Predicting Car Accidents in California, USA

K Thilanjan M Vithana

October 4, 2020

1. Introduction

1.1 Background

The US government collects and publishes detailed information about traffic accidents across the country. This information includes, but is not limited to, geographical locations, weather conditions, type of vehicles, accident severity, number of casualties, etc. What this project intends to accomplish is to investigate if there is a relationship and can we accurately model the published accident conditions to the severity of the accidents in California. The findings of the project will assist users to understand the risks of operating a vehicle in a knowing the expected conditions in California

1.2 Problem

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

1.3 Interest

There is a growing interest in the correlation studies on the impact of weather on roadside accidents. Particularly in a state like California where most of the population commute with their own automotive and far less use of public transport, the YoY there is an increase of traffic accidents. Though there is may be climate wise significant difference in CA vs rest of the States in the US it will be a good study to identify how much relatable a traffic accident in CA to rest of the US weather wise.

2. Data acquisition and cleaning

2.1 Data sources

Data is obtained from Kaggle (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, using 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. Currently, there are about 3.5 million accident records in this dataset*." –

US Accidents (3.5 million records), Kaggle

2.1.1 Data Used for Analysis

The main data set consist of 3.5 Million records. For this project a random sample of 700,000 records are selected to build the models. The main data set is filtered by State='California' and the first 150,000 rows are selected as the data set for California that's used to predict using the built models and evaluate accuracy.

2.2 Data Cleaning

Measurement data such as Temperature (F), Precipitation(in) had many missing values. By grouping the available data by Year, Week (Week number), State, County and obtaining the mean for the measure was the only logical step to get the best possible average at the most granular level seasonal and geography wise. Rest of the missing data was populated with 0 on a as-needed basis.

Categorical data such as Weather Condition had multiple combinations (122) that were re-grouped to four distinguishable groups based on common denominator key words and by inspection and common knowledge about the overall condition of the weather. Unclassifiable weather conditions were grouped as 'Unknown'.

2.3 Feature selection

All features that represent weather and relatable to driver's perception of driving was selected for our analysis.

2.3.1 Main data set and description

Table 1: Main data set

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 impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic
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.

2.3.2 Main Feature set

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

2.3.2.1 Weather Condition reclassified sub-types

Table 2: Weather condition sub-types

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

2.3.3 Main Predicted Variable

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.

3. Exploratory Data Analysis

3.1 Overall distribution of weather conditions and severity

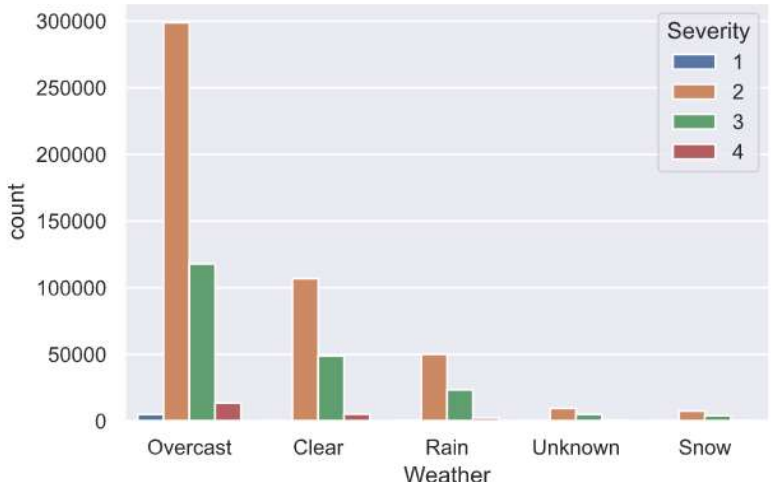

Figure 1: Overall distribution of weather conditions and severity (4 value)

Figure 1 illustrates the total data set and the reclassified weather conditions are distributed according to the 4 values of severity of the accident.

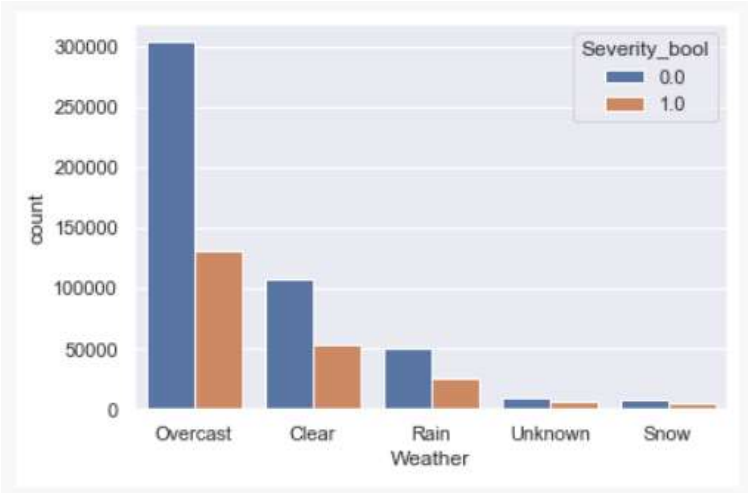

Figure 2: Overall distribution of weather conditions and severity (Boolean)

Figure 2 illustrates the total data set and the reclassified weather conditions are distributed according to the Boolean values of severity of the accident after mapping to 0 and 1 .

3.2 Relationship between Temperature vs Accident severity in newly classified weather conditions

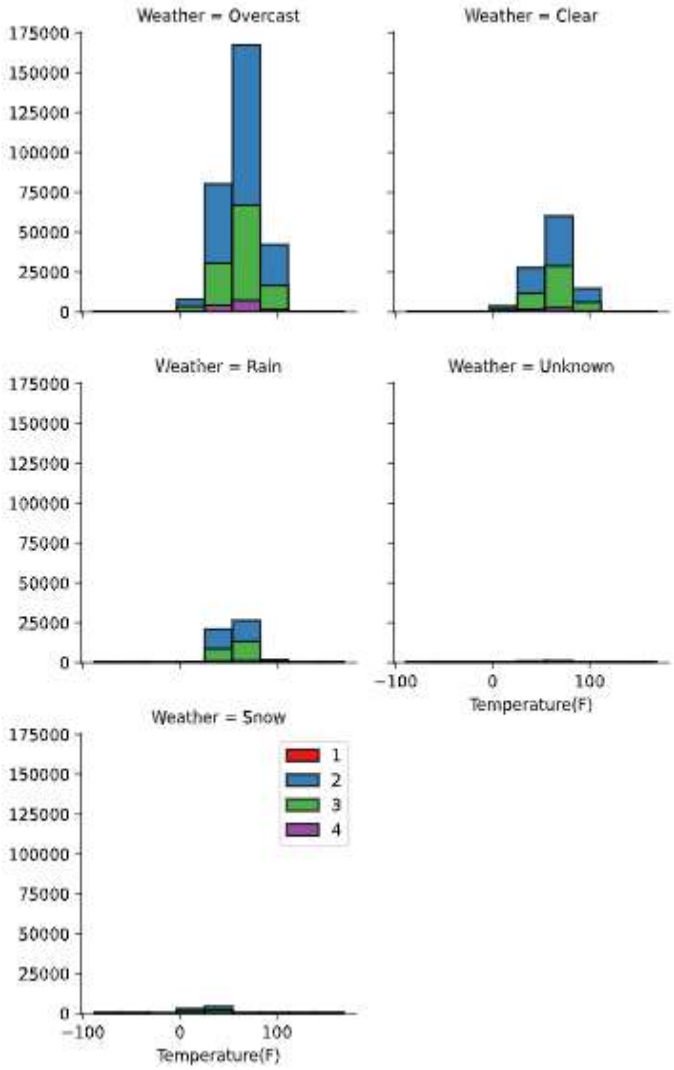

Figure 3: Overall distribution of Temperature per weather conditions and severity

3.2 Relationship between Temperature vs Accident severity in Day or Night, Time of day

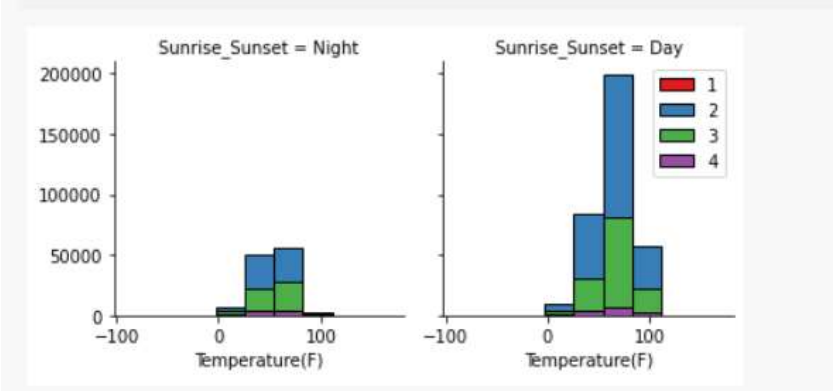

Figure 4

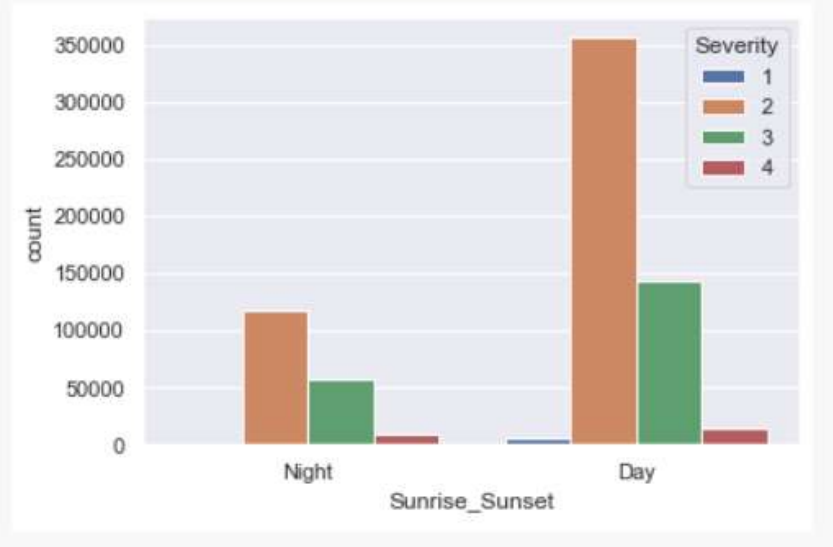

Figure 5

4. Predictive Modeling 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.

- Decision Tree
- Logistic Regression (Solver = 'saga' on larger datasets.
  Reference:https://towardsdatascience.com/dont-sweat-the-solver-stuff-aea7cddc3451)
- K Nearest Neighbor(KNN)
- Support Vector Machine

4.1 Decision Tree Model Results

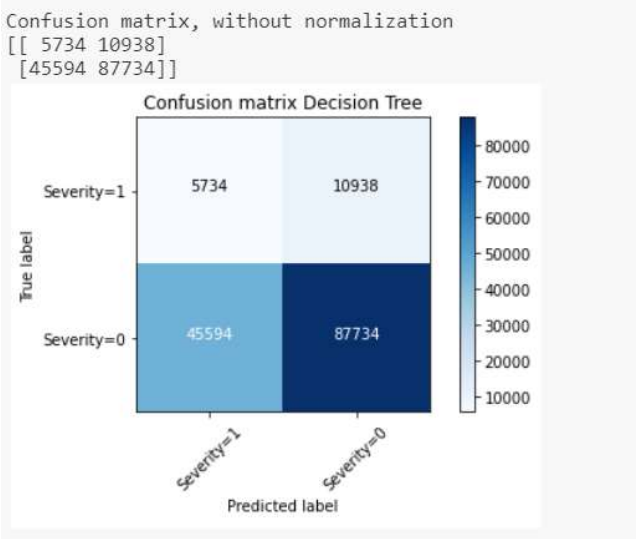

Figure 6: Confusion Matrix for Decision Tree Model

4.1.1 Decision Tree model Evaluation

Jaccard index: 0.09
F1-score: 0.69

As the plot in figure xxx suggest, the model was successful in predicting Severity 0 at a very high accuracy but not Severity 1. This is justifiable as the data visualization section of this report suggest that most of the data lies in the Severity 0 (Severity 1 and 2 in main US traffic data file sample) and a low number of data points lie in the Severity 1 (Severity 3 and 4 in main US traffic data file sample). The models are trained accordingly. The Jaccard Index being extremely low suggests that the predictor variables and the source variables has a very low correlation. F1-score is not too high as much as we like it to be due to the inability to predict a True positive.

4.2 Logistic Regression Model Result

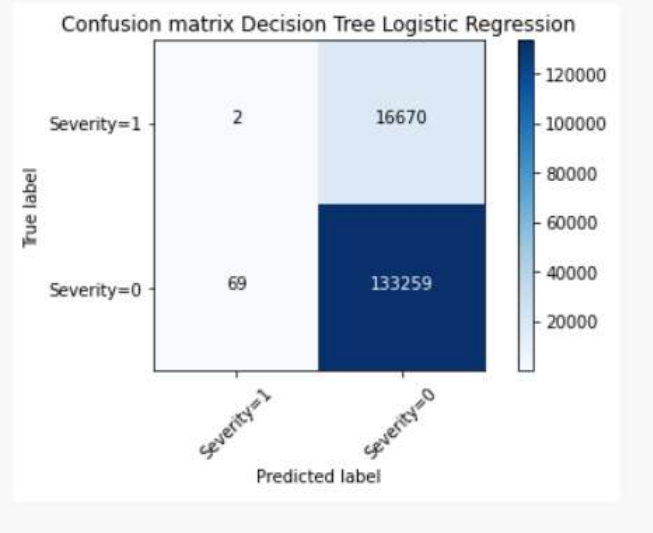

Figure 7: Confusion Matrix for Logistic Regression Model

4.2.1 Logistic Regression model Evaluation

Jaccard index: 0.00
F1-score: 0.84
LogLoss 0.47

The results of the model is similar to the Decision Tree Model in section 4.1.2 but this time the ability to predict a True positive has extremely gone down. The higher F1-score does suggest the ability to produce a True negative using this model. Logloss seems high suggest the accuracy of this model is low is we are expecting the model to predict a true positive

4.3 k Nearest Neighbor Model Result

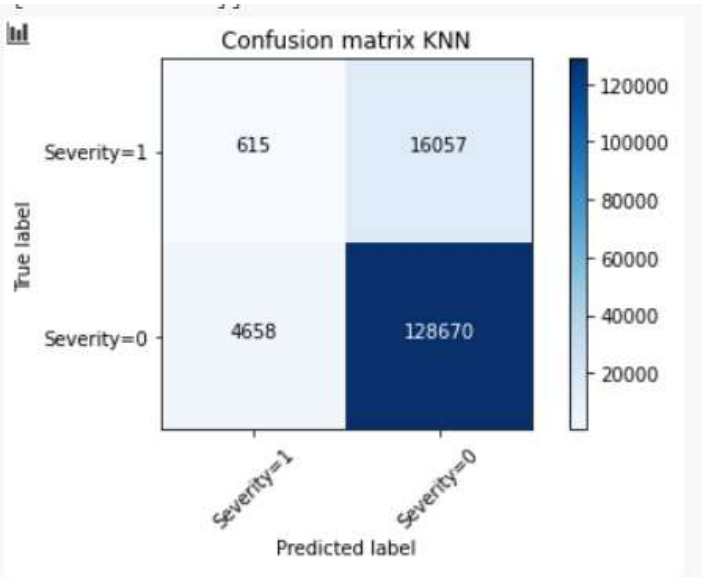

Figure 7: Confusion Matrix for kNN Model

Jaccard index: 0.03
F1-score: 0.83

4.3.1 kNN model Evaluation

This model displays similar characteristics to Logistic Regression model in section 4.2.1.

4.4 Support Vector Machine Model Result

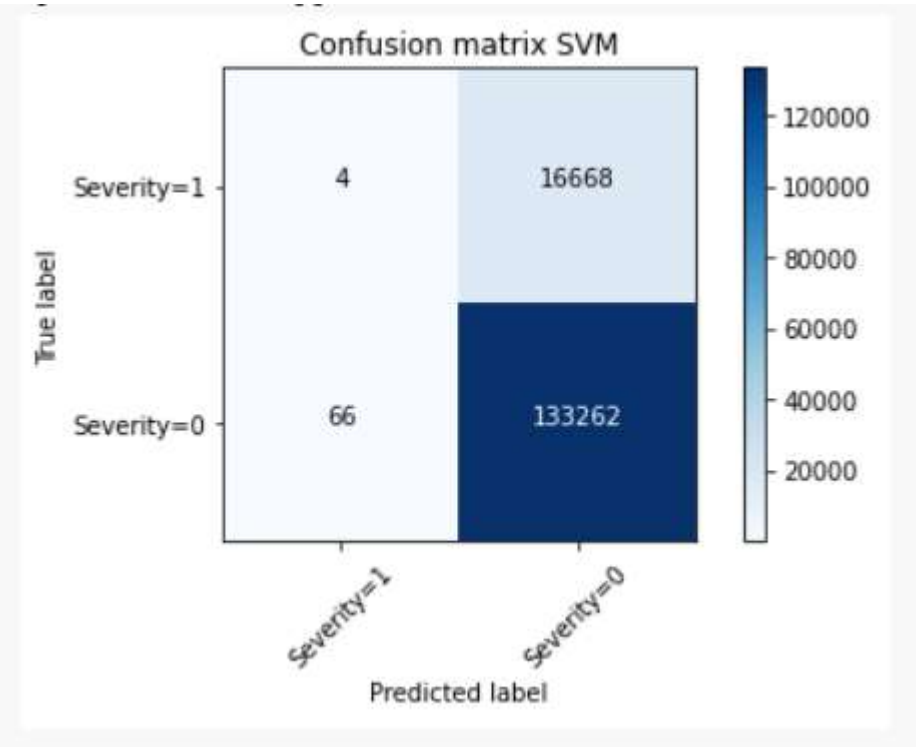

Figure 8: Confusion Matrix for SVM Model

Jaccard index: 0.00
F1-score: 0.84

4.4.1 SVM model Evaluation

This model displays similar characteristics to kNN model in section 4.3.1.

5. Observations Summary

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)

6. Conclusion

Per the four models above, the accuracy levels of predicting a severity or 3 and 4 level accident is at its lowest levels. However, all models suggest that the 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.

7. 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

8. 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/