
# School Budget Line Item Classification

## A Multi-target, Multi-label Classification Problem

Proposal for Capstone Project 2

Springboard, DSCT

01Jun2018

Mark Sausville
saus@ieee.org

# School Budgets: Multi-target, multi-label classification

# Capstone 2 Proposal

## Overview

Budgets for schools and school districts are huge, complex, and unwieldy. Education Resource Strategies is a non-profit that analyzes school budgets using human expert analysts with the goal of letting districts be more effective in their spending.

The task is to correctly label every budget line items in nine different classifications with appropriate labels for each class. These labels allow ERS to understand how schools are spending money and tailor their strategy recommendations to improve outcomes for students, teachers, and administrators.

An effective machine learning model will let human analysts complete work much faster by eliminating a substantial amount of time consuming hand analysis and allow ERS to better serve its core client base (public schools and school districts).

The data consists of 14 columns of free-form text data and 2 numerical columns.

The project is a multi-target, multi-label classification problem with substantial issues in text processing, feature design and engineering, model selection and model tuning.

This task also has the advantage of participation in an active competition so models can be objectively judged for quality.

## Structure

This data is hosted as part of an active competition at Drivendata.org. Along with the competition documentation, DrivenData has created a tutorial demonstrating several relevant modeling techniques.

The goal of this project will be to analyze the performance of the presented models independently produce a high quality model that will be competitive.

This project breaks down into the following steps:

1. Acquire the data.
2. Explore and describe the characteristics of the data.
3. Analyze the structure and performance of the models that DrivenData has presented.
4. Independently produce a competitive model.

# Deliverables

The deliverables for this project will include the following:

- Milestone report
- Final project report
- Project code
- Presentation slide deck