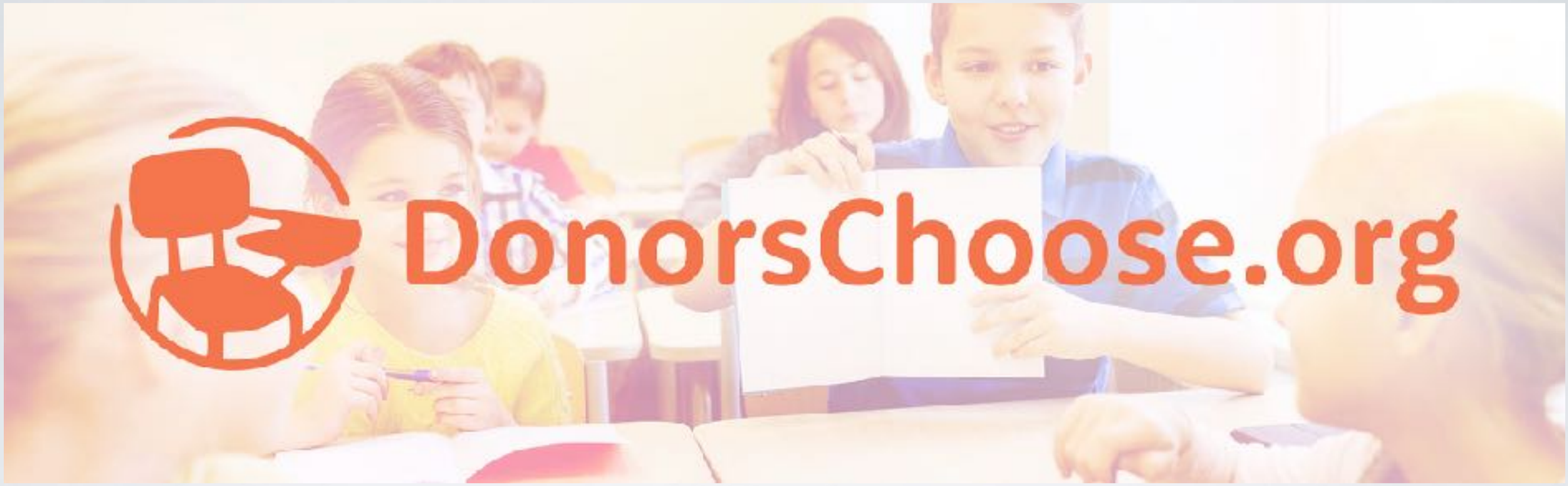


APPLICATION SCREENING

Vishal Kumar
Springboard - Data Science Career Track
Capstone Project

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.
- 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.
- 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

POTENTIAL CLIENTS

- **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.

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 is 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

- The entire project is divided into two parts.

1. EDA - Exploratory Data Analysis:

    - In the EDA part, we have tried to find the approval rates of all the columns with the target column. The overall approval rate is 85%.

    - Time series analysis to find out the relation between approval rates on month, day of the week and period of the day.

    - Derived column: columns generated from calculation on other columns. Like total price, number of words etc.

    - Text Analysis: Most frequently used words in project title, essay and description.

2. Modeling - 3 Models

    - Linear Model - LogisticRegression

    - Naive Bayes Model - MultinomialNB

    - Decision Tree - RandomForest

- Dataset
  - This dataset was chosen because it involves the use of Natural Language Processing (NLP).
  - This is kaggle competition dataset and was downloaded from https://www.kaggle.com/c/donorschoose-application-screening/data

DATA WRANGLING

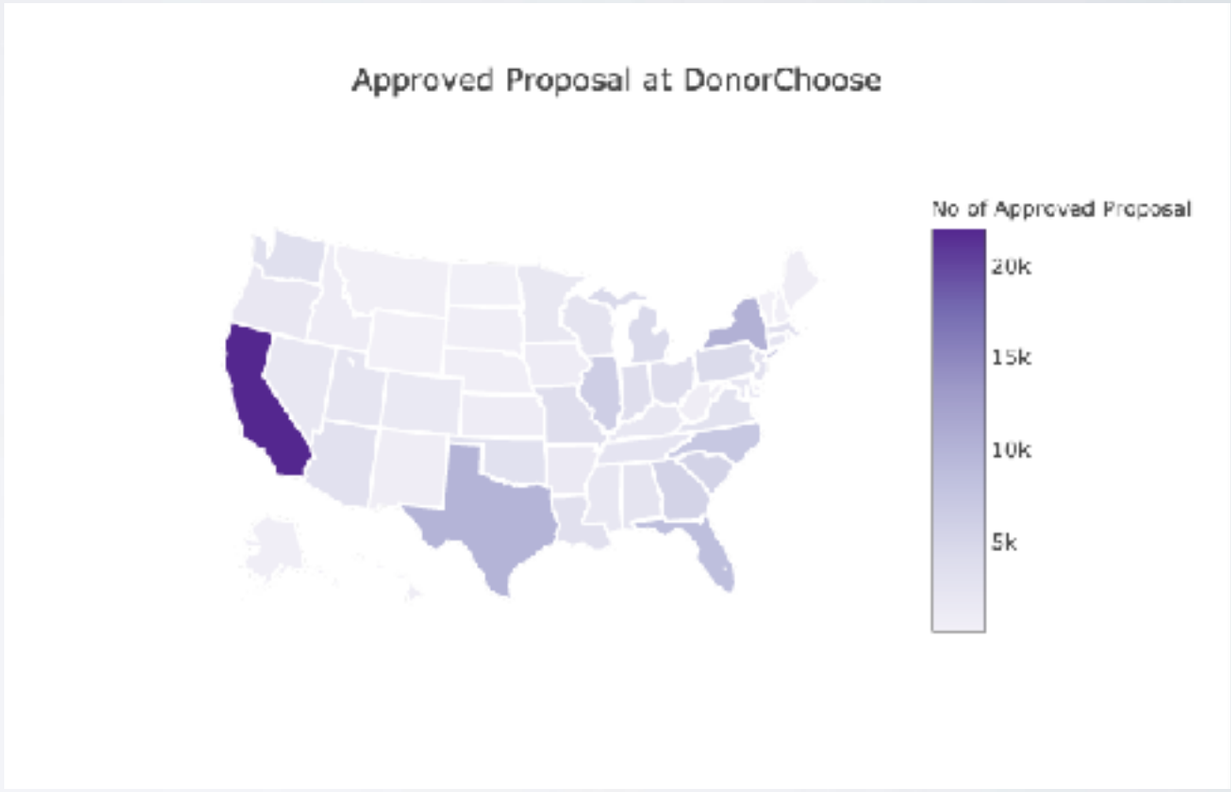
- Bringing the data to desired format
  - For teacher's gender based approval rate, I had to combine doctors and teachers as unknown genders, hence categorizing the entire dataset into three category, Female, Male and Not Known.
  - Used ***split()*** function on project subject category and project subject sub category. By doing this we could have reduced the number of unique subject to 9 and unique subject sub category to 30, which was easy to handle and made more sense for the analysis.
- We derived 5 new columns using datetime data type of python.
- divided the day into 4 categories as below and grouped the data accordingly.
  1. 06:00 - 11:59 Morning
  2. 12:00 - 15:59 Afternoon
  3. 16:00 - 19:59 Evening
  4. 20:00 - 05:59 Night(nights are long)

- To create a price range, used itertool package along with tee, islice, chain of python.
- Used a special pandas method pandas.cut()
- For text and word cloud analysis, I have used some basic string operations of python like .join() and split() method.

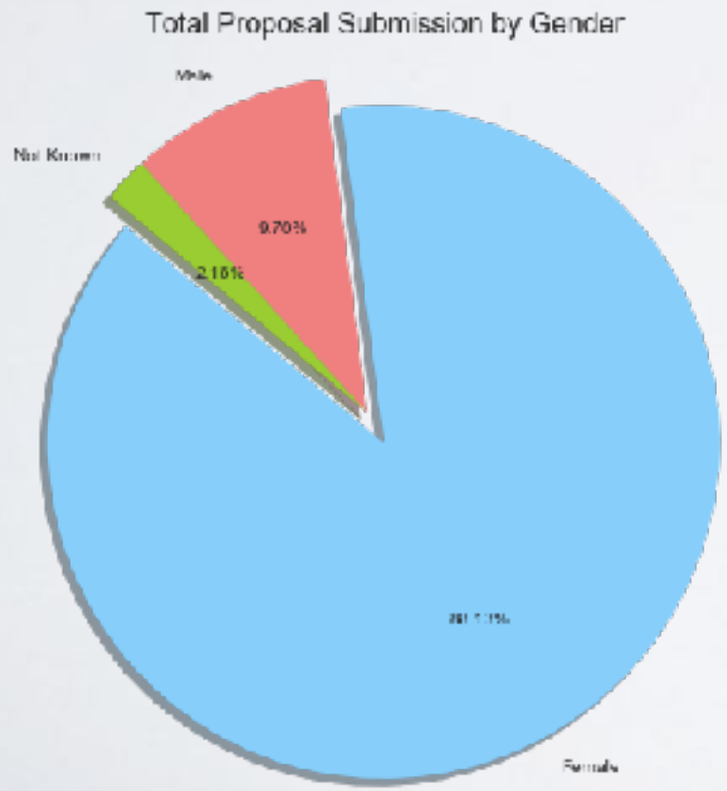
- Data Cleaning: Since this competition dataset most of the data were clean, however we encountered missing values or **NaN** values in some of the columns.
- **NaN** values dealt with fillna() method accordingly to the situations.
- Teacher's prefix column has 4 NaN values, which were filled with "Teacher" prefix
- Price range values were filled with 0(zero) for all **NaN** values.
- Merging essay 4 columns into 2 columns as per the given definitions.

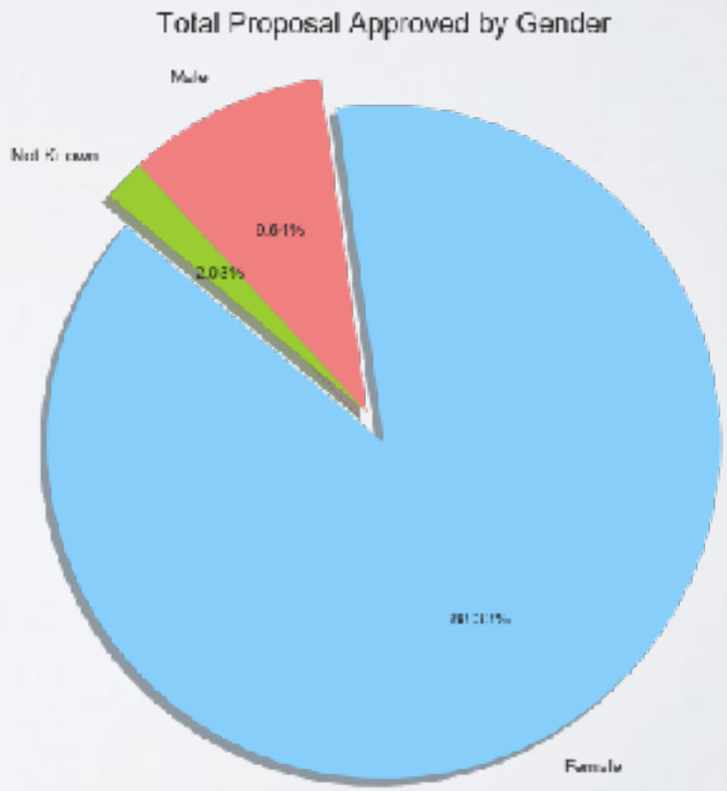
DATA EXPLORATION

- The overall approval rate is almost 85%
- By State:
  - All states has more than 81% approval rate.
  - California has the highest number of application submission and hence has the highest number of approval rate.
  - Delaware has the highest % of approval rate at slightly more than 89%.
  - D.C. Has the highest % of rejection at almost 19%.

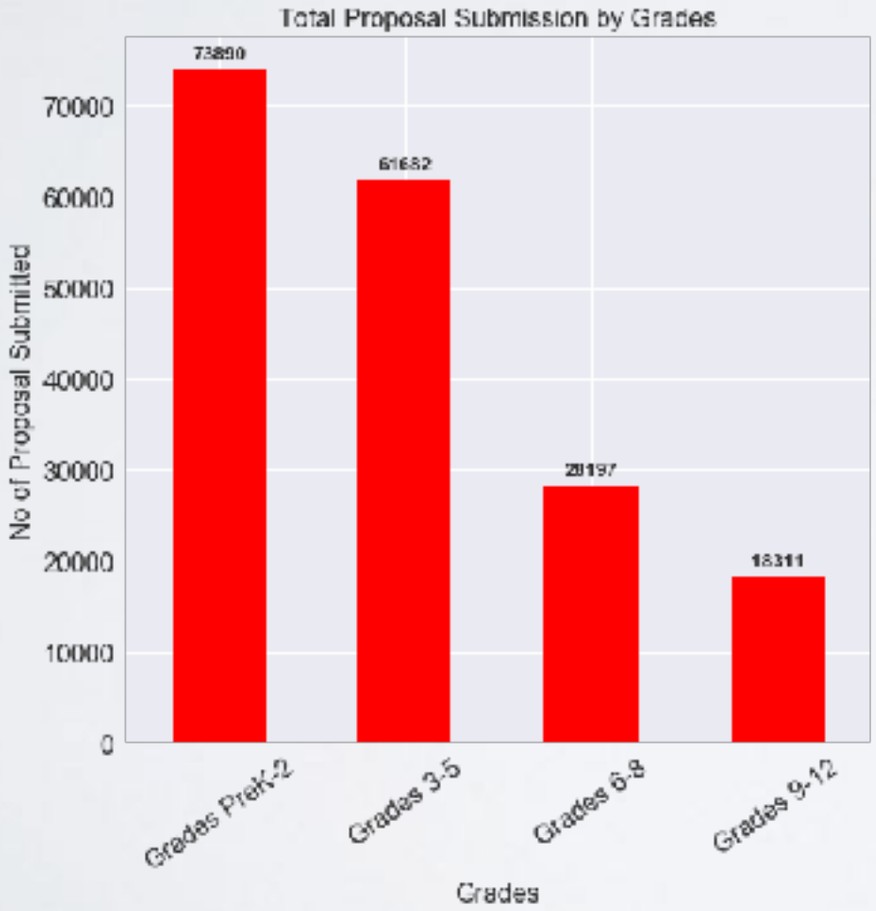


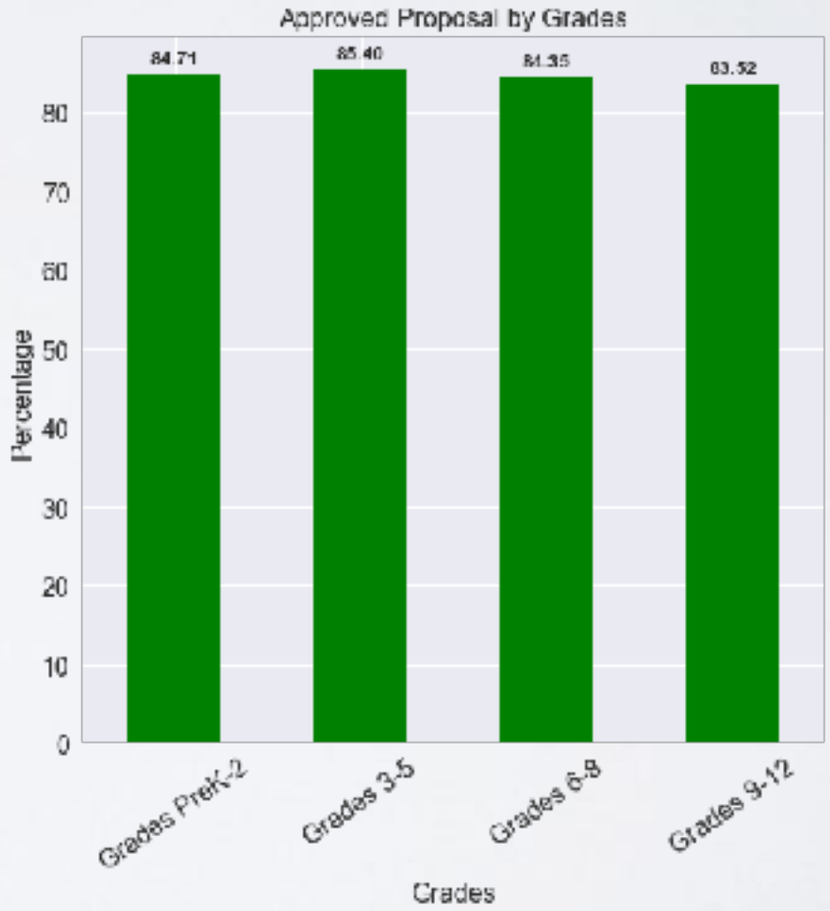
- By Gender
  - Female teachers have the highest number of application submission at 88%, compared to 10% of males and 2 % of unknown gender.
  - Male and Female approval rate is almost same around 84% and 85% respectively.
  - Even teachers with not know gender has a very good approval rate of almost 80%.



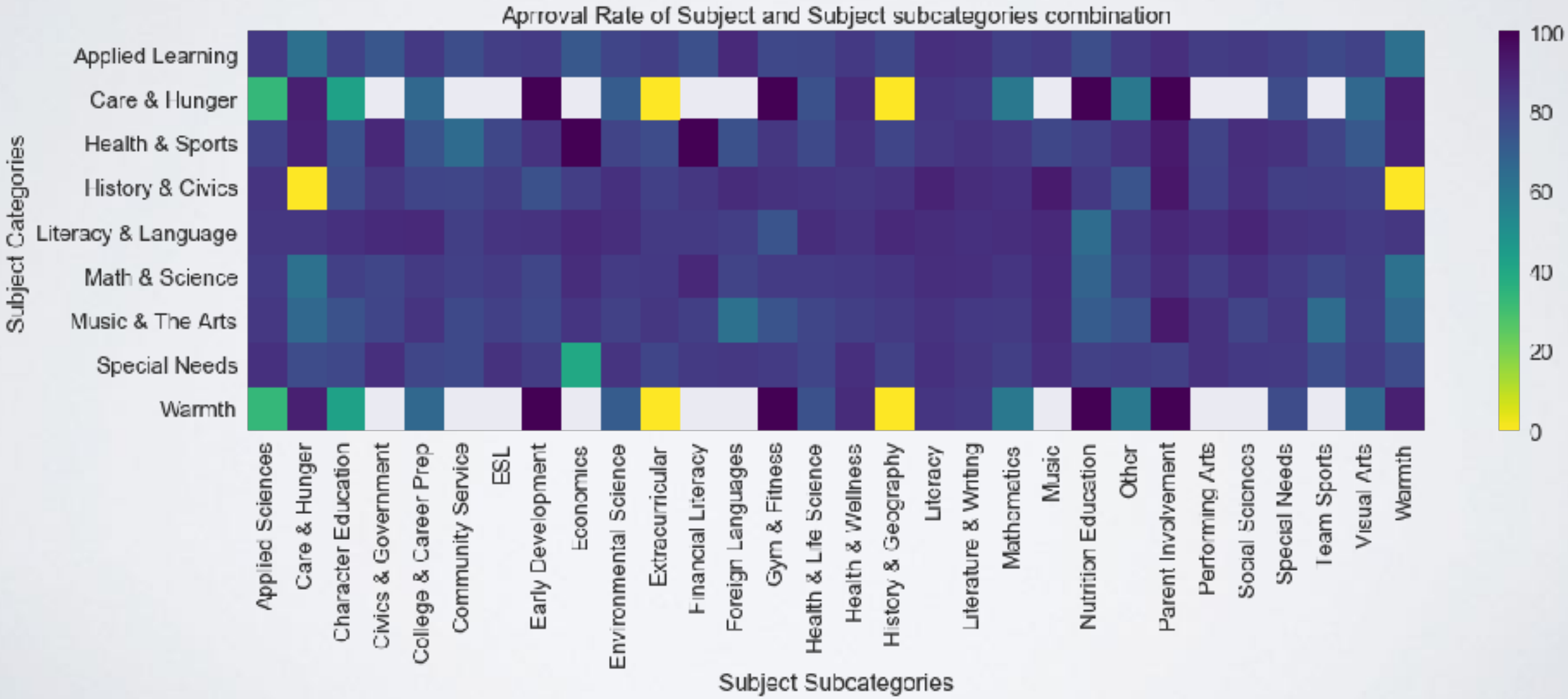


- By Grades
  - The proposal submissions for grades Prep-2 and grades 3-5 is pretty high compared to grades 6-8 and grades 9-12.
  - The approval rate for each of the grade category is between 83%-85%.

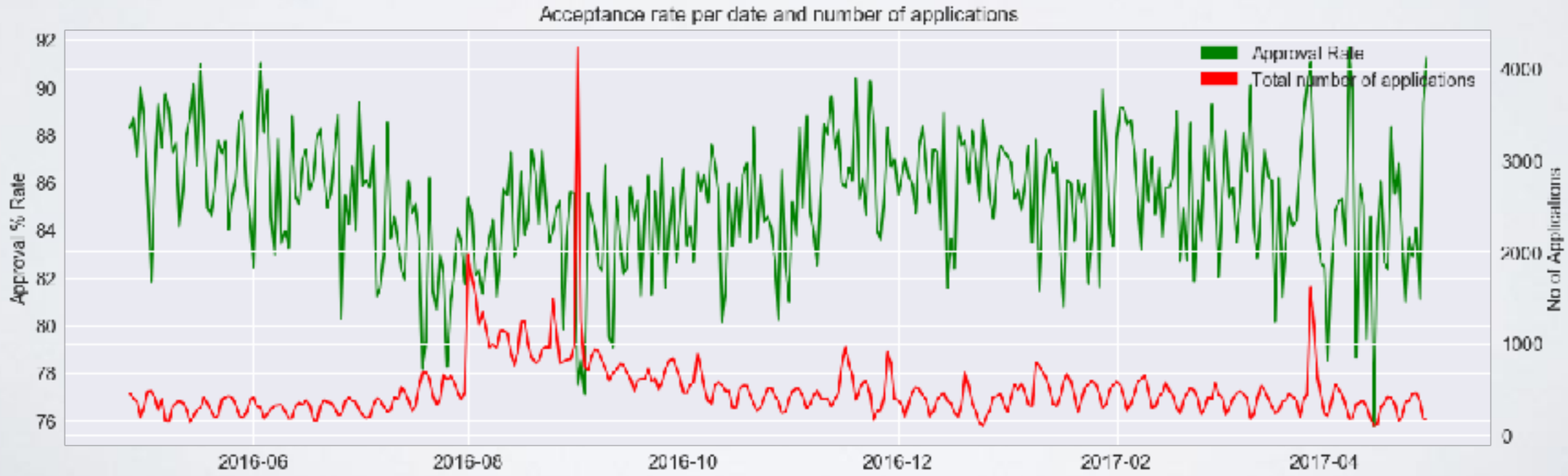




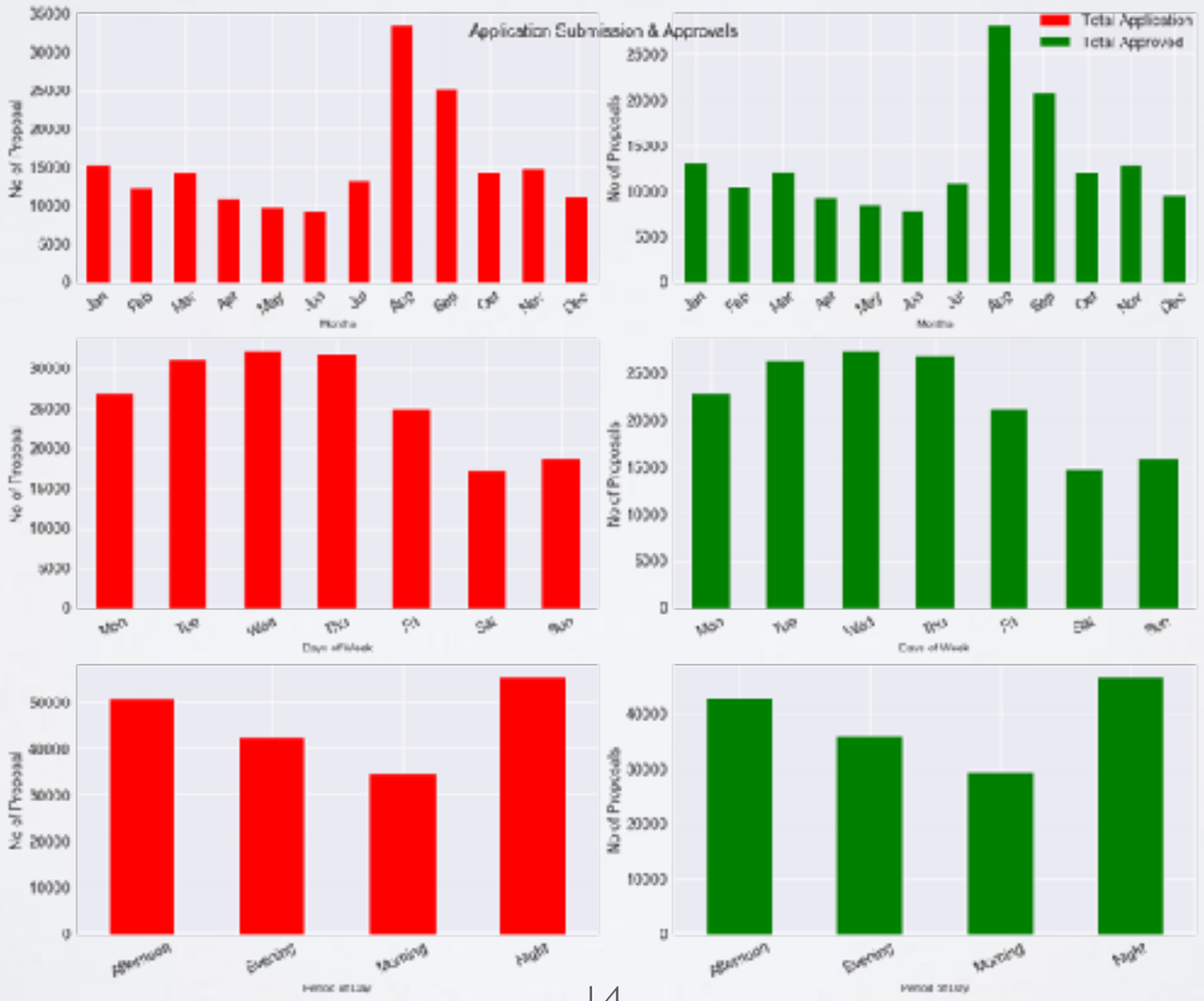
- By Subject and Subject Subcategory
  - Identify unique subjects and subjects subcategory.
  - Subjects like Care & Hunger and Warmth have 0% approval rate for some of the sub-categorical subjects, thats because there is no submission in the sub-categorical subject.
  - Applied Learning, Literacy & Language, Maths & Science subjects have better approval rates ranging from 60% to 80%.



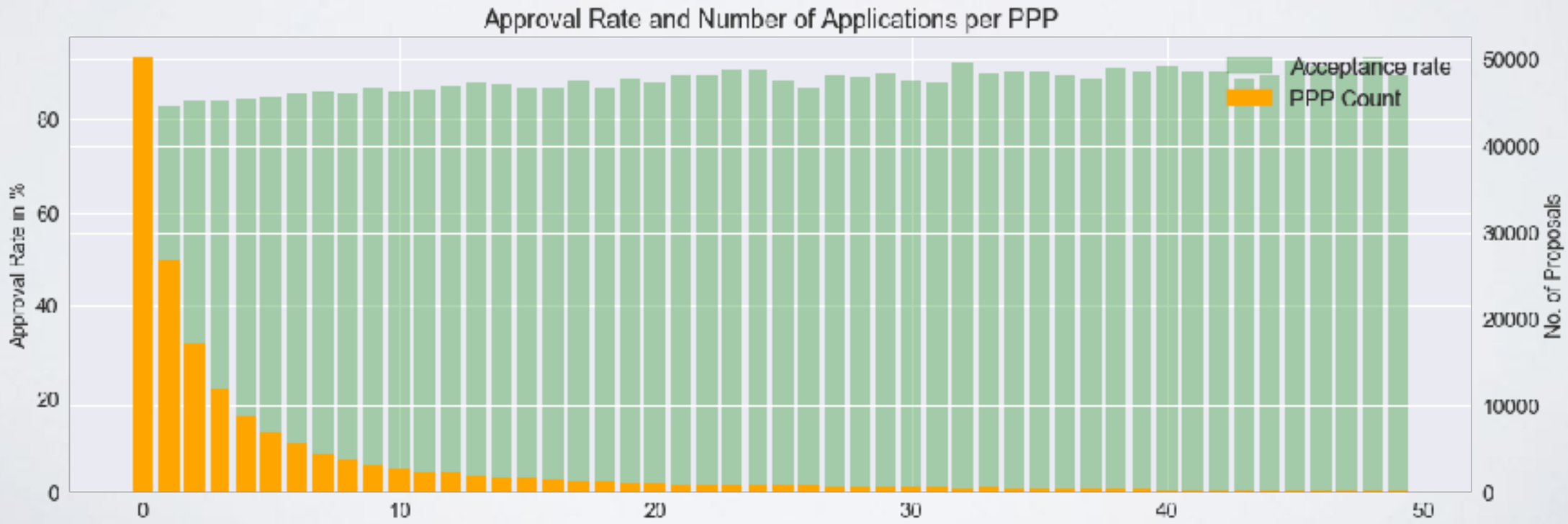
- Time Based
  - We have data for the full year.
  - The high peaks near the month August and September and then falls back to the normal baseline.
  - Approval rates are ranging from as low as 75% to as high as to nearly 92%.



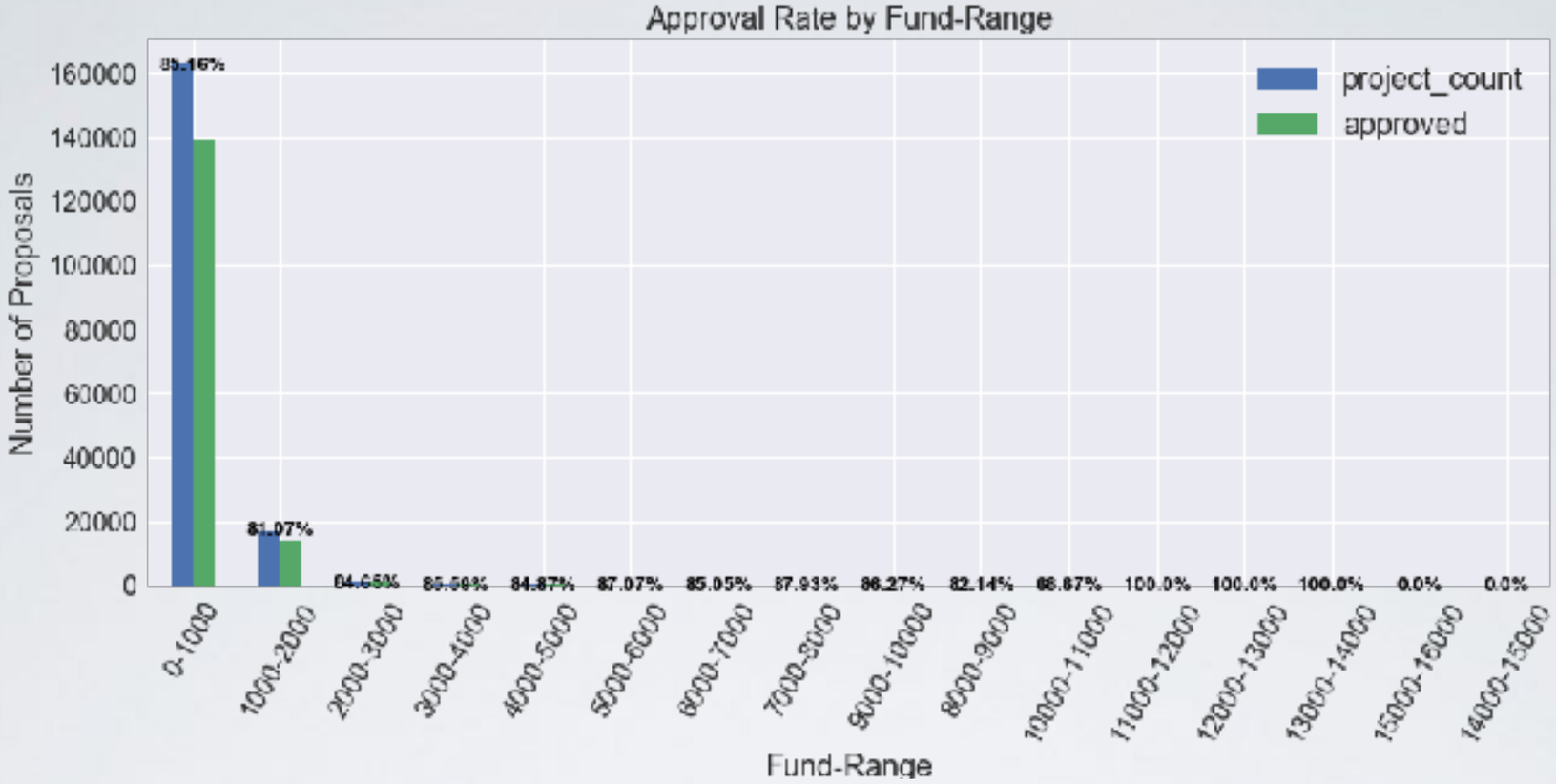
- April, May and June months are dull the compared to other months.
- Slightly more than 50% proposals are submitted on these Tuesday, Wednesday and Thursdays.
- Most teachers choose night time (after 8:00 Pm and before 06:00 Am) to submit a proposal. Morning time from 6am to 12 noon is the least chosen period of day to submit a proposal.
- Number of approved proposals are in the same order as of the total number of submission. The approval rate is around 80%-85% in all category.

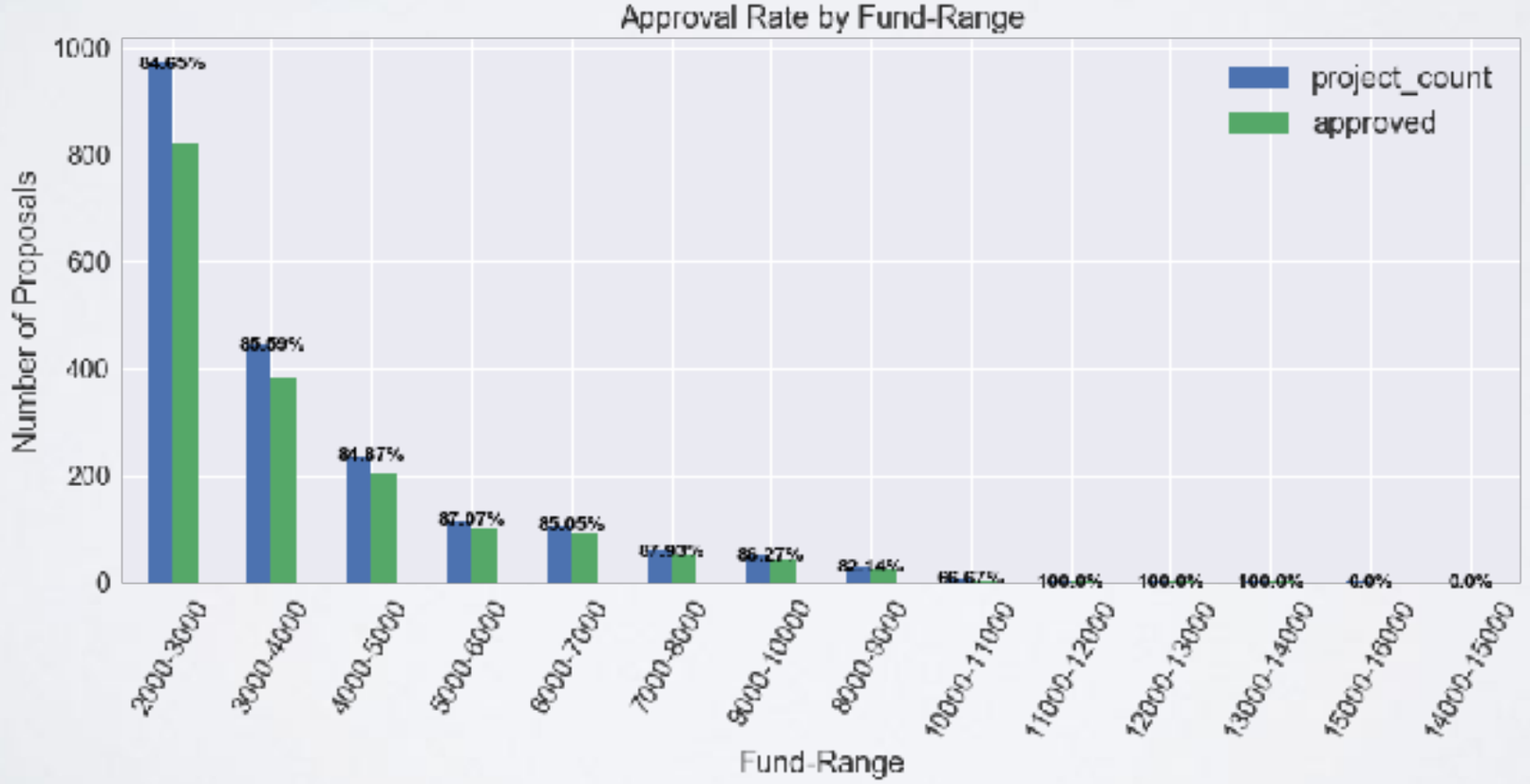


- Previously Posted Project wise
  - Teachers with no record of submitting proposal are more than 50K which is almost like one third of the total number of proposals.
  - The approval rate for all of these teachers are more than 80%.
  - There are few teachers, with very high number of previously posted proposals, with 100% approval rates.



- By Resources and Price
  - Every project has some associated resource requirement.
  - For this project, we have 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 is increased.
  - More than 160K project has the fund requirement of between $0-$1000.
  - The minimum total amount requested for a project is $100 and the maximum is around $15300.
  - 75% projects requested about $690.
  - There are just 1 or 2 projects which have requested for high amount of more than $10K.


Approval Rate by Fund-Range
project_count
approved
Number of Proposals
0
20000
40000
60000
80000
100000
120000
140000
160000
81.07%
0-1000
1000-2000
2000-3000
3000-4000
4000-5000
5000-6000
6000-7000
7000-8000
8000-9000
9000-10000
10000-11000
11000-12000
12000-13000
13000-14000
15000-16000
14000-15000
Fund-Range

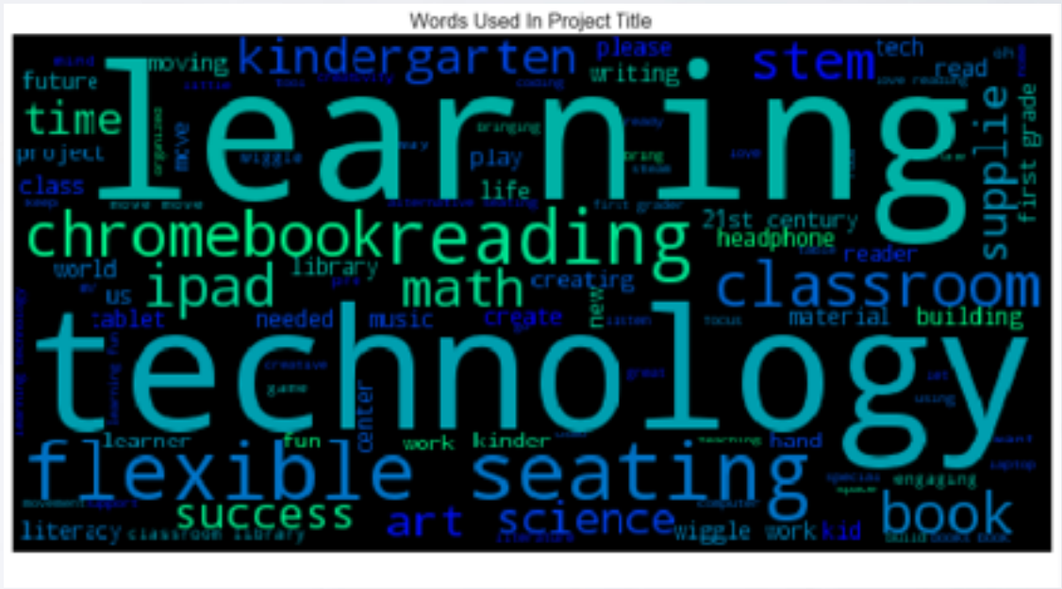

Approval Rate by Fund-Range
project_count
approved
Number of Proposals
0
200
400
600
800
1000
85.09%
84.87%
87.07%
85.05%
86.27%
100.0%
100.0%
100.0%
0.0%
0.0%
2000-3000
3000-4000
4000-5000
5000-6000
6000-7000
7000-8000
8000-9000
9000-10000
10000-11000
11000-12000
12000-13000
13000-14000
15000-16000
14000-15000
Fund-Range

TEXT ANALYSIS WITH WORD CLOUDS - PROJECT TITLE

- Most frequently used words in the project titles are technology, learning.

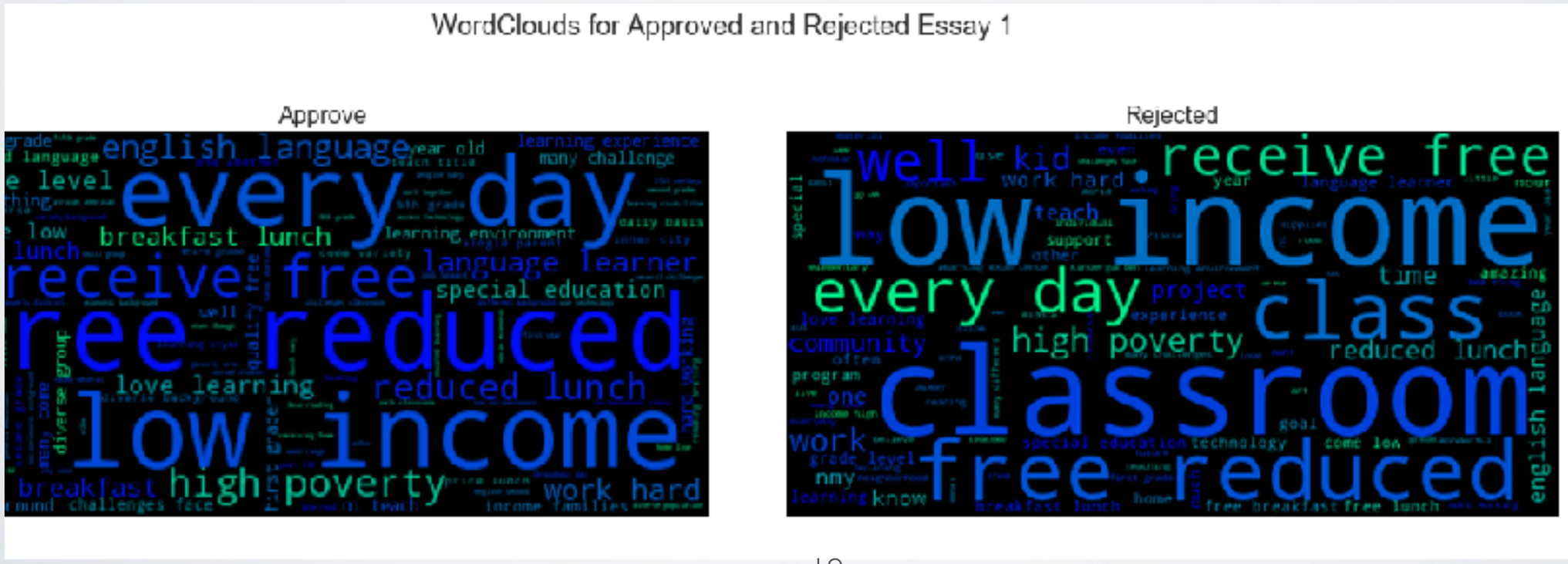
- Science and math are also seems very popular words.

- Words like classroom, books, building, flexible seating reflects the infrastructure need of the students for better learning experience.



TEXT ANALYSIS WITH WORD CLOUDS - PROJECT ESSAY 1

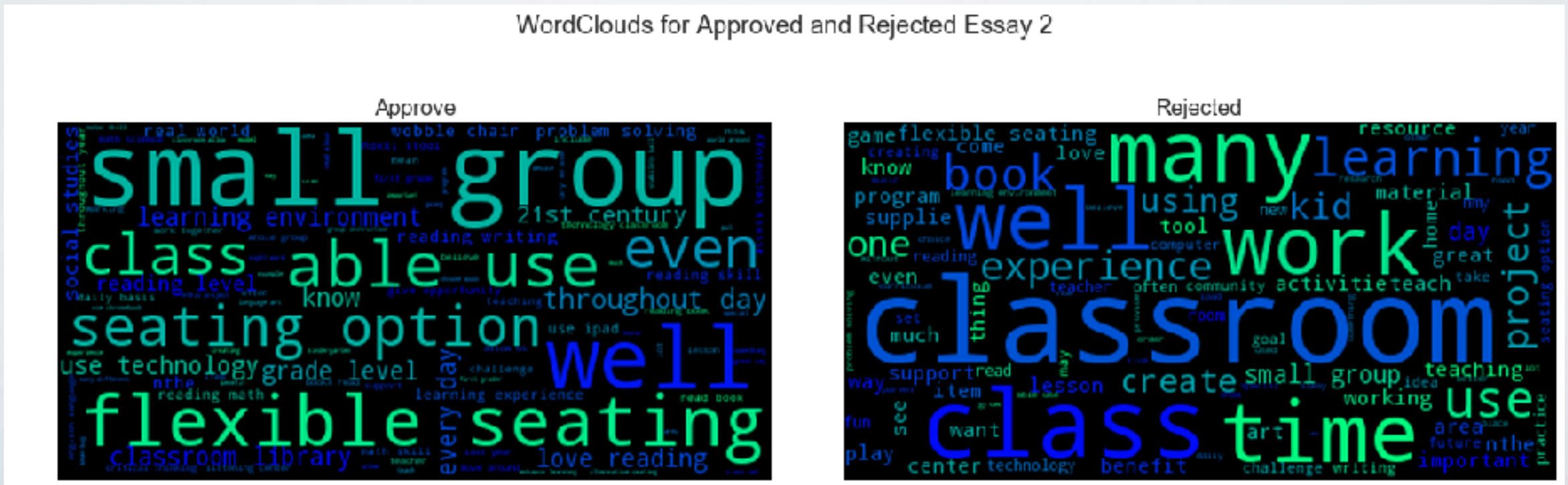
- Most of the frequently used words are related to students as this field required to write down about students and class.
- There is no much difference in the frequently used words in approved and rejected proposals, except word "Classroom" which has used frequently in proposals which has been rejected.



TEXT ANALYSIS WITH WORD CLOUDS - PROJECT ESSAY 2

- Essay 2 is more about problems which students face, things which will help them and how exactly will they help to improve the learning.

- Words like, "small group, seating options, flexible, well, use" are frequently used words. There seems to be many difference in frequency of words used in essay 2 for approved and rejected proposals.



TEXT ANALYSIS WITH WORD CLOUDS - RESOURCE SUMMARY

- The most frequently used words are related to the object, like chair, devices like ipad mini and words that describes these objects.
- The are sports accessories related words 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.

PREDICTIVE MODELING

- Models to predict if the application of project proposal to DonorsChoose.org is approved to be posted on the DonorsChoose.org website or not.
- Compare the models to figure out the model that has the better performance.
- The model with best performance will be our model to predict the real test data.
- 3 steps involved
    1. Feature Engineering or Data Preprocessing
    2. Modeling
    3. Model evaluation and final prediction

FEATURE ENGINEERING

- Categorical Features
  - 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().
- Datetime Features
  - From submission date and time columns, we had converted or extracted each day, month, day of week, and hour of the day.
- Derived Features
  - Features derived/calculated from other columns. E.g. word count of the important text columns, total price and number of unique resources
- Text Data Features
  - Combined all the text columns of the dataset into one column.
  - Used WhitespaceTokenizer and WordNetLemmatizer on the entire text.
  - Used tf-idf(term frequency - inverse document frequency) technology. Then the entire column vectorized using compressed sparse matrix.

MODELING

- 3 Models used for this binary classification
    1. Linear Model: LogisticRegression
    2. Naive Bayes Model: MultinomialNB
    3. Decision Tree Model: RandomForest
- Evaluation Metrics
    - Confusion Matrix
    - Average Precision score
    - ROC AUC score
    - F1 - Score

- LogisticRegression
  - Widely used for the binary classification problems.
  - The linear model with the "sklearn" package.
  - The Confusion Matrix for LogisticRegression:
    [[6601 4493]
    [16308 45430]]
- Average Precision for LogisticRegression: **0.9277**
- The ROC AUC Sore: **0.7272**
- The f1 score for the LogisticRegression model: **0.8137**

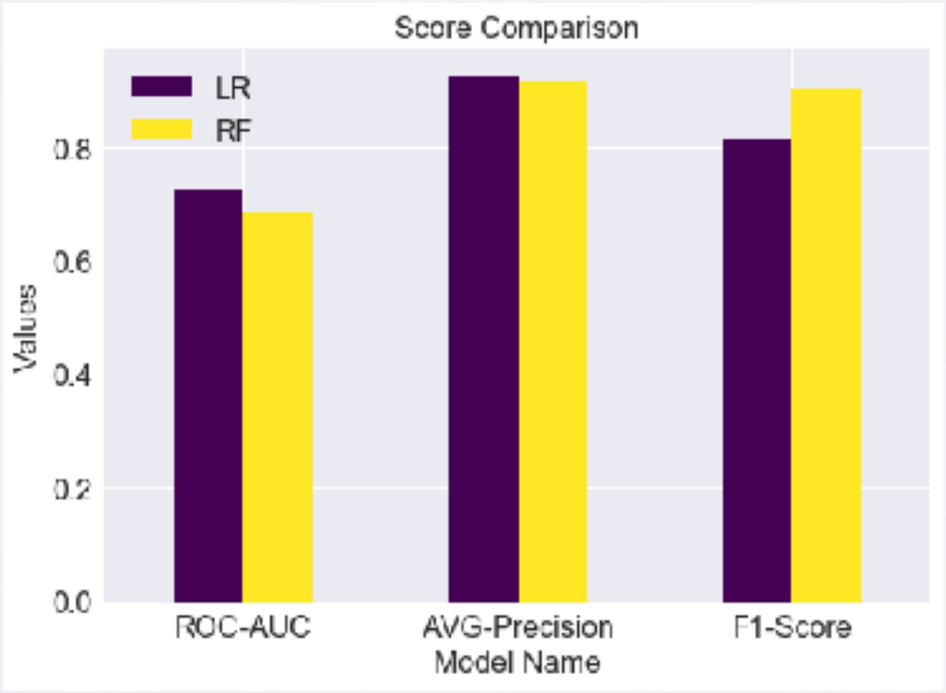
- MultinomialNB
  - Quite similar to the linear models however, they tend to be faster in training.
  - The price paid for this efficiency is that Naive Bayes models often provide generalization performance that is slightly worse than linear classifiers.

- The confusion matrix for MultinomialNB classifier:
  [[6296 4798]
  [28448 33290]]
- The average Precision score for MultinomialNB is: **0.8722**
- The ROC AUC Sore: **0.5716**
- The f1-score for MultinomialNB model is: **0.6669**

- RandomForest
  - Random forests are one way to address the problem of overfitting.
  - Random forests are essentially a collection of decision trees, where each tree is slightly different from the others.
  - The confusion matrix for Random Forest:
    [[1233 9861]
    [2356 59382]]
  - The average precision score for RandomForest is: **0.9173**
  - The ROC AUC Sore: **0.6879**
  - The f1 score for the RandomForest model: **0.9067**
  - From confusion matrix we can see that almost 83% of records were classified correctly.

MODEL COMPARISON AND FINAL PREDICTION

- The MultinomialNB model is the worst performing model among the three models with respect to each of the metrics used.
- As per confusion matrix and f1-score, RandomForest has the better performance.
- Average Precision score and the ROC AUC score suggests that the LogisticRegression model has the better performance.



- Final Prediction
- Used both the LogisticRegression and RandomForest models on real test dataset to predict the classification.
- Created a dataframe having probabilities of both the models together.

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

CONCLUSION

- RandomForest, and LogisticRegression are the two models which are used to predict the real test dataset.

- We have used 4 metrics for model evaluations.

- RandomForest performed better on the 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 the different threshold for approval and rejection condition.

RECOMMENDATIONS

- If the number of applications are huge, we can set the threshold of the probability columns to higher number so that only those project are considered which has more chances of getting approved.

- Teachers can use this model to pre screen if the project they are proposing has the credible chances of getting approved.

PRACTICAL CONSIDERATIONS AND SUGGESTIONS

- Total number of public schools in the U.S. and in each states would give better analysis.

- Financial conditions of students and states would also help in creating better environment for improved learning of the students.

REFERENCES

- Dataset
  - https://www.kaggle.com/c/donorschoose-application-screening/data
- DonorsChoose
  - https://www.donorschoose.org/
- Help
  - https://en.wikipedia.org/wiki/Main_Page
  - https://stackoverflow.com/

THANK YOU

I would like to thank Springboard and my Mentor Srdjan Santic, for his valuable feedback through out this project and keep me motivated to complete the project.