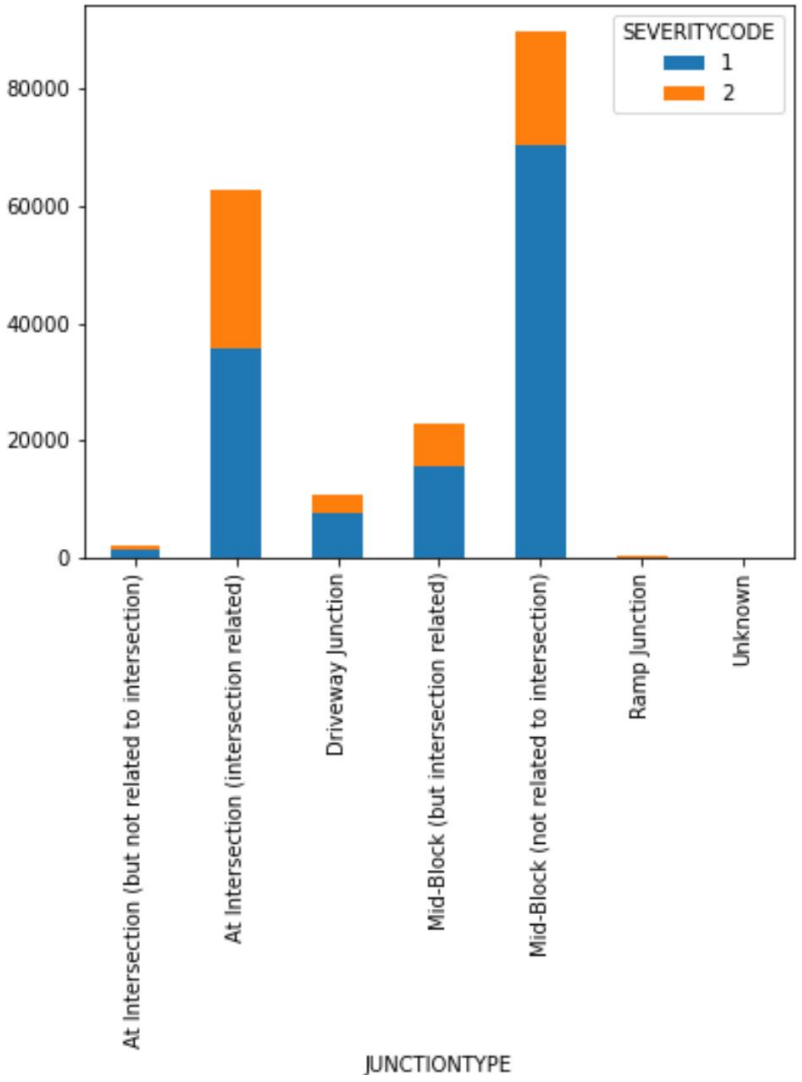
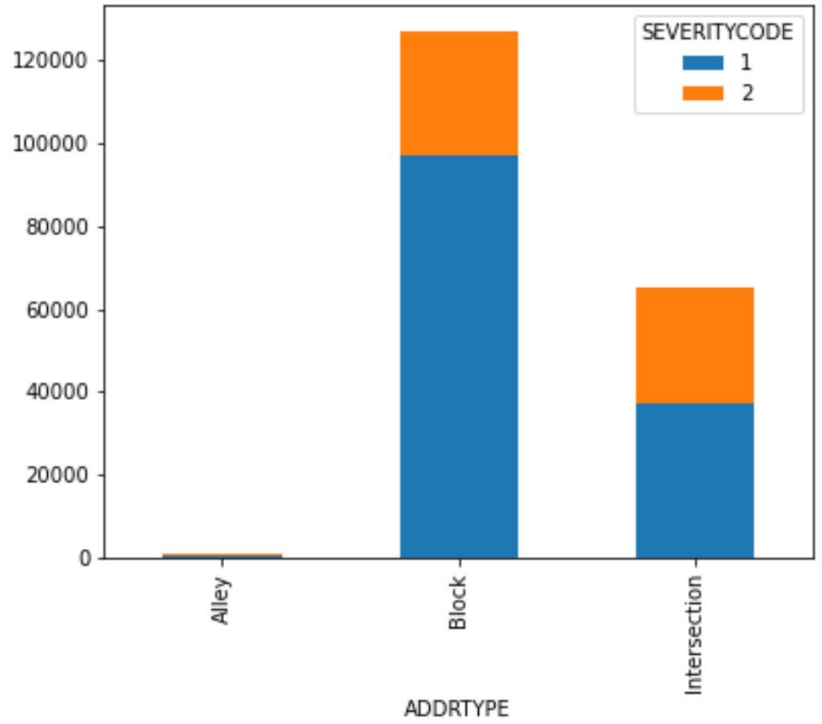
# Coursera Capstone Project
# Predicting accident severity in Seattle

Data from: collisions data from 2004 to 2020 (38 features, 194673 samples)

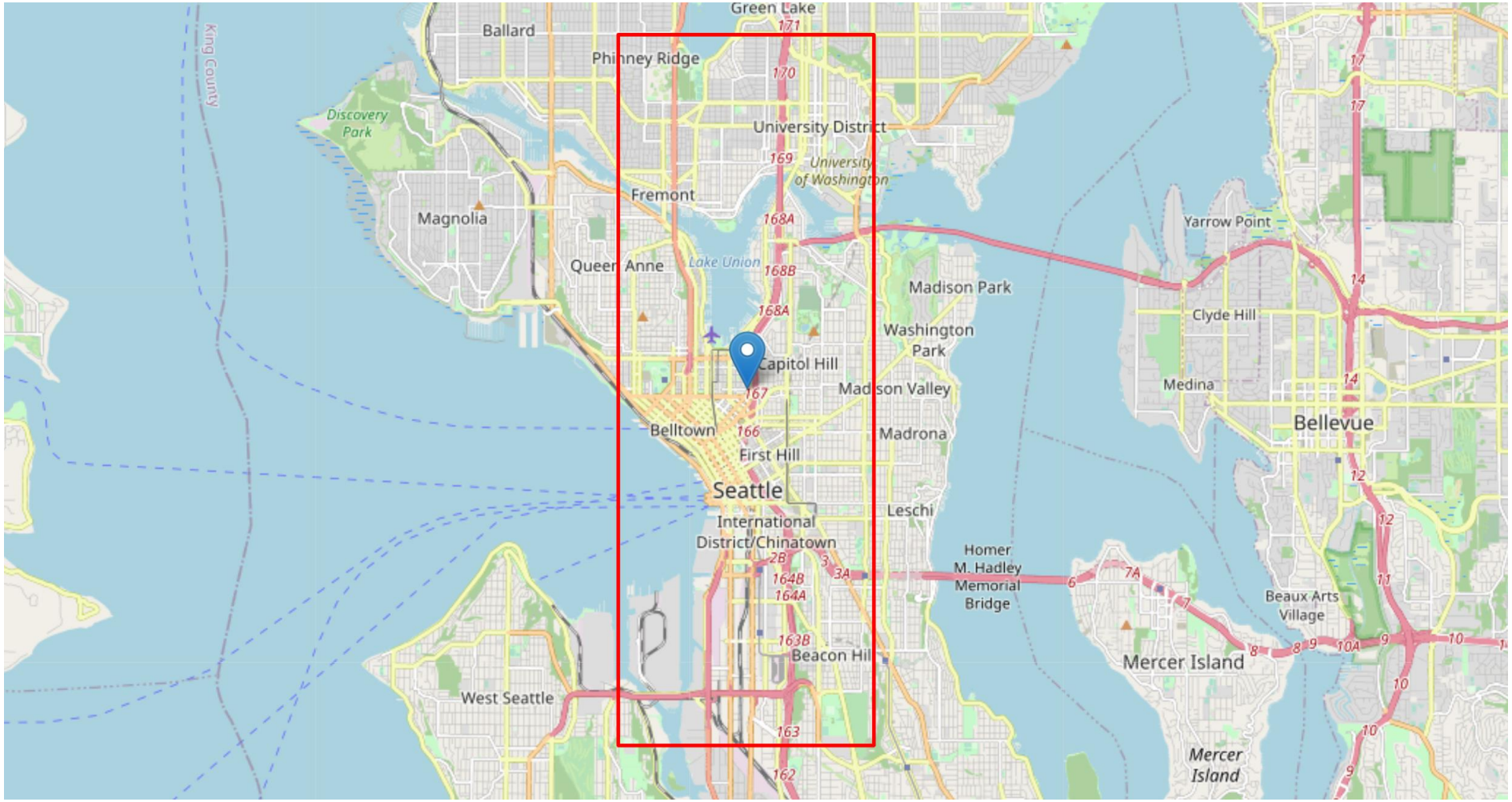
# Exploratory data analysis (1)



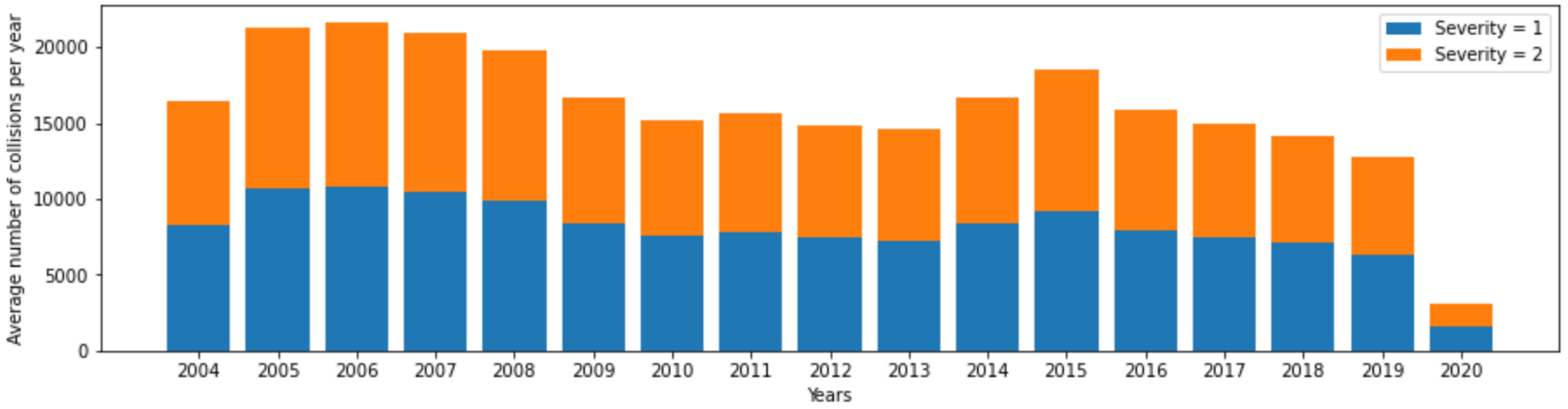

Severity vs junction and address types

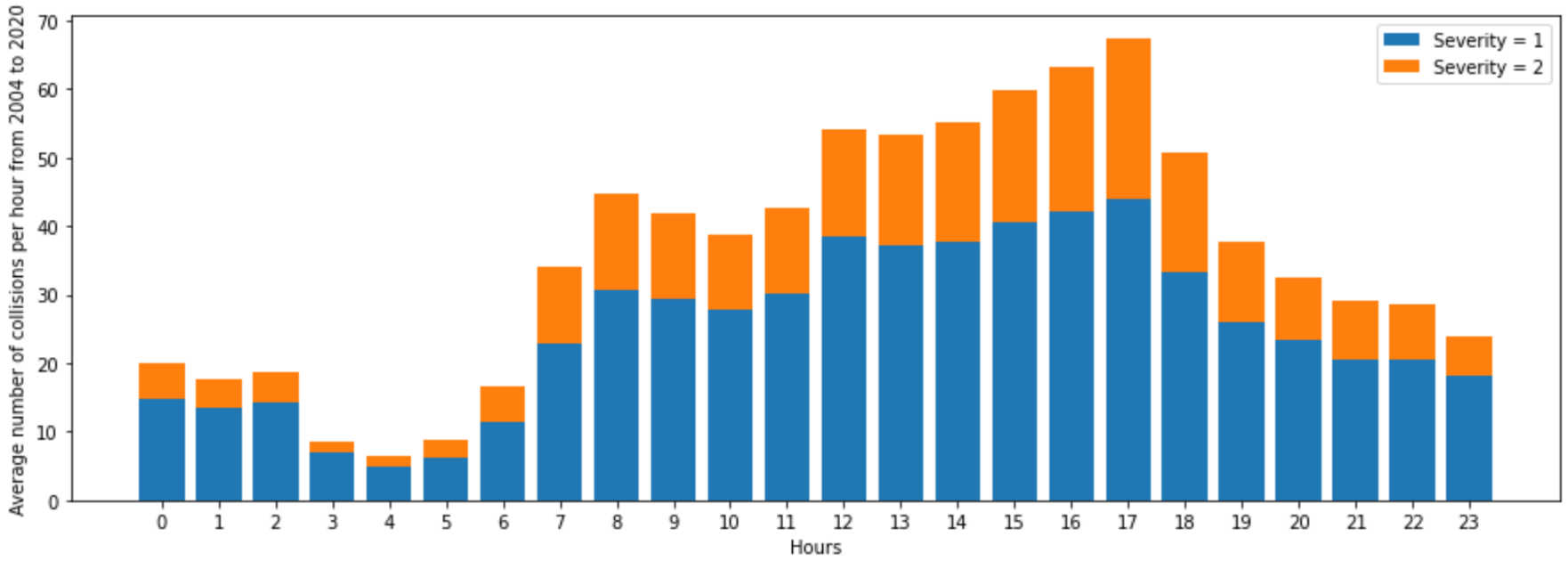
# Exploratory data analysis (2)




Average location of accidents and their standard deviation (calculated by keeping the other dimension constant)

# Exploratory data analysis (3)




Time data analysis:

- Accidents in decline from 2015 onward -> Seattle's Vision Zero initiative started in 2015
- 2 peaks: at around 8am and at around 5pm -> peak hours

# Exploratory data analysis (4)

| COLLISIONTYPE | Severity = 1 | Severity = 2 | Tot_collisions | PEDCOUNT | PEDCYLCOUNT | VEHCOUNT |
|---|---|---|---|---|---|---|
| Angles | 21050 | 13624 | 34674 | 60 | 17 | 71978 |
| Cycles | 671 | 4744 | 5415 | 98 | 5447 | 5295 |
| Head On | 1152 | 872 | 2024 | 0 | 1 | 4305 |
| Left Turn | 8292 | 5411 | 13703 | 22 | 14 | 28117 |
| Other | 17591 | 6112 | 23703 | 71 | 7 | 34276 |
| Parked Car | 45325 | 2662 | 47987 | 72 | 6 | 102983 |
| Pedestrian | 672 | 5936 | 6608 | 6857 | 4 | 6702 |
| Rear Ended | 19419 | 14671 | 34090 | 27 | 7 | 75753 |
| Right Turn | 2347 | 609 | 2956 | 5 | 3 | 5985 |
| Sideswipe | 16103 | 2506 | 18609 | 18 | 5 | 38505 |

Analyses of collisions circumstances, weather, road and light conditions, showing

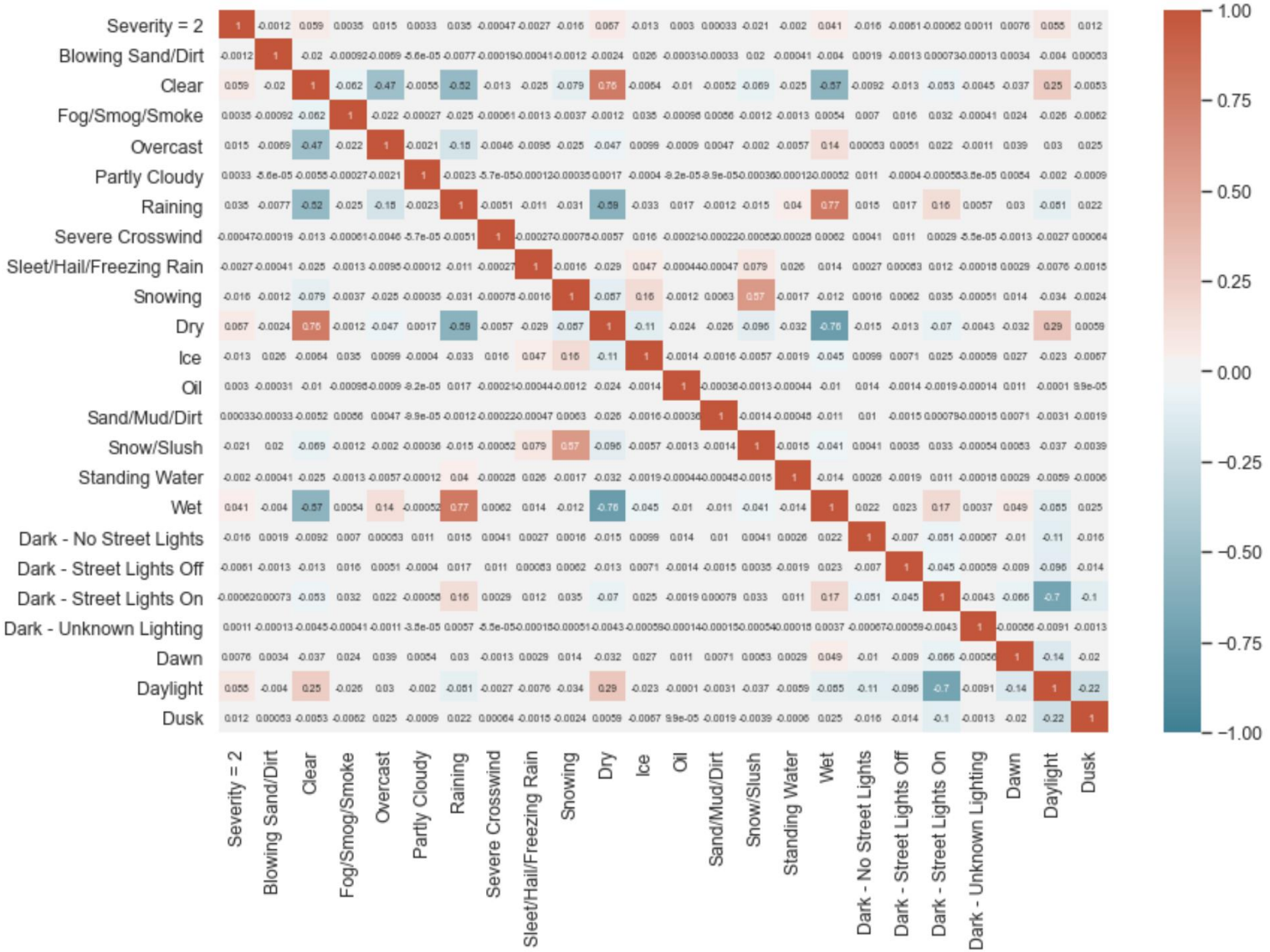
- Majority of accidents involving bicycles and pedestrians are Sev 2 accidents
- Severity 2 accidents are primarily happening with Angles and Rear-ended type collisions
- Sev 2 accidents are mainly occurring with clear weather, dry conditions and in daylight – basically all conditions when one would expect to find the majority of drivers on the roads.

| WEATHER | Severity = 1 | Severity = 2 |
|---|---|---|
| Blowing Sand/Dirt | 41 | 15 |
| Clear | 75295 | 35840 |
| Fog/Smog/Smoke | 382 | 187 |
| Other | 716 | 116 |
| Overcast | 18969 | 8745 |
| Partly Cloudy | 2 | 3 |
| Raining | 21969 | 11176 |
| Severe Crosswind | 18 | 7 |
| Sleet/Hail/Freezing Rain | 85 | 28 |
| Snowing | 736 | 171 |
| Unknown | 14275 | 816 |

| ROADCOND | Severity = 1 | Severity = 2 |
|---|---|---|
| Dry | 84446 | 40064 |
| Ice | 936 | 273 |
| Oil | 40 | 24 |
| Other | 89 | 43 |
| Sand/Mud/Dirt | 52 | 23 |
| Snow/Slush | 837 | 167 |
| Standing Water | 85 | 30 |
| Unknown | 14329 | 749 |
| Wet | 31719 | 15755 |

| LIGHTCOND | Severity = 1 | Severity = 2 |
|---|---|---|
| Dark - No Street Lights | 1203 | 334 |
| Dark - Street Lights Off | 883 | 316 |
| Dark - Street Lights On | 34032 | 14475 |
| Dark - Unknown Lighting | 7 | 4 |
| Dawn | 1678 | 824 |
| Daylight | 77593 | 38544 |
| Dusk | 3958 | 1944 |
| Other | 183 | 52 |
| Unknown | 12868 | 605 |

# Exploratory data analysis (5)




Correlation matrix showing no strong 1 to 1 correlations between Severity 2 collisions and collision type, weather, road or light conditions

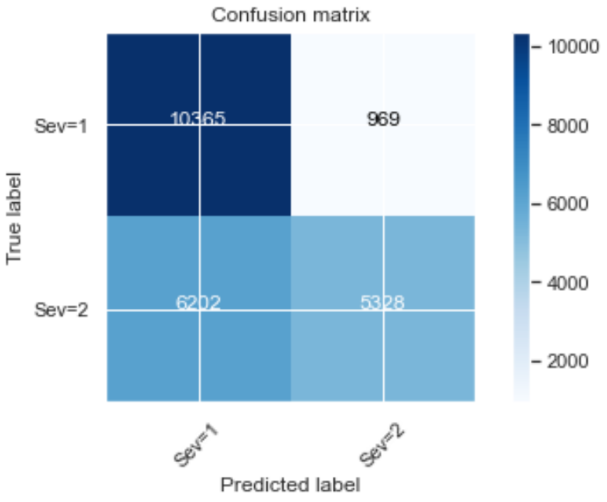
# ML application: Decision tree, random forest, logistic regression (left to right)

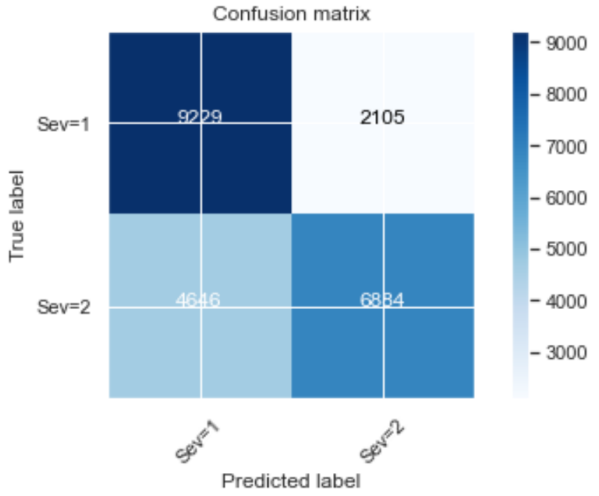
DECISION TREE

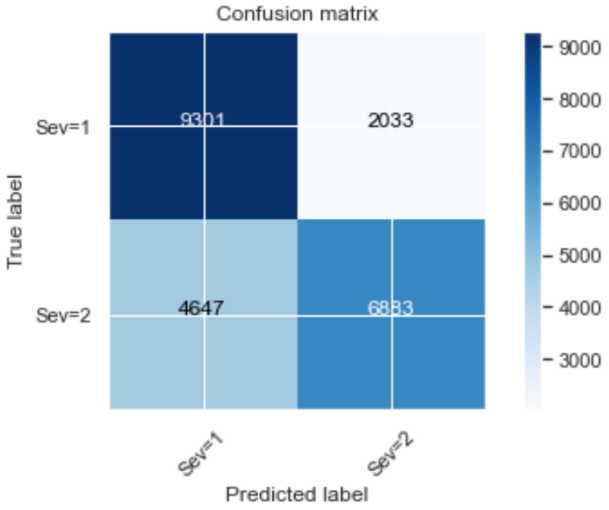
RANDOM FOREST

LOGISTIC REGRESSION

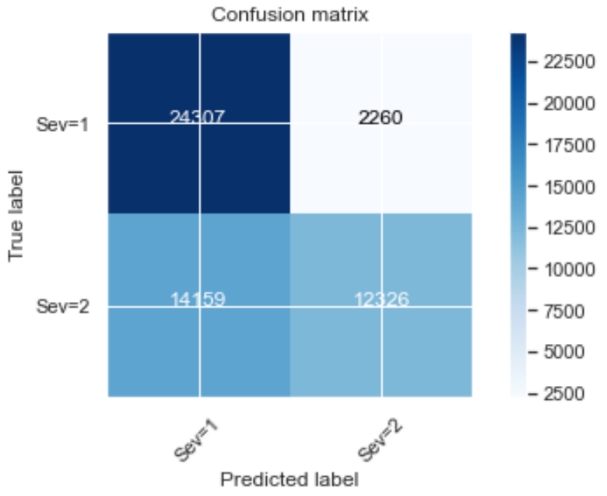
Under-sampling

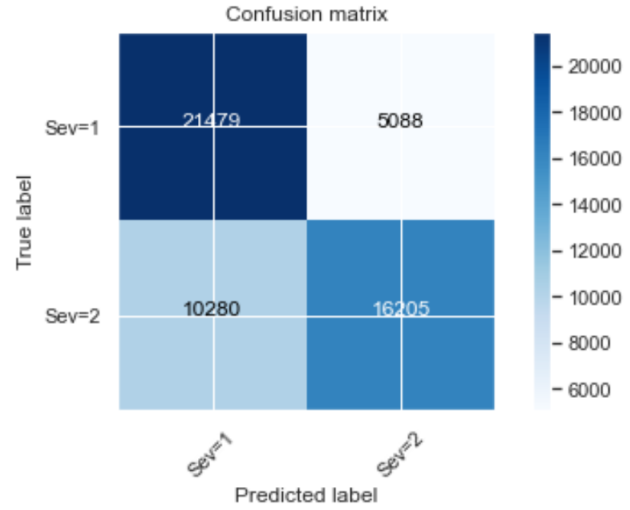


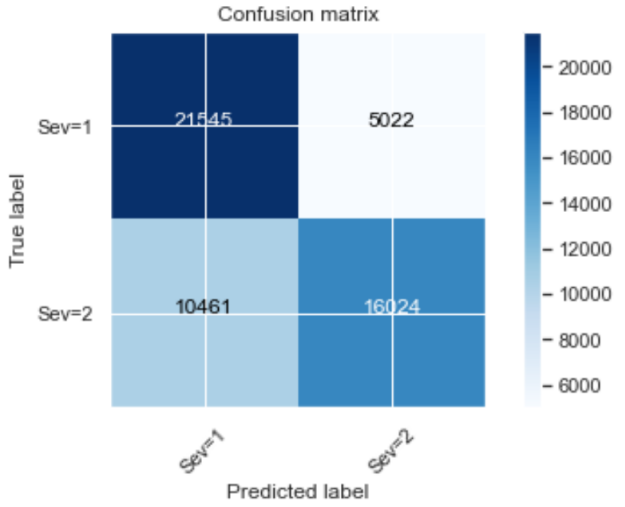




Over-sampling

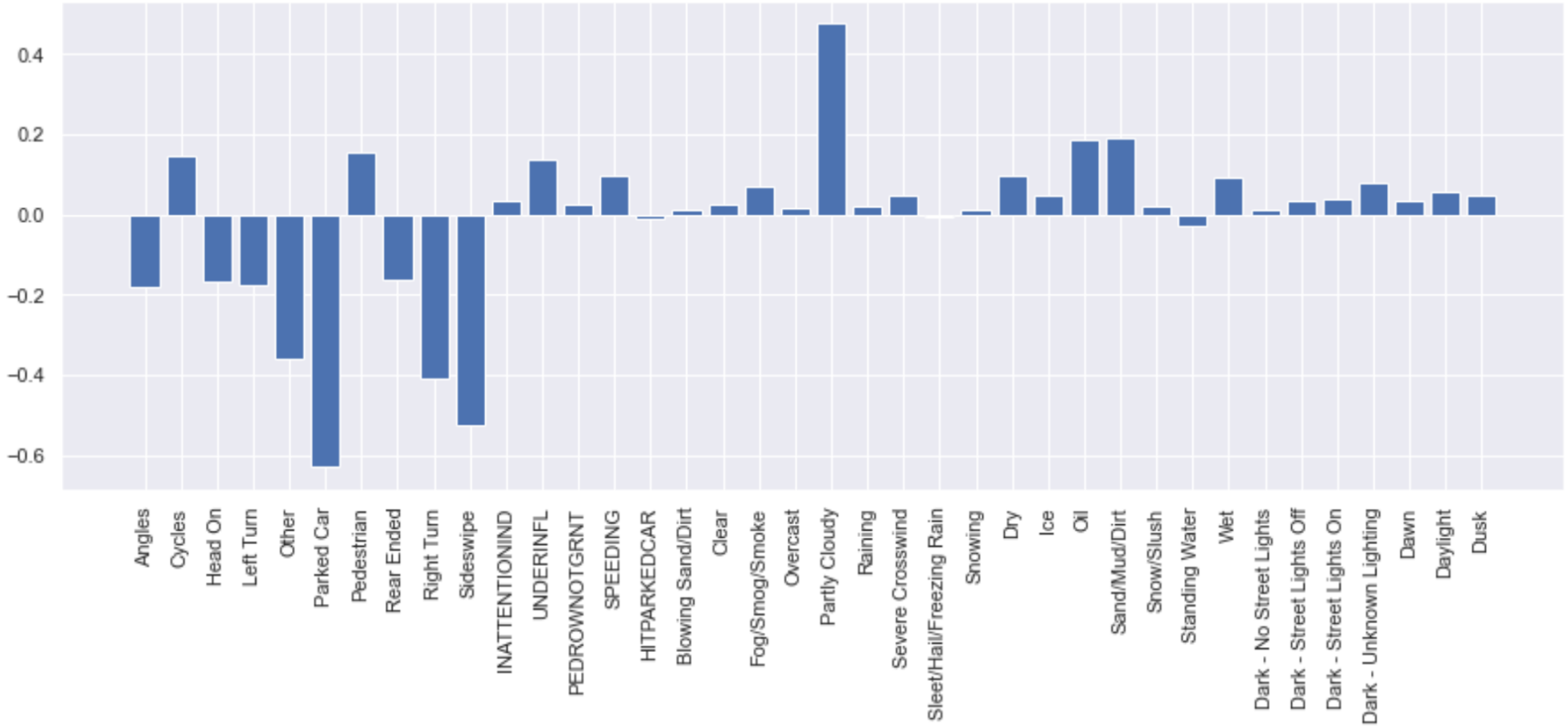



These confusion matrices show the efficacy of each machine learning algorithm adopted. In particular they shown the number of false positives, false negatives, true positives and true negatives found by the ML algorithms. They finally show how random forest and logistic regression perform better than decision tree.

# ML application: Linear regression




This last plot shows the coefficients importance in the linear regression model created to understand inputs' weights on outputs. However, the model shows quite a high mean squared error which means that this linear regression model should be revised and further investigated before drawing a robust conclusion.