Collision Data Analysis - Seattle

Vishal Dongare

September 13, 2020

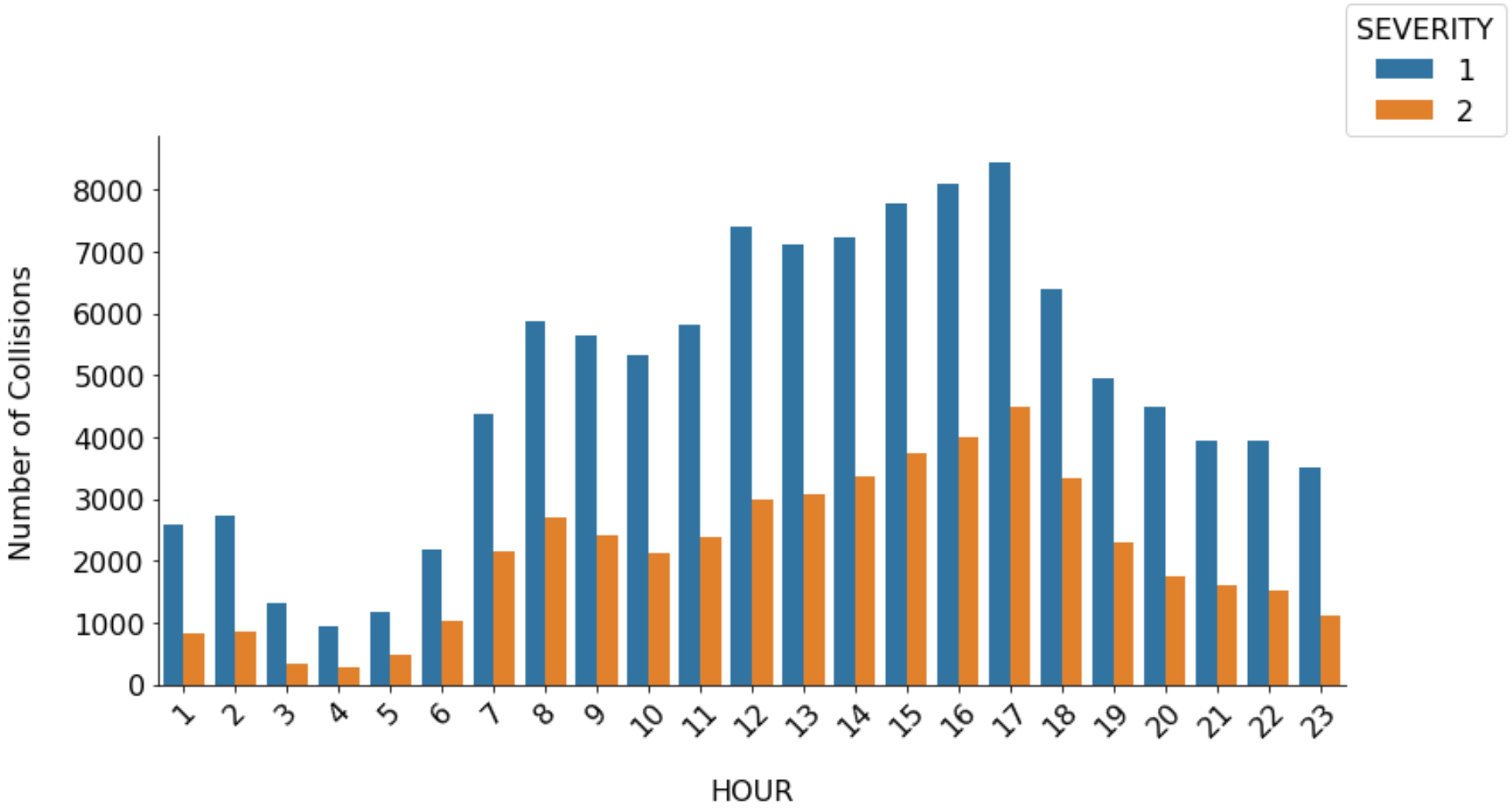
Introduction

- Seattle is well known for its tough driving conditions.
- Collision data analysis can produce some useful insights
- Helpful to Seattle police and traffic departments
- Machine learning model can be built to predict severity of a collision based on various factors, such as weather, road conditions etc.

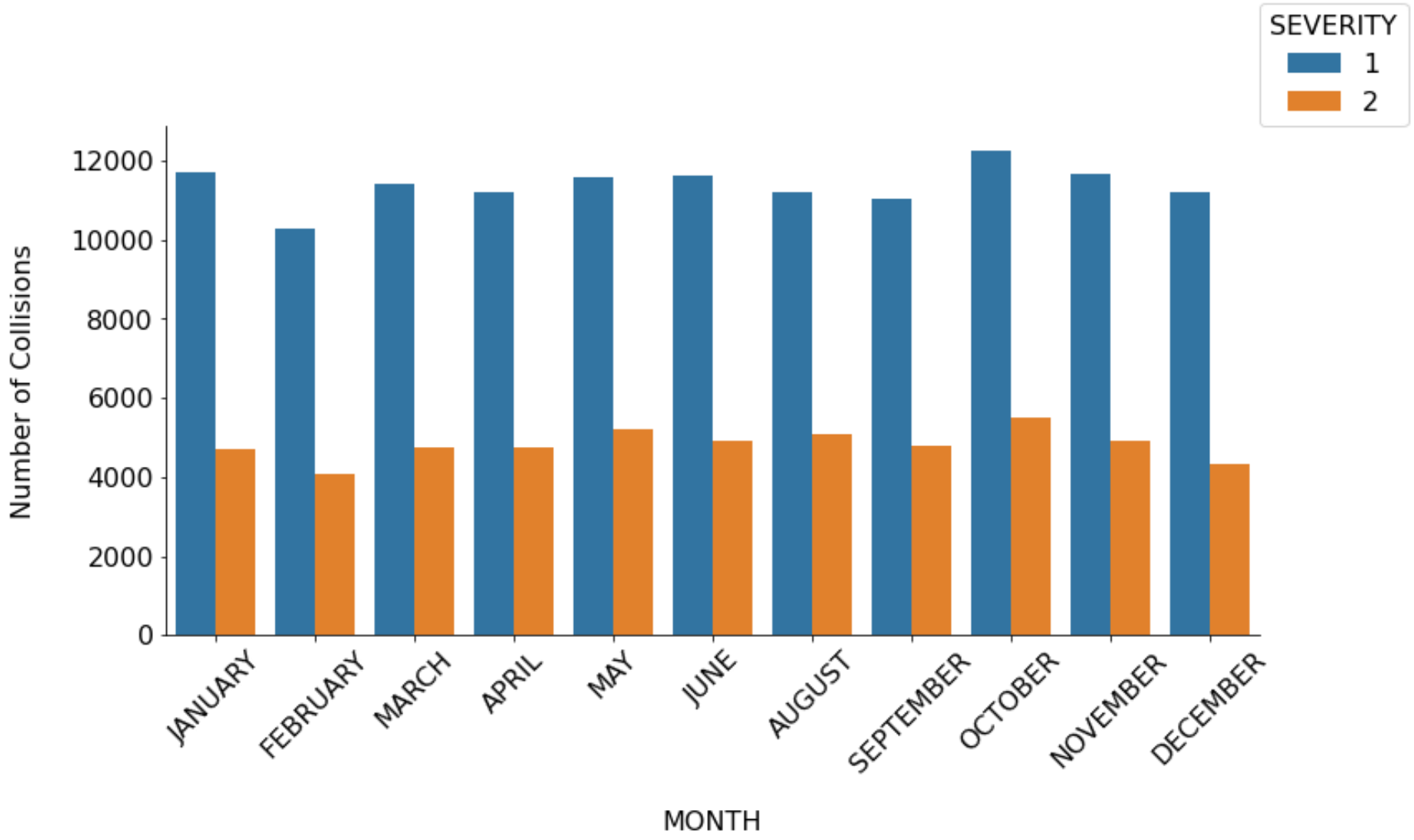
Data

- Sourced from the shared data included in Capstone project
- Around 195000 entries and 38 columns
- Includes many features, such as severity, location, date, time, various driving conditions, collision types etc.
- Unnecessary column were dropped in data cleaning
- Missing values were filled by appropriate categorical values

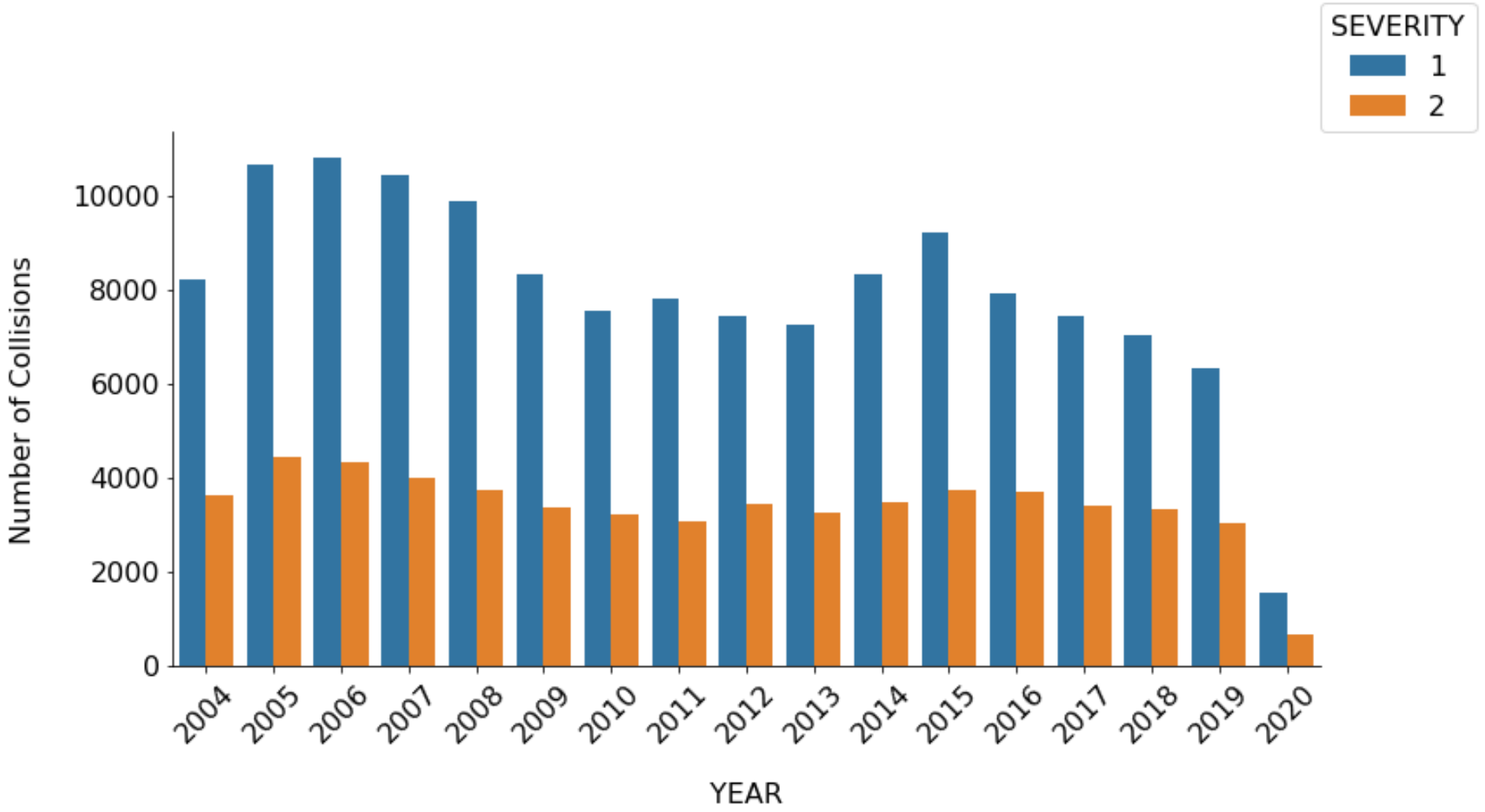
More collisions occur in the afternoon, peak at around 5 pm



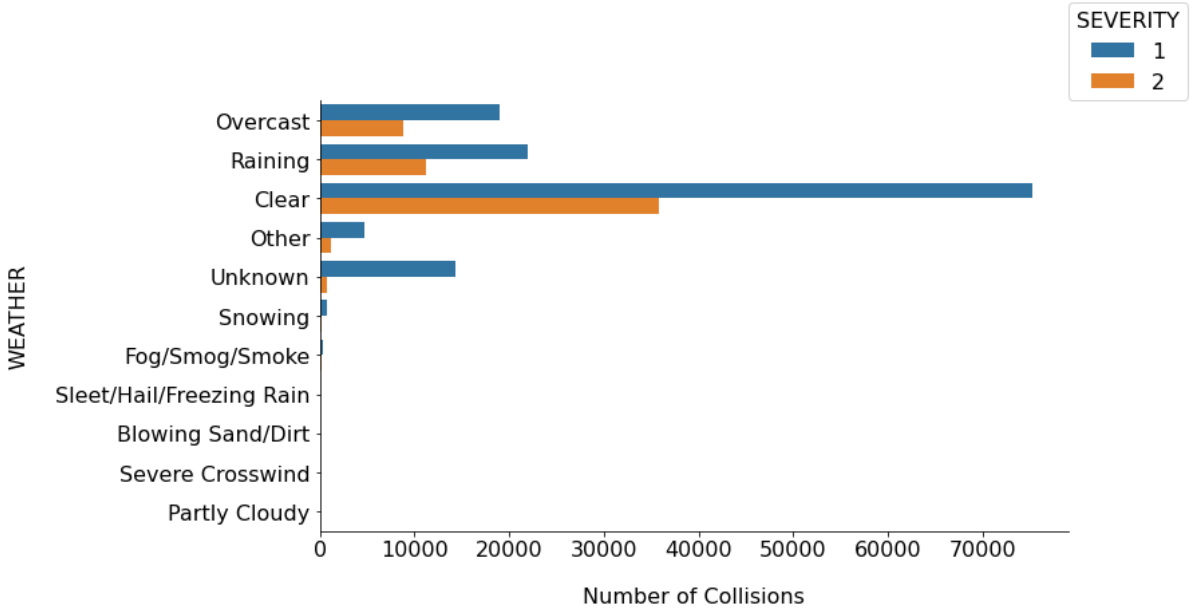
Month has no impact on number of collisions

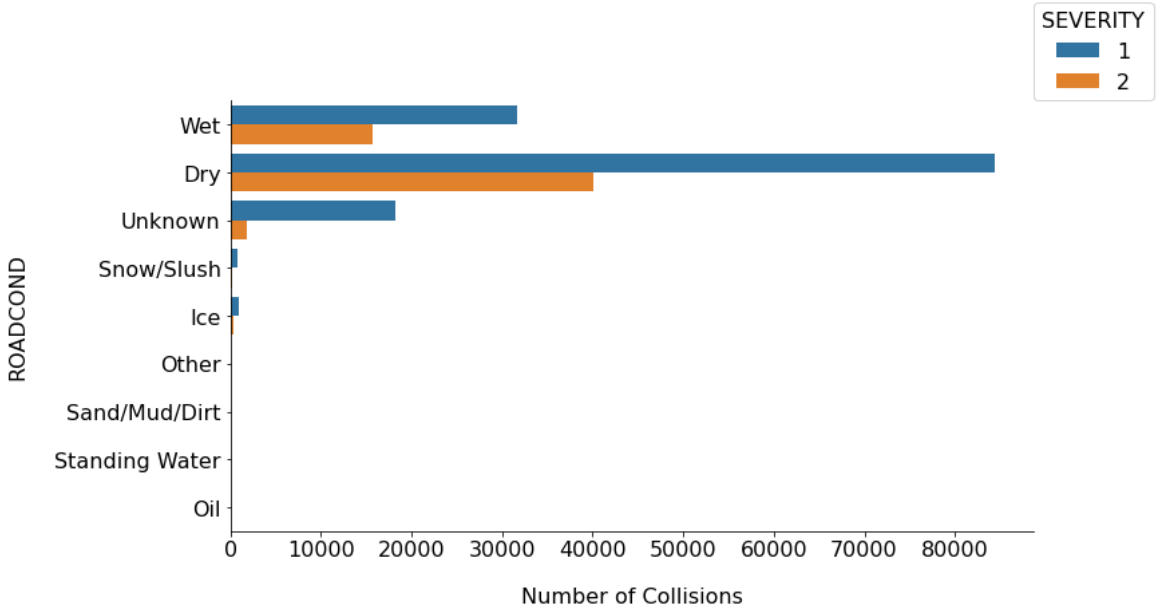


Number of collisions is decreasing year over year



Most collisions occur on clear days and on dry roads





More insights

- ‘Angled’ and ‘rear-ended’ collision types are more severe
- Collisions occurring at the intersection produce severe collisions
- Collisions involving pedestrians and bicycles almost always result in severe collisions.
- Intersection collisions are more severe resulting in injuries to the people involved

Predictive model

- This is a classification problem
- To build a model that can predict severity of a collision
- Four model were used
    1. Logistic regression
    2. Random Forest
    3. Naïve Bayes
    4. Gradient boost
- Accuracy, Precision, Recall and F1 score were used as performance evaluation parameters

Results

Models performed poorly in classifying severity 2

	Logistic regression	Random forest	Naïve Bayes	Gradient boost
Accuracy	**0.76**	0.75	0.73	**0.76**
Precision Severity 1	0.77	0.78	**0.80**	0.77
Recall Severity 1	**0.95**	0.91	0.83	**0.95**
F1 score Severity 1	**0.85**	0.84	0.81	**0.85**
Precision Severity 2	0.72	0.63	0.55	**0.73**
Recall Severity 2	0.32	0.38	**0.49**	0.32
F1 score Severity 2	0.44	0.47	**0.52**	0.44

Results after under-sampling

Three Models performed better in classifying severity 2

	Logistic regression	Random forest	Naïve Bayes	Gradient boost
Accuracy	**0.72**	0.71	0.60	**0.72**
Precision Severity 1	0.76	0.74	0.56	**0.77**
Recall Severity 1	0.65	0.65	**0.98**	0.63
F1 score Severity 1	0.70	0.69	**0.71**	0.69
Precision Severity 2	0.69	0.68	**0.92**	0.68
Recall Severity 2	0.79	0.76	0.21	**0.81**
F1 score Severity 2	0.73	0.72	0.34	**0.74**

Conclusion

- Gradient Boost was the overall winner resulting in unbiased predictions with 72 % accuracy.
- It produced 4366 false positive prediction and only 2189 true negative predictions which is preferable in this type of classification problem
- Other techniques should be considered to deal with unbalanced data
- Date and time data can be included in model building
- Hyper-parameter tuning can be used to further improve model's performance