Analyzing time series on Rossmann sales using Deep neural networks and Support Vector Machines

Chetan Sharma
National College of Ireland
Dublin,Ireland
Divyang Jain
National College of Ireland
Dublin,Ireland

Aarush Sakhuji
National College of Ireland
Dublin,Ireland
Yogesh Sanjay Golecha
National College of Ireland
Dublin,Ireland

***Abstract*- Rossmann is a famous store of drug chain across the European continent. Presently it operates around three thousand stores in seven European countries. If we talk about the daily task of the managers of the stores then it's about to find about the sales of their stores on the daily basis and this will be stretched up to six weeks. This paper is extended by using machine learning techniques such as Support Vector Machine(SVM) and Deep neural networks [LSTM,GRU] and their empirical results are compared to dig out the predictions of the sales and use the same for the Rossmann Store. Deep neural network is famous but a tedious approach to use, but if done with proper concentration it can work as a key to the hidden patterns of the problems and since it consists of multiple layers running through it, hence it can act as a method of giving out satisfactory results.**

Keywords*- *Deep Neural Network, Support Vector Machine, Machine Learning, GRU, LSTM, time series

I. INTRODUCTION

The idea for the project has been grabbed from the competition "Rossman Sales" held by the Kaggle and it holds huge sets of data which is used for the modelling for the prediction purposes and analysis. This type of platform is quite beneficial. These types of companies have large forms of data and this is all used for gaining knowledge for strengthening business of all sorts and some sort of solved questions which remain hidden but are essential for the growth of the business. When similar functions are taken into consideration, these ideas and queries are used to bring out the data which is engraved in Kaggle competition and are used to find out all answers of the questions.

In Kaggle, there is a competition held to analyze and predict the sales for Rossmann stores using different techniques. A lot of techniques have been applied and analyzed to make prediction for the sales for the stores of rossmann all over the world. The data is of 3 years (2013,2014,2015) for the rossmann stores. The data for Rossmann is available at: https://www.kaggle.com/c/rossmann-store-sales.

Research related to forecasting, the rossmann sales were published using various different algorithms and among that few are discussed in this paper that uses machine learning algorithms to predict the rossmann sales. This paper includes many parameters which are used to find the target variable via time series forecasting on sales hence making managers capable enough to tell about sales that were done before or will be done in future. As suggested by [1] and [2], SVM and deep neural network are regarded as efficient predictive model. Hence, this study focuses on enhancing the accuracy of the model by using

SVM and different algorithms of deep neural network. So the question which is formed regarding the same is " *Can the technique of deep neural networks and SVM be used to increase the accuracy of the model by deploying time series forecasting for predicting Rossmann Sales?".*

II. RELATED WORK

This particular section deals with research being carried by us on the Rossmann Sales for predicting sales of first six weeks by deploying different techniques. In this we were able to derive different efficient models for predictions to be carried out using their data set and regarding the same, the case study has been explained below. During later period of time, we will be able to give an insight into the intelligence of the business for the sales of drugs across the stores of the Europe using above explained features and construct an analytically correct query. This is to be carried out by the comparisons to be held.

A study is introduced by [3] who was the winner of the Kaggle competition which suggested that XGBOOST as a tool of decision tree. By using this, the sales of further six weeks was predicted. I focused on three main principles which were data which was recent, information that was temporal and current trends. He explored holdout set of last six week sales history. He also added some features on the number of holidays during the current, last and next week on the same data set. This research includes the ridge regression for penalizing the complexity. Some extra features were added into it to reduce the over fitting.

He focused on the customer rather than the sales and transformed the dependent variables (sales) and did not include zero in the training data set. His interpretation shows the fairly stationary process. To deviate from the research we have used deep neural networks in which we have focused on the sales and for the feature engineering [4], we divided the data set of past five months into various time lags to enhance the accuracy. Moreover [5]. suggested that neural networks can be used in the future research hence we are continuing the research.

Another study by [6] states that he have worked on the same data set by using blended techniques of time series with statistical and traditional approach such as ARIMA, linear model and machine learning like XGBoost and probabilistic approach of Bayesian inference to find the distribution of the model parameters. ARIMA is used for the R package "forecast" and linear regression with LASSO regularization using R package "lars". In this research MSE for the ARIMA model 0.11, MSE for XGBoost model 0.07 and for linear blending of the ARIMA and XGBoost models MSE was 0.093. This research uses Bayesian inference using Markov Chain Monte Carlo (MCMC) algorithm using T-Copula and ensured the different regression coefficients. Conclusion from this research is that when we are using time series forecasting [7] and different techniques are used then we see that efficiency of our approach is increased significantly. SVM has been used in this research as being inspired from [4]. The reason behind the usage of SVM is that it is the efficient technique for time series forecasting, in the above given reference they have used frequency domain regression (FDR) and support vector regression (SVR) in which SVR outperformed FDR using the polynomial kernel with regularization. The three test result of the test error were linear regression: 32.%, Frequency Domain linear Regression: 29.% and Support Vector Regression: 17.%. Hence we will be using SVM by taking the remaining parameters of the data sets.

Proceeding to the next in line of our research, [8] K means is used to group the data and Markov decision process (MDP) to model store output and for this class, gradient boosting technique of decision tree making is used and SVR as a baseline algorithm. In Microsoft time series algorithm (MSTN), which is based in ARIMA, is used and as a result gradient boosting technique show achieved 12.3% RMS error to predict the Rossmann Sales without the customer data. This also motivated us to use the SVM model in our research as SVM was used as a baseline in finding the gradient boosting MSE error.

Long-Short Memory (LSTM) was performed by [9] in which it out performs traditional RNN, the research was based on simple context free and context sensitive languages in which RNN was compared with LSTM. As a conclusion LSTM showed the better results. Hence we will be using LSTM in our research methodology.

As we have discussed LSTM, now we will show comparison between LSTM and GRU to show that which of them is better and by [10]..There are various types of Recurrent units for comparison in RNN, some of them are sophisticated regarding the mechanism and some are not, one of them is LSTM unit and other one is GRU unit we compare them on the basis of modelling. The most common and traditional unit of RNN is tan h unit and we will be taking tan h unit as it can replace sigmoid function also it doesn't get saturated at 0 but it doesn't on +1 and -1 and can show the large range. LSTM works well on sequence based task but GRU is better for translation in machine that's why in some cases GRU is much efficient than LSTM and vice versa [10]. Result of this paper motivate us to compare the techniques in predicting Rossmann sales.

[11]states the Local Control for Linear regression (0.4170), Locally constant solution(0.2199), Locally constant solution (with 3 more features) (0.14221), Linear regressions by groups (0.125), Ensemble of XGBoost models(0.091) and Public Leaderboard for Locally constant solution (with 3 more features)(0.13693) and XGBoost random forests with 300 iterations (0.11520). We will try to get better scores using different techniques.

For the same dataset [5] has tested 5 different models and calculated the results for the same. Techniques which has been used are logistic regression, linear regression, KNN-20 neighbors, Random forest on 1000 trees and gradient boosting (1000 trees) and the scores are 0.33,0.38,0.30,0.11,0.10 respectively which results that random forest and gradient boosting are efficient techniques for this dataset. Similarly, we will be using different machine learning techniques to check and compare the results from deep learning-RNN-LSTM,GRU, neural network and SVM

III. RESEARCH METHODOLOGY

The purpose of the paper is to predict the Rossmann sales for 6 weeks across the 7 countries of Europe. The methodology used here is:

CRISP-DM-Cross industry standard process. Our project is divided into 6 different phases which are as follows:

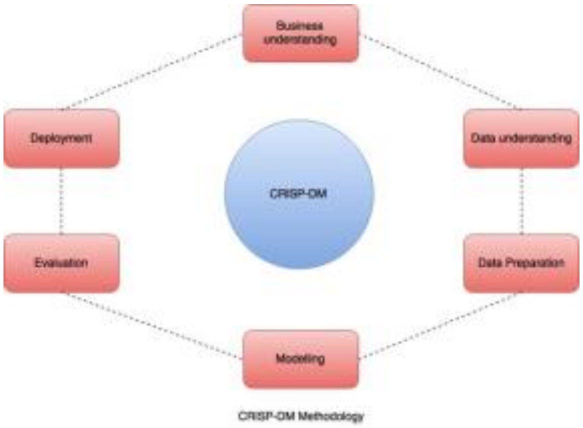
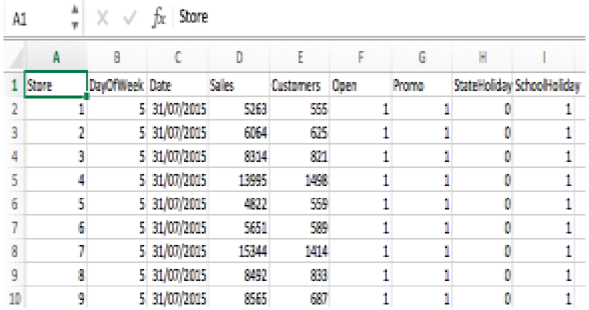


Fig-1 CRISP-DM Methodology
(Made from www. draw.io)

1.Business understanding- The purpose of business understanding is to fulfill the requirements for the business and to analyze the project requirements. The main reason behind business understanding is to find profit for the business. Analyzing sales for Rossmann can be of great help which can contribute towards market and to analyze and predict the resources and sales. `The data that we have collected gives us the deep insights for the store sales and various different aspects. Our analysis is based on predicting the 6 weeks of Rossmann store sales in 7 different geographical location which can be beneficial for the business to take effective measures and constantly improve their services.



	A	B	C	D	E	F	G	H	I
1	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
2	1	5	31/07/2015	5263	555	1	1	0	1
3	2	5	31/07/2015	6064	625	1	1	0	1
4	3	5	31/07/2015	8314	821	1	1	0	1
5	4	5	31/07/2015	13995	1498	1	1	0	1
6	5	5	31/07/2015	4822	559	1	1	0	1
7	6	5	31/07/2015	5651	589	1	1	0	1
8	7	5	31/07/2015	15344	1414	1	1	0	1
9	8	5	31/07/2015	8492	833	1	1	0	1
10	9	5	31/07/2015	8565	687	1	1	0	1

Fig-2 Snippet of the data

2.**Data understanding**-The data that we have collected for sales is from kaggle. As we know that the attributes in the dataset contributes towards analysing, so before predicting we performed feature engineering and performed different statistics like anomaly detection, histograms etc. to get a better understanding of the data. Also, redundant values were removed to maintain the quality of the data.

3.**Data preparation-** This module is of great importance of our analyses. Some variables ,such as sales, year, month, day of week are contributing more for our analyses. The remaining data was also good enough. Also, we found correlation between our variables.

4.**Modelling-** We have used 5 different black box techniques for our modelling,4 from deep neural networks-RNN, LSTM, GRU, and last one is SVM. This techniques are implemented using python's deep learning library called **Keras** which helped us to make our model. Other tool is Rapid miner which is used to model SVM. Prediction was based on the scores and MSE value.

5.**Evaluation-** To analyze our predictive models, the accuracy was checked using mean squared error (MSE).Also, graph were used to analyze our predicted and real values using time series forecasting.

6.**Deployment-** This is the final step for our model, as we have used different techniques. For each technique we got different results and prediction. So in this step,we have analyzed the best technique that fits the model means that technique which have got the minimum MSE value is the best technique and the managers can use that techniques for their stores to manage the stocks in future.

IV. IMPLEMENTATION

This section will be covering different works which will help our Rossmann sales dataset. We have set up our environment and code in **python** for deep neural nets and **rapid miner** for SVM.

1.Anaconda

We got inspired from [12] as Anaconda is a freemium open source dissemination of the Python and R programming dialects for substantial scale information handling, predictive analytics, and logical figuring, that intends to streamline bundle administration and deployment. Anaconda comes loaded with most of the data science libraries that we might need in performing our analysis.

2.Jupyter- Integrated Development Environment

The Jupyter Notebook is an open source web application that can be used as An IDE for our Python Project development and management. The advantage of using Jupyter as an IDE is that documents with the live code can be shared. Visualizations can be performed inline .It constantly performs auto-saves on the code and has a support for over 40 programming languages .Because of this we can easily work and switch between any languages like R or Python.

3.Pandas

Pandas is an essential Python library for Data Science giving easy to use and high performance data structures and data analysis tools that makes working with labeled and relational information both simple and natural.

Here are couple of the things that pandas does well:

Handling missing data in floating and non-floating point data.

Time Series usefulness: generating data range and recurrence change, moving window stats, date moving and slacking, and so on.

4.Keras-Deep learning Library

Keras is a deep learning library written on top of python which can be run on top of Tenserflow, Theano, MXNet. It gives features like user friendliness, modularity, easy configurability and extensibility .It is just an interface and not an end to end machine learning framework but provides intuitive set of high level abstractions irrespective of the backend scientific computing library being used. [14] motivated us to use this library for our project.

5.Matplot library-

In python and especially for the 2D plotting we use a special library which is called Matplotlib which is used for producing figures which are quality based. Matplotlib figure and charts helps in entire data analysis life cycle from data exploration to data visualization and then to report the results.

6.Pyneurgen

Pyraneurgen provides libraries to use in Python projects to fabricate neural networks and genetic algorithms. While the two strategies are helpful in their own particular rights, we are joining the two to empowers more noteworthy adaptability to take care of troublesome issues.

V. INVESTIGATION

In this section, after the implementation we have evaluated our solution of the project which includes different types of methodology .The dataset contains around **1.1 million rows** and **9 columns.**
Business understanding and data understanding has been done and elaborated above. Now it's time for third step of CRISP model that is data preparation.

A.Data Preparation

1.Feature Engineering

Feature engineering is basically the process of using the domain knowledge to tweak features or to introduce new features so that the machine laerning model works in a desired way.
After importing the data into python through jupyter, we summarised the data and looked that daily sales are between 0 and 41551, with the mean of 5774 and promo, open, school holiday always are between 0 and 1. While all other columns are integers and date type state holiday column is objects data type, it is converted to string. We also checked for the missing values and all missing values were dropped. After that various visualizations are performed on our data to get more insight into the data.

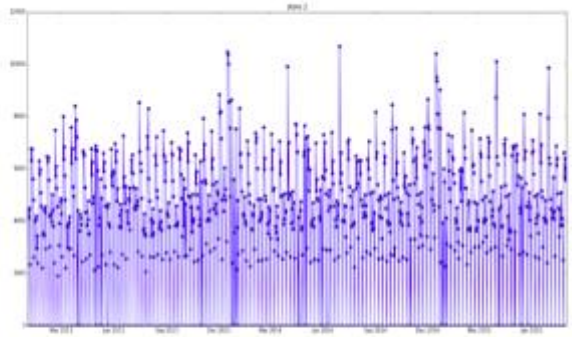


Fig-3 Cyclic Pattern seen

Above figure clearly shows us that seasonality and cycles ARE present in our data.
We can also infer that sales declined on every seventh day as most stores would be closed on Sunday and every alternate week sees a spike. To get a sense of corelation between sales and promo scatter plot is used which can be seen below.

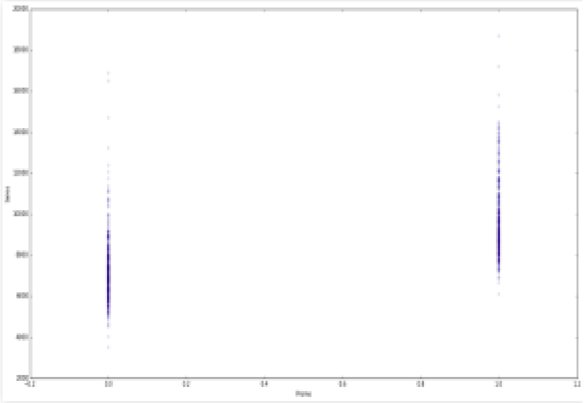


Fig-4 Scatter Plot

Sales got higher apparently when there were Promo on the same day. Now, as we have understood the dataset, we can move towards transforming the data.

Dropping Features: Features that were not needed for our model were dropped from the data. The features are ID, Store, Customers. Store Type, Assortment, Competition Distance, Competition open since month or year, PROMO2, PROMO2 since year or week and PROMO interval. The usage of the encoding namely one hot encoding for day in a week and holidays of the state. We have used binary encoding for OPEN, School Holidays and PROMO attributes by using get dummy() function of the PANDAS to predict the sales.
Log transformation have been applied Sales feature to normalize it. For this we have used mimmax Scalar function of sklearn python library.

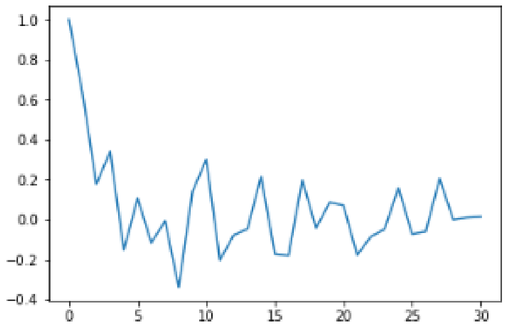


Fig 4.Partial Corelation for 40 time lags

To get better understanding of number of past observations to include in our model Partial Autocorrelation is used which is nothing but the amount of correlation between x_t observation with its lag x_{t-k} that can not be explained by other observations that come in between. We have used statsmodel library for this. Partial auto correlation like normal correlation ranges from +1 o -1 where +1 is a strong positive correlation .

Figure 4 plots the partial autocorrelation for 40 time lags and it is quite visible that past 2 time lags shows some association with current observation.

So we have introduced two new features in our data set Sales_at_TimeLag1 and Sales_at_Time lag2 by performing time lagging on our Sales feature .

And then Log transformation have been applied these two features as well.

2.Modeling

After the data preparation phase we have done our modelling in which we have built the three models using LSTM, GRU and SVM. One of the main thing in modeling deep neural networks is to set the learning rate appropriately and some evidences have been found in literature that supports use of automatic learning rate during stochastic gradient decent that can enhance performance so we have used an automated learning rate for our neural networks models for both LSTM AND GRU by using **rmsprop** algorithm which uses a various learning rate for each update on vector component that it makes. We tried various activation functions like sigmoid and tanh both of which gives you non linearity but our final model uses tanh activation function because of its wider range of -1 to +1 as compared to 0 and 1 range of sigmoid which can sometimes can be more effective for modeling highly non linear problems.

In the deep neural nets there are more number of hidden layers then Artificial neural nets which shows the effective results but before that we have judged the differences between the three years sales difference where there is **no correlation.**

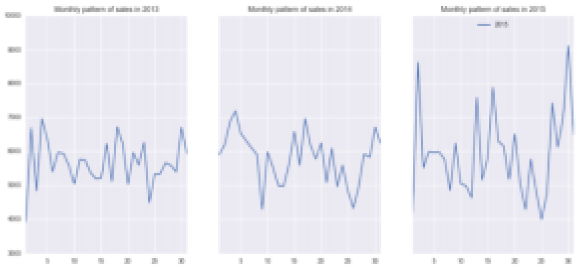
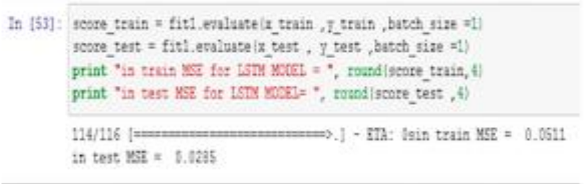


Fig-6 Three Years sales graphs

After looking at the fig 6, we can say there was not much relation between the years and hence we have performed the model for the particular year by dividing the day sales into two slots and same with them predicted the sales.

We have used MSE(mean square error) as our forecasting measure to compare perforomace of our two models made in LSTM ,GRU and SVM.



```
In [53]: score_train = fit1.evaluate(x_train ,y_train ,batch_size =1)
         score_test = fit1.evaluate(x_test , y_test ,batch_size =1)
         print "in train MSE for LSTM MODEL = ", round(score_train,4)
         print "in test MSE for LSTM MODEL= ", round(score_test ,4)

         114/116 [===========================>.] - ETA: 0sin train MSE =  0.0511
         in test MSE =  0.0285
```

Fig-7 MSE from LSTM

Fig 7 shows that the MSE value on which we are judging the accuracy is **0.028** (the value closer to zero is good) which is compared with other two models and evaluated the best model.

There is one graph from LSTM from which we can predict the future 6 weeks sales.

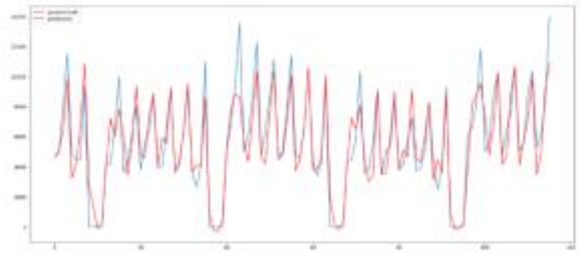
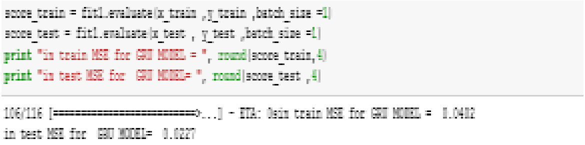


Fig-8 Sales Prediction from LSTM

We can see the accuracy is quite good from LSTM as the red line shows the prediction and blue line is the actual line.Hence, red line matches with blue line except the some points which is showing the good prediction of future 6 weeks.



```
score_train = fit1.evaluate(x_train ,y_train ,batch_size =1)
score_test = fit1.evaluate(x_test , y_test ,batch_size =1)
print "in train MSE for GRU MODEL = ", round(score_train,4)
print "in test MSE for  GRU MODEL= ", round(score_test ,4)
```

```
106/116 [========================>...] - ETA: 0sin train MSE for GRU MODEL =  0.0402
in test MSE for  GRU MODEL=  0.0227
```

Fig-9 MSE from GRU

The MSE value is 0.0227 which is slightly better the LSTM and we can see from the sales prediction graph as well.

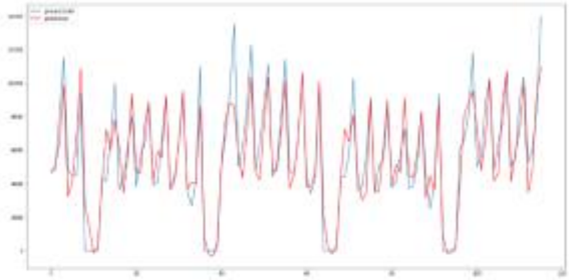


Fig-10 Sales Prediction from GRU

Fig 10 is looking similar to Fig 8 but slightly better prediction for the sales of 6 weeks.This comparison will be covered in next phase of CRISP-DM that is evaluation.

After the two RNN models from python, we have discovered the SVM model through rapid miner.

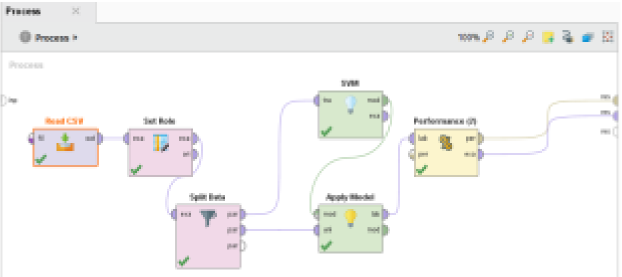
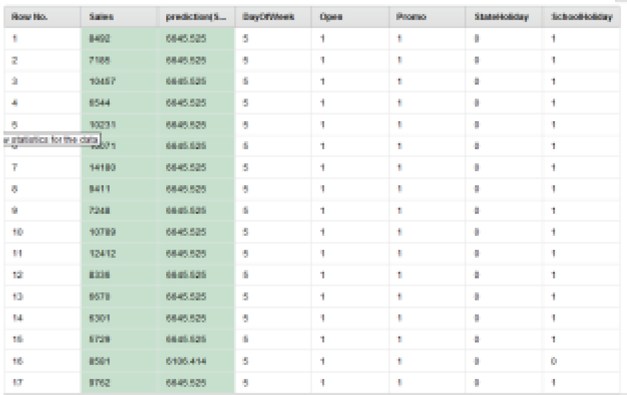


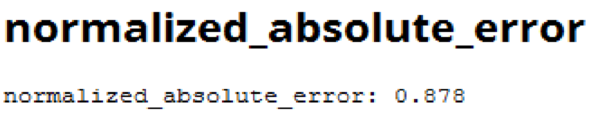
Fig-11 Modelling of SVM through Rapid Miner



Row No.	Sales	prediction(S...	DayOfWeek	Open	Promo	StateHoliday	SchoolHoliday
1	8492	6845.525	5	1	1	0	1
2	7188	6845.525	5	1	1	0	1
3	10457	6845.525	5	1	1	0	1
4	8544	6845.525	5	1	1	0	1
5	10231	6845.525	5	1	1	0	1
[illegible]	[illegible]071	6845.525	5	1	1	0	1
7	14180	6845.525	5	1	1	0	1
8	8411	6845.525	5	1	1	0	1
9	7248	6845.525	5	1	1	0	1
10	10789	6845.525	5	1	1	0	1
11	12412	6845.525	5	1	1	0	1
12	8338	6845.525	5	1	1	0	1
13	8579	6845.525	5	1	1	0	1
14	8301	6845.525	5	1	1	0	1
15	8729	6845.525	5	1	1	0	1
16	8591	6106.414	5	1	1	0	0
17	8762	6845.525	5	1	1	0	1

e statistics for the data

Fig-12 Snippet of the Predicted Sales



normalized_absolute_error

normalized_absolute_error: 0.878

Fig-13 Absolute Error from SVM

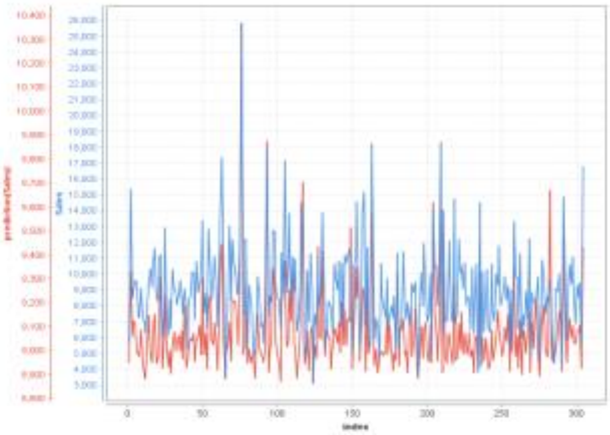


Fig-14 Predicted sales from SVM

As we can see from Fig-13,the error rate is more than LSTM and GRU that is 0.8 and Fig-14 shows the prediction from SVM which shows the correct prediction but varies at more points more than GRU and LSTM.

3.Evaluation

From the above three models we can say GRU was the best model that can be used to predict the sales in Rossmann store.

4.Deployment

The last phase of the CRISP-DM model says that deploy the best model to predict the sales.
Hence GRU is applied and the sales are predicted.

VI CONCLUSION

From the prediction of time series analysis(for six weeks) for the Rossman store sales, using GRU, LSTM and SVM, we can conclude that the best prediction and the minimum mean score value was obtained by GRU, and the techniques we have selected for the analysis have not being tested before. Also, GRU took the least computational time as compared to other models. So, after analyzing the results through Python and Rapidminer, we can say that GRU should be used and implemented for the applications for time series in the Rossman stores.

References:

[1] Claesen, Marc, et al. "Fast prediction with SVM models containing RBF kernels." arXiv preprint arXiv:1403.0736 (2014)

[2] Deng, L., Hinton, G. and Kingsbury, B., 2013, May. New types of deep neural network learning for speech recognition and related applications: An overview. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 8599-8603). IEEE.

*[3] Jacpbusse G.,Goes, december 2015,"Winning Model Documentation
describing my solution for the Kaggle competition "Rossmann Store Sales"*

[4] Guo, C. and Berkhahn, F., 2016. Entity Embeddings of Categorical Variables. arXiv preprint arXiv:1604.06737.

[5] Jia, F., Lei, Y., Lin, J., Zhou, X. and Lu, N., 2016. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mechanical Systems and Signal Processing, 72, pp.303-315.

[6] Pavlyshenko, B.M., 2016, August. Linear, machine learning and probabilistic approaches for time series analysis. In Data Stream Mining & Processing (DSMP), IEEE First International Conference on (pp. 377-381). IEEE.

[7] Muller, Klaus Robert, et al. "Using support vector machines for time series prediction." Advances in kernel methodssup- port vector learning, MIT Press, Cambridge, MA (1999): 243- 254.

[8] Beam, D. and Schramm, M., 2015. Rossmann Store Sales.

[9] Gers, F.A. and Schmidhuber, E., 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. IEEE Transactions on Neural Networks, 12(6), pp.1333-1340.

[10] Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

[11]Sazontyev, V., Rossmann store sales quantity prediction.

[12]Sheppard, K., 2012. Introduction to python for econometrics, statistics and data analysis. Self-published, University of Oxford, version, 2.

[13]Ragan-Kelley, M., Perez, F., Granger, B., Kluyver, T., Ivanov, P., Frederic, J. and Bussonnier, M., 2014, December. The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication. In AGU Fall Meeting Abstracts.

[14]Ketkar, N., 2017. Deep Learning with Python. A Hands-on Introduction.