CHALLENGE 02

GROUP NAME: .COM

SOLUTION TITLE: DATA PYTHONISA

Table of Content

DocumentIdentification

Challenge
02–Big Data – Data Analysis

Group Name
.COM
Group Members
Captain: Raul E. Lopez Briega
Member 2:Stephanie Anglarill
Member 3:Daniel Garac y Gojac

Solution Title			
Data Pythonisa			
Version	**Status**	**Date**	**Classification**
0.1	*Final*	*10/11/2013*	*Internal*

Version	Date	Description
0.1	*10/11/2013*	*Data analysis proof of concept .COM team.*
0.2		

EvaluationSummary

Evaluation Committee
Member 01:
Member 02:
Member 03:
Member 04:
Member 05:

Dimensions and sub-dimensions	Weight	Evaluations					
		01	02	03	04	05	Result
01. Development of the Concept/Idea (up to 1 page)	***25%***						
Business Orientation	*10%*						
Business Applicability/feasibility	*10%*						
Innovation	*5%*						
02. Research Documentation (up to 3 pages)	***25%***						
Relevance of the research to the group´s solution	*10%*						
Documentation structure and assertiveness	*10%*						
Source Documentation (references)	*5%*						
03. POC (proof of concept) implementation & documentation	***40%***						
Data Cleaning, Preprocessing and EDA	*5%*						
Model Development	*10%*						
Optimization Results	*10%*						
Documentation Code/Script/Configuration	*10%*						
Application references and licenses	*5%*						
04. Oral presentation	***10%***						
Presentation Structure	*2%*						
Time (up to 15 minutes)	*2%*						
Defense Quality (was the group able to defend properly its solution?)	*2%*						
Question and Answers (group ability to answer to the evaluators)	*4%*						
EVALUATION RESULT							

SolutionDocumentation

Development of the Concept

1. Introduction

At current times, where we are generating data every time and everywhere; companies are now aware that data can make a difference in an election or business model.

Data analysis is now critical to businesses strategy. Businesses increasingly are driven by data analytics, so there is great professional advantage in being able to interact with the vast amount of data of today world. Understanding the fundamental concepts, and having frameworks for organizing data-analytic thinking not only will allow one to interact competently, but will help to envision opportunities for improving data-driven decision-making, or to see data-oriented competitive threats.

For all the above, in this challenge, we are trying not only to help the NGO to find out the best strategy for its campaign but also to build an ultimate framework for dealing with the data analytics process.

Our principal goal is first building a robust, productive and general purpose aframework that were easy-to-use and effective for data analysis; and then use this framework to perform our analysis and build a response model to help the NGO with its campaign.

In helping the NGO; our goal is to improve the profits for at least 50 % in order to make the difference statistically significant and better than chance.

More information about our data analysis framework and our data mining process can be found in the following pages and in the attached documents.

Research Documentation

2. Steps towards our Research

1. Analysis of different tools for data analysis
2. Why Python.
3. Data cleaning
4. Data exploration
5. Building our model

2.1 Analysis of different tools for data analysis

The first step we did in our research was to choose the tools we are going to use for building our data analysis framework. As we pointed out in the introduction, we wanted a general purpose and easy-to-use framework; so we decided that the key properties of our framework had to be:

◆ Interactivity: our framework must allow interaction with the end-user; we did not want to create a batch process, we wanted something where you can input your questions and receive the immediately answer.

◆ Visualization: when performing Data analysis, the possibility to visualize the results is critical; so we wanted that our framework allows easy and clear visualizations.

◆ Easy-to-use: For us, the learning curve was an important factor; we wanted something easy to understand and learn; so we could start working with it right away.

◆ Reporting: Reporting is another important factor, the results of any analysis worth nothing if no-one read it. So, our framework should give us easy-to-use reporting and sharing capabilities.

◆ Extensions: we did not want a domain specific framework that only allow us to perform one kind of analysis and nothing more; we wanted something that allow us to extend our framework for another purpose and give us the freedom of choosing between different tools.

With all this properties in mind, we started our research and selecting four possible options for our framework:

1. To use a proprietary tool for data analysis, such as SAS or SPSS.
2. To use the open source statistic programming language R, with R studio application.
3. To use the web application Tableau.
4. To use the open source programming language Python, with its modules iPython, pandas, matplotlib, numpy and sci-kit learn.

Research Documentation *(page 2)*

2.2 why Python

After a careful analysis of the four options described above; we decided to use Python, with its extensions modules iPython, pandas, matplotlib, numpy and sci-kit learn.
The reason we choose Python over the other options was principally because we believe that it was the option that better accomplished all the goals we had listed for our framework.

The iPython notebook module, give us the reporting and interactivity capabilities we wanted for our framework.
Pandas, numpy and sci-kit learn modules, give us the data analysis, data manipulation and machine learning tools we needed.
And matplotlib, give us the visualizations capabilities that were essential for our data analysis framework.

Another option, that was close enough to Python, was to use R programming language; the R environment was specifically designed for statistical analysis and graphics, making it perfect for data analysis. R like Python, is an open source project; so we can use it for free.

R is a great tool for data analysis, with great libraries too; but we choose Python because is easier to learn and understand than R. Moreover, if we need some specific functionality from R; we can call it from Python using Rpy2module

2.3 Data Cleansing

Now it is time to start the description of the data mining process we followed for the challenge. The first step in any good knowledge discovery process is to clean the dataset. This is a important process because incorrect or inconsistent data can lead to false conclusions and misdirected actions. Our principal goals of this phase were, to complete the missing values, to detect the outliers and to remove the no necessary information.

For accomplishing this phase goals we use some of the build-in functions that python pandas module offers; we removed the no statistically significant columns from the dataset, created some more descriptive new columns, and identified some of the most important outliers.

2.4 Data Exploration

Once that our dataset is clean, we can continue with the next step; the exploratory phase; here our focus was in detecting the key factors and fields that give us a way to predict the donation behavior.

This phase is quite important because the only way to develop intuition for what is going on in an unfamiliar dataset is to immerse yourself into it.

Research Documentation *(page 3)*

In this phase we made an extensive use of visualizations; our goal for this phase was to get to know the data; we examined some data distributions, validated some assumptions and asked a lot of questions.

Some of the insights and understandings we gained during this phase were:

i. The most significant variables for predicting a customer's donation behavior are the previous donation behavior summaries.
ii. The demographics data turns out to be quite strongly connected to the donation performance of the population.
iii. Identifying donors is a thoroughly different task than maximizing donation. There is an inverse correlation between likelihood to respond and the dollar amount of the gift.

2.5 Building our model

With all the information and knowledge we gained from the exploratory phase, we were ready to start building a model to test our assumptions and try to predict the donor's behavior.

We first started with a single model; to build this model, we have created 7 segmentsfrom the different insight we got from the exploration data analysis. These segments are:

1. MAXRAMNT > 30
2. RAMNTALL > 250
3. HV2 > 1600 and AGE between 30 and 60.
4. EC8 > 12
5. IC4 > 800
6. RAMNT_3 > 3.5
7. STATE in ('CA', 'FL', 'MI')

Applying this single model to the dataset, we got a profits improvement of 50 %.

This are great results, but we did not stop there, then we tried to build a more complex model using the Random forest machine learning algorithm to predict the results.
We use almost the same variables as future selections to apply to the algorithm; with this new model, we got a profits improvement of 650 %.

A more technical information about our data mining process could be found in the attached document (gA Tech Contest - Challenge 02.html).

POC Documentation

The following questions will guide the contender groups in the work and documentation of their POC. For more information about the words in capital letters, please refer to the [Challenge 02 document: REFERENCES, GLOSSARY, ANNEXES]

03.01. Data Cleaning, Preprocessing and EDA (exploratory data analysis)

03.01.01) Please specify the applications/software/solutions used to carry the data cleaning, PREPROCESSING and EDA experiments.

In order to clean and preprocess the data we used Python modules such as:

- *Pandas (used for Data Analysis)*
- *IPython Notebook (used for the Data Reporting and Data Visualization)*
- *Matplotlib (used for graphics)*

03.01.02) Please specify (in minutes) how long (in terms of CPU and people time) it took to complete the data cleaning, PREPROCESSING and EDA tasks.
CPU time (in minutes).....: *2 minutes*
People time (in minutes).....: *900 minutes*

03.01.03) How did you treat the records and variables containing MISSING VALUES? Please specify the corresponding missing value treatment method/technique by ATTRIBUTE TYPE.

Regarding numeric values, we filled it with a zero if we did not have another way to predict it value. For the other types of values, we didn't handled their missing value.

03.01.04) Did you create any additional attributes based on DATA TRANSFORMATIONS? Please summarize.

Yes. We segmented the data by age and donation amounts.

03.01.05) Did you consider treating the outliers? If you did, what general rule did you apply in treating the outliers?

We identified the more important outliers; as general rule, we didn't take them into consideration because there were not statistically significant.

03.01.06) Prior to the application of the data mining algorithms, did you normalize, scale or standardize the input variables?
- The target or the dependent variable?
- The records?

- If you did, which method(s) of scaling have you used?
- If you have scaled the dependent variable during modeling, which format is it in your submitted results?

Yes, we did.

03.01.07) Did you find REDUNDANT or COLLINEAR FEATURES in the data set? If you did, how did you treat them?

Yes, we found some demographic information redundant and we dismiss it.

03.01.08) Did you implement VARIABLE/FEATURE SELECTION? If you did, how did you implement it?

Yes, we selected the more statistically significant variables; we tried to implement the variable selection based on its correlation with donation flag and the donation amounts.

03.02. Model Development & Implementation

03.02.01) Please specify (in minutes) how long (in terms of CPU and people time) it took to complete the data mining tasks.
CPU time (in minutes).....: *4 minutes*
People time (in minutes).....: *900 minutes*

03.02.02) Which software tool(s) and programming language(s), if any, were used to apply the DATA MINING algorithms?

We use Python as programming language and the data mining algorithm we used was the decision tree.

03.02.03) Please consider the learning file you used in generating your results. What was the file size(s) used during learning (total number of records)?

We use the complete LEARNING dataset in our leaning process.

03.02.04) During learning and/or validation, did you:

(a) ARTIFICIALLY EXTEND or INFLATE the data set(s)?
No, we did not.

(b) Use a BALANCED data set(s)?
No, we did not.

(c) Used a related methodology not specified above?
No, we did not.

If you answered a 'yes' to any of the above, please specify how and why?

03.02.05) Which data mining technique(s) or algorithm(s) did you use in deriving your results? If you considered more than one algorithm, which criteria did you use in selecting among competing algorithms?

We use a single white-box model and a decision tree algorithm.

03.02.06) How did you assess the predictive power/accuracy of your model(s)?
Did you develop more than one model?
If you did, which criteria did you use in selecting among competing models?

We only develop two models, a single model and a prediction model; our prediction model obtain better results.

03.02.07) Were you concerned with overfitting?
If you did, how did you safeguard against over-fitting (in other words, make sure that you were getting good generalization)?

The algorithm of the python module we used, has some safeguard controls against over-fitting; we did not take any other action.

03.02.08) Could you please list all relevant statistics pertaining to the architecture or complexity of your final model, i.e., number of weights, number of hidden nodes in a layer, number of layers, number of levels and nodes in decision tree, number of rules, etc.

We use 50 estimates for our Random forest regression model.

03.02.09) How many variables are in your final model?
Please list their names and, if relevant, their relationship (positive or negative) with thetarget variable. If you have a mechanism in determining their importance or impact in the model, please list them by the order of importance and describe your mechanism very briefly.

You can find the details of our data mining process in the attached document (gA Tech Contest - Challenge 02.html).

03.02.10) Does your software tool generate a SCORING CODE (see glossary for more information) that can be used to export the model outside the data mining environment?

No, we did not implement a scoring system.

03.03. Optimization Results

03.03.01) What is the estimated donation flag for record (line) of the VALIDATION SUBSET? What is corresponding estimated dollar amount for each instance (line).

N/A

03.03.03) What is the estimated LIFT CURVE (see glossary) – cumulative percentage of targets in the top quantiles of the file

N/A

03.03.04) What is the receiver operating characteristics (ROC) curve analysis and the area under the ROC curve.

N/A

03.04. Documentation Code/Script/Configuration

03.04.01) Please provide the source code, scripts and/or step-by-step configurations used to perform the POC.

The source code could be found in the attached documents.

The easier way to install all the python modules used in our framework is to download a single binary installer thatcomes with a full Python distribution and a lot of widely used external packages, includingIPython.

Popular distributions include:

- *The Enthought Python Distribution (EPD) and the new Canopy by Enthought:*

http://www.enthought.com/

- *Anaconda by Continuum Analytics:*

http://www.continuum.io/

- *Python(x,y), an open source project:*

http://code.google.com/p/pythonxy/

- A*ctivePython by ActiveState:*

http://www.activestate.com/activepython

All these distributions support Linux, OS X, and Windows, except Python(x,y) whichonly supports Windows. They all offer a free edition (and possibly a commercialedition) and they all contain IPython.

03.05. Application references and licenses

03.05.01) Which tools, applications, solutions were used to perform the task?

The tools we used were:

- *Python.*
- *R.*
- *Matplotlib.*
- *Numpy.*
- *Ipython.*
- *Sci-kit learn.*

03.05.02) What kind of license is required for each used tool?

- *Python. Python Software Foundation License.*

http://docs.python.ort/2/license.html

- *R. GNU General Public License.*

http://www.r-project.org/COPYING

- *Matplotlib. BSD compatible license.*

http://matplotlib.org/users/license.html

- *Numpy. Numpy license.*

http://docs.scipy.org/doc/numpy/license.html

- *Ipython. BSD compatible license.*

http://ipython.org/ipython-doc/dev/about/license_and_copyright.html

- *Sci-kit learn. BSD compatible license.*

http://opensource.org/licenses/BSD-3-Clause

03.05.03) What are the end-user requirements for each used tool?
(a) Marketing/Product/Industry Manager
(b) Business Analyst
(c) Statistician or data mining specialist
(d) Other, please specify

The end-user for our framework are business analyses and statistician or data mining specialist.

03.05.04) Using a scale ranging from 1 to 5, where 1 means that extensive programming effort on the part of the end-user is required to handle the task in question and 5 means that the software tool automatically handles the task with minimal initial input from the user, please indicate the degree of end-user input required to handle each of the tasks listed below.

(Of course, the majority of the software tools provide a capability somewhere in between. For example, qualifying attributes to various tasks by pointing and clicking through a graphical user interface would be, say, a 4 on the automation scale. If the task in question is not applicable to your application/software tool, please check N/A)

(a) DATA PREPROCESSING

(1)------(2)------(3)------(4)-------(5)		N/A
End-user manually programs	Tool fully-automatically handles	Not applicable

(b) Application of the DATA MINING algorithms:

(1)------(2)------(3)------(4)-------(5)		N/A
End-user manually programs	Tool fully-automatically handles	Not applicable

The degree of end-user input required for our framework is of 3 in both phases (Data preprocessing and data mining)