U.S. Road Accidents Forecasting using Sequential Model

Dhruv Arora*
Graduate Student, CSE 573
University of Washington
dharora@uw.edu

Abstract

Forecasting time-series data is an important subject in economics, business, finance and many other domains. The existing popular techniques such as ARIMA, Generalized Additive Models (GAM) and many others statistical methods have dominated the time-series forecasting area. The recent advancement in the area of deep learning, specifically sequential models have evolved to be particularly suitable for time-series forecasting and provide other forms of advantage such as scalability along with forecasting accuracy. The research question investigated in this article is to compare the forecasting accuracy of "Long Short-Term Memory (LSTM)" against the existing statistical methods - ARIMA and GAM. The empirical study is conducted on U.S. road accidents dataset which exhibits suitable temporal dependency for sequential models. The study reports that LSTM outperforms in majority of time-series when building a single model for all the time-series and comparing it to ARIMA and GAM models. The advantages and disadvantages of LSTM are also discussed.

1 Introduction

1.1 Motivation

There have been 4.2M road accidents in the US in last 4 years as reported by the Bing and Map-quest sources. The motivation is to build ML forecast models that can predict road accidents for each state separately. This can help enable data driven decisions for transportation governing agencies to measure and reduce the road accidents. This project aims to forecast road accidents using state of the art ML methods using LSTM-RNN (and its variants) and compare its performance against existing statistical methods such as ARIMA and Generalized Additive Models (GAM). The road accidents time-series data exhibits suitable time-series characteristics for modeling sequential data. The study will also analyze the advantages and disadvantages of using state of the art methods. Due to computational and time restriction, this study focuses on forecasting accidents for the top 10 states suffered by road accidents as observed in the data.

1.2 Forecasting Methodologies

A variety of statistical methods concerning to time-series forecasting have been evolved in literature. ARIMA has been a standard method for time-series forecasting for a long-time as per Hyndman and Athanasopoulos (2018). Even though, ARIMA being very popular method suffers from major limitations. For instance, it cannot handle multiple complex seasonality or handle non-linearity. Furthermore, it is assumed that there is constant variance in errors, which in practice may not satisfied.

*Master in Data Science Student (starting batch - 2019)

AI Project on Sequential Model-CSE 573 (Intro to AI 2021), Seattle, Washington.

GAM models have also become popular but also suffers from similar limitation such as non-linearity. These method tends to work well in uni-variate time-series modeling when trend and seasonality are the main components to learn. However, they tend to be limited in case of multi-variate time-series forecasting.

Recently, new techniques in deep learning have been developed to address the challenges related to the forecasting models. LSTM [Olah (2015)] is special case of Recurrent Neural Network that was initially introduced by [Hochreiter and Schmidhuber (1997)]. A recurrent neural network is a class of artificial neural network which exhibit temporal dynamic behavior. It is derived from the feed forward neural network and specialized in learning sequence data such as multi-variate time series data.

2 Data Analysis and Pre-processing

The dataset used in this empirical study is publicly available as part of the published paper Moosavi et al. (2019b) and Moosavi et al. (2019a). The dataset contains accident data that are collected from February 2016 to Dec 2020 for the contiguous United States. The data is sourced mainly from two APIs- Bing and MapQuest and reports 4.2M accidents in 4 years of time-frame.

2.1 Exploratory Data Analysis

The dataset columns are shown below which are of interest for this empirical study:

	Date	State	Temperature	DailyAccidents	Pressure	Visibility	Weather_Condition
0	2016-02-09	MI	31.000	2	29.590	1.85	Light Snow
1	2016-02-09	PA	29.450	2	29.585	1.25	Light Snow
2	2016-02-10	PA	21.500	3	29.750	1.60	Light Snow
3	2016-02-11	PA	14.725	4	30.285	10.00	Mostly Cloudy
4	2016-02-12	MI	23.000	1	29.960	10.00	Mostly Cloudy

Figure 1: Snippet of dataset

The columns of interest to forecast is Daily Accidents which represents number of road accidents on a given time-period for each State. The other columns are potential features that can be used as input to the models. The main assumption made is that the input features are available for future time frame such as Temperature, Pressure, Visibility and Weather Condition.

To visualize the daily accidents over time, the Daily Accident column is plotted for top 10 States against daily date in Figure 2.

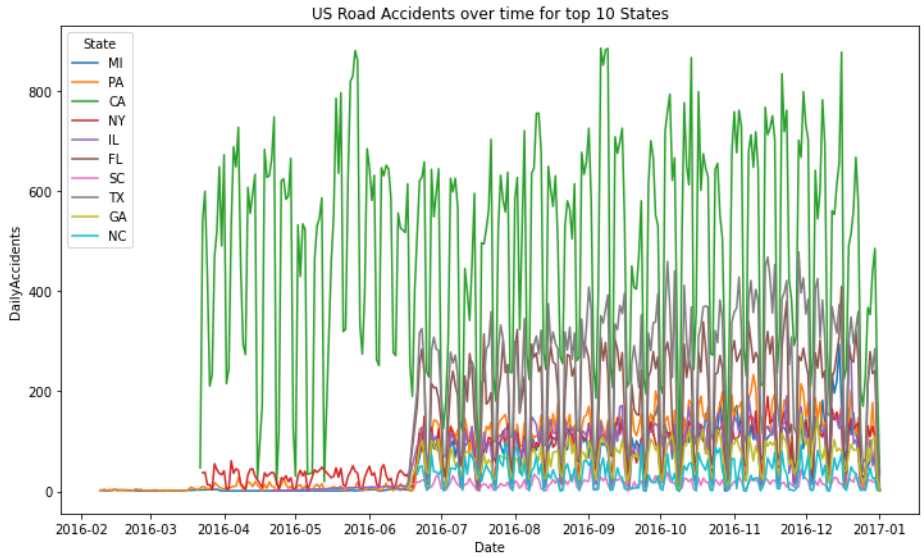

Figure 2: Daily Road Accidents

2.2 Data Pre-processing

Here are the pre-processing steps performed to transform and prepare data for modeling:

1. The dataset contains each accident as one single record which has been rolled-up to daily date and for each State.
2. Filter the dataset to include only top 10 States which suffered the most accidents for modeling.
3. The imputation method for input features use forward fill method
4. Create lag features such as 365 day lag and 7 day lag for inputting seasonality features in LSTM model
5. Split the data into train, test and validation for performance evaluation
6. Prepare environment to install FbProphet (GAM package), statsmodel (ARIMA package) and TensorFlow library (LSTM package) for modeling

The implementation is available on github Arora (2020).

3 Implementation

ARIMA and GAM models were build separately for each of the States whereas a single model was built for LSTM where State was given as input to the model. All of these models uses weather features in one form or another such as Temperature, Pressure, Visibility and Weather Condition. For the sake of reproducibility of implementation, we have hosted the python notebook on github [Arora (2020)].

3.1 Baseline Model

We identified that GAM model [SJ and B. (2017)] consistently outperformed the ARIMA models for all States in category of statistical models. Therefore, we setup the baseline model as GAM for further comparison with LSTM. We have shown the out of sample forecast for California state in Figure 3 below. For other States forecast, please refer to the notebook on github[Arora (2020)].

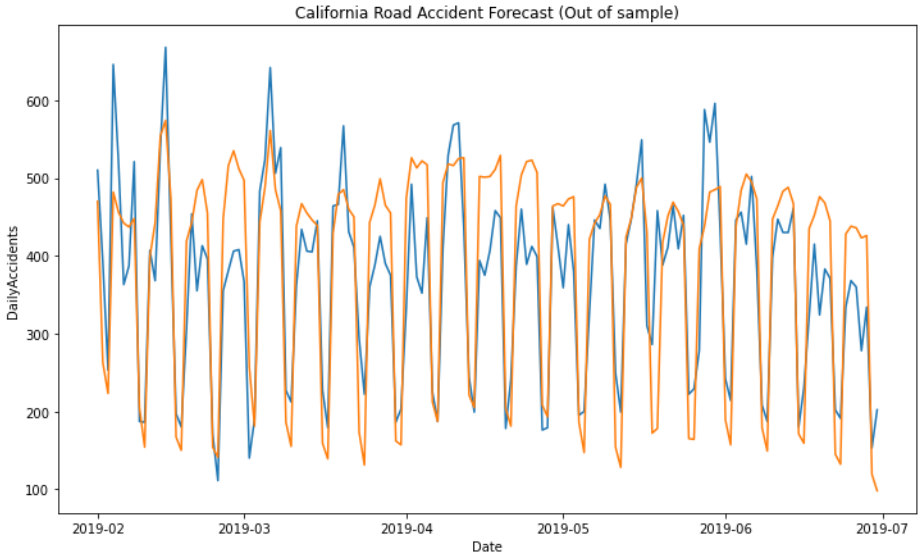

Figure 3: California Road Accident Forecast

3.2 Experimentation

LSTM models [J. (2016)] were evaluated with various variants of hyper-parameters such as epoch, batch size, number of neurons in LSTM layer, number of LSTM layers, dropout etc. It was identified that model with 100 epochs, 64 batch size, 1 layer of LSTM and no dropout performs best among other variants which is further compared against baseline GAM model for conclusion.

4 Result

It is identified that LSTM model outperforms the GAM model in 6 of the 10 States whereas GAM model outperforms the LSTM model in 3 States and there is a tie in one State. It is also to note that LSTM model is one single model trained for all States whereas GAM model is build separately for each of the 10 States.

Out of Sample- Model Performance			
#	**State**	**RMSE (LSTM)**	**RMSE (GAM)**
1	CA	89.2	**77.52**
2	TX	**48.75**	**48.75**
3	FL	**40.61**	45.65
4	SC	**33.83**	67.39
5	NC	**40.54**	73.02
6	NY	**38.17**	40.27
7	GA	23.12	**19.78**
8	IL	**21.84**	23.74
9	MI	24.41	**22.52**
10	PA	**22.98**	24.6

Figure 4: Model Performance

5 Conclusion

LSTM outperforms the GAM model in the empirical study performed on US road accidents dataset. It is found that LSTM outperform other models in forecasting road accidents in majority of the States. GAM model outperforms the LSTM model in only three of the ten states. The other advantage of using LSTM model is building one single model that can forecast ten States whereas statistical models require building separate model for each State. This gives an edge to the LSTM model in terms of scalability of the approach as it can be extended to build forecast for each of the 50 US States in one shot setup and does not require tuning the hyper-parameter for each time-series separately. The effort in building and maintaining the forecast models is also less in LSTM when comparing to statistical models. However, initial round of hyper-parameter tuning may take substantial amount of time, LSTM being computationally expensive. The one of the disadvantage of LSTM is the computation power required for building the models. The LSTM model suffers from interpretability and transparency of the feature whereas statistical models are easy to explain and tweak. The final LSTM model with some future enhancements has potential to enable daily alerting on projected road accidents based on weather forecast and enable data driven decision making in mitigating the road accidents.

References

Arora, D. (2020). Github. *"https://github.com/Arora-Dhruv/U.S-Road-Accidents-Forecasting-using-LSTM"*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

Hyndman, R. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. 2nd edition, OTexts: Melbourne, Australia. Otexts.com/fpp2. Accessed on 2020.03.09.

J., B. (2016). Time series prediction with lstm recurrent neural networks in python with keras. *"https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/"*.

Moosavi, S. et al. (2019a). Accident risk prediction based on heterogeneous sparse data. *"Cornell University"*.

Moosavi, S. et al. (2019b). A countrywide traffic accident dataset. *"Cornell University"*.

Olah, C. (2015). Understanding lstm networks. *"colah.github.io"*.

SJ, T. and B., L. (2017). Forecasting at scale. *"PeerJ Preprints"*.