ANALYZING VEHICLE-PEDESTRIAN COLLISION IN LOS ANGELES

ISHAAN JAIN, CHIATANYA GOKHALE,
SRI HARSHA, HARSHAL BHANGALE,
CHAITANYA NIJHARA, CHENYUE LIN

Executive Summary

Objective	➢ The goal of this project is to build a model that gives us insights into conditions and parameters that lead to severe vehicle-pedestrian traffic collisions.
Methodology & Tools	➢ Clearly define response variable – Severity (1/0) ➢ Business knowledge followed by statistical tools such as logistic regression to identify significant independent variables ➢ Building the model using JMP software
Key Insights	➢ Factors that impact severity of collision include day/time of collision, driver's age and interestingly, pedestrian action

Table of Contents

Introduction

Los Angeles is the second most populous metropolitan area in USA. It has ranked high on the lists of cities with the worst traffic for decades. The average rush hour delay per 30-minute journey is 25 minutes. That adds up to an additional 95 hours behind the wheel each year due to traffic jams, according to the report released by LADOT. Nonetheless, traffic collisions happen more often in rush hours. Based on our data, 1 out of 10 collisions (10.88%) in the LA are involved with a pedestrian.

Hypothesis

To predict the likelihood of severe vehicle-pedestrian collision occurring, where severity is defined as occurrence of either fatality or > 2 injuries in a collision

Methodology

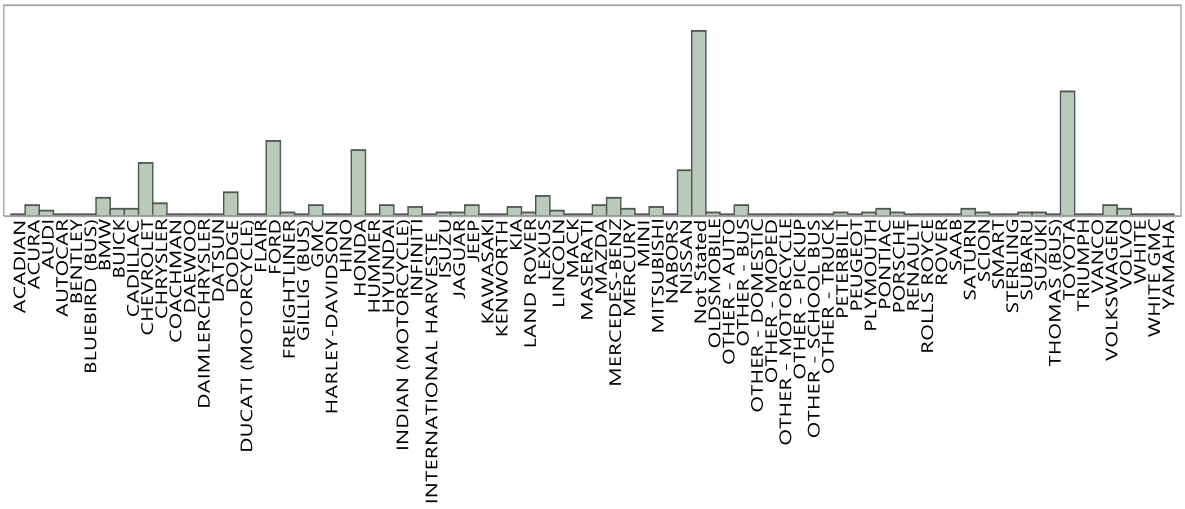
After cleaning the database, selecting only vehicle-pedestrian collisions, and treating missing values, we built the model in 5 broad steps:

1. First, we defined the response variable i.e severity of vehicle-pedestrian collision (1/0). We combined two variables: fatality and number of injuries. Any collision where 1 or more fatalities were involved was considered as severe, and any collision with 2 or more injuries was considered as severe (1) as well. All the other collisions were considered not severe (0).
2. Using domain / business knowledge selecting top 23 variables (out of 170 variables given in the database) that we believe could impact severity of collisions, while discarding those variables that could not possibly play a role in severity of collision (such as driver eye color, hair color, etc)
3. Out of those 23 variables that we selected, we ran logistic regression and got 7 significant variables that could impact severity of a vehicle-pedestrian collision, using which we built our 2nd best model. To beat that model, we then enriched the significant variables and used our business acumen to select top 5 variables that could impact severity of collisions the most.
4. At each step of building models, we used both logistic regression and neural network and selected the one that yielded better results.

5. Finally, to compare models we took an assumption that every severe collision costs the city $10,000. For every severe collision correctly predicted the model gained $10,000 and for every severe collision that the model missed, it lost $10,000

Data & Descriptive Statistics

The data used for the report is provided by LADOT (Los Angeles Department of Transportation).The timeline for the data set is April 2008 to March 2013. It is a very comprehensive dataset with 400,000 rows and over 160 variables. Upon analyzing the data we found that pedestrians are 5 times more vulnerable and likely to be severely injured or killed in a traffic collision. Therefore our key metric is KSI which stands for Killed or Severely Injured. To improve roadside safety we have decided to focus on analyzing collisions involving pedestrians. Each tuple i.e. a row represents a unique collision. There are numerous variables covering every minute detail of the collision ranging from car make and color, driver age, sex, date and time of collision, party at fault and so on.



The fig above shows the distribution of the variable car-make.

Since the data set is a collection of manual entries made by LADOT officials, it is bound to have a lot of missing values as can be seen from the above descriptive statistics.
For Filling up the missing values we followed two different approaches for qualitative and quantitative variables. Quantitative variables were on most occasions replaced by the mean (average).Replacing the missing values using the averages were a rather easy process. Filling up

the missing values for qualitative variables called for the application of domain knowledge and intuition. There were 3 different approaches followed for filling the missing values for qualitative variables.

1) Ignore
2) Replace by “NA” (Not Available)
3) Context Based Substitution.

For few variables like weather conditions there was no need to fill in any values so such tuples were neglected. For variables like location we replaced missing values by NA.
For variables like car make we took context based substitution approach. From the fig above the second highest number was Toyota cars. So we calculated the percentage of Toyota cars in the total dataset and then replaced those percentages of Not Stated values by Toyota. This strategy was followed by considering the significant values in decreasing order until all the missing values were filled. The intuition behind this approach was “Something which is significant in a whole should also be significant in the part that makes up the whole.”
The number of rows was reduced to 12,000 since we wanted to focus only on those collisions which involved pedestrian. The columns were reduced to 23. There were few variables like officer who gave ticket to party at fault after the collision took place which were irrelevant to the analysis which we were going to make hence those were removed altogether from the dataset.

ETL (Extract Transform Load)

The original data was present in the MS Access format. There were multiple related tables. The tables were combined to form one table using Primary Key, Foreign Key constraints. The data was put into Excel format since there are few operations like V-lookup which can be done efficiently in excel. The python script was written for data cleaning task.
The data was finally loaded into JMP in order to build models and make further analysis.

(For detailed description of variables please refer to APPENDIX I)

Analysis

2nd Best Model

Broad Methodology:

1. We started off with 23 variables that we had initially selected based on business insights and created a logistic regression model to determine significant variables.
2. The model gave us 12 statistically significant variables.
3. We chose a combination of the best 7 variables that gave highest R^2 value in the logistic regression model.
4. We then created both logistic regression model & neural network model and compared the profit values from both
5. Finally, we chose the neural network model for higher profit values.

Variables for the 2nd best model:

1. Collision day of the week
2. Driver Age
3. Primary collision factor
4. Party Violation Category
5. Pedestrian Action
6. Hit and run:
7. Direction of travel:

R2 values:

Training data set: 0.0935

Testing data set: 0.0538

(*Please refer to APPENDIX II for more details*)

Best Model

The R-Square for our model was still very low. In order to increase our profit obtained in the second best model we had to rethink our model. We approached our best model by first running the model on the 7 variables that were perceived with our 2nd best model. We grilled our model through a rigorous process of which had enrichment, remodelling and final validation based on the profit.

Enhancement:

The 2nd best model gave us an insight into how our model can be conceived but it still was not very significant based on the R-Square and the profit increase. We applied a bit of an intuition on our next model and further enhanced certain variables, which we thought would improve our model, which can be seen in table H.1.

Table H.1: A sample set of the variables that were enriched to test if these changes made the model better.

	Columns Enhanced	Reason	Enhanced values	Did it make the model better
1	If(Driver-Age > 60 && Driver-Age < 24 then 1 else 0)	The column is to check if there is a relation between Driver Age and severe accident.	1 –High Propensity by age to get into an accident. 0 - Low Propensity.	Yes
2.	If(Primary_Violation_Category == 'DUI' && if(Collision Day of the week == 'Sunday' ||'Saturday') then 1 else 0)	This column was to check if there is a correlation between the sobriety of the driver and the day of the week.	1 –High Propensity based day of the week and sobriety to get into an accident. 0 - Low Propensity.	Yes
3.	If(Driver-Sex == 'Male' || Driver-Sex == 'Female')then 0 else 1.	Since our decision tree showed a high segregation between not stated and actually the sex being stated	1 – High	No. The division happened simply because of high number of unstated cases and data was pretty evenly divided.

Final Variables	p-Value
Collision day of the week	<0.001
Driver Age	<0.041
Associated Factor	<0.001
Party Violation Category	<0.001
Pedestrian Action	<0.001

Remodelling and Testing:

Once we had all the underperforming variables eliminated or enriched, we went ahead to actually testing our model. But yet we only received a slight improvement in the accuracy of 69% from 65%. We further drilled down our model ever so often keeping the R-Square of our model as an indicator an increase in performance, and basing our surmise on our intuition until we reached the R-square shown in table H.2. The model built gave us a significant improvement of about 73% in accuracy from baseline model which only had 45% accuracy. The final list of best elements can be seen and there p-value can be seen below

Table H.2: The summary of the fit of the model.

Summary of Fit	
RSquare	0.152565
RSquare Adj	0.151122
Root Mean Square Error	0.333733
Mean of Response	0.155316
Observations (or Sum Wgts)	9999

Table H.3: The individual elements that were chosen for our final model.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.4248129	0.018349	23.15	<.0001*
Driver Age	-0.000952	0.000301	-3.16	0.0016*
Associated Cause - Factor 1{None Apparent&Other&Previou	-0.0743	0.011243	-6.61	<.0001*
Pedestrian Action{Crossing In Crosswalk Not at Intersection{	-0.104039	0.006747	-15.42	<.0001*
Party Violation Category{Auto R/W Violation&Hit and Run&	-0.064394	0.010102	-6.37	<.0001*
Collision Day of Week 2[0]	-0.018239	0.004234	-4.31	<.0001*

A scatter plot of some of the variables can be seen in the appendix which shows a clear demarcation between the various dimensions when plot against the severity.

Validation:
We validated our model based on the profit that our model produced with the baseline profit and we achieved more that 75 % of our baseline profit. The profit calculation and the accuracy is further discussed in APPENDIX III

Performance Measures

We mainly relied on comparing R^2 values from one model to the other and analysing the confusion matrices for profit calculation. R^2 values gave us a general idea of how our model was performing in terms of how many data points from the testing data base actually lied on the curve projected by our model vs. the training data base. Our basic approach was to maximize R2 for the testing data base, as far as possible.

On the other hand, we utilized confusion matrix to objectively analyse and compare different models by one final numerical value of profit projection in $

Confusion Matrix for Profit Calculation:

	Predicted	
Actual	True Positive	False Negative
	False Positive	True Negative

True Positive: Cases where we correctly predicted a collision to be severe
True Negative: Cases where we correctly predicted a collision to be not severe.
False Positive: Cases where we incorrectly predicted a collision to be severe, but it turned out to be not severe
False Negative: Cases where we incorrectly predicted a collision to be not severe, nut it turned out to be severe.

We decided on concentrating on only 2 of the 4 types of values i.e True positive and False negative, because in case of traffic collisions, in our opinion, these were the two types of data that made most sense, as in cases of True negative the city doesn't need to take any actions, and in case of False positive the city is incorrectly cautioned about a severe collision, but such cautions don't go in vain.

Profit Calculation:

To calculate profit for a model we used the following simple operation:

> Total Profit ($) = (# True positive * $10,000) – (#False Negative * $10,000)

Base Line Model

	Severe	Not Severe		Profit
Severe	1317	403	1720	$ 9,140,000.00
Not Severe	803	185	988	
	2120	588	2708	

2nd best model - Training

	Severe	Non Severe		Profit
Severe	6554	1075	7629	$ 54,790,000.00
Non Severe	800	1571	2371	
	7354	2646	10000	

2nd best model - Testing

	Severe	Non Severe		Profit
Severe	1623	97	1720	$ 15,260,000.00
Non Severe	803	185	988	
	2426	282	2708	

Best model - Training

	Severe	Non Severe		Profit
Severe	6081	790	6871	$ 52,910,000.00
Non Severe	1856	1273	3129	
	7937	2063	10000	

Best model - Testing

	Severe	Non Severe		Profit
Severe	1686	67	1753	$ 16,190,000.00
Non Severe	715	240	955	
	2401	307	2708	

Business Insights

Variable	Business Insights	p-Value
Collision day of the week	Clearly there were more cases of pedestrian accidents on weekends rather than on weekdays which was correlated to sobriety test and we found that 63% of the total cases were drunk on a weekend.	<0.001
Driver Age	Here we observed that the number of accidents caused were minimal around the median and were high on the extremities.	<0.041
Associated Factor	During our model development there was a clear distinction between benign reasons such as Stop & Go Traffic, No apparent reason and violations like defective vehicle equipment, previous collision and violation.	<0.001
Party Violation Category	Party violation category is another variable that is significant for our model. The various types of violations play a crucial role in determining the severity of the accidents. This is a quantitative category and some of the values it can taken are “Pedestrian Violation” , “Automobile Right of Way“ , “Driving under influence” and so on.	<0.001
Pedestrian Action	Since we are considering the involvement of pedestrians in the accident, more often than not even the pedestrians are equally guilty leading up to the accidents. Our intuition is backed up with a statistically significant variable. Some of the values it can take are as follows : “Crossing In Crosswalk at Intersection”, “Crossing Not in Crosswalk”, “In Road” and so on.	<0.001

Improvements

By doing this project we realized how much time it takes to get the database in shape, cleaned and ready for analysis. If we have to do the same project all over again, we would make the following changes in our methodology:

1. Check the data base thoroughly for usability. Perhaps, our most time consuming step was to figure out the response variable for the project. We didn't realize until late that since the dataset only had collision data (i.e all successes) and not data for conditions where collisions didn't occur, we cannot take it as our response variable for a classification project. We then had to spend a lot of time figuring out the most logical response variable to do meaningful analysis
2. A lot much more time to cleaning and pre-processing the database. We didn't a lot sufficient time in the begging to cleaning and pre-processing of the data base. This forced us to rush in the end.
3. Finally, in terms of working methodology, we should have one machine / person dedicated for analysis and handle the database. Our team faced a lot of version control issues, which could've been avoided and helped us save a lot of time and repetition of work.

Conclusion & Recommendations

Based on our experience of working with the database and the insights that we gathered from it, we have recommendations in two specific areas: Data collection, which focuses on how we can collect data more effectively to conduct more efficient analysis in the future; Policy& Infrastructure improvement, which includes the changes we recommend to the city of L.A

1. Data Collection Improvement

- Provide Zip Code or Official Region Names along with GPS Data:
  Zip code and Official region names can help the collision visualization process. Patterns and Clusters in terms of specific administrative districts are more likely to be detected.
- Avoid Ambiguous Names of Roads:
  For example, 101ST STREET, 101ST STREET (W), and 101ST STREET (N) should be re-categorized into only two classes for the accuracy of future data analysis.

- Keep the consistency of Records:
  For example, in injury table, “number killed” are not all records of consecutive years. This lack of records may impede timeline related analysis of injury data.

2. Policy& Infrastructure improvement

   - Rush Hour Policies:
     Temporary limit based on odd-and-even license plate can be implement to improve traffic congestion during rush hours. From another perspective, public transportation should have discount to commuters during rush hours to encourage them choosing public transportation instead of private cars.
   - Public Transportation Encourage:
     Improve the safety environment of public transportation, because safety concern is the biggest reason people do not choose transportation in Los Angeles. The city needs to combine the resources in order to improve the efficiency of the public transportation. The culture of taking public transportation should be promoted through campaign ads.
   - Pedestrian Signals and Signs:
     Based on our analysis, Koreatown and Downtown LA have greatest numbers of severe vehicle-pedestrian collisions, mainly because of traffic signals and signs issue. We suggest to improve the existing signals/signs, as well as install cameras on high collision frequency intersections.
   - Patrol/Police Distribution:
     Based on high risk timeline and location analysis, Patrol/Police should be distributed accordingly to avoid high possibility of severe vehicle-pedestrian collisions

APPENDIX I

Variables

S.no		Description	Data Format	
1	Driver Sex	Gender of the driver	qualitative	
2	Driver Age	Present age of the driver	quantitative	
3	Vehicle Year	Year in which vehicle was made	quantitative	
4	Vehicle Make	Company name of vehicle	qualitative	
5	Vehicle Type	Type of vehicle (Sedan)	qualitative	
6	Direction of Travel	Direction in which vehicle was moving	quantitative	
7	Party Violation Category	What type of violation	qualitative	
8	Associated Cause - Factor 1	The reason behind accident	qualitative	
9	Associated Cause - Factor 2	Next reason behind accident	qualitative	
10	Party Info table_Safety Equipment	Air bag deployed etc.	qualitative	
11	Collision date	Data on which accident happened	quantitative	
12	Collision Time	Time at which accident happened	quantitative	

13	Collision Day of Week	The day at which accident happened	qualitative	
14	Primary Collision Factor	Main reason for accident	qualitative	
15	Hit and Run	Felony, Yes, no etc.	qualitative	
16	Pedestrian Action	Action the person made at the time when accident took place	qualitative	
17	GPS X	Lattitude of the location where accident took place (converted)	quantitative	
18	GPS Y	Longitude of the location where accident took place(converted)	quantitative	
19	LONG	Longitude of the location where accident took place	quantitative	
20	LAT	Lattitude of the location where accident took place	quantitative	
21	Primary Road	Main road name	qualitative	
22	Secondary Road	The nearest small road name	qualitative	
23	Intersection Accident	Name of the intersection where it happened	qualitative	

APPENDIX II

For Baseline Confusion Matrix :

	Severe	Not Severe	
Severe	1317	403	1720
Not Severe	803	185	988
	2120	588	2708

For 2nd Best Model Confusion Matrix :

	Severe	Non Severe	
Severe	6554	1075	7629
Non Severe	800	1571	2371
	7354	2646	10000

	Severe	Non Severe	
Severe	1623	97	1720
Non Severe	803	185	988
	2426	282	2708

Accuracy %	81.25%
True Positive Rate	66.26%
False Positive Rate	14.09%
Sensitivity (True Positive Rate)	66.26%
Specificity (True Negative Rate)	85.91%

Accuracy %	66.77%
True Positive Rate	18.72%
False Positive Rate	5.64%
Sensitivity (True Positive Rate)	18.72%
Specificity (True Negative Rate)	94.36%

Lift Table in Propensity	Training	Testing
Lift with respect to Baseline - JMP Model	2.504118007	2.766884727

Neural

Model NTanH(5)

Training

Severity (1/0)

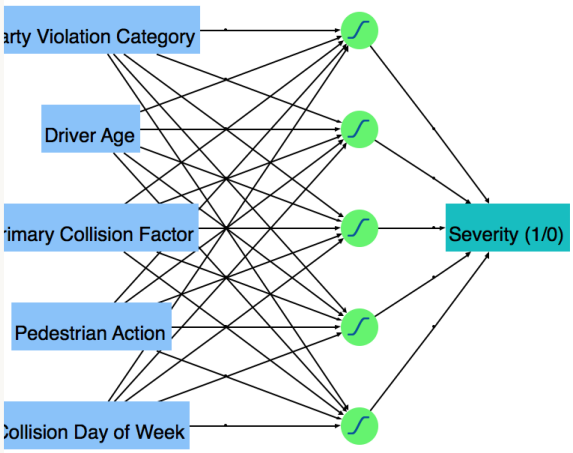
Measures	Value
RSquare	0.0971247
RMSE	0.3402885
Mean Abs Dev	0.2316496
-LogLikelihood	3466.7128
SSE	1177.3004
Sum Freq	10167

Validation

Severity (1/0)

Measures	Value
RSquare	0.1373945
RMSE	0.3322971
Mean Abs Dev	0.2278541
-LogLikelihood	806.03742
SSE	280.58067
Sum Freq	2541

Diagram



APPENDIX III

Best Model Statistics

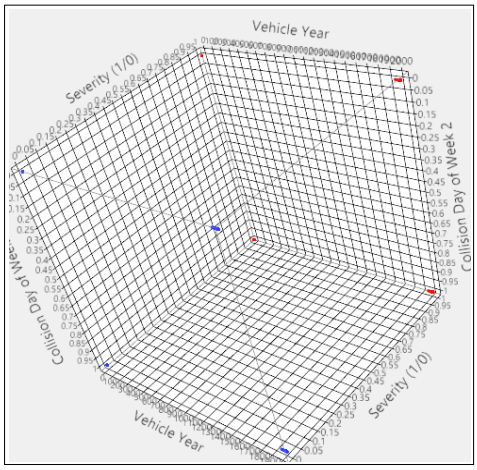
Training:

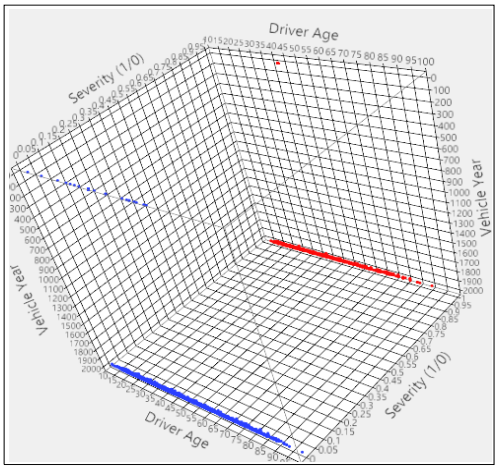
	Severe	Non Severe	
Severe	6081	790	6871
Non Severe	1856	1273	3129
	7937	2063	10000

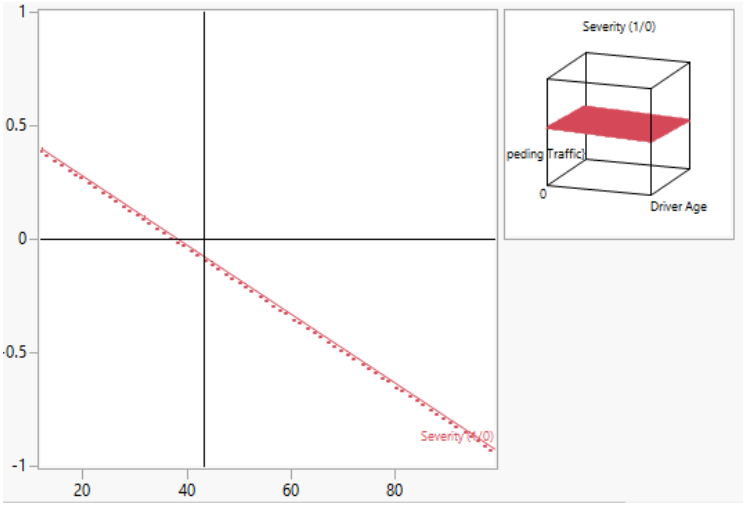
Testing

	Severe	Non Severe	
Severe	1686	67	1753
Non Severe	715	240	955
	2401	307	2708

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	2753	394.9951	0.143478	1.3338
Pure Error	7238	778.5752	0.107568	Prob > F
Total Error	9991	1173.5703		<.0001*
				Max RSq
				0.4065







Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	53.3568	10.6714	84.7459
Error	9994	1258.4623	0.1259	Prob > F
C. Total	9999	1311.8191		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.4248129	0.018349	23.15	<.0001*
Driver Age	-0.000952	0.000301	-3.16	0.0016*
Associated Cause - Factor 1{None Apparent&Other&Previou	-0.0743	0.011243	-6.61	<.0001*
Pedestrian Action{Crossing In Crosswalk Not at Intersection&	-0.104039	0.006747	-15.42	<.0001*
Party Violation Category{Auto R/W Violation&Hit and Run&	-0.064394	0.010102	-6.37	<.0001*
Collision Day of Week 2[0]	-0.018239	0.004234	-4.31	<.0001*