San José Post-Pilot Report

Julien Refour, Data Scientist

March 5th, 2019

1 Introduction

Every year, the City of San José spends approximately $300 000 to study the movement of road users along corridors and at intersections. In September 2018, the City of San José engaged UrbanLogiq in a pilot project to develop a platform with the goal of maximizing the use of some otherwise under-exploited datasets held by the City. Towards this goal, UrbanLogiq produced three main outcomes: First, a platform was created to ingest, visualize and analyze data relating to the use of San José's road network. Second, an exploratory dashboard was created to query and analyze roughly ten years worth of incident data that the City had collected, permitting the user to select incidents according to where and when they happened, and analyze which intersections may be prone to specific types of incidents. Finally, a machine learning model was developed to predict if an intersection would see an incident resulting in a fatal or major injury during a particular period in time. The model was trained using the incident data, as well as geospatial data provided by the City.

This report describes in detail the results of UrbanLogiq's engagement with the City. §2 describes the datasets that were delivered to UrbanLogiq by the City, as well as how each of the datasets was used to develop the solutions put forward by UrbanLogiq. §3 presents the procedure followed to analyze the incident data and create new features using geospatial datasets provided by the City. §4 discusses how a binary classification model was trained using the City's incident geospatial data. The performance of the model is also evaluated. Finally, §5 summarizes the report, and provides some avenues for future work.

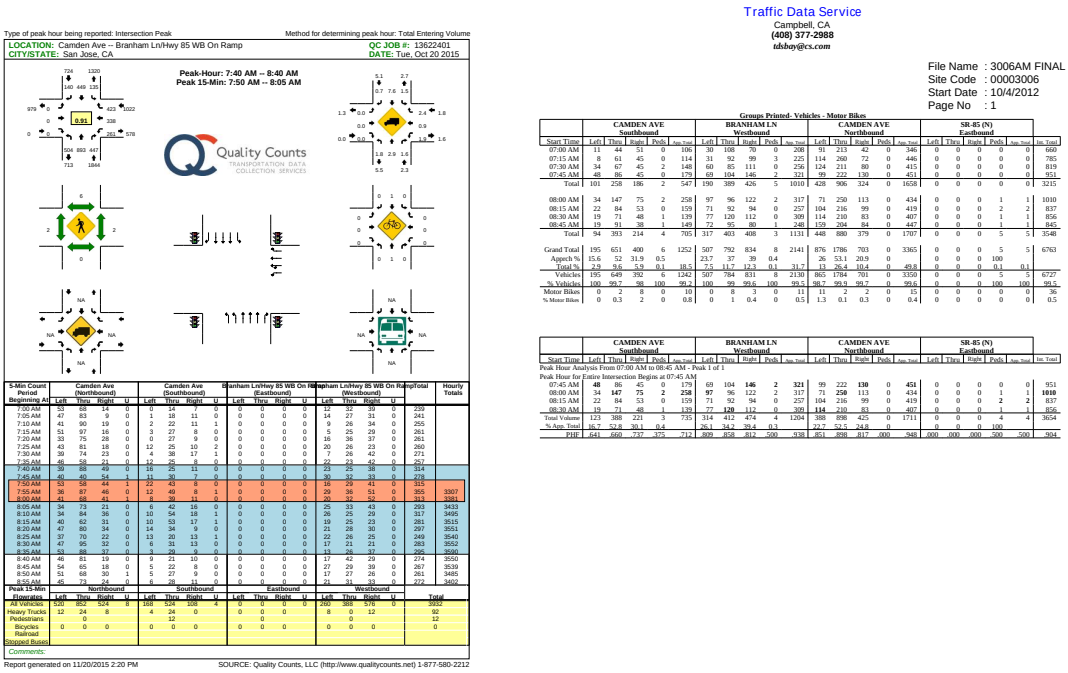
2 Datasets

In total, seven (7) distinct datasets were used to develop the pilot project delivered to the City. In this section, a brief summary of each dataset is given. Additionally, the use of the dataset in the pilot project is described, as well as the challenges and transformations associated with that use.

2.1 Consultant Turning Movement Count (TMC) Reports

2.1.1 TMC Summary

The TMC report, at its base, is a study of how an intersection is used by the surrounding population. More specifically, a TMC report will detail the volume of vehicles that each lane present at an intersection sees, over a specified amount of time. A TMC report is usually created through highly manual processes, such as placing humans at the intersection to actually count the number



(a) **(b)**

Figure 1: Two examples of TMC reports commissioned by the City of San José. These two reports were produced by two different consulting companies. (a) shows a TMC report created by Quality Counts, while (b) shows a similar report produced by Traffic Data Service. There is a wide variation between the two in terms of time increment (5 minutes vs. 15 minutes), time range (7:00 AM-8:55 AM vs. 7:00 AM-8:45 AM) and types of turns (Left-Thru-Right-U vs. Left-Thru-Right).

of vehicles each lanes sees, or recording the intersection using cameras and watching the footage back to count. The results are typically delivered as a static document. Figs. 1a and 1b show examples of TMC reports that were purchased by the City of San José in order to study an intersection in 2012 and again in 2015. There is wide variation between the styles and content of the two reports, as they were produced by different consultation firms. Variations between the types of turns measured, the time increments, the time ranges and the types of road users measured can be seen.

2.1.2 Use of TMC Reports

For the pilot project, UrbanLogiq undertook the task of digitizing five (5) years, from 2013 to 2018, of TMC reports that the City possessed. In total, roughly 300 intersections had at least one TMC report associated with it in the the given time range, amounting to approximately 2500 unique reports marked for ingestion. These 2500 reports were split between 12 different vendors. As Fig. 1 shows, the reports did not have consistent formatting or contents. Even within a single vendor, differences in formatting and contents could be found.

Once digitized, the reports were ingested into a cloud-based platform that allowed the reports to be queried in both space and time. The platform also some typical analysis done on the data from the reports. The data from each report was geoencoded to a point of latitude and longitude referring to the intersection it studied. Fig. 2 shows the results of the geoencoding process. The

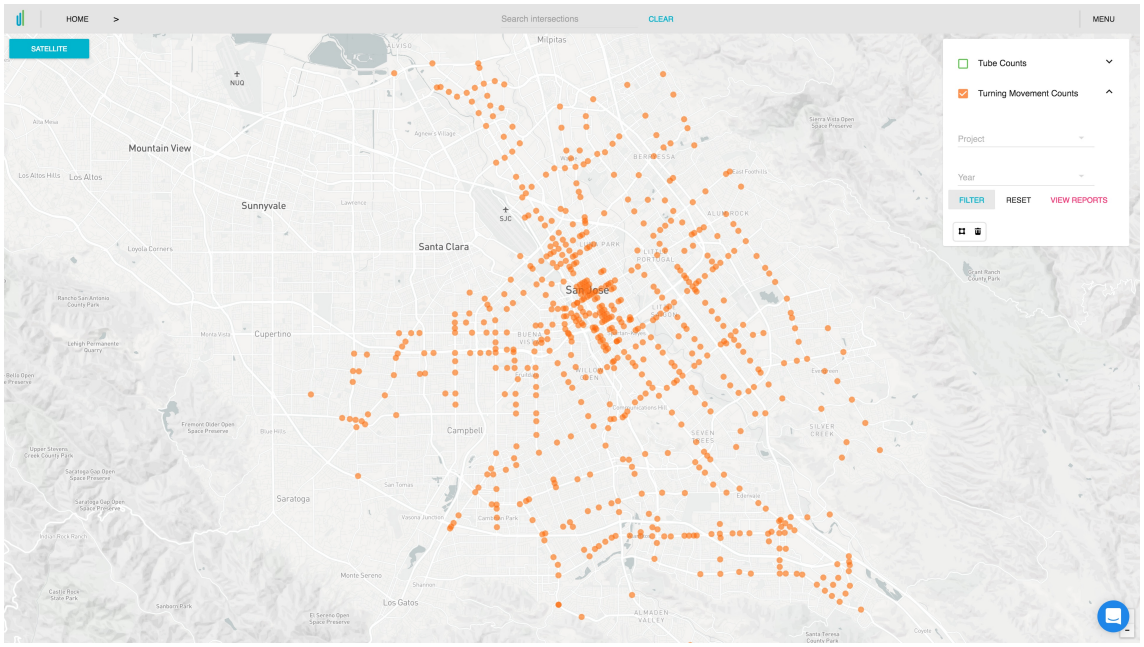


Figure 2: The home screen of the cloud platform delivered to the City of San José. Here, the user can search for TMC reports by location (by clicking on the intersections, represented as orange dots), name (by using the search bar at the top of the page), year and project.

platform interface allows a user to select an intersection based on its location or name, and filters historical reports based on the year and project name. Once an intersection has been selected, the user is brought to the TMC report page itself, shown in Fig. 3, where the volume counts can be viewed for each lane present at the intersection, filterable by time and road user type. The total volume for each lane of the intersection, for each direction and turn, is shown. Peak Hour (PH) and Peak Hour Factor (PHF), which measure the peak traffic hour and what fraction of total traffic is contained in that hour, respectively, are computed and displayed for the chosen data. Selecting the "View Details" button will bring the user to the analysis page, shown in Fig.4, where reports can be generated and exported. Fig. 5 shows the comparison function available when an intersection has more than a single study available. Finally, the original TMC reports were linked to each intersection and made exportable.

2.2 Vehicle Volume Data

2.2.1 Volume Data Summary

TMC reports represent one method of measuring the use of a city's road infrastructure. Another method is the measurement of how many vehicles pass along a corridor within a given time, commonly done using pneumatic tubes laid across a road. When vehicles pass over the tubes, the change in pressure created indicates the passage of a vehicle. In addition to counting vehicles, pneumatic tubes can also be used to measure the speed and type of the vehicles they are counting. Fig. 6 shows two examples of volume studies commissioned by the City of San José.

2.2.2 Use of Volume Data

For the pilot, UrbanLogiq digitized vehicle volume studies produced for 2017, totalling to roughly 300 individual studies at 200 unique locations. Like the TMC reports, the volume studies came in

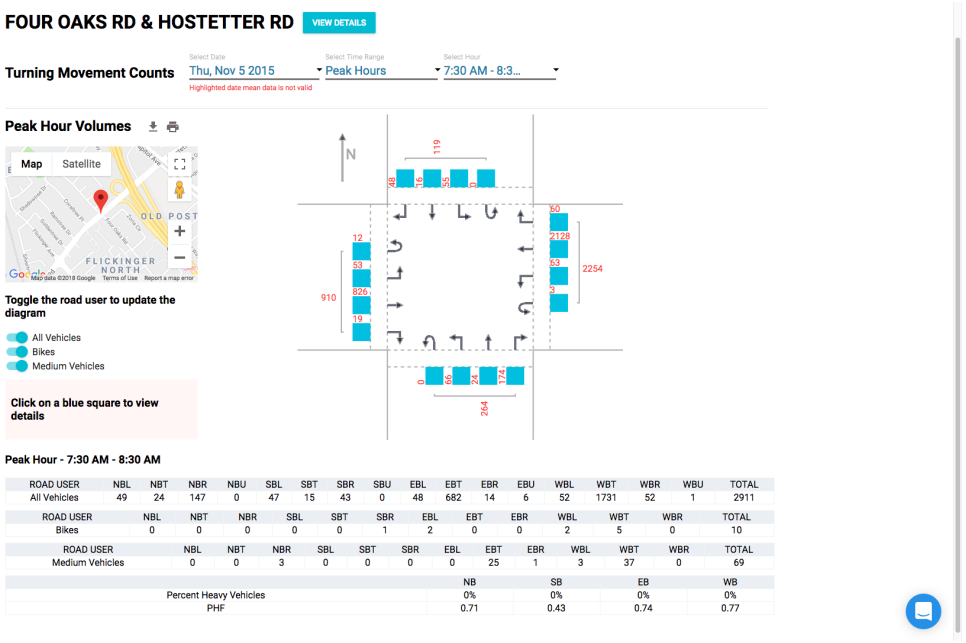


Figure 3: Once a user has selected an intersection from the view offered in Fig. 2, they are brought to this page, where the latest TMC report for the chosen intersection is displayed. Here, the user is presented with the digitized version of the TMC where the volume counts for each lane at the intersection can be viewed filterable by time and by the types of road user counted.

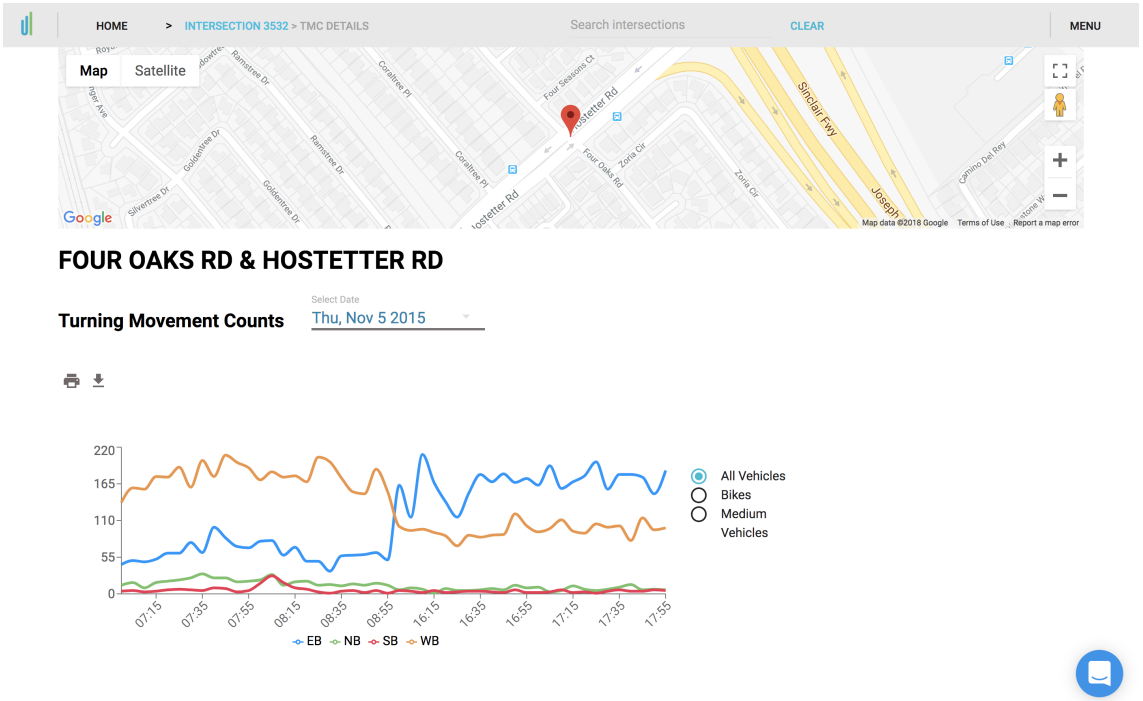


Figure 4: Example of the analysis page available for each intersection. A summary of the chosen date is shown, split by direction of travel and vehicle type. If multiple studies are available, they can be selected and compared, an example of which can be seen in Fig. 5.

a variety of formats. Once digitized, these reports were then geoencoded to the segments they measured and ingested into the same platform as the TMC reports. During the geoencoding process, properties such as the speed limit, road type, and road function were attached to each report. This was made possible due to the existence of street center-line data (detailed later in this section).

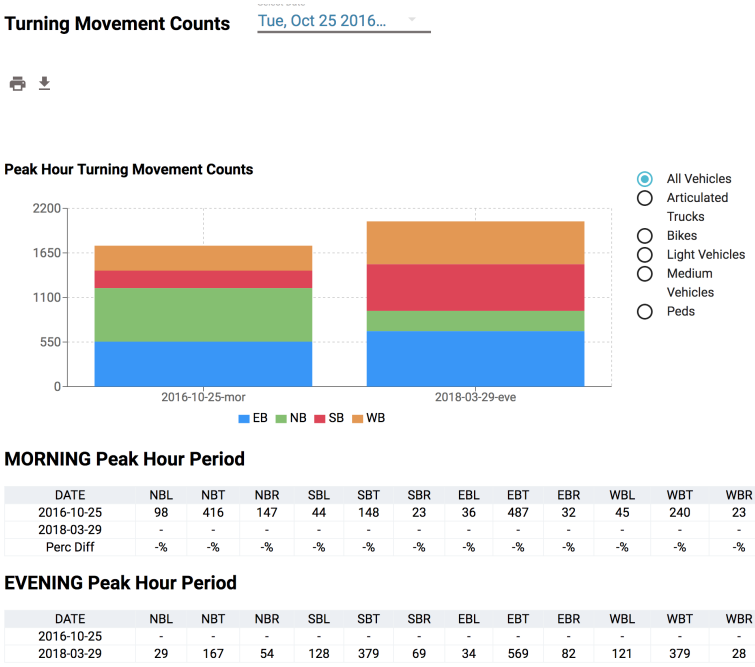


Figure 5: Example of the comparison function available when an intersection has more than a single report associated with it. Peak hour volumes are plotted and, if overlapping measurements are present within the two reports, percent differences are calculated and shown for the morning and evening peak hour periods.

Fig. 7 shows the results of the digitization and geoencoding. Selection of a particular corridor brings the user to a page where volume studies associated with the corridor are displayed. Fig. 8 shows an example of a full report. If available, reports for different directions of travel or different years can be displayed, as well as the PH and PHF.

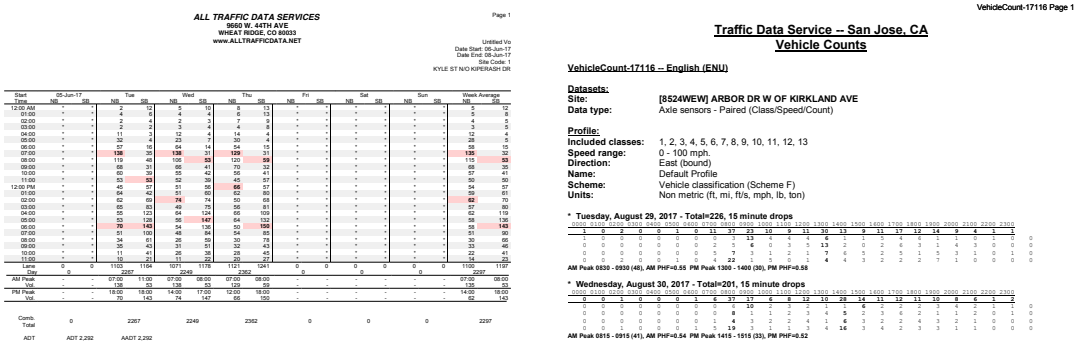
A commonly used statistic derived from corridor volumes is the Average Daily Traffic (ADT). The ADT of a corridor measures the daily traffic expected on the average day. It is an important statistic for determining if a corridor is behaving as expected. A view was provided where corridors could be filtered according to their calculated ADT. This view is shown in Fig. 9. Corridors with high ADT are shown in red, while those with lower ADT are shown in green. The user can also choose to select a range of ADTs to filter the corridors by, where only corridors that fall within that range will be shown.

2.3 San José Road Centerlines

2.3.1 Applications of Road Centerline Data

The City of San José road network is composed of approximately 11 000 unique roads, each with an associated road type (such as Residential or Freeway), speed limit, name and directionality (uni- vs. bi-directional). The entire network is shown in Fig. 10, where each road is coloured according to its speed limit in miles per hour. The road centerlines were used in two parts of the pilot project: attaching road segments to volume reports and augmenting incident data delivered to UrbanLogiq by the City (described in §2.6) with contextual information.

As mentioned in §2.2, each volume report was attached to a road segment, the results of



(a) **(b)**

Figure 6: Two examples of pneumatic tube volume reports commissioned by the City of San José. (a) shows a volume report created by All Traffic Data Services, while (b) shows a similar report produced by Traffic Data Service. As with TMC reports, there is variation in the formatting and contents. For example, All Traffic Data Services contains both directions in one file, while Traffic Data Service splits the report into multiple files, one file per direction of travel.

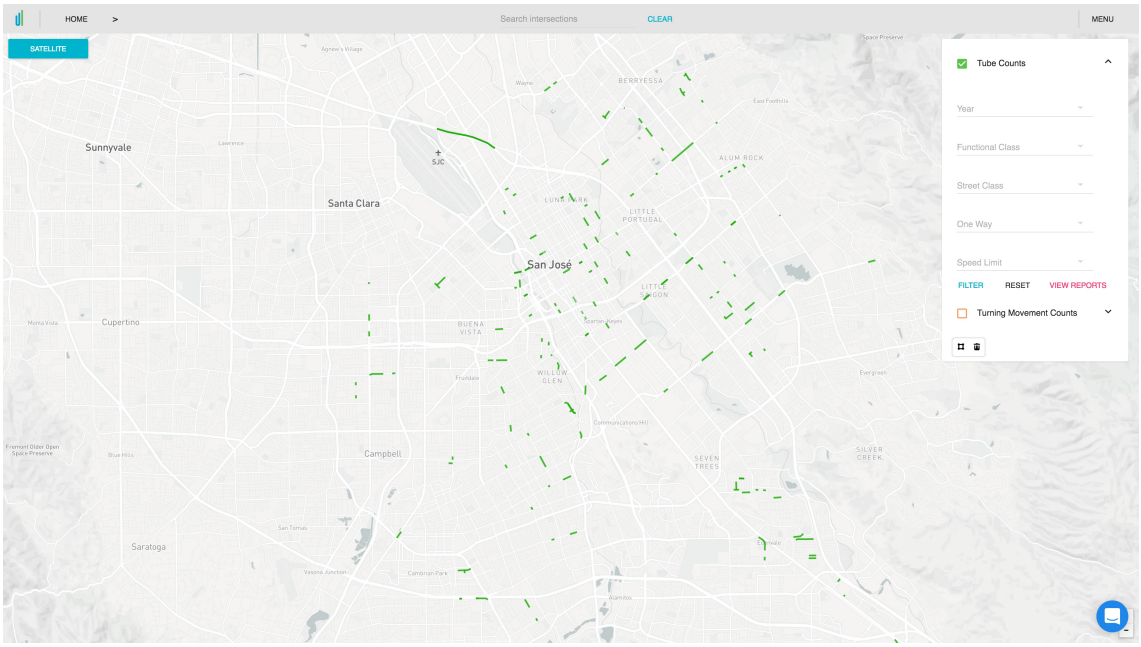
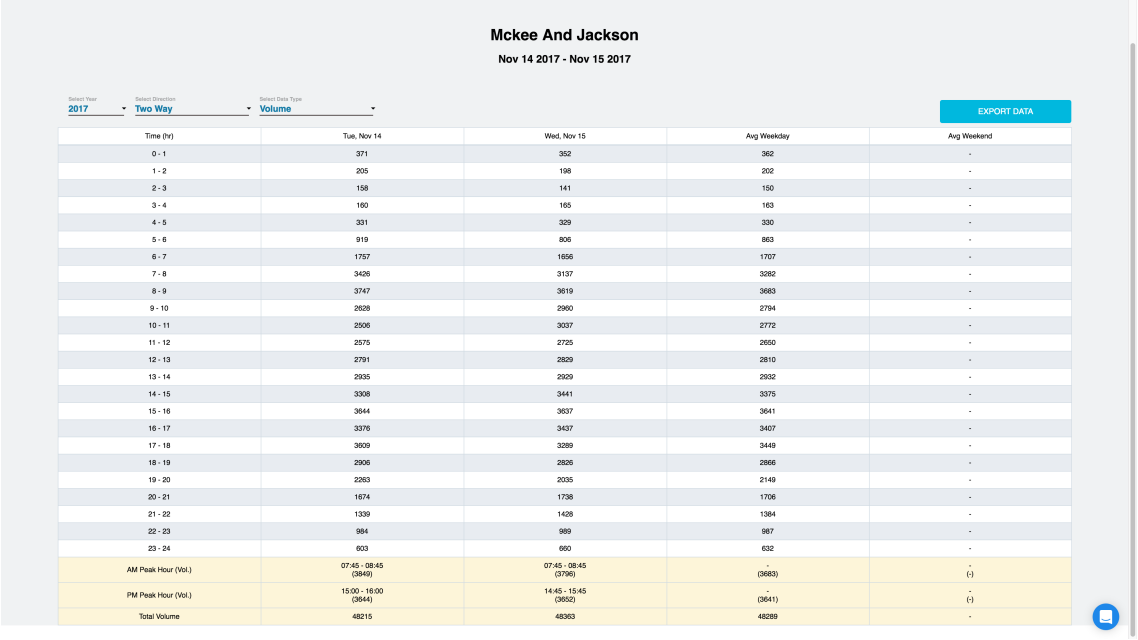


Figure 7: The home screen of the cloud platform delivered to the City of San José. Intersections with TMC reports have been deselected, while corridors with an associated volume study are shown in green. The user can select corridors by their functional class, street class, speed limit and directionality, in addition to the year that the report was created. Selection of a particular corridor brings the user to a page where the digitized reports for that page can be inspected and downloaded, as shown in Fig. 8.

which are shown in Fig. 7. To accomplish this, the text location of the report had to be matched to the road segment. In general, the text location of a report was formatted as "**Road**, east/west/north/south of **Cross Road**", where **Road** and **Cross Road** are the names of the road on which vehicles were counted and an intersecting road, respectively (see Fig. 6b for an example). Differences in naming conventions between the two datasets, such as "Tully Road" vs. "Tully Rd", or "HW 101" vs. "Old Bayshore Highway" presented significant challenges to ensuring that each report was matched with the proper segment. Additional issues, such as ambiguous text locations like "Tully Rd and King Rd", which fails to describe if the counts were to the west or east of King Rd, were also observed.



Time (hr)	Tue, Nov 14	Wed, Nov 15	Avg Weekday	Avg Weekend
0 - 1	371	352	362	-
1 - 2	205	198	202	-
2 - 3	158	141	150	-
3 - 4	160	165	163	-
4 - 5	331	329	330	-
5 - 6	919	806	863	-
6 - 7	1757	1656	1707	-
7 - 8	3426	3137	3282	-
8 - 9	3747	3619	3683	-
9 - 10	2628	2960	2794	-
10 - 11	2506	3037	2772	-
11 - 12	2575	2725	2650	-
12 - 13	2791	2829	2810	-
13 - 14	2935	2929	2932	-
14 - 15	3308	3441	3375	-
15 - 16	3644	3637	3641	-
16 - 17	3376	3437	3407	-
17 - 18	3609	3289	3449	-
18 - 19	2906	2826	2866	-
19 - 20	2263	2035	2149	-
20 - 21	1674	1738	1706	-
21 - 22	1339	1428	1384	-
22 - 23	984	989	987	-
23 - 24	603	660	632	-
AM Peak Hour (Vol.)	07:45 - 08:45 (3849)	07:45 - 08:45 (3796)	- (3683)	- (-)
PM Peak Hour (Vol.)	15:00 - 16:00 (3644)	14:45 - 15:45 (3652)	- (3641)	- (-)
Total Volume	48215	48363	48289	-

Figure 8: Once a corridor has been chosen, all the reports associated with it can be queried by year, direction of travel and report type (Volume, Speed or Class). The AM and PM Peak Hours are calculated and displayed. Once a report has been chosen for further analysis, it can be exported using the "Export Data" function.

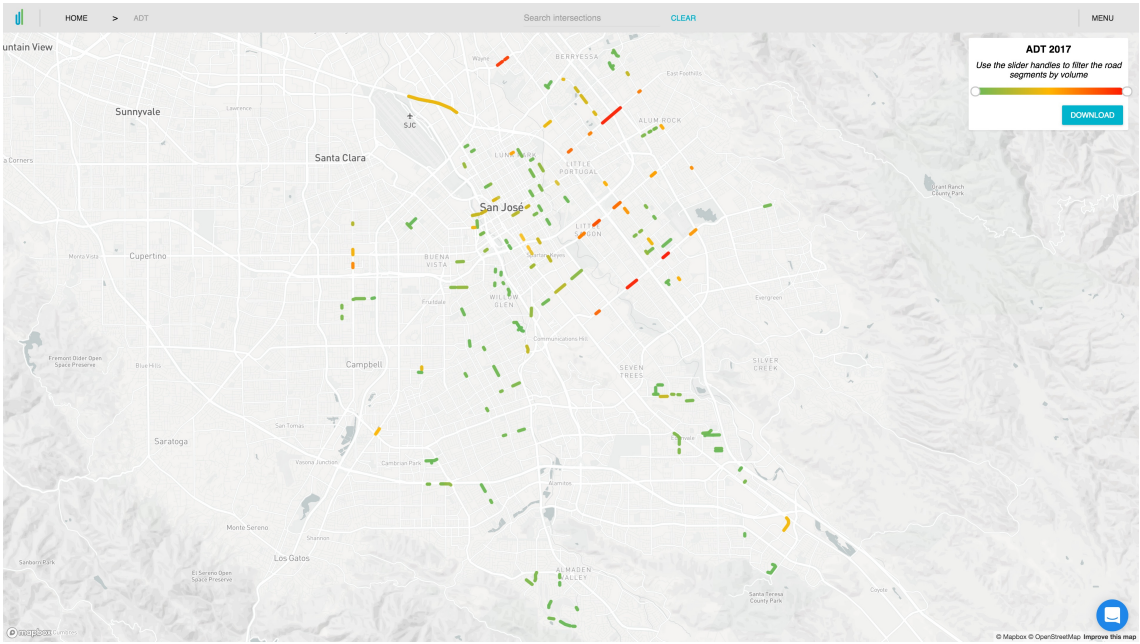


Figure 9: In this view, the user can filter the corridors by the computed ADT. Corridors with high ADT are shown in red, while those with lower ADT are green. The user can also select a range of ADT values to filter by. If a corridor falls outside of that range, it will not be displayed.

2.4 San José Intersections

2.4.1 Intersection Data Summary

San José's 11 000 roads meet at about 27 000 unique locations, shown in Fig. 11. Each intersection was associated with one of the following intersection types: "Intersection", "Ramp", "End", "Muni", "Non-Intersection". Other than the intersection type, there were no other useful

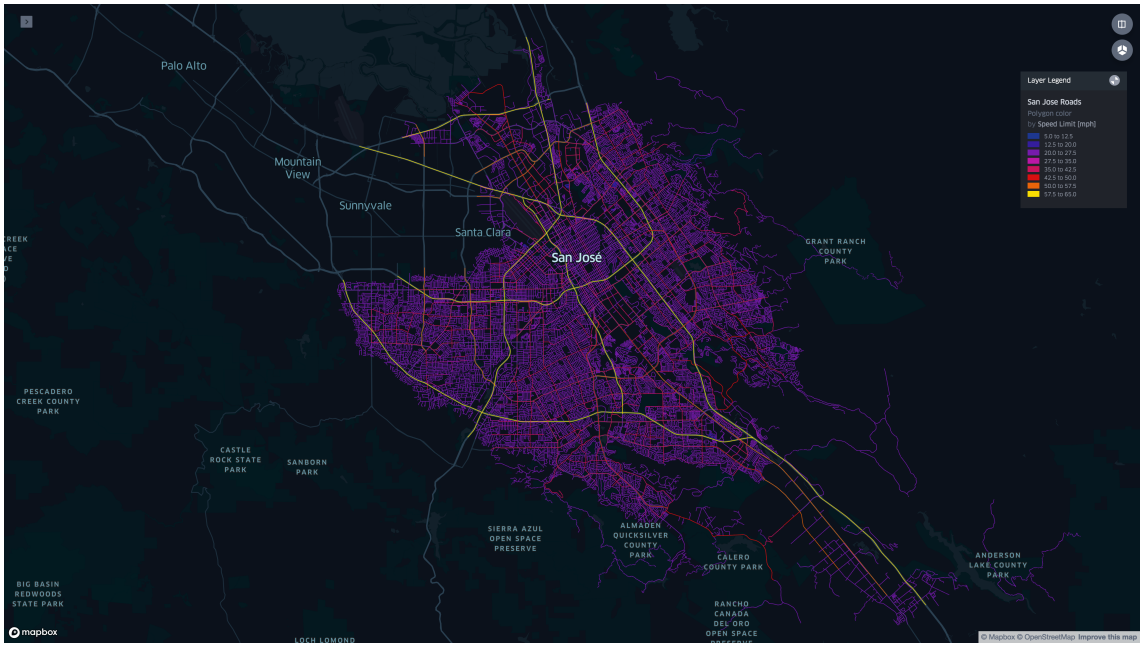


Figure 10: San José's road network (≈ 11 000 individual roads), where each road has been coloured according to its speed limit. Roads on the blue end of the spectrum have lower speed limits, those with higher speed limits are closer to yellow. Expressways and freeways are shown in yellow. A dark background is used to improve contrast. Image produced using *kepler.gl*.

properties associated with each intersection.

2.4.2 Use of Intersection Data

Knowing the locations of every intersection was pivotal in being able to effectively use and explore the incident data collected by the City, a significant part of the pilot project. The incident data provided by the City contained only text descriptions of the location of each incident (i.e. King and Tully). Using the intersection locations and names, it was possible to geoencode each incident to the intersection closest to where it occurred. Geoencoding the incidents was a vital step in not only building the incident dashboard, described in §3, but also generating contextual information for each intersection in San José.

2.5 Signalized Intersection Data

2.5.1 Signalized Intersection Data

1276 (roughly 5%), shown in Fig. 12, of San José's intersections are signalized. This dataset contains the locations of those intersections.

2.5.2 Use of Signalized Intersection Data

The locations of all the signalized intersections in San José were used to determine which intersections in San José were signalized, a property that was not included in the original intersection data.

2.6 Incident Data

Included in the pilot project was the creation of an incidents dashboard, where incident data that was collected by the City could be queried and explored. This section summarizes the incident

Figure 11: San José's intersections (totalling to $\approx 27\,000$), where each intersection is shown as a blue dot. These intersections were used to geoencode the incident data used in the pilot project. Image produced using *kepler.gl*.

Figure 12: The locations of the 1276 signalized intersections in San José. Image produced using *kepler.gl*.

data delivered to UrbanLogiq by the City, as well as some challenges associated with the creation of a dashboard.

2.6.1 Incident Data Summary

The incident data used by UrbanLogiq during this pilot spanned ten (10) years, containing events that occurred as early as 2008 and as late as 2018. Totalling to 74 000 incidents, each incident had 34 features associated with it, the more interesting of which are summarized in Table 1.

Feature	Description
AccidentDateTime	Date and time the incident occurred.
Intersection	Name of the intersection associated with the incident.
CollisionType	The type of collision (Rear End, Broadside, Vehicle/Pedestrian etc.).
Lighting	Lighting present during collision (dark, daylight, etc.).
MovementPrecedingCollision	Movement of vehicle leading up to the incident.
PedestrianAction	The movement of the pedestrian preceding the incident (crossing road, standing etc.).
PartyType	Type of the main involved party (Car, truck etc.).
RoadwayCondition	Condition of road during incident (Construction, reduced width etc.).
ViolationCode	Infraction committed by primary party (Speeding, reckless driving etc.).
Weather	Weather conditions at time of incident (Rain, sunny etc.).
DriverAge	Age of driver
Sobriety	Whether or not the driver had been drinking.
FatalInjuries	Number of deaths associated with incident.
MajorInjuries	Number of major injuries associated with incident.

Table 1: A sample of the features available for each incident. The name, a brief description and in some cases, examples of possible values are provided for each feature. In total, each incident had 34 features associated with it.

2.6.2 Use of Incident Data

For the pilot project, an incident dashboard, shown in Fig. 13, was created to allow the user to explore where and when incidents occurred, as well as some contributing factors to each incident. Three forms of preliminary analysis are given in the dashboard: the top ten intersections as measured by the number of incidents associated with that intersection, analysis of the the factors that caused the incidents, and a time series analysis. Using the dashboard, the user can select subsets of the incident data in three ways. First, the user can select regions of San José using a geofencing tool. Secondly, the user can select a time period for analysis using a time slider at the bottom of the dashboard. Fig. 14 shows the application of these two methods. Finally, the user can select an intersection or incident factor for further analysis. Selection of the intersection with the most incidents will perform the analysis on only the incidents that occurred at that intersection. Selection of a particular factor will in turn select only incidents that were caused by that factor.

In order to provide locations for each incident, the use of the intersection data detailed in §2.4 was required. The original incident data was only located through a text description of the closest intersection. In order to display each incident on a map, these text descriptions had to be matched

Figure 13: Home interface of the incidents dashboard developed using the incidents data delivered to UrbanLogiq. Here, the user can see each intersection coloured by the number of incidents that occurred there during the chosen time period. Intersections with more incidents are coloured red, while those with fewer incidents are coloured white. The top middle and top right panels rank intersections and incident factors by the number of associated incidets, respectively. The time series plot towards the middle shows how the number of incidents has changed over time, while the curve beneath it allows the user to select a particular time period for analysis and display.

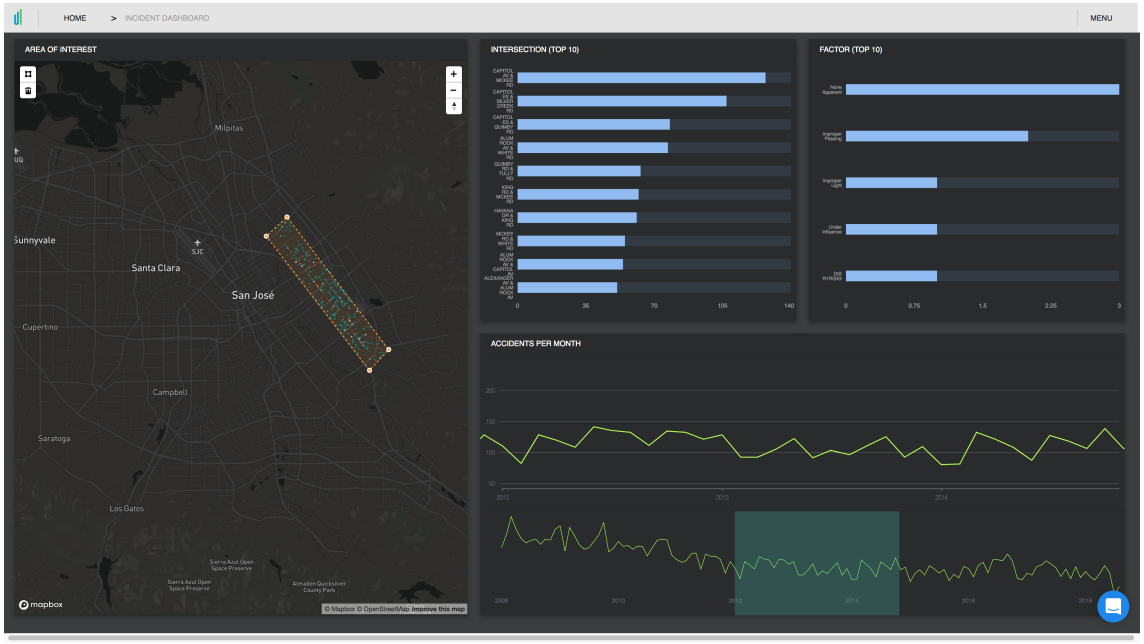


Figure 14: Example of filtering incidents in both time and space. The geofencing tool allows the user to select a spatial region of study, while the time selection tool allows the user to focus on a time period. Only the incidents that fall within the chosen area and time period are displayed. The summarizing statistics are then recomputed for the chosen incidents. In addition to time and space, the user can filter incidents by their cause, by selecting one of the factors in the top right panel, and by the intersection at which they occurred, by selecting a particular intersection for study. Every time a subset of the data is selected, the summarizing statistics are updated accordingly.

to the intersection names given in the intersection data.

Figure 15: The building footprints of each of San José's 32 000 buildings. Each building is coloured by its height, showing the concentration of higher buildings towards the center of San José. 75% of buildings have a height less than 18 ft. Image produced using *kepler.gl*.

2.7 Building Footprint Data

2.7.1 Building Footprint Data Summary

The footprints of roughly 32 000 buildings that are part of the City are shown in Fig. 15. Each building footprint is coloured by its height in feet, showing the concentration of higher buildings towards the downtown area of San José. In addition to the building height, the area of each building footprint is also provided in the dataset.

2.7.2 Use of Building Footprint Data

For the purposes of the pilot, building footprint data was used to describe the number and average height of the buildings located around an intersection, the details of which are described in §3.

2.8 Sidewalk Data

2.8.1 Sidewalk Data Summary

The City of San José contains nearly 87 million square feet of paved pedestrian walkways (Fig. 16 spread across the City.

2.8.2 Use of Sidewalk Data

The locations and areas of the sidewalks present in San Jose were used to augment the incident data with the area of sidewalk present within a radius of each intersection, as described in §3.

3 Geospatial Feature Engineering and Incident Analysis

In this section, the incident data introduced in §2.6 is analyzed and prepared for its use in developing a model designed to predict whether or not an intersection will see an incident resulting in a fatal

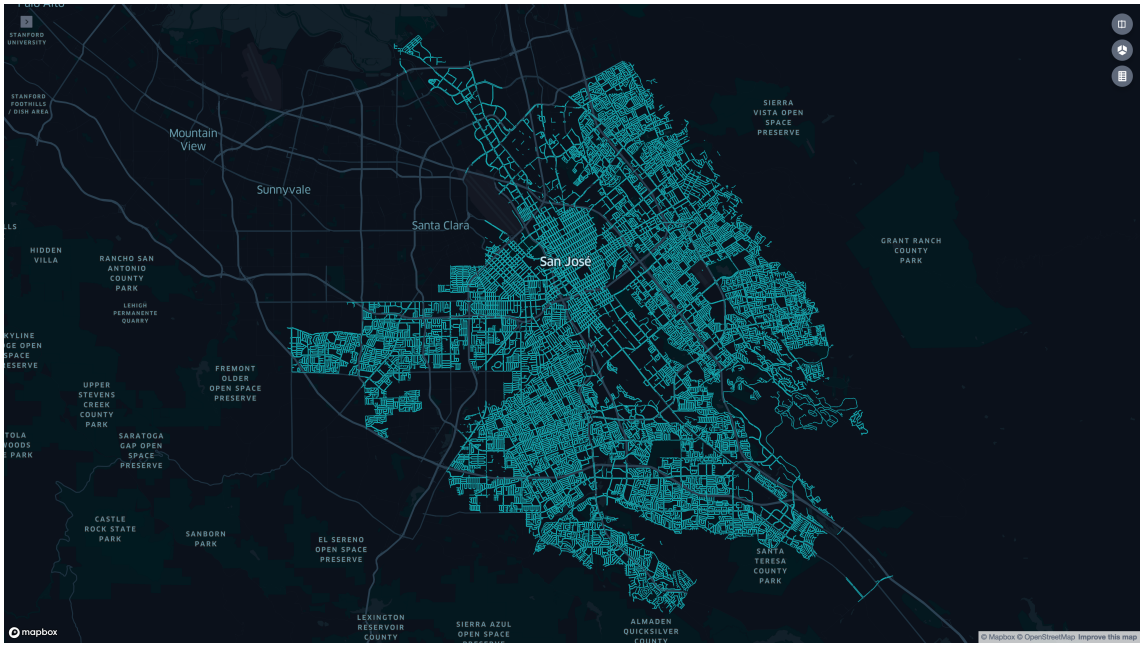


Figure 16: The sidewalks of the City of San Jose, totalling to ≈ 87 million square feet of paved pedestrian walkways. Image produced using *kepler.gl*.

or major injury over a particular time period. §3.1 shows the methodology used to select incidents appropriate for training an intersection-based model, i.e., correcting for long term trends seen in the number of incidents occurring each year and selecting incidents that only occurred near or at an intersection. §3.2 presents the methods used to combine various geospatial datasets introduced in the previous section to augment the chosen incident data. A brief analysis is also provided for each new geospatial feature.

3.1 Preliminary Analysis of Incident Data

The incident data delivered to UrbanLogiq was composed of roughly 74 000 separate incidents, occurring between January 1st, 2008 and September 31st, 2018. Each incident had an associated number of fatal, major, moderate and minor injuries. Incidents with at least one fatal or major injury (approximately 2.5% of total incidents) are compared to incidents with no fatal or major injuries, and incidents that resulted in no injuries. In this section, an subset of incidents are chosen as training data for the model developed in §4.

3.1.1 Time Analysis

Fig. 17 shows the number of incidents per three months in two cases: incidents that had at least one fatal or major injury are shown in red, while all other incidents are shown in blue. When comparing the general trend between the two curves (excluding 2018, since the data was delivered before the end of the year). The total number of incidents with no fatal or major incidents exhibited a significant decrease (approximately 40%) between 2008 and 2012[1], after which the number of incidents remained stable at 1500 ± 100 incidents every three months. Incidents that had at least one fatal or major injury were much more variable with no obvious consistent trends. On average there were 45 ± 10 incidents with at least one fatal or major injury every three months.

[1] After consultation with the City, the reason for this decrease was identified as a reduced response rate to incidents that included only property damage.

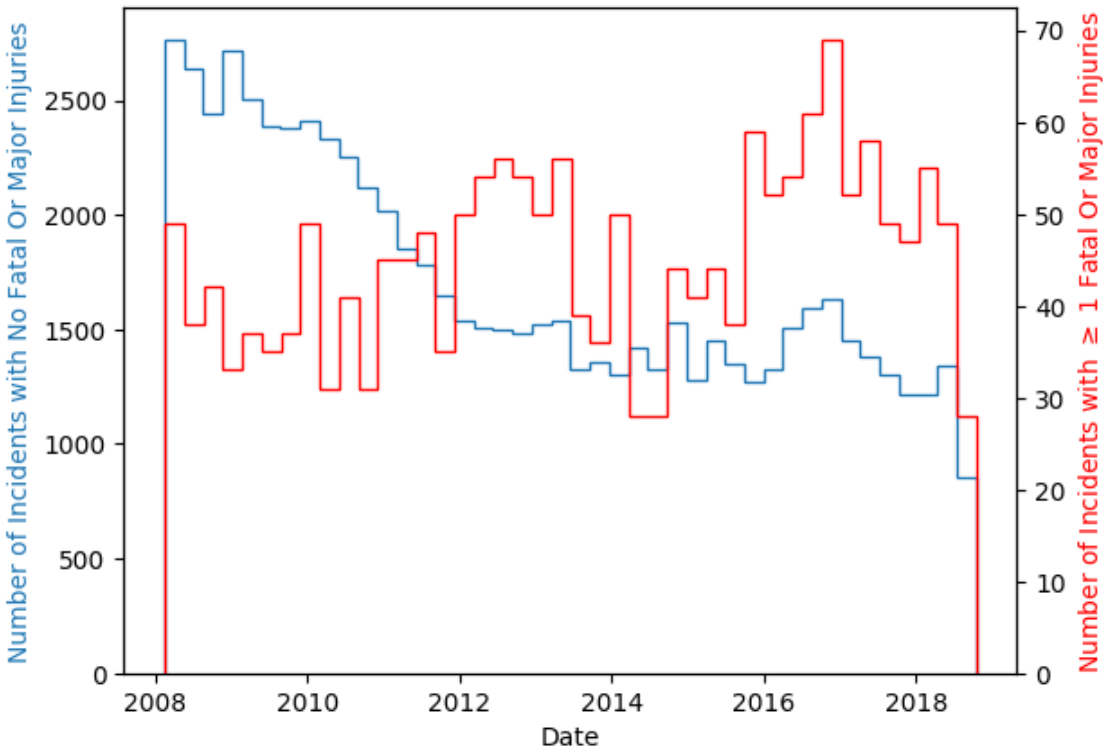


Figure 17: Tri-monthly number of incidents spanning the entire range of the dataset (January 1st, 2008 - September 31st, 2018) for all incidents (blue) and incidents that had at least one major or fatal injury (red). Between 2008 and 2012, the total number of incidents per three months decreased from approximately 2500 to 1500. From 2012 to 2017, however, the number of incidents remained constant at an average of roughly 1500 ± 100. Incidents with at least one fatal or major injury were much more variable over the entire date range, with an average of 45 ± 10 incidents.

3.1.2 Proximity to Intersection Analysis

Each incident had attached to it a measure of its proximity to a particular intersection. Fig. 18 shows the fractional distribution of incidents with respect to the proximity to a nearby intersection in two cases: incidents that had at least one fatal or major injury (red) and incidents that had no fatal or major injuries (blue). In both cases, incidents of the types "Intersection" and "Related" made up the majority ($> 50\%$) of the incidents. In total, 51% ($\approx$ 36 000 incidents with no fatal or major injuries, and $\approx$ 1000 for incidents with at least one fatal or major injury) of all recorded incidents fell into the "Intersection" and "Related" categories. Figs. 19a and 19b show the distribution of the distances from the intersection for incidents with their proximity to intersection categorized as "Related" and "Intersection", respectively. In both cases, the majority ($> 95\%$) of the incidents occurred within 10 ft of an intersection.

3.2 Geospatial Feature Engineering and Analysis

Using the geospatial datasets described in §2, ten (10) new features were created for each intersection of type "Intersection". Table 2 gives a description of each of these features. In addition to the engineered geospatial features, each intersection had attached to it the total number of incidents of proximity class "Intersection" and "Related" occurring between 2012 and 2017. Incidents of other proximity classes were excluded because they were not associated with an intersection. Incidents occurring before 2012 were excluded due to the observed decreasing trend in number of incidents. Incidents occurring during 2018 were excluded as the year had not finished. The goal of

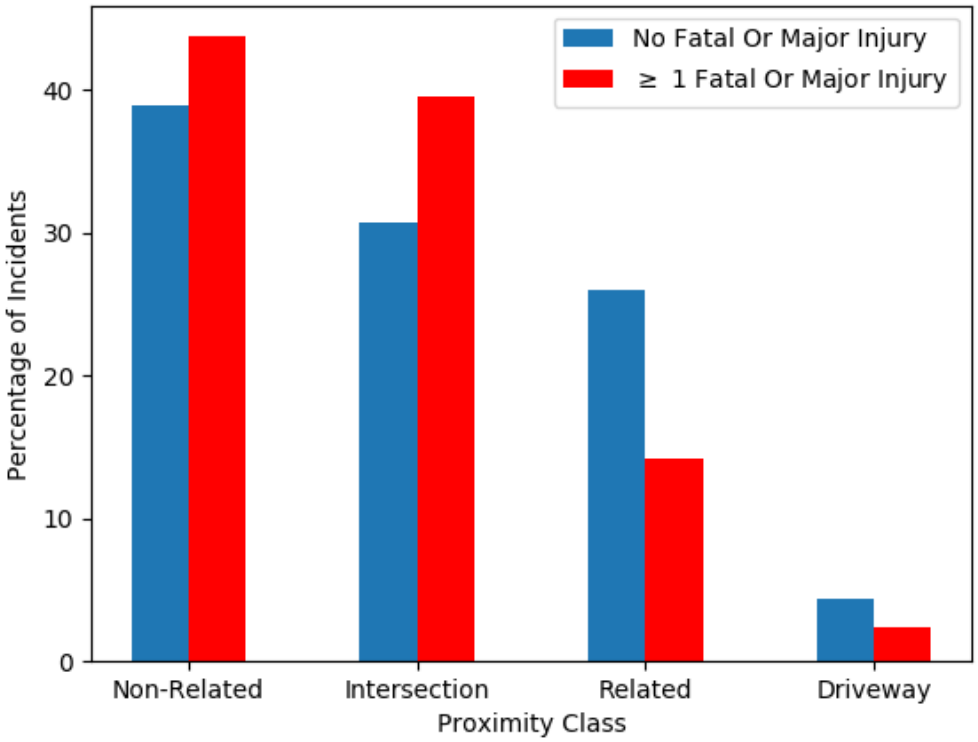
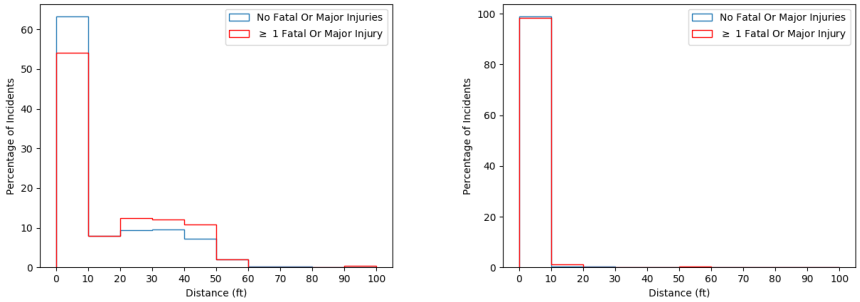


Figure 18: Percentage of incidents that fall into each proximity class. Incidents with no fatal or major injuries are plotted in blue, while those with at least one fatal or major injury are plotted in red. In both cases, incidents of the types 'Intersection' and 'Related' made up the majority ($> 50\%$) of the incidents.



(a) **(b)**

Figure 19: a) The distribution of the distances from an intersection for incidents that were categorized as "Related". Incidents with no fatal or major injuries are shown in blue, while incidents with at least one fatal or major injury are shown in red. In both cases, over 50% of incidents occurred within 10 ft of an intersection. The remaining incidents are between 20 and 60 ft away from an intersection. **b)** The distribution of the distances from an intersection for incidents that were categorized as "Intersection". Incidents with no fatal or major injuries are shown in blue, while incidents with at least one fatal or major injury are shown in red. For both types, more than 95% of incidents occurred within 10 ft of an intersection.

the model was to determine if an intersection was likely to see an incident that resulted in a fatal or major injury, over a period of time, given the properties described in Table 2. The inclusion of incomplete years, or years where the number of incidents was not constant, may have affected the results. When analyzing the results of the geospatial features, intersections are grouped into three

Feature Name	Feature Description
Sidewalk Area	The total area (ft^2) of sidewalks contained within a 200 ft radius of the intersection.
Number of Buildings	The number of buildings within a 200 ft radius of the intersection.
Average Height of Buildings	The average height (ft) of the buildings counted for Number of Buildings.
Standard Deviation of Height of Buildings	The standard deviation from the average of the heights of the buildings counted in Number of Buildings.
Is Signalized	Whether or not an intersection was signalized.
Difference of Speed Limits	The difference in the speed limits of the two intersecting roads that created the intersection.
Total of Speed Limits	The sum of the speed limits of the two intersecting roads that created the intersection.
Angle of Intersection	The angle of intersection (radians) of the two roads that created the intersection.
North Angle	The angle of the road closest to pointing north (radians).
Total Curvature	The total curvature (radians) of the roads within 100 ft of the intersection.

Table 2: Descriptions of the features created using the geospatial datasets described in §2.

categories: intersections that had no incidents resulting in injuries (green), intersections that had no incidents resulting in a fatal or major injury (blue) and intersections that had incidents that resulted in at least on fatal or major injury (red).

3.2.1 Sidewalk Area Analysis

Fig. 20a shows the distribution of sidewalk areas for three categories of intersections. Intersections with no injuries or no fatal or major injuries behave similarly, with 75% of intersections having a surrounding sidewalk area of $< 24\ 000\ ft^2$. Intersections that had at least one fatal or major injury, however, tend towards areas with a higher amount of sidewalks, 75% having sidewalk area $< 27\ 000$.

3.2.2 Number of Buildings Analysis

Fig. 20b shows the distribution of the number of buildings for three categories of intersections. Intersections with no injuries or no fatal or major injuries behave similarly, with 75% of intersections having ~ 30 buildings within 200 ft. Intersections that had at least one fatal or major injury,

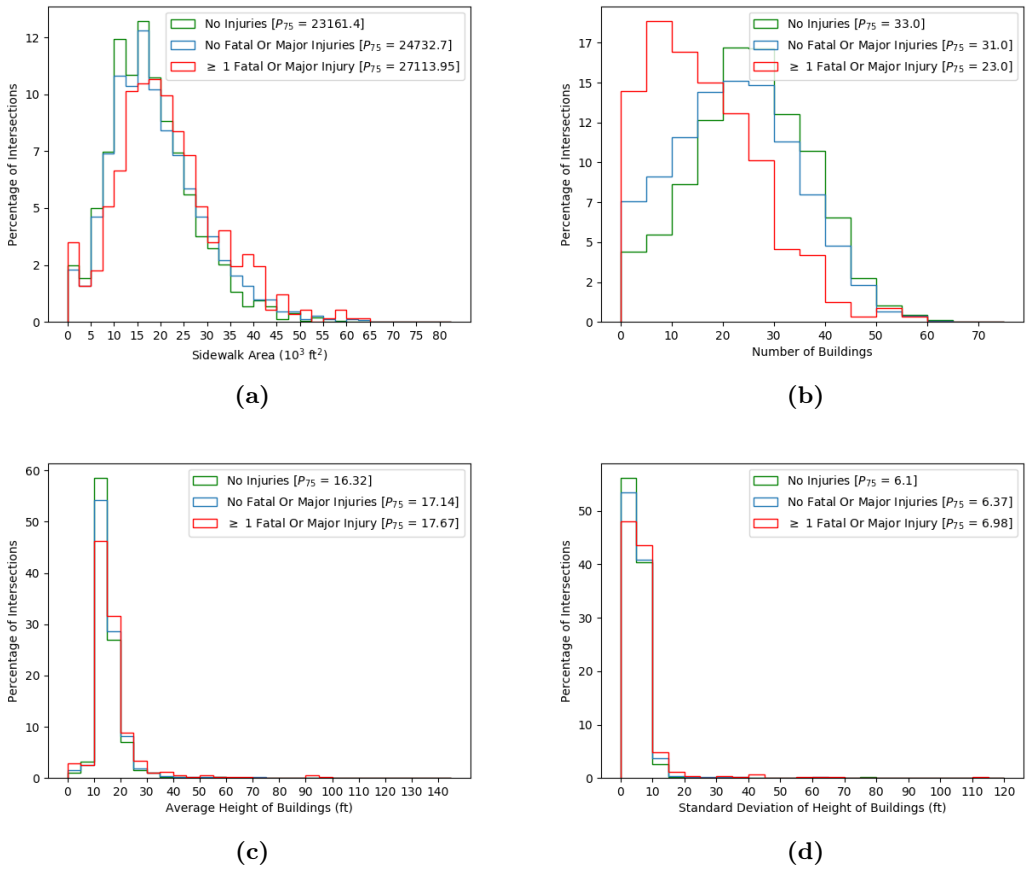


Figure 20: Distributions of geospatial features for three categories of intersection: intersections that had no incidents resulting in injuries (green), intersections that had no incidents resulting in a fatal or major injury (blue) and intersections that had incidents that resulted in at least on fatal or major injury (red). Also included are the 75$^{\text{th}}$ percentiles, where relevant. **a)** Distribution of sidewalk area. Intersections with at least one fatal or major injury tend towards higher (∼27 000 ft^2) amounts of surrounding sidewalk area when compared to the other categories (∼24 000 ft^2). **b)** Distribution of number of buildings within 200 ft. Intersections with at least one fatal or major injury tend towards a lower (∼23) amount of buildings when compared to the remaining categories (∼32). **c)** Distribution of average height of buildings within 200 ft. All three categories of intersections behave similarly. **d)** Distribution of the standard deviations in the height of the buildings within 200 ft an intersection. All three categories behave similarly.

however, are tend towards areas with fewer buildings, 75% having $<$ 23 buildings surrounding them.

3.2.3 Average Height of Buildings Analysis

Fig. 20c shows the distribution of the height of the buildings for three categories of intersections. All three categories behaved similarly.

3.2.4 Standard Deviation of Average Height of Buildings Analysis

Fig. 20d shows the distribution of the standard deviation in the height of the buildings surrounding an intersection for three categories of intersections. All three categories behave similarly.

3.2.5 Signalization Analysis

Fig. 21a shows the distribution of the standard deviation in the height of the buildings surrounding an intersection for three categories of intersections. Intersections where at least one fatal or major

injury occurred showed a roughly even split between signalization and no signalization, while intersections with no injuries and intersections with no fatal or major injuries show splits tending towards no signalization (80/20 and 90/10, respectively).

3.2.6 Speed Limits Analysis

Fig. 21b shows the distribution of the difference in the intersecting roads speed limits for each intersection, while Fig. 21c shows the distribution for the sum of speed limits of the intersecting roads. Intersections with no injuries tend towards lower differences in speed limits and lower speed limit sums ($P_{75} = 5$ MpH and 55 MpH, respectively). Intersections with no fatal or major injuries exhibit higher values for both features ($P_{75} = 10$ MpH and 60 MpH), while intersections with at least one fatal or major injury have a similar difference in speed limits ($P_{75} = 10$ MpH) and a higher sum of speed limits ($P_{75} = 65$ MpH).

3.2.7 Road Angles Analysis

Fig. 22a shows the distributions of the angles of intersecting roads and the angle of the road most closely oriented towards north, respectively. In the case of the angle of intersection between the roads, all three categories peaked at angles of 55 and 80 degrees. Roughly 30% of intersections that saw at least one fatal or major injury have roads intersecting at 55 degrees, compared to 50% of intersections that saw no fatal or major injuries and 70% of intersections that saw no injuries whatsoever. In contrast, roughly 45% of intersections that saw at least one fatal or major injury have roads intersecting at 80 degrees, compared to 30% of intersections that saw no fatal or major injuries and 15% of intersections that saw no injuries whatsoever.

Fig. 22b shows the distribution of the sine of the angle of the road most closely oriented towards North. For the peaks at 0.95, 0.86 and 0.64 (∼70, 60 and 40 degrees), the three classes of intersection behaved similarly. At 0.48 (∼30 degrees), intersections with at least one fatal or major injury were more strongly represented when compared to intersections with no fatal or major injuries, and intersections with no injuries whatsoever (6% vs. 2% and 3%, respectively).

3.2.8 Total Curvature Analysis

Fig. 22c show the distribution of the total curvature of the roads within 200 ft of the intersection. As the severity of the types of crashes increased (no injuries, no fatal or major injuries, at least one fatal or major injury), the total curvature for each class of intersections was seen to increase ($P_{75} = 4.57$, $P_{75} = 10.0$, $P_{75} = 14.87$, respectively).

4 Model Creation

In this section, the features described in §3.2 are used to train a model with the goal of predicting whether or not an intersection is expected to see an incident resulting in a fatal or major injury during the study period (2012 to 2018). For the purposes of this work, only intersections that have seen incidents are used as the training data. As such, an additional feature is added on top of the geospatial features already included: the total number of incidents that an intersection has seen. Fig. 23 shows the distribution of the intersections that have seen an incident during the study period, coloured by the number of incidents resulting in fatal or major injuries. §4.1 describes the outline of the model used to achieve this goal, while §4.2 presents reports on the performance of the resulting trained model.

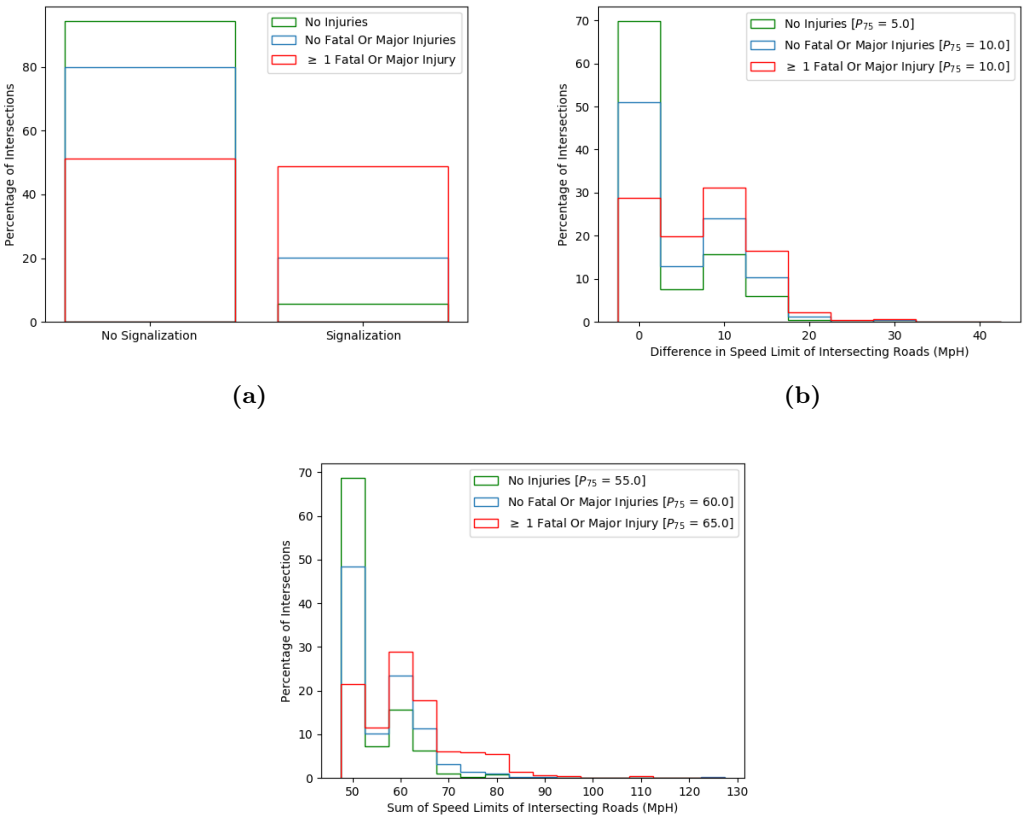


Figure 21: Distributions of geospatial features for three categories of intersection: intersections that had no incidents resulting in injuries (green), intersections that had no incidents resulting in a fatal or major injury (blue) and intersections that had incidents that resulted in at least on fatal or major injury (red). Also included are the 75^{th} percentiles, where relevant. **a)** Distribution of whether or not an intersection was signalized. Intersections with no injuries were strongly distributed towards no signalization (roughly 90/10 split), while intersections where no fatal or major injuries occurred showed a 80/20 split in favour of no signalization. Intersections at which at least one fatal or major injury showed almost an even split between signalized and non-signalized. **b)** Distribution of the difference in speeds of the intersecting roads at each intersection. Intersections with no injuries tended towards a speed difference of 5 MpH, while intersections injuries with showed a strong tendency towards a higher speed difference (10 MpH). 30% of intersections with at least one fatal or major injury had intersecting roads with the same speed limit, compared to 50% and 70% in the cases of no fatal or major injuries, and no injuries, respectively. **c)** Distribution of the sum of the speed limits of the intersecting roads for each intersection. Intersections with no injuries tend towards lower overall speeds (55 MpH) while intersections with no fatal or major injuries and those with at least one fatal or major injury showed a tendency towards higher overall speeds (60 and 65 MpH, respectively).

4.1 Model Schema

4.1.1 XGBoost-An Ensemble of Decision Trees

To solve the problem of predicting whether or not an intersection would see an incident resulting in a fatal or major injury, a class of machine learning models known as decision trees, which can be thought of a series of questions that can be answered with a True or False, (see Fig. 24) was implemented. More specifically, a particular implementation of this class of algorithms, known as XGBoost was used. XGBoost leverages what is known as ensemble learning to train a multitude of "weak predictors", individual decision trees that each use a subset of features to make a prediction. These predictions were then combined to make a final prediction.

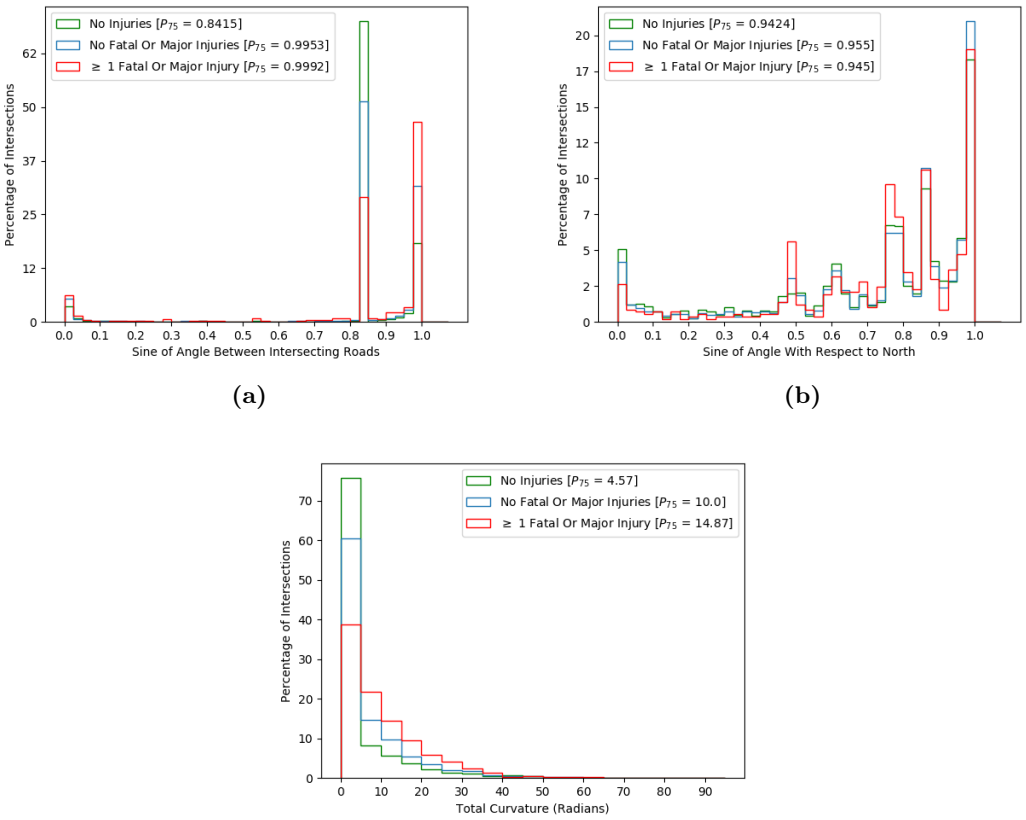


Figure 22: Distributions of geospatial features for three categories of intersection: intersections that had no incidents resulting in injuries (green), intersections that had no incidents resulting in a fatal or major injury (blue) and intersections that had incidents that resulted in at least on fatal or major injury (red). Also included are the 75$^{\text{th}}$ percentiles, where relevant. **a)** Distribution of the sine of the angle of intersection of the roads intersecting at each intersection. All three categories peak at 0.841 and 0.997 (corresponding to angles of approximately 55 and 80 degrees). ∼35% of intersections that saw at least one fatal or major injury have roads intersecting at 55 degrees, compared to ∼50% for intersections that saw no fatal or major injuries and ∼70% for intersections that saw no injuries. 45% of intersections that saw at least one fatal or major injury have roads intersecting at 80 degrees, compared to 35% for intersections that saw no fatal or major injuries and 20% for intersections that saw no injuries. **b)** Distribution of the sine of the angle of the road most closely oriented towards North. All three classes behave similarly at ∼ 0.947, 0.875, and 0.635 (approximately 70, 60 and 40 degrees, respectively). At 0.48 (∼30 degrees), intersections with a fatal or major injury are more strongly represented (∼ 6% of intersections) when compared to intersections with no injuries (∼ 2%) and intersections with no fatal or major injuries (∼ 3%). **c)** Distribution of the total road curvature within 200 ft of an intersection. Intersections with at least one fatal or major injury tend towards higher curvatures (P_{75} = 14.87), when compared to intersections with no fatal or major injuries and intersections with no injuries (P_{75}= 10.0 and 4.57 respectively).

4.1.2 Model Performance Measure

Given that the model is a binary classifier, model performance was measured using a Receiver Operating Characteristic curve, which compares the True Positive Rate (TPR) against the False Positive Rate (FPR) at different prediction thresholds. The Area Under the Curve (AUC), is then computed as the area under the ROC, and is taken as a measure of the models ability to distinguish between the two classes (True and False, or in this case, will an incident with a fatal or major

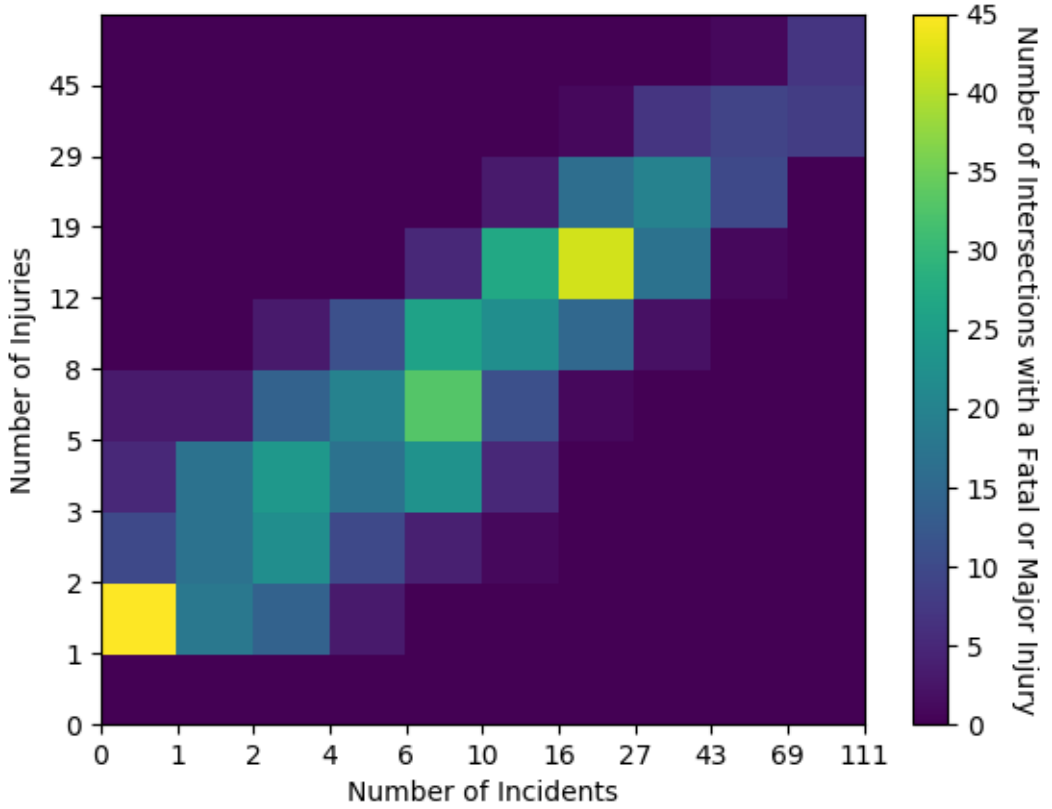


Figure 23: Number of intersections that saw a fatal or major injury as a function of the number of incidents that occurred at that intersection and the number of injuries resulting from those incidents. Both the number of incidents and the number of injuries are plotted on logarithmic scales. Two main areas of interest were observed. Intersections that had few incidents and few injuries (bottom left of the plot), as well as intersections that had between roughly 15 and 25 incidents and between 10 and 20 injuries (center of the plot).

injury be observed at a particular intersection). The TPR is defined as

$$\text{TPR} \equiv \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

where TP is the number of True Positives, in this case the number of intersections that were predicted to have an incident resulting in a fatal or major injury, which turned out to have such an incident occur, while FP is the number of False Positives, the number of intersections predicted to have an incident resulting in a fatal or major injury that, in fact, never saw such and incident. Likewise, the FPR is defined as

$$\text{FPR} \equiv \frac{\text{FP}}{\text{FP} + \text{TP}}. \tag{2}$$

The ideal classifier would have a TPR of 1 (all intersections with a fatal or major injury were correctly predicted as such) for a FPR of 0 (intersections that had no fatal or major injuries were never predicted to have any), as well as an AUC of 1. A random guess, where TPR = FPR has an AUC of 0.5. Any model with an AUC above 0.5 is considered to be better than a random guess.

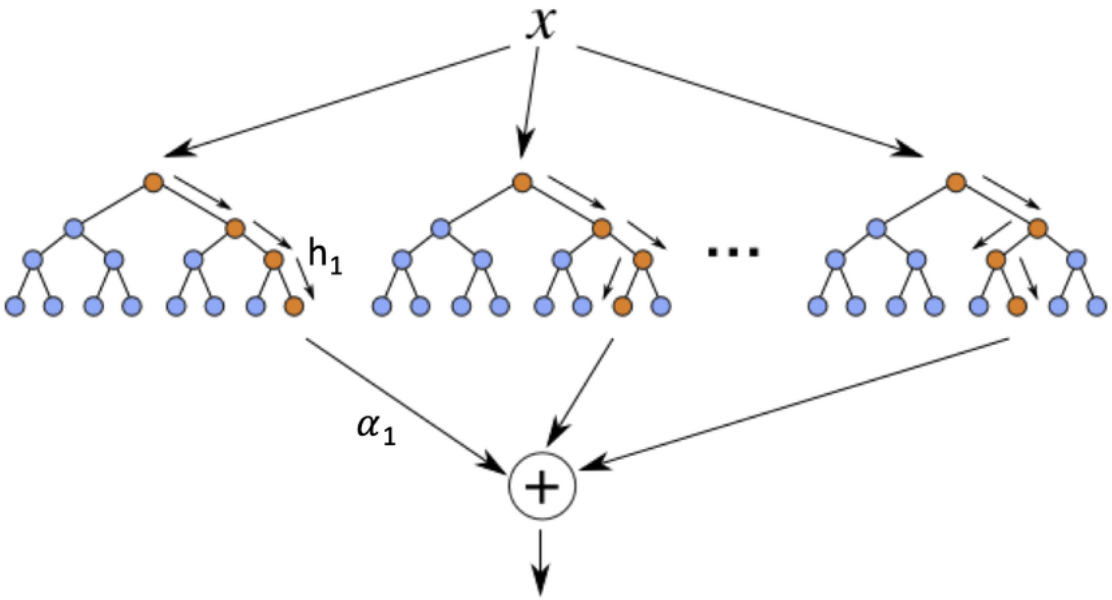


Figure 24: Schematic of an ensemble of decision trees. An example X (in this case, an intersection) is passed to multiple decision trees. Each tree makes a prediction h, which is represented by a number α. The resulting α are then summed together to generate the final prediction for X.[1]

Subset	Features
Road Features	• Difference in Speed Limits • Sum of Speed Limits • Angle of Intersection • North Angle • Total Curvature • Is Signalized
Contextual Features	• Sidewalk Area • Number of Buildings • Average Height of Buildings • Standard Deviation of Height of Buildings
Road and Contextual Features	Road Features + Contextual Features
Including Total Number of Incidents	Road and Contextual Features + Total Number of Incidents

Table 3: Feature subsets used to train the model for each of the cases described in Fig. 25.

4.2 Model Results

The model was trained four times, each time using a different subset of the total feature set. Fig. 25 shows the performance of the model on the test set for each set of features, while Table 3 shows the features included in each subset. When the Road Features subset was used, the AUC was 0.56, indicating that the model outperformed a random guess. However, much better performance was observed when only features belonging to the Contextual Features subset were used, yielding an AUC of 0.67. Combining the two subsets into the Road and Contextual Features subset produced another boost in performance with an AUC of 0.72. Finally, adding in the number of total incidents

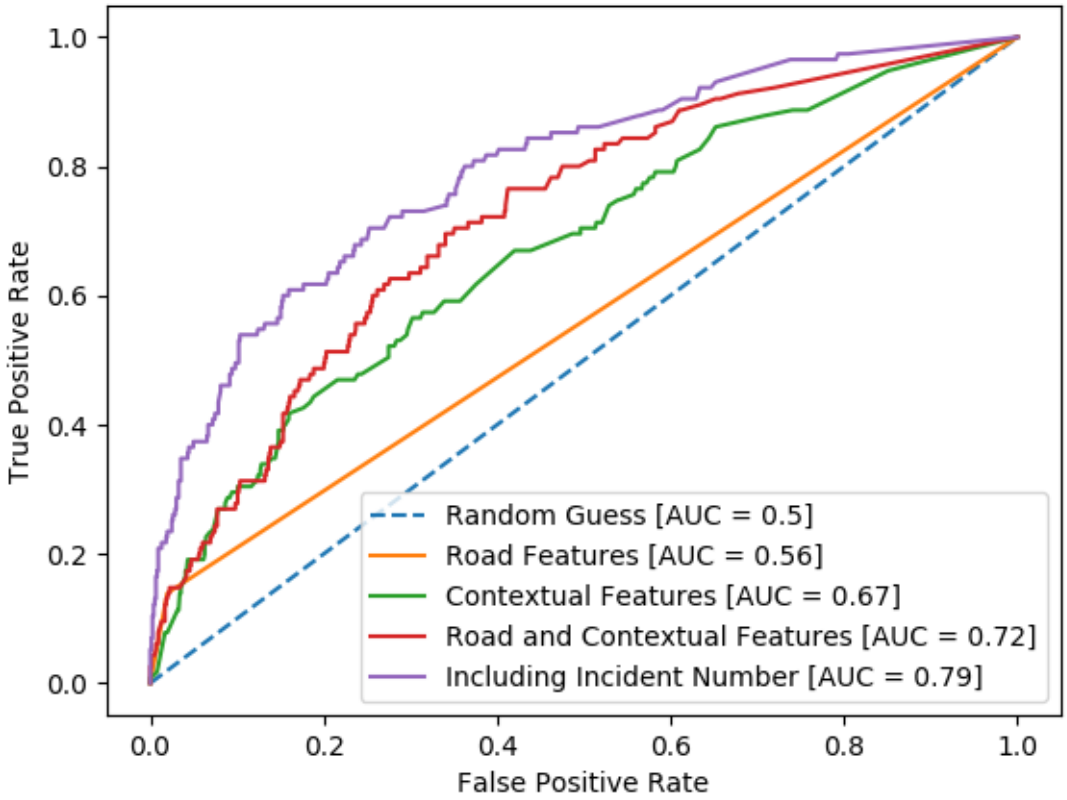


Figure 25: The ROC curve for the model, trained on four subsets of features, described in Table 3. Performance of model is measured by the Area Under the Curve (AUC). Road Features perform better than a random classifier (AUC = 0.56), while Contextual Features boost performance significantly (AUC = 0.67). Both Road and Contextual features yield and AUB of 0.72, and including the total number of incidents brings the AUC up to 0.79.

as a final feature brought the AUC up to 0.79.

For the final model, selecting a threshold of 0.377 yielded a TPR = 0.70 for FPR = 0.25, indicating that 70 percent of intersections that had a fatal or major incident were correctly predicted as such, while 25 percent of intersections that had no fatal or major incidents were incorrectly predicted as having one.

5 Conclusion and Future Work

Over the course of the four-month engagement period for the pilot project between UrbanLogiq and the City of San José, Urbanlogiq combined the different datasets delivered by the City to produce three main deliverables: First, a platform designed to simplify and streamline the City's use of its traffic data by allowing it to be queried and analyzed geospatially and over time.

Second, a dashboard designed to house and filter incident data, where each incident was attached to an intersection in San José, ranging ten (10) years. The dashboard allows the user to query incidents in time and group intersections together spatially. Finally, the incident data was also used, in addition to some other geospatial datasets, to train a binary classifier (based on an

ensemble of boosted decision trees) with the goal of predicting whether or not an intersection would see a fatal or major incident over a five-year time range. Some initial analysis of the number of fatal or major incidents as a function of the geospatial features found that fatal or major incidents tend towards areas with higher sidewalk area ($P_{75} = 27113.95$), lower number of buildings ($P_{75} = 23$), higher sum of speed limits ($P_{75} = 65$ MpH), higher speed limit differences ($P_{75} = 10$ MpH) and higher total road curvature ($P_{75} = 14.87$ radians). The AUC for the ROC curve of the final model was 0.79. A model using features derived from the roads around an intersection produced an AUC of 0.55, while one using features derived from the surrounding infrastructure (sidewalks and building) produced an AUC of 0.67. A model using both the road and surrounding infrastructure features gave an AUC of 0.72. The best model had a TPR of 0.70 for a FPR of 0.25.

5.1 Future Work

5.1.1 Traffic and Incident Data Platform

The initial purpose of the platform was to provide seamless and universal access to historical traffic data that had already been collected by the City. Future work to enhance the utility and functionality of the platform could include:

1. Automated validation of previously collected traffic data as well as immediate validation of newly collected data from both City and third-party sources.
2. Introduction of multi-disciplinary data, such as demographics, events, businesses, infrastructure, weather, econometrics and land use to provide context for observed travel patterns and broader community mobility behaviours.
3. Introduction of third-party data such as connected vehicle telematic data with the goal of providing wider and more up-to-date measurements of traffic behaviours.

5.1.2 Improvements to Machine Learning Model

The model developed in this work was intended to broadly predict whether or not an intersection would see a fatal or major incident within a particular time period. More targeted models focusing on specific types of incidents (such as Vehicle/Pedestrian and Vehicle/Bike) could be developed with the intent of identifying contributors to these incidents. Additionally, a main feature was the number of incidents at each intersection. This feature was intended as a stop-gap for measuring overall intersection usage. A more complete model would have a proper measure of intersection usage such as daily volume as a feature. This could be achieved in two ways: (1) by leveraging existing volume data to predict volume at intersections where it had not been measured; and (2) the introduction of third-party data such as connected vehicle telematics to measure intersection vehicle volumes. Both of these methods could be extended to pedestrian and bicycle volumes.

Ultimately, the incident data collected by the City is of extreme value in terms of informing, data-driven policy making to reduce road-related fatalities. This work has only scratched the surface of what this data can do to save lives.

References

[1] Shoaran, M., Haghi, B., Taghavi, M., Farivar, M. and Emami-Neyestanak, A. (2018). Energy-Efficient Classification for Resource-Constrained Biomedical Applications. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 8(4), pp.693-707.