Teradata Data Challenge

Team Number (or name):

6

Team Members:

Eshan Bhatt

Sonika Rajan

Dhwani Contractor

Stephen Wooley

Course: DSBA/ MBAD 6211- Advanced Business Analytics

Instructor: Dr. Mousavi

Spring 2018

1 Project Report

1.1 Introduction

After 2012, Bike MS participation and revenue have been on the decline despite a retention rate of over 50%. There are not enough new participants joining the series to reverse the damage caused by attenuation. In 2017, Bike MS had 74,000 riders and a total of 6,150 teams. Their goal for 2018 is 80,572 riders (40,000 new who have not participated within the last five years), and 6,489 teams. Hence, one of the goals should be increasing new participant acquisition.

Bike MS's strongest group of donors has been corporate teams. They observed that the teams with 10+ members raise 3x more than teams with fewer than 10 members. In 2017, they had 1,561 teams with 10+ members; 642 of those were corporate 10+ teams. It follows that the top priority will be to attract new corporate teams that will recruit at least 10 members to participate for greatest revenue return.

As part of increasing donations from corporate teams, some key questions must first be answered. We will identify industries that traditionally have strong involvement with Bike MS so that Bike MS can identify new candidates for corporate team acquisition. The factors shared by the top performing corporate teams should also be examined. This information can help predict new corporate team performance, and it can be used to help improve the performance of returning corporate teams. One of the major concerns with retaining and acquiring corporate teams is that Bike MS is fighting with other non-profit organizations for donors. We will attempt to quantify the effects competing events have on Bike MS by looking at trend data.

Another way Bike MS is seeking to improve acquisitions is through digital marketing. Digital marketing has proven to be effective for Bike MS in some markets, but they are overall seeing a decline in acquisitions even with an increase in their digital marketing budget. They are looking to find out where they have seen the largest return on investment and how to make their digital marketing more effective.

1.2 Background

Advances in the machine learning infrastructure at Airbnb has enabled them to lower the cost involved in deployment of machine learning models to production. Here, the author has explained the process with the help of Customer Lifetime Value(LTV) modelling. They have used several tools to put their model in the production. These tool include Zipline, scikit-learn and AutoML-frameworks.

Zipline is Airbnb's internal feature repository, which enables Airbnb to use the existing top features from the repositories or create new features from these already created new features. At a high level, Airbnb use pipelines to specify data transformations for different types of features, depending on whether those features are of type binary, categorical, or numeric. Since

experimenting all the models can be very time consuming, Airbnb uses a AutoML framework which can speed up the process. Based on the results, they realized that XGBoost (Extreme Gradient Boosted Trees) significantly outperformed several other models. Airbnb also used one more tool, called as ML Automator which can transform the Jupyter notebook into Airflow pipeline so that the data scientist can work with ease.

Such approach of using machine learning methods and models can help us in deriving new insights with high accuracy; like for determining the relationships between the digital marketing investments and acquisitions in different years and future strategies for increased coverage.

Airbnb leveraged its empirical data to improve its product recommendation to the users. Here, Airbnb uses the data to generate host preferences. The premise Airbnb is to match the people looking for accommodation to those wanting to rent out their place. Here, author talks about how Bar Ifrach's research project answered question like; what affects hosts acceptance decisions? The initial understanding Ben discovered is, that host are more likely to accept decisions that fall into their calendar, thus minimizing their gaps. This lead to finally working on machine learning research project at Airbnb which could generate 4% more request-to-acceptance rate.

This kind of approach can enable us to understand and identify the top contributors for various events, also identify the top corporate industries that have the highest involvement in the events.

After the release and success of the 2003 book *Moneyball,* sports teams have been realizing that their data is more powerful than they had ever imagined. Using data science and machine learning tactics, Booz Allen's team was able to develop an application for MLB coaches to predict any pitcher's throw with up to 75% accuracy, changing the way that teams prepare for a game. Looking at all pitchers who had thrown more than 1,000 pitches, the team developed a model that takes into account current at-bat statistics, in-game situations, and generic pitching measures to predict the next pitch.

Now, before a game starts, a coach has the ability to analyze an opposing team's lineup and run predictive models to anticipate how to structure his plays, not only adding capability for his team but changing the manner in which the game itself is played.

Founded in 2014, San Francisco-based Bayes Impact is a group of experienced data scientists assisting nonprofits in tackling some of the world's heaviest data challenges. Since its founding, Bayes has helped the U.S. Department of Health make better matches between organ donors and those who need transplants, worked with the Michael J. Fox Foundation to develop better data science methods for Parkinson's research, and created methods to help detect fraud in microfinance. Bayes is also developing a model to help the City of San Francisco harness data science to optimize essential services like emergency response rates. Through organizations like Bayes, data science has the power to make a significant social impact in our data-driven world.

Bike MS seeks to understand what kind of quantifiable impact their competitors are having on their donations. An approach to this problem could be to use Google Trends to see if interest increasing in their competitors causes a decrease in interest in Bike MS and if so, to what extent. One of the ways to model this kind of data would be with a time series like an autoregressive or vector autoregressive model. Seth Stephens-Davidowitz and Hal Varian wrote an article on using Google Correlate to collect data for building a Bayesian structural time series. The article explains how they created their dataset from Google Correlate and how they built their model using R. They compared a standard autoregressive model with the Bayesian structural time series model and were able to show that the additional predictors added from the Google data decreased error by around 23% (Stephens-Davidowitz and Varian, 2015).

1.3 Data

Teradata University Network (TUN) has provided us with eight datasets as follows. All the datasets include data collected by Bike MS for years 2013 through to 2017.

1. Donations - This Dataset contains details of each individual Donations. Each donation will have an unique entry, even if it is from the same participant. Donation details include, Donation ID, Amount, Payment Details/Methods, Event details etc.
2. Participants - this is a Participant Dataset which contains all the details of participants like Participant name, Gender, Contact, Team Details, Donation Amount etc.
3. National Team Activity - this dataset contains information on the national teams such as information regarding the events the teams participated in, contact information for the team, and the primary connection to MS.
4. Affiliate Codes - this dataset contains a list of the different chapters of Bike MS. Each chapter has a 3 bit identification code, and this dataset can be used to track mergers among different chapters.
5. Bike MS Digital Advertising Reports- this dataset contains details about all the digital advertisements along with the market overview and adwords overview. Each of the file have several variables, which includes impressions, cost per registrations, total original budget and total conversions, and all this data can be used to answer questions like opportunities for digital marketing investments and also understand the relationship between the digital marketing and acquisitions.
6. Bike Teams- this dataset contains the details about the teams that have participated in the different events along with other attributes like total gifts , team total goal and previous year total gifts, such data can be used to understand the relationship between different teams and the total gifts for a particular fiscal year.
7. Google analytics - This part contains different data sets regarding Exit pages, Landing pages, Top events, Pages Acquisition, Pages referral traffic, location, location city, location metro, events overview, audience overview. Such datasets can be used to

determine the trend of acquisition and can be used to develop a better registration process.

8. Bike events - This data set contains the information regarding fundraising goals of the event, active/non active participants, Total fees, Total online/offline gifts, self-donors, captain, teams, Average team size and address of the event.

The main problem faced by Bike MS campaign is participant attrition and the consequent loss in donation generated. For Bike MS to address this, we are analyzing the data collected from the past 5 years. These are some of the most critical variables that we feel needs to be focused on or looked at, in order for us to come up with new suggestions for Bike MS. The visualizations have been created using SAS Enterprise Guide and Tableau.

Variable: Participant Gender

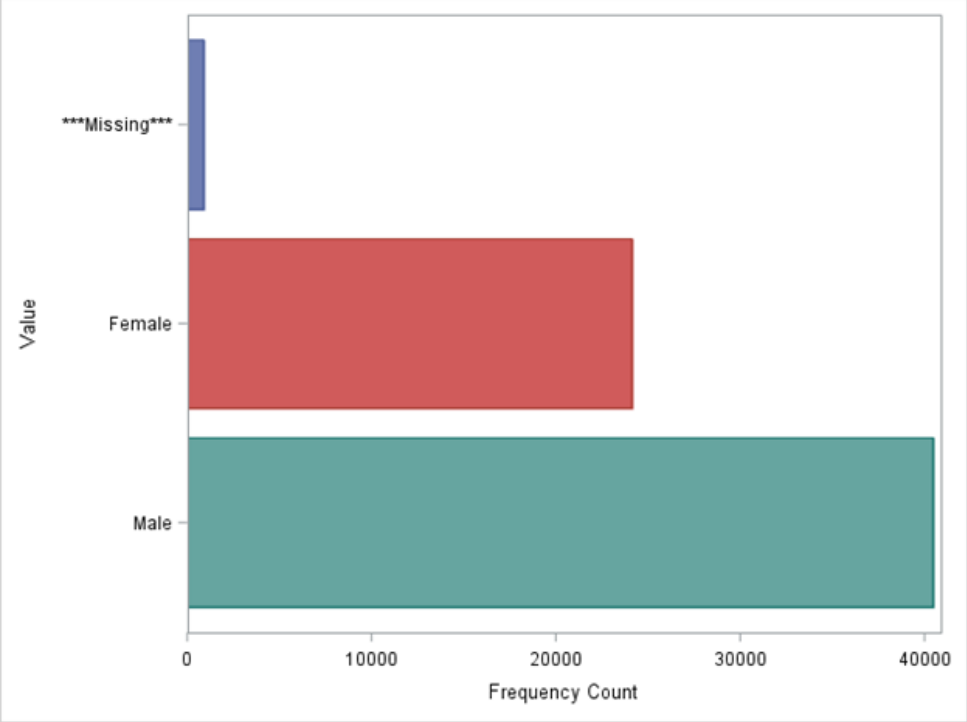

Figure 1: Count of Participants vs Gender

As shown in Figure 1, participants are mostly men. It is important to identify why this is the case, as it provides an opportunity to analyze and understand on how to improve the participation of women.

Variable: Participant Occupation

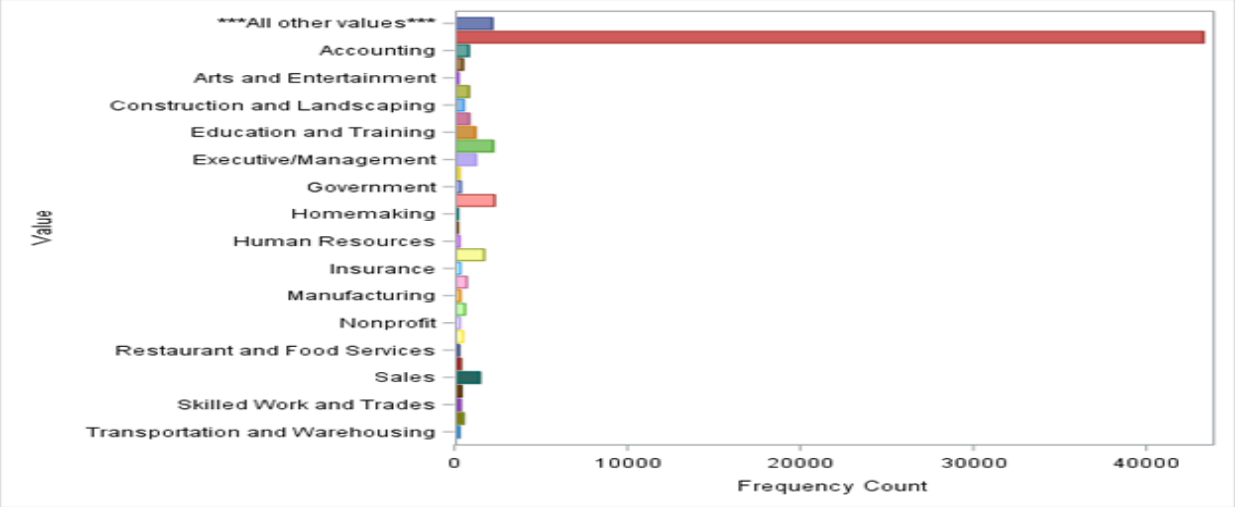

Figure 2: Count of Participants from Different Occupation

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Participant Occupation		***Missing***	43287	66.0880
		Healthcare	2315	3.5344
		Engineering	2225	3.3970
		Information Technology (IT)	1718	2.6229
		Sales	1442	2.2016
		Executive/Management	1224	1.8687
		Education and Training	1199	1.8306
		Consulting	844	1.2886
		Banking and Financial Services	828	1.2641
		Accounting	814	1.2428
		Legal and Paralegal	708	1.0809
		Marketing	606	0.9252
		Construction and Landscaping	526	0.8031
		Student	505	0.7710
		Real Estate, Rental, and Leasing	496	0.7573
		Administrative, Support, and Cle	475	0.7252
		Science and Biotechnology	378	0.5771
		Government	357	0.5450
		Insurance	347	0.5298
		Manufacturing	341	0.5206
		Retail/Wholesale	336	0.5130
		Skilled Work and Trades	329	0.5023
		Nonprofit	328	0.5008
		Fire, Law Enforcement, and Secur	291	0.4443
		Human Resources	279	0.4260
		Transportation and Warehousing	269	0.4107
		Restaurant and Food Services	228	0.3481
		Arts and Entertainment	214	0.3267
		Homemaking	209	0.3191
		Hotel, Gaming, Leisure, and Trav	200	0.3053
		All other values	2181	3.3298

Figure 3: Table showing Participant Occupation Statistics

Figure 3 shows that people from healthcare industry participate in Bike MS campaign more compared to other fields. This might be because people working in healthcare industry will probably be aware of Multiple sclerosis and its effects on patients. Thus, we need to find better ways to create awareness among other sectors. This variable contains large number of missing values and Bike MS should give more focus on collecting this information to do a better analysis of this variable in future.

Variable: Is Prior Participant

Variable	Label	Value	Frequency Count	Percent of Total Frequency
Is Prior Participant		Yes	530459	60.2230
		N/A	340501	38.6571
		Missing	9858	1.1192
		FALSE	4	0.0005
		(	1	0.0001
		TRUE	1	0.0001

Figure 4: Count of Participants who participated in Previous Bike MS Event

The above statistics indicate that only 40% of participants are new registrations. This is very low, and Bike MS should focus on attracting new registrations at the same time holding on to existing members. Also, this variable consists of values like True, False which does not conform to the standard set of values (Yes, N/A) and thus it is better to omit them. This variable also contains missing values and Bike MS should give more focus on collecting this information to do a better analysis of this variable in future.

Variable: Team Division

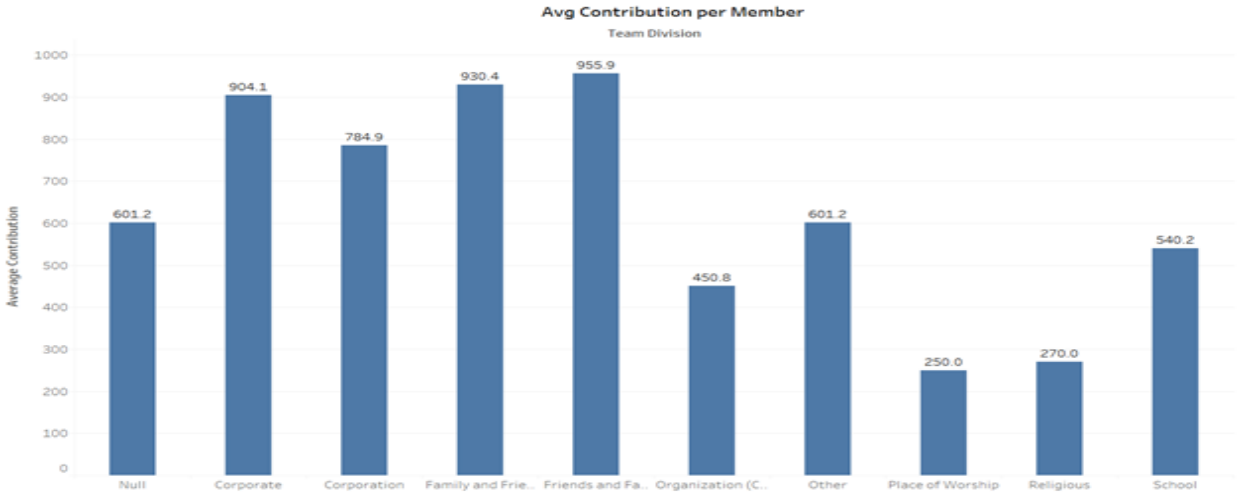

Figure 5: Average Contribution per Member

The Figure 5 shows the average contribution per member grouped by Team Division. Using this we can see that per member contribution of Corporates and Friends and Family are much higher than other groups. Hence, Bike MS should be focusing on these categories the most to generate maximum donations. The labels of Team Division variable seem to be repetitive. For example, Corporate vs. Corporation, Friends and Family vs. Family and Friends.

1.4 References

[1]Stephens-Davidowitz, S., and Varian, H. 2015. "A Hands-on Guide to Google Data," Google, Inc.

[2] Thomas, Hetlevik. 2017. "Digital Marketing: The analytical edge", https://towardsdatascience.com/digital-marketing-the-analytical-edge-e2fe733f82f7

[3] Sherice, Jacob. 2017. "How Airbnb Uses Data Science to Improve Their Product and Marketing", https://blog.kissmetrics.com/how-airbnb-uses-data-science/

[4] ScienceImmersive, V. 2018. "Data Science Immersive", *Generalassemb.ly*, (available at https://generalassemb.ly/education/data-science-immersive; retrieved February 20, 2018).