Video Game Sales Prediction

[*]Jiten Sidhpura
Department of Computer Science
Sardar Patel Institute of Technology
Mumbai, India
jiten.sidhpura@spit.ac.in

[*]Rudresh Veerkhare
Department of Computer Science
Sardar Patel Institute of Technology
Mumbai, India
rudresh.veerkhare@spit.ac.in

***Abstract*—The gaming industry is certainly one of the booming industries of the modern age, with market size valued at USD 151.06 billion in 2019 and is expected to grow at a Compound Annual Growth Rate (CAGR) of 12.9% from 2020 to 2027. With the availability of technologies like AR/VR in consumer products like gaming consoles and even smartphones, the gaming sector shows great potential. There are various factors responsible for the success of any video game, but traditionally success is associated with sales of a particular video game. There is a tremendous amount of analytics data available in the gaming industry, and a great deal of value can be derived from it. To leverage this, data science is a powerful tool. So, the goal of this study is to analyse the data to predict the sales i.e the success of a given video game, which in turn will help video game companies to make informed decisions. We have performed a details analytical study of the data available, used various machine learning models like Linear Regression, Random Forest, XGBoost and CatBoost for modeling the data. Also to interpret the model predition we have analyzed the SHAP values. Finally we have achieved the best score root mean squared error of 1.65 by using CatBoost model.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In the technology world, people from various ages ranging from small kids to adults play video games. This has led to the spike in the sales of video games nowadays. Video games are released by major publishers across many popular hardware platforms. It provides the only real experience of interactive entertainment offered by modern technologies. It also provides a rapidly growing form of entertainment and is being used for educational as well as business purposes. In the previous decade, several major video gaming releases have raised the bar of conventional entertainment goods in terms of revenues earned. There are several types of blockbuster video games on sale today such as Grand Theft Auto IV' by RockStar Games and Call of Duty' Series by Activision. These types of video games have produced a series of annual records for revenues over the course of a three-year period. Predictive modelling has long been the goal of many individuals and organizations. This science has many techniques, with simulation and machine learning at its heart. Aside from the potential of simulations, machine learning techniques are known for their ability to uncover hidden data trends. While the choice of algorithms used in each study may differ, they all have one common similarity, they give choices that human track experts made and are able to use the data to create arbitrage opportunities. For example, we used Back-Propagation Neural network with Principal Component Analysis to predict the weekly video games sales. Other than that, transformation of classification trees into a decision-analytic model has also been used in solving the value-maximizing game development policy. The results showed that the compact predictive models created by data mining algorithms can help to make decision-analytic feed-forward control feasible, even for large, complex problems

II. Literature Survey

Alice Yufa et al [1] used machine learning techniques to predict the global sales of video games. Authors of the paper have used Video Game Sales with Ratings from Kaggle. The database has in total 17 features and 4 different features indicating sales of the game: North American sales, European sales, Japanese sales and the rest of the world. 4 Sales variables were ignored during the training process and they performed step wise regression. Finally their best model achieved R Square Error (R2 = 0.126). This indicates that the model is estimated to explain about 12.7% of the variability in Global Sales price prediction.

Julie Marcoux and Sid-Ahmed Selouan [2] use a new methodology based on connections and subspace decomposition approaches to tackle the problem of sales forecasting. A tool has been created to assist firm management in determining predicted sales figures. The weekly sales of a video game are predicted using neural network trained with a back-propagation technique. An ideal topology is determined and a time-sensitive neural network is developed for this purpose. As inputs, we considered a variety of influencing signs and characteristics. We use Principal Component Analysis to pre-process the data in order to determine the importance of these factors. The proposed system's performance is assessed and compared to baseline reference sales. Accuracy metric is used for evaluation

In this study, Jeffry Babb et al [3] looks at video game sales by platform in the global market from 2006 to 2011. They attempt to evaluate which parts of the video game market have the biggest impact on sales. This question is especially pertinent given the video game industry's maturing, which has seen top game publishers and developers attempt both vertical integration and horizontal expansion in the hopes of

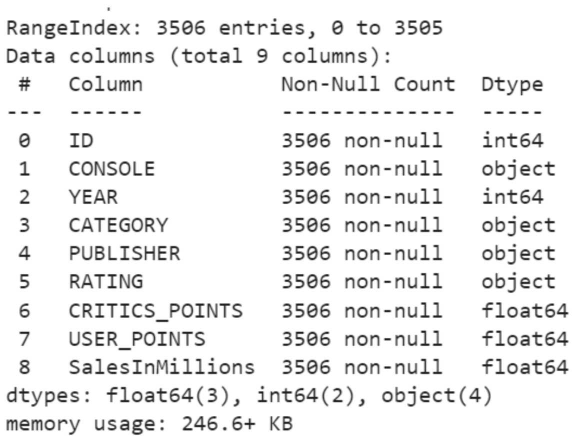
[*]These two authors have contributed equally to this work.

properly establishing the sector. The Kruskal-Wallis test is used to compare eight distinct gaming platforms in this study. The results show that Nintendo's Wii was the best-selling global platform, followed by the Nintendo DS, Xbox 360, Sony PlayStation 3, and the personal computer (PC); Sony PlayStation 2 and Sony PSP are in the fourth tier; and the retired sixth generation Nintendo GameCube is in the lowest sales tier.

TM Geethanjali et al [4] used video games sales data from Kaggle that had been collected from years 1983 and 2016. In total there are 7 features in their dataset. They have used a simple linear regression model to predict the sales of the video games in the North American region. Before feeding the data to the model, they performed standard preprocessing techniques such as removing redundant data, missing and duplicate data. After training their model, they received a p score of (0.00032) which is significantly less than the significance threshold of 0.05 and hence their model is statistically significant.

III. Dataset

A competition was conducted to analyse the success of video games based on the attributes like platform, user ratings, critics ratings, year of publication, rating, etc. In this competition the organizing company has provided a dataset containing 4509 datapoints. This dataset covered a wide variety of video games starting from 1997 to 2019 having diverse distribution. Given the variety of the dataset it will be sufficient to derive insights for our task.



```
RangeIndex: 3506 entries, 0 to 3505
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               3506 non-null   int64
 1   CONSOLE          3506 non-null   object
 2   YEAR             3506 non-null   int64
 3   CATEGORY         3506 non-null   object
 4   PUBLISHER        3506 non-null   object
 5   RATING           3506 non-null   object
 6   CRITICS_POINTS   3506 non-null   float64
 7   USER_POINTS      3506 non-null   float64
 8   SalesInMillions  3506 non-null   float64
dtypes: float64(3), int64(2), object(4)
memory usage: 246.6+ KB
```

Fig. 1. Dataset Info

IV. Methodology

We will be using the Python Programming Language and Scikit Learn module extensively for our study. We will be using the traditional Linear Regression model and with that tree based models such as Random Forest, XGBoost, CatBoost. We will use boxcox transformation to reduce skewness of features. Depending on the number of unique values of categorical variables we will either use LabelEncoder or OneHotEncoder. The detailed description of Machine Learning Models is given below:

A. Linear Regression

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

B. Random Forest

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of over-fitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method,[2] which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

C. XGBoost

XGBoost [5] initially started as a research project by Tianqi Chen as part of the Distributed (Deep) Machine Learning Community (DMLC) group. Initially, it began as a terminal application which could be configured using a libsvm configuration file. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.

D. CatBoost

Catboost [6] is a gradient-boosting tree-based model with out of box support for categorical features. Because of this, there is no need to perform any preprocessing on categorical features. The algorithm converts the categorical features into numerical values with the help of target statistics. It handles categorical data by computing target statistics of the features from it's target feature.

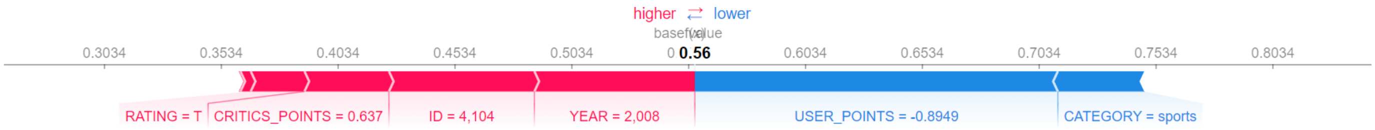


Fig. 2. Shap Text Plot for a Single Data point

Designed by a Russian company Yandex and has applications in the fields of self-driving cars, weather prediction, and more. The CatBoost algorithm has performed better than other state-of-the-art gradient boosting algorithms such as XGBoost, LightGBM on Epsilon, Amazon, and other standard machine learning datasets.

V. Model Prediction Analysis

Today machine learning plays a very critical role in many indudstries. In any business, the reason for prediction by a model is equally important as the model predicting accurately. The Sales Department of the companies can take appropriate policies or decisions if they know how features of the game affect it sales. For this, we have computed SHAP (Shapley Additive Expla- nations) values of the CatBoost model. SHAP values are used to explain why a model has predicted for the given sample of features in a certain way. Popular profession networking website LinkedIn also uses this SHAP Analysis to improve it's machine learning algorithms performance

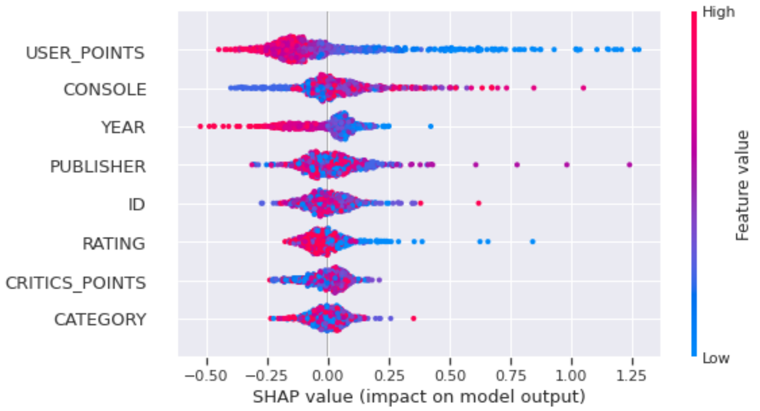


Fig. 3. Summary plot

From the figure 4, we can conclude that user points is the most important feature and category is the least important feature while making prediction. Because shap summary plot by default sorts the features in descending order of importances. It can be seen that as USER_POINTS increase it decreases the model's prediction score. In case of CONSOLE feature we can see that consoles (x360 = 14) and (xone = 15) have comparatively higher sales than console (3ds = 0) and (ds = 1). In case of YEAR feature, it can be easily seen that as year increases sales in general decreases. For Publishers such as (Zoo Digital Publishing = 93) and (Zoo Games = 94) their sales are very high as shown by few red dots at the extreme right. Features such as RATING, CRITIC_POINTS, and CATEGORY have majority of their plots placed at center only and thus they do not have a specific relationship with target feature. Density of red (high) and blue (low) points is approximately the same.

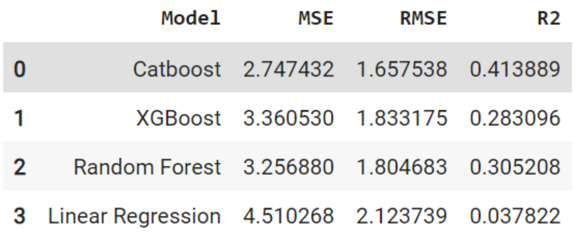
Sometimes we may be interested to know how our model is performing for individual cases. We can also analyze the model prediction using SHAP values. In figure 2, we have considered one example from the test dataset. From the graph it can be easily concluded that features such as YEAR, ID, CRITIC_POINTS, RATING caused model output to increase whereas features such as USER_POINTS, CATEGORY have made the model output to decrease.

VI. Experimental Setup

We performed our proposed methodology extensively on Google Colaboratory. Our environment used Python (v3.7.12), scikit-learn (v1.0.2), CatBoost (v1.0.4), XGBoost (v0.90). While training, the model that gave the best performance on the evaluation dataset is selected in all the experiments.

VII. Results

The results of the study are compaired below. CatBoost has given the best result amongst all.



	Model	MSE	RMSE	R2
0	Catboost	2.747432	1.657538	0.413889
1	XGBoost	3.360530	1.833175	0.283096
2	Random Forest	3.256880	1.804683	0.305208
3	Linear Regression	4.510268	2.123739	0.037822

Fig. 4. Final Results

VIII. Conclusion

Sales Prediction of a product is very crucial for game development companies because it can help them make better decisons based on features gathered from the end users and the market.

Our machine learning solution can help such companies to predict their sales and help them compare their projected values with the model's predictions. By performing Exploratory Data Analysis (EDA), we transformed few features to a new format so their skewness is reduced. We used Catboost model, a special algorithm that takes care of the categorical data processing by itself and it tends to give better performances over other models that apply labelencoding and one hot encoding. Among Catboost, XGBoost, Random Forest and Linear Regression the catboost model was the best performer.

These companies generally need models that can also justify their predictions so that they can no where to improve. Hence, model should not only be robust with great accuracy but we should be able to justify it's prediction. Hence to achieve this we have used SHAP analysis to perform our model's predicition analysis.

Our catboost model gave a MSE score of 1.2222. By gathering more data and More features such as text review of a user we can increase the performance of the models. If data is significantly large then neural networks can also be used as they generally work best with enormous data.

References

[1] Amar Aziz , Shuhaida Ismail, Muhammad Fakri Othman, Aida Mustapha," Empirical Analysis on Sales of Video Game: Data mining Approach "Amar Aziz et al 2018 J. Phys.: Conf. Ser. 1049 012086

[2] Marcoux, Julie, and Sid-Ahmed Selouani. "A hybrid subspace-connectionist data mining approach for sales forecasting in the video game industry." 2009 WRI World Congress on Computer Science and Information Engineering. Vol. 5. IEEE, 2009.

[3] Babb, Jeffry, Neil Terry, and Kareem Dana. "The Impact Of Platform On Global Video Game Sales." International Business Economics Research Journal (IBER) 12.10 (2013): 1273-1288.

[4] TM Geethanjali, Ranjan D, Swaraj HY, Thejaskumar MV, Chandana HP, "Video Games Sales Analysis: A Data Science Approach." International Journal of Creative Research Thoughts (IJCRT). 2020 IJCRT — Volume 8, Issue 5 May 2020.

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[6] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural informa- tion processing systems, 31.

[7] Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.