INSY 662: Data Mining and Visualization (Fall 2021)

Individual Project – Modelling the Kickstarter Dataset

Task 1: Classification (Supervised) Model

Final Classification Model: Gradient Boosting Classifier with minimum samples split of 4.

Explanation of Model:

When the dataset was pre-processed, all the variables that make invalid predictors at the time a project is launched were removed. For example, the pledged variable was removed because at the time of launch, the amount that will be pledged during the funding period is unknown. Please refer to the Python file for a step-by-step description of the data pre-processing.

The final model uses project success as the target variable, and 22 predictor variables. The predictor variables describe nine characteristics of a project: funding goal, country, category, description blurb, funding deadline, project creation date, launch date, days between creation and launch, and days between launch and deadline.

Justification of why predictors we included or excluded:

Exclusion of predictors:

In addition to the removal of invalid predictors, other redundant or irrelevant predictors were removed from the dataset. Currency was removed because it is duplicative of (98% correlated with) the country variable since the currency used to fund a project corresponds to the country of the project. The day of the week variables were also removed because the dataset already includes the day, month, and year of each project milestone. Lastly, the USD exchange rate variable was removed because it is only useful to convert two invalid predictors ('pledged' to 'usd_pledged'), which were already removed.

Inclusion of predictors:

The characteristics of variables that were included as predictors in the final model share two useful features: they are numerical (in the original dataset, or after feature engineering), and they have a logical impact on people who are deciding whether to pledge money for a project. A

project is successful if it reaches its funding goal before the deadline. Thus, the success of a project is entirely dependent on its ability to convince people to pledge money. The description, category, and country of a project are going to influence one's decision to pledge for a project. Therefore, the final model includes variables that capture these characteristics. Additionally, the decision to pledge could be influenced by timing. For example, if a project is far from its goal, but the deadline is soon, someone would be less likely to pledge money. Consequently, variables that capture the funding timeline were also included in the model.

Is this model realistic and/or useful in a business context?

This model could help prospective project owners improve their project's likelihood of success. For instance, an owner could use the model to determine the project category, description, funding goal, and period to choose for their project. These choices are realistic for a project owner; however, it could be unrealistic for the owner to change the country of their project. Consequently, this is a limitation of the model since project owners may not be able to vary the country of their project to improve the likelihood of its success.

Task 2: Clustering (Unsupervised) Model

Final Clustering Model: K-Means clustering with k=3.
Selected variables: state_successful, Length_funding, spotlight_True, and launched_at_yr

state_successful: Binary dummy variable that equals 1 if project was successful and 0 if it failed.
Length_funding: Numeric variable representing the length of a project's funding period in days.
spotlight_True: Binary dummy variable that equals 1 if project was featured on the Kickstarter spotlight page and 0 if it was not.
launched_at_yr: The year the project was launched

Explanation of Cluster Characteristics:

Cluster Centers Table

Cluster #	# of projects	state_successful	Length_funding	spotlight_True	launched_at_yr
Cluster 1	2280	-0.688379407	1.692977264	-0.688379407	-0.125417574
Cluster 2	4666	1.42752534	-0.152747437	1.42752534	-0.043960235
Cluster 3	7268	-0.700512959	-0.433030906	-0.700512959	0.067566115

The projects in **Cluster 1** were unsuccessful and thus unlikely to have been featured on the spotlight page. Relative to the projects in the other clusters, they were launched longer ago.

Therefore, the projects in **Cluster 1** tend to be the oldest projects. Finally, the **Cluster 1** projects have a much longer funding period than the projects in the other clusters. The projects in **Cluster 2** were successful and thus likely to have been featured on the spotlight page. The projects in **Cluster 2**, tend to be older than the projects in **Cluster 3**, but not as old as the projects in **Cluster 1**. The funding period of the **Cluster 2** projects is longer than the **Cluster 3** projects, but shorter than the **Cluster 1** projects. The **Cluster 3** projects are very unsuccessful and thus very unlikely to have been featured on the spotlight page. Relative to the projects in the other clusters, they were launched most recently. The **Cluster 3** projects also tend to have longer funding periods than the **Cluster 2** projects, but shorter than the **Cluster 1** projects. These characteristics of each cluster seem abstract, but the characteristics about each cluster can be used to gain business insights about Kickstarter projects.

Is this model realistic and/or useful in a business context?

This clustering model provides insights about what makes a project successful, and how Kickstarter projects have evolved overtime. The model indicates that projects with a very long funding period are likely going to be unsuccessful. This insight is useful because project owners set the deadline for their funding period when they launch a project, so it would be useful for them to know that if they choose a funding period that is very long, compared to the periods of past projects, their project will likely be unsuccessful. The model also indicates that older projects tend to have longer funding periods (compared to newer projects), and that recent projects are more likely to be unsuccessful, despite that fact that the length of the funding period for the more recent projects tends to be lower than for the older projects. This information is useful to Kickstarter's management because it would allow them to see how the success and funding period of projects have evolved over time.