Data Mining and Predictive Analytics

Project Title: Video Game Pricing Analytics

ORIGINAL WORK STATEMENT

I certify that the actual composition of this proposal is original work.

I. Executive Summary

The primary focus of this project is to investigate the impact of customer sentiment on video game pricing. Our analysis is based on a comprehensive dataset obtained from SteamDB, a prominent video game database, which includes information such as release dates, reviews, and prices. Leveraging sentiment analysis, we predict the pricing premium associated with video games. This study holds substantial significance in the competitive video game market as it aims to offer data-driven recommendations for the optimization of pricing, discounting, and game development strategies.

Key findings of our work include the following:
In reviews left for Counter Strike Global Offensive, the most common words used were “game”, “play”, and “get”. Meanwhile, when running an LDA model we find the majority of topics discussed regarding the game include the gameplay itself and the “Russian” aspect and themes of the game. The most common words used in reviews for Persona 5 Royal include “game”, “persona”, and “time”. From the LDA model we find that much of the reviews discuss the wide variety of character offerings within the game and the time it takes to fully complete the game. Call of Duty: Black Ops 3 saw reviews commonly using words such as “game”, “zombies”, and “map”. Using the LDA model, we see much of the reviews discuss the multitude of maps available to play through the online multiplayer as well as the zombies mode.

We can conclude that overall, the random forest model has better predictive performance than linear regression for forecasting the prices of video games based on customer reviews. We arrived at this conclusion by analyzing the Random Forest and Linear Regression RMSE under three different sentiment analysis methods: Afinn, Bing, and Syuzhet. By comparing RMSE of different methods, we also concluded that Syuzhet is the most efficient method for sentiment analysis.

I. Data Description

The review dataset for 3 video games - Call of Duty : Black Ops 3, Persona 5 Royal and Counter Strike: Global Offensive was taken through a web scrape of SteamDB [https://steamdb.info/] which is a large repository for game related data such as release dates, reviews, prices, and more.
In the initial scrape, each individual game has two files - customer reviews (Count: 100 reviews) and price time series data.

To obtain data on the reviews of the selected video games, we performed web scraping using R software. The code we developed extracts comments and dates from Steam store pages of different games. It utilizes the RSelenium package, which allows controlling a web browser programmatically, and the rvest package, which is used for web scraping. Additionally, the lubridate package is included for handling dates.
The code begins by loading the necessary libraries, including RSelenium, rvest, and lubridate. It then defines a vector called game_names, which contains the unique identifiers of the games on the Steam store. The subsequent for loop iterates over each game name, so that we obtain data for every game we want, and in the future if our code is used for other games, the only change that will need to be made is adding the name of the game to the vector “game_names”
Within the loop, a Selenium WebDriver is initialized to control the Firefox browser. Firefox was chosen over Chrome as the browser for web scraping due to encountering frequent errors and security issues with Chrome. By opting for Firefox, the code aims to achieve a more stable and secure scraping process, ensuring smoother data extraction from the Steam store pages. The code navigates to the Steam store page of the current game using the remDr$navigate() function. To ensure all comments are loaded, the page is scrolled to the bottom by sending the "end" key to the web element representing the page. After waiting for a few seconds to allow comments to load, the code finds the review boxes under "most helpful reviews" and "recently posted" sections.

For each review box, the code extracts the comment text and the corresponding date by locating specific elements within the review box. The comments and dates are stored in separate lists. These lists are then combined into a data frame called comments_df. The data frame contains two columns: "Comment" and "Date," representing the extracted information. The code generates a unique filename for each game by replacing the forward slash in the game name with an underscore and prepending "comments_". Finally, the comments are saved to a CSV file using the write.csv() function.
At the end of each iteration, the browser and R session are closed. This process is repeated for each game in the game_names vector. The resulting CSV files will contain the comments and dates extracted from the Steam store pages for every game individually, providing valuable data for further analysis or reporting purposes.

The customer reviews dataset contains the date that the review was posted and the review text, while the price dataset contains the date that the price was changed and the price on that date. In order to clean and prepare the data we first start by sectioning the data in excel. After scraping, our csv file fits each review in one row with the date. We split the data, separating date and review, allowing them to have separate columns. Luckily scraping the price separated price and date, so after the separating we just made sure that every file had similar column names.

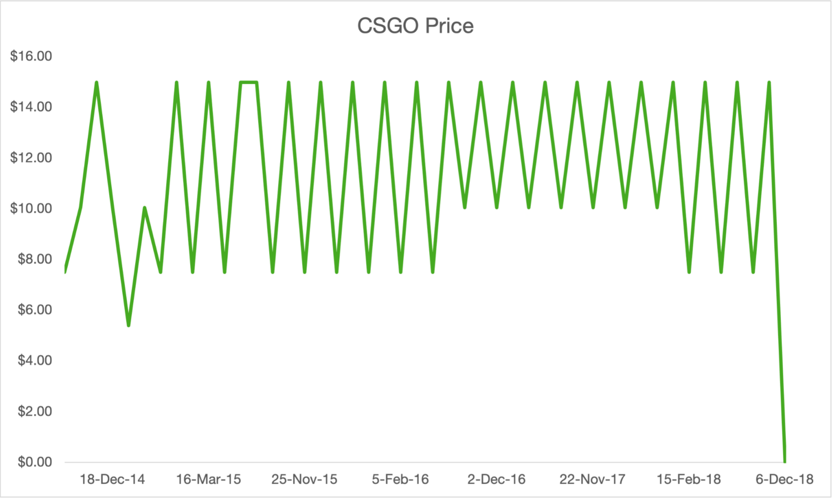
After, we use R to finish the cleaning. Each game has a separate file for prices and review, so each of the prices is converted into a continuous time series by extending the previously available price for each date. Then the price dataset is combined with its respective in R on the common date column using left join. The resulting dataset for each game contains four columns - game name, date, reviews and price. From there, we allow the user to select the game they would like to view.

Within the dataset selected, we begin to clean the reviews within the dataframe. We remove punctuation, links, symbols and select words we find may come up often but wouldn’t be helpful as results. Afterwards, we create a corpus and remove numbers and stopwords, strip white spaces and make words lowercase. We use the corpus for plotting word frequencies and to create an LDA model while we use the cleaned dataset to run our sentiment analysis.

III. Data Exploration

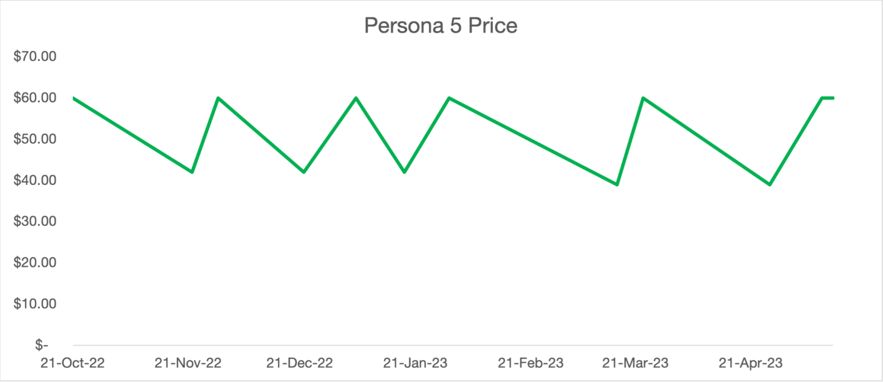
During the data exploration phase, we conducted a comprehensive analysis of customer reviews by employing word frequency counting and topic modeling using Latent Dirichlet Allocation (LDA). By transforming the reviews into individual words and determining their frequencies, we construct informative bar charts highlighting the most common words associated with each game. Also, leveraging LDA, we uncover underlying topic clusters that emerge from customer reviews, enabling us to gain profound insights into customer preferences for each game. This approach facilitates a comprehensive understanding of customer sentiments and preferences, serving as a foundation for strategic decision-making and enhancing the overall customer experience

Counter Strike: Global Offensive



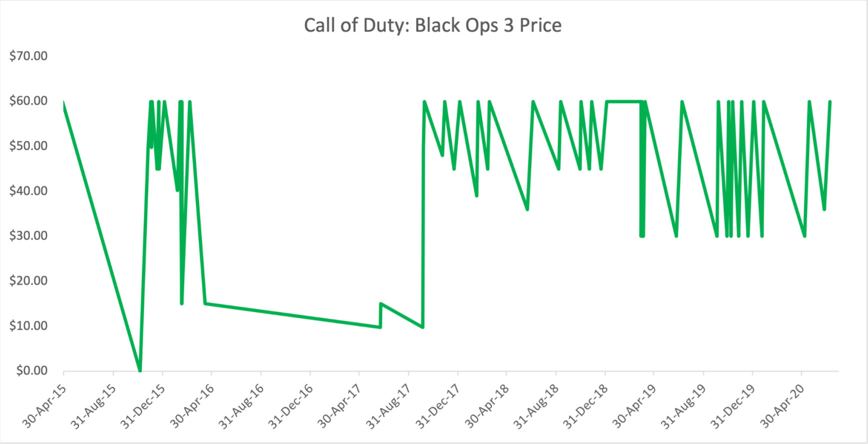
For Counter Strike: Global Offensive, a multiplayer tactical first-person shooter game that transitioned to free-to-play in 2020, customers extensively discuss the game's "Russian" aspect and their gameplay hours. The surge in reviews during 2020 suggests its popularity as a preferred pastime during the COVID-19 lockdown.

Persona 5 Royal -



Persona 5 Royal, an action RPG game, elicits common themes in reviews such as its diverse character offerings, the game's completion time, and the availability of different personas for players to acquire and utilize.

Call of Duty: Black Ops 3 -



Call of Duty: Black Ops 3 generates customer discussions about its relatively high price, while receiving appreciation for its zombie mode, weapons, and maps. While video games often experience price reductions over time, particularly during sale periods, this game exhibited a notable trend. After initial price fluctuations between 2015 and 2017, it predominantly maintained its original price of $60. Through these insights, developers and publishers can gain valuable knowledge about customer sentiments, allowing them to tailor their strategies to align with customer expectations and optimize their game offerings.

V. Research Questions

1. How does customer sentiment towards a video game impact its pricing dynamics in the competitive video game market?

2. Can sentiment analysis of customer reviews provide insights into the optimal pricing and discounting strategies for video games?

3. What role do customer reviews and sentiment play in the long-term viability and commercial success of video games, and how can developers leverage this information to create more successful products?

III. Methodology (1 page)

The methodology of this project involves several steps to analyze customer reviews and predict the prices of video games using sentiment analysis, linear regression, and random forest techniques. The following is a detailed explanation of each step:

1. Web Scraping: The project begins with the web scraping of customer reviews from video game websites, specifically SteamDB. This data collection process allows us to gather a large volume of customer reviews for analysis.

2. User Input and Dataset Selection: The code prompts the user to select a game for analysis from three options: Counter Strike: Global Offensive, Persona 5 Royal, and Call of Duty: Black Ops 3. The chosen dataset will be used for further analysis and modeling.

3. Review Corpus Creation: Once the dataset is selected, the algorithm creates a "review corpus." A review corpus is a collection of text data that includes all the customer reviews related to the chosen game. This review corpus serves as the basis for data exploration and analysis.

4. Data Exploration: The algorithm explores the most common terms used in customer reviews within the review corpus. This step involves analyzing the frequency of words to identify the most commonly used terms and gain insights into customer preferences and sentiments.

5. Topic Modeling using Latent Dirichlet Allocation (LDA): The methodology employs Latent Dirichlet Allocation (LDA), a topic modeling technique, to cluster the commonly used words in customer reviews into different topics. LDA helps identify common themes or topics present in the reviews, enabling a deeper understanding of customer perceptions and preferences for each game.

6. Sentiment Analysis: Sentiment analysis is performed on the cleaned dataset using three methods: afinn, bing, and syuzhet. Sentiment analysis aims to determine the sentiment or emotional polarity of each review, whether it is positive, negative, or neutral. The afinn, bing, and syuzhet methods provide sentiment scores that quantify the sentiment expressed in the reviews.

7. Data Splitting: The dataset is split into training and test sets. This division allows for the evaluation of predictive models on unseen data, ensuring unbiased performance assessment.

8. Predictive Modeling: Two predictive modeling techniques, namely linear regression and random forest, are utilized to predict the prices of video games. The sentiment scores obtained from the sentiment analysis, along with the date of the review, are utilized as predictor variables. The price of the video game on the review date serves as the independent variable.

a. Linear Regression: Linear regression is a statistical modeling technique that establishes a linear relationship between the predictor variables (sentiment scores and date) and the dependent variable (price). It allows us to predict the price of video games based on the given predictors.

b. Random Forest: Random forest is a machine learning algorithm that builds an ensemble of decision trees. It utilizes multiple decision trees to make predictions and provides a robust and accurate prediction model.

The combination of sentiment analysis, linear regression, and random forest allows us to gain insights into customer sentiments, predict video game prices, and evaluate the influence of sentiment scores on pricing. This methodology facilitates informed decision-making in the video game industry and helps optimize pricing strategies and marketing efforts.

IV. Results and Finding (varies considerably in length depending on study)

The analysis below gives an overview of the RMSE values and their trends across different sentiment analysis methods and prediction models for the chosen video games.

RMSE for each method -

Sentiment Method	Afinn		Bing		Syuzhet	
Prediction Model	Random Forest	Linear Regression	Random Forest	Linear Regression	Random Forest	Linear Regression
Counter Strike: Global Offensive	0.1950569	17.10915	0.3765424	16.85598	0.2074857	17.11717
Persona 5 Royal	95.29412	95.29412	95.29412	95.29412	95.29412	95.29412
Call of Duty: Black Ops 3	13.374	131.7615	1.957361	131.3781	3.338852	128.766

The sentiment analysis methods used are Afinn, Bing, and Syuzhet. These methods analyze the sentiment or emotion expressed in a given text.

The prediction models employed are Random Forest and Linear Regression. These models are used to predict or estimate a numeric value based on input variables.

The table is structured in rows representing our selected video games and columns representing the combination of sentiment analysis method and prediction model.

The RMSE values represent the average difference between the predicted sentiment score (or value) and the actual sentiment score (or value). Lower RMSE values indicate better predictive accuracy.

From analyzing each entry in the table, we can observe that:

- The Afinn sentiment method generally performs better with the Random Forest prediction model compared to Linear Regression.
- The Bing sentiment method has lower RMSE values with Random Forest compared to Linear Regression.
- The Syuzhet sentiment method produces similar RMSE values regardless of the prediction model used.
- Persona 5 Royal has consistently high RMSE values across all combinations, indicating poor predictive performance. The reason for this is that the data available for Persona 5 Royal is limited in years.
- Call of Duty: Black Ops 3 shows varying RMSE values across different combinations, indicating mixed predictive performance.

We can conclude that overall, the random forest model performs better than linear regression.

V. **Conclusion**

In summary, our study employed sentiment analysis methods, including Afinn, Bing, and Syuzhet, to analyze the sentiment or emotion expressed in the customer reviews. We also utilized prediction models, namely Random Forest and Linear Regression, to estimate sentiment scores based on input variables. The table presented in our analysis structured the results by video game and the combination of sentiment analysis method and prediction model.

Upon analyzing the entries in the table, several observations can be made. Firstly, the Afinn sentiment method generally demonstrated better performance when paired with the Random Forest prediction model compared to Linear Regression. Similarly, the Bing sentiment method yielded lower RMSE values with Random Forest than with Linear Regression. On the other hand, the Syuzhet sentiment method produced similar RMSE values regardless of the prediction model used.

Regarding specific games, Persona 5 Royal consistently exhibited high RMSE values across all combinations, indicating poor predictive performance. This can be attributed to the limited availability of data for Persona 5 Royal in terms of years. On the other hand, Call of Duty: Black Ops 3 displayed varying RMSE values across different combinations, indicating mixed predictive performance.

In conclusion, our findings suggest that the Random Forest model generally outperformed the Linear Regression model in terms of predictive accuracy. This information can guide developers and publishers in choosing the most effective sentiment analysis method and prediction model for optimizing their pricing, discounting, and game development strategies.

VI. Appendix (Any additional information to be submitted):

Acknowledgements

We would like to express our sincere gratitude for the completion of the Data Mining and Predictive Analytics final project at the University of Maryland, College Park, under the guidance and supervision of Professor Kislaya Prasad.

We extend our heartfelt appreciation to Professor Prasad for his invaluable expertise, unwavering support, and commitment throughout this project. His extensive knowledge and guidance played a pivotal role in shaping our understanding of data mining techniques and predictive analytics methodologies.

Furthermore, we would like to express our gratitude to our fellow classmates who engaged in meaningful discussions and shared their insights, contributing to a collaborative and enriching learning experience.

Finally, we would like to acknowledge the Smith School for providing us with the opportunity to pursue this course and gain valuable knowledge and skills in the field of data mining and predictive analytics.

Overall, this project has been an enlightening and fulfilling experience, and we are grateful for the support and guidance throughout this journey.

References:

https://towardsdatascience.com/nlp-with-lda-latent-dirichlet-allocation-and-text-clustering-to-improve-classification-97688c23d98

https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

https://www.tidytextmining.com/sentiment.html

https://afit-r.github.io/sentiment_analysis