Aidan Soares
012042436

D214 – Data Analytics Graduate Capstone

Executive Summary Report

Research Problem and Hypothesis

The research question posed for this study is: "Can a predictive model be constructed using logistic regression to determine the likelihood of a Kickstarter campaign's success?" utilizing historical data from other projects.

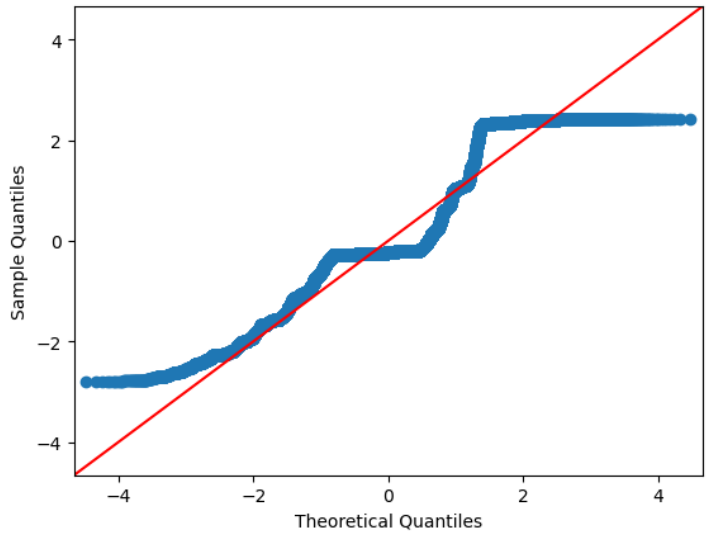
Null Hypothesis, H_0: A predictive Logistic Regression model cannot be constructed from the research dataset with a model accuracy above 70%.

Alternative Hypothesis H_1: A predictive Logistic Regression model can be constructed from the research dataset with a model accuracy above 70%.

For new entrepreneurs, sourcing the capital needed to take a business venture is difficult without a support network. Traditional financing could require multiple applications, a lengthy approval process, and added risk of taking on debt. Kickstarter is a platform that allows for project creators to source funding directly from consumers, reducing traditional barriers to entry in funding new startup projects/businesses. Running a Kickstarter campaign is not guaranteed to achieve the funding goal established, as there are a multitude of factors that can influence a successful outcome. With a predictive model built from historical campaign data, creators can assess the likelihood of their project succeeding based on current parameters such as funding goal, campaign duration, and prospective backer count to adjust the scope of the project during the planning phase. This would undoubtedly improve the efficiency of campaign design, maximizing the likelihood of receiving the startup capital needed, and would save the creator time and resources.

Data Analysis Process

Data was extracted from a public dataset sourced from Kaggle and stored within a dataframe for manipulation. Following the data preparation phase, parametric testing was conducted on the Duration variable to assess normality of distribution using both a Q-Q plot and a Shapiro-Wilk test.



```
ShapiroResult(statistic=0.8532191514968872, pvalue=0.0)
```

Despite a conclusion that the Duration variable did not mathematically display evidence of normal distribution, the Central Limit Theorem asserts that samples of this population should be normally distributed, given the large sample size of the dataset (Turney, 2022). Following this theorem, a t-test was conducted on the means of Duration for each project State to ascertain if a project's established duration has statistically significant variance between Successful and Failed projects.

```
Ttest_indResult(statistic=-62.938657517178946, pvalue=0.0)
```

The t-test resulted in a p-value less than 0.05, concluding that the duration of a project's campaign does have influence on the outcome of the project as the mean duration between both population samples are statistically different.

Following the above analysis, a logistic regression model was created for the dependent variable State using the independent variables Backers, Country_US, Funding Goal (in $USD), Project Duration, and all Project Category Types.

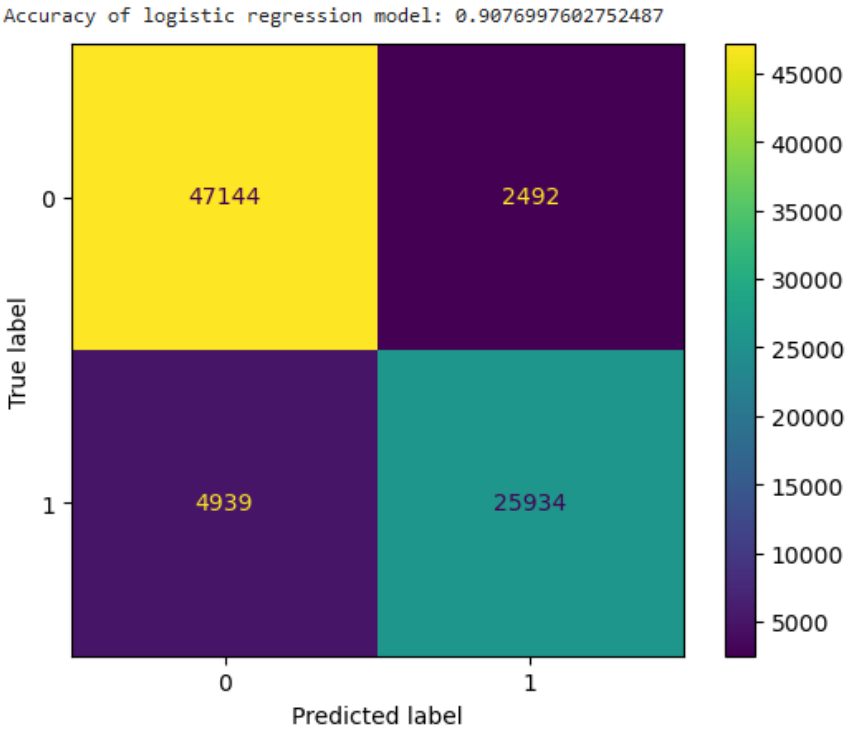
Logit Regression Results

Dep. Variable:	state	No. Observations:	268361
Model:	Logit	Df Residuals:	268342
Method:	MLE	Df Model:	18
Date:	Sat, 06 Jul 2024	Pseudo R-squ.:	0.5855
Time:	00:39:11	Log-Likelihood:	-74052.
converged:	True	LL-Null:	-1.7866e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
backers	4.1464	0.019	223.506	0.000	4.110	4.183
is_us	0.0371	0.016	2.280	0.023	0.005	0.069
usd_goal_real	-3.2500	0.019	-171.812	0.000	-3.287	-3.213
duration	-0.1669	0.007	-25.239	0.000	-0.180	-0.154
main_category_Comics	-0.9338	0.045	-20.911	0.000	-1.021	-0.846
main_category_Crafts	-0.6340	0.042	-15.136	0.000	-0.716	-0.552
main_category_Dance	0.7054	0.056	12.621	0.000	0.596	0.815
main_category_Design	-0.8517	0.036	-23.916	0.000	-0.921	-0.782
main_category_Fashion	-0.5570	0.033	-16.667	0.000	-0.623	-0.492
main_category_Film & Video	0.3208	0.024	13.231	0.000	0.273	0.368
main_category_Food	-0.4255	0.035	-12.198	0.000	-0.494	-0.357
main_category_Games	-1.8147	0.038	-47.729	0.000	-1.889	-1.740
main_category_Journalism	-0.5475	0.063	-8.756	0.000	-0.670	-0.425
main_category_Music	0.1049	0.025	4.226	0.000	0.056	0.154
main_category_Photography	-0.3633	0.039	-9.231	0.000	-0.440	-0.286
main_category_Publishing	-0.5523	0.028	-19.971	0.000	-0.607	-0.498
main_category_Technology	-0.6362	0.039	-16.142	0.000	-0.713	-0.559
main_category_Theater	0.7323	0.037	19.622	0.000	0.659	0.805
const	-0.8556	0.024	-34.950	0.000	-0.904	-0.808

The resulting model established the coefficients of all statistically significant independent variables, denoting the magnitude of influence they apply to the likelihood of project success. This model was then utilized to generate predictions, based on 70% of the dataset being utilized as training data. The predictions were evaluated against the remaining 30% of the dataset in the confusion matrix below.

Aidan Soares
012042436



Outline of Findings

In addition to the results of the t-test, the logistic regression model demonstrates significant efficacy in the prediction of a successful Kickstarter campaign. The calculated accuracy of the model is very strong at 90.7%, demonstrating its proficiency through the True Positive and True Negative outcomes predicted.

The results of the logistic model also provided insight into the magnitude of influence each independent variable introduces to the likelihood of a successful campaign. To summarize, keeping all things constant:

- One unit of increase in backer count **increases** the log odds of campaign success by 4.146
- The campaign originating in the US **increases** the log odds of campaign success by 0.037
- One unit of increase in funding goal **decreases** the log odds of campaign success by 3.25
- One unit of increase in duration **decreases** the log odds of campaign success by 0.167
- A project created within the Comics category **decreases** the log odds of campaign success by 0.934
- A project created within the Crafts category **decreases** the log odds of campaign success by 0.634
- A project created within the Dance category **increases** the log odds of campaign success by 0.705
- A project created within the Design category **decreases** the log odds of campaign success by 0.852
- A project created within the Fashion category **decreases** the log odds of campaign success by 0.557
- A project created within the Film & Video category **increases** the log odds of campaign success by 0.321
- A project created within the Food category **decreases** the log odds of campaign success by 0.426
- A project created within the Games category **decreases** the log odds of campaign success by 1.815
- A project created within the Journalism category **decreases** the log odds of campaign success by 0.548
- A project created within the Music category **increases** the log odds of campaign success by 0.105
- A project created within the Photography category **decreases** the log odds of campaign success by 0.363
- A project created within the Publishing category **decreases** the log odds of campaign success by 0.552
- A project created within the Technology category **decreases** the log odds of campaign success by 0.636
- A project created within the Theater category **increases** the log odds of campaign success by 0.732

Aidan Soares
012042436

Limitations of the Techniques and Tools Used

Throughout my analysis I ran into some performance-based problems due to computational load. While I knew one of the advantages of utilizing Python and Logistic Regression Modelling is that these tools can support analysis of large quantities of data, I encountered performance limitations due to the computational complexity of the data being processed. In early stages of my analysis, I kept getting errors from the logistic regression model stating that it could not iterate upon data due to exponential errors, as computations being performed on values such as $100,000,000 from the Funding Goal column were stalling the model. Due to this, the data needed to be scaled to reduce the strain on my logistic regression model.

Regarding the dataset, the information was collected from Kickstarter's website in 2018, containing information on a total of about 380,000 projects hosted on the site. According to Kickstarter's website, that number is approximately 626,000 today (Kickstarter, 2024), but from the searches I conducted online, there was no recent version of this dataset available for download. As such, the analysis was limited to projects made up to 2018, and the analysis conducted may misalign with favorability of specific project categories in projects made today.

Proposed Actions

Given the predictive accuracy of my logistic regression model, I propose that creators utilize the model to assess the likelihood of their campaign succeeding. Creators should use this as a tool to indicate when adjustments of campaign parameters need to be made in the planning stage, allowing them to maximize chances of success before launching on Kickstarter.

Furthermore, the model suggests two key areas to focus on in maximizing chances of success. Ideally, creators should aim to establish funding goals within reason, requesting only the costs that are necessary in taking the project/business off the ground. I propose that creators also strongly focus on increasing the backer count for their project, as it directly relates to the funding their campaign will receive. This could be done through online marketing, connections through a network of friends and family, or establishing milestone rewards to attract more supporters.

Expected Benefits of the Study

The expected benefit of this study is aimed at assisting entrepreneurs in attaining the funding needed to launch a project or a business venture. As an example, through the proposed actions above a prospective creator could identify if a project within the Game category having a proposed funding goal of $70,000 USD, estimated backers of 200, and campaign duration of 32 days is likely to succeed or not, to an accuracy of 90.7%. If not, these parameters can be tweaked until the model predicts a successful output, giving the creator ideal parameters to strive towards. The best part is that this can be done *before* launching a campaign on Kickstarter, preventing a creator from having to try their luck with the platform, potentially failing, and having to start over again, saving them time and money in achieving the funding they need to fund their venture.

Aidan Soares
012042436

Citations

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

Turney, Shaun. (June 22, 2023). *Central Limit Theorem | Formula, Definition & Examples*. Scribbr. Retrieved July 5, 2024, from https://www.scribbr.com/statistics/central-limit-theorem/