Impact of Construction Projects on the City of Kingston - Impact IQ

CISC 451 – Topics in Data Analytics

Project - Final Report

Aurko Bhattacharyya 20115983, Nikolay Alabi 20099148, Vedant Srinivasan Kartik Srinivasan 20090733

Problem Definition

Kingston urban planning often disrupts the daily lives of citizens as the construction zones often disturb the flow of traffic, contribute to an increase in the number of hazards in an area, and result in more complaints to local government. Some construction projects are more carefully planned and cause less of a disruption to the lives of surrounding residents than others. These projects should be prioritised, and more disruptive projects should be modified so that they can cause less of a disruption. Our solution proposes a model that can predict if a project is going to be disruptive. This problem will have to be solved in two steps. First, we must determine metrics for "disruptiveness" which we will quantify by comparing service requests, traffic incidents, hazardous incidents, and other similar indicators between periods where construction projects were and were not in progress. Construction projects with minimal changes to the disruptive metrics in their surrounding areas are deemed to be minimally disruptive. Projects with significant changes in disruptive metrics are deemed to be disruptive. We will then create a model that can predict whether or not a project will be disruptive based on the project's characteristics such as type of project, location, proposed period, and other factors. This model can identify disruptive projects prior to them ever being implemented and can help companies and government's modify the project plan to make it less disruptive and improve the lives of their constituents.

Detailed Data Description

Four datasets were used in this project:

- Capital Planning (Points) and Capital Planning (Lines) from the Kingston Open Data Portal. This data set shows the coordinate location, date, type, and status of capital projects undertaken by the city in .csv format. The points data set shows projects that occur around one point (such as building construction) and the lines data set shows projects that stretch over more space (such as sewage line construction). This dataset had over 300 projects ranging from 2016 to the present.
- Service Requests dataset from the Kingston Open Data Portal. This data set includes information on service requests such as the type of request, dates requests were opened and resolved, request status, request location, and other minutiae about the nature of the requests in .csv format. Service requests include a wide variety of municipal services requested by residents such as requests concerning garbage pickup, the bus service, vehicle parking, and general complaints and inquiries. Not all service requests, such tax inquiries or similar requests, are not relevant to our project so irrelevant requests had to be filtered out. This dataset includes service request data from October 2019 to the present as it is regularly updated.

- AADT (Annual Average Daily Traffic volume) for Kingston Intersections, provided to us by Craig Hollingsworth on behalf of the city of Kingston. This dataset shows the AADT from each cardinal direction for Kingston intersections in a given year in a .csv format. So for instance, one record would be the intersection for Princess St. and University St. There would be 4 columns corresponding North, South, East and West with the annual average daily traffic volume for each respective direction as values. This dataset contains 676 records spanning from 2010 to 2021. It should be noted that AADT data was not collected on the same intersection every year so there is some variability in the intersections represented from year to year.
- Kingston Traffic Collision dataset showing the date, time, and intersection of collisions in Kingston in a .csv format. This dataset also includes numerical codes for features describing the collision such as the vehicle types involved and road conditions. This dataset included 5,519 records and included collision data from January 2016 to the end of 2018 in December. This dataset was also provided to us by Craig Hollingsworth.

Visualisations and Statistical Measures

Due to the geographic nature of this problem, visualising our datasets on a map was an important graphic to produce. When evaluating the disruptiveness, it is important to have a sense for where projects are taking place and their spread.

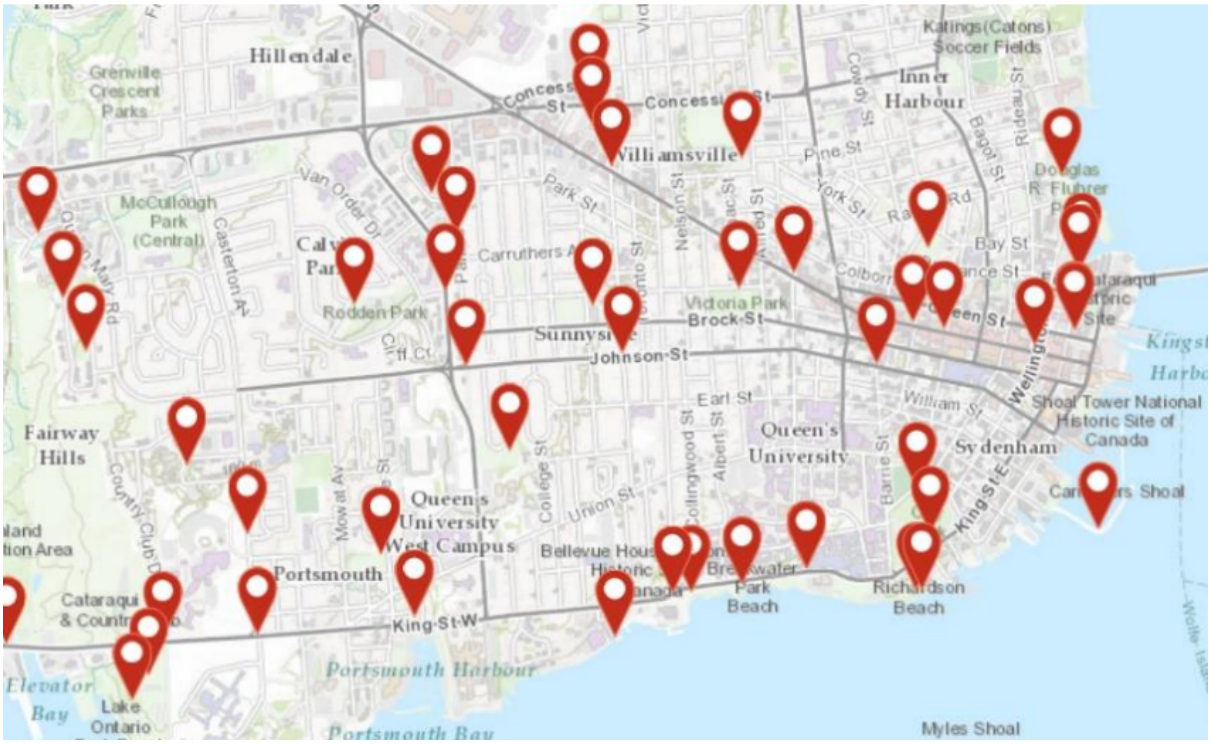

Figure 1: Map showing some projects outlined in Capital Planning (Points) dataset

As we hypothesised that construction projects would increase some types of service requests, such as road blockages, utility repairs, noise complaints, etc., we found it important to view the geographic distribution of service requests.

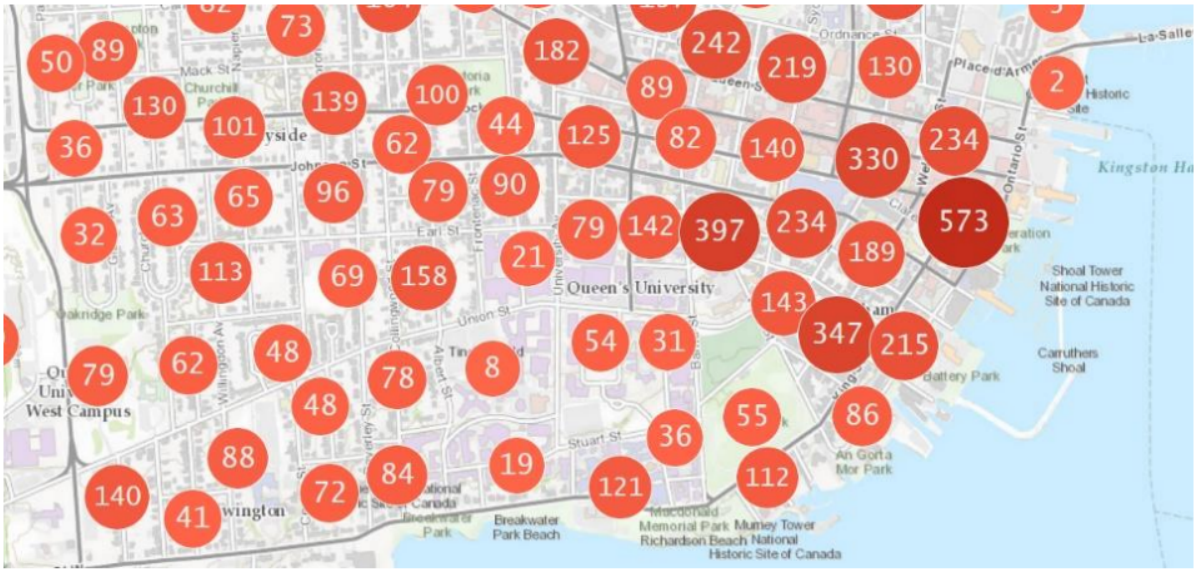

Figure 2: Map showing concentration of service requests from Service Requests dataset.

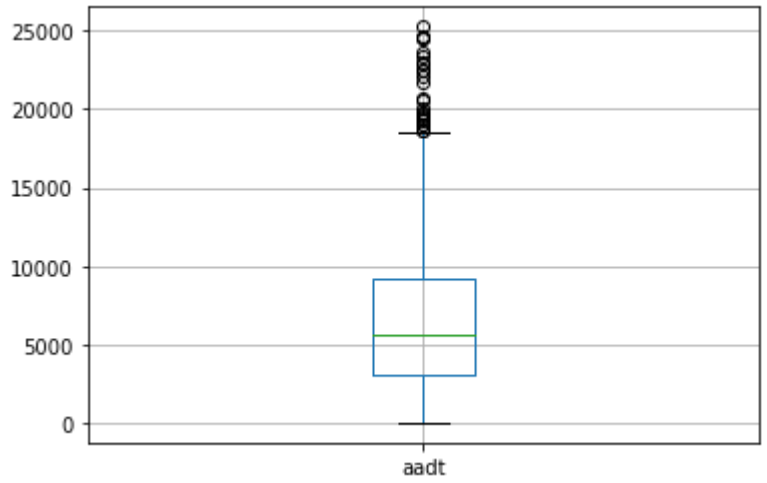

Figure 3: Boxplot showing AADT averaged across all directions in each intersection

Challenges

1. The first challenge that came up in this project was obtaining good traffic data for the Kingston area. We were unable to find free historical traffic data on the Kingston Open Data Portal or other publicly available sources so we had to reach out to the city of Kingston directly to obtain this data. Our original plan was to quantify the traffic changes before and during a construction project in order to classify the project as disruptive or not. We were able to reach out the City of Kingston who provided us with traffic data, the dataset was not ideal because it measured traffic from each cardinal direction at specific intersections which necessitated lots of pre-processing on the dataset as well as a lot of assumptions to make it workable with the rest of our data. Some of the assumptions we had to make included the length of each direction from each intersection. For instance, the dataset only indicated that 300 cars came from the North direction at Princess and University intersection. However, this

did not indicate how far north this could be extrapolated to. Therefore, we had to make the assumption that each direction was 1.5km, as this is the average distance between intersections in Kingston.

2. Another challenge we faced was associating geographic locations amongst datasets. While the Open Kingston datasets provided geo coordinates of the locations of the construction projects as well as the service requests, the AADT and collisions dataset did not provide explicit coordinates. Instead they provided the intersection names. Thus, we had to utilise a Google Maps API to search the intersection name and return the coordinates.

3. Another key challenge was the incompleteness of our data. While we had construction projects from 2016-2022, only the AADT dataset actually contained reported traffic that spanned the range of the construction projects. However, as mentioned the AADT dataset was also not complete in that certain years were missing some data on some intersections. Because the service requests dataset contained data that started in Oct 2019, we decided to multiply the number of service requests in 2019 by 4 to make it representative of an entire year. Under that assumption, we had service requests data from 2019 till the present. And lastly, the collisions dataset only contained collisions from 2016-2018. Thus, our approach was to aim to get at least some type of information about each project whether it was collisions, requests, or traffic, or ideally a combination of several. Important to keep in mind is that since we were comparing the before and during a construction project, we needed data for two years of a certain disruptive metric. For instance, if a project began construction in 2019, we had collision data for the year prior but not for the year during, making collision data unusable. For the construction year 2019 we had service requests, however we did not have requests for the year prior, also making requests unusable. In the end for some projects we were only left with traffic data. So essentially our approach was to get disruptive information for each project, whether it came from one source or multiple. However, another issue arose: because on certain years some intersections were included in the dataset and other years they were excluded this could lead to a bias in our calculations. For instance, if the previous year of a construction project had 3 intersections worth of traffic data and the year during only 1 intersection worth of data, if we simply summed the traffic before and during, this would lead to a decrease in the before and during, which presents a calculation bias. To counteract this, we took the average traffic based on the numbers of intersections and streets that were considered.

4. Another challenge was in determining which projects were disruptive versus minimally disruptive based on the before and during disruption metrics. While we calculated the percent change for every disruption metric, we recognized we were going to have to make some assumptions as to what is considered disruptive. While it is sensible that an increase in service requests and collisions is disruptive, what level of increase is considered to be very disruptive and not just a small change from year to year. Moreover, the question remained of whether an increase or a decrease in traffic is considered disruptive. In the end we decided that decreases in traffic is considered disruptive as less cars would be able to travel through a disruptive project. Secondly, we decided to take the following approach in classifying

disruptive projects. For any of the 3 disruptive metrics: collisions, traffic, and requests, any projects that were above median percent change were considered to be disruptive. However, as mentioned, for some projects we have more than one metric of disruptiveness, for those projects we looked to see if both metrics were in agreement with disruptiveness, if so, the project was labelled disruptive, if not, then it was labelled non-disruptive. While this assumption does have its flaws in equally weighting metrics, if simplified the preprocessing work.

Methodology

Labelling Projects as Disruptive vs. Non-Disruptive

Due to the complexity of location based data and incongruence between how location is denoted across datasets, we had to find a way to associate relevant items to the same geographic area and timeframe. To solve this we first created a 2.25km^2 square around completed and in-progress construction projects in the “Points” dataset. Since degrees of latitude are of a constant size (1/360th of the Earth’s 40,075 km circumference) , we can convert from degrees of latitude to kilometres by simply dividing the circumference by 360. Our square has sides of 1.5 km so we can calculate that the horizontal sides of the square will be approximately 0.0135 degrees of latitude (The length of a longitudinal line is dependent on the latitude and after some research, we found that a line of longitude at a latitude l can be expressed in kilometres as $\cos(l) * 40075$ km. To get this in km/degree of longitude we can simply divide the previous expression by 360. To implement this in code, we split the coordinates of the construction project points into separate latitude and longitude columns in the dataframe. Then we wrote a function to convert kilometres to degrees of latitude and longitude in order to draw lines of degrees equivalent to 1.5 km in the positive and negative x and y directions. Connecting the outer points of each line produces a 2.25 km^2 square that is considered the area of the project.

For the preprocessing for the “Lines" dataset a similar process was used, however, because the project is a line and not a point the following approach was taken. A coordinates-based rectangle was created by taking the smallest and largest longitude and latitude values from the geo coordinates for each project. Once the rectangle was created, it was expanded 1.5 km in every direction using a similar process outlined in the preprocessing of the points dataset.

Now that the area for each construction project had been defined, the next task was to define the area for the other items that could indicate the effects of construction in that area, namely service requests, collisions, and traffic volume. Using the Google Maps API, we fed the intersection names from the AADT and collisions dataset to get the location of the intersection in terms of latitude and longitude from Google Maps. This allowed us to view the traffic flow and collision data in the context of their location relative to the area of construction projects. Since the service request location data is already in terms of latitude and longitude, all of the relevant factors we were looking at are now comparable.

Some additional preprocessing was done on the datasets. For the construction projects dataset, we excluded projects that had not yet begun construction, leaving us with projects who had been completed and those currently under construction. For the service requests dataset, we removed any records without geo coordinates. Additionally, because the service requests dataset contained data that started in Oct 2019, we decided to multiply the number of service requests in 2019 by 4 to make it representative of an entire year. For the traffic dataset, a geo coordinate line was created between the coordinates of the intersection and 1.5km in every direction (North, South, West, and East). The dataset was restructured (melted) so that the value of each column (North, South, West, and East) became its own record associated with a year.

We created functions that did the following:

1. For each project, look at the year before construction.

    1.1 Look through the service requests dataset for service requests that occurred in the year before construction and whose geo coordinates were within the geo coordinate square/rectangle of the construction project. Sum the number of requests that met the criteria and denote this as the number of requests before.

    1.2 Look through the collisions dataset for collisions that occurred in the year before construction and whose geo coordinates were within the geo coordinate square/rectangle of the construction project. Sum the number of collisions that met the criteria and denote this as the number of collisions before.

    1.3 Look through the traffic dataset for traffic that occurred in the year before construction and whose geo coordinates were within the geo coordinate square/rectangle of the construction project. Sum the number of traffic values that met the criteria and divide the summed value by the number of streets contributing to the value and denote this as the average traffic before.

2. For each project, look at the year during construction and repeat steps 1.1 - 1.3 but during the year of construction.

For each project, we had the before and during disruptive metrics. We then calculated the percent change for each metric where applicable (As mentioned earlier, some projects did not have complete data for all metrics, so we only considered complete metrics those with data for the before and during years). For any of the 3 disruptive metrics: collisions, traffic, and requests, any projects that were above median percent change were considered to be disruptive (Figures 3-5).

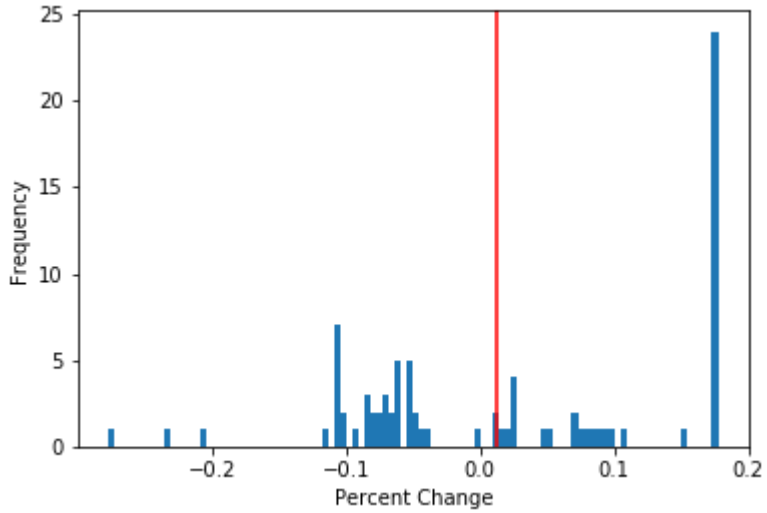

Figure 4. Collisions percent change histogram. Red line indicating median.

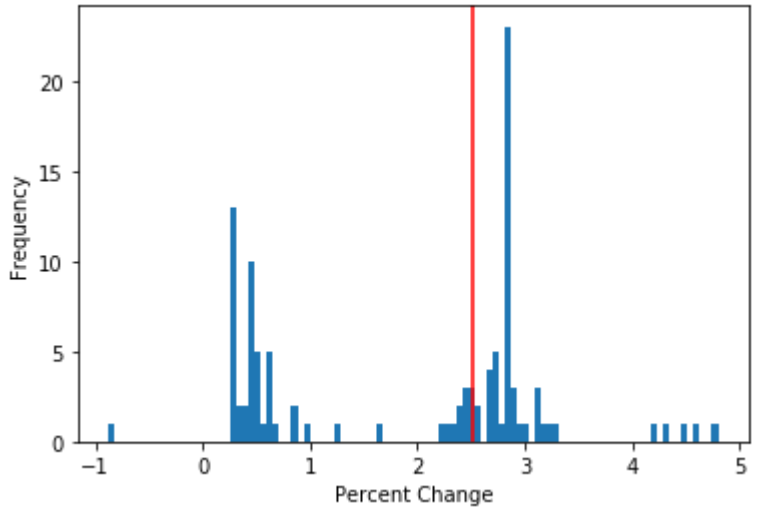

Figure 5. Requests percent change histogram. Red line indicating median.

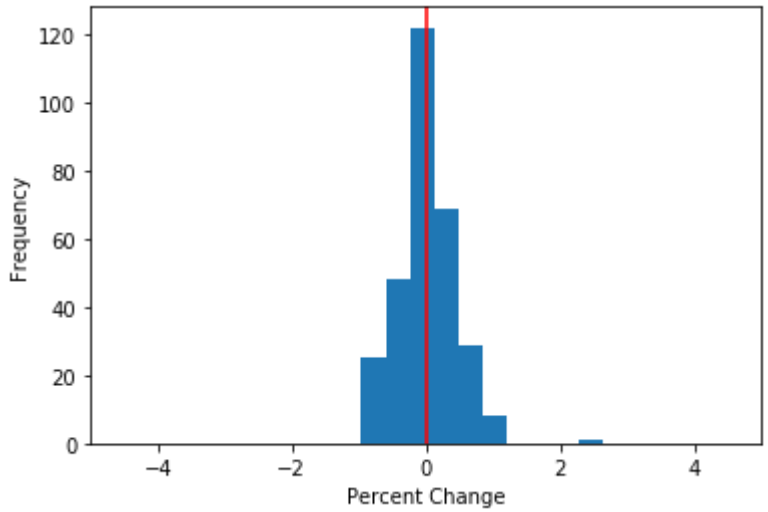

Figure 6. Traffic percent change histogram. Red line indicating median.

As mentioned, for some projects we have more than one metric of disruptiveness, for those projects we looked to see if both metrics were in agreement with disruptiveness, if so, the project was labelled disruptive, if not, then it was labelled non-disruptive. We labelled 95 projects disruptive and 218 projects minimally disruptive.

After labelling the projects as disruptive or not, we removed any data relating to the future of each project: "during" metrics and information on its completeness. Removing this data could allow us to treat the modelling step as realistic, since in the future we would not have any data about a project's future, only a project's plan and the current information prior to construction.

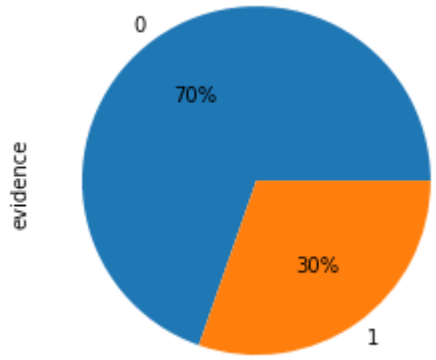

Figure 7: Pie-chart showing ratio of minimally disruptive (0) to disruptive (1) projects

Model Development

Once we have the labels for each row, we can start transforming all the columns to make sure they can fit with the model. For categorical features, we used one-hot encoding. We also had to change the location (lat, lon) features into categorical features somehow. So we performed k-means clustering to group lat and lon to different regions. To find the optimal number of clusters, we used inertia and silhouette scores (Figure 8). After clustering, each lat-lon was mapped to a centroid. We also performed scaling on all the numerical columns. At the end of this step, all values were between 0 and 1.

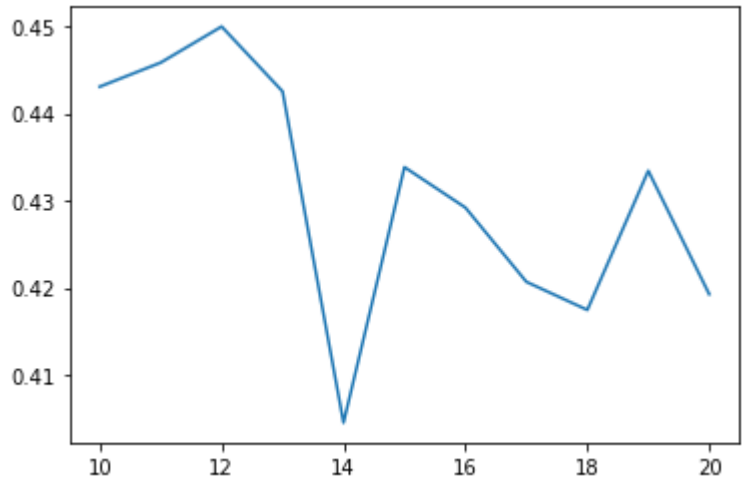

Figure 8. Silhouette score for different number of clusters

For the actual modelling step, we divided the data into a stratified 75-25 train-test split. We elected to focus on the logistic regression and SVM classification models as they were most appropriate given the type of data we had. Particularly because we had only ~200 records to train on, many other models such as random forest and naive bayes require a larger amount of records for adequate performance. Moreover, SVM and logistic regression are ideal for datasets with an imbalance in classes, such as our own. Lastly, we utilised randomised cross-validation in the training set for hyperparameter tuning.

Recreating Results

To recreate the experiment, please use the code in this order: raw_label_creation, feature_eng, modelling. The data is provided in the data folder. In order to use any previously trained models, please train the data on the Pickle dumps provided in the models folder.

Evaluation

After modelling, we used accuracy, recall, precision, and F1 scores. Logistic regression gave better accuracy scores on the testing dataset.

Metric/Model	**Logistic Regression**	**SVM**
Accuracy	72.15%	59.49%
Precision	66.67%	25.00%
Recall	16.67%	16.67%
F1	26.67%	20.00%

*Accuracies may vary ± 5%

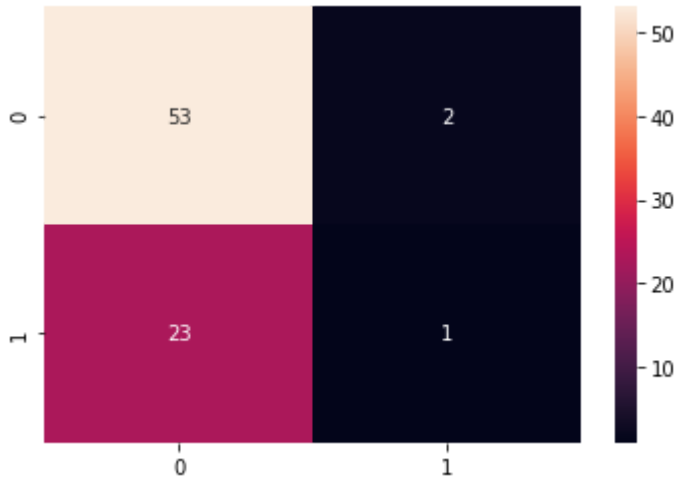

Figure 9. Confusion matrix of testing set for LogReg. 1 indicated disruptive project, 0 indicates minimally disruptive project.

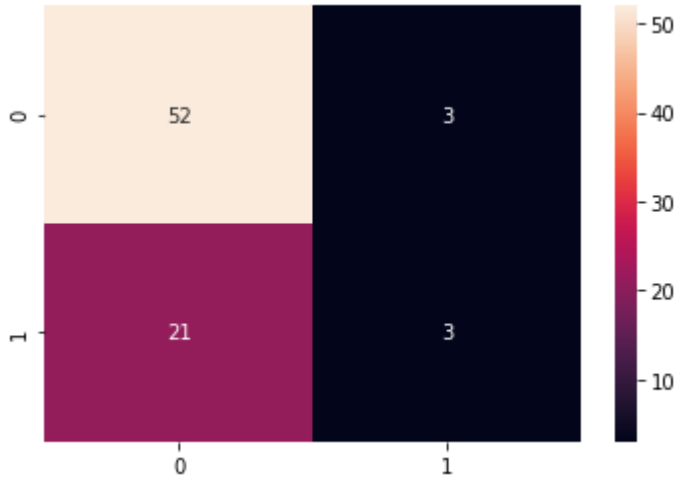

Figure 10. Confusion matrix of testing set for SVM. 1 indicated disruptive project, 0 indicates minimally disruptive project.

Conclusions

Though several assumptions were made in the preprocessing steps of gathering disruptiveness metrics of construction projects and then labelling projects as disruptive or not, we were successfully able to create a dataset that a ML model could learn from to predict if a construction project would be disruptive or not. We were able to build a successful model that could predict a project's disruptiveness with 72% accuracy. Because a model was able to identify patterns and trends within the data, it validates a lot of the assumptions we had taken in labelling the data. If we had made very poor assumptions in labelling the construction projects as disruptive or not, a model would have not been able to predict on the data. Secondly, the model's success also indicates the predictive power of the features of the dataset. It stands to reason that more feature engineering and obtaining more data about each project can help improve the model's accuracy.

Future Work

There is lots of future work and iteration that could be done on this project. Firstly, doing the project again and seeing how the omittance of some assumptions that were made would affect the final results. It would also be interesting to see this process repeated using data from cities larger than Kingston as there will be more data to train and evaluate the model and construction projects have the potential to be more deleterious in high-density areas. For example in a population dense area like Toronto, the impact of construction projects may be seen more dramatically through features such as service requests and/or collisions since there are more people affected by the construction. Conducting analysis on construction projects in this type of environment may further elucidate the impacts of construction projects on the surrounding areas.

The traffic data we received for free was very limited so it would also be interesting to see if paying for more complete traffic data and other geographic datasets would yield better results. We were unable to ascertain the full impact of construction projects on traffic because the AADT datasets did not have the same intersections recorded every year. Although our model was evaluated to have relatively high accuracy, there is always further refinement to be done in terms of selecting the most predictive features or adding other features that we missed or were unable to add. For example, additional features such as data on foot traffic or emergency services in an area could help create a more accurate model and give more holistic insight into the impact of construction projects in an area. Overall this project would benefit from additions that we were unable to make due to the technical and financial limitations we were under. Such additions such as a larger area being evaluated (i.e. A larger city or multiple municipalities rather than just Kingston) or access to municipal data that is not publicly available for free would be promising avenues for future work on this project. This project lays a promising groundwork for future, more expansive analysis.