



Machine Learning Project (using Scikit) : Accurately predict “Injury” and “Property Damage” variable

BY VARUN GREWAL

E-MAIL: VGREWAL.TECH@GMAIL.COM

[LINKEDIN](#) | [GITHUB](#)

Table of Contents

- ▶ Introduction (3-6)
- ▶ Statistical Visualization (6-8)
- ▶ Model 1 – Simple Regression (9-10)
- ▶ Model 2 – Multiple Regression and Test/Train Scenarios (11-14)
- ▶ Model 3 - Decision Tree and Confusion Matrix Results (15-17)
- ▶ Future Applications (18)

Introduction: Key Information

► **Project Goal:**

- Take a deep dive into the "Collisions" historical dataset for the city of Seattle
- Discover the statistical relationships between different variables
- Build Regression and Decision Tree model to deliver accurate predictions for "Injury" and "Property Damage" variable

► **Data Description:**

Source: Seattle Police Department | **Year:** 2004-2020 | **Dataset Size:** 195K rows X 38 columns

► **Tools Used:**

Tool: Jupyter Notebook | **Language:** Python






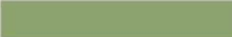










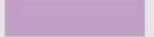



Packages: NumPy | Pandas | Scikit-Learn | SciPy | Matplotlib | Seaborn | iPywidgets

Introduction: City of Seattle

- ▶ Seattle is a growing city. In 2019, It had a net gain of about 11,400 people (+1.5% growth), reaching a total population of 753,700.
- ▶ Population Density of 7.9K/mile which makes it the no. 10 spot among the 50 most populous cities in the US
- ▶ High density results in "Traffic" which in turn creates ideal conditions for "Collisions"

Seattle crashes top 10 for density

For the first time, Seattle ranks in the Top 10 for population density among the 50 biggest U.S. cities.

CITY	POPULATION DENSITY PER SQUARE MILE, 2014		DENSITY CHANGE SINCE 2010	
1. New York City	28,056		3.9 %	
2. San Francisco	18,187		5.9	
3. Boston	13,586		6.2	
4. Miami	11,997		7.7	
5. Chicago	11,959		1.0	
6. Philadelphia	11,635		2.2	
7. Washington, D.C.	10,793		9.5	
8. Long Beach, Calif.	9,416		2.4	
9. Los Angeles	8,383		3.6	
10. Seattle	7,962		9.8	

Source: U.S. Census Bureau

GARLAND POTTS / THE SEATTLE TIMES

Introduction: Need for Analysis?

"Traffic Collisions" has multiple negative consequences for society:

- ▶ It can result in loss of human life or a life altering serious injury
- ▶ It can result in property damage/financial loss to the people involved and the city
- ▶ It can cause traffic jams lasting hours which can result in billions of dollars in lost productivity
- ▶ It creates unsafe road conditions for other drivers

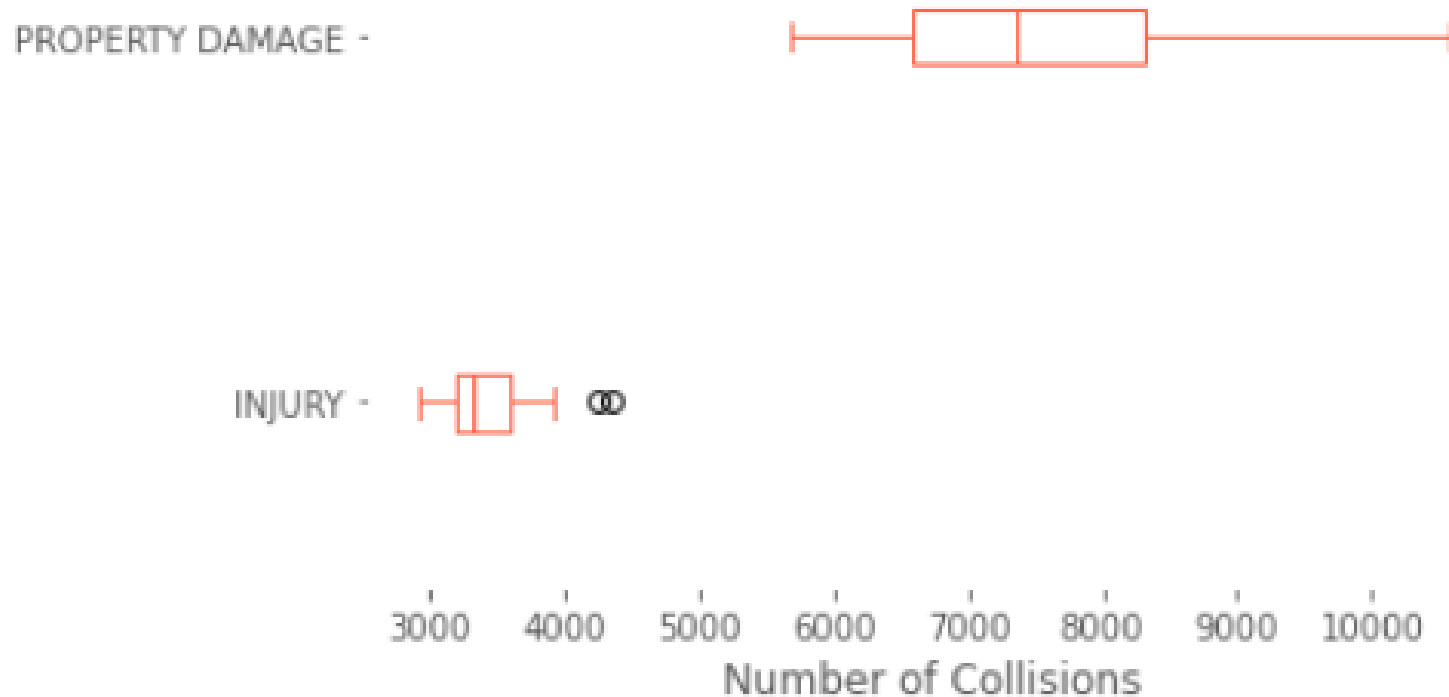
Introduction: Audience for the Analysis?

- ▶ Potential Employers
- ▶ Fellow Data Scientists/ Enthusiasts
- ▶ Public Office Holders (Mayor, City Council Members, etc.)
- ▶ City/Transportation Planners
- ▶ Emergency service providers such as Police, Fire and Medical Technicians
- ▶ Non-Profit Organizations
- ▶ Residents of the City

Statistical Visualization 1:

'Property Damage' collisions have a much wider range

Box plot of Collisions



Statistical Visualization 2:

In aggregated data, both “Injury” and “Property Damage” collisions show a **STRONG** correlation with YEAR. “Property Damage” seem to have a steeper slope

Regression plot of Injuries from 2004-2019

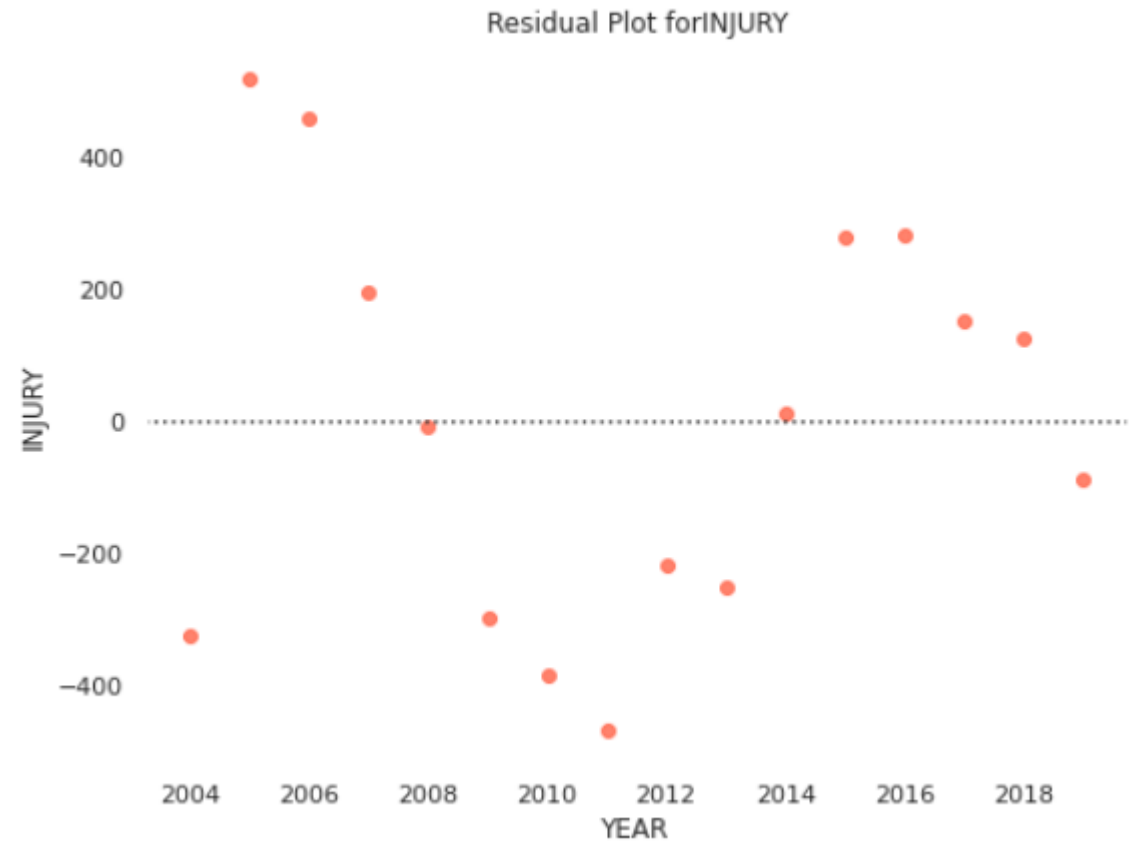
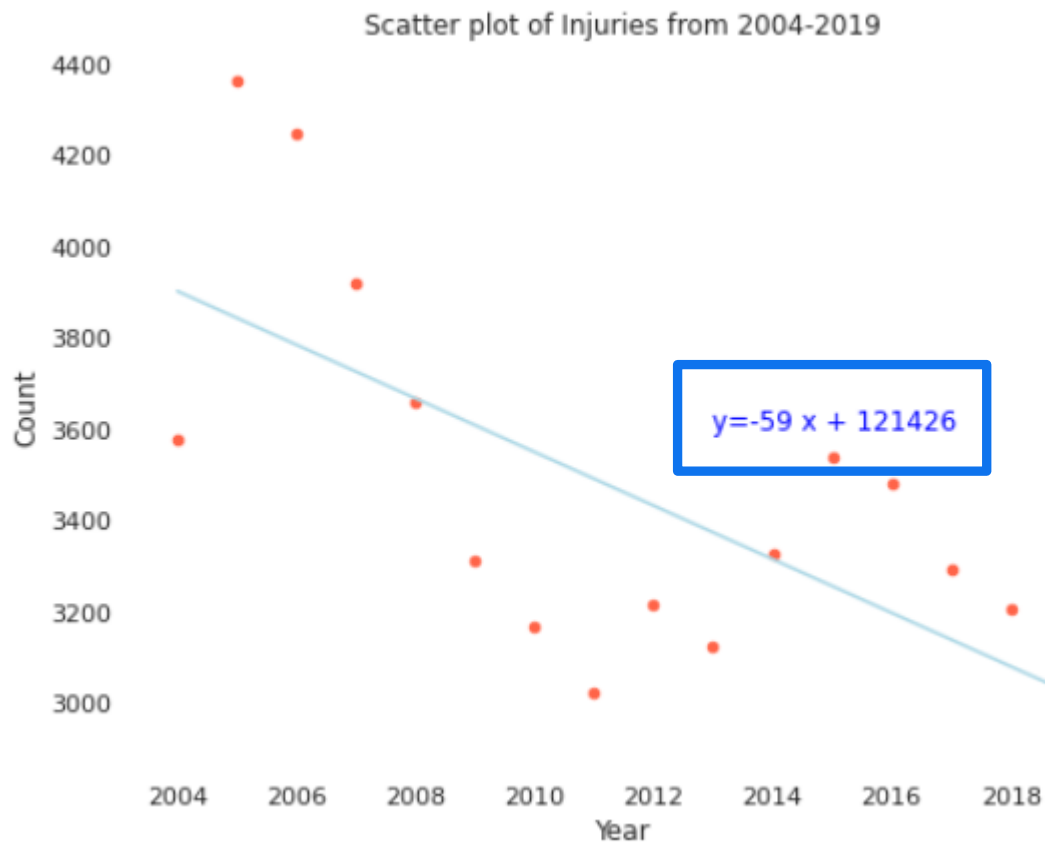


Regression plot of Property Damage from 2004-2019



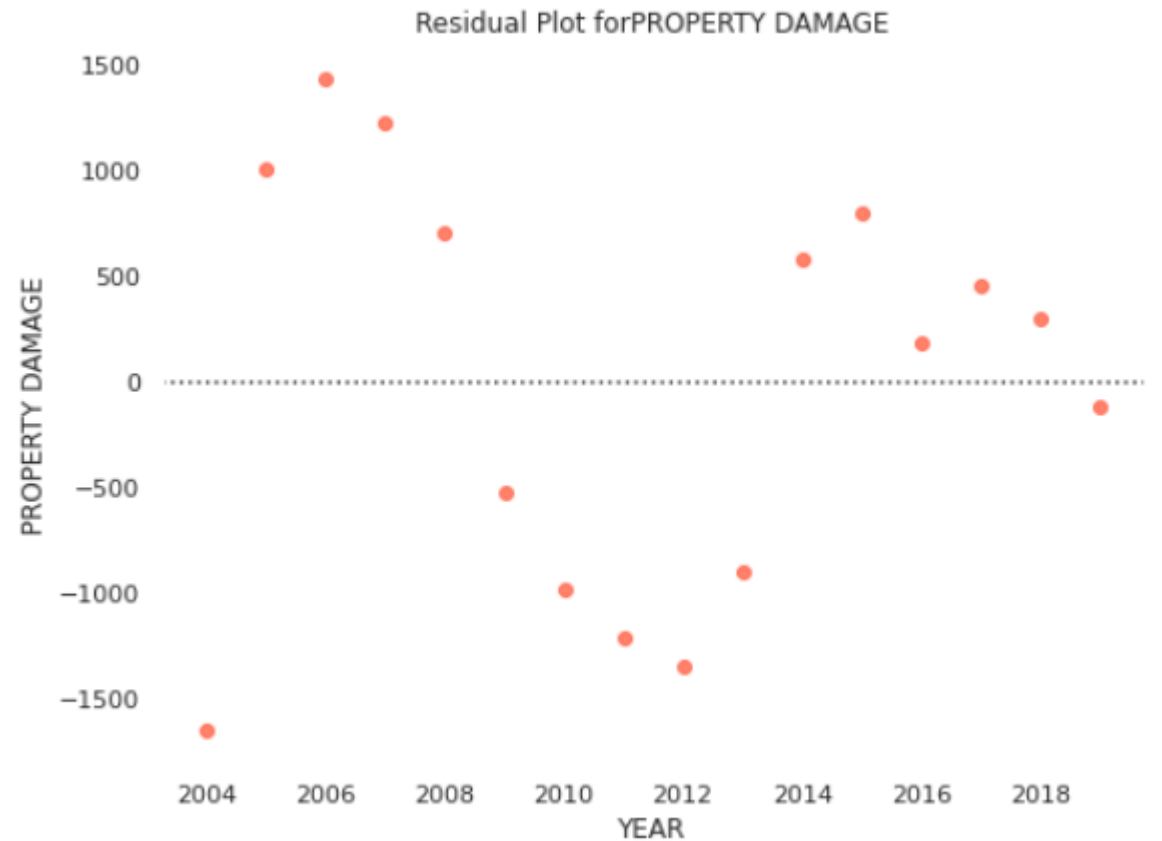
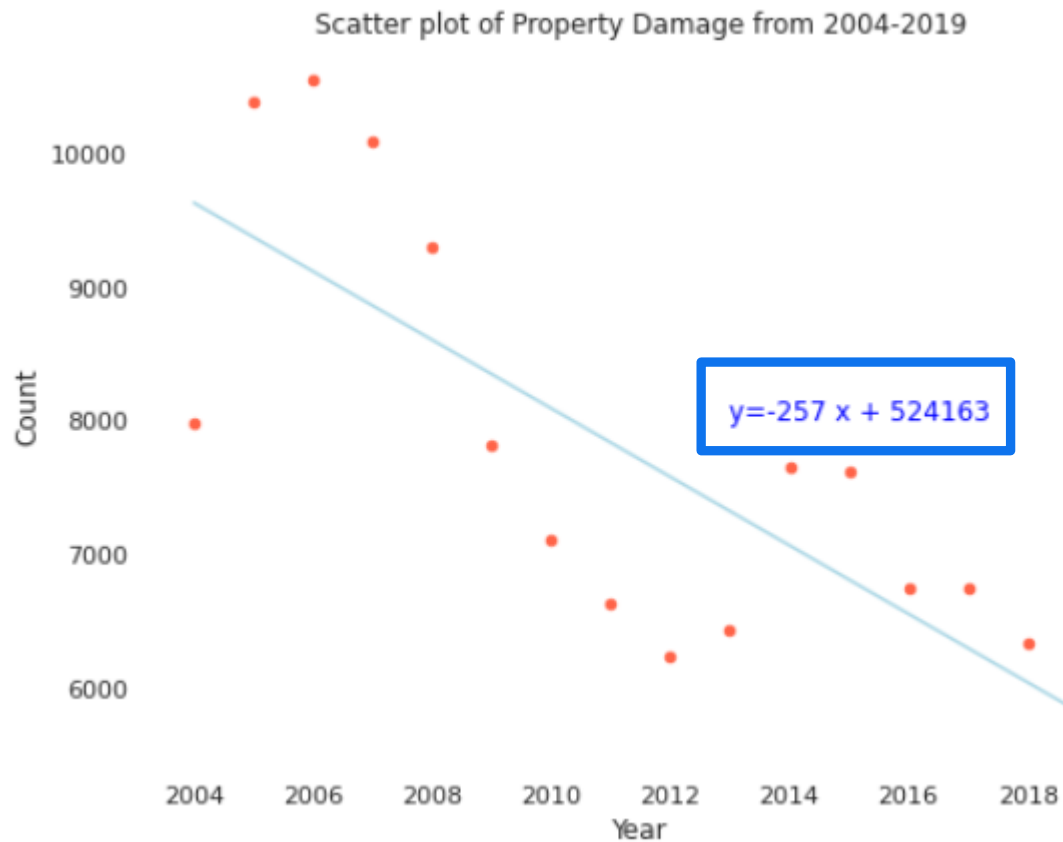
Simple Regression Results (Injury):

As expected, relationship is strong, but the residual values show that accuracy has room for improvement



Simple Regression Results (Property Damage):

As expected from Regression plot, “Property Damage” does have a steeper slope as compared to “Injury”. Accuracy has room for improvement.



Multiple Regression Results (Injury):

Variables ('DRY','CLEAR', 'AVGTEMP-F', 'MONTH','DAY','YEAR') show a STRONG relationship with "Injury" collisions and hence can be used to build a HIGH accuracy prediction model.

PEARSON CORRELATION :

Intercept: [45996.24525687]

Co-efficients: [[-1.59471644e-01 5.10139160e-01 2.47274532e-02 3.15819823e-02
-2.21726910e+01 -1.14035659e-02]]

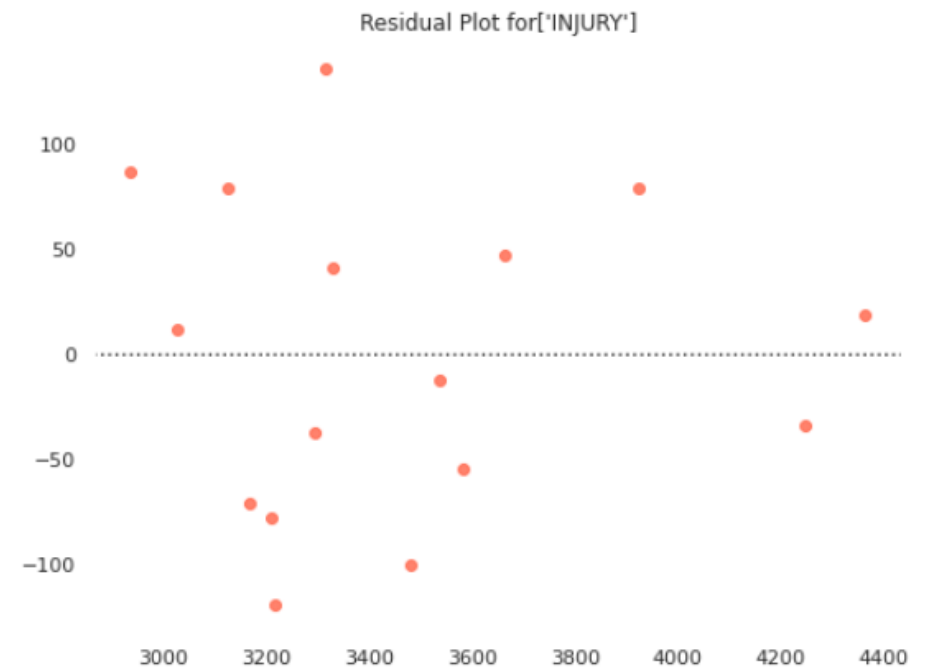
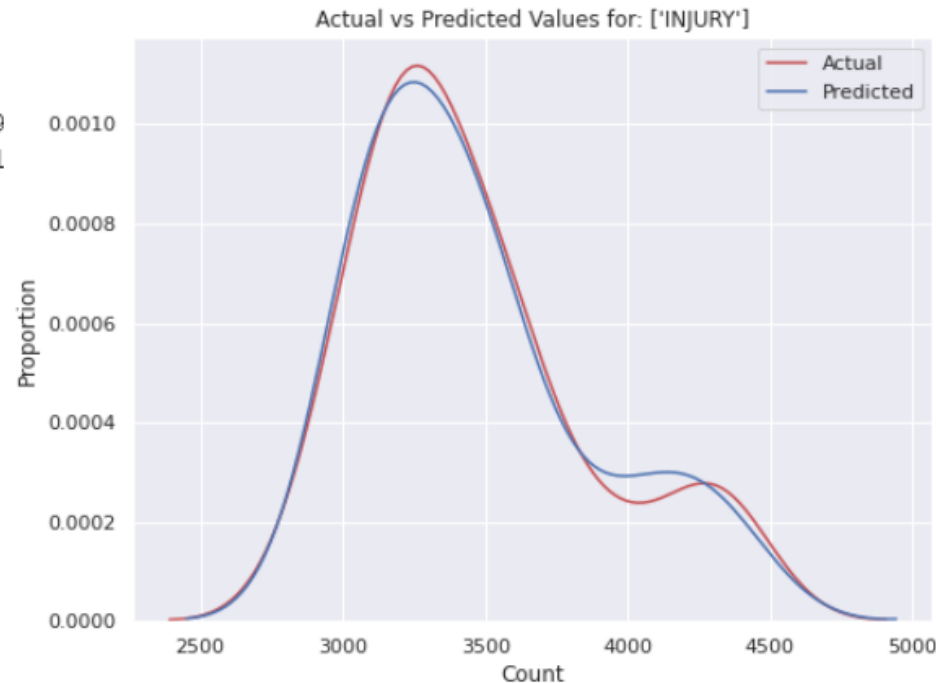
ERROR :

Mean Absolute Error (MAE): 64.59

Mean Square Error (MSE): 5461.91

ACCURACY :

R2-score: 0.97



Test/Train Scenarios(Injury) :

Results show a good fit for the selected variables

TestSize 0.20

TEST/TRAIN RESULTS :

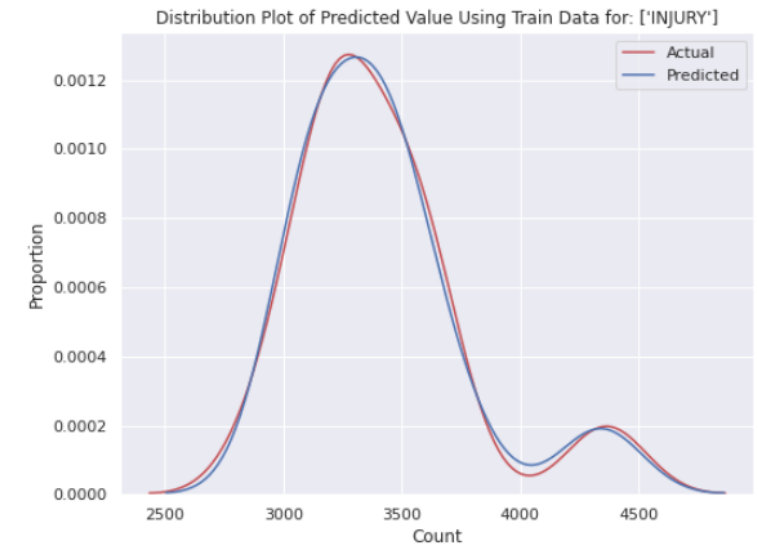
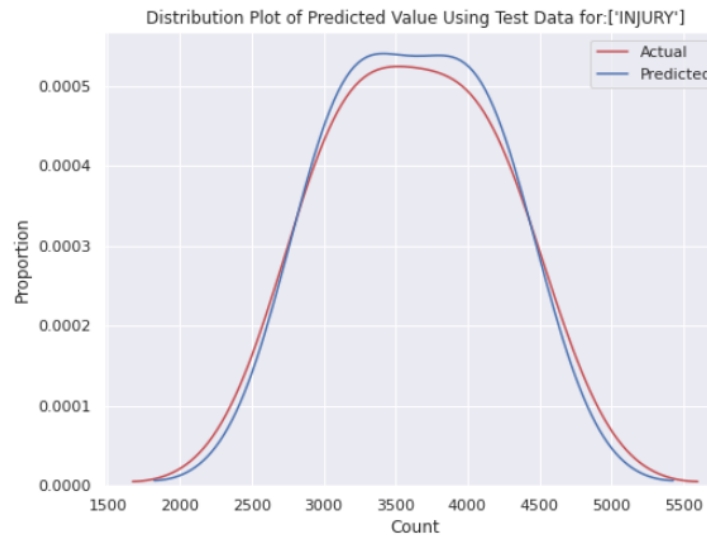
	DESC.	TEST	TRAIN
0	Samples	4.00	12.00
1	R2-Score	0.98	0.95

TEST VALUES :

PREDICTION : [3982.0, 3245.0, 3095.0, 4157.0]
ACTUAL : [3923.0, 3293.0, 3025.0, 4246.0]

TRAIN VALUES :

PREDICTION : [3120.0, 3368.0, 3724.0, 4341.0, 3150.0]
ACTUAL : [3169.0, 3330.0, 3662.0, 4364.0, 3209.0]



Multiple Regression Results (Property Damage):

Variables ('AVGRAINFALL-INCHES', 'DRY', 'CLEAR', 'MONTH', 'YEAR') show a STRONG relationship with "Property Damage" collisions and hence can be used to build a HIGH accuracy prediction model.

PEARSON CORRELATION :

Intercept: [109678.59333081]

Co-efficients: [[1.89793187e-01 -4.26614555e-01 5.50592876e-01 2.00895059e-02
-5.46452633e+01]]

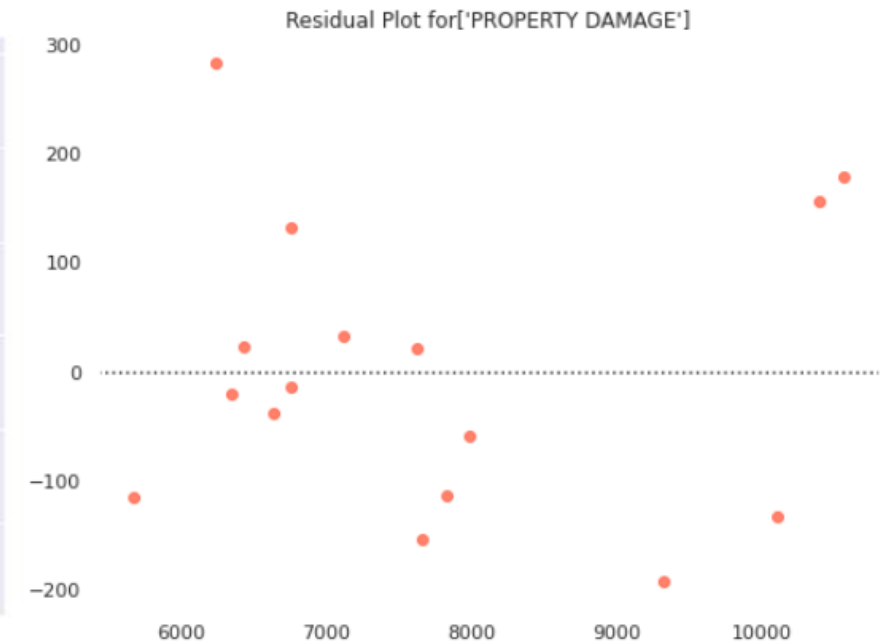
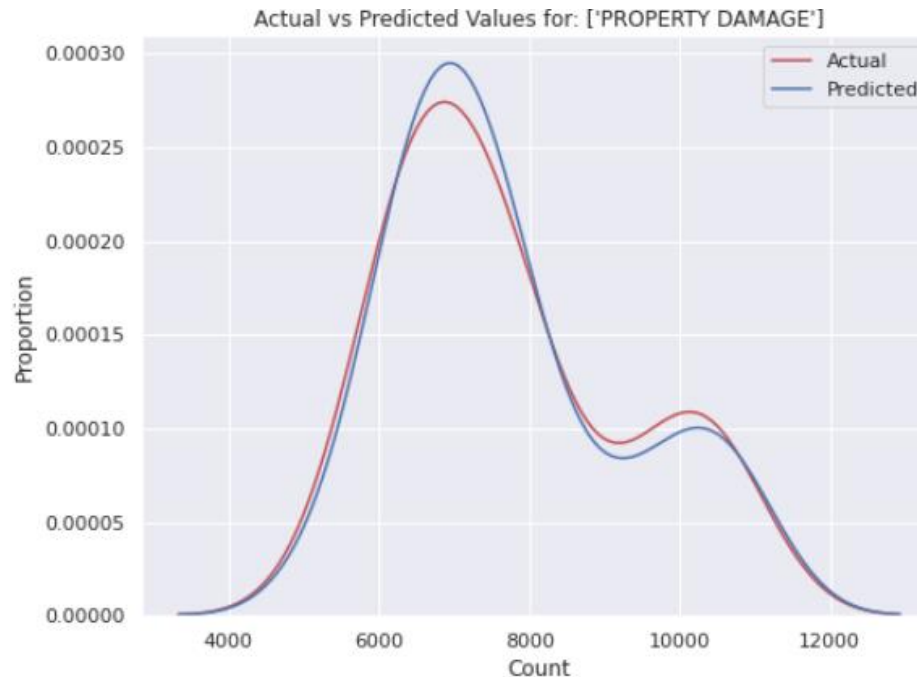
ERROR :

Mean Absolute Error (MAE): 103.1

Mean Square Error (MSE): 16750.2

ACCURACY :

R2-score: 0.99



Test/Train Scenarios(Property Damage) :

Results show a good fit for the selected variables

TestSize 0.20

TEST/TRAIN RESULTS :

	DESC.	TEST	TRAIN
0	Samples	4.00	12.00
1	R2-Score	0.99	0.99

TEST VALUES :

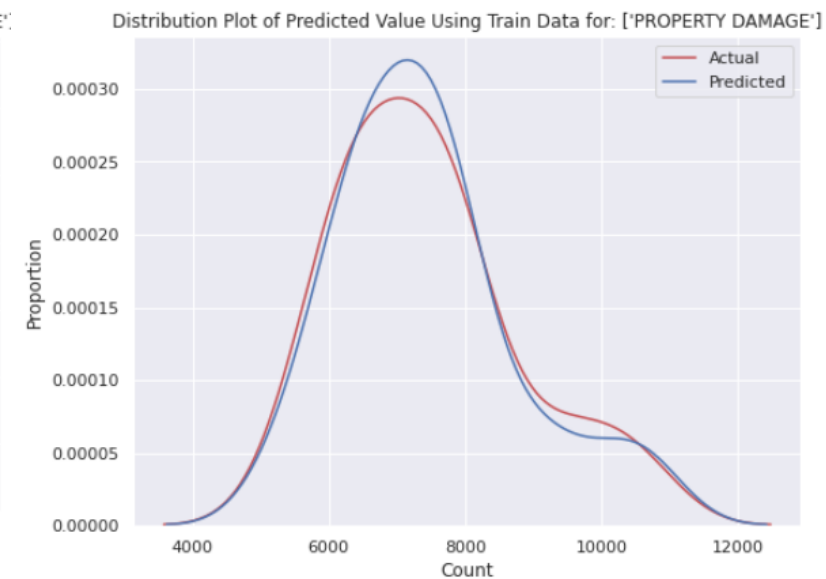
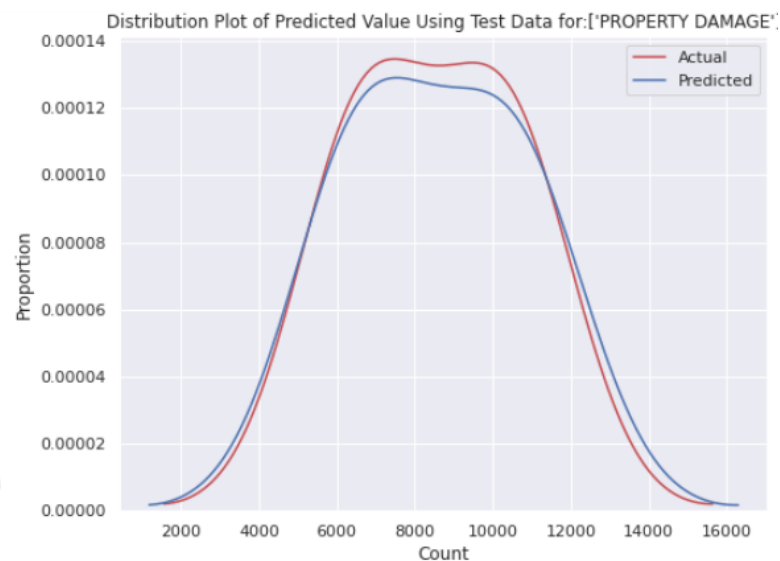
PREDICTION : [10033.0, 6773.0, 6558.0, 10922.0]

ACTUAL : [10102.0, 6759.0, 6635.0, 10563.0]

TRAIN VALUES :

PREDICTION : [7127.0, 7490.0, 9156.0, 10505.0, 634

ACTUAL : [7117.0, 7659.0, 9315.0, 10395.0, 6344.0]



Decision Tree Results:

“Property Damage” has a higher accuracy score.

Best Score is for TESTSIZE: 0.35

Train Sample: (116240, 8) (116240, 1)

Test Sample: (62591, 8) (62591, 1)

DecisionTrees's Accuracy : 0.7403

Jaccard Similarity Score : 0.7403

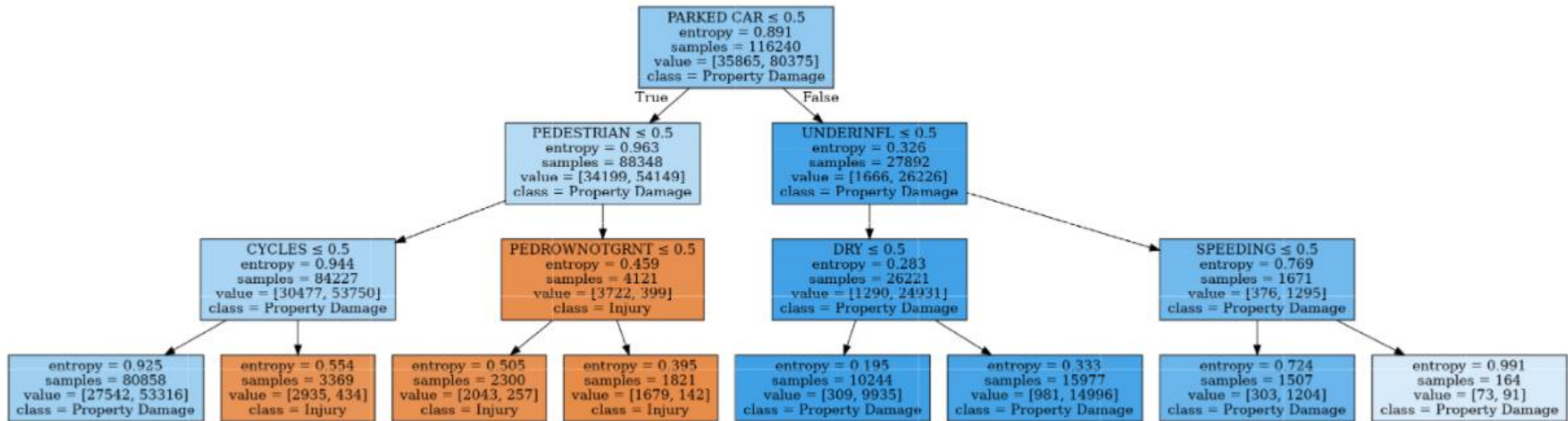
F1 Score : 0.6763

Log Loss : 0.523

Classification Report :

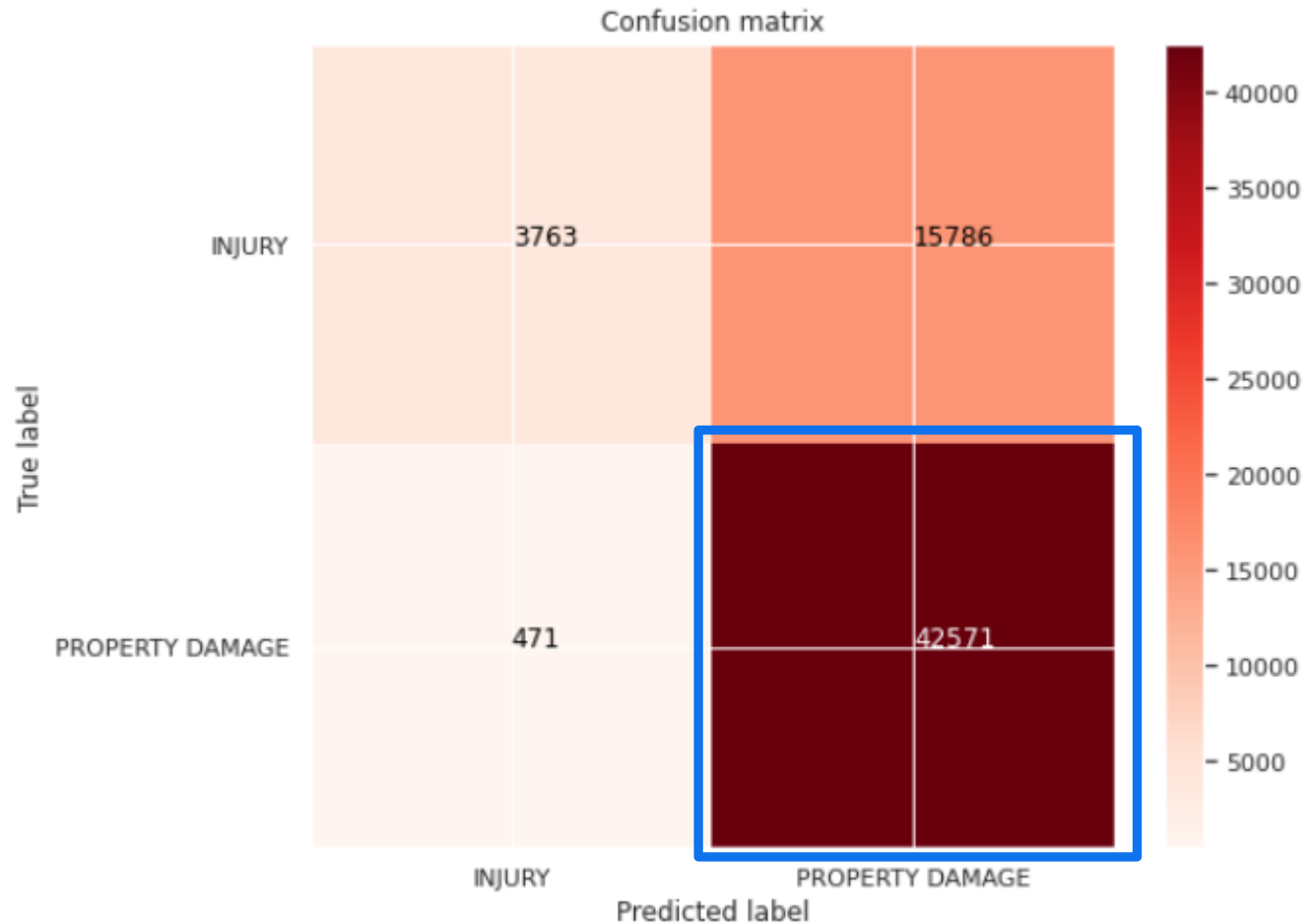
	precision	recall	f1-score	support
Injury	0.89	0.19	0.32	19549
Property Damage	0.73	0.99	0.84	43042
micro avg	0.74	0.74	0.74	62591
macro avg	0.81	0.59	0.58	62591
weighted avg	0.78	0.74	0.68	62591

Decision Tree Results (contd.):



Confusion Matrix:

As expected, "Property Damage" shows higher accuracy



Future Applications:

This analysis has shown a path forward for a technological solution that can analyze the data in real time and can deploy a forecast engine that can generate accurate forecast for "Injury" and "Property Damage" collisions for next 10-20 days. Based on forecast, government can deploy its limited resources to areas where they can get maximum return on investment and effort.

Thank You! 😊

QUESTIONS/COMMENTS/FEEDBACK:

E-MAIL: VGREWAL.TECH@GMAIL.COM

[LinkedIn](#) | [GITHUB](#)