NKF Project Write Up

The National Kidney Foundation (NKF) of Florida hosts ten one mile walk events a year. These events have two goals raise awareness for the causes and symptoms of kidney disease and raise funds to assist those financially struggling due to the challenges presented by kidney disease. However these events have widely varying performance outcomes both in fundraising (min $4008, max $33,095) and number of participants (min 103, max 626). The purpose of this project is to assist the National Kidney Foundation of Florida by decomposing the factors which influence event reach, measured by number of new participants; event engagement, measured by returning participants; and event financial performance measured the total revenue generated by the event.

The approach we employed was to dig down from exploring aggregate event level effects down to individual participant level drivers of those effects. Therefore this report is organized as follows: first we provide a brief overview of the dataset and the highlights from our exploratory analysis, then we provided a concise description of the new features which created to support our model development, after which we present our modeling choices, validation plan, and results for our investigation for drivers at the event level for whether an event is increasing in its fundraising capability and reach, as well as individual level participant drivers pertaining to repeat participation, fundraising from others, self donation, and individual goal attainment.

The dataset for this project is comprised of all event participants and their fundraising productivity for all ten NKF events for three years 2017, 2018, and 2019 drawn from the online registrant and donation tracking tool employed by the NKF of Florida.

Exploratory Data Analysis

I’ll a here of the events some of my prior graphs.

Chart, bar chart

Description automatically generated

Chart

Description automatically generated

Methodology:

This project employed the methodology proposed by Shrestha et al. (2020) where the application of machine learning (ML) methods are used to support induction for theory building. Algorithm supported induction can enable the identification of robust data patterns while producing findings that support theory building and management communication. The approach employed by the analysts follows three stages:

***Stage 1: Splitting the sample*** First the analyst split the dataset provided by the NKF into training (i.e., sample I) and test datasets (sample II).

***Stage 2: Detection of robust and interpretable patterns (i.e. data reduction)*** In two-stage is a feature selection process using the training dataset (sample I), which aims to balance predictive accuracy with interpretability and often is completed in two steps:

In step 1, the analyst identifies robust associations within sample I by applying ML methods without concern for interpretability or complexity to the data to identify highly influential variables and uses cross validation for hyperparameter turning.

In step 2, the analysis identifies highly influential variables as well as interaction effects (depending on the analysts appetite for model complexity) from the modeling results in step one and applies a more interpretable feature selection algorithm such as LASSO or RIDGE regression to obtain a smaller set of important features and interprets those variables with coefficients greater than zero.

***Stage 3: Building and testing hypothesis*** In stage-three the analysis uses the hold out data (sample II) to test their features for exploratory power using OLS regression or other exploratory regression approaches.

Feature engineering: in support of model development several new features were created. Probably the single most important new feature is the calculation of distance traveled by event participants which was measured by using the participants zip code and measuring the distance from the middle of that zip code to the zip code of the event location. However additional features were also constructed including the following binary variables: whether the participant made a personal donation, whether the participant elicited donations from others, whether the participant set a fundraising goal for themselves different than the system default of 250 dollars, whether the participant was on a team, updated their personal fundraising page at any time, sent fundraising emails, provided a reason for their participation, and fully completed all registration information.

**Event Level Analysis**:

Identification of Growth Events Revenue Raised:

* Modeling Choice:
  + Classification: [insert model name]
    - Model Validation: [insert validation measures and results]
  + Data reduction: [insert model name]
    - Model Validation: [insert validation measures and results]
  + Performance against Holdout: [Test variables with explanatory regression (OLS or other) significant variables identified from data reduction model on the holdout sample and report results (R^2, RMSE etc.) ]
  + Results: [provide synthesis of results what is interesting from the ML what is interesting from LASSO how well does it fit the hold out]
* Use total revenue as your DV

Identification of Growth Event Number of Participants:

* Modeling Choice:
  + Classification: [insert model name]
    - Model Validation: [insert validation measures and results]
  + Data reduction: [insert model name]
    - Model Validation: [insert validation measures and results]
  + Performance against Holdout: [Test variables with explanatory regression (OLS or other) significant variables identified from data reduction model on the holdout sample and report results (R^2, RMSE etc.) ]
* Results: [provide synthesis of results what is interesting from the ML what is interesting from LASSO how well does it fit the hold out]
* Use n as your dv.
* **Individual Level Analysis**:

Fundraising Individual:

* Modeling Choice:
  + Classification: Generalized Boosted Regression Model (GBM)
    - Model Validation:

Chart

Description automatically generatedChart

Description automatically generated

Attempted several regression techniques along with some ensemble models. The BAG and GBM resulted in the highest Kappa values, with the GBM having the highest Accuracy. GBM was selected for the best overall performance in both categories.

Text

Description automatically generated

A picture containing graphical user interface

Description automatically generated

* + Data reduction: Logit Regression
    - Model Validation:

MARS, LASSO and Logit Regression techniques were utilized for developing models.

MARS performance against train data:

Accuracy: 0.90207

Kappa: 0.120

Logit performance against train data:

Accuracy: 0.897

Kappa: 0.3639

LASSO performance against train data:

Add results

* Performance against Holdout

Text

Description automatically generated

* Results: There was all around poor performance on predicting individual fundraising by this model. The overall low participation of volunteers that actually fundraised (804/7306= 0.11) make this particularly challenging. XXXX More XXXXXXX
* Donating Individual:
* Modeling Choice:
  + Classification: [insert model name]
    - Model Validation: [insert validation measures and results]
  + Data reduction: [insert model name]
    - Model Validation: [insert validation measures and results]
  + Performance against Holdout: [Test variables with explanatory regression (OLS or other) significant variables identified from data reduction model on the holdout sample and report results (R^2, RMSE etc.) ]
* Results: [provide synthesis of results what is interesting from the ML what is interesting from LASSO how well does it fit the hold out]

Self Goal Individual:

* Modeling Choice:
  + Classification: [insert model name]
    - Model Validation: [insert validation measures and results]
  + Data reduction: [insert model name]
    - Model Validation: [insert validation measures and results]
  + Performance against Holdout: [Test variables with explanatory regression (OLS or other) significant variables identified from data reduction model on the holdout sample and report results (R^2, RMSE etc.) ]
* Results: [provide synthesis of results what is interesting from the ML what is interesting from LASSO how well does it fit the hold out]

Classification of a Repeat Participant:

* Modeling Choice:
  + Classification method: bagged forest
    - Model Validation:
  + Data reduction: binomial Lasso
    - Model Validation:
* Results: