Accident Severity Prediction

BUSINESS PROBLEM:

- More than 38,000 people die every year in crashes on U.S. roadways. The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants. An additional 4.4 million are injured seriously enough to require medical attention. Road crashes are the leading cause of death in the U.S. for people aged 1-54.

- Road accidents are increasing day to day all over the world due to several reasons. These accidents may happen due to either avoidable circumstances or un-avoidable circumstances.

1. Avoidable scenarios: Speeding, not wearing safety gear such as helmets or seat belts, breaking traffic rules.
2. Un avoidable scenarios: Bad roads, weather conditions, lighting conditions, traffic conditions.

METHODOLOGY

- Data cleaning is the first step by detecting and dealing with null/missing values present in the dataset.
- We will perform EDA (Exploratory data analysis) and feature engineering on most of the data attributes.
- Once the data is ready and in appropriate format, we build a machine learning model using decision tree and logistic regression.
- Finally, we test the accuracy of the model with different metrics.

ACKNOWLEDGEMENT:

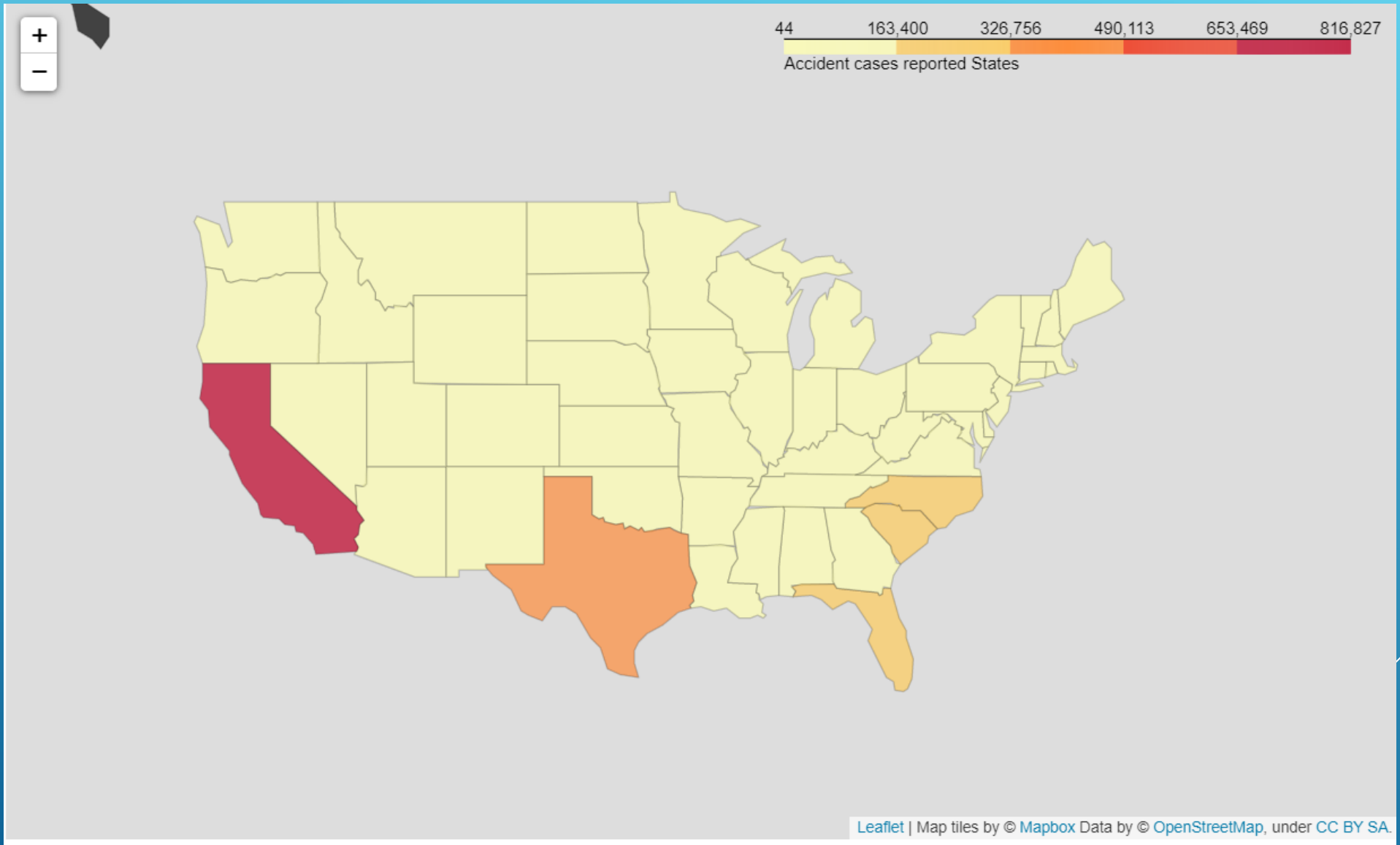
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

DATA DESCRIPTION:

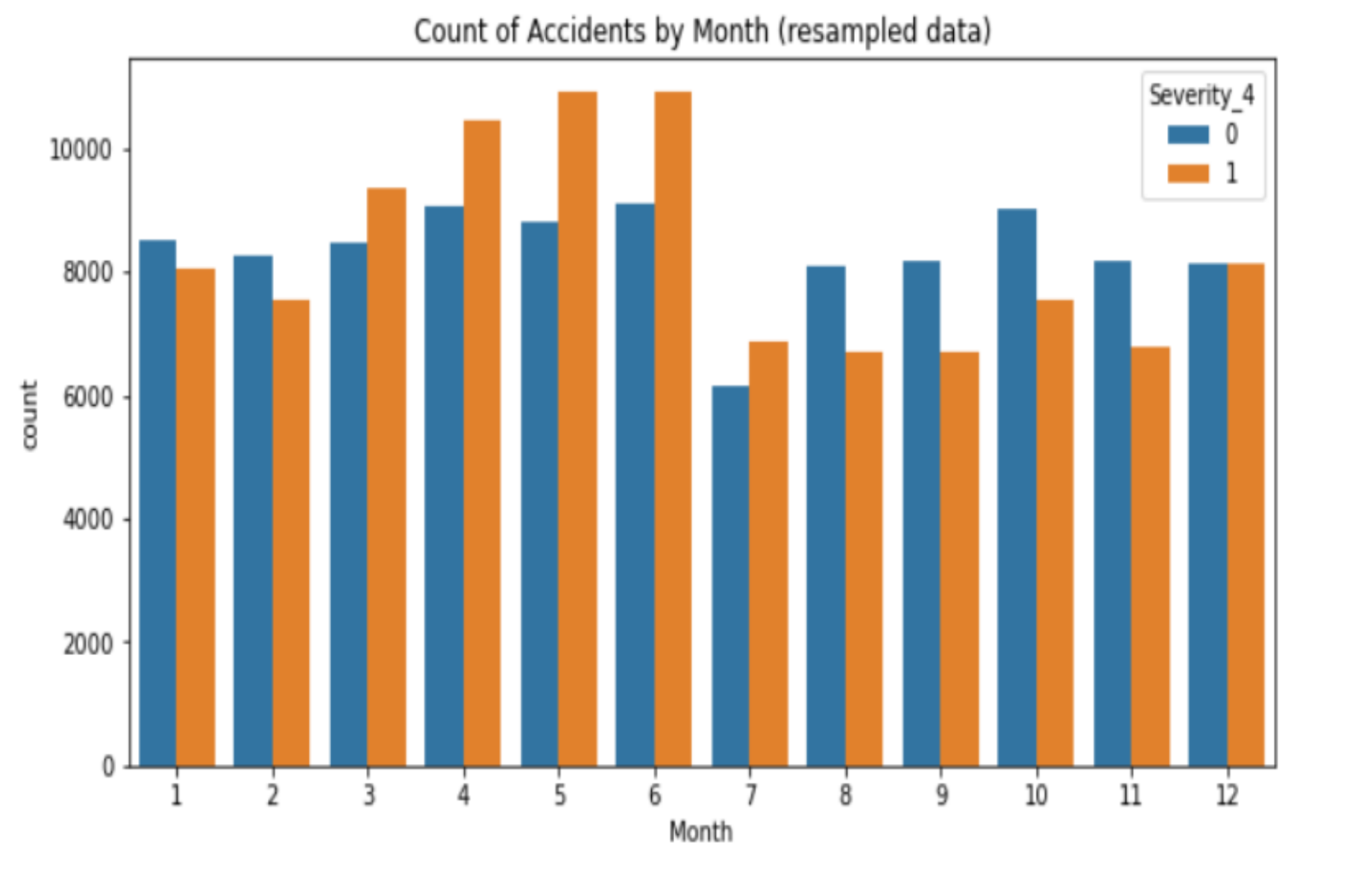
- Initially, dataset contained 49 features.
- Missing values and un important features are dropped from the feature after thorough analysis.
- Suitable transformation was applied on to the dataset to make it simpler and compatible with ML algorithms.
- Dataset was unbalanced initially and therefore it needed to be resampled to 100000.
- Analysis is performed against Less Severe 0 and More Severe accidents 1.

ANALYSIS

ACCIDENT CASES REPORTED ACROSS USA

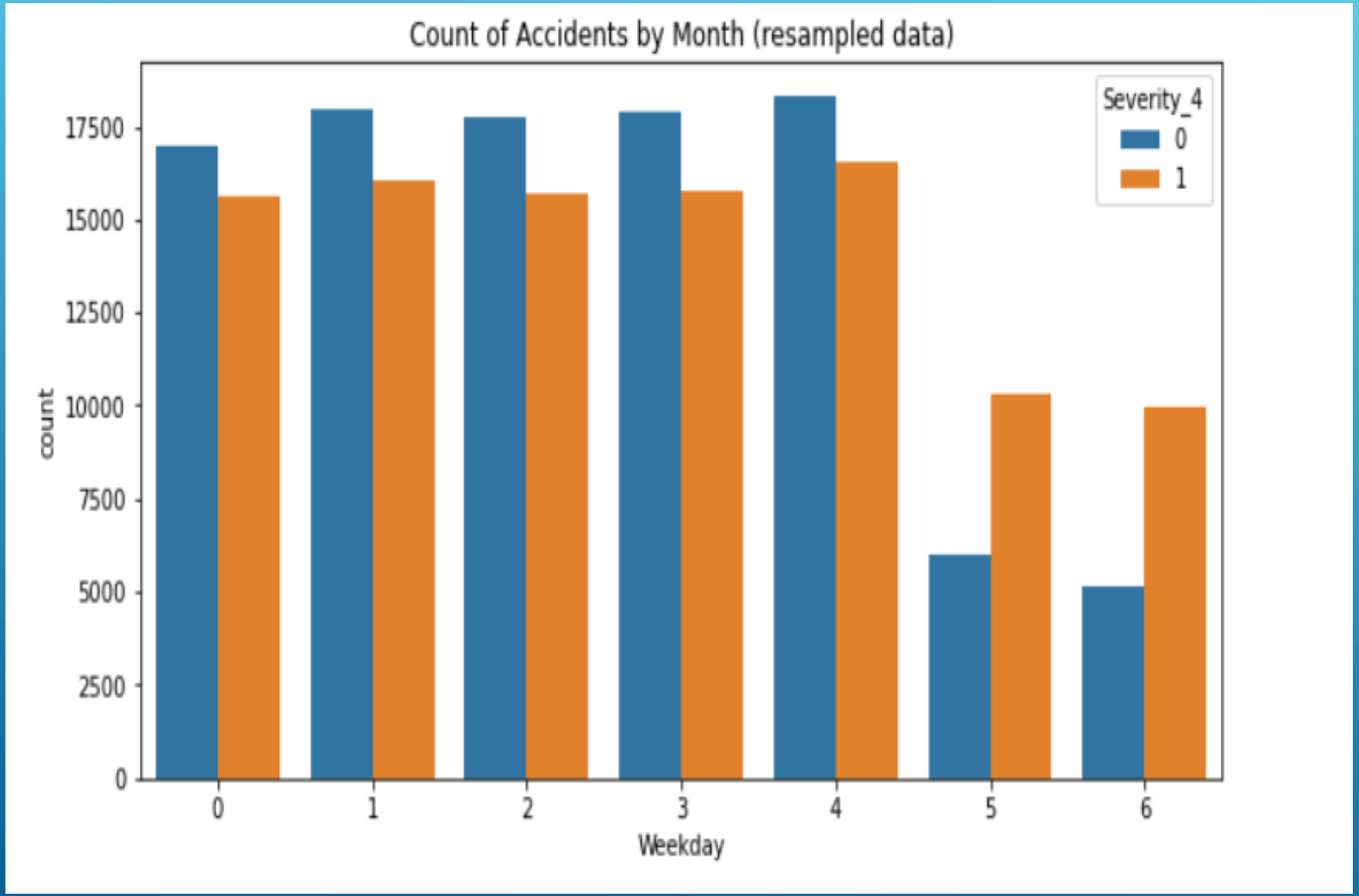


COUNT OF ACCIDENTS BY MONTH



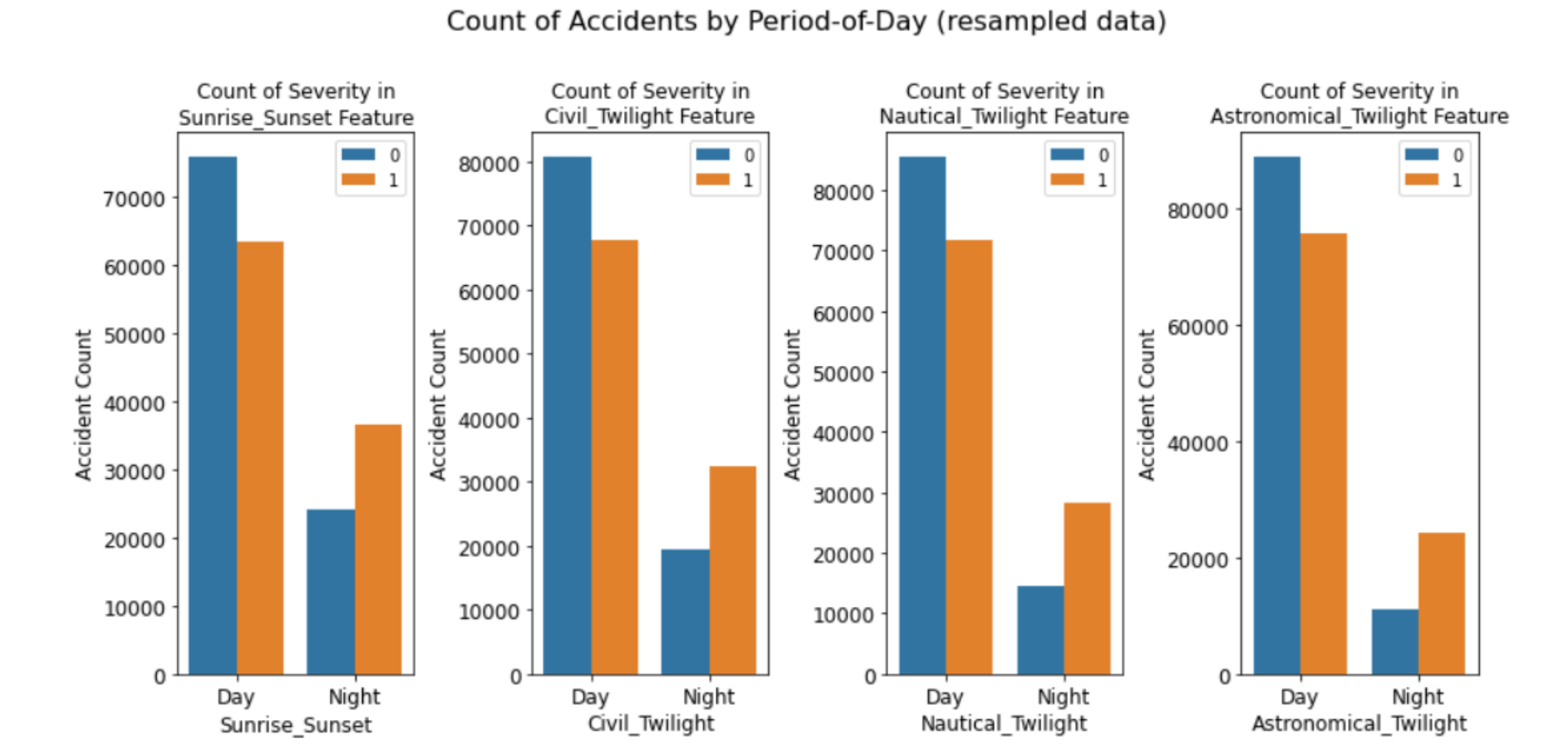
We see that accidents gradually increases from January to June and then rapidly falls down during July and again increases until December.

ACCIDENT COUNT BASED ON MONTH



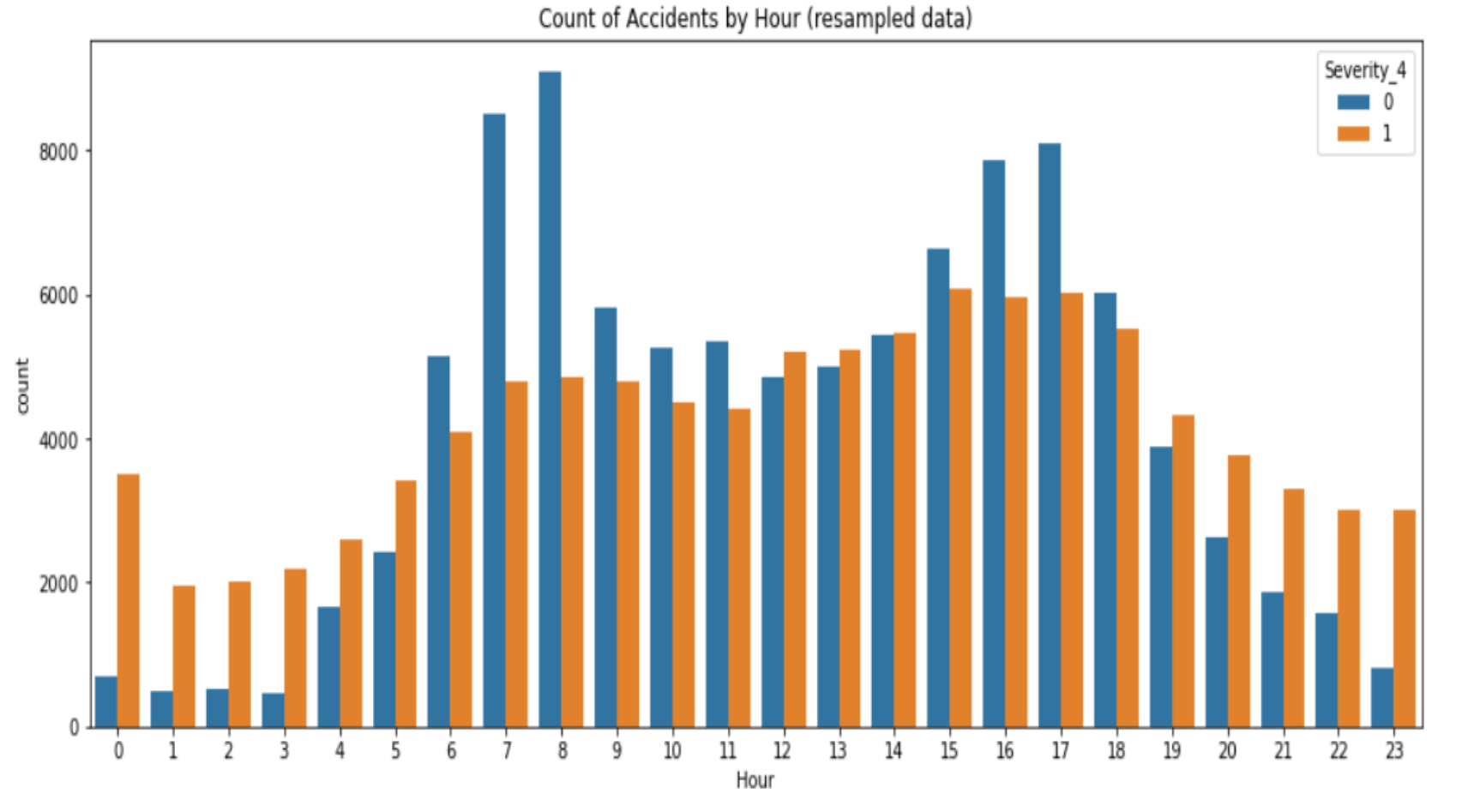
We see that number of accidents are less during weekends although the severity of accidents is high on weekends than weekdays.

ACCIDENT COUNT DURING PERIOD OF DAY



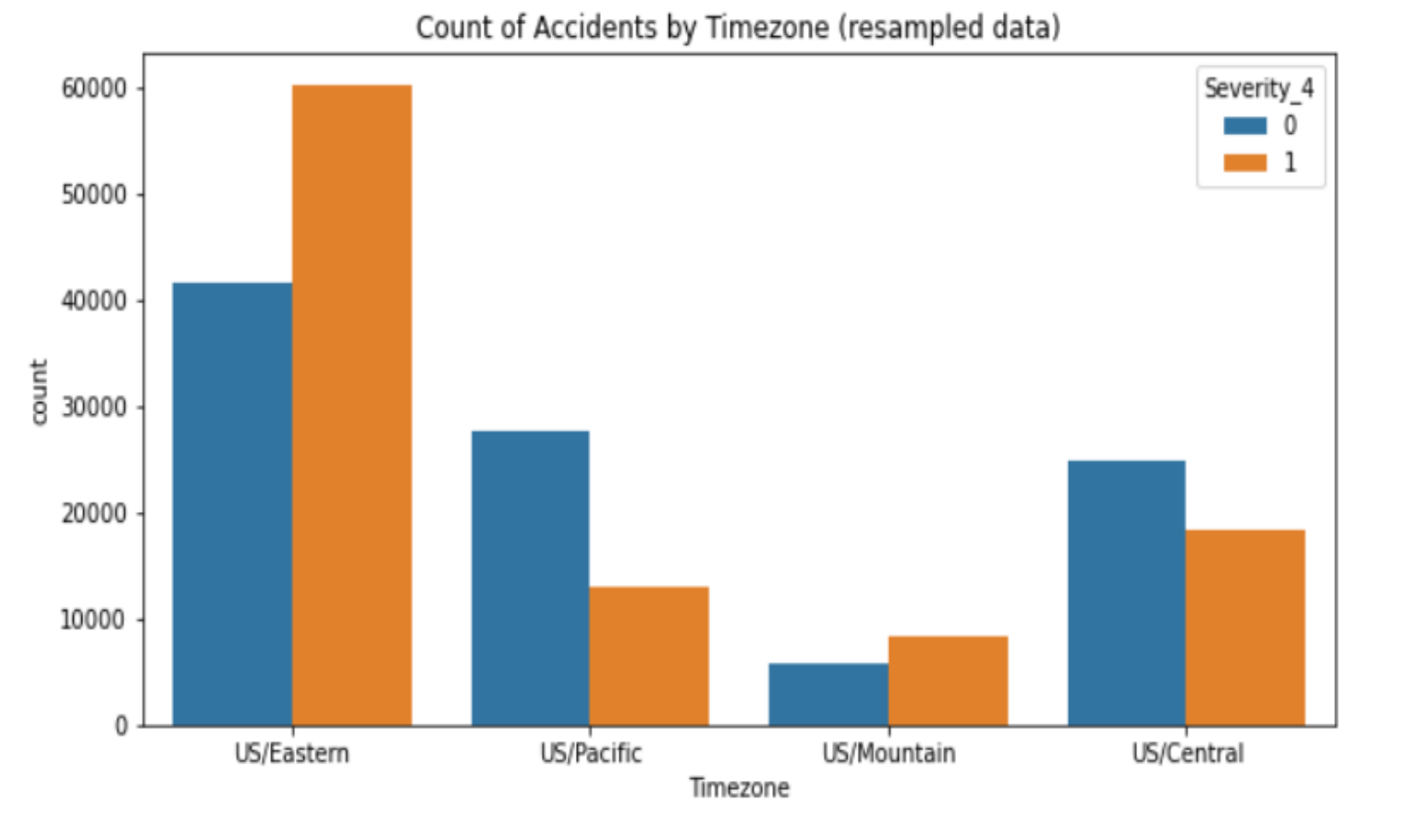
Although, the number of accidents was higher during day time severity of accidents was higher during night.

ACCIDENT COUNT BY HOUR (THROUGH A DAY):



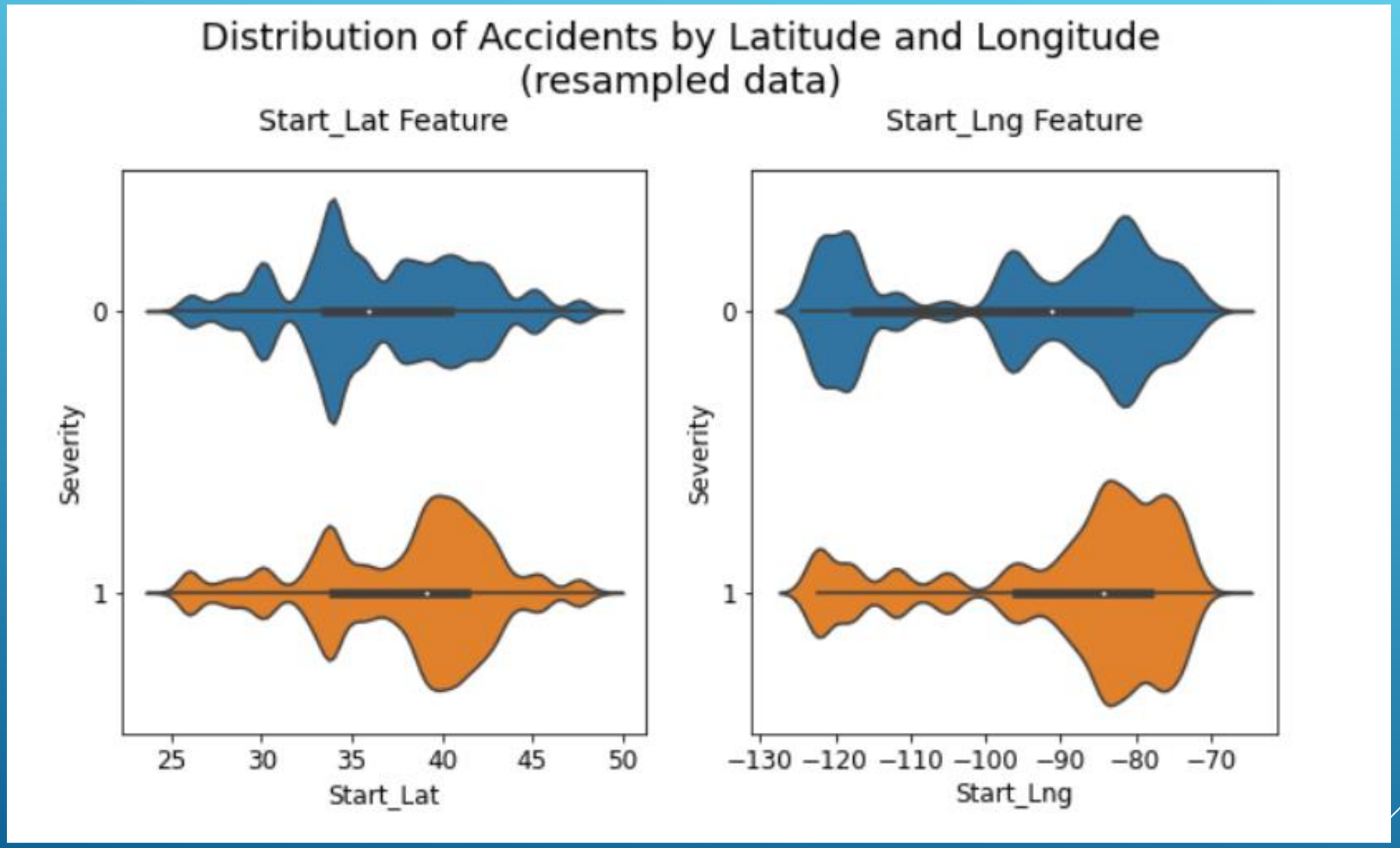
Although accidents are higher during early morning may be due to huge population travelling to work, accidents during night tend to be more severe.

ACCIDENT CASES ACROSS DIFFERENT TIME ZONES

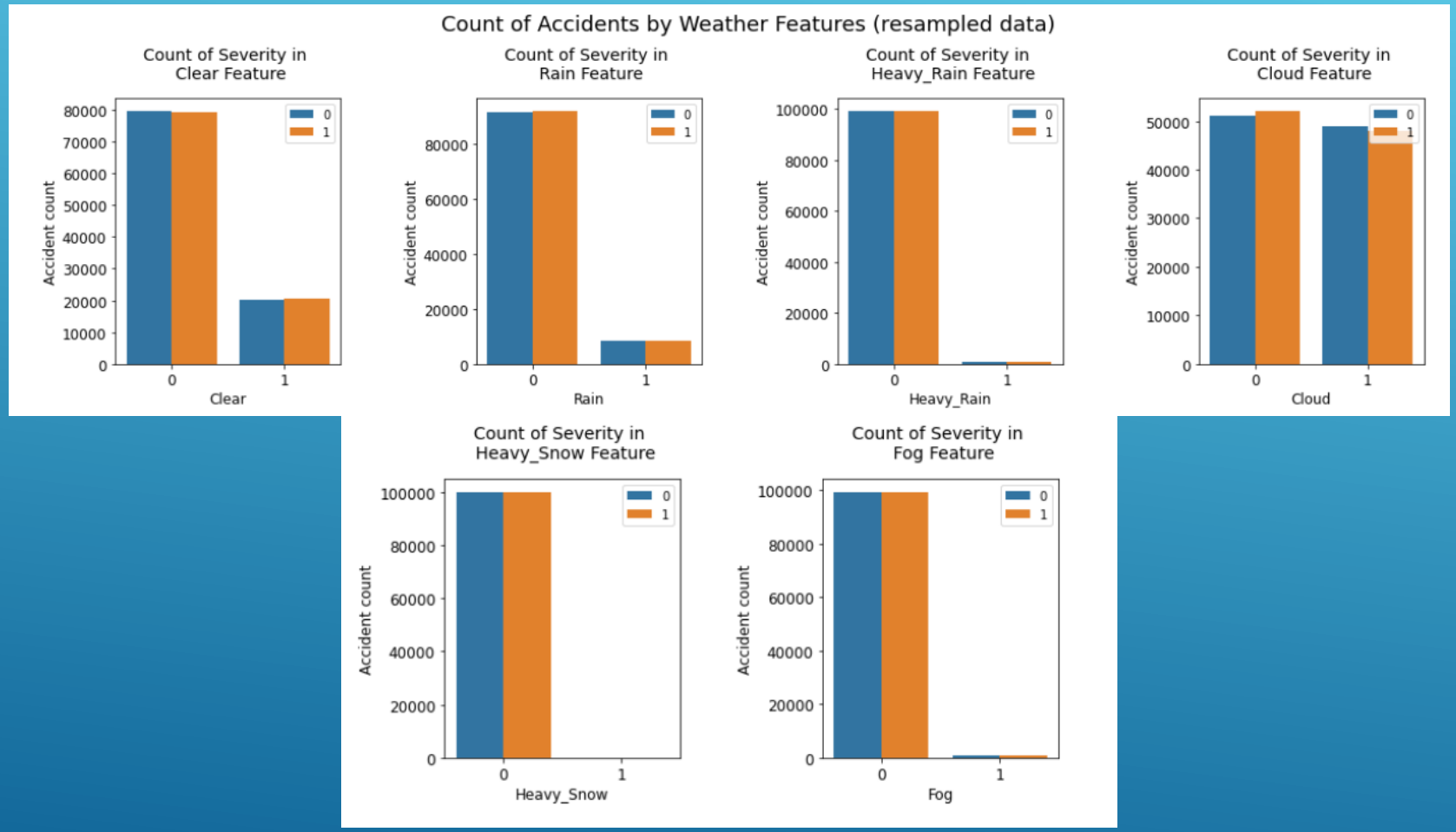


We can see that highest number of accident cases occur
in **US/Eastern** Timezone.

DISTRIBUTION OF ACCIDENTS BY LATITUDE AND LONGITUDE

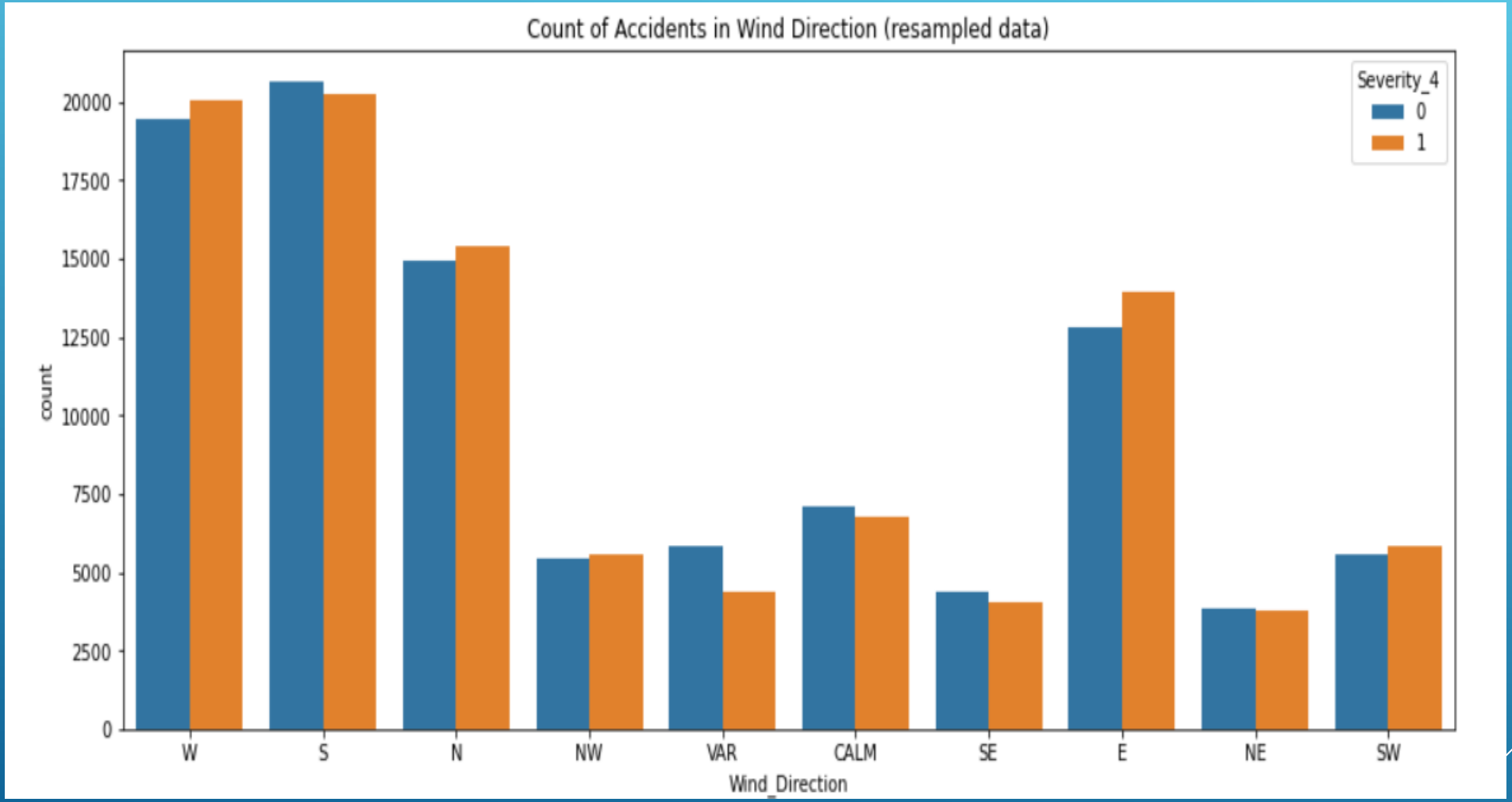


ACCIDENT COUNT DURING DIFFERENT WEATHER CONDITIONS



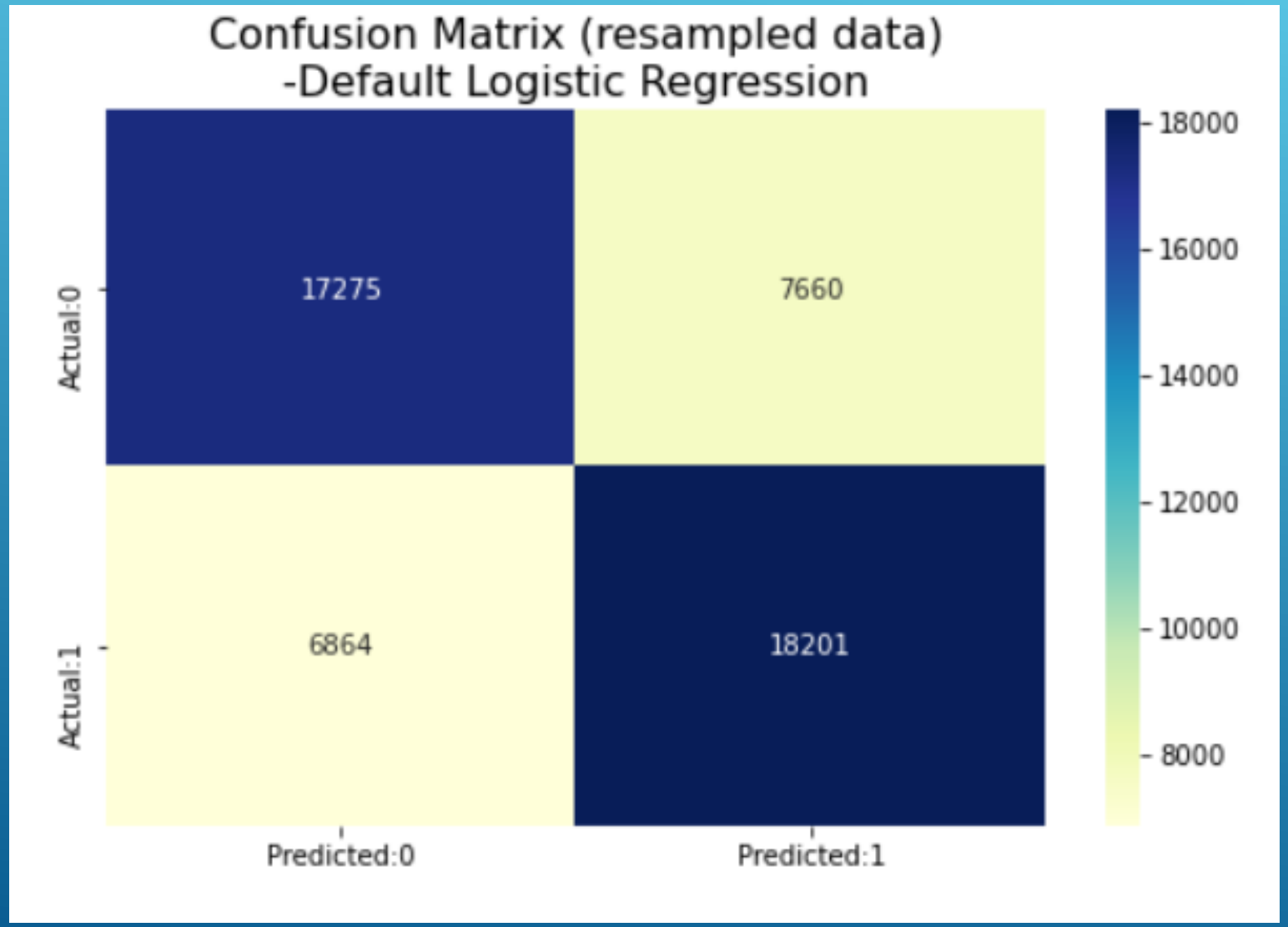
We can see that accidents occur to be more serious during Heavy Rain/Snow rather than a Cloudy weather

ACCIDENT COUNT BASED ON WIND DIRECTION



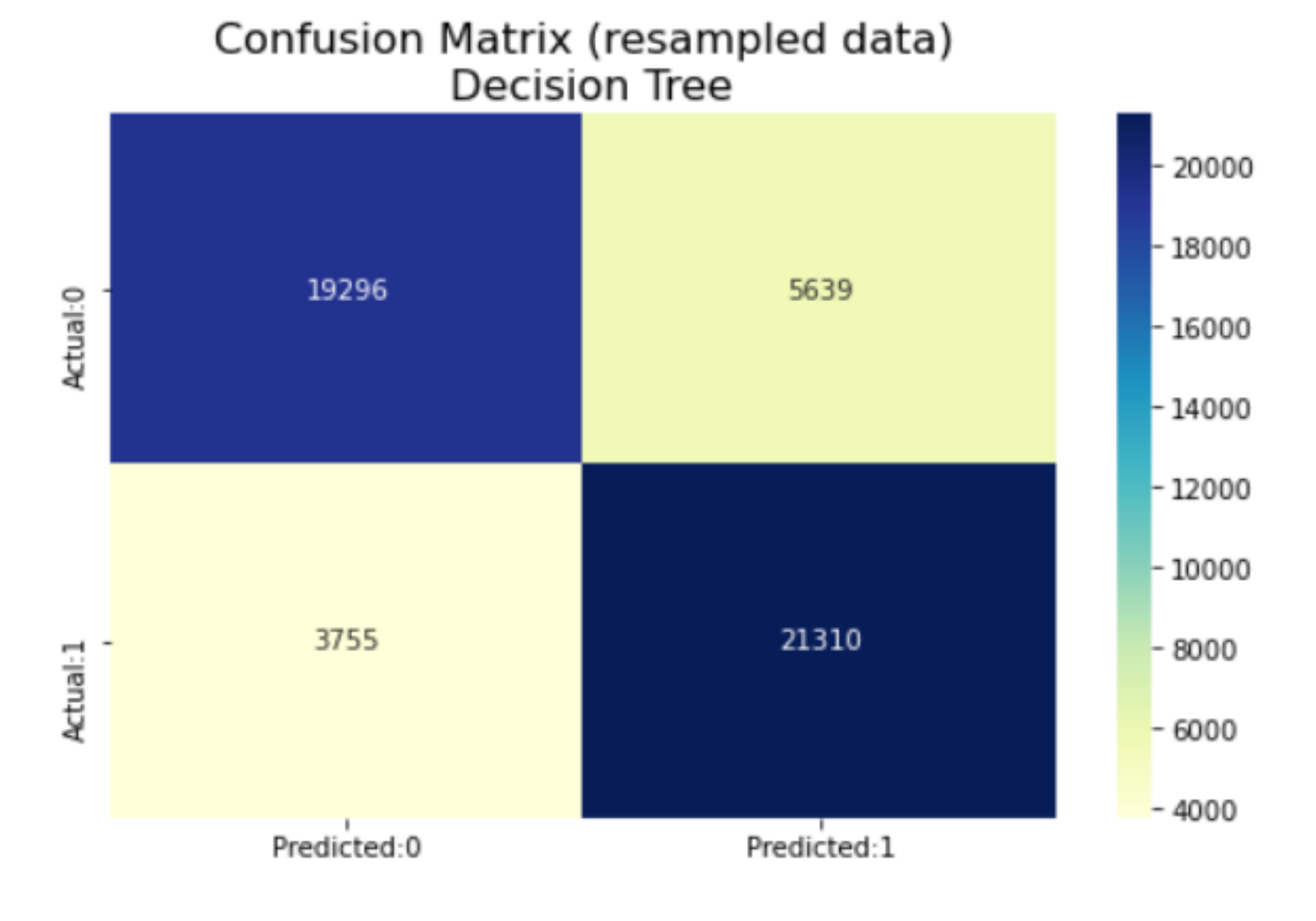
Accidents were higher with Wind directions from West, South and North.

CONFUSION MATRIX FROM LOGISTIC REGRESSION MODEL:



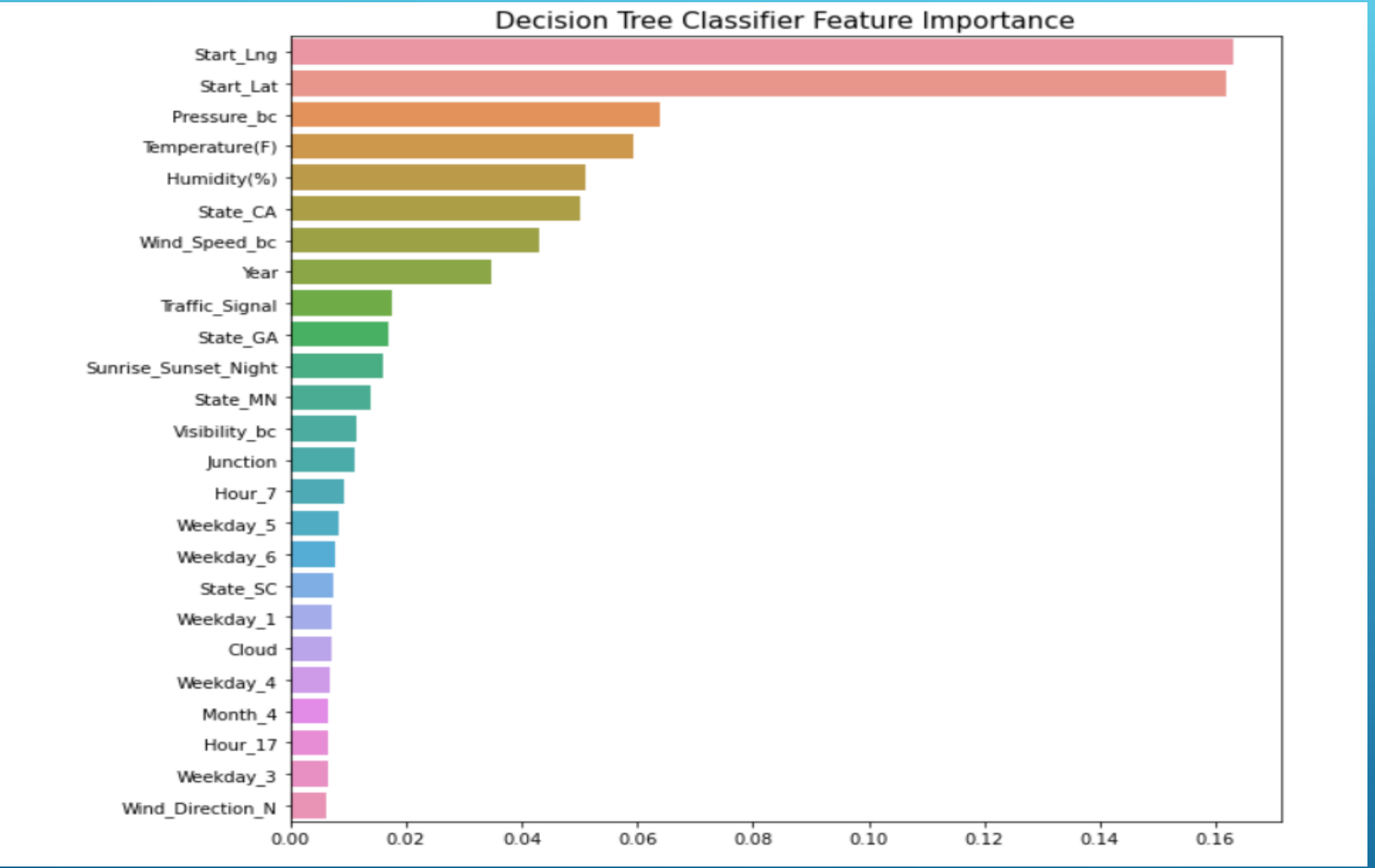
Train Accuracy was 70.9%.
Test Accuracy was 71.1%

CONFUSION MATRIX FROM DECISION TREE MODEL:

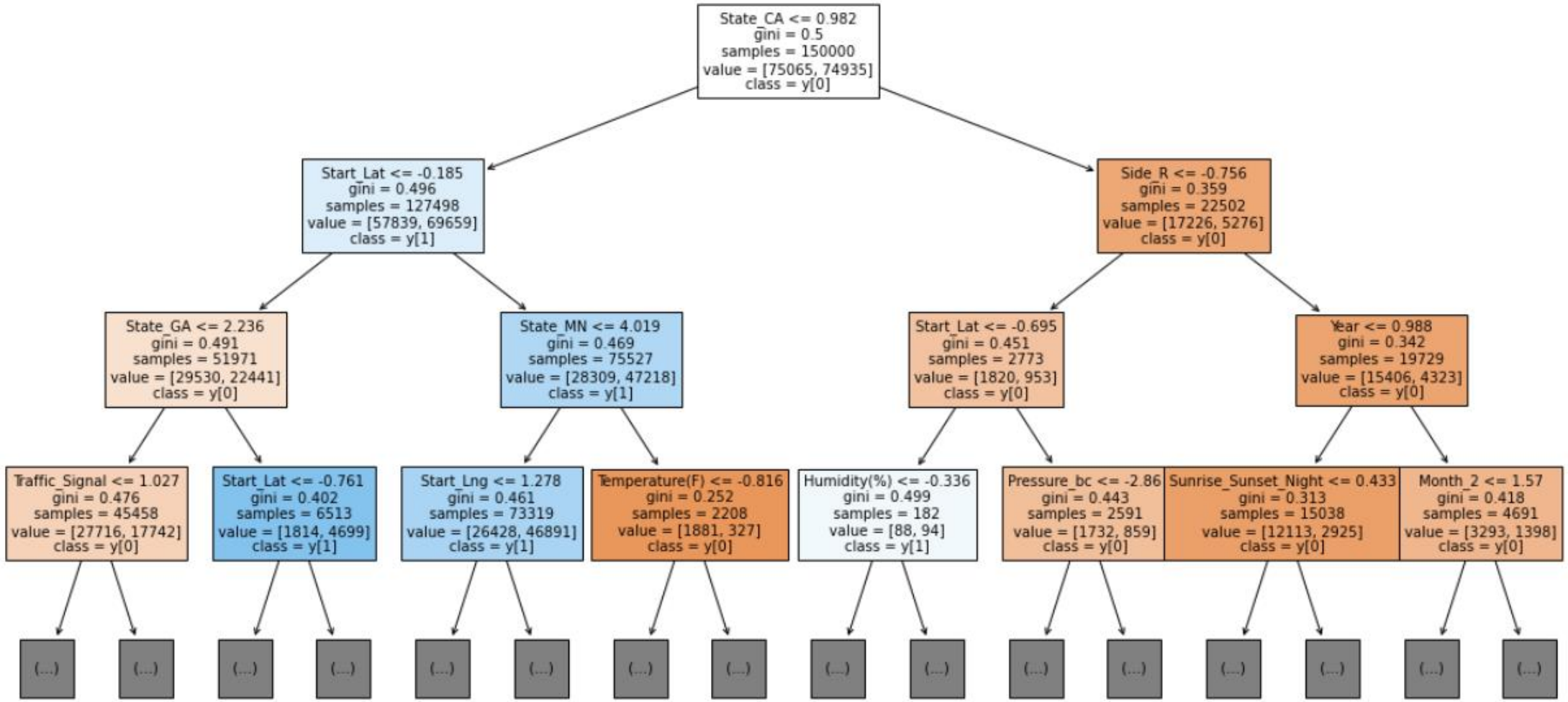


Train Accuracy was 99.5%.
Test Accuracy was 81.2%

DECISION TREE FEATURE IMPORTANCE:



Plot shows us the 25 important features useful in severity prediction.


State_CA <= 0.982
gini = 0.5
samples = 150000
value = [75065, 74935]
class = y[0]
Start_Lat <= -0.185
gini = 0.496
samples = 127498
value = [57839, 69659]
class = y[1]
Side_R <= -0.756
gini = 0.359
samples = 22502
value = [17226, 5276]
class = y[0]
State_GA <= 2.236
gini = 0.491
samples = 51971
value = [29530, 22441]
class = y[0]
State_MN <= 4.019
gini = 0.469
samples = 75527
value = [28309, 47218]
class = y[1]
Start_Lat <= -0.695
gini = 0.451
samples = 2773
value = [1820, 953]
class = y[0]
Year <= 0.988
gini = 0.342
samples = 19729
value = [15406, 4323]
class = y[0]
Traffic_Signal <= 1.027
gini = 0.476
samples = 45458
value = [27716, 17742]
class = y[0]
Start_Lat <= -0.761
gini = 0.402
samples = 6513
value = [1814, 4699]
class = y[1]
Start_Lng <= 1.278
gini = 0.461
samples = 73319
value = [26428, 46891]
class = y[1]
Temperature(F) <= -0.816
gini = 0.252
samples = 2208
value = [1881, 327]
class = y[0]
Humidity(%) <= -0.336
gini = 0.499
samples = 182
value = [88, 94]
class = y[1]
Pressure_bc <= -2.86
gini = 0.443
samples = 2591
value = [1732, 859]
class = y[0]
Sunrise_Sunset_Night <= 0.433
gini = 0.313
samples = 15038
value = [12113, 2925]
class = y[0]
Month_2 <= 1.57
gini = 0.418
samples = 4691
value = [3293, 1398]
class = y[0]
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

CONCLUSION:

- Accident Severity can be predicted with few attributes such as time, period of day, location and weather.
- Accidents during night time are much sever than day time accidents.
- Accidents are much likely to occur on the right side of the road.
- Wind Directions also play an important role in accident severity prediction.
- Number of accidents are higher during early hours of the day.

FUTURE WORK:

- ML models created in this project can be incorporated in real time to predict accidents.
- Detailed *relations* between different important factors can be further studied.