Severity and Risk of Traffic Accidents

Taisha Ferguson &
Cristina Giraldo

Outline

- Problem Statement
- Literature Review
- **Part I – Predicting Accident Severity**
    - Data and Preprocessing
    - EDA
    - Models
    - Results
- **Part II – Predicting Risk**
    - Data and Preprocessing
    - EDA
    - Models
    - Results
- **Part III – Demo**
- Conclusions
- References

Problem Statement

According to the Centers for Disease Control and Prevention (CDC), car accidents are one of the leading causes of death in the U.S., causing around thirty-five thousand deaths per year. While there is some understanding of the factors that contribute to accident risk and severity, there is a need for further exploration as to how these factors together influence accident severity and risk. In our research we intend to use the different factors associated with car accidents to predict the severity and risk of an accident.

Literature Review

Iranitalab, Amirfarrokh, and Aemal Khattak. **“Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction.”** 2017

- Tested the performance of 4 Different Machine Learning and Statistical Models
- Multinomial Logit, Nearest Neighbor Classification, Support Vector Machines, and Random Forest
- Proposed a new Crash Cost Based Approach to measure performance Accuracy
- **Conclusion** – Best Was Nearest Neighbor and different models were better accuracy for different levels of severity

H. Ren, Y. Song, J. Wang, Y. Hu and J. Lei, **"A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction,"** 2018

- Predicted Accident Risk
- Compared the results of LSTM to other Baseline ML models: Lasso, SVM, Decision Tree Regression, and Autoregressive Moving Average Model (ARMA)
- Focused on the fact that traffic accidents have a temporal (time) component that cannot be fully explored by other models

Our Contribution

- Compare the accuracy of 5 different ML models to in order to predict accident severity
- Compare a SARIMA and Convolution Neural network in order to predict the risk of a traffic accident

Part I

Predicting Severity

Data and Preprocessing

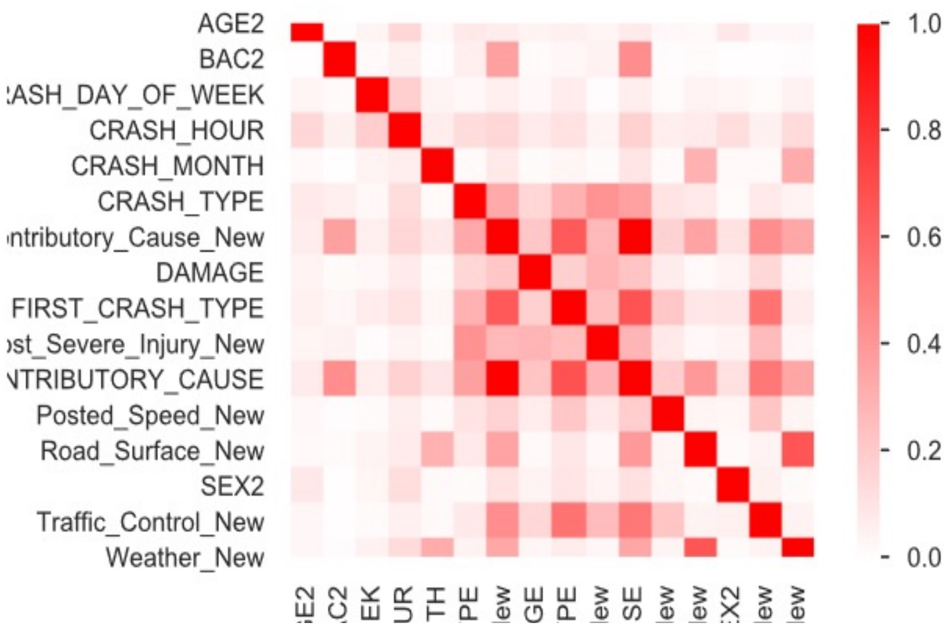
- 3 Datasets
- 148 variables
- Reduce Categories (I.e. weather and road condition)
- Limit crashes to two units
- Feature Selection to reduce dimensionality
- Populate missing values (I.e. age according to sex)
- Encoding
- Train set 70% - Test set 30%

Exploratory Data Analysis

Contributory_Cause_New is highly correlated with PRIM_CONTRIBUTORY_CAUSE	High Correlation
PRIM_CONTRIBUTORY_CAUSE is highly correlated with Contributory_Cause_New	High Correlation
Posted_Speed_New has 15351 (4.6%) missing values	Missing
Traffic_Control_New has 10947 (3.3%) missing values	Missing
Weather_New has 15116 (4.5%) missing values	Missing
Road_Surface_New has 23012 (6.9%) missing values	Missing
SEX2 has 80622 (24.2%) missing values	Missing
BAC2 has 80622 (24.2%) missing values	Missing
AGE2 has 80622 (24.2%) missing values	Missing
BAC2 is highly skewed (γ1 = 42.47060866)	Skewed
CRASH_HOUR has 5600 (1.7%) zeros	Zeros
BAC2 has 252537 (75.7%) zeros	Zeros

CRASH_TYPE
Categorical

Value	Count	Frequency (%)
NO INJURY / DRIVE AWAY	277094	83.1%
INJURY AND / OR TOW DUE TO CRASH	56434	16.9%



Models

MODEL	SCORE
xgboost.sklearn.XGBClassifier	0.8926300092422869
sklearn.ensemble._forest.RandomForestClassifier	0.8904018349753059
sklearn.ensemble._gb.GradientBoostingClassifier	0.8704127726626192
sklearn.ensemble._weight_boosting.AdaBoostClassifier	0.8441930870606278
sklearn.tree._classes.DecisionTreeClassifier	0.8426586344216769
sklearn.neighbors._classification.KNeighborsClassifier	0.8194304733856299
sklearn.linear_model._logistic.LogisticRegression	0.6842805574005306
Catboost without oversampling	0.84
Catboost Oversampled	0.79

Results

XGB Classifier		
Injury	3091	13755
No Injury	2318	80895
	Injury	No Injury

CatBoost Classifier No Resampled		
Injury	2538	14308
No Injury	1652	81561
	Injury	No Injury

CatBoost Classifier Resampled		
Injury	9733	7113
No Injury	22134	61079
	Injury	No Injury

Estimator	Value	Precision	Recall	F1	Support
XGBClassifier Oversampled	Injury	0.57	0.18	0.28	16846 83213
	No Injury	0.85	0.97	0.91	
CatBoost No Sampled	Injury	0.61	0.15	0.24	
	No Injury	0.85	0.98	0.91	
CatBoost OverSampled	Injury	0.31	0.58	0.40	
	No Injury	0.90	0.73	0.81	

Part II
Predicting Risk

Data and Preprocessing

Data was transformed from a Classification problem **to a Time Series Regression** problem

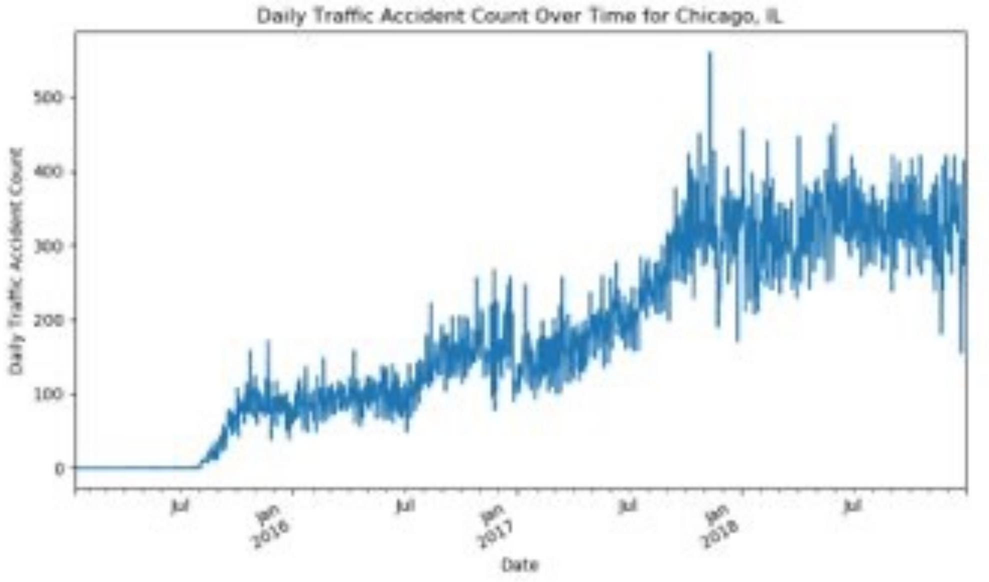
Traffic Accident records were reformatted to Daily Counts

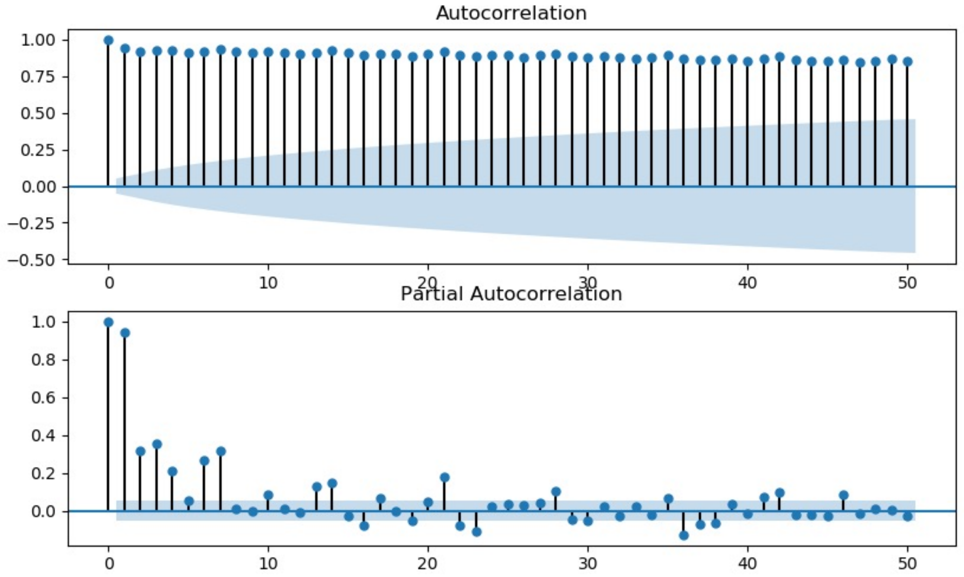
Time Period: Jan 10 2015 – December 31, 2018

Train Set: 2015-2017

Test Set: 2018

Exploratory Data Analysis





Models

Seasonal Autoregression Moving Average Model (SARIMA)

- Season – Weekly
- Autoregressive Order – 3
- Moving Average Order 0
- Trend - Linear

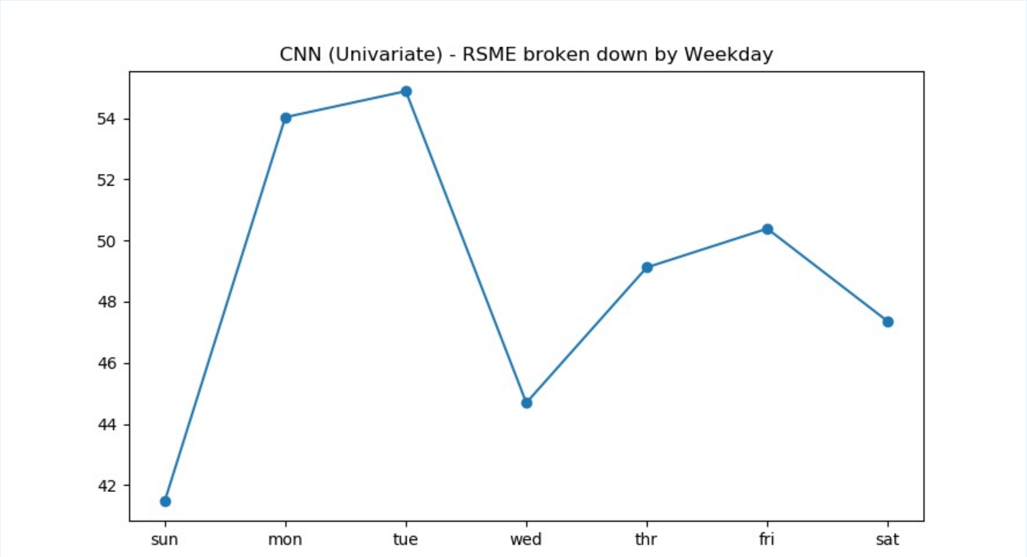
Convolutional Neural Network – Univariate

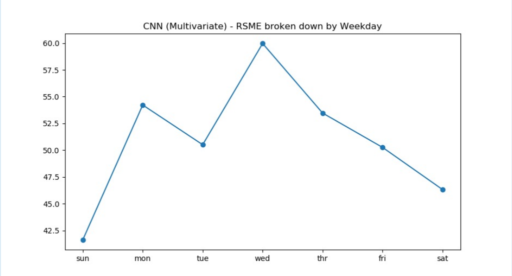
- 2 Convolutional Layers
- Max Pooling Layer
- One Fully Connected Layer
- Activation Relu
- Loss Function MSE

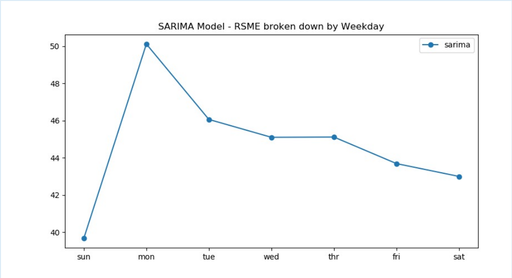
Convolutional Neural Network – Multivariate

- Added Daily Temperature and Total Precipitation

Model Results







Model	Average RMSE
SARIMA	44
CNN (Univariate)	48
CNN (Multivariate)	52

Demo

Conclusions

Conclusions

- The variables with the strongest relationship to accident severity were age, hour, month, week, day of the week, first crash type, primary contributory cause, sex, speed, traffic control, weather and road surface
- Accident Risk had a upward linear trend and a weekly seasonal pattern.
- CatBoost was the best ML for severity
- For predicting Acident Risk the baseline model SARIMA outperformed the Neural Network CNN

Limitations and Feature Research

Limitations and Feature Research

- Time
- Do more hyperparameter tuning for CNN and shallow models.
- Try LSTM and other RNN
- Use spacial features in analysis not just time elements
- Monte Carlo Simulation to calculate probabilities under specific conditions
- Expand app to have real-time risk calculations. So people can know their traffic accident risks before heading on a trip.

References

References

1. Iranitalab, Amirfarrokh, and Aemal Khattak. **“Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction.”** *Accident Analysis & Prevention*, vol. 108, Nov. 2017, pp. 27–36, doi:10.1016/j.aap.2017.08.008.
2. Pradhan, Biswajeet, and Maher Ibrahim Sameen. **“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*, edited by Biswajeet Pradhan and Maher Ibrahim Sameen, Springer International Publishing, 2020, pp. 111–17, doi:10.1007/978-3-030-10374-3_9.
3. H. Ren, Y. Song, J. Wang, Y. Hu and J. Lei, **"A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction,"** 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, 2018, pp. 3346-3351.
4. M. Zheng *et al.*, **"Traffic Accident’s Severity Prediction: A Deep-Learning Approach-Based CNN Network,"** in *IEEE Access*, vol. 7, pp. 39897-39910, 2019.


Questions?