Fall 2022

Traffic Accidents Analysis & Prediction in UAE

Majid Alhosani

Traffic Accidents Analysis & Prediction in UAE

by

Majid Alhosani

A Capstone Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

Fall 2022

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Majid Alhosani**

Graduate Capstone Title**: Traffic Accidents Analysis & Prediction in UAE**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr.Ioannis Karamitsos **Date:**

Member of committee

Acknowledgments

First of all, I would like to praise Allah, the most compassionate and Merciful.

I would like to acknowledge and give my warmest thanks to my mentor Dr. Ioannis Karamitsos who made this work possible. His guidance, patience and advice carried me through all the stages of writing my capstone project. I would also like to thank Dr. Sanjay Modak, Chair of Graduate Programs and Research at RIT Dubai, for his suggestions and effective comments during the program.

Moreover, I would also like to give special thanks to my wife, family and coworkers as a whole for their continuous support and understanding when undertaking my research and writing my capstone project. Your prayer for me was what sustained me this far.

Finally, I would like to extend my thanks to RIT university and all the teaching faculty for their continuous support.

Abstract

UAE has state-of-the-art roads and traffic infrastructure, and yet there has been a rapid increase in the number of traffic accidents. As per National Agenda 2021, the UAE traffic departments are working to minimize the rate of road traffic death per 100 thousand people. The Traffic and Patrols Directorate in Abu Dhabi launched a road safety management plan. This aims at decreasing the fatalities from traffic accidents to 3 per 100,000 inhabitants by 2021. According to police, distracted driving, sudden swerving, entering a road without ensuring that it is clear, tailgating and speeding without considering the road conditions has caused serious accidents so far. The disturbing figure has led to the amendment of the federal traffic law which now imposes hefty penalties. This study aims to identify the trends and evaluate the information available on possible causes.

In the due course of the report, I would like to understand these trends around road accidents and their inherent causes. This would help the government bolster road safety measures in order to reduce the number or even avoid accidents on the road. There has been much research work done on the above to determine causes and driving factors of accidents on the road.

Keywords: Road Safety, Machine Learning, Data Analysis, Traffic Accidents

Table of Contents

Chapter 1 - Introduction

1.1 Introduction

Traffic accident is defined as an accident that occurs unintentionally between one or more vehicles. These are classified according to severity into two categories: report accidents and accidents which happen to go to the court. Traffic accidents could include collision of moving vehicles, hitting pedestrians, overturning of vehicles, hitting stationary objects and animals, and also passengers falling from a moving vehicle. The cases that go to the court include accidents which result in fatalities and serious damage to property. These also include accidents caused by drunk driving and driving without a valid license. Over 1.2 million lives are lost and over 50 million injuries are sustained annually due to road accidents worldwide. The Gulf Cooperation Council (GCC) region has been particularly impacted by the large number of road accidents and injuries. Although the UAE has shown a declining road-injury and fatality records, yet road accidents continue to take an economic toll on the country. UAE's amended federal traffic law came into effect on 1st of July 2017 that aims to further protect the lives of road users and reduce traffic casualties from about 6 per 100,000 people to 3 per 100,000. In an attempt to prevent road accidents and reduce its impact on lives and properties, different government entities in the UAE have launched initiatives and awareness campaigns to lessen the impact of traffic fatalities. Some of these initiatives are: Abu Dhabi strategic traffic safety plan, Road safety audit, road safety awareness, Central traffic control system, Speed management strategy.

1.2 Project goals

The scope of my project involves collecting a dataset from a reliable data source. In the real-world case, the data would be collected in databases and repositories which can later be used for analyses and insights generation. When it comes to traffic accidents, there will be different data sources and data collection like vehicle registration data, traffic signal datasets, user reports and emergency reports with respect to vehicle accidents. Now, in my case, I would be using the dataset from a government website and I plan to use the dataset to perform exploratory data analysis to understand various accident trends and some of the most accident-prone cities and neighborhoods. Once I have a clear understanding of the road accidents, I could use the information to build models for forecasting/predicting future traffic accidents. This would involve creating machine learning models like Regression, Decision Trees or Random Forest and some other classifiers or predictors.

1.3 Research Questions

Through the course of the research, I identified the different driving factors and patterns of road accidents in the UAE. By leveraging data, I identified the main points and the underlying insights behind the problem statement. Some of the questions that I can answer using data are -

- What are the age groups, gender, education level or some other demographics information of the drivers involved in road accidents?
- How severe are the reported accidents, and how is the severity related to other factors in the dataset?
- Which type of vehicles are involved in the accidents and how severe are they?

Some of the above questions will help me understand various drivers of road accidents and the severity of these accidents. I also studied numerous papers and related research to understand if there is any pattern behind accidents in the UAE, and understand what solutions have been provided by other researchers in the same problem domain.

Chapter 2 - Literature Review

2.1 Literature Review

As per a study conducted by Hassan (2012), awareness, education and training, infrastructure, vehicle and law enforcement are some of the main factors contributing to traffic safety in the country. The 18-35 age group turned out to be the highest rate of traffic fatalities and injuries. The less than 15 age group were considered to be the next highest and the elderly had the lowest traffic injury rate. The traffic departments in Abu Dhabi play an important role in reducing the numbers of traffic fatalities and injuries. According to the manager of traffic departments in Abu Dhabi, the present traffic campaigns - which involve using speed cameras - are crucial for reducing traffic injuries and fatalities.

In an article published by BMC (2020), the per capita income of UAE is higher than many other industrialized countries and has strict laws with regards to alcohol use, strict safety vehicle licensing and high quality of roads. Although the fatality rate has fallen over the years, it continues to be higher than countries with similar socio-demographic indicators. One-quarter of its population belongs to the age group of younger than 25 years old and in fact the leading cause of death in this age group is road injuries. Hence, prevention and mitigation of road traffic accidents for younger age groups needs to be prioritized. The youth are not restrained and have more risky driving practices, driving with no license, drug/alcohol use and distracted driving and needs to be monitored more than anybody else.

According to a research study conducted by Francisco and Dina (2019), there have been some improvements in the road safety field in the region of UAE. It has been found that over 1.2 million road traffic accidents have occurred in the Emirate of Abu Dhabi (which covers around 87% of the UAE) between 2012 and 2017 and approximately 99% of these accidents did not result in an injury. The study also found that side-impact and rear-end collisions were the most commonly observed crash types, while pedestrian-related collisions were significantly more violent encounters. Pedestrian collisions accounted for only 0.23 percent of all reported crashes; yet, they accounted for one-quarter of all fatal crashes. The majority of the reasons for crashes include tailgating, reversing, and reckless driving. However, running on the red light and drunk driving are far more violent crashes as compared to the rest. 65% of accidents happened on roads with speed limits of up to 40 kph and only 5% of the accidents happened on roads with speed limits of 100-120 kph. It is interesting to note that a small percentage of drivers at fault were minors who are less than 18 years.

According to a study conducted by Akmal and Saif (2015), it has been revealed that there is a decrease in the total number of accidents between 2008 and 2009 which can be attributed to the new traffic law. The new traffic law allows the police to withdraw a driving license if the drivers are involved in serious traffic violations. According to the new system, a driving license is withdrawn if the driver obtains 24 traffic points within a 12-month period. These violations include drunk driving, driving a vehicle without a license plate, not stopping after causing an accident with injuries, and dangerous overtaking. While some other violations result in a penalty of 12 points such as "racing", reckless driving, speeding above 60 km/h and running away from a policeman.

The number of road crashes in the Gulf Cooperation Council (GCC) countries have been devastating for the nation and the total research conducted around this issue has not been too prevalent in the area. Due to this reason, decision makers have not been able to understand the problems of road accidents and have been able to mitigate the same through effective solutions. In their paper, Dina (2021) have implemented multivariate logistic regression to determine the factors that act as the contributing factors to road accidents in the Emirate of Abu Dhabi. The data has been collected between the time frame 2012 and 2017, and since this data has been collected within the GCC, it can be used for other countries like Saudi Arabia, Qatar, Bahrain etc. The traffic guidelines and road setups are pretty similar in the GCC countries and hence the results from one country can be replicated to the other regions with ease. Based on the entire study, there were certain recommendations that the research team had including law enforcement especially on weekends, punishments for red-light skipping, road and transport design aligned towards pedestrian safety and many others. Some of the above measures would enable reckless drivers and drivers breaking traffic rules to be dealt with punishment or severe fines, so that the number of such occurrences are brought down.

In another research work by Omar (2018) and his team determined that even with the increase in road safety measures and traffic rules, accidents have been constant in the Emirate of Abu Dhabi. These road accidents account to almost 5,500 people being killed in the UAE in the course of 6 years, which accounts to about two people every day. This brings Abu Dhabi to the spotlight in terms of road accidents being quite prevalent in the region. Researchers have been able to prove the fact that 80% of accidents are due to human negligence. In his research proposal, the team has developed a telematics system which can be used to reduce fatalities and severe injuries from road accidents. This would help in improving the road safety and better life on the roads in Abu Dhabi. Some of the services of the telematics system monitor real time driver behavior to report irrational or anomalous driving behavior, which may result in prevention of severe accidents down the road if caught before.

Due to the plaguing of the Gulf Cooperation Council (GCC) countries, there have been numerous researches carried around in this area to mitigate the same in the future. The research carried out by Albuquerque

(2020) is not only relevant to the UAE but also other GCC countries given the fact that there are many similarities between these countries in terms of road design, vehicle fleet and the driving culture. The study identified 1.26 million road accidents, 9,327 injuries of which 1,305 were fatal during the given time period. They also identified an interesting insight from the research that the accidents that happened between 22:00 and 5:59 produced the most fatal injuries, which happened during adverse weather situations involving drunk drivers and the collisions happening at high speed past the speed limits of the highway. Some of the main causes of the accidents occurred due to tailgating and reckless driving which constituted the 50% of the accident reports and of all, more than half of the people involved in an injury did not wear any seat belts. There were also many recommendations provided by the researchers around improving road safety in the Emirate of Abu Dhabi.

In research by Hamad (2016), the problem of road traffic accidents is elaborated which particularly is prominent in the Emirate of Sharjah which is the third largest emirate in UAE. This study was carried out with data between 2001 and 2014 and the researchers had several findings which were discussed in the due course of the paper. It was identified that Sharjah had done something better than the other Emirates combined due to which the annual number of road accidents decreased by more than half, when there was an increase of the Sharjah population by two times. It was also observed that the accidents and injuries per 100,000 population decreased annually, while the fatality per 100,000 population decreased only marginally. This was observed right after the new traffic law of UAE was introduced in 2008. To summarize, Sharjah had seen a lower fatality rate per 100,000 population as compared to the other Emirates or the whole of UAE combined. Hence, the researchers wanted the paper as a reference to be used to improve traffic rules and guidelines in the other regions. It has been observed from the above researchers that while one Emirate faces challenges in this problem area, another one thrives due to better reforms and regulation. It is very crucial for one to learn from the other so that the Emirates can thrive as a whole to curb these problems in the longer run.

One of the key factors for road accidents is human drivers due to which it is very important to understand the characteristics of drivers involved in the road accidents in UAE. Alkheder (2013) and his team used a dataset from UAE traffic accidents between 2007 and 2010 which was used to determine the relationship between traffic accidents and the citizenship of the drivers for the different types of vehicles involved in the accidents. Due to the citizenship of the drivers being different, special care can be given to these nationality drivers to prevent road accidents. It was observed in the study that the majority of the accidents belonged to the Emirati citizens (30%), while second in line are the Pakistanis (21%) and the rest are the Indian drivers (11%). One of the key points in the research is the involvement of the dataset analysis which was analyzed after the new traffic law introduction in UAE.

2.2 Takeaways from Literature Review

Based on numerous articles and research works, it is evident that traffic road accidents are a daunting problem to be solved. These research works allow me to have an in-depth understanding of the subject before moving ahead with my own analyses and projects. Literature Survey allows me to get a better understanding of the subject matter before even moving ahead with the actual work. Below are some of the key takeaways based on the readings and numerous research articles -

- Some of the most key problems learning to road accidents is traffic safety awareness, vehicle and law enforcement as well as lack of safety measures and infrastructure
- As most of the accidents are road accidents prevalent in the age group ~25, it is imperative that these age group drivers are given adequate training before handing out licenses
- Moreover, during recent times the road accidents have reduced and the ones that do happen are mostly non-fatal. This is conveyed as per recent studies and indicates that road accidents are reducing and safety measures are improving

Chapter 3 - Research Methodology

3.1 Methodology

CRISP-DM methodology will be selected for this capstone project. This is one of the most popular frameworks used for solving data science problem statements. It follows a cycle of steps which are done in sequence to maintain the flow of the end outcomes from the entire solution approach.

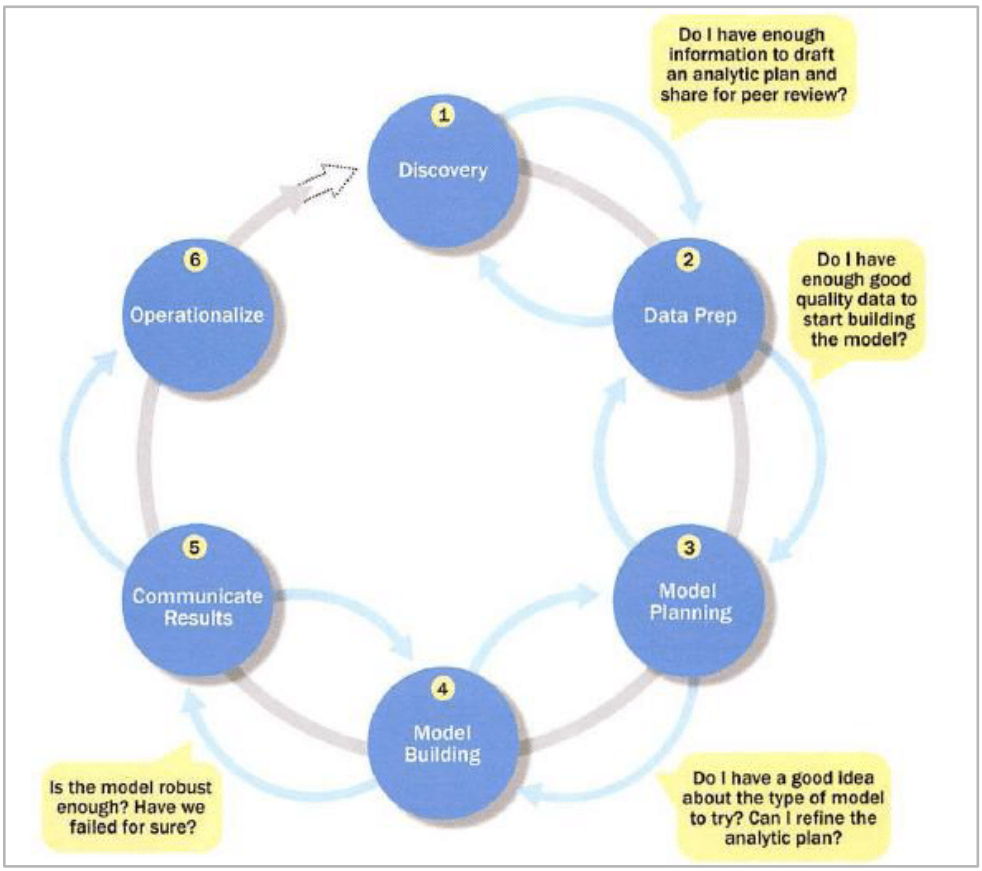


Fig.2.

As given on the image above, a typical data analytics lifecycle contains six different steps of progression, starting from data discovery. Once I determine the dataset and understand it, I can use it for the provided case/scenario. In the next step of the phase is the data preparation step wherein the data is prepared as per my use case. The dataset is cleaned and manipulated for appropriate context of the use case. For example, I would need to clean the missing values and remove other data inconsistency within the dataset. Once that is done, I might need to one-hot encode some of the categorical features to make it simple for the model to

understand. The next phase of the problem is the model planning, which entails choosing the right model as per my requirement. After that is done, the dataset is then split to train and test set to be able to train the model. The trained model is then fed with the test or valuation dataset to understand the accuracy of the model. The final steps are communicating the results to the stakeholders and operationalizing the model for production use. Production deployment of the model allows different teams and stakeholders to use the results and the model independently.

3.3 Dataset Description

In this project a traffic accidents dataset has been selected from Kaggle, which contains all the details about when an accident happened and what were the victims' details. This dataset helped me understand accident trends which I then used to create a forecasting model to determine future accidents. The dataset contains the following 23 attributes and 12316 rows in total and it is used to analyze and fit a forecasting model on. These different attributes define the accident for someone at that time and day, and what traits did the victim have with respect to their age, sex, etc.

- Time
- Day of week
- Age band of driver
- Sex of driver
- Educational level
- Vehicle driver relation
- Driving experience
- Owner of vehicle
- Area accident occurred
- Lanes or Medians
- Road alignment
- Types of Junctions
- Road surface type
- Road surface conditions
- Light conditions
- Weather conditions
- Type of collision

- Number of vehicles involved
- Number of casualties
- Vehicle movement
- Pedestrian movement
- Cause of accident
- Accident severity

I cleaned the dataset and made the dataset consistent to be able to fit it into a forecasting model. This involved checking for missing values, datatype issues as well as data inconsistencies in each attribute throughout the data. Once the data is clean, it is easier for the model to interpret all the values and be able to forecast accurately.

For this project, the R/Python programming languages have been used for the following tasks:

- Data manipulation and preparation
- Data analysis and visualization
- Machine Learning and/or forecasting

Chapter 4 - Data Analysis

4.1 Data Preparation

In this phase, the process of procuring and processing the dataset is performed so that the dataset is usable for EDA and modeling in the subsequent steps. Moreover, the dataset has been explored briefly to indicate the features that will be dealt with.

	Time	Day_of_week	Age_band_of_driver	Sex_of_driver	Educational_level	Vehicle_driver_relation	Driving_experience
0	17:02:00	Monday	18-30	Male	Above high school	Employee	1-2yr
1	17:02:00	Monday	31-50	Male	Junior high school	Employee	Above 10yr
2	17:02:00	Monday	18-30	Male	Junior high school	Employee	1-2yr
3	1:06:00	Sunday	18-30	Male	Junior high school	Employee	5-10yr
4	1:06:00	Sunday	18-30	Male	Junior high school	Employee	2-5yr

Type_of_vehicle	Owner_of_vehicle	Service_year_of_vehicle	Defect_of_vehicle	Area_accident_occured	Lanes_or_Medians
Automobile	Owner	Above 10yr	No defect	Residential areas	NaN
Public (> 45 seats)	Owner	5-10yrs	No defect	Office areas	Undivided Two way
Lorry (41?100Q)	Owner	NaN	No defect	Recreational areas	other
Public (> 45 seats)	Governmental	NaN	No defect	Office areas	other
NaN	Owner	5-10yrs	No defect	Industrial areas	other

Road_allignment	Types_of_Junction	Road_surface_type	Road_surface_conditions	Light_conditions	Weather_conditions
Tangent road with flat terrain	No junction	Asphalt roads	Dry	Daylight	Normal
Tangent road with flat terrain	No junction	Asphalt roads	Dry	Daylight	Normal
NaN	No junction	Asphalt roads	Dry	Daylight	Normal
Tangent road with mild grade and flat terrain	Y Shape	Earth roads	Dry	Darkness - lights lit	Normal
Tangent road with flat terrain	Y Shape	Asphalt roads	Dry	Darkness - lights lit	Normal

Type_of_collision	Number_of_vehicles_involved	Number_of_casualties	Vehicle_movement	Casualty_class	Sex_of_casualty
Collision with roadside-parked vehicles	2	2	Going straight	na	na
Vehicle with vehicle collision	2	2	Going straight	na	na
Collision with roadside objects	2	2	Going straight	Driver or rider	Male
Vehicle with vehicle collision	2	2	Going straight	Pedestrian	Female
Vehicle with vehicle collision	2	2	Going straight	na	na

Age_band_of_casualty	Casualty_severity	Work_of_casuality	Fitness_of_casuality	Pedestrian_movement	Cause_of_accident
na	na	NaN	NaN	Not a Pedestrian	Moving Backward
na	na	NaN	NaN	Not a Pedestrian	Overtaking
31-50	3	Driver	NaN	Not a Pedestrian	Changing lane to the left
18-30	3	Driver	Normal	Not a Pedestrian	Changing lane to the right
na	na	NaN	NaN	Not a Pedestrian	Overtaking

In the below section, the number of missing values for all the features has been checked, and it has been observed that the highest missing values in the feature ‘Defect of vehicle’, followed by ‘Service year of vehicle’.

Time 0	Light_conditions 0
Day_of_week 0	Weather_conditions 0
Age_band_of_driver 0	Type_of_collision 155
Sex_of_driver 0	Number_of_vehicles_involved 0
Educational_level 741	Number_of_casualties 0
Vehicle_driver_relation 579	Vehicle_movement 308
Driving_experience 829	Casualty_class 0
Type_of_vehicle 950	Sex_of_casualty 0
Owner_of_vehicle 482	Age_band_of_casualty 0
Service_year_of_vehicle 3928	Casualty_severity 0
Defect_of_vehicle 4427	Work_of_casuality 3198
Area_accident_occured 239	Fitness_of_casuality 2635
Lanes_or_Medians 385	Pedestrian_movement 0
Road_allignment 142	Cause_of_accident 0
Types_of_Junction 887	Accident_severity 0
Road_surface_type 172	
Road_surface_conditions 0	

The above missing value treatment can be done using missing value imputation (mode imputation) for all features. Since the features are categorical in nature, mode imputation should fix the problem of missing values.

4.2 Exploratory Data Analysis

In this section of the report, the data has been leveraged and python was used to shape the dataset. After the data is processed and ready for use, various visualization techniques have been used such as univariate and bivariate analysis using different packages. This will help in understanding some of the core aspects of the dataset along with some understanding like data distribution, relationship between different features in the data as well as many other details. After determining an understanding of the dataset, the information and insights are used to move on to the next phase of the analysis, i.e., modeling.

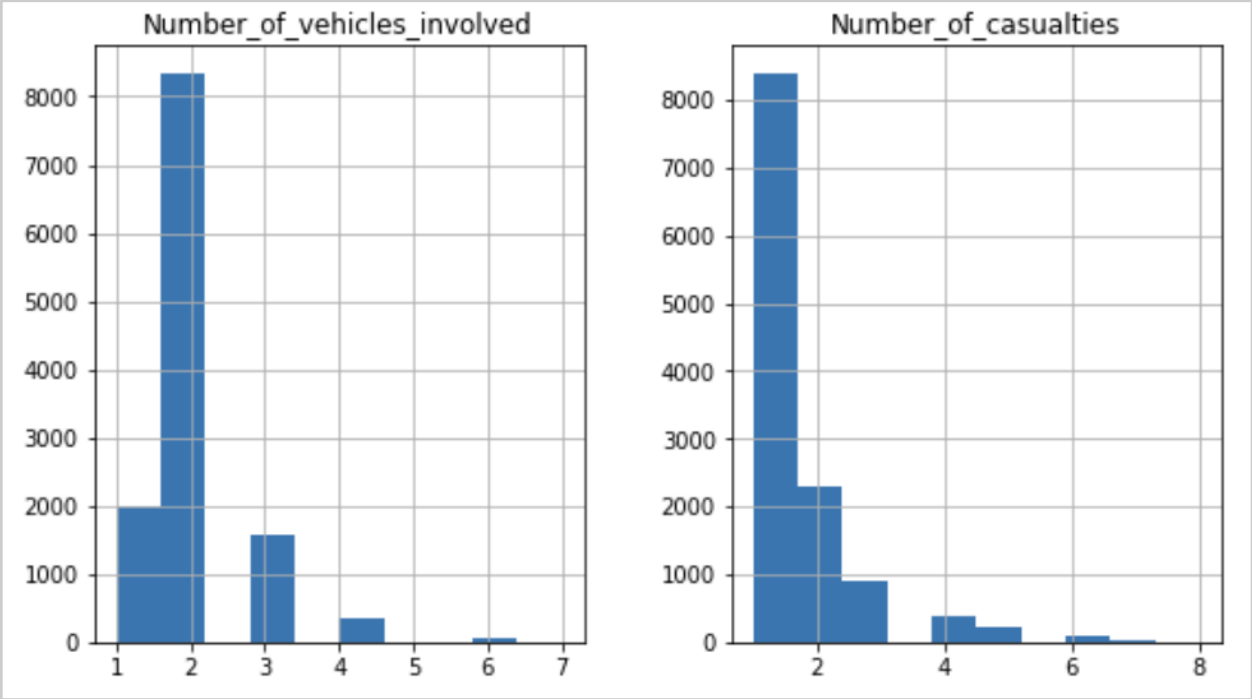


Fig.3. Vehicles vs causalities

The above chart is a representation of the data points with respect to the number of vehicles involved in an accident and the number of casualties in a particular accident. Figure 3 shows that the greatest number of vehicles involved in an accident are either 1 or 2. As for the number of casualties, it has been observed that the histogram is right skewed with a long tail on the right end. Most accidents are within the 1-2 range of the x-axis indicating that the number of people involved in an accident are 1 or 2 mostly.

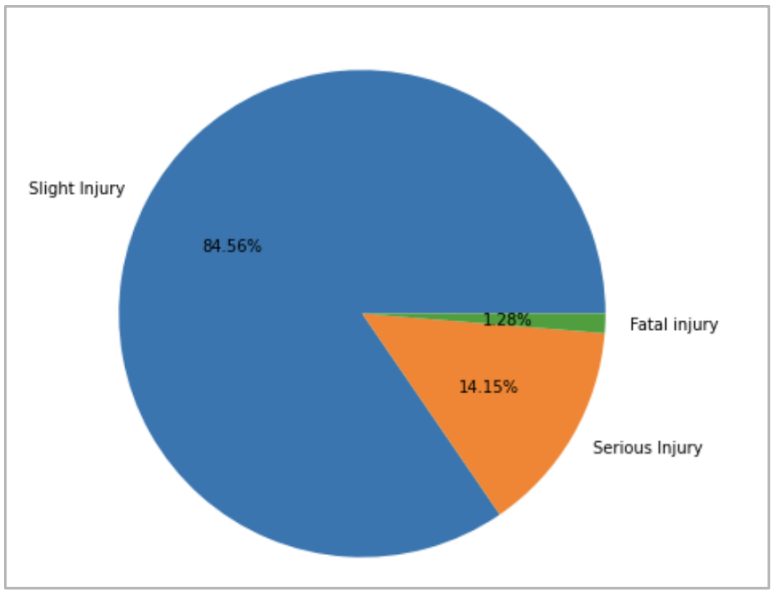


Fig.4 type of accidents

Figure 4 shows the distribution of the type of accidents in which 84.5% of the accidents resulted in minor injuries, the remaining percentage of ~15% are either serious injuries or fatal injuries resulting from the accidents. This feature distribution is understood from the above representation of the dataset and can be leveraged in the subsequent steps of my reporting and modeling.

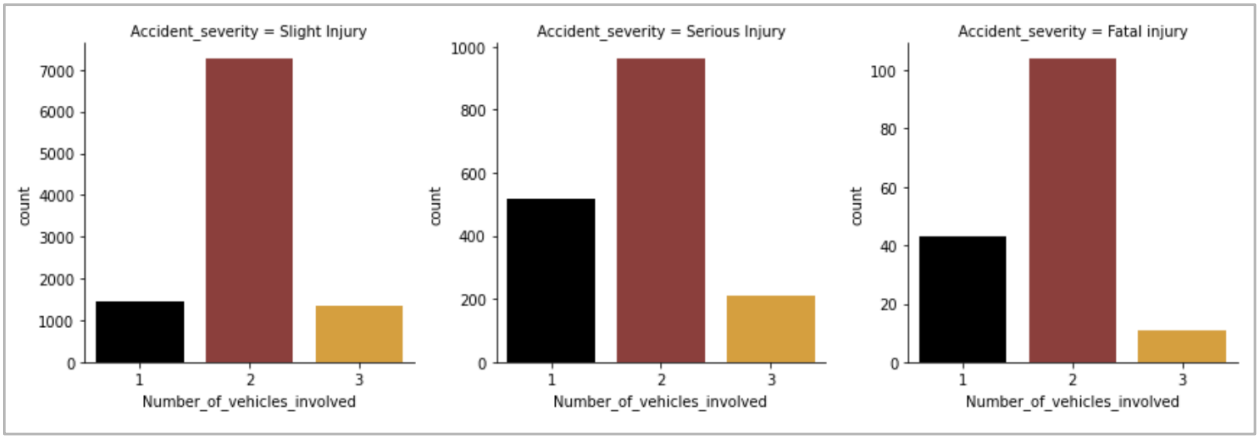
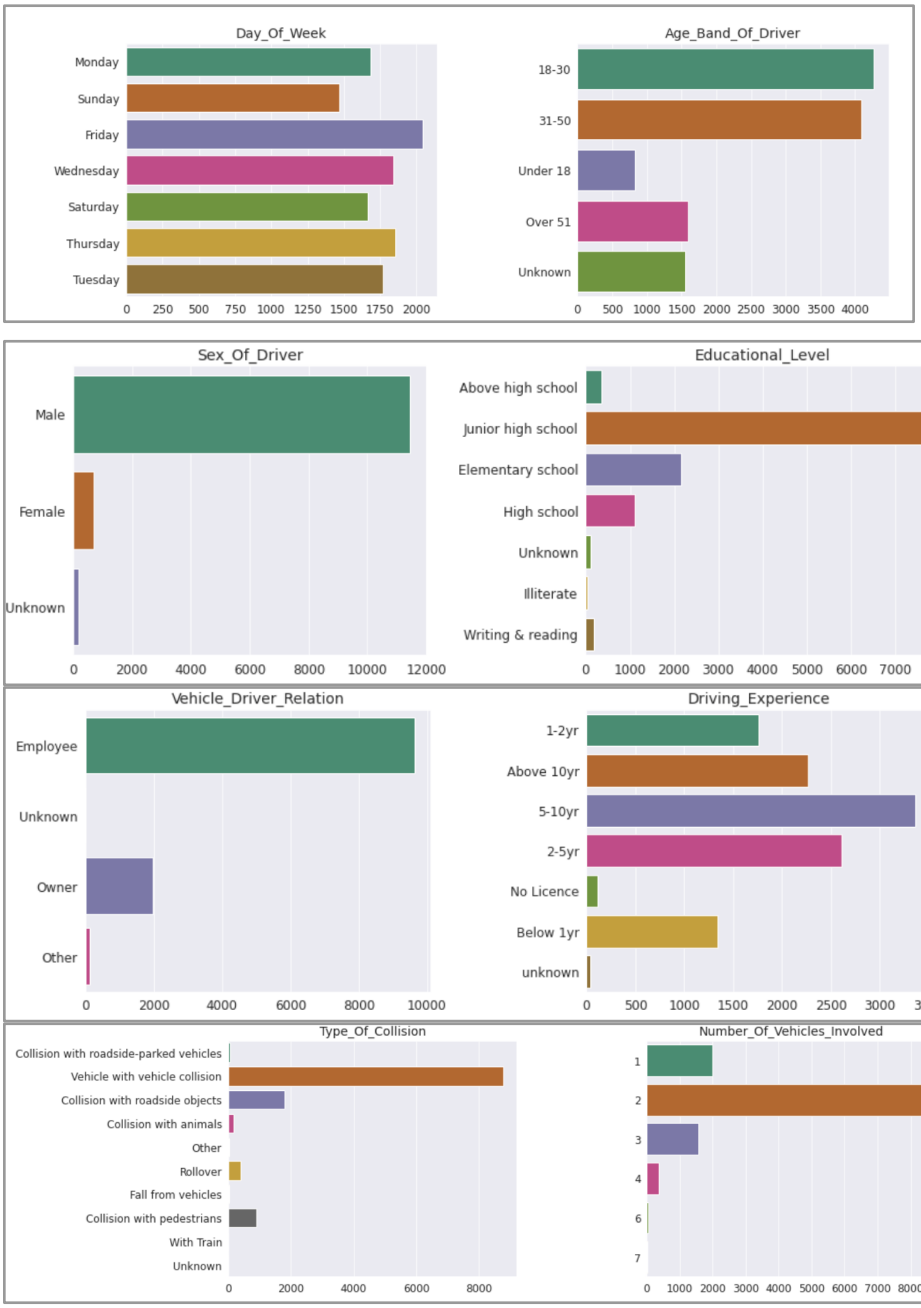


Fig.5.vehicle count vs severity

Figure 5 shows the vehicle's involvement counts along with the severity of the accidents. It is seen that for serious or fatal injury related accidents the number of vehicles involved are 2, which is the case for all types of accidents. But in some cases, a single vehicle has resulted in either serious or fatal injuries more than lower intensity injuries.

In the next section, the total accidents for different traits of the dataset like day of week, sex of driver, age group etc. has been plotted, the total accidents have been plotted with respect to the different

features. This process of visualizing the dataset with different features in the data is generally termed as univariate or multivariate analysis which helps in understanding the distribution of the feature in the data. For example, if it is required to find which gender is more involved in accidents compared to the other, it can be done through visualization or univariate/multivariate analysis.



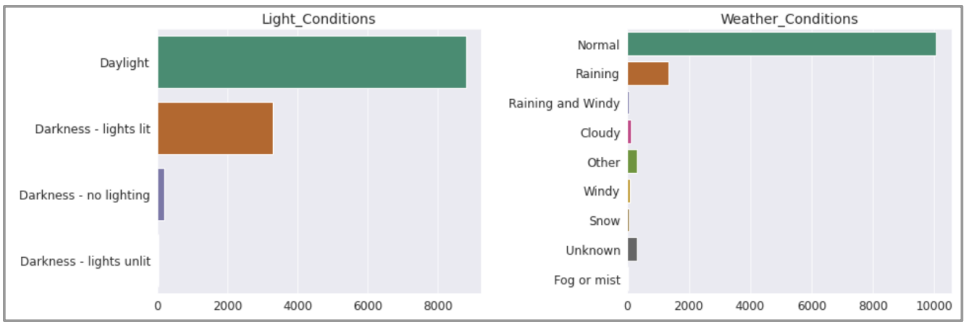


Fig.6.

In the above sequence of plots, the following findings have been derived -

- Fridays and Wednesdays see comparatively a greater number of accidents than other days of the week
- 18-30 and 31-50 age group has observed relatively higher number of accidents compared to the other age groups
- Males have been observed to be involved in more accidents than the other gender, and junior high school
- It is also observed that the vehicle driver relationship with employee were the most involved in accidents with the driver experience with 5-10 years being involved in the most accidents
- Coming to the type of collision of people, vehicle with vehicle collision is the most prominent and the number of vehicles involved is mostly 2
- Most of the accidents happen during daylight with the weather condition being mostly normal followed by some during rains

As shown in Figure 7, the an. So, it is imperative to tighten traffic rules during these hours and road safety and emergency response teams can be better prepared for faster response during these times of the day to mitigate the same.

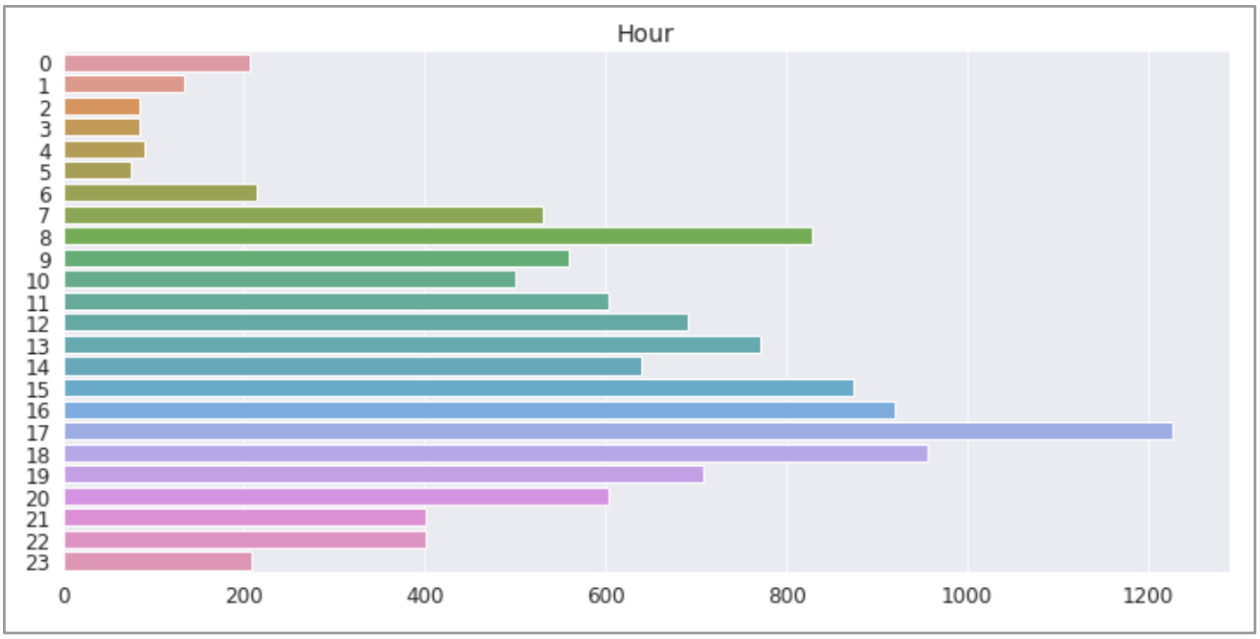


Fig.7.Accidents vs hour

Based on the above findings and insights, a lot more about the accident patterns and the person details who are most frequently involved in accidents is understood. This helps in using the dataset better for the modeling phase of the experiment.

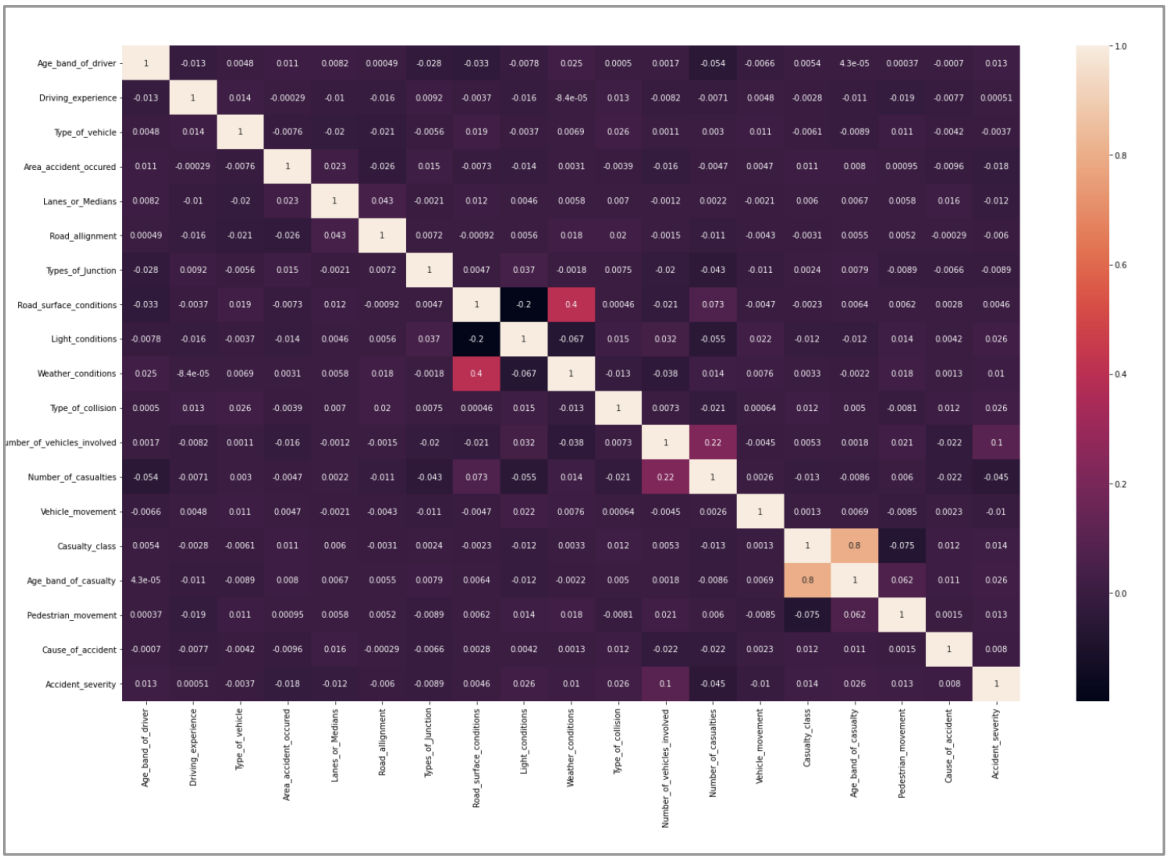


Fig.8. Correlation plot

A correlation plot is generated for all features to understand the relationship between all the features. Typical correlation values are between 1 and -1. Values close to 1 indicate high correlation or positive correlation while values towards -1 indicate negative correlation between the features. Based on the functionality and interpretation of a correlation plot, the features Casualty Class and Age Group of Causality have high positive correlation. This information helps in the subsequent steps as to which features impact each other with the increase or decrease in value, or if they are linear dependent.

4.3 Model Description

For the ease of understanding of the readers, different machine learning models has been explained, which will be used in the subsequent sections. The different models are used to train and predict on a test set of the data. This will provide accurate scores with which are used to determine the best fit model for the exercise. The purpose is to determine the severity of an accident through predictive analytics.

- Gradient Boosting. This is a machine learning technique which involves the use of regression and classification and forms collections of decision trees which is an ensemble. From these collections, the best tree is picked to derive the results in turn
- Random Forest Classifier. Random trees are an ensemble of many decision trees and this is a classification and prediction method that is developed on both classification and regression tree methods. (Pierre, 2006) This training method of the model uses partitioning which is recursive in manner and splits the records with similar output values into segments
- Decision Tree Classifier. A decision tree model works by recursive partitioning to classify whether the accident type was highly severe, moderate or mild based on past trends. The model can be trained with different cross validations, complexity parameter controls how deep the tree grows, a small value allows for the splitting of even smaller nodes which doesn't improve prediction fit by a significant amount.
- Extra Trees Classifier. This modeling tree uses a class and a meta estimator that fits a number of randomized decision trees on different sub-samples of the dataset. It uses the method of averaging to improve the predictive accuracy to also control the problem of over-fitting
- Logistic Regression. A logistic regression model was built to predict whether the severity of an accident was high, medium or low. The model assumes a linear relationship between the dependent variables and logs the odds of default/charge off. (Michael L., 2008) The prediction equation is:

$$\text{Logit} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

For the above listed models, I will derive the accuracy for each one of them and then identify the best model based on accuracy. This process will help me identify the best model for my use case in solving the problem statement, wherein I want to determine the severity of an accident preemptively through predictive analysis.

4.4 Modeling

In this section of the study, several machine learning models were used to compare and understand the predictive capabilities of the models. For the implementation of these models different Python programming packages are imported. Before implementing the modeling, a split ratio of 70:30 has selected for the training and testing dataset respectively, and the 'Accident Severity' feature is considered as the target variable for the models. Ordinal encoding of the categorical features was performed for the model to understand the data better. Encoding helps in streamlining the process of model interpretability during the subsequent phases.

In the below section, the train and test dataset has been split based on the predictor variable 'Accident Severity'. First off, a Random Forest modeling was used to determine the factor importance of all the features impacting the target variable.

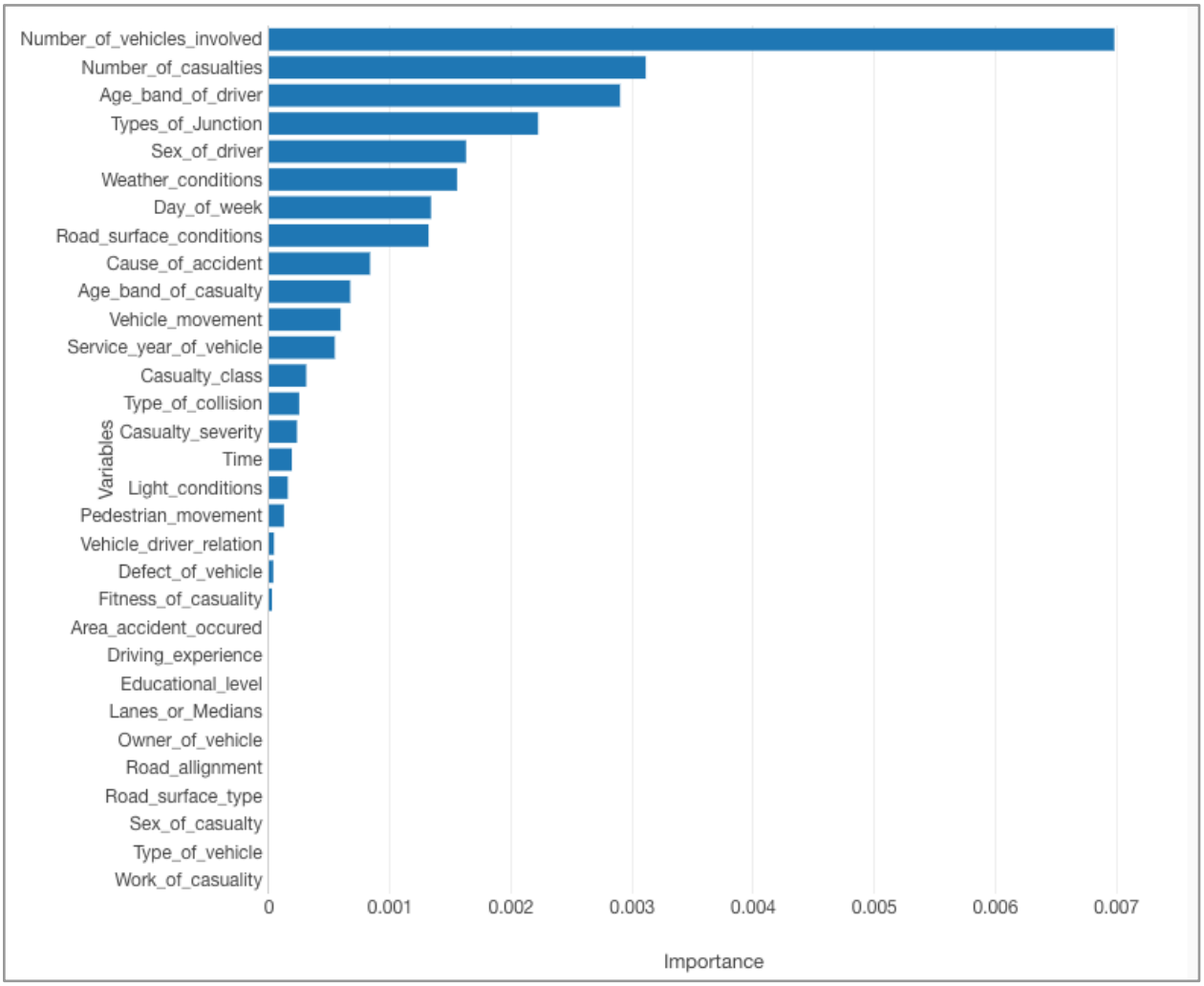


Fig 9. Factor significance derived from Random Forest

In Fig.9 the Number of Vehicles involved, number of casualties, age band of the driver and types of junctions play a key role in determining the severity of the accident. The above factor importance scores help in understanding the different features that play an important role in affecting the target variable at the end.

```
y_test = ordinal_encoder(pd.DataFrame(y_test, columns = ['Accident_severity']), pd.DataFrame(y
_test, columns = ['Accident_severity']).columns)['Accident_severity']
y_train = ordinal_encoder(pd.DataFrame(y_train, columns = ['Accident_severity']), pd.DataFrame
(y_train, columns = ['Accident_severity']).columns)['Accident_severity']
```

In the next sections, the confusion matrix for different ML were presented.

```
gbc = GradientBoostingClassifier(random_state = 0, learning_rate=0.45)
rfc = RandomForestClassifier(random_state = 0)
lr = LogisticRegression(random_state = 0)
dtc = DecisionTreeClassifier(random_state = 0)
svc = SVC(random_state = 0)
extree = ExtraTreesClassifier()


lst = []

for i in(gbc,rfc,dtc,lr,svc,extree):
    i.fit(X_train, y_train)

    i_pred = i.predict(X_test)

    cm = confusion_matrix(y_test, i_pred)

    cr = classification_report(y_test, i_pred)
    i_acc = round(accuracy_score(y_test, i_pred), 4)

    lst.append(i_acc)

    print(i,':\n','The confusion matrix:\n',cm,'\n')

    print('The classification report:\n',cr,'\n')

    print('-'*60)
```

The above section implements various models like Gradient Boosting, Random Forest, Logistic Regression, Decision Tree, SVC and Extra Trees which are first initialized on the top. The following section is a loop built to fit all these models one by one and then the corresponding scores for accuracy and the classification report is plotted. Using the for loop instead of manual runs is an ideal way of modularization of codes and

best practices for data analysis reports. Functions and codes should be reusable and hence the above steps were used to perform the modeling process. Based on the modeling steps, it is required to evaluate the performance of the same through the above steps and check the accuracy of the models along with the classification report.

```
GradientBoostingClassifier(learning_rate=0.45, random_state=0) :
 The confusion matrix:
 [[   5    5   42]
 [   5  111  436]
 [  37  297 2757]]

The classification report:
              precision    recall  f1-score   support

           1       0.11      0.10      0.10        52
           2       0.27      0.20      0.23       552
           3       0.85      0.89      0.87      3091

    accuracy                           0.78      3695
   macro avg       0.41      0.40      0.40      3695
weighted avg       0.75      0.78      0.76      3695
```

```
RandomForestClassifier(random_state=0) :
 The confusion matrix:
 [[   3    8   41]
 [   0  113  439]
 [   2  257 2832]]

The classification report:
              precision    recall  f1-score   support

           1       0.60      0.06      0.11        52
           2       0.30      0.20      0.24       552
           3       0.86      0.92      0.88      3091

    accuracy                           0.80      3695
   macro avg       0.58      0.39      0.41      3695
weighted avg       0.77      0.80      0.78      3695
```

```
DecisionTreeClassifier(random_state=0) :
 The confusion matrix:
 [[  11   15   26]
 [  16  198  338]
 [  80  653 2358]]

The classification report:
              precision    recall  f1-score   support

           1       0.10      0.21      0.14        52
           2       0.23      0.36      0.28       552
           3       0.87      0.76      0.81      3091

    accuracy                           0.69      3695
   macro avg       0.40      0.44      0.41      3695
weighted avg       0.76      0.69      0.72      3695
```

```
LogisticRegression(random_state=0) :
 The confusion matrix:
 [[  26   11   15]
 [ 110  152  290]
 [ 552  755 1784]]

The classification report:
              precision    recall  f1-score   support

           1       0.04      0.50      0.07        52
           2       0.17      0.28      0.21       552
           3       0.85      0.58      0.69      3091

    accuracy                           0.53      3695
   macro avg       0.35      0.45      0.32      3695
weighted avg       0.74      0.53      0.61      3695
```

```
ExtraTreesClassifier() :
 The confusion matrix:
 [[   0    4   48]
 [   0   68  484]
 [   0  153 2938]]

The classification report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        52
           2       0.30      0.12      0.18       552
           3       0.85      0.95      0.90      3091

    accuracy                           0.81      3695
   macro avg       0.38      0.36      0.36      3695
weighted avg       0.75      0.81      0.78      3695
```

Having performed the modeling steps above, the different performance KPIs has been derived for all the models like precision, recall, F1-Score and Support. And then a table with all the accuracy scores of the models below has been constructed to be able to compare better.

```
Table = pd.DataFrame({'Model':['Gradient Boosting Classifier','Random Forest Classifier','Logist
ic Regression','Decision Tree Classifier','SVC','ExtraTreesClassifier'],
                      'Acc_Score': lst})

Table.sort_values('Acc_Score', ascending = False)
```

:

	Model	Acc_Score
5	ExtraTreesClassifier	0.8135
1	Random Forest Classifier	0.7978
0	Gradient Boosting Classifier	0.7775
2	Logistic Regression	0.6947
4	SVC	0.5667
3	Decision Tree Classifier	0.5310

It is observed that of all the models, ExtraTreesClassifier performed the best with respect to predicting the severity of an accident with an accuracy of 81%. Second in line is the Random Forest Classifier with an accuracy of 79%.

To conclude, various Machine Learning models has been implemented and the best fit model has been chosen based on accuracy in the above section. There is scope to choose many more models based on the appropriateness with the problem statement and such should be experimented further for future reference.

Chapter 5 - Conclusion

5.1 Conclusion

In the entire experiment, I picked a problem statement and designed an approach with a solution approach to predict the accident severity using a machine learning model. The goal of the experiment was to use one of the popular methods of solving data science problems called CRISP-DM with the help of which I was able to define and understand the problem and dataset in-depth, prepare the data along with deep-dive exploration. Finally, I determined the candidate models for the same before moving ahead with the modeling and prediction phase. This helped me derive the model performance scores like accuracy which helps me understand the best model based on their accuracy.

While exploring the dataset, I was able to identify certain trends and based on these patterns I was able to determine some important findings and insights from the same. This helped me understand the features within the dataset in depth and be able to recommend accordingly. Exploratory Data Analysis is an essential step in the entire project cycle, wherein the dataset is used to perform univariate and multivariate analysis along with using certain complex visualizations to represent the data in an intuitive manner to the readers and audience. Through the various bar charts, pie charts, and correlation plots that I plotted in the previous sections, I was able to determine essential insights which helped me in the modeling phase as well as the recommendations from my dataset.

During the course of the project, I also researched various experiments and researches done in the past by other researchers which helped me understand the width and depth of related research work. It is imperative to study related work done in the domain so as to have a better understanding of the problem statement that I am trying to solve.

5.2 Recommendations

From the research on the above topic, it was determined that Extra Trees Classifier and Random Forest Classifier performed the best in terms of their accuracy in determining the severity of an accident from the given dataset. It was also determined from the data exploration that Wednesday and Fridays between hours 15 and 19 see the greatest number of accident occurrences and hence these times and days should be closely monitored and emergency response should be available around the areas of accident proneness. There are numerous insights and findings from the above exercise which can further be extended to newer horizons which I have discussed in the future works section. These recommendations should help my readers with taking the appropriate measures during future studies in the same problem space and while using the same dataset.

5.3 Future Work

In the scope of future work, I can collect more data points and features using advanced technologies like the Internet of Things to have a comprehensive understanding about the reason behind these accidents and take necessary actions accordingly. Moreover, I could use many more Machine Learning algorithms to determine the predictive capability and do an even more comprehensive analysis of these model performance KPIs to have an even more robust model for predictions. There are numerous researches performed in this domain and such research work can be studied to expand my view and understanding of this problem space even further.

Bibliography

1. Kumar, A. (2021, May 8). Revealed: Top causes of road accidents in UAE. Khaleej Times. Retrieved April 7, 2022, from https://www.khaleejtimes.com/transport/revealed-top-causes-of-road-accidents-in-uae
2. Reporter, A. S. (2019, July 25). *Road Accidents UAE's second biggest killer*. Uae – Gulf News. Retrieved April 7, 2022, from https://gulfnews.com/uae/road-accidents-uaes-second-biggest-killer-1.393653
3. UAE Government, Road safety - the official portal of the UAE Government. (n.d.)., from https://u.ae/en/information-and-services/justice-safety-and-the-law/road-safety
4. *Statistics ... It needs good data to foster change!* Road Safety UAE. (2022, March 16)., from https://www.roadsafetyuae.com/statistics/
5. AlKetbi, L. M. B., Grivna, M., & Al Dhaheri, S. (2020, August 31). *Risky driving behavior in Abu Dhabi, United Arab Emirates: A cross-sectional, survey-based study - BMC Public Health*. BioMed Central. Retrieved April 7, 2022, from https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-09389-8
6. Akmal, Abdelfatah, Mohamed, Saif, AlZaffin, Waleed, & Hijazi. (1970, January 1). *[PDF] trends and causes of traffic accidents in Dubai: Semantic scholar*. undefined. Retrieved April 7, 2022, from https://www.semanticscholar.org/paper/Trends-and-Causes-of-Traffic-Accidents-in-Dubai-Akmal-Abdelfatah/e8469517d6d066f176414f2850b4da16131248ce
7. Dina M.Awadallaa, Francisco Daniel, B.de Albuquerque (3rd February 2021), *Fatal Road Crashes in the Emirate of Abu Dhabi: Contributing Factors and Data-Driven Safety Recommendations*, https://doi.org/10.1016/j.trpro.2021.01.030
8. Omar Kassem Khalil (12 Feb 2018), *A Study on Road Accidents in Abu Dhabi Implementing a Vehicle Telematics System to Reduce Cost, Risk and Improve Safety,* 2017 10th International Conference on Developments in eSystems Engineering (DeSE), https://doi.org/10.1109/DeSE.2017.41
9. Michael P. LaValley, Logistic Regression, Aha Journals, 6 May 2008, https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.106.682658
10. Pierre G., Damien E., Louis W., Extremely randomized trees, Springer, 2 March 2006, https://link.springer.com/article/10.1007/s10994-006-6226-1
11. Francisco Daniel B.de Albuquerquea, Dina M.Awadallaa (15 Sep 2020), *Characterization of road crashes in the emirate of Abu Dhabi*, Elsevier, https://doi.org/10.1016/j.trpro.2020.08.136

12. Khaled Hamad (22 June 2016), *Road Traffic Accident Trends in Sharjah, United Arab Emirates during 2001-2014*, International Journal of Vehicle Safety, https://doi.org/10.1504/IJVS.2016.077151
13. Sharaf A. Alkheder, Reem Sabouni, Hany El Naggar, Abdul Rahim Sabouni (April, 2013), *Driver and vehicle type parameters' contribution to traffic safety in UAE*, Journal of Transport Literature, https://www.scielo.br/j/jtl/a/dVnQsJk73syKqQcVdZtjvbM/abstract/?lang=en#
14. Madampath Sankaran-Kutty, Abdulbari Bener, Kulangara P. Muralikuttan, and Michael Sebastian, 1 July 1998, *Road Traffic Accident Admissions in the United Arab Emirates,* https://doi.org/10.5144/0256-4947.1998.349
15. A.Bener, J.C.Murdoch, N.V.Achan, A.H.Karama, L.Sztriha, September 1996, *The effect of epilepsy on road traffic accidents and casualties*, Elsevier, https://doi.org/10.1016/S1059-1311(96)80039-2
16. Zhenhua Zhang, Qing He, Jing Gaod, Ming Ni, January 2018, *A deep learning approach for detecting traffic accidents from social media data,* Elsevier, https://doi.org/10.1016/j.trc.2017.11.027
17. Robert W. Thomas, José M. Vidal, 9 January 2017, *Toward detecting accidents with already available passive traffic information,* IEEE Xplore, https://doi.org/10.1109/CCWC.2017.7868428
18. Nejdet Dogru, Abdulhamit Subasi, 31 May 2018, *Traffic accident detection using random forest classifier,* IEEE Xplore, https://doi.org/10.1109/LT.2018.8368509
19. Fotios Zantalis, Grigorios Koulouras, Sotiris Karabetsos, Dionisis Kandris, 10 April 2019, *A Review of Machine Learning and IoT in Smart Transportation,* Future Internet, https://doi.org/10.3390/fi11040094
20. Singh, J., Singh, G., Singh, P., Kaur, M. (2019). *Evaluation and Classification of Road Accidents Using Machine Learning Techniques*. In: Shetty, N., Patnaik, L., Nagaraj, H., Hamsavath, P., Nalini, N. (eds) Emerging Research in Computing, Information, Communication and Applications. Advances in Intelligent Systems and Computing, vol 882. Springer, Singapore. https://doi.org/10.1007/978-981-13-5953-8_17
21. Steven Haynes, Prudencia Charles Estin, Sanela Lazarevski, Mekala Soosay, Ah-Lian Kor, 25 July 2019, *Data Analytics: Factors of Traffic Accidents in the UK,* IEEE Xplore, https://doi.org/10.1109/DESSERT.2019.8770021
22. Seok-Lyong Lee, 10 October 2016, *Assessing the Severity Level of Road Traffic Accidents Based on Machine Learning Techniques*, Ingenta Connect, https://doi.org/10.1166/asl.2016.8006

23. Pawłowicz, B., Salach, M., Trybus, B. (2019). *Smart City Traffic Monitoring System Based on 5G Cellular Network, RFID and Machine Learning*. In: Kosiuczenko, P., Zieliński, Z. (eds) Engineering Software Systems: Research and Praxis. KKIO 2018. Advances in Intelligent Systems and Computing, vol 830. Springer, Cham. https://doi.org/10.1007/978-3-319-99617-2_10
24. Jonghak Lee, Taekwan Yoon, Sangil Kwon, Jongtae Lee, 23 December 2019, *Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study,* Applied Sciences, https://doi.org/10.3390/app10010129
25. Bulbula Kumeda, Fengli Zhang, Fan Zhou, Sadiq Hussain, Ammar Almasri, Maregu Assefa, 21 November 2019, *Classification of Road Traffic Accident Data Using Machine Learning Algorithms,* IEEE Xplore, https://doi.org/10.1109/ICCSN.2019.8905362