Capstone Proposal

Group Members: Daniel Orrego (PM), Diego Andre Colon, Gregory Torrez, Jacob T Dorfman, Namig Alakbarzade

Project Title:

Data-Driven Insights for E-commerce: Analyzing and Optimizing Sales Prediction

Project Pitch:

Our research project aims to create a data-driven framework to help understand what determines E-Commerce sales and customer retention and what variables impact them most significantly. We will focus on analyzing several factors of E-Commerce, such as seasonal trends, historical sales, and user engagement data. We will focus on individual E-Commerce brands' overall sales, not singular products. With this information, we will create a linear regression model using R to predict sales and compare our results with other models to compare performance. With our results, we can then recommend strategies for optimal sales based on the predictions of our model and show the most impactful variables for customers. We will present our findings through an R Shiny interface that will let E-Commerce brands better understand what would be the most effective for certain brands to use to increase sales through different graphs. The goal is for our project to help E-commerce companies better understand their customer base, ultimately improving customer sales based on our findings.

Names and general skills:

- Daniel Orrego: Data Analysis & Visualization using Python, R, Excel, Tableau, and Access
- Diego Andre Colon: Data collection with Python (Pandas), Data Cleaning, Data visualization with R
- Gregory Torrez: Python, R, Data mining, and database management to organize and compile data
- Jacob T Dorfman: Statistical Analysis, Data collection, and preprocessing (Python)
- Namig Alakbarzade: Data collection with Python (Pandas), Data Cleaning, Data Analysis with R

Names and data/assets required:

- Daniel Orrego: Collection of E-commerce trends, Data management, Project visualization
- Diego Andre Colon: Sales data collection, Data cleaning, and Create prediction model.
- Gregory Torrez: Research previously used algorithm models, contribute to data cleaning, prediction model and research analysis
- Jacob T Dorfman: Collect data on purchase history, finalize data sets Determine best visualizations to present
- Namig Alakbarzade: E-commerce Market Trends Research, Data Collection with Python, Data Analysis (R), Contribution to E-commerce Algorithm Development and Research Analysis.
- Data world: https://data.world/datasets/ecommerce This website provides several data sets that provide information from product data from different retailers and E-commerce sales divided by merchandise category.
- Statista: https://www.statista.com/topics/871/online-shopping/#topicOverview Statista is a database that contains information on E-commerce and consumer behavior. The data presented includes trends, patterns, and insights about online shopping and its evolution.
- Kaggle: https://www.kaggle.com/datasets This platform contains several E-commerce related datasets. The data includes retail transactions, shipping information, and seasonal trends.
- BigCommerce: https://www.bigcommerce.com/articles/ecommerce/ecommerce-trends/ This article provides expert insights on the biggest e-commerce trends that are powering online retail forward. It also includes a biweekly audio series where global thought leaders discuss all things e-commerce, from industry news and trends to growth strategies and success stories.

Literature/Market Review:

Yotpo website (Competing product)

https://www.yotpo.com/ecommerce-trends/

Description: Yotpo focuses primarily on marketing information for e-commerce businesses. Their platform provides insights into customer retention strategies that e-commerce businesses can implement to stand out in the market.

Differentiation: While Yotpo offers e-commerce marketing services, our project will analyze and optimize sales predictions using data-driven insights. Our emphasis will be on predicting sales outcomes rather than purely marketing, which differentiates our approach from Yotpo's primary services.

HubSpot article (Strategies)

https://blog.hubspot.com/service/customer-retention-metrics

Description: The article underscores 10 strategies that companies employ to retain their customers, essentially laying a foundation for metrics to determine customer retention.

Differentiation: While the HubSpot article offers strategies to maintain a customer base, our project intends to predict sales based on data analytics. We will provide insights from customer behavior patterns, which is not the focus of the HubSpot article.

Shopify Blog (Information)

https://www.shopify.com/plus/commerce-trends

Description: This blog provides insights into the latest trends in E-commerce, including the rise of social commerce, the importance of customer experience, and the growing use of artificial intelligence and machine learning. The article also discusses the impact of the COVID-19 pandemic on E-commerce and provides tips for businesses looking to adapt to the changing landscape.

Differentiation: Our project will emphasize data-driven sales predictions. While Shopify discusses trends and impacts, we will be using these insights to refine and optimize predictive algorithms for e-commerce businesses.

Forbes Article (Information)

https://www.forbes.com/sites/forbesmarketplace/2023/03/21/the-future-of-e-commerce-trends-to-watch-in-2023/

Description: This article discusses the latest trends and strategies in E-commerce, including the importance of personalization, the rise of mobile commerce, and the growing use of social media for marketing. The article also provides insights into the impact of emerging technologies such as augmented reality and blockchain on the future of E-commerce.

Differentiation: Our project will use these trends to refine our predictive model. While the Forbes article provides a more comprehensive view of the e-commerce landscape, our model will aim to translate these insights into data models that predict sales.

BigCommerce Article (Strategies)

https://www.bigcommerce.com/articles/ecommerce/ecommerce-trends/

Description: This article lists 14 trends in E-commerce on how people can make their business stand out via AI and voice search to help better personalize the shopping experience for their customers.

Differentiation: Our project will use the information about trends to understand better what drives E-commerce and what factors will affect it. BigCommerce can give this insight because its platform offers scalability, flexibility, speed, and agility in sales to E-commerce brands.

Timeline:

ISTA 498 Capstone

Group Members: Diego Andre Colon, Daniel Orrego(PM), Greg Sean Torrez, Jacob T Dorfman, Namig Alakbarzade

Week		6	7	8	9	10	11	12	13	14	15
Data Research /Collection											
	Collect data regarding sales	1	1	1							
	Research previously used models	2	2								
	Collect data on customer demographics	5	5	5							
	Collect data on customer trends	3	3								
	Collect data on purchase history	4	4	4							
Data Cleaning											
	Cleaning and structuring collected data			All	All	5					
	Data management to select which ones to use			3	1						
	Finalize data sets				1 & 4	1 & 4					
Data Modeling											
	Defining clear variables			All	All	5					
	Define best analysis methods to use			All							
	Create prediction model				2 & 4	2 & 4					
	Troubleshooting prediction model					2 & 4	2 & 4				
Visualizaiton											
	Choose best visualizations & check progress				3 & 1			3 & 1			
	Create visualizations R Shiny & troubleshooting				1, 3 & 4	1, 3 & 4	1, 3 & 4	1, 3 & 4	1, 3 & 4	1, 3,4 & 5	
Presentation											
	Create presentation & poster									All	All

Legend	Number	Color
Diego Andre Colon	1	
Greg Sean Torrez	2	
Daniel Orrego	3	
Jacob T Dorfman	4	
Namig Alakbarzade	5	
All or Some	Multiple	
Amount of Time	No number	

Gantt Chart 2.0

Final Deliverable and Specifications

The goal of our product is a sales algorithm that maximizes RShiny's capability to display our data-driven framework aimed at improving the customers' E-commerce platform sale strategies. It will utilize all historical and transactional data to create a data-driven application that outputs the findings through visualizations. This will allow sales and marketing teams to make better decisions by gaining the ability to control their desired variables and analyzing its effect. We will be presenting our findings through a PowerPoint presentation and will include the conclusions of our prediction model as well as three graphs to summarize the results of our model.

Minimum Deliverables

- Our sales prediction model will be executed in RShiny and will allow the user to interactively toggle between the selected variables in the algorithm.
- Data analysis of the information will be available for the users.
- A visual representation of the information will be displayed through created graphics.
- The sales prediction algorithm will be visually enhanced through RShiny.

Extra Deliverables

- This interactive model will include more measurable references to test the accuracy of the given data through interchangeable variables.
- Our users will have the ability to measure the model against other machine learning types to compare data, increasing the range in which our model can be applied.
- The Shiny web application UI will be created to include toggling and control widgets actions for increased control over the data.

Final Presentation Format

The final presentation will be conducted through various media. A presentation will provide the details regarding our project and the methods used to create the final product. After the presentation, we will provide a quick demonstration of how our application performs through RShiny. We will provide a quick demonstration of how our application performs through RShiny. This application will be made available through a GitHub link.

What Analysis Is Being Run?

Our analysis will be based on the E-commerce datasets that we will have obtained through various e-commerce datasets and surveys. After we have visualized the variables within the data set based on the R programming language, we will analyze the correlation between influencing factors. A linear regression model will allow us to predict the possibility of sales and that will be compared to other models to test compare performance. This aims to evaluate the factors that can help improve customer satisfaction by predicting sales.

What Accuracy Is Expected?

The application is expected to have at least a 90% accuracy rating based on similar models that are currently available today. If the data we will be analyzing has too little or too much information that varies, we could expect a lower percentage rating closer to 75%.

What if the Analysis doesn't work?

In the case that the Analysis doesn’t work we will adjust the expectations of the outcome to a lower accuracy percentage.

What if the Data Isn't Available?

If the data is not available for one of our variables, we will look for a similar variable that can replace the previous one without compromising the result of the analysis, however, if several of our variables are unavailable, we will resort to simulating the data required.

Data Ethics Checklist

#	Question	Generally	Data Breach	Example
1	Could a user sell drugs or other illegal items on your platform?	N	N	https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1072&context=sph_pubs
2	Could a user of your platform engage in sex trafficking?	N	N	https://stmuscholars.org/craigslist-and-backpage-sex-trafficking-at-your-fingertips/
3	Could a user sell class notes or cheat on their homework on your platform?	N	N	https://www.edsurge.com/news/2021-02-23-more-students-are-using-chegg-to-cheat-is-the-company-doing-enough-to-stop-it
4	Could a stalker use your project to find someone?	N	Y	https://www.nytimes.com/2018/05/19/technology/phone-apps-stalking.html
5	Could your app be used to spy on or track individuals?	N	Y	https://www.nytimes.com/2019/12/22/us/politics/totok-app-uae.html
6	Could your app/software access the camera or microphone and record things without users being aware?	Y	Y	https://gizmodo.com/these-academics-spent-the-last-year-testing-whether-you-1826961188
7	If someone uses your platform, could they be re-traumatized or have their mental health impacted in some way?	N	N	https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

8	Could your algorithm promote material that would traumatize or upset individuals?	Y	Y	https://www.theguardian.com/technology/2021/oct/16/tiktok-eating-disorder-thinspo-teens
9	Would your users be upset if the data you collect was given to someone else?	Y	Y	https://www.businessinsider.com/stolen-data-of-533-million-facebook-users-leaked-online-2021-4
10	Could a data leak potentially lead to identity theft?	N	Y	https://www.ftc.gov/enforcement/cases-proceedings/refunds/equifax-data-breach-settlement
11	If your site was hacked, would users of that product potentially lose their job, spouse, or family?	N	N	https://www.forbes.com/sites/zakdoffman/2019/08/23/ashley-madison-is-back-with-30-million-cheating-spouses-signed-since-the-hack/?sh=22f1ba5c3878
12	Should there be an age limitation on your product?	N	N	https://www.bbc.com/news/technology-48925623
13	Could someone use your product to find, contact, and potentially commit elder abuse?	N	Y	https://www.nbcnews.com/health/aging/genetic-testing-scam-targets-seniors-rips-medicare-n1037186
14	If the data on your platform was breached, could it be used to blackmail the users?	N	N	https://www.wired.com/story/parler-hack-data-public-posts-images-video/
15	Does the existence of your project imply that a particular racial group, gender, religion or other protected category is inherently bad, gross, or unwanted?	N	N	https://www.distractify.com/p/pinky-gloves-dragged

16	Could your product be used to commit hate crimes against a specific group?	N	N	https://ibmandtheholocaust.com/
17	Does the primary content of your game or algorithm focus on something considered deeply unethical?	N	N	https://www.quora.com/What-is-the-most-unethical-video-game-ever-created
18	Does your game or software contain race, gender, or other stereotypes?	N	N	https://en.wikipedia.org/wiki/List_of_controversial_video_games
19	Could users of your app scam other individuals?	N	N	https://dailyiowan.com/2021/06/21/opinion-kickstarter-scams-are-on-the-rise/
20	Is your particular algorithm biased towards predicting correctly only for one race, gender, or other group?	N	N	https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm
21	Are the users of your project, players of your game, or those being surveyed for your data aware of how their data will be used?	N	N	https://www.computerweekly.com/news/252464048/Many-search-engine-users-unaware-of-personal-data-collection
22	What are the possible misinterpretations of your results? For example - would a white supremacist or misogynist be stoked about your results if they misinterpreted it?	N	N	https://www.nature.com/articles/s41467-020-19723-8

23	Does the use or purchase of your data potentially contribute to a dangerous group or regime?	N	N	https://vertpaleo.org/svp-sends-letter-to-paleontological-community-on-myanmar-amber/
24	Could your virtual reality environment cause injury to the user?	N	N	https://bonejoint.net/blog/eight-things-you-should-know-about-virtual-reality/
25	Are your study participants or game players aware that their data will be collected and used?	Y	Y	https://www.polygon.com/features/2019/5/9/18522937/video-game-privacy-player-data-collection
26	Does your game or app contain addictive design elements without benefit to the user?	N	N	https://searchsoftwarequality.techtarget.com/tip/5-examples-of-ethical-issues-in-software-development
27	Does your survey contain an aspect of compulsion or unusually large incentive, that would command users to take it even if it was to their detriment?	N	N	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4214066/
28	Could your research outcomes harm an individual or entity?	N	N	https://rutgersaaup.org/rutgers-budget-system-is-a-mess-what-will-president-holloway-do-about-it/

Data Ethics Responses

- 4: Data breaches are always going to be an expected issue, and the leak of sensitive data will be handled with the utmost importance. The listed application will provide a variety of safeguards to ensure that all data listed in our storage will be encrypted and stored on trusted servers. These safeguards will include authentication, authorization, encryption, logging, and application security testing.

1. **Authentication**: Users will be required to authenticate themselves through secure login mechanisms, such as multi-factor authentication, before accessing any sensitive data or performing actions that could impact the integrity of data.

2. **Authorization**: We will use role-based access control to restrict access to data and system functionalities based on users' roles and responsibilities. Only authorized users will have access to specific data sets.
3. **Encryption**: This ensures that even if data is intercepted during transmission or in the event of a breach, it remains unreadable without the appropriate decryption keys.

4. **Logging**: We will implement robust logging mechanisms to track and monitor all system activities. This includes logging user interactions, system events, and access attempts. If there is an event of a breach, these logs will be invaluable for forensic analysis.

5. **Application Security Testing**: Our system will undergo regular security assessments and penetration testing to identify weaknesses. Vulnerability scanning, code reviews, and penetration testing by certified professionals will be conducted periodically to mitigate potential threats.

5: Our application would have our customers sensitive information so in the instance there is a data breach, there is a possibility that leaked information could be used to spy or track individuals. However, as previously stated, there will be proper security measures in place to prevent or mitigate potential breaches.

- 6: To provide a better and unique experience for each user, allowing for permission to access the camera or microphone can influence their personalized marketing experience. If users do not fully read or understand the terms and conditions, there is a possibility that they would be unaware that these items can be used.
- 8: Item searches that may include sensitive topics can be utilized to market these items directly to the user. Depending on the topic it can trigger an unintentional negative impact on their mental health.
- 9: Personal data can be subject to distribution to other parties to improve the marketing effectiveness of everyone. Since the information can be sold, it can create scenarios where our customers would be upset if their data were given to someone, they did not choose to release it to. This would prove difficult to prevent.
- 10: The user's home address, payment methods, contact information, etc. would all be present within our stored databases, and because of this, there is a potential this information could be stolen and used for identity theft. Ensuring additional security is added to each account would prove to minimize this.

- 13: As listed above, the information isn't completely secure and will always be subject to potential threats. In this scenario, there could be a chance where an elder would be made vulnerable should their information be leaked. Including additional security measures might prove useful in limiting hacking attempts.
- 25: The purpose of the application is to obtain information from our users to provide a customized shopping experience for them. This information is necessary for the application to function as intended.