

Data Visualization on video games global sales analysis & Predictive analysis on Real Estate pricing in Boston

Tabdil Sai Akhil ^[1], Navjot Kaur ^[2], B Surendranath Reddy ^[3],
B Vinay ^[4], K Nanda Kishore ^[5]

Computer Science and Engineering Department
Lovely Professional University - Punjab

ABSTRACT

This project is about giving user an interactive dashboard on various analysis (using various graphs and parameters) on given dataset (user input) and giving predictive analysis depending on user requirements and implementing on a website. These users may be of various organizations belonging to such as supermarket, mobile store, videogames store etc. In this project we are taking a Videogames manager and real estate's manager as users (sample) who provides data sets videogame sales data set (for data Visualization) and real estate housing price data (for house price prediction) which are sample datasets taken as input from kaggle.com These are implemented on our website with various parameters such as login page, Authentication etc. dashboard is created using tableau online tool.

Keywords: - Data visualization, Predictive analysis, Dashboard

I. INTRODUCTION

1.1 Videogames data set (sales visualization):

Video games are electronic gadgets that involve user interaction or input device in the form of keyboard, joystick, motion sensor device etc. Through the past decade many videogames have been published through various platforms such as ps2, ps3, WII etc. [1]. These games published would belong to different age group (ex: E, E10+, M, RP etc.) published by different publication all over the world.

Through this dataset we can get to see an interactive dashboard which shows sales of videogames, top games played and different comparisons for the *past 40 years* through various charts such as pie chart, bar graph, heat map etc.

The data used in the visualizations was gathered from kaggle.com which is generated through vgchartz.com

1.2. Real Estate dataset (Predictive Analysis)

A prophecy is a concept about a future event [2]. One can only predict about any future event only based on its past and present stats or data. Study of these past data and predicting its future data can be said as predictive analysis. It is use of statistical data, algorithms and machine learning teachings or methods to identify the future outcomes. Its goal is to go beyond what happened in past to provide best outcomes what comes in future.

II. VIDEOGAMES DATA SET (SALES VISUALIZATION)

2.1 Data Visualization:

Data visualization is the practice of translating data into a visual context, such as a map or graph or chart, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify trends, patterns, and outliers in large *data_sets*. The term is often used interchangeably with others, including information graphics, information visualization and statistical graphics.

Data visualization is one of the steps of the *data science* process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made. It is also an element of the broader Data Presentation Architecture (DPA), which aims to, locate, identify, manipulate, format and deliver data in the most efficient way possible

A Line charts. This is one of the most basic and common techniques used. Line charts can display how features can change over time.

B Area charts. This visualization method is a variation of a *line chart*; it displays multiple values in a time series or a sequence of data collected at consecutive, equally spaced points in time.

C Scatter plots. This technique displays the relationship between two variables. A *scatter plot* takes the form of an x and y axis with dots to represent data points.

D Heat map displays magnitude of phenomenon as a color in 2d.

2.2 DASHBOARD

A dashboard is a collection of several views, letting you compare a variety of data simultaneously through various filters and parameters.

If you have a set of views that you can review every day, you can create a dashboard that displays all the views at once, rather than go through separate worksheets.

Similar to worksheets, you access dashboards from tabs at the bottom of a workbook.

Data in sheets and dashboards are connected; so that when you modify a sheet, any dashboards containing it change, and vice versa.

Both sheets and dashboards update with the latest data available from the data source (ex. Excel sheet)

How has the Video game representation in digital world evolved through the years?

2.2.1 Dashboard 1:



Fig.1. analysis of global sales

Through the above(fig.1) dashboard we can see the distribution of no of games published each year, [3] and number of games published for different age groups and country wise sales. To make the above dashboard more interactive area chart is made as a filter through which we can sales of any country for a particular year. For example in the year 1999 North american sales are highest with 126.1 million dollars

2.2.2 Dashboard 2:



Fig.2. classification by genre and platform

Similarly through the above(fig.2) dashboard we can see distribution of games by genre using pie chart distribution, number of games published through different platforms and also number of games published by a publisher through heatmap. Similarly as above to make dashboard more interactive genre is used as filter. Using this dashboard some of the conclusions that can be derived are action games are played most, most number of adventure games are produced through DS platform by idea factory publications.

2.2.3 Dashboard 3:

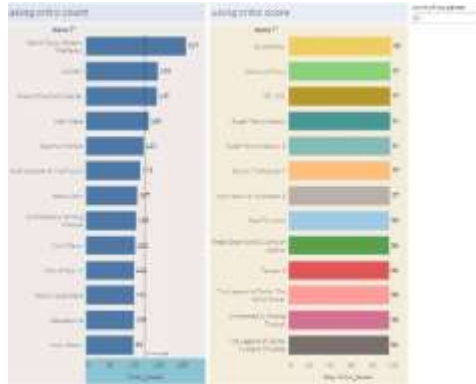


Fig.3. top games using critic count and score

Through this dashboard the critic rate and critic count can be viewed. Through these bar charts top games can be viewed based on critic rate and also critic count. To be more dynamic for number of top games to be displayed parameters are used. Through this dashboard some of facts that can be inferred are call of duty is top game based on critic count and soul caliber is top game based on critic score.

III. REAL ESTATE DATASET (PREDICTIVE ANALYSIS)

This dataset consists of Boston [4]housing data obtained by StatLib library which is maintained at Carnegie Mellon University. This dataset contains data of value of owner occupied homes based on

various features such as age of house, per capita crime rate of town, nitric oxides concentrations etc. This dataset has total 511 instances(houses) and 14 features.

This data consists in the form of only numerical variables. we split our dataset in to train test dataset in the ratio 70:30.

3.1 Exploratory data analysis(EDA):

EDA is mainly an approach to analyzing data to summarize its main characteristics in the form of graphs such as boxplot, scatterplot etc. For this dataset we have taken boxplot of various features to determine outliers of various features.

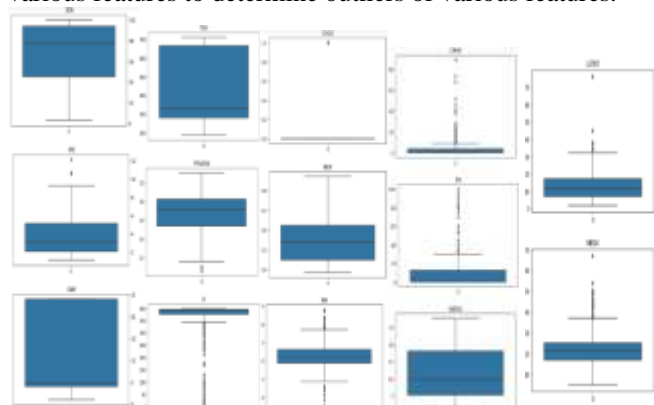


Fig.4. Boxplot to display outliers

3.1.1 Data cleanup:

First we determined there are no null values in our data and all data is in same format, we left original data as it is.

3.1.2 Multi collinearity:

collinearity in regression refers to the correlation of various features with each other [5]. These correlation coefficients are obtained using correlation matrix. Here multi collinearity is a concept of having multiple variables not only correlated to response variable but also highly correlated to each other. This results to having redundant variables which thereby may affect accuracy of final prediction.

In this data set RAD and tax columns are highly correlated with each other so to avoid multi collinearity we dropped RAD column from table.

3.2: Models

The models we used in this are linear regression, ridge regression, elastic net regression.

3.2.1 Linear regression:

Linear regression is an approach to model relationship between variables where if single independent feature relationship is simple linear regression and if multiple explanatory variables approach is multiple linear regression.

3.2.2 Ridge regression:

Ridge regression is special case of linear regression in which all parameters are regularized (L2) equally.

3.2.3 Elastic regression:

Elastic net regression is a regularized approach in which it combines L1 and L2 penalties of Lasso and Ridge linearly.

3.2.4 Random Forest Regression:

This model is basically an ensemble modelling technique where it uses multiple decision trees and output the class that is mean/average ie regression.

3.3 Performance Optimization:

3.3.1 Cross validation:

In this project we used the k-fold cross validation by taking K number of samples to train data. Here we took value of K=5.

3.3.2 Hyper parameter optimization:

Hyper parameter is user defined parameter which is used to control the learning process of a model.

Here we used different hyper parameters such as cv (k-fold cross validation), alpha(regularization)

And used grid search cv method to determine best hyper parameters.

3.3.3 Dimensionality Reduction

For dimensionality reduction we used Principle Component analysis that explain best variance ratio of top n estimators where n is variable and used this hyper parameter too for building the

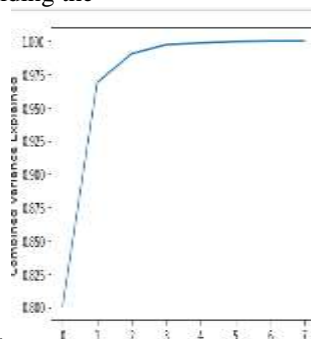


Fig.5. Principle component Analysis

model.

3.4. Modelling and Results:

By using this data set we first started with building linear regression model and obtained an accuracy of 50.03% then regularized the regression model (I e) built ridge regression model and determined best hyper parameters using gridsearchCV method (I e) alpha=1 and obtained r2 value 49.9%[6]. Then we proceeded with L1 and L2 regularization that are regularized linearly (I e) elastic net regression model and determined best parameters as alpha=0.01 and obtained r2 score 50.1% which are almost equal. By further modelling using other modelling techniques (I e) random forest regression best hyper parameters are determined as n_estimators=200 and max important features as 5 thereby we got best r2 score in it by 76.07%.

Then we proceeded with Principle component analysis for dimensionality reduction and obtained the hyper parameter best value as 5 through which target variable can be explained by using there 5 best estimators.

IV. CONCLUSION

4.1 Video game sales data set (Data visualization):

Through the analysis of visualizations, we can conclude that production of games in the past 40 years has arisen to peak in year 2008 but there is downfall since then due to increase in the production of smart phones through which various games can be played. Northern America has highest sales volume for the past 40 years with 4401 million dollars. Most games are

Literature Survey:

1. *Olgun Kitapci* researched the Use of Artificial Neural Networks (ANN) in **Forecasting Housing Prices in Ankara**, Turkey. In this study, an AI-based decision support method for estimation is proposed Home prices. This method aims to assist decision-makers in analyzing the market's structure. Compare recent prices with your findings. In the form, 15 different entries are the price of the home to your site, they are used for input. The output of the system is the future price of the homes. [7]

2. Predicting housing prices using advanced regression techniques

Bharathi researched using techniques Data collection, Data processing, Data Analysis, Feature Engineering, Modelling. In this work, LASSO regression technique was implemented for house price forecasting. The procedure for analyzing the data set and finding the correlation between the parameters was mentioned. [8]

3. Real Estate Price Prediction

The goal of this project is to develop a classification model that is able to accurately estimate the value of a home in terms of features. Used Classification, Regression, Performance Optimization. House prices have been categorized into price categories and will be need to perform multi-category classification to forecast house prices in these categories. [9]

4. A Hybrid Fuzzy Risk Evaluation Model for Real Estate Investments [10]

Alp Ustundag researched a hybrid fuzzy risk model for real estate investments. In this study, a hybrid fuzzy risk evaluation model is proposed for real estate investments and as a base price is determined based on the properties of a real estate by a FRBS that uses the output of FAHP method as one of the inputs for it.

5. ANALYZING THE HOUSING RATE IN A REAL ESTATE: PREDICTION ANALYSIS

Aditya Narhari Khobragade researched about the powerful ways for house selling price and used Artificial Neural Network to predict house sales price and compare with linear regression. At point when Artificial Neural Network approach has been applied to predict the housing value, the outcome is better than linear regression and random forest in the predicting. ANN shows the capacities of the network to learn the patterns. [11]

designed by keeping in view that they can be played by all age groups.

4.2 Real Estate Dataset (predictive analysis):

Through the regression problem, we can conclude that Random Forest Regression works best to predict housing price with accuracy of 75%. Through PCA problem accuracy is increased to 76%. Which also is best among all algorithms with PCA analysis

6. Predicting Global Video-Game Sales

Alice Yufa researched prediction of global video sales analysis by using the dataset from kaggle that was released in 2016 and used data from the review site, metacritic. By using regression techniques, he carried on to build model which resulting in accuracy of 12.7% by using critic score and critic count as independent variables.

7. Video Games

Gregory Smith used The LGBTQ Video Game Archive. Sexualities-distribution of video games based on sexualities using histogram. Gender presentation and identity- using this representation division of non-binary characters over years is displayed.

8. Mining Tracks of Competitive Video Games

Elsevier B. V researched and got a detailed track based system for competitive video games and studied its potential for a strategic analysis of the team play. He introduced low-level topological clues, allowing for characterizing the space structure of a MOBA. He showed that these clues were relevant for predicting the outcome of the match. [12]

9. Analysis of Strategy Video Game Play in StarCraft

Patrick Trinh researched and present the first ever large scale analysis of actual strategy video game play. Using a corpus of over 2,000 StarCraft: Brood War games from international tournaments, he analyses every player's actions to determine which of their capacities lead to success. Both APM and SVA are useful in StarCraft due to specific components of its design. With a high APM, one can better reduce the stochasticity of unit movement and take advantage of the unlimited opportunity for action StarCraft offers. With a high SVA, a player reduces the amount of uncertainty due to incomplete information by acting on and revealing more areas of the map. [13]

10. Learning, Attentional Control, and Action Review Video Games

C.S. Green based on analysis of games and chronological perspective, deducted a pattern emerging when moved through digital game library from *Pong* to *Sims*. The early games do not contain as many as 17 functional categories as of now. From this he hypnotized that there may also more in future.

11 Improved Probabilistic Inference as a General Learning Mechanism with Action Video Games

Through action videogames *Alexandre Pouget* researched on Visual Motion Direction Discrimination, Auditory Tone Location Discrimination, Video Game Training, Ruling out Alternative Accounts—Critical Duration Study: Experiment. He inferred that Action game experience results in more

efficient use of sensory networks, these improvements may not be limited to visual modality, but appear in the auditory modality as well.

Comparison paper: (from literature survey)

1. Used ANN for forecasting future house prices	1. Used linear regression and random forest regression for predicting house price based on age of home and other parameters.
2. Used lasso regression technique for house price prediction	2. Used random forest regression for house price estimation
3. Used classification model to estimate value of house in terms of features	3. Used regression model house price estimation
4. Used fuzzy for real estate investments and used method for input.	4. Used random forest regression method for price estimation.
5. Aditya Narhari Khobragade researched about the powerful ways for house selling price and used Artificial Neural Network to predict house sales price and compare with linear regression. ANN shows the capacities of the network to learn the patterns.	5. Researched house price estimation in Boston using random forest regression, cross validation, hyper parameter optimization.

6. Researched prediction of global video sales analysis by using the dataset from kaggle that was released in 2016. By using regression techniques, he carried on to build model which resulting in accuracy of 12.7% by using critic score and critic count as independent variables.	6. Displayed visualization through an interactive dashboard of videogame sales globally with various other categories such as genre, publishers, etc. through which some results such as yearly sales of country can be inferred through it.
7. Used The LGBTQ Video Game Archive. Sexualities-distribution of video games based on sexualities using histogram. Gender presentation and identity-using this representation division of non-binary characters over years is displayed.	7. used video games sales data which is generated through vgchartz.com through which video game gales globally each year, top video games through critic score and critic count can be inferred.
8. <i>Elsevier B. V</i> researched and got a detailed track based system for competitive video games and studied its potential for a strategic analysis of the team to play.	8. Displayed visualization through an interactive dashboard of videogame sales globally with various other categories such as genre, publishers, etc. through which some results such as yearly sales of country can be inferred through it.
9. <i>Patrick Trinh</i> researched and present the first ever large scale analysis of actual strategy video game play. He analyzed every player’s actions to determine which of their	9. used video games sales data which is generated through vgchartz.com through which video game gales globally each year, top video games through critic score and critic count can be

capacities lead to success.	inferred.
10. C.S. Green based on analysis of games and chronological perspective, deduced a pattern emerging when moved through digital game library from <i>Pong</i> to <i>Sims</i> . From this he hypnotized that there may also more in future.	10. Displayed visualization through an interactive dashboard of videogame sales globally with various other categories such as genre, publishers, etc. through which some results such as yearly sales of country can be inferred through it.
11. Alexandre Pouget researched on Visual Motion Direction Discrimination, Auditory Tone Location Discrimination, He inferred that Action game experience results in more efficient use of sensory networks, these improvements may not be limited to visual modality, but appear in the auditory modality as well.	11. Displayed visualization through an interactive dashboard of videogame sales globally with various other categories such as genre, publishers, etc. through which some results such as yearly sales of country can be inferred through it.

price prediction. Proceedings of the 2009 International Conference on Machine Learning and Cybernetics. 2. 970 - 974. 10.1109/ICMLC.2009.5212389.

6. Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6), 2928-2934

7. Rochard J. Cebula. "The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District". In: *The Review of Regional Studies* 39.1 (2009), pp. 9–22.

8. Wang, Ting, Yan-Qing Li, and Shu-Fei Zhao. "Application of SVM based on rough set in real estate prices prediction." *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*. IEEE, 2008.

9. Khobragade, Aditya Narhari, N. Maheswari, and M. Sivagami. "Analyzing The Housing Rate In A Real Estate Informative System: A Prediction Analysis." *Int. J. Civil Engine. Technol* 9.5 (2018): 1156-1164.

10. Sarip, Abdul Ghani & Hafez, Muhammad Burhan & Daud, Md. (2016). APPLICATION OF FUZZY REGRESSION MODEL FOR REAL ESTATE PRICE PREDICTION. *Malaysian Journal of Computer Science*. 29. 15-27.

11. Khobragade, Aditya Narhari, N. Maheswari, and M. Sivagami. "Analyzing The Housing Rate In A Real Estate Informative System: A Prediction Analysis." *Int. J. Civil Engine. Technol* 9.5 (2018): 1156-1164.

12. François Rioult, Jean-Philippe Métivier, Boris Helleu, Nicolas Scelles, Christophe Durand, Mining Tracks of Competitive Video Games, AASRI Procedia, Volume 8, 2014,

13. Lewis, Joshua, Patrick Trinh, and David Kirsh. "A corpus analysis of strategy video game play in starcraft: Brood war." Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 33. No. 33. 2011.

REFERENCES

1. Babb, Jeffry & Terry, Neil & Dana, Kareem. (2013). The Impact Of Platform On Global Video Game Sales. *International Business & Economics Research Journal (IBER)*. 12. 1273. 10.19030/iber.v12i10.8136.

2. Xiaolong, Hu & Ming, Zhong. (2010). Applied research on real estate price prediction by the neural network. 2010 2nd Conference on Environmental Science and Information Application Technology, ESIAT 2010. 2. 10.1109/ESIAT.2010.5567321.

3. Sacranie, & John, & Leekley, & Advisor, Faculty & Robert,. (2021). Consumer Perceptions & Video Game Sales: A Meeting of the Minds. John Wesley Powell Student Research Conference.

4. Wu, J., Gyourko, J., & Deng, Y. (2012). Evaluating conditions in major Chinese housing markets. *Regional Science and Urban Economics*, 42(3), 531-543.

5. Li, Da-Ying & Xu, Wei & Zhao, Hong & Chen, Rong-Qiu. (2009). A SVR based forecasting approach for real estate