NTCC report on

Regression Models Using RNN

AMITY University, Kolkata

for the partial fulfillment of the award of the degree

MASTER OF COMPUTER APPLICATION

Saransh Kumar Karan

A91400717011

Industry Guide: Mr. P. S. Mukherjee

Faculty Guide: Prof. S. Chaterjee

AMITY INSTITUTE OF INFORMATION TECHNOLOGY

AMITY UNIVERSITY, KOLKATA

November 2019

DECLARATION

I hereby declare that the dissertation entitled “Regression Model using Recurrent Neural Networks” submitted by me in partial fulfillment of the requirement for the Degree of MCA to AMITY UNIVERSITY, KOLKATA is based on the experiments and studies carried out by me. This work is original and has not been submitted in part or full for any other degree or diploma of any university or institution.

Date: 22/11/2019

Place: Kolkata

Saransh Kumar Karan

A91400717011

CERTIFICATE

The research work embodied in this dissertation entitled "Regression Model using Recurrent Neural Networks" submitted by Saransh Kumar Karan, A91400717011 in partial fulfillment of the requirements for the award of the Degree of Master of Computer Application to the AMITY UNIVERSITY, KOLKATA is based on the experiments and studies carried out by him. This work is original and has not been submitted in part or full for any other degree or diploma of any university or institution.

Date: 22/11/2019

Place: Kolkata

Prof S. Chaterjee

Acknowledgement

I would like to thank my institute AIITK and ISI Kolkata for providing me with this oppurrtunity. I would also like to thank my faculty guide and industry guide for showing me the way.

Table of Contents

1. ABSTRACT

Machine Learning aims to find algorithms that help in gaining insights from data. Regression is a type of machine learning problem where the input features (attributes) are being mapped to continuous values. The input features may be categorical or continuous. The problem is solved by firstly, representing a model to represent the problem. Then a cost function is used to measure the accuracy of the prediction made by the model. The next step is to reduce the error to improve the accuracy of the model based on available data. This is called training. Then the model can be used to predict unseen data.

Time series data is data that is stored in chronological order with respect to time. For example, the price of a stock during consecutive days. Recurrent Neural Network is a Deep Learning Technique that is developed to work on time series data. They have been efficiently used to predict the stock price of a company for companies.

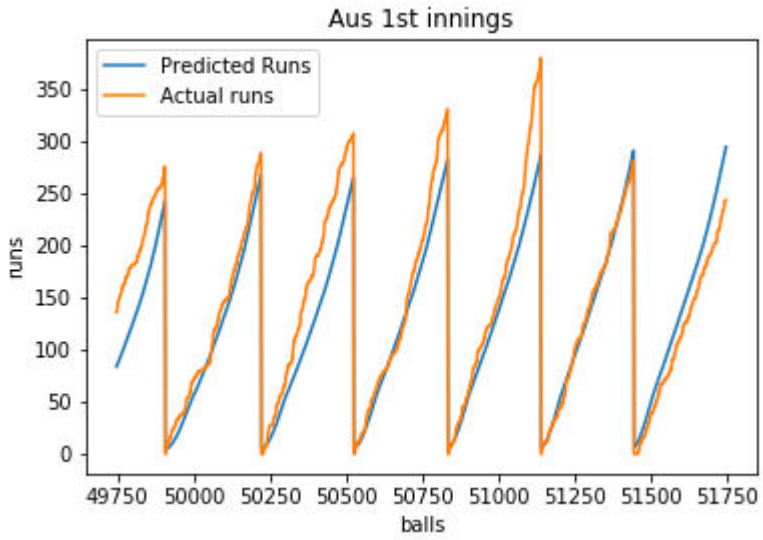
This project aims to speculate final score of cricket matches. Long Short Term Memory, which is a special type of RNN was used to model the problem. The inputs where index number and ball number. The data set was downloaded from cricksheet.org which was in yaml format. The data was parsed into separate csv files for the top four teams, consisting only information about a specific innings. This csv file was indexed which was used as an input.

It was found that certain trends such as secular trends were predicted by the model. However, the random movements were not identified by

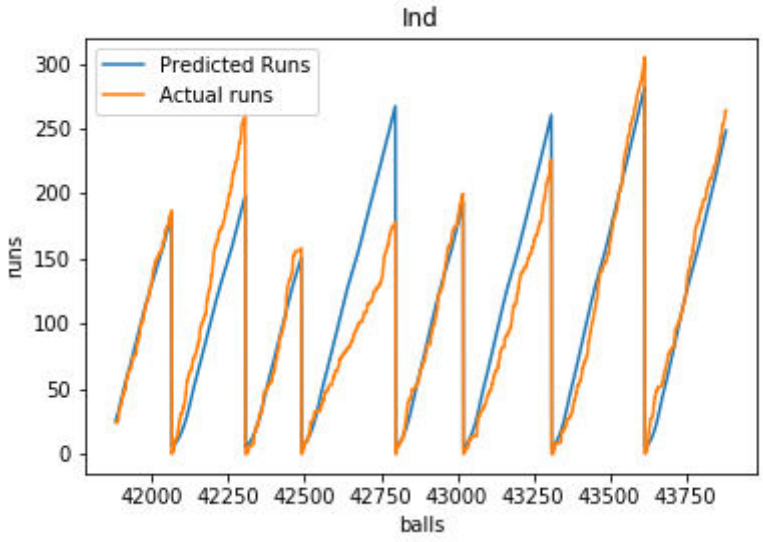
the model.The performance of the model may be improved by increasing the number of input features to the model. For example, the weather and pitch conditions, the batsman and bowler in the current ball etc.

Analysis

‘Keras’ was used to develop the LSTM network on the transformed data. The model was trained on 28000 balls to observe seasonal fluctuations. The accuracy did not improve greatly when the number of balls was decreased to 8000. The following graph was used to validation:



There is a secular trend initially the model did not effectively predict it. However the loss was not large.



The above validation

set seems to be unpredictable. There is a decrease
in total runs scored after the second match. The model accurately predicted that. Furthermore, the predicted final score seems to be very close to the actual final score. Then there is a steady increase in the final score, this trend is also predicted with lesser accuracy by the model.

Limitations

The model takes into account the ball number for training. This trait is learned by the model. Therefore, the change of innings is also being considered in the model. To avoid this the ball number attribute can be removed or a fixed sequence of balls may be used to predict the score(say 300).

Furthermore, the model does not take into consideration important attributes such as the fall of wickets.

The future work is to improve the accuracy of the model may be improved by increasing the number of input features.

References

1. https://towardsdatascience.com/

2. LONG SHORT-TERM MEMORY Neural Computation 9(8):1735{1780, 1997 Sepp Hochreiter Fakult□at f□ur Informatik Technische Universit□at M□unchen 80290 M□unchen, Germany hochreit@informatik.tu-muenchen.de http://www7.informatik.tu-muenchen.de/~hochreit J□urgen Schmidhuber IDSIA Corso Elvezia 36 6900 Lugano, Switzerland juergen@idsia.ch http://www.idsia.ch/~juergen