Predicting Traffic Accident Risk and Severity

Cristina Giraldo and Taisha Ferguson

George Washington University

Data Science Capstone

DATS 6501-11

April 30, 2020

Table of Contents

Glossary of Terms

Machine Learning - Machine learning modeling is statistical modeling using computer algorithms that use sample data to learn patterns and relationships in order to make predictions (Machine Learning, 2020).

Deep Learning – Branch of Machine Learning that uses Artificial Neural Networks (ANN) models that are inspired by the biological neural networks (Deep Learning, 2020).

Time Series – Data that is ordered by time (Time Series, 2020).

Traffic Accident Severity Prediction – The prediction of the level of severity of traffic accidents. In this research there are two levels of severity: 1) no injury 2) at least one person involved in accident had some injury.

Traffic Accident Risk Prediction – Prediction of the likelihood of traffic accident occurring. In the scope of this research we are predicted daily risk of a traffic accident.

MSE – The mean of the squared error of the predicted values versus actual values.

RMSE – Square root of the mean squared error.

Introduction

Traffic accidents are responsible for 1.25 million deaths worldwide and is the fifth leading cause of death in the US (Global Status Report on Road Safety 2015, 2015). Injuries and costs associated with traffic accidents have a large impact on the individuals involved and the communities they live in. While there is a lot of current research on different modeling techniques in the Machine Learning and Deep Learning disciplines, there is still a need for further exploration in how these modeling techniques compare to one another when predicting accident risk and severity.

This project aims to test several machine learning, time series, and deep learning models in order to decipher which modelling techniques have the best outcomes for predicting traffic accident risk and severity. Armed with the best modelling techniques, stakeholders such as city officials, insurance companies, and hospitals can make more informed and perhaps improved decisions related to traffic accident preparedness. The remainder of this article is as follows. The second section is a discussion of current research related to this area of study. The third section discusses methodology including a description of the dataset and data preprocessing techniques. The fourth section discusses the results of the different modeling techniques and the final section summarizes our conclusions and ideas on further research.

Literature Review

Traffic accident risk and prediction is currently a very active area of study. Much of the research focuses on either risk prediction or severity prediction. The distinction between risk and severity prediction dictates the type of modelling techniques used to solve the problem. Traffic accident severity is a classification problem where the predicted outcome may belong to two or more categories. In 2017 researchers compared 4 different machine learning models to determine which method achieved better accuracy when predicting severity (Iranitalab, 2017). In their paper a custom measurement technique was proposed, incorporating the actual costs associated with the crashes in order to measure the performance of each model. The results of the analysis demonstrated that different models can be better at predicting different levels of severity. In similar research, a comparison of the machine learning model Support Vector Machines (SVM) and the neural network Multilayer Perceptron (MLP) was done showing that SVM outperformed MLP for predicting accident severity (Pradhan, 2020). Another group of researches demonstrated the prediction power of the Convolution Neural Network (CNN) when predicting accident severity (M. Zheng, 2019).

The prediction of traffic accident risk is the prediction of the likelihood of a traffic accident occurring. Unlike the classification of severity, the prediction of risk is continuous and therefore has different modeling strategies associated with its prediction. Researchers in China compared machine learning, time series, and deep learning methods for the prediction of traffic accident risk and found that the deep learning model LSTM was the most effective method for predicting traffic accident risk (H. Ren, 2018). Similarly, another group of researchers used LSTM modeling for the prediction of traffic accident risk and explored the use of spatial dependence of

traffic accidents in their modeling (Amir Bahador Parsa, 2019). It is the aim of this project to compare several machine learning and deep learning methods discussed in this review as well as additional methods that were not explored in their research such as the machine learning algorithm CatBoost and the time series method Holt-Winter.

Methodology

The analysis conducted in this paper is based on traffic accident data from the City of Chicago. That data and information was obtained from the city website data.cityofchicago.org. This section discusses the different data and modelling methods used in the prediction of severity and risk of traffic accidents.

Data Description

The dataset used for this report is a collection of traffic accidents in the Chicago metropolitan area collected from March 2013 to March 2020. The dataset lists over 300,000 observations and includes information on crash date, primary contributory cause, first crash type, damage, number of units involved in a crash, day of the week, month, hour, weather, road surface, most severe injury, longitude, latitude, sex and age. Crash date includes the date and the time of the crash. Primary contributory cause contains the main cause of the crash. First crash type indicates type of collision. Damage contains the values caused by the accidents. Road surface indicates if it was wet, dry, snowy or with other conditions. Most severe injury indicates the severity of the accident as fatal, incapacitating, or non-incapacitating. And longitude and latitude refer to the location of the crash.

Data Collection

The dataset contains information for all areas under the jurisdiction of the Chicago police department (CDP). When a crash occurs, the police record the traffic accident by using software called E-Crash, which is an electronic crash reporting system. In this system, the CDP records all the information available about the accident, such as speed, weather condition, day, hour and month of the crash, number of vehicles involved, and other important variables about the traffic accident, except personal identifiable information. Once the report is saved, E-Crash interfaces with data.cityofchicago.org to post the information about the traffic accident and make the information publicly available (Chicago, 2020). The information on data.cityofchicago.org can be visualized on the website but is also posted in CSV format and available to download. That data is divided into three datasets: people, crashes and vehicles. The description about the features contained in each dataset can be found in the Appendix 1 of this paper.

Data Preprocessing

SEVERITY PREDICTION

In preparation for modeling traffic accident severity, three datasets from the city website were joined. After joining the datasets, there were 148 features and more than 800,000 data samples.

A review of the data revealed that data the contained crash records involving vehicles, pedestrians, and bicycles, and animals. Our analysis also revealed high cardinality among the variables. For example, the feature “contributory cause” contained 19 different values and “road surface” showed six different conditions. For purposes of our project, we limited the dataset to

crashes involving just two vehicles. Additionally, features with high cardinality were reduced. For example, “contributory cause” was reduced from 19 categories to six. This was done by combining similar categories.

Additional formatting was done to variables at the person and vehicle levels in order to merge with accident event. For example, a new category for “sex” was created to capture the sex of both drivers. The categories were: both male, both female, both other, male female, male other and female other. Furthermore, features with no data were dropped from the dataset.

RISK PREDICTION

For the prediction of daily traffic accident risk, daily counts were generated between the time period of January 1, 2017 through December 31, 2019. Once the daily counts were generated, they were further split into weekly increments for modelling and prediction comparison. The risk modelling was split into 80% for training and 20 % for testing.

Exploratory Data Analysis

SEVERITY PREDICTION

Once the dataset was formatted for modelling, an EDA was performed to help gain additional insights. The EDA revealed some variables were highly correlated and some of the columns had missing values, as shown in *Figure 1* below.

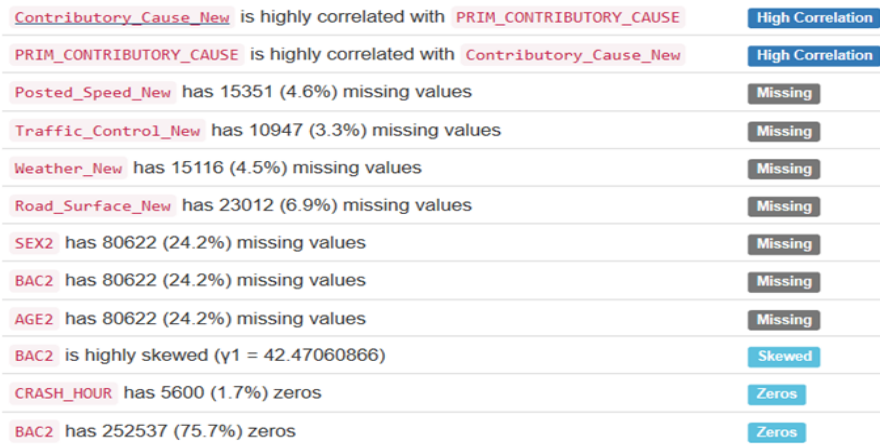

Figure 1. Data Exploratory Analysis

For the highly correlated data, one of the redundant columns was dropped. To process the missing values, the dataset was populated according to other features and was then averaged. For instance, the column age was populated according to sex.

The EDA also revealed that the distribution of the variable to predict was imbalanced, as shown in *Figure 2* below.

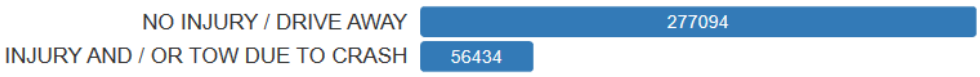

Figure 2. Imbalanced Target - Crash Type

Likewise, it was noticed that other columns provided extra information according to their distribution. For example, the feature “crash hour” indicated to us that most accidents tend to occur between 3 p.m. and 5 p.m. This distribution can be seen in *Figure 3* below.

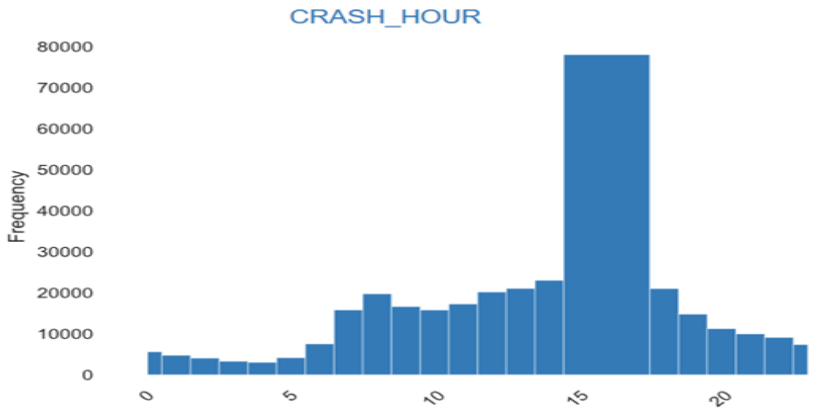

Figure 3. Crash Hour Distribution

Moreover, when feature importance was performed, we observed that this variable "crash type" along with "primary contributory cause" and "first crash type" were of more relevance than the others. *See Figure 4* below.

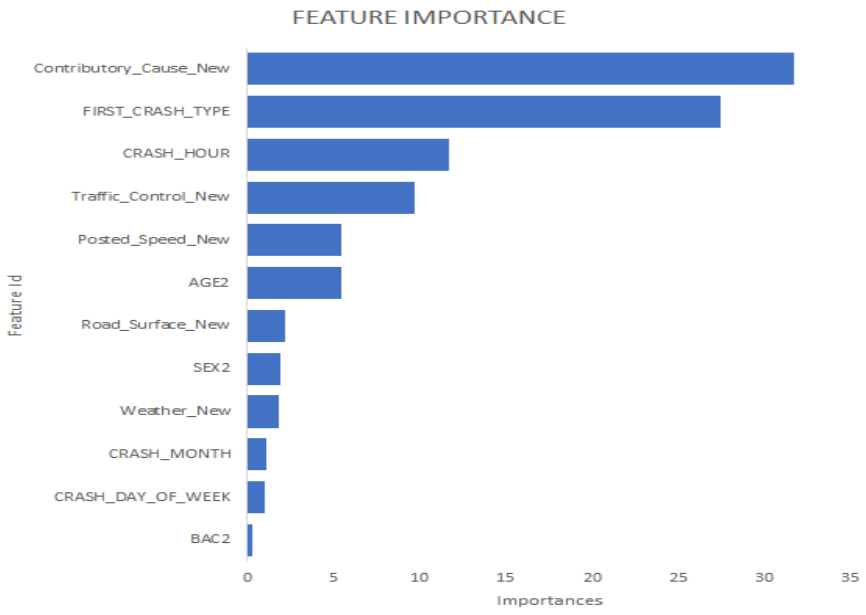

Figure 4. Feature Importance

RISK PREDICTION

In preparation for modelling traffic accident risk, the following graphs were created: time plot of daily values versus time (*Figure 5*), autocorrelation of daily count lags (*Figure 6*), and seasonal decomposition (*Figure 7*). *Figure 5* and *Figure 7* show that there is a small linear trend in the dataset between 2017 and 2018. *Figure 7* also shows there is a weekly trend in the data. *Figure 7* also reveals some of the weekly seasonality in the autocorrelation at the different lag values. The linear trend and seasonality components will be addressed in the modeling of the data.

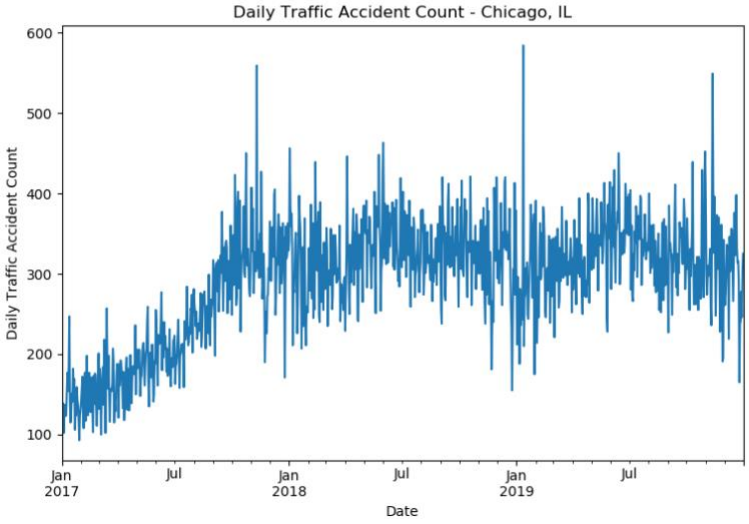

Figure 5. Daily Traffic Accident Counts

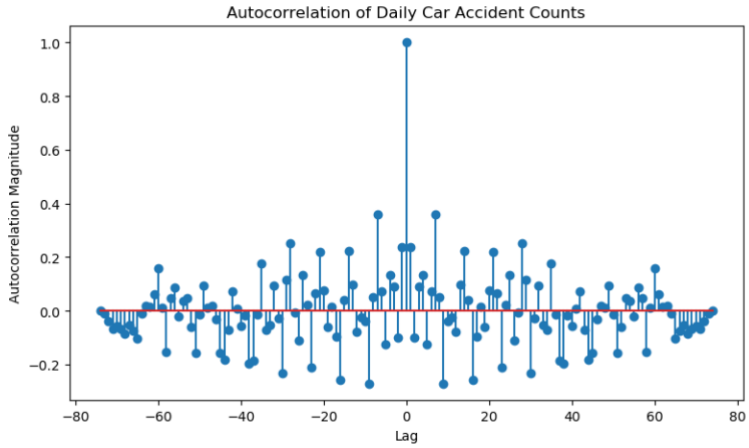

Figure 6. Autocorrelation of Daily Accident Counts

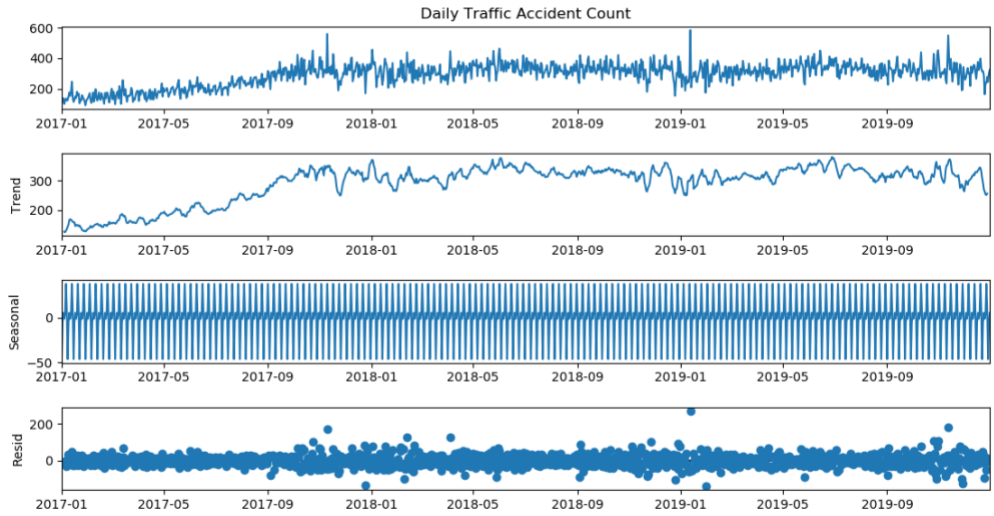

Figure 7. Seasonal Decomposition

Data Modelling and Visualizations

SEVERITY PREDICTION

For severity prediction, several algorithms were implemented, specifically Classical Machine Learning algorithms and Deep Learning algorithms. For Classical Machine Learning algorithms, the data was split into 70 percent training and 30 percent testing.

Among the Classical Machine Learning algorithms we applied were Supporting Vector Machines (SVM), Random Forest, Key Nearest Neighbors (KNN) and Logistic Regression. However, the most relevant results came from XGBoost and CatBoost, which were implemented as well. Nonetheless, the analysis in this section was performed by using CatBoost rather than XGBoost given that it provides several advantages, as follows:

1. "CatBoost is a high-performance open source library for gradient boosting on decision trees" (Yandex, 2020). This quality makes CatBoost easier to understand.
2. CatBoost allows you to obtain good results without fine-tuning hyperparameters.
3. CatBoost helps to reduce overfitting given that it uses symmetrical trees, ordered boosting and random permutations.
4. CatBoost permits using categorical features without one hot encoding them.
5. CatBoost can be run in GPU which helps to speed up the model training.

Since, the target was highly imbalanced, we applied an oversampling method to help balance the predictive class and, therefore, obtain better predictions. The approach used to increase the minority class when the information is a combination of categorical and nominal data is called SMOTENC.

SMOTENC is an approach where synthetic data is created in order to perform data augmentation and thus balance the classes (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

RISK PREDICTION

For the prediction of traffic accident risks, we compared the modeling techniques of the Average method, the Holt-Winter method, Seasonal Autoregressive Integrated Moving Average (SARIMA), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM). The details of each model technique are explained below.

Average Method

The Average method is a baseline time series modeling technique that can be used as a benchmark for more sophisticated modeling techniques. The Average method takes the average of the test dataset and uses that as the prediction for all for future values. For this dataset the average value is 282.93. The RMSE for the test set predictions using the Average method was 66.87.

Holts-Winter Method

Holt-Winter method is a time series forecasting method that is an extension of the Simple Exponential Smoothing (SES). Exponential smoothing is method that uses the weighted averages of previous timesteps, and the weights decrease exponentially as the timesteps move further toward past time steps. The Holt-Winter method, unlike SES, accounts for seasonality and linear trends. The formula for the additive Holt-Winter method is shown below.

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$$
$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

Where $\widehat{y_{t+h|t}}$ is the predicted value with lag h, l_t is the equation for the level factor, b_t is the equation for the linear trend, and s_t is equations for seasonality

The period frequency of 7 was chosen due to the weekly seasonality discovered in the Exploratory Data Analysis. After the model parameters were estimated using the Holt-Winter function from the Stats model library, the model was tested on the test data. The RMSE of the test set was 47.55 which is significantly better than Average method.

SARIMA Model

Seasonal Autoregressive Integrated Moving Average (SARIMA) is a time series model used to forecast future values of time series data. The seasonal component indicates the period of the seasonality, AR is estimated using the prior values of the dependent variable, I represents the degree of integration, MA is based on prior error terms, and SARIMA models is a combination of all three. The general form for an ARIMA(n_a, d, n_b) is shown below:

$$(1 + a_1 q^{-1} + \cdots + a_{na} q^{-na})(1 - q^{-1})^d y(t) = (1 + b_1 q^{-1} + \cdots + b_{nb} q^{-nb}) \in (t)$$

Where n_a, is the order of the AR process, n_b is the order of the MA process, and d is the degree of differencing.

A SARIMA model was fitted with the same ARIMA parameters of (6,1,0). The seasonality frequency of 7 was added with an AR order of 3. The results RMSE for prediction of the test set was 43.34.

Convolution Neural Network (CNN)

Convolution Neural Networks (CNN) is a deep learning model that was originally created for image data. The power of the convolution network lies in the convolution operation that extracts features or patterns from the dataset (Brownlee, 2018). Once features or patterns are derived in the convolutional layers pooling are typically used to extract more obvious patterns.

For this project, three convolution layers were used, 2 max pooling layers, and a final fully connected layer. Relu was used as the activation function throughout and the layer and mean square error was chosen as the loss function. After training and testing the CNN model over 5 trials, it received an average RMSE of 51.02.

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is Recurrent Neural Network (RNN). RNNs, and especially in the case of LSTMs, were created to help solve time-related modelling problems (Brownlee, 2018). RNNs are built with memory gates in each model unit that allow the model to "remember" sequences of data. For this project we built a 2-layer neural network with a LSTM layer of 200 units and 1 fully connected layer. The results from 5 trails of the LSTM prediction yielded a RMSE of 49.13.

Results and Analysis

SEVERITY PREDICTION

Several different machine learning models were tested in order to compare their performance in traffic accident severity prediction. After comparing the results of each model, it was shown that CatBoost outperformed XGBoost and all other models. Although the overall accuracy of XGBoost seemed higher, CatBoost was 30 percent more accurate in predicting whether people were injured in traffic accidents. It is worth mentioning that the measure that we used was the "recall" score. Given that for this problem, identifying injuries caused by crashes has a higher relevancy. The results can be seen in *Figure 8* and *Figure 9*, respectively.

	precision	recall	f1-score	support
INJURY AND / OR TOW DUE TO CRASH	0.57	0.18	0.28	16846
NO INJURY / DRIVE AWAY	0.85	0.97	0.91	83213
accuracy			0.84	100059
macro avg	0.71	0.58	0.59	100059
weighted avg	0.81	0.84	0.80	100059

```
[[ 3091 13755]
 [ 2318 80895]]
```

Figure 8. XGBoost Results

	precision	recall	f1-score	support
INJURY AND / OR TOW DUE TO CRASH	0.31	0.58	0.40	16846
NO INJURY / DRIVE AWAY	0.90	0.73	0.81	83213
accuracy			0.71	100059
macro avg	0.60	0.66	0.60	100059
weighted avg	0.80	0.71	0.74	100059

```
[[ 9733  7113]
 [22134 61079]]
```

Figure 9. CatBoost Results

As noted previously, we implemented other algorithms, but we achieved results using such algorithms that were lower than expected. (To give one example, the overall accuracy for Logistic Regression was 68 percent.)

We initially expected that many different features comprising the dataset will influence the model. But running and obtaining the feature importance, we reached a number of conclusions, some of which were surprising. For instance, “contributory cause,” “first crash type” and “crash hour” are more likely to influence the model for traffic accidents. Contrary to

our thinking, “weather” and “day of the week” appeared to be less relevant and did not influence the model as much as expected. (This information can be seen in *Figure 4*.)

RISK PREDICTION

Figure 10 shows the performance of the tests for all the models in this analysis. The Holt-Winter method had the best performance overall and the LSTM model had the best performance for the neural network models. *Figure 11* shows the predicted values versus the true values for the Holt-Winter method. These graphs demonstrate the ability of the Holt-Winter method to capture the weekly seasonality of the dataset. *Figure 12* shows the autocorrelation of the Holt-Winter model residuals. The residuals appear to be close to zero, which is an indication that the model is capturing much of the variability of the dataset.

Model	Average RMSE
Average	66.27
Holt-Winter	46.68
SARIMA	48.34
CNN	51 (Avg 5 trails)
LSTM	49 (Avg 5 trails)

Figure 10. Risk Predictions

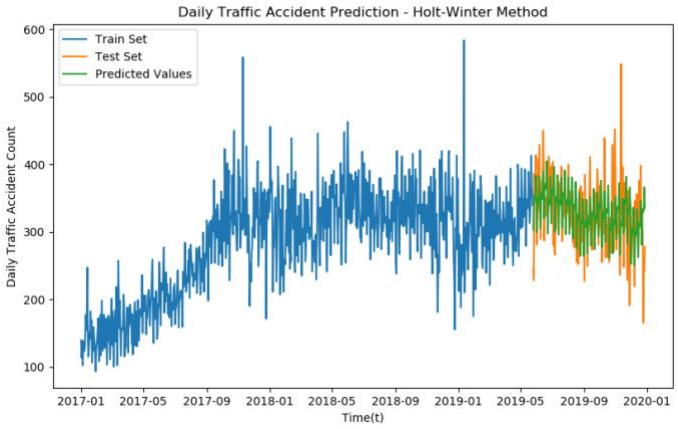

Figure 11. Holt-Winter Predictions

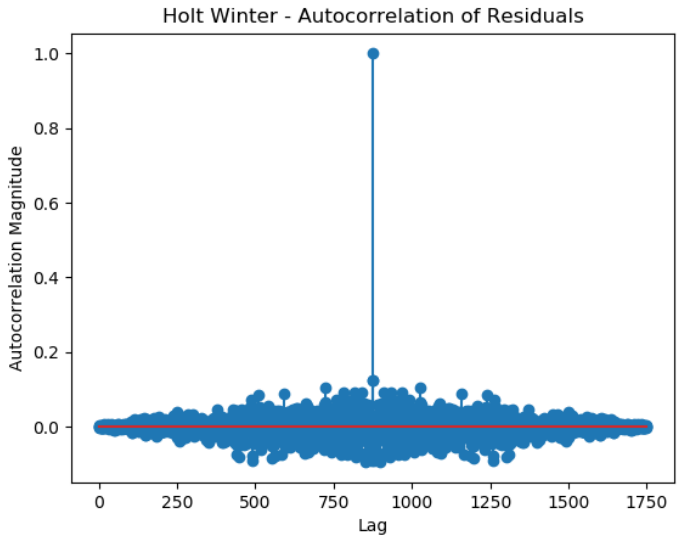

Figure 12. Autocorrelation of Residuals

Conclusion

This research aimed to test several machine learning, time series, and deep learning models in order to determine which modelling techniques had the best performance for predicting traffic accident risk and severity. Based on our analysis using the City of Chicago's traffic dataset, we concluded that the machine learning algorithm CatBoost had the best

performance when predicting traffic accident severity, with 58% accuracy with injury predictions.

The traditional time series Holt-Winter method had the best performance for accident risk prediction with an average RMSE of 46.68.

To advance this research we would recommend implementing additional Machine Learning algorithms such as Monte Carlo. For the prediction of risk, we assess that a multivariate analysis using additional traffic variables such as special features and weather data will likely improve model performance.

References

CatBoost - State-of-the-Art Open-Source Gradient Boosting Library with Categorical Features Support. (2020, April 30). Retrieved from https://catboost.ai

Amir Bahador Parsa, R. S. (2019). Applying Deep Learning to Detect Traffic Accidents in Real Time Using Spatiotemporal Sequential Data. *University of Illinois at Chicago*.

Brownlee, J. (2018). *Deep Learning for Time Series Forecasting, Predict the Future with MLPs, CNN and LSTM.*

Chawla, N. V. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence*, 321–357.

Chicago, C. o. (April, 2020 26). *Traffic Crashes - Crashes: City of Chicago: Data Portal*. Retrieved from https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if

Deep Learning. (2020, April 27). Retrieved from Wikipedia: https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=953563597

Global Status Report on Road Safety 2015. (2015). *World Health Organization*.

H. Ren, Y. S. (2018). A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3346-3351.

Iranitalab, A. a. (2017). Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction. *Accident Analysis & Prevention,*, 27–36.

M. Zheng, T. L. (2019). Traffic accident's severity prediction: A deep-learning approach-basedCNN network. *IEEE Access, vol. 7*, 39897–39910.

Machine Learning. (2020, April 30). Retrieved from Wikipedia: https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=954000510

Pradhan, B. a. (2020). Modeling Traffic Accident Severity Using Neural Networks and Support Vector Machines. *Laser Scanning Systems in Highway and Safety Assessment: Analysis of Highway Geometry and Safety Using LiDAR*, 111–17.

Time Series. (2020, April 25). Retrieved from Wikipedia: https://en.wikipedia.org/w/index.php?title=Time_series&oldid=953132528

Appendix 1

VEHICLES		
COLUMN	**DESCRIPTION**	**TYPE**
CRASH_UNIT_ID	A unique identifier for each vehicle record.	Number
CRASH_RECORD_ID	This number can be used to link to the same crash in the Crashes and People datasets. This number also serves as a unique ID in the Crashes dataset.	Plain Text
RD_NO	Chicago Police Department report number. For privacy reasons, this column is blank for recent crashes.	Plain Text
CRASH_DATE	Date and time of crash as entered by the reporting officer	Date & Time
UNIT_TYPE	The type of unit	Plain Text

NUM_PASSENGERS	Number of passengers in the vehicle. The driver is not included. More information on passengers is in the People dataset.	Number
VEHICLE_ID		Number
CMRC_VEH_I		Plain Text
MAKE	The make (brand) of the vehicle, if relevant	Plain Text
MODEL	The model of the vehicle, if relevant	Plain Text
LIC_PLATE_STATE	The state issuing the license plate of the vehicle, if relevant	Plain Text
VEHICLE_YEAR	The model year of the vehicle, if relevant	Number
VEHICLE_DEFECT		Plain Text
VEHICLE_TYPE	The type of vehicle, if relevant	Plain Text
VEHICLE_USE	The normal use of the vehicle, if relevant	Plain Text

TRAVEL_DIRECTION	The direction in which the unit was traveling prior to the crash, as determined by the reporting officer	Plain Text
MANEUVER	The action the unit was taking prior to the crash, as determined by the reporting officer	Plain Text
TOWED_I	Indicator of whether the vehicle was towed	Plain Text
FIRE_I		Plain Text
OCCUPANT_CNT	The number of people in the unit, as determined by the reporting officer	Number
EXCEED_SPEED_LIMIT_I	Indicator of whether the unit was speeding, as determined by the reporting officer	Plain Text
TOWED_BY	Entity that towed the unit, if relevant	Plain Text
TOWED_TO	Location to which the unit was towed, if relevant	Plain Text
AREA_00_I		Plain Text

AREA_01_I		Plain Text
AREA_02_I		Plain Text
AREA_03_I		Plain Text
AREA_04_I		Plain Text
AREA_05_I		Plain Text
AREA_06_I		Plain Text
AREA_07_I		Plain Text
AREA_08_I		Plain Text
AREA_09_I		Plain Text

AREA_10_I		Plain Text
AREA_11_I		Plain Text
AREA_12_I		Plain Text
AREA_99_I		Plain Text
FIRST_CONTACT_POINT		Plain Text
CMV_ID		Number
USDOT_NO		Plain Text
CCMC_NO		Plain Text
ILCC_NO		Plain Text
COMMERCIAL_SRC		Plain Text

GVWR	gross vehicle weight rating (GVWR)	Plain Text
CARRIER_NAME		Plain Text
CARRIER_STATE		Plain Text
CARRIER_CITY		Plain Text
HAZMAT_PLACARDS_I	Department of Transportation (DOT) mandates that Hazmat Placards be used when transporting hazardous materials and dangerous goods in the United States.	Plain Text
HAZMAT_NAME		Plain Text
UN_NO		Plain Text
HAZMAT_PRESENT_I		Plain Text
HAZMAT_REPORT_I		Plain Text

HAZMAT_REPORT_NO		Plain Text
MCS_REPORT_I	MOTOR CARRIER IDENTIFICATION?	Plain Text
MCS_REPORT_NO		Plain Text
HAZMAT_VIO_CAUSE_CRASH_I		Plain Text
MCS_VIO_CAUSE_CRASH_I		Plain Text
IDOT_PERMIT_NO		Plain Text
WIDE_LOAD_I		Plain Text
TRAILER1_WIDTH		Plain Text
TRAILER2_WIDTH		Plain Text
TRAILER1_LENGTH		Number

TRAILER2_LENGTH		Number
TOTAL_VEHICLE_LENGTH		Number
AXLE_CNT		Number
VEHICLE_CONFIG		Plain Text
CARGO_BODY_TYPE		Plain Text
LOAD_TYPE		Plain Text
HAZMAT_OUT_OF_SERVICE_I		Plain Text
MCS_OUT_OF_SERVICE_I		Plain Text
HAZMAT_CLASS		Plain Text

CRASHES

COLUMN	DESCRIPTION	TYPE
CRASH_RECORD_ID	This number can be used to link to the same crash in the Vehicles and People datasets. This number also serves as a unique ID in this dataset.	Plain Text
RD_NO	Chicago Police Department report number. For privacy reasons, this column is blank for recent crashes.	Plain Text
CRASH_DATE_EST_I	Crash date estimated by desk officer or reporting party (only used in cases where crash is reported at police station days after the crash)	Plain Text
CRASH_DATE	Date and time of crash as entered by the reporting officer	Date & Time
POSTED_SPEED_LIMIT	Posted speed limit, as determined by reporting officer	Number
TRAFFIC_CONTROL_DEVICE	Traffic control device present at crash location, as determined by reporting officer	Plain Text
DEVICE_CONDITION	Condition of traffic control device, as determined by reporting officer	Plain Text

WEATHER_CONDITION	Weather condition at time of crash, as determined by reporting officer	Plain Text
LIGHTING_CONDITION	Light condition at time of crash, as determined by reporting officer	Plain Text
FIRST_CRASH_TYPE	Type of first collision in crash	Plain Text
TRAFFICWAY_TYPE	Trafficway type, as determined by reporting officer	Plain Text
LANE_CNT	Total number of through lanes in either direction, excluding turn lanes, as determined by reporting officer (0 = intersection)	Number
ALIGNMENT	Street alignment at crash location, as determined by reporting officer	Plain Text
ROADWAY_SURFACE_COND	Road surface condition, as determined by reporting officer	Plain Text
ROAD_DEFECT	Road defects, as determined by reporting officer	Plain Text
REPORT_TYPE	Administrative report type (at scene, at desk, amended)	Plain Text

CRASH_TYPE	A general severity classification for the crash. Can be either Injury and/or Tow Due to Crash or No Injury / Drive Away	Plain Text
INTERSECTION_RELATED_I	A field observation by the police officer whether an intersection played a role in the crash. Does not represent whether or not the crash occurred within the intersection.	Plain Text
NOT_RIGHT_OF_WAY_I	Whether the crash begun or first contact was made outside of the public right-of-way.	Plain Text
HIT_AND_RUN_I	Crash did/did not involve a driver who caused the crash and fled the scene without exchanging information and/or rendering aid	Plain Text
DAMAGE	A field observation of estimated damage.	Plain Text
DATE_POLICE_NOTIFIED	Calendar date on which police were notified of the crash	Date & Time
PRIM_CONTRIBUTORY_CAUSE	The factor which was most significant in causing the crash, as determined by officer judgment	Plain Text

SEC_CONTRIBUTORY_CAUSE	The factor which was second most significant in causing the crash, as determined by officer judgment	Plain Text
STREET_NO	Street address number of crash location, as determined by reporting officer	Number
STREET_DIRECTION	Street address direction (N,E,S,W) of crash location, as determined by reporting officer	Plain Text
STREET_NAME	Street address name of crash location, as determined by reporting officer	Plain Text
BEAT_OF_OCCURRENCE	Chicago Police Department Beat ID. Boundaries available at https://data.cityofchicago.org/d/aerh-rz74	Number
PHOTOS_TAKEN_I	Whether the Chicago Police Department took photos at the location of the crash	Plain Text
STATEMENTS_TAKEN_I	Whether statements were taken from unit(s) involved in crash	Plain Text
DOORING_I	Whether crash involved a motor vehicle occupant opening a door into	Plain Text

	the travel path of a bicyclist, causing a crash	
WORK_ZONE_I	Whether the crash occurred in an active work zone	Plain Text
WORK_ZONE_TYPE	The type of work zone, if any	Plain Text
WORKERS_PRESENT_I	Whether construction workers were present in an active work zone at crash location	Plain Text
NUM_UNITS	Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory.	Number
MOST_SEVERE_INJURY	Most severe injury sustained by any person involved in the crash	Plain Text
INJURIES_TOTAL	Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries as determined by the reporting officer	Number

INJURIES_FATAL	Total persons sustaining fatal injuries in the crash	Number
INJURIES_INCAPACITATING	Total persons sustaining incapacitating/serious injuries in the crash as determined by the reporting officer. Any injury other than fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities they were capable of performing before the injury occurred. Includes severe lacerations, broken limbs, skull or chest injuries, and abdominal injuries.	Number
INJURIES_NON_INCAPACITATING	Total persons sustaining non-incapacitating injuries in the crash as determined by the reporting officer. Any injury, other than fatal or incapacitating injury, which is evident to observers at the scene of the crash. Includes lump on head, abrasions, bruises, and minor lacerations.	Number
INJURIES_REPORTED_NOT_EVIDENT	Total persons sustaining possible injuries in the crash as determined by the reporting officer. Includes momentary unconsciousness, claims of injuries not evident, limping, complaint of pain, nausea, and hysteria.	Number

INJURIES_NO_INDICATION	Total persons sustaining no injuries in the crash as determined by the reporting officer	Number
INJURIES_UNKNOWN	Total persons for whom injuries sustained, if any, are unknown	Number
CRASH_HOUR	The hour of the day component of CRASH_DATE.	Number
CRASH_DAY_OF_WEEK	The day of the week component of CRASH_DATE. Sunday=1	Number
CRASH_MONTH	The month component of CRASH_DATE.	Number
LATITUDE	The latitude of the crash location, as determined by reporting officer, as derived from the reported address of crash	Number
LONGITUDE	The longitude of the crash location, as determined by reporting officer, as derived from the reported address of crash	Number

LOCATION	The crash location, as determined by reporting officer, as derived from the reported address of crash, in a column type that allows for mapping and other geographic analysis in the data portal software	Point

PEOPLE		
Column	**Description**	**Type**
PERSON_ID	A unique identifier for each person record. IDs starting with P indicate passengers. IDs starting with O indicate a person who was not a passenger in the vehicle (e.g., driver, pedestrian, cyclist, etc.).	Plain Text
PERSON_TYPE	Type of roadway user involved in crash	Plain Text
CRASH_RECORD_ID	This number can be used to link to the same crash in the Crashes and Vehicles datasets. This number also serves as a unique ID in the Crashes dataset.	Plain Text

RD_NO	Chicago Police Department report number. For privacy reasons, this column is blank for recent crashes.	Plain Text
VEHICLE_ID	The corresponding CRASH_UNIT_ID from the Vehicles dataset.	Plain Text
CRASH_DATE	Date and time of crash as entered by the reporting officer	Date & Time
SEAT_NO	Code for seating position of motor vehicle occupant: 1= driver, 2= center front, 3 = front passenger, 4 = second row left, 5 = second row center, 6 = second row right, 7 = enclosed passengers, 8 = exposed passengers, 9= unknown position, 10 = third row left, 11 = third row center, 12 = third row right	Plain Text
CITY	City of residence of person involved in crash	Plain Text
STATE	State of residence of person involved in crash	Plain Text
ZIPCODE	ZIP Code of residence of person involved in crash	Plain Text
SEX	Gender of person involved in crash, as determined by reporting officer	Plain Text

AGE	Age of person involved in crash	Number
DRIVERS_LICENSE_STATE	State issuing driver's license of person involved in crash	Plain Text
DRIVERS_LICENSE_CLASS	Class of driver's license of person involved in crash	Plain Text
SAFETY_EQUIPMENT	Safety equipment used by vehicle occupant in crash, if any	Plain Text
AIRBAG_DEPLOYED	Whether vehicle occupant airbag deployed as result of crash	Plain Text
EJECTION	Whether vehicle occupant was ejected or extricated from the vehicle as a result of crash	Plain Text
INJURY_CLASSIFICATION	Severity of injury person sustained in the crash	Plain Text
HOSPITAL	Hospital to which person injured in the crash was taken	Plain Text
EMS_AGENCY	EMS agency who transported person injured in crash to the hospital	Plain Text
EMS_RUN_NO	EMS agency run number	Plain Text

DRIVER_ACTION	Driver action that contributed to the crash, as determined by reporting officer	Plain Text
DRIVER_VISION	What, if any, objects obscured the driver's vision at time of crash	Plain Text
PHYSICAL_CONDITION	Driver's apparent physical condition at time of crash, as observed by the reporting officer	Plain Text
PEDPEDAL_ACTION	Action of pedestrian or cyclist at the time of crash	Plain Text
PEDPEDAL_VISIBILITY	Visibility of pedestrian of cyclist safety equipment in use at time of crash	Plain Text
PEDPEDAL_LOCATION	Location of pedestrian or cyclist at the time of crash	Plain Text
BAC_RESULT	Status of blood alcohol concentration testing for driver or other person involved in crash	Plain Text
BAC_RESULT VALUE	Driver's blood alcohol concentration test result (fatal crashes may include pedestrian or cyclist results)	Number
CELL_PHONE_USE	Whether person was/was not using cellphone at the time of the crash, as determined by the reporting officer	Plain Text