Application of Machine Learning in Kickstarter Dataset

The project leverages various Machine Learning techniques to predict the success and failure of a crowdfunding project using various Machine Learning techniques. Knowing the outcome of the project can enable the organization to proactively plan and follow a plan that would increase the chances of success of the project.

Supervised Machine Learning Techniques

Machine Learning technique	Hyperparameters used	Accuracy
Logistic Regression	Max_iter = 5000	71.8%
K Nearest Neighbors	n_neighbors =45	67.5%
Decision Tree	max_depth=5	71.01%
Random Forest	max_features=50,max_depth=5,n_estimators=29	73.00%
Gradient Boosting	max_depth = 22, min_samples_split = 800, min_samples_leaf =800, n_estimators = 900	73.04%

Tree-based algorithms are performing comparatively better compared to other algorithms. Using cross-validation techniques the hyperparameters are tuned to enhance the accuracy of the model.

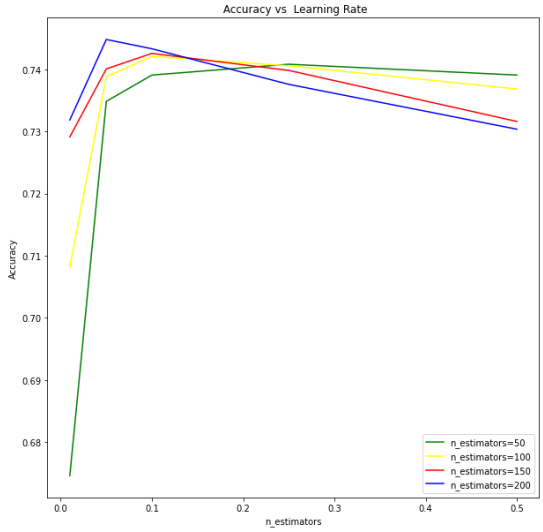

Figure: Accuracy vs Learning Rate for Gradient Boosting Algorithm

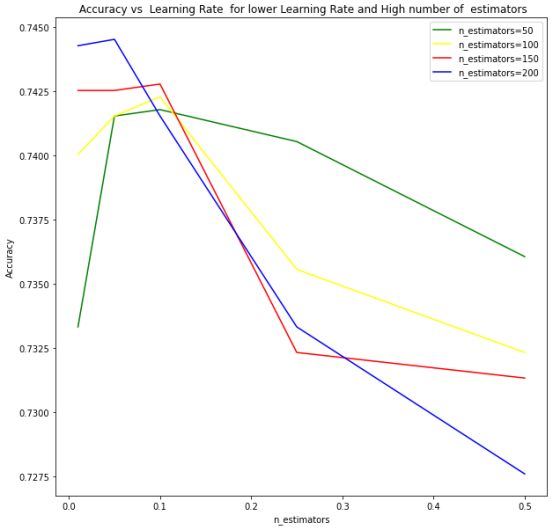

Figure: Accuracy vs Learning Rate for Gradient Boosting Algorithm for low learning rate and high number of estimators

Using tree-based methods would be more appropriate for the use case because a model that can capture a high level of variance would be more useful. Sometimes more diverse projects can be more promising. Hence, capturing this variance would be very useful.

Unsupervised Machine Learning Techniques

K Means clustering is applied to the dataset to obtain the various segments. The variables used for clustering were 'goal','disable_communication', 'static_usd_rate', 'category','name_len', 'name_len_clean', 'blurb_len', 'blurb_len_clean', 'goal_usd', 'create_to_launch_days', 'launch_to_deadline_days', 'create_to_deadline' where all the variables except, 'create_to_deadline' are defined in the data dictionary

'create_to_deadline' = 'create_to_launch_days' + 'launch_to_deadline_days'

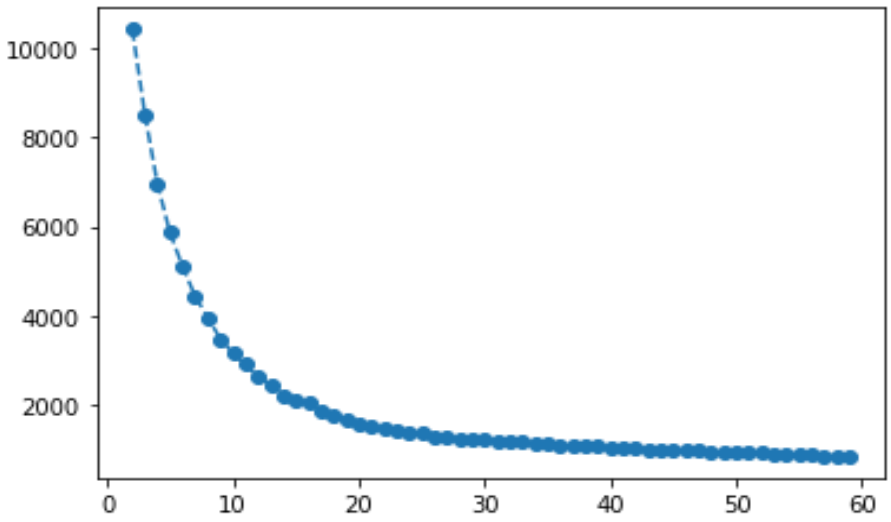

Figure: Using the elbow method to find the optimal K value to capture the model inertia

Approach: The state variable is not added while clustering the data. After clustering is done, we create a table on much each of the clusters consisting of failed projects and successful projects. Based on this we can predict the cluster a particular set of predictor variables will

belong to and we can make a probabilistic inference on the probability of the success of the project.

Cluster_label	Failure %	Success %
0	71.22	28.777
1	71.74	28.261
2	71.12	28.877
3	67.57	32.432
4	72.42	27.582
5	73.19	26.807
6	66.67	33.333
7	71.91	28.093
8	69.16	30.841
9	68.45	31.553
10	77.6	22.396
11	70.35	29.65
12	69.32	30.682
13	69.61	30.395
14	67.84	32.164
15	71.03	28.966
16	71.73	28.266
17	70.21	29.787
18	70.57	29.433
19	68.25	31.753

Figure: Report on the percentage of success and failure in a particular cluster for k =14

For example, If we get a new project, we can assign the variables to a cluster and can get an insight into the probability of the success of the project. Since only the variables that are available at the time of project creation are used in clustering, this method can be effective in the prediction of the success or failure of the project based on the cluster it is assigned. For example, cluster 10 has very less chance of success, and cluster 6 has a high chance of success.

Future Work

Although the machine learning techniques that are used in the project are fairly accurate in predicting the outcome of the project. The predictive analytics task would provide even better results if we leverage some text analytics and natural language processing in the description of the project as it would be able to classify projects on more granular levels and some more interesting insights can be uncovered.