Machine Learning

Predicting Road Accident Severity

UCSC Extension DBDAx408 (Winter 2022)

Chien-Yu Huang, Sameer Sainani & Richard Railton

UCSC Silicon Valley Extension

Introduction

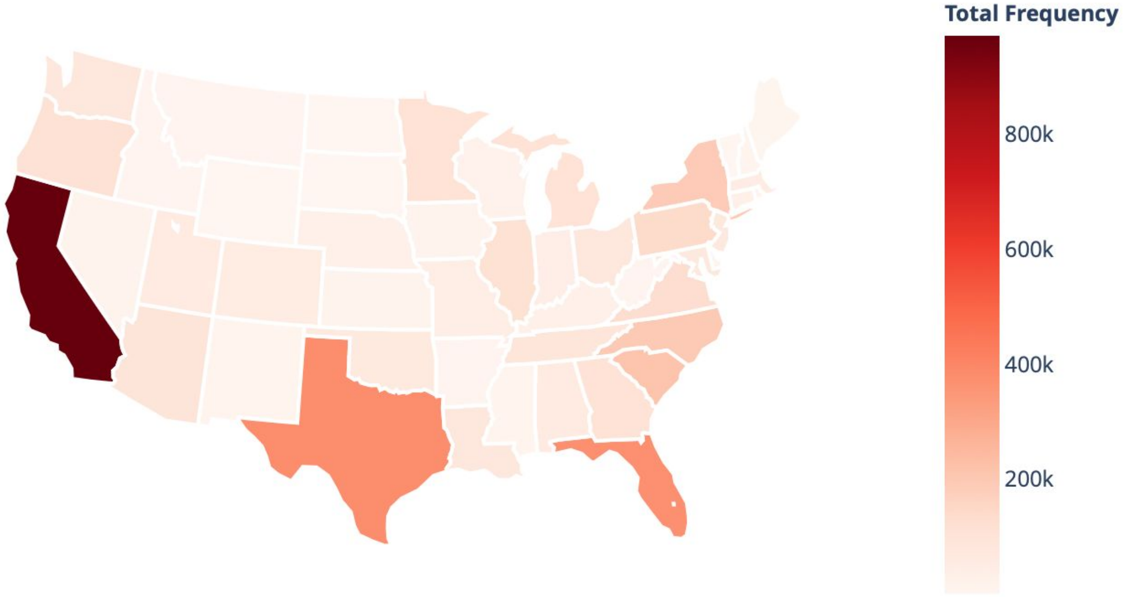
- US Accidents (2016 - 2021)
  - US Accidents (2016 - 2021) | Kaggle
- Whole dataset - 2.85 million rows
- California 2021 - 388,838 rows

Why we picked this data set

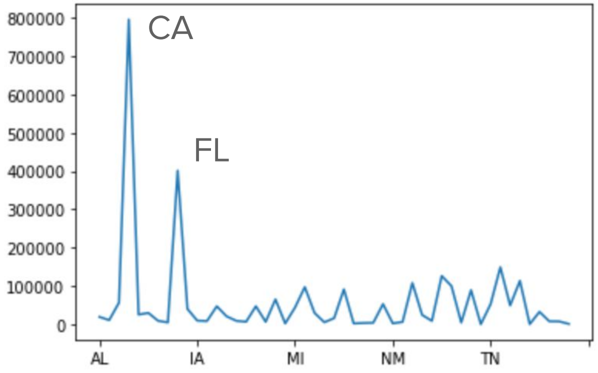
- 38,680 deaths in motor vehicle traffic crashes in 2020[1]
- Real world application
- Potential to Improve the human experience
- Large number of observations

[1] https://www.nhtsa.gov/press-releases/2020-fatality-data-show-increased-traffic-fatalities-during-pandemic

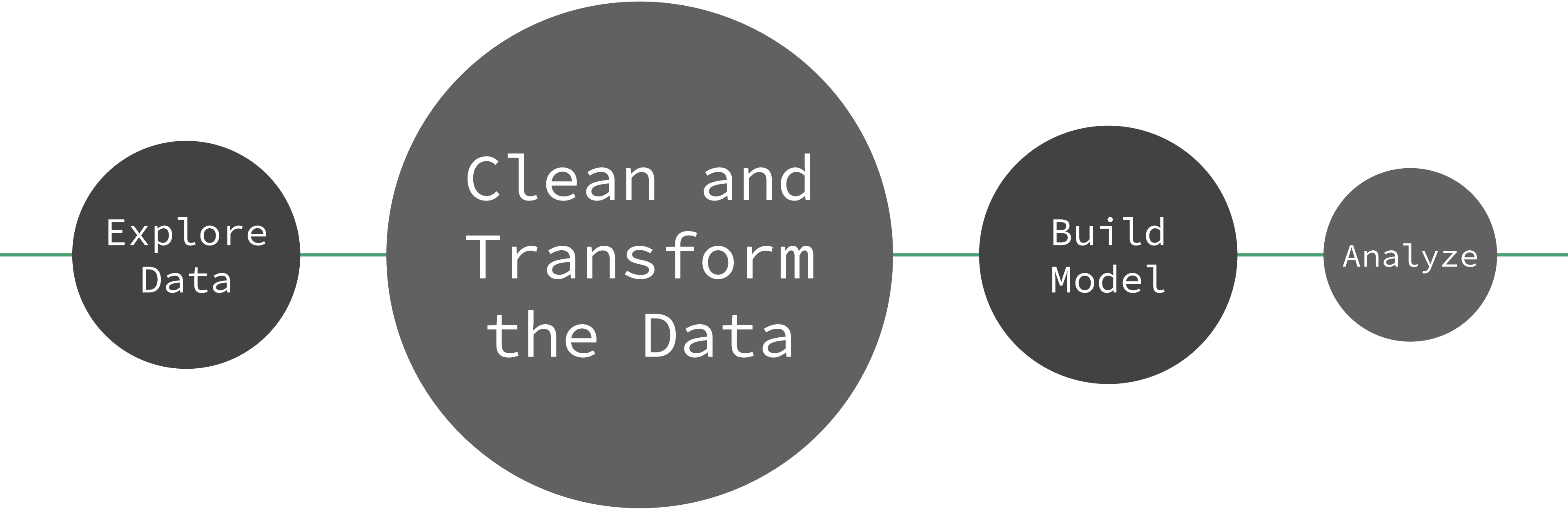
Frequency Distribution of US-Accidents (Dec 2020)
(Hover for breakdown)



Accidents by State 2016 - 2021



Phases


Explore Data
Clean and Transform the Data
Build Model
Analyze

Traffic Data Collection

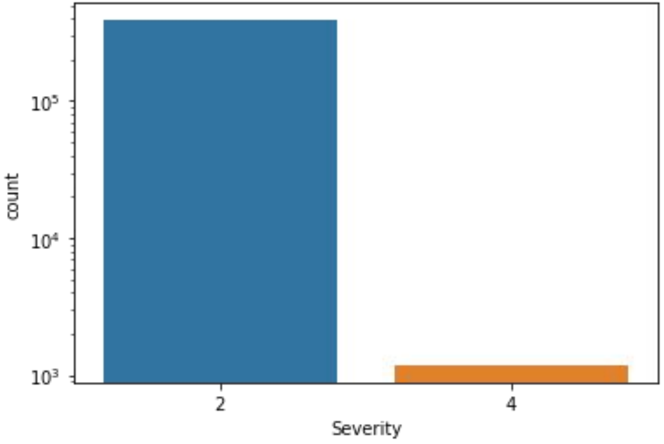
2021 streaming traffic data was collected from mainly Bing Maps whoes API

> “broadcast traffic events (accident, congestion, etc.) captured by a variety of entities - the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.”

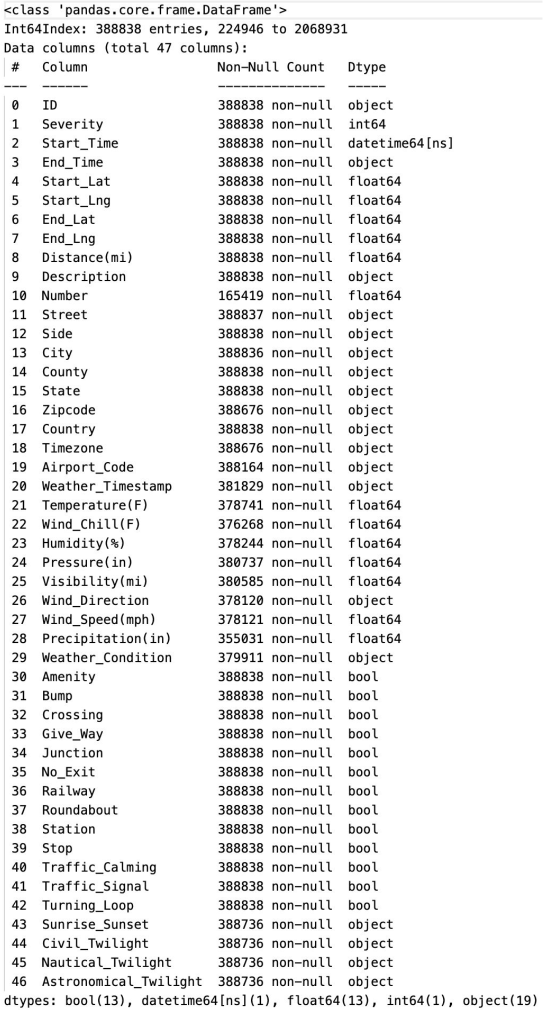
https://arxiv.org/pdf/1906.05409.pdf

Data Exploration

Severity Counter({2: 387655, 4: 1183})

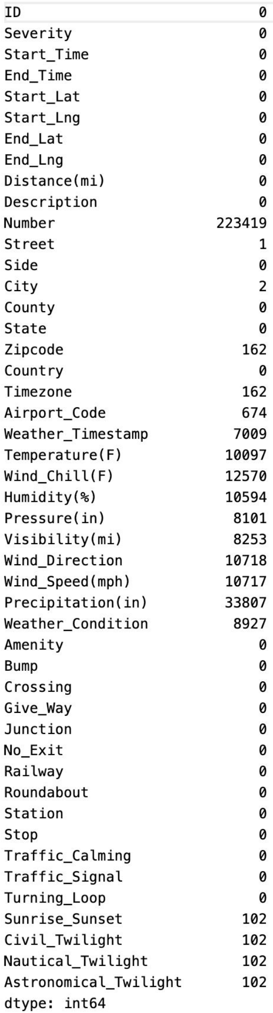


47 features



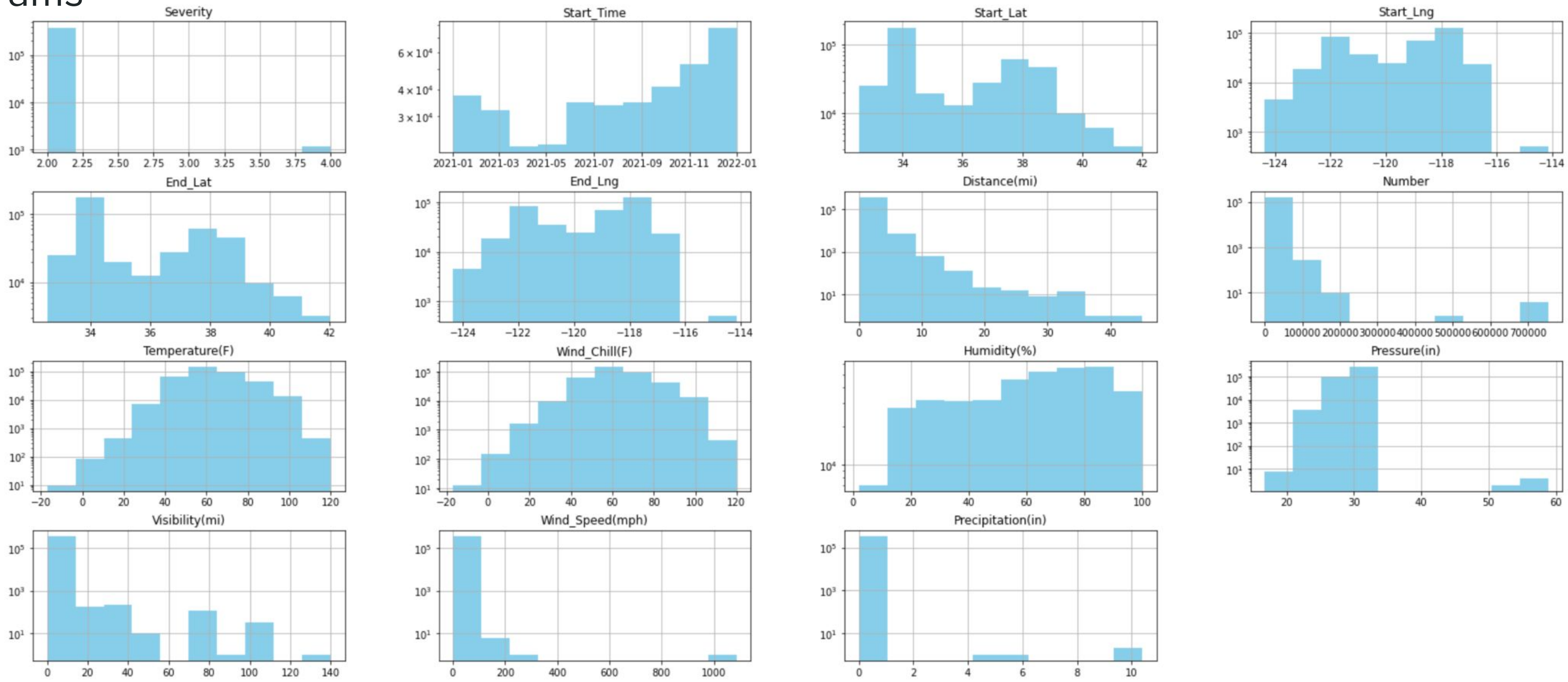
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 388838 entries, 224946 to 2068931
Data columns (total 47 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   ID                     388838 non-null  object
 1   Severity               388838 non-null  int64
 2   Start_Time             388838 non-null  datetime64[ns]
 3   End_Time               388838 non-null  object
 4   Start_Lat              388838 non-null  float64
 5   Start_Lng              388838 non-null  float64
 6   End_Lat                388838 non-null  float64
 7   End_Lng                388838 non-null  float64
 8   Distance(mi)           388838 non-null  float64
 9   Description            388838 non-null  object
 10  Number                 165419 non-null  float64
 11  Street                 388837 non-null  object
 12  Side                   388838 non-null  object
 13  City                   388836 non-null  object
 14  County                 388838 non-null  object
 15  State                  388838 non-null  object
 16  Zipcode                388676 non-null  object
 17  Country                388838 non-null  object
 18  Timezone               388676 non-null  object
 19  Airport_Code           388164 non-null  object
 20  Weather_Timestamp      381829 non-null  object
 21  Temperature(F)         378741 non-null  float64
 22  Wind_Chill(F)          376268 non-null  float64
 23  Humidity(%)            378244 non-null  float64
 24  Pressure(in)           380737 non-null  float64
 25  Visibility(mi)         380585 non-null  float64
 26  Wind_Direction         378120 non-null  object
 27  Wind_Speed(mph)        378121 non-null  float64
 28  Precipitation(in)      355031 non-null  float64
 29  Weather_Condition      379911 non-null  object
 30  Amenity                388838 non-null  bool
 31  Bump                   388838 non-null  bool
 32  Crossing               388838 non-null  bool
 33  Give_Way               388838 non-null  bool
 34  Junction               388838 non-null  bool
 35  No_Exit                388838 non-null  bool
 36  Railway                388838 non-null  bool
 37  Roundabout             388838 non-null  bool
 38  Station                388838 non-null  bool
 39  Stop                   388838 non-null  bool
 40  Traffic_Calming        388838 non-null  bool
 41  Traffic_Signal         388838 non-null  bool
 42  Turning_Loop           388838 non-null  bool
 43  Sunrise_Sunset         388736 non-null  object
 44  Civil_Twilight         388736 non-null  object
 45  Nautical_Twilight      388736 non-null  object
 46  Astronomical_Twilight  388736 non-null  object
dtypes: bool(13), datetime64[ns](1), float64(13), int64(1), object(19)
```

#na

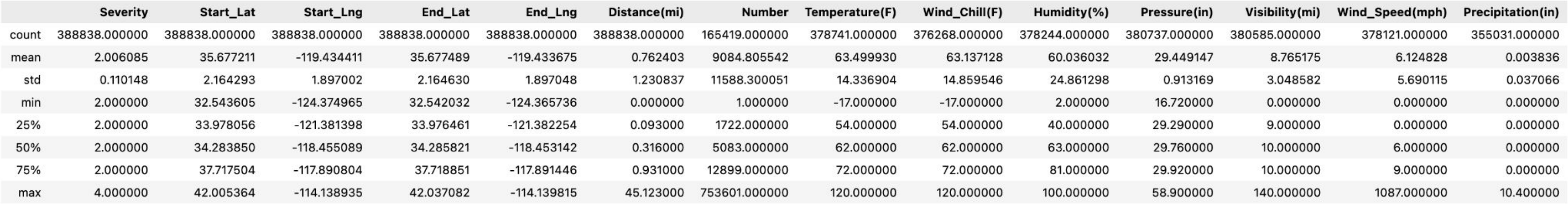


```
ID                            0
Severity                      0
Start_Time                    0
End_Time                      0
Start_Lat                     0
Start_Lng                     0
End_Lat                       0
End_Lng                       0
Distance(mi)                  0
Description                   0
Number                   223419
Street                        1
Side                          0
City                          2
County                        0
State                         0
Zipcode                     162
Country                       0
Timezone                    162
Airport_Code                674
Weather_Timestamp          7009
Temperature(F)            10097
Wind_Chill(F)             12570
Humidity(%)               10594
Pressure(in)               8101
Visibility(mi)             8253
Wind_Direction            10718
Wind_Speed(mph)           10717
Precipitation(in)         33807
Weather_Condition          8927
Amenity                       0
Bump                          0
Crossing                      0
Give_Way                      0
Junction                      0
No_Exit                       0
Railway                       0
Roundabout                    0
Station                       0
Stop                          0
Traffic_Calming               0
Traffic_Signal                0
Turning_Loop                  0
Sunrise_Sunset              102
Civil_Twilight              102
Nautical_Twilight           102
Astronomical_Twilight       102
dtype: int64
```

Histograms



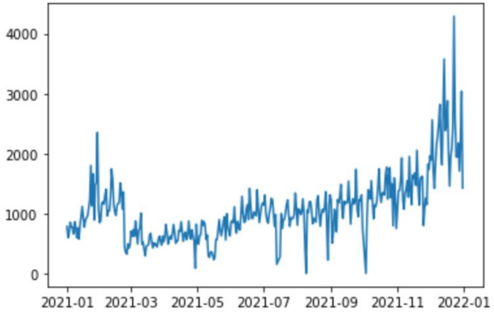
Descriptive Statistics



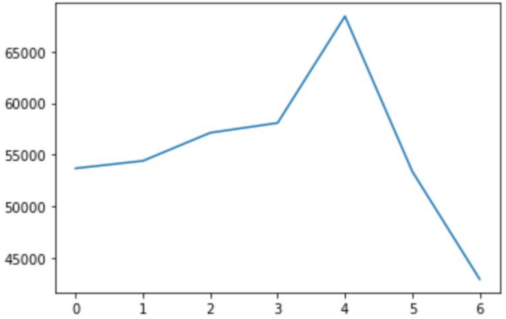
	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Number	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
count	388838.000000	388838.000000	388838.000000	388838.000000	388838.000000	388838.000000	165419.000000	378741.000000	376268.000000	378244.000000	380737.000000	380585.000000	378121.000000	355031.000000
mean	2.006085	35.677211	-119.434411	35.677489	-119.433675	0.762403	9084.805542	63.499930	63.137128	60.036032	29.449147	8.765175	6.124828	0.003836
std	0.110148	2.164293	1.897002	2.164630	1.897048	1.230837	11588.300051	14.336904	14.859546	24.861298	0.913169	3.048582	5.690115	0.037066
min	2.000000	32.543605	-124.374965	32.542032	-124.365736	0.000000	1.000000	-17.000000	-17.000000	2.000000	16.720000	0.000000	0.000000	0.000000
25%	2.000000	33.978056	-121.381398	33.976461	-121.382254	0.093000	1722.000000	54.000000	54.000000	40.000000	29.290000	9.000000	0.000000	0.000000
50%	2.000000	34.283850	-118.455089	34.285821	-118.453142	0.316000	5083.000000	62.000000	62.000000	63.000000	29.760000	10.000000	6.000000	0.000000
75%	2.000000	37.717504	-117.890804	37.718851	-117.891446	0.931000	12899.000000	72.000000	72.000000	81.000000	29.920000	10.000000	9.000000	0.000000
max	4.000000	42.005364	-114.138935	42.037082	-114.139815	45.123000	753601.000000	120.000000	120.000000	100.000000	58.900000	140.000000	1087.000000	10.400000

Time Series Plots - CA 2021

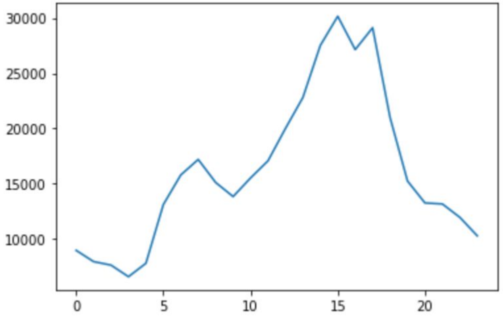
Count by Date



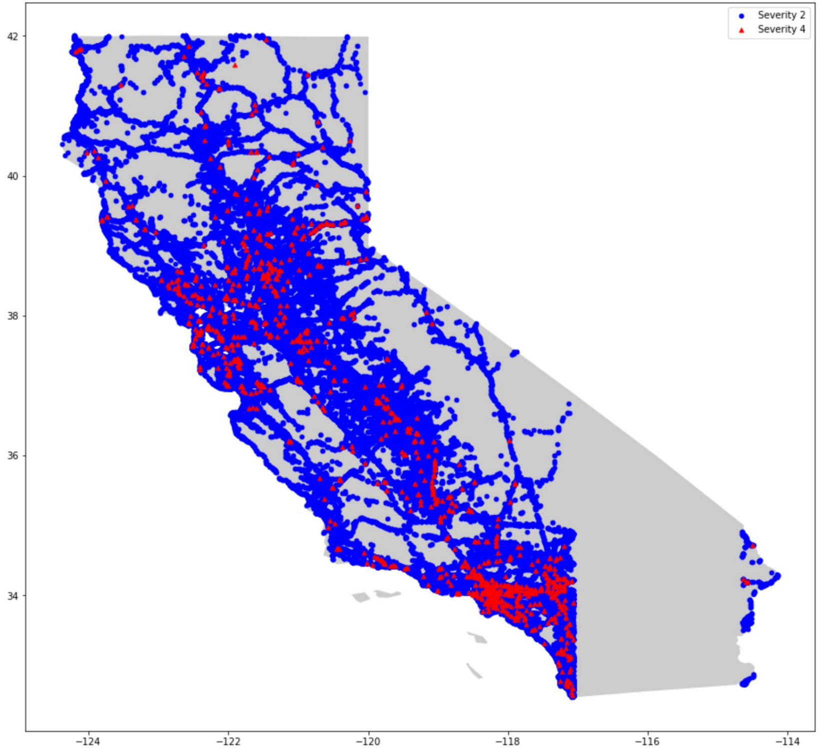
Count by Weekday



Count by Hour



Map of Accidents - CA 2021



- Seem to be missing some data from the South East of California.

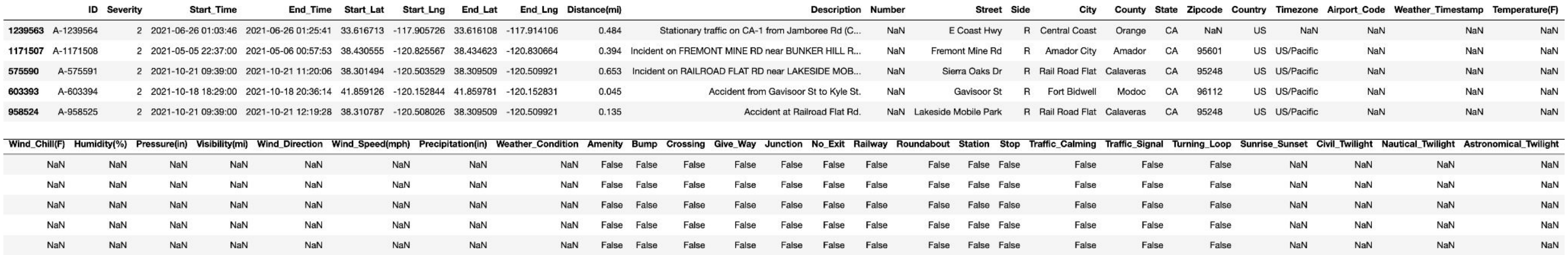
- Severity 4 accidents are clustered around urban areas.

Data Cleaning

- Check for duplicates
- Drop irrelevant columns
  - 'ID', 'Description', 'Number', 'End_Time', 'End_Lat', 'End_Lng', 'Country', 'Turning_Loop', 'Weather_Timestamp'
- Convert Start_Time to datetime
- Create new date features
  - 'Date', 'Year', 'Month', 'Day'
- Fill na’s for weather float features
  - Use medians and group on Month and Airport_Code/City/County
  - Drop remaining na’s
- Fill na’s for weather object features
  - Use modes and group on Month and Pressure_binned
  - Drop remaining na’s
- Drop remaining na’s (<1000 rows)
- Remove obvious outliers

Data Cleaning

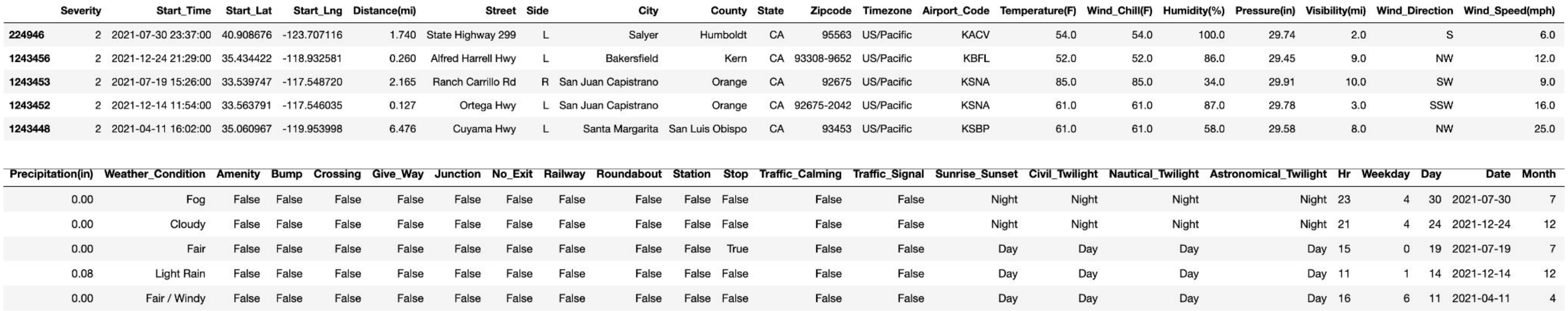
Before Cleaning



	ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Description	Number	Street	Side	City	County	State	Zipcode	Country	Timezone	Airport_Code	Weather_Timestamp	Temperature(F)
1239563	A-1239564	2	2021-06-26 01:03:46	2021-06-26 01:25:41	33.616713	-117.905726	33.616108	-117.914106	0.484	Stationary traffic on CA-1 from Jamboree Rd (C...	NaN	E Coast Hwy	R	Central Coast	Orange	CA	NaN	US	NaN	NaN	NaN	NaN
1171507	A-1171508	2	2021-05-05 22:37:00	2021-05-06 00:57:53	38.430555	-120.825567	38.434623	-120.830664	0.394	Incident on FREMONT MINE RD near BUNKER HILL R...	NaN	Fremont Mine Rd	R	Amador City	Amador	CA	95601	US	US/Pacific	NaN	NaN	NaN
575590	A-575591	2	2021-10-21 09:39:00	2021-10-21 11:20:06	38.301494	-120.503529	38.309509	-120.509921	0.653	Incident on RAILROAD FLAT RD near LAKESIDE MOB...	NaN	Sierra Oaks Dr	R	Rail Road Flat	Calaveras	CA	95248	US	US/Pacific	NaN	NaN	NaN
603393	A-603394	2	2021-10-18 18:29:00	2021-10-18 20:36:14	41.859126	-120.152844	41.859781	-120.152831	0.045	Accident from Gavisoor St to Kyle St.	NaN	Gavisoor St	R	Fort Bidwell	Modoc	CA	96112	US	US/Pacific	NaN	NaN	NaN
958524	A-958525	2	2021-10-21 09:39:00	2021-10-21 12:19:28	38.310787	-120.508026	38.309509	-120.509921	0.135	Accident at Railroad Flat Rd.	NaN	Lakeside Mobile Park	R	Rail Road Flat	Calaveras	CA	95248	US	US/Pacific	NaN	NaN	NaN

Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Direction	Wind_Speed(mph)	Precipitation(in)	Weather_Condition	Amenity	Bump	Crossing	Give_Way	Junction	No_Exit	Railway	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop	Sunrise_Sunset	Civil_Twilight	Nautical_Twilight	Astronomical_Twilight
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False	False	False	False	False	False	False	False	False	False	False	False	False	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False	False	False	False	False	False	False	False	False	False	False	False	False	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False	False	False	False	False	False	False	False	False	False	False	False	False	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False	False	False	False	False	False	False	False	False	False	False	False	False	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False	False	False	False	False	False	False	False	False	False	False	False	False	NaN	NaN	NaN	NaN

After Cleaning

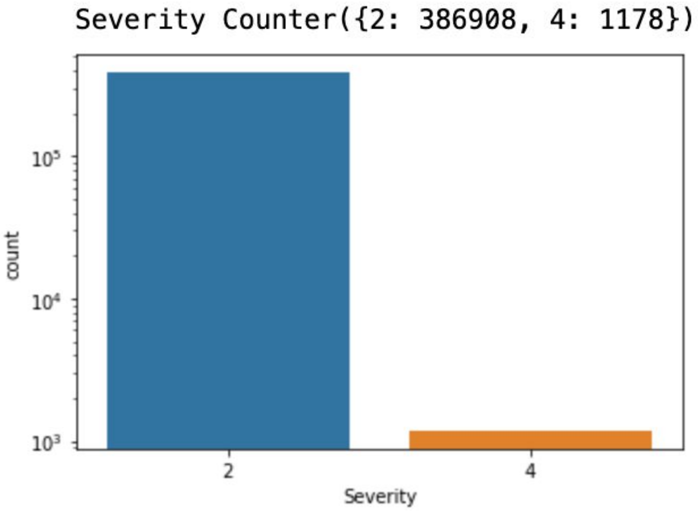


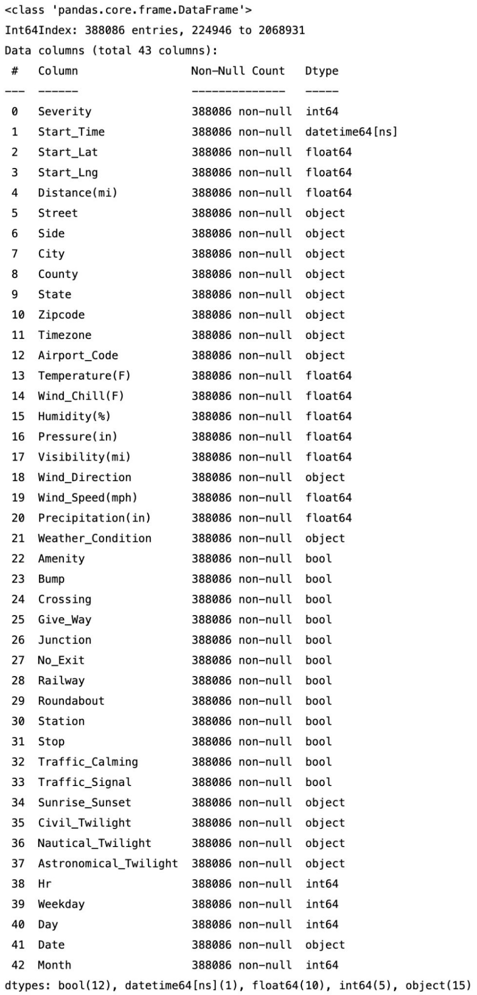
	Severity	Start_Time	Start_Lat	Start_Lng	Distance(mi)	Street	Side	City	County	State	Zipcode	Timezone	Airport_Code	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Direction	Wind_Speed(mph)
224946	2	2021-07-30 23:37:00	40.908676	-123.707116	1.740	State Highway 299	L	Salyer	Humboldt	CA	95563	US/Pacific	KACV	54.0	54.0	100.0	29.74	2.0	S	6.0
1243456	2	2021-12-24 21:29:00	35.434422	-118.932581	0.260	Alfred Harrell Hwy	L	Bakersfield	Kern	CA	93308-9652	US/Pacific	KBFL	52.0	52.0	86.0	29.45	9.0	NW	12.0
1243453	2	2021-07-19 15:26:00	33.539747	-117.548720	2.165	Ranch Carrillo Rd	R	San Juan Capistrano	Orange	CA	92675	US/Pacific	KSNA	85.0	85.0	34.0	29.91	10.0	SW	9.0
1243452	2	2021-12-14 11:54:00	33.563791	-117.546035	0.127	Ortega Hwy	L	San Juan Capistrano	Orange	CA	92675-2042	US/Pacific	KSNA	61.0	61.0	87.0	29.78	3.0	SSW	16.0
1243448	2	2021-04-11 16:02:00	35.060967	-119.953998	6.476	Cuyama Hwy	L	Santa Margarita	San Luis Obispo	CA	93453	US/Pacific	KSBP	61.0	61.0	58.0	29.58	8.0	NW	25.0

Precipitation(in)	Weather_Condition	Amenity	Bump	Crossing	Give_Way	Junction	No_Exit	Railway	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Sunrise_Sunset	Civil_Twilight	Nautical_Twilight	Astronomical_Twilight	Hr	Weekday	Day	Date	Month
0.00	Fog	False	False	False	False	False	False	False	False	False	False	False	False	Night	Night	Night	Night	23	4	30	2021-07-30	7
0.00	Cloudy	False	False	False	False	False	False	False	False	False	False	False	False	Night	Night	Night	Night	21	4	24	2021-12-24	12
0.00	Fair	False	False	False	False	False	False	False	False	False	True	False	False	Day	Day	Day	Day	15	0	19	2021-07-19	7
0.08	Light Rain	False	False	False	False	False	False	False	False	False	False	False	False	Day	Day	Day	Day	11	1	14	2021-12-14	12
0.00	Fair / Windy	False	False	False	False	False	False	False	False	False	False	False	False	Day	Day	Day	Day	16	6	11	2021-04-11	4

Data Cleaning

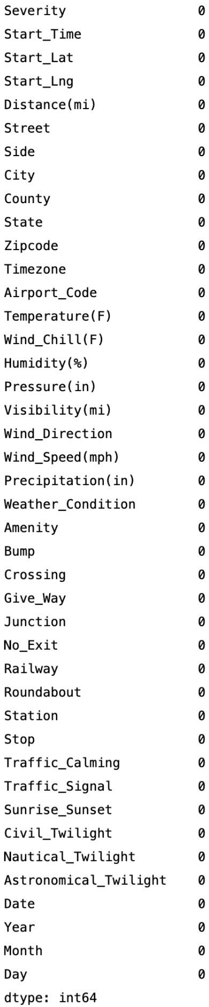
43 features remaining


Severity Counter({2: 386908, 4: 1178})



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 388086 entries, 224946 to 2068931
Data columns (total 43 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   Severity               388086 non-null  int64
 1   Start_Time             388086 non-null  datetime64[ns]
 2   Start_Lat              388086 non-null  float64
 3   Start_Lng              388086 non-null  float64
 4   Distance(mi)           388086 non-null  float64
 5   Street                 388086 non-null  object
 6   Side                   388086 non-null  object
 7   City                   388086 non-null  object
 8   County                 388086 non-null  object
 9   State                  388086 non-null  object
 10  Zipcode                388086 non-null  object
 11  Timezone               388086 non-null  object
 12  Airport_Code           388086 non-null  object
 13  Temperature(F)         388086 non-null  float64
 14  Wind_Chill(F)          388086 non-null  float64
 15  Humidity(%)            388086 non-null  float64
 16  Pressure(in)           388086 non-null  float64
 17  Visibility(mi)         388086 non-null  float64
 18  Wind_Direction         388086 non-null  object
 19  Wind_Speed(mph)        388086 non-null  float64
 20  Precipitation(in)      388086 non-null  float64
 21  Weather_Condition      388086 non-null  object
 22  Amenity                388086 non-null  bool
 23  Bump                   388086 non-null  bool
 24  Crossing               388086 non-null  bool
 25  Give_Way               388086 non-null  bool
 26  Junction               388086 non-null  bool
 27  No_Exit                388086 non-null  bool
 28  Railway                388086 non-null  bool
 29  Roundabout             388086 non-null  bool
 30  Station                388086 non-null  bool
 31  Stop                   388086 non-null  bool
 32  Traffic_Calming        388086 non-null  bool
 33  Traffic_Signal         388086 non-null  bool
 34  Sunrise_Sunset         388086 non-null  object
 35  Civil_Twilight         388086 non-null  object
 36  Nautical_Twilight      388086 non-null  object
 37  Astronomical_Twilight  388086 non-null  object
 38  Hr                     388086 non-null  int64
 39  Weekday                388086 non-null  int64
 40  Day                    388086 non-null  int64
 41  Date                   388086 non-null  object
 42  Month                  388086 non-null  int64
dtypes: bool(12), datetime64[ns](1), float64(10), int64(5), object(15)
```

#na



```
Severity                 0
Start_Time               0
Start_Lat                0
Start_Lng                0
Distance(mi)             0
Street                   0
Side                     0
City                     0
County                   0
State                    0
Zipcode                  0
Timezone                 0
Airport_Code             0
Temperature(F)           0
Wind_Chill(F)            0
Humidity(%)              0
Pressure(in)             0
Visibility(mi)           0
Wind_Direction           0
Wind_Speed(mph)          0
Precipitation(in)        0
Weather_Condition        0
Amenity                  0
Bump                     0
Crossing                 0
Give_Way                 0
Junction                 0
No_Exit                  0
Railway                  0
Roundabout               0
Station                  0
Stop                     0
Traffic_Calming          0
Traffic_Signal           0
Sunrise_Sunset           0
Civil_Twilight           0
Nautical_Twilight        0
Astronomical_Twilight    0
Date                     0
Year                     0
Month                    0
Day                      0
dtype: int64
```

Feature Engineering

- Hour Category
  - Create Weekday and Hour and use them to create Rush, Day, Night, Weekend_Day and Weekend_Night features.
- Wind Direction
  - Simplified to 'S', 'W', 'CALM', 'VAR', 'E', 'N', 'SE', 'SW', 'NW', 'NE'
- Wind Condition
  - Simplified to Clear, Cloud, Rain, Heavy_Rain, Snow, Heavy_Snow, Fog
- Street
  - Created Boolean features for the 40 most common words in street name
- Drop redundant columns
  - 'Start_Time', 'Start_Lat', 'Start_Lng', 'Distance(mi)', 'Street', 'City', 'State', 'Zipcode', 'Airport_Code', 'Wind_Direction', 'Weather_Condition', 'Date'

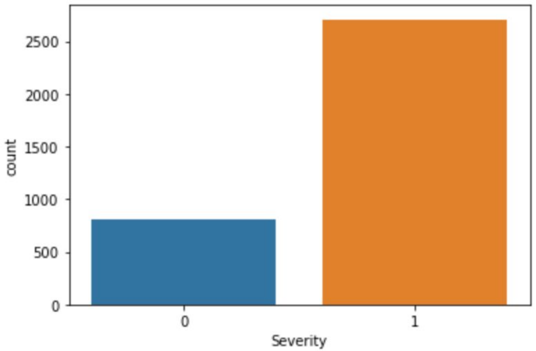
Challenges

- Data Collection Method Issues
- Severity value discrepancies between sources
- Unbalanced target variable (Severity)
- Features of Large Urban Populations may have undue influence

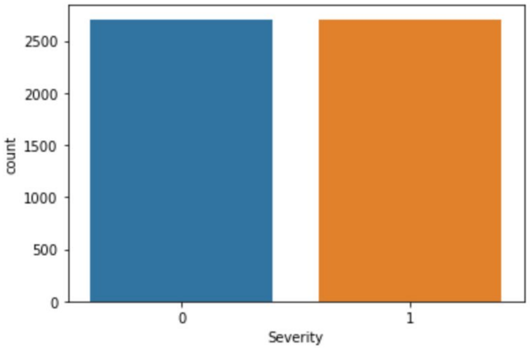
Data Preprocessing and Methodology

- Standardize numeric features
- Split dataset into Training and Testing (7:3)
- Handling Imbalanced training data
  - UnderSampling
  - OverSampling
- Testing data remains imbalanced

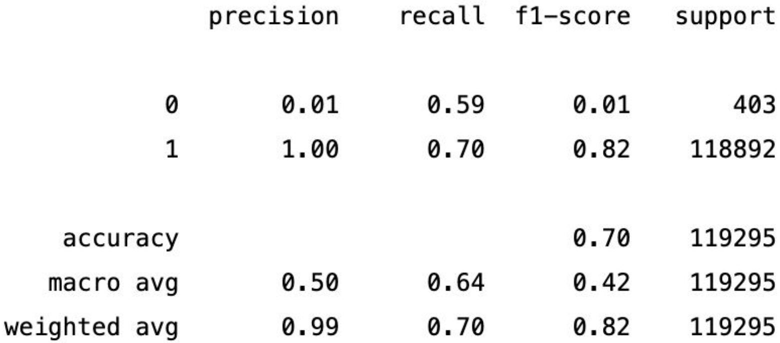
```
Severity Counter({1: 2708, 0: 810})
```



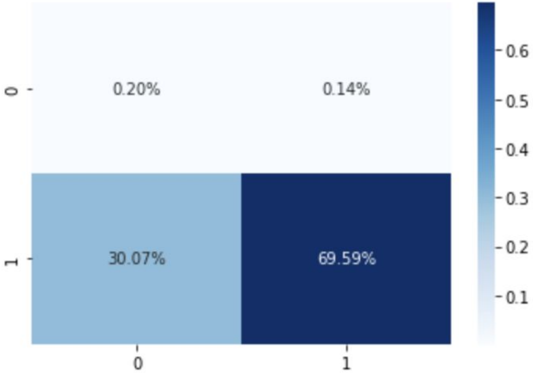
```
Severity Counter({1: 2708, 0: 2708})
```



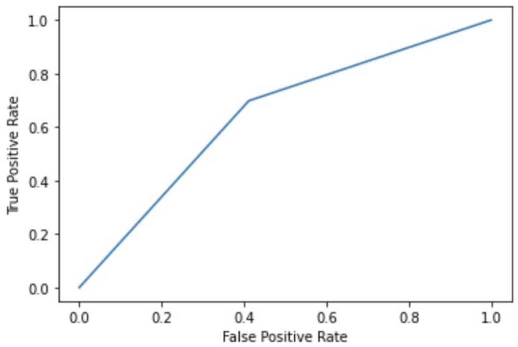
kNN Model



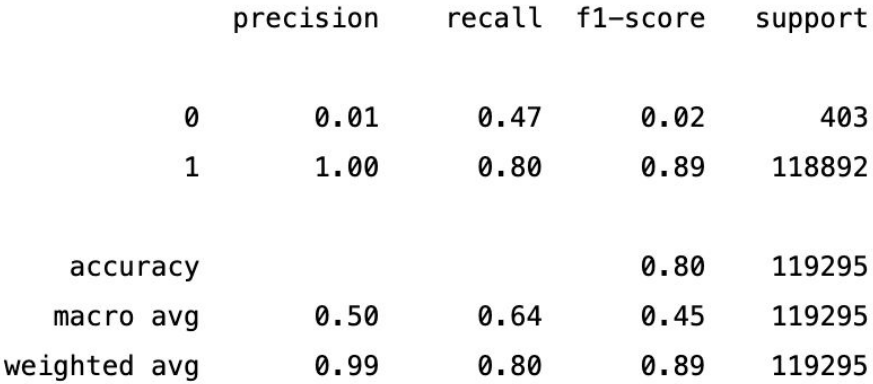
```
              precision    recall  f1-score   support

           0       0.01      0.59      0.01       403
           1       1.00      0.70      0.82    118892

    accuracy                           0.70    119295
   macro avg       0.50      0.64      0.42    119295
weighted avg       0.99      0.70      0.82    119295
```



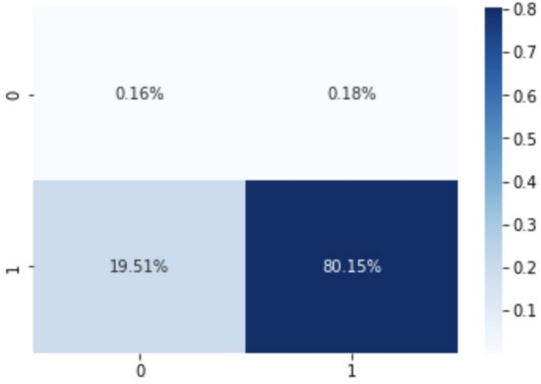
```
ROC_AUC score: 64.32%
```



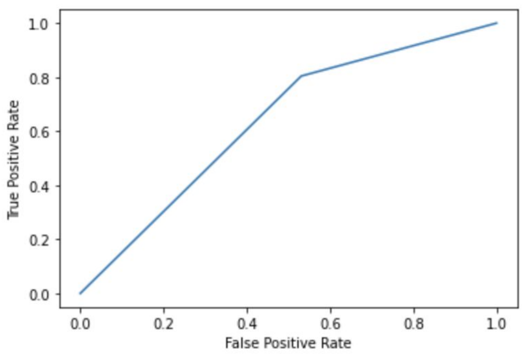
SVM - Classifier Model



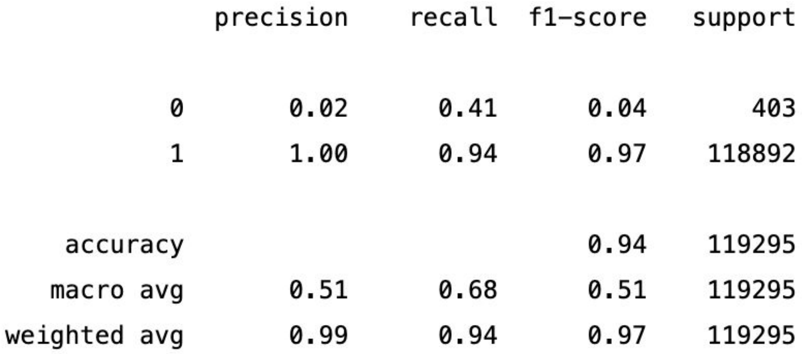
```
              precision    recall  f1-score   support

           0       0.01      0.47      0.02       403
           1       1.00      0.80      0.89    118892

    accuracy                           0.80    119295
   macro avg       0.50      0.64      0.45    119295
weighted avg       0.99      0.80      0.89    119295
```



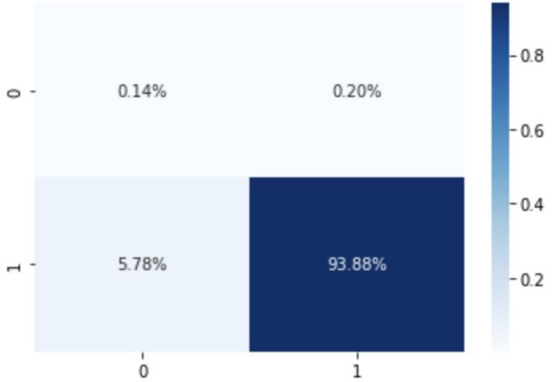
```
ROC_AUC score: 63.66%
```

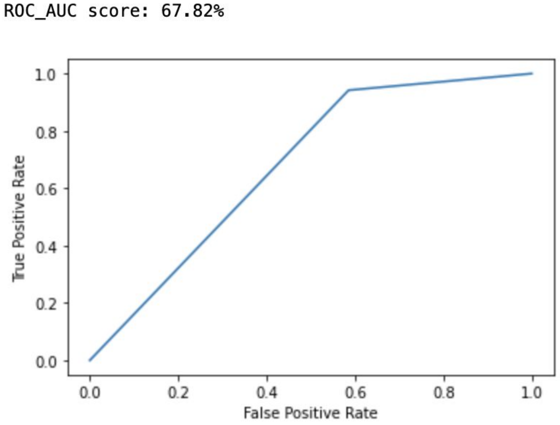


Random Forest Model



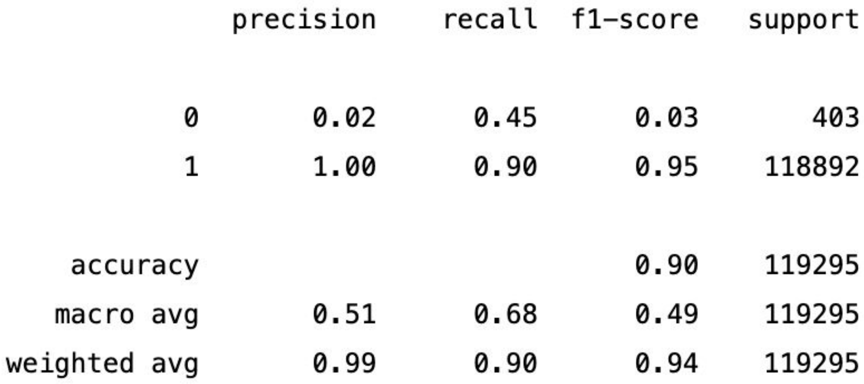
```
              precision    recall  f1-score   support

           0       0.02      0.41      0.04       403
           1       1.00      0.94      0.97    118892

    accuracy                           0.94    119295
   macro avg       0.51      0.68      0.51    119295
weighted avg       0.99      0.94      0.97    119295
```



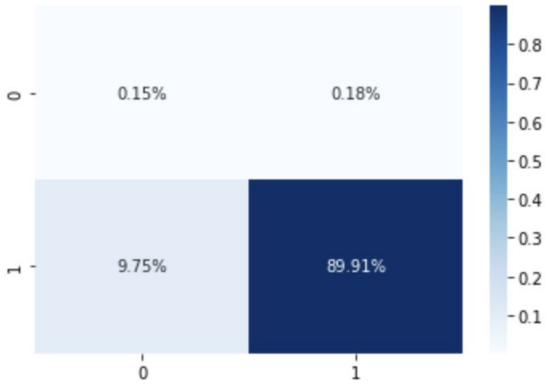


```
ROC_AUC score: 67.82%
```

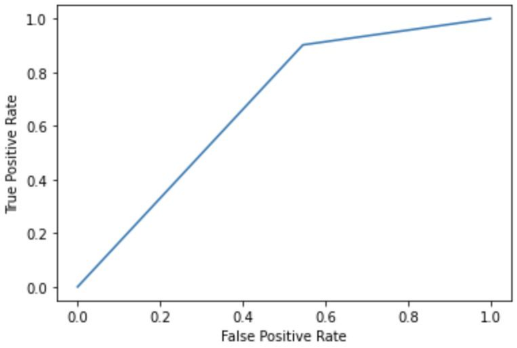
XGBoost Model



```
              precision    recall  f1-score   support

           0       0.02      0.45      0.03       403
           1       1.00      0.90      0.95    118892

    accuracy                           0.90    119295
   macro avg       0.51      0.68      0.49    119295
weighted avg       0.99      0.90      0.94    119295
```



ROC_AUC score: 67.81%



Comparing the Models

Model	**ROC_AUC**	**Precision**	**Recall**	**F1**
KNN	64.32%	99.8%	69.83%	82.17%
SVM	63.66%	99.78%	80.42%	89.06%
Random Forest	67.82%	99.79%	94.2%	96.92%
XGboost	67.81%	99.8%	90.22%	94.76%

Analyzing the Results

- Initial results had very high accuracy but we suspected it due to imbalanced data
  - With under and over sampling we hoped to eliminate this issue

- Random forest and XGBoost model perform better than KNN and SVM
  - RF and XGB are tree based algorithms work better with imbalanced data.
  - For imbalanced data, recall and F1 score are more important than precision and accuracy

Other Possible Algorithms / Techniques

Data Cleaning

- KNN for imputation
- Multiple imputation using a Normal Distribution

Machine Learning

- Logistic Regression
  - We could have used logistic regression as a baseline[1]
- DNN

Feature Selection

- Variance, PCA and Random Forest attempted prior to undersampling and oversampling.

[1] https://arxiv.org/pdf/1909.09638.pdf (Section 6.2)

Conclusions

- Data collection out of our control - requires review.
- Model is only as good as the underlying data - the key feature, Severity, was poorly defined.
- Difficult to separate the signal from the noise
- Given more time we would investigate predicting the number of accidents rather than severity.
- Dataset Remorse

Thank You!

Chien-yu Huang
Richard Railton
Sameer Sainani