Vishal Kumar
Data Science Career Track
April 4, 2018

Contents

1. Introduction

DonorsChoose.org is a United States–based, 501(c)(3), 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 crowdfunding platforms of its kind. The organization has been given Charity Navigator's highest rating every year since 2005. In January 2018, they announced that 1 million projects had been funded. In 77% of public schools in the United States, at least one project has been requested on DonorsChoose.org. Schools from wealthy areas are more likely to make technology requests, while schools from less affluent areas are more likely to request basic supplies. It's been noted that repeat donors on DonorsChoose typically donate to projects they have no prior relationship with, and most often fund projects serving financially challenged students.

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.

In 2003, the Oprah Winfrey Show featured Charles Best in a segment on Innovative Teachers, which aired on June 20. Traffic generated from the show crashed the site, but viewers donated $250,000 to classroom projects. The sustained exposure from the Oprah Winfrey Show allowed DonorsChoose.org to expand from the New York City region to key metropolitan areas across the country.

In 2006, following the destruction of hurricanes Katrina and Rita, the site opened to public school teachers in Louisiana, Mississippi, Alabama, and Texas. In 2007, the site opened to every public school in the United States.

As of September 2016, donors have contributed over $400 million to fund more than 750,000 classroom projects posted on the site, reaching 19 million students in public schools across the United States

1.1 Problem Definition

DonorsChoose.org receives hundreds of thousands of project proposals each year for classroom projects in need of funding. Right now, a large number of volunteers are needed to manually screen each submission before it's approved to be posted on the DonorsChoose.org website.

Next year, DonorsChoose.org expects to receive close to 500,000 project proposals. As a result, there are three main problems they need to solve:

1. How to scale current manual processes and resources to screen 500,000 projects so that they can be posted as quickly and as efficiently as possible
2. How to increase the consistency of project vetting across different volunteers to improve the experience for teachers

3. How to focus volunteer time on the applications that need the most assistance

1.2 Potential Clients

This model can be used by any fund raiser companies as a filter to select certain qualifying projects to be posted on the internet. For example, online fund raiser companies would get hundreds of request every day and they must analyze projects importance and need, before they can be posted on the website, this model will help them reduce the time spent on analyzing these project manually.

Causes, Crowdrise, DonateNow/Network for Good, FirstGiving, FundRazr are few companies which can make use of this model with little bit of update with the model.

This model can also be used by public school teachers across the U.S. to prepare better proposal and submit it to the DonorsChoose.org.

2. Data Acquisition

The dataset has been downloaded from kaggle.com (link is given below). This is a kaggle competition dataset for DonorsChoose.org. We found this dataset interesting because of Natural Language Processing (NLP) involved in modeling.

The dataset contains information from teachers' project applications to DonorsChoose.org including teacher attributes, school attributes, and the project proposals including application essays.

https://www.kaggle.com/c/donorschoose-application-screening/data

File descriptions

- train.csv - the training set
- test.csv - the test set, this is the dataset where we do not have the known target value.
- resources.csv - resources requested by each proposal; joins with test.csv and train.csv on id
- sample_submission.csv - a sample submission file in the correct format

3. Data Wrangling and Cleaning

Generally we do not get the data in the format best suited for our problems. We always need to do some data cleaning and wrangling to bring the data to the desired format. Since this is an ongoing competition dataset, the dataset is pretty much clean. But to do some of the analysis we have to wrangle some of the data like time series and text based analysis.
Most of the dataset is clean but we did find some missing values. NaNs, and there were some NaNs which got generated from calculations. These values dealt with ***fillna()*** method accordingly to the situations.

- Teacher's prefix column has 4 NaN values, which were filled with "Teacher" prefix which was more suitable.
- During calculation of price range, there were price ranges in which there were no submission, for calculation of approval rate these values got NaN values, I filled it with 0 (zero).
- The 3rd instance of NaN values were due to process change while filing the application. The 4 essay parts became only 2. Hence NaN values in columns essay_3 and essay_4. We have combined all 4 parts to just 2 parts according to the given definitions of all the 4 essays.

Data Wrangling

Here, will discuss more about the data wrangling steps I have taken to perform my analysis. Some of the techniques have been used throughout the notebook.

- During analysis based on gender of the teachers, I had to combine doctors and teachers as unknown genders, hence categorizing the entire dataset into three categories, Female, Male and Not Known.
- I have used ***split()*** function on project subject category and project subject sub category. By doing this we have reduced the number of unique subjects to 9 and unique subject sub categories to 30, which was easy to handle and made more sense for the analysis.
- Time Series analysis, we derived 5 new columns using ***datetime*** data type of python. These new columns are date, month, year, day of the week, time. This helped me analyze which time during a year has the highest number of submission and what time it remains relatively dull. Then I divided the day into 4 categories as below and grouped the data accordingly.
  - 06:00 - 11:59 Morning
  - 12:00 - 15:59 Afternoon
  - 16:00 - 19:59 Evening
  - 20:00 - 05:59 Night(nights are long)
- To create a price range, I have used ***itertool*** package along with ***tee, islice, chain*** of python. I have also used a special pandas method ***pandas.cut()***. The ***cut*** function is useful for going from a continuous variable to a categorical variable.
- For text and word cloud analysis, I have used some basic string operations of python like ***.join() and split()*** method.

Apart from this, ***round() and lambda()*** methods have been used in almost all the section to bring the floating point data in to 2 decimal places, which is more readable in percentage calculation.

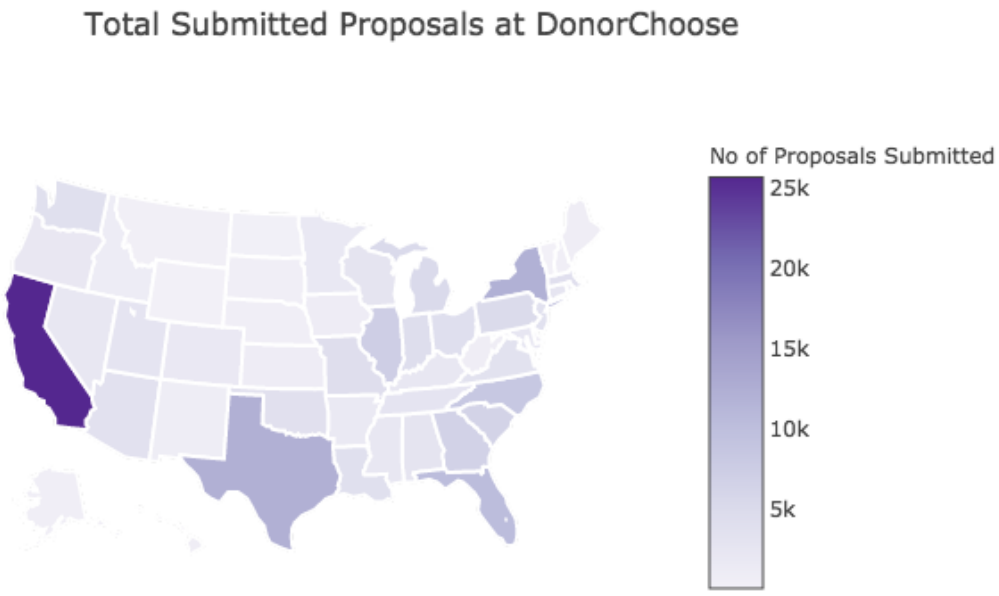
4. Data Exploration

As per the training dataset, the overall approval rate of a project, going to be posted on DonorsChoose.org is at **84.77%** which is very good. It means that any teacher's proposal at the time of submission has **85%** chance that his proposal will be approved to be posted on the website.

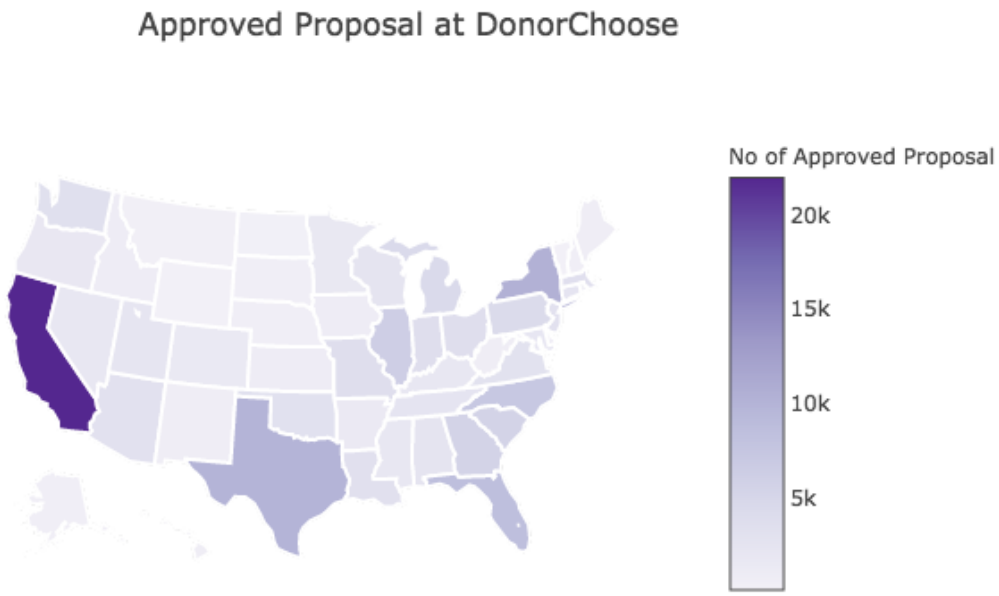
Now, let's see the individual features and their approval rates based as per the given dataset.

4.1 States wise

Since the DonorsChoose organization is spread all over the U.S., we have analyzed the training dataset to find which of the states have submitted most number of proposals to the DonorsChoose and what the approval rate for those states is. Let's check out the number of submissions first.



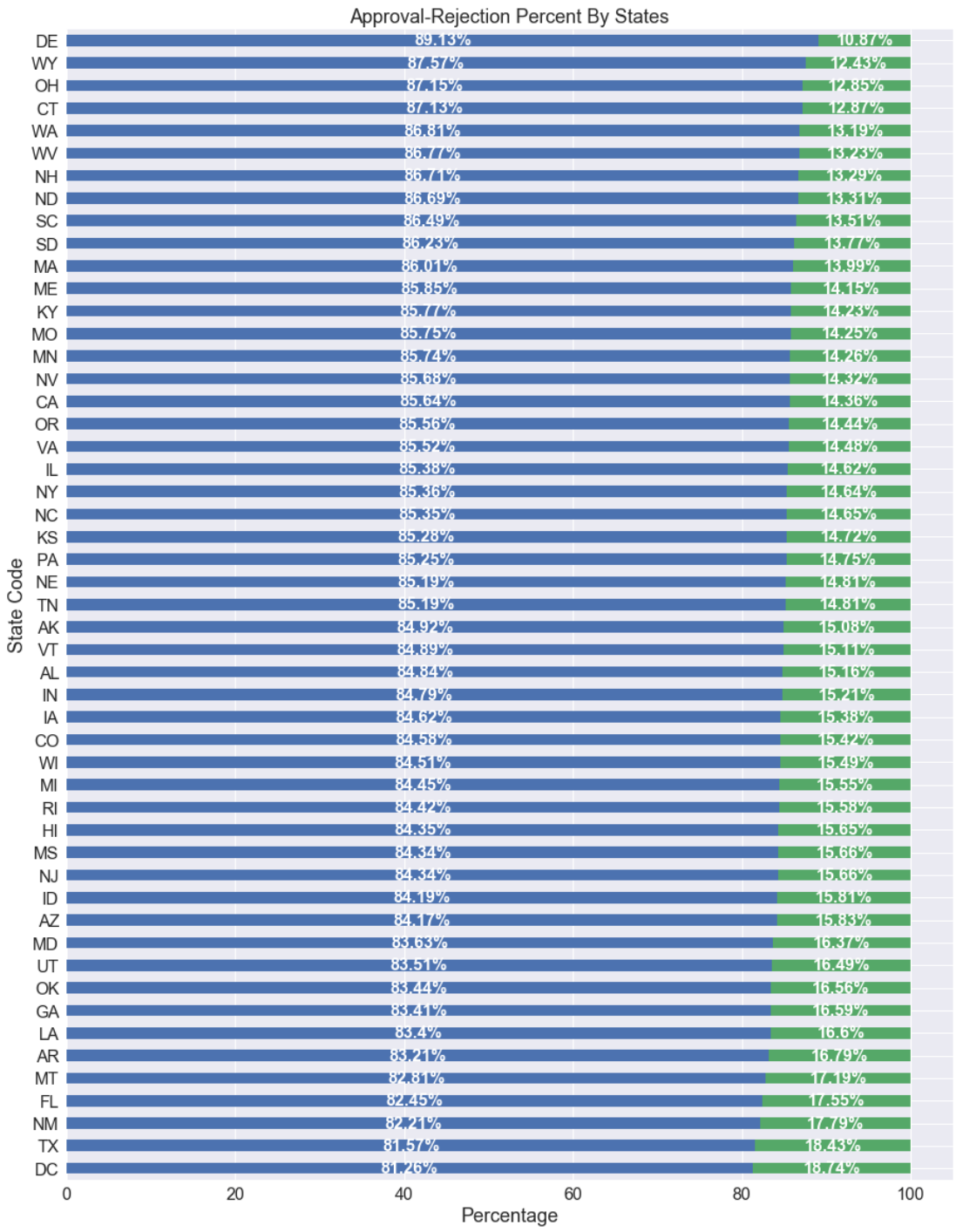
As we can see on the above map that **California** has the highest number of submissions in the U.S. **Texas, New York**, and **Florida** are on 2nd, 3rd and 4th position respectively. Let's check out the approval rates of the states.



California also tops the list of most number of approved projects, New York and Texas to follow. So, when the number of submissions is high, the number of approved proposals is also high.

Below are some of the interesting findings related to the approval rates of all the states.

- All the states have more than **81%** of approvals of the total proposals submitted to DonorsChoose.org
- **Washington D.C, Texas** and **New Mexico** have slightly higher rejections compared to other states.
- A very small state in east, **Delaware,** has the highest approval rate followed by **Wyoming** and **Ohio.**
- We don't know the number of public schools available in each state, we are assuming that states with bigger geographical size and relatively large population will have more number of public schools and hence will see more number of submissions in those states.
- But the approval-rejection rate of each state would be different.



As we can see from the above chart that California may have large number of submissions. Hence, the approval but the approval-rejection rate is not better than Delaware a small state in east.

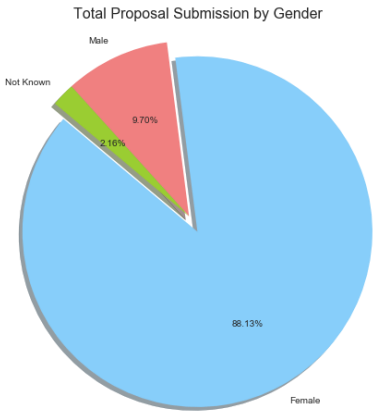
4.2 Teacher's Gender wise

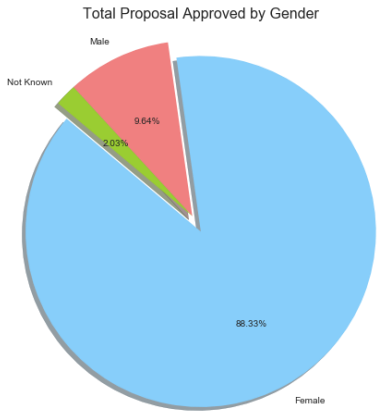
The approval rate by gender of the teacher and the relationship between male and female approval rate. We will also see, if the gender of the teacher is not known then the proposal is approved or not.

Table 1

Gender	Total	Approved	Approval Rate
Female	160471	136338	84.96
Male	17667	14876	84.2
Not Known	3942	3132	79.45

Female category has the highest number of submission and hence approved proposals from Female teachers are very high. Proposal submitted by **female** teachers **88%** compared to male which is only **10%** of the total submission. **Male** and **Female** approval rate is almost same around **84%** and **85%** respectively. Even teachers with not know gender has a very good approval rate of almost **80%.** There are around **2%**of teachers who can't be identified as male or female.

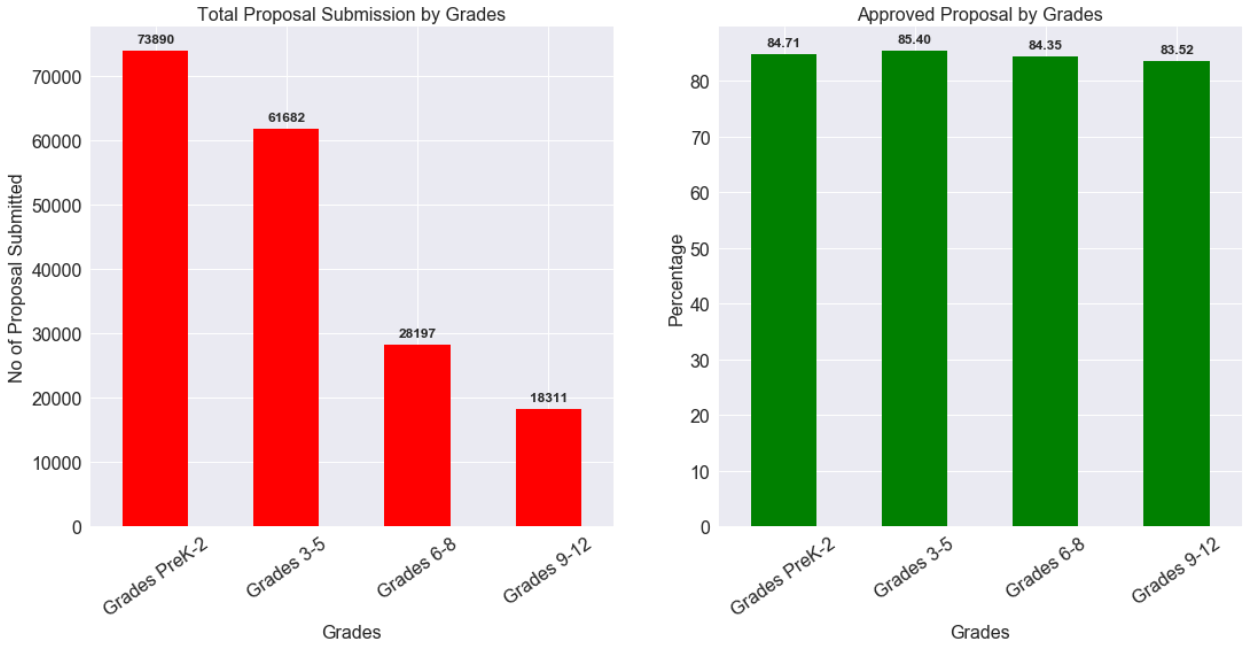




As we can see the dominance of female teacher in submission of proposal and hence the number of approvals is also very high, however the approval rate for male and female is similar and the approval rate of all the unknown gender is also very impressive at 80%. So we can safely say the approval doesn't depend on the gender of the teacher.

4.3 Grade Wise

Now it's time to check the approval rate of proposals by the grades, meaning which grade has the highest number of proposals and approvals.

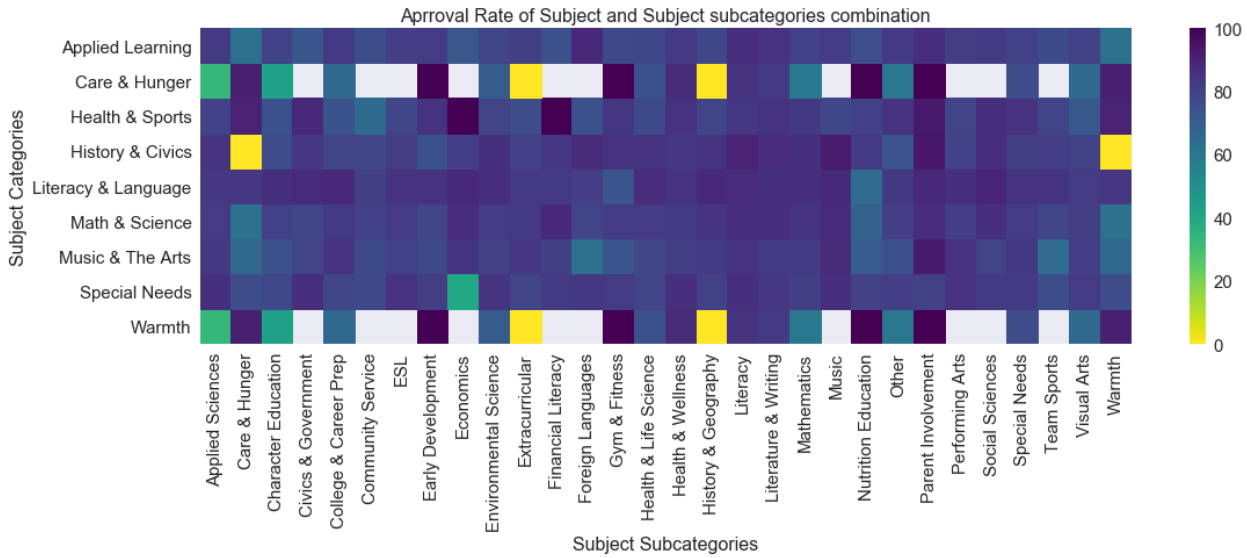


The proposal submissions for grades **Prep-2** and grades **3-5** is pretty high compared to grades **6-8** and grades **9-12**. However the approval rate for each of the grade category is between **83%-85%**, which is good. Grades also seem to have no effect on the approval of the proposals.

4.4 Project Subject and Subject Sub-Category wise

The number of unique subject category and sub category is very big. So it was important for us to bring this down to some manageable numbers as there were combinations of repeated **subjects** and **subject subcategories** and there are repetitions of some of the subjects with different combination of other subject sub- category. For example we found that '**Music & The Arts**' has been repeated twice but with other subject combinations like '**Music & The Arts, Warmth, Care & Hunger'**, and again with '**Music & The Arts, History & Civics**'. So we want to get these unique subjects just once.

Then we can check out the approval rate of the subjects and subject sub-categories.



Subjects like **Care & Hunger** and **Warmth** have 0% approval rate for some of the sub-categorical subjects, that's because there is no submission in the sub-categorical subject. **Applied Learning, Literacy & Language, Maths & Science** subjects have better approval rates ranging from **60% to 80%.**

4.5 Time Series

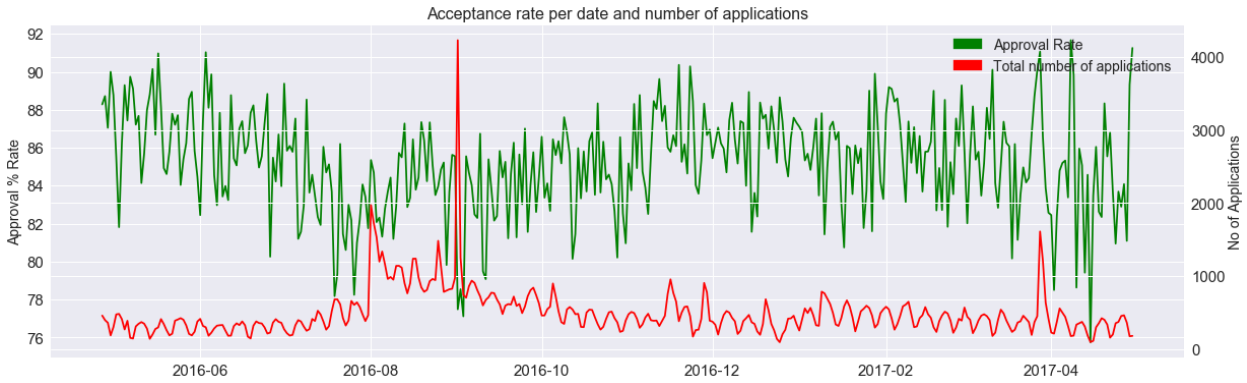
In this time series analysis, we have done analysis on month that has the highest number of project proposals submitted, day of the week that has the highest number of submissions, period of the day that teachers prefer submission of project proposals and the approval rate of those mentioned above.

Created a range of the day, i.e. decide hours for morning, afternoon, evening and night. I came up with the below

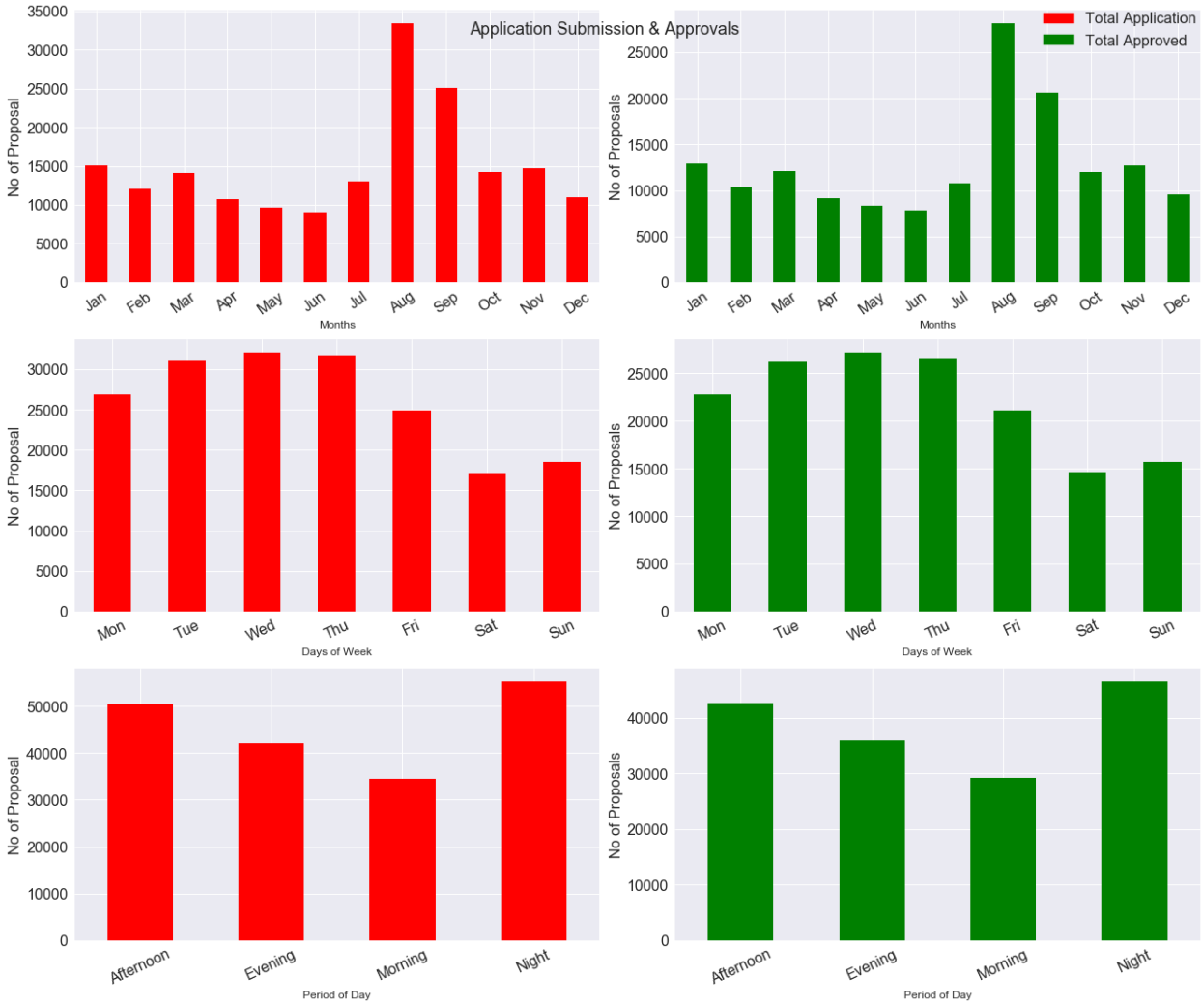
- 06:00 - 11:59 Morning
- 12:00 - 15:59 Afternoon
- 16:00 - 19:59 Evening
- 20:00 - 05:59 Night(nights are long)

This categorization was used to find out which period of the day teachers submit their proposal. And then we will have analysis of the approval rate for the same. I would be interesting to see month, day of week and period of the day a teacher prefer to submit the proposals.

So let's start by checking the approval rate of each day for the entire dataset.



As we can see the high peaks near the month of **August and September** and then it falls back to the normal baseline is due to the fact that this time range is the start of the school year. Hence this time of the year more project proposals are submitted. Approval rates are ranging from low to high as **75%** to nearly **92%.** Now let's check the number of months, day of week and period of the day submission and approval rate.



April, May and June months are dull; the DonorsChoose.org does not get large number of proposals compared to other months.

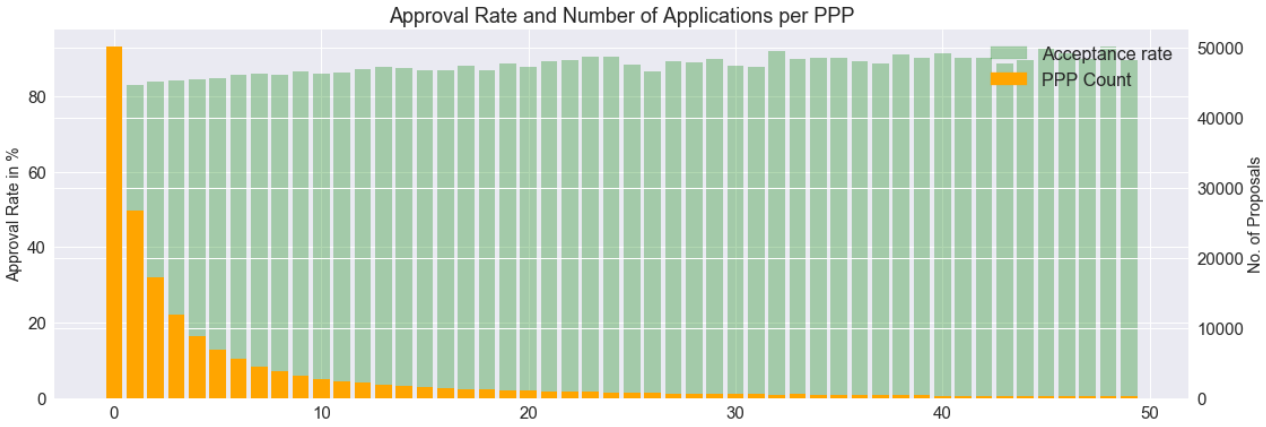
Tuesdays, Wednesday, Thursday are most preferred days of the week to submit the proposal. In fact slightly more than **50%** proposals are submitted on these 3 days. **Saturdays and Sundays** are the least preferred day of the week to submit a new proposal.

Most teachers choose **night time** (after 8:00 Pm and before 06:00 Am) to submit a proposal. Morning time from 6am to 12 noon is the least chosen period of day to submit a proposal.

Number of approved proposals is in the same order as of the total number of submissions. The approval rate is around **80%-85%** in all categories.

4.6 Previously Posted Proposals wise

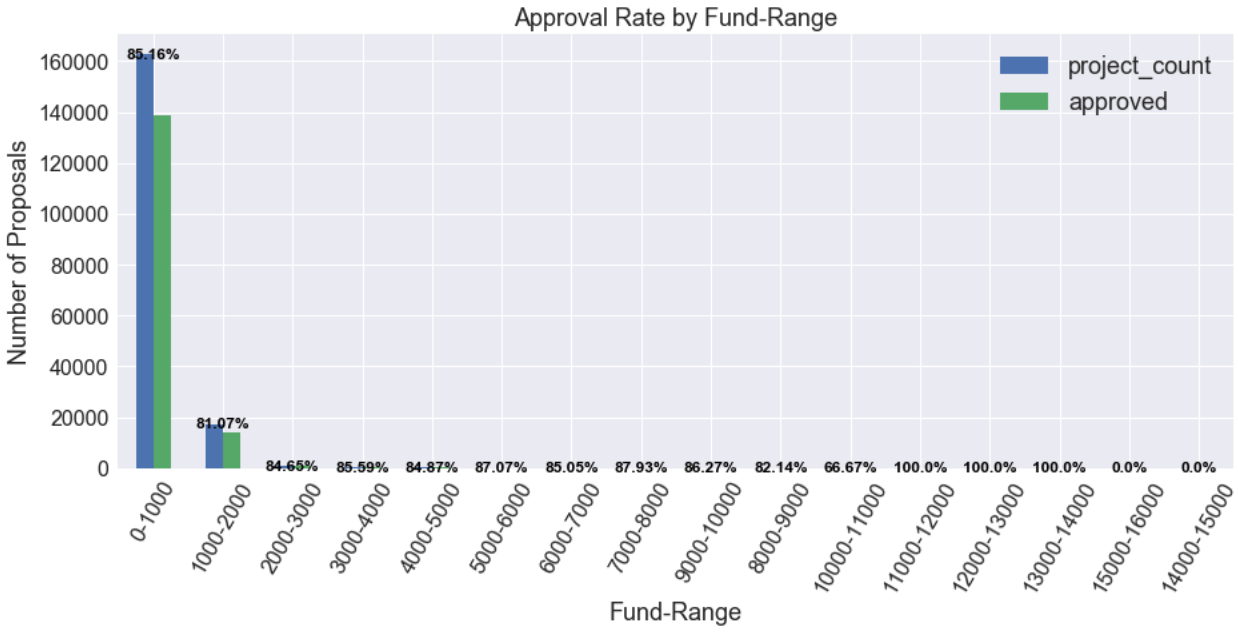
In this section we have done analysis on how the numbers of previously posted proposals by teachers affect the approval rate of the proposals. There are 401 unique submissions done in the past. It will be difficult for us to use these many records for bar plot. So, we will take only first 50 records. These records will include the number of proposals submitted by a **new teacher** i.e. who has no record of submitting any proposal in the past.



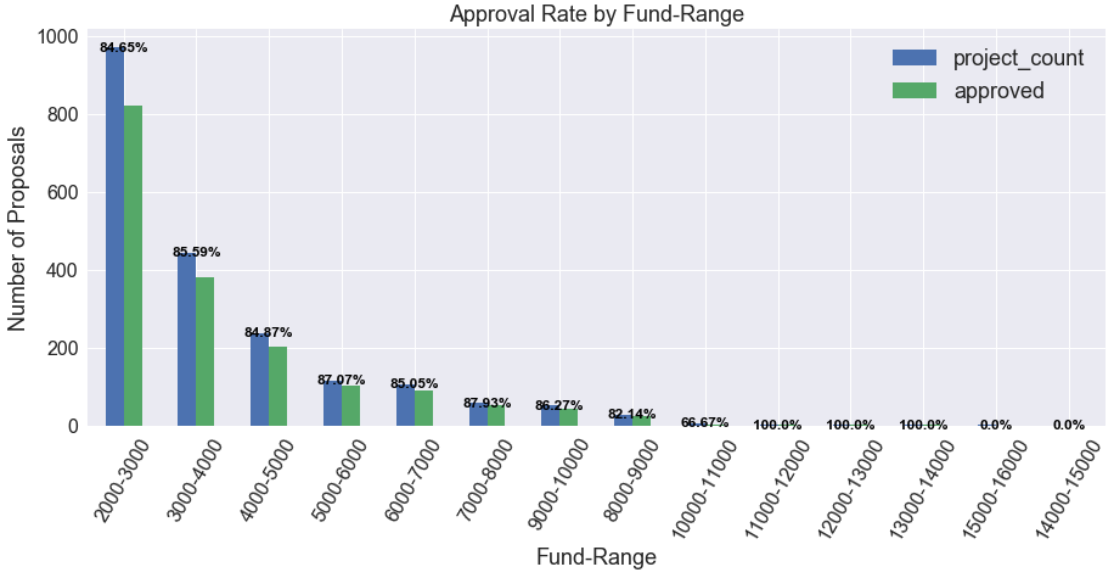
The number of proposals by **new teachers**, that is, teachers with no record of submitting proposal, is more than **50K** which is almost like **one third** of the total number of proposals. And the approval rate for all of these teachers is more than **80%**. So it doesn't look like the number of previously posted projects has any effect on approval of the project applications. There are teachers very few, with very high number of previously posted proposals, with **100%** approval rates.

4.7 Resources and Price wise

Every project has some associated resource requirement. Here analyzed the **total price range** and **approval rate**. Each project has a total cost associated with it. We will categorize these prices into range of 1000 dollars, starting from 0 to maximum of the total price.



The number of proposals decreases significantly with the fund required for the project increased. More than 160K project has the fund requirement between $0 and $1000. Let's try to zoom in little bit on the above graph to check the numbers if possible.



The minimum total amount requested for a project is **$100** and the maximum is around **$15300. 75%** projects requested about **$690.** As the price **increases**, the number of projects **decreases** significantly. There are just 1 or 2 projects which requested for high amount of more than **$10K.**

4.8 Text analysis and WordClouds

This analysis was done on all the text columns of the dataset to find the approval rates and to find if there are any particular set of words whose approval rate is higher than the rejected proposals.

4.8.1 Project Title

Project title is one of the most important text columns for any project. This tells the concise description of the project. We found some funny project titles with surprisingly great success rate. Below are the top 5 project titles and their success rate.

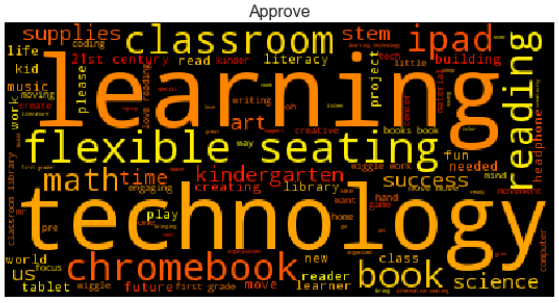
Table 1-1

Project_title	Project Count	Approved	Approval Rate
Flexible Seating	377	298	79.05
Wiggle While You Work	150	137	91.33
Wiggle While We Work	149	136	91.28
Can You Hear Me Now?	144	135	93.75
Wiggle While You Work!	134	122	91.04

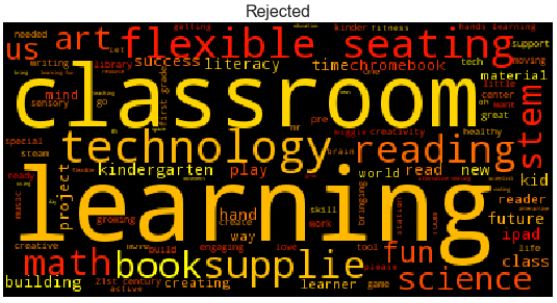
We can see the use of world "Wiggle" and the approval rate of more than 90%.

WordClouds for Approved and Rejected Project Titles

Approve



Rejected

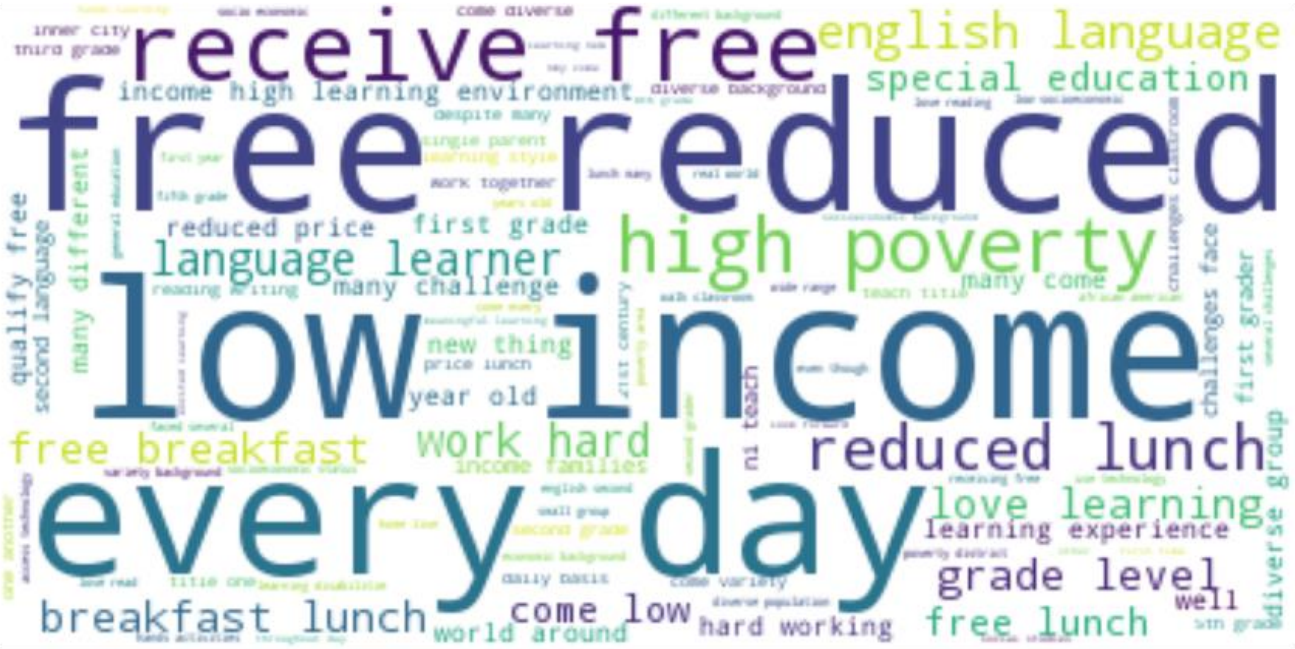


Technology and related words are most used words in the project title. Science and math also seems very popular words. Words like **classroom, books, building, flexible seating** reflects the infrastructure need of the students for better **learning**. There is not much difference in the titles of approved and rejected proposals.

4.8.2 Project Essay

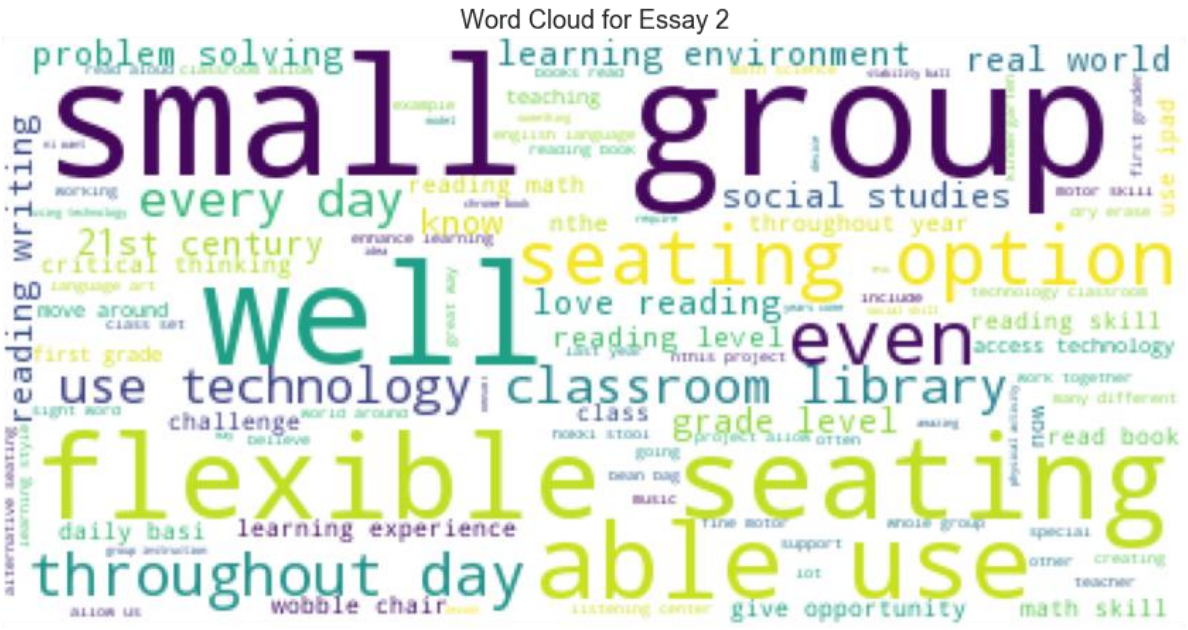
Essay gives opportunity to the teachers to tell their story about the class, students, projects and how the project will improve **learning**. It would be interesting to see what kind of words teachers have used in the submitted, approved and rejected proposals.

Project Essay 1



Most frequently used words are related to **students** as this field is required to write down about students and class. There is not much difference in the frequently used words in approved and rejected proposals, except word "**Classroom"** which has been used frequently in proposals which were rejected.

Project Essay 2

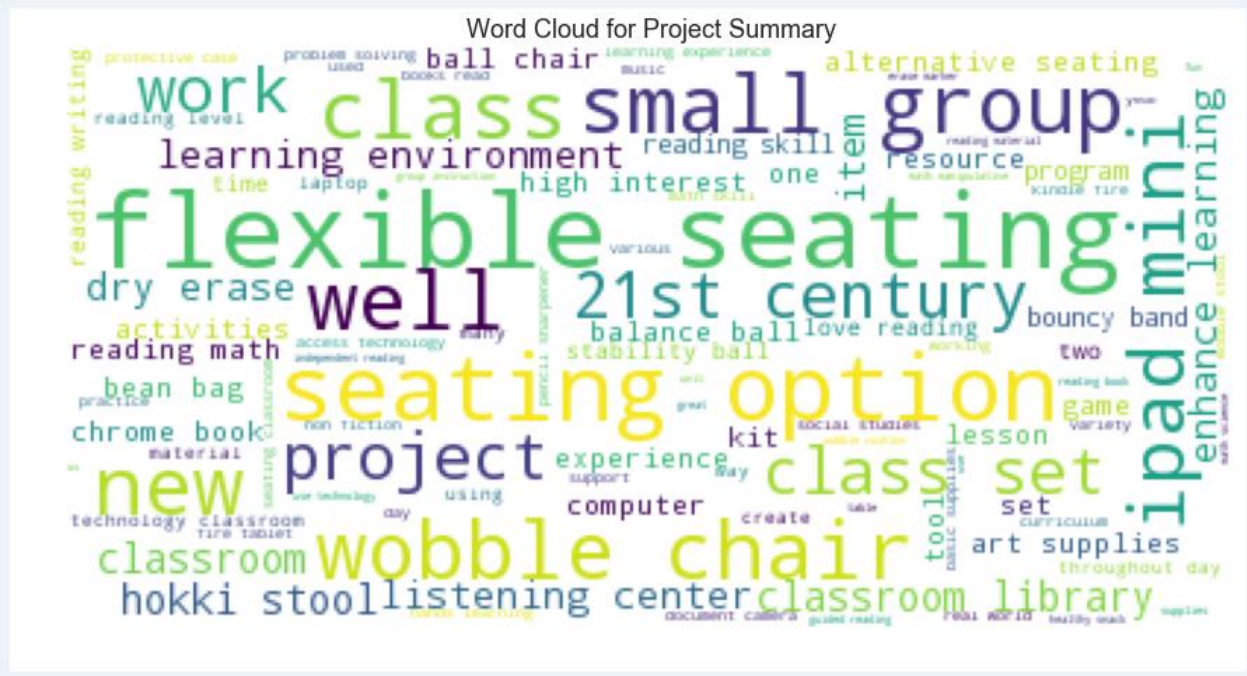


Essay 2 is more about problems which students face, things which will help them and how exactly they will help improve learning. Words like, **"small group, seating options, flexible, well, use"** are frequently used words. There seems to be many differences in frequency of words used in essay 2 for approved and rejected proposals.

4.8.3 Resource Summary

Summary of the resources needed for the project. All proposals also include resources requested. Each project includes multiple requested resources. Describing about project resource is equally important part of the proposal.

The most frequently used words are related to the object, like **chair**, devices like **ipad mini** and words that describes these objects. As summary is about the resources requested, we would expect something like this to improve the learning environment. There are sports accessories related words that are also frequently used in describing the resources. The words are almost same in both, approved and rejected proposals, just frequency of these words are different. We can see the word cloud in the next page.



5. Predictive Model

We will be developing predictive models for classification. These models will predict if the application of project proposal to **DonorsChoose.org** is approved to be posted on the **DonorsChoose.org** website or not. We will also compare the models and try to figure out which model has the better performance. The model with best performance will be our model to predict the real test data for which we have no actual target values.

There are 3 steps involved in this part.

1) Feature Engineering: Feature engineering is one of the most important and tricky process in machine learning. With the dataset we have numerical, categorical and text data. We will create some features with text data available in different files and columns and convert them into numerical values; categorical data will be engineered into numerical values as well.

2) Modeling: Once we are done with the feature engineering part, modeling will be much easier. We will choose one model from each of the **Linear, Naive Bayes and Decision Trees models** for classification. We will evaluate these models and compare the models on their performance.

3) **Model evaluations and final prediction:** We will compare all of the 3 models mentioned above and will use the best model to predict the real test dataset.

5.1 Data Pre-Processing

Before we start feeding the data into the machine learning algorithms, we need to bring the data into desired format. This includes converting **categorical** values into numerical values, **date**

time column converted into different related numerical values, derived features from other columns like **word count** in the project title, **total price** of the resources required in the proposals and **Word Vectorization** for all the text columns.

5.1.1 Feature Engineering - Categorical Features

All the categorical columns like *teacher_id', 'teacher_prefix', 'school state', 'project grade category','project subject categories', 'project_subject_subcategories'* available in the dataset got converted into the numerical values using **LabelEncoder()**. In some classification models LabelEncoder() performs better than the OneHotEncoder().

5.1.2 Feature Engineering - Datetime Features

There is one column which has submitted date and time and we had converted or extracted each day, month, day of week, and hour of the day. The ***datetime*** library helped extract these values.

5.1.3 Feature Engineering - Derived Features

Some of the features were derived from other columns, for example word count of the important text columns, the total price and number of unique resources, quantities required for the proposal from the resource dataset.

5.1.4 Feature Engineering - Text Features

This is one of the most important features of the dataset. We have combined all the text columns of the dataset, used ***WhitespaceTokenizer*** and ***WordNetLemmatizer*** on the entire text to get only the words which are needed and to avoid English stop words, we have used tf-idf(term frequency - inverse document frequency) technology. Then the entire columns were vectorized using compressed sparse matrix.

5.2 Modeling

As we have discussed earlier, we will be using 3 different models of three different kinds.

1) Linear Model: LogisticRegression
2) Naive Based model: MultinomialNB
3) Decision Tree: RandomForest

We have used GridSearchCV to tune hyperparameter for better performance of the model.

Model Evaluation

We used the below 4 metrics to evaluate all the models to find the best model to predict the real test dataset.

1) Confusion Matrix
2) Average Precision score
3) ROC AUC score
4) F1- score

5.2.1 LogisticRegression

This is the linear model with the "sklearn" package in python for data science. This model is widely used for the binary classification problems. As discussed earlier, we have used GridSearchCV to tune the hyperparameter and for this model the best parameters are ***Best Parameter: {'C': 100}, Best score: 0.734377159641.*** with c=100 the model produces the score of .7344 on the training dataset.

The **model evaluation** with all the metrics discussed above, resulted as below.

1) The Confusion Matrix for LogisticRegression:

```
[[6601  4493]
[16308 45430]]
```

2) Average Precision for LogisticRegression: **0.927741703189**

3) The ROC AUC Sore: **0.727257741346**

4) The f1 score for the LogisticRegression model: **0.813712934686**

The model was able to classify total of **52081** records correctly, which is like 72% correct prediction. We want our model to predict approved i.e. class 1, so we would like to see that the number 45430 gets as big as possible. The other scores would make more sense when we compare the models.

5.2.2 MultinomialNB

Naive Bayes classifiers are a family of classifiers that are quite similar to the linear models. However, they tend to be faster in training. The price paid for this efficiency is that Naive Bayes models often provides generalization performance that is slightly worse than linear classifiers. The reason that Naive Bayes models are so efficient is that they learn parameters by looking at each feature individually, and collect simple per-class statistics from each feature. There are three kinds of Naive Bayes classifiers implemented in scikit-learn, GaussianNB, BernoulliNB and MultinomialNB. We will be using **MultinomialNB.**

When we used GridSearchCV on the model to tune the alpha hyperparameter, we got Best parameter: {'alpha': 0.001}

```
Best score: 0.561342432619
```

The best parameters on the test dataset produce score of around .56, which is no better than LogisticRegression. Let's have a look on other metrics.

1) The confusion matrix for Naive Bayes Multinomial classifier:

```
[[ 6296  4798]
[28448 33290]]
```

2) The average Precision score for MultinomialNB is: **0.872202932123**
3) The ROC AUC Sore: **0.571602972261**
4) The f1-score for MultinomialNB model is: **0.66696051129**

As per the confusion matrix the total number of correct predictions is **39572,** which is only around 55%. This is clear that the model is no better than LogisticRegression when we compare all the 4 metrics. Now it's time to move to the next model.

5.2.3 RandomForest

A main drawback of decision trees is that they tend to overfit the training data. Random forests are one way to address this problem. Random forests are essentially a collection of decision trees, where each tree is slightly different from the others. The idea of random forests is that each tree might do a relatively good job of predicting, but will likely overfit on part of the data. If we build many trees, all of which work well and overfit in different ways, we can reduce the amount of overfitting by averaging their results. This reduction in overfitting, while retaining the predictive power of the trees, can be shown using rigorous mathematics. For this model we had to manually set the hyperparameter and train the model because of the computational limitations of my system. We found that the **n_estimators = 100** and **max_depth =100** gave the best f1 score on training dataset compared to other hyperparameter settings. Below are the other scores of metrics used for model evaluation.

1) The confusion matrix for Random Forest:

```
[[ 1233  9861]
[ 2356 59382]]
```

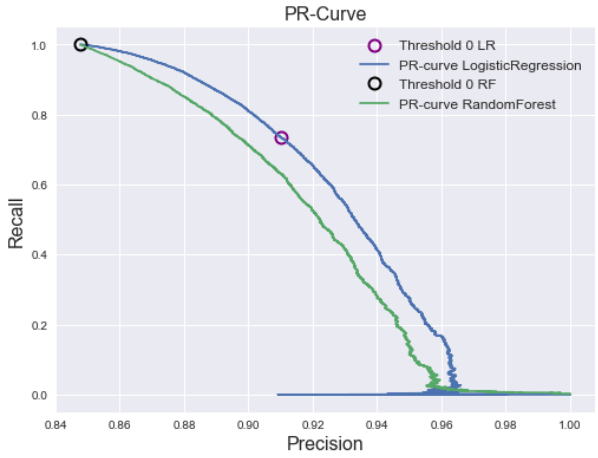
2) The average precision score for RandomForest is: **0.917358842643**
3) The ROC AUC Sore: **0.687922987604**
4) The f1 score for the RandomForest model: **0.90672692986**

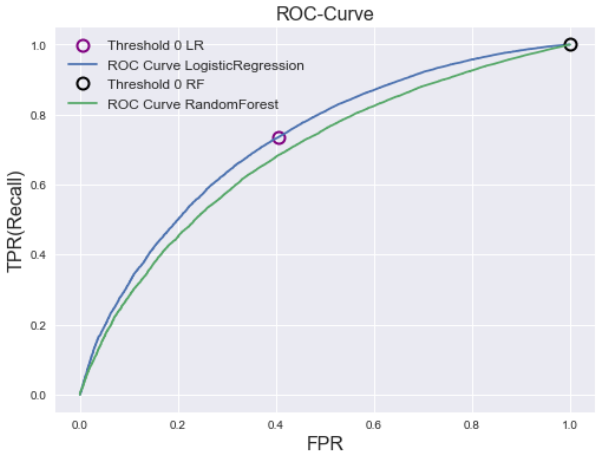
From confusion matrix we can see that almost 83% of records were classified correctly, which is better than other two models.

5.2.4 Model Comparison

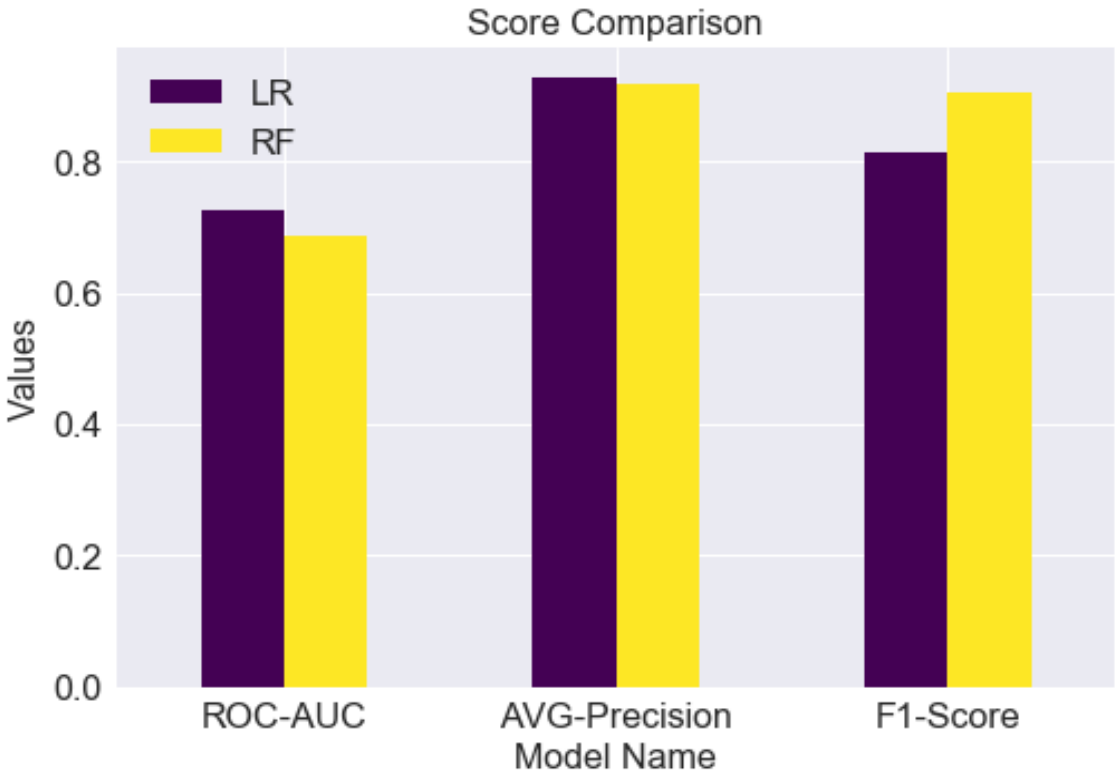
We compared the three models based on the 4 different metrics discussed above and found that the **Multinomial Naive Bayes** model is the worst performing model among the three models with respect to each metrics.

The other two models, **LogisticRegression** and **RandomForest** have a tie, 2 of the 4 metrics namely **confusion matrix and f1-score** to measure the performance of the model suggests that **RandomForest** model has the better performance whereas the other two metrics, **Average Precision score** and **ROC AUC score** suggests that the **LogisticRegression** model has the better performance.





Score comparison



Hence, we decided to use both the models on real test dataset to predict the classification and create a dataframe having probabilities of both the models together. First few records of the resulted dataframe are as below.

Final Dataframe - Real dataset

id	lr_prob_0	lr_prob_1	LR_Approved	rf_prob_0	rf_prob_1	RF_Approved
p233245	0.089648	0.910352	1	0.179338	0.820662	1
p096795	0.383522	0.616478	1	0.140699	0.859301	1
p236235	0.013738	0.986262	1	0.246652	0.753348	1
p233680	0.099697	0.900303	1	0.198845	0.801155	1
p171879	0.190225	0.809775	1	0.169755	0.830245	1

So we have the final dataframe, which can be transformed into any desired format for submission. Also, we can **increase/decrease** the threshold for the probability class for better screening. For example, if the applications submitted to the DonorsChoose.org are more, they can increase the probability threshold of "approved or 1" class from default .5 to something like

.6 or .65, to regulate the number of project proposals so that volunteers can spend time on really good proposals which has more chances of getting approval and to be posted on **DonorsChoose.org**

6. Limitations

This being a competition dataset, the chance of having different analysis and solving a different problem with the same data is very rare. But hypothetically, I believe if we had the data about number of public schools in each state, number of students in each class or the class for which the proposal was made, and data like the state provided funds to schools etc. would make this analysis much more productive in terms of assessing the requirement of the students and help prioritizing the proposals by the class teachers.

Execution time, i.e. time it takes to train the models with train dataset and with best hyper parameters for the models, especially RandomForest with n_estimators = 100 and max_depth =100, is higher on personal computers as the number of core CPU is limited.

7. Other Data and Future Work

As mentioned in the limitations section, more data about the public schools in each state, with the financial conditions of the schools and students, will make analysis better and improve the learning environment for the students. I would love to work on solving problems with different grades and requirements for the students. Also identify some metrics to measure the improvement after the class teacher gets his/her proposal approved and gets the required resources. Like how the learning is different in a better way, than previously when the students lacked resources.

8. Conclusion

8.1 EDA

During the EDA (Exploratory Data Analysis), we tried to find the relation between different columns of the dataset with the target columns. The overall approval rate of the training dataset is around 85%, which means most of the data proposals which are applied to DonorsChoose.org get approved to be posted on the website for fund raising.

We found some interesting and funny project titles of the proposals submitted like **Wiggle While You Work!,** and these titles have the approval rate of more than 90%. Also class teachers uses words like **classroom, free, low income everyday** etc. which describes the condition of the students and class for the essay 1 which is about the students and classroom, and uses words like **learning, small groups, flexible seating** etc. to describe what problem students are facing and how the requested resources would help them improve their learning.

We also found that most of the teachers submit the project during the month of **Aug and Sep,** which is the starting of school semester.

8.2 Machine Learning Models

RandomForest, and **LogisticRegression** are the two models which are used to predict the real test dataset. **GridSearchCV** had helped finding best hyperparameter for the different models. We have used 4 metrics for model evaluations. RandomForest performed better on two metrics and LogisticRegression performed better on the rest two, hence we used both the models to predict the real test dataset.

By doing so, we can set different thresholds for approval and rejection condition. Like a project is considered to be approved, if it has more than 60% probability for getting approved. This will help in better screening when number of applications is really big like in millions which is the expectation of **DonorsChoose.org.**