closed as 0.

Impact variables considered : latitude, longitude, source\_City Worker App, source\_Constituent Call', 'source\_Employee Generated', 'source\_Self Service'

Model 1: Logistic Regression on imbalanced dataset

Results:

F1\_ score : 0.8825

Accuracy: 0.7898

Precision: 0.7898

Recall: 1.0

Confusion Matrix:

Pred:0 Pred:1

Actual:0 0 4664

Actual:1 0 17530

This model could only give TP & FP. TN & FN values could not be identified.

This was probably on the account of imbalanced dataset.so, we tried with another modelling approach.

**Model 2: Logistic Regression on balanced dataset**

**Results:**

**F1\_ score : 0.5473**

**Accuracy: 0.4991**

**Precision: 0.5024**

**Recall: 0.6010**

**Confusion Matrix:**

**Pred:0 Pred:1**

**Actual:0 6915 10565**

**Actual:1 7082 10670**

This model helped us predict the TN & FN but the F1 score and accuracy was compromised. This led us to try another modelling approach.

Model 3: Random Forest on imbalanced dataset

Results:

F1\_ score : 0.8724

Accuracy: 0.7821

Precision: 0.8112

Recall: 0.9436

Confusion Matrix:

Pred:0 Pred:1

Actual:0 816 3848

Actual:1 988 16542

This model gave balanced result and could be deployed for predicting the cases closure within the deadline.

Recommendation: 311 agency can use this model to feed in the cases upon their receipt and find out whether this issue can be closed within the target closure time. For the cases where the model predicts that it won’t be closed on time, special attention could be given or system may be put in place to reduce the delay. This will help them improve their efficiency.

Problem 2

How can we leverage **seasonality patterns** from historical data to make accurate **predictions of the duration** it takes to **close a case?**

In problem 1 , we worked on predicting the closure within deadline based on the submitted photo and the source of complain. Now, we wanted to check the effect of season and predict the no. of days it will take for the closure of the complaint in different season.

Approach : Usage of multiple regress**ion** on the dataset.

Impact Variables: dept\_responsible, team\_responsible, neighborhood, submittedphoto ,latitude, longitude, source\_City Worker App, source\_Constituent Call, source\_Employee Generated, source\_Self Service , season.

Flagging: A new column capturing the season was created which was further dummified.

Model 1: Multiple regression model on the impact variables.

**RMSE: 375.54**

The impact variables upon dummification gave 65 predictor variables which made the model complex resulting in very high RMSE. So, we had to try another approach.

Model 2: Multiple regression with Forward Feature selection.

**RMSE: 19.38**

This gave comparatively much better result and the features that were selected for the prediction were reduced to **20**. The features along with their coefficient are

|  |  |  |
| --- | --- | --- |
| **S.No** | **Feature** | **Coefficient** |
| **1** | latitude | -16.88 |
| **2** | dept\_responsible\_Inspectional Services | -1.97 |
| **3** | dept\_responsible\_Parks & Recreation Department | 3.95 |
| **4** | dept\_responsible\_Property Management | 80.99 |
| **5** | dept\_responsible\_Public Works Department | -3.73 |
| **6** | team\_responsible\_Building | 15.06 |
| **7** | team\_responsible\_Highway Maintenance | 3.78 |
| **8** | team\_responsible\_Housing | 3.83 |
| **9** | team\_responsible\_Recycling | 11.37 |
| **10** | team\_responsible\_Signs & Signals | 8.95 |
| **11** | team\_responsible\_Street Cleaning | 1.10 |
| **12** | team\_responsible\_Street Lights | 19.25 |
| **13** | team\_responsible\_Traffic Management & Engineering | 9.54 |
| **14** | team\_responsible\_Trees | 28.47 |
| **15** | neighborhood\_Dorchester | 1.48 |
| **16** | neighborhood\_Roxbury | 1.09 |
| **17** | neighborhood\_South Boston / South Boston Water... | -1.03 |
| **18** | season\_Spring | 3.53 |
| **19** | season\_Summer | 1.75 |
| **20** | season\_Winter | 5.70 |

Positive value of the coefficient indicates that it will positively affect (increase) the count of days and negative is vice versa. The magnitude of coefficient defines the strength by which the no. of days will be affected due to the respective coefficient.

latitude: The coefficient for latitude is negative, which suggests that as the latitude of the case location increases, the expected number of days to close the case decreases. This could be due to factors such as weather conditions, demographics, or different workloads for the departments that handle the cases in different locations.

dept\_responsible: The coefficients for the different departments are positive or negative depending on the department, which suggests that some departments take longer than others to close cases. For example, the Property Management department has the highest coefficient, indicating that cases handled by this department take significantly longer to close than cases handled by other departments.This comparison is wrt Animal control department which has been taken as reference.

team\_responsible: Similar to the dept\_responsible coefficients, the coefficients for the different teams suggest that some teams take longer than others to close cases. For example, the Trees team has the highest coefficient, indicating that cases handled by this team take significantly longer to close than cases handled by other teams. This comparison is wrt Administrative & General Requests team which has been taken as reference.

neighborhood: The coefficient for Dorchester is significantly positive, indicating that cases located in this neighborhood take longer to close than cases located in other neighborhoods. This could be due to various factors such as the population density, income levels, or infrastructure of the neighborhood. This is in reference to Allston.

season: The coefficients for the different seasons are positive and are of different magnitude, indicating that cases reported during different seasons take different amount of time to close. For example, cases reported during the Fall season take the shortest time to close which are our reference variable, while cases reported during the winter take longer. This could be due to various factors such as weather conditions, workload of the departments, or staffing levels during different seasons.

Based on these insights, the Boston 311 agency can take following steps to improve upon their service efficiency:-

1. They need to focus on exploring the root cause for the delay of the issues in the area when we move move from west to east across Boston.
2. Property department current working system needs to be reviewed as it is taking the longest time for the case closure.
3. Similarly for the teams struggling to keep up with the average closure time like that of team\_responsible trees, a special monitoring for the identification of probable root cause need to be in place.
4. Cases belonging to Dorchester & Roxbury needs special attention. The reasons for delay in the case resolution over there can be studied for improvement.
5. For Winter season , additional manpower can be hired for timely response to the increased no. of cases and special working gears should be incorporated to avoid any delay due to bad weather.

