ALY 6040: Assignment-4 Logistic Regression Decision Tree and Random Forest

Spring 2022
Instructor: Justin Grosz
Apr 24th ,2022
Hang Wu

Table of Contents

Business Problems

What factors are more likely (more odds) to cost the injured?

Are the more important odds to cost injured inline with the findings in other methods?

Model Building and Analysis

The target for all algorithms is set to be a binary variable IsInjured. the IsInjured column is engineered from the "Crash Severity" column where "Possibly Injury", "Incapacitating Injury", "Killed" are set to be 1 and "Not Injured" set to be 0. The column represents whether **the person is injured**. As for the predictors, all Boolean columns are converted to 1/0 binary integer for easier classification. Dummy variable are created for all categorical columns of the Bike Crash dataset. That increases the predictors amount to 48. The 48 predictors will be reduced to 10 or 11 depending on the algorithms after the multicollinear analysis. The multicollinear issues are solved with Variance Inflation Factor (VIF) score, and VIF scores with VIF = inf or VIF > 10 will be eliminated. 39 dummy features are eliminated due to the VIF score being infinity(Table 1), and one feature "At Intersection Flag" being eliminated because the VIF score is greater than 10 (822,442). Based on the correlations matrix (Figure 7), Person Helmet_Worn" and "Damaged','Person Helmet_Worn, Not Damaged' are also dropped due to high correlation with the existed columns, which leads to the singularity matrix error in the Logistic Regression model. Then, 3 models will be created. The first is the Logistic Regression Modelling using the 12 features left after column deletion. The second model which is the Decision Tree model is fitted with depth to be 3, and Random Forest Model is fitted with depth to be 4 and the minimum trees to be 500.After all, check the prediction accuracy for precision, recall of the 3 classifiers. Use the F1-score which combines the metrics prior for benchmarking. AIC is not used because both Decision Tree and Random Forest algorithm don't have AIC module, AUC (Area Under Curve) is used to support the ranking of models and it's aligned with the F1-score. Yet we can't get a breakdown of the precision and recall. Since "Crash Total Injury Count" becomes such an important feature, we will rank the rest of the feature importance with the feature removed. Confusion Matrix are used to compare prediction accuracy and more importantly the False prediction.

Interpretation and Recommendation

Part 1: Logistic Regression

The logistic model presents a prediction accuracy of 99.44%, the AIC is 72.45 and the BIC is 139.48(Table 2).With FP=2,TN=32,FN=0,TP=459(Figure 1a) as the confusion matrix, since the False Negative result is 0 and the False Positive results is 2, meaning that there aren't any injured people who is predicted as non-juried and only two cases of non-injured case predicted as injured. We have a near perfect case of prediction rate.The confusion matrix is important in terms of knowing what kind of prediction errors we are facing during the prediction. Knowing that we would know whether the model should be re-engineered or not. "Crash Total Injury Count" has the most importance followed by Person "Helmet Not worn", "unknown" and "worn with damage". on the coefficients(Table 2) and the Odds Ratio (Table 3), "Person Helmet_Not Worn"," Person Helmet_Unknown If Worn", and "Person Helmet_Worn, Unk Damage" have more odds to predict the injuries than the other 9 features.

The overall prediction accuracy and the TP prediction in Confusion Matrix are the two parts that Logistic Regression outperform the others. However, it lacks the True Negative Prediction, and it's interpretation, both Odds Ratio, Coefficient, z-score, don't tell a better story than the Tree based classification system. We don't have a decision path that we can use to make decisions as well as building a profile. The interaction effect of the predictors is missing.

Part 2: Decision Tree

Decision Tree model facilitates decision-making by breaking down business questions into a series of questions. Choosing the correct questions and the best tree would be the most important in terms building the correct tree. With the best 98.78% of prediction accuracy can be seen with a single tree of depth 3. "Crash Total Injury Count"(Figure 3) and "$1000 to any damaged property" are the most important features though Crash Count is far more important(97.17%) comparing to others. The decision tree is also presented in Appendix(Figure 2). With

FP=2,TN=39,FN=4,TP=447(Figure 1b), It has more False Negative Cases than other models. 7 more True Negative cases which favors predicting Non-Injured Cases, which is needed for this dataset because it's heavily skewed toward Injured cases.

With gini index to be 0, which suggests a pure gini/perfect split, indicating that we are 100% certain if there are injured person in the accident, the injured status would definitely be yes. Conversely, a gini of 0.469 in the third branch (Speed limit < 37.5)suggest an equal distribution of all elements in the injured class, and less impurity (0.245) is spotted in the non-injured class, suggesting that speed limit is a better predictor for non-injured case. In fact, both speed limit less than 37.5 and crash time greater than 239.5 could use to predict non-injured case. while higher speed limit can't be used to predict injured case. Low crash time has higher chances to predict non-juried case than lower speed limit and less than $1000 property damage, but crash count is still the only pure node to predict injuries convincingly. Comparing to the Random Forest and the Logistic Regression, it's only advantage is the precision prediction accuracy of the NonInjuried case, which is 0.95 comparing the 0.93(RF) and 0.91(LR). Classification Report(Table 6) is the only benchmarking metrics that Decision Tree outperforms other models in one area. However, that's hardly enough because F1-score of the Decision Tree (0.94) is less than LR (0.96) and RF(0.97). Decision Tree is the least favorible model of the 3 by confusion matrix, F1-score and overall prediction accuracy 98.7% vs the 99.39% for FP and 99.5% for Logistic Regression.

Part 3: Random Forest

A Random Forest algorithm is a classification algorithm consisting of many decision trees(*Yiu, T. 2021)*, It tries to build uncorrelated models(trees) with depth be 4 and minimum tree be 5000, being run 3 times with number of trees set to 5000,9000 and 30000. The optimal number of trees is between 5000 and 9000, further test needed to optimize the algorithm. Best accuracy is 99.39%, and the worst to be 99.18%. With FP=3,TN=43,FN=0,TP=447(Figure 1c), it has more True Negative(Non-injured) cases than both Decision Tree and the Logistic Regression, 1 more False Positive case than LR. It's focus on the True Negative prediction balance the prediction results as the heavily skewed dataset have 10 times more Injured cases than Non-Injured ones.

Since False Negative prediction, that a person is indeed injured but classified as non-injured can potentially harm a person's wellbeing and possibly life. RF's ability to minimize FN prediction makes more favorable than Decision Tree which has the most FN cases. It has the same prediction accuracy as the LR but it's recall prediction of both TP and TN are 0.95 and 0.99 against the 1 and 0.99 with Logistical Regression.

As for feature importance comparisons, if the Crash Total Injury Count feature is dropped, then Crash time and the speed limit are the most important features. Figure 8 shows a sample single tree out of the 9000 trees in the Random Forest. The sample tree shows a more evenly distributed branches for each classification category, and has better splits and lower gini impurity of some of the node. In fact, the highest gini impurity is on the node which is 0.136, which is about 1/3 of the 0.469 in the decision tree model. This indicates that Random Forest is a more balanced model and the fact that there are another 8999 of them helps us to choose the most optimal Decision Tree automatically. A automatic performance evaluation model can be used(*Płoński, P. 2020*) to find the optimal number of trees in the forest to get the best results via 100 box plots. The code is included in the script without a successful run due to the long computational time. However, it suggests the possibility to further optimize the model.

Feature Importance diagram(Figure 5) indicates a more uniform representation of importance of all features than the Feature Importance of the Decision Tree(Figure 4). As DT model heavily lean towards one feature, overfit could happen as a result especially considering the skewness of the dataset. Comparing to the Odds Ratio(Table 3) of the Logistic Regression also heavily lean towards Crash count and the other 5 features. It has a clear difference between rankings, making feature elimination easier should we ever eliminate more features for computational speed.

Random Forest has less overall prediction accuracy than Logistic Regression, 99.39% and 99.5% respectively. The sample tree would suggest a more balanced classification and decision path than the LR model.

Interpretation

The decision tree offers a simple yet effect story, if we were looking at a case that costs more than $1000 damage more than 1 occasion, and the speed limit of the area being greater than 37.5%. We would have a probability of injuries albeit the gini impurity is extremely close to 0.5, making it an impure node to end the split.

Random Forest offers a better prediction rate than Decision Tree when the depth is controlled in 3-4. With computation time way more than other algorithms, RF is expected to outperform both Logistic Regression and Decision Tree. Logistic Regression has prediction accuracy better than Random Forest. The most important feature that all 3 models to predict injured status suggest is the "Crash Total Injury Count", which is 16 times to 36 times more than the second most important feature, of course if the injured count is more than 0, then the injured status is 1. In order to understand the importance ranking of the other features, Crash Total Injury Count is deleted and the Feature Importance graphs are computed again. Without reposting all tables and graphs again, we will cut to the conclusion that logistic regression favors the school zone flag(Odds Ratio to be 1.71x10^11) then whether a helmet worn. Decision tree (Figure 4) and random forest(Figure 6) favor both crash time and speed limit with opposite ranking. They are better indicators of injured case than whether a helmet is worn or whether it's a school zone in both algorithms. It's clear that in all algorithms only crash count can convincingly predict injuries, while the other predictors would predict non-injured case.

Conclusion

Logistic regression focuses on features that can predict the injured case, while decision tree and by extent the random forest focuses on predicting the non-injured cases. Crash count is definitely has the perfect certainty of predicting the injured status, but helmet not worn or damaged helmet can definitely increase the odds to predict injuries. Injured cases are not happened in low-speed zone or when the crash doesn't occupy as much time. FI-score and overall prediction accuracy suggests that the macro average of Random Forest and Logistic Regression However Random

Forest has better precision and recall accuracy as well as a more balanced split and purer nodes. Random Forest is the superior model of the 3.

Reference

Yiu, T. (2021, September 29). *Understanding random forest*. Medium. Retrieved May 8, 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Płoński, P. (2020, June 30). *How many trees in the random forest?* MLJAR. Retrieved May 8, 2022, from https://mljar.com/blog/how-many-trees-in-random-forest/

Appendix

Table 1: VIF Score of all engineered features

index	VIF	Column
24	Infinity	Surface Condition_Other (Explain In Narrative)
21	Infinity	Roadway Part_Service/Frontage Road
23	Infinity	Surface Condition_Ice
42	Infinity	Traffic Control Type_Stop Sign
25	Infinity	Surface Condition_Sand, Mud, Dirt
26	Infinity	Surface Condition_Standing Water
27	Infinity	Surface Condition_Unknown
28	Infinity	Surface Condition_Wet
29	Infinity	Traffic Control Type_Bike Lane
30	Infinity	Traffic Control Type_Center Stripe/Divider
31	Infinity	Traffic Control Type_Crosswalk
32	Infinity	Traffic Control Type_Flagman
33	Infinity	Traffic Control Type_Flashing Red Light
34	Infinity	Traffic Control Type_Flashing Yellow Light
35	Infinity	Traffic Control Type_Marked Lanes
36	Infinity	Traffic Control Type_No Passing Zone
37	Infinity	Traffic Control Type_None
38	Infinity	Traffic Control Type_Officer
39	Infinity	Traffic Control Type_Other (Explain In Narrative)
22	Infinity	Surface Condition_Dry
20	Infinity	Roadway Part_Other (Explain In Narrative)
41	Infinity	Traffic Control Type_Signal Light With Red Light Running Camera
11	Infinity	Day of Week_Tuesday
44	Infinity	Traffic Control Type_Yield Sign
43	Infinity	Traffic Control Type_Warning Sign
6	Infinity	Day of Week_Friday
7	Infinity	Day of Week_Monday
8	Infinity	Day of Week_Saturday
9	Infinity	Day of Week_Sunday
19	Infinity	Roadway Part_Main/Proper Lane
10	Infinity	Day of Week_Thursday
12	Infinity	Day of Week_Wednesday
13	Infinity	Intersection Related_Driveway Access
14	Infinity	Intersection Related_Intersection
15	Infinity	Intersection Related_Intersection Related
16	Infinity	Intersection Related_Non Intersection
17	Infinity	Intersection Related_Not Reported
18	Infinity	Roadway Part_Entrance/On Ramp
40	Infinity	Traffic Control Type_Signal Light
45	4.1348715	Person Helmet_Not Worn
46	3.02264282	Person Helmet_Unknown If Worn
48	2.6233016	Person Helmet_Worn, Not Damaged
47	1.78444934	Person Helmet_Worn, Damaged
3	1.20658784	Crash Total Injury Count
4	1.08978675	Crash Year
5	1.07481924	Speed Limit
2	1.0454867	Crash Time
1	1.00787239	Construction Zone Flag
0	822442	At Intersection Flag

Table 2: Logistical Regression Results

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.014698
         Iterations: 35
                           Logit Regression Results
==============================================================================
Dep. Variable:              NoInjured   No. Observations:                 1970
Model:                          Logit   Df Residuals:                     1960
Method:                           MLE   Df Model:                            9
Date:                Sat, 07 May 2022   Pseudo R-squ.:                 0.9402
Time:                        06:08:33   Log-Likelihood:                -28.955
converged:                      False   LL-Null:                       -484.27
Covariance Type:            nonrobust   LLR p-value:                3.170e-190
=============================================================================================================
                                                coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
const                                       -28.3184   5.61e+04     -0.001      1.000    -1.1e+05     1.1e+05
$1000 Damage to Any One Person's Property     2.5585      0.730      3.503      0.000       1.127       3.990
Active School Zone Flag                     -22.9002   4.65e+05  -4.93e-05      1.000   -9.11e+05    9.11e+05
Construction Zone Flag                      -11.2970   2524.496     -0.004      0.996   -4959.219    4936.625
Crash Time                                   -0.0002      0.001     -0.348      0.728      -0.001       0.001
Crash Total Injury Count                     52.9567   5.61e+04      0.001      0.999    -1.1e+05     1.1e+05
Speed Limit                                   0.0177      0.021      0.843      0.399      -0.023       0.059
Person Helmet_Not Worn                       25.0465   5.61e+04      0.000      1.000    -1.1e+05     1.1e+05
Person Helmet_Unknown If Worn                24.0359   5.61e+04      0.000      1.000    -1.1e+05     1.1e+05
Person Helmet_Worn, Unk Damage                8.4233   5.59e+04      0.000      1.000    -1.1e+05     1.1e+05
=============================================================================================================

Possibly complete quasi-separation: A fraction 0.94 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Figure 1a: Confusion Matrix for the Logistic Regression

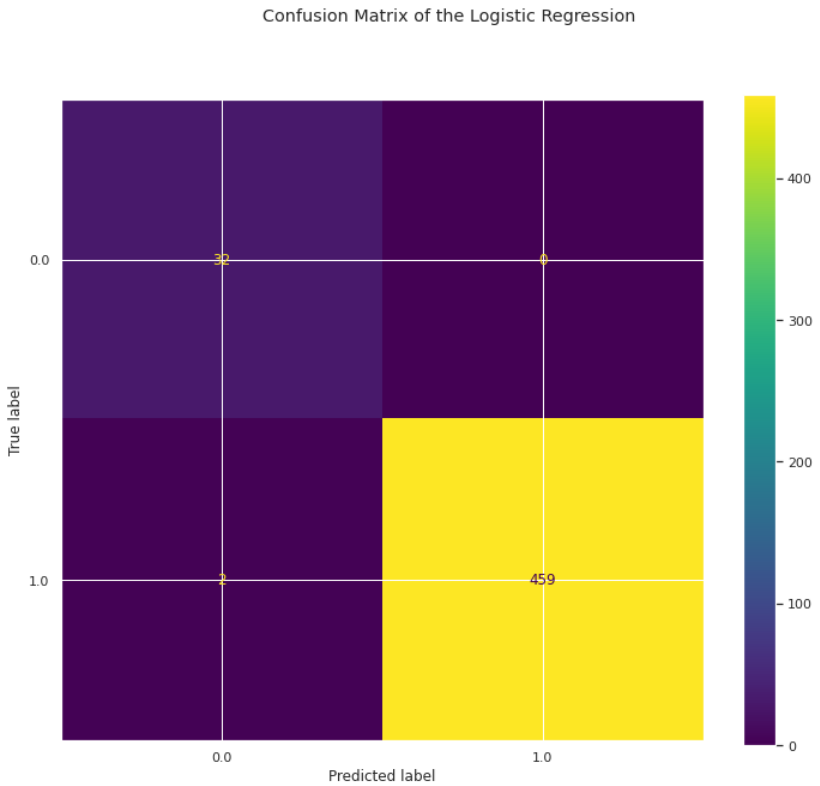

Figure 1b: Confusion Matrix for the Decision Tree

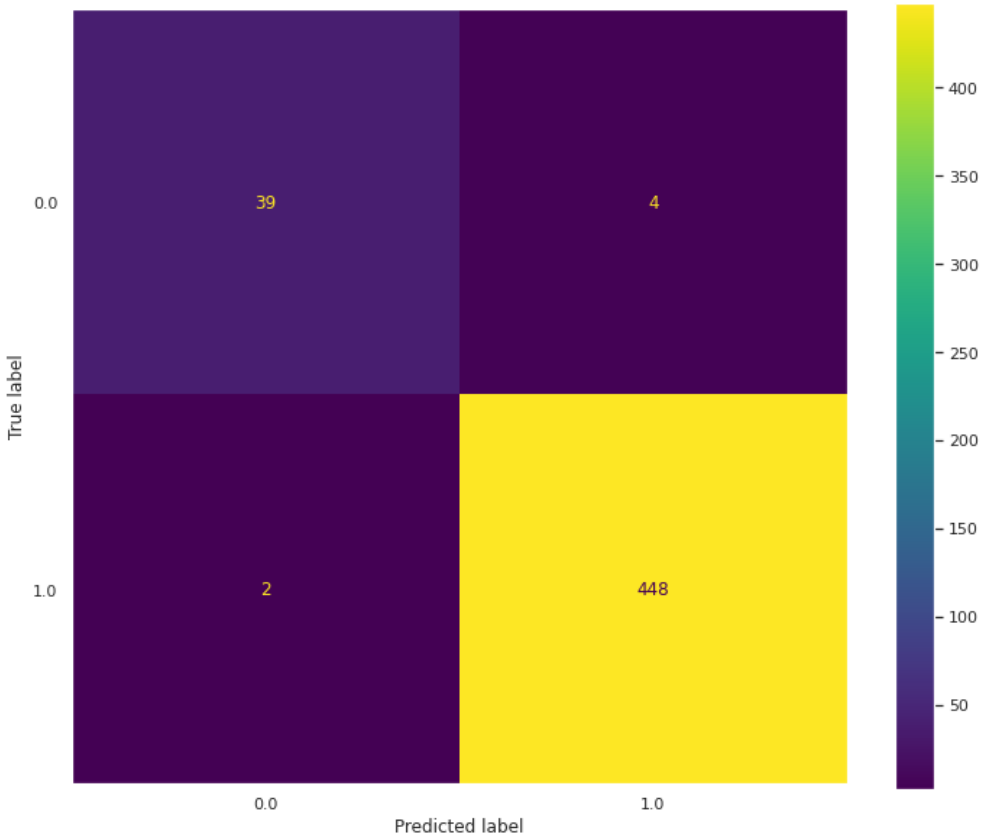

Figure 1c: Confusion Matrix for the Random Forest

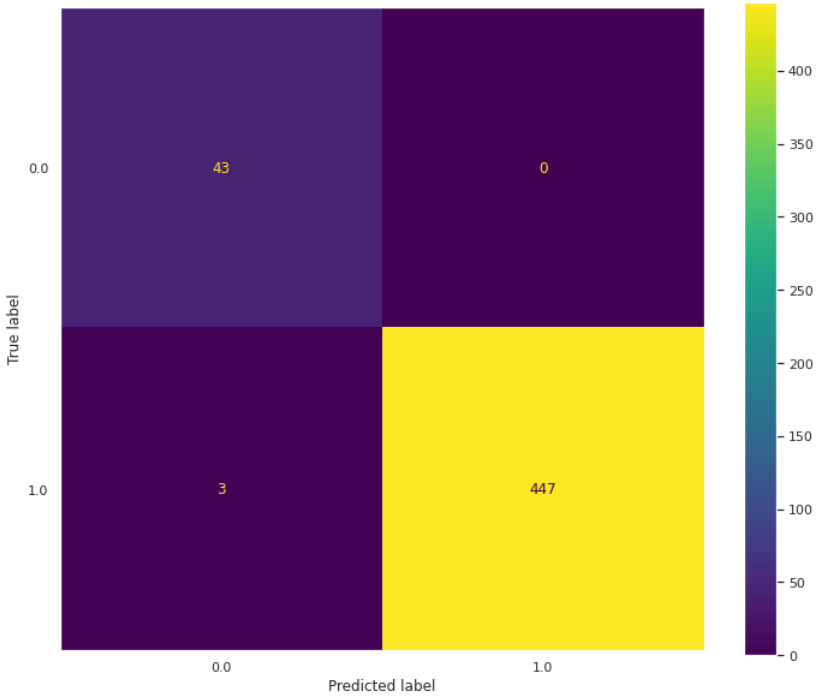

Table 3: Odds Ratio Table of Logistic Regression

index	OR	z-value	2.50%	97.50%
const	4.08E-159	0.95806615	0	Infinity
$1000 Damage to Any One Person's Property	17.36379481	0.00013623	4.00644482	75.2540927
Active School Zone Flag	3.77E-15	0.99872978	0	Infinity
Construction Zone Flag	1.09E-06	0.9985604	0	Infinity
Crash Time	0.999968115	0.95037183	0.99896455	1.00097269
Crash Total Injury Count	1.50E+24	0.99752299	0	Infinity
Crash Year	1.185044007	0.25279879	0.88585339	1.5852841
Speed Limit	1.022696636	0.29996589	0.98020347	1.06703194
Person Helmet_Not Worn	172756520.1	0.99781609	0	Infinity
Person Helmet_Unknown If Worn	76787010.73	0.99790945	0	Infinity
Person Helmet_Worn, Unk Damage	178405343.8	0.99781238	0	Infinity

Figure 2. Decision Tree with 4 branches

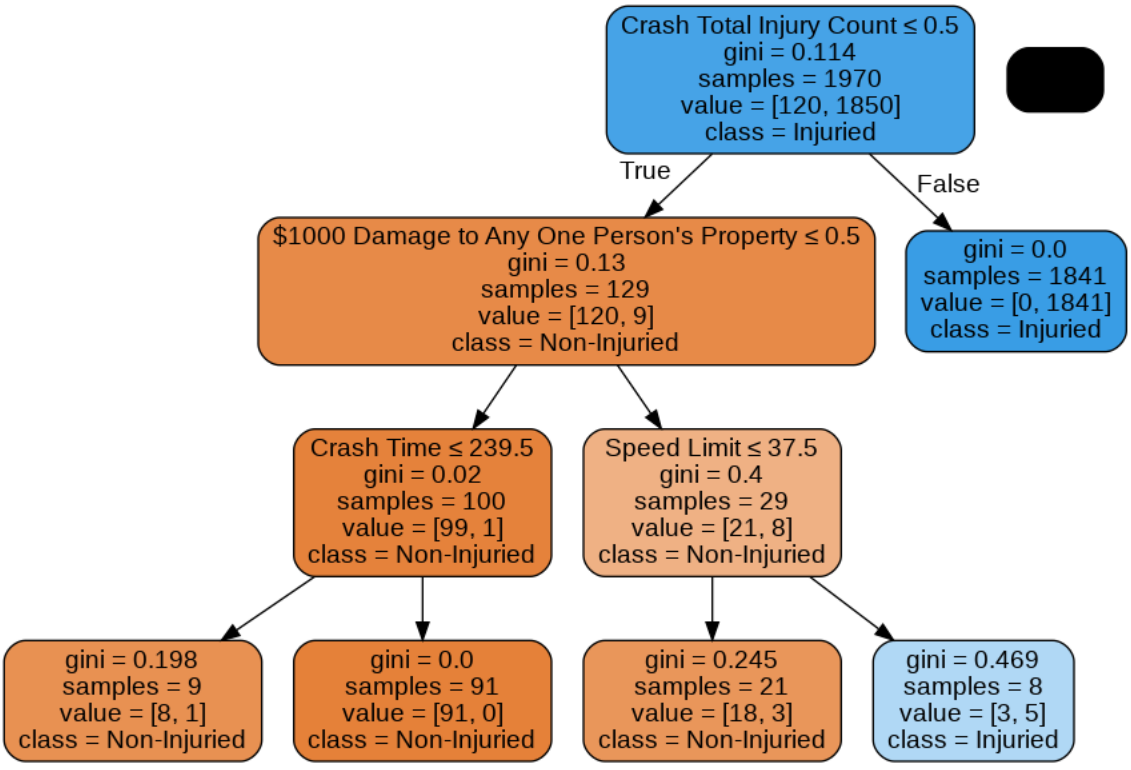

Figure 3: Feature Importance for Decision Tree

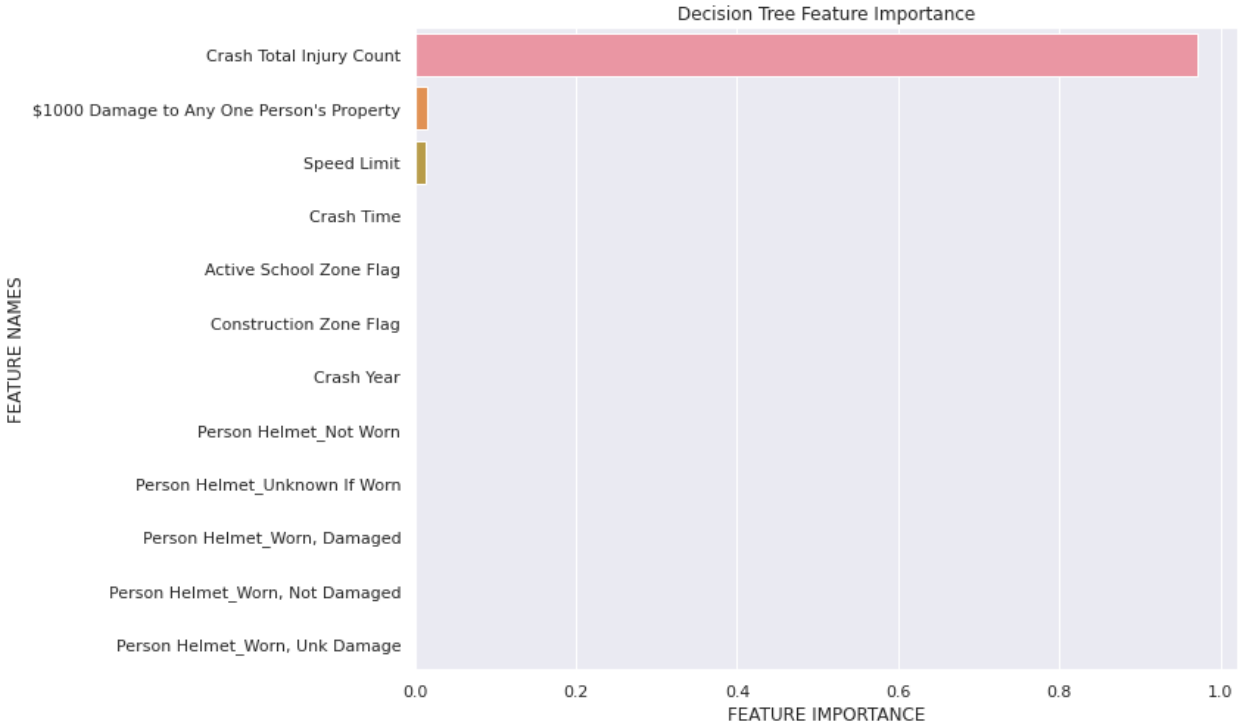

Figure 4: Feature Importance without Crash Count for Decision Tree

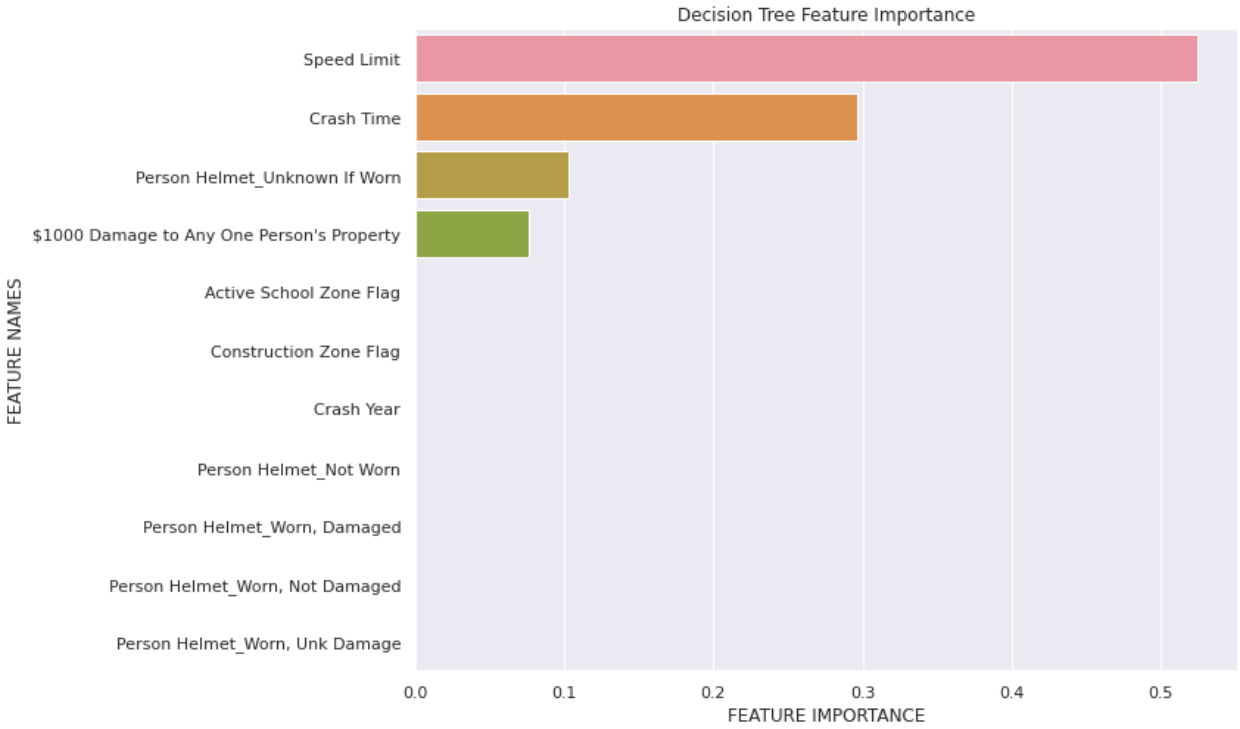

Figure 5: Feature Importance for Random Forest

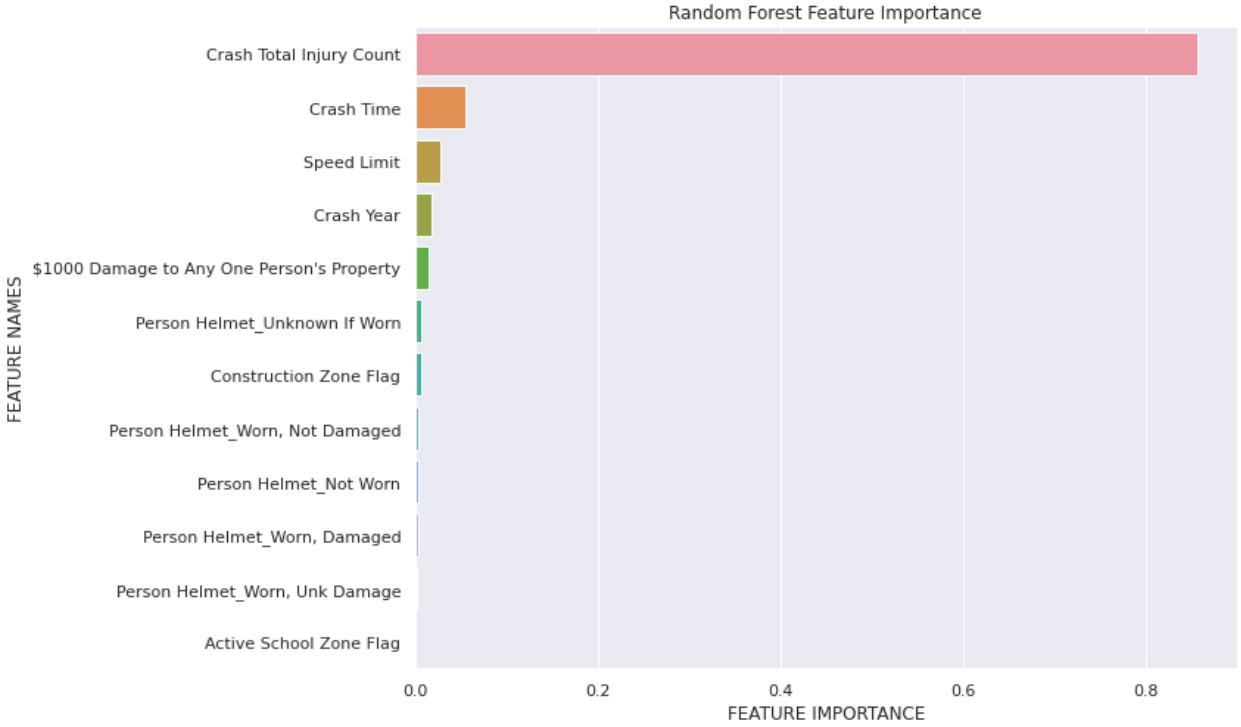

Figure 6: Feature Importance without Crash Count for Random Forest

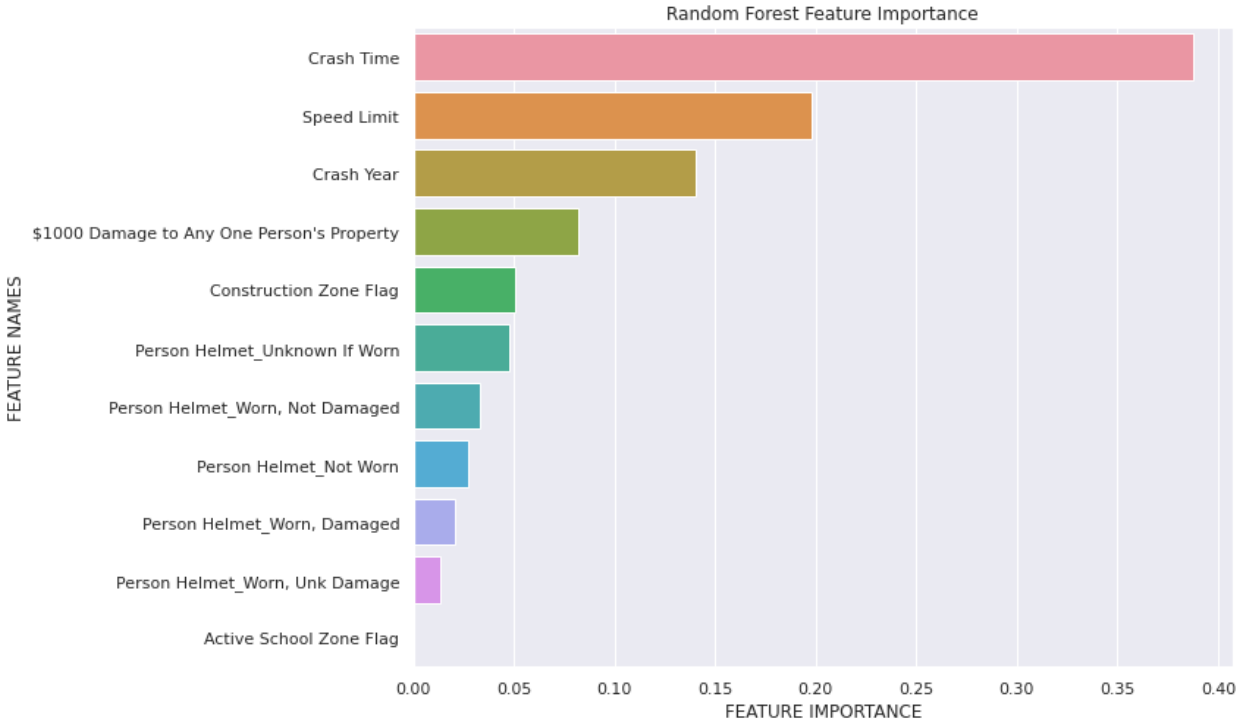

Figure 7: Correlation Matrix of the Remaining Engineered Features

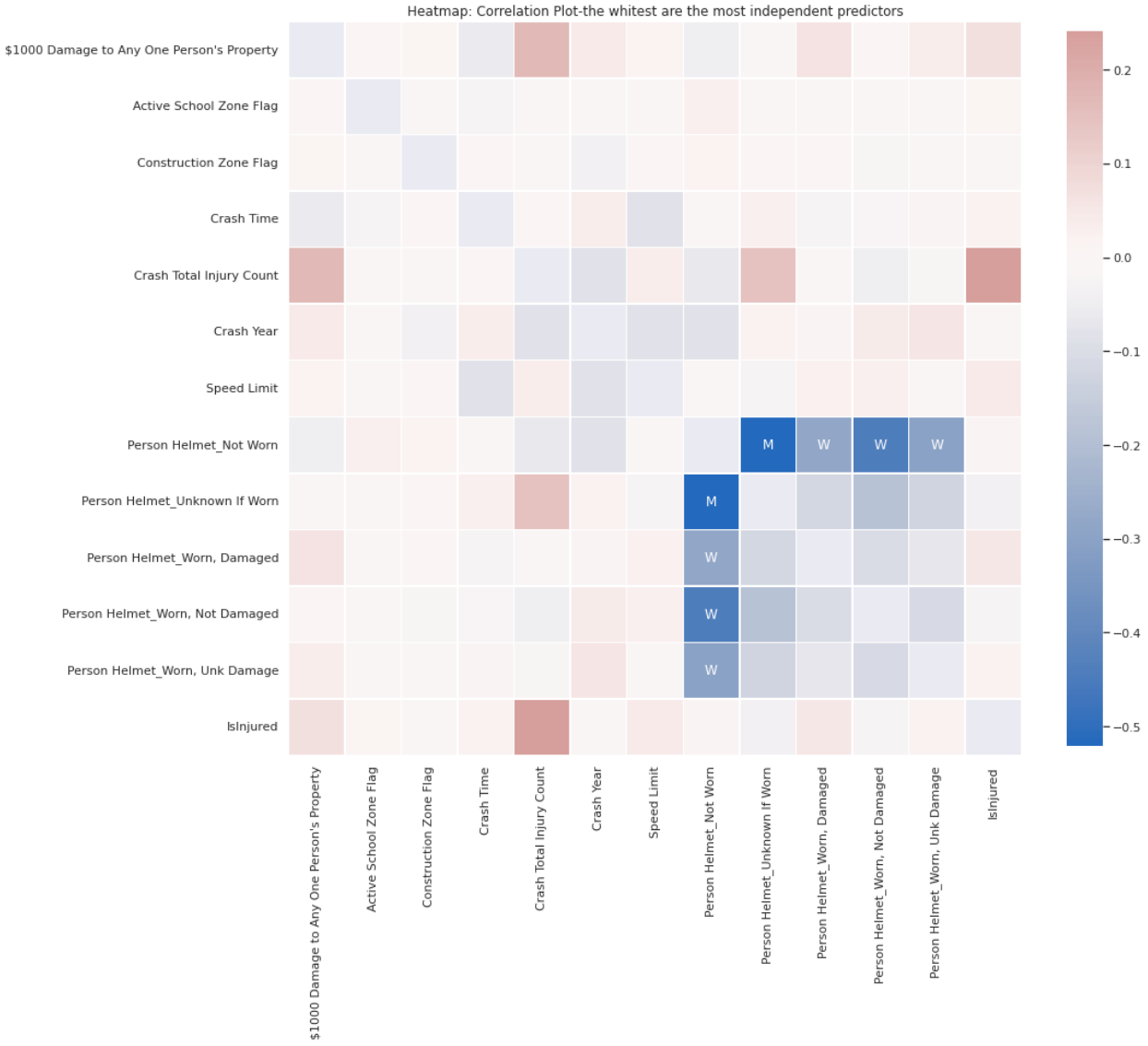

Figure 8: Single Tree from the Random Forest

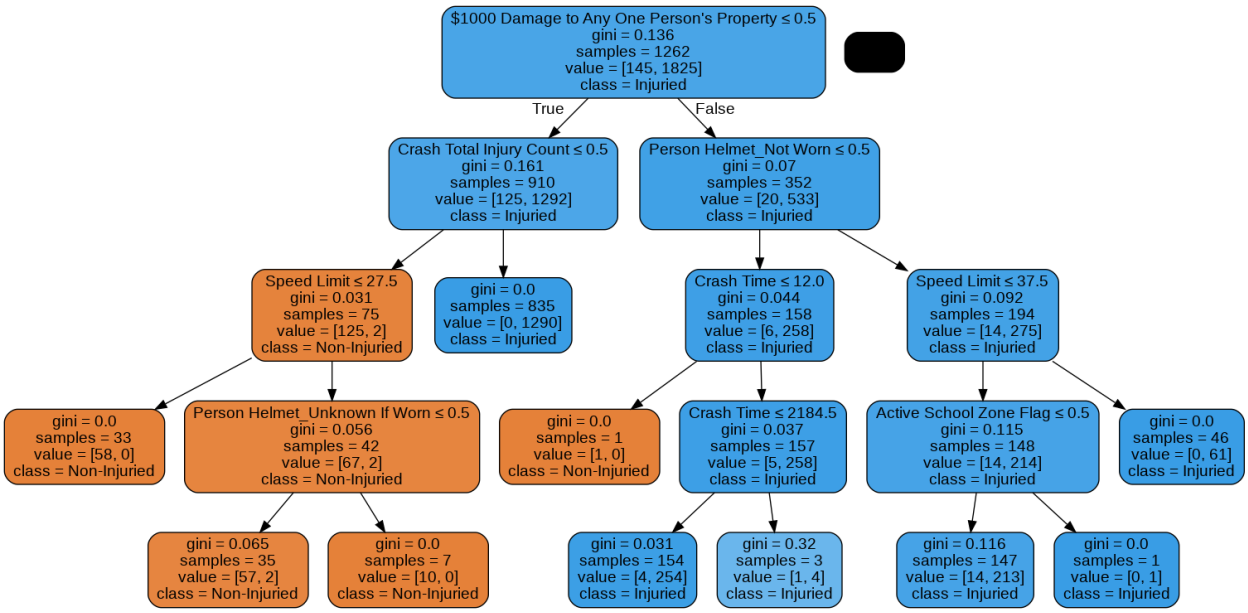
$1000 Damage to Any One Person's Property ≤ 0.5
gini = 0.136
samples = 1262
value = [145, 1825]
class = Injuried
True
False
Crash Total Injury Count ≤ 0.5
gini = 0.161
samples = 910
value = [125, 1292]
class = Injuried
Person Helmet_Not Worn ≤ 0.5
gini = 0.07
samples = 352
value = [20, 533]
class = Injuried
Speed Limit ≤ 27.5
gini = 0.031
samples = 75
value = [125, 2]
class = Non-Injuried
gini = 0.0
samples = 835
value = [0, 1290]
class = Injuried
Crash Time ≤ 12.0
gini = 0.044
samples = 158
value = [6, 258]
class = Injuried
Speed Limit ≤ 37.5
gini = 0.092
samples = 194
value = [14, 275]
class = Injuried
gini = 0.0
samples = 33
value = [58, 0]
class = Non-Injuried
Person Helmet_Unknown If Worn ≤ 0.5
gini = 0.056
samples = 42
value = [67, 2]
class = Non-Injuried
gini = 0.0
samples = 1
value = [1, 0]
class = Non-Injuried
Crash Time ≤ 2184.5
gini = 0.037
samples = 157
value = [5, 258]
class = Injuried
Active School Zone Flag ≤ 0.5
gini = 0.115
samples = 148
value = [14, 214]
class = Injuried
gini = 0.0
samples = 46
value = [0, 61]
class = Injuried
gini = 0.065
samples = 35
value = [57, 2]
class = Non-Injuried
gini = 0.0
samples = 7
value = [10, 0]
class = Non-Injuried
gini = 0.031
samples = 154
value = [4, 254]
class = Injuried
gini = 0.32
samples = 3
value = [1, 4]
class = Injuried
gini = 0.116
samples = 147
value = [14, 213]
class = Injuried
gini = 0.0
samples = 1
value = [0, 1]
class = Injuried

Table 5: Logistic Regression Classification Report

Logistic Regression Classification Report

	precision	recall	f1-score	support
0.0	0.91	1.00	0.96	43
1.0	1.00	0.99	1.00	450
accuracy			0.99	493
macro avg	0.96	1.00	0.98	493
weighted avg	0.99	0.99	0.99	493

Table 6: Decision Tree Classification Report

Decision Tree Classification Report

	precision	recall	f1-score	support
0.0	0.95	0.93	0.94	43
1.0	0.99	1.00	0.99	450
accuracy			0.99	493
macro avg	0.97	0.96	0.97	493
weighted avg	0.99	0.99	0.99	493

Table 7: Random Forest Classicisation Report

Random Forest Classification Report

	precision	recall	f1-score	support
0.0	0.93	1.00	0.97	43
1.0	1.00	0.99	1.00	450
accuracy			0.99	493
macro avg	0.97	1.00	0.98	493
weighted avg	0.99	0.99	0.99	493

Figure 9: Missing Number Graph