Business Scenario

1. To predict the amount pledged at the end of each project, near the time of launch of the project
2. To predict the state of the project at its deadline, at the time of launch of the project.

Assumptions: It is assumed that the Number of Backers present in the 'Kickstarter' dataset is present only during the deadline of a project, and hence cannot be used as a powerful predictor to predict "Pledged Amount" and "State".

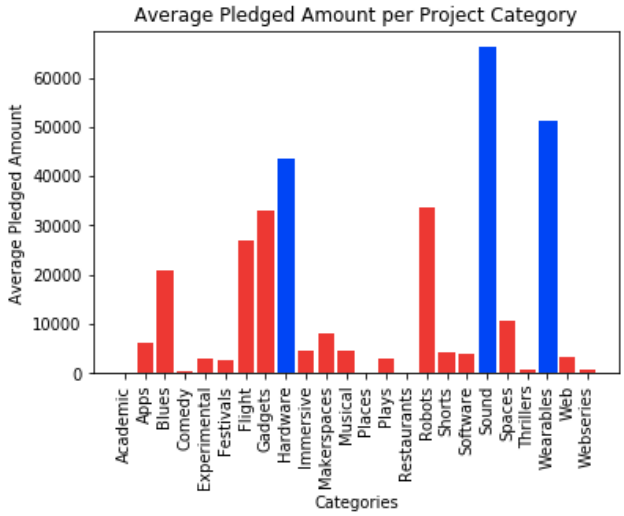
Summary Statistics

1. Spotlight vs State

Spotlight allows project creators to make a home for their project on Kickstarter after they've been successfully funded. Hence, if a project is on 'Spotlight', then State is always Successful and vice versa.

Spotlight	False	True
State		
canceled	2051	0
failed	9535	0
live	426	0
successful	0	4679
suspended	188	0

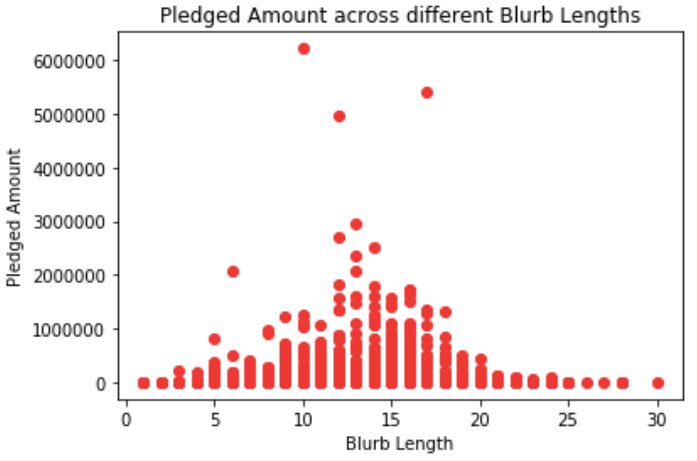
2. Average Pledged Amount per category:

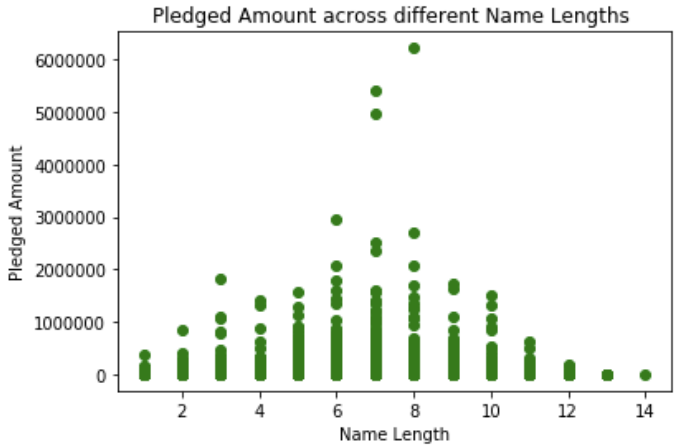


The plot shows that projects with Category 'Sound', 'Hardware' and 'Wearables' gets more pledged amount on an average than other categories. Hence, we can say that projects with

these categories are more likely to be pledged with a large amount and can be a good indicator for Pledged amount. Category ‘Web’ has the least proportion of ‘Successful’ projects, and hence can be a great predictor of the project state.

3. Blurb Length and Name Length

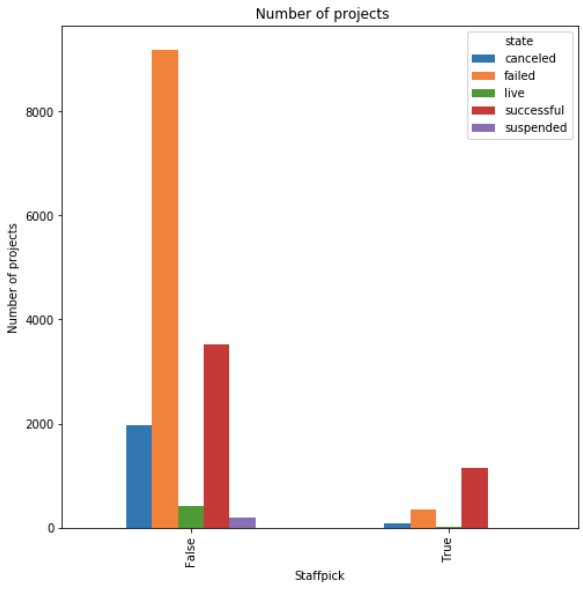




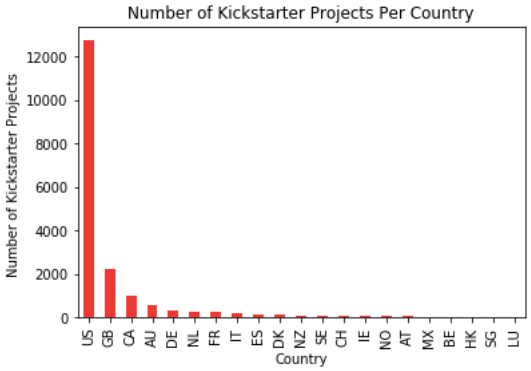
An interesting observation seen from the adjacent figure is that Pledged amount tends to be higher for those projects with a Blurb length of 10-20 (Considered to be the optimal length for description). Smaller Length and very large descriptions tend to get lesser pledged amount. A similar trend is seen across name length where Pledged amount increases initially, reaches a peak at an optimum, and continues to decrease.

4. Staff Pick

Staff Pick is a very important predictor in determining the success or failure of a project. A project has great chances of Failure if the project is not Staff Picked and great chances of a success if it is Staff Picked.



5. Country



The countries that have most number of Kickstarter projects are US, Great Britain (GB) and Canada (CA), and hence were considered important in the process categorization of countries.

Regression model

Initially extra variables like "Season" were created (by grouping 'months') and a random forest was run to perform feature selection to select the variables that have the most predictive power, based on Gini Coefficient. Unfortunately, "Season" did not surface as one of the important variables during the feature selection process. Top predictors from Feature selection, along with predictors obtained from Exploratory Analysis (such as Categories and Countries) were chosen to build predictive models to predict 'usd_pledged amount'. Below are the Mean Squared Errors (MSE) obtained as a result of each of the models.

Regression Algorithm	**MSE**
Linear Regression	13,373,911,730
Random Forest	8,282,961,949
Support Vector Regressor	14,558,517,960
Decision Tree Regressor	20,267,618,537

The best regression model (the model with the lowest MSE) was a Random forest regressor with 100 trees, with a max depth of 5, which included the following predictors:

usd_goal,create_to_launch_days,blurb_len_clean,category_Sound,deadline_month,deadline_day,staff_pick_False,staff_pick_True,category_Web,category_Hardware,country_US,country_GB,country_Other,country_CA,launch_to_deadline_days,name_len_clean,created_at_yr,deadline_yr

Goal is a good indicator of the amount pledged, especially when a project is successful(Since goal=pledged most of the times when a project is successful). If a project is not successful, then goal may differ from pledged amount. Blurb length and Name Length are very good

predictors of pledged amount as it was noticed that the amount of pledged amount had a triangular symmetry with increasing values of blurb and name length. Even though there wasn't any visible patterns of the correlation between Pledged amount and predictors like *create_to_launch_days, launch_to_deadline_days, deadline* etc, it gave a high Gini Coefficient during the process of feature selection using Random Forest, and hence were included in the set of predictors.

Classification model

The column 'spotlight' is True for all successful projects, because only Successful projects come under spotlight. Since Spotlight is a phenomenon that occurs only after State change, it could not be used as one of the predictors to predict spotlight. Similarly, the column 'pledged' could also not be used because it always occurs during the change of state of a project. With the exception of 'spotlight' and 'pledged', predictors were chosen with the help of feature selection and exploratory analysis. The list of predictors included were the same as the predictors used for regression with the addition of other categories like Category_Hardware and Category_Web.

The final list of predictors chosen were :

usd_goal,create_to_launch_days,blurb_len_clean,category_Sound,deadline_month,deadline_day,staff_pick_False,staff_pick_True,category_Web,category_Hardware,country_US,country_GB,country_Other,country_CA,launch_to_deadline_days,name_len_clean,created_at_yr,deadline_yr,category_Plays

Four models were generated to predict 'state' and their accuracy is as follows:

Classification Algorithm	**Accuracy Score**
Logistic Regression	0.666
Random Forest	0.734
Support Vector Regressor	0.730

Hence by going with the most accuracy score, the obvious choice was to choose a Random Forest Classifier.

Clustering

The main aim of clustering was to cluster the data points and find patterns in them. For this purpose, K-Means Clustering was implemented with a cluster size of 4 which resulted in a mean Silhouette score of 0.62. The differences and similarities among the clusters is summarized in the following table:

Predictors	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Label for Clusters	Old Projects released in Fall	New projects released in Summer	Not very old, not very new projects released in Summer (Outlier Class)	Recent projects released in mid-Spring
Launched Year	<=2014	>=2013	>=2013 <=2015	>=2015
Launched Month	Between August-December	May,June,July	June,July,August	March,April,May
Successful/Failed Projects	Mix of both	Mix of both	All failures	Mix of both
Mean Goal Amount in USD	$ 59,935.00	$ 69,179.00	$ 5,66,06,003.00	$ 74,703.00
Mean Name Length	5.29	5.1	4.5	5.35
Category	Hardware and Software	Web and Software	Hardware, Software, Web	Web and Gadgets

Conclusion

The regression model used is pretty good in predicting Pledged Amount with a Standard Deviation of $ 89442.7191, given that the model is not considering important factors such as ‘backers count’ (since it occurs at the end of the project), thus boosting the applicability of the model in a real-life Business scenario. Similarly the classification model does a fairly good job in predicting success of a project by considering only the factor relevant to project launch.