International Journal of Advance Research in Science and Engineering
Volume No. 11, Issue No. 02, February 2022
www.ijarse.com
IJARSE
ISSN 2319 - 8354

COMPARATIVE ANALYSIS OF SUPPORT VECTOR AND QUNATUM SUPPORT VECTOR REGRESSION ON THREE DIFFERENT UNIVARIATE TIME SERIES DATASETS

Arvind Rathee[1], Kamaldeep Joshi[2], Ashish Nandal[3], Rainu Nandal[4]

[1,2,3,4] *Department of Computer Science and Engineering, UIET, Maharshi Dayanand University (MDU), Rohtak 124001, Haryana, India*

Abstract

In this paper, a study on three different univariate time-series datasets is conducted using the SVR and QSVR algorithms. The datasets used in this study were not univariate, so they were transformed into univariate time series using python. Additionally, these datasets were scaled to get three more datasets. These univariate datasets were divided into training and testing sets of ratios 4:1. Each regression algorithm was trained on every data set separately and results are predicted corresponding to every data set using all trained regression models. The findings were recorded and plotted to get an overall picture of how QSVR performed in comparison to SVR for different kinds of datasets and if it performed better for scaled or unscaled datasets.

Keywords: Quantum Support Vector Regression, Time Series, Support Vector Regression, Univariate Time Series

Introduction

In this ever-growing digital world, data is being produced at an exponential speed. It was estimated that at dawn of 2020 the world datasphere had 44 zettabytes and by 2025 it is expected to grow to 125 zettabytes**.** This data is like gold for companies, it is used by them to extract information and make decisions based on that information to attract more customers, increase productivity, optimize work flow and do many more critical tasks. To do this with less manpower, use of Machine Learning has increased in variety of fields such as education, chemistry, medicine etc. As Machine learning have strong mathematical foundation, it helps to develop reliable application that can run in real time benefiting everyone without throwing any errors[1], [2]. One such example being used widely is email spam filtering.

The data being produced can be of any dimension and form. One of these is Time series data. Time series can be defined as series of data points that have been recorded over a period of time may it be a second or centaury. Time series help us to find different patterns in data by providing data ordered using time (i.e Time series gives us a time dimension). Generally, in machine learning datasets collection of observations are treated equally to predict future outcomes. Whereas in Time series Time dimension is used to make accurate prediction. Typically, it's assumed that the gathered time series is recorded in regular time interval but sometimes data may be in irregular time intervals. An example of regular time series is Daily Temperature and ATM withdrawals come under irregular time series [3], [4].

Machine Learning and Data Analysis use statistical approaches to examine data and offer computers with learning capabilities based on observations made from that data. Machine Learning uses different kind of learning styles and based on those learning styles machine learning is divided into following categories: Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, and Reinforcement Learning. Computation time and storage are two of the most well-known drawbacks of employing machine learning algorithms, especially when dealing with large volumes of data [5]. Furthermore, when existing deep learning techniques are applied, the training time can be extended much further. Researchers developed a viable alternative by harnessing the power of quantum computers, which are clever enough to cut storage and computing time.

Most prominent and most widely used algorithm for time series forecasting is autoregressive integrated moving average (ARIMA). Because ARIMA models are limited by the assumption of linear inter-temporal dependencies, they struggle to capture nonlinear stylized facts, which are ubiquitous in real-world circumstances [6], [7]. To overcome this limitation many models have been developed like autoregressive conditional heteroscedastic (ARCH) model, bilinear model, generalized autoregressive conditional heteroscedastic (GARCH) model, smooth transition autoregressive (STAR) model and many more but unfortunately none of them are applicable to every kind of problem [8], [9]. Due to growth of computational power, the use of Machine learning models has been rising. Machine learning provides additional benefit that as compared to classical time series analysis, it does not require any prior knowledge of underlying structure of data. Many researchers have successfully used Machine learning algorithms to get excellent results [10]–[12]. Support Vector Regression, a regression algorithm that works on principle of SVM that is finding best fit line for a given dataset. It is frequently used to get satisfactory outcomes in case of Time series data. [13]–[15].

Quantum computing, a quantum mechanics-based sort of computer that interacts with a chaotic and uncertain physical reality, is a buzzword in today's technology market. Because quantum mechanics is a more general model of physics than classical mechanics, it leads to quantum computing, a more generic model of computing that can solve issues that classical computing can't. Unlike ordinary computers, which use binary bits 0 and 1 to store and process data independently, quantum computers use their own quantum bits, also known as 'Qubits,' to store and alter data [16]. Quantum Machine Learning is an emerging new inter-disciplinary theoretical field. It uses the properties of Quantum Computing along with theory of Machine Learning. It is basically used to implement Classical Machine Learning algorithms in quantum sphere by using quantum properties like superposition and entanglement [17].

In this paper, we aim to explore the field of Time Series using the quantum algorithms. Specifically, the target of this paper is to compare Classical Support vector regression and Quantum Support Vector regression in case of univariate Time series using base line settings.

Datasets

For this study, following three datasets were acquired from Kaggle:

1. US Monthly Unemployment Rate 1948 – 2019 [18]

The raw data acquired was in the form shown in Table 1. The dataset had 13 columns and 72 rows. Each row contained year and 12-month unemployment data for a particular year. This kind of dataset is example of multivariate Time Series. This data was converted into univariate Time Series as shown in Table 2, that contained 2 columns, column 1 had Month and year in it and for that corresponding time period column 2 consisted the unemployment data.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1948-01-01	3.4	3.8	4.0	3.9	3.5	3.6	3.6	3.9	3.8	3.7	3.8	4.0
1949-01-01	4.3	4.7	5.0	5.3	6.1	6.2	6.7	6.8	6.6	7.9	6.4	6.6
1950-01-01	6.5	6.4	6.3	5.8	5.5	5.4	5.0	4.5	4.4	4.2	4.2	4.3
1951-01-01	3.7	3.4	3.4	3.1	3.0	3.2	3.1	3.1	3.3	3.5	3.5	3.1
1952-01-01	3.2	3.1	2.9	2.9	3.0	3.0	3.2	3.4	3.1	3.0	2.8	2.7

Table 1 - Raw US Unemployment dataset data.

Date	Data
1948-01-01	3.4
1948-02-01	3.8
1948-03-01	4.0
1948-04-01	3.9
1948-05-01	3.5

Table 2 - Univariate Time Series for US Unemployment data

2. Microsoft Stock [19]

The structure of acquired Microsoft Stock data was in the form as shown in Table 3. The dataset consisted of 6 columns and 1151 rows. Each row contains daily date, open, high, low, close and volume data of Microsoft. For study in this paper only open data along with dates was extracted as a univariate time series as shown in Table 4.

Date	Open	High	Low	Close	Volume
2015-04-01 16:00:00	40.60	40.76	40.31	40.72	36865322
2015-04-02 16:00:00	40.66	40.74	40.12	40.29	37487476
2015-04-06 16:00:00	40.34	41.78	40.18	41.55	39223692
2015-04-07 16:00:00	41.61	41.91	41.31	41.53	28809375
2015-04-08 16:00:00	41.48	41.69	41.04	41.42	24753438

Table 3 - Daily Microsoft Stock dataset.

Date	Open
2015-04-01 16:00:00	40.60
2015-04-02 16:00:00	40.66
2015-04-06 16:00:00	40.34
2015-04-07 16:00:00	41.61
2015-04-08 16:00:00	41.48

Table 4 - Univariate Time Series of Open Price of Microsoft Stock.

3. Climate Change: Earth Surface Temperature Data [20]

Third and final dataset was in the form as shown in Table 5. The dataset consisted of 9 columns and 3192 rows. Out of 3192 rows which for LAT and LATU consisted data in 3180 rows had data and for remaining columns i.e., LMT, LMTU, LMIT, LMITU, LOAT, and LOATU had data in 1992 rows. The remaining cells did not have data for corresponding date in date column. The dataset was cleaned by removing empty cells and a univariate series was constructed for LAT data of the form shown in Table 6.

dt	LAT	LATU	LMT	LMTU	LMIT	LMITU	LOAT	LOATU
1750-01-01	3.034	3.574	NaN	NaN	NaN	NaN	NaN	NaN
1750-02-01	3.083	3.702	NaN	NaN	NaN	NaN	NaN	NaN
1750-03-01	5.626	3.076	NaN	NaN	NaN	NaN	NaN	NaN
1750-04-01	8.490	2.451	NaN	NaN	NaN	NaN	NaN	NaN
1750-05-01	11.573	2.072	NaN	NaN	NaN	NaN	NaN	NaN

Table 1 - Raw data of Global Temperature dataset.

dt	LAT
1750-01-01	3.034
1750-02-01	3.083
1750-03-01	5.626
1750-04-01	8.490
1750-05-01	11.573

Table 2 - Univariate Time Series for LandAverageTemperature in Global Temperature dataset.

Abbreviation	Actual Name
LAT	LandAverageTemperature
LATU	LandAverageTemperatureUncertainty
LMT	LandMaxTemperature
LMTU	LandMaxTemperatureUncertainty
LMIT	LandMinTemperature
LMITU	LandMinTemperatureUncertainty
LOAT	LandAndOceanAverageTemperature
LOATU	LandAndOceanAverageTemperatureUncertainty

Table 3 - Abbreviation for Global Temperature Dataset.

Methodology

The entire process used in this paper can be summarized in three steps those being Data Acquisition and Transformation, Model Training and Evaluation, and Plotting. Figure 1 depicts the same process with all sub tasks. Entire Process can be explained as below:

Step 1: First and foremost, task was to acquire data, for this particular paper three different dataset were acquired.

Step 2: After the data acquisition, same data was cleaned. Here only one basic operation was performed that being removal of empty cells.

Step 3: Once data was cleaned, a new dataset was created using normalization.

Step 4: Both datasets were split into separate training and testing sets.

Step 5: Training datasets from step 4 are used to regression algorithms (SVR and QSVR) to obtain a trained model.

Step 6: Trained models from step 5 are tested against respective testing sets created in step 4.

Step 7: Based on results in step 6, RMSE score were calculated.

Step 8: Two variants of one each from SVR and QSVR were selected with least RMSE score.

Step 9: Selected models from step 8 were plotted against actual data separately.

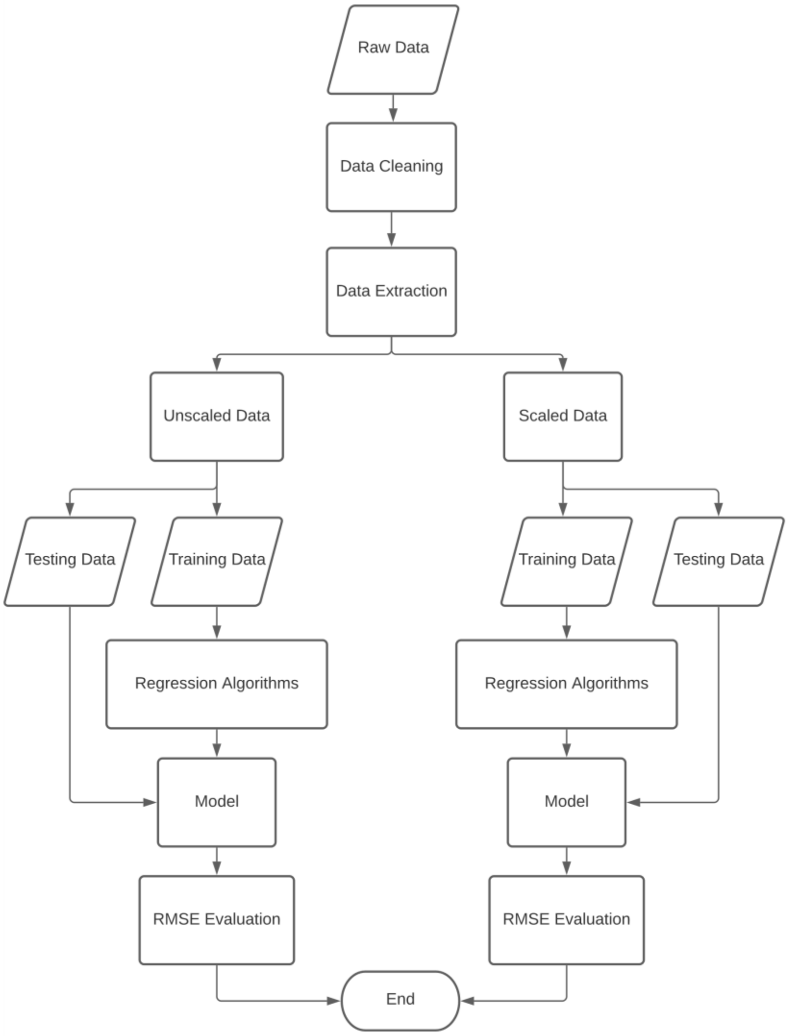

Fig.1 - Methodology Flow Diagram.

Results

Based on Table 8 it can observations can be made

- Scaling the dataset improved the accuracy of the model.

- Out of SVR and different QSVR variants based on no. of qubits, it is noted that SVR performed better in all three datasets.
- In QSVR variants a trend can be seen, increasing the no. of qubits either increases or decreases accuracy of model.
- In case of unscaled data, SVR outperformed QSVR with good margin.

Dataset Name	Type	SVR(Poly Kernel)	One Qubit	Two Qubit	Four Qubit	Six Qubit	Eight Qubit	Ten Qubit
US Monthly Unemployment Rate 1948 – 2019	Unscaled	0.171516	2.119085	2.285849	2.40917	2.432553	2.429157	2.459904
	Scaled	0.170237	0.183304	0.187507	0.212265	0.207455	0.183257	0.17088
Microsoft Stock	Unscaled	100.909092	131.748994	130.621682	130.214795	129.950459	129.466839	129.051765
	Scaled	9.434313	25.231973	39.439135	60.470557	69.661607	74.152382	77.30001
Climate Change: Earth Surface Temperature Data	Unscaled	2.325771	3.978409	4.072693	4.074688	4.100872	4.10366	4.114682
	Scaled	2.109456	2.095314	2.094846	2.104333	2.115063	2.129042	2.139778

Table 4 - RMSE Score for all Datasets.

These recordings were used to determine plots for QSVR (model with least RMSE was plotted) whereas in case of SVR model with Poly kernel is plotted.

US Monthly Unemployment Rate 1948 – 2019

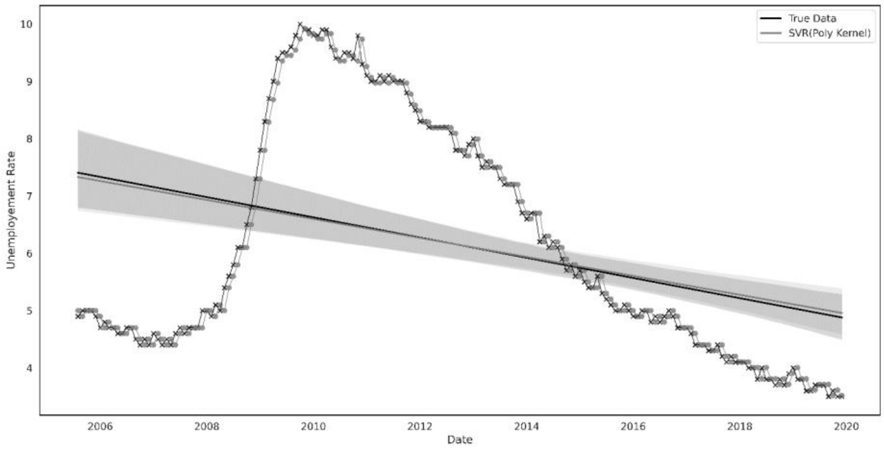

Fig.2 - SVR chart for Scaled Unemployment Dataset.

International Journal of Advance Research in Science and Engineering
Volume No. 11, Issue No. 02, February 2022
www.ijarse.com
IJARSE
ISSN 2319 - 8354

The RMSE for SVR in both scaled and unscaled unemployment dataset is quite similar, in fact the graphs are nearly same too. From graph in Figure 2 it can be observed that the regression line of both predicted values and true data are nearly same with some variance. Similar behavior is seen in graphs of unscaled unemployment dataset.

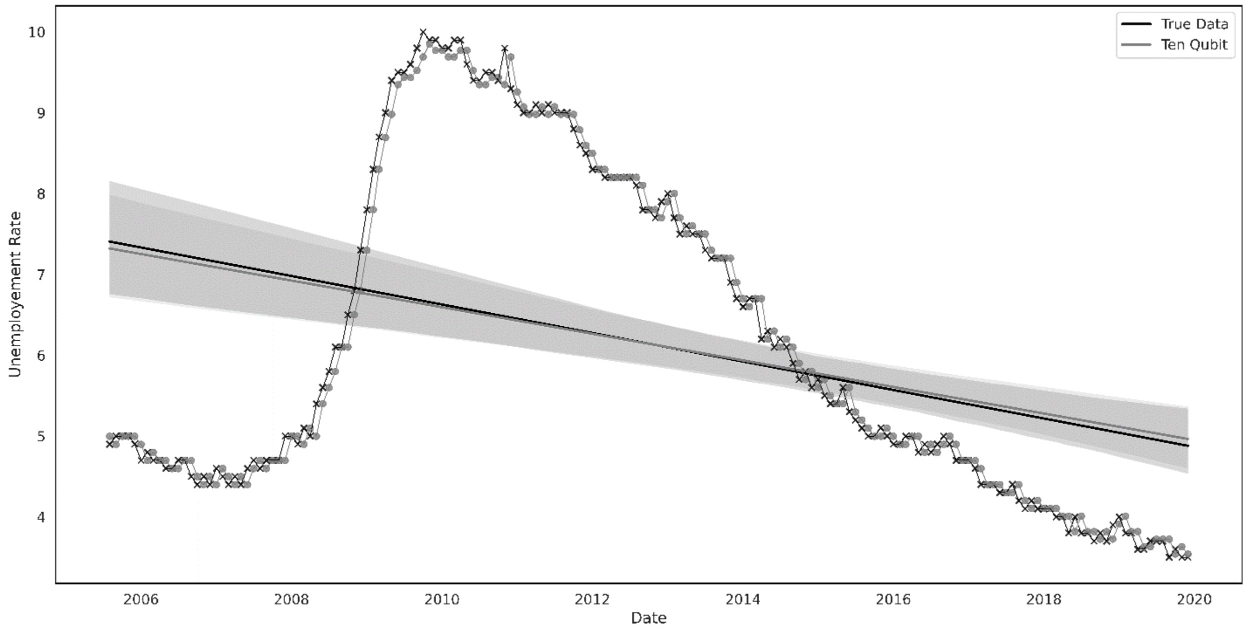

Fig.3 - QSVR chart for Scaled Unemployment Dataset.

As compared to SVR, QSVR showed significant improvement in results when dataset was scaled. The QSVR model gave poor results in case of unscaled dataset. The results were inverted when the data started trending. Whereas, in case of scaled dataset the results as shown in Figure 3 were similar to SVR model. In case of unscaled dataset with increasing qubits the accuracy was also decreasing whereas in case of scaled dataset the accuracy was increasing with increasing qubits.

Microsoft Stock

IJARSE
ISSN 2319 - 8354

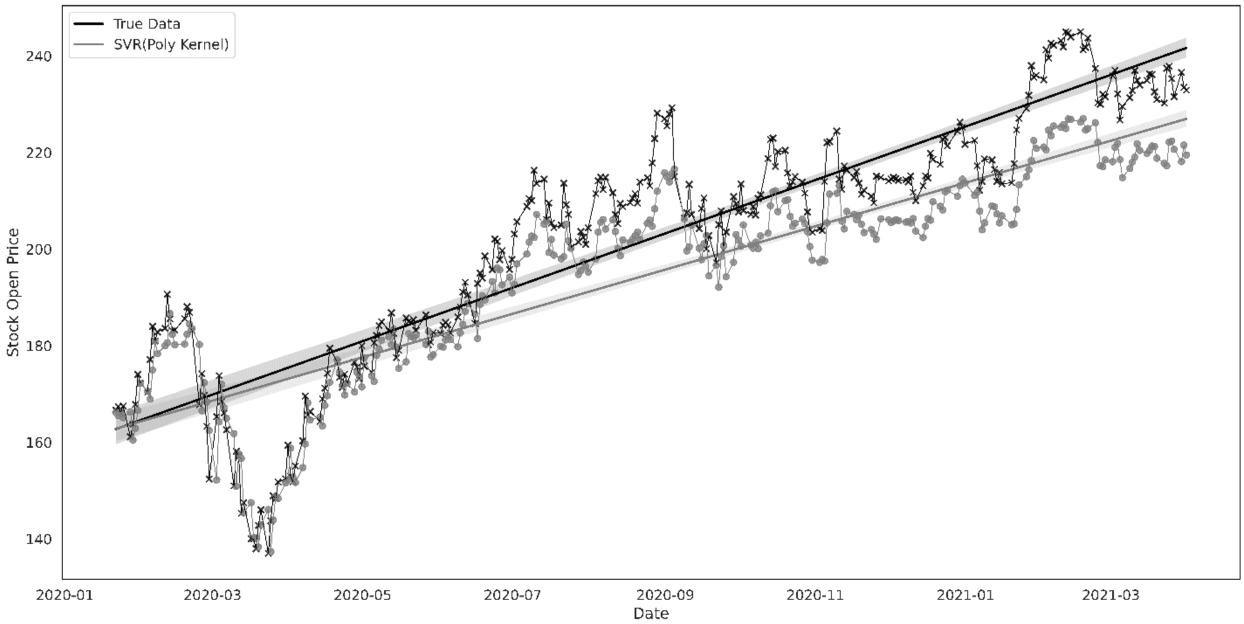

Fig.4 - SVR chart for Scaled Stock Dataset

In this dataset, only open prices were used for analysis. So, Figure 4 shows predicted values along with actual values of open prices over a period. In case of unscaled dataset, SVR gave quite high RMSE values and when plotted the graph showed no trend at all, most of the predicted values were constant. Whereas, in scaled dataset as it can be seen in Figure 4, the results are good. In addition to that the RMSE is low as well. It can be seen with time even though the predicted values are diverging from actual data, but the trend is maintained.

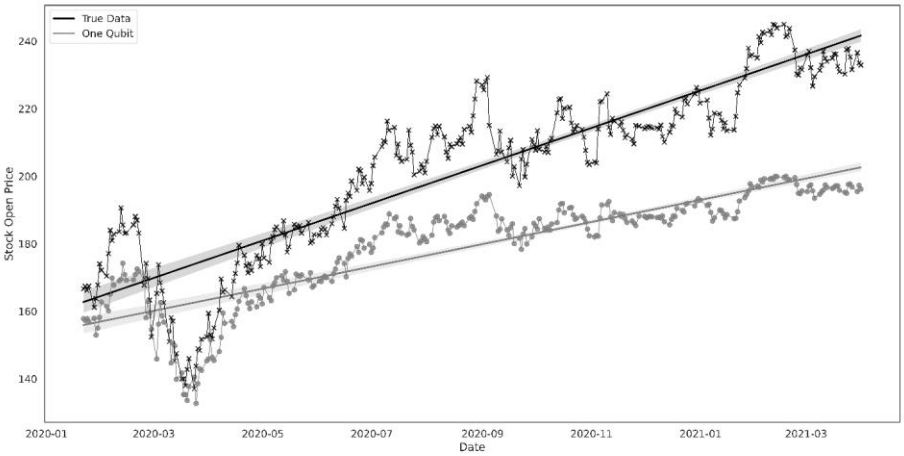

Fig.5 - QSVR chart for Scaled Stock Dataset

As for QSVR, RMSE improved as compared to unscaled data. It can be observed that the regression line is diverging similarly to SVR chart. Even though RMSE values are low, the graph indicates that over time the divergence will increase and along with-it accuracy will be affected. As compared to SVR chart where in beginning the results were quite similar in QSVR the results are varying from start and with time the variance increases. It can be also observed that with time the results start to converge along regression line. In QSVR as no. of qubits are increased the accuracy showed increasing trend in scaled dataset and decreasing trend in unscaled dataset.

Climate Change: Earth Surface Temperature Data

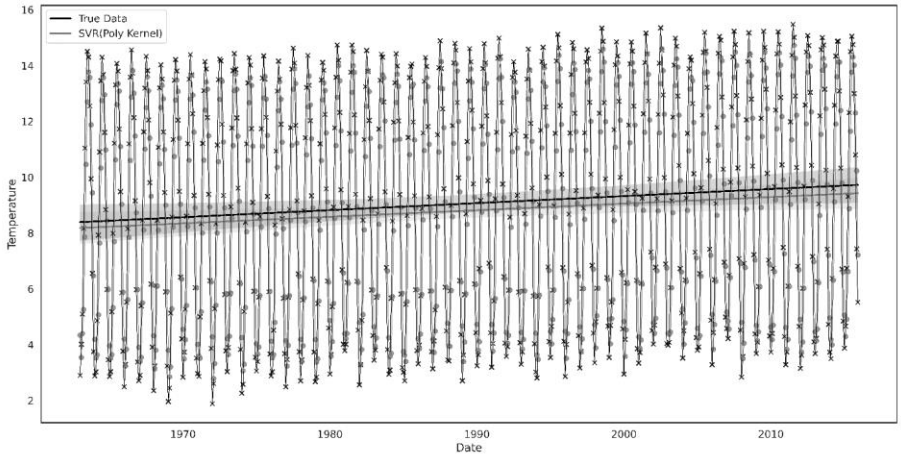

Fig.6 - SVR chart for Scaled Temperature Dataset

In climate change dataset, the dataset is present for monthly Land Average Temperature. As we know that temperature shows a certain trend in a year. This trend is highlighted in the graph by a zigzag wave pattern. RMSE in case of SVR was approximately equal for scaled and unscaled dataset. It can be observed that the regression lines in Figure 6 parallel to each other meaning the prediction were in same trend and were off by little amount. Similar observations can be made in case of unscaled dataset.

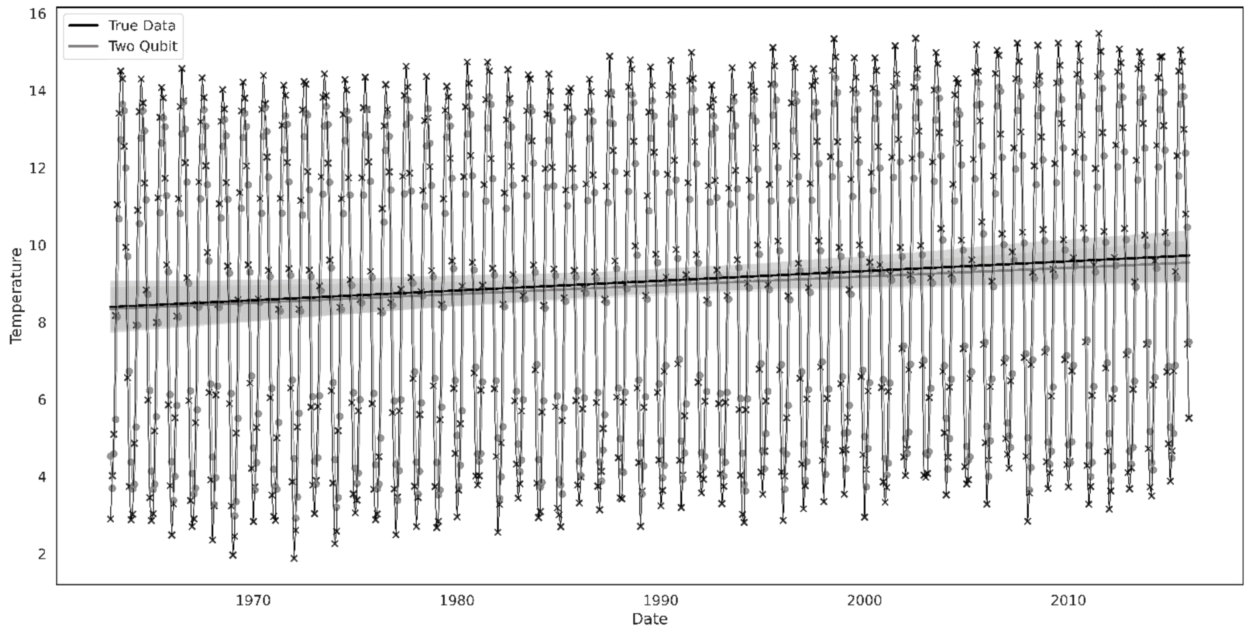

Fig.7 - QSVR chart for Scaled Temperature Dataset

QSVR showed good improvement in accuracy for scaled dataset as compared to unscaled dataset. RMSE for unscaled is approximately double RMSE of scaled dataset. It can be seen from Figure 7 that regression line overlaps each other in the start and diverge over time. Even though divergence is not too much but it may increase if time span is increased. Whereas for unscaled data the predicted values were off by a lot. RMSE showed no consistent trend with increasing qubits for both scaled and unscaled dataset.

Discussion

Several studies have been done in field of time series forecasting using many different techniques. Out of those techniques SVM (Support Vector Machine) is one of the famous techniques. Several studies have been done using SVM to forecast time series datasets. For instance, Chen et al. [21] used SVM to make load prediction for a EUNITE competition database. The focus of paper was to forecast daily load demand for a month using data from the previous two years. It was observed that without taking temperature into account, the MAPE for the winter season was 1.95 percent, and for the months of January and February, it was 2.54 percent. To forecast a regional yearly electric load in Taiwan Hong in his two studies developed a SVM regression model using immune algorithm (IA). The forecasted MAPE for the four regions is all less than 2.45 percent, according to the SVRIA model. Another study to forecast building energy consumption in a tropical region Dong et al. [22] used SVM and found that prediction results have a CV less than 3% and percentage error less than 4%. In one study four modelling strategies were described by Li et al. for the forecast of hourly cooling demand for an office building in Guangzhou, China [23]. The SVR model's mean relative error (MRE) was less than 1.02 percent in four months of testing. To forecast Turkey's electricity usage, Ogcu et al. used ANN and SVM. For the testing dataset, the MAPE for ANN and SVM was 3.9 percent and 3.3 percent, respectively [24]. Apart of studies listed above, many more studies have been made using SVM for time series forecasting.

In this paper an analysis was conducted using base parameters for SVR and QSVR. The SVR findings were comparable with previous studies. The RMSE score was low indicating the models were accurate. As many studies have been done using SVM, so it is considered as basis for QSVR performance. In most of the cases QSVR performed either on par or worse than SVM. Also, the results can be said good as the models were trained and tested on three different kind of time series and of variable length. Based on the results explained above, it was noted that when data is scaled it gives better results as compared to raw data. Apart from that in QSVR, increasing the number of qubits results in a trend in RMSE score it was either increasing or decreasing with one exception of two qubit in climate change dataset.

The study can be easily replicated using appropriate libraries and dataset. To be sure to get similar results, one should follow steps explained in dataset transformation and methodology section. This study was purely focused in univariate time series.

Conclusion

It is concluded that whether the univariate time series data is scaled or unscaled, the SVR algorithm always performs similarly or better than the QSVR algorithm for time series forecasting. It was expected that the QSVR algorithm would give better results as compared to SVR, but the findings didn't support our expectations. As Quantum computing and quantum machine learning are developing rapidly, soon it can be expected that with quantum hardware at our disposal, the QSVR algorithm may perform better than the SVR algorithm. Soon, we may have several other classical machine learning algorithms that will be implemented on quantum hardware and will use the full potential of quantum computers, which may lead to extraordinary results.

References

[1] S Shalev-Shwartz and S Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014. Accessed: Oct. 07, 2021. [Online]. Available: https://scholar.google.com/scholar_lookup?title=Understanding%20Machine%20Learning%3A%20From%20Theory%20to%20Algorithms&publication_year=2014&author=S.%20Shalev-Shwartz&author=S.%20Ben-David

[2] E. Alpaydin, *Introduction to machine learning*. 2020. Accessed: Oct. 07, 2021. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=tZnSDwAAQBAJ&oi=fnd&pg=PR7&ots=F3USbV6rDg&sig=gyOlGOu1hFjd0O0rWG0Ga7f54nA

[3] J. D. Scargle, "An introduction to chaotic and random time series analysis," *International Journal of Imaging Systems and Technology*, vol. 1, no. 2, pp. 243–253, 1989, doi: 10.1002/IMA.1850010213.

[4] A. Banerjee, G. Gallo, and E. Otranto, "Frontiers in time series analysis: Introduction," *Oxford Bulletin of Economics and Statistics*, vol. 68, no. SUPPL. 1, pp. 679–682, Dec. 2006, doi: 10.1111/J.1468-0084.2006.00451.X.

[5] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, "Introduction to Machine Learning, Neural Networks, and Deep Learning," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 14–14, Jan. 2020, doi: 10.1167/TVST.9.2.14.

[6] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:: The state of the art," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35–62, Mar. 1998, doi: 10.1016/S0169-2070(97)00044-7.

[7] J. G. de Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443–473, Jan. 2006, doi: 10.1016/J.IJFORECAST.2006.01.001.

[8] R. F. Engle, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, vol. 50, no. 4, p. 987, Jul. 1982, doi: 10.2307/1912773.

[9] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, Apr. 1986, doi: 10.1016/0304-4076(86)90063-1.

[10] D. Pradeepkumar and V. Ravi, "Forecasting financial time series volatility using Particle Swarm Optimization trained Quantile Regression Neural Network," *Applied Soft Computing Journal*, vol. 58, pp. 35–52, Sep. 2017, doi: 10.1016/J.ASOC.2017.04.014.

[11] M. Situm, "Time Series Volatility Forecasting Using Linear Regression and GARCH," *SSRN Electronic Journal*, Dec. 2011, doi: 10.2139/SSRN.1549385.

[12] H. Levenbach, "Time series forecasting using robust regression," *Journal of Forecasting*, vol. 1, no. 3, pp. 241–255, 1982, doi: 10.1002/FOR.3980010304.

[13] N. Nava, T. di Matteo, and T. Aste, "Financial time series forecasting using empirical mode decomposition and support vector regression," *Risks*, vol. 6, no. 1, Mar. 2018, doi: 10.3390/RISKS6010007.

[14] R. I. Rasel, N. Sultana, and P. Meesad, "An efficient modelling approach for forecasting financial time series data using support vector regression and windowing operators," *International Journal of Computational Intelligence Studies*, vol. 4, no. 2, p. 134, 2015, doi: 10.1504/IJCISTUDIES.2015.071180.

[15] K. J. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1–2, pp. 307–319, 2003, doi: 10.1016/S0925-2312(03)00372-2.

[16] E. Rieffel and W. Polak, "An introduction to quantum computing for non-physicists," *ACM Computing Surveys*, vol. 32, no. 3, pp. 300–335, 2000, doi: 10.1145/367701.367709.

[17] M. Schuld, I. Sinayskiy, and F. Petruccione, "An introduction to quantum machine learning," *Contemporary Physics*, vol. 56, no. 2, pp. 172–185, Apr. 2015, doi: 10.1080/00107514.2014.964942.

[18] "US Monthly Unemployment Rate 1948 - Present | Kaggle." https://www.kaggle.com/tunguz/us-monthly-unemployment-rate-1948-present (accessed Oct. 08, 2021).

[19] "Microsoft Stock- Time Series Analysis | Kaggle." https://www.kaggle.com/vijayvvenkitesh/microsoft-stock-time-series-analysis (accessed Oct. 08, 2021).

[20] "Climate Change: Earth Surface Temperature Data | Kaggle." https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv (accessed Oct. 08, 2021).

[21] B. J. Chen, M. W. Chang, and C. J. Lin, "Load forecasting using support vector machines: A study on EUNITE Competition 2001," *IEEE Transactions on Power Systems*, vol. 19, no. 4, pp. 1821–1830, Nov. 2004, doi: 10.1109/TPWRS.2004.835679.

[22] B. Dong, C. Cao, and S. E. Lee, "Applying support vector machines to predict building energy consumption in tropical region," *Energy and Buildings*, vol. 37, no. 5, pp. 545–553, May 2005, doi: 10.1016/J.ENBUILD.2004.09.009.

[23] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. Mochida, "Applying support vector machine to predict hourly cooling load in the building," *Applied Energy*, vol. 86, no. 10, pp. 2249–2256, 2009, doi: 10.1016/J.APENERGY.2008.11.035.

[24] G. Oğcu, O. F. Demirel, and S. Zaim, "Forecasting Electricity Consumption with Neural Networks and Support Vector Regression," *Procedia - Social and Behavioral Sciences*, vol. 58, pp. 1576–1585, Oct. 2012, doi: 10.1016/J.SBSPRO.2012.09.1144.