U.S. Traffic Accidents

Data Science Career Track

Sara Satti

August 2020

designed using resources from PoweredTemplate.com

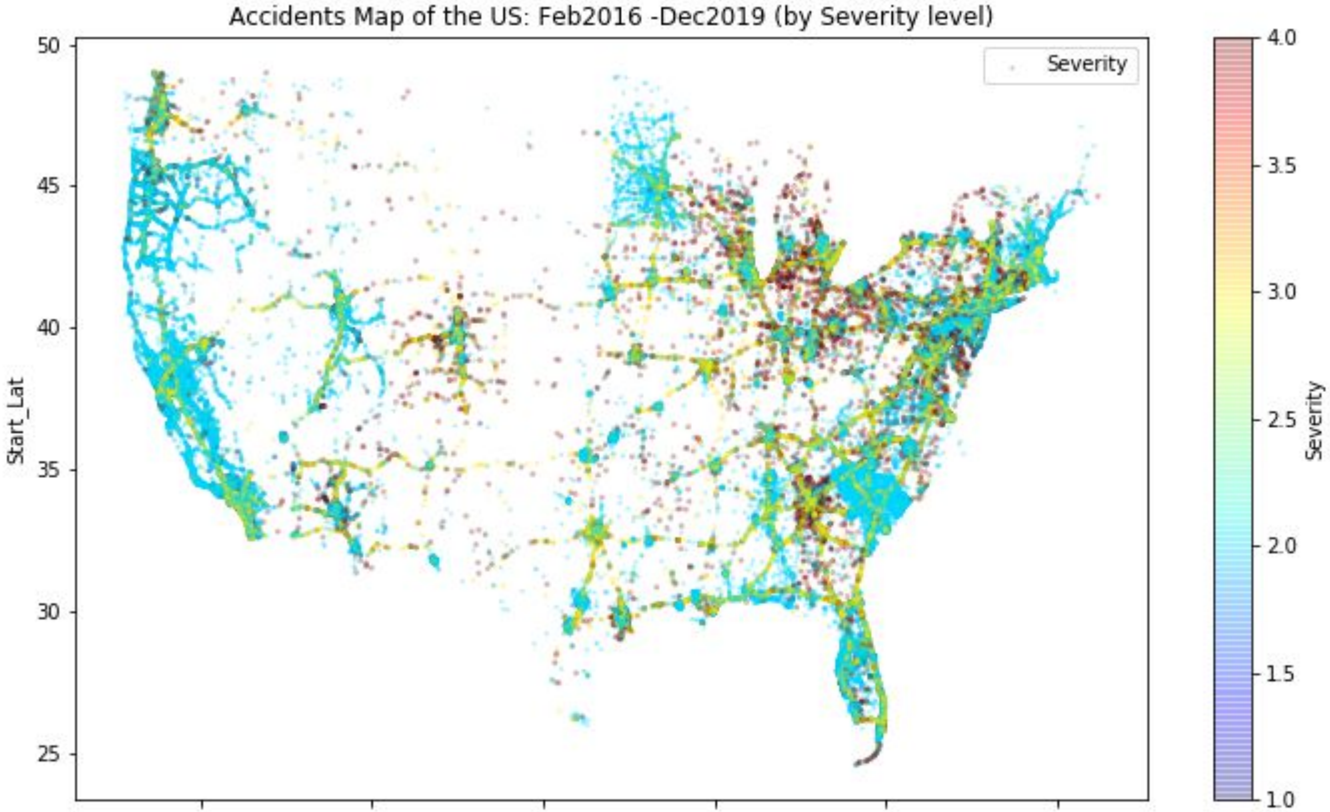
Objective:

Predict road accident severity given weather conditions, location, points of interest, and other engineered features.

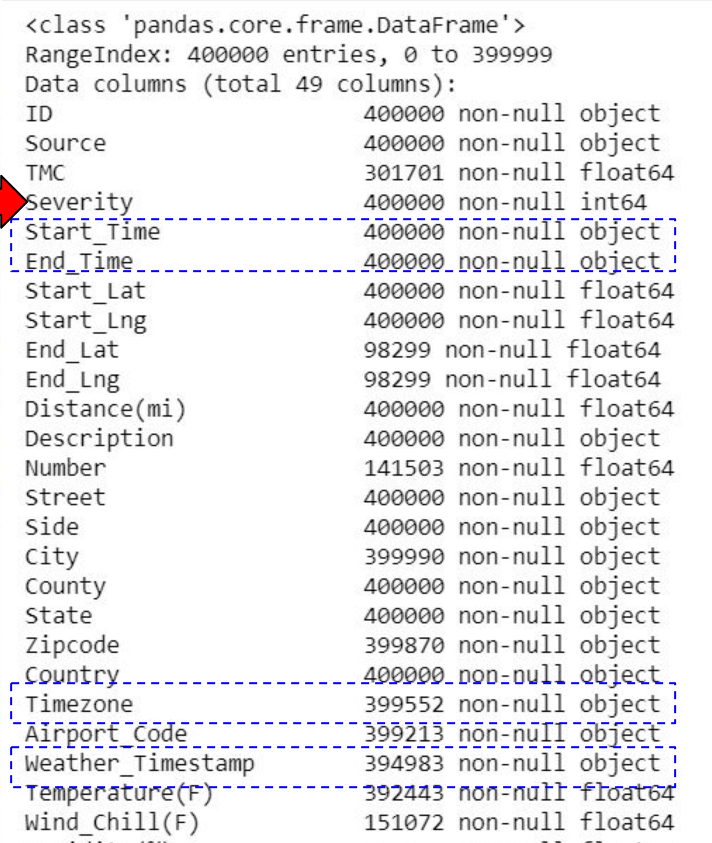
Outline

- Data Wrangling
  - Dataset
  - Data Manipulation
    - Missing Data Handling
    - Feature Engineering
- Exploratory Data Analysis
- Modeling
- Conclusions and Recommendations

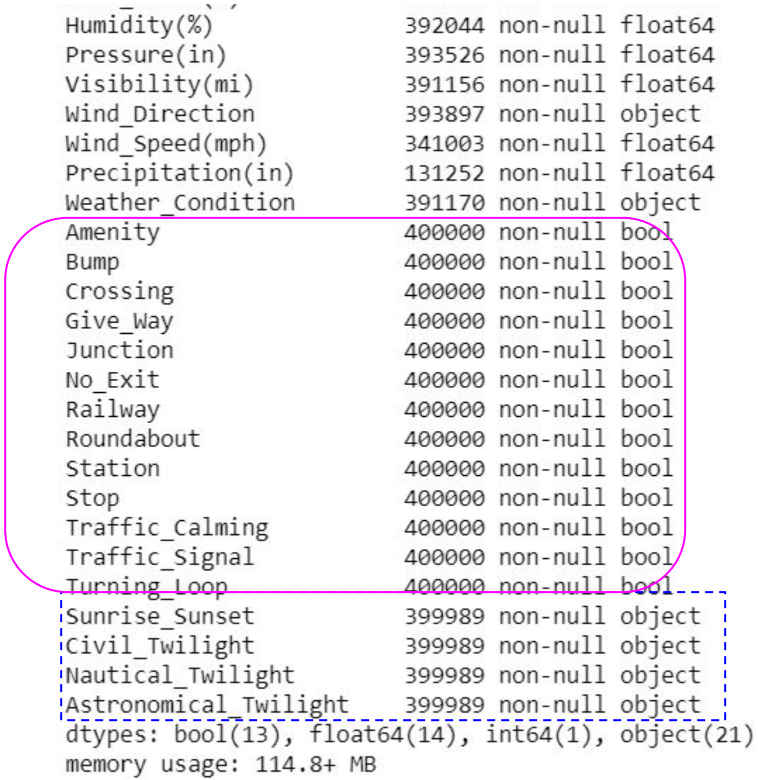
Data Wrangling


Accidents Map of the US: Feb2016 -Dec2019 (by Severity level)
Severity
Start_Lat
50
45
40
35
30
25
Severity
4.0
3.5
3.0
2.5
2.0
1.5
1.0

Dataset:



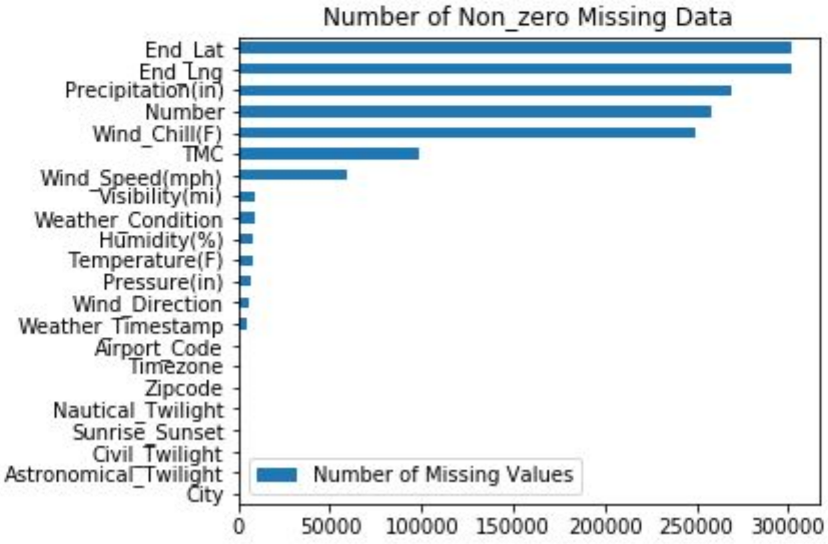
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400000 entries, 0 to 399999
Data columns (total 49 columns):
ID                       400000 non-null object
Source                   400000 non-null object
TMC                      301701 non-null float64
Severity                 400000 non-null int64
Start_Time               400000 non-null object
End_Time                 400000 non-null object
Start_Lat                400000 non-null float64
Start_Lng                400000 non-null float64
End_Lat                  98299 non-null float64
End_Lng                  98299 non-null float64
Distance(mi)             400000 non-null float64
Description              400000 non-null object
Number                   141503 non-null float64
Street                   400000 non-null object
Side                     400000 non-null object
City                     399990 non-null object
County                   400000 non-null object
State                    400000 non-null object
Zipcode                  399870 non-null object
Country                  400000 non-null object
Timezone                 399552 non-null object
Airport_Code             399213 non-null object
Weather_Timestamp        394983 non-null object
Temperature(F)           392443 non-null float64
Wind_Chill(F)            151072 non-null float64
```



```
Humidity(%)              392044 non-null float64
Pressure(in)             393526 non-null float64
Visibility(mi)           391156 non-null float64
Wind_Direction           393897 non-null object
Wind_Speed(mph)          341003 non-null float64
Precipitation(in)        131252 non-null float64
Weather_Condition        391170 non-null object
Amenity                  400000 non-null bool
Bump                     400000 non-null bool
Crossing                 400000 non-null bool
Give_Way                 400000 non-null bool
Junction                 400000 non-null bool
No_Exit                  400000 non-null bool
Railway                  400000 non-null bool
Roundabout               400000 non-null bool
Station                  400000 non-null bool
Stop                     400000 non-null bool
Traffic_Calming          400000 non-null bool
Traffic_Signal           400000 non-null bool
Turning_Loop             400000 non-null bool
Sunrise_Sunset           399989 non-null object
Civil_Twilight           399989 non-null object
Nautical_Twilight        399989 non-null object
Astronomical_Twilight    399989 non-null object
dtypes: bool(13), float64(14), int64(1), object(21)
memory usage: 114.8+ MB
```

Data Manipulation - 1

Missing data handling:



- ➔ isnull() count, and wasnull column addition.
- ➔ Missing precipitation = 0
- ➔ Start Lat/lng instead of end Lat/Lng.
- ➔ Wind_Chill: through linear regression.
- ➔ Others by infilling with Mean/median.

Data Manipulation - 2

Feature Engineering

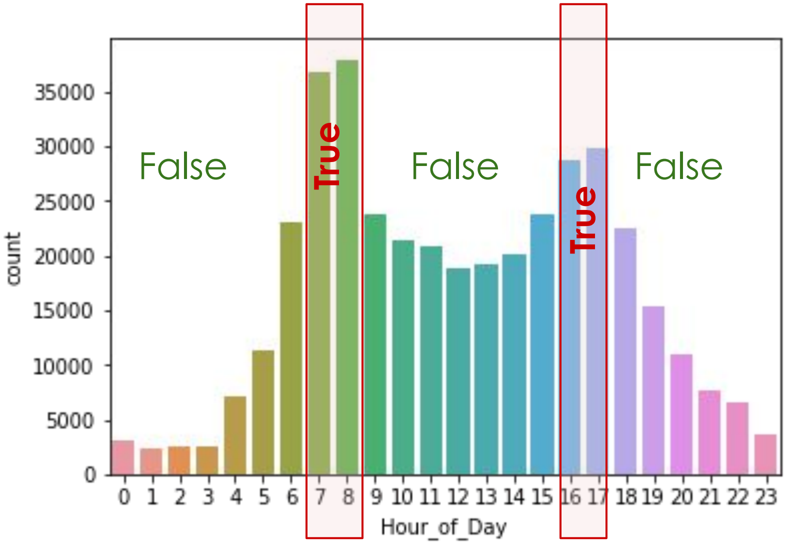
1. Wasnull columns → tag null values within DF

from datetime data

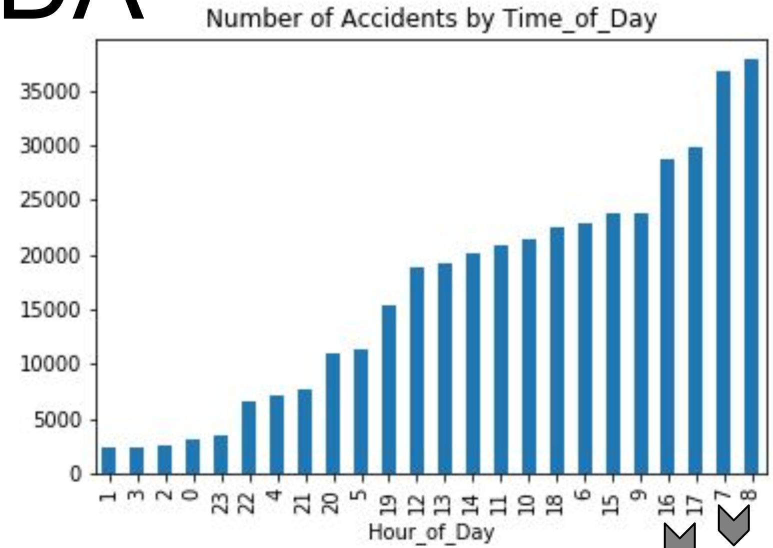
2. Day_of_Week

3. Hour_of_Day

Rush Hour

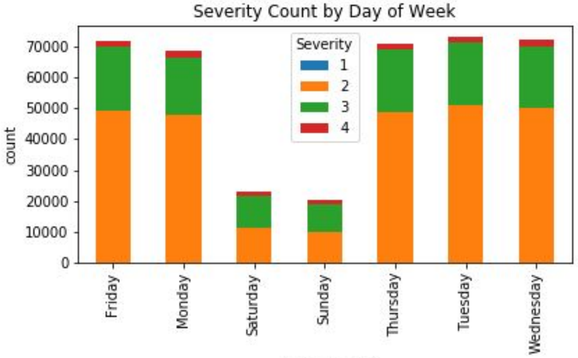


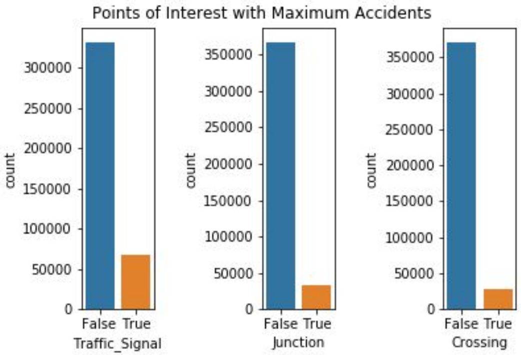
EDA

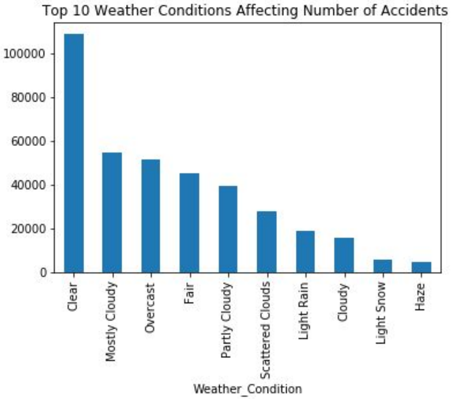


Morning R.H.

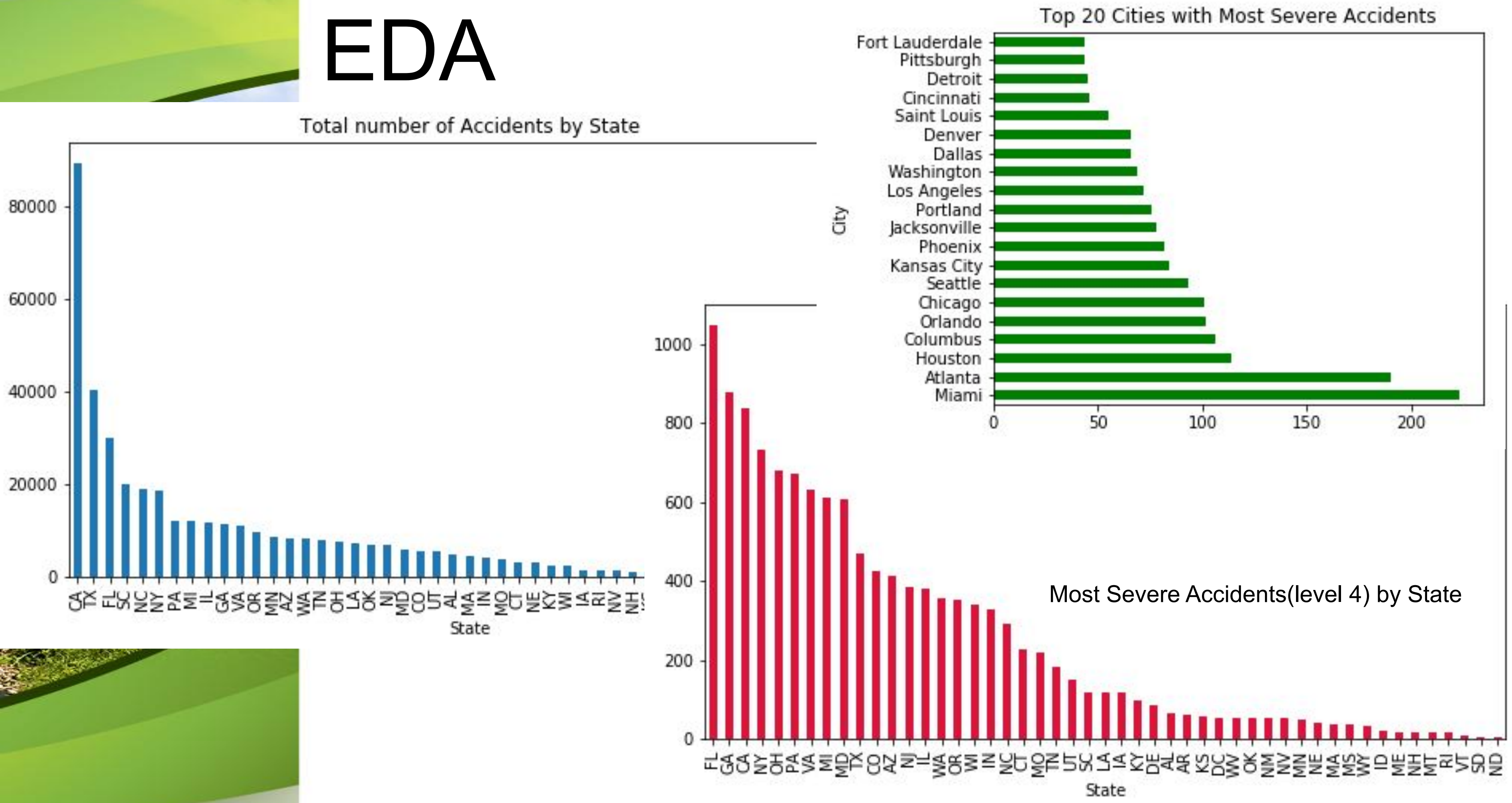
Evening Rush hour







EDA



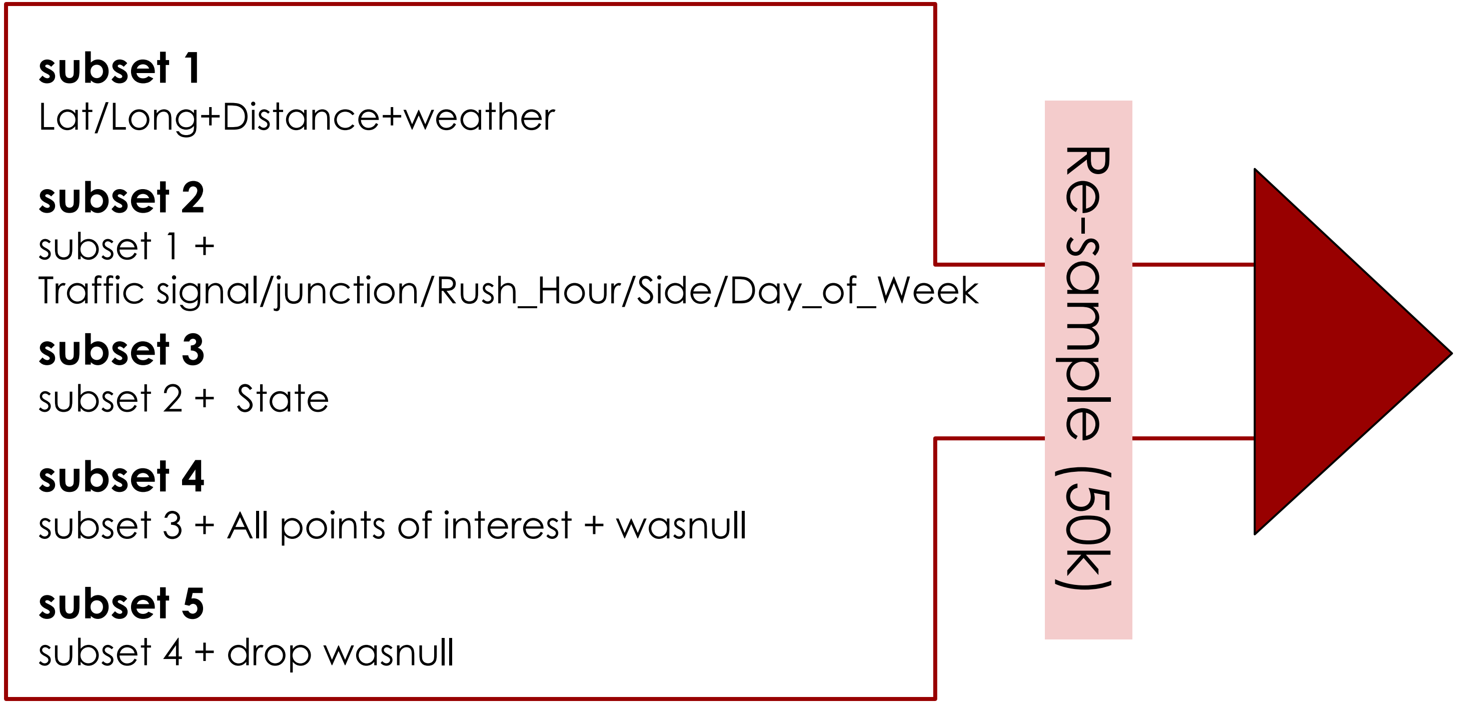
Total number of Accidents by State

Top 20 Cities with Most Severe Accidents

Most Severe Accidents(level 4) by State

Data Manipulation - 3

Modeling prep



subset 1
Lat/Long+Distance+weather

subset 2
subset 1 +
Traffic signal/junction/Rush_Hour/Side/Day_of_Week

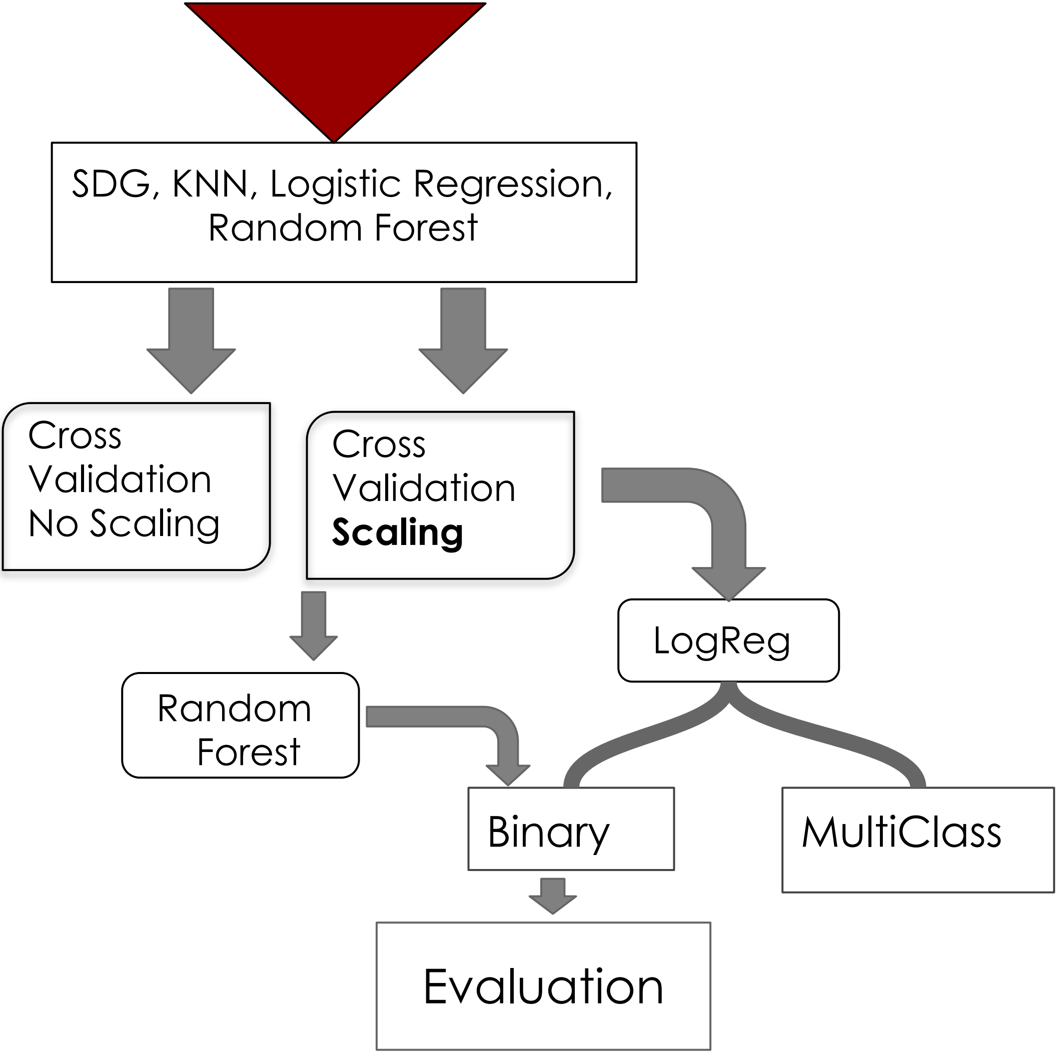
subset 3
subset 2 + State

subset 4
subset 3 + All points of interest + wasnull

subset 5
subset 4 + drop wasnull

Re-sample (50k)

Modeling



SDG, KNN, Logistic Regression, Random Forest

Cross Validation No Scaling

Cross Validation **Scaling**

Random Forest

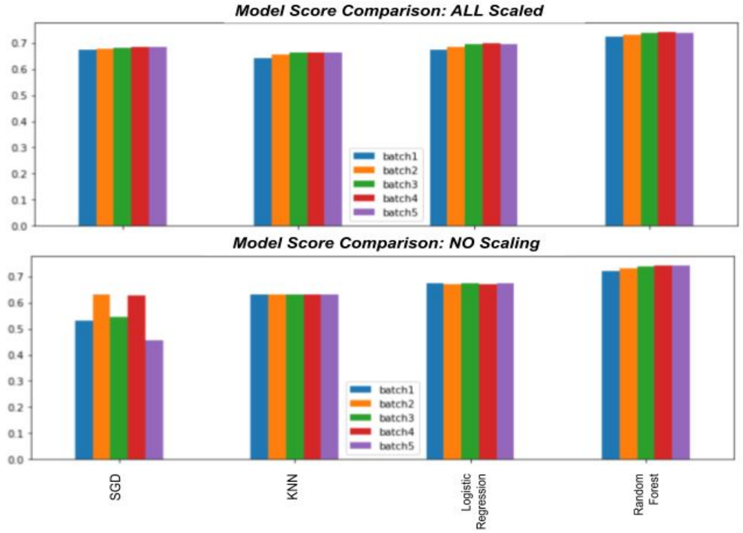
LogReg

Binary

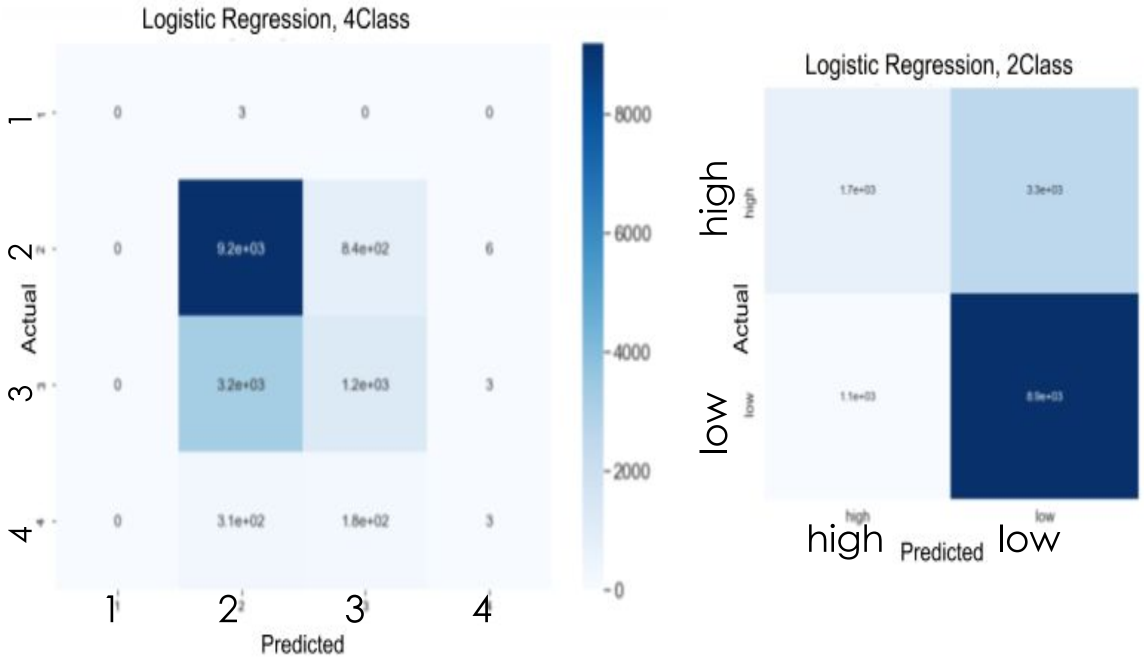
MultiClass

Evaluation

Modeling



Confusion Matrices: MultiClass & Binary



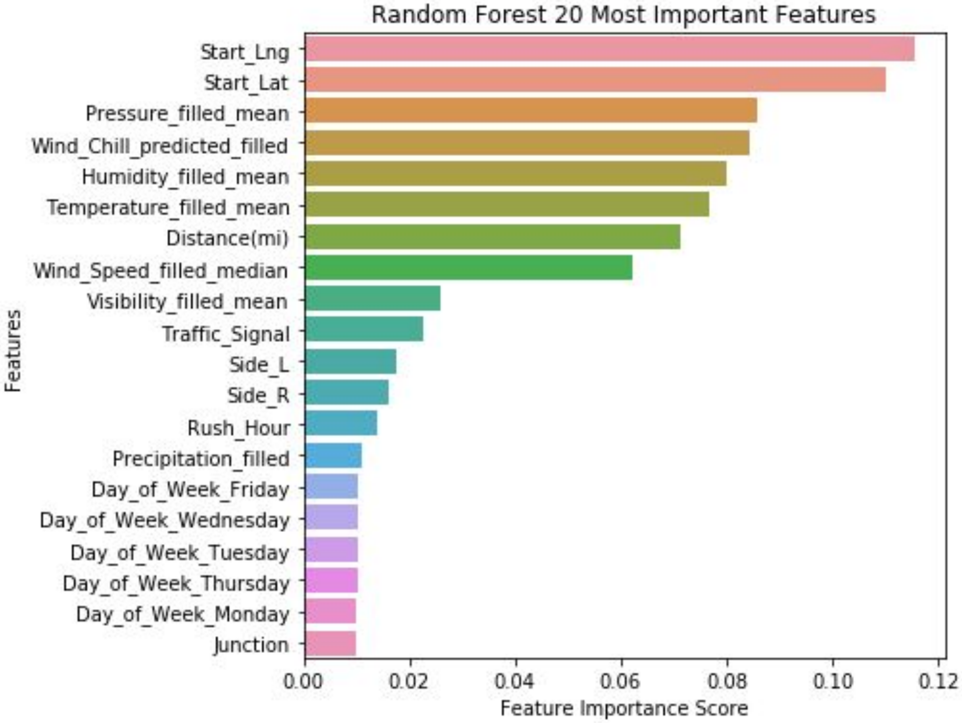
MultiClass and Binary Classification Reports

Logistic Regression

	precision	recall	f1-score	support
1	0	0	0	3
2	0.72	0.92	0.81	10024
3	0.55	0.28	0.37	4481
4	0.25	0.01	0.01	492
	precision	**recall**	**f1-score**	**support**
high	0.61	0.34	0.43	4973
low	0.73	0.89	0.8	10027

Modeling

Random Forest



- **78% accuracy.**

- **f1 scores:**
  - **'low' = 85%**
  - **'high'=78%**

Conclusions/ Recommendations

- Of the four machine learning algorithms used Random Forest gave the best results (78% accuracy).

- Room for improvements through:

more feature engineering and advanced machine learning.

The End

Q&A