Spark Project Deliverable 0 Vibons Group 2

Andrew Whittum: awhittum@bu.edu

Jaewoo Kang: timg5599@bu.edu

Giacomo Ieronutti: giaiero@bu.edu

Background:

- Project Description: https://docs.google.com/document/d/1I9l0J7LRW71cGJv7-vqgY26hpC-G7vswavwm_7L_NfQ/edit
- Data Set: http://52.23.136.25/raw_data.csv
    - Over 1 million data points
    - Activation Date: the time that the content was delivered to the user
    - Activity Date: The time that the user accessed the content
    - Action: Percent completion of the content
- Primary Goal:
    - Analyze the dataset to find the optimal time to send content to users in order to increase the completion rate
    - Personalize these recommendations, i.e. find the optimal time to send the content to users
- Secondary Goal:
    - Use the information from the existing users to find the best time to deliver content to new hires who have just been signed up for Vibons
- Background reading for time optimization: https://blog.robly.com/2019/07/16/the-science-behind-send-time-optimization/

Process:

- Data Cleaning/Transformation:
    - Data seems to be pretty well formatted and organized upon a first examination
    - Some values in the “Name” and “Journey Name” columns seem to have some formatting issues, will have to discuss with client about how to rectify this problem
    - Values in the Channel column are mostly “Direct Connection” but there are a few others such as “From Web Referrer,” “Mobile Connection,” and “From Email,” we need more information about what exactly these terms mean
    - Vibons said they would add the time zone of the user and the date that users were signed up for Vibons service – need to add this to the dataset
    - Perform a dimensionality reduction
    - Add day of the week to the dataset
    - Adjust times for time zones of users
- Approach for answering strategic questions:
    - Perform exploratory data analysis:
        - Perform a clustering analysis to see what insights can be extracted from the data
        - Perhaps group data by completion rate to see what they have in common

        - Group 100%
        - Group 0% together
        - Group completion rate between 0 and 100 in buckets (i.e. 0-20, 21-40, etc.)
    - Group data by user, to see the characteristics of users with high/low completion rates
    - Consider if we should use time series analysis
    - Check for thinks like homoscedasticity, multi-collinearity, autocorrelation
- Perform data analysis using a variety of techniques and compare each technique to answer primary and secondary goals
    - Split data into training and test sets
        - Decide what size training and test sets we want – there are over 1 million data points, do we need to use all or even most of them? Or should we take random samples from the data?
    - Generate models using linear regression, neural networks, etc.
    - Compare models using various techniques: RMSE, SHAP, etc.
- Put together results in order to present them to the client
    - Use various media such as PowerPoint, Jupyter notebooks
    - Generate visuals and graphics to communicate results