CUNY MSDS DATA 698 Capstone Project

CUNY SCHOOL OF PROFESSIONAL STUDIES

Good Roads

Using machine learning to measure how socio and non-socio-economic factors differently influence traffic accident injury rates

---

By:

PRIYA SHAJI

Table of Contents

Abstract

This project is based on analyses of socio-economic factors in conjunction with traditional traffic accident, collision and injury data to analyze if they play a meaningful role in predicting motor vehicle-related injury rates and thus be used to help improve traffic safety. Road and census variables are analyzed for socio economic factors. Factual reasons for accidents and vehicle types are analyzed for non-socio-economic factors. Project is divided in two parts consisting of socio and non-socio factors which includes data exploration, prediction modeling and feature extraction. Total casualties per person is taken as target variable and remaining factors as explanatory variables. For socio economic part, two prediction models are created in which prediction curves are compared to perfect ranking curve. Selected model is then tested on LA and DC datasets followed by feature extraction using XGBoost and random forest. Population density and road length are top reasons for total casualties. Relationship between important features and total casualties is analyzed using OLS regression. For non-socio-economic part, two models are considered, from which random forest has a better normally distributed residual vs fitted curve and hence was selected as the final model. Driver inattention and following too closely are the top causes of vehicular accidents.

Introduction

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. However, cities with limited budgets, inadequate data, and constrained resources need help prioritizing focus areas and gathering insights that can lead to effective interventions.

Road traffic injuries must be counted among the less equal compared to other health crisis. Their global health burden of causing more than 1.2 million deaths per year, corresponding to 2.4% of all global deaths, establishes them as a major global health threat, in a similar league with tuberculosis(1).

The combined effort of the global community towards funding road safety is roughly estimated to be between US$10–25 million per year(2), a fraction of the sums spent on other public health issues of comparable significance. International health organizations have never attempted eradication, have invested only few research funds, in fact, road safety has only recently become issue for world health organizations (3).

My objective for project is to use socio and non-socio-economic factors to determine how differently they influence road traffic injuries.

This project can serve as a starter for revisiting and revising road safety measures taken to achieve lower road casualty rates. This analysis can also be applied to other cities in US to ensure that we are taking every factor to achieve a better road safety model.

Literature Review

In the course of the twentieth century road traffic injuries (RTIs) became a major public health burden. RTI deaths first increased in high-income countries and declined after the 1970s, and they soared in low- and middle-income countries from the 1980s onwards (4).

As motorization took off in North America and then spread to Europe and to the rest of the world, discussions on RTIs have reflected and influenced international interpretations of the costs and benefits of development, as conventionally understood.

This project would seek to help cities to get started with road traffic safety measures considering both socio and non-socio-economic factors using data science approach.

- By using predictive analytics to assist cities without robust collision data and to understand socio-economic factors, example: population density, median age etc. and non-socio-economic factors, example: turning improperly, following too closely etc. which can help them to invest in those areas.
    - Example: If there are socio-economic features in parts of New York city that are similar to those in say, New Jersey, 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.

- With this ranking of socio and non-socio-economic feature consideration, city officials can make more informed decisions about where to allocate resources and learn about the in formativeness of the variables that produce it.

Hypothesis

Do socio-economic factors – in conjunction with traditional traffic accident, collision and injury data – play a meaningful role in predicting motor vehicle-related injury rates and thus be used to help improve traffic safety?

Data and Variables

Data Source and Data Acquisition

1) For first part of the project where socio-economic factors contributing to road accidents and casualties will be taken into consideration, I have used datasets which are aggregate of census tract and road characteristic variables from the following source: GitHub Dataset

- Currently, I will be focusing on the NYC datasets i.e. NYC census variables and NYC road characteristic datasets for training and I will test the results on LA and DC dataset that is provided in the same GitHub repository.

- 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.

2) For second part of the project where non-socio-economic factors contributing to road accidents and casualties will be taken into consideration, I have used NYPD motor vehicle collision summary dataset which is available in NYC open data portal.

- This dataset consists of details of motor vehicle collisions in New York City provided by the Police Department (NYPD).

- The historical data about vehicular collisions in NYC indicates that an average of 586 collisions happen in New York City every day. These incidents have increased the demand for emergency services in the city.

Data Variables

Socio economic factors

Two main datasets are considered under this category:

1) NYC census dataset

This dataset consists of 38 variables and 2114 rows.

Few variables that this dataset consists are as follows:

“GEOID”: Geo ID on basis of which all boroughs or locations in the dataset are identified

“pop dens”: Population density of the area

“race minority”: Number of people belonging to minority race.

“female”: Number of females

“widowed”: Number of People who are widowed

“median age”: Median age of the person

“not us citizen”: Number of people who are no US citizens

“median earnings”: Median earning of people

“Casualties Per Pop Dens”: Casualty per population density

2) NYC road characteristic dataset

This dataset consists of 53 variables and 2114 rows. Few variables of this dataset are same as NYC census dataset as described above

Few variables that this dataset consist of are as follows:

“Road maxlength”: Maximum length of the road

“Road maxlanes”: Maximum lanes on the road

“Road pavewidth”: Pavement width of the road

“Road maxspeed”: Maximum speed of the road

“Below pov”: Number of people below poverty line

Non-Socio-economic factors

In this category motor vehicle collision dataset has been used which covers most of the factual reasons for a crash.

It has 29 variables and 1720935 rows

Few variables that this dataset consist of are as follows:

“Latitude and Longitude”: Location where the crash took place

"Number of persons injured/killed": Two columns describing count of persons injured/killed

"Number of pedestrians injured/killed": Two columns describing count of pedestrians injured/killed

"Number of cyclist injured/killed": Two columns describing count of cyclist injured/killed

"Number of motorist injured/killed": Two columns describing count of motorist injured/killed

"Borough": Boroughs of NYC

"Contributing factor": There are 5 columns of factors which are identified as contributors for the accident

"Vehicle type": There are 5 columns for different vehicle types which are identified as contributors for the accident

Statistical Methods

The project consists of two parts. First part refers to how socio-economic factors affect road accidents and casualties.

Second part refers to how non-socio-economic factors affect road accidents and casualties.
Each part I will explain how data is prepared and cleaned for analysis followed by exploratory analysis methods used followed by feature extraction and also address hypothesis.

Part 1: Using solely socio-economic factors such as city's census and road characteristics variables and find models which can predict the factors for vehicular collisions better.

Data Preparation

Data manipulation, cleaning and visualizing

- Two datasets namely `census` and `collisions` are left joined by a common variable, `GEOID`, to create a master dataset which has both census and collision data variables. All NA's are imputed with mean of the column using `dplyr` package in R.
- This dataset is used as a basis for exploring and analyzing 4 potential target variables each of them grouped by city:

1. Injuries (excluding pedestrians)
2. Pedestrian Injuries
3. Collisions
4. Deaths

Exploratory Analysis and Data Visualization:

Let's visualize the few variables of socio-economic factors:

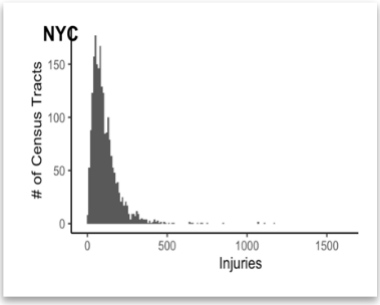

Fig 1.1

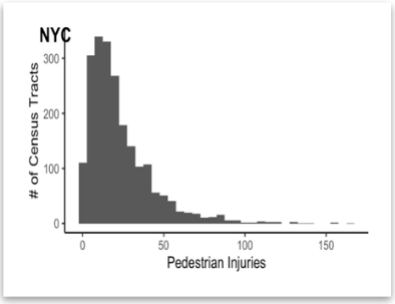

Fig 1.2

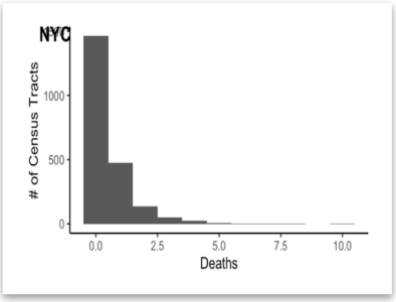

Fig 1.3

1. Injuries by number of census tracts (Fig1.1): In NYC around 50 census tracts have collision injuries in range from 1 to 300.

2. Pedestrian Injuries (Fig1.2): Pedestrians from more than 300 census tracts have collision injuries in range from 1 to 30.

3. Collisions: Collision per census tract is same a Fig 1 .1 where around 50 census tracts have collision injuries in range from 1 to 300.

4. Deaths (Fig1.3): Around 500 census tracts in NYC have collision which leads to deaths in range from 0.5 to 2

5. Race groups per census tracts grouped by city: Dataset has 5 race groups variables i.e. White, Black, Native, Asian, Hispanic. These groups count as per census tract is plotted in histograms.
   Around 1 to 1500 Hispanics live in 20 census tracts. 1 to 1000 Asians live in 40 census tracts. 1 to 1500 Blacks live in 40 census tracts, around 2000 Whites live in 10 census tracts.

6. Total Pedestrian Injuries per race grouped by city

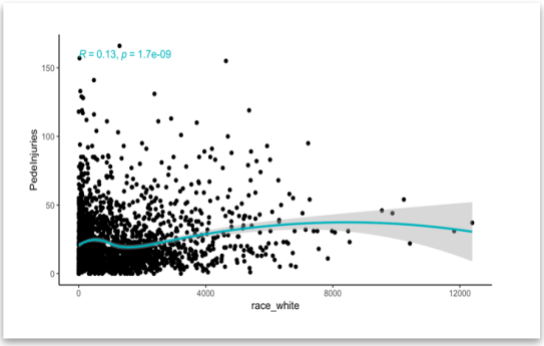

Fig 1.4

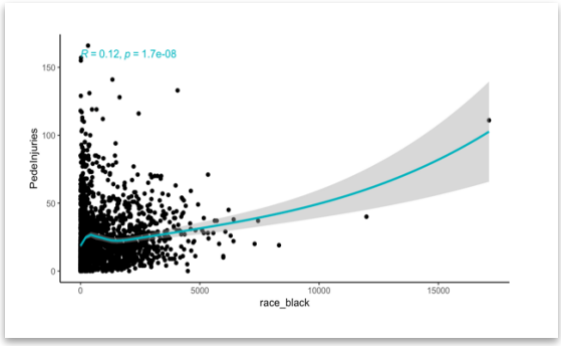

Fig 1.5

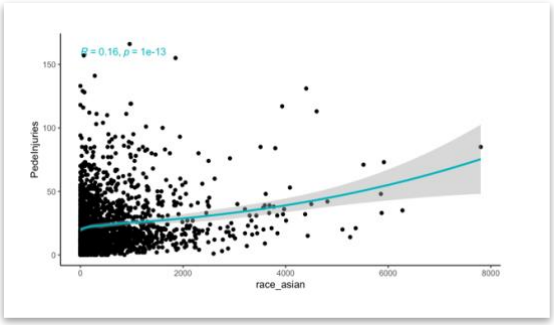

Fig 1.6

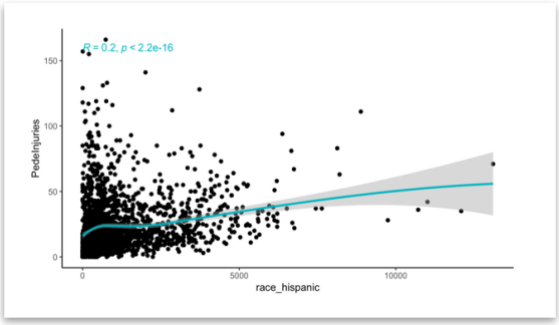

Fig 1.7

- According to the fitted models above, in Fig 1.4, we see that people who belong to white race have lesser case of pedestrian injuries, with most of the cases below 50.

- In Fig 1.5, we see that people who belong to black race have more cases of pedestrian injuries, cases ranging from 20 to 100.

- Similarly, for Asian race in Fig 1.6, more people have cases of pedestrian injury, cases ranging from 20 to 90. And in Hispanic, medium number of cases, around 10000 people have cases of pedestrian injuries, cases ranging from 20 to 50.

A generalized plot (Fig1.8, Fig1.9) which I created in Power BI platform is described below:

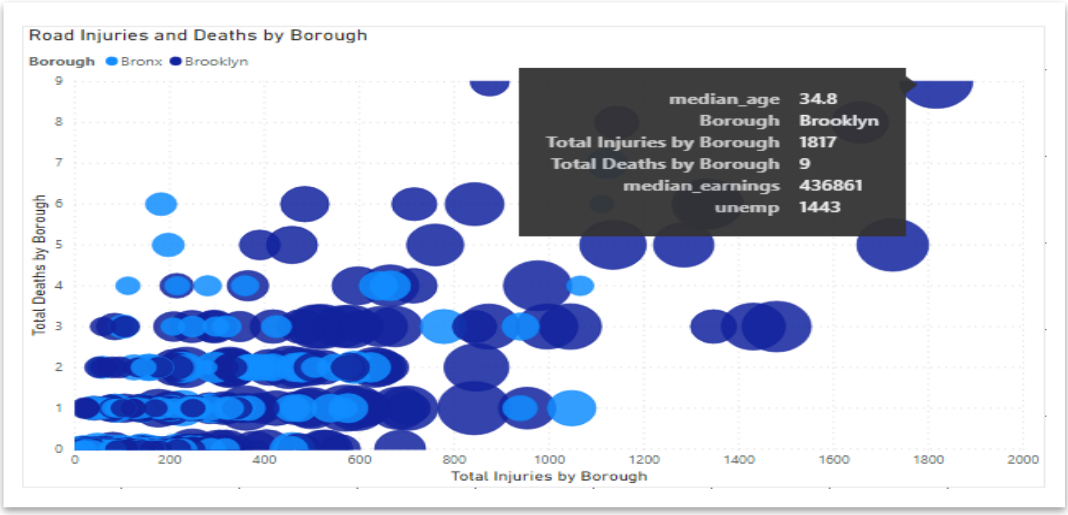

Fig 1.8

- The above visual (Fig1.8) shows “road injuries and deaths by borough”. Diameter or size of each bubble shows median salary of people whose median age is 34.8(selected bubble) who live in Brooklyn and are unemployed, have increase in number of injuries which leads to death.

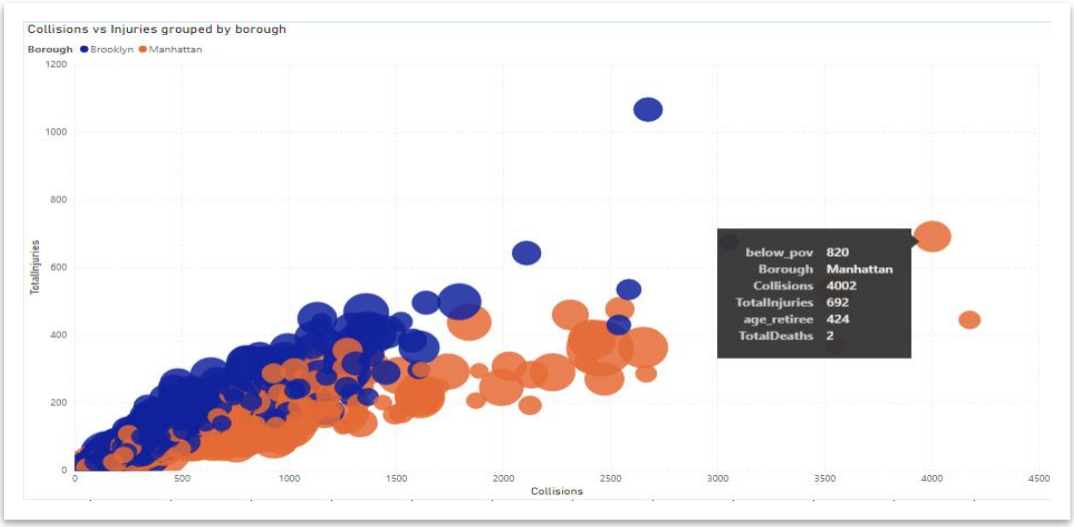

Fig 1.9

- The above visual shows (Fig1.9) shows "road injuries and collisions by borough". Diameter or size of each bubble shows number of people who are retired who live in Manhattan and are below poverty line, have increase in number of injuries in road collisions.

This analysis will be followed by binomial model, k-nearest neighbors, random forest, xgboost, also feature importance will be tabulated based on these models. Followed by ranking performance chart based on how well models performed compared to perfect ranking.

Part 2: Using solely non-socio-economic factors i.e. factual reasons for accidents such as alcohol involvement while driving, fatigue etc. and find models which can predict factors for vehicular collisions better.

Data Preparation

Data manipulation, cleaning and visualizing

- The dataset was downloaded from NYC open data. In columns , `number of people injured`/`killed` I removed values of 0 to get the required data for analyses because with data exploration, I found that there were a lot of entries where nobody was killed or injured but the collision was noted in the database.

- Thus, to remove these entries we perform data cleaning and obtained the refined data set which will be helpful for exploratory data analysis.

- First, I selected location variables and created subset with latitude, longitude and zip code. This dataset was then imputed using deterministic regression imputation via `mice` function. By this method all the NA's were imputed, means these imputed values were drawn from a distribution.

- Simulating random draws from a distribution does not include uncertainty in model parameters. Therefore, to perform data imputation for these variables I used MICE package. MICE package uses PMM which stands for Predictive Mean matching and performs linear regression matching algorithm to impute the data.

Exploratory Analysis and Data Visualization

Let's visualize the few variables of non-socio-economic factors:

1) Summarizing according to hour

- The timeline of injuries caused in NYC are plotted borough wise across the 5 boroughs of New York City

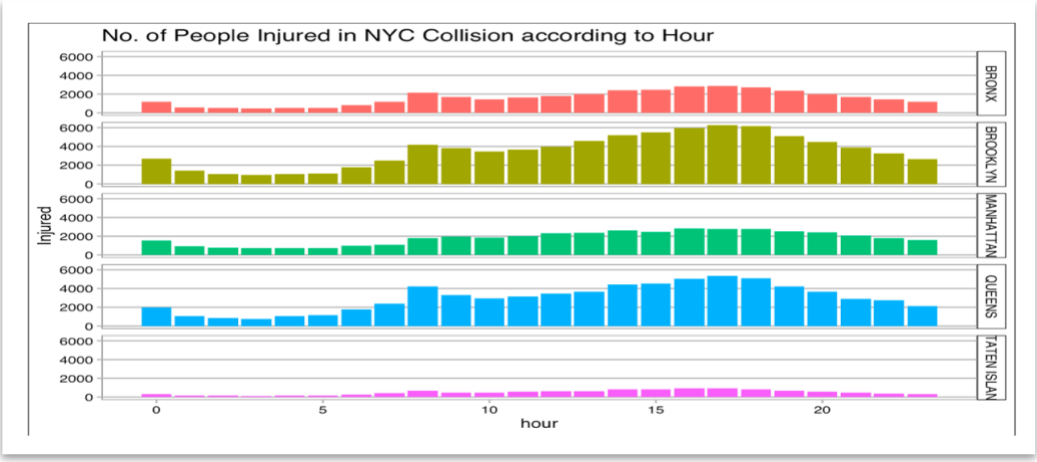

Fig 2.1

- The second and fourth section from top in plot above shows number of people injured between 15th and 20th hour of the day were highest in Brooklyn and Queens as well. Injured numbers are relatively less in Bronx followed by Staten Island. In general, it shows that maximum number of accidents occur during rush hours of 10AM to 5PM.

2) Summarizing according to reasons of accidents

- Using the five given column of contributing factor category, I tabulated (Fig2.2) top 10 reasons which caused accidents. I find that driver inattention, failure to yield the right way, following too closely where the top reasons while Alcohol involvement was 10th which is interesting because we usually are under impression that drink and drive cause maximum collisions.

	Reason	Total_accidents
1	Driver Inattention/Distraction	68828
2	Failure to Yield Right-of-Way	33633
3	Following Too Closely	19623
4	Traffic Control Disregarded	10886
5	Fatigued/Drowsy	8177
6	Other Vehicular	8139
7	Unsafe Speed	6241
8	Backing Unsafely	5889
9	Turning Improperly	5287
10	Alcohol Involvement	5108

Fig 2.2

3) Summarizing number of people injured according to vehicle type

- Using 5 given column of vehicle type category which were involved in the accidents, I tabulated (Fig2.3) the vehicles that caused maximum injuries which are shown in the following data table. I find that sedan, passenger vehicle, sports utility vehicle were involved in maximum injuries in a vehicular collisions.

	Vehical_type	Injured
1	Sedan	234110
2	PASSENGER VEHICLE	213705
3	Station Wagon/Sport Utility Vehicle	182126
4	Taxi	25271
5	Bike	16415
6	TAXI	16342
7	UNKNOWN	15298
8	Pick-up Truck	11075
9	VAN	10640
10	OTHER	8830

Fig 2.3

Prediction Modeling

In this section, I will analyze prediction models that performs well and select a model that can predict the crashes.

Part 1: Socio-economic factors

First, I created a ranking model that would help us to determine the extent to which census and road features are helpful to predict casualties. Then, I evaluate models' performance based on how well the predicted ranking stacks up to the actual ranking.

The regression algorithms I used are as follows:

1. Negative Binomial Regression (NB): Negative binomial regression is a generalization of Poisson regression which loosens the restrictive assumption that the variance is equal to the mean made by the Poisson model [5]. As crash data is non-negative count data, NB is the generalized linear model most frequently used in previous research [6].

2. K-Nearest Neighbor Regression (KNN): 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 [7].

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 [8]. I incorporate this calculation into our model to understand which features are driving the model prediction results.

4. Extreme Gradient Boosting Regression (XGBoost): 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 [9].

I build a model workflow that passes the training data through each of these four algorithms. In addition, I use the feature importance calculations from RF and XGBoost to identify the key features that are driving the results.

In the figure 3.1 below, we see perfect ranking line plotted along total casualties count versus rank position which ranks census tracts with highest number of casualties.

Train & Test on NYC with census features

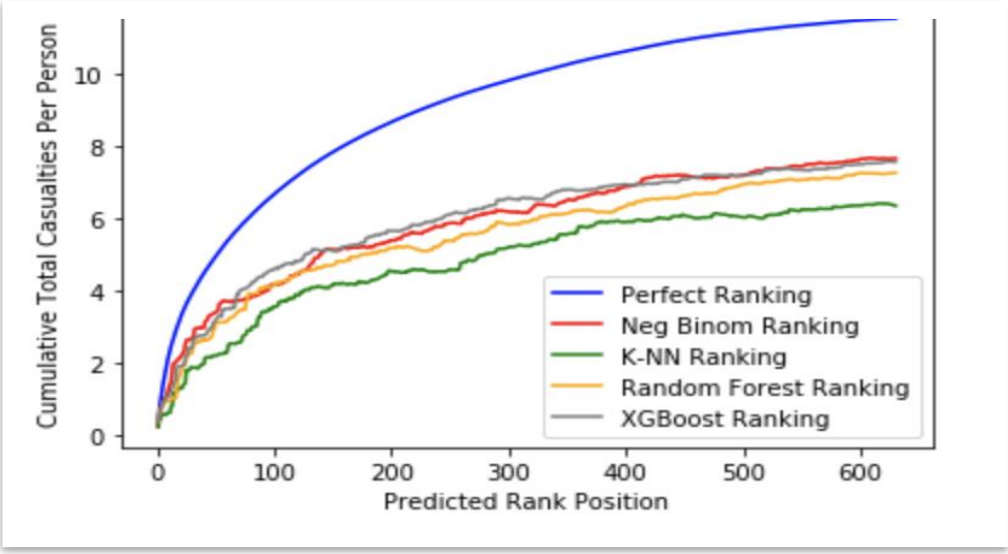

Fig 3.1

As we analyze fig 3.1, I used only census features to create perfect ranking and all other model rankings.

Even though there is a substantial gap between the perfect ranking curve and all four of the predicted ranking curves, fig 3.1 suggests that there is some predictive possibility within census features that relate to the count of casualties at the census tract level.

Train & Test on NYC with census & road features

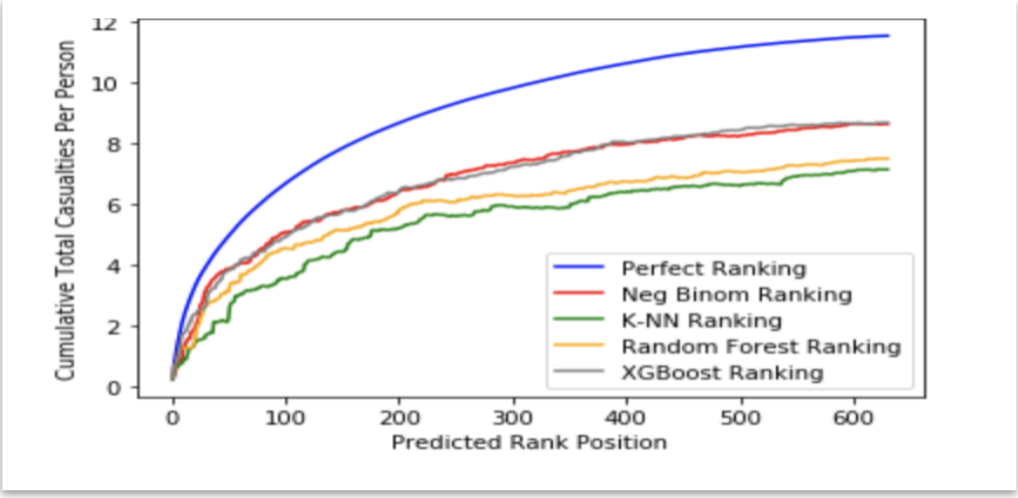

Fig 3.2

In the above fig 3.2, I incorporate census and road features to create perfect ranking and all other model ranking curves. Doing so improved the prediction ranking of four models. Also, gap between perfect ranking curve and other prediction ranking curves have decreased.

Now, as I mentioned before, I will be focusing on the NYC datasets i.e., NYC census variables and NYC road characteristic datasets for training and I will test the results on LA and DC dataset that is provided in the same GitHub repository. So, since we are done with training the datasets, let us analyze test results.

Train on NYC and test on LA data

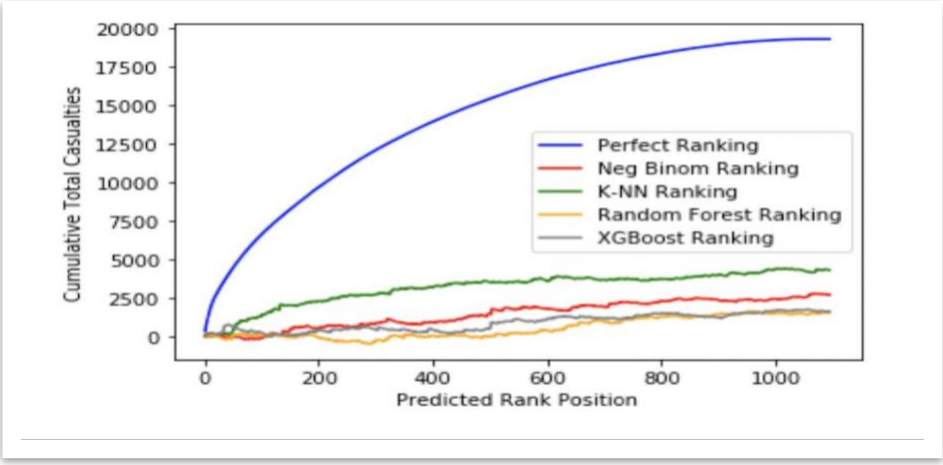

Fig 3.3

Fig 3.3 shows model trained on NYC data and tested on LA data looks less predictive when compared to results from previous fig 3.2. The model's curves or lines are closer to x-axis. KNN performed well compared to other models. It shows that predictive ability of socio factors from NYC is not much when deployed on LA.

On the other hand, fig 3.4 I have used NYC data to test on DC dataset. Compared to fig 3.3, this model shows an improvement in predictive performance of models as predictive ranking curves from all four algorithms are noticeably closer to the perfect ranking curve.

Train on NYC and test on DC data

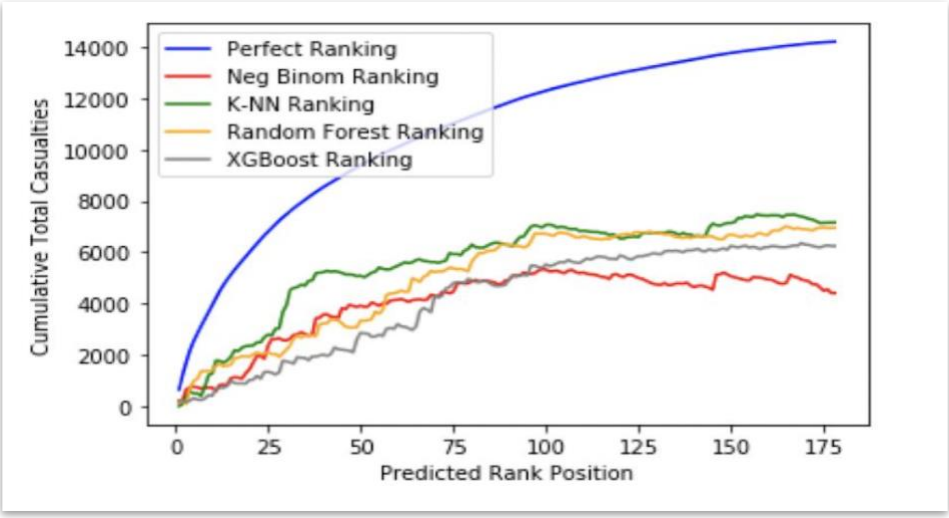

Fig 3.4

Now let us analyze why ranking models between the LA data versus DC data is significantly different. It might be because parts of NYC could be more similar to DC than LA, in terms of the way people travel throughout the city and the characteristics of the neighborhoods.

Feature Importance

Now let's analyze important features extracted by XGBoost and random forest algorithms which will be helpful to analyze top features which drive prediction results.

As we see in fig 3.5, population density is top feature among census and road characteristics followed by total length of road.

Feature *importance from RF and XGBoost with casualties per person as target variable*

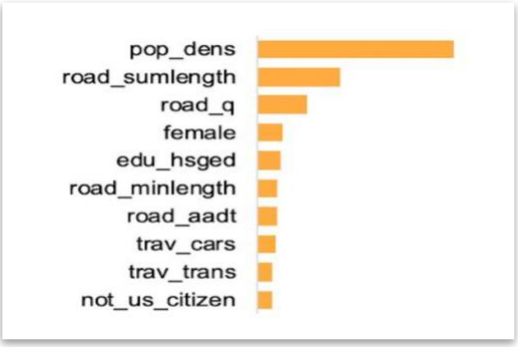

Fig 3.5

Now, let us analyze relationship between population density and casualties. I apply ordinary least squares (OLS) regression to gain further insight into how Census and road variables relate to casualties (fig 3.6).

OLS regression with important census and road variables

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

Fig 3.6

As we analyze from fig 3.6 that population density has a negative relationship with casualties per person, which means that as the number of people per square mile goes up, the average casualties per person goes down. This brings us to a counterintuitive question: Why would census tracts with more people per square mile have fewer casualties per person?

We can postulate that census tracts with higher population density have more street signage, lower speed limits and therefore less count of casualties.

Part 2: Non-socio-economic factors

ARIMA Model

I perform auto regressive integrated moving average model i.e. ARIMA statistical method to forecast number of vehicular collisions that would occur in NYC.

To perform this method, I converted data object to time series format. After data conversion, I test the time series data for its stationarity by performing augmented dickey-fuller test to check the stationarity in which p-value was less than 0.05 which confirmed that data is stationery.

Then data is split into test and train datasets to obtain model performance. Then the residuals are plotted (Fig 3.1) to find that they are normally distributed around zero with less variance which will be a good sign suggesting that model is performing well.

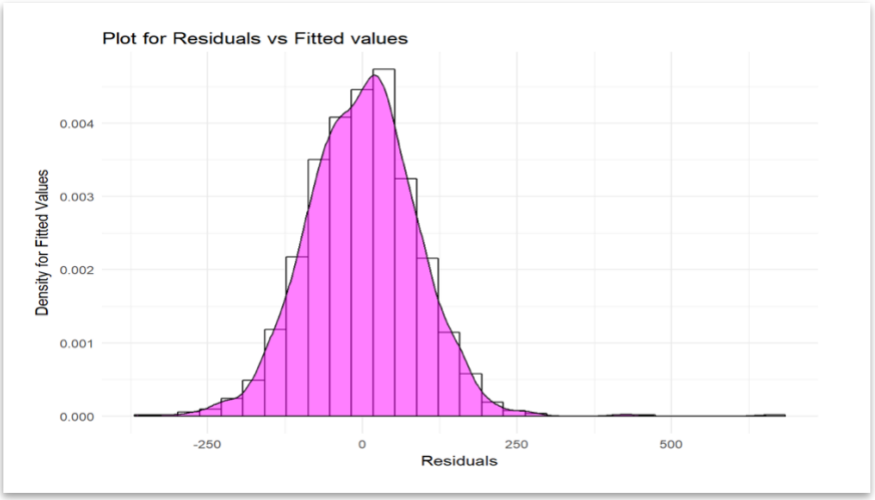

Fig 3.6

In the above figure 3.6, we see that ARIMA model has performed pretty well with its mean centered around zero but RMSE or root mean square error for ARIMA model is too high to consider it an optimized model. Therefore, let's create a random forest model to compare the performance of both the models.

Random Forest Model

Random Forest is an ensemble method which is used to obtain decision tree type prediction system. For time series, random forest cannot detect seasonality and stationarity on its own [10].

To help with analyses, I create attributes such as Day, Month, Weekday and year from the date and summary of vehicular crashes that had happened on that specific day. In this manner, I create weights for different Day, Weekday, Month & Year combinations which helps random forest predict time series. Then residuals vs fitted graph is plotted for this model (fig 3.7).

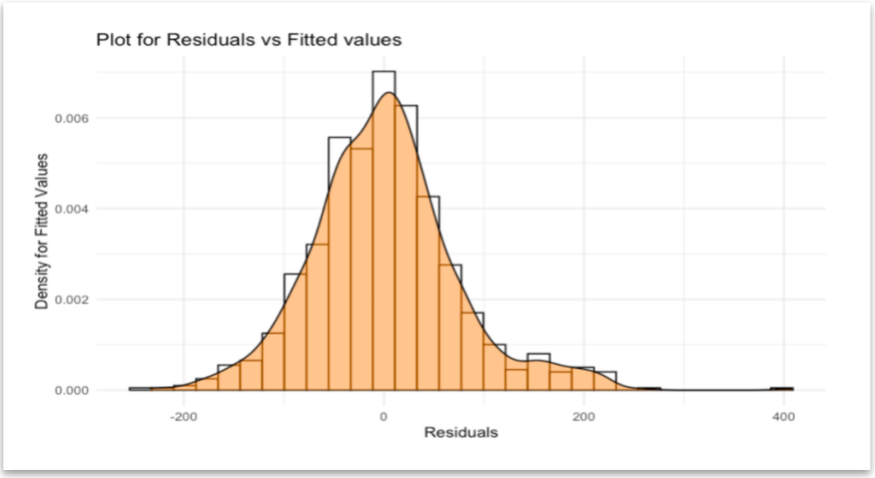

Fig 3.7

In the above figure 3.7, we see that random forest model has performed well with a better normality distribution compared to ARIMA model, with its mean centered around zero. Also, RMSE value for random forest model is 80.7 which is much less when compared to ARIMA model. Therefore, random forest model has performed better than ARIMA model

Discussion of Results

In this section, I will summarize results and insights that I derived from analyses of this project.

First part focuses solely on socio-economic factors such as city's census and road characteristics variables and to find models which can predict the factors for vehicular collisions better. The selected model uses four algorithms i.e., Negative Binomial, K-Nearest Neighbor, Random Forest, Extreme Gradient Boosting on census and road features.

The predicted ranking curve performed well in predicting casualties at the census tract level. After training NYC data, it is tested on LA and DC datasets from which I analyzed that NYC model works better on DC because predictive ranking curves from all four algorithms were noticeably closer to the perfect ranking curve and also 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.

Important features from XGBoost (fig 4.2) and random forest (fig 4.1) with casualties per person as target variable were extracted and as acquired from RF top three census features are population density, female and education level. Top three road features are total length of road, road quality measure and minimum length of road.

RF Top 10 Features	
pop_dens	0.29590
road_sumlength	0.11350
road_q	0.07833
female	0.05335
edu_hsged	0.04142
road_minlength	0.03834
road_aadt	0.03164
trav_cars	0.02526
not_us_citizen	0.02241
age_genx	0.01991

Fig 4.1

XGBoost Top 10 Features	
pop_dens	0.18084
road_sumlength	0.12689
road_minlength	0.07404
not_us_citizen	0.05179
trav_cars	0.03259
road_q	0.03155
trav_home	0.03147
divsep	0.03008
age_boomer	0.02806
road_aadt	0.02775

Fig 4.2

From XGBoost, top three census features are population density, not US citizen, traveling by car and top three road features are total road length, minimum road length and road quality.

It is important to note that the features that are highlighted to be significant for RF may be different from that of XGBoost. This is to be expected because of the inherent distinction in how the two algorithms perform predictions.

Second part focuses on non-socio-economic factors i.e., factual reasons for accidents such as alcohol involvement while driving, fatigue etc. and to find models which can predict factors for vehicular collisions better.

Two models i.e., ARIMA and random forest were used to predict casualties in which random forest performed better with a lower RMSE value compared to ARIMA model. Important features were tabulated based on reasons for highest count of injuries. Top three reasons for accidents were driver

inattention/distraction, failure to yield right of way, following too closely. Also, important features were tabulated based on vehicle types. Top three vehicle types that caused maximum number of injuries were sedan, passenger vehicle, station wagon/sports utility vehicle.

Conclusion and Future Work

In conclusion, selected model for socio economic factors can predict the ranking of census tracts by casualties per person. Furthermore, we find that with both census data and road characteristic data, the accuracy of the model increases meaningfully. Limitations in this project can be lack of road characteristic dataset in cities other than New York. In order to implement insights of one city to another, similar level of datasets must be available.

Also, I find relationship between important census and road features wherein we infer that population density is a major factor affecting casualty's count. Whereas non-socio-economic factors which cause most accidents is driver inattention/distraction.

According to the analysis cities should not only implement solutions such as constructing more street signage, implementing lower speed limits, developing road length and road quality to the highest ranked census tracts but also engage the local community in conversation and debate when it comes to transportation design and investment decisions since from data exploration, we see that immigrant communities are more prone to road injuries. Having a diverse viewpoint combined with data analysis will lead to better solutions which can lessen the curve of predicted road accidents.

I believe these ranking models or any future work on them should help cities better understand the communities with whom they wish to collaborate.

Sources Cited

1. https://apps.who.int/iris/bitstream/handle/10665/39723/WHO_PHP_12.pdf;jsessionid=29378E331316A4203D337D38B569B26D?sequence=1

2. https://www.who.int/roadsafety/decade_of_action/plan/plan_english.pdf?ua=1

3. https://www.who.int/violence_injury_prevention/publications/road_traffic/saving_millions_lives_en.pdf?ua=1

4. https://read.qxmd.com/read/23393405/road-traffic-injuries-social-change-and-development

5. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative_Binomial_Regression.pdf

6. Lord, Dominique, Mannering, Fred. "The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives." pp14. (2010).

7. https://scikit-learn.org/

8. " 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.

9. https://xgboost.readthedocs.io

10. https://machinelearningmastery.com/random-forest-for-time-series-forecasting/