1) Abstract

Bike MS (nonprofit organization) is a national fundraiser. Which benefits the NMS Society (nonprofit organization) by conducting charity cycling series across the country. Charity cycling series events are held throughout the year across country. Funds are raised through registration fees, corporate donations and individual donations. Our objective is to help Bike MS raise more money, attract new participants and maintain its status as top fundraising cycling event.

2) Problem and Motivation

Of the Business Priorities, we decided to focus on the question: "What occupations were responsible for most of BIKE MS fundraising?" As we understand, BIKE MS organizes variety of events at the national level to raise funds and it uses these funds to help the people suffering from the disease – "multiple sclerosis". We wanted to help BIKE MS by identifying the top occupations which generates maximum funds each year so that BIKE MS can target those industries and continue to engage them in future events as these occupations have potential to contribute immensely to achieve BIKE MS's goal.

3) Approaches

We used datasets 2013-2017 Bike MS Participants and 2013-2017 Bike Teams data and created a new dataset.

Preprocessing: As part of preprocessing and cleaning data

1. We merged above two data sets to create new file that can help us to satisfy the requirements.
2. We have done Exploratory Data Analysis on the required parameters of the new dataset - The results of the EDA are as follows.

Exploratory Data Analysis:

Step 1: We did a frequency graph of **Event type**. We found that there are negligible number of records in Cycling Event. So we ignored them.

Step 2: We researched on **Email Id of Team Captains**. Our findings specify that most team leads used their personal Email Id's and some did not have email id's. Very less people tend to give their corporate email id. We need to come up with innovative methods to get the corporate email id of the participants. Also, for better motivation and engaging the team leads we have to have their email ids wherein we can give exclusive offers/newsletters/event milestones until the next available event for them.

Step 3: We did analysis on **Team Captain Accept Mail ID**. We see that most of the team leader were accepted through mail invitation and ¼ of the leaders came through some other invitation. This inference can be used to encourage participants to invite friends and family by offering incentives.

Step 4: We did analysis on variable **Team Division**. We can understand that, most of the teams came along with the friends and family. Corporate team's participation tops the second in the list. This proves that we have a good number of participants from corporate companies. The proportion of team division of Corporate versus Friends and family was analyzed using this.

Step 5: We analyzed variable **Participant Occupation**. We have few fields that were not genuine or recognizable, which have been removed. Some such fields are - IT Director, Key Accounts Manager, FALSE, Photojournalist etc. Also we have removed rows where the occupation count is less than 6, as we have considered those to be outliers - Data which might be entered wrongly by the participants.

Step 6: We have aggregated the **Revenue generated by each occupations** for the last 5 year. This helped us to select the top 5 performing occupation fields responsible for most of the revenue. It was understood that Engineering, Executive/Management, Healthcare, Information Technology (IT), Sales etc. contributes towards most of the fund raise. We start the modeling process from this stage.

Modeling:

Step 1: The revenue generated from each occupation type in the past 5 years (2012 to 2017) is computed.

Step 2: The total revenue generated vastly depends on the top 12 occupations and in order to determine the revenue generated for 2018 we are concentrating on these occupations.

Step 3: The Amount contributed by top 5 occupations each year is computed. Engineering occupation generates the most revenue each year even though there is a decreasing revenue by year. Executive/management occupations in addition with Health care consistently generated revenue each year and stayed in top 3. Legal and paralegal, Information technology stayed in next 3 places all through the years.

Step 4: The top performing occupations on the overall 5 years is selected. We develop a linear regression model to predict the revenue that will be generated for each occupation for the year 2018. Developed model with linear regression and did predictions on. We have not used any open data from any external sources.

1) Tools & Analytics

We used R language for data preprocessing, analysis and modeling. We used Exploratory Data Analysis to understand the variables and their importance in answering our question. We used Regression for model building. Tableau is used to generate visual plots.

2) Results

From EDA analysis we infer that we need Email Id's for engaging and motivating the team leads. Team Division is dominated by Family and Friends teams. So we need to encourage the participation of such teams. Funds raised by each Occupation is calculated and we found that Engineering, Executive/Management, Healthcare, Information Technology (IT), Sales etc. contributes towards most of the funds raised. When we predicted future funds raised we found there to be a decrease in trend of revenue generated. We have planned to develop a multilinear regression modelling technique to analyses the reasons for decrease in trend for each occupation. This will help us focus the factors to be considered for each occupation.