Analysis of Traffic violations in Maryland County

Anirudh BS
Computer Science
PES UNIVERSITY
Bangalore
anibs2171@gmail.com

KS Abhisheka
Computer Science
PES UNIVERSITY
Bangalore
ksabhisheka@gmail.com

Abhishek D
Computer Science
PES UNIVERSITY
Bangalore
abhishekdinesh21042001@gmail.com

Abdul Rahman Shigihalli
Computer Science
PES UNIVERSITY
Bangalore
abdulrahmanshigihalli@gmail.com

***Abstract*— The number of fatalities, property damages, personal injuries, and traffic violations has multiplication since the increase in the production of vehicles. Much chaos can be created with improper and irresponsible usage of vehicles. Along with the damage, many of the vehicles have been charged for violating the traffic rules and penalized for the same. In this article, we go through the data collected from Maryland County and mention various insights. A brief literature review also shows the works that have been done in similar areas of interest. A time series model is built to forecast the accident statistics in the future.**

Keywords—traffic, seatbelts, race, damage, fatality, alcohol consumption, time series.

I. INTRODUCTION

In this world of growing technology and the full speed of developing products, many of the companies have become successful in producing their products in no time. In this context, in a small period, many car companies were determinant in producing a large number of cars which had its advantages and disadvantages. Due to the increase in the number of cars in the past decades, it has been noticed that traffic violations have also increased steadily. Traffic violations are defined as an inappropriate act of any citizen seen to break the rules imposed by the government on traffic. It will include not wearing a seatbelt, drink and drive cases, speeding, registration, etc. The dataset we have chosen contains the traffic rules violated by the driver in Maryland Count from the year 1964 to 2018. Initially, this dataset contained ten lakhs plus entries. It was difficult to process the data. Therefore, we decided to reduce the count of rows in the dataset by breaking the dataset into two data frames based on values of Alcohol and Belts features and then concatenating both the data frames randomly to get a count of around one lakh rows.

The number of damages or casualties, be it property damage or personal fatality keeps increasing with the increase in traffic violations year by year. It can either be because of drinking and drinking or not wearing a seat belt. As we have raced in our dataset as one of the features we would also make use of that to check for racial disparity in traffic violations. It can be considered Racial profiling. Using the race feature we can find out which race is violating the traffic rules frequently and being subjected to charges given by the police for doing so. We would also like to work on figuring out which city the driver belongs to by using the geolocation feature in the dataset.

II. Literature Review

In the paper, Investigating the Relationship between Traffic Violations and Crashes at Signalized Intersections: An Empirical Study in China [1], the researchers investigated the relationship between traffic violations and traffic crashes. Various types of data were collected from thirty-one different signalized intersections for one year. The data included the traffic volume, traffic violation, and traffic crash data. A White test was used to test the homoscedasticity of the data and a multiple linear regression model was employed to investigate the relationship between traffic crashes and violations.

The quantitative research on the relationship between unsafe acts and accidents can be traced back to the classic Heinrich's Law[2].

In his 1931 book "Industrial Accident Prevention, A Scientific Approach", Herbert W Heinrich put forward the following concept that became known as Heinrich's Law: in a workplace, for every accident that causes a major injury, there are 29 accidents that cause minor injuries and 300 accidents that cause no injuries. Although this conclusion is summarized for a factory workplace, Heinrich's work is still claimed as the basis for the theory of behavior-based safety.

In another research conducted by Bjørnskau and Elvik in the paper "Can road traffic law enforcement permanently reduce the number of accidents?"[3], shows that road users tend to abide by the law if they are being observed by the police, and violate if there are no police around. This means that attempts at any on-site enforcement will not have long-lasting effects, either on the road-user behavior or on the crashes.

Another research was done by Retting, Ferguson, and Hakkert in the paper "Effects of red light cameras on violations and crashes: a review of the international literature" [4], suggested that electronic police enforcement can generally reduce violations by an estimated 40–50%.

In the paper, Analysis of the Impact of Traffic Violation Monitoring on the Vehicle Speeds of Urban Main Road: Taking China as an Example [5], analysis of the vehicle speeds nearby road traffic violation monitoring area on urban main roads has been done. It focussed on finding the impact of road traffic violation monitoring on vehicle speeds.

According to the authors of "A Geographical Approach to Racial Profiling", minorities are more likely to be stopped and ticketed/arrested for traffic violations. Furthermore, some studies suggest that crimes are more likely to be committed by minorities(Gaines,2006). The macro-level analysis by the authors discovered the unfavorable treatment of police towards people in the region where Blacks or Hispanics resided. Findings of this study suggested that minority drivers may be stopped, searched, arrested, and charged with a felony because they are more likely to drive in high crime areas where they reside, and more vigorous law enforcement is a common practice(Robinson Matthew, racesRoh Sughoon,2009)

III. Problem Statement

Using the traffic violations dataset with its available features, we will analyze the relationship between seat-belts and accidents, thus signifying the importance of seat belts, and also explore which group of people are most responsible among others to wear seat belts. With the given data, we have analysed the damage rates of various car makers. The advantage of analyzing such a statement gives a clear note on which model to choose from the right makers for personal transportation and various other transportation. Finally, we are

predicting the violation statistics for the later months using forecasting.

Data preprocessing

The dataset mostly consists of records that show no alcohol consumption and no belt violations, which are not of much significance for our goal. Thus we conduct dimensional reduction, wherein we have selected only a fraction of such records. After dimensionality reduction the number of records reduced from 12092399 to 108863. The missing data existing in the dataset were imputed using mode (for attributes like Make, Model, Color) or using mean (for attributes like Year). We also noticed that the values for the attribute Make (car maker) had many inconsistencies. To remove the inconsistencies we used an autocorrect function, which compares the values with a list of corrected names using a similarity function.

Exploratory data analysis

Role of belts :

As evident from the figures Fig.1 and Fig.2, we can see that when there is any personal injury or property damage, there is belt violation involved. From this, we can realize the importance and the role that seat belts have in any road accidents.

Belts and personal injury

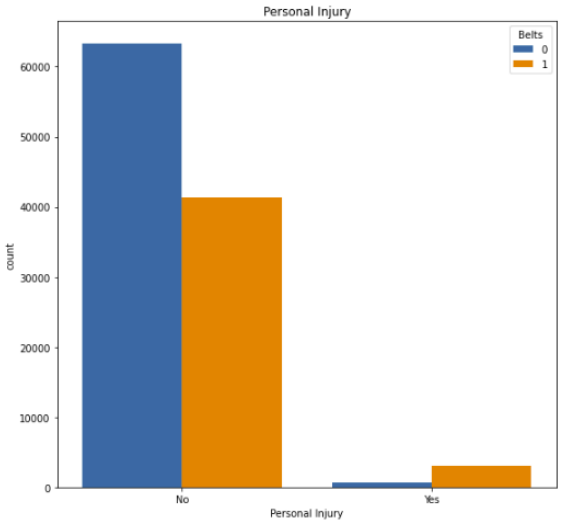

Fig. 1

Belts and property damage

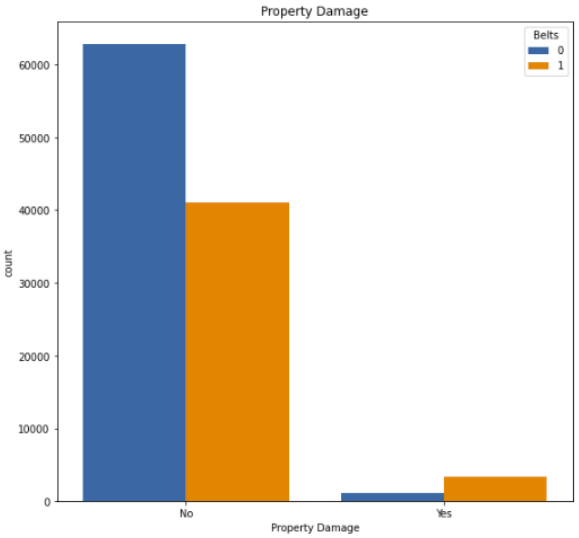

Fig. 2

Belts and Gender

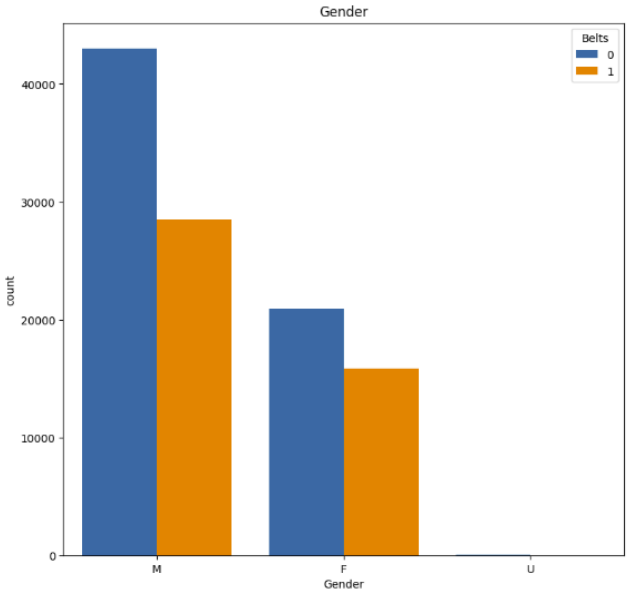

Fig. 3

From the above plot, the ratio of belt violations is greater with females.

Belts and Race

Fig. 4

From the above coxcomb chart, we see that Asians have a record of most belt violations whereas the Native Americans have the lowest ratio of belt violations.

Make and Personal Injury

We also analysed the ratio of personal injury to the total number of traffic violations recorded for each car maker.

```
['0.5', 'PLYMOTH'],
['0.17777777777777778', 'Chester Built'],
['0.16853932584269662', 'WildFire'],
```

Fig. 5

Fig. 5, shows the makers which have been the most involved in accidents involving personal injury.

```
['0.013968775677896467', 'Infiniti'],
['0.0136986301369863', 'SAAB'],
['0.013636363636363636', 'Jaguar'],
['0.010582010582010581', 'Mini'],
['0.009174311926605505', 'Landrover']
```

Fig. 6

Fig. 6, shows the makers which have the lowest cases of personal injury.

Time Series analysis and forecasting

We created a new attribute, namely number of violations. We assigned its value to 1 for every record present. The data frame was then resampled month-wise which also reflected the number of violations every month from the January of 2012 to the April of 2018.

The time series data of the number of violations in a calendar month from 2012 to 2018 is shown in Fig. 7.

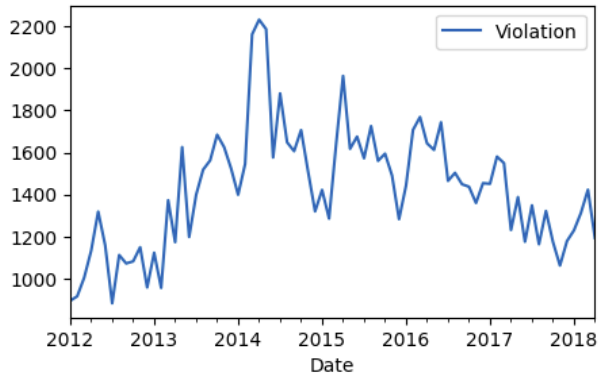

Fig. 7

Exploring Trend, Stationarity and Residual components

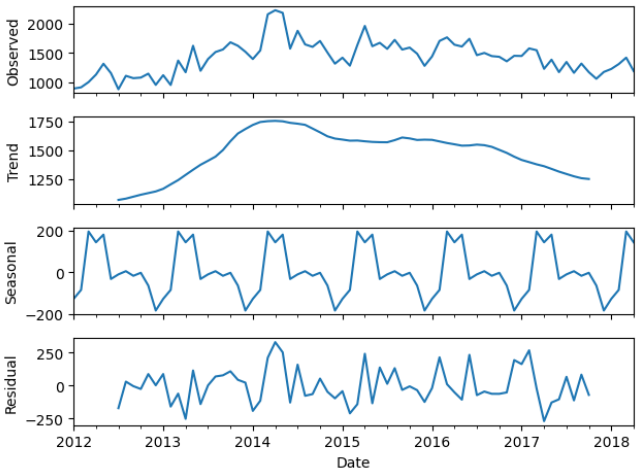

Fig. 8

Testing for Stationarity

We have used the ADF test (Augmented Dickey Fuller) to check for stationarity.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} .. + \phi_p \Delta Y_{t-p} \cdot$$

The test results are shown in Fig. 9.

The p value is greater than 0.05. And the ADF statistic is higher than two of the critical values. We accept the null hypothesis that . In order to transform the time series to a stationary series, we use differencing.

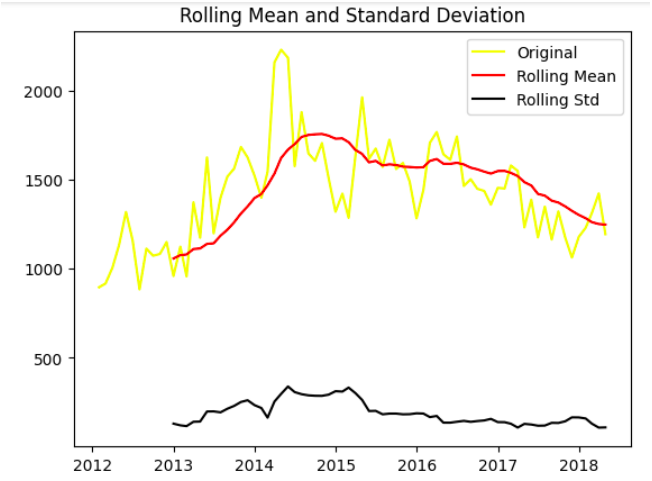

Fig. 9

Since there is a seasonal component, we will use SARIMA.

Before that, we decided to see how the ARIMA model will perform.
Fig. 10 shows true values (blue) and predicted values from mid 2017.

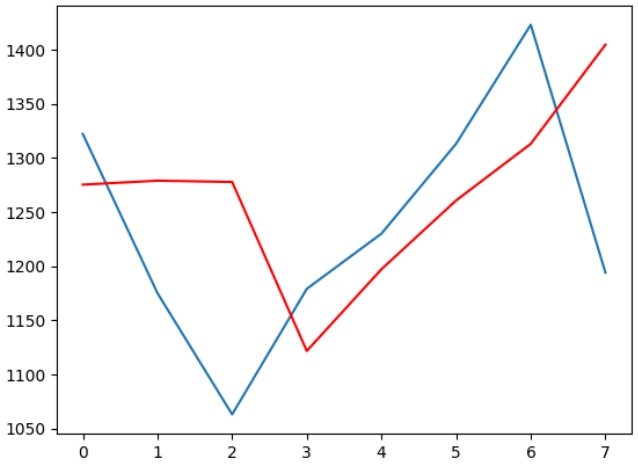

Fig. 10

The model was evaluated by using mean absolute percentage error, which was obtained as 0.087

We used BIC(Bayesian Information Criteria) and AIC(Akaike's Information Criteria) to evaluate SARIMA models.

$$AIC_i = -2\log L_i + 2p_i$$

$$BIC_i = -2\log L_i + p_i \log n$$

L = Likelihood function
p = number of parameters
n=sample size

Of all the SARIMA models, we chose the one with lowest BIC and AIC values.

	pdq	pdqs	bic	aic
545	(2, 2, 1)	(2, 2, 1, 12)	338.248947	330.002570
524	(2, 2, 0)	(2, 2, 1, 12)	341.649963	334.581640
544	(2, 2, 1)	(2, 2, 0, 12)	346.373420	339.305097

Fig. 11

The optimum p,d and q values are 2,2 and 1 as shown in Fig. .

Statespace Model Results

Dep. Variable:	Accident	No. Observations:	76
Model:	SARIMAX(2, 2, 1)x(2, 2, 1, 12)	Log Likelihood	-158.001
Date:	Sun, 05 Dec 2021	AIC	330.003
Time:	11:13:36	BIC	338.249
Sample:	01-31-2012 - 04-30-2018	HQIC	332.190
Covariance Type:	opg		

Fig. 12

The normal QQ plot for the model built is shown in the plot in Fig. 13.

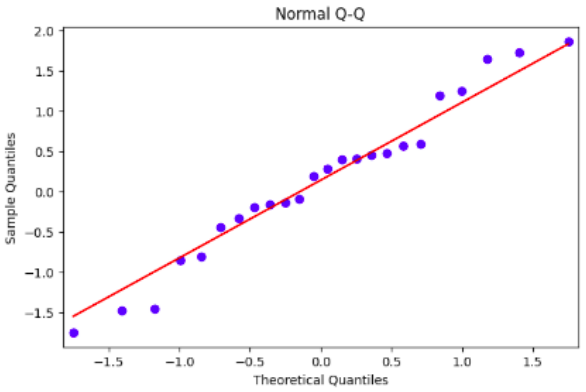

Fig. 13

The model performs well as can be seen in the Fig. . The figure shows One-step ahead forecasts from mid 2017.

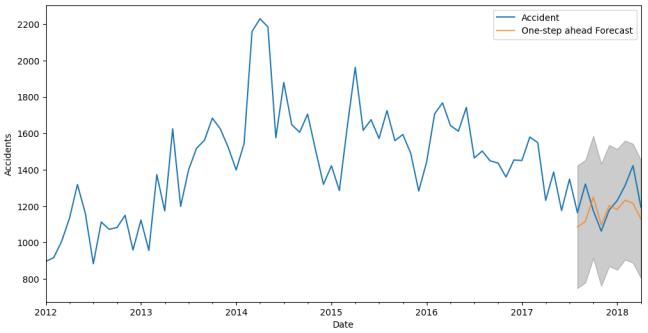

Fig. 14

The above model's goodness can be measured using metrics like RMSE, MAE etc. We have considered the Mean Absolute Percentage Error for evaluating the model. The mean absolute percentage error of our forecasts is 0.071.

$$MAPE = \frac{\sum \frac{|A-F|}{A} \times 100}{N}$$

A = Actual value
F = Forecasted value
N = sample size

Future forecasting

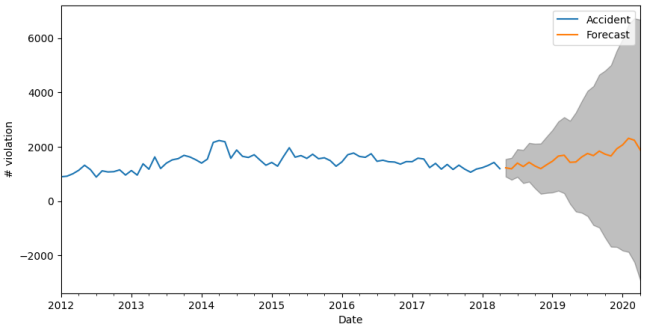

Fig. 15

Fig. 15. shows the forecasted values till 2020.

Conclusion and Future Work

We can infer from the above analysis that SARIMA performs significantly better than ARIMA when there is a seasonal component involved. The future forecasts that were conducted can be tested with the newly released data, and evaluated using Mean absolute percentage error to test the accuracy.

References

[1] Ze-Hao Jiang, Xiao-Guang Yang, Tuo Sun, Tao Wang, Zheng Yang, "Investigating the Relationship between Traffic Violations and Crashes at Signalized Intersections: An Empirical Study in China", *Journal of Advanced Transportation*, vol. 2021, Article ID 4317214, 8 pages, 2021. https://doi.org/10.1155/2021/4317214

[2] H. W. Heinrich, *Industrial Accident Prevention: A Scientific Approach*, McGraw-Hill, New York, NY, USA, 1941.

[3] T. Bjørnskau and R. Elvik, "Can road traffic law enforcement permanently reduce the number of accidents?" *Accident Analysis & Prevention*, vol. 24, no. 5, pp. 507–520, 1992.

[4] R. A. Retting, S. A. Ferguson, and A. S. Hakkert, "Effects of red light cameras on violations and crashes: a review of the international literature," *Traffic Injury Prevention*, vol. 4, no. 1, pp. 17–23, 2003.

[5] Fuquan Pan, Yongzheng Yang, Lixia Zhang, Changxi Ma, Jinshun Yang, Xilong Zhang, "Analysis of the Impact of Traffic Violation Monitoring on the Vehicle Speeds of Urban Main Road: Taking China as an Example", *Journal of Advanced Transportation*, vol. 2020, Article ID 6304651, 11 pages, 2020. https://doi.org/10.1155/2020/6304651

[6] Roh, S.; Robinson, M. (2009). *A Geographic Approach to Racial Profiling: The Microanalysis and Macroanalysis of Racial Disparity in Traffic Stops.*