Enhancing the Prediction accuracy of Video Game Products via Machine Learning and Deep Neural Networks Learning Approaches

***Abstract*— The market of video games has seen a tremendous growth since its inception in 1970s. Now-a-days, video games have become a daily form of entertainment for people of all ages around the world and hence it is a highly profitable market. The video game sales in U.S. spiked at $16.5 billion in 2017. In comparison, the film industry sold $29.2 billion in 2017 in the U.S. Forecasting is one of the most important activities for enhancing productivity and improving quality in the whole organizational function. This is due to the fact that most forecasting results are very influential in making managerial decisions and evaluating performance of the company products. The accurate demand forecasting is the fundamental aspect of businesses with supply chain management. This paper contributes to enhance the forecasting accuracy by using emerging, promising and powerful machine-learning techniques such as Random Forest, Naive Bayes, Logistic Regression, Deep Neural Network and ensembled model. We have implemented DNN(Deep Neural Networks) and ensembled model as base line methods for the purpose of comparing the results. Empirical results of these methods are used to find out the best predictions of the sales which can be further recommended to the video gaming industry. The report starts with introduction and related work. Further sections follow CRISP-DM process.**

Keywords*– *Random Forest, Naive Bayes, Logistic Regression, Deep Neural Network, Ensembled model, Video Gaming Industry.

I. INTRODUCTION

Video Game is an electronic game played on electronic medium devices such as personal computers, television screen, gaming consoles or mobile phones. Sometimes, the video game industry is called the interactive entertainment industry. There are many variations in input devices and game controllers across the platforms. In addition to keyboards and mouse, a common controller includes game pad, joysticks, the touch screens of mobile devices, buttons, etc. Players typically view the game on a video screen or TV screen and there are often incorporated game sounds from loudspeakers.

Video Game development has a long history since 1970's. The recent past, due to revolution of the smart phones and tablets, there has been new categories of video games such as mobile and social games. Virtual reality has been playing great roles to popularise video games and the developers are introducing various technologies and methodologies in the computing system to make them more interesting and interactive.

Forecasting is one of the most important activities for enhancing productivity and improving quality in the whole organizational function. It is due to the fact that most forecasting results are very influential in making managerial decisions and evaluating performance of the company products. The accurate demand forecasting is the fundamental aspect of supply chain management.

The idea for the project is taken from kernels held by Kaggle. Kaggle provide huge datasets, which are used for data modelling and analysis. Many techniques have been applied so far to make prediction for the sales for video gaming products all over the world. The data in the dataset is present from the year 1980 to 2017. The data for Video Gaming Sales is available at: https://www.kaggle.com/jruots/forecasting-video-game-sales/data

The motive behind this project is to visualize the data set and experiment with exploratory analysis. For the better understanding, this project have analysed the data by some histograms and plots, which help us to know the trend of the industry. Some statistical methods are also used to fit the data set during data pre processing.

So far, there are many different approaches of demand forecasting applied in a variety of areas. In early times, the most commonly used approaches are mainly statistical methods such as Trend Analysis and Extrapolation. The way to use this kind of method is quite simple and its cost is rather low. However, this method is only effective when factors such as industry trends and technology evolution are not considered. Practical reality is different and eventually the approaches with statistical methods have been weeded out. Hence, the business objective and research question is - ***Can deep learning and machine learning approaches be used to enhance the predictive performance of the model for forecasting Video Gaming Sales?***

II. RELATED WORK

Many experts and scholars have carried out comprehensive research on logistics concepts from modern logistics and economics, but still do not have a satisfactory conclusion. In this work, we have studied primary product demand logistics, which is defined as the social and economic activities of a certain period leading to the emergence of demand in form of space, time and cost requirements. It is based on configuration of primary products in production. This paper considers following research work done in product sale prediction.

As suggested by [2] and [14], deep learning neural networks are regarded as efficient predictive models, our study focuses performing experiments to enhance accuracy using deep-learning and machine-learning based algorithms. For comparison purpose, we are implementing these classifiers and deep-learning method along with ensembled model.

Chetan S. J. et al [1] describes prediction of sales with dataset from Rossman stores. Rossmann is a famous store of drug chain across the European continent. It operates around three thousand stores in seven European countries at present. The daily tasks of the managers in the stores comprises of finding sales of their stores on daily basis which is further stretched up to six weeks. The work is extended in this paper by using machine learning techniques such as Decision tree and Deep neural networks like LSTM, GRU and SNN. The empirical results are compared to dig out the predictions of the sales and recommended the same for the Rossman Store. The authors concluded that although deep neural network is famous it is tedious to use. If this approach is reconciled with proper concentration it can work as a key to the hidden patterns of the problems and it can act as a method of giving out satisfactory results as it consists of multiple layers running through it. We are inspired by the methods and results on time series prediction of sales data Rossman store and verify its applicability for prediction of gaming products sale.

Liu Y. et al [11] worked on Demand Forecasting by Using Support Vector Machine. Demand forecasting plays a crucial role for supply chain management of retail industry. The future demand for a certain product constructs the basis of its relevant replenishment system. In this research, the technique of Support Vector Machine (SVM) is employed for demand forecasting. Various factors that affect the product demand such as seasonal and promotional factors have been taken into consideration in the model. Meanwhile, different other approaches such as Statistical Model, Winter Model and Radius Basis Function Neural Network (RBFNN) are also used for comparison and evaluation.

Guo, C. et al [3] presented techniques of Demand Forecasting and Price Optimization. They presented the work with an online retailer, an example of how a retailer can use its wealth of data to optimize pricing decisions on a daily basis. It is in the online fashion sample sales industry, where they offer extremely limited-time discounts on designer apparel and accessories. One of the retailer's main challenges is pricing and predicting demand for products that it has never sold before, which account for the majority of sales and revenue. To tackle this challenge, they use machine-learning techniques to estimate historical lost sales and predict future demand of new products. The nonparametric structure of our demand prediction model, along with the dependence of a product's demand on the price of competing products, pose new challenges on translating the demand forecasts into a pricing policy. They have developed an algorithm to efficiently solve the subsequent multi-product price optimization that incorporates reference price effects. They implemented this algorithm into a pricing decision support tool for Rue La La's daily use. They conducted field experiments and found that a sale does not decrease due to implementing tool recommended price increases for medium and high price point products. They estimated an increase in revenue of the test group by approximately 9.7% with an associated 90% confidence interval of [2.3%, 17.8%].

Taweepol suesat et al [13] worked on Demand forecasting approach for inventory control for warehouse automation. This paper presented a technique to control the inventory level on the warehouse automation by using continuous reviewing and forecasting system. The product demand that stored in the warehouse depends on orders from customers. For each product, the demand follows a Poisson process and the delivery lead-time is known. The demand forecast can predict the trend of order and redefine the reorder point (RP) of inventory that linked to automatic warehouse control system. The functions of computer integrated manufacturing system (CIMS) included production planning, material requirement planning, work order generation, process control, quality control, shipping planning, warehouse and inventory management and material cost accounting.

Wang X. et al [14] carried out experiments for Demand Prediction Based on Neural Network Theory. Primary product logistics shares the challenges of other logistical problems, but also possesses many unique features which preclude the application of usual methods of the logistics of primary products. In particular, it is not possible to accurately forecast demand. To overcome the limitations of single logistics demand forecasting techniques and the difficulties in primary products logistics that exist currently. The authors have reported the use of neural network theory to establish a predictive model of the demand in primary products logistics based on a back-propagation (BP) neural network. The BP Algorithm used in the learning process includes two processes: forward computing of data stream and backward propagation of error signals, which make the output vector closer to the expected output vectors by continuous adjusting of weights, thus improving the accuracy of the logistics forecasting. Primary products demand and example Analysis verify the accuracy of this BP neural network-based prediction model for primary product demand.

The promising directions for forecasting of sales products are based on best prediction model. Hence, we build our methodology for different machine learning techniques. We use Deep Neural Network and ensembled model for comparison purpose to make our predictions.

III. METHODOLOGY

The methodology used here is **Cross-industry standard process for data mining,** commonly known by its acronym **CRISP-DM** (shown in fig 1.1). It is a data mining process model that describes commonly used approaches that data mining experts use to tackle problems. It is the leading methodology used by industry data miners who decide to respond to the survey. CRISP-DM breaks the process of data mining into six major phases as described below with reference to our work.

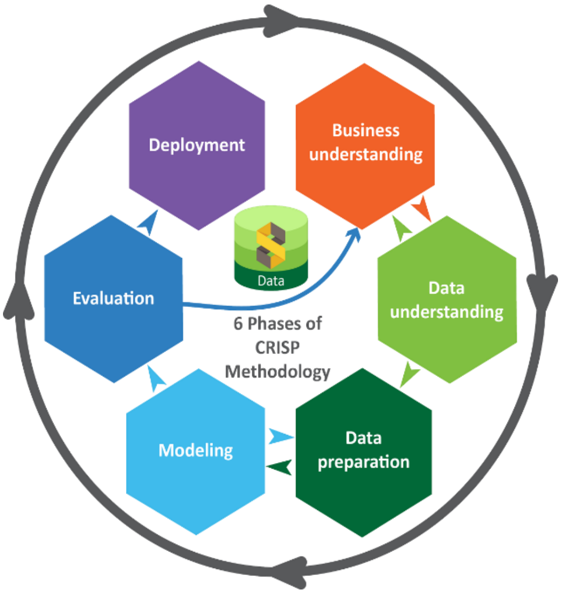
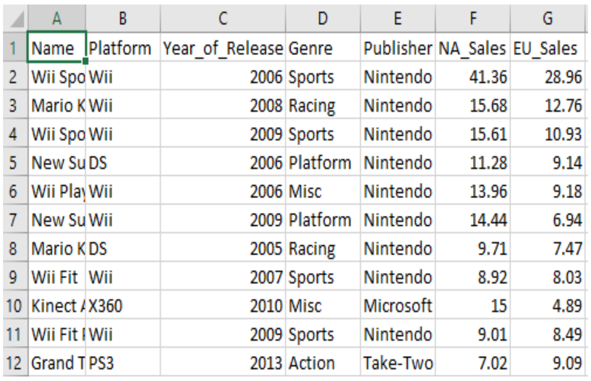


Fig. 1. Crisp-DM Methodology
(Source: predictt.net)

1. Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives. Analyzing sales for Gaming Product sale (Fig. 1) 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 gaming products platforms, year of release, genre and various different aspects. Our analysis is based on predicting from year 1985 to 2016 sales in 3 different geographical locations which can be beneficial for the business to take effective measures and constantly improve services in gaming product industry.



	A	B	C	D	E	F	G
1	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales
2	Wii Spo	Wii	2006	Sports	Nintendo	41.36	28.96
3	Mario K	Wii	2008	Racing	Nintendo	15.68	12.76
4	Wii Spo	Wii	2009	Sports	Nintendo	15.61	10.93
5	New Su	DS	2006	Platform	Nintendo	11.28	9.14
6	Wii Pla	Wii	2006	Misc	Nintendo	13.96	9.18
7	New Su	Wii	2009	Platform	Nintendo	14.44	6.94
8	Mario K	DS	2005	Racing	Nintendo	9.71	7.47
9	Wii Fit	Wii	2007	Sports	Nintendo	8.92	8.03
10	Kinect A	X360	2010	Misc	Microsoft	15	4.89
11	Wii Fit P	Wii	2009	Sports	Nintendo	9.01	8.49
12	Grand T	PS3	2013	Action	Take-Two	7.02	9.09

Fig. 2. Snippet of the Dataset

2. Data Understanding

The data-understanding phase starts with an initial data collection. It proceeds with activities in order to get familiar with the data. It is about to identify data quality problems and to discover the insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

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.

3. Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data.

This module deals with first sight analysis of data and make it ready for further model building. Some variables such as User count, rating, developer, genre, year of release, NA Sales, JP Sales etc. are contributing more for our analyses. A few instances are removed due to some missing values or null values. Also, we found correlation between our variables. This way, pre-processed data is partitioned into two parts – one for model building and the other for model testing and evaluation.

4. Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type.

In this project, we have used four techniques for modeling of Random Forest, Naive Bayes, Logistic Regression, Deep Neural Networks and Ensembled model. These techniques are implemented using Python's libraries. RapidMiner tool is used for processing ensembled model. Prediction is based on accuracy of scores and mean square error (MSE) values.

5. Evaluation

In this section, we will be covering the execution environment for modeling our Video Gaming Sales dataset. For the sake of reliability and better packages processing, we are using Anaconda, a distribution of Python and R. Python scripting is used for processing our models using machine learning techniques.

6. Deployment

This is the last step in CRISP-DM process and we have identified the best techniques. Here we have arrived at the solution of problem for business organizations. We have analysed the best technique that best fits the model. The model with the minimum MSE model is the best technique. We recommend that Random Forest should be used and implemented for the applications for forecasting gaming product sales.

IV. IMPLEMENTATION

In this section, we will be setting up the execution environment to implement different models on our Video Gaming Sales dataset. We will be using Python to perform Deep Neural Networks, Random Forest, Naive Bayes, Logistic Regression and RapidMiner for ensembled modeling which is a combination of Naive Bayes, Random Forest and KNN.

A. Anaconda

Anaconda is a free and open source distribution of Python programming language to perform machine-learning related applications. It is used for large-scale data processing and predictive analytics that intends to streamline business administration. This environment sets the best for our Video Gaming Sales as we are focusing on prediction of sales which will be used by Python as free source analytics.

B. Pandas

Pandas is an open-source Python library with fast access, adaptive and easy-to-use data structures used for data analysis. It has an enhanced objective of transforming into the most effective control tool used in any language. With the help of Python with Pandas, the following things were accomplished.

- Customized indexing of data and handling missing values.
- NumPy libraries for muti-dimensional objects into data frames.
- Performed operations for aggregation and transformations on data.
- Time Series usefulness: generating date range and recurrence change, date moving and slacking and so on.

C. Matplot library

Matplotlib is a Python 2D plotting library. It is used for drawing quality figures in different hardcopy formats and interactive environments across various platforms. We have used it in python scripts for easy understanding via histograms and other charts. It has numerical mathematical extensions NumPy. It has object-oriented API for embedding plots using GUI toolkits.

D. Scikit-learn

Scikit-learn, a free software machine learning library is used for programming in Python. It supports various classification, regression and clustering algorithms including random forest, k-means, etc. It operates well with Python numeric and scientific libraries viz., NumPy and SciPy.

E. Seaborn

This is a Python visualization library which is based on matplotlib that provides high level interface for representing statistical graphics. We are using this library for generating the confusion matrix.

V. INVESTIGATION

Once the implementation part is completed, the next step is to evaluate our solution of the project which incorporates different types of methodology. The dataset contains around 6922 rows 16 columns. We have already discussed about business understanding and data understanding in above sections. Now, we will investigate third step of CRISP-DM model i.e. data preparation.

A. Data Preparation

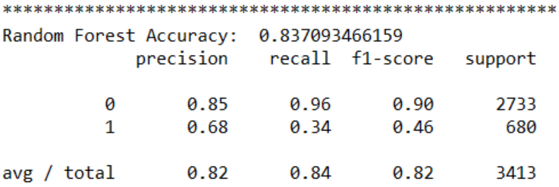
1.Feature Engineering

Feature Engineering is most important process in which the domain knowledge dependent on the data is used to create a design that would aid in selecting algorithms which are related to machine learning in a suitable format. Now, in order to predict the global sales of video games from the given data, we made transformation which will be used for machine learning algorithms i.e. Random forest, Logistic regression and Naive Bayes. We have converted polynomial to integer data via python code. The aim of the model is to predict whether the global sales hits the success rate or not, for this we used classification technique thus data is converted into 1 or 0. If global sales is greater than 1 million then it indicates 1 and if global sales is less than 1 million then it indicates 0. The logic has been implemented in Python. Note that the data for the global sales is in million. The performance of the implemented models is summarised using a confusion matrix., i.e. mainly used for classification problems.

2.Modeling

As of now, we have done with data preparation phase after this, we have built four models namely, Random forest, Logistic Regression, Naive Bayes and Deep Neural Network. RF, LR and NB are mainly used for classification problem as in our case we are predicting whether global sales is going to hit or not.

We are performed random forest algorithm for obtaining the accuracy of our global sales.



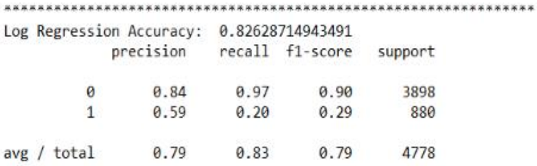
```
****************************************************
Random Forest Accuracy:  0.837093466159
             precision    recall  f1-score   support

          0       0.85      0.96      0.90      2733
          1       0.68      0.34      0.46       680

avg / total       0.82      0.84      0.82      3413
```

Fig. 3. Accuracy from Random Forest

From the above diagram, we could see that the accuracy of random forest is 83%. Here, we have also compute other performance factor for our model such as MSE is 0.16(value closer to zero is good), precision, recall etc are shown. The above result will be comparing with other models.



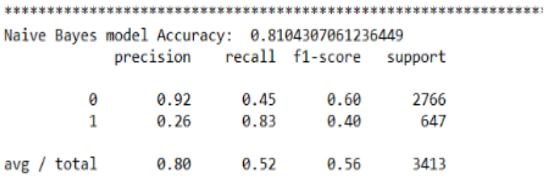
```
****************************************************************
Log Regression Accuracy:  0.82628714943491
             precision    recall  f1-score   support

          0       0.84      0.97      0.90      3898
          1       0.59      0.20      0.29       880

avg / total       0.79      0.83      0.79      4778
```

Fig. 4. Accuracy from Logistic Regression

Similarly, we have developed logistic regression on the same dataset. The accuracy for this model is 82% which is slightly less than the random forest and MSE is 0.17 which is also higher than the previous model. Again, we are implemented Naive Bayes model which is developed for predicting the global sales of video sales. The below figure shows the results of our model: As we can see that accuracy for this model is 81% which are again less than the previous models and MSE from Naive Bayes is 0.5 which are much higher error value than the random forest and logistic regression.



```
*****************************************************************
Naive Bayes model Accuracy:  0.8104307061236449
             precision    recall  f1-score   support

          0       0.92      0.45      0.60      2766
          1       0.26      0.83      0.40       647

avg / total       0.80      0.52      0.56      3413
```

Fig. 5. Accuracy from Naive Bayes

Another approach has been considered for comparing the above results with Neural Network. Hence, we have implemented Neural Network in Python for forecasting global sales for video gaming products. The accuracy obtained from Neural Network is 79% which is less than the machine learning algorithms.

Thus, we can say that for predicting the global sales for video game for this dataset the random forest algorithm is best algorithm.

As explained above we have implemented four models for our project in python. Now, we are going to discover ensembled model via RapidMiner. This result would be compared with all other models which we had implemented before. The ensembled model can be formed by combining two or more different algorithms and after that producing the output in a single score. This technique can be useful to improve the accuracy of predicting the global sales for video games. The novel approach is that ensembled model is not being used before for predicting global video game sales especially for this dataset.

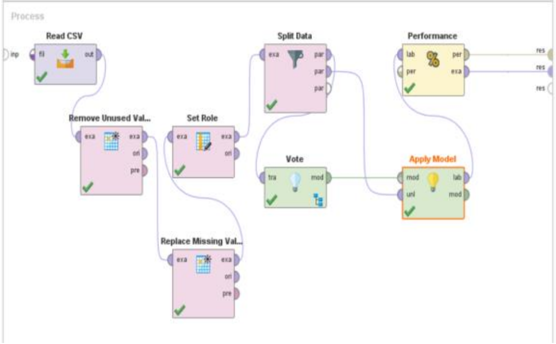


Fig. 6. Ensembled model through RapidMiner (Process Flow 1)

The above figure shows the ensemble model for predicting the global sales. First, we are reading .csv file with help of read csv operator. Once file is read, it then passes it to remove unused value operator for removing the missing and unused values. Set role operator is used for labelling the data. Here, we are splitting our data into 70:30 ratio. The important operator is vote which has three models in it i.e. Naive Bayes, Random Forest and K-NN which is used for predicting inner layers as shown in fig. Apply model operator is used to create a new model from train and test data. Here, our aim is to obtain a prediction on invisible data by applying a prepossessing model.

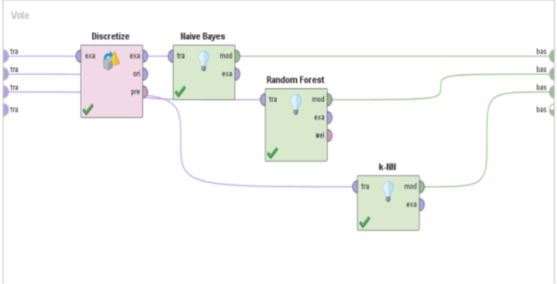
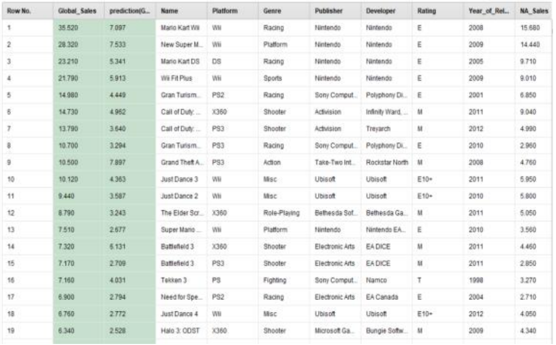


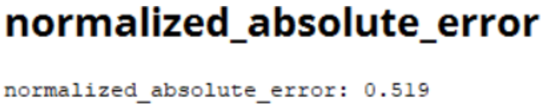
Fig. 7. Process Flow 2

It is comparing all the three techniques and find out the best result that would predict the global sales. Here, we have evaluated all the three models that would generate best performance metrics of all.



Row No.	Global_Sales	prediction(G...	Name	Platform	Genre	Publisher	Developer	Rating	Year_of_Rel...	NA_Sales
1	35.520	7.097	Mario Kart Wii	Wii	Racing	Nintendo	Nintendo	E	2008	15.680
2	28.320	7.533	New Super M...	Wii	Platform	Nintendo	Nintendo	E	2009	14.440
3	23.210	5.341	Mario Kart DS	DS	Racing	Nintendo	Nintendo	E	2005	9.710
4	21.790	5.913	Wii Fit Plus	Wii	Sports	Nintendo	Nintendo	E	2009	9.010
5	14.980	4.449	Gran Turism...	PS2	Racing	Sony Comput...	Polyphony Di...	E	2001	6.850
6	14.730	4.962	Call of Duty: ...	X360	Shooter	Activision	Infinity Ward, ...	M	2011	9.040
7	13.790	3.640	Call of Duty: ...	PS3	Shooter	Activision	Treyarch	M	2012	4.990
8	10.700	3.294	Gran Turism...	PS3	Racing	Sony Comput...	Polyphony Di...	E	2010	2.960
9	10.500	7.897	Grand Theft A...	PS3	Action	Take-Two Int...	Rockstar North	M	2008	4.760
10	10.120	4.363	Just Dance 3	Wii	Misc	Ubisoft	Ubisoft	E10+	2011	5.950
11	9.440	3.587	Just Dance 2	Wii	Misc	Ubisoft	Ubisoft	E10+	2010	5.800
12	8.790	3.243	The Elder Scr...	X360	Role-Playing	Bethesda Sof...	Bethesda Ga...	M	2011	5.050
13	7.510	2.677	Super Mario ...	Wii	Platform	Nintendo	Nintendo EA...	E	2010	3.560
14	7.320	6.131	Battlefield 3	X360	Shooter	Electronic Arts	EA DICE	M	2011	4.460
15	7.170	2.709	Battlefield 3	PS3	Shooter	Electronic Arts	EA DICE	M	2011	2.850
16	7.160	4.031	Tekken 3	PS	Fighting	Sony Comput...	Namco	T	1998	3.270
17	6.900	2.794	Need for Spe...	PS2	Racing	Electronic Arts	EA Canada	E	2004	2.710
18	6.760	2.772	Just Dance 4	Wii	Misc	Ubisoft	Ubisoft	E10+	2012	4.050
19	6.340	2.528	Halo 3: ODST	X360	Shooter	Microsoft Ga...	Bungie Softw...	M	2009	4.340

Fig.8. Snippet of the predicted Sales



normalized_absolute_error

```
normalized_absolute_error: 0.519
```

Fig. 9. Absolute error from Ensemble model

From the figure we can see that the error rate is more than random forest and Logistic regression that is 0.52. Here, we are not only focuses on the absolute error in our model but also considered other performance criteria which are essential for judging the best predictive model such as root mean squared error which is to evaluate the accuracy of continuous variable in our case we obtained 1.2, absolute error is 0.38 relative error 178.71% +/- 633.31%, correlation 0.772, root relative squared error 0.729 etc.

3. Evaluation:

From the above four models, we can suggest that random forest was the best model that would be valuable to predict the global sales for video game industry.

4. Deployment

This is the last phase of the CRISP-DM methodology that includes deploying the best model to predict the video sales in global market. In our project, random forest is applied, and the best results for Random Forest were attained.

VI. CONCLUSION AND FUTURE WORK

The market of video game product sale has been increasing since its inception. Forecasting the sale of video products is useful for the growth of gaming industry. We have surveyed and selected promising ones for prediction of and demand forecasting. The variations of deep learning approaches are gaining popularity in this domain. We implemented simple deep neural network and machine learning algorithms viz., Random Forest, Naive Bayes and Logistic Regression. Deep Neural Networks and ensembled model were used as base line methods for comparison. Among the many techniques, we conclude that the best prediction and the minimum mean score value was obtained by Random Forest. In addition, Random Forest took the least computational time as compared to other models. Hence, we can say that Random Forest should be used and implemented for the applications for forecasting gaming product sales.

The future scope of study can be extended to fetch better accuracy with recent neural network architectures like spike neural networks and learning mechanisms. Support vector machines are also promising alternatives as baseline methods. The deep learning with neural network is also computationally sensitive. The nature of learning algorithm and architecture can also be parallelized and implemented on CUDA to improve further computing time.

VII. REFERENCES

[1] Chetan Sharma Jain, Aarush Sakhuji, Yogesh Sanjay Golecha [2017], Analysing time series on Rossmann sales using Deep neural networks and Support Vector Machine.

[2]] Deng L., Hinton G., Kingsbury B. [2013], New types of deep neural network learning for speech recognition and related applications: an Overview, IEEE Int. Conf. on ICASSP, pp. 8599-8603.

[3] Guo C, Berkhahn F [2016], Entity embeddings of categorical variables, arXiv preprint arXiv: 1604.06737.

[4] Jacpbusse G, Goes [2015], Winning model documentation describing my solution for kaggle competition, Rossman stores sales.

[5] Wang X., Zhao Kun [2014], Demand Prediction Based on Neural Network Theory.

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

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

[8] Hamdy A. Taha, [1997], “Operations Research- An Introduction”, Prentice-Hall Inc, USA.

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

[10] Shenoy G. V., U K Srivastava, Subhash C Sharma [1991], Operations Research for management, Wilay Eastern Limited, New Delhi, India.

[11] Liu Yue, Yin Yafeng [2015], Demand forecasting using support vector machine, School of Computer Engineering & Science, Shanghai University.

[12] Kris Johnson Ferreira, Bin Hong Alex Lee, David Simchi-Levi [2015], Analytics for an Online Retailer: Demand Forecasting and Price Optimization.

[13] Taweepol Suesut, Suphan Gulphanich, Phongchai Nilas, Prapas Roengruen KittiTirasesth [2016], Demand forecasting approach: inventory control for warehouse automation.

[14] Jia F, Lei Y, Lin J, Zhou X, Lu N [2016], Deep Neural Networks: A promising tool for fault characteristic mining and diagnosis of rotating machinery with massive data, Mech. Sys. And signal processing, pp. 303-315.