

Report by Avengers

Proposing a Recommendation system for DonorsChoose.org

Dataset chosen: Donors Choose.org

SAS **GLOBAL**FORUM

IMAGINE. CREATE. INNOVATE.
LAS VEGAS | APRIL 19-21, 2016

sas



Resources on the Donors Choose website

CURRENT FEATURES

News on the Donors Choose

<http://www.donorschoose.org/>

TIMELINE

Detailed timeline of their history

<http://www.donorschoose.org/about/story.html>

API AND OPEN DATA

There is a complete data available free for download. There is data available for five major categories: **Projects:** All classroom projects that have been posted so far on the website, including lots of school info such as its NCES ID (government-issued), latitude/longitude, and city/state/zip.

Donations: All donations, including donor city, state, and partial-zip. **Gift cards:** All website-purchased gift cards, including donor and recipient city, state, and partial-zip. **Project resources:** All materials/resources requested for the classroom projects, including vendor name. **Project written requests / essays:** Full text of the teacher-written requests accompanying all classroom projects.

<http://data.donorschoose.org/open-data/overview/>

<http://data.donorschoose.org/open-data/overview/>

<http://data.donorschoose.org/open-data/overview/>

BLOG

Interesting articles and success stories

<http://www.donorschoose.org/blog/>

HOW IT WORKS

Brief video on how this works

<http://www.donorschoose.org/about>

IMPACT

Informative visualizations on the impact of Donors Choose

<http://www.donorschoose.org/about/story.html>

PARTNERS

A complete list of their partners and supporters

http://www.donorschoose.org/about/partnership_center.html

VOLUNTEER WITH DONORSCHOOSE

Volunteers review the thank-you notes to protect student privacy, check them into our computer system, repackage them, and mail them off to eager donors — all within 24 hours.

<http://www.donorschoose.org/volunteer>

STAFF AND BOARD

Detailed list of people associated with Donors Choose – The Team

http://www.donorschoose.org/about/meet_the_team.html

OPEN PROJECTS

Comprehensive list of projects that needs funding

<http://www.donorschoose.org/donors/search.html>



Table of Contents

INTRODUCTION	4
PROBLEM STATEMENT	5
DATA CLEANING AND VALIDATION	5
VISUALIZATION	6
ANALYSIS	6
PREDICTIVE MODELING	6
DONOR SEGMENTATION	7
RECOMMENDATION SYSTEM.....	8
SUGGESTION FOR FUTURE STUDY.....	9
CONCLUSIONS	9
APPENDICES	10
APPENDIX A: DATA EXTRACTION AND CLEANING	10
APPENDIX B: EXPLORATORY ANALYSIS.....	12
APPENDIX C: PREDICTIVE MODELING AND SEGMENTATION RESULTS	15

DonorsChoose.org is a United States–based nonprofit organization that allows individuals to donate directly to public school classroom projects. Founded in 2000 by former public school teacher Charles Best, DonorsChoose.org was among the first civic crowd funding platforms of its kind. The organization has received Charity Navigator’s highest rating every year since 2005.

Charles Best, a social studies teacher at Wings Academy in the Bronx, New York, founded donorsChoose.org. Charles and his colleagues often spent their own money on school supplies for their students, and discussed materials they wished they could afford in the teachers’ lunchroom.

Charles envisioned a platform for individuals to connect directly with classrooms in need, providing materials requested by teachers. With the help of his students, he built the first version of the site in his classroom and invited colleagues to post material requests. Charles anonymously funded the first 10 project requests to demonstrate the effectiveness of the site.

INTRODUCTION

Donorschoose.org is a nonprofit organization that allows individuals to donate directly to public school classroom projects. Teachers from public schools post request for funding project with a short essay describing it. Donors all around the world can look at these projects when they login to Donorschoose.org and donate to projects of their choice. The idea is to have personalized recommendation webpage for all the donors, which will show them the projects, which they prefer, like and love to donate. Implementing the recommender system for DonorsChoose.org website will improve user experience and help more projects to meet their funding goals. It also will help us in understanding the donors' preferences and delivering to them what they want or value. One type of recommendation system can be designed by predicting projects that will less likely to meet funding goal, segmenting and profiling the donors and using that information for recommending right projects when the donors login to DonorsChoose.org.

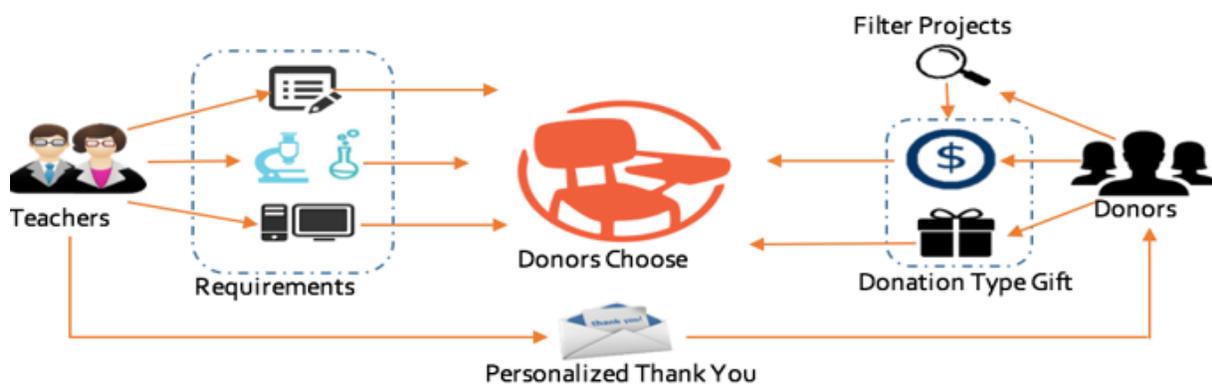


Figure 1: Process overview of how DonorsChoose.org

DATA UNDERSTANDING

DonorsChoose.org is a United States based 501 nonprofit organization that allows individuals to donate directly to public school classroom projects. **Figure 1** gives a process overview of DonorsChoose.org. The official website has a schema diagram that gives a brief understanding of the relationships among the datasets. There are five different datasets related to the Donorschoose.org. They are "Projects", "Essays", "Gift cards", "Resources" and "Donations". The "**Projects**" dataset contained school level, teacher level, project level and demographic information about the project and data about when a certain project was posted online, the completion date and the date when thank you packet was sent to the donors. It also provided the funding status of the projects. It contained around 900K records and 50 different variables. "**Donation**" dataset contained several details about donors such as demographics (city, state), behavior (donation message), and donation amount. Donation dataset had 5 Million observations and around 20 variables. "**Essays**" datasets contained details about the essays, which were posted by teachers regarding their projects. Essays dataset had 900K observations and 7 variables out of which project ID is the unique identifier and rest 6 variables were plain text. "**Resources**" dataset contained information regarding resources required for a certain project. "**Gift cards**" dataset contained information regarding the gift cards bought on the website along with the mode of payment, city-state and other details. Initial analysis suggested that most of the important information was contained in Projects, donations and essays datasets and hence greater emphasis was placed on working with these

three datasets. A field project ID was used as the unique identifier for joining the projects and essays dataset.

PROBLEM STATEMENT

It is important that DonorsChoose.org understands what Donors like donating to and their preferences. This in turn will help in getting a higher number of projects meet their funding goals. Literature review on Donorschoose.org dataset suggested that substantial amount of research has been done on ranking the projects that are posted on the website. There has been very little work done for building any kind of recommendation system. Exploratory research revealed that approximately 70% of the projects meet the funding goal where as 30% do not. There is a need to understand factors, which affect the funding status of projects. In order to build a recommendation system for DonorsChoose.org, the team came up with three-step process.

1. Analyze the impact of different factors on project funding and try to build a predictive model to find whether a project will get funding or not.
2. Attempt to understand the donors' preferences and needs to group them into clusters. Segmenting, profiling the existing donors helps us to understand the new donors as well
3. Target the specific donor segments with the predicted unfunded projects [from 1st part], so that number of funded projects increases.

By implementing this recommendation system, projects that are not likely to meet its funding goal can be recommended to Donors based on their preferences.

DATA CLEANING AND VALIDATION

The datasets were in .csv format. There were many punctuations marks in the text fields which included characters such as “,”,”/”,”//” and “.”. The issue was addressed with the help of UNIX shell scripting and SAS codes. A composite delimiter “\$|” and dummy indicator “<new _ line>” was included to maintain the data consistency. In order to convert.csv files into SAS datasets, DLMSTR='\$|’ option was used. After the SAS datasets were created, validation of data was done to confirm that none of the observations and variables were truncated. Many new variables were created in SAS datasets. The steps followed to convert the .csv files in SAS files can be viewed in **Appendix A: Data cleaning and validation part**.

Projects dataset: Exploratory analysis showed that 67.45% of the projects were completed, 28.83% had expired and the rest were into other (re-allocated and go-live. Basic descriptive statistics show that variables such as *students _ reached*, *total _ price _ excluding _ price* have an extremely high skewness and kurtosis values. Looking at the box plots and histogram of these variables, suggested that there were many outliers. For example, the highest value in *total_price_excluding_optional* was 10,250,017 whereas the 99th percentile was only 2,567.49. Therefore, the data was filtered and transformed in order to reduce the skewness and kurtosis to permissible values (closer to zero). **Essays dataset:** It contained many text variables. We made sure that the text was imported successfully into the SAS datasets. **Donations dataset:** This dataset contained data regarding donations by donors. Several variables such as *donation_amount* had to be cleaned after examining their distributions. **Resources and gift cards:** These datasets revealed information about resources for each project and various gift cards given. Several high performance procedures like HPIMPUTE, HPSUMMARY were used to clean donations, essays, resources, projects and gift cards datasets.

VISUALIZATION

The visualizations done in SAS studio can be seen in **Appendix B: Exploratory Analysis**

ANALYSIS

Analysis and approach to this problem statement is three fold.

1. Predict the projects which will not meet funding goals; involves the finding the factors responsible for predicting the project funding status and building a predictive model for the same.
2. Segment donors based on the preferences and behavior by looking at their past donations.
3. Suggest the unfunded projects to the donors who may like it and fund it.

The first part of the analysis involves the finding the factors responsible for predicting the project funding and building a predictive model. The step-by-step analysis done can be seen below in **figure 2**.



Figure 2: Steps followed to solve the problem statement

Predicting if a project meets its funding goal or not:

In order to predict if a project will reach its funding goal or not, we created a target variable called as '*funding_coded*'. This takes a value of '1' if a particular project meets its funding goal and takes a value of '0' if it is expired. There were four categories of funding status in the project. Completed, expired, live and re-allocated. We have a value of '1' for completed projects and a value of '0' for expired and re-allocated projects. The live projects were filtered and used as scoring dataset. Initial distribution of projects, which met its funding goal, and those, which expired, was approximately 70% and 30% respectively. Stratified sampling was done with 60% for training the model, 20% for validating the model and 20% for testing purpose. Prior to predictive modeling variables were imputed using various kinds of imputation techniques. Filtering and transformation techniques were used to deal with outliers, reducing skewness and kurtosis of different variables. Various kinds of variable selection and variable reduction techniques were used. Few important variables we choose from projects dataset were *resource_type*, *poverty_level*, *grade_level*, *number_of_students_reached* and *total_price_excluding_optional*.

PREDICTIVE MODELING

Various kinds of predictive models such as Logistic regression, Decision trees, Random forest, Neural Networks, Ensemble and Rule-based models were built in order to predict if a given project will meet its funding goal or not. Statistics such as ROC index and Misclassification rate were used to pick the best models. Input variables used were structured data from projects dataset, text clusters, text topics created from essays datasets and a combination of these together. Rule-based models were built using essays text field.

Dealing with text variables: Essays dataset contained many text fields such as *title*, *short_description*, *need statement* and *essays*. These are the text fields that are filled up by the teachers when they fill out the projects on donorschoose.org. The way these essays were written can affect if the project would reach its funding goal or not. Initially the dataset was parsed and interactive filtering was done to drop

few words and view concept links. We used the HPTMINE node in text miner to build text cluster and text topics. These text topics and clusters were used as input variables in the predictive modeling.

Predictive Modeling results: Analysis reveals that for every \$10 increase in cost of the project, the odds of project reaching funding goal is likely to reduce by 10%. The odds for grades level 6-8 are 12 % lower than the odds of pre k-2 schools. Whereas the odds of pre k-2 are 8% higher than the projects belonging to grades 3-5. The odds of project belonging to applied sciences primary subject is 22% higher than the odds of visual arts. The odds of projects belonging to literature & writing are 25% lower than the projects belonging to visual arts. Similarly, the odds of projects, which have primary subject of music, is 37% higher than that of visual arts. We observe that projects posted in summer have higher probability of project reaching funding goal that those posted in the rest of the year. The odds for projects posted in May are 20% higher compared to the projects posted in the month of December. The projects which have a teacher prefix of 'Mrs.' have 13 % lower odds compared to those which have a teacher prefix of 'Ms.'. Technology projects have a 60% lower odds compared to projects which have resource type as 'Books'. Projects belonging to schools from highest poverty area have 21 % higher odds compared to projects belonging to moderate poverty. **Modeling results can be seen in Appendix 4.4**

The best model turned out to be Neural Network with a Misclassification rate of 27.68% and a ROC index of 0.74. Model performance was verified in order to avoid any kind of over-fitting and under-fitting of models. In the final model the variables used were *resource_type*, *grade_level*, *poverty_level*, total cost of project excluding optional, number of students reached, month when project was posted, state to which the project belonged to and the cluster membership of various text variables in essays dataset.

Scoring the live projects: There are currently around 26,000 projects which are under live phase in projects dataset. The best model stated above, was used to score this dataset in order to predict if these projects will reach its funding goal or not. **The analysis results can be seen in appendix C.**

DONOR SEGMENTATION

The next step done was segmenting the existing donors from the Donors dataset. There were around 5 million records in the Donors dataset. Aggregation was done at a donor level using *Donor_ID* as the primary key and various new variables were created. There are approximately 1.6 million unique donors for DonorsChoose.org.

New variables created: Many new variables such as number of total donations, average time between each donation, number of times donor donated to projects in same state, number of times donor donated to each kind of grade level, poverty level, resource needed to projects etc. These variables were used as input variables for donor segmentation. All the necessary variable transformation for the clustering and segmentation is done and details are attached in the **appendix C**. Segmentation and segment profiling was done. The entire donors in Donations dataset were divided into 7 segments.

Preferences											
Segment	Grades 3 5	Grades 6 8	Grades 9 11	Grades PreK2	high poverty	highest poverty	books	supplies	technology	Literacy	Math science
1	7	433	8	8	6	137	36	124	114	43	152
2	131	106	95	65	115	90	460	7	6	222	4
3	68	105	75	149	372	5	35	165	62	52	147
4	297	12	10	16	9	144	38	168	60	51	136
5	6	8	485	7	35	139	36	104	133	36	135
6	182	29	19	121	155	32	8	6	254	165	65
7	10	8	7	334	8	152	87	126	71	131	63
Overall	100	100	100	100	100	100	100	100	100	100	100

Table 1: Table explaining the weighted importance of each variable with respect to its segments

Segments	Care about
Segment 1 [Middle school & math lovers]	Care more about 6th-8th graders; subject Maths & Science
Segment 2 [Books & literacy lovers]	Care only about book resources; subject literacy & languages
Segment 3 [Concerned about poverty & supplies]	Care more toward high level poverty; resource supplies and subject Maths & Science
Segment 4 [Upper elementary grade lovers]	Care more about 3rd-5th graders; resource supplies
Segment 5 [High school lovers]	Care more about 9th-11th graders; highest level poverty
Segment 6 [Technology lovers]	Care more about technolog resources; 3rd-5th graders
Segment 7 [Pre-kindergarten lovers]	Care mode about preK2 graders; highest level poverty

Table 2: Table showing the different segments and the values they care about

Details of the seven segments and the characteristics in each of these segments can be seen in **table 1** and **table 2**. Names were assigned to each of these segments in order to uniquely define them.

RECOMMENDATION SYSTEM

Approximately 75% of the Donors have donated only once. Therefore, there are not enough data points to create a recommender system on an individual donor level. That is the reason why segmentation of donors was done. Using the characteristics of the segments, of suitable projects were recommended. For example, consider this case: the donor [ID: 00000ce845c00cbf0686c992fc369df4] belongs to the Technology lovers' segments. Based on the results of the predictive models, a list of live projects, which are likely not going to meet the funding goal, was generated. These were filtered with the conditions based on the characteristics from Technology lover segments and sorted it by the probability of the project being unfunded. Then, based on the location where the project exists, recommendations were given to the donor when he or she login to DonorsChoose.org. For the donorID: 00000ce845c00cbf0686c992fc369df4 [Technology lover], the recommended live projects title is shown in the table. In this way, the donor does not have to put much effort to find the

Recommended Projects for Donor ID:00000ce845c00cbf0686c992fc369df4
Teaching Music Fundamentals with Technology!
Instant Learning On iPads
Technology for All!
ZSpace Virtual Reality: See it\, Feel it\, Build
Enter Your Response Please
Using Technology To Create A Paperless Classroom
Sometimes It's the Big Things
No More Textbooks!
An Apple a Day...
Chromebooks for Our Class
Chrome Books for Literacy
Promethean Board Promises Learning!

Table x: Table showing the recommended projects for a donor

project, which donor likes, and the projects, which are in the verge of unfunded, will get full funding. This is just an example of 12 projects recommended to the donor segment 6, who love donating to projects involving technology. This way we created a recommendation for all the seven segment of donors. A few projects might exist in one or more recommendation for segments. This way after the donor has donated one time, the next time he or she login to DonorsChoose.org, based on previous donation history, segmentation can be done as shown above and recommendation of projects can be done. This recommendation system works for donors who have done a prior donation. Before a donor donates for the first time, no prior account is created, nor is the donor asked to give any preferences. Hence, with the existing process in operation of the website, it is difficult to build a recommendation system for first time donors.

SUGGESTION FOR FUTURE STUDY

In the existing system, a recommendation cannot be given to donor who is about to donate for first time as there is no data about their preferences. Instead the system can be slightly modified such that prior to first donation, preferences of donor can be asked. Based on preferences, we can map the donor to one of the segments demonstrated above and a recommendation can be made. Using click stream data, the results of predictive as well as segmentation models used can be improved. There are groups of donors who donate only because they are teachers in the school, or their children study in the school, or they know the teacher personally, or the teacher is a relative. The donation message can be text mined to improve the segmentation results.

CONCLUSIONS

Wherever a transaction is done in online, everyone come across “You may like this”, “Users Who Viewed This Item Also Viewed” section. It makes the users’ life easier by showing the related products which they may like and help them in choosing the right product effectively. If this can be done for a monetary benefit in all domains, why can’t we do the same for a good cause as well? This project is a simple prototype for our idea. Using the best model, the live projects were scored. 1.6 million donors were segmented into seven segments, and based on the characteristics of each segment, recommendations were made from the set of live projects to each segments. These recommended projects match the likes, and preferences of donors based on previous donation data. If the donor has donated previously then this recommendation system can be used to recommend projects that donors like. This recommendation system gives a good experience to donors, by reducing the time taken to search for projects they like. By implementing this recommendation system, we can understand what the donor is looking at and what they like to donate for. It can be made better by adding the details mentioned in the scope for the future section. As demonstrated in this document, DonorsChoose.org website can be made better by implementing this recommendation system. It will not only help in donors having a better experience, but also can increase the percentage of projects that meet funding goal.

APPENDICES

APPENDIX A: DATA EXTRACTION AND CLEANING

1.1: Code used to extract and clean csv files and convert them into SAS files

UNIX shell scripting for data preparation.

We have used following scripts to address data issues of .csv files. We have converted the .csv files to text files.

We had a unique issue of presence of new line characters in the text columns. We replaced these new line characters with <new_line> to facilitate SAS dataset creation.

Additionally the comma delimiter is replaced by the composite delimiter \$|

```
//////////////////// Code begins for opendata_projects.csv //////////////////////
```

```
cat opendata_projects.csv | awk 'gsub(/\r/, "") {printf "%s" , $0;next} {print}' > new.txt
```

```
// this is to remove additional new line characters in the data
```

```
cat new.txt | awk '{ if (gsub(/\$/ , " ")) printf "%s", $0; else print }' > 1.txt
```

```
sed 's/\\\\\\\\\\\\\\\\/ <new_line> /g' 1.txt > 2.txt
```

```
// this replaces the comma delimiter by $| to address the parsing issue of any .csv files
```

```
sed 's/"/, "/g' 2.txt > 3.txt
```

```
sed '1s/\/, $|/g' 3.txt > opendata_projects.txt
```

```
//////////////////// Code ends //////////////////////
```

Similar code is used for modifying the other 6 datasets.

The Donation dataset has quantitative data pertaining to donations as well as the donations message sent by a donor while donating to a project. We have separated the textual information (donation message) from the quantitative data by creating two separate datasets as per below.

```
cut -d "," -f1-22 opendata_donations.txt > donations.txt
```

```
cut -d "," -f1,23 opendata_donations.txt > donations_message.txt
```

1.2: Code used to extract files from and push files into HDFS



```
SAS Studio
Program 1 x Program 2.sas x
CODE LOG RESULTS
1
2 proc hadoop username="herambjoshi" password="*****" verbose;
3 hdfs copyfromlocal = '/home/EC2.INTERNAL/herambjoshi/sasuser.v94/donations_pro_messages.sas7bdat'
4 out = '/contest/team06/data/heramb/donations_pro_messages.sas7bdat' overwrite;
5 run;
6
```

1.3: Code used to clean the datasets and convert variables into appropriate data types

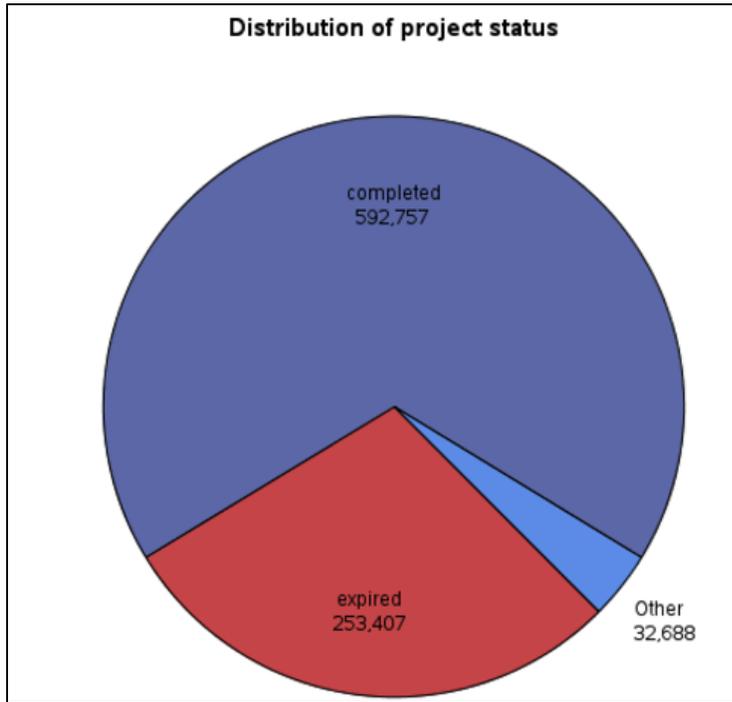
```
data new.projects_c;  
set '/home/EC2.INTERNAL/sandeepchittoor/sasuser.v94/projects_cleaned';  
_projectid = compress(_projectid , '');  
_teacher_acctid = compress(_teacher_acctid , '');  
_schoolid = compress(_schoolid , '');  
school_ncesid = compress(school_ncesid , '');  
school_latitude = compress(school_latitude , '');
```

```
data '/home/EC2.INTERNAL/vigneshdhanabal/sasuser.v94/projects_c_new';  
set '/home/EC2.INTERNAL/sandeepchittoor/sasuser.v94/projects_c';  
day_dp = input(substr(date_posted, 9,2),2.);  
month_dp = input(substr(date_posted, 6,2),2.);  
year_dp = input(substr(date_posted, 1,4),4.);  
date_post = mdy(month_dp, day_dp, year_dp);  
date_post = input(date_post, 20.);  
format date_post date9.;
```

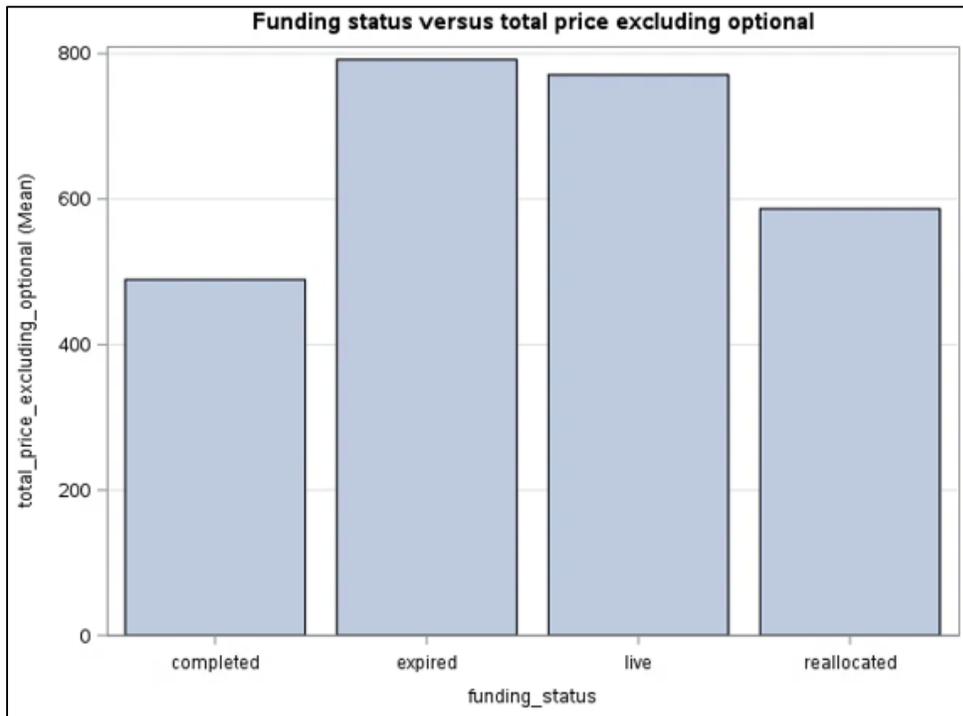
```
data new.projects_c;  
set '/home/EC2.INTERNAL/sandeepchittoor/sasuser.v94/projects_cleaned';  
_projectid = compress(_projectid , '');  
_teacher_acctid = compress(_teacher_acctid , '');  
_schoolid = compress(_schoolid , '');  
school_ncesid = compress(school_ncesid , '');  
school_latitude = compress(school_latitude , '');
```

APPENDIX B: EXPLORATORY ANALYSIS

2.1: Distribution of project status in projects dataset



2.2: Average total price excluding optional for different project status



2.3: Cross tab between the *resource_type* needed for project and the project funding status

Frequency Expected Row Pct Col Pct	Table of funding_status by resource_type						
	funding_status	resource_type					
		Books	Other	Supplies	Technology	Trips	Visitors
completed	133558	63817	213331	174092	6880	1051	592727
	120778	64537	202908	197088	6356.2	1082.5	
	22.53	10.77	35.99	29.37	1.16	0.18	
	74.58	66.69	70.91	59.58	73.01	65.48	
expired	39714	27373	75765	107714	2319	503	253388
	51632	27589	86742	84245	2717.3	462.77	
	15.67	10.80	29.90	42.51	0.92	0.20	
	22.18	28.61	25.18	36.87	24.61	31.34	
live	4100	3630	8816	7827	141	45	24559
	5004.3	2674	8407.2	8165.2	263.36	44.853	
	16.69	14.78	35.90	31.87	0.57	0.18	
	2.29	3.79	2.93	2.68	1.50	2.80	
reallocated	1701	866	2926	2546	84	6	8129
	1656.4	885.1	2782.8	2702.7	87.173	14.846	
	20.93	10.65	35.99	31.32	1.03	0.07	
	0.95	0.91	0.97	0.87	0.89	0.37	
Total	179071	95666	300838	292179	9424	1605	878803
Frequency Missing = 49							

2.4: One-way frequencies of different secondary focus area in the project

The FREQ Procedure

secondary_focus_area	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Applied Learning	86438	14.20	86438	14.20
Health & Sports	21658	3.56	108096	17.76
History & Civics	47489	7.80	155585	25.56
Literacy & Language	214862	35.30	370447	60.86
Math & Science	155947	25.62	526394	86.48
Music & The Arts	45402	7.46	571796	93.94
Special Needs	36894	6.06	608690	100.00
Frequency Missing = 270162				

2.5: Summary of missing values of different categorical variables in projects dataset

school_state	Frequency	Percent
Non-missing	878852	100.00%

Primary_focus_area	Frequency	Percent
Missing	42	0.00%
Non-missing	878852	100.00%

Secondary_focus_area	Frequency	Percent
Missing	270162	30.74%
Non-missing	608690	69.26%

poverty_level	Frequency	Percent
Non-missing	878852	100.00%

Grade_level	Frequency	Percent
Missing	44	0.01%
Non-missing	878808	99.99%

2.6: Summary statistics of different continuous variables in dataset

Variable	Mean	Std Dev	Minimum	Maximum	N
total_donations_	396.8973764	495.4140851	0	9999.00	878852
num_donors_	4.3579135	6.2881867	0	848.0000000	878852
payment_processing_charges_	7.9915385	17.5595281	0	9715.00	843771
total_price_excluding_optional	585.2855661	11034.18	0	10250017.00	878852
total_price_including_optional	693.8735535	13453.73	0	12500020.70	878852
sales_tax_	18.1459315	54.8642519	0	13797.00	843771
vendor_shipping_charges_	16.7857364	61.5523265	0	38861.00	843771

Variable	Mean	Std Dev	Minimum	Maximum	N	Skewness	Kurtosis	99th Percentile
students_reached	0	800	0	9999999	878700	476.79	232348.75	800
total_price_excluding_optional	0	2600	0	10250017	878852	912.88	847160.48	2567.49

APPENDIX C: PREDICTIVE MODELING AND SEGMENTATION RESULTS

3.1: Data partition result for modeling purpose (60:20:20)

Summary Statistics for Class Targets					
Data=DATA					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
funding_coded	.	0	261536	30.6144	
funding_coded	.	1	592756	69.3856	
Data=TEST					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
funding_coded	.	0	52308	30.6145	
funding_coded	.	1	118552	69.3855	
Data=TRAIN					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
funding_coded	.	0	156921	30.6143	
funding_coded	.	1	355653	69.3857	

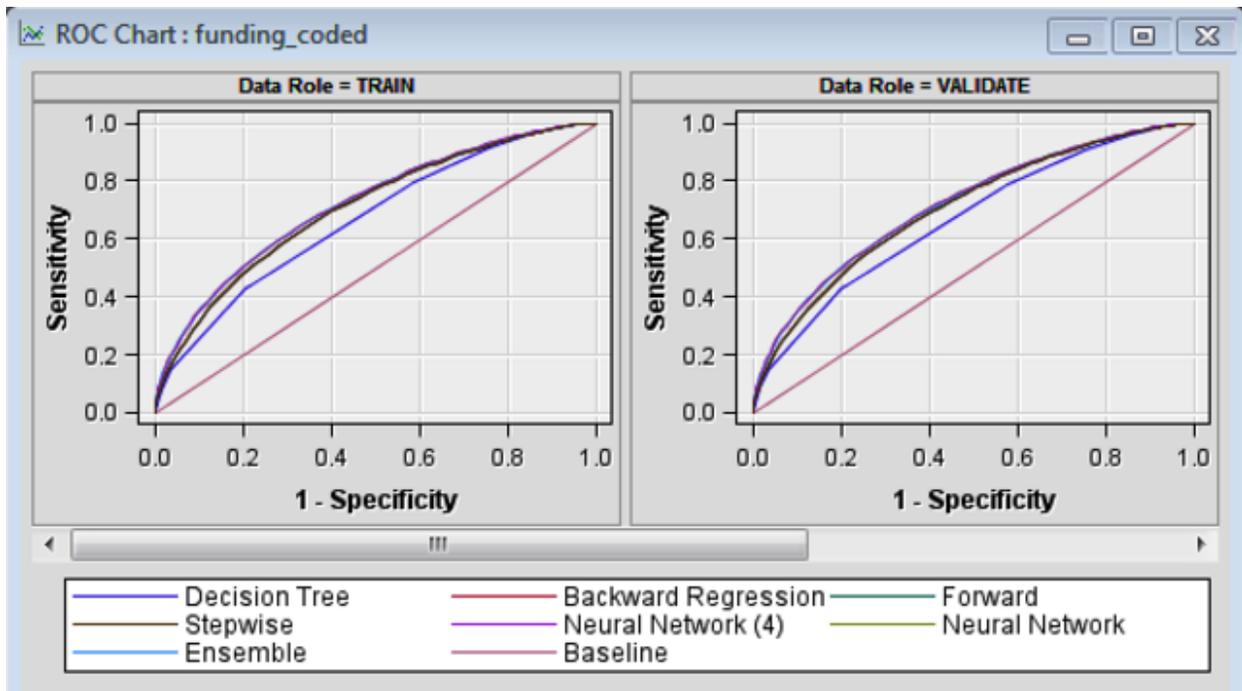
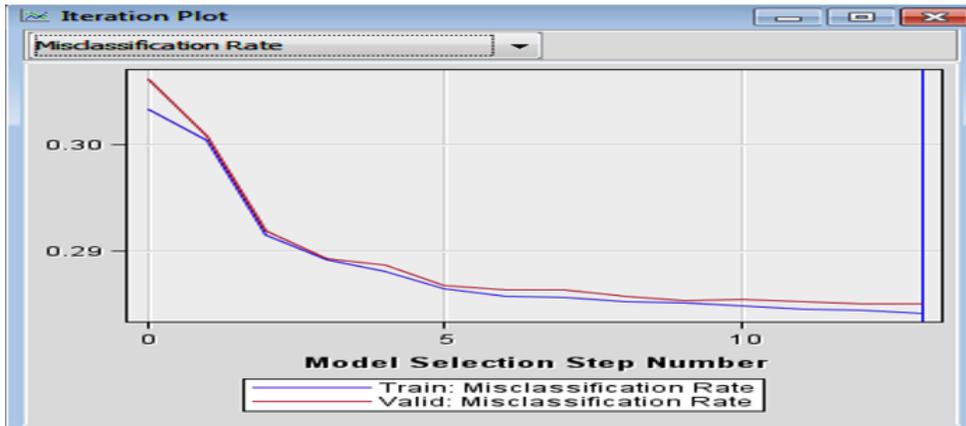
3.2: Imputation results of different categorical variables

Variable name	Imputation method	Impute Value
grade_level	Count	Grades PreK-2
primary_focus_subject	Count	Literacy
resource_type	Count	Supplies
school_metro	Count	urban
teacher_prefix	Count	Mrs.

3.3: Filtering bounds for different continuous variables

Variable	Minimum	Maximum
students_reached	0	800
total_price_excluding_optional	0	2600

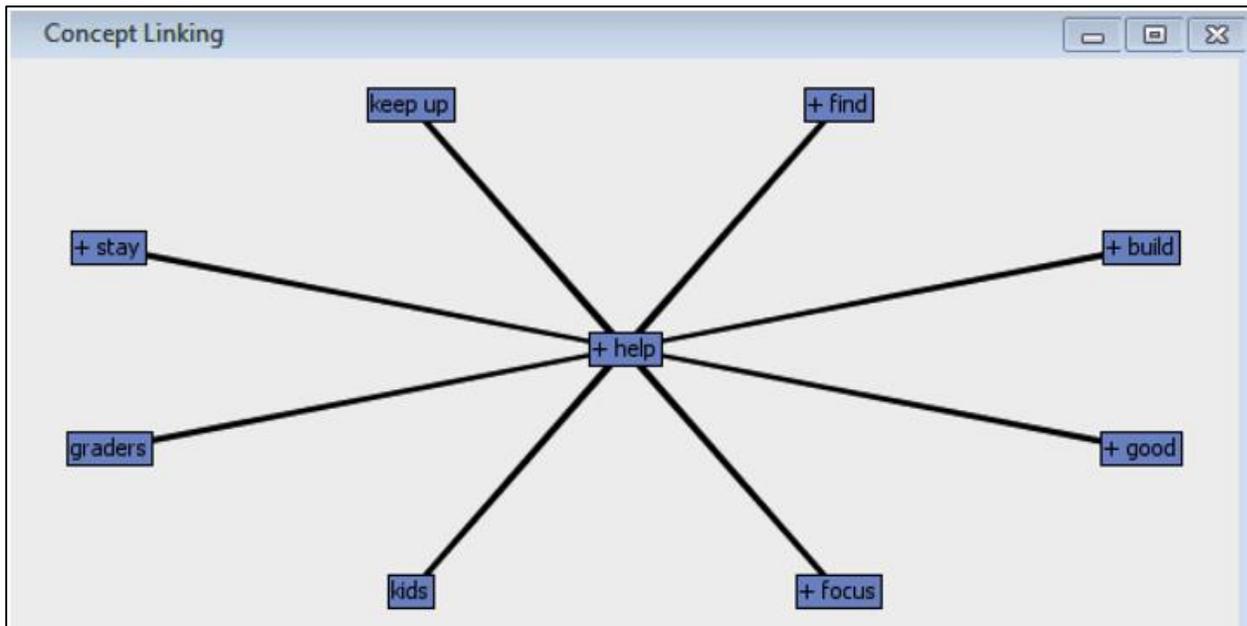
3.4: iteration plot and ROC index of logistic regression, used to predict if a given project meets funding goal



3.5: Model comparison of different models built

Input variables used and modeling technique	Misclassification rate	ROC INDEX
Rule-based models(Essay as input field)	35.92%	
Sturctured variables only (Best model: Neural Network)	28.95%	0.66
Clusters membership only (Neural Network)	29.58%	0.67
Text Topics only (Generated from title, short description,essays)	31.12%	0.63
Structured variables and text topics together (Best model:Decision tree)	30.30%	0.64
Structured variables and cluster membership together (Best model: Neural networks)	27.68%	0.74

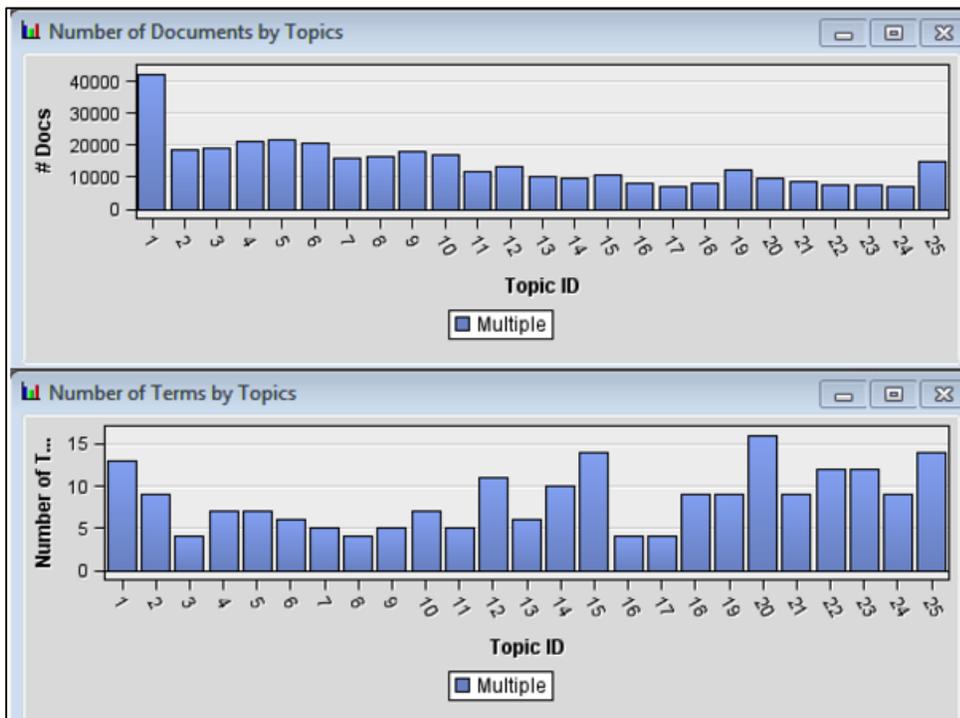
3.6: Concept links for word help present in essays variable of essays dataset



3.7: Different words present in text topics built for essays text variable in essays dataset

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	# Docs	Number of Terms
Multiple	1	0.132	0.018	+help,+build,+organize graders,+hear	42365	13
Multiple	2	0.101	0.018	+learning,+help,english,+play,+read	18531	9
Multiple	3	0.103	0.018	+technology,+classroom,+century,+book,+enhance	18844	4
Multiple	4	0.104	0.018	+book,+good,+good book,+big,+big book	21000	7
Multiple	5	0.101	0.018	+read,+ready,+succeed,+extra,+set	21456	7
Multiple	6	0.102	0.018	+classroom,+help,+create,+computer,+library	20401	6
Multiple	7	0.092	0.018	+listen,+center,+learn,+help,+learning	16001	5
Multiple	8	0.095	0.018	+math,hands-on,+hands-on math,+master,+math center	16343	4
Multiple	9	0.093	0.018	+reading,+guide,+rug,+classroom,+cozy	17915	5
Multiple	10	0.085	0.018	+center,+kindergarten,+literacy center,+math center,+classroom	16970	7
Multiple	11	0.081	0.018	+science,hands-on,+learning,+hands-on science,+stem	11702	5
Multiple	12	0.081	0.018	+learn,+play,+learning,+ready,+place	13591	11
Multiple	13	0.072	0.018	+write,+right,+stuff,+learning,+help	9993	6
Multiple	14	0.071	0.018	+bring,+century,+help,+music,+history	9707	10
Multiple	15	0.070	0.018	+create,readers,+century,+help,+building	10926	14
Multiple	16	0.068	0.018	+art,+music,+art supply,+supplies,+smart	8024	4
Multiple	17	0.068	0.018	+time,+keep,+help,kids,+learning	7212	4
Multiple	18	0.069	0.018	+love,+read,+kindergarten,+inspire,+foster	7935	9
Multiple	19	0.068	0.018	+world,+color,+explore,+open,+day	12169	9
Multiple	20	0.067	0.018	students,engaging,+inspire,+help,+keep	9898	16
Multiple	21	0.066	0.018	+literacy,financial,+financial literacy,+literacy center,+music	8598	9
Multiple	22	0.066	0.018	+want,+hear,+know,readers,+play	7743	12
Multiple	23	0.066	0.018	+building,readers,+library,+future,+stem	7588	12
Multiple	24	0.066	0.018	fun,+games,+math,+help,+classroom	7017	9
Multiple	25	0.064	0.018	+future,+project,+teaching,+stem,+kindergarten	14899	14

3.8: Number of documents by topics and number of terms by topics analysis



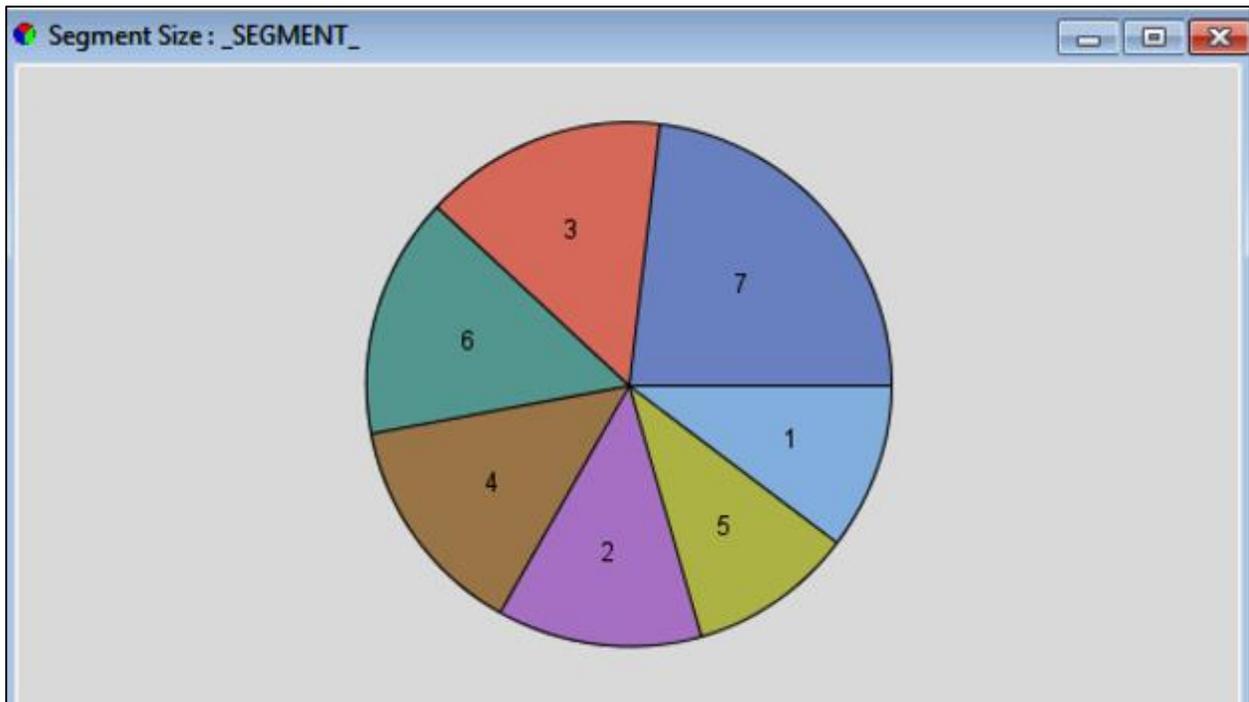
3.9: Different rules for rule-based models built to predict the funding status of different projects

Target Value	Rule #	Rule	Precision	Recall	F1 score	Valid Precision	Valid Recall	Valid F1 score	True Positive/Total	Valid True Positive/Total
1	1	tornado & destroy & start	98.97%	0.03%	0.05%	91.67%	0.03%	0.06%	96/97	33/36
1	2	soil & compost & ~mobile & environmental	99.34%	0.04%	0.08%	88.33%	0.04%	0.09%	54/54	20/24
1	3	year & ~technology & ~camera & ~rural & ~sil & ~center & ~organize & ...	99.49%	0.06%	0.11%	88.06%	0.05%	0.10%	46/46	6/7
1	4	unit & ~storage & ~technology & ~digital & ~rural & ~video & holocaust ...	99.57%	0.06%	0.13%	87.18%	0.06%	0.11%	35/35	9/11
1	5	soil & worm & school garden	99.63%	0.08%	0.15%	88.64%	0.07%	0.13%	40/40	10/10
1	6	literature & ~technology & ~center & ~rural & english & ~fluency & ~inte...	99.67%	0.08%	0.17%	86.67%	0.08%	0.15%	32/32	13/17
1	7	sharpener & ~technology & sharpen & ~center & mathematician	99.70%	0.09%	0.18%	86.61%	0.08%	0.16%	30/30	6/7
1	8	novel & ~technology & graphic novel & engage book	99.47%	0.10%	0.21%	85.71%	0.09%	0.18%	44/45	11/14
1	9	bully & ~technology & high & city	98.71%	0.13%	0.26%	86.45%	0.11%	0.23%	85/89	26/29
1	10	engaging & ~technology & ~center & ~camera & ~chair & independent ...	98.82%	0.14%	0.28%	86.31%	0.12%	0.24%	45/45	12/14
1	11	understanding & ~technology & ~camera & ~tablet & school garden	98.45%	0.16%	0.32%	86.32%	0.14%	0.28%	70/73	19/22
1	12	continue & ~technology & ~camera & ~center & ~rural & last & ~century ...	98.52%	0.17%	0.34%	85.65%	0.15%	0.30%	27/27	15/19
1	13	immigrant & ~technology & ~lcd & ~storage & ~website & ~desk & sha...	98.58%	0.18%	0.35%	85.52%	0.16%	0.32%	27/27	10/12
1	14	immigrant & ~technology & ~lcd & ~storage & ~website & life & public	98.53%	0.19%	0.38%	85.11%	0.17%	0.34%	49/50	11/14
1	15	neighborhood & ~technology & ~camera & bully	97.30%	0.24%	0.48%	84.75%	0.21%	0.42%	197/211	51/61

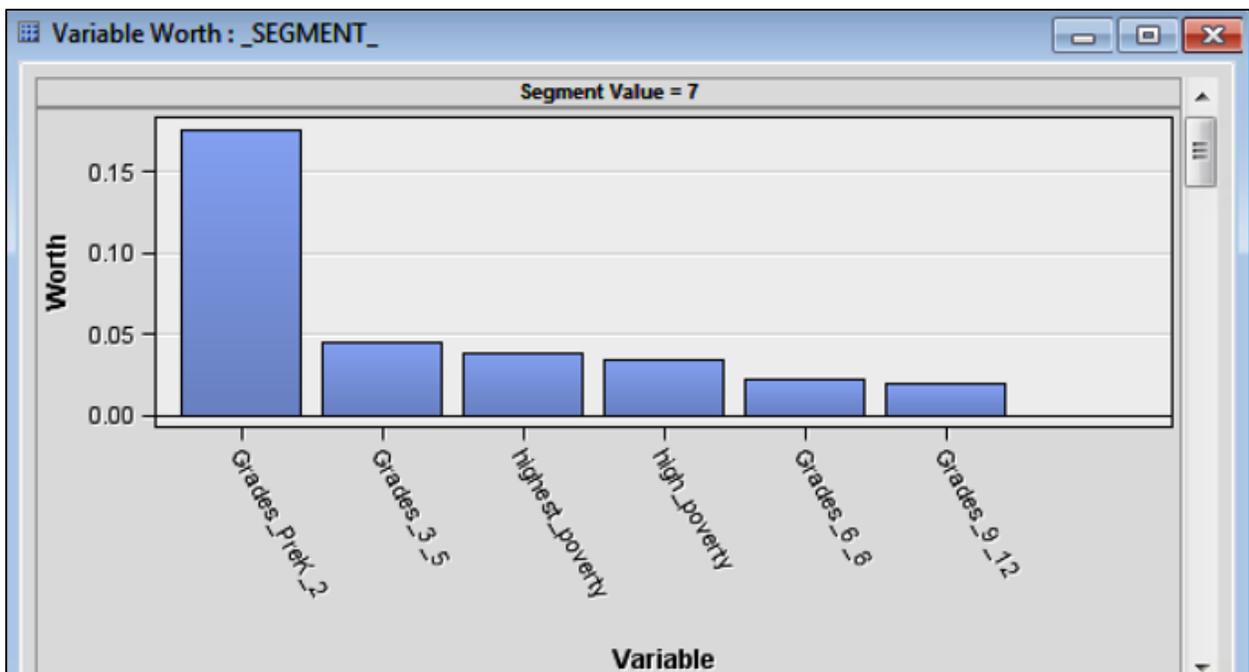
4.0: Fit statistics of rule-based models

Fit Statistics				
Target=funding_coded Target Label=' 1 '				
Fit				
Statistics	Statistics Label	Train	Validation	Test
<u>_ASE_</u>	Average Squared Error	0.07	0.07	0.07
<u>_DIV_</u>	Divisor for ASE	1025148.00	341716.00	341720.00
<u>_MAX_</u>	Maximum Absolute Error	0.60	0.60	0.60
<u>_NOBS_</u>	Sum of Frequencies	512574.00	170858.00	170860.00
<u>_RASE_</u>	Root Average Squared Error	0.27	0.27	0.27
<u>_SSE_</u>	Sum of Squared Errors	74716.66	24988.47	24948.73
<u>_DISF_</u>	Frequency of Classified Cases	512574.00	170858.00	170860.00
<u>_MISC_</u>	Misclassification Rate	0.36	0.37	0.37
<u>_WRONG_</u>	Number of Wrong Classifications	184313.00	63474.00	63735.00

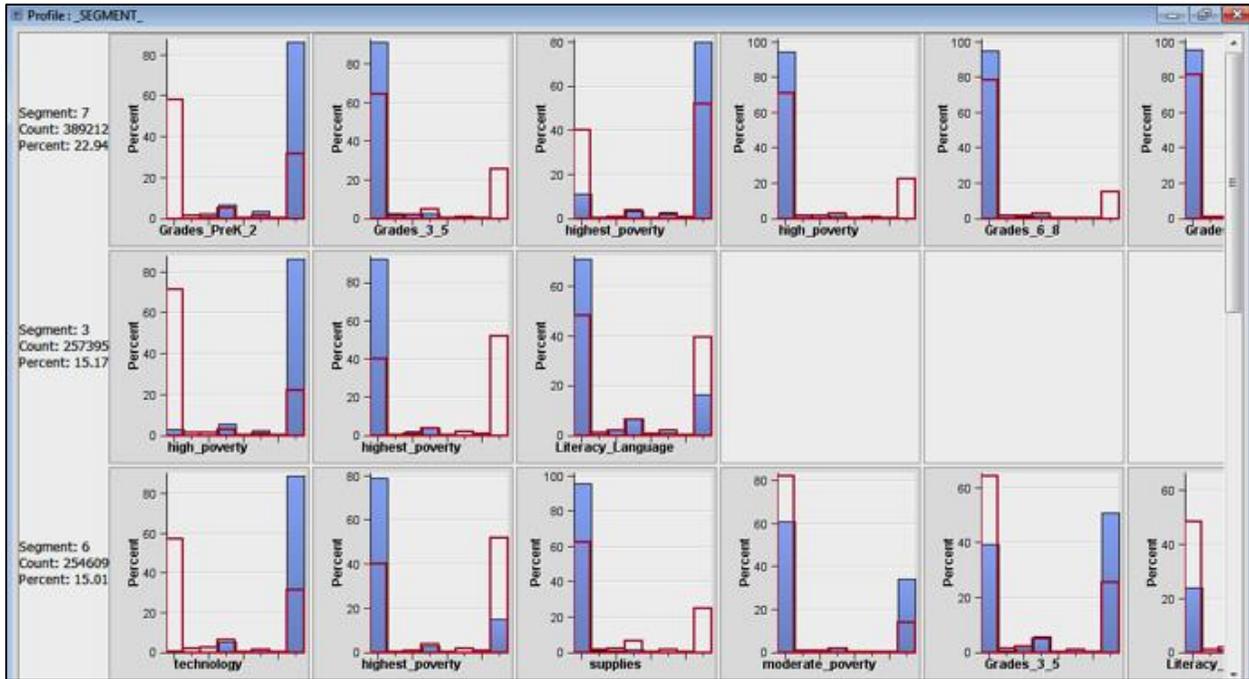
4.1: Segmentation of donors output



4.2: Variable worth for of different variables in various segments of donors



4.3: Donor segment profiling output



4.4: SAS Enterprise Miner modeling diagram

