US Traffic Accident

Project by
Amirreza Eshraghi - A20451901
Anneke Soraya Hidayat - A20406957 (Team Leader)
Namitha Venkataramanan - A20453185
Sourav Yadav - A20450418

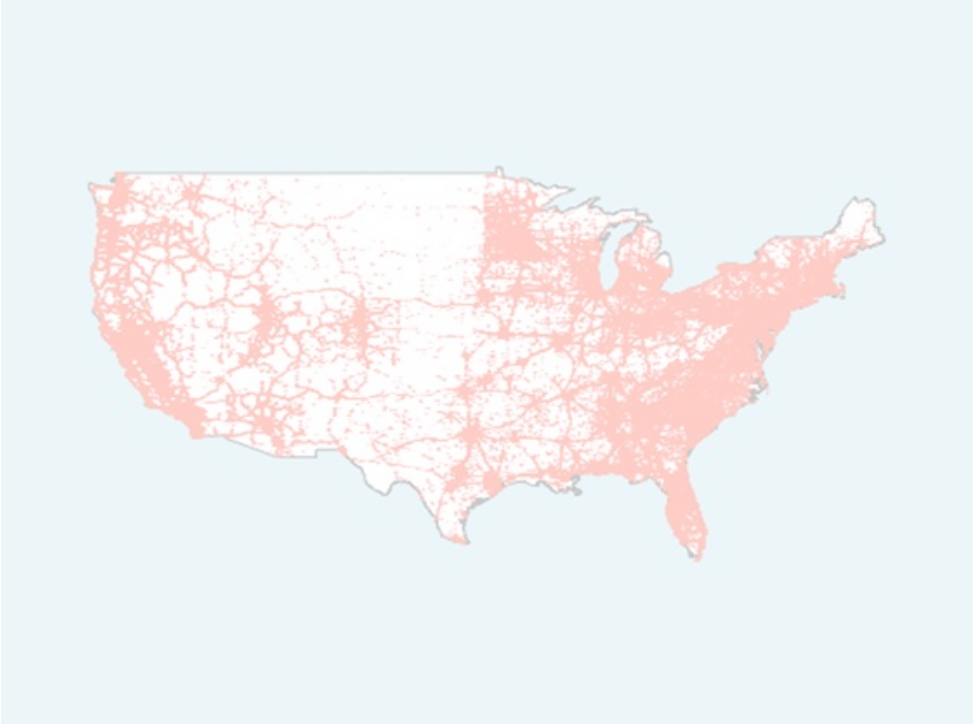
Outline

1. Overview
2. Data Preprocessing
3. Data Exploration
4. Model Implementation
5. Results
6. Discussion
7. Conclusion

OVERVIEW

Overview

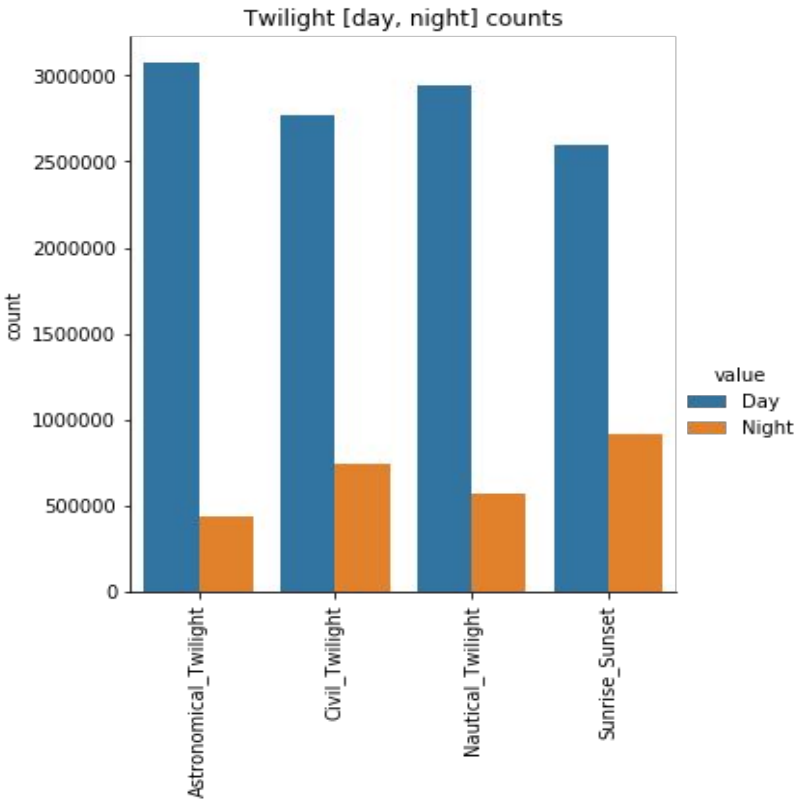
- Traffic Accident in the United States
- What factors does influence the severity of the traffic?
  - Weather?
  - POI?
  - Location?
- Total accident VS Accident Rate
- How accurate the model represent the general population?



DATA PREPROCESSING

Data Adjustments

- Consistency across similar features - `Civil_Twilight, Nautical_Twilight, Astronomical_Twilight, Sunrise/Sunset`
- No significant influence when these values are different to each other
- Drop variables that have unique values - `Source, Country, ID`
- Drop variables that have granular attributes that may share similar attributes with other variables - `Airport_Code, Side, Number, Weather_Timestamp, Timezone, Precipitation(in)`
- Drop values that only has one value, or NaN - `End_Lat, End_Lng, Turning_Loop`



DATA EXPLORATORY ANALYSIS

Accident Distribution

Based on `Start_Lat and End_Lng`

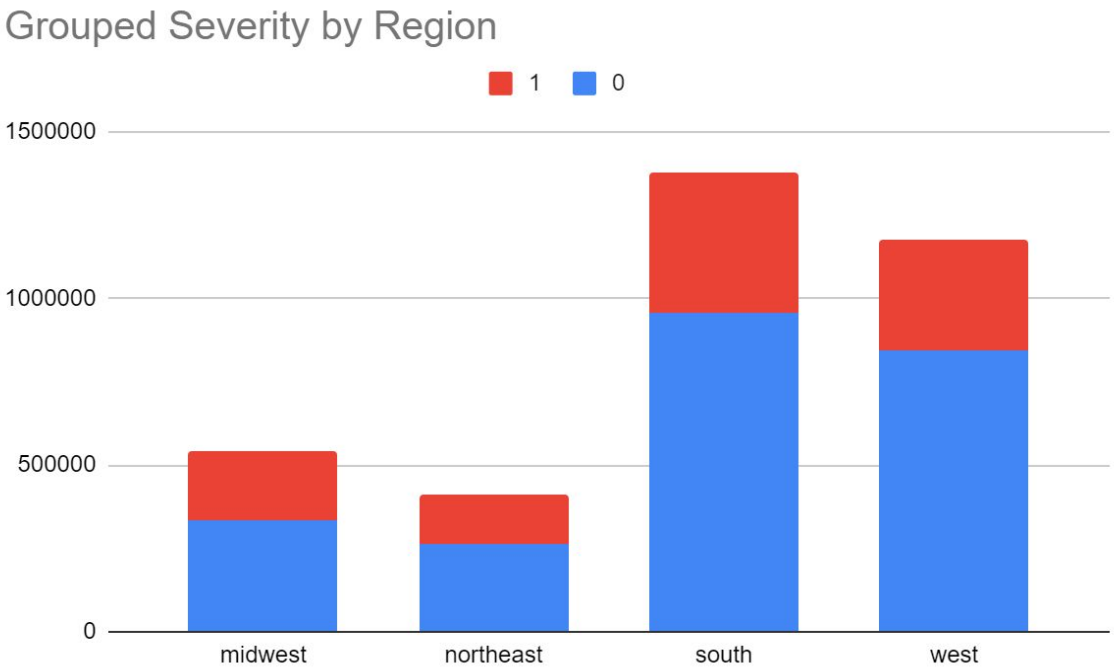
High concentration regions - West coast (SW), East coast (NE)

Traffic Accident Rate

- Identifying states with high number of accidents
- US Population by State
- US total car number by State

	Traffic Accident		Accident Rate by Population		Accident Rate by Total number of car	
1	CA	816804	SC	0.034	SC	0.095
2	TX	329284	OR	0.021	OR	0.061
3	FL	257974	CA	0.021	UT	0.055
4	SC	173277	UT	0.016	CA	0.054
5	NC	165955	NC	0.016	NC	0.049
6	NY	160787	OK	0.015	OK	0.046
7	PA	106787	MN	0.015	LA	0.044
8	IL	99691	LA	0.013	MN	0.041
9	VA	96075	NE	0.012	TX	0.040
10	MI	95983	FL	0.012	NE	0.035

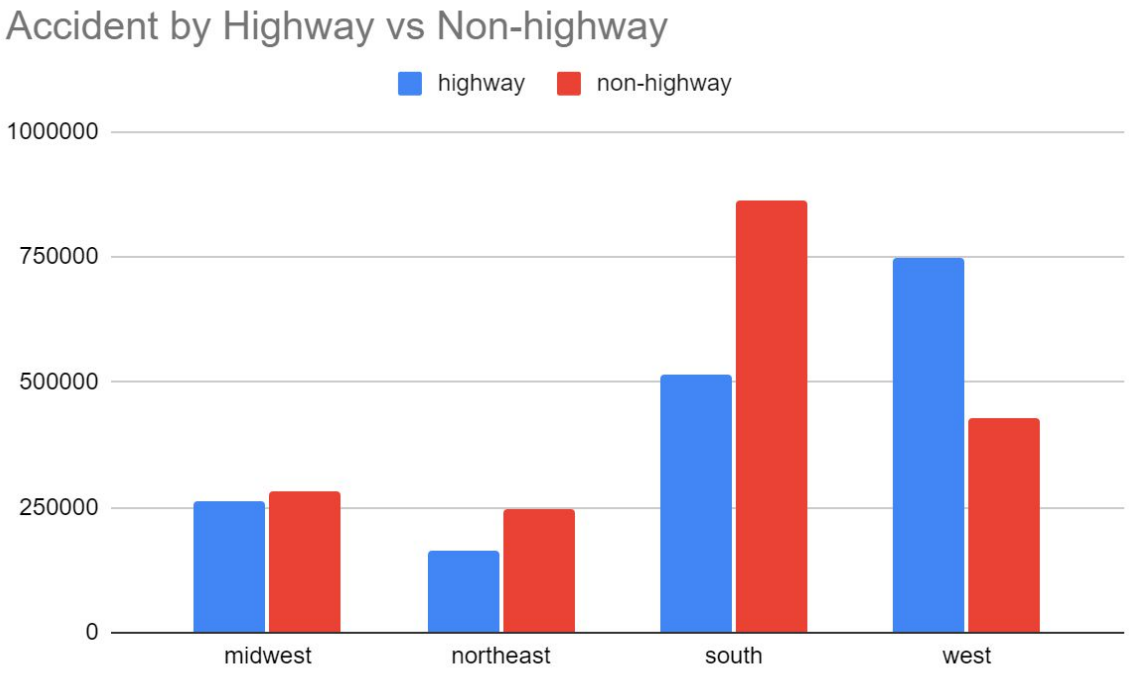
Severity by Region



Highest number of accidents - South region

Accident by Highway vs Non-Highway

- Highest accidents in Highways - West
- Highest accidents in Non-Highways - South

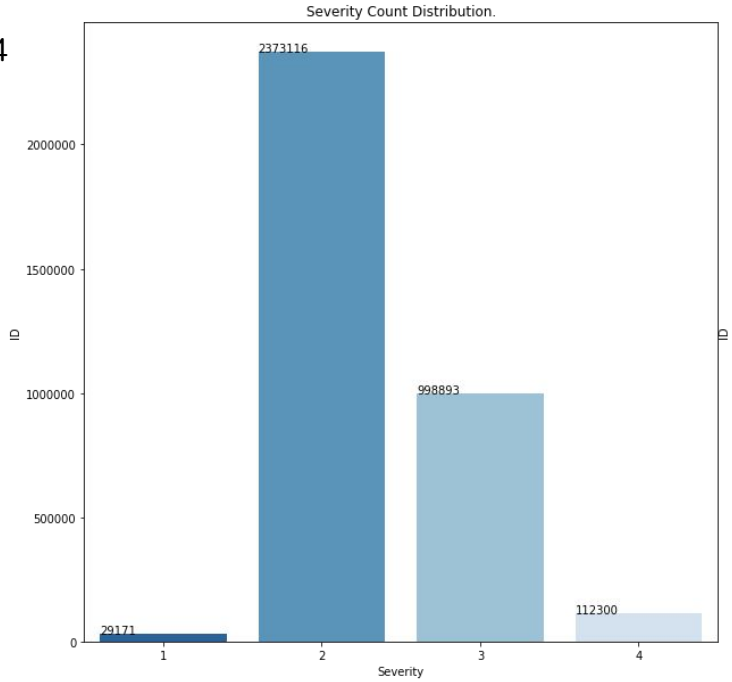


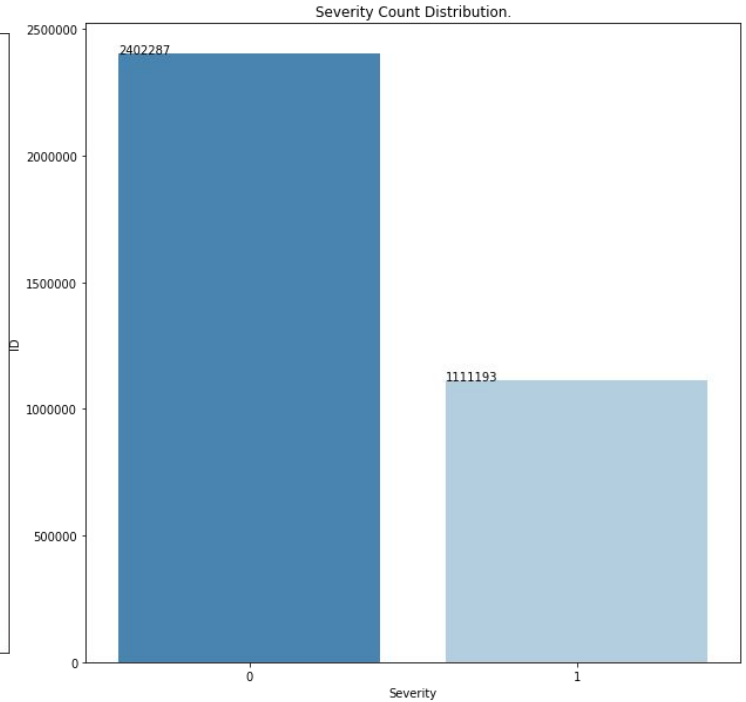
Target Value - Severity

4 integer values: 1,2,3 and 4
1 indicates traffic is not affected by the accident (or less affected) and 4 indicates traffic is highly affected by the accident

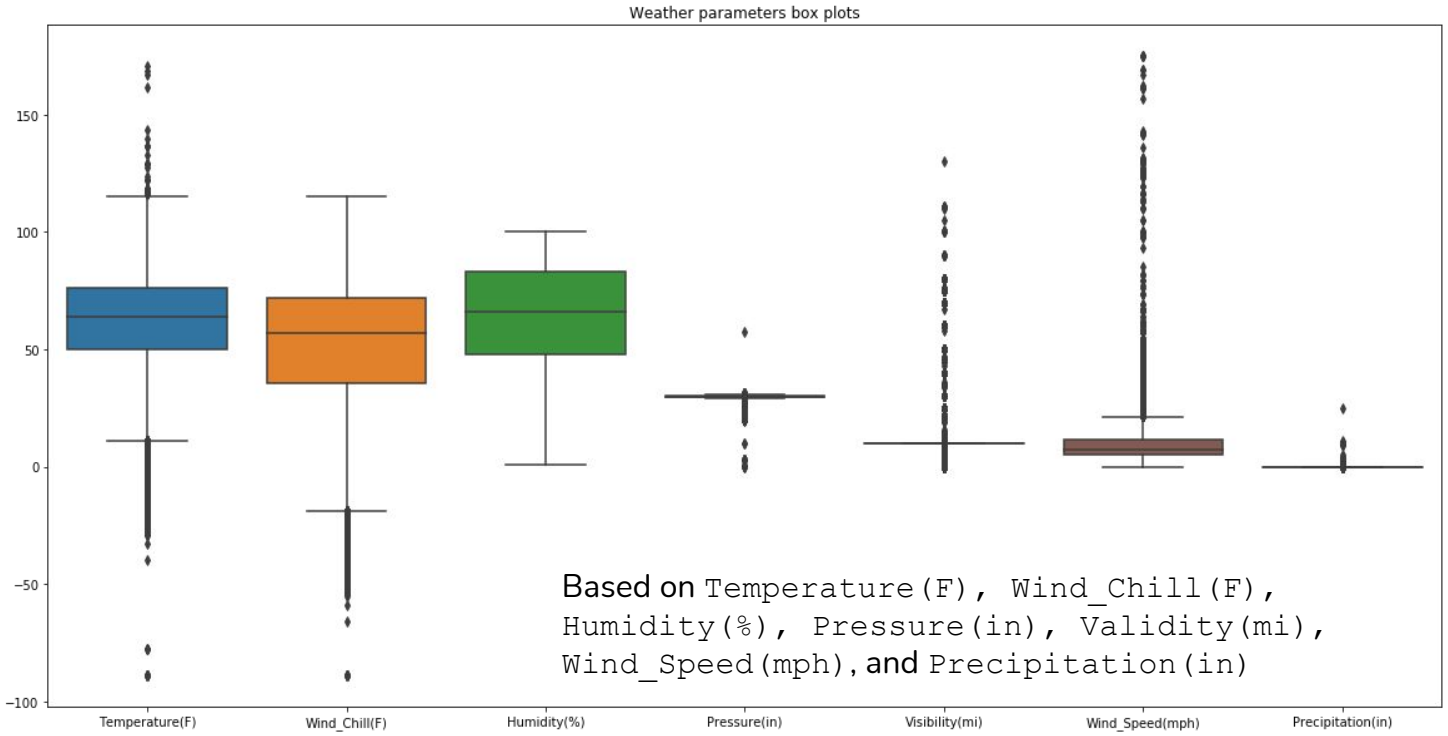
Mapping the severity classes of “1 and 2” as 0 and “3 and 4” as 1

68% of data belongs to 0 class and 32% belongs to 1 class





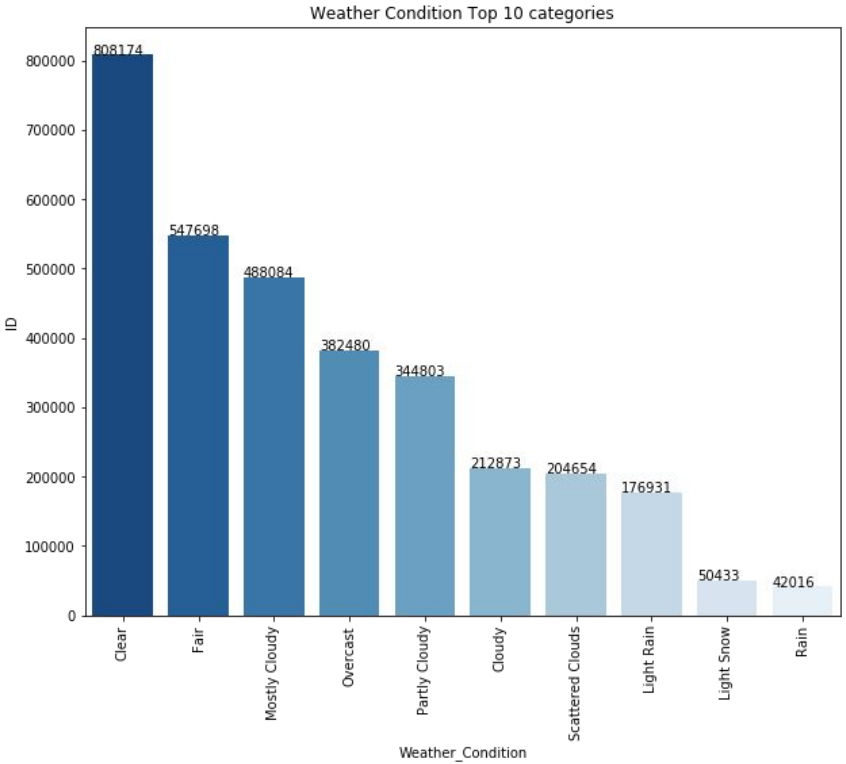
Weather Features



Based on `Temperature(F)`, `Wind_Chill(F)`, `Humidity(%)`, `Pressure(in)`, `Validity(mi)`, `Wind_Speed(mph)`, and `Precipitation(in)`

Weather Condition

- 127 unique strings
- Describes `weather` in phrases



Overcast
Light Snow
Light Snow
Scattered Clouds
Overcast
Overcast
Partly Cloudy
Clear
Light Snow
Overcast

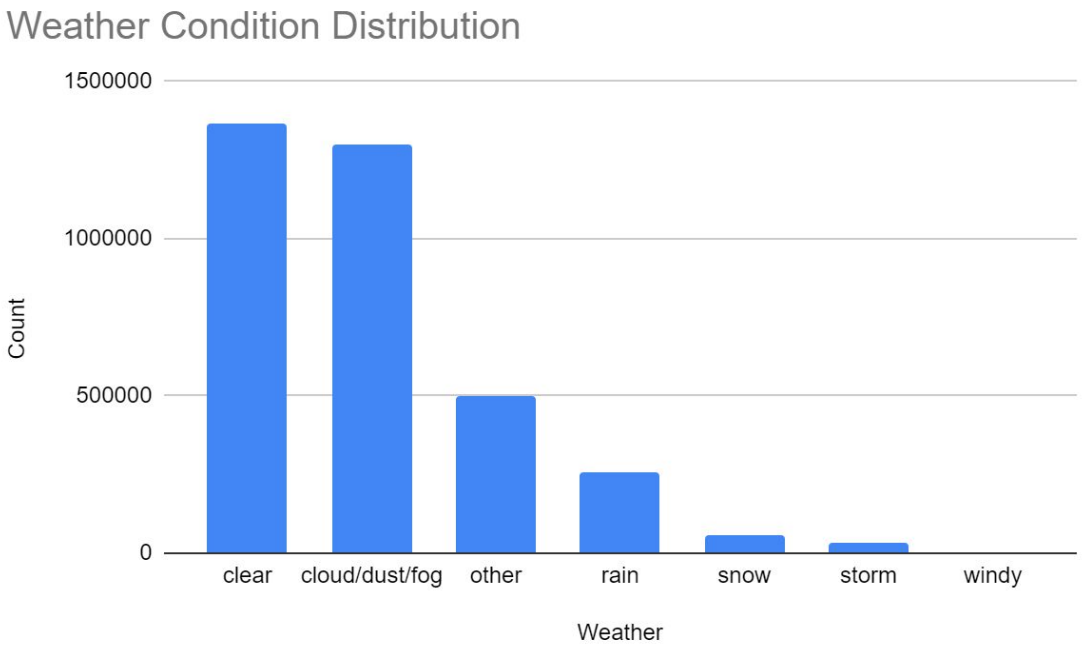
Weather Condition - Preprocessing

```
If [“storm”, “thunder”, “smoke”, “tornado” in Weather_Condition:
        Assign “storm”
elif [“clear”, “fair”] in Weather_Condition:
        Assign “clear”
Elif [“rain”, “drizzle”] in Weather_Condition:
        Assign “rain”
Elif [“cloud”, “dust”, “fog”] in Weather_Condition:
        Assign “cloud/dust/fog”
Elif [“snow”, “ice”] in Weather_Condition:
        Assign “snow”
Elif [“wind”] in Weather_Condition:
        Assign “windy
Else:
        Assign “other”
```

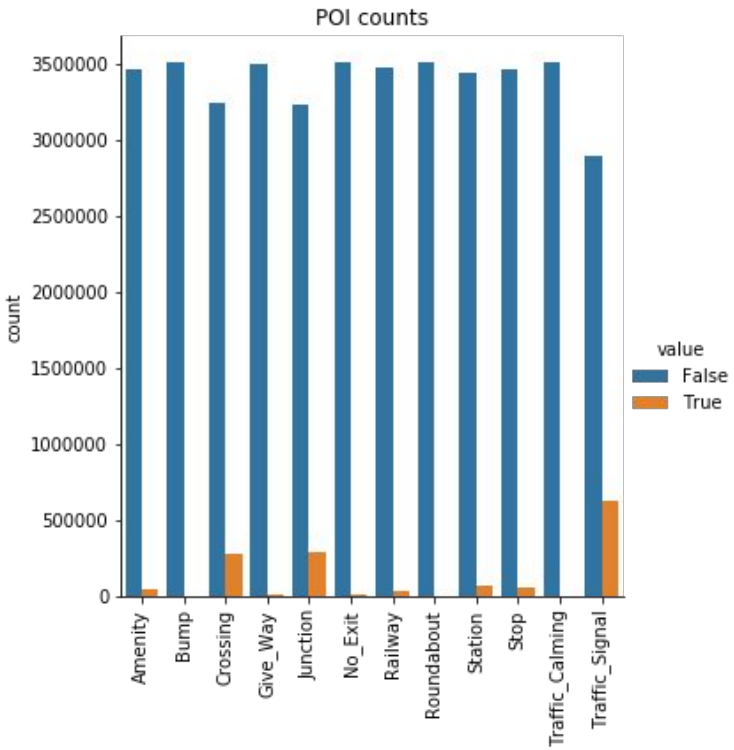
Bucketed values into smaller categories

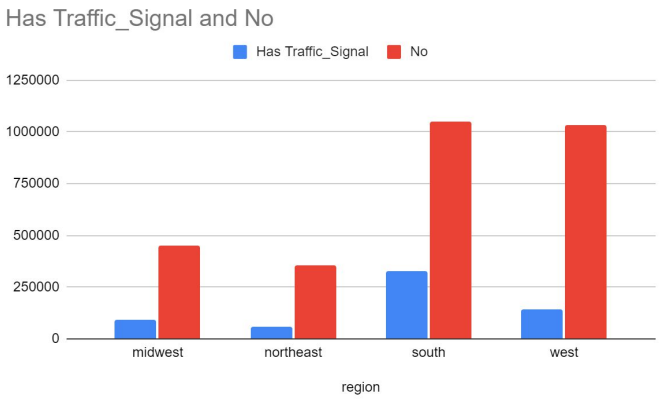
Weather Condition - Preprocessing

- 7 values that can treated as factors
- Most accidents occurrence - *clear* or *cloud* weather
- Extreme severity of traffic accident by separating data on this feature



Point of Interest (POI)





- Not good predictors
- Can be used for grouping variables for further analysis

TMC

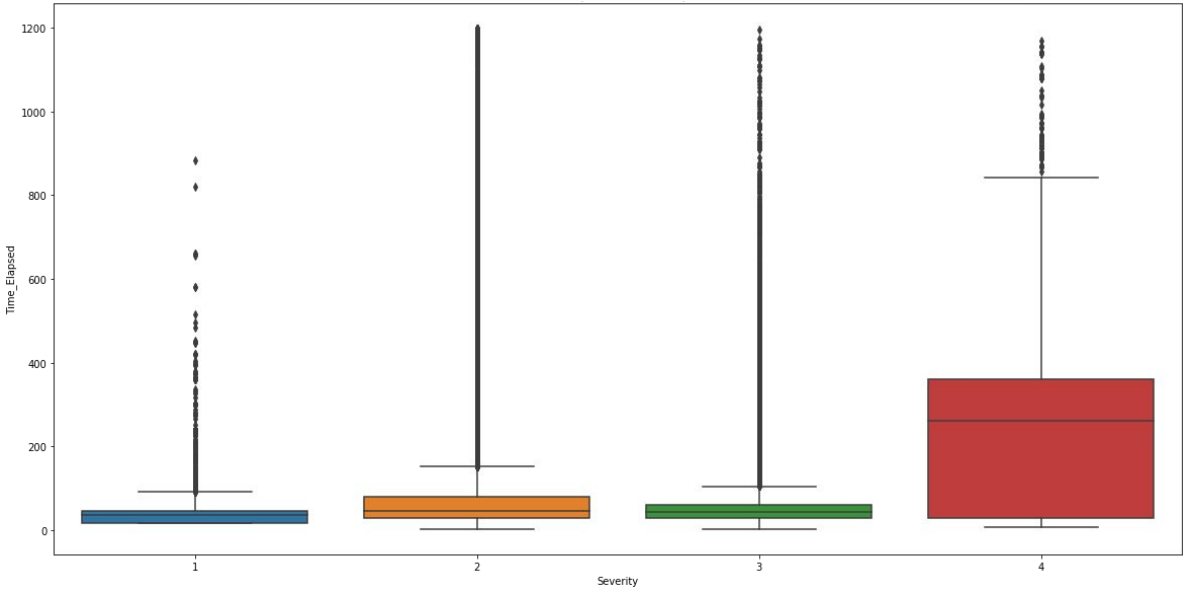
“Traffic Message Code”
201: accidents
241:(Q) accident(s). Right lane blocked
245:(Q) accident(s). Two lanes blocked
229:(Q) accident(s). Slow traffic
203:multi-vehicle accident (involving Q vehicles)
...

- Code of communication between traffic department and police
- TMC is so important and related to Severity!!!
- But too many null values(more than 1 million)
- Transfer to categorical variable (Even NAs!)

Traffic Features

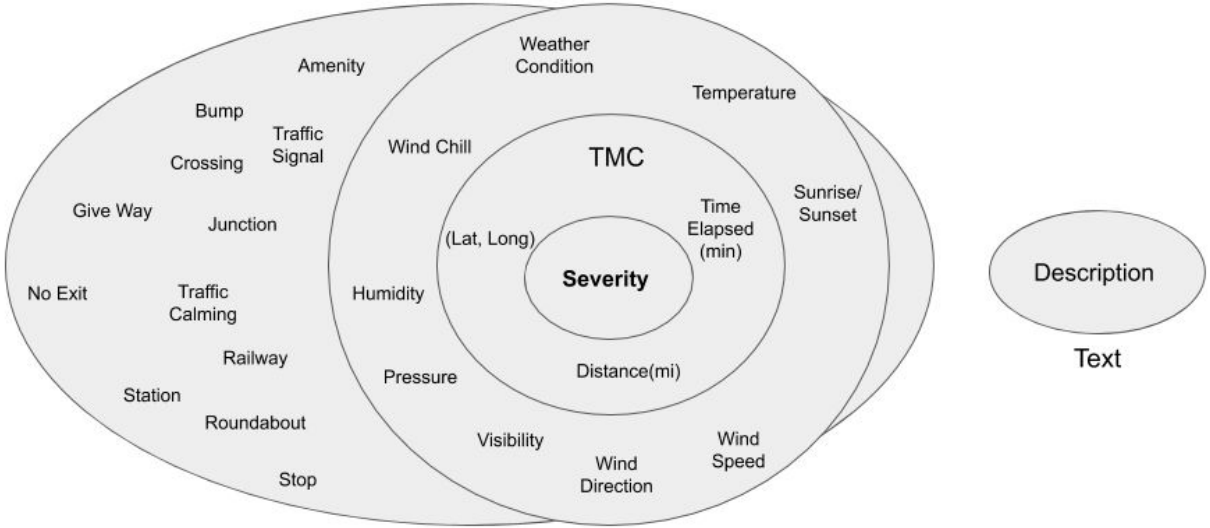
TMC Code	Description	Freq
201:	accidents	2080341
241:	(Q) accident(s). Right lane blocked	249852
245:	(Q) accident(s). Two lanes blocked	40338

- The traffic features cover `TMC`, `Distance(mi)` and `Time_Elapsed`
- 33% of data points that have missing value on `TMC`
- Move to next traffic feature due to large proportion of missing values



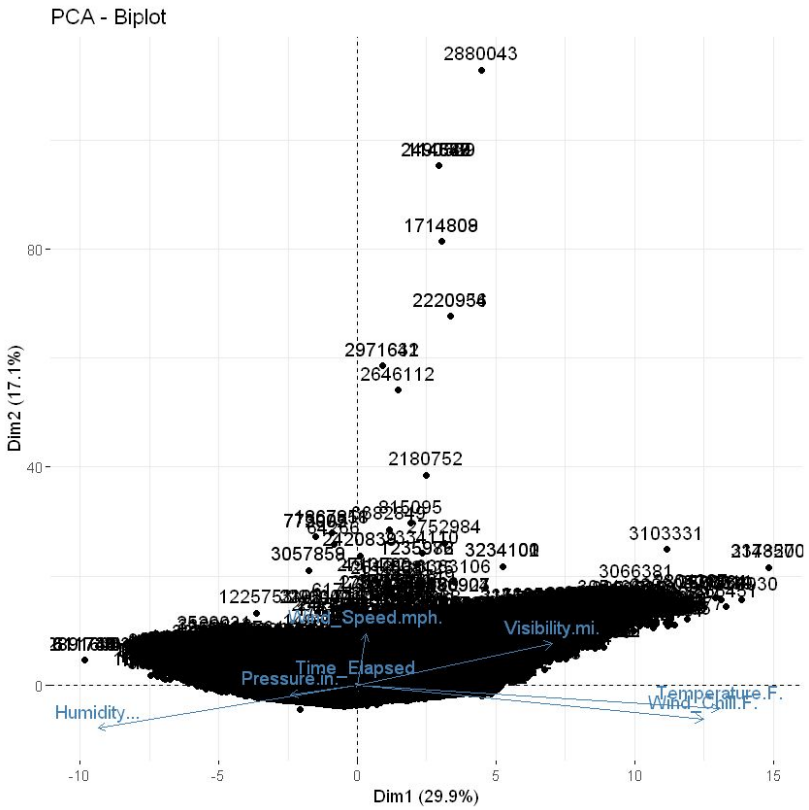
Feature Selection and Analysis

- `Severity` as the target variable located at center
- Followed by traffic features, weather features, `POI`, and `description`
- The text `description` is grouped as independent - does not carry information to predict `severity`, but only as additional variables that explain the `severity`.



Principal Component Analysis (PCA)

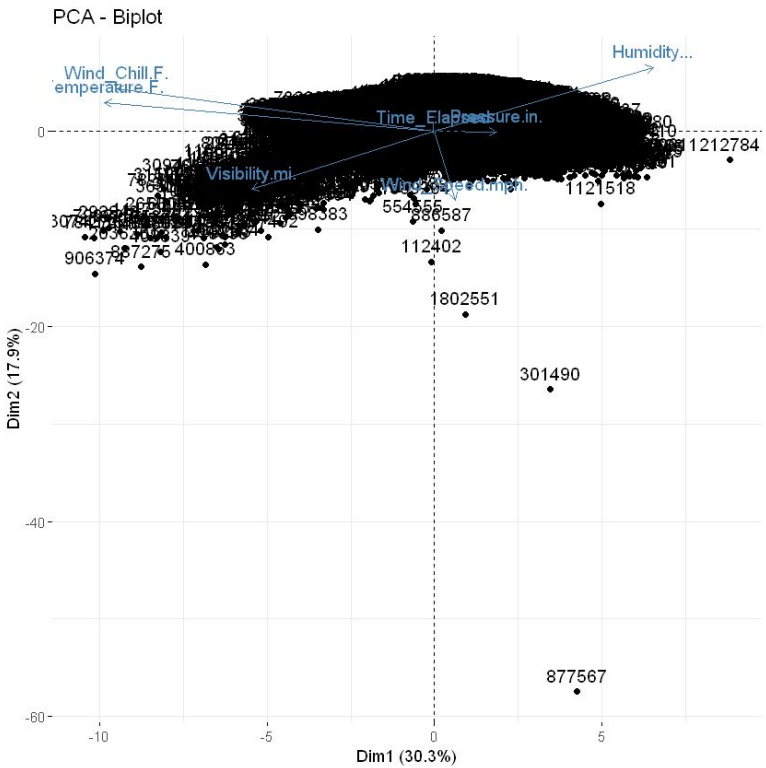
- Correlation between numerical features
- First principal component - 30% variation explained
- Positive correlation - `Temperature` and `Wind chill`
- Negative correlation - `Humidity` and `Visibility`



MODEL IMPLEMENTATION

Principal Component Analysis (PCA)

- Sample size computation by (N*p) where N is the total samples and p is the number of predictors
- 50,000 records from each region across US
- Final sample size - 200,000 records
- PCA - similar feature response as whole dataset



Model Details

- Target - classify severity of traffic accident mapped asa binary response
- Our approach - feature selection from different models' perspective
- Predictors Importance by dimensionality reduction to see each features interaction with how much of data explained in the lower dimension
- GLM with binomial family
- GLMNET for feature selection in Lasso
- Decision Tree to interpret rule path
- SVM for performance comparison

Performance Metric - Confusion Matrix

	Models	# Param	Param		Accuracy	
					Training	Testing
1	GLM	12	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.", "Visibility.mi.", "Wind_Direction",	"Wind_Speed.mph.", "Sunrise_Sunset", "is_highway", "region", "Time_Elapsed", "Weather"	0.6939	0.6917
2		12	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.", "Visibility.mi.", "Wind_Speed.mph.",	"Sunrise_Sunset", "is_highway", "region", "Weather" "Traffic_Signal", "Junction"	0.6974	0.6994
3		11	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.", "Visibility.mi.", "Wind_Speed.mph.",	"Sunrise_Sunset", "is_highway", "region", "Weather" "Traffic_Signal", "Junction"	0.6971	0.699
4		11	"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset", "Distance.mi"	"is_highway", "region", "weather" "Traffic_Signal" "TMC."	0.7334	0.7338
5		10	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.", "Visibility.mi.",	"Wind_Speed.mph.", "Sunrise_Sunset", "is_highway", "region", "Weather"	0.6923	0.6915
6		10	"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset",	"is_highway", "region", "weather" "Traffic_Signal" "Distance.mi."	0.6979	0.6991
7		8	"Temperature.F.", "Wind_Chill.F.", "Humidity...", "Pressure.in.",	"Wind_Speed.mph.", "Sunrise_Sunset", "is_highway", "region",	0.6954	0.6964
8		9	"Temperature.F.", "Humidity...", "Pressure.in.", "Wind_Speed.mph.", "Sunrise_Sunset",	"is_highway", "region", "weather" "Traffic_Signal" "Distance.mi."	0.6977	0.6979
9		8	"Temperature.F.", "Pressure.in.", "Wind_Speed.mph.", "Sunrise_Sunset",	"is_highway", "region", "weather" "Traffic_Signal	0.6962	0.6971

For **GLM** - Best Model
Features - Temperature.F., Pressure.in., Wind_Direction, Wind_Speed.mph., Sunrise_Sunset, Distance.mi, is_highway, region, weather, Traffic_Signal, TMC

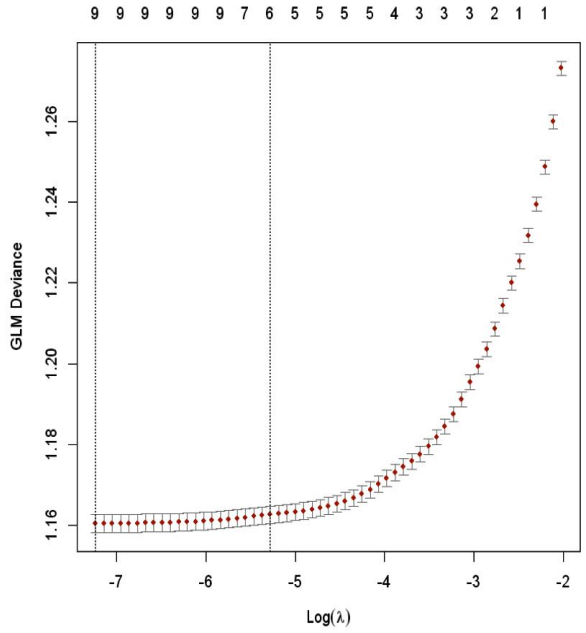
# of Parameters - 11

Prediction Accuracy
Training - 73.34%
Testing - 73.38%

Support GLM Analysis with Lasso

- GLM - 4th is the best performance among GLM models.
- GLM model's features - forms basis of features that we are applying on other models
- GLM with Lasso for feature selection - obtain best lambda
- Five important features - `Pressure.in.`, `Wind_Speed.mph.`, `Distance.mi`, `is_highway`, `Traffic_Signal`

(Intercept)	-3.869037259
Temperature.F.	.
Wind_Chill.F.	.
Humidity...	.
Pressure.in.	0.081668916
Visibility.mi.	.
Wind_Speed.mph.	0.008904345
Distance.mi.	0.111850245
is_highway	0.969244253
Traffic_Signal	-0.743849349



Performance Metric - Confusion Matrix

Decision Tree	11	"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset", "Distance.mi"	"is_highway", "region", "weather" "Traffic_Signal" "TMC."	0.7351	0.7344
	3	"Traffic_Signal" "is_highway", "TMC"	-	0.7351	0.7344
*SVM**	*11*	*"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset", "Distance.mi"*	*"is_highway", "region", "weather" "Traffic_Signal" "TMC."*	*0.7259*	*0.5721*

For **Decision Tree** - Best Model
Features - Temperature.F., Pressure.in., Wind_Direction, Wind_Speed.mph., Sunrise_Sunset, Distance.mi, is_highway, region, weather, Traffic_Signal, TMC

# of Parameters - 11

Prediction Accuracy

Training - 73.51%
Testing - 73.44%

Performance Metric - Confusion Matrix

Decision Tree	11	"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset", "Distance.mi"	"is_highway", "region", "weather" "Traffic_Signal" "TMC."	0.7351	0.7344
	3	"Traffic_Signal" "is_highway", "TMC"	-	0.7351	0.7344
*SVM**	*11*	*"Temperature.F.", "Pressure.in.", "Wind_Direction", "Wind_Speed.mph.", "Sunrise_Sunset", "Distance.mi"*	*"is_highway", "region", "weather" "Traffic_Signal" "TMC."*	*0.7259*	*0.5721*

*For **SVM** - Best Model ****
Features - Temperature.F., Pressure.in., Wind_Direction, Wind_Speed.mph., Sunrise_Sunset, Distance.mi, is_highway, region, weather, Traffic_Signal, TMC

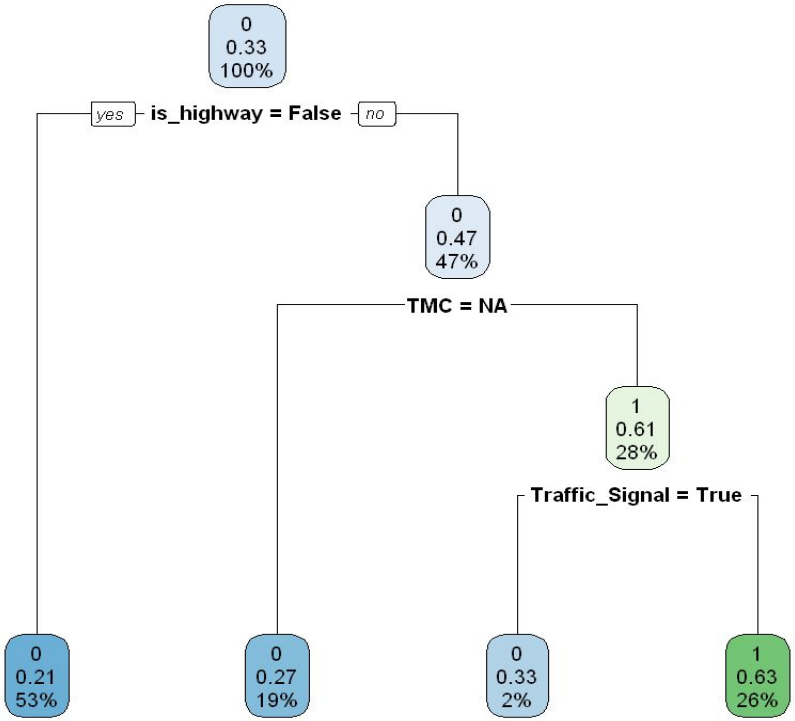
# of Parameters - 11

Prediction Accuracy

Training - 72.59%
Testing - 57.21%

Decision Tree

- Input - 11 features
- Output - 3 major features on its split
- Three important features - `is_highway, TMC, Traffic_Signal`
- Decision tree model provides both performance and the interpretability of rule based visualization of the model, with accuracy of 73%



SVM

- Sample 10,000 data points from both training and testing dataset
- Only applicable as model comparison but cannot take into account since the comparison is **not an apple-to-apple comparison**
- With minimum number of data, it can classify the training dataset up to 72% accuracy only with 10,000 samples
- This approach overfits the training data, since the testing performance dropped to 57% accuracy
- Reason for constraint due to computational limit

DISCUSSION, LIMITATION AND CONCLUSION

Discussion

- Our objective - to perform inference analysis to which predictors that explain data the best

- Model analysis output - small number of important features separates the severity the best
  `Is_highway`: boolean
  `TMC`: factors, including NA as factor.
  `Traffic_Signal`: boolean

- Biggest Discovery - Decision Tree!

- 63% chance that the traffic accident is more severe

Limitation and Conclusion

Limitation:

Computational resources

- Large dataset, 49 feature columns and various types of features (numerical, unique, categories, boolean, timestamp),
- requires more time on data exploration compared to modeling

Conclusion:

Under the assumption of random sampling that represents the general population, we conducted analysis through a classification model. Our final result indicates that among default features from the dataset, top-3 features are enough to separate the severity into two categories with a significant performance.

THANK YOU!

Questions?