INSY 662 – Data Mining & Visualization
Individual Project Summary Report
Submitted by Robert Prattico (260867390)
Monday, December 12, 2022

Task 1: Classification Model

The first step for developing the classification model was to examine the given dataset. From the first glance, it was clear that some of the variables within the dataset would be irrelevant to predicting the *state* of the project. Variables related to the changing of the project's state were dropped due to the fact that this information would not be available at the launch of a project. It was also assumed that the likelihood of a project reaching its funding goal is directly correlated to the project quality. Thus, the model was to focus on "quality-reflective" characteristics of the project. As such, "non-quality-reflective" characteristics, such as *id, name_len, currency,* and those related to date details were also dropped from the dataset. However, *deadline_month* and *launch_month* were kept in order to capture any seasonality at the start and end of a project's fundraising period.

After dummifying the retained predictors, the next step was to standardize the data. To complete this stage, robust scaling was used. Doing so made it possible to account for any outliers that could be in the dataset and that could have a detrimental impact on the predictive performance of the model. The selection of robust scaling was validated during testing.

A decision needed to be taken on which modeling methods would be tested. Given that the purpose of this task is to build a model with the highest possible accuracy, logistic and ridge regression were not taken into consideration for testing. Furthermore, it could not be assumed that the data was linear, therefore linear SVM was also eliminated from consideration. That left 4 methods to consider: KNN, Random Forests, Neural Networks, and kernel RBF.

Feature selection was a point of contention: it was unclear which feature selection should be used, if any at all. With that said, it was only applicable to one of the remaining model

methods – KNN. Therefore, three versions of KNN were built and tested: one with PCA selection, one with Random Forest selection, and another with none feature selection.

The table below outlines the different versions of models that were built and tested as well as their accuracy score on the grading set.

	Model Type	Feature Selection Method	Scaling Method	Accuracy Score	Accuracy Score on Grading
M1	KNN	None	Robust	0.984	0.7
M2	KNN	Random Forest	Robust	0.666	0.648
M3	KNN	PCA	Robust	0.962	0.75
M4	Random Forest	None necessary	None necessary	1	1
M5	Neural Network	None necessary	Robust Scaling	1	0.7
M6	SVM RBF	None necessary	Robust Scaling	0.985	0.68
M7	KNN	PCA	Standard Scaling	0.974	0.72
M8	KNN	PCA	MinMax Scaling	0.99	0.53

As per the table, a KNN model with robust scaling and PCA feature selection that captured 95% of the data's variability was the highest performing model, generating an accuracy score 0.75 on the grading set. This model is realistic as it bases its predictions on the characteristics one would expect to have the highest influence on the state of a project's funding. Moreover, an accuracy rate of 75% indicates that the model is fairly reliable. In a business context, it means that project owners can input the key features of their projects into the model and predict the outcome of their funding campaign prior to its launch with 75% certainty. This can help them coordinate contingency plans for funding.

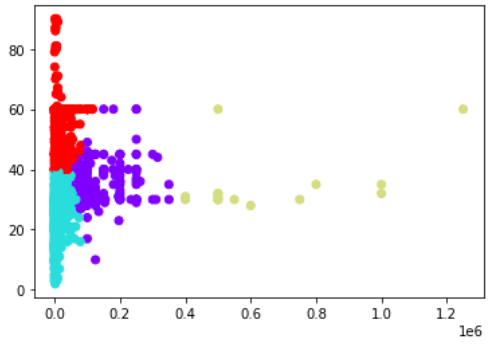
Task 2: Clustering Model

Task 2 consisted of building a clustering model that would provide valuable insights to a project owner. The model should be reflective of a use case that could directly impact a project's fundraising success. The use case chosen is one that answers the following question: how many

days should a project owner place between the fundraising launch and deadline dates in order to maximize the likelihood that the project reaches its fundraising goal? Allocating the right amount of time to tasks is an integral part of every element of project management. For a component as essential to a project as fundraising, not allocating the appropriate amount of time can prove disastrous both to the outcome of this step and to the organization of the rest of the project. As such, knowing how many days to allocate to the fundraising campaign can greatly benefit the successful completion of the project.

For this task, the variables *goal* and *launch_to_deadline_days* were selected as predictors. Because the purpose of the model was to indicate how many days are necessary for a *successful* fundraising campaign, the dataset was stripped to only *state=successful* observations.

The next step was to standardize the data. Because this isn't a predictive model, outliers are not a concern. Thus, applying a StandardScaler was acceptable. Then, a loop was run to determine how many clusters between 2 and 10 would generate the highest silhouette score. This was observed to be 4 clusters. The K-Means model that was subsequently built yielded a commendable silhouette score of 0.57.



From the clustered scatter plot, we can infer the following insights: projects with goals as high as $78k are successfully funded if they are launched 39 days before their deadline. Projects with goals up to $116k need between 39 and 90 days to be funded. Projects with goals between $60k and $350k need between 10 and 60 days to be successfully funded, while those with goals between $400k and $1.25M need 28 to 60 days.

	Cluster 0 Size		Cluster 1 Size		Cluster 2 Size		Cluster 3 Size	
	Goal	LTD	Goal	LTD	Goal	LTD	Goal	LTD
Min	$60,000	10	$1	2	$400,000	28	$1	39
Max	$350,000	60	$78,000	39	$1,250,000	60	$116,000	90