PURPOSE OF PROJECT

The main objective of this project was to determine two main factors for an individual or organization looking to create and publish their idea on Kickstarter.com.

1. To predict the amount in USD that would be pledged to your project using a regression model
2. To classify the factors/variables that would determine whether the project would have a success or failure state

This project consisted of two datasets; one to train the model followed by another dataset to test the model. The dataset required some preprocessing to ensure that the model that would be built will be relevant and will not be affected by factors such as correlation and overfitting. An initial filtering of 'State' variable was done, to ensure that the datasets only consisted of projects that were in either a 'Successful' or 'Failed' status as these were the Kickstarter projects that were focused on.

Furthermore, a large part of the preprocessing phase was dedicated to removing variables that were deemed irrelevant or too correlated both through intuition and methods of Feature Selection that will be described in detail below.

REGRESSION

FEATURE SELECTION

Most variables were removed based on intuition since these variables were deemed irrelevant and did not provide any logical reason to be included in the model, these included. Variables such as; staff pick, backers count, and spotlight were removed as they are not factors that can be determined prior to publishing a project on Kickstarter, as a result would not serve any predictive purpose. Furthermore, variables that were in the created 'category', 'state', and 'state changed' were removed as well as they did not provide much insight and di not make sense to include in the model building as they do not provide any predictive power to the model. Variables such as; name_len_clean, blurb_len_clean were removed as I did not feel that they provided much insight over having name_len, and blurb_len as backers will not care about the length without spaces. Furthermore, I decided to remove all deadline variables from my model as it was not providing a good MSE to my model, and there is also a maximum period of 60 days that you can list on Kickstarter. Therefore, I believe that the 'launched'

variables serves as a better predictor of the USD_pledged variable. Finally, the variable 'goal' was removed as it is in the currency that the project is launched at, as a result will not provide a good prediction; to resolve this problem I created a new column called 'goal_usd' which is a multiplication of 'goal' and 'static_usd_rate', and this was included in the model.

Following this process of eliminating variables through intuition I prepared the dataframe for a feature selection method, and to do this I began by identifying all the categorical variables and creating dummies, followed by standardizing my X variables. Once this process was completed, I used the LASSO method to select my variables as this method also is good at regularizing the data. For this dataset I wanted to narrow the amount of predictors I was going to use to train the model, and after multiple rounds of testing, I found that having an Alpha of 2000 provided me with the best number of predictors that were not only countable, but also provided the best insight in terms of intuition.

Goal_usd	Name_len
Country_US	Category_flight
Category_gadgets	Category_hardware
Category_sound	Category_wearables
Category_web	Launched_at_weekday_Tuesday
Launched_at_weekday_Wednesday	Launched_at_yr_2013
Launched_at_yr_2016	Launched_at_hr_2
Launched_at_hr_4	Launched_at_hr_6
Launched_at_hr_7	Launched_at_hr_8
Launched_at_hr_10	

With a business perspective this made the most sense as well, as it is evident some of the most important predictors that were revealed after the LASSO selection made most sense. When you look at the categories that were considered the most important were all technology related, and over the years Kickstarter has been the foundation for technology related projects that are looking funding, it only makes sense that these will provide the best prediction for the outcome of how much money will be pledged after launching on the website.

MODEL SELECTION

The main basis for selecting the best model was done through multiple rounds of testing, and then selecting the model that gave the lowest MSE. Below is a list of all the models that I used to test the regression and the relevant MSE that was given by using the models.

Model	MSE
Linear Regression	12,223,226,677
KNN Regressor	12,529,695,569
Random Forest Regressor	11,550,342,085

From the above it was clear that the Random Forest Regressor was the best algorithm to use for my model, and as a result this was the model; I decided to use to predict the outcome for future projects posted on Kickstarter. The Random Forest Regressor's hyperparameters were tested using loops to find the ideal model in order to minimize the MSE, and as a result the model is as follows;

Rf = RandomForestRegressor (random_state=0, n_estimators=100, max_features=4, max_depth= 8, min_samples_split=20, min_samples_leaf=7)

Dataset	MSE
Training Model	11,550,342,085
Testing Model	15,828,903,770

CLASSIFICATION

FEATURE SELECTION

The feature selection for classification was constructed differently from the regression feature selection, mainly because the target variable changed from 'usd_pledged' to 'state'. For this I decided to use 'state_Successful' as my target variable and drop 'state_Failure', thereby classifying the all 0's as projects that failed, and 1's as the projects that succeeded. The only other variable I dropped prior to doing a Feature Selection algorithm was usd_pledged as this is not a factor we would know prior to launching the project on Kickstarter.

For classification I decided to go with the Random Forest Classifier as my feature selection algorithm, which resulted in giving me the following predictors to use to train my classification model;

Goal_usd	Static_usd_rate
Blurb_len	Name_len
Create_to_launch_days	Category_Plays
Category_Software	Category_Web

MODEL SELECTION

For classification I used the model that provided the best accuracy score for my model, and as a result I went with the KNN Classifier. I constructed a loop to find the ideal number of n_neighbors and as a result my KNN Classifier model is as follows;

knn_final = KNeighborsClassifier(n_neighbors = 14).fit(X_train, y_train)

Dataset	Accuracy Score
Training Model	0. 7270808909730363
Testing Model	0. 6695652173913044

CONCLUSION

Following extensive testing and modelling it is important to remark on the business implications based on the results that have been collected thus far. It is important to note that my model provided a high MSE, but an acceptable accuracy score. This drives me to the conclusion that Regression might not be the best method to predicting the amount of money that will be pledged for the project. However, it is evident that the feature selection provided a clear insight on some of the most important factors that will determine a successful project on Kickstarter. After concluding this analysis, my advise to someone hoping to launch their product on Kickstarter is to pay attention to the type of project they intend on launching, there is strong evidence that the backers that are active on Kickstarter are those interested in technology related projects, from the US where the exchange rate will play a strong part in determining the amount of money that can be raised. Furthermore, it is interesting to note that prospective backers will pay attention to the length of your project name.

In terms of the classification model, it is evident that the exchange rate, blurb length, and the amount of time spent from creating a project to launch (amount of time spent to prepare the project prior to launching) will pave way to determining whether a project will be successful or not.

In conclusion, I believe that my two models will provide some insight to a prospective entrepreneur who wants to launch their product on Kickstarter, but from my final results I would say that my report will help an entrepreneur better understand whether his product will be a success or not, but it will be difficult to provide an accurate estimate on how much the project will raise. However, if a prospective entrepreneur can structure his project to put more emphasis on the variables I have chosen for my classification model, he/she will be able to have a successful project on Kickstarter which will most likely result in having fulfilled the target pledged amount.