Predicting Car Accidents in California, USA

K Thilanjan M Vithana

10/04/2020

Business Problem

How accurately can we model the severity of accidents in California State based on past accident data covering the US?

- There is a growing interest in the correlation studies on the impact of weather on roadside accidents.
- In a state like California, USA most of the population commute with their own automotive and far less use of public transport.
- YoY there is an increase of traffic accidents.
- It will be a good study to identify how much relatable a traffic accident in CA to rest of the US weather wise.

Data Acquisition and Cleaning

Data is obtained from Kaggle repository(https://www.kaggle.com/sobhanmoosavi/us-accidents) .

- This is a countrywide car accident dataset, which covers 49 states of the USA.
- The accident data are collected from February 2016 to June 2020.
- Two APIs that provide streaming traffic incident (or event) data.
- These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.
- 3.5 million accident records in this dataset.

Data Used for Analysis

- A Random sample of 700,000 records out of the 3.5 Mn are selected to build the models.
- The main data set is filtered by State='California' and the first 150,000 rows used to predict and evaluate model accuracy for California.

Data Acquisition and Cleaning (Contd)

Independent Variables

- Measurement data such as Temperature (F), Precipitation(in) had many missing values.
- Group the data by Year, Week (Week number), State, County and obtaining the mean at the most granular level to backfill missing data.
- Rest of the missing data was populated with 0 on a as-needed basis.
- Categorical data such as “Weather Condition” had multiple scenarios (122) that were re-grouped to four distinguishable groups (Clear, Overcast, Severe_rain, Severe_Snow)
- Unclassifiable weather conditions were grouped as ‘Unknown’.

Main Feature set

- Temperature(F) / Humidity(%) / Pressure(in) / Visibility(mi) / Wind_Speed(mph) /Precipitation(in) / Weather_Condition / Sunrise_Sunset

Main Data set fields

Field	Description
ID	This is a unique identifier of the accident record
Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least imp traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffi
Description	Shows natural language description of the accident.
Number	Shows the street number in address record.
Street	Shows the street name in address record.
City	Shows the city in address record.
County	Shows the county in address record.
State	Shows the state in address record.
Zipcode	Shows the zip code in address record.
Country	Shows the country in address record.
Temperature(F)	Shows the temperature (in Fahrenheit).
Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
Humidity(%)	Shows the humidity (in percentage).
Pressure(in)	Shows the air pressure (in inches).
Visibility(mi)	Shows visibility (in miles).
Wind_Direction	Shows wind direction.
Wind_Speed(mph)	Shows wind speed (in miles per hour).
Precipitation(in)	Shows precipitation amount in inches, if there is any.
Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.

Data acquisition and cleaning...Contd

Weather Condition breakdown

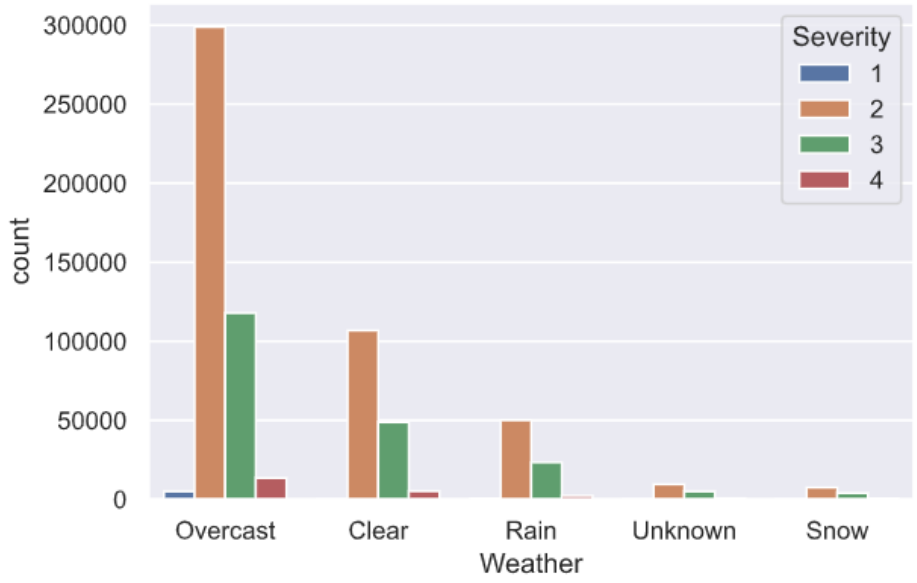
New Classified 'Weather condition'	' Weather Condition ' String contains
Clear	'Clear'
Overcast	'Overcast'
Severe_Snow	'Ice', 'Snow', 'Wint', 'Freeze', 'Hail', 'Sleet'
Severe_Rain	'Rain', 'Thunder', 'Storm', 'Heavy', 'Fog', 'Mist', 'Wind', 'Haze', 'Shower', 'Drizzle', 'T-'
Unkonwn	Rest of the data that cannot be classified

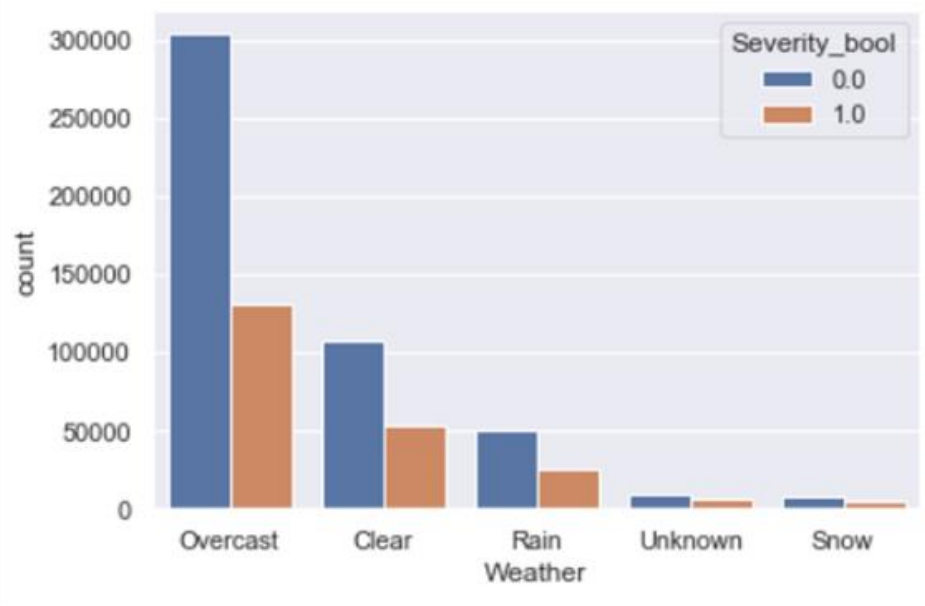
Dependent variables

- The main predicted variable is Severity. It holds the values 1 thru 4 based on ascending order of severity of the accident.
- For this exercise severity is re-classified to binary format where Severity 1, 2 are mapped to 0 and 3, 4 are mapped to 1 in the data pre-processing stage.

Data Visualization

Histograms for Weather Condition vs Severity

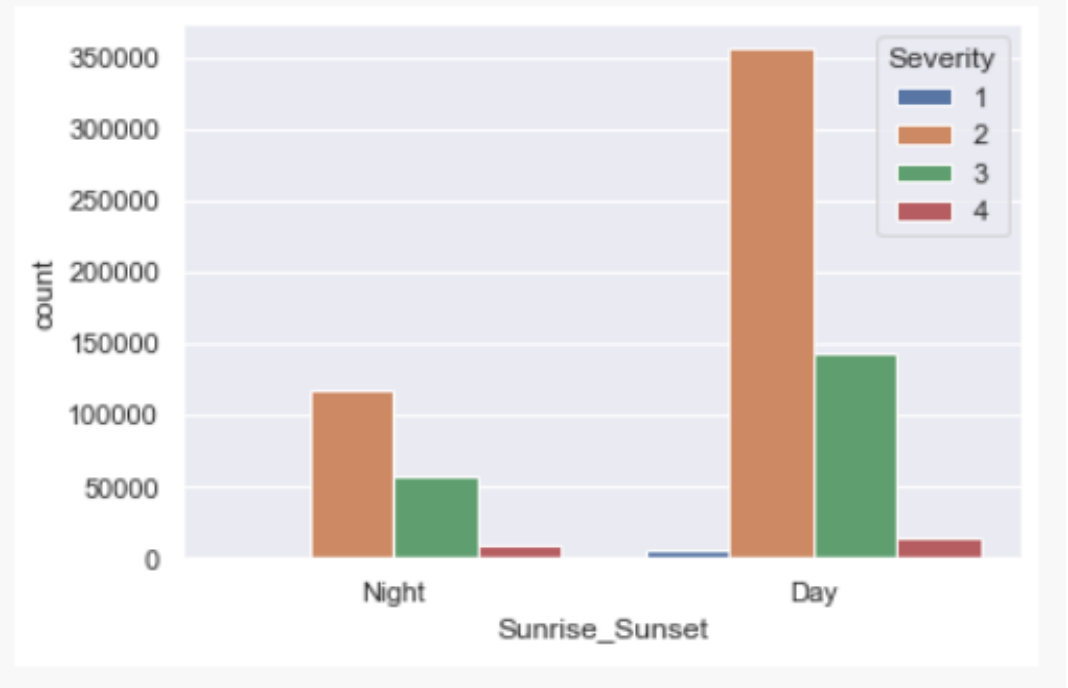




- Most of the data points lie on Severity 1 and 2 (Severity 0 in binary format).
- Please refer Appendix for continuous variable data visualization plots

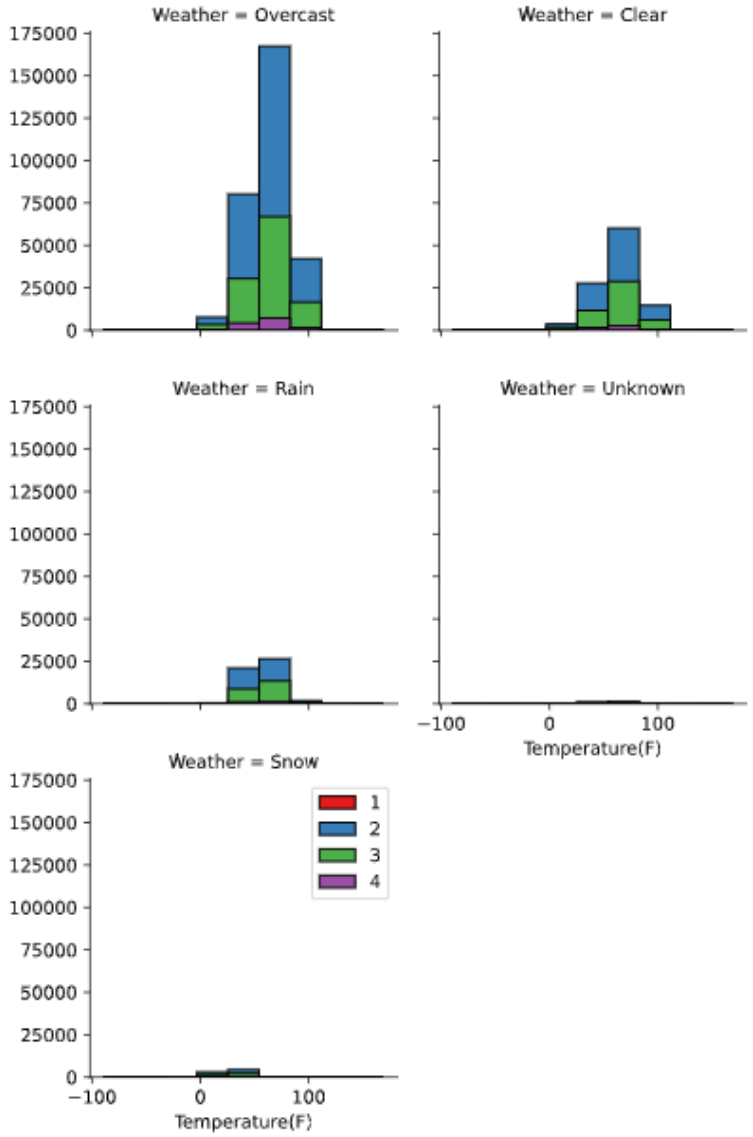
Data Visualization (Contd)

Histograms for Sunrise_Sunset vs Severity



- Illustrates most of the accidents are severity 1 and 2 (Severity 0 in binary format) and happens during the day

Continuous Variable example



Temperature Histogram

This plot suggest the severity and quantity of the accidents increase with temperature when it is overcast

As observed in Histograms illustrated on previous slides, most of the data lie in Overcast Weather condition

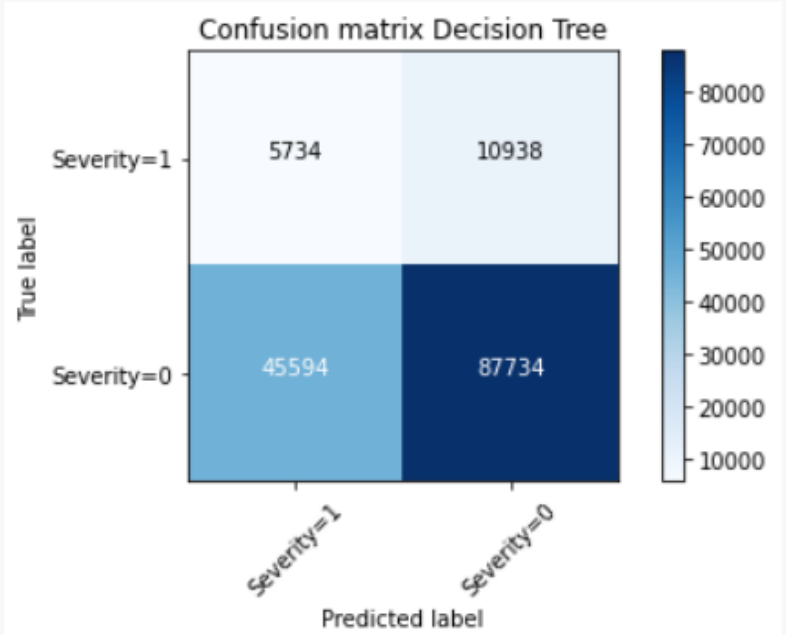
Predictive Modelling and Results

This is a binary classification problem, and the following models were selected to create multiple models and evaluate each of their performance/accuracy of predicting accidents for CA. These models were selected as there is a large data set to process .

- Decision Tree
- Logistic Regression
- K Nearest Neighbor(KNN)
- Support Vector Machine

Decision Tree Model and Performance

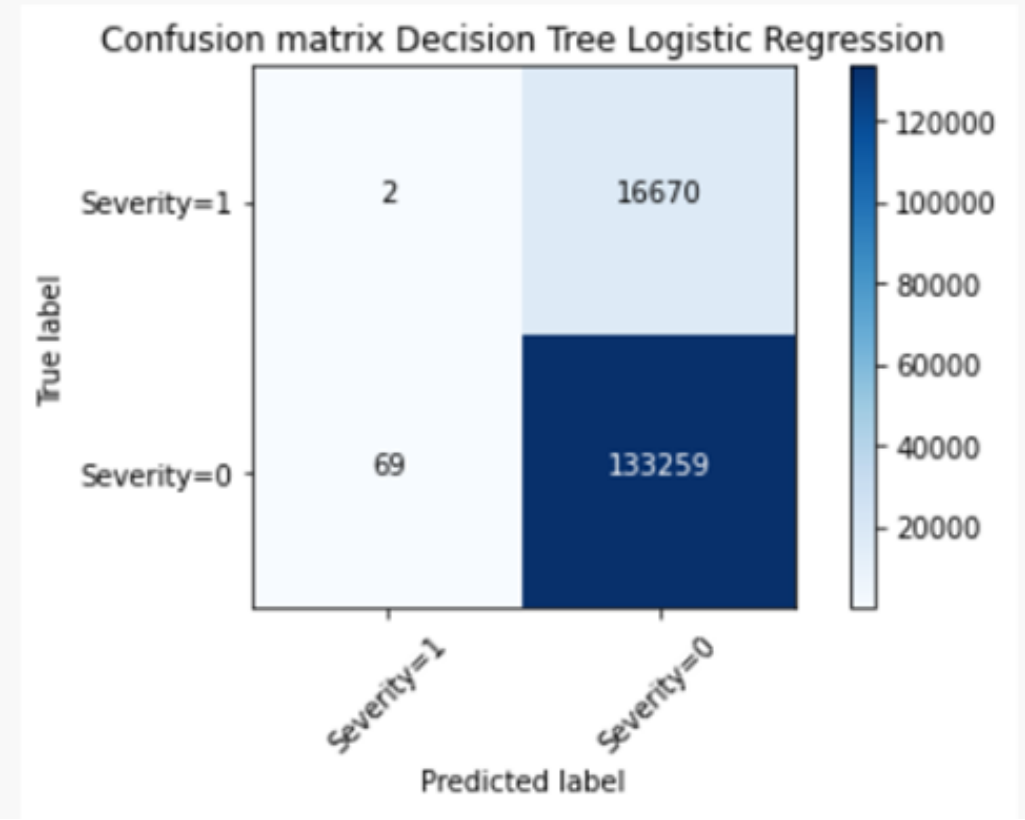
```
Confusion matrix, without normalization
[[ 5734 10938]
 [45594 87734]]
```



Jaccard index: 0.09

F1-score: 0.69

Logistical Regression Model and Performance

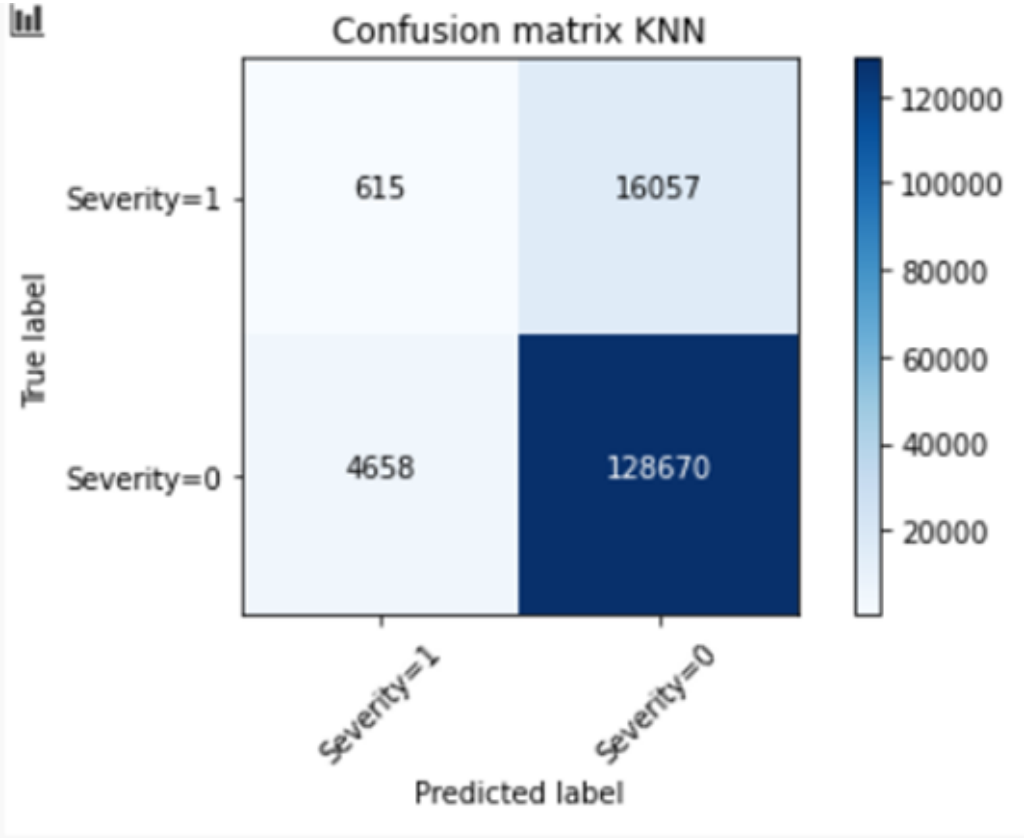


Jaccard index: 0.00

F1-score: 0.84

LogLoss 0.47

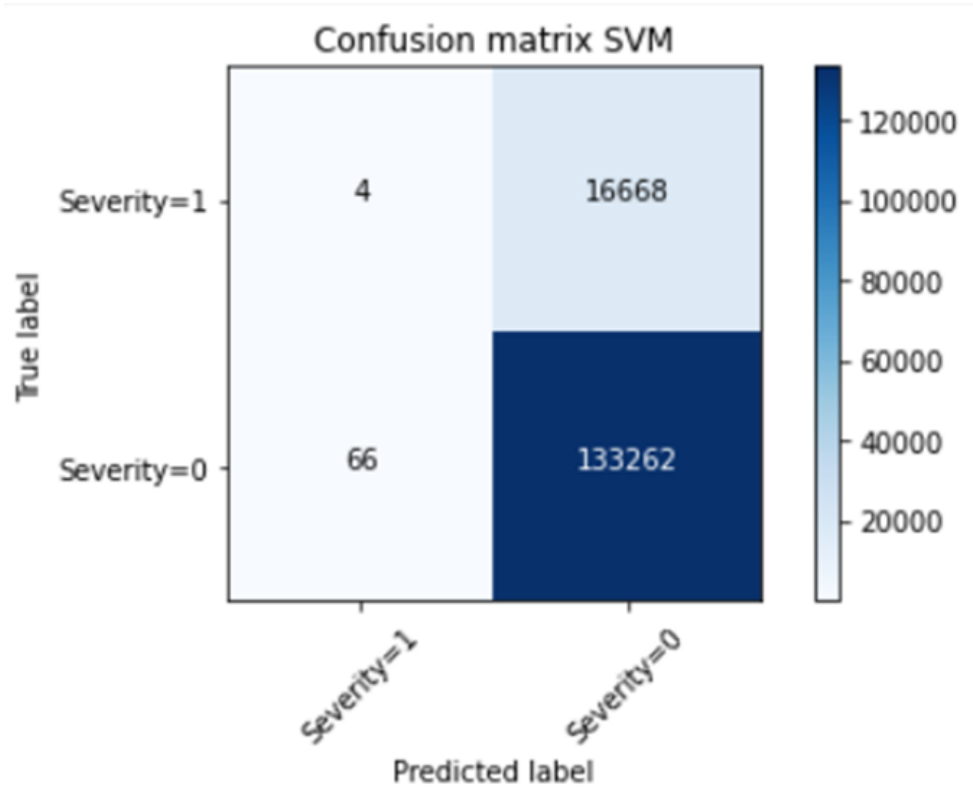
k Nearest Neighbor Model and Performance (k=15)



Jaccard index: 0.03

F1-score: 0.83

Support Vector Machine Model and Performance



Jaccard index: 0.00

F1-score: 0.84

Observations Summary

	Decision Tree	Logistic Regression	K Nearest Neighbor	Support Vector Machine
Jaccard index	0.09	0.00	0.03	0.00
F1-score	0.69	0.84	0.83	0.84
LogLoss		0.47		

- **Decision Tree Model** -Jaccard index: 0.09 / F1-score: 0.69 -Slightly better model to estimate a severe accident but a better model to estimate less severe accident due to weather condition

- **Logistic Regression** -Jaccard index: 0.00 LR F1-score: 0.84 LR LogLoss: 0.47 - Low accuracy predicting a severe accident (3, 4) but higher accuracy for predicting a less severe accident (1, 2)

- **K Nearest Neighbor** - Jaccard index: 0.03 KNN F1-score: 0.83 - Low accuracy predicting a severe accident (3, 4) but higher accuracy for predating a less severe accident (1, 2)

- **Support Vector Machine** - Jaccard index: 0.00 SVM F1-score: 0.84 - Low accuracy predicting a severe accident (3, 4) but higher accuracy for predicting a less severe accident (1, 2)

Conclusion

- Accuracy levels of predicting a severity or 3 and 4 level accident is at its lowest levels.
- Accuracy of predicting a severity 1 or 2 is very high.

- In conclusion, this project can be summarized as follows to answer the business problem.
  - ❖ yes, it is possible to model and predict accident severity of California based on US National data up to a severity of low to medium.
  - ❖ Best model to use to predict less severe accidents would be
  - ✓ Logistic Regression
  - ✓ kNN Model
  - ✓ Support Vector Machine

Future Directions

Following suggestions can be made based on the data and the predictions

- Use localized samples for train model. E.g.: CA only train/test data
- Increase the number of samples out of US National data
- Increase the number of samples to predict from CA
- Revisit the weather condition grouping for better grouping.
- Run separate model for measurable values such as Temperature / Pressure etc.
- Run separate model for categorical values of weather condition vs Severity

References

- https://medium.com/@prashantchaturvedi2020/accuracy-of-classifier-db5d9cff0a21
- https://stackoverflow.com/questions/46391128/pandas-fillna-using-groupby
- https://clay-atlas.com/us/blog/2019/10/27/python-english-tutorial-solved-unicodeescape-error-escape-syntaxerror/