Group 204: Classify and Recommend Projects to Donors

First Name	Last Name	Email
Yeshwanthi	Jayaraman Durairaj	yurairaj@hawk.iit.edu

Table of Contents

1. Introduction

Donorschoose.org is a nonprofit organization that allows individuals from all over the world to donate to the projects proposed by students studying in public schools. This website consists of list of projects along with the information about the type of the project and resources utilized by the project. The aim of the project is to analyze the projects posted in Donorschoose.org and identify the projects which are less likely to be funded. These projects are then recommended to the donors based on their choice of interest. This will help the students to get funding for their projects which are not funded or less likely to be funded. This will also help donors to easily choose projects, that they wish to fund based on their preference.

Motive of segmenting donors: Donors are segments to different groups based on their choice of interest. The motive of segmenting donors is to send notification to segment of customers having same interest about a project, if the project fails to meet its funding requirement for a long time and nearing project expiry date.

2. Data

Kaggle Data Set

The Data set for this project belongs to Donorschoose.org and the data source is available in Kaggle website. There are six different data sets comprising data about the students and the project they are working, and the teachers and donor's data related to them. An overview of each data set is listed below.

Projects: Projects data set consist of data related to each project like project id, School id, Teacher id, Project posted date, Project completed date, Project Expiration Date, Project Donation Total amount, Project funding status, Project grade level, Project resource type etc.

Resources: Resources dataset comprises of resources required for a project. Each record can be joined with their project using the project id. Resource item name, Resource quantity, Resource unit price, Resource vendor name are the other features in this data set.

Donations: Donations data set consist of donation details of each project. Every record consists of donation id and donor id which helps to join with other data sets. Donation amount, cart sequence, receive date are the other features in this data set.

Donor: This data set consist of details of the donor like donor city, donor state and donor zip.

Student: This data set represents unique schools, and this is joined with the school id in the project data set. School name, School metro type, School state, School zip, School city, School county and School districts are the other features in this data set.

Teacher: This data set represents unique teachers, and this is joined with the school id in the project data set. School name, School metro type, School state, School zip, School city, School county and School districts are the other features in this data set. Teacher id, Teacher prefix and Teacher first project posted date are the other features in this data set.

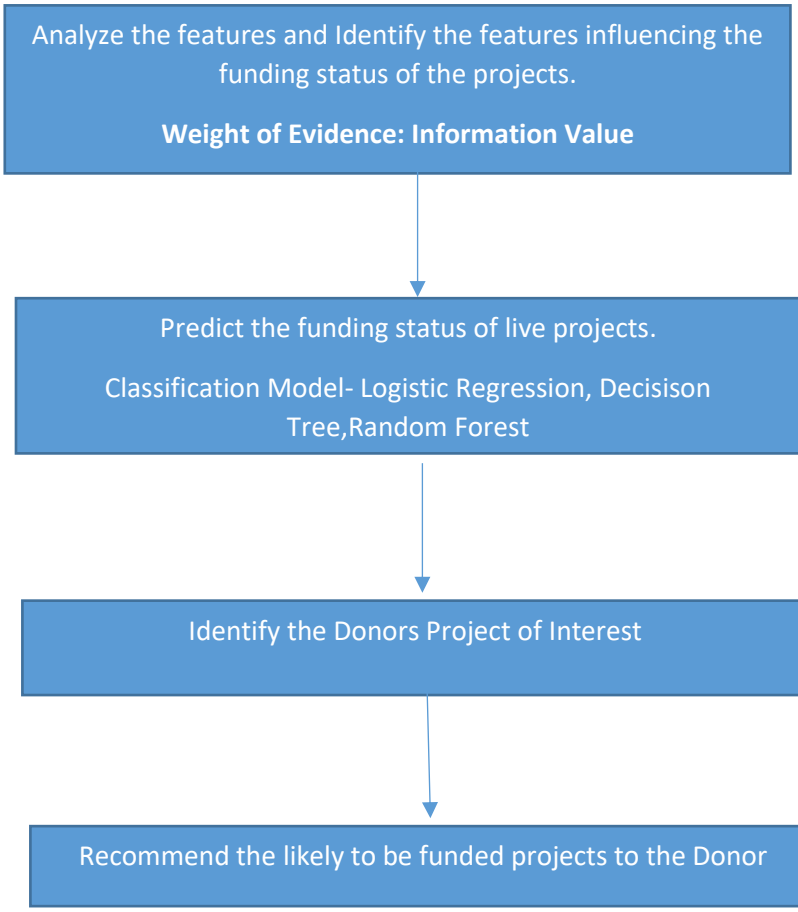
The joined data set consist of 16 GB of data and 27 variables.

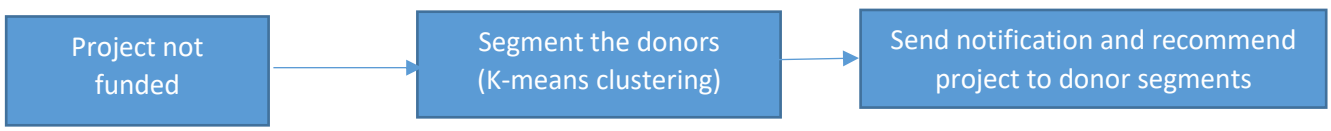
3. Problems to be Solved

The below mentioned Objectives are met in this project.

Analyze and identify the projects that are less likely to be funded based on the project category, resource category, students' grade, school region, project posted season. Fund goal is the target variable, 0 indicates that the project needs funding, 1 indicates project met funding requirements.

Recommend projects which are less likely to be funded to the donors based on the donor's choice of interest from their previous history of donations. This will make sure that more projects which are less likely to be funded are funded by the donors and help students to complete their project successfully.





4. KDD

4.1. Data Processing

Data Processing consisted of initial steps like data merging, data cleaning, data transformation, grouping of levels in categorical data. Analysis of target data.

4.1.1 Data Merging

Project, Resources and school data set was merged to form a consolidated data set to build classification models. Donations, Donors and Project data set was merged to form an aggregated data set at donor level. This merged data set is then aggregated by donor id to find number of projects the donor has donated, total amount of donation done, total number of donations to projects based on different subject category, total number of donations to projects based on different resource category, total number of donations to projects done based on the type of school the student is studying, total number of donations to projects based on the grade in which the student is studying.

Project:Resource:School

```
In [4]: # Joining Project and school data set
        project_school = pd.merge(projects,schools, on='School ID')
```

```
In [5]: project_school.head()
```

Out[5]:

	Project ID	School ID	Teacher ID	Teacher Project Posted Sequence	Project Type	Project Title	Project Essay	Pr Descri
0	7685f0265a19d7b52a470ee4bac883ba	e180c7424cb9c68cb49f141b092a988f	4ee5200e89d9e2998ec8baad8a3c5968	25	Teacher-Led	Stand Up to Bullying: Together We Can!	Did you know that 1-7 students in grades K-12 ...	Di kno stude grades
1	e11cc13f5f29c1a92c1a396a8c7471fb	e180c7424cb9c68cb49f141b092a988f	4ee5200e89d9e2998ec8baad8a3c5968	26	Teacher-Led	Recipe For Happiness: Children Need Support!	Did you know that bullying takes many forms? I...	Di kno bu takes form
2	9e367907db2c9b5c0b23fa6c1318eced	e180c7424cb9c68cb49f141b092a988f	4ee5200e89d9e2998ec8baad8a3c5968	27	Teacher-Led	Follow the Money!	What is the price kids pay when they do not un...	What pric pay they c

Donations:Donor:Project

```
In [47]: donor_agg_data.head()
```

Out[47]:

Donor ID	Donor City	Donor State	Donor Is Teacher	Donation Amount	Project ID	Applied Maths & Science	Extracurricular	History & Civics	Literacy & Language	Art Instruments	BASICS	E
00000ce845c00cbf0686c992fc369df4	Evanston	Illinois	No	50.00	1	0.0	0.0	0.0	1.0	0.0	0.0	
00002783bc5d108510f3f9666c8b1edd	Appomattox	other	No	99.00	1	0.0	0.0	1.0	0.0	0.0	0.0	
00002d44003ed46b066607c5455a999a	Winton	California	Yes	425.26	8	1.0	1.0	5.0	1.0	0.0	7.0	
00002eb25d60a09c318efbd0797bffb5	Indianapolis	Indiana	No	210.00	5	2.0	0.0	0.0	3.0	0.0	5.0	
00004e32a448b4832e1b993500bf0731	Stamford	Connecticut	No	25.00	1	0.0	0.0	0.0	1.0	0.0	1.0	

4.1.2 Data Cleaning

Missing Values:

There were few missing values for project subject category and project resource category. These missing values are identified are filled by the mode of the feature values.

```
In [53]: # Check for missing values
print(req_project_school.isnull().sum(axis = 0))
# View Records having null values
#req_project_school[req_project_school['Project Subject Category Tree'].isnull()]
# Fill missing values with mode for Project Resource Category Tree and Project Subject Category Tree
req_project_school['Project Resource Category']=req_project_school['Project Resource Category'].fillna(req_project_school['Project
req_project_school['Project Subject Category Tree']=req_project_school['Project Subject Category Tree'].fillna(req_project_school[
# Remove Projects with null Expiry date
```

```
Project ID                           0
Teacher Project Posted Sequence      0
Project Type                         0
Project Subject Category Tree       29
Project Grade Level Category         0
Project Resource Category           36
Project Expiration Date             14
Project Posted Date                  0
Project Fully Funded Date       283244
Project Cost                         0
School Metro Type                    0
Project Current Status               0
dtype: int64
```

Dropping Duplicates:

There were few duplicate records for the same project ID. These records are identified and dropped from the data set.

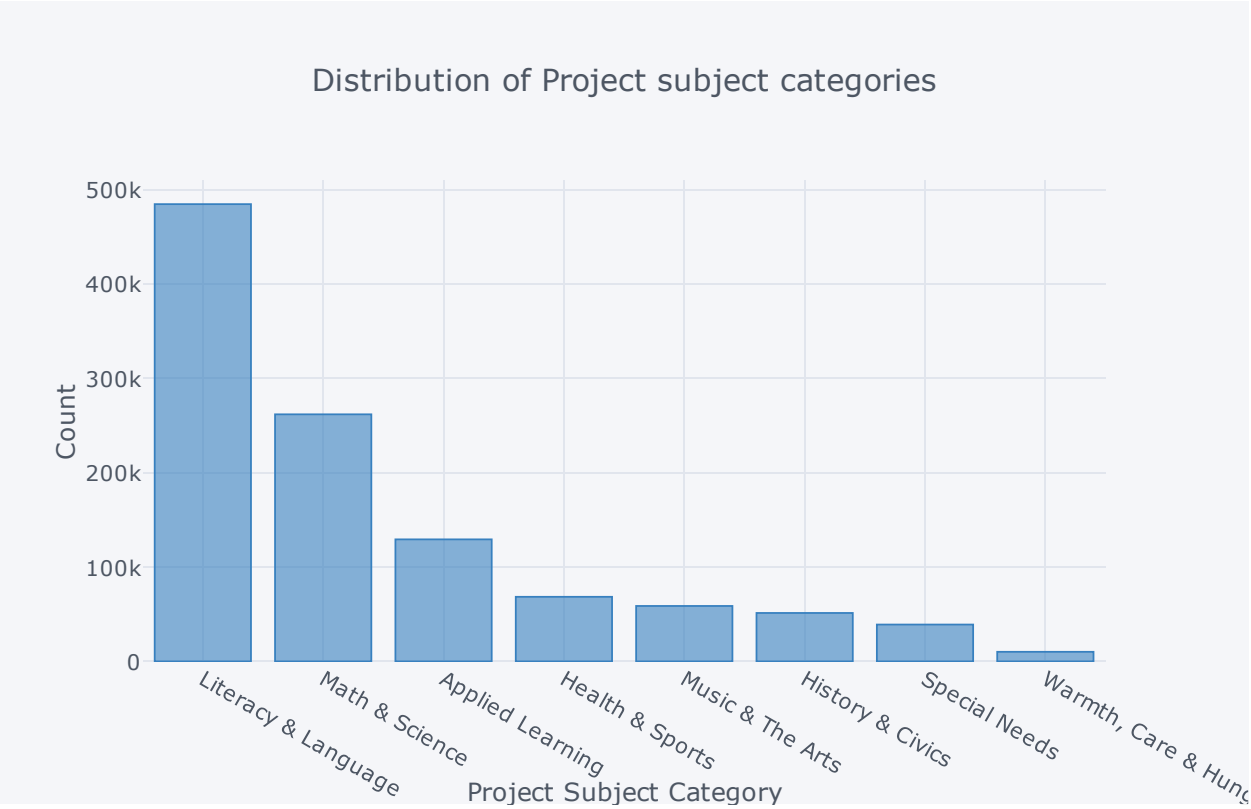
	Project ID	School ID	Teacher ID	Project Posted Sequence	
22650	c940d0e78b7559573aca536db90c0646	5e1132a3628825b31f329a70fcec7030	69b63ccd52afd09ef3963b4cc3b15d00	1	T
22651	c940d0e78b7559573aca536db90c0646	5e1132a3628825b31f329a70fcec7030	69b63ccd52afd09ef3963b4cc3b15d00	1	I C
399918	99c07777fdcf63d3a0fdb4a0deb4b012	de69b6ee78c74d4356bf3302a649fb1e	05f09b29bc8fd4f3cc10fc75eccbb573	11	T
399919	99c07777fdcf63d3a0fdb4a0deb4b012	de69b6ee78c74d4356bf3302a649fb1e	05f09b29bc8fd4f3cc10fc75eccbb573	11	I C

Grouping levels of categorical data:

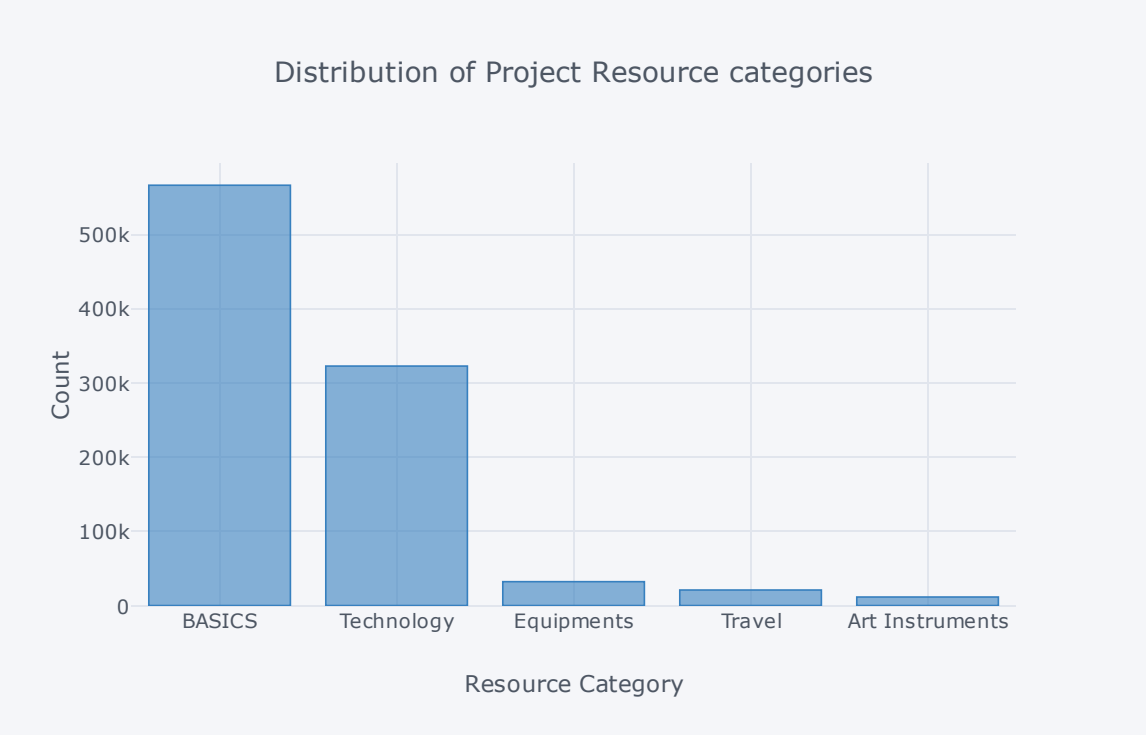
Project Subject Category, Project Resource Category are categorical features with many levels. These many levels were very much hard to interpret so the levels are grouped and reduced to lesser number of levels.

Project Subject Category:

Project Subject Category had around 35 categories. I analyzed these data and grouped and reduced to seven different categories based on the subject specialization.

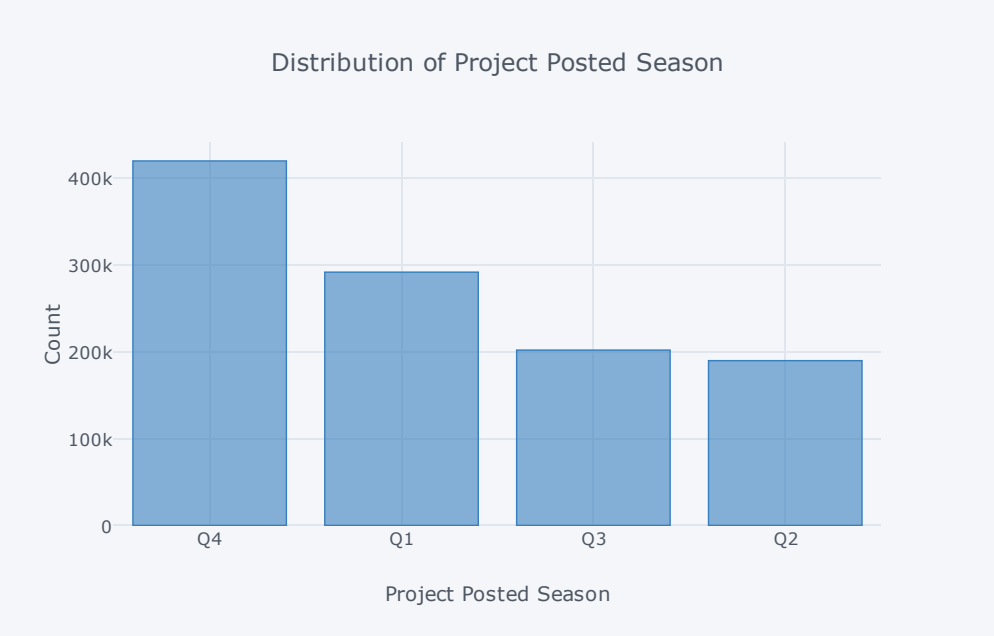


Project Resource Category: Project Subject Category had around 35 categories. I analyzed these data and grouped and reduced to seven different categories based on the subject specialization.



Deriving new features: Features are derived from the existing features to have make use of the features in the predictive models which may influence the label outcome.

Project Posted Season: Project Posted Season was derived from the month of the Project Posted Season which will help in classifying the less likely to be funded projects based on the Project Posted Season.



Project Duration:

Project Duration was derived from the Project Expiry date and Project Posted Date. This features number of days allocated for completing a project which will help in classifying the less likely and more likely to be funded projects.

4.1.3. Segmentation- Processing Donor Data:

Donors data are processed and aggregated based on Donor ID. The aggregated data set consist of number of projects done be the donor, total donation amount donated by a donor, number of donations for the projects done in the subjects separately, number of donations for the projects done using the resource category separately, number of donations made for project done by a student studying in rural school, number of donations made for project done by a student studying in urban school, number of donations made for projects done by a student studying in lower grades, number of donations made for projects done by a student studying in middle grades, number of donations made for projects done by a student studying in higher grades.

```
In [47]: donor_agg_data.head()
```

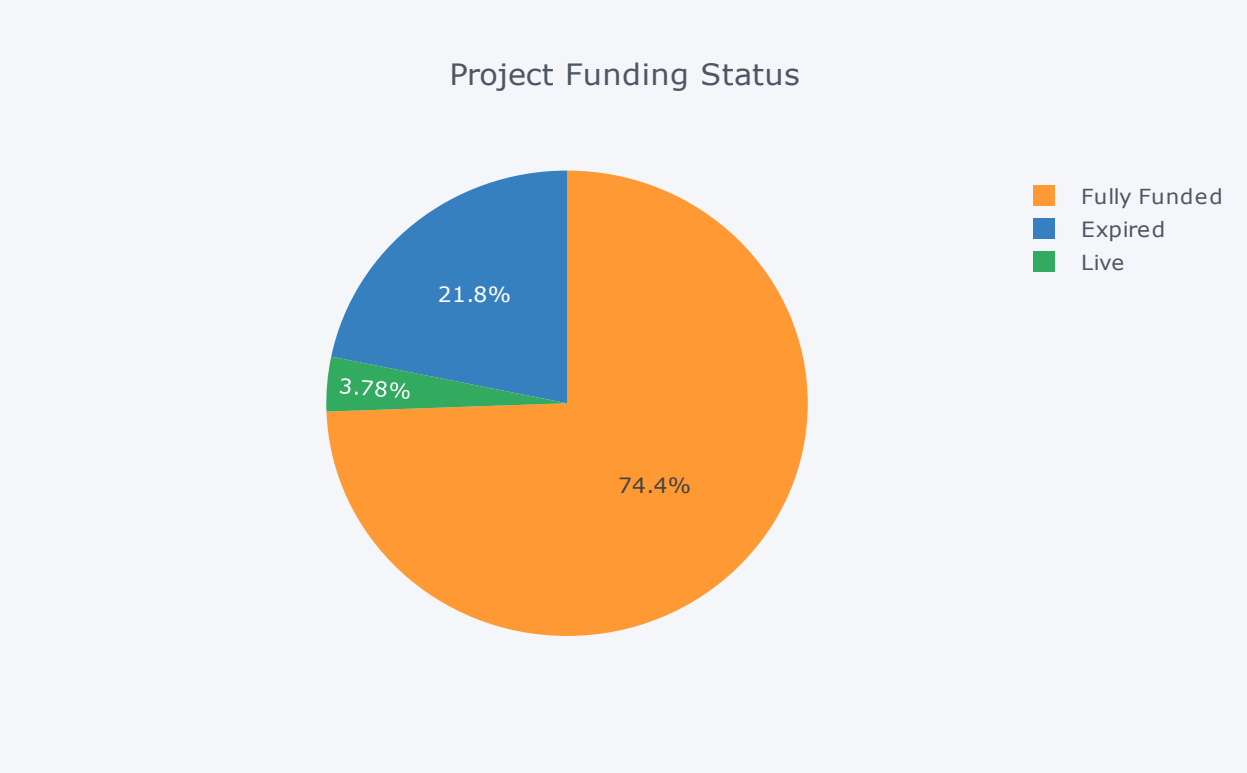
Out[47]:

Donor ID	Donor City	Donor State	Donor Is Teacher	Donation Amount	Project ID	Applied Maths & Science	Extracurricular	History & Civics	Literacy & Language	Art Instruments	BASICS	E
00000ce845c00cbf0686c992fc369df4	Evanston	Illinois	No	50.00	1	0.0	0.0	0.0	1.0	0.0	0.0	
00002783bc5d108510f3f9666c8b1edd	Appomattox	other	No	99.00	1	0.0	0.0	1.0	0.0	0.0	0.0	
00002d44003ed46b066607c5455a999a	Winton	California	Yes	425.26	8	1.0	1.0	5.0	1.0	0.0	7.0	
00002eb25d60a09c318efbd0797bffb5	Indianapolis	Indiana	No	210.00	5	2.0	0.0	0.0	3.0	0.0	5.0	
00004e32a448b4832e1b993500bf0731	Stamford	Connecticut	No	25.00	1	0.0	0.0	0.0	1.0	0.0	1.0	

4.2. Data Mining Methods and Processes

4.2.1 Predictive Modelling

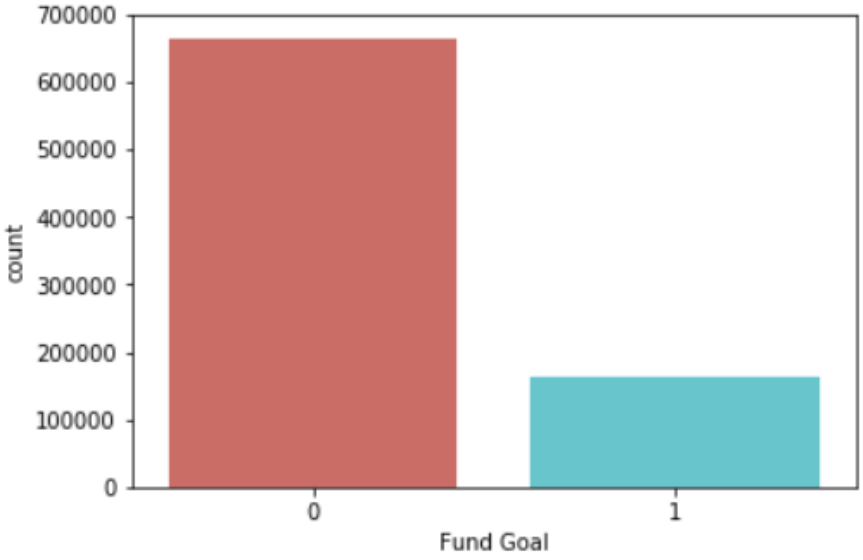
Class Label: A target label called as **Fund Goal** was created based on the Project current Status. All expired projects are labelled as 1 (Not Funded Projects) and projects which are funded are labelled as 0.



Splitting scoring and modelling data: Data set was divided into modelling and scoring data set based on the project current status. Expired and Fully funded projects are our model data set and live projects are our scoring data set. Scoring data set will be predicted using the predictive model which we will build using our modelling data set.

Class Imbalance Issue: Exploratory analysis on target label (Fund Goal) was done and I discovered that there is class imbalance issue with the target label. 80 % of the projects are funded and 20 % of the projects are expired in our model data set.

```
percentage of funded is 80.15029135662098
percentage of expired 19.849708643379014
```

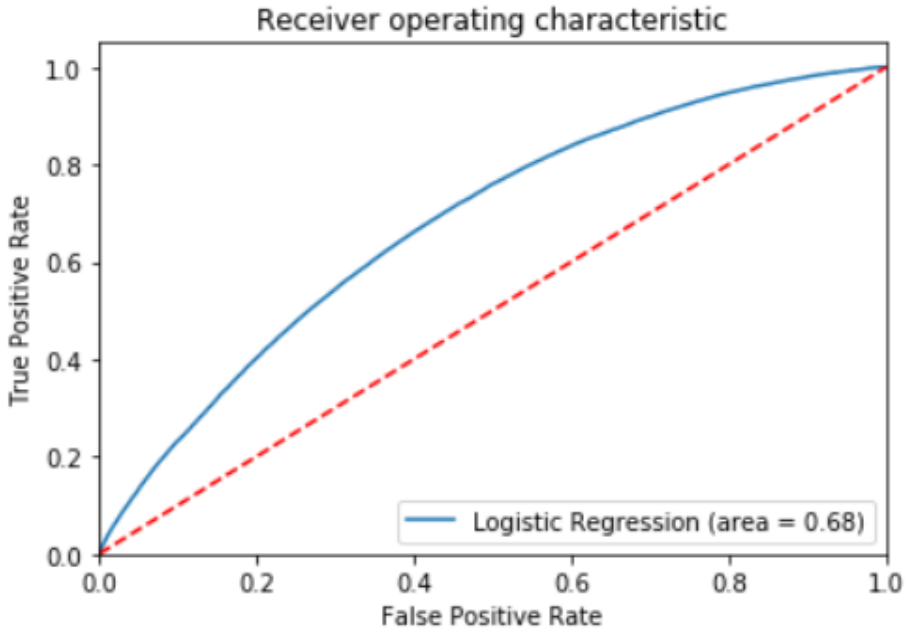


Over Sampling Data: Expired projects are over sampled in the training data set to overcome the class imbalance issue. Oversampling was done by creating similar samples of expired projects from its nearest neighbors.

```
length of oversampled data is  931056
Number of expired projects in oversampled data 465528
Number of fully funded 465528
Proportion of expired data in oversampled data is  0.5
Proportion of fully funded data in oversampled data is  0.5
```

Model -1: Logistic Regression:

```
Accuracy of logistic regression classifier on test set: 0.65
```



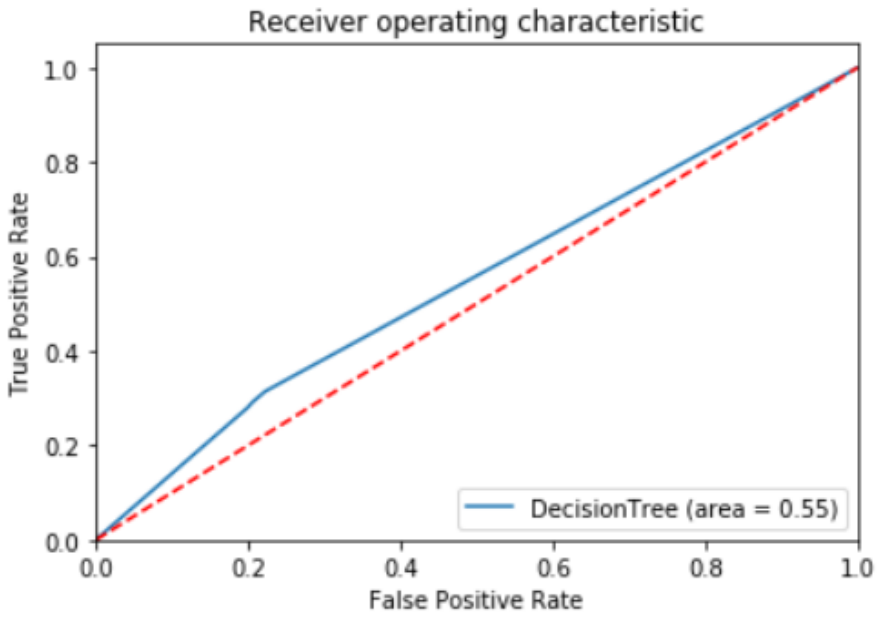

```
[[132801  67122]
 [ 20270  28884]]
Misclassification rate is  0.3508633876271193
              precision    recall  f1-score   support

           0       0.87      0.66      0.75    199923
           1       0.30      0.59      0.40     49154

   micro avg       0.65      0.65      0.65    249077
   macro avg       0.58      0.63      0.58    249077
weighted avg       0.76      0.65      0.68    249077
```

Model 2- Decision Tree

```
Accuracy of Decision Tree classifier on test set: 0.69
```



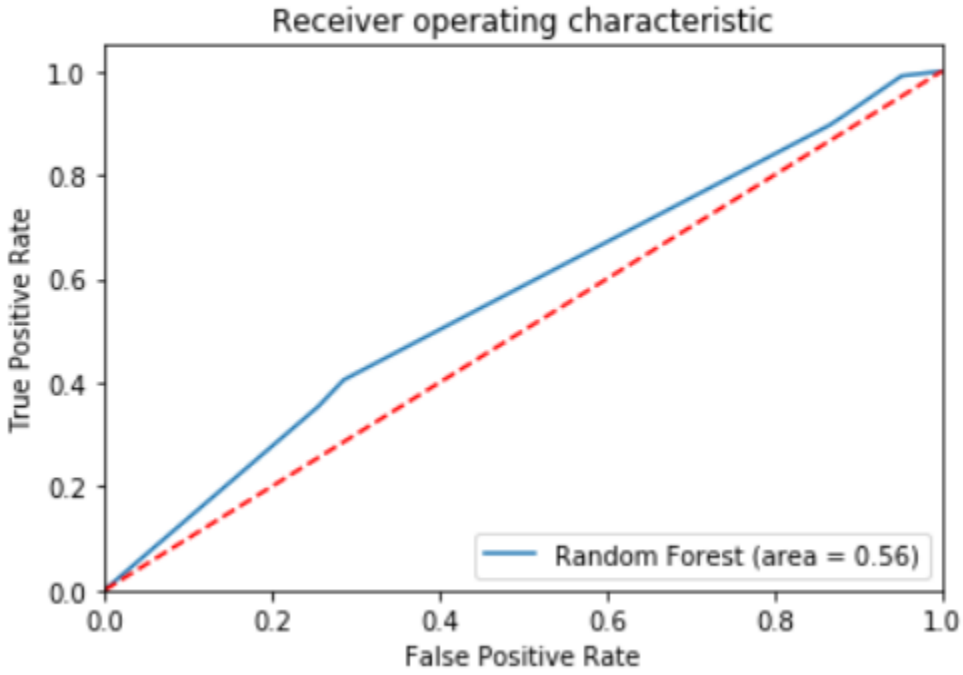

```
[[158477  41446]
 [ 34721  14433]]
Misclassification rate is  0.3057970025333531
              precision    recall  f1-score   support

           0       0.82      0.79      0.81    199923
           1       0.26      0.29      0.27     49154

   micro avg       0.69      0.69      0.69    249077
   macro avg       0.54      0.54      0.54    249077
weighted avg       0.71      0.69      0.70    249077
```

Model 3 - Random Forest

```
Accuracy of Random Forest classifier on test set: 0.67
```




```
[[148835  51088]
 [ 31744  17410]]
Misclassification rate is  0.3325557959988277
              precision    recall  f1-score   support

           0       0.82      0.74      0.78    199923
           1       0.25      0.35      0.30     49154

   micro avg       0.67      0.67      0.67    249077
   macro avg       0.54      0.55      0.54    249077
weighted avg       0.71      0.67      0.69    249077
```

Feature Selection: Feature selection was done to get the most optimum subset of features which influences in determining the funding status of the projects. This improves the accuracy of the model and reduces noise and eliminates outliers for the predictive models. Information value and weight of Index are used to identify the value of the features and select the features that influences Fund Goal.

Weight of evidence and Information value: Weight of evience for each feature is calculated by finding the log value of every feature occurrence with the event class divided by every feature occurrence with the nonevent class.

Out[50]:

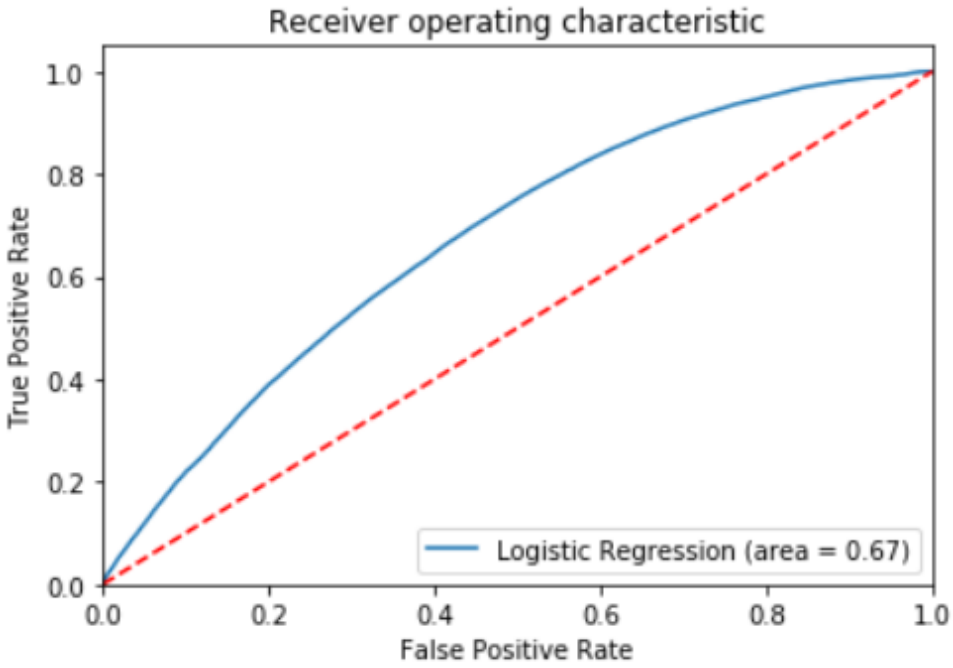
	index	IV
2	Project Cost	0.48
8	Resource Category	0.16
9	School Metro Type	0.01
1	Project Category	0.00
3	Project Current Status	0.00
7	Project_Posted_Season	0.00
5	Project Type	0.00
0	Grade Level	0.00
6	Project_Duration	0.00
4	Project ID	0.00

Based on the weight of Index method, the Information value for each feature are displayed above. Project Cost, Resource Category and School Metro Type are the important features that influences the target variable Fund Goal.

Note – I used all feature selection and feature reduction methods and the variables suggested by weight of evidence improved the accuracy of the model.

Model 4- Logistic Regression (After Feature Selection):

```
Accuracy of logistic regression classifier on test set: 0.66
```



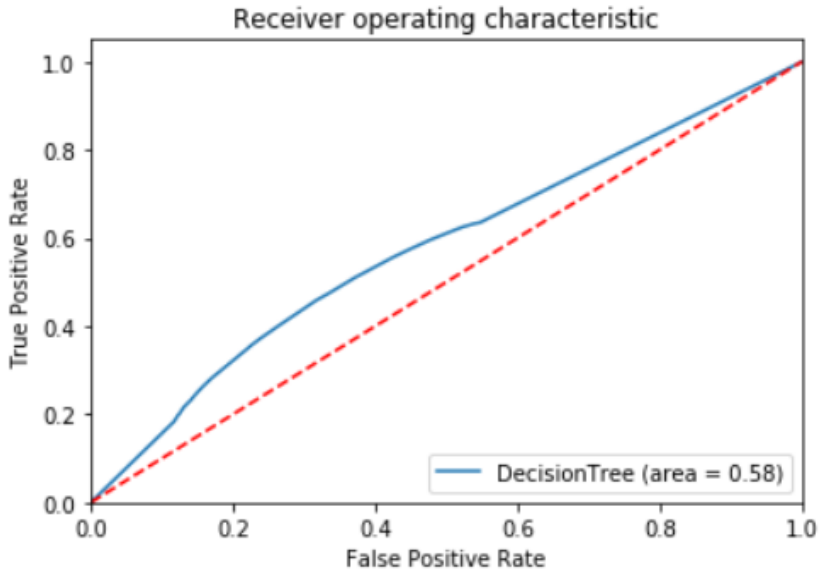

```
[[136904  63019]
 [ 22305  26849]]
Misclassification rate is  0.34256073423078004
              precision    recall  f1-score   support

           0       0.86      0.68      0.76    199923
           1       0.30      0.55      0.39     49154

   micro avg       0.66      0.66      0.66    249077
   macro avg       0.58      0.62      0.57    249077
weighted avg       0.75      0.66      0.69    249077
```

Model 5 - Decision Tree (After Feature Selection):

```
Accuracy of Decision Tree classifier on test set: 0.72
```



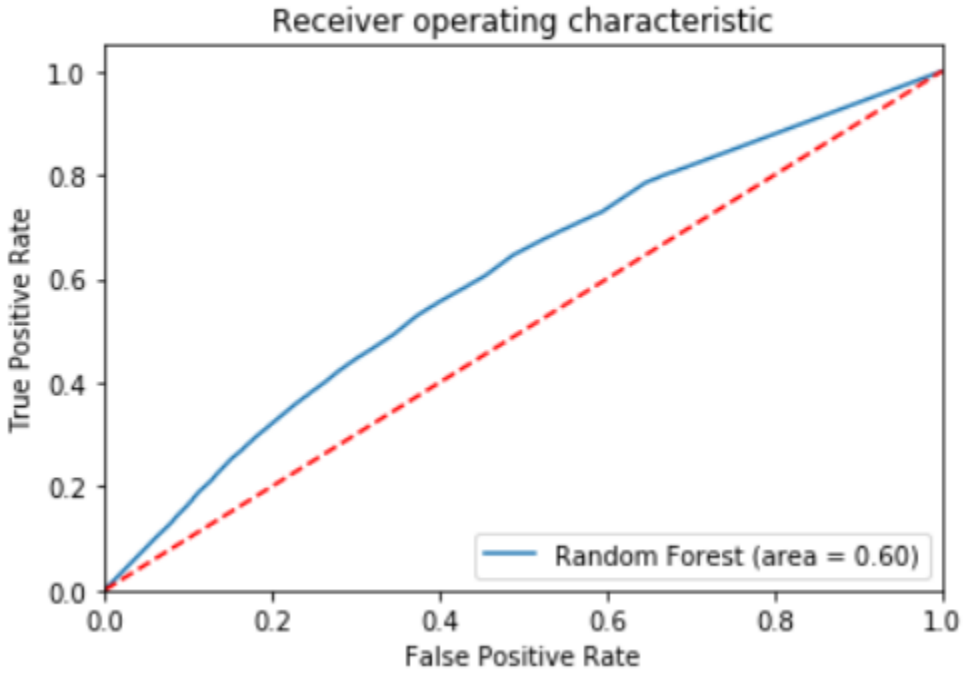

```
[[166127  33796]
 [ 35354  13800]]
Misclassification rate is  0.2776249914685017
              precision    recall  f1-score   support

           0       0.82      0.83      0.83    199923
           1       0.29      0.28      0.29     49154

   micro avg       0.72      0.72      0.72    249077
   macro avg       0.56      0.56      0.56    249077
weighted avg       0.72      0.72      0.72    249077
```

Model 6 - Random Forest (After Feature Selection):

```
Accuracy of Random Forest classifier on test set: 0.71
```




```
[[160095  39828]
 [ 33414  15740]]
Misclassification rate is  0.29405364606125817
              precision    recall  f1-score   support

           0       0.83      0.80      0.81    199923
           1       0.28      0.32      0.30     49154

   micro avg       0.71      0.71      0.71    249077
   macro avg       0.56      0.56      0.56    249077
weighted avg       0.72      0.71      0.71    249077
```

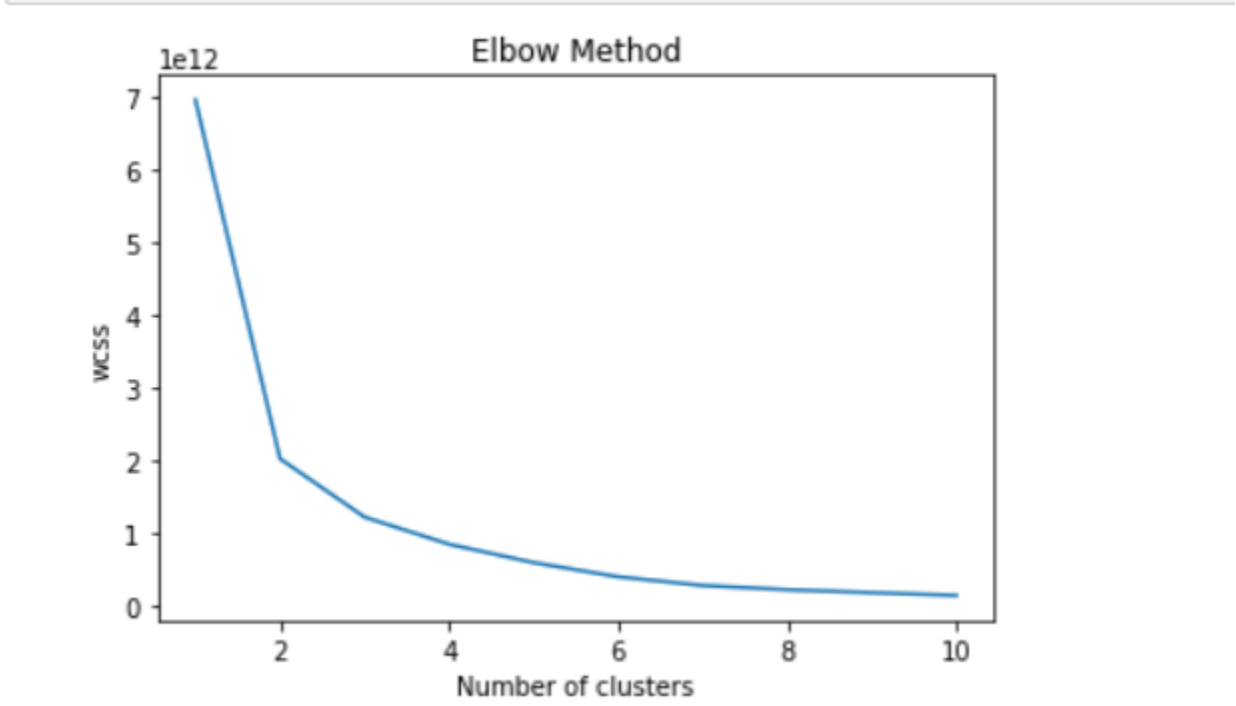
Segmentation of Donors:

Donors are segmented into different groups based on their choice of Interest from previous history of donations. Donors having similar interest in projects are clustered into same group and donors having different interest are clustered into another group.

K- Means Clustering:

K- Means clustering is used to cluster the donors. The numbers of clusters required for the aggregated data set is decided by using Elbow method.

Elbow method:



The elbow method suggests 4 clusters.

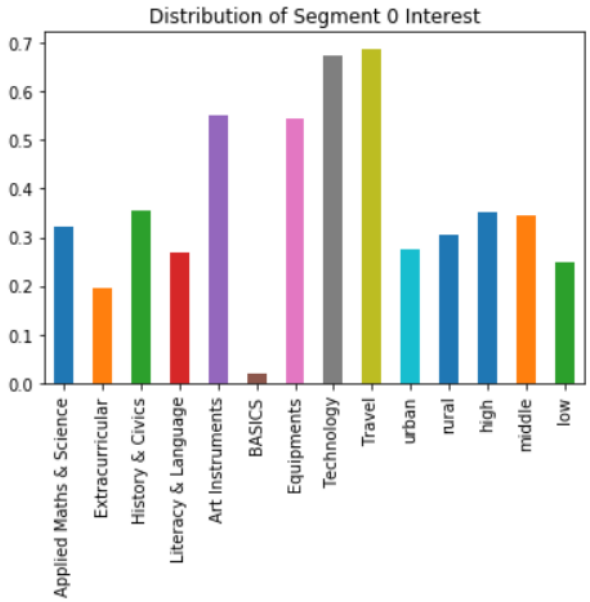
Segmentation of Donors- Importance of Variables with respect to segments.

Donors are segmented to fours segments based on their interest in donating towards specialization of project, resource used by the projects, location of the school, grade of the students doing the project.

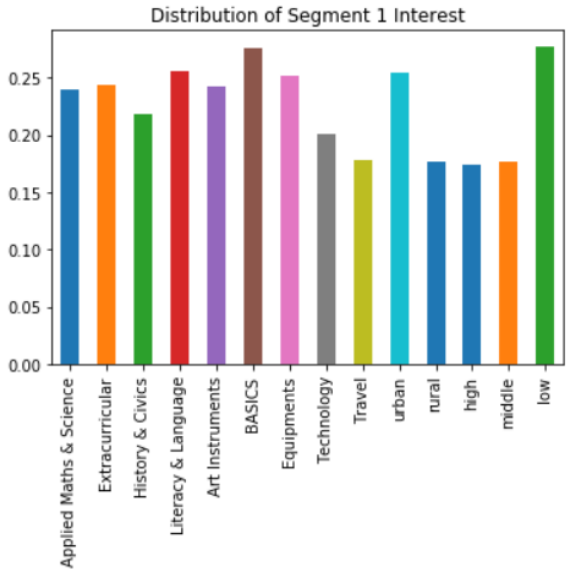
Out[48]:

ients	Applied Maths & Science	Extracurricular	History & Civics	Literacy & Language	Art Instruments	BASICS	Equipments	Technology	Travel	urban	rural	high	middle	low
0	0.322574	0.194569	0.354776	0.267776	0.549640	0.018179	0.543436	0.672918	0.687514	0.275702	0.305367	0.353053	0.344007	0.248960
1	0.239643	0.243381	0.218066	0.255874	0.242812	0.276245	0.251755	0.201240	0.178131	0.254543	0.176700	0.173435	0.176740	0.277263
2	0.295473	0.416198	0.293990	0.327758	0.058009	0.543203	0.043445	0.005298	0.047840	0.317298	0.435587	0.374967	0.370323	0.310899
3	0.142310	0.145852	0.133168	0.148592	0.149539	0.162374	0.161363	0.120544	0.086515	0.152456	0.082347	0.098545	0.108930	0.162877

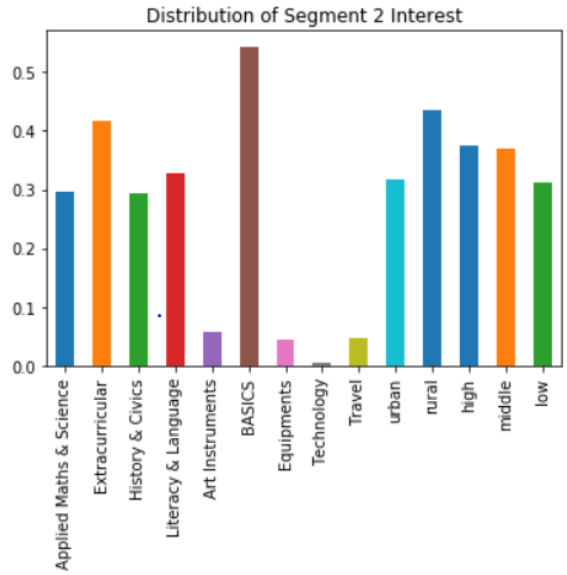
Segment 0: Donors in segment 0 are interested in History and civics projects using Travel and Technology resources.



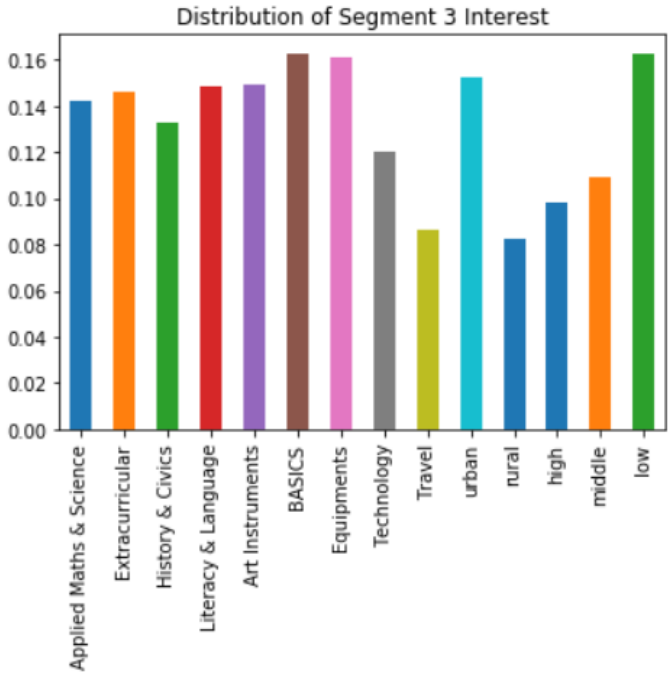
Segment 1: Donors in segment 1 are highly interested with students in lower grades studying in urban school in History and Language specialization.



Segment 2: Segment 2 donors are highly interested in Basics resources for students studying in rural schools.



Segment 3: Segment 3 students are interested in students studying in lower grades with Equipments and basics resources.



5. Evaluations and Results

5.1. Evaluation Methods

Evaluation of Predictive Models

The various predictive models built using modelling data set is evaluated based on the **mis classification** rate and the model with better misclassification rate is Decision Tree built after applying WOE feature selection.

Model	Accuracy	Misclassification rate
Logistic Regression	65%	35%
Decision Tree	69%	30%
Random Forest	67%	33%
Logistic Regression with feature selection	66%	34%
Decision Tree with feature selection	73%	27%
Random Forest with feature selection	71%	29%

Optimum model – Decision Tree after applying feature selection.

Class Label: Fund Goal.

Independent Features: Project Cost, Project Resource Category, Project School Metro Type.

5.2. Results and Findings

Predictive model Feature Importance:

Feature Importance

The analysis of predictive model says the below findings.

	Feature_names	Importance
0	Project Cost	0.652881
5	School Metro Type_urban	0.032616
3	Resource Category_Technology	0.012464
2	Resource Category_Equipments	0.011640
1	Resource Category_BASICS	0.008863
4	Resource Category_Travel	0.001563

Project Cost: Project cost is the important feature which determines whether project will be funded or not. As the project cost increases, chance of funding decrease.

School Metro Type Urban: Projects done by the students studying in urban school are less likely to be funded.

Resource Category Technology and Equipment's: Projects done using Technology and Equipment's are less likely to be funded

Scoring Funding Status of Live Projects:

The Funding Status of the live projects are scored using the Decision Tree model built after feature selection. There are totally 32,412 projects out of which 5890 projects are predicted to less likely to be funded and 26522 projects are predicted to be more likely to be funded.

```
In [18]: print("Number Of LiveProjects is ",len(scoring_data))
         print("Number of Projects Less Likely to be Funded ",len(scoring_data[scoring_data['Fund Goal']==1]))
         print("Number of Projects More Likely to be Funded",len(scoring_data[scoring_data['Fund Goal']==0]))
```

```
Number Of LiveProjects is  32412
Number of Projects Less Likely to be Funded  5890
Number of Projects More Likely to be Funded 26522
```

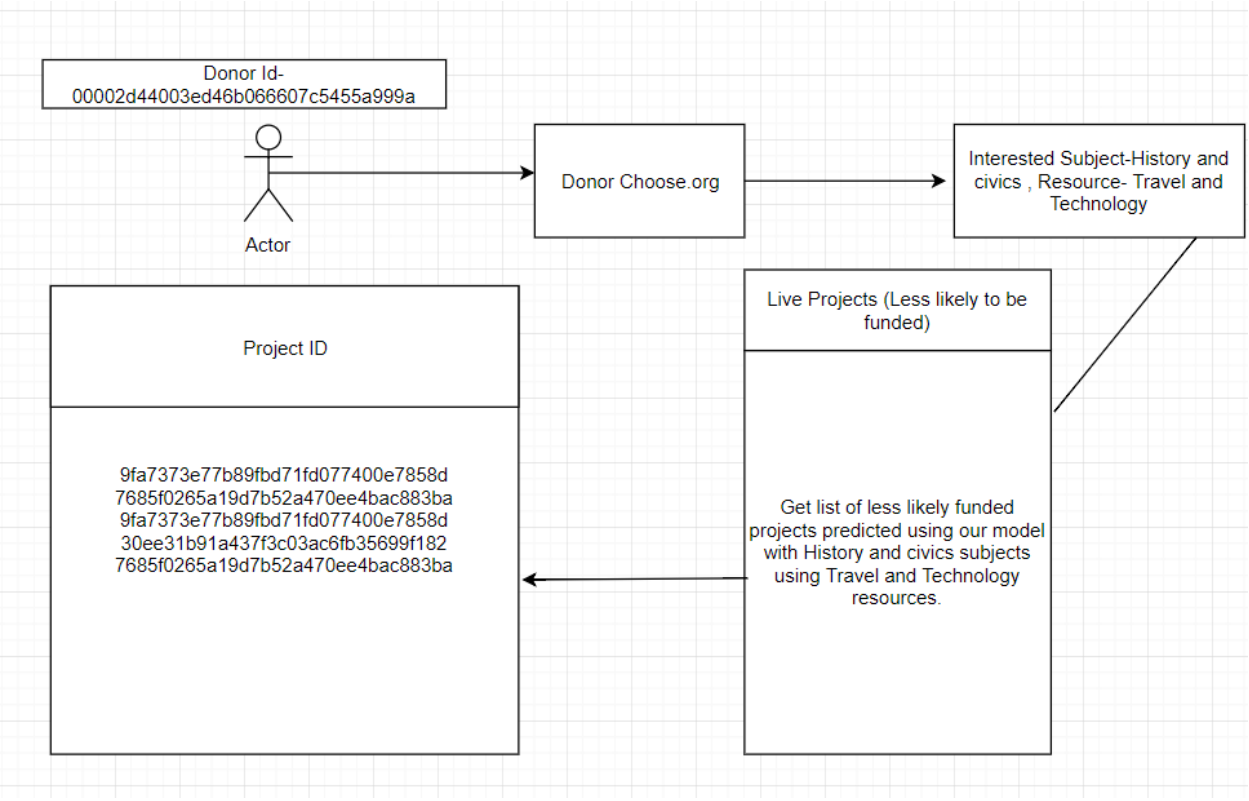
Segmentation of Donors:

Segments	Interest
0	Applied Maths and Science, Technology and Travel
1	low grades,urban schools, History and civics specialization
2	Basics resource, rural schools
3	low grades ,Equipments and basics resources, History and civics

Recommend Projects to Donors

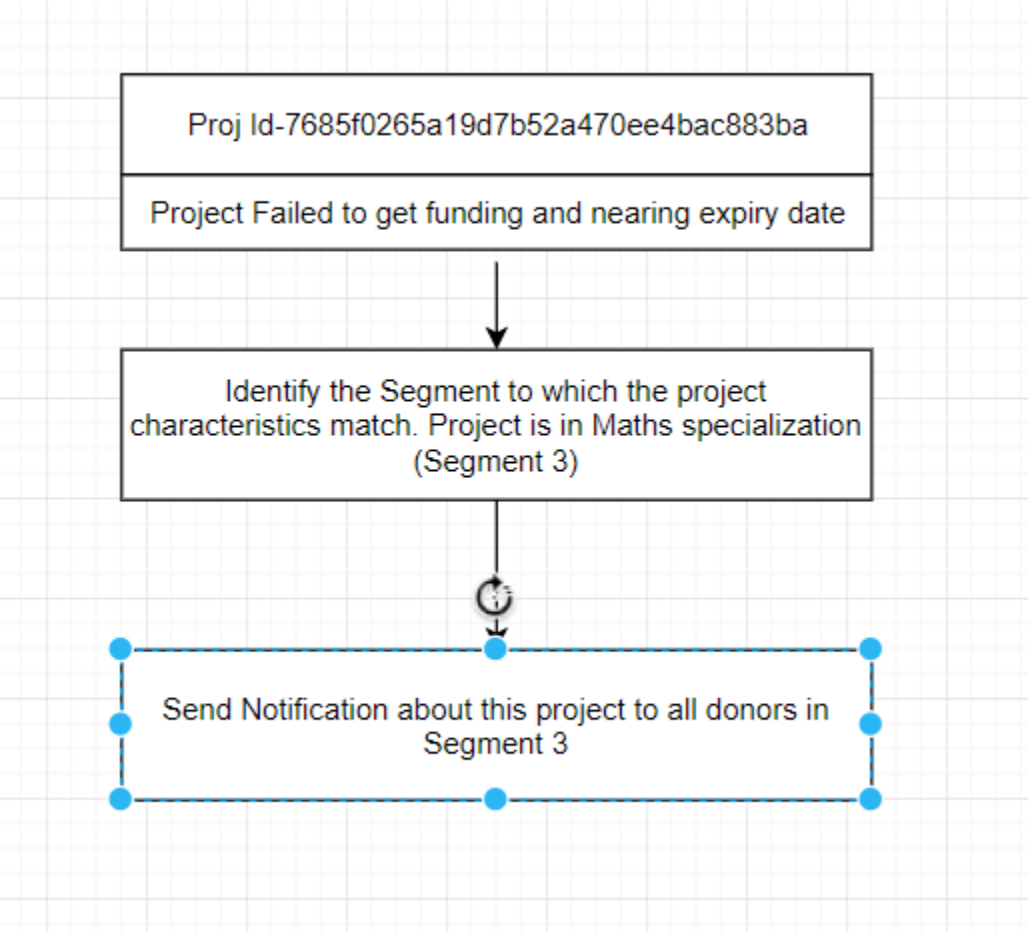
Projects that are less likely to be funded can be recommended to the donors based on their choice of interest from their previous history of donations. This will increase the chance of project which are less likely to be funded.

For example, if a Donor ID login to donor choose.org, his interest in projects are identified based on his previous history of donations. Then list of projects which are under the donor interest and predicted as less likely to be funded are recommended to the donor.



Recommend Projects to segment of donors

Projects which are donated by any donors can be recommended to the donors by sending notifications to all donors belonging to a cluster to which the project is characterized upon. This is the main motive behind segmenting donors based on their choice of interest.



6. Conclusions and Future Work

6.1. Conclusions

Thus, Donors choose website can be made more useful and better by incorporating these objectives in their website. This will not only help donors to identify their projects that they want to donate but also ensures that projects that are less likely to be funded are donated. Students will get benefited through this classification models and recommendation system.

6.2. Limitations

This project is limited to donors who already donated through Donors Choice.org and not for any first-time donors. We cannot recommend projects to first time donors as we don't know anything about their choice of interest.

6.3. Potential Improvements or Future Work

As mentioned above in the limitations, we cannot recommend projects to first-time donors. This can be enhanced by asking their choice of interest when the donor's login for first time and recording their interest in the respective segments. This will help in recommending projects for first time donors. Text mining will be done on the project brief description to predict the project funding status.