Machine Learning Engineer Nanodegree

Capstone Project

Omar Kamal

December 10th, 2019

I. Definition

Project Overview

Donorschoose.com is an organization that helps teachers funding their projects based on crowd funding through their website. The process is that a teacher submits his project proposal online then one of the workers in Donorschoose.com review it, see if it’s a good project then decides to accept it to be posted on the website or not. Once it’s posted on the website people can donate to it.

My project will be to help them decide if they should accept the project or not. I will use the data provided by DonorsChoose.com through kaggle:
https://www.kaggle.com/c/donorschoose-application-screening/data
I will use train.csv & resources.csv.

Problem Statement

So, as mention above our goal will be to build a model that can decide whether a teacher’s proposal for a project should be approved or not. To do so, I will build a binary classification model that takes in some features then predicts the status of the project.

First, I will build a baseline model using logistic regression & I will use only teacher_number_of_previously_posted_projects feature as It seems to be an important feature. Second, I will try a gradient boosting based model & a neural network based model to see which performs better.

Metrics

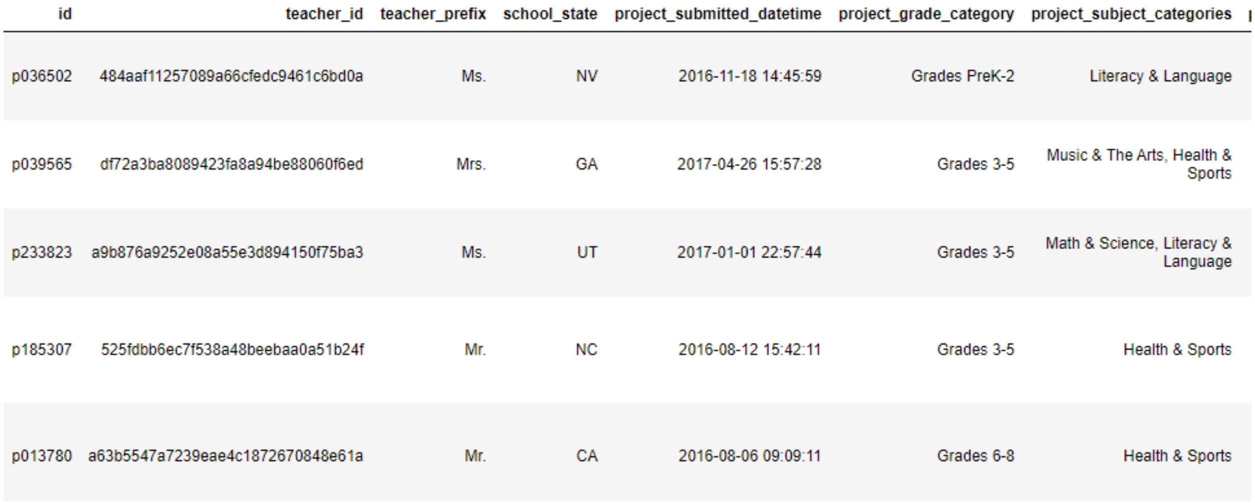
To evaluate our models we will use ROC curve plot & the area under the ROC curve.

A Receiver Operating Characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

II. Analysis

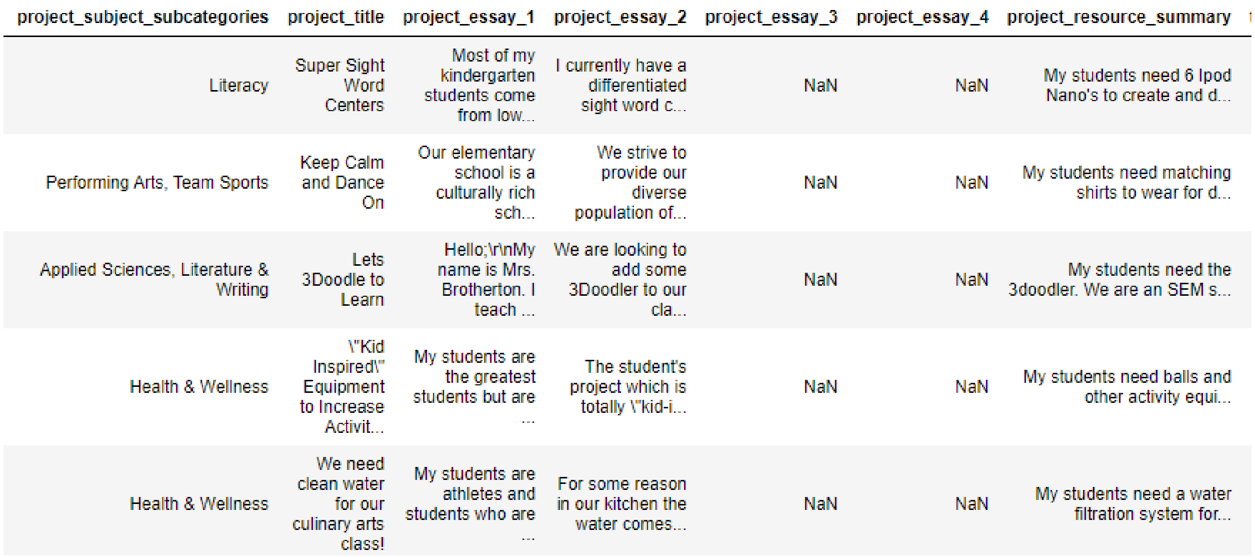
Data Exploration & visualization

First, let's take a look at the train_data:



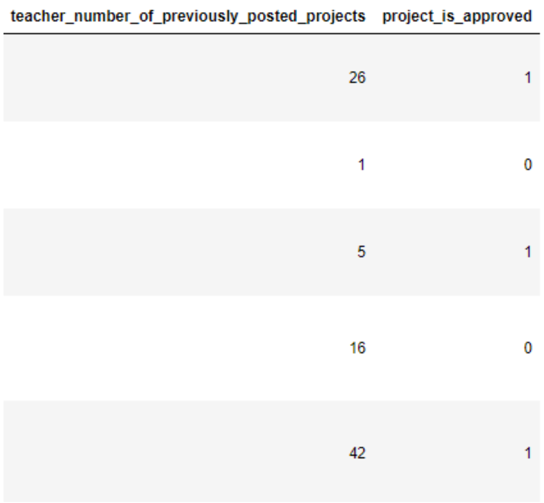
id	teacher_id	teacher_prefix	school_state	project_submitted_datetime	project_grade_category	project_subject_categories
p036502	484aaf11257089a66cfedc9461c6bd0a	Ms.	NV	2016-11-18 14:45:59	Grades PreK-2	Literacy & Language
p039565	df72a3ba8089423fa8a94be88060f6ed	Mrs.	GA	2017-04-26 15:57:28	Grades 3-5	Music & The Arts, Health & Sports
p233823	a9b876a9252e08a55e3d894150f75ba3	Ms.	UT	2017-01-01 22:57:44	Grades 3-5	Math & Science, Literacy & Language
p185307	525fdbb6ec7f538a48beebaa0a51b24f	Mr.	NC	2016-08-12 15:42:11	Grades 3-5	Health & Sports
p013780	a63b5547a7239eae4c1872670848e61a	Mr.	CA	2016-08-06 09:09:11	Grades 6-8	Health & Sports

Figure 1



project_subject_subcategories	project_title	project_essay_1	project_essay_2	project_essay_3	project_essay_4	project_resource_summary
Literacy	Super Sight Word Centers	Most of my kindergarten students come from low...	I currently have a differentiated sight word c...	NaN	NaN	My students need 6 Ipod Nano's to create and d...
Performing Arts, Team Sports	Keep Calm and Dance On	Our elementary school is a culturally rich sch...	We strive to provide our diverse population of...	NaN	NaN	My students need matching shirts to wear for d...
Applied Sciences, Literature & Writing	Lets 3Doodle to Learn	Hello;\r\nMy name is Mrs. Brotherton. I teach ...	We are looking to add some 3Doodler to our cla...	NaN	NaN	My students need the 3doodler. We are an SEM s...
Health & Wellness	\"Kid Inspired\" Equipment to Increase Activit...	My students are the greatest students but are ...	The student's project which is totally \"kid-i...	NaN	NaN	My students need balls and other activity equi...
Health & Wellness	We need clean water for our culinary arts class!	My students are athletes and students who are ...	For some reason in our kitchen the water comes...	NaN	NaN	My students need a water filtration system for...

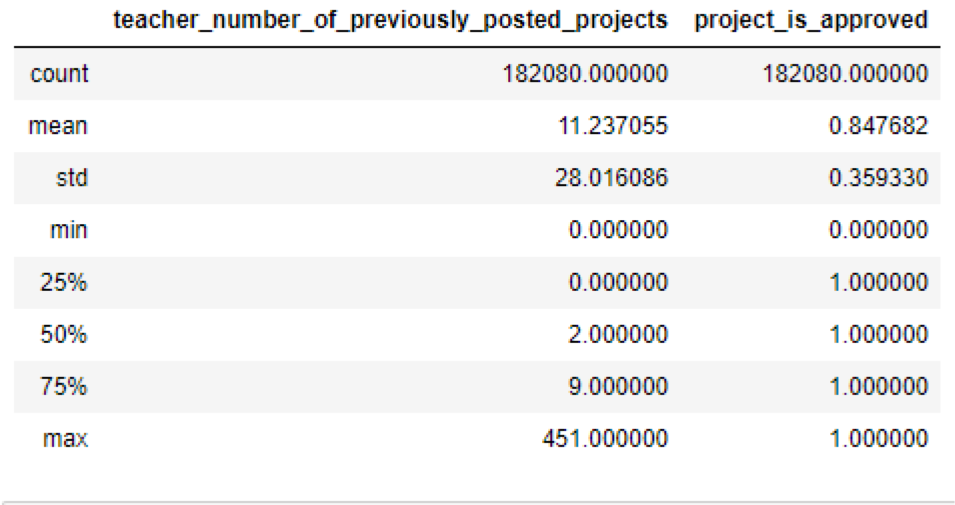
Figure 2



teacher_number_of_previously_posted_projects	project_is_approved
26	1
1	0
5	1
16	0
42	1

Figure 3

We can see that most of our data is either categorical or text so lots of feature engineering will probably be required here. Also we can see that our target is project_is_approved . Let’s take a further look:



	teacher_number_of_previously_posted_projects	project_is_approved
count	182080.000000	182080.000000
mean	11.237055	0.847682
std	28.016086	0.359330
min	0.000000	0.000000
25%	0.000000	1.000000
50%	2.000000	1.000000
75%	9.000000	1.000000
max	451.000000	1.000000

Figure 4

From the mean we can see that our dataset is highly imbalanced almost 85% of the projects are approved.

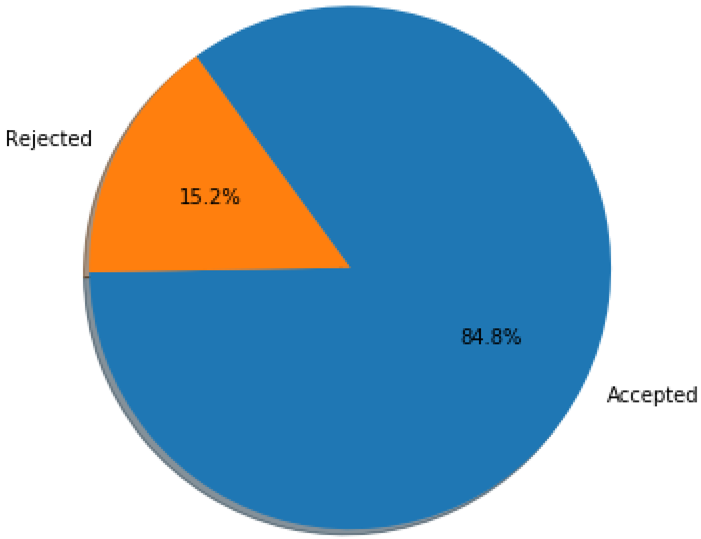
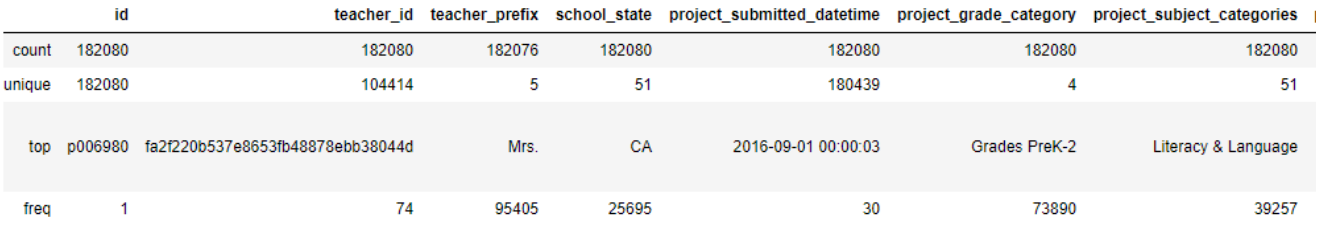
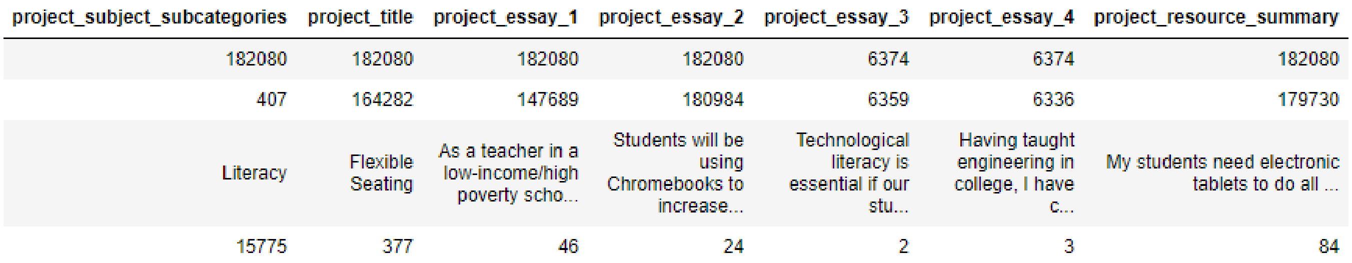


Figure 5



	id	teacher_id	teacher_prefix	school_state	project_submitted_datetime	project_grade_category	project_subject_categories
count	182080	182080	182076	182080	182080	182080	182080
unique	182080	104414	5	51	180439	4	51
top	p006980	fa2f220b537e8653fb48878ebb38044d	Mrs.	CA	2016-09-01 00:00:03	Grades PreK-2	Literacy & Language
freq	1	74	95405	25695	30	73890	39257

Figure 6

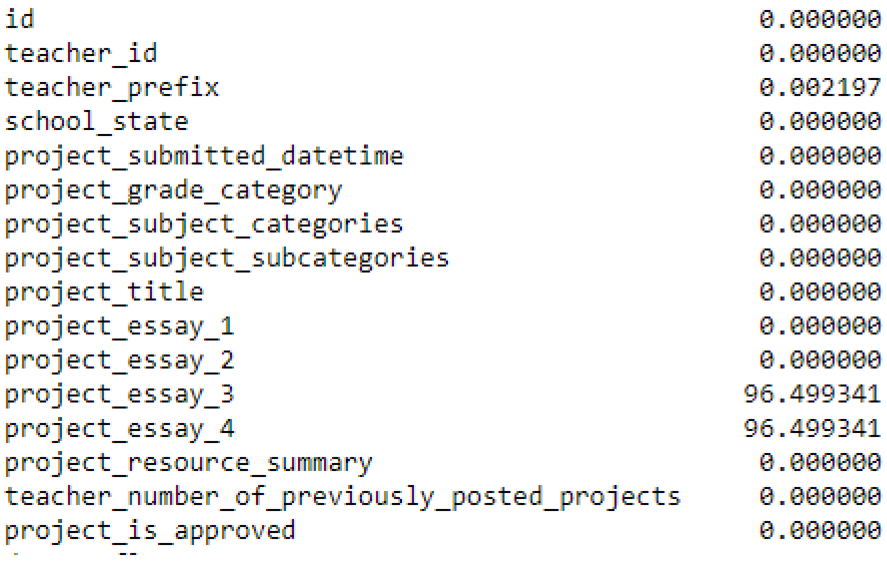


project_subject_subcategories	project_title	project_essay_1	project_essay_2	project_essay_3	project_essay_4	project_resource_summary
182080	182080	182080	182080	6374	6374	182080
407	164282	147689	180984	6359	6336	179730
Literacy	Flexible Seating	As a teacher in a low-income/high poverty scho...	Students will be using Chromebooks to increase...	Technological literacy is essential if our stu...	Having taught engineering in college, I have c...	My students need electronic tablets to do all ...
15775	377	46	24	2	3	84

Figure 7

We have categorical features with low number of categories like teacher_prefix & project_grade_category, those with moderate number of categories like school_states & project_subject_categories, with high number of categories like project_subject_subcategories and those with very high number of categories like teacher_id. We also have a datetime feature which we will need to deal with.

In the resources data we can find the quantity & the price of an item requested by a proposal. One proposal may have multiple entries in the resources table, they are linked by the id column.



```
id                                              0.000000
teacher_id                                      0.000000
teacher_prefix                                  0.002197
school_state                                    0.000000
project_submitted_datetime                      0.000000
project_grade_category                          0.000000
project_subject_categories                      0.000000
project_subject_subcategories                   0.000000
project_title                                   0.000000
project_essay_1                                 0.000000
project_essay_2                                 0.000000
project_essay_3                                96.499341
project_essay_4                                96.499341
project_resource_summary                        0.000000
teacher_number_of_previously_posted_projects    0.000000
project_is_approved                             0.000000
```

Figure 8

We can see that essay 3 & 4 contain 96.5% NaN values & all other features nearly contain 0 NaNs.

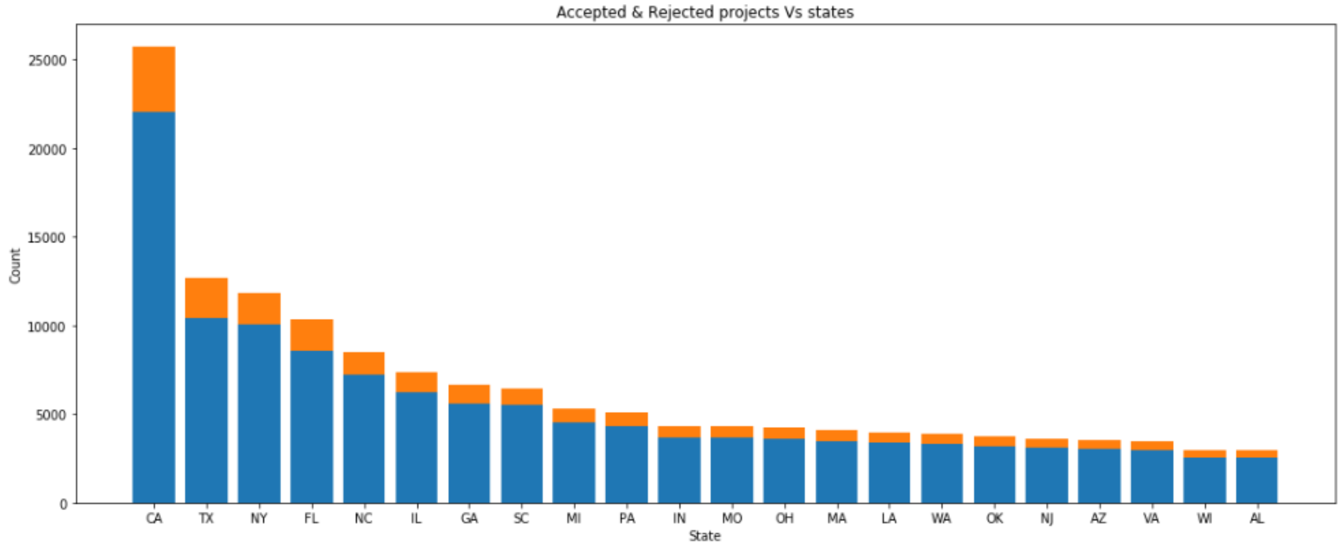


Figure 9

We can see that large number of proposals come from California. We can also see that in all states the approval dominates the rejection.

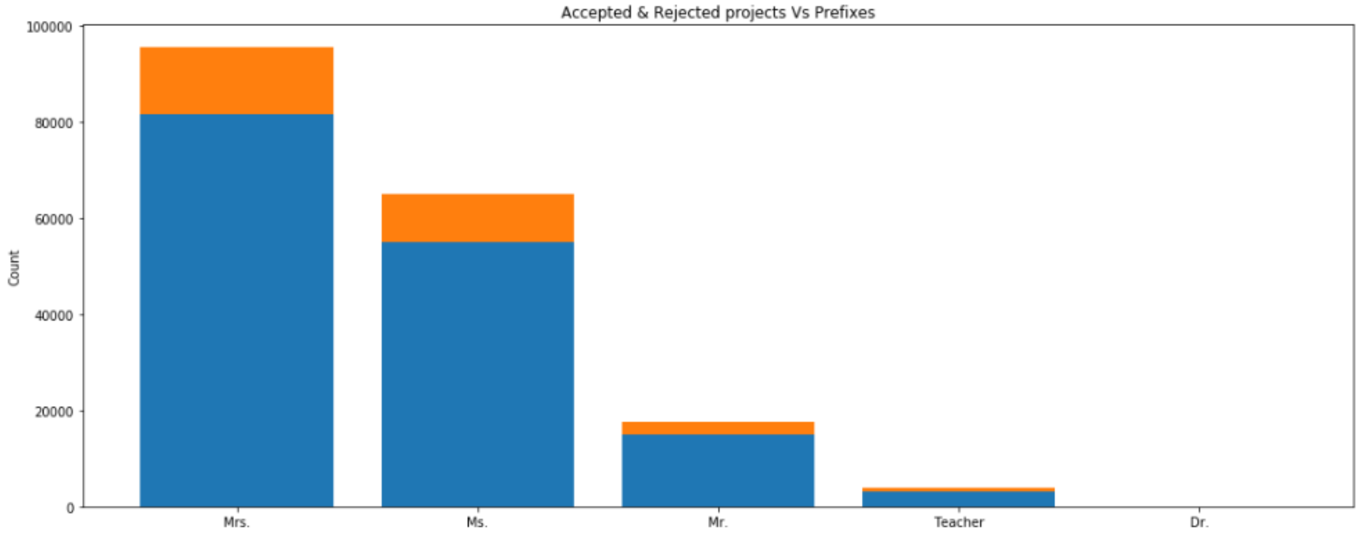


Figure 10

Most of proposals are submitted by females. We can also see that for all titles the approval dominates the rejection & that there are nearly no proposals by a Dr..

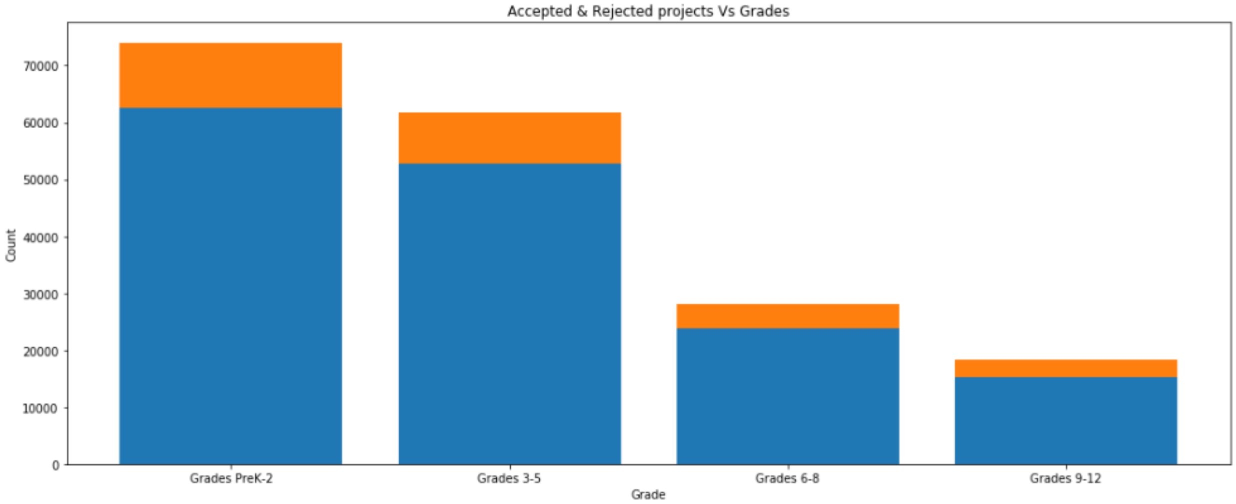


Figure 11

The number of proposals declines as the grade of the students increase.

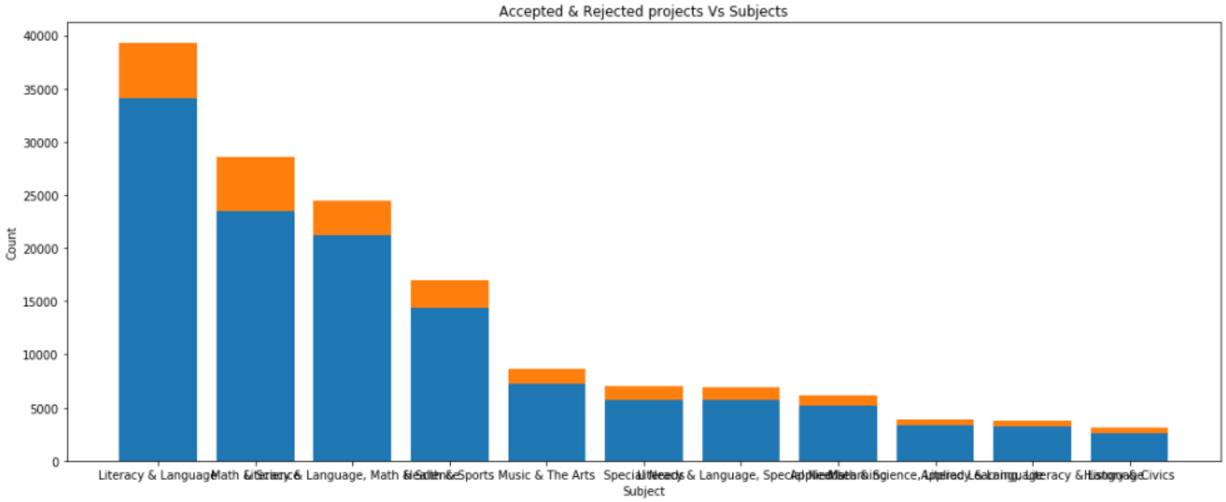


Figure 12

People care more about Literacy & Languages than Maths & science.

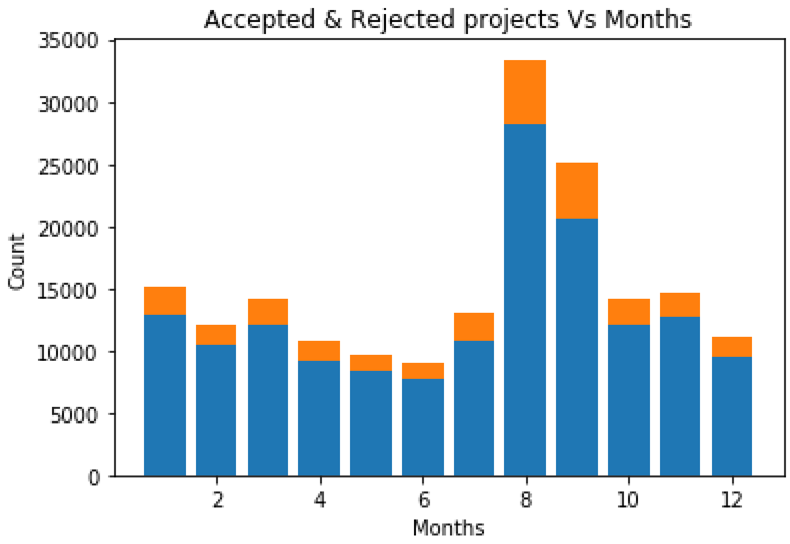


Figure 13

Number of proposals peaks in August & September.

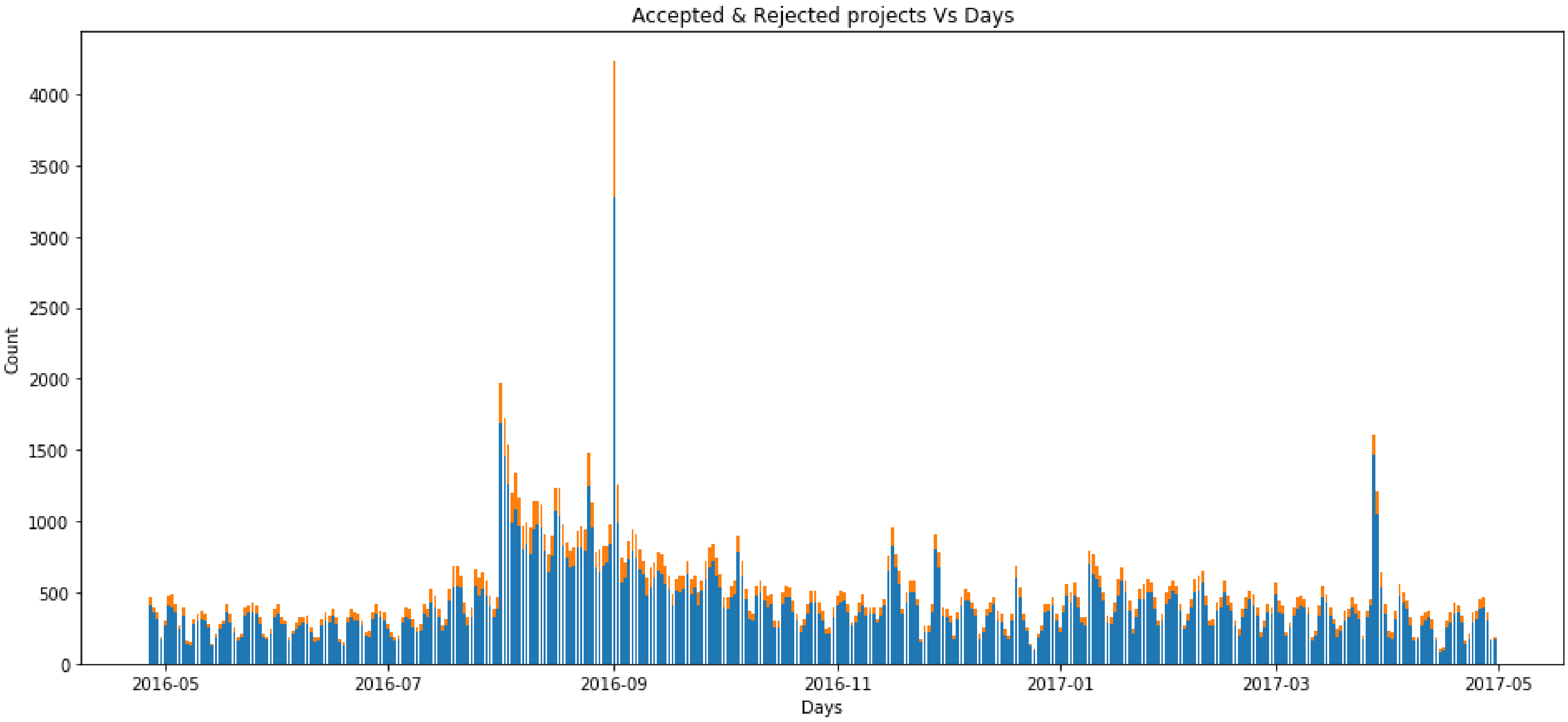


Figure 14

We have data for a full year.

Algorithms and Techniques

From the above analysis we can see that we need an algorithm that can deal with imbalanced data, categorical data & textual data well. Before exploring the data my plan was to try gradient boosting algorithms and neural networks algorithms but now due to the above mentioned nature of the dataset I think neural networks will be a poor choice so I will go with a gradient boosting algorithm.

After checking the available gradient boosting algorithms I decided to go with lightgbm.

LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel and GPU learning.
- Capable of handling large-scale data.

Benchmark

As a benchmark, I will use logistic regression model with input feature teacher_number_of_previously_posted_projects. It gives an AUC value of 0.56 & plots the following ROC curve

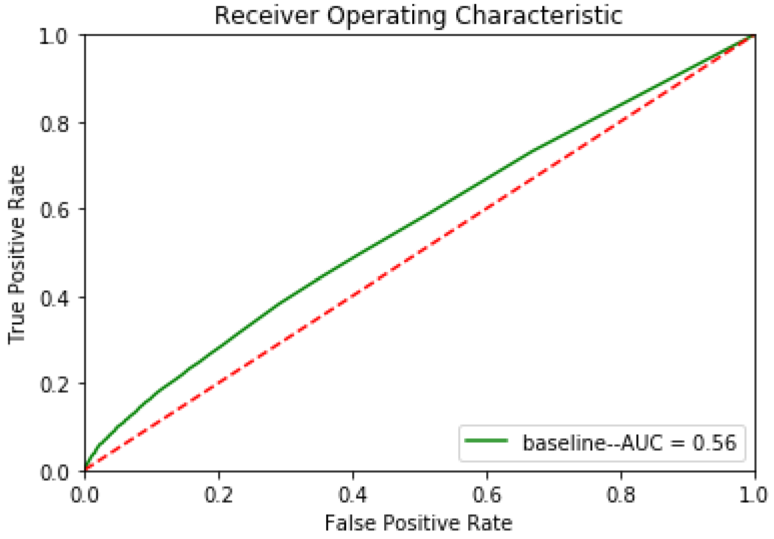


Figure 15

III. Methodology

Data Preprocessing

We will handle categorical data in 2 approaches:

- Low no. of categories: One-hot encoding.
- Moderate & high no. of categories: category frequency (we replace each category with its count in the whole data).

For the textual data we will extract the following:

- Length.
- Word count.
- Univariate Word density.
- Bivariate Word density.

For the datetime we will break it to:

- Year
- Month
- dayOfWeek
- hour.

Then we will treat them as categories. To avoid any overlap between extracted features we add prefixes.

Implementation

Most of the work done was in the feature engineering part, first I extracted the year, month, weekday & hour from the project_submitted_datetime. Then the length & word count of essay 1, essay 2, project title & project_resource_summary.

Then I built some custom functions to process the text features to extract the Univariate Word density, Bivariate Word density. Let's take essay 1 for example, I concatenate all the entries in it into two strings, one for approved proposals & one for rejected proposals, then I clean it from special characters & stopping words (common words that mostly conveys no meaning) then I calculate the densities of all words in it, choose the top 10 words for the approved entries & the top 10 words for the rejected entries and use them to extract new stopping words (words that exist in both lists with similar densities so they provide few information) then I use that list to get new top 20 words for approved and 20 for rejected that doesn't contain those words then I merge the the two lists removing any duplicates and them as features to the data. I did that for both univariate & bivariate word densities & for all above mentioned text columns.

For the model I used LGBMClassifier. I started with the default parameters, then I started manually tuning the num_leaves, learning_rate, n_estimators & max_depth to optimize the model between overfitting & underfitting till I reached the current state.

Refinement

At first, I didn't use Bivariate density words & I didn't take in consideration that redundant words like students may be in the top 10 words but they convey no information gain.

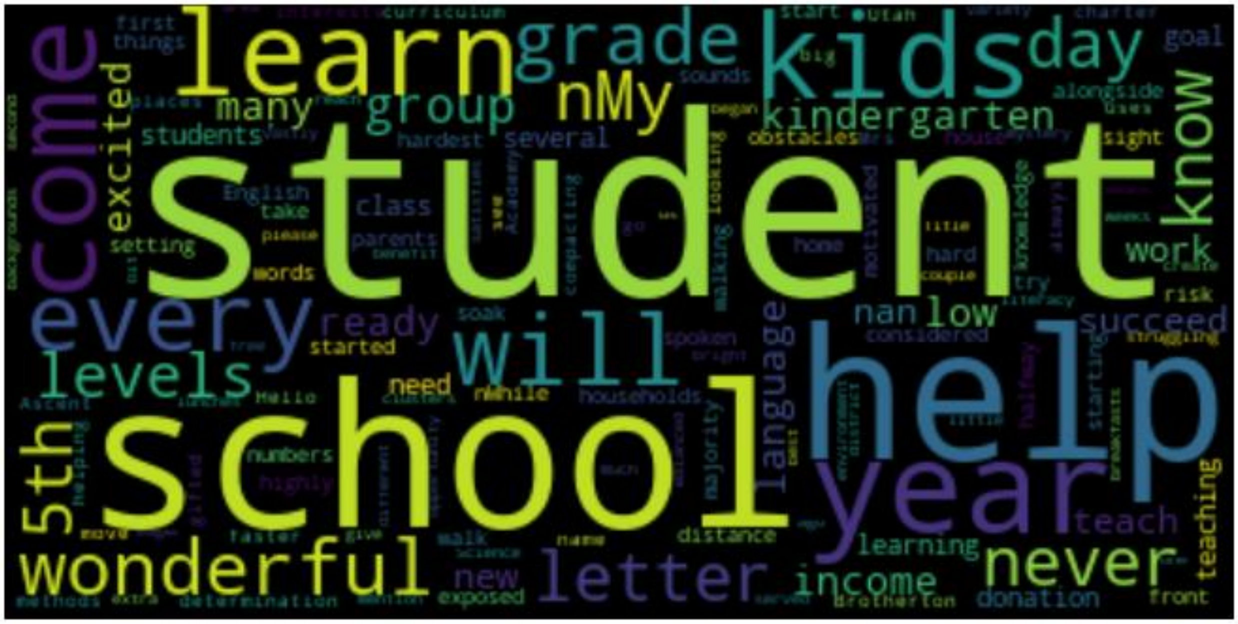


Figure 16 – Word cloud for essay 1 in Approved proposals

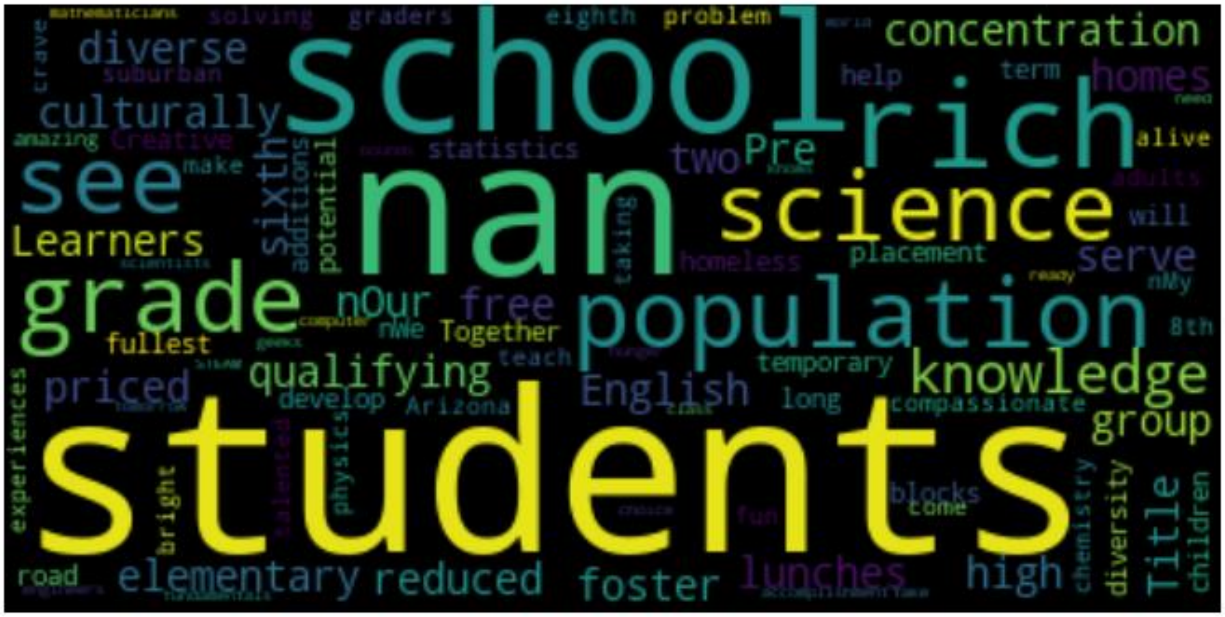


Figure 17 -- Word cloud for essay 1 in Rejected proposals

But then as mentioned above I fixed that. Also as mentioned above when I went with the default parameters of LGBMClassifier I didn’t get good results so I had to go through each parameter and learn how it affects the model and started tuning them manually. It would have been better to go with gridsearchCV but it would take lots of time & computing power which I can’t afford.

IV. Results

Model Evaluation and Validation

The data was split randomly into 20% testing & 80% training to ensure robustness & generalization. LGBMClassifier scored 0.7428 AUC, 0.993 recall score & 0.85 precision score which is considered a great solution given the ambiguous nature of the problem. So when the model rejects a proposal it’s completely trusted, however accepted proposals may need some more checking.

Justification

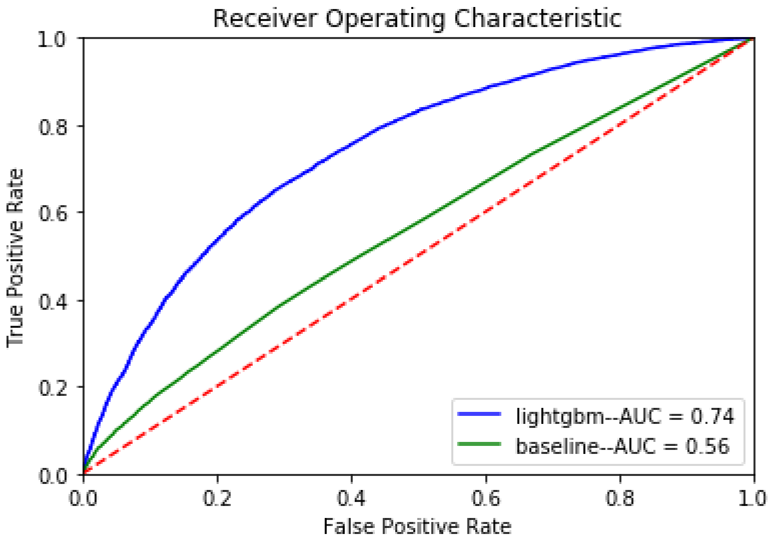


Figure 18

We can clearly see that my solution outperforms the benchmark model.

V. Conclusion

This is only a brief summary, for more details please go through my notebook.html.

- Data is highly imbalanced that is approx. 85 % projects were approved and 15 % projects were not approved. Majority imbalanced class is positive.
- Out of 50 states, **California (CA)** having higher number of projects proposal submitted **approx. 14 %** followed by **Texas (TX) (7 %)** and **Tennessee (NY) (7 %)**.
- Out of 4 school grade levels, Project proposals submission in school grade levels is higher for **Grades Prek-2** which is approximately **41 %** followed by **Grades 3-5** which has approx. **34 %**.
- Out of 51 Project categories, Project proposals submission for project categories is higher for **Literacy & Language** which is approx. **21.5 %** followed by **Math & Science** which has approx. **15.7 %**.
- Higher number of project proposal submitted by **married women** which is approx. **53 %** followed by **unmarried women** which has approx. **37 %**.

- Project proposal submitted by **Teacher** which is approx. **2 %** is very low compared to **Mrs., Ms., Mr**.
- If Number of previously posted applications by the submitting teacher was **Zero,** we have more acceptance rate.
- Female having more count which is approx. **88 %** than Male which has **10 %** in terms of projects proposals submissions.
- Most of the prices requested for resources are from 0 to 2k dollar.
- If price per project is low, then we have more chances of approval.
- Projects with less number of items have better chances of approval.
- Project Submission Time Analysis :
  - **September month** has the second highest number of proposals.
  - The number of proposals decreases as we move towards the end of the week.
  - Looks like we have approximately one years' worth of data (May 2016 to April 2017) given in the training set.
  - There is a sudden spike on a single day (Sep 1, 2016) with respect to the number of proposals (may be some specific reason related to the schools season?)

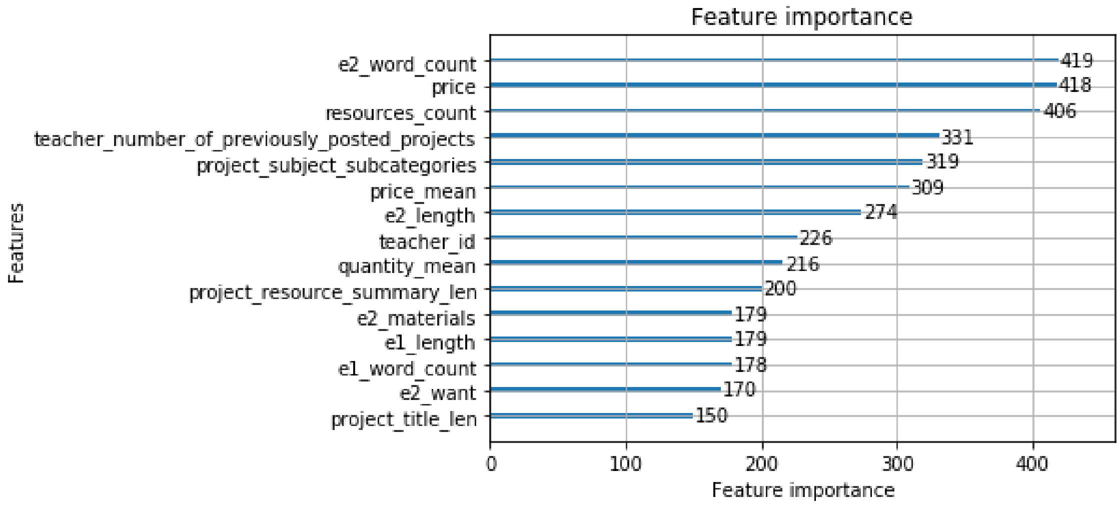


Figure 19

Above is the most important 15 features as extracted by the lightgbm. We can see that my assumption in building the benchmark model that teacher_number_of_previously_posted_projects is an important feature is true, also we can see that I synthesized lots of important features, some of them even more important than teacher_number_of_previously_posted_projects.

Improvement

To improve our solution further we can add a new feature to the data which is the prediction of the model then we separate the data that was predicted to be accepted, then use this data to train another lgbm model. So we now have a pipeline of two models. Also we can do gridsearchCV for both of them to optimize the hyperparameters.