Using Predictive Analytics to Uncover Traffic Safety Insights

NYU STERN

MS in Business Analytics

Photo credit: Tamara Kempf, May 2018

Noah Branham
David Huang
Tamara Kempf
Michael Marra
Aaron Smith

Advisor: Panos Ipeirotis

May 2019

Executive Summary

Every year, traffic-related injuries and deaths cause enormous human and economic harm in the United States. To address the crisis, city governments have adopted Vision Zero, a set of data-driven safety strategies rooted in engineering, enforcement, and education with the goal to reduce traffic fatalities to zero. New York City has led the way so far, committing $1.6 billion to the initiative through 2021.[1] However, cities with limited budgets, inadequate data, and constrained resources need help prioritizing focus areas and gathering insights that can lead to effective interventions.

Our Capstone project seeks to help cities who are new to Vision Zero 1.) prioritize which local areas to focus their efforts on and 2.) glean insights from other cities as to how socio-economic and road design data might impact traffic injuries / deaths on a local level. We collected public collision data from New York City, Washington D.C., and Los Angeles from 2013 to 2017 and aggregate injuries and deaths at the census tract level (a unit of analysis about the size of 2–3 city blocks). We then pair each census tract's collision data with its corresponding socio-economic and road inventory data. To our knowledge, this dataset is the first of its kind to join public collision and Census data *across* cities.

We then use predictive analytics to assist cities without robust collision data understand where they can have the greatest impact. Rather than build a model that predicts total count of casualties (injuries + deaths) in each census tract, we are interested in how well we can predict the *ranking* of census tracts in a city from highest to lowest casualties over a five year period. With this ranking, city officials can make more informed decisions about where to allocate resources, and learn about the informativeness of the variables that produce it.

We find that, using solely a city's Census data, we can predict the ranking of census tract casualties, from highest to lowest, better than random chance. Performance is further improved with road inventory data. To a certain extent, we can also transfer a model built on New York City Census data to another city, and achieve ranking results better than random chance. This suggests that there is something predictive about socio-economic and road inventory data, and we use linear regression techniques to further explore their relationship to casualties.

Special Thanks

We would like to extend gratitude to the following industry experts, who shared their domain knowledge and provided guidance throughout our Capstone journey.

Julia Kite, Director of Strategic Initiatives (Vision Zero) at New York City Department of Transportation

Kelliann Beavers, Operations Manager, Urban Data Geeks Lab Director at State of Place

We'd like to thank **Lou Riccio** for connecting us to Vision Zero in the first place.

And we are immensely grateful to **Anindya Ghose**, **Mike Pinedo**, and our advisor **Panos Ipeirotis** for their thoughtful ideas, emotional support, and steady patience, while we figured out what exactly we wanted to share with the world.

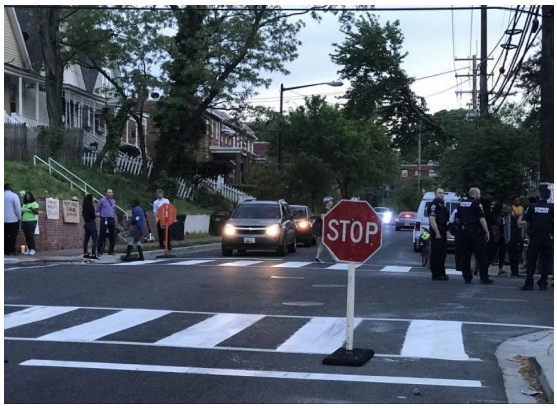
Table of Contents

Background

On April 24, 2019, a group of activists held a vigil on the corner of V Street and 16th Street Southeast in Washington D.C.[2] At this same intersection just days before, a speeding driver ran through a stop sign and struck another vehicle, which then spun out and killed a pedestrian crossing the street.[3] The vigil turned into a rallying cry. "We're burying far too many young people in our community for nothing," a council member told the crowd, "Look at those little stop signs. If you look right there you can't see them while driving."[4] While the District Department of Transportation planned to institute an all-way stop with clearer signage, activists urged more comprehensive action: "We should not be implementing policies based on [where deaths happened]. That should not be the catalyst for change. We should be proactively fixing the streets, safety fixtures, and road conditions."[5]

Figure 1:
The intersection of V St and 16th St SE[6]

Figure 2:
A view of the stop sign at V St and 16th St SE[7]



The tragedy at V and 16th Street is not an isolated one. In 2018, more than 40,000 people died from motor vehicle crashes[8] and 4.5 million more were seriously injured.[9] D.C., along with many other cities, has tried to proactively address the crisis by drawing on a radical approach to traffic safety. They've introduced a program called Vision Zero, which aspires to reduce traffic fatalities down to zero. First launched in Sweden in the 1990s, Vision Zero has gained traction worldwide, and has seen positive results in countries across Europe.[10] The philosophy behind Vision Zero conceptualizes traffic deaths and serious injuries not as unavoidable "accidents," but as preventable tragedies. According to the Vision Zero website:

> *While traditional approaches to transportation safety have prioritized reducing or preventing collisions, Vision Zero instead advocates for the focus to be preventing injuries. Instead of asking "Why did that person crash?" the Vision Zero framework examines "Why was that person so seriously injured in the crash?"*[11]

Vision Zero thus seeks to systematically address the *severity*, rather than the frequency, of collisions. It does so by implementing a variety of traffic interventions related to 1.) *engineering*, such as street redesign, traffic cameras, wider sidewalks, and bike lanes 2.) *enforcement*, such as policing and fines and 3.) *education*, such as community engagement and outreach. Importantly, no

type of intervention is sufficient in and of itself; Vision Zero requires a holistic approach to improving safety conditions.

In 2014, New York City became the first American city to adopt Vision Zero.[12] Mayor Bill de Blasio made the initiative a linchpin of his platform, calling on multiple agencies to work together to lower speed limits, introduce protected bike lanes, and ramp up traffic enforcement.[13] Queens Boulevard was the first major undertaking. Its multiple lanes, stretching as wide as 300 feet in places, made it nearly impossible for pedestrians to cross safely.[14] Between 1990 and 2015, 186 people were killed on what became known as the "Boulevard of Death."[15] Vision Zero assiduously installed bike lanes to protect cyclists, added video cameras to catch speeding drivers, and extended the crosswalk signals. The efforts paid off. Since 2014, no pedestrian or cyclist has been killed on the boulevard, which Mayor de Blasio promptly renamed the "Boulevard of Life."[16]

Vision Zero NYC has made considerable progress over the past four years, reducing total traffic fatalities by 28% in 2018.[17] But pressure for faster results remains. "[Vision Zero] is an achievement, and it should be celebrated," a leader of a public advocacy group told *The New York Times,* "But I think the bigger question is: Why aren't we doing more against the enormity of the epidemic? We're only really taking baby steps."[18]

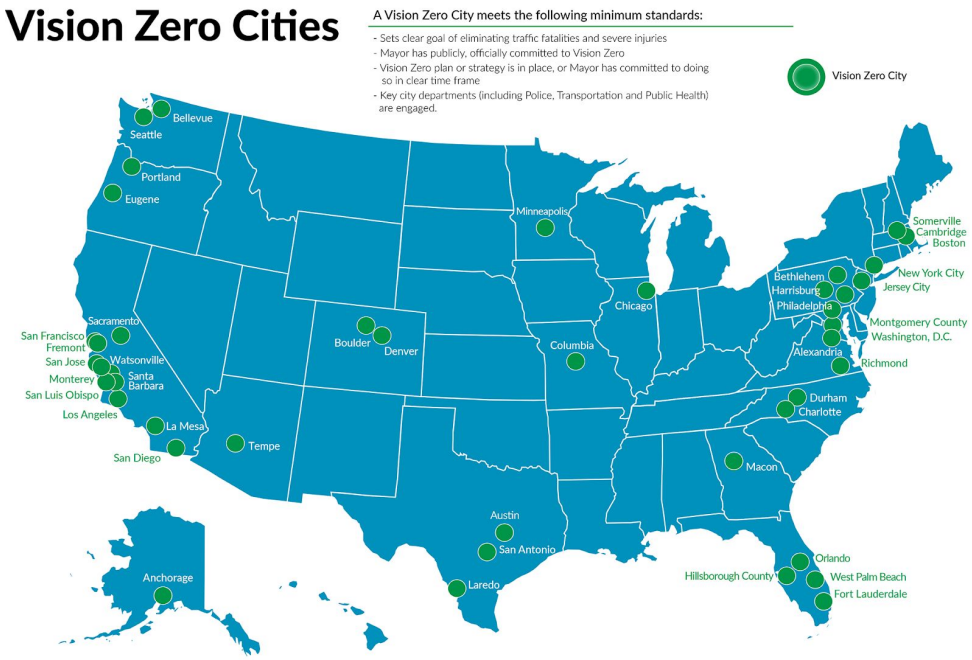
To accelerate momentum, New York has fully committed to Vision Zero as a transportation philosophy. The city has funded the initiative with $1.6 billion through 2021,[19] while also organizing and structuring its collision data to support more data-driven decisions.

Other cities, especially smaller ones, have struggled to do more with less. Limited resources, political barriers, and constrained budgets have hamstrung meaningful progress.[20] And while Vision Zero has been celebrated for its bold ambition, it has also been criticized for being too vague in practice. As Jon Orcutt, a director at Vision Zero Los Angeles, explained, "We had no safety policy before Vision Zero. It's pretty common across the United States. There's nothing to build on and you have established this really ambitious goal. It's hard for everyone to figure out."[21]

Business Understanding

In the years since New York's foray into Vision Zero, other U.S. cities have adopted the program as well. In 2015, mayors in cities such as Portland, Boston, and Denver announced goals of reducing traffic fatalities to zero by 2025.[22] See **Figure 3** for a map of Vision Zero cities.

Figure 3: Map of Vision Zero cities



Given that many cities are still in the nascency stage with Vision Zero, there is a clear and urgent need to extract insights from data and motivate action that can save lives. To help with this process, our Capstone team aims to use predictive analytics to address a primary and secondary objective:

- ***Primary Objective:*** **Where does a city new to Vision Zero allocate resources?** With limited resources and sub-optimal data, it is difficult for a city to know which neighborhoods to prioritize traffic resources and proactive interventions.

- ***Secondary Objective:*** **How can Vision Zero cities better learn from each other?** With Vision Zero mostly operating as a local city initiative, it is difficult to uncover "cross-city" insights, which could strengthen coordination and unify efforts nationwide.

These two objectives are related, but we'll elaborate on each in turn.

1.) Where does a city new to Vision Zero allocate resources?

Before a city can make changes, it must identify locations that should be "treated" with road safety measures. According to the National Cooperative Highway Research Program (NCHRP), the primary process by which this happens is to pinpoint high frequency collision locations and create a list of problematic corridors and intersections. From there, further analysis determines which type of

intervention to design, and how to implement it.[23] We believe there are several challenges preventing cities from doing this effectively.

The first problem is that road safety treatments can be expensive and time consuming. Queens Boulevard, for instance, took $100 million and 24 months to re-construct.[24] By way of comparison, Toronto has approved $22 million for Vision Zero initiatives by 2021, Boston $5 million for 2019, and many other cities under $1 million per year.[25] While New York can afford to transform a multitude of corridors and arterial streets at once, other cities need to be more scrupulous and precise with their target locations.

Another problem has to do with data quality. Although many cities keep historical crash records in their systems, this data can be easily disjointed, misreported, or inconsistent over time.[26] Crash under-reporting, for example, can happen if private settlements occur with insurance companies or if there are problems with transcribing paper files into digital systems.[27] A selection bias has also been observed, as certain kinds of crashes (i.e. more severe or dramatic) are more likely to be reported to the police.[28] Even when "motor vehicle accident reports" are filed, documents can be lost in translation as they flow across agency departments.[29]

Based on our domain knowledge and expert interviews, we believe New York City, D.C., and Los Angeles have the most complete (though still imperfect) repositories for robust and publicly available collision data.[30] These cities have refined the crash data collection process over the past five years, attaching location data to each instance, and reviewing and updating old entries. When the New York City Council, for example, passed Local Law #11 in 2011, it meant that crash data would be collected every month and reviewed by the TrafficStat Unit before being posted on the NYPD website.[31] We'll use this data to build predictive models that may be used in "unseen" cities, who are new to Vision Zero but may not have the infrastructure in place to make proactive treatment decisions.

2.) How can Vision Zero cities better learn from each other?

The secondary goal of our project is to motivate dialogue across Vision Zero cities. To date, cities operate in silos, in part because of the assumption that each city is too different and unique for insights to be transferable. Our hypothesis, however, is that there are shared characteristics — for example, socio-economic attributes — that exist across urban areas and contribute to collision severity. If there are socio-economic features in parts of D.C. that are similar to those in say, Chicago, with respect to their relationship to collision outcomes, then there are meaningful conversations to be had across city lines about how to work together on solving difficult problems.

The academic literature has explored contributing factors to casualty counts in terms of socio-demographic, road design, and land use variables (See **Table 1** for common variables).

Table 1: Commonly used explanatory variables by category

Socio-demographic	Road / Infrastructure / Traffic	Land Use
• Population density • Age (young vs old) • Education • Employment • Income, poverty	• Vehicle Miles Traveled (VMT) / Vehicle Kilometers Traveled (VKT) • Road length • Average annual daily traffic volumes • Bikes lanes	• Retail • Commercial • School

Studies have found any number of these variables to be statistically significant in predicting traffic injuries and deaths. Most often, population has been shown to have a positive relationship with collisions.[32] Other studies have observed relationships between collision outcomes and income[33] , race[34], age[35] , vehicle miles driven[36] , traffic volume[37] , and commercial land use.[38] Very little consensus exists in pinpointing the nature of the relationship, however. For example, even as researchers find positive relationships between population and traffic related injuries[39], others still show negative relationships.[40] Road features are generally thought to be predictive of collision outcomes.[41] See **Exhibit 1** for a complete breakdown of the literature.

Previous studies also analyze collision severity by state or city.[42] Our project follows suit with research that examines collision severity on a more local scale[43], making use of a geographic subdivision called a census tract.[44] Census tracts are small, relatively permanent areas that are made up of a few city blocks, and usually contain anywhere between 1,000 to 8,000 residents. Census tracts are useful because they allow researchers to match socio-economic Census data to very granular and precise areas. We'll describe census tracts further in our "Unit of Analysis" discussion.

The target variable for traffic safety research spans pedestrian, cycling, and motorist injuries and deaths. For our purposes, we'll draw on "casualties" as our primary target, defined as the sum of deaths and injuries in a given area (we'll treat deaths as the most severe injury for reasons we discuss in the "Target Variable" section).

Project Overview

We have selected New York, Los Angeles, and Washington D.C. as our cities of interest between 2013-2017, though we will train our models mostly on New York. Our goal is not only to predict the *count* of casualties that might occur in any given census tract, but to produce a predicted *ranking* of census tracts from highest to lowest casualties, over the five year period. In the absence of robust collision data, a city could then use our ranking model to determine the order of census tracts with which to activate long term, proactive interventions. To explore the feasibility of such a use case, we test three hypotheses:

***Hypothesis 1*:** Using solely a city's Census data, we can predict the ranking of census tract casualties better than random chance.

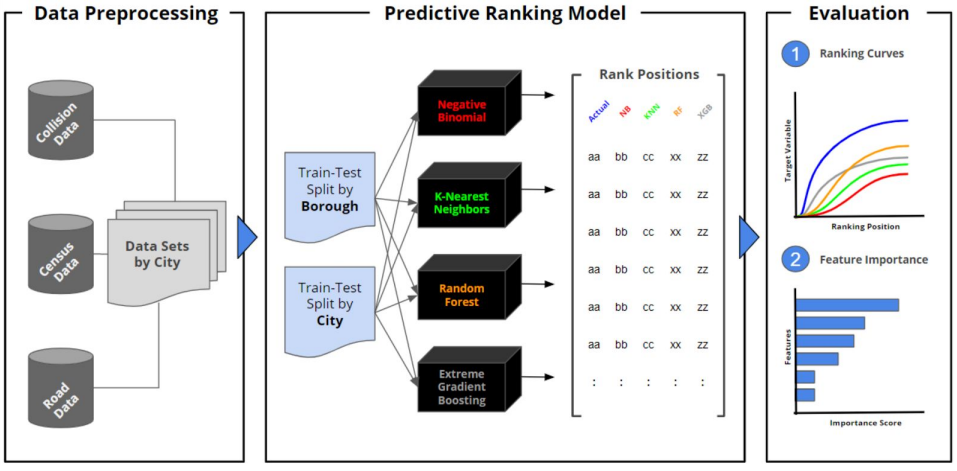
***Hypothesis 2*:** Using a city's Census data *and* road characteristics, we can predict the ranking of census tract casualties better than if we had used Census data alone.

***Hypothesis 3*:** We can predict the ranking of census tract casualties by applying the model built in Hypothesis #1 to another city and achieve results better than random chance.

If we can provide evidence to support any or all of these hypotheses, we have reason to believe that there is something predictive about the socio-economic environment with which casualties occur. This could have implications for where cities decide to focus traffic treatment and inform the nature of the intervention they design. Although we cannot be prescriptive about which specific intervention to develop in any given tract, we can provide macro insights to inspire closer inspection into the role of socio-economic variables in predicting collision severity.

In summary, our project involves joining Census and collision data from three major cities. We apply machine learning techniques to predict the count of casualties per census tract, and then rank the census tracts from highest to lowest casualties. We evaluate performance based on how well our predicted ranking compares to the actual ranking (and random chance), and review each model's feature importance. We present an overview of the workflow in **Figure 4**; we will go into greater detail in the pages that follow.

Figure 4: Workflow diagram of Capstone project



The second part of our analysis investigates “cross-city” insights that can foster greater collaboration across the Vision Zero network. We draw on ordinary least squares (OLS) regression to explore the relationships between socioeconomic variables and casualty counts over a five year period. Here, our intention is to gain a better understanding of the factors driving our predicted ranking, and observe any common statistically significant variables across cities.

Data Understanding

To prepare our data for analysis, we aggregate collision records from NYC, LA, and DC from 2013 - 2017 into one dataset. A collision record in each city represents an individual "motor vehicle accident report," containing the number of injuries and deaths (if any at all) that occurred. The data also provides the latitude and longitude coordinates of each crash, the date, and the cross street name or intersection. Personally identifiable information is not provided.

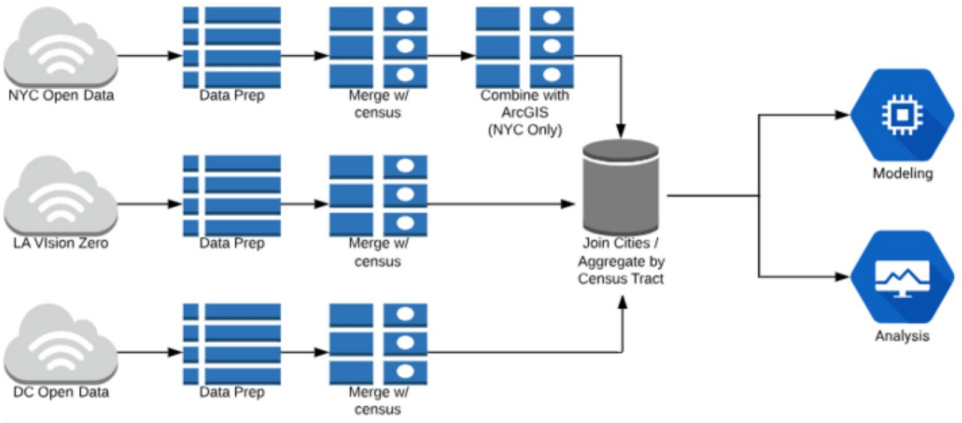
For all three cities, we engaged in an extensive Extract, Transform, and Load (ETL) process. We observed that each city had an independent method for classifying, storing, and defining variables for each collision event, which we had to grapple with, as well as varying time ranges for which data was collected. We also dropped incidents with missing values or incorrectly recorded data elements (**Table 2**). DC took the most cleaning, as some values had incorrect dates in the early stages of their data collection process, and since many observations were outside of our intended date range altogether.

Table 2: Percentage of missing values per city

City	Start	End	% Observations Dropped
NYC	1,389,580	1,209,865	12.9%
DC	197,447	139,170	29.5%
LA	192,027	191,247	0.4%

We matched each crash record's latitude/longitude with a census tract ID (GEOID), which we elaborate further in the Units of Analysis discussion. This allows us to draw boundaries around crash records in order to summarize the total casualties within each census tract. Then, using the Census API, we merge socio-economic data from the American Community Survey 2013 - 2017 to each GEOID, allowing us to fully connect U.S. Census data to total collision data over the five year period. **Figure 5** summarizes our overall ETL process as it pertains to collisions and Census variables.

Figure 5: Workflow diagram of ETL process



Collision Open Data

We collected motor vehicle collision data from the Open Data network of DC, Los Angeles, and New York.[45] We selected these cities because 1.) they represent diverse cosmopolitan areas in terms of size and breadth, socioeconomic background, culture, and road design 2.) they have the most complete information on injuries and deaths for cyclists, pedestrians, and passengers over the time period 2013 - 2017 and 3.) they have both latitude and longitude available for the majority of their instances. A summary of the data is in **Table 3**.

Table 3: Summary of collision data per city or borough

City	Borough	2017 Population	2013-2017 Collisions	2013-2017 Injuries	2013-2017 Deaths
New York	Manhattan	1.7M	211K	38K	182
	Bronx	1.5M	119K	35K	149
	Queens	2.4M	250K	66K	326
	Brooklyn	2.7M	263K	76K	300
Washington D.C.	D.C.	700K	11K	30K	108
Los Angeles	LA	4M	176K	31K	686

Each city structures their data in slightly different ways. While all three cities provided total deaths and injuries for Motorist, Cyclist, and Pedestrians, we recognize that an "injury" in one city could very well mean something different in another city. We noticed, for example, that DC and LA broke down the severity of injuries into "minor," "major", or "unknown" categories.

To keep things consistent, we decided to sum all injuries from DC and LA — major and minor — to create more of an "apples to apples" comparison with New York. In addition, we examined descriptive data to compute the collision to injury ratio. If a city had a drastically different definition of an injury, we'd expect this ratio to be substantially different from another city. However, as shown in **Table 4**, the ratio appears to be comparable. We further discuss our approach to processing the data in a standardized format in the "Data Preparation" section.

Table 4: Ratio of collision to injury per city

City	Collision to Injury Ratio
NYC	0.26
LA	0.18
DC	0.27

American Community Survey (2013-2017)

The American Community Survey (ACS) is an ongoing survey produced by U.S. Census Bureau. The surveys are based on a monthly sample size of 295,000 addresses monthly or 3.5 million addresses annually.[46] Census data is available in 1-year, 3-year, and 5-year estimates. In the context of our business case - to inform long term (multi-year) proactive interventions - we determined that the 5-year Census data was the most appropriate resource. At the time of our project, the most recent 5-year Census data spanned from 2013 - 2017.

The next decision was how to narrow down the ACS's exhaustive list of socio-economic variables. As represented in its 130-page data dictionary, the ACS covers everything from the obvious requisites (i.e. education level, median income) to the seemingly obscure (i.e. World War II participation status and condo fees). Ultimately, we isolated 26 out of the possible 2,620 variables that we determined would reflect the socio-economic environment of a given census tract, including population, race, age, median earnings, modes of commuting, and education (**See Exhibit 2 and 3**).

ArcGIS Online & Roadway Inventory

While much of our initial modeling was focused on Census data, we were able to incorporate additional road variables for New York City. We acquired the data from ArcGIS Online, a cloud based platform from a geographic mapping company called Esri. The New York Roadway inventory data is collected by the New York Department of Transportation, which uses software to analyze traffic flow and surveys to evaluate road quality and pavement condition.[47] This data set includes:

- Road speed limit
- Daily annual traffic
- Road length
- Number of road bumps
- Road pavement width
- Number of road lanes
- "Perceived safety score"
- Road quality index

Using ArcGIS's "Summarize" tool, we applied a census tract layer to the road inventory data. This allowed us to obtain the average, maximum, minimum, and sum statistic of a given road feature within the census tract zone. We then merged the road inventory variables with the socio-economic variables to complete the dataset.

Data Preparation

Moving forward with ETL, we ensured that when combining cities, dates were altered to a standardized format and all fields were coerced to consistent class types.

Each collision record's latitude and longitude was passed through a combination of functions available through R packages "Tigris" and "sp" in order to return the census tract GEOID. We used these packages to make a spatial join, applying census tract shape boundaries to summarize the crash records within each tract. This process was extensive, as we needed to ensure each crash record was uniquely contained in a census tract, and not duplicated in another tract.

Finally, we used the R package "Tidycensus", which "allows users to interface with the US Census Bureau's decennial Census and five-year American Community APIs and return tidyverse-ready data frames."[18] Using Tidycensus, we retrieved Census information by explicitly passing in our pre-selected Census variables as well as the three U.S. states associated with our collisions. This returned the Census variables for all census tracts in each of the three states.

For each city, we performed a left join with collisions and Census Variables using GEOID as the foreign key, thereby merging the collision records with the corresponding Census data variables.

Lastly, we appended the three cities together, now with their Census variables, to create one master data set for the purpose of analysis and modeling.

We believe this dataset alone is a meaningful contribution to Vision Zero, which can use it to make cross city comparisons and analyses. As best we can tell, no previous research has attempted to merge public collision data across cities, in part, we believe, because of the laborious effort required to join and merge them cohesively.

Unit of Analysis

Our project relies on census tracts as our unit of analysis. Census tracts can vary based on population in a given zone. For example, a rural area with low population might have geographically larger, widely dispersed census tracts ranging several miles in diameter, while metropolitan areas with high populations might have geographically smaller, tightly packed census tracts spanning only a few city blocks (See **Figure 6**).[48]

Figure 6: Census tract geography for DC, LA, and the five boroughs of NYC

Target Variable

Given our business context, we chose to focus our modeling on total casualties (sum of deaths and injuries) per census tract. While our data set affords us the opportunity to model many collision outcomes, we wanted to choose a target variable that would align with Vision Zero's mission of improving public safety. For this reason we chose to focus on casualties per census tract as our target variable, which emphasizes the public safety implications of collisions, as opposed to the raw number of collisions themselves.

The overwhelming majority of incidents (1.2M) in our data involved zero injuries and deaths. As illustrated in the histogram below, approximately 80% of collision incidents had zero injuries or deaths for all three cities (**Figure 7**). We also observed that total census tract collision counts were positively correlated with census tract injury counts, but the relationship was less clear with census tract deaths (**Figure 8**).

Figure 7: Total Casualties per Collision, 2013 - 2017

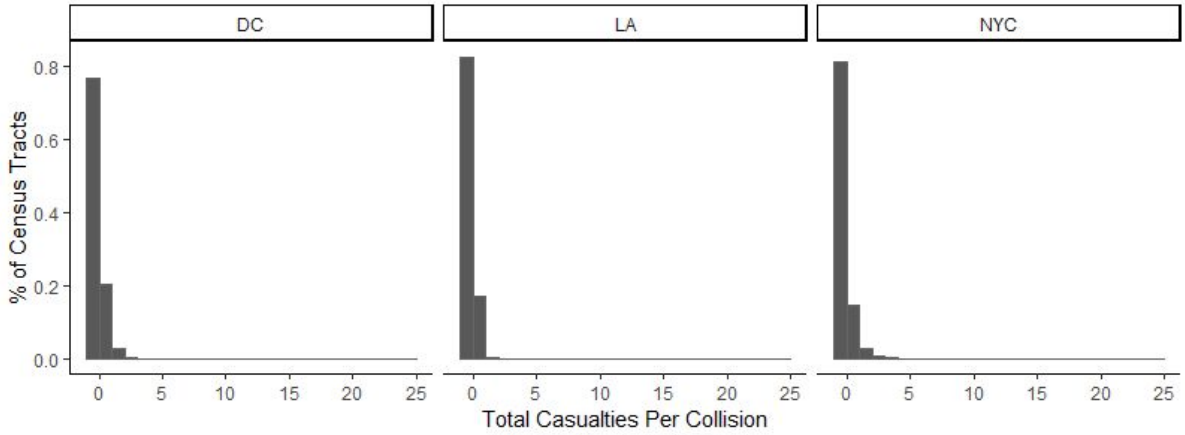
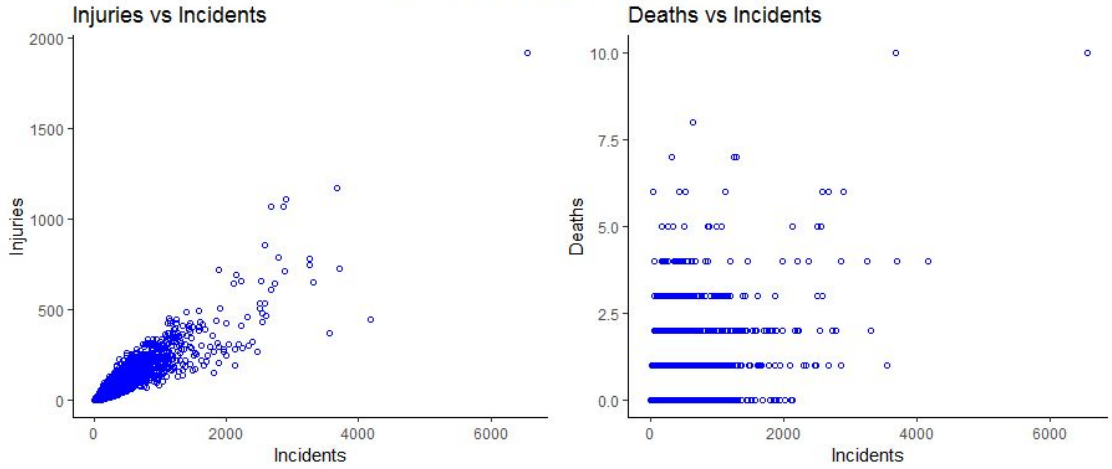
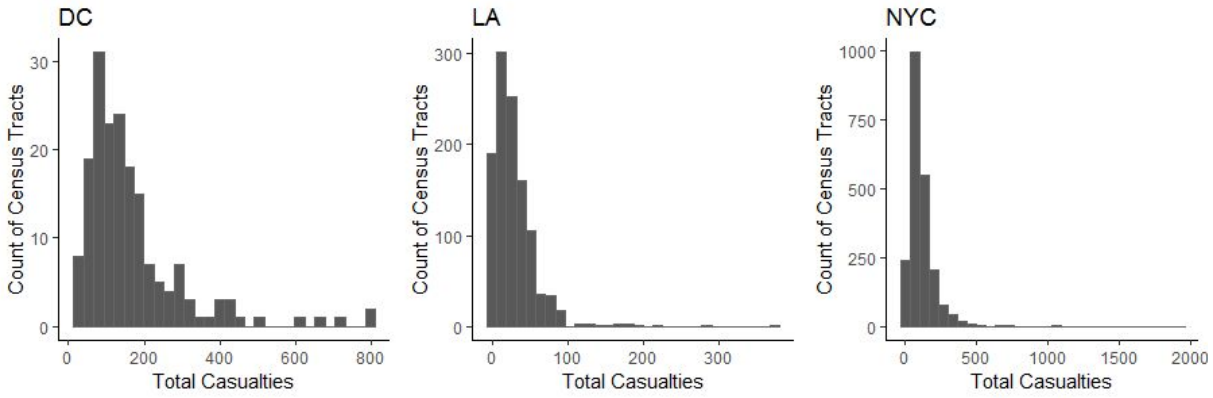


Figure 8: Injuries & deaths vs. incidents (all census tracts combined: NYC, LA, DC), 2013 - 2017



In all cities we observed a power law distribution in incidents, injuries, and deaths by census tract, which indicated that for these metrics, most census tracts had low counts, while a small percentage of census tracts had significantly higher counts. This was particularly clear for total casualties (**Figure 9**).

Figure 9: Distribution of count of casualties per census tract

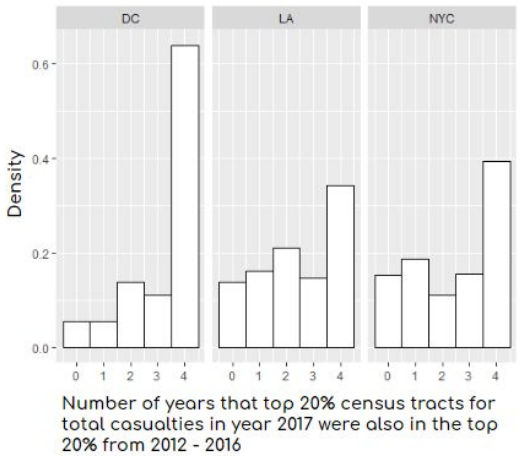


The insight that census tract injuries and deaths follow a power law distribution is valuable in the context of the business problem because it suggests that cities could make drastic improvement in public safety by targeting a small amount of concentrated areas and steadily roll out proactive interventions.

As an additional step, it was important to determine whether top census tracts by total casualties generally *remained* among the top ranks for multiple consecutive years. This would help ensure our predictions were based on stable observations, as opposed to volatile or random swings each year. It would also help ensure that long term interventions (road improvements, bike lanes, etc) deployed as a result of our modeling would impact census tracts that would have otherwise remained high risk.

We ranked census tracts by total casualties relative to all other tracts in their city — highest to lowest — for each year from 2013 - 2017. We then identified whether each census tract was among the top 20% in each year, for incidents, injuries, and deaths. Using the top 20% of tracts from 2017 as a base, we calculated how many times these same tracts were also in the top 20% each of the previous 4 years (**Figure 10**). In all three cities, at least one third of census tracts that ranked in the top 20% of casualty count in 2017 were also among the top 20% in all four of the previous years, and over 50% remained among the top for at least three out of the four prior years.

Figure 10



Predictive Analytics Using Ranking Methods

So far, we've discussed how our project seeks to help cities 1.) prioritize traffic resources at the census tract level and 2.) uncover insights that can be transferred across cities.

It is important to choose modeling techniques compatible with our business problem. Therefore, we intend to *rank* census tracts by number of casualties, and recommend greater investment in tracts at the top of the ranked list. That way, Vision Zero would not be evaluating each tract separately, but rather taking action on the top *n* tracts above a certain threshold. One could imagine Vision Zero descriptively looking at past data to establish such a ranking. However, historical data would not reveal the informativeness of different attributes in producing the ranking; it would not explain *why* the ranking is the way it is, which is crucial to designing interventions across cities.

In our context, we use predictive analytics, not necessarily to forecast a future event, but rather to estimate unknown values (census tract ranks) in the present.[49] The value, then, is in the understanding gained by making predictions rather than the actual predictions themselves. For example, we can learn about the relative importance of socio-demographic characteristics and road features in neighborhoods wherein collisions occur. Recall our three hypotheses:

> ***Hypothesis 1*:** Using solely a city's Census data, we can predict the ranking of census tract casualties better than random chance.
>
> ***Hypothesis 2*:** Using a city's Census data *and* road characteristics, we can predict the ranking of census tract casualties better than if we had used Census data alone.
>
> ***Hypothesis 3*:** We can predict the ranking of census tract casualties by applying the model built in Hypothesis #1 to another city and achieve results better than random chance.

We'll examine each in turn. By building a ranking model within and across cities, we can determine the extent to which Census and road features are predictive of casualty rankings, rather than just casualty counts. Our methodology to test each hypothesis is as follows:

- Predict count of casualties by census tract
- Rank census tracts from highest to lowest predicted casualties
- Compare predicted ranking with an actual and random ranking

We evaluate our models' performance based on how well the predicted ranking stacks up to the actual ranking, and on its lift above a random ranking.

Predicting Count of Casualties by Census Tract

In order to produce a ranking, our model must first generate a predicted casualty count for each census tract. To balance complexity, performance, and explainability, we draw on four regression algorithms with different strengths:

1. **Negative Binomial Regression (NB)[50]:** As crash data is non-negative count data, NB is the generalized linear model most frequently used in previous research.[51] In our data, we choose NB, as opposed to a Poisson derivative, because the mean and variance are not the same.

2. **K-Nearest Neighbor Regression (KNN)[52]**: KNN is a machine learning algorithm that, unlike generalized linear models, does not require any assumptions about how the underlying data is distributed. In the simplest terms, KNN looks at how similar the features of the test set are to those of the training set. A predicted value is then determined through this logic.

3. **Random Forest Regression (RF)**: RF is a popular machine learning algorithm that, as the name implies, creates a "forest of multiple decision trees." This ensemble-type tree model produces multiple decision trees, with each tree randomly selecting a subset of features. This process improves prediction accuracy because it is based on the results of many randomly structured decision trees. In addition, the RF algorithm in the scikit-learn module also produces information about how important each feature is in producing the predictions. This is called "feature importance," which is measured by how much a particular feature decreased "impurity" when used by the classification trees.[53] We incorporate this calculation into our model to understand which features are driving the model prediction results.

4. **Extreme Gradient Boosting Regression (XGBoost)[54]**: XGBoost is another powerful ensemble-type tree-based machine learning algorithm. The strength of XGBoost is that instead of drawing from the combined strength of multiple randomly structured decision trees, each tree constructed by XGBoost is sequentially learning "from the mistakes" of the previously constructed trees in order to produce the best predictions. Similar to RF, XGBoost also produces its own feature importance calculation, which we also use as part of model evaluation.

We build a model workflow that passes the training data through each of these four algorithms (with certain hyperparameters[55]), then produces four separate predictions for each census tract in the test data. In addition, we use the feature importance calculations from RF and XGBoost to identity the key features that are driving the results.

Ranking as Evaluation Metric

Traditionally, the metrics most often used to evaluate predictions, such as Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE), produce scores based on the difference between a predicted value and its actual corresponding value. This approach works well if all instances are of equal importance to us.

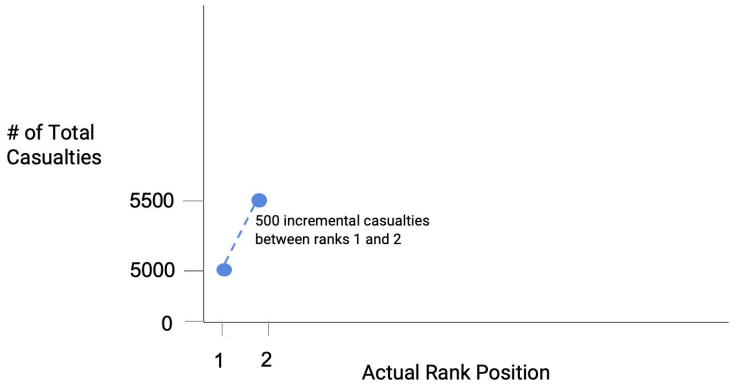
However, as previously stated, we want to contribute to Vision Zero's effort by helping cities prioritize census tracts where resources are needed the most. Similar to search engine algorithms, of which ranking methodologies are heavily rooted, Vision Zero is likely only concerned with the top *n* % of census tracts for long term interventions, and less so in the remaining tracts.

The ranking methodology can be broken down into three parts: the actual ranking, predicted ranking, and random ranking.

Actual Ranking

First, we need to define what a “perfect ranking” would be. Since we want to prioritize the census tracts with the highest number of casualties, we define “perfect” as the cumulative sum of casualties for each rank position (sorted by casualties in descending order). A simple hypothetical example helps to illustrate the methodology. In **Figure 11**, the census tract with the highest rank (#1) represents 5,000 actual casualties. The second highest ranked census tract has 500 casualties, which makes the cumulative sum of casualties 5,500 at rank position two.

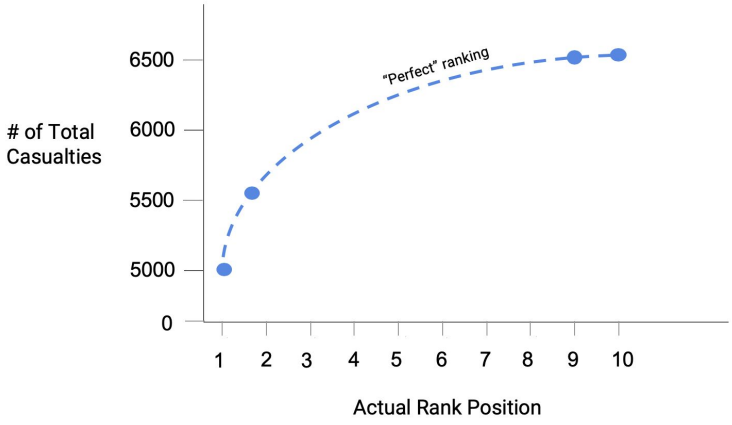
Figure 11: Incremental change in actual ranking



This calculation continues until all census tracts are accounted for. Since the perfect ranking will reflect the power law distribution of casualties, we see the curve have a steep slope early on, but eventually flatten out, as shown in **Figure 12**. This is because the incremental increase in the cumulative number of casualties decreases as we go down to the lower-ranked census tracts, producing fewer numbers of incremental casualties.

It is important to note that we include ties in this ranking. For example if census tract A had 11 casualties, tract B had 10 casualties, and tract C had 10 casualties, tract A would get a rank of 1, while both tracts B and C would get a rank of 2. This is common in the actual results but not for our predicted results.

Figure 12: Incremental changes make up the predicted ranking curve



Random Ranking

We incorporate a "random ranking" to provide a baseline level of performance. The intuition behind a random ranking is that if a predicted ranking is not noticeably different compared to the random ranking, then we can say the predictions are as good as randomly guessing.

The random ranking methodology is done in several steps:

1. We order the the census tracts by the predicted casualties, as produced by the machine learning and negative binomial models. In a hypothetical example, let's say Random Forest predicts a casualty count of 7,000 for the highest rank.

2. For each rank position, we subtract the predicted number of casualties (7,000) from the actual number of casualties (5,000). In this case, the difference is 2,000. This gives us the "number of remaining casualties" unaccounted for at that given rank position.

3. We determine what a "random" count of casualties might be. We do this by taking the "number of remaining casualties" at a given rank (2,000) and dividing it by the number of rank positions left in the overall ranking sequence. The output of this is the "gain" over the random prediction at each rank, representing the difference between the "perfect" score and the "random" score.

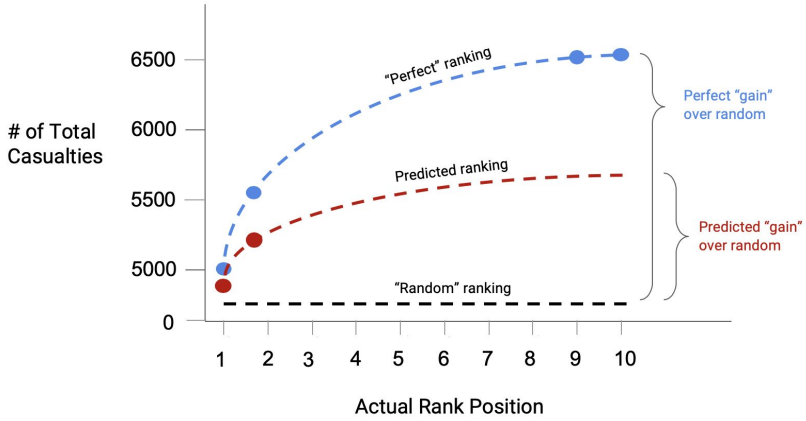
Predicted Ranking

Lastly, we incorporate the predicted values into ranking curves. Similar to the random method, we first ordered the census tracts from highest to lowest predicted casualties. From there, at each rank we add up the actual values based on the corresponding predicted rank. For example, if there are three predicted values of 7000, 500, and 200 and the corresponding actual values are 5000, 500, and 300, then the new column, "total", would have values of 5000, 5500, and 5800 (the cumulative gain of the actual values). We sum the values from the actual scores that correspond to the predicted rank.

Next, for each rank we sum the total actual casualties and subtract the previous rank's "total." This is essentially the "remaining" casualties after the prediction has been made. Similar to the "perfect" score, we calculate what the "random score" would be, then the "gain" over random at each rank, and finally the cumulative "total gain" over the random values.

As shown in **Figure 13** below, once we have values from the perfect ranking and the random ranking, we simply plot them. Intuitively, the closer the predicted ranking curve is to the the perfect ranking curve, the greater the lift over random chance. In the context of comparing multiple predicted ranking curves, as we do in assessing our model results and the algorithms, the curve that is closest to the perfect ranking curve is deemed the strongest performing algorithm.

Figure 13: Using ranking curves as model evaluation metric



There may be situations where it may make sense to ignore values after a certain rank position. For this exercise, however, we will be evaluating the algorithms on the total performance throughout all ranking positions. Vision Zero officials can select whichever threshold they deem appropriate.

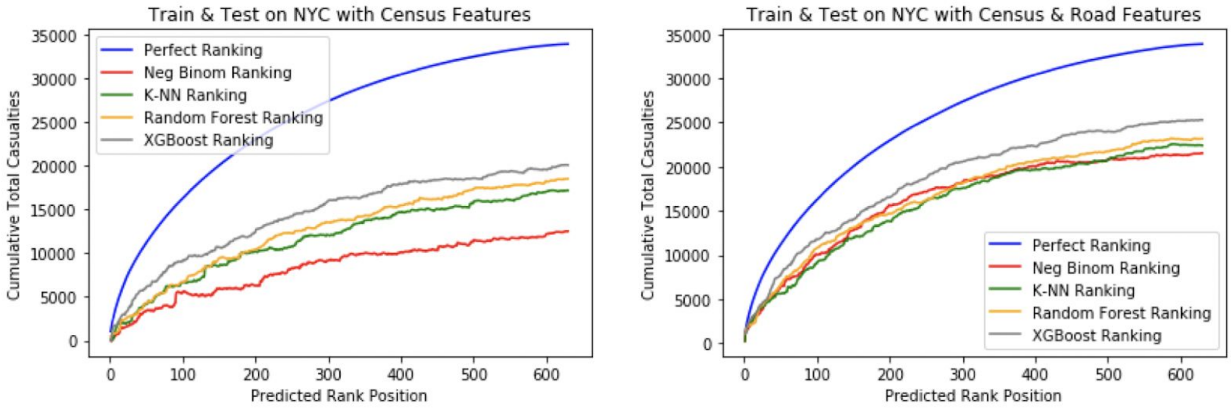
Predicting the Rank of Casualties Within a City (Hypothesis #1)

To test the first hypothesis, that we can predict the ranking of census tracts based on Census data, we split the NYC data into a training set and a test set. We use the stratified method to split the data set to ensure that our training and test data are representative of each New York City borough in terms of casualties. As shown on the left side of **Figure 14**, each of the four algorithms in our predictive model produce a predicted ranking curve that is better than random. Even though there is a substantial gap between the perfect ranking curve and all four of the predicted ranking curves. This suggests that there is some predictive power within Census features that relate to the count of casualties at the census tract level.

Predicting the Rank of Casualties with Road Features (Hypothesis #2)

The second hypothesis states that we can improve the ranking predictions if we incorporate road condition data. To do this, we use the same census tracts in the prior training data set with the addition of road inventory features (again, NYC only). The results, as shown on the right side of **Figure 14**, demonstrate that road condition features can indeed improve ranking predictions. All four predicted ranking curves are now closer to the perfect ranking curve. In addition, the performance of each algorithm seem to almost converge with one another at each ranking position.

Figure 14: NYC ranking curves with and without road features



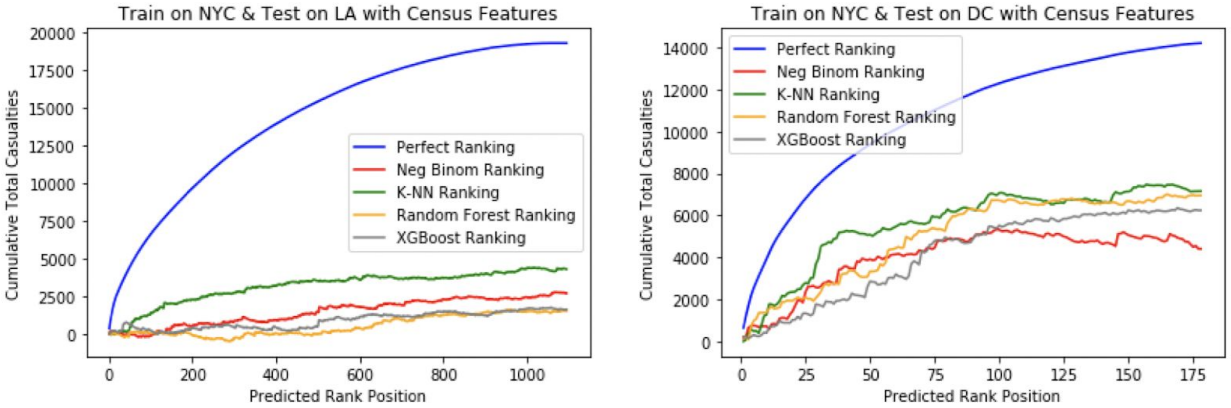
Predicting the Rank of Casualties Across Another City (Hypothesis #3)

The last hypothesis we test is whether a model trained on one city's Census variables can be useful to predict another city's casualties at the census tract level. That is to say, if the predicted ranking in this case is better than random, then we can make a case that Census variables do have predictive power even at a cross-city setting. This has real business implications for organizations like Vision Zero, as the robustness of one city's data may be useful to another city.

For this test, the training and test data sets are different compared to the previous two tests. In the prior tests, the training and the test sets come from NYC. In this test, however, we train the model on the entire NYC data set, and test it on the LA data set, as well as on the DC data set.

As shown on the left side of **Figure 15**, the model trained on NYC data and deployed on LA data seems less predictive when compared to the results from the previous tests. The algorithms performed barely better than random, as the curves are closer to the X-axis. The exception is K-NN, which performed meaningfully better than others. This tells us that the predictive power of Census variables, trained from NYC, is less clear when deployed on LA data.

Figure 15: Predicted LA and DC ranking curves with NYC training data



In the case of the DC data set, however, the deployment of the same model shows an increased relative performance. As shown on the right side of **Figure 15**, the predicted ranking curves from all four algorithms are noticeably closer to the perfect ranking curve. This may suggests that, in this case, the predictive power of Census variables on census tract casualties is non-trivial.

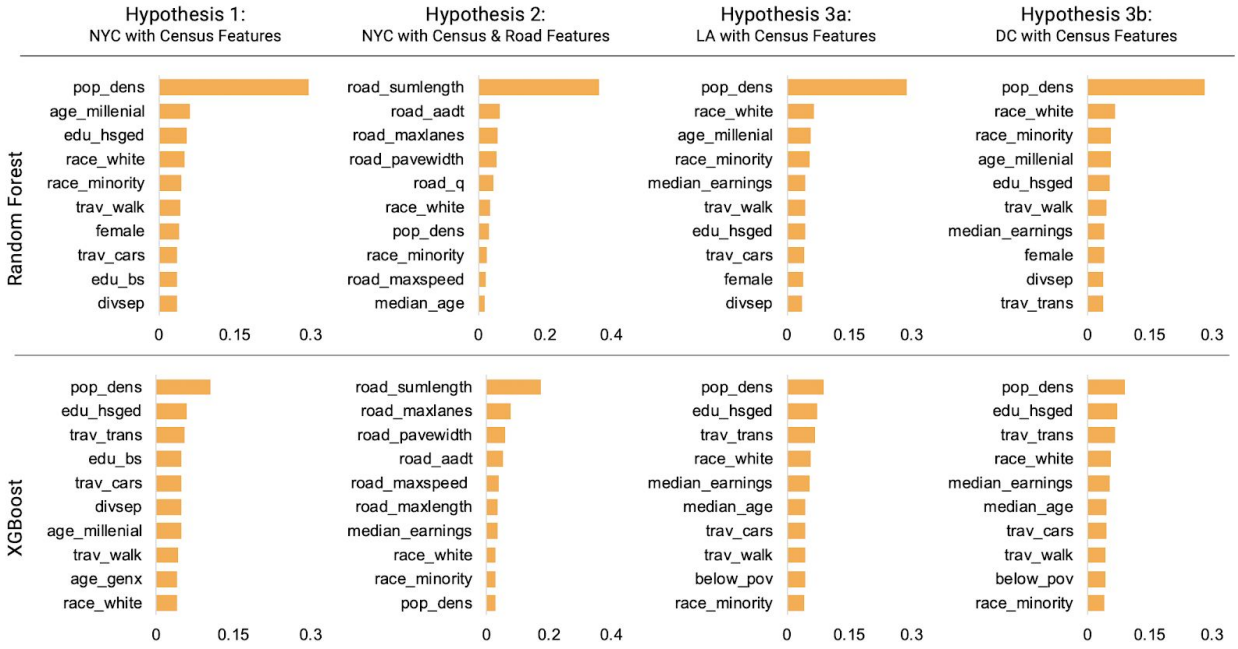
The difference in results between the LA data versus DC data begs the question of why the model performs significantly better for DC than it does for LA. Recall from the second test that the inclusion of road condition data can greatly increase the performance of our predictive model. Since we do not have access to the same road condition data for LA and DC, we could not test whether we would see a similar lift in model performance. However, it stands to reason that parts of NYC could be more similar to DC than LA, in terms of the way that people travel throughout the city and the characteristics of the neighborhoods.

Feature Importance

As previously mentioned, both RF and XGBoost algorithms produce feature importance calculations. This is helpful for us to pinpoint the top features that drive prediction results. It is important to note that the features that are highlighted to be important for RF may be different from that of XGBoost. This is to be expected because the inherent distinction in how the two algorithms perform predictions.

As demonstrated in **Figure 16**, a set of bar charts the top 10 features by the respective algorithms for each of the hypothesis tests, shows that *pop_dens*, or population density[56] is the single most important feature in tests that do not include road condition data.

Figure 16: Feature importance from RF and XGBoost with count of casualty as the target variable



The importance of a feature describes only its level of informativeness, not the relationship that it bares on the target variable. In other words, we can not know from the model results whether

higher *pop_dens* means more casualties, or vice versa. We touch on this subject in our secondary business objective.

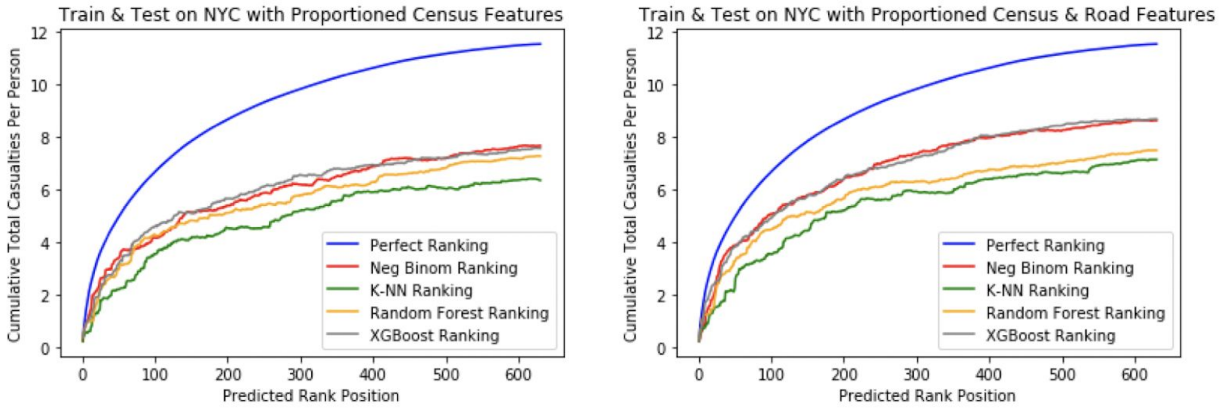
Testing the Three Hypotheses with Normalized Data

Given the feature importance results, we question whether casualties are simply a function of the population size of a census tract. As a robustness check, we control the effects of population size of the census tracts by normalizing both the target variable and the Census feature variables by the population size. We rerun the model with a normalized target variable, *casualties per person*, with the Census variables transformed as proportion of population.

In addition to normalization, we also remove any tracts that have under 200 count population for New York City and Los Angeles, which accounts for less than 1% of the tracts in the respective cities. This is to ensure that the normalized values do not create outliers that would skew the model results.

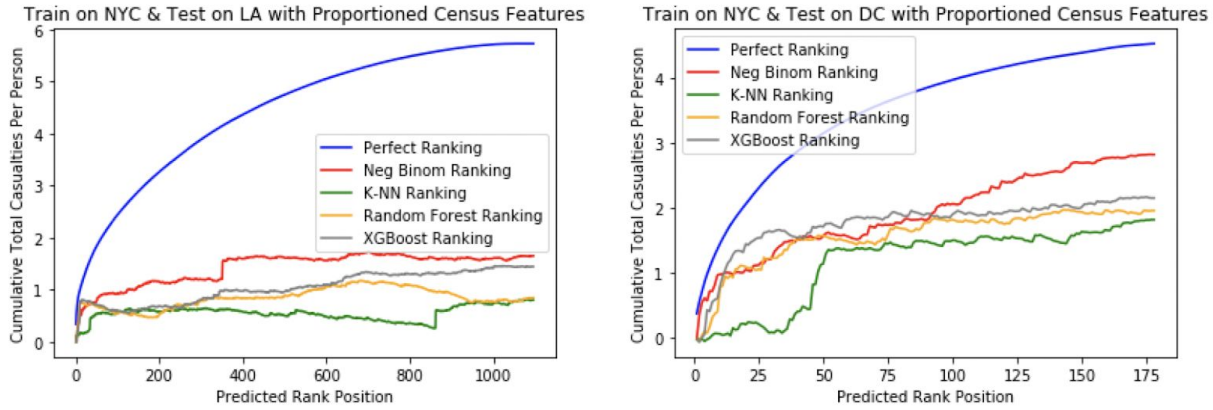
As shown in **Figure 17**, all four algorithms predict rankings of casualties per person better than random with normalized target variable and features. We see a similar result in that the model's performance remain strong when we add road condition data, albeit with less noticeable lift as compared to the previous iteration of the test (refer back to **Figure 14**).

Figure 17: NYC predicted ranking curves with proportioned features



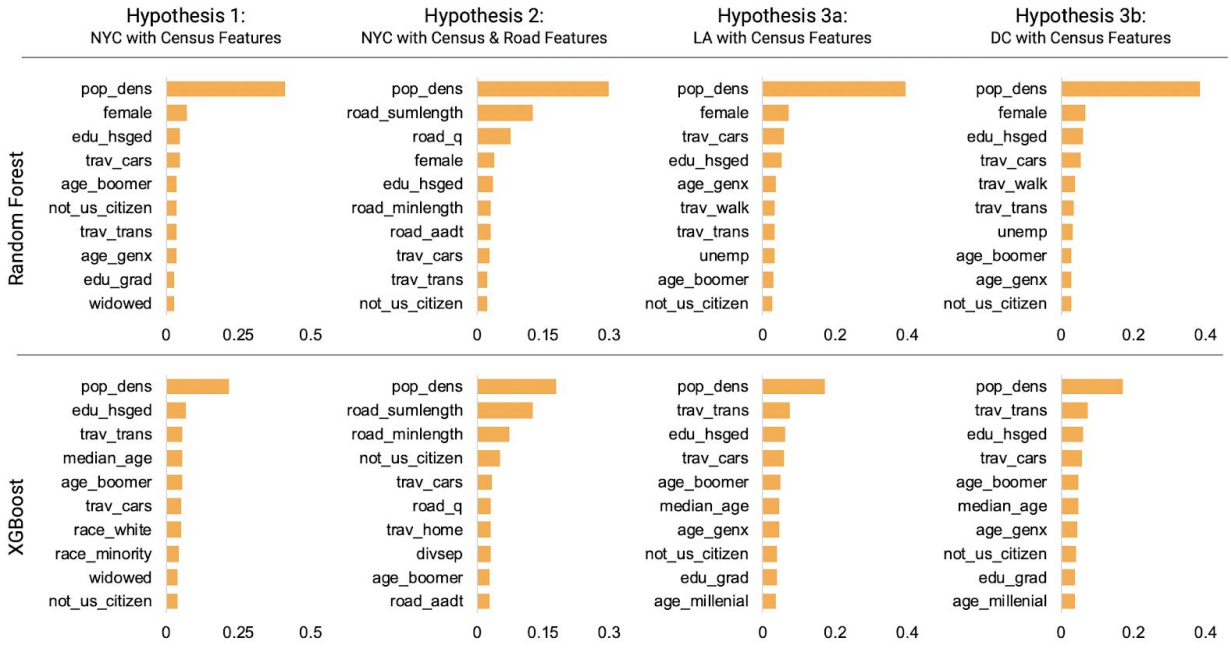
Furthermore, as shown in **Figure 18**, applying the model outside of NYC data, to LA and DC, we see the model's performance relatively unchanged compared to the previous, unnormalized iterations (refer back to **Figure 15**). These results tell us that even when controlling the effects of population size at the tract level, the predictability of the model continues to exhibit better-than-random accuracy.

Figure 18: LA and DC predict ranking curves with proportioned features



In examining the importance of the normalized features, we see that *pop_dens* continues to be the single feature that drives the prediction results. As shown in **Figure 19**, even in hypothesis #2 where both Census and road condition features are trained in the model, *pop_dens* exceeds the importance of any and all road condition features.

Figure 19: Feature importance from RF and XGBoost with casualties per person as target variable



Implication of Modeling Results

In conclusion, our model can predict the ranking of census tracts both by total casualties and casualties per person with non-trivial accuracy performance. Furthermore, we find that with both Census data *and* road characteristic data, the accuracy of the model increases meaningfully. These findings satisfy the first business objective, which is to demonstrate that we can indeed help Vision Zero prioritize and focus resources on the census tracts that need them the most.

In the context of our second business objective, to extract transferable insight from one city to another, we manage to uncover that population density is a predictive feature that may work across cities. In order for this insight to be actionable, however, we take a step further in our analysis to examine the relationship that may exist between population density and casualties at the census tract level.

Uncovering Relationships & Cross-City Insights

Our ranking model helps us understand which variables contain predictive power, but it does not reveal the direction of the relationships between explanatory variables and casualties, nor in the magnitude of their effect. We apply ordinary least squares (OLS) regression to gain further insight into how Census and road variables relate to casualties.

We begin by including explanatory variables for which we could come up with a plausible story as to how they might impact casualties. For example, variables for gender and education may hold predictive value, but it is hard to rationalize how they would matter in ways that would not be captured by other variables. In addition to our own intuition, we relied on the literature to inform our variable decisions, noticing that most studies identified features in **Table 5** as significant (Also see **Exhibit 1**).

Table 5: Frequently cited statistically significant explanatory variables in relation to collision casualties

Census Variables	Road Variables
• Population density • Median age • Median earnings • % who commute via public transit • % who commute via walking • % who commute via biking • % who commute via driving • % minority race	• Max number of lanes • Annual average daily traffic • Perceived road safety • Road pavement width • Max speed limit

To avoid skewing the results due to population size, we choose “casualties per population” as our target variable. We focus attention on census tracts in New York City from 2013-2017, since NYC has both Census and road features available. After removing census tracts with missing values, we observe the following summary statistics in **Table 6**.

In looking at the distributions of each variable, we notice casualties per person, population density, and average annual daily traffic features exhibit right skewness. We log these variables to normalize them for OLS. Our regression results (**Figure 20**) produced an adjusted R squared of 0.54, with several significant variables. We look for multicollinearity as a robustness check (see **Exhibit 5**) but even when co-correlated variables are removed, the overall R squared holds and statistically significant variables do not drastically change.

Table 6: Summary statistics of select features

Summary Statistics	NYC n = 1954				LA n = 1097				DC n = 178			
Variables	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Casualties Per Population	0.06	0.49	0.00	18.02	0.02	0.46	0.00	15.20	0.05	0.05	0.00	0.42
Pop. Density	53912	36383	7.96	241500	16717	13520	0.62	96800	17720	11983	2208	66344
Median Age	36.37	6.06	12.60	65.7	36.08	6.50	17.50	68.20	34.93	6.08	19.70	50.20
Median Earnings	38461	18621	3059	155030	32576	16252	2499	101528	48045	21707	2499	110872
% Car Commuters	0.13	0.08	0.00	0.54	0.37	0.09	0.02	0.65	0.21	0.07	0.01	0.37
% Pub. Trans. Commuters	0.27	0.10	0.00	0.80	0.05	0.05	0.00	0.41	0.19	0.08	0.001	0.41
% Bike Commuters	0.01	0.01	0.00	0.10	0.01	0.01	0.00	0.20	0.02	0.02	0.00	0.15
% Walking Commuters	0.04	0.05	0.00	0.66	0.02	0.02	0.00	0.28	0.07	0.09	0.00	0.38
% Below Poverty Line	0.19	0.13	0.00	0.75	0.21	0.13	0.00	0.81	0.18	0.12	0.01	0.63
Max Speed Limit	20.11	14.23	0.00	55	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Max Number of Lanes	2.04	1.55	0.09	10.00	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Annual Avg. Daily Traffic	6485	10690	0.20	118413	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Road Quality Index	25.67	1.97	17.65	30.34	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
% Minority Race	0.69	0.29	0.00	1.00	0.71	0.27	0.08	1.00	0.66	0.29	0.16	1.00
Pavement Width	4.23	3.86	0.01	36.32	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Figure 20: OLS regression with select Census & road variables, NYC

Source	SS	df	MS	Number of obs.	=	1,953
				F(10, 1942)	=	234.800
Model	837.566	10	83.757	Prob > F	=	0.000
Residual	692.744	1942	0.357	R-squared	=	0.547
				Adj R-squared	=	0.545
Total	1530.310	1952	0.784	Root MSE	=	0.597

Log Casualties Per Pop.	Coef.	Std. Error	t	P>t	[95% Conf.	Interval]
Log Popularion Density	-0.786	0.019	-40.66	0.000	-0.824	-0.748
Median Age	-0.010	0.003	-3.76	0.000	-0.015	-0.005
Median Earnings	0.000	0.000	4.56	0.000	0.000	0.000
% Car Commuters	-3.458	0.242	-14.27	0.000	-3.933	-2.983
% Pub. Trans. Commuters	0.271	0.175	1.55	0.122	-0.073	0.614
Max Speed Limit	-0.006	0.002	-3.94	0.000	-0.010	-0.003
Max Lanes	0.010	0.012	0.83	0.409	-0.014	0.034
Log Annual Avg. Daily Traffic	0.146	0.014	10.33	0.000	0.118	0.174
% Minority Race	0.707	0.061	11.58	0.000	0.587	0.827
% Walk * Bike to Work	31.983	10.085	3.17	0.002	12.203	51.762
Intercept	3.669	0.272	13.50	0.000	3.136	4.203

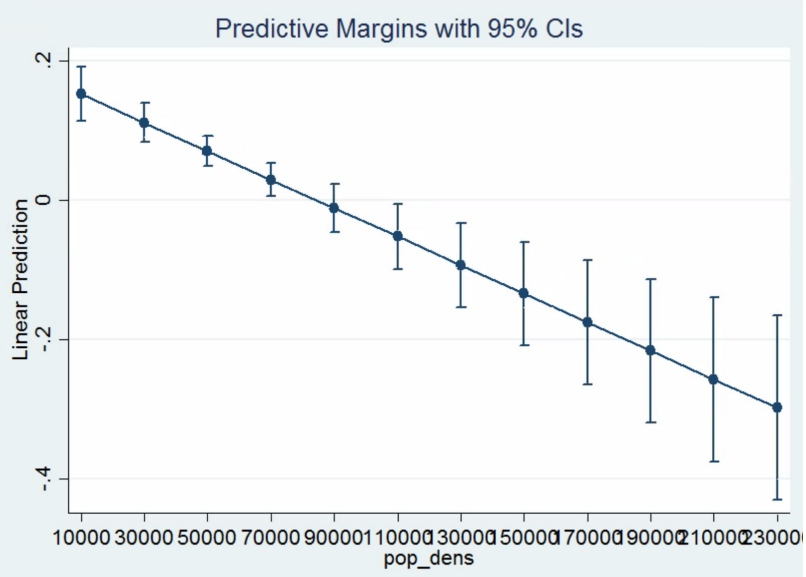
Population Density

We find that population density has a negative relationship with casualties per person, meaning that as the number of people per square mile goes up, the average casualties per person goes down. This was a surprising result, and slightly counterintuitive. Why would census tracts with more people per square mile (i.e. many of those in Manhattan) have, generally speaking, fewer casualties per person? We might postulate that census tracts with higher population density have more street signage, lower speed limits, and "safety in numbers." But perhaps this is just the case of New York City, biased by Manhattan's unique urban structure. After all, there is presumably a strong difference between Manhattan's day- and night-time population.

We ran dozens of other regressions with the same variables on other boroughs, such as the Bronx, Brooklyn, Queens, Staten Island, and other cities, such as Los Angeles and D.C. In all cases, population density had a negative relationship with casualties per person. We changed the target variable to number of casualties, removed census tracts with less than 200 people, and tried various combinations of Census and road features. Still, population density exhibited a negative relationship to casualties.

To illustrate population density's impact, consider **Figure 21** for a marginal analysis of the first NYC regression. The chart holds all other explanatory variables constant in our model, while changing population density in 20,000 increments from 10,000 to 300,000 people, and produces the resulting impact on the casualties per person. As you can tell, the greater the population density, the lower the expected number of casualties per person (the bands are the confidence intervals).

Figure 21: Marginal Analysis, Population density vs. Linear prediction of casualties per person, NYC



We speculate this result could be a function of urban areas specifically, where there is a large portion of the population walking and biking to work, as opposed to suburbs or small towns, where most people are driving everywhere they go. It could be that in more rural areas, population density has the opposite effect than the one we're finding here.

Cross-city Insights

Our regression analysis can also help cities ask better questions. Which types of areas are prone to highest casualties per person across cities? Population density was not the only variable that stood out. We created an interaction variable in our regression because we believed there was a multiplicative effect in the percent biking and walking to work. In other words, we believe the people who are traveling to work via bike or walking live nearby their place of employment and frequently opt for either one. The interaction was significant in New York, LA, and DC (**Table 7**).

Table 7: Cross-city OLS regression results with select Census variables only

DV: Log Casualties Per Person	NYC			LA			DC		
IV	Coef.	Std. Error	P-Value	Coef.	Std. Error	P-Value	Coef.	Std. Error	P-Value
Intercept	5.745	0.243	0.000	-3.028	0.552	0.000	0.631	0.975	0.518
Log Pop. Density	-0.820	0.019	0.000	-0.192	0.047	0.000	-0.463	0.093	0.000
Median Age	-0.008	0.003	0.005	-0.013	0.007	0.054	0.030	0.011	0.008
Median Earnings	-0.001	0.000	0.082	-0.000	0.000	0.944	-0.001	0.000	0.045
% Pub. Trans. Commuters	0.440	0.185	0.017	6.708	0.821	0.000	1.205	0.792	0.130
% Car Commuters	-3.856	0.254	0.000	-0.738	0.424	0.082	-2.831	0.918	0.002
% Walk * Bike to Work	29.540	10.702	0.006	164.521	24.017	0.000	53.793	18.387	0.004

While the percent traveling to work via walking or biking has a positive relationship to casualties per person, the percent commuting to work via cars has a negative relationship with casualties per person, across all cities. This makes us wonder if crashes are impacting people near where they live. If so, then the people impacted would be those who are lower income, working class populations who commute to work without cars. Average annual daily traffic has a positive relationship with casualties per person and is statistically significant in the Bronx, Queens and Brooklyn. This provides further evidence that neighborhoods that are often driven through, perhaps with traffic enroute to the city, are associated with higher casualties per person (**Table 8**).

Table 8: Cross-borough OLS regression results

DV: Log Casualties Per Person	Queens			Bronx			Brooklyn		
IV	Coef.	Std. Error	P-Value	Coef.	Std. Error	P-Value	Coef.	Std. Error	P-Value
Intercept	3.162	0.524	0.000	3.914	0.574	0.000	3.565	0.519	0.000
Log Pop. Density	-0.823	0.038	0.000	-0.774	0.037	0.000	-0.783	0.038	0.000
Median Age	0.000	0.005	0.916	-0.006	0.008	0.464	-0.012	0.005	0.009
Median Earnings	-0.000	0.000	0.003	-0.000	0.000	0.000	-0.000	0.000	0.400
% Car Commuters	-0.932	0.502	0.064	-0.346	0.672	0.607	-2.725	0.498	0.000
% Pub. Trans. Commuters	2.119	0.387	0.000	0.883	0.554	0.112	0.562	0.300	0.062
Max Speed Limit	-0.001	0.002	0.625	-0.003	0.003	0.405	-0.011	0.003	0.001
Max Number of Lanes	0.035	0.019	0.066	-0.026	0.027	0.339	0.036	0.025	0.142
Annual Avg. Daily Traffic	0.138	0.022	0.000	0.122	0.029	0.000	0.192	0.026	0.000
% Minority Race	0.651	0.121	0.000	0.635	0.244	0.010	0.725	0.086	0.000
% Walk * Bike to Work	3.162	0.524	0.924	-132.011	231.723	0.569	-4.805	37.426	0.898

Although we would stop short of calling our findings “causal,” we believe there is evidence to suggest that higher casualty census tracts are associated with “exurban” areas or “commuter towns” which are industrial in nature but still home to thousands of residents. Exurban census tracts are less dense in population, have higher speed limits, less street signage, and potentially lower quality infrastructure. Vision Zero should pay special attention to these areas with regards to engineering, enforcement, and education.

Implications & Recommendations

Ethical Considerations

Since there is potential to impact protected populations and minority groups with our analysis, we must discuss ethical concerns with deployment. Ample research suggests that people of color are more likely to be targeted by the police than other races.[59] As we wrote in our "Data, Privacy, and Ethics" Capstone paper, "If a city dedicates stronger law enforcement to "high risk" census tracts, it could exacerbate racial profiling and unfair ticketing, which could have unintended consequences. If indeed Vision Zero amps up police action in certain neighborhoods, would the perceived benefits of safety be worth the potential harm of any mistrust and animosity that propagates because of it?"[60]

We wish to be mindful that African American and immigrant communities have historically been marginalized and under-resourced, and as such, have not had a seat at the table when it comes to transportation design and investment decisions.[61] We believe cities should not unilaterally implement interventions to the highest ranked census tracts, but rather engage the local community in conversation and debate. It's worth noting that Vision Zero is highly cognizant of these issues and has made "equity" part and parcel to their initiative.[62]

When it comes to our deployment, therefore, we believe our ranking model should help cities better understand the communities with whom they wish to collaborate. For example, in D.C., 5 of the top 10 census tracts in terms of casualties per person had populations comprised of more than 70% minorities. Rather than intensify enforcement here, the city could involve the community in designing solutions collaboratively.

Deployment

Many municipalities have sought to reduce traffic-related casualties for decades. Unfortunately, the problem will only grow in complexity as road infrastructure ages and city budgets tighten. Our deployment aims to contribute one major piece to a much larger puzzle by helping cities identify where their limited resources should best be used in order to make the largest impact on traffic safety. We return to the primary and secondary objective of this work.

Where Does a City New to Vision Zero Allocate Resources?

The data engineering, analysis, modeling, and visualization results from our project would primarily target Vision Zero members and municipal policy makers who wish to focus on proactive interventions in small geographic areas and communities. While collision data can be problematic in terms of quality, every U.S. city has Census data at their disposal, presumably collected with the same standards.[63] They can use our models to understand how accurate their actual ranking might be compared to our predicted ranking, and learn from the important features driving the results.

It should be noted that our project scope focuses on where interventions should be deployed given limited resources, and not the interventions themselves. Once cities have decided where they should focus efforts, they have a wide range of potential solutions to choose from, including traffic

cameras, bike lanes, and road improvements to name a few. However, the efficacy and subsequent recommendation of such interventions represent a significant body of work, one that is separate from our project scope. Put simply, our efforts are to identify where cities should focus efforts, and then allow experts to decide specifically what should be done as a result.

How Can Vision Zero Cities Better Learn From Each Other?

When we presented our findings to a Director at Vision Zero, and told them about our hypotheses, we learned that the biggest problems occur where there is competing land usage — that is, semi trucks may be coming in and out of the city, while residents are trying to go to school, work, or shopping centers.[64] As a path for future research, we believe data can help discover which parts of corridors and intersections have the highest "competition" over land use. These places could either be transformed to be safer for pedestrians and cyclists, or blocked off from them entirely, so as to mitigate risk. We believe Vision Zero cities can use our socio-economic insights to inform policy conversations in how to manage land usage wisely.

Conclusion

Our Capstone comes at a time when political and social pressure for traffic safety is mounting. There could be an opportunity for us to act as consultants to municipal and city officials, helping them to extract insights from data to save lives. In that sense, this project may be just the beginning of traffic analytics services. We might imagine using A.I. to process images of Google Street View to determine street safety risk, or in exploring telematics[65] datasets to understand individual driver risk. Regardless of the analytics possibilities, however, we must remember the human costs at stake.

On April 26th, 2019 - - two days after the vigil on V and 16th Street — hundreds of protesters marched down Pennsylvania Avenue in Washington D.C., bringing traffic to a halt.[66] People carried signs above their head that read "Safe Streets for All" and "Broken Promises Cost Lives." Some activists physically lay on the ground in front of city hall, their bikes flat on the ground by their sides. Over 800 letters were sent to Mayor Bowser's administration, according to one activist, urging investments in safer road infrastructure. During the demonstration, Lucinda Babers, deputy mayor for operations and infrastructure, stepped out to address the crowd: "We understand you, and we will take bold action to work towards Vision Zero. We are going to do better. We can, we should, and we will."[67]

Appendix

Exhibit 1: Summary review of related research

Geographic Focus	Variables	Unit of Analysis	Model(s) & Methodology
Manhattan, NY, U.S.[68]	**Dependent Variables** - Pedestrian injury crash costs. **Independent Variables** - *Socio-demographic*: population, race, population under 14, population over 65, median age, median income, employment status, and gender. *Road Characteristics*: traffic volumes, VMT, truck ratio, length sidewalk, length bike path, bust stop density. *Land Use*: commercial ratio, residential ratio, and mixed use park ratio.	Grid Cells	Tobit Model
Florida, U.S.[69]	**Dependent Variables** - Crash frequency. **Independent Variables** - *Socio-demographic*: population density, families without vehicle, school enrollment, urbanization, employment status, commute type (cycle, walk or public), % families without vehicle, log population density, log number of commuters by public transportation, and log number of commuters by cycling. *Road Characteristics*: VMT, % heavy vehicle mileage in VMT, length of road, number signalized intersections, length bike lanes, length sidewalks, and log signalized intersection density. *Land Use*: Number of hotels/motels.	TAZ	1) Zero Inflated Negative Binomial (ZINB), 2) Hurdle Negative Binomial (HNB), and 3) Spatial Spillover
Angeles,CA U.S.[70]	**Dependent Variables** - Pedestrian crash frequency. **Independent Variables** - *Socio-demographic*: population density, age, gende, % Latino, % below poverty, education, household size, and vehicles per household. *Road Characteristics*: average annual daily traffic (AADT), and road density. *Land Use*: % educational, % public, % industrial, % commercial, % low density residential, and % medium & high density residential.	census tract	Negative Binomial
Orange County, CA, U.S.[71]	**Dependent Variables** - Pedestrian injuries. **Independent Variables** - *Socio-demographic*: age, poverty, education attainment, speak other language at home, and population density.	census tract	Negative Binomial
Buffalo, NY U.S.[72]	**Dependent Variables** - Bicycle and pedestrian crash frequencies. **Independent Variables** - *Socio-demographic*: population density, education attainment, race, income, % pedestrians, % cyclists, mean travel time, poverty, age, and mean travel time. *Road Characteristics*: intersection to street ratio and signalized intersections. *Land Use*: business density, retail density, universities, and schools.	census tract	Ordinary Least Squares
U.S. (48 States)[73]	**Dependent Variables** - Classifying high vs. low risk areas based on fatal accidents. **Independent Variables** - *Road Characteristics*: speed limit, light conditions, surface conditions, and collision type. *Other*: weather and EMS arrival time.	U.S. State	1) Apriori Algorithm, 2) Naive Bayes Classification Model, and 3) K-Means Cluster Model
New York, NY, US[74]	**Dependent Variables** - High frequency crash intersections. **Independent Variables** - *Socio-demographic*: population, age, income, commuting patterns employment, and commute. *Road Characteristics*: sidewalks, roadbeds, truck routes, planimetrics, Citi bikes, road networks, traffic signals, traffic counts, crash records, and taxi/limousine data. *Land Use*: % retail, % office, and % garage. *Other*: 311 calls, turnstile records, and parking violations.	Traffic intersections and circular buffers	1) Random Forest and 2) Logistic Regression
NYC, NY U.S.[75]	**Dependent Variables** - Pedestrian Crash frequencies. **Independent Variables** - *Socio-demographic*: population, race, median	census tract	Random Parameter Negative Binomial

	age, and education attainment. *Road Characteristics*: number of all way stop, intersections, number of signalized intersections, road length, number of subway stations, and number of bus stops. *Land Use*: Industrial, commercial, park, and schools.		
Melbourne, Australia[76]	**Dependent Variables** - Pedestrian and cyclist crash frequencies. **Independent Variables** - *Socio-demographic*: population density, % of commuters cycling or walking to work, % of households without motor vehicles, and young population elderly population. *Road Characteristics*: VKT, bike crashes, and pedestrian crashes. *Land Use*: % residential area, % industrial area, % commercial area, and land use mix.	SA1, SA2 (statistical areas)	1) Random parameter negative binomial (RPNB), 2) Non-spatial Negative Binomial (NB) model, and 3) Poisson-Gamma-CAR model
Chicago, IL U.S.[77]	**Dependent Variables** - Crash frequency. **Independent Variables** - *Road Characteristics*: average annual traffic volume, pavement surface, and work zones. *Other*: weather, lighting, and time of day.	Traffic intersections	Random Parameter Poisson
Japan[78]	**Dependent Variables** - Injury and death frequency. **Independent Variables** - *Socio-demographic*: total population, age, income, unemployment, degree of urbanization, alcohol consumption, and college entrance rate. *Road Characteristics*: length of road. *Other*: registered vehicles, licensed drivers, number of physicians, and number of patients transferred by ambulance.	Prefectures	Stepwise Multivariate regression

Exhibit 2: Original select Census variables

Census Groups	Variables Selected as Features[79]	
Demographic (12)	• Population • Male • Female • White • Black • Native American • Asians	• Hawaiian & Pacific Islander • Hispanic or Latino • Other • Two or More Races • Not US Citizen
Age (12)	• Age 5 or Under • Age 5 Through 17 • Age 18 Through 24 • Age 25 Through 34 • Age 35 Through 44 • Age 45 Through 54	• Age 55 Through 59 • Age 60 Through 61 • Age 62 Through 64 • Age 65 Through 74 • Age 75 and Above • Median Age
Marital Status & Family Structure (9)	• Never Married • Married • Divorced • Separated • Widowed	• Unmarried Household • Family Household • Non-Family Household • Non-Family Group Quarters
Socioeconomic (10)	• Median Earnings • Household Income • Receiving Social Benefits • Unemployed • Below Poverty Line	• Income-to-Poverty Ratio (IR) Below 1.0 • IR Between 1.0 to 1.9 • IR Above 2.0 • Does Not Own a Car • Gini Equality Index
Means of Transportations To & From Work (9)	• By Cars • By Car Alone • By Carpool • By Public Transit • By Taxi	• By Motorcycle • By Bicycle • By Walking • Working From Home

Education (8)	• No Formal Education • Some High School • High School Diploma • GED	• Some College • Undergraduate Degree • Graduate Degree • Doctorate Degree

Exhibit 3: Post-engineered Census variables[80]

Census Groups	Variables Selected as Features	
Demographic (5)	• Population Density • Female • White	• Minority • Not US Citizen
Age (6)	• Gen Z • Millennial • Gen X	• Boomer • Retiree • Median Age
Marital Status & Family Structure (2)	• Divorced or Separated	• Widowed
Socioeconomic (3)	• Median Earnings • Unemployed	• Below Poverty Line
Means of Transportations To & From Work (6)	• By Cars • By Transportation • By Motorcycle	• By Bicycle • By Walking • Working From Home
Education (4)	• Low Education • High School / GED	• Undergraduate Degree • Graduate Degree or Higher

Exhibit 4: Ranking model algorithm hyperparameter settings

K-Nearest Neighbor	Random Forest	XGBoost
• K = 6 • Number of folds for cross validation: 5	• N decision trees: 500 • Max N nodes per tree: 5 • Max % of features used: 50% • K-folds for cross validation: 5	• Learning rate: 0.08 • N decision trees: 500 • Max N nodes per tree: 5 • Max % of features used: 50% • K-folds for cross validation: 5

Exhibit 5: Correlation matrix

	Log Casualties Per Pop.	Log Popularion Density	Median Age	Median Earnings	% Car Commuters	% Pub. Trans. Commuters	Max Speed Limit	Max Lanes	Log Annual Avg. Daily Traffic	% Minority Race	% Walk * Bike to Work
Log Casualties Per Pop.	1.000										
Log Popularion Density	-0.595	1.000									
Median Age	-0.023	-0.236	1.000								
Median Earnings	0.056	-0.077	0.357	1.000							
% Car Commuters	0.033	-0.553	0.398	-0.044	1.000						
% Pub. Trans. Commuters	-0.006	0.317	-0.100	0.255	-0.552	1.000					
Max Speed Limit	0.292	-0.337	0.079	0.065	0.125	-0.109	1.000				
Max Lanes	0.247	-0.234	0.100	0.136	0.033	-0.022	0.686	1.000			
Log Annual Avg. Daily Traffic	0.341	-0.246	0.039	0.102	0.024	-0.056	0.732	0.557	1.000		
% Minority Race	0.069	0.117	-0.204	-0.629	-0.044	-0.028	-0.056	-0.057	-0.107	1.000	
% Walk * Bike to Work	0.107	0.033	-0.028	0.342	-0.233	0.178	0.006	0.009	0.038	-0.212	1.000

Endnotes

1. Trottenberg, Polly. “The New York City Model for Vision Zero Progress.” Medium, November 13, 2018. https://medium.com/vision-zero-cities-journal/the-new-york-city-model-for-vision-zero-progress-8353a45cb76b
2. Jordan, George Kevin. “A community mourns Abdul Seck and demands safer roads east of the Anacostia.” Greater Greater Washington, April 25, 2019. .https://ggwash.org/view/71883/a-community-mourns-abdul-seck-and-demands-safer-roads-east-of-the-anacostia
3. Metropolitan Police Department. “Traffic Fatality: Intersection of 16th Street and V Street, Southeast..” DC.Gov. April 22, 2019. https://mpdc.dc.gov/release/traffic-fatality-intersection-16th-street-and-v-street-southeast
4. Jordan, https://ggwash.org/view/71883/a-community-mourns-abdul-seck-and-demands-safer-roads-east-of-the-anacostia
5. Kurzius, Rachel. “As Changes Come To Intersection Where Pedestrian Was Killed, Community Raises Funds For Burial.” DCist, April 25, 2019.https://dcist.com/story/19/04/25/as-changes-come-to-dangerous-intersection-where-pedestrian-was-killed-his-family-struggles-to-bury-him/
6. “V St and 16th St SE D.C.”, Google Maps, https://www.google.com/maps/place/V+St+SE+%26+16th+St+SE,+Washington,+DC+20020,+USA/@38.8654726,-76.9817121,125a,35y,196.07h,45t/data=!3m1!1e3!4m5!3m4!1s0x89b7b9b94dbb2a4d:0x7e2360fac8e69b12!8m2!3d38.8644743!4d-76.9824308
7. Jordan, https://ggwash.org/view/71883/a-community-mourns-abdul-seck-and-demands-safer-roads-east-of-the-anacostia
8. Ryan Beene, “Traffic Deaths in U.S. Exceed 40,000 for Third Straight Year.” Bloomberg, February 13, 2019. https://www.bloomberg.com/news/articles/2019-02-13/traffic-deaths-in-u-s-exceed-40-000-for-third-straight-year ;
9. Collins, Sam PK. “Car Accidents Send 2.5 Million Americans To The Emergency Room Every Year.” Think Progress, October 10, 2014. https://thinkprogress.org/car-accidents-send-2-5-million-americans-to-the-emergency-room-every-year-b81b191a09b8/ ; https://www.nsc.org/road-safety/safety-topics/fatality-estimates
10. Shahum, Leah. “Safe Streets: Insights on Vision Zero Policies from European Cities.” The German Marshall Fund of the United States 27, October 20, 2017: 6. http://www.gmfus.org/publications/safe-streets-insights-vision-zero-policies-european-cities
11. https://visionzeronetwork.org/ ; Leber, Jessica. “U.S. Cities Want To Totally End Traffic Deaths–But There Have Been A Few Speed Bumps.”Fast Company, August 23, 2016. https://www.fastcompany.com/3062492/us-cities-want-to-totally-end-traffic-deaths-but-there-have-been-a-few-speed-bumps
12. Shahum, pp 6. http://www.gmfus.org/publications/safe-streets-insights-vision-zero-policies-european-cities
13. Trottenberg, Polly. “The New York City Model for Vision Zero Progress.” Medium, November 13, 2018. https://medium.com/vision-zero-cities-journal/the-new-york-city-model-for-vision-zero-progress-8353a45cb76b
14. Hu, Winnie. “No Longer New York City’s ‘Boulevard of Death.’” New York Times, December 3, 2017. https://www.nytimes.com/2017/12/03/nyregion/queens-boulevard-of-death.html
15. Ibid
16. Ibid
17. The Official Website of the City of New York. New York City Mayor’s Office of Operations. “Vision Zero Year Four Report.” March 2018. https://www1.nyc.gov/assets/visionzero/downloads/pdf/vision-zero-year-4-report.pdf
18. Fitzsimmons, Emma G. “Traffic Deaths in New York City Drop to 200, a Record Low.” New York Times,
19. Trottenberg, Polly. “The New York City Model for Vision Zero Progress.”
20. Walker, Alissa, “City Council votes to dedicate $27M to Vision Zero, its plan to end traffic deaths.” Los Angeles Curbed, May 18, 2017. https://la.curbed.com/2017/5/18/15660082/vision-zero-walking-biking-budget-bonin ; Rogers, Jonathan M. “The Story for Vision Zero in Washington D.C.” Metropolitan Washington Council of Governments, District Department of Transportation, Vision Zero Washington DC. pp 24.
21. Laker, Laura. “Vision Zero: has the drive to eliminate road deaths lost its way?” The Guardian, September 17, 2018. https://www.theguardian.com/cities/2018/sep/17/vision-zero-has-the-drive-to-eliminate-road-deaths-lost-its-way
22. Badar, Inshaal. “Vision Zero: The Road Safety Movement Taking Over the World.” Geotab, September 26, 2018. https://www.geotab.com/blog/vision-zero/; Small, Andrew. “Dangerous Streets? Take the Bus.” Citylab, September 11, 2018.
23. “Identification of High Pedestrian Crash Locations.” US Department of Transportation Federal Highway Administration, No No. FHWA-HRT-17-107 (March 2018): 28-32. https://www.fhwa.dot.gov/publications/research/safety/17107/17107.pdf
24. Bliss, Laura. “The Incredibly Cheap Street Fix That Saves Lives.” Citylab, January 26, 2018. https://www.citylab.com/transportation/2018/01/the-incredibly-cheap-street-fix-that-saves-lives/551498/ ; Kuntzman, Gersh. “First Death on Queens Boulevard Since 2015 Vision Zero Fix.” Streets Blog NYC, December 17, 2018. https://nyc.streetsblog.org/2018/12/17/first-death-on-queens-boulevard-since-2015-vision-zero-fix/
25. State of Place. “YOUR ABSOLUTELY ESSENTIAL EVERYTHING YOU NEED TO KNOW ABOUT VISION ZERO GUIDE!” July 10, 2018. http://www.stateofplace.co/our-blog/2018/7/ultimate-vision-zero-guide
26. Shahum, pp 20. http://www.gmfus.org/publications/safe-streets-insights-vision-zero-policies-european-cities ; Hinds,Kate. “The NYPD’s Crash Data is Bad and There’s Not Enough of It.” WNYC, October 10, 2013.

27. Imprialou, Marianna and Quddus, Mohammed. “Crash data quality for road safety research: current state and future directions.” Transport Studies Group, School of Civil and Building Engineering, Loughborough University, February 22, 2017. Pp2.
https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/24464/1/Imprialou%20%26%20Quddus%20Crash%20data%20for%20road%20safety%20research_current%20state%20and%20future%20directions.pdf
28. Ibid.
29. Kite, Julia, Director of Strategic Initiatives. Vision Zero New York City. Phone interview April 15, 2019, 3pm EST.
30. Miller, Stephen. “NYPD Crash Data Now Easier to Use and Updated Daily.” Streets Blog NYC, May 7, 2014.
31. “NYPD Motor Vehicle Collisions.” NYC Open Data.
https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95
32. Xie,Kun, Ozbay,Kaan, Kurkcu,Abdullah, Yang,Hong . “Analysis of Traffic Crashes Involving Pedestrians Using Big Data: Investigation of Contributing Factors and Identification of Hotspots.” (2017)
33. Chakravarthy,Bharath, Anderson,Craig L., Ludlow,John, Lotfipour,Shahram, Vaca,Federico E.. “The Relationship of Pedestrian Injuries to Socioeconomic Characteristics in a Large Southern California County.”(2010).
34. Ukkusuri,Satish, Hasan,Samiul, Aziz, H. M. Abdul. “Random Parameter Model Used to Explain Effects of Built-Environment Characteristics on Pedestrian Crash Frequency”. (2011); Cahill Delmelle,Elizabeth, Thill,Jean-Claude F., Ha,Hoehun. “Spatial epidemiologic analysis of relative collision risk factors among urban bicyclists and pedestrians.” (2011).
35. Xie, Ozbay, Kurkcu,Yang. (2017); Chakravarthy, Anderson, Ludlow, Lotfipour, Vaca. (2010); Nagata,Takashi, Takamori,Ayako, Berg,Hans-Yngve, Hasselberg,Marie. “Comparing the impact of socio-demographic factors associated with traffic injury among older road users and the general population in Japan.” (2012).
36. Xie,Ozbay, Kurkcu,Yang. (2017); Cai,Qing, Lee,Jaeyoung, Eluru,Naveen, Abdel-Aty,Mohamed. “Macro-level Pedestrian and Bicycle Crash Analysis: Incorporating Spatial Spillover Effects in Dual State Count Models.” (2016); Amoh-Gyimaha,Richard, Saberia,Meead, Sarviba,Majid. “Macroscopic Modeling of Pedestrian and bicycle crashes: A cross comparison of estimation methods” (2016).
37. Roshandeh,Arash M., Agbelie,Bismark R.D.K., Lee,Yongdoo. “Statistical modeling of total crash frequency at highway intersections.” (2016); Loukaitou-Sideris,Anastasia,Liggett,Robin, Sung,Hyun-Gun. “Death on the Crosswalk: A Study of Pedestrian-Automobile Collisions in Los Angeles.” (2005);Akred,Erin, Dowd, Michael, Weiser,Jackie, Kashuk,Sina . “Analyzing Traffic Crashes in New York City Using Machine Learning Based Approaches.” (2017).
38. Xie, Ozbay, Kurkcu,Yang. (2017); Ukkusuri, Hasan, Aziz. (2011); Loukaitou-Sideris, Liggett, Sung. (2005);
39. Xie,Ozbay, Kurkcu, Yang. (2017); Ukkusuri, Hasan, Aziz. (2011); International Transport Forum. “Why Does Road Safety Improve When Economic Times Are Hard?” OECD, (2015): pp 31.
https://www.itf-oecd.org/sites/default/files/docs/15irtadeconomictimes.pdf
40. Cahill Delmelle, Thill,J, Ha. (2011).
41. Nagata,Ayako, Hans-Yngve, Marie. (2012); Ukkusuri,Hasan,Aziz.(2011);Akred, Dowd,Weiser, Kashuk. (2017); Cai,Lee,Eluru,Abdel-Aty. (2016); Xie,Ozbay, Kurkcu,Yang. (2017).
42. “Identification of High Pedestrian Crash Locations.” US Department of Transportation Federal Highway Administration, No No. FHWA-HRT-17-107 (March 2018): 28-32. https://www.fhwa.dot.gov/publications/research/safety/17107/17107.pdf
43. Cai,Qing, Abdel-Aty,Mohamed, Lee,Jaeyoung, Eluru,Naveen “Comparative analysis of zonal systems for macro-level crash modeling.” Journal of Safety Research No 61 (2017): 157.
https://drive.google.com/file/d/1LqnMSb47nu7y0pKSiK45HudtjfidQou4/view
44. United States Census Bureau. “Understanding and Using American Community Survey Data.” US Department of Commerce, July 2018. pp 12
https://www.census.gov/content/dam/Census/library/publications/2018/acs/acs_general_handbook_2018.pdf
45. NYC Open Data: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95). Los Angeles Open Data: http://visionzero.geohub.lacity.org/datasets/ladot::los-angeles-collisions-2013through2018. DC Open Data: http://opendata.dc.gov/datasets/crashes-in-dc)
46. United States Census Bureau. “Understanding and Using American Community Survey Data.” pp 1-5
47. New York State. Department of Transportation. https://www.dot.ny.gov/highway-data-services
48. Human Resources & Services Administration. “Defining Rural Population.“ December 2018.
https://www.hrsa.gov/rural-health/about-us/definition/index.html
49. Provost, Foster, and Tom Fawcett. *Data Science for Business: What You Need To Know About Data Mining and Data-Analytic Thinking*. " O'Reilly Media, Inc.", 2013.
50. We use the StatsModels API, an open source Python library for statistical modeling and analysis. Detailed documentations can be found here: https://www.statsmodels.org
51. Lord,Dominique, Mannering, Fred. “The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives.” pp14. (2010).
52. We use the Scikit-Learn API, an open source Python library for general purpose machine learning modeling, for both KNN and RF algorithms. Detailed documentations can be found here: https://scikit-learn.org/
53. “Chapter 7. Ensemble Learning and Random Forest.” Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tool, and Techniques to Build Intelligent Systems, by Aurelien Geron, O’Reilly, 2018.
54. We use the XGBoost API, an open source Python library specialized for using the XGBoost algorithm. Detailed documentations can be found here: https://xgboost.readthedocs.io
55. Exhibit 4 lists the hyperparameters set for RF and XGBoost, such as number of trees, maximum depths, and maximum percentage of features used per tree, etc.
56. Population density is defined as population per square-mile in a census tract.
57. Lord and Mannering, pp 14.

58. LaScala, Elizabeth A., Daniel Gerber, and Paul J. Gruenewald. "Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis." Accident Analysis & Prevention 32, no. 5 (2000): 651-658; Cahill Delmelle,Elizabeth, Thill,Jean-Claude F., Ha,Hoehun. "Spatial epidemiologic analysis of relative collision risk factors among urban bicyclists and pedestrians." (2011).
59. Bliss,Laura. "Vision Zero's Troubling Blind Spot." City Lab, September 1, 2016. https://www.citylab.com/transportation/2016/09/black-lives-matter-and-vision-zero/497495/ ; Fox,Jenn, Shahum,Leah. "VISION ZERO EQUITY STRATEGIES for Practitioners." http://visionzeronetwork.org/wp-content/uploads/2017/05/VisionZero_Equity.pdf
60. Branham,Noah, Huang,David, Kempf,Tamara, Marra,Michael, Smith,Aaron. "Data, Privacy, & Ethics - Capstone Paper." Data, Privacy & Ethics course, NYU Stern, 2019.
61. Bliss, https://www.citylab.com/transportation/2016/09/black-lives-matter-and-vision-zero/497495/
62. Fox, Shahum, http://visionzeronetwork.org/wp-content/uploads/2017/05/VisionZero_Equity.pdf
63. We want to be clear that Census estimates of income, education, population, and so on are indeed approximations, and they should not be treated as an absolute reflection of reality. There are different kinds of potential bias in Census data, including sampling and response bias, that could exist, but they are beyond the scope of this paper.
64. Kite. Phone Interview April 15, 2019, 3pm EST.
65. A method for monitoring a vehicle with a GPS system and on-board diagnostics.
66. Lazo, Luz. 'It's absolutely unacceptable': Protesters demand safer streets in District." The Washington Post. April 26, 2019. https://www.washingtonpost.com/local/trafficandcommuting/its-absolutely-unacceptable-protesters-demand-safer-streets-in-dc/2019/04/26/e6f91ce2-652f-11e9-a1b6-b29b90efa879_story.html
67. Ibid.
68. Xie,Kun, Ozbay,Kaan, Kurkcu,Abdullah, Yang,Hong. "Analysis of Traffic Crashes Involving Pedestrians Using Big Data: Investigation of Contributing Factors and Identification of Hotspots." (2017).
69. Cai,Qing, Lee,Jaeyoung, Eluru,Naveen, Abdel-Aty,Mohamed . "Macro-level Pedestrian and Bicycle Crash Analysis: Incorporating Spatial Spillover Effects in Dual State Count Models." (2016).
70. Loukaitou-Sideris,Anastasia, Liggett,Robin, Sung,Hyun-Gun. "Death on the Crosswalk: A Study of Pedestrian-Automobile Collisions in Los Angeles." (2005).
71. Chakravarthy,Bharath, Anderson,Craig L., Ludlow,John, Lotfipour,Shahram, Vaca,Federico E. "The Relationship of Pedestrian Injuries to Socioeconomic Characteristics in a Large Southern California County."(2010).
72. Cahill Delmelle,Elizabeth, Thill,Jean-Claude F., Ha,Hoehun. "Spatial epidemiologic analysis of relative collision risk factors among urban bicyclists and pedestrians." (2011).
73. Li,Liling, Shrestha,Sharad, Hu,Gongzhu."Analysis of road traffic fatal accidents using data mining techniques." (2017)
74. Akred,Erin, Dowd,Michael, Weiser,Jackie, Kashuk,Sina. "Analyzing Traffic Crashes in New York City Using Machine Learning Based Approaches." (2017).
75. Ukkusuri,Satish, Hasan,Samiul, Aziz,H. M. Abdul. "Random Parameter Model Used to Explain Effects of Built-Environment Characteristics on Pedestrian Crash Frequency". (2011).
76. Amoh-Gyimaha,Richard, Saberia,Meead, Sarviba,Majid. "Macroscopic Modeling of Pedestrian and bicycle crashes: A cross comparison of estimation methods" (2016).
77. Roshandeh,Arash M., Agbelie,Bismark R.D.K., Lee,Yongdoo. "Statistical modeling of total crash frequency at highway intersections." (2016).
78. Nagata,Takashi, Takamori,Ayako, Berg,Hans-Yngve, Hasselberg,Marie. "Comparing the impact of socio-demographic factors associated with traffic injury among older road users and the general population in Japan." (2012).
79. With the exception of Median Age, Median Earnings, Household Income, and Gini Equality Index, all other variables are in the form of count.
80. The 60 original features from the raw Census data has been reduced to 26 through combining similar features and removing potentially redundant ones.