Data Analytics Capstone Topic Approval Form

Student Name: Aidan Soares

Student ID: 012042436

Capstone Project Name: Parametric testing and Logistic Regression on Kickstarter Project Dataset

Project Topic: Predictive Model for Kickstarter Project Success

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: Can a predictive model be constructed using logistic regression on the research dataset?

Hypothesis: H_0: A predictive logistic regression model cannot be constructed from the research dataset.
H_1: A predictive logistic regression model can be constructed from the research dataset with a model accuracy above 70%.

Context: The contribution of this study to the field of Data Analytics is to build a predictive model that can determine the likelihood of a Kickstarter project in succeeding so a startup venture can tweak the composition of their product/offering prior to launch, without increased risk of failure (Kaggle, 2024). With this model, new entrepreneurs can assess their project's feasibility in attaining the startup capital needed to first begin a venture based upon chosen category, funding timeframe, capital goal, and country of origin. This study will utilize a Logistic Regression Model to analyze statistical significance in predictive variables to isolate relevant features that best predict an outcome of success. Per IBM documentation, logistic models (often referred to as logit models) are utilized for classification and predictive analysis to demonstrate relationships between independent variables, and the probability of a dependent variable between 0 and 1. Several studies have found success utilizing logistic regression and other classification methods for effective prediction of Kickstarter success on similar datasets, hypothesizing that features such as category, duration of campaign, funding goal, and location all play statistically significant roles in the determination of success as the outcome (Huang, 2021; Marques, 2018; Zhou, 2017). As such, an analysis of the relationship between independent and dependent variables will present ideal parameters for entrepreneurs to focus on during the planning phase.

Data: The dataset collected for this study is a public Kickstarter dataset sourced from Kaggle. It contains 378,661 rows of projects sourced from the Kickstarter platform and are dated up to January of 2018 (Kaggle, 2018). The projects vary in country of origin, duration, and goal, with status denoted in terms of success, failure, cancellation, undefined, and currently live.

The dataset consists of publicly available information that can be found on Kickstarter's website. In addition to the above, variables found within the dataset include main category, category, name, project ID, currency, launch date, deadline, goal in original currency, goal in USD, backer count, and pledge total. However, for the purposes of this study, the predictor variables selected for the initial feature selection process of the analysis are as follows:

https://www.kaggle.com/datasets/kemical/kickstarter-projects

Variable	Data Type
Main Category	Categorical
Campaign Duration	Continuous
Goal in USD	Continuous
Country	Categorical

These variables will be utilized in predicting the logit outcome of the State of a project (that of success or failure), which is a categorical variable. While the initial selection seems small, the other available features such as Title, ID, pledged, and backers do not demonstrate an implicit relationship with the outcome of a project. Furthermore, while logistic regression is easy to implement and is effective on large datasets, it in inclined towards overfitting of data in high dimensional datasets (Rout, 2024), thus it would be more appropriate to maintain as small of a variable pool as possible.

The dataset operates under the CC BY-NC-SA 4.0 License granted by Creative Commons; an internationally active non-profit organization that provides free licenses for creators. Under this license, users of the dataset are free to share and adapt the material (Creative Commons, n.d.). Furthermore, while this dataset is not

provided by the Kickstarter platform directly, all project information found within the dataset is publicly available information sourced from Kickstarter's website. There is no data within the dataset that reveals any personal information of the creators of the projects. On Kickstarter's Terms of Use page, they state that the platform grants a non exclusive license that allows for the reproduction of content that covers "both Kickstarter's own protected content and user-generated content on the Site", for non commercial use (Kickstarter, 2024). As the dataset is not being utilized for commercial purposes, this dataset, and the information contained within is eligible to utilized for this study.

Limitations: The dataset contains total pledge amounts for each project in USD, but success of a project is determined solely through achieving the goal established by the project creator. Whether the pledge amount exceeds or falls short of the established goal bears no relevance in the determination of a successful project, as there is no outcome for a project being "almost a success" or a "wildly popular success". As such, this study can only make assertions of projects that are likely to binarily succeed or fail. Additionally, the dataset contains information on projects predominantly originating within the United States, data on other countries may be limited and the prediction model may not have enough data to efficiently train on for regions outside of the US. The data is also not particularly recent. Information was obtained through synthesizing data on approximately 370,000 projects from Kickstarter's website in January 2018, but at the present time approximately 626,000 projects are noted to exist on the platform (Kickstarter, 2024), current market trends and favorability for specific project categories may not be the same as they were 6 years ago.

Delimitations: The dataset will be delimited by removing any null or erroneous data such as duplicate entries or incorrect country codes. Additionally, project state can vary from the binary Success/Failure status, containing outcomes such as "cancelled", "undefined", and "live". These outcomes bear no relevance to our analysis in a project's success as undefined projects contain erroneous data, live projects have neither succeeded for failed, and cancelled projects may be done so at the decision of the creator due to external influences from outside the dataset (Thomas, 2021). Finally, the launch date and deadline will be re-expressed as a numerical length of days for the project's runtime.

Data Gathering: Data will be downloaded as a publicly available csv file from Kaggle, which shows data on projects sourced from Kickstarter's website in January 2018 (Mouillé, 2018). Any entries with missing or erroneous fields will be removed as they can negatively skew estimates in the resulting predictive model (Mishra, 2024). The dataset appears to reflect all appropriate data that can be found from the official Kickstarter website, even though the dataset was not distributed by the platform itself. It contains both Qualitative and Quantitative data, as well as categorical and continuous data. Python is the chosen programming language that will be used for generating a dataframe to house the dataset, removing and performing imputation on null data if any exists, as well as re-expression of data such as duration of the campaign. Furthermore, project states that are not Successes or Failures will be dropped from the dataframe as well. Following this, categorical variables will be converted to a binary format, with features having more than 3 outcomes re-expressed as dummy variables. From a cursory glance, data appears very stable with no null values, but a few projects were noted to have invalid Country codes which will need to be removed. Data sparsity is < 1%.

Data Analytics Tools and Techniques: Exploratory Data Analysis will be performed for insights into any pre-existing correlations and trends that can be seen between the independent and dependent variables, as shown through univariate and bivariate statistics/graphs within the presentation layer.

A Shapiro-Wilk test and Q-Q plot were used to determine normality of the continuous data.

Kickstarter officially states that projects lasting longer than 60 days generally do not see success, and that they instead suggest running a campaign for less than 30 days to increase the likelihood of meeting the funding goal (Kickstarter, 2024). To test this assumption, a two-sample t-Test will be conducted on the mean and distribution of the project duration between projects that have Succeeded and those that have Failed to identify if the project duration does play a statistically significant role in a project's likelihood of success.

A Logistic Regression Model does not require an assumption of normality (University of Wisconsin-Madison, n.d.) but will be run to further corroborate statistical significance in project duration as well as significance in other factors from the list of independent variables above. Feature selection will be performed through stepwise analysis of VIF values for reducing multicollinearity (Bhandari, 2024), and isolation of features with p-values less than 0.1. The dataset will then be split into a training set containing 70% of the entries, and testing set containing 30% of the entries for accuracy assessment of the model. Following the above steps, a confusion matrix will be constructed to depict and identify the accuracy of the model and to ultimately reject or accept the null hypothesis (Rosen, 2019). Additionally, a formal regression equation will be constructed for interpretation and proposal of suggested parameters for a higher likelihood in project success.

Justification of Tools/Techniques: Python will be utilized for the creation of the logistic regression model as python has several packages available to conduct the analysis needed, and it is a simple programming language to work with. It can handle cleaning and processing of large quantities of information well (BeyondVerse, 2023) and when used in conjunction with the programming environment Jupyter Notebooks, it allows for a structured organization of markdown notes alongside code for a linear organization of analysis process (Day, 2024).

Project Outcomes: The project will generate a predictive model that is capable of determining the success or failure of a Kickstarter project based on the given parameters prior to launching, with an accuracy of 70% or above. Support for the alternative hypothesis has been found in other studies (Huang, 2021) that logistic regression models can be built upon the existing data in predicting Kickstarter campaign success with acceptable accuracy. Additionally, the project will result in a logistic regression equation that will provide insights into the features of a project that are most likely to influence a successful campaign, providing parameters to focus on the during the planning phase of a Kickstarter project.

Projected Project End Date: 07/17/2024

Sources:

Bai, R., Chung, S., Hou, E., Tang, A. (December 24, 2020). *Uncovering What Traits Make a Kickstarter Campaign Successful.* Medium. Retrieved June 29, 2024, from https://ucladatares.medium.com/uncovering-what-traits-make-a-kickstarter-campaign-successful-dd4b62df9eb4

BeyondVerse. (July 29, 2023). *The Role of Python in Big Data and Analytics*. Medium. Retrieved July 1, 2024, from https://medium.com/@beyond_verse/the-role-of-python-in-big-data-and-analytics-2da818c4cbf

Bhandari, Aniruddha. (June 13, 2024). *Multicollinearity | Causes, Effects and Detection Using VIF (Updated 2024).* Analytics Vidhya. Retrieved July 1, 2024, from https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Creative Commons. N.d. *Attribution-NonCommercial-ShareAlike 4.0 International*. Creative Commons. Retrieved June 29, 2024, from https://creativecommons.org/licenses/by-nc-sa/4.0/

Day, Faithe. (June 6, 2024). *Top 5 Uses for Jupyter Notebook*. NobleDesktop. Retrieved July 1, 2024, from https://www.nobledesktop.com/classes-near-me/blog/top-uses-for-jupyter-noteboook

Huang, Crystal. (May 16, 2021). *ML Classification Model to Predict Kickstarter Campaign Success.* Medium. Retrieved June 29, 2024, from https://crystaldatasy.medium.com/ml-classification-model-to-predict-kickstarter-campaign-success-128c8358f0d3

Kickstarter. (2024). *Kickstarter Support; Getting Started*. Kickstarter. Retrieved June 29, 2024, from https://help.kickstarter.com/hc/en-us/articles/115005128434-What-is-the-maximum-project-duration

Kickstarter. (April 15, 2024). *Terms of Use*. Kickstarter. Retrieved June 30, 2024, from https://legal.kickstarter.com/policies/en/?name=terms-of-use

Marques, Joey. (2018). *The Use of Modern Statistical Methods to Predict the Successfulness of a Kickstarter*. Georgia College and State University. Retrieved June 29, 2024, from https://www.gcsu.edu/sites/files/page-assets/node-808/attachments/marques.pdf

Mishra, Tanmoy. (May 2, 2024). *How to Handle Missing Data in Logistic Regression?* Geeksforgeeks. Retrieved June 30, 2024, from https://www.geeksforgeeks.org/how-to-handle-missing-data-in-logistic-regression/

Mouillé, Mickaël. (February 8, 2018). *Kickstarter Projects*. Kaggle. Retrieved June 29, 2024, from https://www.kaggle.com/datasets/kemical/kickstarter-projects/data

Rosen, Tara. (December 7, 2019). *Understanding Classification Evaluation Metrics*. Medium. Retrieved July 1, 2024, from https://medium.com/@t.rosen2101/my-previous-blog-focused-on-classification-and-logistic-regression-and-while-i-mentioned-4b6578a86844

Rout, Amiya. (24 June, 2024). *Advantages and Disadvantages of Logistic Regression.* Geeksforgeeks. Retrieved June 29, 2024, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Social Science Computer Cooperative. (n.d.). *Logistic Regression.* University of Wisconsin-Madison. Retrieved June 30, 2024, from https://sscc.wisc.edu/sscc/pubs/RegDiag-R/logistic-regression.html

Thomas, David. (November 2, 2021). *Why Cancel a Kickstarter.* Absurdist Productions. Retrieved June 29, 2024, from https://www.absurdistproductions.com/why-cancel-a-kickstarter/

Zhou, Peter. (December 5, 2017). *Predicting the Success of Kickstarter Campaigns.* UC Berkeley. Retrieved June 29, 2024, from https://www.stat.berkeley.edu/~aldous/157/Old_Projects/Haochen_Zhou.pdf

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 7/1/2024

Reviewed by: 

Comments: Click here to enter text.