Chapter  7
A Comparison of Machine Learning Algorithms of Big Data for Time Series Forecasting Using Python

Son Nguyen
Bryant University, USA

Anthony Park
Bryant University, USA

ABSTRACT

This chapter compares the performances of multiple Big Data techniques applied for time series forecasting and traditional time series models on three Big Data sets. The traditional time series models, Autoregressive Integrated Moving Average (ARIMA), and exponential smoothing models are used as the baseline models against Big Data analysis methods in the machine learning. These Big Data techniques include regression trees, Support Vector Machines (SVM), Multilayer Perceptrons (MLP), Recurrent Neural Networks (RNN), and long short-term memory neural networks (LSTM). Across three time series data sets used (unemployment rate, bike rentals, and transportation), this study finds that LSTM neural networks performed the best. In conclusion, this study points out that Big Data machine learning algorithms applied in time series can outperform traditional time series models. The computations in this work are done by Python, one of the most popular open-sourced platforms for data science and Big Data analysis.

DOI: 10.4018/978-1-7998-2768-9.ch007

Copyright © 2020, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
INTRODUCTION

Machine learning models have proven their usefulness for predicting classification and regression outcomes in cross-sectional data for many years now. Banks, doctors, insurance firms, social media companies, and more use machine learning to detect fraud, increase consumer retention, target the right audiences, manage risk, and identify diseases. In the research field, extensive studies have been done on machine learning model types, resampling methods, evaluation metrics, feature selection, and variable manipulation. One place research and applications of machine learning techniques have been underrepresented is in time series forecasting though. Like machine learning, time series forecasting is performed in various fields and industries including for macroeconomic variables, retail sales, stock market returns, and production. If done accurately, these analyses can give a business or organization the edge over competition by helping them increase profits and reduce risk.

To sufficiently study an underrepresented field of study and to ensure the findings of this study are robust, three Big Data sets are used: unemployment rate, bike rental sales, and car traffic. These three data sets were selected to ensure the machine learning methods analyzed in this study consistently produce similar results across a wide range of industries and time series data types. Furthermore, each data set requires different pre-processing technique to be applied in order to transform the data into a machine learning problem. These techniques and further details on these data sets will be discussed further in the Data Exploration and Pre-Processing section.

Through analysis and comparison of many data models, we find that machine learning models can produce reliable forecasts on a wide range of time series data sets that vary in the number of observations and the frequency of the observations. Models used in this research (Holt-Winter’s, vector autoregression, ARIMA, regression trees, SVM Regressions, Recurrent Neural Networks, long short-term memory) and the evaluation metrics (Root Mean Squared Error, R-squared on the testing data, and Mean Absolute Scaled Error) will be explained before the results of the research are presented in the last two sections.

BACKGROUND AND LITERATURE REVIEW

Forecasting

Time series forecasting has largely been based around the Box-Jenkins method of Autoregressive Integrated Moving Average (ARIMA) developed in the 1960s and the Holt-Winters method of exponential smoothing developed in the 1950s. These methods, especially Box-Jenkins, rely heavily on the forecaster to visualize the time
Introduction to the Popular Open Source Statistical Software (OSSS)
www.igi-global.com/chapter/introduction-to-the-popular-open-source-statistical-software-osss/248874?camid=4v1a