Georgia State University

CSC4780/6780– Fundamentals of Data Science

[Spring 2022]

Project Progress Report

[Group Name]: Inferno

Hemanth Chebrolu (002-67-7427)

Ravi Teja Tatikonda (002-67-3746)

Someswara Rao Gollangi(002-67-7428)

Sri Latha Jammigumpula (002-67-6916)

Supriya Parvatham (002-67-1494)

Table of Contents

1. Business Understanding

DonorsChoose.org receives a number of project proposals each year for classroom projects that are in need of funding. Right now, we need a large number of volunteers. These volunteers must manually screen each submission before it's approved to be posted on the DonorsChoose.org website. Next year there will be around 500,000 applications that need to be screened manually. It is not economically scalable to hire all these volunteers and screen the projects manually.

1.1 Business problem

We have three major problems that we need to resolve.

- We must scale the current manual process and the resources. We must automate the entire process so that the projects can be screened quickly and efficiently so that they can be posted on DonorsChoose.org as soon as possible.
- We must increase the consistency of the project vetting across different volunteers to improve the experience for teachers
- We need to focus the energy of the volunteers on the applications that need the most assistance.

1.2 Dataset:

- The dataset we used to predict whether a DonorsChoose.org project proposal submitted by a teacher will be approved is from Kaggle.
- We are using two data sets which we will be using together for our project. Train.csv and resources.csv. There are 15 descriptive features in train.csv and 3 descriptive features resources.csv. Many projects require multiple resources. The 'id' value in resources.csv corresponds to 'project_id' in train.csv and it will be used as a key to retrieve resources.
- There are a total of 109248 rows in our dataset
- 'project_is_approved' attribute is our target variable

1.3 Proposed Analytics Solution

We plan on applying three different classifiers to our proposed project namely K-Nearest Neighbor classifier, Naïve Bayes Classifier, Linear regression classifier. After training the data with all the three classifiers we will then choose the one that has highest accuracy for our data.

2. Data Exploration and Preprocessing

2.1 Data quality Report

The dataset that we chose has 17 features out of which 3 of them are Continuous and 14 of them are Categorical. Most of these categorical features are string data type and some are nominal data scale and some of them are ordinal data scale. 2 out of 3 continuous features have int data type and one is float datatype.

Data quality report for Categorical Variables:

DATA QUALITY REPORT FOR CATEGORICAL VARIABLES:

	Feature	Count	% of Missing	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode Perc
0	id	109248	0.00	109248	p000002	1	0.00	p189294	1	0.00
1	teacher_id	109248	0.00	72168	fa2f220b537e8653fb48878ebb38044d	44	0.04	1f64dcec848be8e95c4482cc845706b2	42	0.04
2	teacher_prefix	109248	0.00	5	Mrs.	57269	52.42	Ms.	38955	35.66
3	school_state	109248	0.00	51	CA	15388	14.09	TX	7396	6.77
4	project_submitted_datetime	109248	0.00	90885	9/1/2016 0:00	227	0.21	9/1/2016 0:01	121	0.11
5	project_grade_category	109248	0.00	4	Grades PreK-2	44225	40.48	Grades 3-5	37137	33.99
6	project_subject_categories	109248	0.00	51	Literacy & Language	23655	21.65	Math & Science	17072	15.63
7	project_subject_subcategories	109248	0.00	401	Literacy	9486	8.68	Literacy, Mathematics	8325	7.62
8	project_title	109248	0.00	100850	Flexible Seating	234	0.21	Wiggle While You Work	93	0.09
9	project_essay_1	109248	0.00	94319	As a teacher in a low-income/high poverty scho...	32	0.03	I teach a special day class that is filled wit...	21	0.02
10	project_essay_2	109248	0.00	108831	Students will be using Chromebooks to increase...	15	0.01	Students will be using Chromebooks to increase...	5	0.00
11	project_essay_3	109248	96.56	3755	Daily, students will check out their chrome bo...	2	0.00	I love to make learning fun for my students, h...	2	0.00
12	project_essay_4	109248	96.56	3750	As I analyze this year's student data, compreh...	2	0.00	This project will help my students to develop ...	2	0.00
13	project_resource_summary	109248	0.00	108323	My students need electronic tablets to do all ...	48	0.04	My students need Chromebooks to do all the thi...	15	0.01

Data quality report for Continuous Variables:

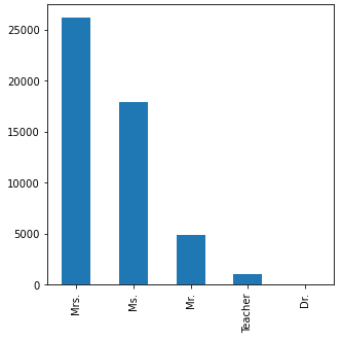
DATA QUALITY REPORT FOR CONT VARIABLES:

	Feature	Count	Cardinality	% of Missing	Min	Max	Mean	Median	Std Dev	Q1	Q3
0	teacher_number_of_previously_posted_projects	109248	374	0.0	0.00	451.0	11.153165	2.00	27.777154	0.00	9.0
1	price	109248	50675	0.0	0.66	9999.0	298.119343	206.22	367.498030	104.31	379.0
2	quantity	109248	332	0.0	1.00	930.0	16.965610	9.00	26.182942	4.00	21.0

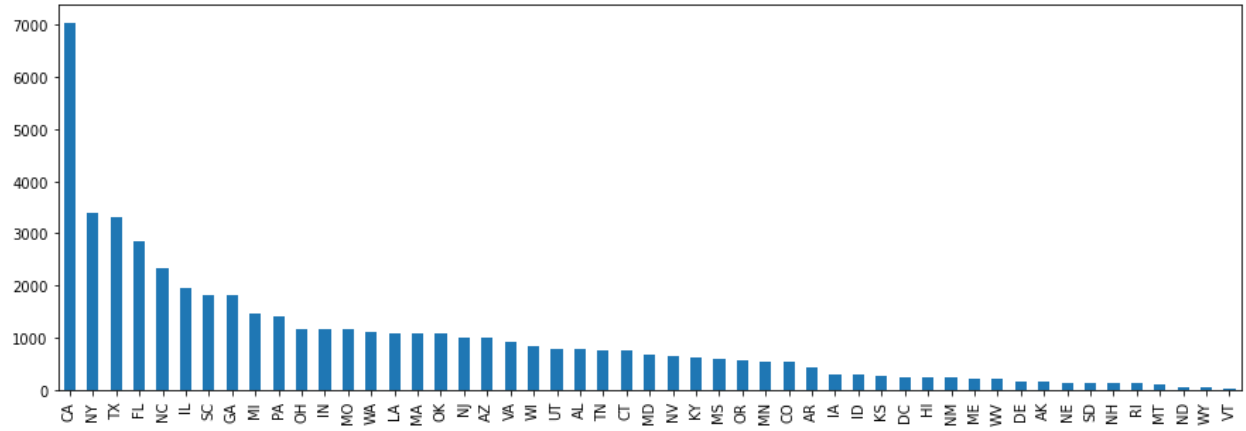
Figure 1. Visualizations of Categorical Features in Dataset:

Bar Plots:

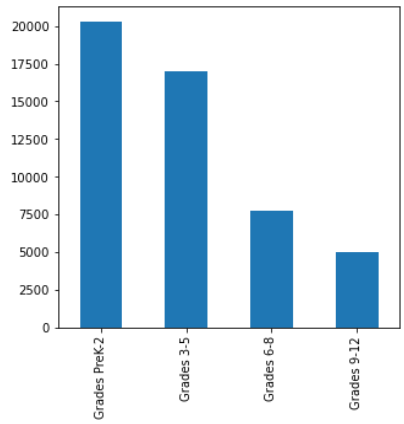
teacher_prefix



school_state



project_grade_category:



project_subject_categories

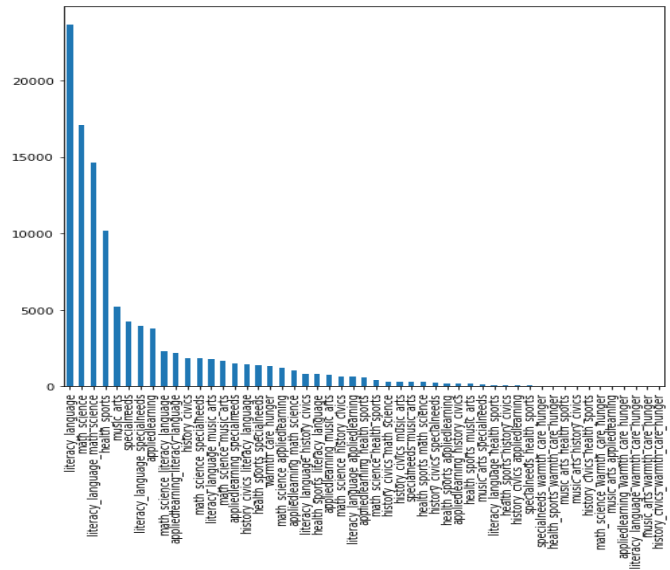
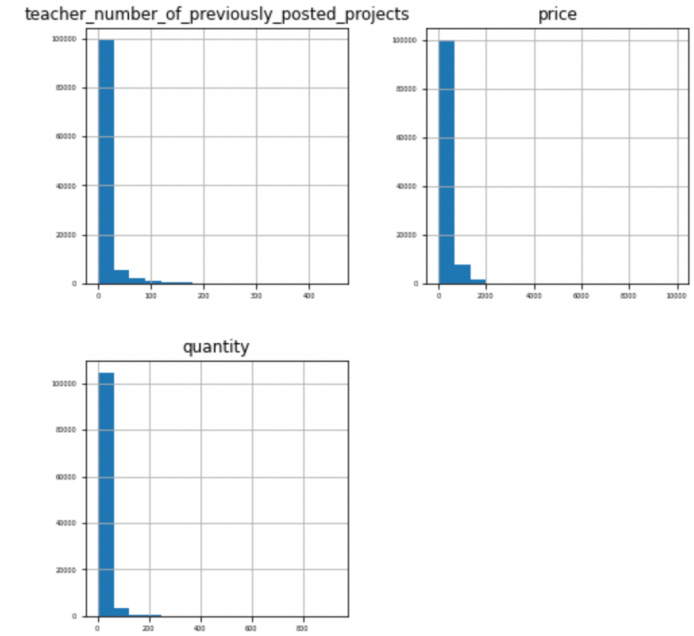


Figure 2. Visualizations of Continuous Features in Dataset:

Histograms:

```
array([[<AxesSubplot:title={'center':'teacher_number_of_previously_posted_projects'}>,
        <AxesSubplot:title={'center':'price'}>],
       [<AxesSubplot:title={'center':'quantity'}>, <AxesSubplot:>]],
      dtype=object)
```



2.2 Missing values and Outliers:

- In the data set we chose, there are missing value in the categorical feature named "**teacher_number_of_previously_posted_projects".** There are only 3 missing values in this feature and we have removed these missing values by replacing them with the element that repeated the most "Mrs."
- We removed outliers for all the 3 continuous variables. Lower range Upper Range and InterQuartile Range(IQR). For teacher_number_of_previously_posted_projects Lower Range: -13.5 Upper Range: 22.5 IQR: 9.0. For price Lower Range: -307.72499999999997 Upper Range: 791.035 IQR: 274.69. For quantity Lower Range: -21.5 Upper Range: 46.5 IQR: 17.0

2.3 Normalization:

- We have used Minmax scaler from sklearn library. Minmax scaler normalizes all values in the dataset to values in the range -1 to 1.
- The Minmax scaler applied on 3 attributes namely price,quantity, teacher_number_of_previously_posted_projects

2.4 Transformations:

- We have applied Ordinal encoding to all the 14 categorical features in the dataset.

- We have used ordinal encoder from the sklkearn library to encode the categorical features