RISK PREDICTION OF COLLISIONS IN TORONTO

by

Shailendra Khadka Yadav

A Major Research Project

presented to Ryerson University

in partial fulfillment of the requirements for the degree of

Master of Science

in the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2017

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Project. This is a true copy of the MRP, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Shailendra Khadka Yadav

RISK PREDICTION OF COLLISIONS IN TORONTO

Shailendra Khadka Yadav

Master of Science 2017

Data Science and Analytics

Ryerson University

ABSTRACT

Collision prediction models are used for a variety of purposes; most frequently to estimate the expected accident frequencies from various roadway entities like aggressive driving, traffic control, road class, speeding etc and also to identify factors that are associated with the occurrence of accidents. In this study, the Decision Tree, Random Forest and ARIMA time series model are implemented and analyzed over the Killed or Seriously Injured (KSI) Traffic Data so as to predict the severity of injury type, number of collisions in Toronto for future 12 months. The ARIMA model gives accuracy of 85% for the prediction of number of collisions. The Decision Tree using CART and Radom Forest models returns accuracy of 57% and 67% respectively for the classification of injury types.

Keywords: Collision Prediction, ARIMA, Decision Tree, Random Forest

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Risk Prediction of Collisions in Toronto

Shailendra Khadka Yadav*, Ryerson University*

I. INTRODUCTION

Collision prediction models are used for a variety of purposes like estimating the expected collision frequencies from various roadway entities such as aggressive driving, traffic control, road class, speeding. Determining risks of collision and factors associated with the road accidents has always been a focus of research.

In this project, a model for forecasting road accidents in Toronto city is developed and analyzed with several factors. The objective of this project is to predict risk of collision in city in future and the severity of injury of collision. Additionally, I have attempted to identify which factors are affecting for an accident in the city. In this study, the decision tree and Random Forest are used for classification of injury type to minimal, minor, major, fatal or none. Additionally, Random forest is used in my study to determine most important variable for collision of pedestrian, cyclist, automobile, motorcycle and truck. For the prediction of collision using the time series data, I have used ARIMA time series model which has been a quite appropriate for the future risk prediction of collision.

II. BACKGROUND AND LITERATURE REVIEW

There have been a number of significant researches done for predicting collision patterns. Meng Chen et.al. [1] have pointed the drawback of previous method for the next location prediction problems. Earlier models use historical trajectories of individual object or all objects to predict the next movement pattern of individual object or all objects respectively and based on mining frequent patterns or association rules to predict movement patterns of objects. The authors proposed a hybrid model to this problem. Regression technique has been used to develop a hybrid model regression. Time factor is another improvement on this model, it affects the movement pattern of individual and it increased the accuracy of model output.

The authors in [2] aim to develop a combined model of information mining model that extract feature from dataset and a prediction model to predict traffic accident based on machine learning techniques. It has been claimed that among these four methods; Statistical methods: Naïve Bayes (NB), Bayesian Network (BN), Linear Discriminant Analysis (LDA), Decision Trees: C4.5, CART, Artificial intelligence: Back Propagation Network (BPN), PNN, machine learning approaches: Bagging, RF, Random Committee (RC), the PNN has 97% accident severity predication rate, superior than others, when 18 parameters were input.

In [3], the authors have used autoregressive integrated moving average (ARIMA) time series intervention models to identify the effect of the raise in speed limit on fatalities, serious injuries, and case fatality. They have calculated case fatality rates (CFR), which are the risk of death, among all individuals injured in road crashes. Using the ARIMA time-series models, they have illustrated that a small increase in the speed limit resulted in an immediate, substantial and persistent increase in road deaths.

Miao Chong et al. in [4] have compared the performance of four machine learning algorithms; namely, neural networks trained using hybrid learning approaches, support vector machines, decision trees and a concurrent hybrid model involving decision trees and neural networks. With experimental result, among the four approaches considered, the hybrid decision tree-neural network approach outperformed the other approaches.

In [5], the authors have used Gompertz growth model to project vehicle ownership and the calculation of road accident death rate was done using Autoregressive Integrated Moving Average (ARIMA) model with transfer noise function. The Gompertz model forecasted that vehicle ownership would be equal to 0.4409 by the year 2010. The road accident death rate is estimated to decrease to 4.22 in year 2010, at an average downfall rate of 2.14% per year.

In [6], Rami Harb et al studied accident analysis using trees. They have explored the drivers, vehicles, and environments characteristics associated with crash avoidance maneuvers. In this study, rear-end collisions, head-on collisions, and angle collisions are analyzed separately using decision trees. The random forest method is also used to identify the importance of the drivers, vehicles, and environments characteristics on crash avoidance maneuvers.

Three binary classification trees were developed. The analysis showed that close to 30% of the drivers do not take evasive action on critical traffic conditions approaching to motor vehicle accidents.

Paraskevi Michalaki et al. in [7] explored Vector Autoregressive (VAR) model to examine the leaning of hard shoulder (HS) and motorway collisions over the same period. It is observed that hard shoulder collisions are much more severe than main carriageway (MC) collisions. In their result they mentioned that 10.7% of HS collisions are fatal and 2.3% of MC collisions being fatal

Authors in [8] studied the significance of Data Mining classification algorithms in predicting the vehicle collision patterns occurred in training accident data set. They have used a data mining tool named TANAGRA for the required experimentations. They have implemented and analyzed various classification algorithms namely, C4.5, Decision List, ID3 and RndTree are few to name. The results showed that in all the cases the Random Tree performs best among all of the other classifiers.

In [9], the authors have used data of accidents records consists of continuous and categorical data. They have used artificial neural network for analyzing the continuous while the categorical data are analyzed using decision tree. Under neural network approach, they have used the algorithms based on multilayer perceptron and radial basis function. Similarly, the authors have used ID3 and Function Tree approach for decision tree Performance analysis. The overall study in their research showed that the decision tree approach outperforms the neural network approach.
Authors in [10], performed an extensive review of literatures on accident prediction modeling. Their research is a survey based on questionnaires from several National Road Administrations (NRAs) in Europe, US and Australia. The aim of the survey was to collect detailed information like on Accident Prediction Models (APMs) developed and used by them in those regions. They covered the availability, quality and definitions of relevant data like crash data, traffic data, road design data and other related data in the questionnaire survey.

III. METHODOLOGY

The models used for collision prediction of the Toronto city are briefly discussed in this section. There are four components for this research project methodology that explains every step for risk prediction of collision. Descriptive analysis and predictive analysis is done for different experiments in Experiments section.

A. Data Description

Most of the variables are factor type and these variables contain multiple categories and some of variables contain binary values. For example, one variable called **INJURY** contains 6 factors and one variable like **REDLIGHT** is a binary variable and it contains 2 factors and so on.

Feature Selection:

The selection of inputs variable is the most important part of building a useful prediction. It represents all of the information that is available to the model for the prediction.

- **Important Variables:** After feature selection I am getting AG_DRIV, REDLIGHT, SPEEDING, ALCOHOL, LIGHT, ROAD_CLASS, TRAFFCTL, RDSFCOND and VISIBILITY these variables are mostly important.
- **Variable Selection Process**: To select those variables I have used some statistical algorithms and find these variables have very significant impact on collision. Here I have used some statistical techniques like, correlation analysis, frequency distribution plot, Factor analysis, Weight of Evidence and Random Forest.

B. Model Selection: Machine Learning Algorithm

To get the final result I have used the following models:

Time Series Model: In time series model I have used total number of collisions for year and month combinations. Here the collision records are present from 2006 to 2016. I have aggregated it into year and month combination. Also I have used overall level seasonal index as a regression variable to predict future collision numbers.

Autoregressive integrated moving average (ARIMA) models are best known Time series models. So knowledge about the time series algorithms is very necessary to predict future collisions.

Mathematical equation for ARIMA:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

where,

$\hat{y}_t$ = Forecasted Series at time t.

μ = Constant value

ϕ_1=Coefficient for yt-1

y_{t-1} = Actual value for t-1 th time point

θ_1 = Coefficientfor e_{t-1}

e_{t-1} = error for t-1 time period

and so on

Decision Tree Algorithm: Decision tree is applied for classification and there are many decision tree algorithms are available. This is mainly rules based algorithm, it captured most optimized rule for classification. In our case if I want to build rules to predict what is the injury type, is it minimal, major, none or fatal etc. So, our objective is to get the path why the injury is major or minor? To solve these types of classification problem need to build decision tree algorithm, then we can see for which factors injury is minimum or injury in maximum. And this problem is also explained with Random Forest.

To predict injury type I have used some variable like AG_DRIV, REDLIGHT, SPEEDING, ALCOHOL, LIGHT, ROAD_CLASS, TRAFFCTL, RDSFCOND, INVAGE and VISIBILITY and so on. And then I have considered the INJURY as dependent variable and other variables as independent variable. Then I am getting the final classification after running the both algorithms and I am also able to capture the important factors for INJURY type.

Random Forest: Random forest is another machine learning algorithm for classification. It builds many decision trees on that data for different samples and then it uses the voting algorithm and then combined the result. And it works better than decision tree because it taking decision from different trees and the result is more reliable. In my project, I have used Random Forest to identify the most important variables for motorcycle collision, automobile collision, truck collision etc. I have used this algorithm to identify the major factors for different types of collisions. Like which are the major factors for Truck collision and which are the major factors for motorcycle collision and so on. Here the dependent variable is binary variable and the independent variables are the other factor variables that I have used before in decision tree algorithm.

Weight of Evidence (WoE) and Information Value: I have used WoE to capture the important variables for different types of collision. With this model, it is possible to determine the variable importance with score and which component of a variables affecting more. For categorical response variable, WoE is the most important technique to identify the most important variables for classification. Here I am using this as a challenger process of Random Forest.

Factor Analysis: I have also applied factor analysis on that data to check the direction and association of those variables. Which variable are most similar and falling into similar group.

C. Model Validation

I am using cross validation technique for model validation that I have create 2 partitions of the dataset. I am using 1st part of the data for training the model and the 2nd part for validation. For training purpose, I have captured 80% of the data and rest 20% for validation.

D. Analytics and Experiments

Data Analysis

The goal of data analysis is to apply techniques of Data Visualization and Data Mining to analyze collisions in Toronto from the year 2006 to 2016. My objective is to explore some key insights like where the accident mostly happening and what is the key reason for that accident. I am trying to find some key insights from the data using some exploratory statistical techniques.

i. Data Collection and Description

This Killed or Seriously Injured (KSI) dataset has been collected from public source for Toronto Police Service site.

KSI dataset has 13,173 records and 54 variables. Most of them are categorical in nature. Few of those variables are explaining about geographical locations, some variables are explaining about injuries, vehicles and causes.

ii. Data Understanding

To understand the data we have to check the quality of the data. For this purpose we have to check missing value, variable distribution, and duplication in data set.

Table 1 Missing Value Report I

Missing Value Report

Field.Name	Description	No_of_Missing	Pct_of_Missing
ACCLASS	Classification of Accident	0	0
ACCLOC	Accident Location	5498	41.7
AG_DRIV	Aggressive and Distracted Driving Collision	6454	49
ALCOHOL	Alcohol Related Collision	12567	95.4
AUTOMOBILE	Driver Involved in Collision	1207	9.2
CYCACT	Cyclist Action	12646	96
CYCCOND	Cyclist Condition	12647	96
CYCLIST	Cyclists Involved in Collision	11792	89.5
CYCLISTYPE	Cyclist Crash Type - detail	12632	95.9
DATE	Date Accident Occurred	0	0
DISABILITY	Medical or Physical Disability Related Collision	12794	97.1
District	City District	0	0
Division	Police Division	0	0
DRIVACT	Apparent Driver Action	6771	51.4
DRIVCOND	Driver Condition	6773	51.4
EMERG_VEH	Emergency Vehicle Involved in Collision	13158	99.9
IMPACTYPE	Initial Impact Type	1	0
INITDIR	Initial Direction of Travel	4132	31.4
INJURY	Severity of Injury	1603	12.2
INVAGE	Age of Involved Party	0	0
INVTYPE	Involvement Type	4	0
LATITUDE	Latitude	0	0
LIGHT	Light Condition	0	0
LOCCOORD	Location Coordinate	98	0.7
LONGITUDE	Longitude	0	0
MANOEUVER	Vehicle Manouever	5804	44.1
MOTORCYCLE	Motorcyclist Involved in Collision	12224	92.8
PASSENGER	Passenger Involved in Collision	8280	62.9
PEDACT	Pedestrian Action	11064	84
PEDCOND	Condition of Pedestrian	11064	84
PEDESTRIAN	Pedestrian Involved In Collision	7955	60.4
PEDTYPE	Pedestrian Crash Type - detail	11055	83.9
RDSFCOND	Road Surface Condition	0	0
REDLIGHT	Red Light Related Collision	12042	91.4
ROAD_CLASS	Road Classification	0	0
SPEEDING	Speeding Related Collision	10819	82.1
TRAFFCTL	Traffic Control Type	29	0.2
TRSN_CITY_VEH	Transit or City Vehicle Involved in Collision	12362	93.8
TRUCK	Truck Driver Involved in Collision	62	0.5
VEHTYPE	Type of Vehicle	1383	10.5
VISIBILITY	Environment Condition	0	0
YEAR	Year Accident Occurred	0	0

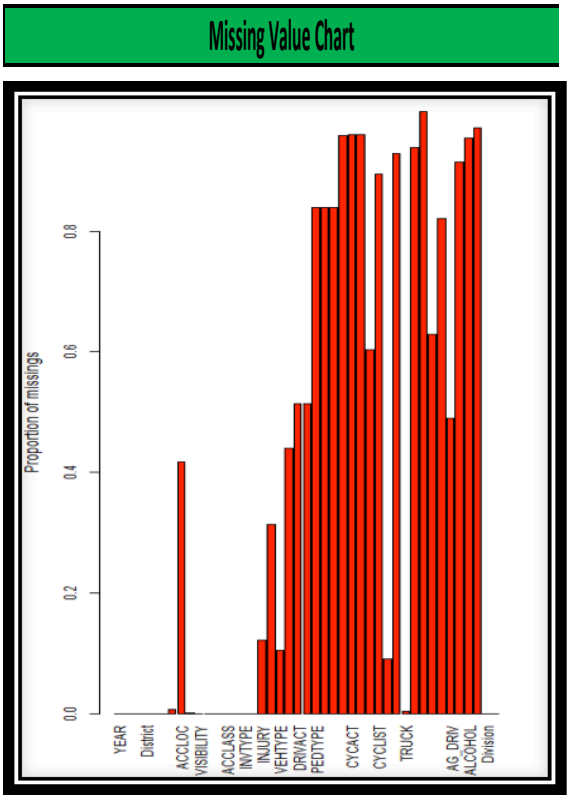

Figure 1 Missing Value Chart I

It can be seen from above table, EMERG_VEH having 99.9% of missing value, DISABILITY having 97.1% of missing values, CYCLIST_TYPE having 95.7% of missing values. These are few examples where as INTDIR having average missing value i.e. 37.4%, ACCLOC having 41.7% of missing value. ACCLASS, DATE and more having no missing values.

In the case of missing data analysis, either we can delete the cases with missing data or try to estimate the value of missing data.

Missing value Treatment:

I have taken some logical steps as well as some statistical techniques to impute missing values.

So, we need to overcome with such problems by imputing missing values. This is first level of treatment for missing value by imputation. In this phase I have imputed missing value variables having binary class like PEDESTRIAN, CYCLIST, MOTORCYCLE etc. by "No" class value.

Table 2 Missing Value Report II

Missing Value Report

Field.Name	Description	No_of_Missing	Pct_of_Missing
ACCLASS	Classification of Accident	0	0
ACCLOC	Accident Location	5498	41.7
AG_DRIV	Aggressive and Distracted Driving Collision	0	0
ALCOHOL	Alcohol Related Collision	0	0
AUTOMOBILE	Driver Involved in Collision	0	0
CYCACT	Cyclist Action	12646	96
CYCCOND	Cyclist Condition	12647	96
CYCLIST	Cyclists Involved in Collision	0	0
CYCLISTYPE	Cyclist Crash Type - detail	12632	95.9
DATE	Date Accident Occurred	0	0
DISABILITY	Medical or Physical Disability Related Collision	0	0
District	City District	0	0
Division	Police Division	0	0
DRIVACT	Apparent Driver Action	6771	51.4
DRIVCOND	Driver Condition	6773	51.4
EMERG_VEH	Emergency Vehicle Involved in Collision	0	0
IMPACTYPE	Initial Impact Type	1	0
INITDIR	Initial Direction of Travel	4132	31.4
INJURY	Severity of Injury	1603	12.2
INVAGE	Age of Involved Party	0	0
INVTYPE	Involvement Type	4	0
LATITUDE	Latitude	0	0
LIGHT	Light Condition	0	0
LOCCOORD	Location Coordinate	98	0.7
LONGITUDE	Longitude	0	0
MANOEUVER	Vehicle Manouever	5804	44.1
Month		0	0
MOTORCYCLE	Motorcyclist Involved in Collision	0	0
PASSENGER	Passenger Involved in Collision	0	0
PEDACT	Pedestrian Action	11064	84
PEDCOND	Condition of Pedestrian	11064	84
PEDESTRIAN	Pedestrian Involved In Collision	0	0
PEDTYPE	Pedestrian Crash Type - detail	11055	83.9
RDSFCOND	Road Surface Condition	0	0
REDLIGHT	Red Light Related Collision	0	0
ROAD_CLASS	Road Classification	0	0
SPEEDING	Speeding Related Collision	0	0
TRAFFCTL	Traffic Control Type	29	0.2
TRSN_CITY_VEH	Transit or City Vehicle Involved in Collision	0	0
TRUCK	Truck Driver Involved in Collision	0	0
VEHTYPE	Type of Vehicle	1383	10.5
VISIBILITY	Environment Condition	0	0
YEAR	Year Accident Occurred	0	0

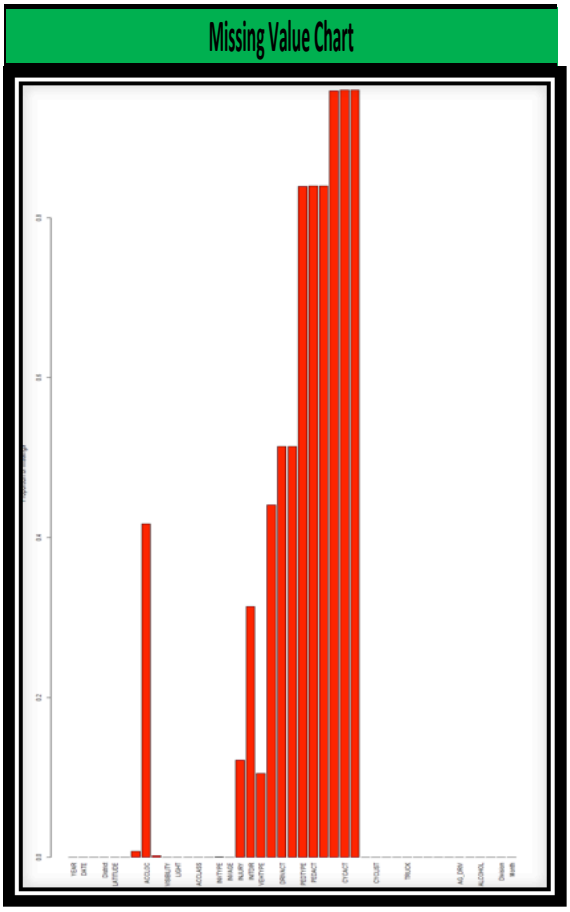

Figure 2 Missing Value Chart II

From the above table there is a still missing value present in the few variables and I have done 2nd level missing value imputation that is observed in the below table.

If PEDESTRIN involvement in collision is "No", so it is clear that variable related to PEDESTRIAN like PEDTYPE, PEDCOND and PEDACT are also not present, which is logical. Similarly if CYCLIST involvement in collision is "No", so it is also clear that variable related to CYCLIST like CYCLISTYPE, CYCCOND and CYCACT are also not present and so on.

Table 3 Missing Value Report III

Field.Name	Description	No_of_Missing	Pct_of_Missing
ACCLASS	Classification of Accident	0	0
ACCLOC	Accident Location	5498	41.7
ACCNUM	Accident Number	0	0
AG_DRIV	Aggressive and Distracted Driving Collision	0	0
ALCOHOL	Alcohol Related Collision	0	0
AUTOMOBILE	Driver Involved in Collision	0	0
CYCACT	Cyclist Action	0	0
CYCCOND	Cyclist Condition	0	0
CYCLIST	Cyclists Involved in Collision	0	0
CYCLISTYPE	Cyclist Crash Type - detail	0	0
DATE	Date Accident Occurred	0	0
DISABILITY	Medical or Physical Disability Related Collision	0	0
District	City District	0	0
Division	Police Division	0	0
DRIVACT	Apparent Driver Action	0	0
DRIVCOND	Driver Condition	0	0
Driver		4	0
Driver_Flag		0	0
EMERG_VEH	Emergency Vehicle Involved in Collision	0	0
IMPACTYPE	Initial Impact Type	1	0
INITDIR	Initial Direction of Travel	4132	31.4
INJURY	Severity of Injury	1603	12.2
INVAGE	Age of Involved Party	0	0
INVTYPE	Involvement Type	4	0
LATITUDE	Latitude	0	0
LIGHT	Light Condition	0	0
LOCCOORD	Location Coordinate	98	0.7
LONGITUDE	Longitude	0	0
MANOEUVER	Vehicle Manouever	5804	44.1
Month		0	0
MOTORCYCLE	Motorcyclist Involved in Collision	0	0
PASSENGER	Passenger Involved in Collision	0	0
PEDACT	Pedestrian Action	0	0
PEDCOND	Condition of Pedestrian	0	0
PEDESTRIAN	Pedestrian Involved In Collision	0	0
PEDTYPE	Pedestrian Crash Type - detail	0	0
RDSFCOND	Road Surface Condition	0	0
REDLIGHT	Red Light Related Collision	0	0
ROAD_CLASS	Road Classification	0	0
SPEEDING	Speeding Related Collision	0	0
TRAFFCTL	Traffic Control Type	29	0.2
TRSN_CITY_VE	Transit or City Vehicle Involved in Collision	0	0
TRUCK	Truck Driver Involved in Collision	0	0
VEHTYPE	Type of Vehicle	1383	10.5
VISIBILITY	Environment Condition	0	0
YEAR	Year Accident Occurred	0	0

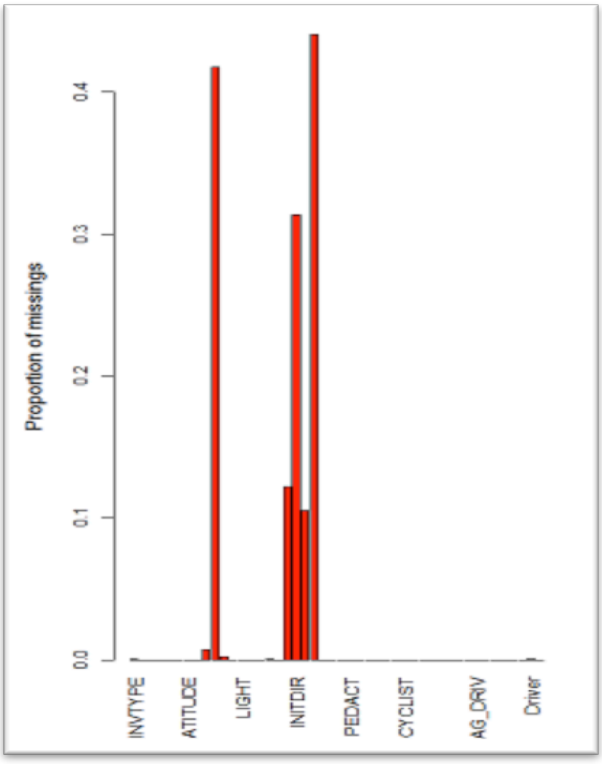

Figure 3 Missing Value Chart III

Similarly, there is a condition when INVTYPE is any type of Driver and the related value for the variables DRIVACT and DRIVCOND are present, otherwise this is a missing value for those variables. So, I have imputed those missing values as a "Not_Applicable".

There are still few variables having missing values but we can't impute further. For example ACCLOC we can't impute any values and INITDIR, VEHTYPE also having same issue.

Duplicate Records Treatment:

Data redundancy is also a major issue in any dataset to deal with. I have found most of the records are repeating in this dataset. There are some variables responsible for duplicate records, for e.g., ACCNUM is same for multiple records.

Accident Frequency: There are 13,173 records in KSI data but these are not actual number of collision. For 1 collision there is more classes for variable INVTYPE; pedestrian, cyclist, automobile, motorcycle, trucks and some other vehicles are involved. So, it creates more records for a particular 1 collision. For example, if the INVTYPE has 3 values (automobile, motorcycle, trucks) then there are 3 duplicate recodes have same accident number (ACCNUM).

District frequency: District is repeating multiple times for 1 collision. If we want to predict which location is most accident prone then this repeated frequency may give wrong interpretation. So, we need to remove duplication in the dataset for some predictions.

A comparison is shown in below between unique records and duplicate records of collision in different districts.

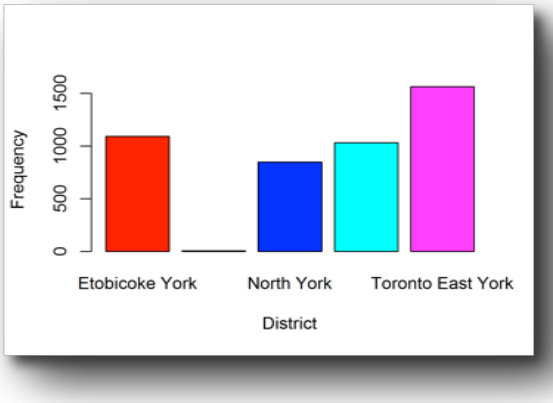

Figure 4 District wise Collision Records with Unique Records

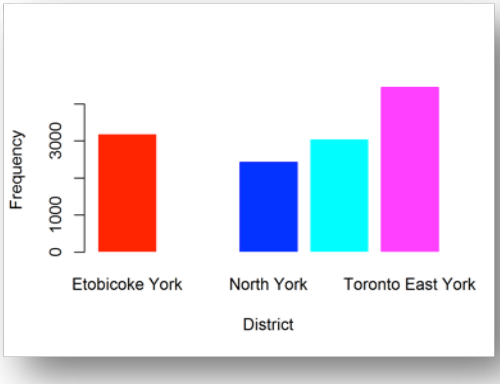

Figure 5 District wise Collision Records with Duplicate Records

There are few variables accountable for duplication of the records; ACCNUM, DATE, ROAD_CLASS, District, LATITUDE, LONGITUDE, LOCCOORD, TRAFFCTL, VISIBILITY, LIGHT, RDSFCOND, ACCLASS, IMPACTYPE etc. So, I have taken into account on these variable and observed new dataset of 4,534 unique records only. It is very less record compare to the original dataset of 13,173 records.

For Second Part

From INVTYPE we can classify vehicle types or all involvement parties in a collision like Motorcycle, Cycle, Truck, Pedestrian, Automobile etc. If INVTYPE is Motorcycle Driver and Motorcycle Passenger then we can say Motorcycle is present on that collision. And from INVTYPE fields related to Truck Driver, we can say Truck is present for that collision. In INVTYPE contains Bicycle and Moped then we can easily classify Cyclist.

For one collision, INVTYPE contains different factors or values like Driver, Passenger, Pedestrian, Motorcycle Driver, Cyclist, Truck Driver etc. Similarly INVAGE, INJURY, INITDIR, VEHTYPE, MANOEUVER,

DRIVACT, DRIVCOND, PEDTYPE, PEDACT, PEDCOND, CYCLISTYPE, CYCACT, CYCCOND these variables contains different factors or values. But PEDESTRIAN, CYCLIST, AUTOMOBILE, MOTORCYCLE, TRUCK, TRSN_CITY_VEH, EMERG_VEH, and PASSENGER contain 'Yes', 'No' values and directly related to INVTYPE.

So, from the combination of (INVTYPE and PEDESTRIAN), (INVTYPE and CYCLIST), (INVTYPE and MOTORCYCLE), (INVTYPE and TRUCK), (INVTYPE and TRSN_CITY_VEH), (INVTYPE and EMERG_VEH), (INVTYPE and PASSENGER), I can get different subset with unique records.

If INVTYPE is 'Truck Driver' then it gives subset of all Truck related accidents; where also have to check the Truck column only containing 'Yes'. After that we are able to capture all unique records for Truck collision.

Similarly, INVTYPE is 'Pedestrian' then it gives subset of all Pedestrian related accidents, where also have to check the Pedestrian column only containing 'Yes'. After that we are able to capture all unique records for Pedestrian collision. In this way I am able to capture the rest of the INVTYPE.

iii. Data Cleaning

After getting the raw data, it is one of the mandatory part is to check the quality of the data. There are several different methods are available to clean the data and after cleaning the data we are able to improve the quality of the data and then we can use the data for further analytics.

In this project I have cleaned the data using missing value treatment. In this data we have seen lot of missing values and without cleaning those missing values it should be difficult to run analytics on that data. I have imputed missing values at 3 levels. First I have rejected some values logically then created missing value report based on those variables. At first level I have creating the report based on 43 basic variables. And from the below report we can see almost 20 variables have some serious problem. Also I have cleaned the Date field from text format to date format.

I have cleaned some duplicate data and normalize the dataset. It has observed that the actual accident is very less than the recorded incidences, which is quite meaningful for experiment.

iv. Feature selection/variable selection

a) Factor Analysis

From this analysis we can identify those variables that are associated with each other. Similar variables are grouped together. From the below table we can see these variables are falling into same group.

Table 4 Variable Grouping

From Variable Grouping table we can conclude that which variables are correlated to each other. Here I am creating 4 groups. Which variables are falling into same group that means those variable are associated to each other.

Variable Grouping

Variables	Description	dim.1	dim.2	dim.3	dim.4	Classification
ACCLASS	Classification of Accident	0.2	0.0	0.0	0.0	Dim1
MANOEUVER	Vehicle Manouever	6.7	4.1	3.4	2.6	Dim1
PEDTYPE	Pedestrian Crash Type - detail	8.0	3.5	7.6	7.5	Dim1
PEDACT	Pedestrian Action	8.0	3.5	7.6	7.5	Dim1
PEDCOND	Condition of Pedestrian	8.0	3.5	7.5	7.5	Dim1
CYCLIST	Cyclists Involved in Collision	5.2	4.9	1.5	1.6	Dim1
TRUCK	Truck Driver Involved in Collision	0.1	0.0	0.0	0.0	Dim1
EMERG_VEH	Emergency Vehicle Involved in Collision	0.0	0.0	0.0	0.0	Dim1
INVTYPE	Involvement Type	10.4	13.4	10.0	12.8	Dim2
VEHTYPE	Type of Vehicle	8.1	12.9	4.3	3.9	Dim2
DRIVACT	Apparent Driver Action	3.8	8.8	2.7	2.8	Dim2
DRIVCOND	Driver Condition	3.6	8.7	1.7	1.1	Dim2
CYCLISTYPE	Cyclist Crash Type - detail	5.3	7.6	2.0	4.7	Dim2
CYCACT	Cyclist Action	5.3	7.6	1.9	4.6	Dim2
CYCCOND	Cyclist Condition	5.3	7.5	1.9	4.6	Dim2
ROAD_CLASS	Road Classification	0.1	0.1	0.6	0.1	Dim3
LOCCOORD	Location Coordinate	0.0	0.0	0.3	0.2	Dim3
TRAFFCTL	Traffic Control Type	0.1	0.0	0.2	0.1	Dim3
VISIBILITY	Environment Condition	0.2	0.0	0.3	0.1	Dim3
LIGHT	Light Condition	0.2	0.0	0.3	0.1	Dim3
RDSFCOND	Road Surface Condition	0.2	0.1	0.3	0.1	Dim3
IMPACTYPE	Initial Impact Type	9.0	5.0	11.6	2.7	Dim3
INITDIR	Initial Direction of Travel	2.0	1.5	7.6	5.8	Dim3
PEDESTRIAN	Pedestrian Involved In Collision	6.3	0.1	7.5	0.5	Dim3
PASSENGER	Passenger Involved in Collision	0.1	0.1	5.3	0.1	Dim3
SPEEDING	Speeding Related Collision	0.1	0.1	2.1	0.3	Dim3
AG_DRIV	Aggressive and Distracted Driving Collision	0.1	0.3	0.7	0.4	Dim3
REDLIGHT	Red Light Related Collision	0.1	0.1	0.6	0.2	Dim3
ALCOHOL	Alcohol Related Collision	0.0	0.0	0.6	0.1	Dim3
DISABILITY	Medical or Physical Disability Related Collision	0.0	0.0	0.2	0.1	Dim3
District	City District	0.2	0.5	0.6	0.7	Dim4
ACCLOC	Accident Location	0.1	0.0	0.2	0.6	Dim4
INVAGE	Age of Involved Party	1.3	0.9	3.2	9.8	Dim4
INJURY	Severity of Injury	1.3	4.2	4.3	14.1	Dim4
AUTOMOBILE	Driver Involved in Collision	0.0	0.2	0.4	0.4	Dim4
MOTORCYCLE	Motorcyclist Involved in Collision	0.3	0.2	0.4	1.2	Dim4
TRSN_CITY_VEH	Transit or City Vehicle Involved in Collision	0.1	0.0	0.0	0.4	Dim4
Division	Police Division	0.2	0.5	0.7	0.8	Dim4

For example, DISABILITY, ALCOHOL, REDLIGHT, AG_DRIV, SPEEDING etc are fallen into dimension 3.It says that there is a relationship among those variables. Similarly there is relationship between 4 different groups.

b) Variable Importance using Weight of Evidence and Random Forest

Here I have done some statistical analysis like WOE (Weight of Evidence) and IV (Information Value) and also did some analysis like Random Forest to capture important factors for different types of collision.

Here I have capture 2 tables. First table is observed after running WOE technique and the 2nd table from Random Forest. Both tables are showing the important factors for automobile collision. From the first table we can see most of the automobile collisions are happening for aggressive driving, red light, Speeding and also logically these are correct factors for collision. And from the 2nd table light, aggressive driving and road class are 3 most important factors for collisions. From both tables we can conclude that aggressive driving is the most important factor for automobile collision.

Table 5 Automobile_WOE

AUTOMOBILE_WOE		
Variables	**IV**	**Importance**
AG_DRIV	0.265	Medium
REDLIGHT	0.234	Medium
SPEEDING	0.133	Low
ALCOHOL	0.106	Low
LIGHT	0.062	Low
TRAFFCTL	0.037	Low
ROAD_CLASS	0.025	Low
VISIBILITY	0.006	Low
LOCCOORD	0.005	Low
RDSFCOND	0.004	Low
DISABILITY	0.001	Low

Table 6 Automobile_VARIMP_RF

AUTOMOBILE_VARIMP_RF	
Variables	**Overall**
LIGHT	36.30418871
AG_DRIV	35.71590665
ROAD_CLASS	33.75430765
TRAFFCTL	32.398788
RDSFCOND	23.35231947
VISIBILITY	21.43700282
LOCCOORD	13.83209178
REDLIGHT	11.20660688
DISABILITY	10.51038992
SPEEDING	10.27622147
ALCOHOL	6.524021526

Variable Importance for AUTOMOBILE Collisions: Using Weight of Evidence (WoE) and Random forest, experiment shows that Light, Aggressive and Distracted Driving, Traffic Control Type, Road Surface Condition, Visibility, Location Coordinate, Red Light, Speeding, Alcohol are important factors.

Variable Importance for CYCLIST Collisions: Here experiment shows that Light, Traffic Control Type, Road Classification, Aggressive and Distracted Driving, Speeding, Road Surface Condition, Visibility, Disability, Red Light, Alcohol are more significant factors for collision.

Table 7 Motorist_WOE

MOTORIST_WOE		
Varname	**IV**	**Importance**
RDSFCOND	0.162	Low
VISIBILITY	0.112	Low
TRAFFCTL	0.103	Low
DISABILITY	0.08	Low
SPEEDING	0.072	Low
LIGHT	0.068	Low
AG_DRIV	0.056	Low
ROAD_CLASS	0.03	Low
ALCOHOL	0.016	Low
REDLIGHT	0.015	Low
LOCCOORD	0.002	Low

Table 8 Motorist_VARIMP_RF

MOTORIST_VARIMP_RF	
Variables	**Overall**
LIGHT	34.64284102
ROAD_CLASS	28.21676282
TRAFFCTL	22.05957356
RDSFCOND	17.15552664
SPEEDING	14.46242674
LOCCOORD	14.05754574
ALCOHOL	11.17583861
VISIBILITY	10.76946675
AG_DRIV	10.01811487
REDLIGHT	4.604355137
DISABILITY	4.361768155

Variable Importance for MOTORCYCLE Collisions: Here experiment shows that Light, Road Classification, Traffic Control Type, Speeding, Location Coordinate, Alcohol, Visibility, Aggressive and Distracted Driving, Red Light, Disability are more significant factors for collision.

Table 9 Cyclist_WOE

CYCLIST_WOE		
Varname	**IV**	**Importance**
SPEEDING	0.195	Low
AG_DRIV	0.12	Low
LIGHT	0.114	Low
RDSFCOND	0.104	Low
VISIBILITY	0.076	Low
ALCOHOL	0.047	Low
REDLIGHT	0.047	Low
LOCCOORD	0.044	Low
TRAFFCTL	0.041	Low
ROAD_CLASS	0.026	Low
DISABILITY	0.001	Low

Table 10 Cyclist_ VARIMP_RF

CYCLIST_VARIMP_RF	
Variables	**Overall**
LIGHT	53.23351022
TRAFFCTL	44.22375406
ROAD_CLASS	39.01231853
AG_DRIV	28.30692894
SPEEDING	24.06554545
RDSFCOND	19.07136832
LOCCOORD	16.0658909
VISIBILITY	13.74696087
DISABILITY	11.66553235
REDLIGHT	10.99900887
ALCOHOL	10.0631745

Variable Importance for CYCLIST Collisions: Here experiment shows that Light, Traffic Control Type, Road Classification, Aggressive and Distracted Driving, Speeding, Road Surface Condition, Visibility, Disability, Red Light, Alcohol are more significant factors for collision.

Table 11 Pedestrian_WOE

PEDESTRIAN_WOE		
Varname	**IV**	**Importance**
SPEEDING	0.224	Medium
REDLIGHT	0.122	Low
AG_DRIV	0.091	Low
VISIBILITY	0.062	Low
ROAD_CLASS	0.059	Low
RDSFCOND	0.057	Low
DISABILITY	0.052	Low
ALCOHOL	0.043	Low
TRAFFCTL	0.034	Low
LIGHT	0.026	Low
LOCCOORD	0.006	Low

Table 12 Pedestrian _ VARIMP_RF

PEDESTRIAN_VARIMP_RF	
Variables	**Overall**
SPEEDING	231.4337196
LIGHT	195.8484149
TRAFFCTL	170.789672
ROAD_CLASS	166.558863
REDLIGHT	150.8348261
VISIBILITY	124.0402368
RDSFCOND	105.5380497
AG_DRIV	95.00251425
DISABILITY	83.01823875
LOCCOORD	71.39858477
ALCOHOL	60.23901612

Variable Importance for PEDESTRIAN Collisions: Here experiment shows that Speeding, Light, Traffic Control Type, Road Classification, Red Light, Visibility, Road Surface Condition, Aggressive and Distracted Driving, Disability, Location Coordinate, Alcohol,are more significant factors for collision.

Table 11 Truck_WOE

TRUCK_WOE		
Varname	**IV**	**Importance**
AG_DRIV	0.531	High
TRAFFCTL	0.401	High
ROAD_CLASS	0.399	High
SPEEDING	0.277	Medium
LIGHT	0.176	Low
RDSFCOND	0.149	Low
VISIBILITY	0.108	Low
LOCCOORD	0.074	Low
REDLIGHT	0.008	Low
ALCOHOL	0.002	Low
DISABILITY	0.001	Low

Table 12 Truck_ VARIMP_RF

TRUCK_VARIMP_RF	
Variables	**Overall**
ROAD_CLASS	1.745742159
LIGHT	1.61930301
TRAFFCTL	1.252908736
VISIBILITY	0.865128483
RDSFCOND	0.797520691
LOCCOORD	0.792917555
AG_DRIV	0.686638659
SPEEDING	0.301078618
ALCOHOL	0.080679695
DISABILITY	0.043596978
REDLIGHT	0.031411332

Variable Importance for TRUCK Collisions: Here experiment shows that Road Classification, Light, Traffic Control Type, Road Surface Condition, Location Coordinate, Aggressive and Distracted Driving, Speeding, Alcohol, Disability, Red Light are more significant factors for collision.

Visualization: After cleaning the data I have created some charts for exploratory data analysis and also to visualize the model result and it helps us interpret difficult business problems in simpler way for other persons. In this project I am using 10 years trend map to identify pattern of the collision is happening. Also I have created some plot to see the basic distribution of categorical variables and as well as time series plot to see time wise collision pattern.

I. Capturing some meaningful insights from the original data

a) Pedestrian involved in collisions

View traffic related collisions data involving Pedestrians. These events include any serious or fatal collision where a Pedestrian is involved.

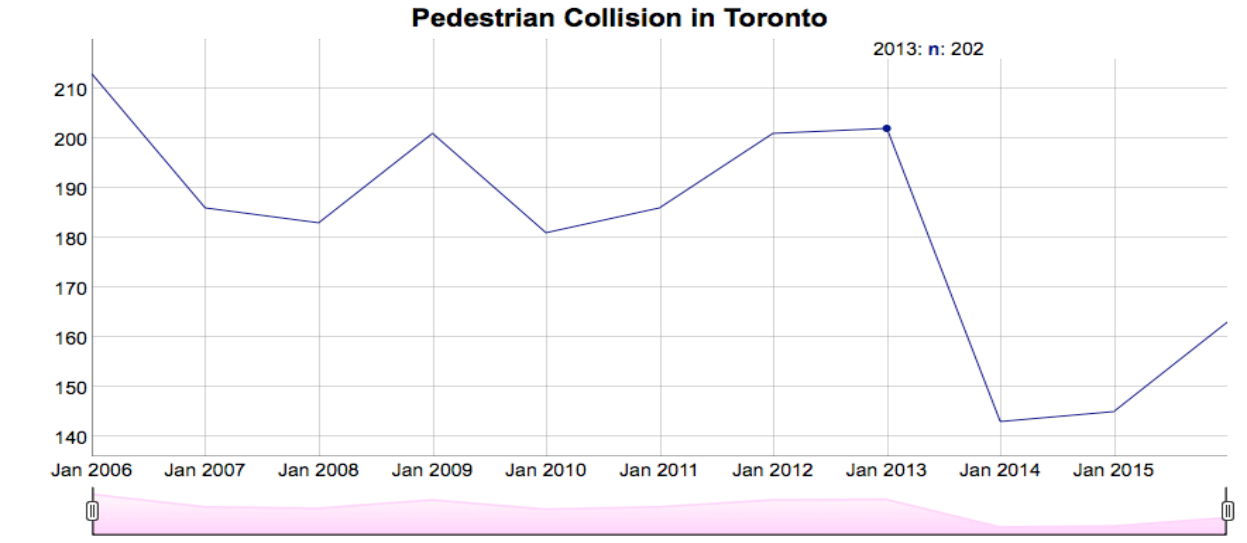

Figure 6 10 Year Trend of Pedestrian Collision

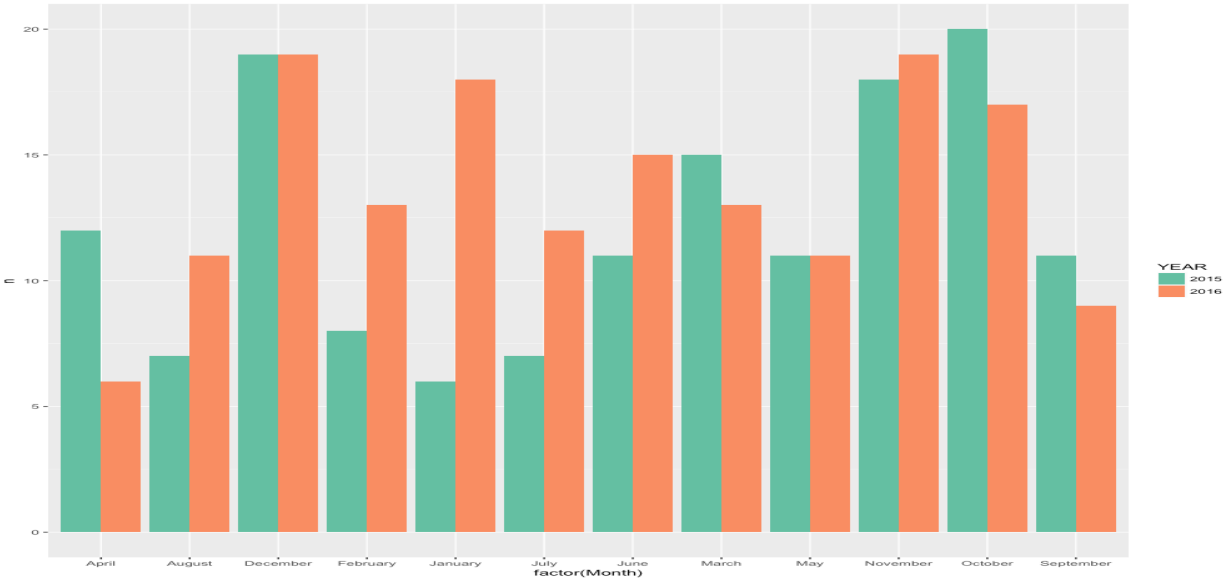

Figure 7 Total KSI by Month of Pedestrian Collision

b) Cyclist involved in collisions

View traffic related collisions data involving Cyclists. These events include any serious or fatal collision where a cyclist is involved.

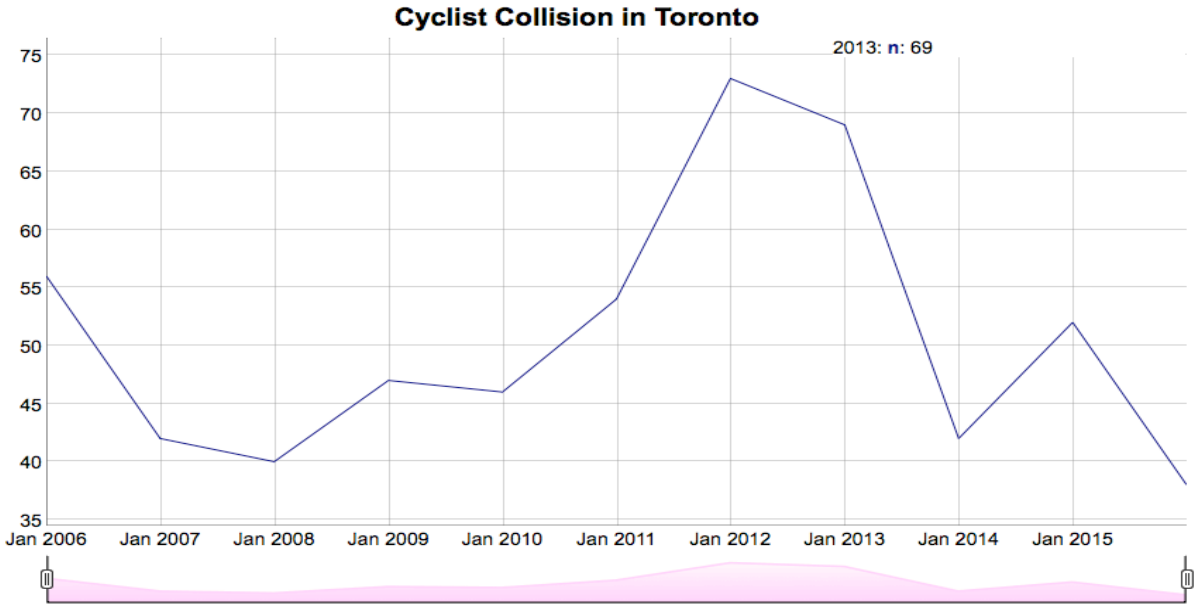

Figure 8 10 Year Trend of Cyclist Collision

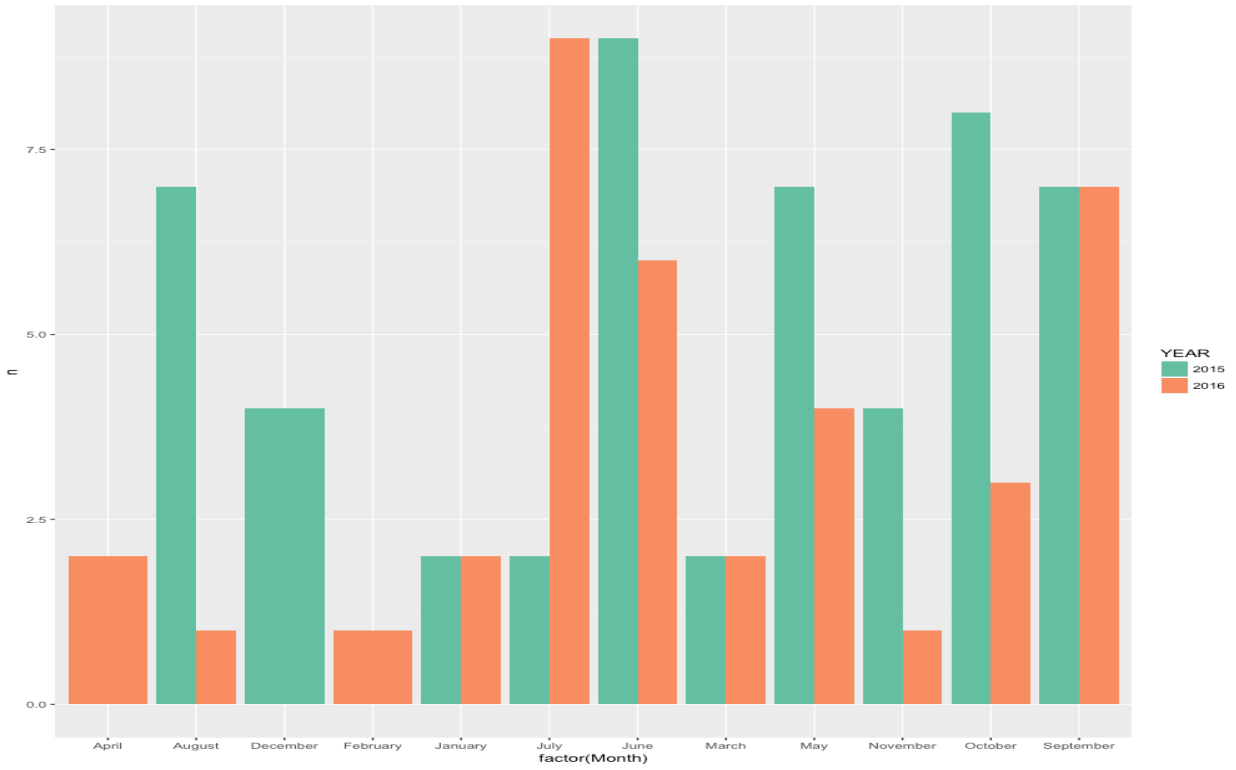

Figure 9 Total KSI by Month of Pedestrian Collision

c) Automobile involved in collisions

View traffic related collisions data involving Automobile. These events include any serious or fatal collision where a automobile is involved. And from this graph we can see there is a decreasing trend of collision in Toronto.

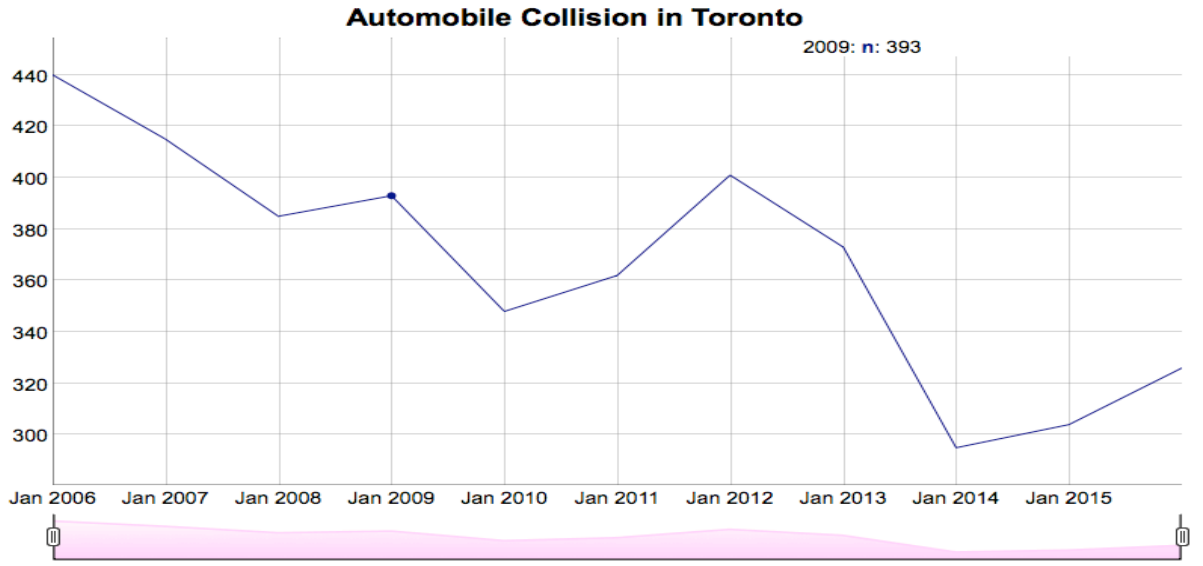

Figure 10 10 Year Trend of Automobile Collision

d) Motorcyclists involved in collisions

View traffic related collisions data involving Motorcyclists. These events include any serious or fatal collision where a motorcyclist is involved. And from this graph we can see there is a increasing trend of collision in Toronto.

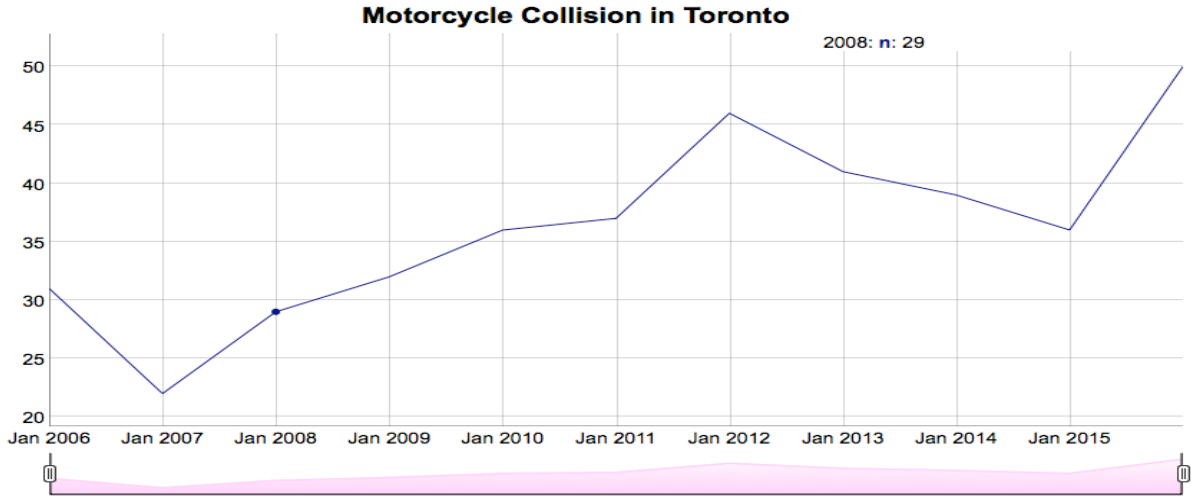

Figure 11 10 Year Trend of Motorcycle Collision

e) Truck involved in collisions

View traffic related collisions data involving Truck. These events include any serious or fatal collision where a truck is involved. We can see there is an increasing trend of collision in Toronto from the graph.

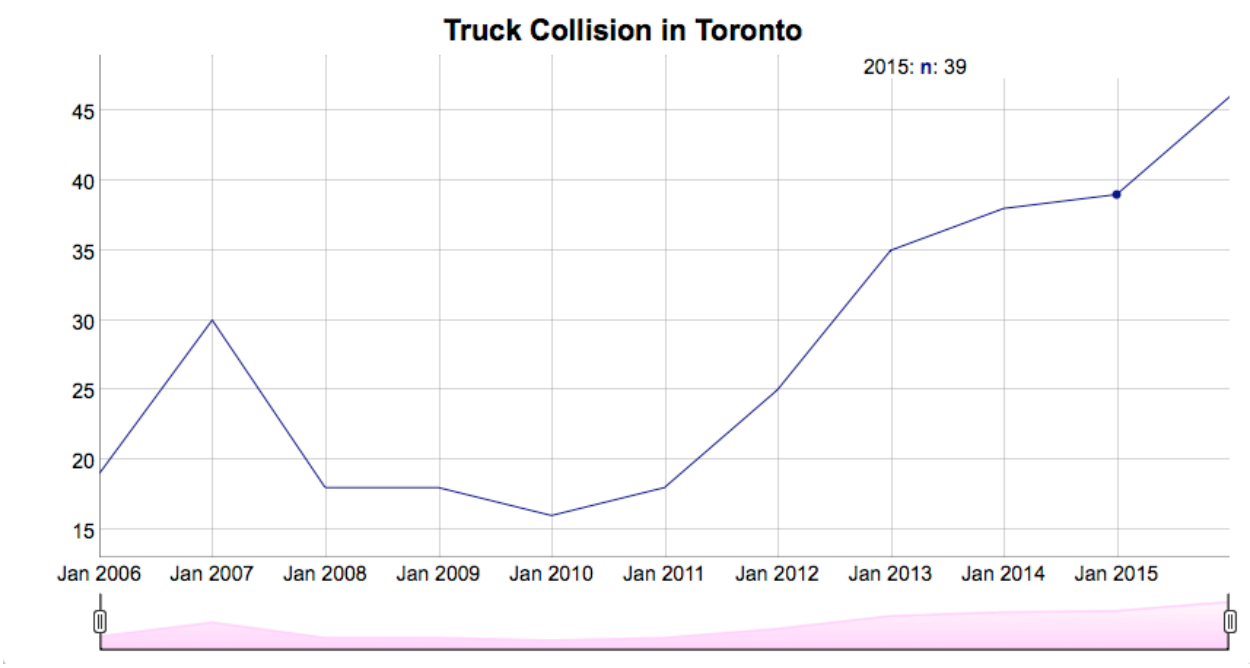

Figure 12 10 Year Trend of Truck Collision

II. Visualize some events that should happen in future days

a) Year and Month Wise Collision Plot

Year and Month wise Collision Pattern Plot in Toronto

I can see from the plot time wise collision distribution in Toronto. Also from the graph we can see there is a seasonality pattern. If we can see most of the accidents are occurring from June to October month.

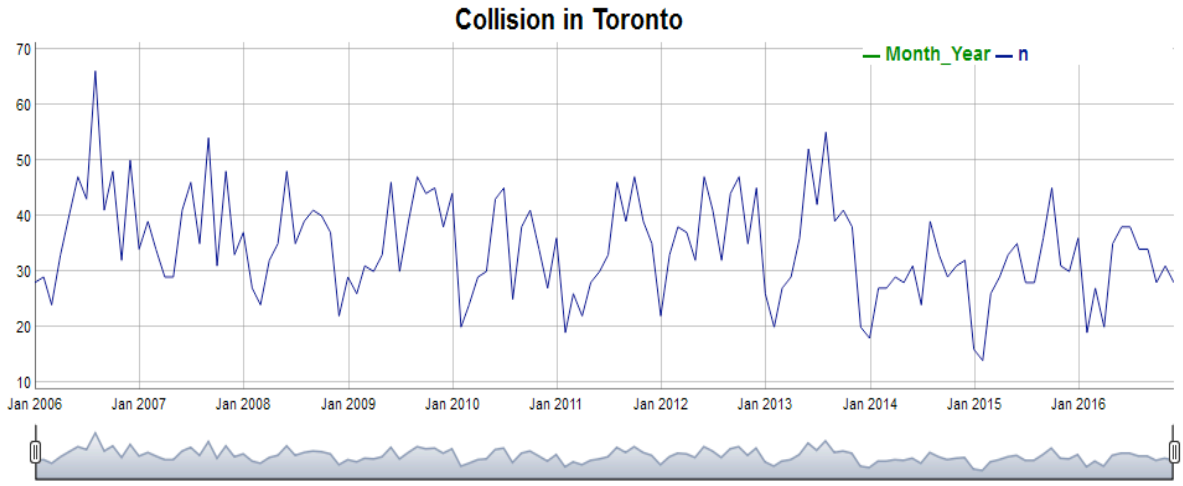

Figure 13 Year and Month Wise Collision Plot

b) Year Wise Collision Plot

Year wise Collision Pattern Plot in Toronto

I can see from the below graph year wise collision pattern. And from this graph we can see there is a decreasing trend of collision in Toronto. We can say now the collision rate is lesser compare to the previous years.

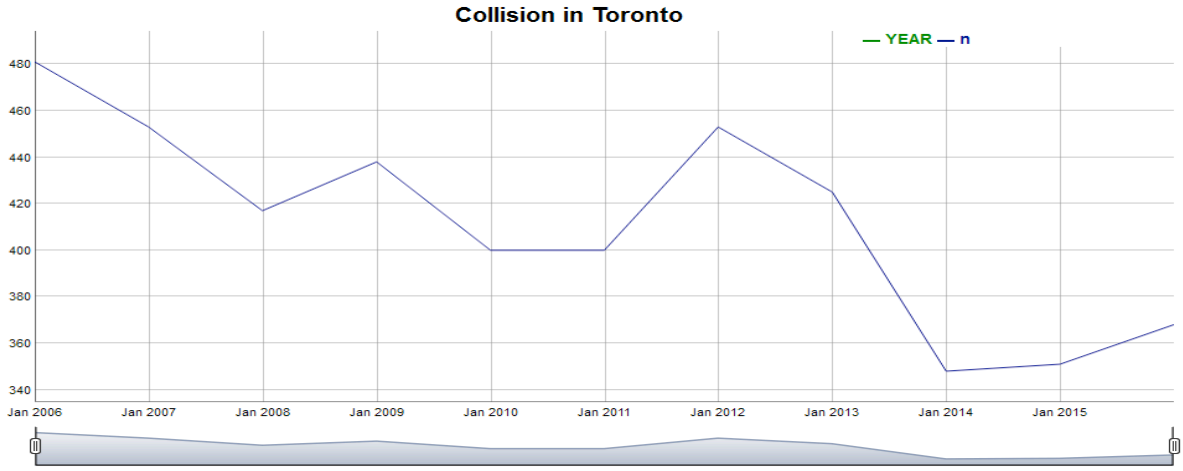

Figure 14 Year Wise Collision Plot in Toronto

c) Monthly Seasonal Index Plot

Month wise Collision Seasonal Index Plot in Toronto

From the below chart we can clearly see there is a seasonal pattern of accident is present. From June to October we can see significantly higher pattern of collision and decreasing pattern of collision from November. Collisions are significantly less in the month of February, March and in April.

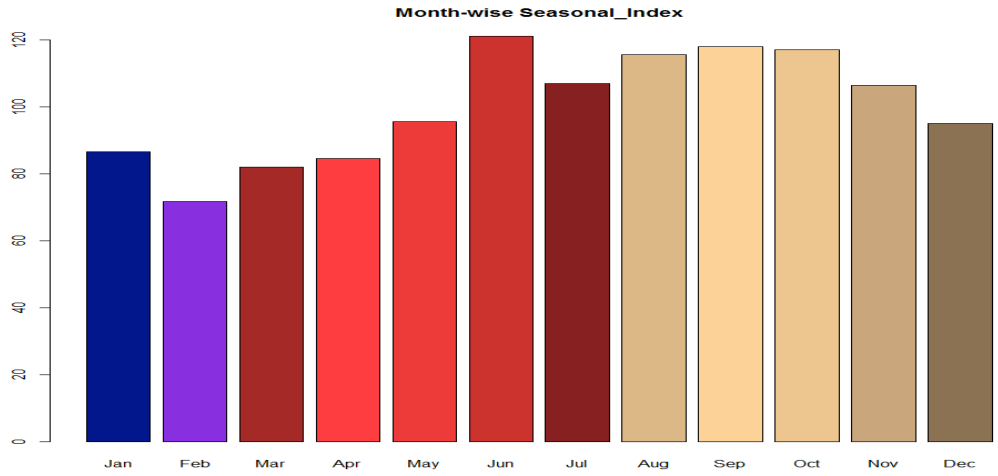

Figure 15 Monthly Seasonal Index Plot

d) Hour wise Collision Pattern Plot

Hour wise Collision Plot in Toronto

From the below graph we can see hour wise collision pattern. In which time of a day collisions are mostly occurred. From this graph we can see a significantly higher rate of collisions occurring from 13th hour of the day to 20th hour of the day and then collision rate is decreasing. And from the 0th hour to 11th hour of a day part collision rate is very less and from 12th hour the collision rate is increasing. The highest collisions are happening on 18th hour.

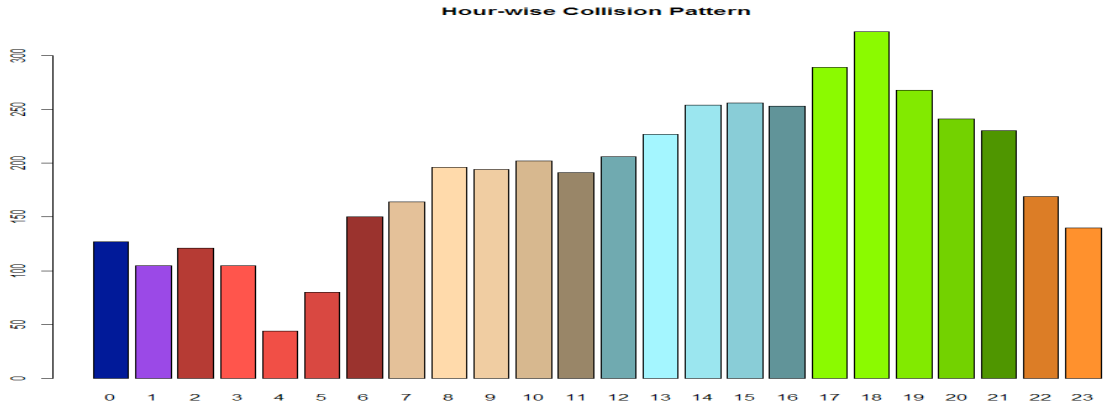

Figure 16 Hour-wise Collision Plot

IV. RESULTS AND DISCUSSION

Prediction of collisions in future days or months

I am trying to forecast possible number of collision in future i.e., for next 12 months and used ARIMA model to forecast the future collision. Black line is considering the historical collision part and the red line signifies the forecasted number of collision. In this graph I have captured historical data from January 2006 to December 2016.

I have run some tests like ACF (Autocorrelation Function), PACF (Partial Autocorrelation Function) and ADF test (Augmented Dickey-Fuller Test). From the ACF and PACF test I am able to decide the Moving Average (MA) parameter and Auto Regressive (AR) parameter. And from the ADF test I am able to decide the stationary cut-off, in which difference is suitable to make the series stationary. After running those tests I have decided that ARIMA model will be good fit for this series.

I have checked the non-stationary in variance. The assumption of a time series model is that the series must be stationary first. That means each subset of a series have similar mean, variance and correlation will be high. In this data I have seen original data is not stationary after checking the ADF test. Then I had seen at first order difference the data become stationary, i.e., similar means, variance and strong correlation between different subset of that series.

Here I have used ARIMA but I have also used seasonal index as a regressor variable. That's why I am able to capture the seasonality and trend. Here seasonal index means average monthly collision pattern that I have calculated from 10 years collision data. So, the model is not only capturing the trend and also able to capture the seasonality too.

(a)Time Series Forecasting

After running the ARIMA model to forecast the future number of collision should happen in next 12 months, then I am getting the result from the model. Here I have used 80% data set as training and 20% data set as test. I am getting 83% accuracy on training data and 85% accuracy on testing data. In the below 1st table I am showing the models actual time series vs forecasted series. In the forecasted series contains 27 and 12 future points i.e. year and month combination, that means total 39 forecasted points and in training it contains 106 points. In the first table the black line indicates the history period and red line indicating forecasted period collision. And the data points are in monthly level.

Actual and Forecasted Combined plot

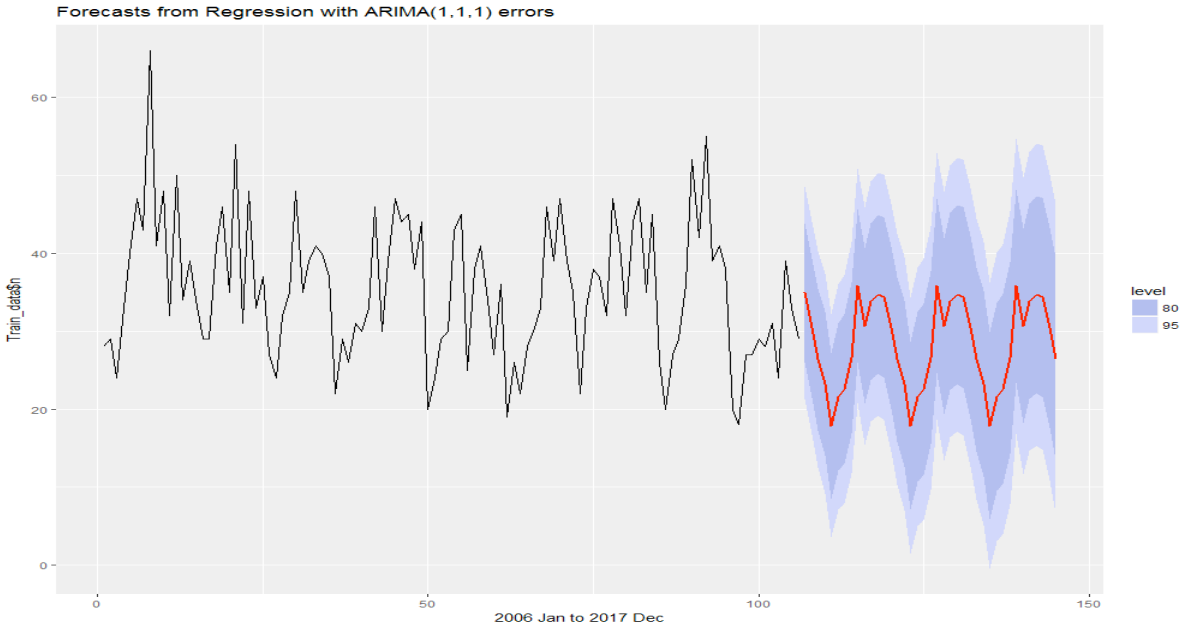

Figure 17 Actual and Forecasted Combined plot

(b) Time Series Model Performance

In the below plot I am trying to show actual collisions Vs model prediction for collisions. Blue line showing the actual collisions and the red line is showing model prediction of collision. In actual data we can see some certain spikes but my model is not able to capture the unexpected spikes.

From the below graph I can see actual collision and forecasted collision for training from January 2006 to October 2014, testing from November 2014 to December 2016 and forecasted series from January 2017 to December 2017. I have highlighted the training and testing part in the figure.

Actual Vs. Forecasted Time Series Plot

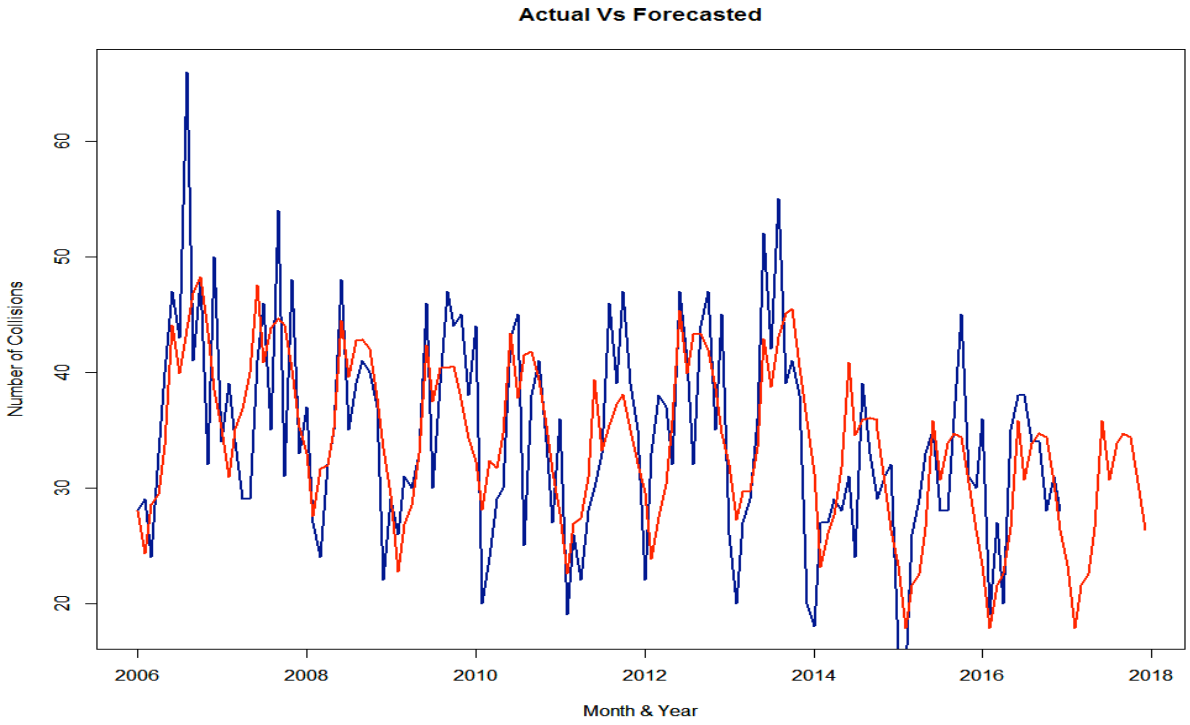

Figure 18 Actual Vs. Forecasted Time Series Plot

In the below table is showing the actual and Forecasted Collision. I have captured some part of result only. APE (Absolute Percentage Error) is given and from the APE we can explain about the % deviation from actual to forecasted result.

It is shown that for a particular year and month from September 2015 to December 2016 there is actual collision, forecasted collision and error percentage from the time series model. And forecasted collision from January 2017 to December 2017 is obtained that mean in every month within year 2017, there is possibility of happening number of collisions.

Table 15 Actual and Forecasted Result

Month_Year	Total Collisions	Forecasted Collisions	flag	APE
9/1/15	36	35	Test	4%
10/1/15	45	34	Test	24%
11/1/15	31	30	Test	2%
12/1/15	30	26	Test	12%
1/1/16	36	23	Test	35%
2/1/16	19	18	Test	6%
3/1/16	27	22	Test	20%
4/1/16	20	23	Test	13%
5/1/16	35	27	Test	24%
6/1/16	38	36	Test	6%
7/1/16	38	31	Test	19%
8/1/16	34	34	Test	1%
9/1/16	34	35	Test	2%
10/1/16	28	34	Test	23%
11/1/16	31	30	Test	2%
12/1/16	28	26	Test	6%
1/1/17		23	Forecast	
2/1/17		18	Forecast	
3/1/17		22	Forecast	
4/1/17		23	Forecast	
5/1/17		27	Forecast	
6/1/17		36	Forecast	
7/1/17		31	Forecast	
8/1/17		34	Forecast	
9/1/17		35	Forecast	
10/1/17		34	Forecast	
11/1/17		30	Forecast	
12/1/17		26	Forecast	

Here I have also captured the summary result of ARIMA model in R where I am getting satisfactory result. And the model is ARIMA(1,1,1) model and overall MAPE is 17% i.e., error percentage. That shows 100%-17% = 83% accuracy rate.

Predict severity of injury in traffic accidents

Here I have build 2 machine learning models to identify the major factors for injury type. After building those models we can see Random Forest model is most efficient and produce best result and getting almost 90% Accuracy. But from the Decision Tree model I am getting 65% accuracy for both training and testing set.

I have mentioned all the confusion matrices for both models that give actual injury type vs predicted injury type. From Decision Tree Training table, Actual Injury having major factor is 0+2322+0+0+31+741=3094. Predicted Injury having Major factor is 341+3222+277+412+1+517=4770. Now we can say perfectly classified for Actual Major is 2322 out of 3094, 31 is misclassified with Missing and 741 is misclassified with None. So the accuracy for Actual Major is 2322/3094= 75% i.e. 25% is error.

Decision Tree Results

Table 16 Decesion Tree Training

Decision Tree Training							
	Predicted						
Actual	Fatal	Major	Minimal	Minor	Missing	None	Error
Fatal	0	341	0	0	0	44	100
Major	0	2322	0	0	31	741	25
Minimal	0	277	0	0	14	266	100
Minor	0	412	0	0	16	312	100
Missing	0	1	0	0	1136	10	1
None	0	517	0	0	289	2492	24.4

Table 17 Decesion Tree Testing

Decision Tree Testing							
	Predicted						
Actual	Fatal	Major	Minimal	Minor	Missing	None	Error
Fatal	0	107	0	0	0	10	100
Major	0	479	0	0	5	147	24.1
Minimal	0	48	0	0	5	59	100
Minor	0	81	0	0	2	53	100
Missing	0	1	0	0	229	0	0.4
None	0	120	0	0	64	565	24.6

From Random Forest Training table, Actual Injury having major factor is 0+3003+7+16+14+54=3094. Predicted Injury having Major factor is 118+3003+49+74+1+57=3302. Now we can say perfectly classified for Actual Major is 3003 out of 3094, 7 is misclassified with Minimal, 16 is misclassified with Minor, 14 is misclassified as Missing and 54 is misclassified with None. So the accuracy for Actual Major is 3003/3094= 97.05% i.e. 2.9% is error.

Random Forest Results

Table 18 Random Forest Training

Random Forest Training							
	Predicted						
Actual	Fatal	Major	Minimal	Minor	Missing	None	Error
Fatal	254	118	0	7	0	6	34
Major	0	3003	7	16	14	54	2.9
Minimal	1	49	442	11	12	42	20.6
Minor	0	74	4	600	9	53	18.9
Missing	0	1	0	0	1138	8	0.8
None	0	57	6	10	177	3048	7.6

Table 19 Random Forest Training

Random Forest Testing							
Actual	Predicted						
	Fatal	Major	Minimal	Minor	Missing	None	Error
Fatal	3	106	1	3	0	4	97.4
Major	3	499	5	16	4	104	20.9
Minimal	0	36	7	11	5	53	93.8
Minor	3	66	5	20	1	41	85.3
Missing	0	1	0	0	226	3	1.7
None	0	96	5	14	59	575	23.2

Unbalanced data a problem

Most machine learning classification algorithms are sensitive to unbalance in the predictor classes. In TPS dataset, we have 588 Fatal, 4340 Major, 811 Minimal, 1039 Minor, 4792 None, 1603 Missing samples in INJURY type. A model that has been trained and tested on such a dataset could now predict "None" for all samples and still gain a very high accuracy. An unbalanced dataset will bias the prediction model towards the more common class.

To balance data for modeling

The basic theoretical concepts behind over sampling is very simple: we randomly duplicate samples from the class with fewer instances or we generate additional instances based on the data that we have, so as to match the number of samples in each class. While we avoid loosing information with this approach, we also run the risk of over-fitting our model as we are more likely to get the same samples in the training and in the test data, i.e. the test data is no longer independent from training data. This would lead to an overestimation of our model's performance and generalizability.

In reality though, we should not simply perform over-sampling or under-sampling on our training data and then run the model. We need to account for cross-validation and perform over-sampling or under-sampling on each fold independently to get an honest estimate of model performance.

a) Modeling the original unbalanced data

I randomly divide the data into training and test sets (stratified by class) and perform Random Forest modeling with 10 x 10 repeated cross-validation. Final model performance is then measured on the test set.

b) Modeling the over-sampling data

From Decision Tree table,

Confusion Matrix and Statistics

Table 20 Decesion Tree Training Actual Result

	Reference					
Prediction	Fatal	Major	Minimal	Minor	Missing	None
Fatal	219	717	30	46	0	40
Major	103	1235	12	38	0	13
Minimal	38	604	330	315	0	867
Minor	51	417	126	271	0	139
Missing	0	0	0	0	1104	225
None	1	65	70	58	19	2071

Overall Statistics

Accuracy : 0.567

Statistics by Class:

	Class: Fatal	Major	Minimal	Minor	Missing	None
Sensitivity	0.53155	0.4065	0.58099	0.37225	0.9831	0.6173
Balanced Accuracy	0.71851	0.6898	0.68513	0.64299	0.9777	0.7905

From the Decision Tree Training table, Actual Injury having Fatal class is 219+103+38+51+0+1=412. Predicted Injury having Fatal class is 219+717+30+46+0+40=1052. Now we can say perfectly classified for Actual Fatal is 219 out of 412, 103 are misclassified with Major, 38 are misclassified with Minimal, 51 are misclassified as minor and 1 is classified as None.

From Random Forest Training table,

Confusion Matrix and Statistics

Table 21 Random Forest Training Actual Result

	Reference					
Prediction	Fatal	Major	Minimal	Minor	Missing	None
Fatal	301	704	20	40	0	25
Major	63	1520	14	39	0	43
Minimal	17	356	408	157	0	427
Minor	31	393	88	461	0	302
Missing	0	0	0	0	1116	227
None	0	65	38	31	7	2331

For Accuracy %

Table 22 Random Forest Training % Accuracy

	Reference					
Prediction	Fatal	Major	Minimal	Minor	Missing	None
Fatal	3.263	7.621	0.217	0.434	0.000	0.271
Major	0.683	16.490	0.152	0.423	0.000	0.466
Minimal	0.184	3.859	4.412	1.702	0.000	4.629
Minor	0.336	4.261	0.965	4.998	00.000	3.274
Missing	0.000	0.000	0.000	0.000	12.099	2.461
None	0.000	0.705	0.412	0.336	0.076	25.271

Overall Statistics

Accuracy : 0.6653

Statistics by Class:

	Class: Fatal	Major	Minimal	Minor	Missing	None
Sensitivity	0.73058	0.5003	0.71831	0.63324	0.9938	0.6948
Balanced Accuracy	0.82052	0.7373	0.80388	0.76872	0.9829	0.8354

From Random Forest Training table, Actual Injury having Fatal class is 301+63+17+31+0+0=412. Predicted Injury having Fatal class is 301+704+20+40+0+25=1090. Now we can say perfectly classified for Actual Fatal is 301 out of 412, 63 are misclassified with Major, 17 are misclassified with minimal, 31 are misclassified as minor. So the accuracy for Actual Fatal is 301/412= 73.05%.

From Random Forest Testing table,

Confusion Matrix and Statistics

Table 23 Random Forest Testing Actual Result

	Reference					
Prediction	Fatal	Major	Minimal	Minor	Missing	None
Fatal	73	338	8	30	0	21
Major	63	565	26	32	0	50
Minimal	12	174	93	107	0	204
Minor	24	175	73	105	0	178
Missing	0	0	0	0	475	95
None	4	50	43	37	5	889

For Accuracy %

Table 24 Random Forest Testing % Accuracy

	Reference					
Prediction	Fatal	Major	Minimal	Minor	Missing	None
Fatal	1.849	8.559	0.203	0.760	0.000	0.532
Major	1.595	14.307	0.658	0.810	0.000	1.266
Minimal	0.304	4.406	2.355	2.710	0.000	5.166
Minor	0.608	4.432	1.849	2.659	00.000	4.507
Missing	0.000	0.000	0.000	0.000	12.028	2.406
None	0.101	1.266	1.089	0.937	0.127	22.512

Overall Statistics

Accuracy : 0.5571

Statistics by Class:

	Class: Fatal	Major	Minimal	Minor	Missing	None
Sensitivity	0.41477	0.4339	0.38272	0.33762	0.9896	0.6186
Balanced Accuracy	0.65478	0.6847	0.62430	0.60696	0.9811	0.7817

From Random Forest Testing table, Actual Injury having Fatal class is 73+63+12+24+0+4=176. Predicted Injury having Fatal class is 73+338+8+30+0+21=470. Now we can say perfectly classified for Actual Fatal is 73 out of 176, 63 are misclassified with Major, 12 are misclassified with Minimal, 24 are misclassified as Minor, 4 are misclassified as None.

The data analysis results shows that the random forest returns an accuracy of 67% and decision tree returns an accuracy 57% after balancing the data for the classification of injury type.

Variable Importance

The variable importance table shows the most important factor for severity of injury.

Table 25 Variable Importance for Injury

Variables	Importance
INVTYPE	1897.172027
IMPACTYPE	909.155131
VEHTYPE	212.4665612
INVAGE	131.656953
DRIVACT	91.8237658
DRIVCOND	20.6270778
TRAFFCTL	7.3357375
LIGHT	6.9938604
MANOEUVER	4.3698221
VISIBILITY	1.2616529
RDSFCOND	0.7866979
ROAD_CLASS	0.3730606

From the variable importance table we can see INVTYPE is the most important factor for severe injury. Also we can see IMPACTYPE, VEHTYPE, INVAGE, DRIVACT have significant effect on injury type. On the basis of variable importance we can say what could be the severity of injury type i.e.; fatal, major, minimal, minor and none.

V. CONCLUSION AND FUTURE WORK

From the time series analysis using ARIMA, a collision pattern for the historical data is observed and collision forecast is illustrated. Here, accuracy of 85% is achieved for prediction of collisions, which is quite significant in this model. This result reflects the dependency of collision with previous collision with the time value. Also from the second model based on decision tree and random forest it is observed that the severity of injury is clearly explained by variables like Involvement Type, Impact Type, and Vehicle Type, Age of Involved Party, Driver Action etc. The variable Involvement Type (INVTYPE) is found to be most important variable for the severe injury. Under this model, I am able to identify the key factors related to severity of injury leading to the injury classes like fatal, major, minimal, minor and none. The data analysis results shows that the random forest returns an accuracy of 90% for the classification of injury while the decision tree results 65% of accuracy on the injury type. But after balancing the data the random forest returns an accuracy of 67% while the decision tree results 57% of accuracy for the classification of injury type.

In future, Recurrent Neural Network and Hidden Markov Models can be used to achieve better level of accuracy on risk prediction of collisions over the time series data. Similarly decision tree method C5.0 can also be used to classification of injury types.

REFERENCES

[1] Meng Chen, Xiaohui Yu, Yang Liua, Mining moving patterns for predicting next location, ELSEVIER, (2015).

[2] Tatiana Tambouratzis, Dora Souliou, Miltiadis Chalikias, Andreas Gregoriades, Maximizing accuracy and efficiency of traffic accident prediction combining information mining with computational intelligence approaches and decision trees, JAISCR, (2014), Vol. 4, No. 1, pp. 31- 42

[3] Lee S Friedman, Paul Barach, Elihu D Richter, Raised speed limits, case fatality and road deaths: a six year follow-up using ARIMA models, Injury Prevention 13.3 (2007): 156-161.

[4] Miao Chong, Ajith Abraham and Marcin Paprzycki, Traffic Accident Analysis Using Machine Learning Paradigms, Informatica 29 (2005) 89–98

[5] LAW, T.H. RADIN UMAR, R.S. WONG, S.V. ,The Malaysian government's road accident death reduction target for year 2010, IATSS research 29.1 (2005): 42-49.

[6] Rami Harb, Xuedong Yan, Essam Radwan, Xiaogang Su, Exploring precrash maneuvers using classification trees and random forests, Accident Analysis & Prevention 41.1 (2009): 98-107.

[7] ParaskeviMichalaki, Mohammed Quddus, David Pitfield, Andrew Huetson, A time-series analysis of motorway collisions in England considering road infrastructure, socio-demographics, traffic and weather characteristics, Journal of Transport & Health 3.1 (2016): 9-20.

[8] S.Shanthi, R.GeethaRamani, Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms, International Journal of Computer Applications 35.12 (2011): 30-37.

[9] Olutayo V.A, Eludire A.A, Traffic Accident Analysis Using Decision Trees and Neural Networks, International Journal of Information Technology and Computer Science (IJITCS) 6.2 (2014): 22.

[10] George Yannis ,AnastasiosDragomanovits , Alexandra Laiou , Thomas Richter ,Stephan Ruhl , Francesca La Torre , Lorenzo Domenichini ,Daniel Graham , NioviKarathodorou , and Haojie Li, Use of accident prediction models in road safety management – an international inquiry, Transportation Research Procedia 14 (2016) 4257 – 4266.

[11] http://data.torontopolice.on.ca/datasets/ksi/data