IST 718 : Big Data Analytics

Group 7

Severity Analysis of US Accidents

Under
Prof. Williamson

Harshita Asnani
Manan Meghani
Rajvee Shah
Yaksh Shobhawat

Problem Statement

- Despite exercising safety rules, the number of deaths caused due to road accidents is pretty high.
- With this project, we are trying to find the states with the highest accident rates and hottest spots for accidents in those states.
- We plan on analyzing the conditions that make these spots more accident prone as compared to others
- We will also be building classification models to predict the severity of accidents

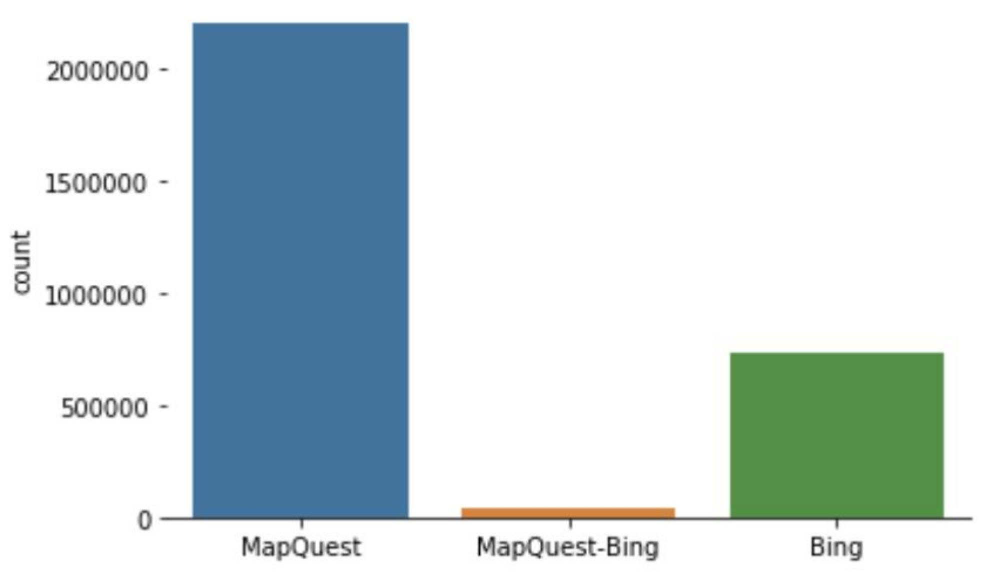
Project description

- Countrywide car accident dataset, which covers 49 states of the United States of America
- Consists of around 3 million accident occurrences and 49 variables
- Data is collected from February 2016 to December 2019, using several data providers, including two APIs that provide streaming traffic incident data.
- US-Accidents can be used for numerous applications and we have implemented casualty analysis and have tried to study the impact of environmental stimuli on accident occurrence

Project Overview

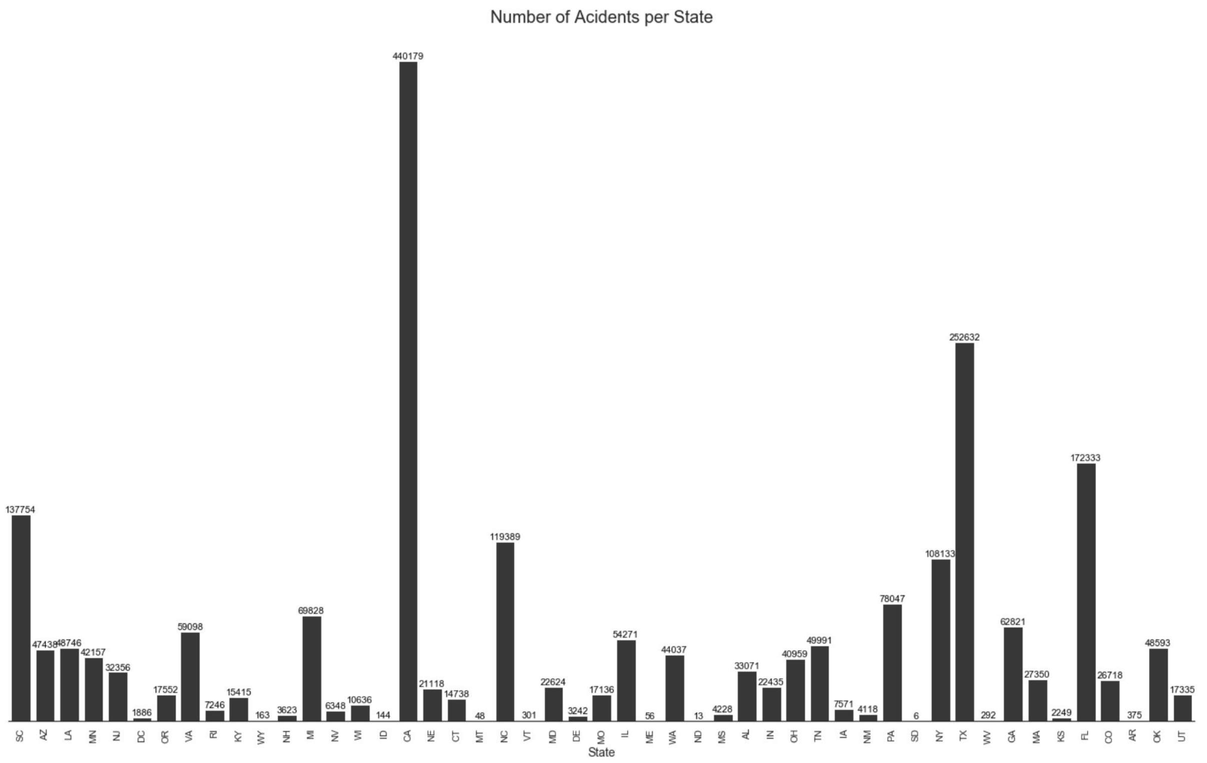
- About the data, data preprocessing
- Data visualization
- Model building
- Results/Summary

Visualizations



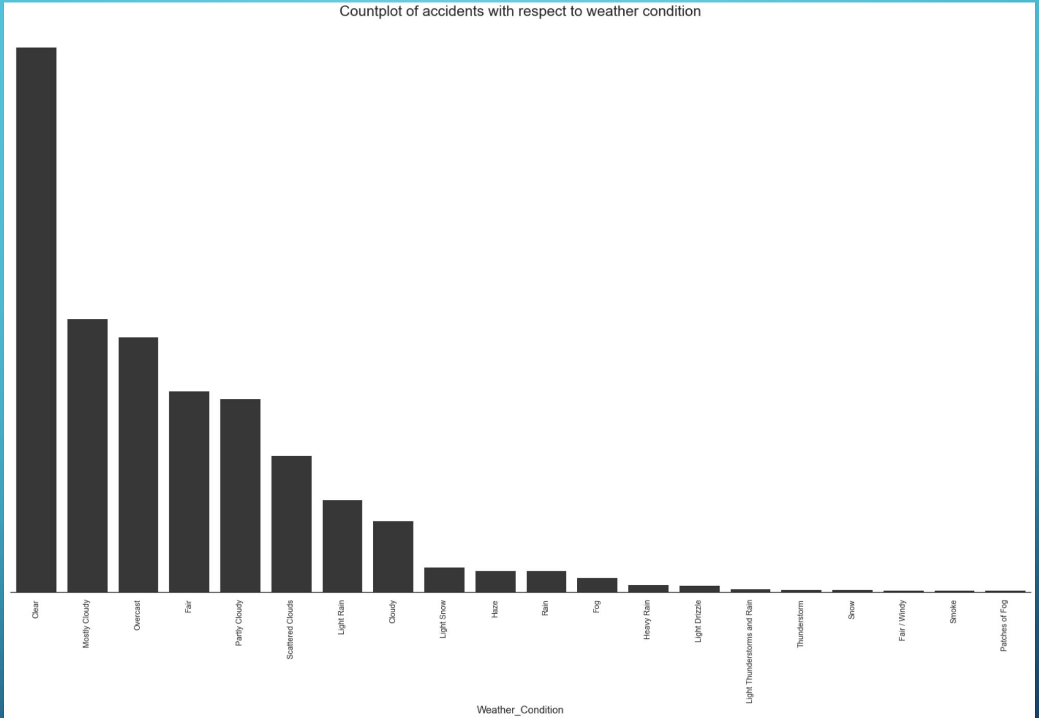
- Three API sources reported the accidents
- Most of the accidents (around 1,700,000) were reported by MapQuest, followed by Bing.

Countplot of accidents w.r.t State



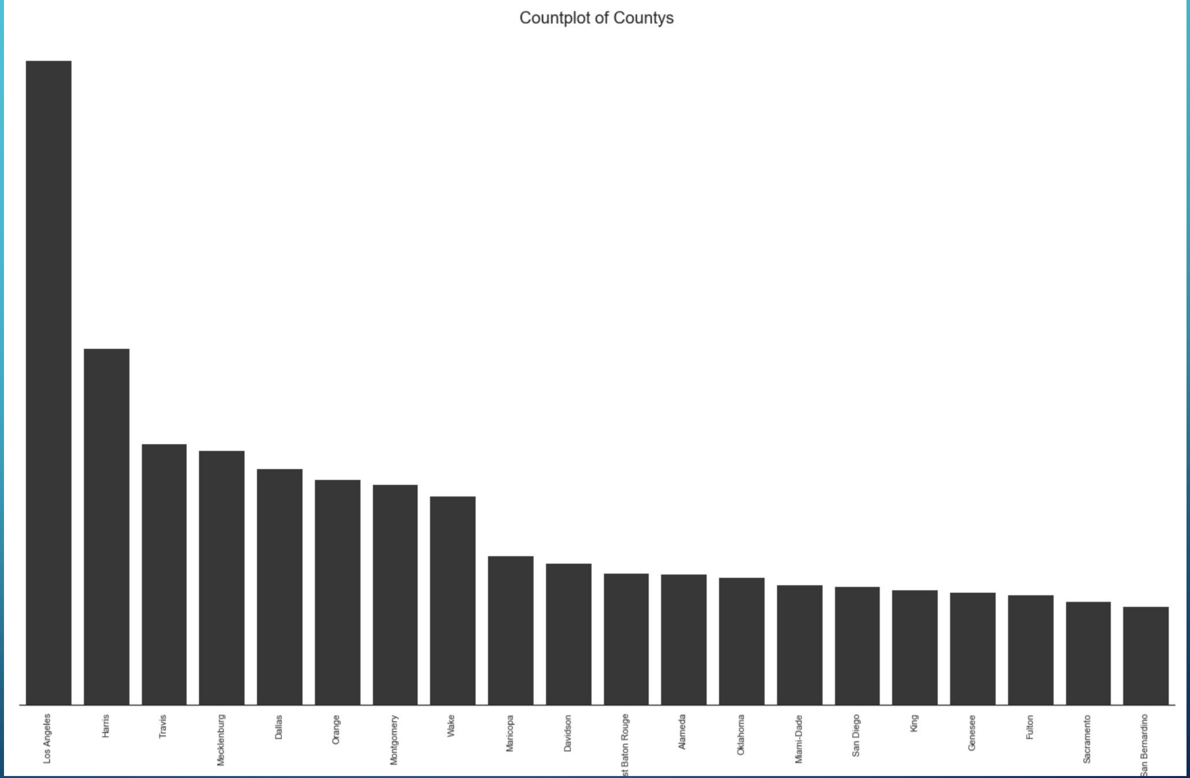
California > Texas > Florida

Countplot of accidents w.r.t Weather_Condition



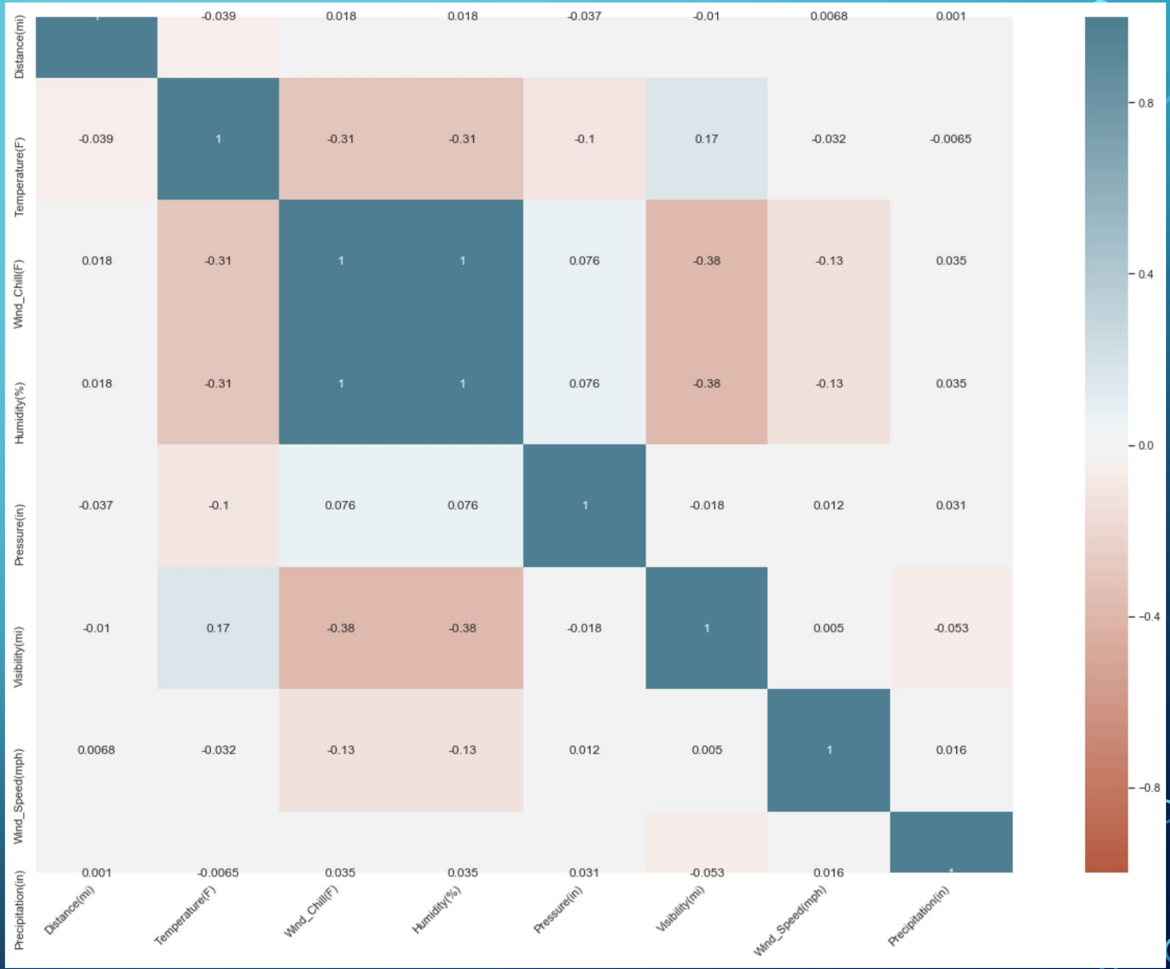
‘Mostly Cloudy’ and ‘Overcast’ conditions mostly responsible for accidents

Countplot of accidents w.r.t County

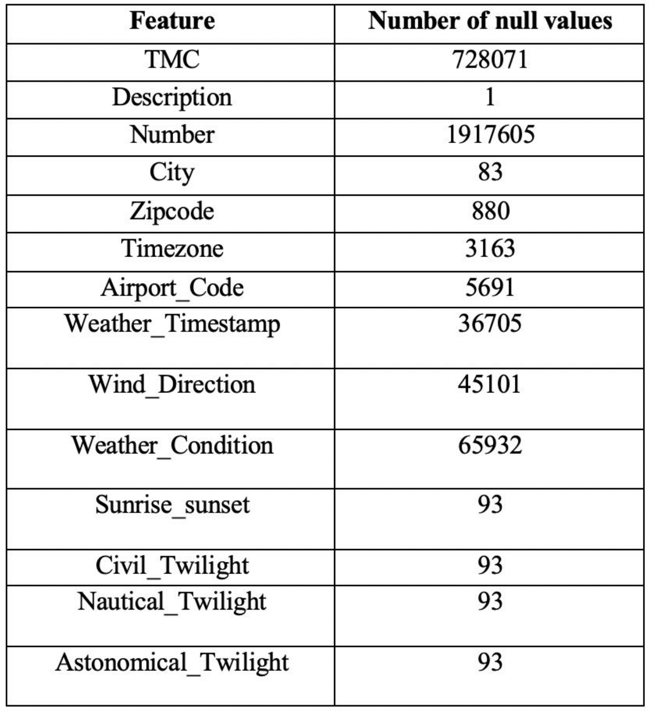


Los Angeles > Harris > Travis

Correlation Matrix



Data Preparation



Feature	Number of null values
TMC	728071
Description	1
Number	1917605
City	83
Zipcode	880
Timezone	3163
Airport_Code	5691
Weather_Timestamp	36705
Wind_Direction	45101
Weather_Condition	65932
Sunrise_sunset	93
Civil_Twilight	93
Nautical_Twilight	93
Astonomical_Twilight	93

- Dataset is huge
- Removed Null values and NAs
- Filtered the variables
- Converted certain categorical variable features to numerical variables for analysis

Decision Trees

Parameters	Accuracy	Runtime
maxDepth=3	67.76%	39.98 mins
maxDepth=5, maxBins=32	65.70%	51.13 mins
maxDepth=6, maxBins=32	65.59%	58.79 mins
maxDepth=7, maxBins=50	63.97%	1 hr 9 mins

Naive Bayes

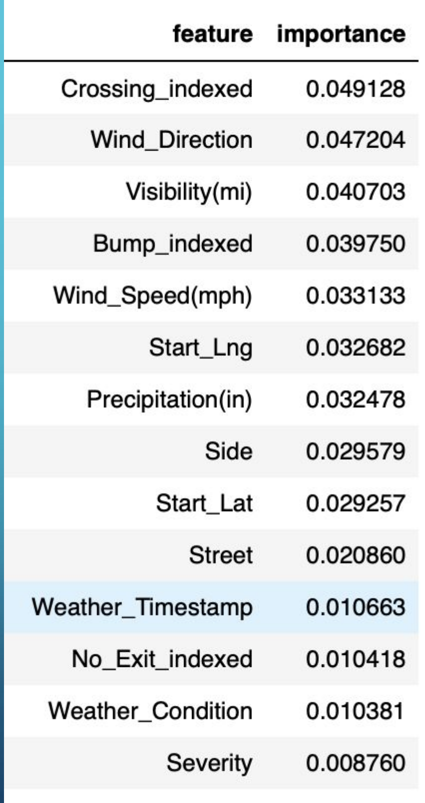
Parameters	Accuracy	Runtime
smoothing=1.0	82.8%	4.61 mins
smoothing=0.5	83.21%	4.62 mins
smoothing=0.4	83.35%	4.51 mins
smoothing=0.2	83.7%	4.79 mins
smoothing=0.1	84%	4.67 mins

Logistic Regression

Parameter	Accuracy
maxIter = 100, elasticNetPram= 0.0, netParam = 0.0	90.08%
maxIter = 100, elasticNetPram= 0.3, netParam = 0.2	73.84%
maxIter = 100, elasticNetPram= 0.1, netParam = 0.3	78.07%

Random Forest

Parameters	Accuracy
Default Parameters	79.35%
numTrees = 100, maxDepth = 5, impurity = entropy	81.18%
numTrees = 100, maxDepth = 5, impurity = gini	81.22%
numTrees = 200, maxDepth = 6, impurity = gini	81.86%

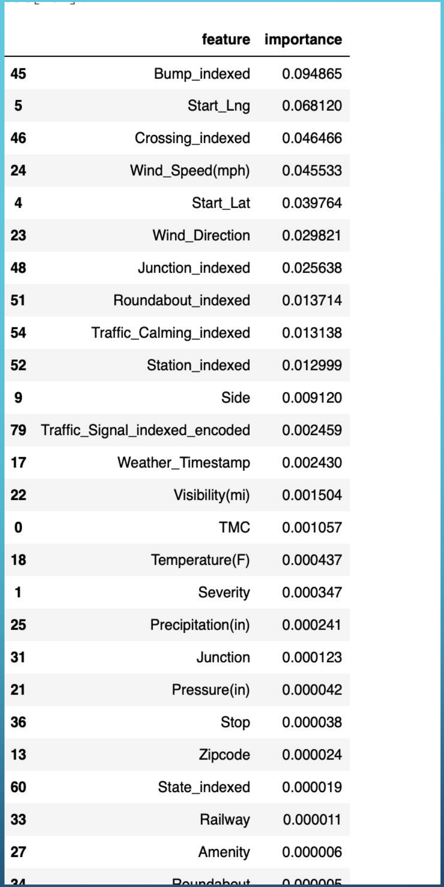


feature	importance
Crossing_indexed	0.049128
Wind_Direction	0.047204
Visibility(mi)	0.040703
Bump_indexed	0.039750
Wind_Speed(mph)	0.033133
Start_Lng	0.032682
Precipitation(in)	0.032478
Side	0.029579
Start_Lat	0.029257
Street	0.020860
Weather_Timestamp	0.010663
No_Exit_indexed	0.010418
Weather_Condition	0.010381
Severity	0.008760

Gradient Boosting

Parameters - Default

Accuracy - 82.41



	feature	importance
45	Bump_indexed	0.094865
5	Start_Lng	0.068120
46	Crossing_indexed	0.046466
24	Wind_Speed(mph)	0.045533
4	Start_Lat	0.039764
23	Wind_Direction	0.029821
48	Junction_indexed	0.025638
51	Roundabout_indexed	0.013714
54	Traffic_Calming_indexed	0.013138
52	Station_indexed	0.012999
9	Side	0.009120
79	Traffic_Signal_indexed_encoded	0.002459
17	Weather_Timestamp	0.002430
22	Visibility(mi)	0.001504
0	TMC	0.001057
18	Temperature(F)	0.000437
1	Severity	0.000347
25	Precipitation(in)	0.000241
31	Junction	0.000123
21	Pressure(in)	0.000042
36	Stop	0.000038
13	Zipcode	0.000024
60	State_indexed	0.000019
33	Railway	0.000011
27	Amenity	0.000006
34	Roundabout	0.000005

Models Used

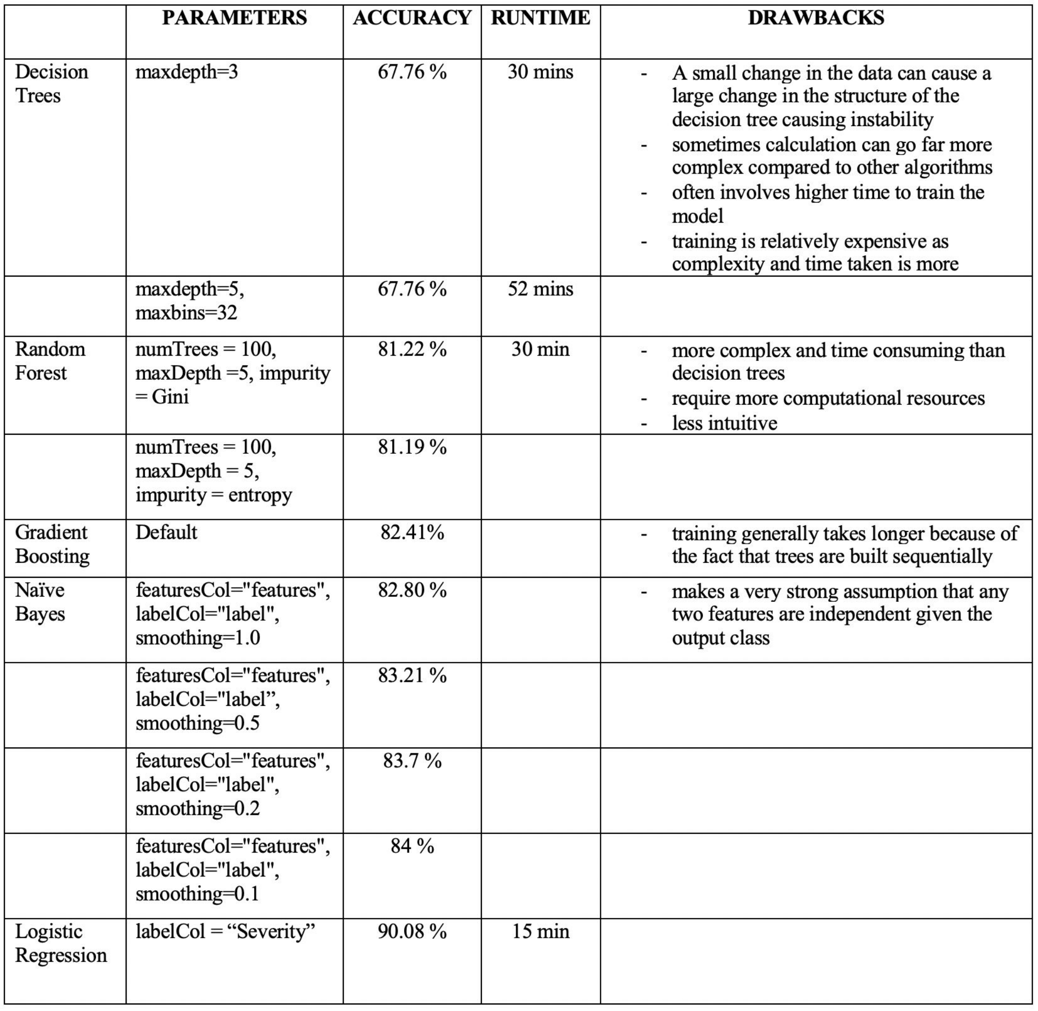
Naïve Bayes

Decision Trees

Random Forest

Gradient Boosting

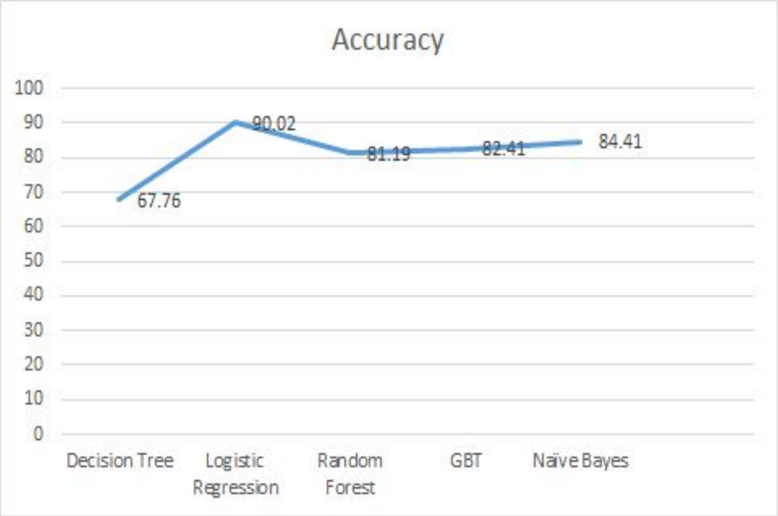
Logistic Regression



	PARAMETERS	ACCURACY	RUNTIME	DRAWBACKS
Decision Trees	maxdepth=3	67.76 %	30 mins	- A small change in the data can cause a large change in the structure of the decision tree causing instability - sometimes calculation can go far more complex compared to other algorithms - often involves higher time to train the model - training is relatively expensive as complexity and time taken is more
	maxdepth=5, maxbins=32	67.76 %	52 mins	
Random Forest	numTrees = 100, maxDepth =5, impurity = Gini	81.22 %	30 min	- more complex and time consuming than decision trees - require more computational resources - less intuitive
	numTrees = 100, maxDepth = 5, impurity = entropy	81.19 %		
Gradient Boosting	Default	82.41%		- training generally takes longer because of the fact that trees are built sequentially
Naïve Bayes	featuresCol="features", labelCol="label", smoothing=1.0	82.80 %		- makes a very strong assumption that any two features are independent given the output class
	featuresCol="features", labelCol="label”, smoothing=0.5	83.21 %		
	featuresCol="features", labelCol="label", smoothing=0.2	83.7 %		
	featuresCol="features", labelCol="label", smoothing=0.1	84 %		
Logistic Regression	labelCol = “Severity”	90.08 %	15 min	

Results Summary

Algorithm	Decision Tree	Random Forest	Gradient Boosting	Naïve Bayes	Logistic Regression
Parameter	maxDepth = 3	numTrees=100 maxDepth = 5 impurity = gini	Default parameters	Smoothing= 0.1	Default
Accuracy	67.76%	81.22%	82.41%	84%	90.02%
Runtime	40 mins	39.98 mins	2 hrs	5 mins	15mins



Problems encountered

- Size of the data
- Proper representation of the main data
- Handling Outliers
- High Volume of NaN Values.
- Algorithm Execution

Credits

Data Preprocessing	Rajvee, Yaksh
Data Visualization	Rajvee, Yaksh
Logistic Regression	Yaksh, Harshita
Naive Bayes, Decision Tree	Manan
Random Forest, Gradient Boosting	Harshita

Thank You