Table of Contents

1. Installations

2. Project Motivation

3. File Descriptions

4. Results

5. Licensing, Authors, Acknowledgements

Installation

The code should run with no issues using Python 3. The following libraries are used in the code:

\* pandas

\* json

\* numpy

\* math

\* matplotlib

\* sklearn

Project Motivation

In this project, I would like to apply the skills that I learned from Udacity Data Scientist course to analyze the data files provided by Starbucks to build a model to predict a customer’s response to an offer or promotion.

File Descriptions

There are only two files for the project:

- Starbucks\_Capstone\_notebook.ipynb

- README.md

The project has four parts:

1. The first part is exploring the three data files provided by Starbucks: portfolio.json, profile.json, transcript.json.
2. The second part is to pre-process these three data files by separating multiple values into different columns, dropping the NaN values, and replacing some values with category values in order to do analysis. Then merge the three cleaned data files into one data file and analyze the data.
3. The third part is to build a predict model then validate the model with different algorithms, I choose two algorithms to compare the results: Logistic Regression, Random Forest. To run the test with each model, I use RandomizedSearchCV with 10 iterations to optimize the training time, and create a parameters dictionary to tune the model.
4. The fourth part is to write a blog and publish it on medium.

Comparing both accuracy and f1score results, Random Forest performs better than Logistic Regression. The accuracy difference between Random Forest and Logistic Regression is about 0.04(4%).

Licensing, Authors, Acknowledgements

Thanks to Udacity and Starbucks for providing the data sets. Thanks to Vish Blze for his post which provides me ideas on where and how I should improve my work on this project.