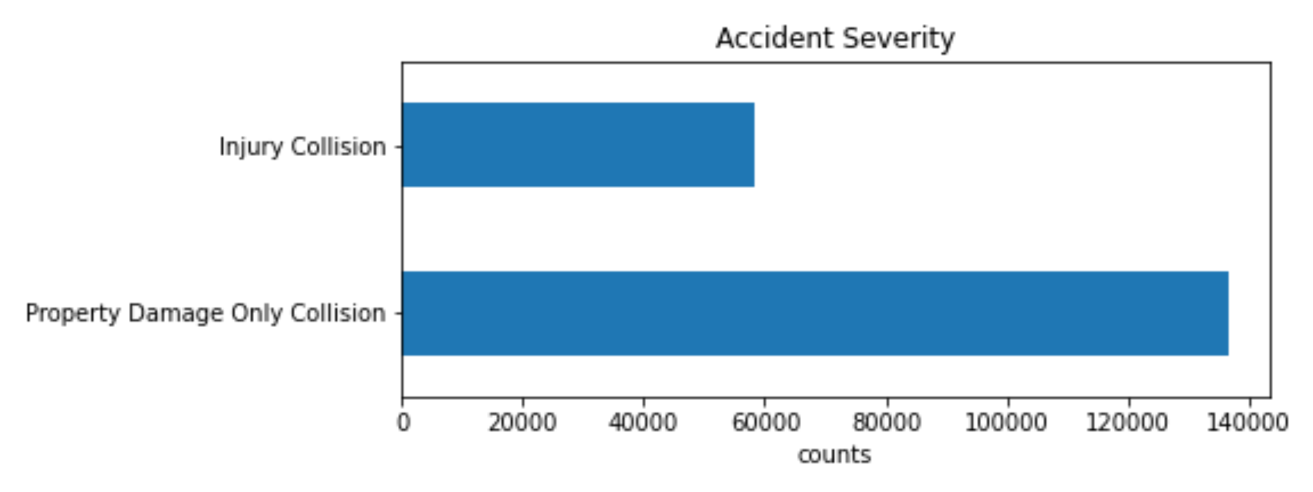
PREDICTING ACCIDENT SEVERITY

Analisa Hill

September 20th 2020

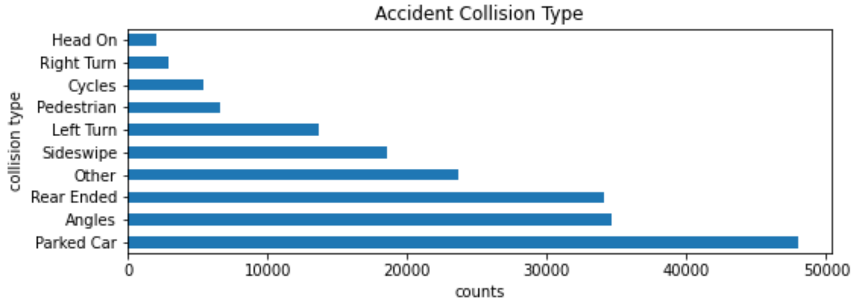
ACCIDENT SEVERITY BREAK-DOWN

- Non-injury collision occurs ~70% of the time versus an injury collision ~30% of the time.
- Unbalanced dataset
- Future: add data to make it more balanced



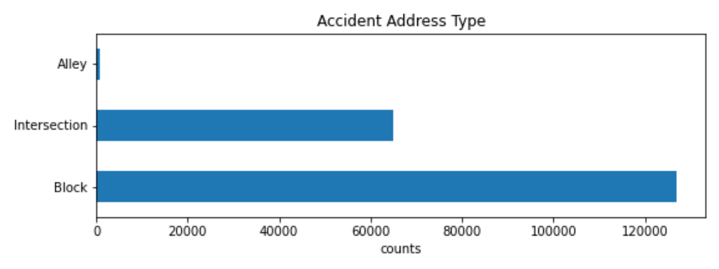
SUMMARY OF ACCIDENT COLLISION TYPE

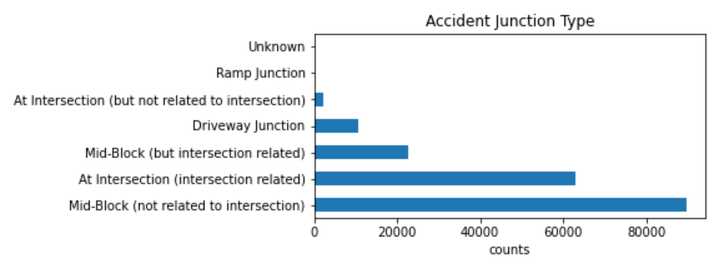
- Collision types for this dataset are labeled as: head on, right turn, unknown, cycles, pedestrian, left turn, sideswipe, other, rear ended, angles and parked car.
- Accidents involving parked cars occur 25.3%
- Accidents at an angle (18.3%), and then rear-ended (~18.0%).
- The lowest collision types were head-on (1.0%), right-turn (1.6%), and cycles (2.9%).



ROADWAY JUNCTION AND ADDRESS BREAK DOWN

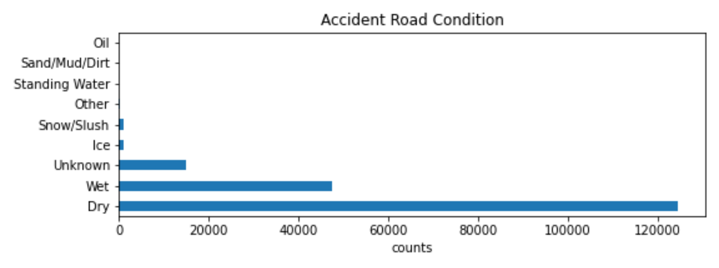
- 66% of the time the accident occurs on a block
- 34% at an intersection
- Less than 1% in an alley
- 48% of these accidents are mid-block and not related to intersections, which coincides with the address type data
- Second prevalent is at an intersection or intersection related, which is also confirmed by the address

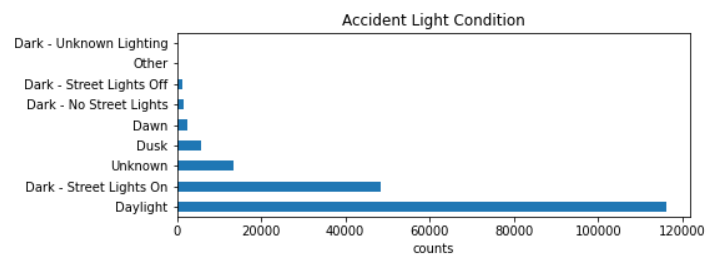




WEATHER, ROAD, AND LIGHT CONDITIONS

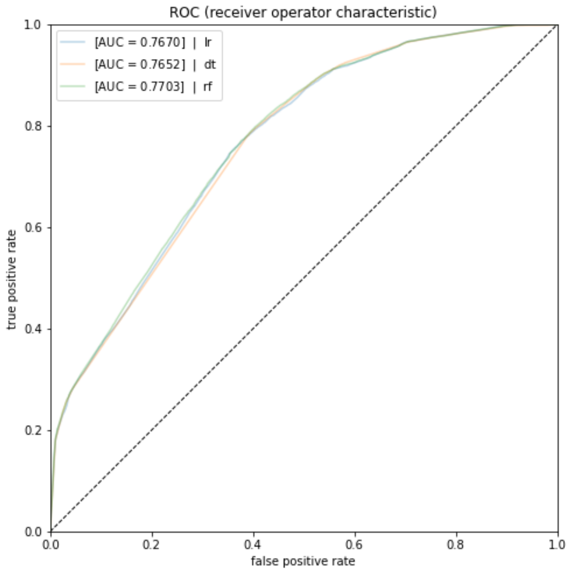
- Surprisingly, 66% of the accidents occurred when the road was dry and 25% of the time when the road was wet.
- The weather conditions also confirmed what is seen in the road conditions.
- 59% of the accidents occurred during clear weather and 17% occurred when there was rain.
- The weather is a little more informative than the road condition data because it also gives the option of overcast which contributes almost 15% of the accidents..

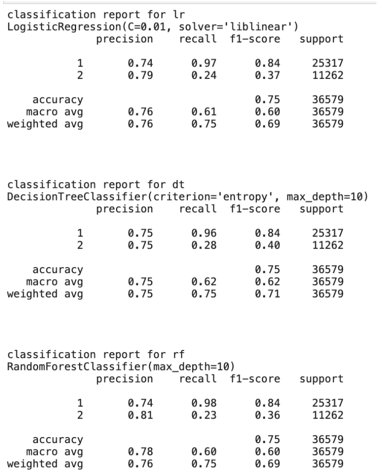




CLASSIFICATION REPORT FOR DIFFERENT MODELS

- Models used: logistic regression, decision tree, random forest
- For each model, the ROC (receiver operator characteristic) curves and AUC (area under the curve) were evaluated.
- The ROC curve calculates the true positive rate versus false positive rate of our classifier. The ROC of each model are comparable.
- The AUC sheds a little more information. The closer the AUC is to one, the better classifier it is.
- Thus, the random forest is a little better than the decision tree and logistic regression with an AUC coming in at 0.7703.

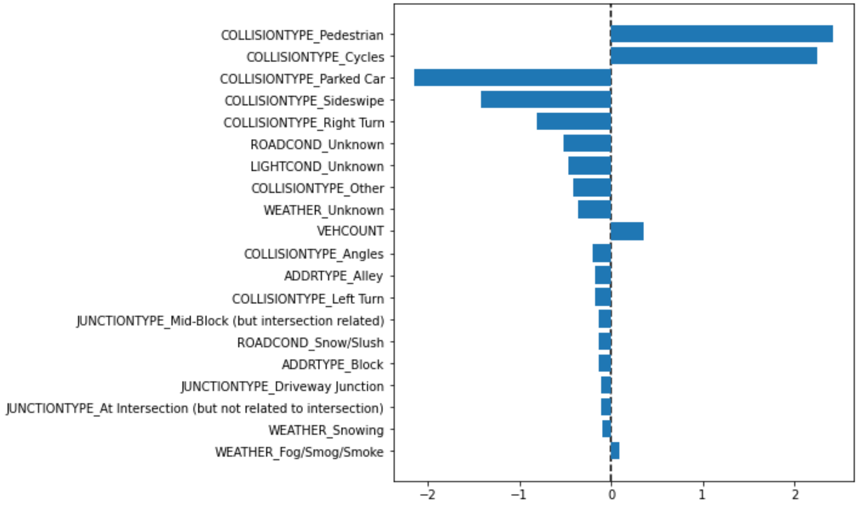


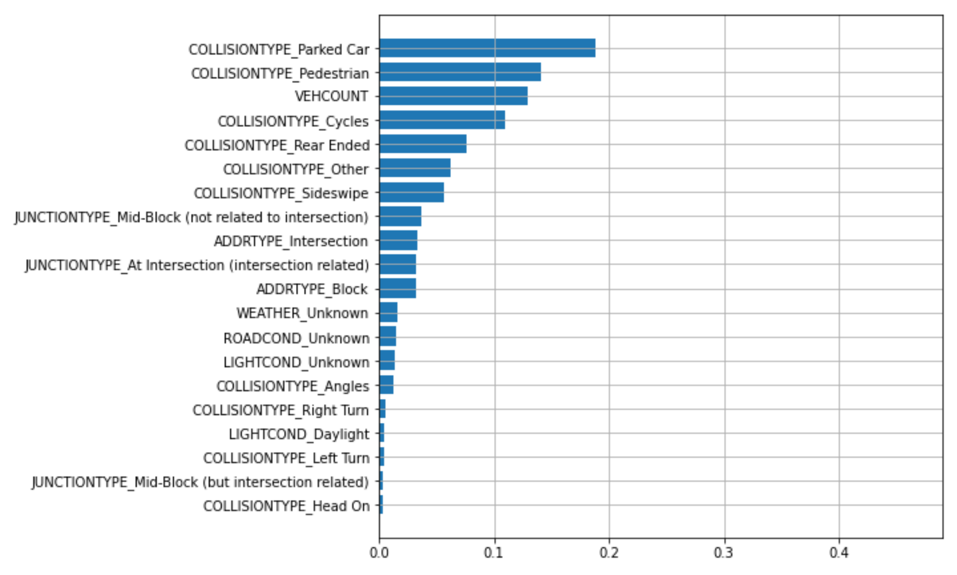


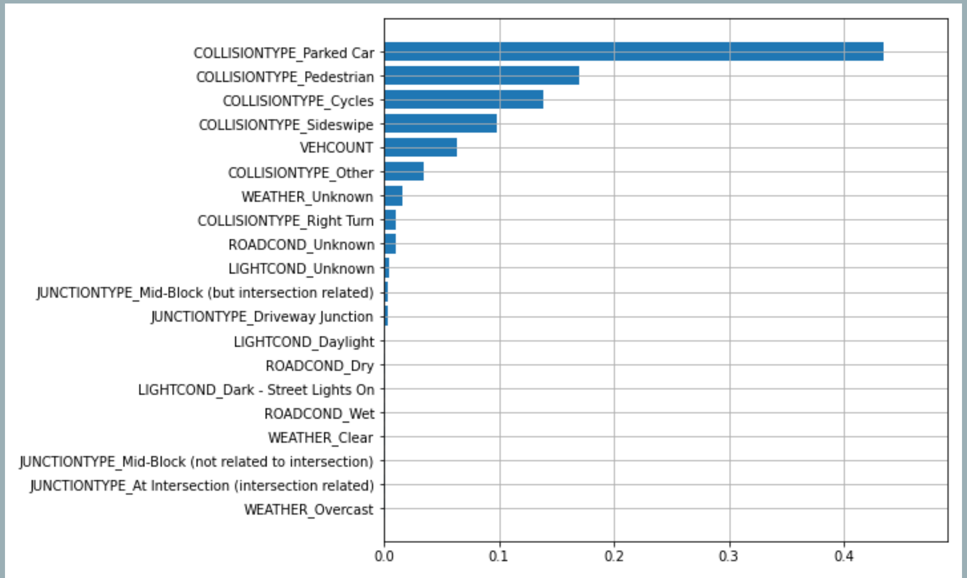
```
classification report for lr
LogisticRegression(C=0.01, solver='liblinear')
              precision    recall  f1-score   support

           1       0.74      0.97      0.84     25317
           2       0.79      0.24      0.37     11262

    accuracy                           0.75     36579
   macro avg       0.76      0.61      0.60     36579
weighted avg       0.76      0.75      0.69     36579



classification report for dt
DecisionTreeClassifier(criterion='entropy', max_depth=10)
              precision    recall  f1-score   support

           1       0.75      0.96      0.84     25317
           2       0.75      0.28      0.40     11262

    accuracy                           0.75     36579
   macro avg       0.75      0.62      0.62     36579
weighted avg       0.75      0.75      0.71     36579



classification report for rf
RandomForestClassifier(max_depth=10)
              precision    recall  f1-score   support

           1       0.74      0.98      0.84     25317
           2       0.81      0.23      0.36     11262

    accuracy                           0.75     36579
   macro avg       0.78      0.60      0.60     36579
weighted avg       0.76      0.75      0.69     36579
```

IMPORTANT FEATURES ACQUIRED FROM LOGISTIC REGRESSION MODEL

- Injury accidents are caused when the collision type involves a pedestrian or cyclist.
- Third largest contributing factor is the vehicle count. These are shown by the positive values on the bar plot.
- It makes logical sense that the more vehicles involved in an accident means that there is a higher probability of getting injured because there are more people involved.
- It also makes sense that a non-injury incident happens most with parked cars.
- This non-injury accident is determined by the negative value on the bar plot.

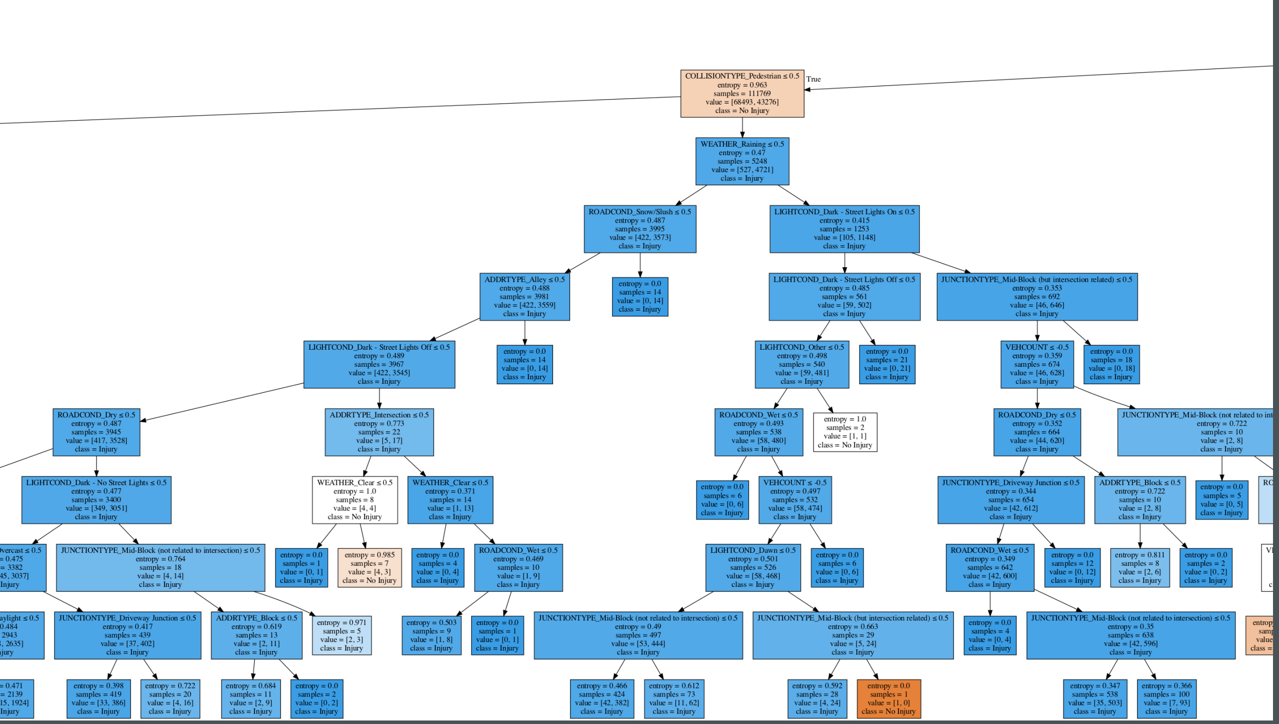


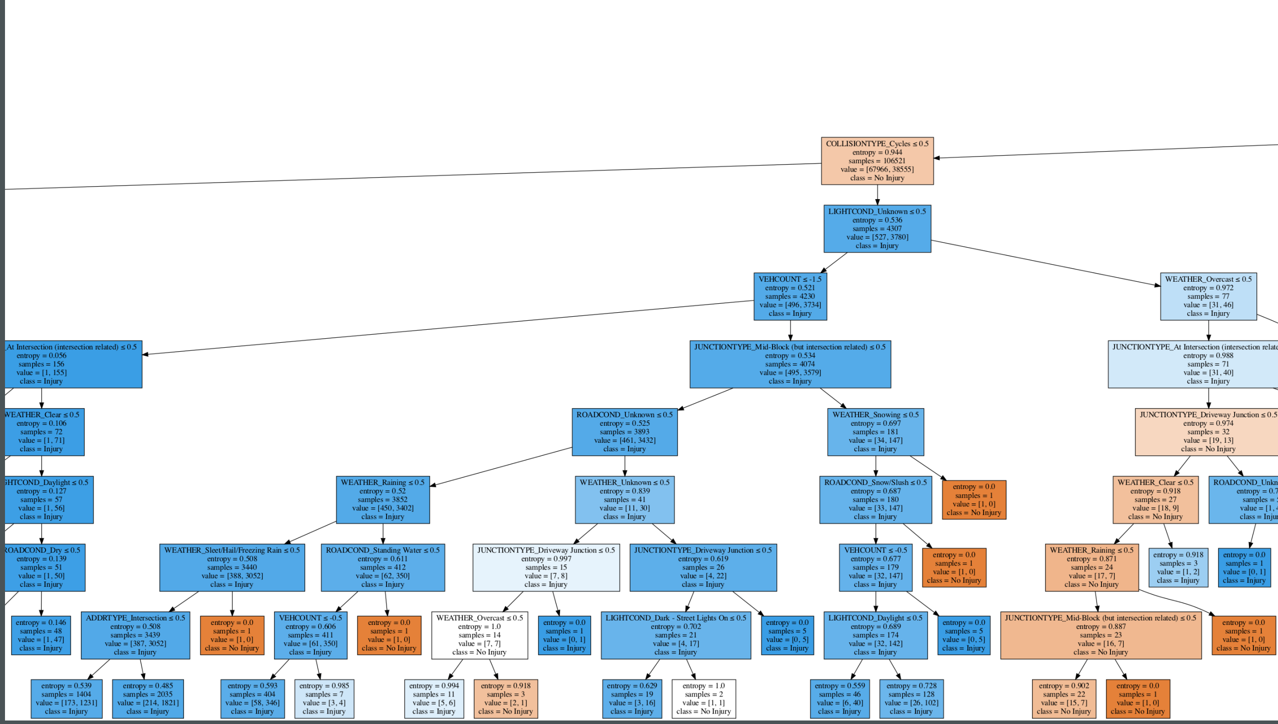




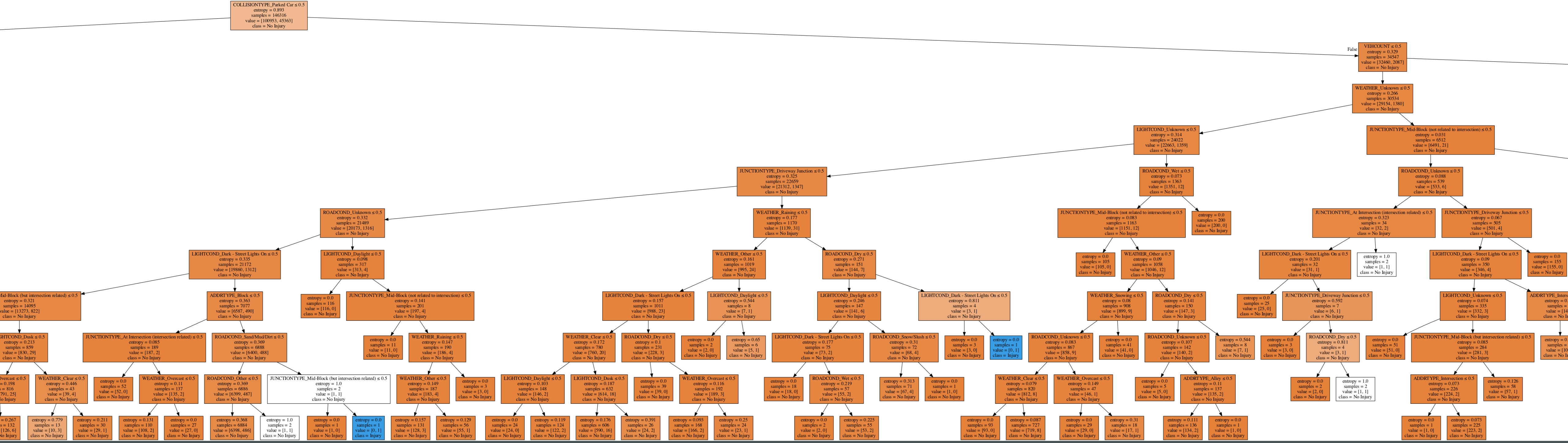
IMPORTANT FEATURES ACQUIRED FROM DECISION TREE AND RANDOM FOREST MODELS

- These bar plots tell us that parked car and pedestrian collision type plays the largest role in deciding whether it was an injury or no injury accident.
- This is similar to the decision tree and almost nearly the same, but are slightly different because weighted of the ensemble nature of the model, but gives us the same trend as with the decision tree.
- This confirms what we found in the logistic regression model, although the logistic regression model tells us more about which features contribute directly from it's negative and positive values in the bar plot.





DECISION TREE AND WHAT LEADS TO INJURY


COLLISIONTYPE_Parked Car ≤ 0.5 entropy = 0.893 samples = 146316 value = [100953, 45363] class = No Injury
False
VEHCOUNT ≤ 0.5 entropy = 0.329 samples = 34547 value = [32460, 2087] class = No Injury
WEATHER_Unknown ≤ 0.5 entropy = 0.266 samples = 30534 value = [29154, 1380] class = No Injury
LIGHTCOND_Unknown ≤ 0.5 entropy = 0.314 samples = 24022 value = [22663, 1359] class = No Injury
JUNCTIONTYPE_Mid-Block (not related to intersection) ≤ 0.5 entropy = 0.031 samples = 6512 value = [6491, 21] class = No Injury
JUNCTIONTYPE_Driveway Junction ≤ 0.5 entropy = 0.325 samples = 22659 value = [21312, 1347] class = No Injury
ROADCOND_Wet ≤ 0.5 entropy = 0.073 samples = 1363 value = [1351, 12] class = No Injury
ROADCOND_Unknown ≤ 0.5 entropy = 0.088 samples = 539 value = [533, 6] class = No Injury
ROADCOND_Unknown ≤ 0.5 entropy = 0.332 samples = 21489 value = [20173, 1316] class = No Injury
WEATHER_Raining ≤ 0.5 entropy = 0.177 samples = 1170 value = [1139, 31] class = No Injury
JUNCTIONTYPE_Mid-Block (not related to intersection) ≤ 0.5 entropy = 0.083 samples = 1163 value = [1151, 12] class = No Injury
entropy = 0.0 samples = 200 value = [200, 0] class = No Injury
JUNCTIONTYPE_At Intersection (intersection related) ≤ 0.5 entropy = 0.323 samples = 34 value = [32, 2] class = No Injury
JUNCTIONTYPE_Driveway Junction ≤ 0.5 entropy = 0.067 samples = 505 value = [501, 4] class = No Injury
LIGHTCOND_Dark - Street Lights On ≤ 0.5 entropy = 0.335 samples = 21172 value = [19860, 1312] class = No Injury
LIGHTCOND_Daylight ≤ 0.5 entropy = 0.098 samples = 317 value = [313, 4] class = No Injury
WEATHER_Other ≤ 0.5 entropy = 0.161 samples = 1019 value = [995, 24] class = No Injury
ROADCOND_Dry ≤ 0.5 entropy = 0.271 samples = 151 value = [144, 7] class = No Injury
entropy = 0.0 samples = 105 value = [105, 0] class = No Injury
WEATHER_Other ≤ 0.5 entropy = 0.09 samples = 1058 value = [1046, 12] class = No Injury
LIGHTCOND_Dark - Street Lights On ≤ 0.5 entropy = 0.201 samples = 32 value = [31, 1] class = No Injury
entropy = 1.0 samples = 2 value = [1, 1] class = No Injury
LIGHTCOND_Dark - Street Lights On ≤ 0.5 entropy = 0.09 samples = 350 value = [346, 4] class = No Injury
entropy = 0.0 samples = 155 value = [155, 0] class = No Injury
Mid-Block (but intersection related) ≤ 0.5 entropy = 0.321 samples = 14095 value = [13273, 822] class = No Injury
ADDRTYPE_Block ≤ 0.5 entropy = 0.363 samples = 7077 value = [6587, 490] class = No Injury
entropy = 0.0 samples = 116 value = [116, 0] class = No Injury
JUNCTIONTYPE_Mid-Block (not related to intersection) ≤ 0.5 entropy = 0.141 samples = 201 value = [197, 4] class = No Injury
LIGHTCOND_Dark - Street Lights On ≤ 0.5 entropy = 0.157 samples = 1011 value = [988, 23] class = No Injury
LIGHTCOND_Daylight ≤ 0.5 entropy = 0.544 samples = 8 value = [7, 1] class = No Injury
LIGHTCOND_Daylight ≤ 0.5 entropy = 0.246 samples = 147 value = [141, 6] class = No Injury
LIGHTCOND_Dark - Street Lights On ≤ 0.5 entropy = 0.811 samples = 4 value = [3, 1] class = No Injury
WEATHER_Snowing ≤ 0.5 entropy = 0.08 samples = 908 value = [899, 9] class = No Injury
ROADCOND_Dry ≤ 0.5 entropy = 0.141 samples = 150 value = [147, 3] class = No Injury
entropy = 0.0 samples = 25 value = [25, 0] class = No Injury
JUNCTIONTYPE_Driveway Junction ≤ 0.5 entropy = 0.592 samples = 7 value = [6, 1] class = No Injury
LIGHTCOND_Unknown ≤ 0.5 entropy = 0.074 samples = 335 value = [332, 3] class = No Injury
ADDRTYPE_Inters entropy = 0. samples = value = [14 class = No In
GHTCOND_Dusk ≤ 0.5 entropy = 0.213 samples = 859 value = [830, 29] class = No Injury
JUNCTIONTYPE_At Intersection (intersection related) ≤ 0.5 entropy = 0.085 samples = 189 value = [187, 2] class = No Injury
ROADCOND_Sand/Mud/Dirt ≤ 0.5 entropy = 0.369 samples = 6888 value = [6400, 488] class = No Injury
entropy = 0.0 samples = 11 value = [11, 0] class = No Injury
WEATHER_Raining ≤ 0.5 entropy = 0.147 samples = 190 value = [186, 4] class = No Injury
WEATHER_Clear ≤ 0.5 entropy = 0.172 samples = 780 value = [760, 20] class = No Injury
ROADCOND_Dry ≤ 0.5 entropy = 0.1 samples = 231 value = [228, 3] class = No Injury
entropy = 0.0 samples = 2 value = [2, 0] class = No Injury
entropy = 0.65 samples = 6 value = [5, 1] class = No Injury
LIGHTCOND_Dark - Street Lights On ≤ 0.5 entropy = 0.177 samples = 75 value = [73, 2] class = No Injury
ROADCOND_Snow/Slush ≤ 0.5 entropy = 0.31 samples = 72 value = [68, 4] class = No Injury
entropy = 0.0 samples = 3 value = [3, 0] class = No Injury
entropy = 0.0 samples = 1 value = [0, 1] class = Injury
ROADCOND_Unknown ≤ 0.5 entropy = 0.083 samples = 867 value = [858, 9] class = No Injury
entropy = 0.0 samples = 41 value = [41, 0] class = No Injury
ROADCOND_Unknown ≤ 0.5 entropy = 0.107 samples = 142 value = [140, 2] class = No Injury
entropy = 0.544 samples = 8 value = [7, 1] class = No Injury
entropy = 0.0 samples = 3 value = [3, 0] class = No Injury
ROADCOND_Dry ≤ 0.5 entropy = 0.811 samples = 4 value = [3, 1] class = No Injury
entropy = 0.0 samples = 51 value = [51, 0] class = No Injury
JUNCTIONTYPE_Mid-Block (but intersection related) ≤ 0.5 entropy = 0.085 samples = 284 value = [281, 3] class = No Injury
entropy = 0. samples = value = [10 class = No In
Overcast ≤ 0.5 = 0.198 = 816 791, 25] No Injury
WEATHER_Clear ≤ 0.5 entropy = 0.446 samples = 43 value = [39, 4] class = No Injury
entropy = 0.0 samples = 52 value = [52, 0] class = No Injury
WEATHER_Overcast ≤ 0.5 entropy = 0.11 samples = 137 value = [135, 2] class = No Injury
ROADCOND_Other ≤ 0.5 entropy = 0.369 samples = 6886 value = [6399, 487] class = No Injury
JUNCTIONTYPE_Mid-Block (but intersection related) ≤ 0.5 entropy = 1.0 samples = 2 value = [1, 1] class = No Injury
WEATHER_Other ≤ 0.5 entropy = 0.149 samples = 187 value = [183, 4] class = No Injury
entropy = 0.0 samples = 3 value = [3, 0] class = No Injury
LIGHTCOND_Daylight ≤ 0.5 entropy = 0.103 samples = 148 value = [146, 2] class = No Injury
LIGHTCOND_Dusk ≤ 0.5 entropy = 0.187 samples = 632 value = [614, 18] class = No Injury
entropy = 0.0 samples = 39 value = [39, 0] class = No Injury
WEATHER_Overcast ≤ 0.5 entropy = 0.116 samples = 192 value = [189, 3] class = No Injury
entropy = 0.0 samples = 18 value = [18, 0] class = No Injury
ROADCOND_Wet ≤ 0.5 entropy = 0.219 samples = 57 value = [55, 2] class = No Injury
entropy = 0.313 samples = 71 value = [67, 4] class = No Injury
entropy = 0.0 samples = 1 value = [1, 0] class = No Injury
WEATHER_Clear ≤ 0.5 entropy = 0.079 samples = 820 value = [812, 8] class = No Injury
WEATHER_Overcast ≤ 0.5 entropy = 0.149 samples = 47 value = [46, 1] class = No Injury
entropy = 0.0 samples = 5 value = [5, 0] class = No Injury
ADDRTYPE_Alley ≤ 0.5 entropy = 0.11 samples = 137 value = [135, 2] class = No Injury
entropy = 0.0 samples = 2 value = [2, 0] class = No Injury
entropy = 1.0 samples = 2 value = [1, 1] class = No Injury
ADDRTYPE_Intersection ≤ 0.5 entropy = 0.073 samples = 226 value = [224, 2] class = No Injury
entropy = 0.126 samples = 58 value = [57, 1] class = No Injury
= 0.267 s = 132 [126, 6] No Injury
entropy = 0.779 samples = 13 value = [10, 3] class = No Injury
entropy = 0.211 samples = 30 value = [29, 1] class = No Injury
entropy = 0.131 samples = 110 value = [108, 2] class = No Injury
entropy = 0.0 samples = 27 value = [27, 0] class = No Injury
entropy = 0.368 samples = 6884 value = [6398, 486] class = No Injury
entropy = 1.0 samples = 2 value = [1, 1] class = No Injury
entropy = 0.0 samples = 1 value = [1, 0] class = No Injury
entropy = 0.0 samples = 1 value = [0, 1] class = Injury
entropy = 0.157 samples = 131 value = [128, 3] class = No Injury
entropy = 0.129 samples = 56 value = [55, 1] class = No Injury
entropy = 0.0 samples = 24 value = [24, 0] class = No Injury
entropy = 0.119 samples = 124 value = [122, 2] class = No Injury
entropy = 0.176 samples = 606 value = [590, 16] class = No Injury
entropy = 0.391 samples = 26 value = [24, 2] class = No Injury
entropy = 0.093 samples = 168 value = [166, 2] class = No Injury
entropy = 0.25 samples = 24 value = [23, 1] class = No Injury
entropy = 0.0 samples = 2 value = [2, 0] class = No Injury
entropy = 0.225 samples = 55 value = [53, 2] class = No Injury
entropy = 0.0 samples = 93 value = [93, 0] class = No Injury
entropy = 0.087 samples = 727 value = [719, 8] class = No Injury
entropy = 0.0 samples = 29 value = [29, 0] class = No Injury
entropy = 0.31 samples = 18 value = [17, 1] class = No Injury
entropy = 0.111 samples = 136 value = [134, 2] class = No Injury
entropy = 0.0 samples = 1 value = [1, 0] class = No Injury
entropy = 0.0 samples = 1 value = [1, 0] class = No Injury
entropy = 0.073 samples = 225 value = [223, 2] class = No Injury

DECISION TREE AND WHAT LEADS TO NO INJURY

CONCLUSIONS

- In this study, I looked at accident severity to predict if someone who was in an accident what is the chance that they were injured.
- The important features that I identified injury accidents are caused when the collision type that involves a pedestrian or cyclist.
- The next contributing factor is the vehicle count.
- I built a logistic regression and decision tree models which turned out to be not as robust as the random forest.
- The random forest model can help first responders determine whether they should dispatch paramedics, police, or none.
- Future directions could be finding a method to determine how many responders should be sent depending on the accident severity.