

# Don't cross that stop line: Characterizing Traffic Violations in Metropolitan Cities

Shashank Srikanth\*

IIIT Hyderabad  
shashank.s@research.iiit.ac.in

Aanshul Sadaria\*

IIIT Hyderabad  
aanshul.sadaria@students.iiit.ac.in

Himanshu Bhatia\*

IIIT Hyderabad  
himanshu.bhatia@students.iiit.ac.in

Kanay Gupta\*

IIIT Hyderabad  
kanay.gupta@students.iiit.ac.in

Pratik Jain\*

IIIT Hyderabad  
pratik.jain@students.iiit.ac.in

Ponnurangam Kumaraguru<sup>†</sup>

IIIT Delhi  
pk@iiitd.ac.in

## Abstract

In modern metropolitan cities, the task of ensuring safe roads is of paramount importance. Automated systems of e-challans (Electronic traffic-violation receipt) are now being deployed across cities to record traffic violations and to issue fines. In the present study, an automated e-challan system established in Ahmedabad (Gujarat, India) has been analyzed for characterizing user behaviour<sup>1</sup>, violation types as well as finding spatial and temporal patterns in the data. We describe a method of collecting e-challan data from the e-challan portal of Ahmedabad traffic police and create a dataset of over 3 million e-challans. The dataset was first analyzed to characterize user behaviour with respect to repeat offenses and fine payment. We demonstrate that a lot of users repeat their offenses (traffic violation) frequently and are less likely to pay fines of higher value. Next, we analyze the data from a spatial and temporal perspective and identify certain spatio-temporal patterns present in our dataset. We find that there is a drastic increase/decrease in the number of e-challans issued during the festival days and identify a few hotspots in the city that have high intensity of traffic violations. In the end, we propose a set of 5 features to model recidivism in traffic violations and train multiple classifiers on our dataset to evaluate the effectiveness of our proposed features. The proposed approach achieves 95% accuracy on the dataset.

## 1 Introduction

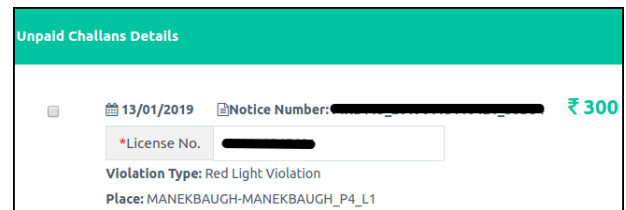
Traffic accidents were responsible for over 1 million deaths all over the world in the year 2016 [1]. Of these accidents, more than 90% occur in developing countries. Previous research in the field of behaviour studies regarding traffic rule violations has shown that in greater than 70% of the cases, the role of human behaviour is one of the causes [2]. Most of these accidents can be prevented if the traffic rules are properly followed and as a result, the traffic police across states in India are adopting automated traffic management systems to promote adherence to traffic rules [3]. These automated systems are capable of varied tasks like capturing violations and issuance of e-challan (Electronic traffic-violation receipts) without any human intervention. These systems can also generate e-challans along with photo evidence and send it to violators through SMS/email/post. Figure 1 shows an example of an e-challan generated in Ahmedabad.

The e-challan consists of several details of a given traffic violation like the time, place of violation and other information like the violation type and corresponding fine amount. Thus, automated traffic management systems like those in Ahmedabad can be mined to extract traffic violations data and used for estimating the possibility of repeat offenses. Such datasets can also be used to characterize the type of traffic violations in a city and identify the spatial and temporal patterns of the traffic violations. The Ahmedabad traffic police launched their automated traffic management system in 2015 and it leverages a network of 6,000 video surveillance cameras (dedicated to detect *red light violations*) installed across 130 traffic junctions. The system has been an enormous success and a total of 1.27 million *stop line violation* e-challans were issued in the year 2018 [4].

Despite the ever-increasing amount of traffic violations data being available, there has not been a systematic analysis of such violations. Such an analysis would be particularly useful for the government which is responsible for framing laws and the police which is responsible for making the roads safer for citizens. In this work, we carry out a longitudinal study of e-challan receipts in the city of Ahmedabad and investigate the effectiveness of the above system in reducing repeat offences. Unlike some earlier works [5, 6], which analyze road accident data, we restrict our analysis only to traffic violations. In order to understand traffic violations from the prism of big-data analysis, we address the following questions.

### 1.1 Research Questions

Different types of traffic violations require different varieties of preventive measures. Similarly, not all people are equally prone to



**Figure 1: A sample e-challan. It contains several important details regarding the traffic violations such as the time and location of violation along with the type of violation and corresponding fine amount. It also includes data about whether the e-challan has been paid or not. The license number and notice number have been hidden to protect privacy.**

\*These authors contributed equally to work.

<sup>†</sup>Major part of the work was done during a year long sabbatical at IIIT Hyderabad.

<sup>1</sup>Note that we have used the term "user". We assume that each vehicle is associated with a unique individual and henceforth will consider them to be equivalent.

committing traffic violations, and quite a few of them can be serial offenders. Unlike traditional challan systems, where payments were made at the time of issue itself, e-challan systems allow people to pay the fines later. Thus, people often delay paying their fines unless forced to or are simply unaware of the fines being issued to them [7]. Therefore, we ask our first research question:

**RQ1:** *What are the characteristics of the users who are issued the e-challans and their corresponding traffic violations?*

The e-challan issue date and fine payments date exhibit some interesting temporal patterns. Besides, the information from the spatial dimension will also be useful to clusters of regions in the city where violations are more likely to occur. Thus, we are interested in solving:

**RQ2:** *What are the temporal and spatial patterns of traffic violations?*

A lot of research has been conducted on recidivism cases amongst prisoners and sex offenders [8, 9]. Most of these works focused on identifying factors that are strongly related to recidivism. Using similar methods to predict recidivism in the case of traffic violations would be very useful for traffic police and government authorities. Thus, we propose the following research question:

**RQ3:** *Can instances of repeat offences be predicted with reasonable accuracy based on just the history of traffic violations of a person?*

## 1.2 Contributions

The main contributions of this work are:

- We describe a method to collect large-scale data of traffic violations from the e-challan portal of Ahmedabad and collect a dataset of over 3 million e-challans.
- We characterize the user behaviour in our dataset with respect to repeat offences and fine payment. We show that users with more violations are less likely to pay their fines and users, in general, prefer to pay the smaller fines as compared to the larger ones. Moreover, we also show that a few of the violations account for most of the e-challans.
- We analyze the temporal distribution of the traffic violation data over all the 4 years and show that there is a drastic increase/decrease in the number of fines issued during the festival days. And we show the emergence of new traffic violation hot-spots over time by performing spatial and temporal analysis simultaneously.
- We create a dataset consisting of over 600,000 users and their corresponding recidivism history. We identify and describe several attributes that can be inferred from a user's past violation history and be used for predicting recidivism.

## 1.3 Privacy and Ethics

We collect data from Ahmedabad traffic police's e-challan portal<sup>2</sup> and all the data used is publicly available. Additionally, we do not use any personally identifiable information for our analysis.

The rest of this paper is organized as follows: We discuss the related work in Section 2 and Section 3 discusses in detail the aspects of data collection. We characterize the user behaviour and the type of violations in our dataset in Section 4 and in Section 5, we identify the spatial and temporal patterns of traffic violations. We describe

the proposed features and machine learning methods to predict recidivism in Section 6. The work is concluded in Section 7 and Section 8 deals with possible extensions of the project.

## 2 Related Work

We structure the discussion of related work into two main themes: related work on traffic and road violations data, and work concerning predicting recidivism amongst convicts.

**Traffic Accidents and Violations:** Traffic accidents account for more than a million deaths each year across the world according to WHO [10] and a lot of research has been conducted on data concerning road accidents. Sanjay et al. [11] analyze the road accidents data across India at a national, state and metropolitan city level and show that distribution of road accident deaths and injuries vary according to age, gender, month and time. They also show that more than 50% of the cities face higher fatality risks as compared to their rural counterparts. Another set of approaches to model road accidents involve identifying road accident hotspots using GIS (Geographical Information System) technologies. Choudhary et al. [5] geocoded 5 years of road accident locations over the digital map of Varanasi and clustered accidents using a spatial heatmap. In contrast to these approaches, which analyze and model road accidents data, we analyze only the traffic violations data of the city of Ahmedabad. The insights gained from analyzing traffic violations would allow the appropriate agencies to take suitable preventive measures. Qiqi et al. [12] show that traffic violations amongst bus drivers are associated with the date, weather, and presence of traffic cameras at bus stations. The analysis done by Qiqi et al. [12] is the most similar to our work, but their dataset size is much smaller than ours in terms of size, and they do not focus on predicting recidivism.

**Predicting Recidivism:** A lot of research has been done on the task of predicting recidivism amongst convicts of crime. There have been several approaches to predicting recidivism which range from using survival time models [13] and machine learning methods [14]. Work by Gerald et al. [15] proposes several independent characteristics like demographics, offense type, location and spatial contagion to model recidivism. They use GIS (Geographic Information System) and logistic regression modelling to show that the likelihood of re-incarceration was increased with male gender, offense type and certain locations. Random forest models have also been applied successfully to predict recidivism in [14] by leveraging a large number of features. Our work differs from all these methods as it aims to predict the recidivism of traffic violations, unlike other criminal activities. Similar to the work by Nancy et al. [14], we also utilize a random forest model to predict recidivism of users based on their past violation history.

## 3 Dataset collection and description

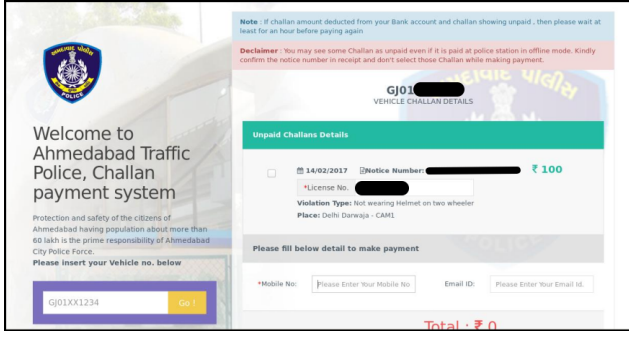
For this work, we collect traffic violations data (e-challan) from the e-challan portal of Ahmedabad traffic police<sup>3</sup>.

### 3.1 Data collection

Ahmedabad traffic police have an online e-challan portal that is used by people to check the status of fines and also make the relevant payments online. A brief description of the portal usage and its features has been shown and discussed in Figure 2. The

<sup>2</sup><https://payahmedabadchallan.org>

<sup>3</sup><https://payahmedabadchallan.org>



**Figure 2: Ahmedabad traffic police’s e-challan portal - on entering the vehicle registration number in the bottom left text box, the details of unpaid e-challans are displayed in the right window pane.**

portal does not require a user to have a login ID or password, and one needs to enter their vehicle registration ID to get the details of all e-challans issued to that vehicle. This is in contrast to various other cities like Hyderabad, Mumbai, etc., where login ID with a password is mandatory. Thus, here we can view the set of e-challans issued to other vehicles as well. The portal does not reveal any personally identifiable information unless a user demands a receipt of payment.

For obtaining the data through the e-challan portal, we leverage the Selenium headless web browser to make millions of web requests to their server and collect over 3 million e-challans. In each new request, we provide a new vehicle registration number in the corresponding header files and obtain traffic violations data for that vehicle. As we do not have access to the list of registered vehicles in the city of Ahmedabad, we make a request to the portal with all possible vehicle registration numbers in Ahmedabad. Vehicle registration IDs across cities in India have a total of 10 characters and barring a few exceptions look like the following: GJ-01-AB-1234. The registration IDs are divided into four main parts:

- First two letters indicate the state/union territory in which the vehicle is registered
- The next two numbers are sequential numbers of the registration district
- The third portion represents the ongoing series of an RTO (Regional Transport Office) which issues the registration IDs
- The last part consists of a 4 digit number, which combined with the previous three parts, is unique to each registered vehicle.

Thus, each vehicle is uniquely identified by its registration ID, and two vehicles cannot have the same ID. To get the violation data of all the vehicles in Ahmedabad, we enumerate all possible combinations of vehicle registration numbers that begin with the prefix of GJ01, which is one of the prefix codes for the city of Ahmedabad. We add a two-letter suffix and a 4 digit number after the letters GJ01 to get a total of  $6,760,000 (26 * 26 * 10^4)$  vehicle registration IDs combinations and query the server with these IDs.

One can also get the payment date of a paid e-challan by requesting for a receipt of the payment in the portal. Downloading and

parsing millions of e-challan receipts in PDF format is computationally expensive, so we randomly sample 4,500 users from our dataset that paid at least one of their fines and collect their fine payment data. The payment receipt is provided in the form of a PDF, and we use Tabula to parse the PDF into JSON format suitable for analysis.

### 3.2 Dataset

We collect a total of over 3 million e-challans over a period of 4 years from 2016 – 2019. There were a total of 1,177,695 unique vehicles with one or more e-challans and they together account for 808,004,725 in fines to the government. For each unique e-challan in our dataset, we have five major attributes - 1) Date of the violation, 2) Location of the violation, 3) Type of violation, 4) Fine amount, 5) Paid or unpaid e-challan (Boolean). Using this data, we compute the total amount of fines paid and owed to the government by computing the sum of all the fines associated with each e-challan in our dataset for paid and unpaid e-challans, respectively. Table 1 provides a quantitative description of our dataset.

Description	
Total number of e-challans	3,571,341
Number of paid e-challans	1,082,132
Number of unpaid e-challans	2,489,209
Number of unique vehicles	1,177,695
Date of first e-challan in our dataset	29th September 2015
Date of last e-challan in our dataset	22st August 2019
Total amount owed to the government	808,004,725
Total amount paid already	215,574,325
Total amount unpaid till 22nd August	592,430,400
Number of types of violations	18
Number of unique locations	135
Number of fine denominations	28

**Table 1: A brief description of the dataset statistics.**

### 3.3 Data Preprocessing

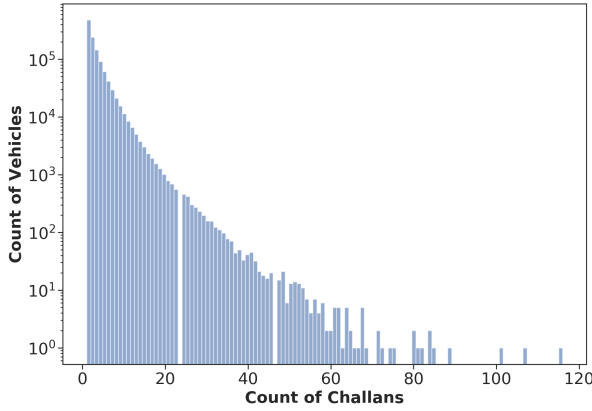
Due to the presence of multiple cameras in a single location, two cameras may have the same geographical location but a different name in the location field of the e-challan. Thus, we cleaned the data by updating the location field of all e-challans to a specific geographic location to perform the spatial analysis. The data after the cleaning process had 135 unique locations across the city and was used for further analysis.

## 4 Characterizing users and types of violations

In this section, we characterize the user behaviour with respect to e-challan payment and repeat offences. We also analyze the distribution of different types of violations in our dataset and discuss their spatial plots later.

### 4.1 Repeat offences

Figure 3 shows the distribution of users and the number of e-challans issued to each one of them. We compute the median number of e-challans issued to all the users, and this comes out to be around 2. Thus, at least half the individuals in our dataset have more than 2 e-challans issued to them. This, suggests that the users are prone to commit traffic violations repeatedly. Figure 3 also reveals



**Figure 3: Count of people who have committed specific number of violations. It shows that large number of vehicles have same number of challans.**

that there are users with even 80 or more e-challans, suggesting that many of them might not be aware of the e-challans being issued to them. This is further confirmed by the fact that out of the 11 users with more than 80 e-challans in our dataset, 5 of them have not paid even one of their e-challans.

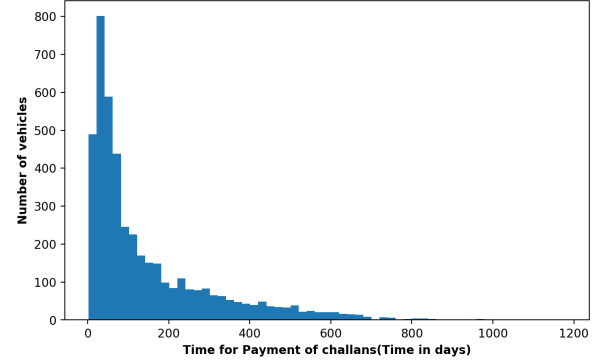
Violation Type	Number of e-challans
Red Light Violation	2,102,105
No Helmet Violation	1,011,263
Improper Parking	206,765
Stop Line Violation	128,658
Driving Without Seatbelt	39,202
Total Number of e-challans	3,571,341

**Table 2: Distribution of e-challans with violation types.**

#### 4.2 Characterizing paid e-challans

Table 1 shows that the total number of paid e-challans is much less than the unpaid e-challans. From Table 1, we see that the ratio of unpaid to paid e-challans is 2.3, and the ratio of fine amount of the unpaid e-challans to that of paid e-challans is 2.75. This suggests that the majority of paid e-challans consist of lower fine amounts as the fine payment ratio is 1.19 times the issued challans ratio. We further analyze the distribution of paid e-challans ratio with respect to the fine amount and find that fines with lesser amounts like Rs. 100, and 50 are more likely to be paid as compared to those of higher denominations like 2,000. The difference in the ratio is drastic, with almost 37.37% of the Rs. 100 fines being paid as compared to 15.44% of Rs. 2000 fines. In an ideal setting, the distribution of ratio of paid e-challans is expected to be uniform with respect to the type of violation, given that the fine amount is same. However, we find that the e-challans of *overspeeding violation* have drastically higher ratio (0.60) of paid e-challans as compared to 0.12 for *brts lane violation* where their average fine amount is Rs. 1017.67 and Rs. 1265.2 respectively.

To characterize user behavior, we analyze the users who had both paid and unpaid e-challans issued to their vehicles. Out of the



**Figure 4: Count of vehicles with their e-challan payment time.**

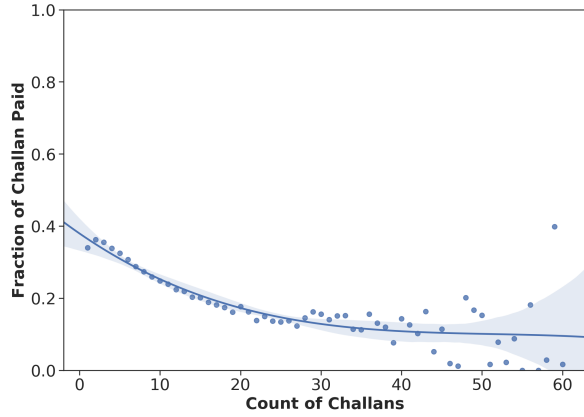
total vehicles as shown in Section 3.2, 229,042 vehicles have both paid and unpaid e-challans issued. Ideally, it is expected that the unpaid e-challans for any vehicle are the recently issued e-challans, but interestingly 31.5% of those vehicles had at least one unpaid e-challan before the last paid e-challan. We characterize these users and find that the fine amounts such as Rs. 50 contribute much less than the higher denomination fines. In terms of violation types, we find that *no helmet violations* constituted more than 50% of the total e-challans in this category which differs drastically from the normal distribution as shown in Table 2. We conclude from this that the violation type and fine amount of the e-challans play a significant role in characterizing fine payment.

To analyze the fine payment behaviour of the users, we compute the average payment time for 4,500 random users sampled from our dataset, as discussed in Section 3. We find that on an average, the users paid back their fines in 153 days. A histogram of the average fine payment time of each of the 4,500 users is given in Figure 4. It indicates that there is a lack of incentive to pay the fines early. A reason for the high average fine payment time could be the lack of awareness amongst the users about the e-challan system, and this needs to be adequately tackled for the system to be effective.

We also analyze the impact of the number of e-challans that a user has and their fine payment ratio. We plot the distribution of the number of e-challans issued and the average ratio of e-challan payment in Figure 5. For Figure 5, we removed the data of 42 users who had an exceptionally high number of violations as they were outliers (less than 0.01% of the users). We can see that people with a lot of e-challans are less likely to pay the fines back as compared to those with fewer e-challans. The distribution fits a third-order curve that decreases with increasing e-challans. The above analysis suggests that the fine payment is highly dependent on the user, the fine amount, and the type of violation associated with the e-challan.

#### 4.3 Violation types analysis

The number of e-challans issued for different types of violations is presented in Table 2. We present the numbers for the top 5 violation types in our dataset that account for more than 1% of all the e-challans. We can infer from Table 2 that the top 2 violations by themselves account for more than 87% of all the e-challans issued in our dataset. Thus, specific targeted measures can be taken to

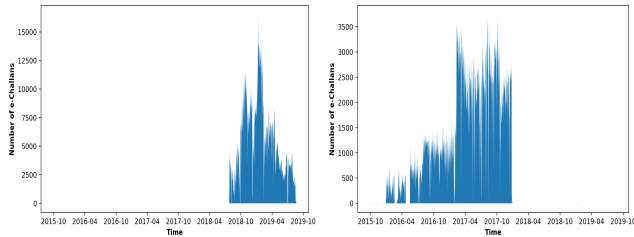


**Figure 5: Fraction of e-challans paid by people decreasing with number of violations.**

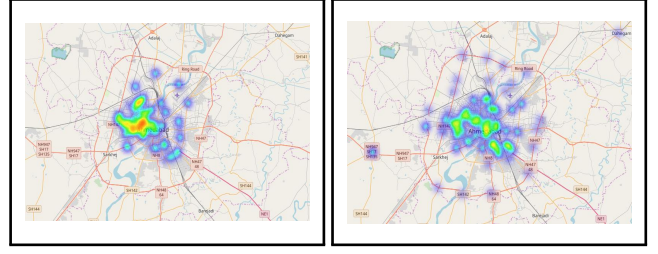
reduce the number of such violations committed as compared to general measures that target all violations equally.

We analyze the violation types from a temporal perspective and show in Figure 6 that *no helmet violations* are more dominant in the first half of the time-series, which drastically drops after 2017. On the other hand, *red light violations* are prevalent from the later half of 2018. Similar temporal bias is observed for other violation types as well, which raises questions on the efficiency of the system and motivation behind it.

We also analyze the violation types from a spatial perspective. We can see in Figure 7 that the hot-spots of different violation type varies according to the location. From Figure 7, we see that *no helmet violations* are more widely distributed across the city as compared to *red light violations* which are concentrated in few regions. Most of the red light violations in our dataset occurred in two specific areas - Navarangpura, a university region and Dariyapur, which lies near the railway station. The implication of the above analysis being, location of a violation type is a specific function of that violation type and are likely to be clustered at certain regions of the city.



**Figure 6: Timeseries plots for red light violation (left) and wearing no helmet on two-wheeler violation (right) in our dataset.**



**Figure 7: Heatmaps for red light violation (left) and wearing no helmet on two-wheeler violation (right) in our dataset.**

## 5 Characterizing Temporal and Spatial Patterns

In this section, we characterize the spatial and temporal patterns that emerge from our data. Spatial analysis is useful to understand the presence of large as well as a small cluster of locations that account for most of the traffic violations. Temporal analysis, on the other hand, allows us to analyze the general trends in the occurrence of traffic violations and when it is more prevalent.

Location	Number of e-challans
Shyamal	232,149
Paldi	218,555
Vijay Cross Roads	150,358
Shashtrinagar	149,631
Sardar Patel Statue	147,743
Total Number of e-challans	3,571,341

**Table 3: Distribution of e-challans with location.**

### 5.1 Spatial analysis

We use spatial analysis to investigate the presence of certain hotspots where traffic violations are more likely. In Table 3, we show that the top 5 locations in terms of the number of e-challans issued account for approximately 25% of all the e-challans data collected. *Shyamal* region had the most number of e-challans in our dataset and by itself accounted for approximately 7% of all the e-challans in our dataset. Thus, the data reveals that most of the traffic violations are concentrated in only a few regions of the city.

To locate the smaller regions which also account for a lot of e-challans, we plot the heatmaps of the traffic violations across the city from 2016 to 2019 collectively as shown in Figure 8. We can infer from Figure 8 (d) that traffic violations are concentrated in a few regions across the city. Also, most of the violations occurred in the central regions of the city and were more concentrated on the left side of the *Sabarmati river*. These insights can be used to take targeted intervention measures for different regions of the city. From Figure 8, we note that the e-challan system in Ahmedabad has been progressively covering more regions every year, which shows the efficiency of this system.

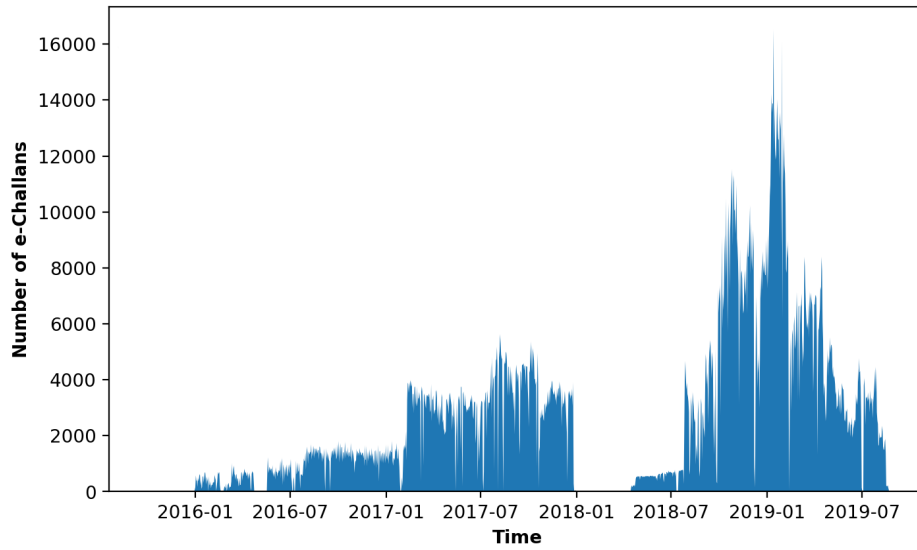
### 5.2 Temporal Analysis

Temporal analysis of the data allows us to find patterns about the occurrence of traffic violations in general. Figure 9 shows the





**Figure 8: Heat map of e-challan progression from 2016 to 2019. The figures shows e-challan distribution till year 2016, 2017, 2018, and 2019 in the sub-figures (a), (b), (c), and (d) respectively. It shows clearly how different regions emerge over time along with the increase of e-challans in some regions.**



**Figure 9: The time series plot of all the e-challans in our dataset. There is a steep increase/decrease in the number of e-challans issued during festival days.**

distribution of all the e-challans issued for 4 years. It has certain empty regions between January and April 2018 because we do not have data of that period [16]. One trend that can be discerned from Figure 9 is that the number of e-challans is continuously increasing over the years with the maximum number of e-challans issued during January 2019. We also analyze the plots to see if there are any spikes around the festival days. In general, there is a steep drop or increase in the number of e-challans issued during festival days. We observe that 2 – 3 days before Rath Yatra (Chariot Procession) - July 14, 2018, and July 4, 2019 - the number of e-challans issued is zero as the police personnel were on security duty. This suggestion is further strengthened by the fact that during a few other festivals like Muharram, Eid-ul-Fitr and Diwali, there is a dip in the number of e-challans issued. However, on some other popular festivals in Gujarat, such as Navratri, Rakshabandhan, Janmashtmi and Ganesh Chaturthi, there is a notable rise in the number of e-challans issued. The highest number of e-challans (16,500) issued on a single day in our dataset was on January 13, 2019, which is a day before Makar

Sakranti, one of the most widely celebrated festivals in Ahmedabad. The underlying trend of the high number of e-challans issued a few days before certain festivals continues during the day of the festival as well.

## 6 Predicting Recidivism

The analysis in previous sections describes in detail the presence of users who repeat offences several times. There have been several papers that have tried to predict the instances of recidivism amongst criminals, youth convicts and sex offenders[8, 14]. Similar to some previous approaches, we also identify factors strongly related to recidivism. We use machine learning to train a binary classifier that predicts recidivism amongst users. We experiment with multiple types of classification algorithms and perform detailed ablation on multiple combinations of the features. We first describe the dataset that we used to estimate recidivism, followed by the experiments on several classification models and ablation studies.

## 6.1 Recidivism dataset

To get the ground truth data for recidivism, we consider all the people who had an e-challan before April 2019 and assign them a label for recidivism. If a person received another e-challan in our dataset after April 2019, then we consider it to be a case of recidivism and hence assign it the corresponding label. In our dataset, 636,482 people had been issued an e-challan before April 2019. Of these people, 468,731 of them had committed a violation after this date.

We thus model our problem as a two-class classification problem with 468,731 positive samples (people who repeat violations) and 167,751 negative samples (people who do not repeat violations).

## 6.2 Features

We identify several features based on a person’s past violation history that could be useful for predicting recidivism. We provide a detailed description of these features and their importance below.

- (1) **Number of paid e-challans (Paid):** We compute the number of paid e-challans that each person had before April 1st and use it as a feature for our model. As discussed in Section 4, there exists the possibility of people being unaware of the e-challans being issued to them. Thus, payment of the fine amount for an e-challan implies that the user is aware of the fines being levied and would be less likely to commit more traffic violations.
- (2) **Number of unpaid e-challans (Unpaid):** The number of unpaid e-challans along with the number of paid e-challans gives us the total number of e-challans issued to a person. A person with a very high number of e-challans issued is more likely to re-offend as compared to someone who has less number of violations.
- (3) **Mean time difference between consecutive violations (Frequency):** This metric provides the frequency/regularity in which a person commits traffic violations. Let  $D_i$  and  $D_{i+1}$  represent the days on which a person received e-challan number  $i$  and  $i + 1$  respectively. Also,  $T$  is the total number of violations for a given person. We compute the mean time difference between consecutive violations ( $A$ ) as:

$$A = \frac{\sum_{i=1}^{T-1} |D_{i+1} - D_i|}{T - 1}$$

Lower values for this metric signify that a person violates traffic rules very frequently, thus increasing their likelihood of violating again. Similarly, higher values for this metric mean that the person is involved in a traffic violation in long intervals of time only, suggesting that he/she committed the violation by mistake.

- (4) **Number of days since the last e-challan (Recency):** A user may have committed a lot of traffic violations a few years ago but may have become more careful since then. Thus, our model should also be conditioned on the time in which a person last committed a traffic violation.
- (5) **Entropy of traffic violation types (Entropy):** We also compute the entropy of all traffic violation types for each person in our dataset. For each user, we first compute the set  $T$  of traffic violation types that a user committed earlier. Let  $T_i$  be the number of violations of type  $i \forall i \in T$ . Then, for

each person, we compute the entropy ( $E$ ) as:

$$E = \sum_{i \in T} \left[ \frac{T(i)}{\sum_{i \in T} T_i} \right] \left[ \log \frac{T(i)}{\sum_{i \in T} T_i} \right]$$

The entropy for each user thus measures the variation in the type of violations that each user committed in the past. A person committing a single type of violation like *Driving without helmet* is more likely to commit the same violation again as compared to someone who has different types of violation.

We use these features to predict recidivism and also conduct an ablation study on the most important features out of these. One of the main benefits of these features discussed above is that they can be easily computed for a new source of data. These features do not leverage any personally identifiable information like age, gender and religion, and are less likely to be biased by such factors.

## 6.3 Classification Techniques

We experimented with the following classifiers: Logistic Regression, Linear Support Vector Machines (SVM), Multi-Layered Perceptron (MLP), Random Forest[17] and XG-Boost [18]. We present the models and their corresponding features and hyper-parameters below:

- **Logistic Regression:** We train a logistic regression model with L2 regularization. We set the value of  $C$  (Inverse of regularization strength) to 1.
- **Linear SVM:** We train a Support Vector Machine with linear kernel and L2 regularization. We set the value of  $C$  to 1 and train it for a total of 5,000 iterations.
- **Multi-Layer Perceptron:** We train a multilayer perceptron (Neural Network) with two hidden layers consisting of 32 and 64 neurons respectively. We train the network for  $X$  epochs and use the Adam [19] optimizer.
- **Random Forest:** We train a random forest classifier on the training data with 200 base learners and a max depth of the tree fixed at 30.
- **XG-Boost:** We use XG-Boost to learn the gradient boosted decision tree models. We train it with 147 base learners, max depth of 5 for the trees and a learning rate of 0.28.

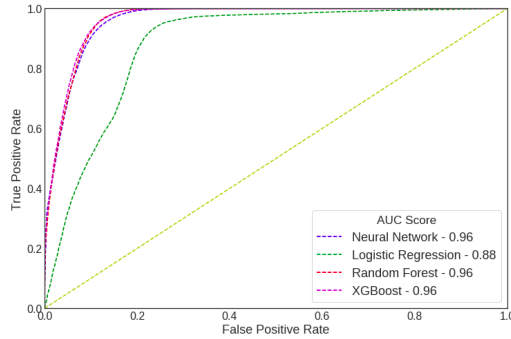
## 6.4 Analysis

To ensure that the results are not a false indication of the performance of the models, we make sure at the time of train-test split that data is properly shuffled. We use 80% of the data for training and the remaining 20% of data for testing purposes. We also report the precision, recall and F1 score apart from testing accuracy to make sure that the insights obtained are correct and report them in Table 4.

The random forest and XG-Boost classifier perform equally well on our test dataset and have a very high F1 score of 95% and 95% respectively. On the other hand, logistic regression and linear SVM classifiers perform comparatively worse. Figure 10 shows the receiver operating characteristic (ROC) curve for all the models except the linear SVM, which does not provide any probabilistic estimate. The AUC (Area Under Curve) of random forests and XG-Boost is quite high, suggesting that the classifier can separate the two classes well.

Classifier	Test Accuracy	Precision	Recall	F-1 Score
Logistic Regression	0.88	0.88	0.88	0.87
Linear SVM	0.85	0.86	0.85	0.84
MLP	0.94	0.94	0.94	0.94
Random Forest	<b>0.95</b>	<b>0.95</b>	<b>0.95</b>	<b>0.95</b>
XG-Boost	0.95	0.95	0.95	0.95

**Table 4: Performance of various classification methods on our test dataset.**



**Figure 10: AUC of different classification techniques used to predict recidivism of traffic violations. Higher AUC score denotes better classification ability of the model.**

## 6.5 Ablation Study

We perform ablation studies to find out the features that contribute most to the test accuracy and F1 score. We use random forest classifier to estimate the feature importance as it was one of the best performing models. The tree-based strategies used by random forest naturally rank by how well they improve the purity of the node. We simply remove one feature at a time from our dataset and train the model on the new set of features and compute its test accuracy. From the results of the ablation study in Table 5, we can see that the *entropy* feature has the most impact on the model, followed by *number of unpaid e-challans*. Removing the entropy feature from our training data decreases the test accuracy drastically from 95% to around 76%. The Table 5 also shows that temporal features like *time since the last e-challan* and *mean frequency of e-challans* do not have a significant impact on the test accuracy.

## 7 Discussion

We created a dataset of e-challans in the city of Ahmedabad and analyzed the data to gain some insights about traffic violations in the city. We perform spatial analysis to show that there are regions in the city with high intensity of traffic violations. The temporal analysis depicts that the number of e-challans issued show a sharp spike or drop around festival days. We show that few violations types account for most of the e-challans and a significant percentage of people are repeat offenders. This suggests that we need to take special targeted intervention measures to gauge these cases. We also show that the payment of e-challans is influenced by the fine amount and the type of violation associated with it. We describe

the features used to predict recidivism in traffic violations and train a random forest model to predict it. We also carry out detailed ablation study and show that the features related to the type of violations matter more than the temporal features and the absolute number of e-challans. The cornerstone of our work is the detailed analysis of traffic violation data to gain unique insights. These can be used by traffic police and government to make the roads safer.

Feature removed	Test Accuracy	Precision	Recall	F-1 Score
<i>Paid</i>	0.88	0.88	0.88	0.88
<i>Unpaid</i>	0.86	0.86	0.86	0.86
<i>Frequency</i>	0.91	0.92	0.91	0.91
<i>Entropy</i>	<b>0.76</b>	<b>0.74</b>	<b>0.76</b>	<b>0.74</b>
<i>Recency</i>	0.94	0.94	0.94	0.94

**Table 5: Results of the ablation study on the features. Removal of *entropy* feature affects the model worst.**

## 8 Future Work

The study is currently limited to only one city, and thus the inferences made can not be generalized to other cities across the world. We can perform a similar analysis on traffic violation data of other cities to make bolder and more general claims. The current analysis does not leverage any form of demographic data like gender or age due to lack of such data. Studies on road accident data have shown that the demographics of the people also significantly affect the fatality and the same is true for our analysis. Thus, analyzing traffic violations from the prism of demographics characteristics is another interesting line of work.

## References

- [1] WHO. The top 10 causes of death. *WHO*, 2018. URL <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>.
- [2] DK Jha, S Vibha, CB Tripathi, and G Naveen. Traffic rule violation: A weak link in prevention of road traffic accidents. *Clin Surg*. 2017; 2, 1589.
- [3] Business Standard. Delhi traffic police launch e-challan and e-payment system. *Business Standard*, 2019. URL [https://www.business-standard.com/article/pti-stories/delhi-traffic-police-launch-e-challan-and-e-payment-system-119071901559\\_1.html](https://www.business-standard.com/article/pti-stories/delhi-traffic-police-launch-e-challan-and-e-payment-system-119071901559_1.html).
- [4] Mihir Ved. Don't cross that stop line. you'll be 'shot at sight'. *Ahmedabad Mirror*, 2019.
- [5] Jayvant Choudhary, Anurag Ohri, and Brind Kumar. Identification of road accidents hot spots in varanasi using qgis. *Organized By Department of Civil Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi-221005 Uttar Pradesh, India*, page 7, 2015.
- [6] Honglei Ren, You Song, Jingwen Wang, Yucheng Hu, and Jinzhi Lei. A deep learning approach to the citywide traffic accident risk prediction. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3346–3351. IEEE, 2018.
- [7] Sarfaraz Shaikh. Ahmedabad: Handy scanners to net e-challan jumpers. *The Times of India*, 2019.
- [8] R Karl Hanson and Monique T Bussiere. Predicting relapse: a meta-analysis of sexual offender recidivism studies. *Journal of consulting and clinical psychology*, 66(2):348, 1998.
- [9] Grant T Harris, Marnie E Rice, and Catherine A Cormier. Psychopathy and violent recidivism. *Law and human behavior*, 15(6):625–637, 1991.
- [10] WHO. Road traffic injuries. *WHO*, 2018. URL <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.
- [11] Sanjay Kumar Singh. Road traffic accidents in india: issues and challenges. *Transportation research procedia*, 25:4708–4719, 2017.
- [12] Qiqi Wang, Wei Zhang, Rendong Yang, Yuanxiu Huang, Lin Zhang, Peishan Ning, Xunjie Cheng, David C Schwebel, Guoqing Hu, and Hongyan Yao. Common traffic violations of bus drivers in urban china: An observational study. *PLoS one*, 10(9):e0137954, 2015.



- [13] Peter Schmidt and Ann Dryden Witte. Predicting criminal recidivism using ‘split population’ survival time models. *Journal of Econometrics*, 40(1):141–159, 1989.
- [14] Nancy Ritter. Predicting recidivism risk: New tool in philadelphia shows great promise. *National Institute of Justice Journal*, 271(February):4–13, 2013.
- [15] Gerald J Stahler, Jeremy Mennis, Steven Belenko, Wayne N Welsh, Matthew L Hiller, and Gary Zajac. Predicting recidivism for released state prison offenders: Examining the influence of individual and neighborhood characteristics and spatial contagion on the likelihood of reincarceration. *Criminal justice and behavior*, 40(6):690–711, 2013.
- [16] Times of India. Traffic cops to expand ambit of e-challans in gujarat. *Times of India*, 2019. URL <https://timesofindia.indiatimes.com/city/ahmedabad/traffic-cops-to-expand-ambit-of-e-challans/articleshow/70624464.cms>.
- [17] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [18] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.
- [19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.