**DONORS CHOOSE PROJECT: ANALYSIS PLAN**

1. **DATA PREPARATION**
   1. **Variables to be not included in the analysis**

All other dimensions of the dataset below were not used are either likely to be customer identifiers, cause the model to overfit, or has no predictive power

| **Projects data** | **Donations data** |
| --- | --- |
| Projectid | Donationid |
| Teacher\_acctid | Projecteid |
| Schooid | Donor\_acctid |
| School\_ncesid |  |
| School\_latitutde |  |
| School\_longitude |  |

* 1. **Data quality issues**
     1. Check if all variables data types are correctly read into the R session
     2. Validate inherent column aggregations in the data like is the sum of donation\_to\_project column and the donation\_optional\_support truly equal to the donation\_total column?
     3. Validate whether there are zero or negative donations and remove such observations
     4. Remove observations with missing donor id
     5. Check for data duplication
  2. **Variable transformations**
     1. Convert all logical data types to factors
     2. Dummy encodes all categorical variables
     3. Standardize variables for all predictive models except for Decision Trees

1. **EXPLORATORY ANALYSIS**

**Questions to explore in exploratory analysis for prediction**

* 1. What is the distribution of exciting versus non-exciting projects for the entire data and by historical years? Is there any concern about class imbalance in the data?
  2. Does a high proportion of great messages influence exciting projects?
  3. Does a high proportion of teacher referrals influence exciting projects?
  4. Does poverty level influence exciting projects?
  5. Do projects eligible for double your impact tend to be more exciting?

**Questions to explore in exploratory analysis for understanding donor profiles?**

* 1. Which cities are our top ten donors? Which cities are our bottom ten donors?
  2. Which payment method is most used by donors?
  3. Are there instances of donations coming from a city other than the school’s city? How does the quantum of these donations compare to donations coming from the same city as the school?
  4. What is the distribution of the quantum of donations by other categorical variables like giving page, gift card, and account credit?

1. **ANALYTICAL METHODOLOGY**
   1. **Predicting exciting projects**
      1. Type of problem: Classification
      2. Model search space: Decision Tree, Random Forest, and Support Vector Machines
      3. Model selection metric: receiver operator area under the curve
      4. Business metric: Increase the annual rate of fully funded projects beyond the current level of 15%
   2. **Understanding donors**
      1. Segment donors with K-means clustering
      2. Use decision trees to understand important donor attributes
      3. Profile donors based on an understanding of cluster descriptive statistics
      4. Evaluate whether cluster sizes are evenly distributed
      5. Evaluate the within-cluster sum of square errors