Capstone Project Proposal
Transforming Education System

1. Domain Background:

As the cost of education per student is increasing in US still it lags behind the global competitors in educational outputs. For the last many years, there is almost no correlation between dollars spending per student and their performance. This may be because the money has been spent on the areas which does not contribute or add to the student achievement and performance. This problem can be solved by finding a better way to distribute the resources for schools and districts and clarifying their spending habits. And finally implementing strategies to use money on education in a more effectively and efficient way.

This problem has been actively handled by ERS, a non-profit consulting firm established in 2004. Since 2004, They have been working with more than 20 school systems across US, including 16 of the 100 largest urban districts, on topics such as teacher compensation and career path, funding equity, school design, central office support, and budget development. In the last few years they have partnered with states as well. So, in this project I will use the dataset provided by ERS to create a strategy on how to better spend the money to improve the education system.

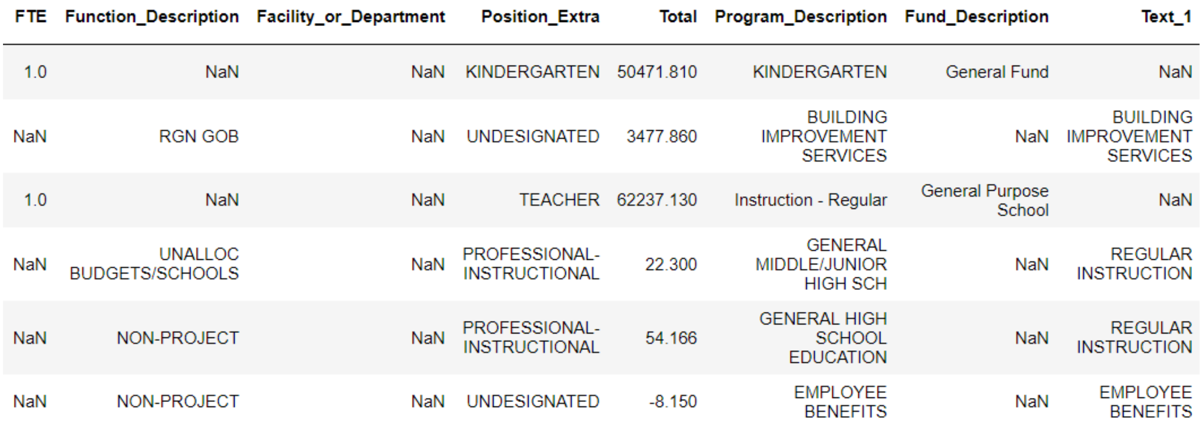
2. Problem Statement:

Budgets for schools and school districts are huge, complex, and unwieldy. It's not easy task to find out where and how schools are using their resources. ERS is a non-profit organization which tackles just this task with the goal of letting districts be smarter, more strategic, and more effective in their spending. So, the problem statement in this project **to solve a multi-class-multi-label classification problem with the goal of attaching canonical labels to the free form text in budget line items** (I will explain the predictor and target variables in details under the section **Datasets and Inputs**). These labels let ERS understand how schools are spending money and tailor their strategy recommendations to improve outcomes for students, teachers, and administrators.

3. Datasets and Inputs

The aim of this project is to predict the probability that a certain label is attached to a budget line item. Each row in the budget has mostly text features, except for the two inputs (FTE and Total) that are noted as float. Any of the fields may or may not be empty and if empty then the values are written as NaN.

The training dataset contains 400277 rows and 26 columns of which only two variables viz FTE and Total are numerical in nature while all other variables are text. Below is a screenshot for the training dataset. Below is a screenshot showing some of the variables and their values in the dataset.



FTE	Function_Description	Facility_or_Department	Position_Extra	Total	Program_Description	Fund_Description	Text_1
1.0	NaN	NaN	KINDERGARTEN	50471.810	KINDERGARTEN	General Fund	NaN
NaN	RGN GOB	NaN	UNDESIGNATED	3477.860	BUILDING IMPROVEMENT SERVICES	NaN	BUILDING IMPROVEMENT SERVICES
1.0	NaN	NaN	TEACHER	62237.130	Instruction - Regular	General Purpose School	NaN
NaN	UNALLOC BUDGETS/SCHOOLS	NaN	PROFESSIONAL-INSTRUCTIONAL	22.300	GENERAL MIDDLE/JUNIOR HIGH SCH	NaN	REGULAR INSTRUCTION
NaN	NON-PROJECT	NaN	PROFESSIONAL-INSTRUCTIONAL	54.166	GENERAL HIGH SCHOOL EDUCATION	NaN	REGULAR INSTRUCTION
NaN	NON-PROJECT	NaN	UNDESIGNATED	-8.150	EMPLOYEE BENEFITS	NaN	EMPLOYEE BENEFITS

Below are the data dictionary of some of the variables -

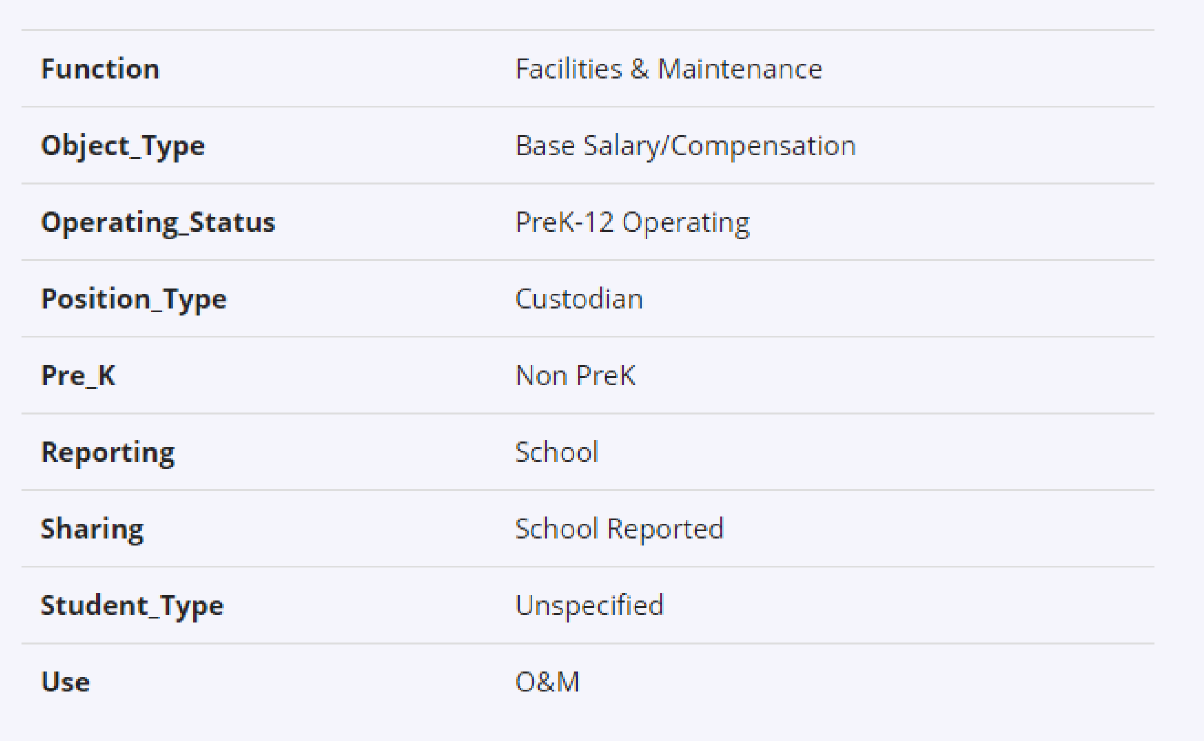
- **FTE**- If an employee, the percentage of full-time that the employee works.
- **Facility_or_Department**- If expenditure is tied to a department/facility, that department/facility.
- **Function_Description**- A description of the function the expenditure was serving.
- **Fund_Description**- A description of the source of the funds.
- **Job_Title_Description**- If this is an employee, a description of that employee's job title.
- **Location_Description**- A description of where the funds were spent.
- **Object_Description**- A description of what the funds were used for.
- **Position_Extra**- Any extra information about the position that we have.
- **Program_Description**- A description of the program that the funds were used for.
- **SubFund_Description**- More detail on **Fund_Description**
- **Sub_Object_Description**- More detail on **Object_Description**
- **Text_1**- Any additional text supplied by the district.
- **Text_2**- Any additional text supplied by the district.
- **Text_3**- Any additional text supplied by the district.
- **Text_4**- Any additional text supplied by the district.
- **Total**- The total cost of the expenditure.

Whereas the labels (target variable) for this problem are the one label from each of 9 different categories–

- Function (contains 37 different Labels)
- Object_Type (contains 11 different Labels)
- Operating_Status (contains 3 different Labels)
- Position_Type (contains 25 different Labels)
- Pre_K (contains 3 different Labels)
- Reporting (contains 3 different Labels)
- Sharing (contains 5 different Labels)

- Student_Type (contains 9 different Labels)
- Use (contains 8 different Labels)

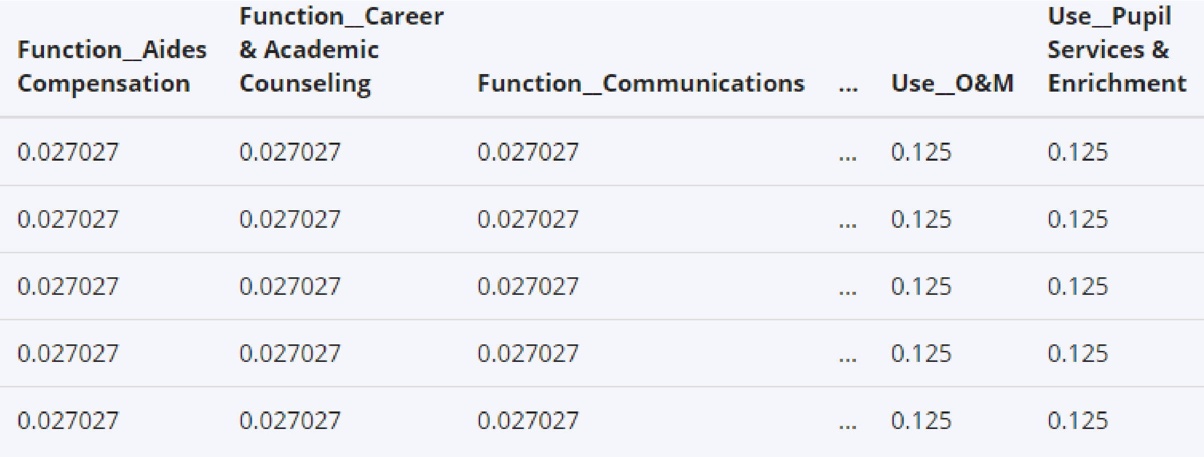
Therefore, there is a total of (37 + 11 + 3 + 25 + 3 + 3 + 5 + 9 + 8 = 104 Labels). And also, there is a hierarchical relationship for these labels. For example, if a line is marked as Non-Operating in the Operating_Status category, then all of the other labels should be marked as **NO_LABEL**. A target variable (label) example is-



Function	Facilities & Maintenance
Object_Type	Base Salary/Compensation
Operating_Status	PreK-12 Operating
Position_Type	Custodian
Pre_K	Non PreK
Reporting	School
Sharing	School Reported
Student_Type	Unspecified
Use	O&M

4. Solution Statement:

In the solution to the problem I will predict the probabilities of each class in 9 different categories and sample output for the problem is –



Function__Aides Compensation	Function__Career & Academic Counseling	Function__Communications	...	Use__O&M	Use__Pupil Services & Enrichment
0.027027	0.027027	0.027027	...	0.125	0.125
0.027027	0.027027	0.027027	...	0.125	0.125
0.027027	0.027027	0.027027	...	0.125	0.125
0.027027	0.027027	0.027027	...	0.125	0.125
0.027027	0.027027	0.027027	...	0.125	0.125

So, the expected output for the problem is the probability for all different labels (a total of 104 Labels) from 9 different categories.

5. Benchmark Model:

There is no benchmark model available as of now. This problem is a part of Machine Learning Competition that I am participating in. So, if I need to create a benchmark just by random guessing then I will give equal probabilities for each label within 9 different categories and that would give me

- Label probabilities for variable **Function =** 1/37 = 0.027
- Label probabilities for variable **Object_Type =** 1/11 = 0.091
- Label probabilities for variable **Operation_Status =** 1/3 = 0.33
- Label probabilities for variable **Position_Type =** 1/25 = 0.04
- Label probabilities for variable **Pre_K =** 1/3 = 0.33
- Label probabilities for variable **Reporting =** 1/3 = 0.33
- Label probabilities for variable **Sharing =** 1/5 = 0.20
- Label probabilities for variable **Student_Type =** 1/9 = 0.11
- Label probabilities for variable **Use =** 1/8 = 0.125

Using these equal probabilities for all the labels across 9 different categories we will get the log loss as **2.0455** (which is my benchmark model) and my aim will be reduce the log loss in comparison to the benchmark log loss by creating an improved model.

6. Evaluation Metric:

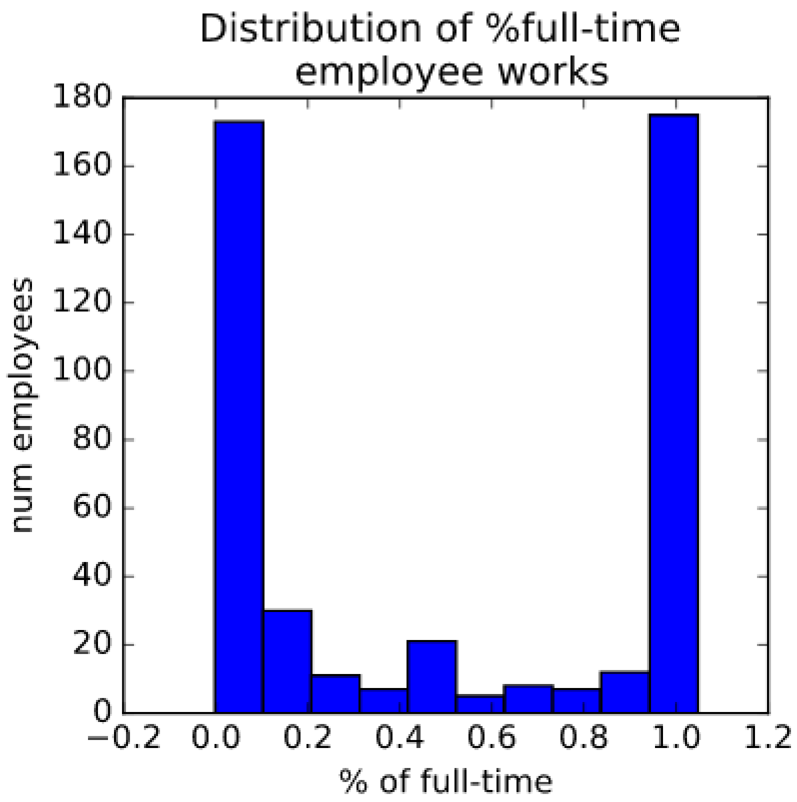
The evaluation metric that I will use in this project is the log_loss. This is the loss function used in (multinomial) logistic regression and it is also used in other ML algorithms such as neural networks, defined as the negative log-likelihood of the true labels given a probabilistic classifier's predictions. The log loss is only defined for two or more labels. For a single sample with true label yt in {0,1} and estimated probability yp that yt = 1, the log loss is-

$$-\log P(yt|yp) = -(yt \log(yp) + (1 - yt) \log(1 - yp))$$

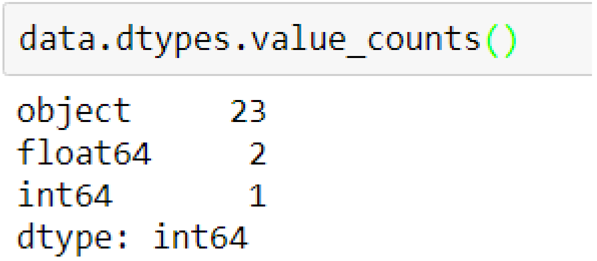
7. Project Design:

Below is the description of the project as a whole

1. **Loading the Data:** Dataset is loaded as a first step.
2. **Exploratory Data Analysis**: Here I will do some EDA on the dataset to know the relationship between the variables. For example, below distribution shows number of fulltime employees in different schools. It looks like the FTE column is bimodal. That is, there are some part-time and some full-time employees in those schools. Similar insights can be drawn for the other variables.



3. **Data Pre-Processing**: Here all different data pre-processing techniques will be used. For example, if we see the data types of all the variables in the dataset we get to see the following



```
data.dtypes.value_counts()
```

```
object     23
float64     2
int64       1
dtype: int64
```

i.e. all the text columns have been stored as an object which we need to convert them into categorical variables.

Also, I will use text tokenization and create bag of words for all the text data in 23 text variables which will be required for model building.

4. **Missing Value Treatment:** I will also use Imputer() Imputation transformer to fill in the missing values in the data.

5. **Model Building:** I will use the Pipeline feature in sklearn module to build the model which will create Machine Learning pipeline for the whole problem at once including missing value treatment till applying a specified classification algorithm to the dataset.

I will use different classification techniques like Logistic Regression, Decision Trees and Random Forest etc.

6. **Parameter Tuning:** Once the model is build I will try to tune the different parameters to achieve a better model.