George Mason University
SYST 699: Masters Capstone Project
Spring 2014

Project Proposal:
SAP Big Data Analytics on Mobile Usage

Inferring age and gender of a person through his/her phone habits

February 11, 2014

Arturo Buzzalino
Justin Nguyen
Mitul Patel
Tanner Suttles

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Background

SAP Mobile Services is developing a new product, Consumer Insight 365 (CI365). The purpose of this product is to enhance a business' ability to expand its market, and provide a tool to perform meaningful analysis of consumer patterns. CI365 will analyze large amounts of global mobile carrier data. This mass analysis will extend to a large number of countries across the world and cover millions of people.

The goal of this project is to provide businesses with an additional, powerful means to expand their markets by focusing their growth efforts on specific regions and demographics. Data visualization and statistical techniques will be used to determine patterns among: socio-demographics, gender, age, URL click stream categories, geo-location and texting / calling habits. Below is a sample of what a CI365 custom report might look like (Figure 1):

Figure 1: *Sample of the type of report a business might be using*

The data from the carriers is received in anonymized form. This is important because SAP does not want to breach the privacy of the consumers. There is no way for SAP to trace the number of a person to determine who he or she is or what is his or her home address.

The data provided will include the users' daily data generation through the handset. The age and gender is provided when the user has a plan with the carrier. For those users who are roaming on the network, or own pay as you go and prepaid plans, this data is not available. This is where SAP wants to use Big Data analytics to be able to infer a user's gender and age based on his or her phone habits.

1.2 Problem Statement

Focusing on a small carrier's mobile user data, determine correlations between texting / calling habits, URL categories and geo-location with user gender / age. SAP is interested in having the ability to determine the gender and general age of the mobile user based on his/her phone habits.

1.3 Scope

This project is very large and can encompass many different aspects. The project team will focus on the small carrier's data detailed in the project description.

Within Scope:

- Constructing model capable of inferring age and gender
- Model should only consider text / call data, URL traffic categories, and geo-location
    - Text and call data will further be broken down once project team has access to the data.
        - Number of texts in 1 hour, length of call, number of calls in 1 day
    - Geo-location will be considered last, as it is the most difficult.
        - Because a user's location changes hourly throughout the day, this will require further investigation
- Only consider the small carrier's data provided by SAP
- Test model and conduct sensitivity on the data for which the age and gender is NOT provided

Outside of Scope:

- Data provided outside of the categories mentioned above (i.e. daily tweets, Facebook likes)
- Data not pertaining to the carrier above

1.4 Document Overview

The remainder of this document details the requirements, technical approach, expected results and project management plan. These sections will describe how the team plans on providing a solution to the problem statement, and achieve the scope designated for the semester.

2 PRELIMINARY REQUIREMENTS

Requirement 1: The team shall utilize data provided by mobile carrier
The type of data sent by the mobile carrier includes metadata about user's texting and calling habits, points of interest frequented (geo-location), and URLs visited

Requirement 2: The team shall develop methods to identify patterns in cell phone usage by gender
The methods used will be developed based on statistical and data-mining principles and techniques. These methods shall consistently identify pattern in cell phone usage by gender.

Requirement 3: The team shall develop methods to and identify patterns in cell phone usage by age group
The methods used will be developed based on statistical and data-mining principles and techniques. These methods shall consistently identify patterns in cell phone usage by age group.

Requirement 4: The team shall develop a model for classifying a subscriber's gender
The team will develop a model within SAP HANA for classifying the gender of a subscriber.

Requirement 5: The model shall predict the gender of an anonymized user as male or female
The model will be developed based on patterns identified to predict the gender of a user.

Requirement 6: The model shall predict the age group of an anonymized user
The model will be developed based patterns identified techniques to predict the age group of a user.

Requirement 7: The model shall provide accuracy for each classification
The model will produce accuracy of its classification result for each subscriber.

3 TECHNICAL APPROACH

3.1 Analysis

In order to better understand and implement big data analytics for the project, multiple factors need to be considered. Research on phone usage behaviors, gender and cultural patterns, and socio-demographics of phone traffic data will need to be conducted as well. The pros and cons of what software would be the most optimal and appropriate to analyze and interpret the data needs to be weighed out. The team will also need to research data clustering and data mining techniques in order to figure how the model should be developed. All of these factors have to be analyzed before developing a method or guideline of inferring what the gender and age is based on the characteristics of the data given.

3.2 Requirements Development

The model requirements will be broken down into multiple categories such as function, system, input/output, operations, and interface. The preliminary requirements will be further developed once the team has access to the phone data so a better understanding of the factors being dealt with can be formulated. The determination of the metrics/rules and statistical methods of how the results will be implicated will also dictate the outcome of the model's requirements. The requirements will guide the model development phase and will be verified in the testing and evaluation phase

3.3 Model Development

The first step in model development will be the selection of features as inputs to the model. The data will be analyzed to look for features that distinguish users by their gender. Distributions in visitation of websites and rates of utilization of numeric metrics will be plotted to show patterns. Further model development will follow an agile approach where the model will be successively refined through short evolutions of design and testing. Development iterations are expected to be based on modification of the feature set, algorithm selection, and algorithm parameters.

The main software that the team is planning on using to analyze the data is HANA's PAL analytics. This tool that HANA provides should be more than enough to build a model to infer the gender and age of the phone user. A major obstacle that the rule set engine of the model must overcome is recognizing the fundamental difference between pre-paid and full plan holders. If the model does not adjust for differences in the populations, the result may be skewed. Another issue will be that the model is constrained to one market rather than a universal setting due to the limited input data.

3.4 Testing and Evaluation

An evaluation of whether the data model's gender and age implication is accurate will be performed. Two forms of testing will be testing of accuracy and performing sensitivity analysis. The testing of accuracy will be conducted on a holdout sample set of data where the gender and age is already known. The team will input that set of data into the model calculation without looking at the age and gender and verify that the model consistently outputs accurate results. The sensitivity analysis will vary features input to the model and see how the output results change. This will reduce the amount of uncertainty in the model as well as increase the understanding of relationships between the input variables and output result. The testing stage will also ensure that all the requirements are being fulfilled and me by the data model.

3.5 Delivery

The final delivery of results will be presented in a summary form rather than individual data results. It will contain information on age and genders ranges with regards to multiple factors such as location, popular interests, or URL categories. The results will be organized in a manner that best satisfies SAP requirements and expected deliverables. The results will be given in both presentation and paper format to the customers, SAP and George Mason University.

4 EXPECTED RESULTS

The project will yield three major deliverables:

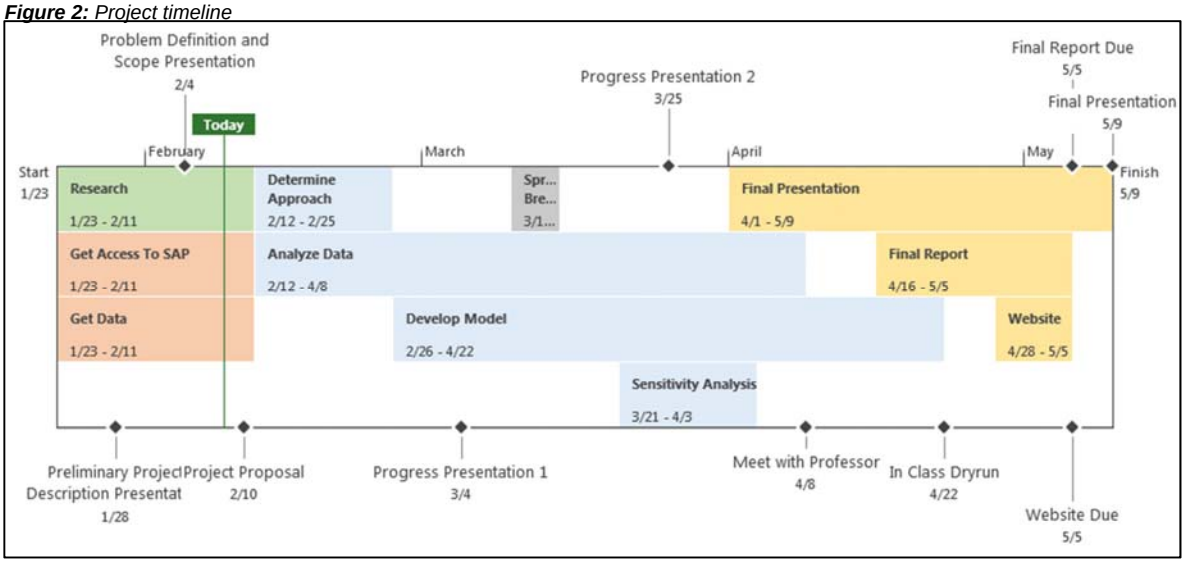
- Methods developed to identify patterns in mobile usage
  - The team will submit a detailed narrative on what methods were used, how they were chosen and how they were applied to identify the mobile usage patterns.
- Model to predict age group and gender of user
  - The team will submit a model that will infer the age group and gender of a mobile user with proven confidence. A detailed narrative about how the analytic methods were used to develop the model, and the model's inputs and outputs will be provided.
- Sensitivity Analysis of model
  - The team will perform a sensitivity analysis of the model to aid in its validation. A narrative of how the analysis was performed and its results will be submitted.

5 MANAGEMENT APPROACH

5.1 Project Plan

The project has been divided into four task areas: project management, research, model development, and final deliverables. The timeline shown in Figure 2 shows the research phase in blue, model development in green and final deliverables in yellow. The project management task covers initial project activities like kickoff and problem definition, along with progress presentations throughout the project. The introduction of new tools and concepts for the team in the area of big data also necessitates a period of research in which the team will learn about big data analytics and current cell phone usage research. The majority of the project will be spent in data analysis and model development. Near the end of the semester the team will start to look toward the final deliverables, and time spent on model development will be shared with drafting the presentation. The majority of the time for the final deliverables will be spent on the final presentation. The work on the final presentation starts earlier in the project with professor reviews, and will help organize the information for the final report. At the end of the project the final report and presentation will be delivered and the project closed out.

Figure 2: *Project timeline*



An abbreviated work breakdown structure (WBS) has been included below in Table 1. The abbreviated version shows major tasks and milestones in the project schedule. The full WBS is included in Appendix A.

Table 1: *Project work breakdown structure*

Outline Number	Task Name	Duration	Start	Finish
1	Project Management	44 days	1/23	3/25
2	Research	14 days	1/23	2/11
2.1	Mobile Phone Use Demographics	14 days	1/23	2/11
2.2	Big Data Tools	14 days	1/23	2/11
3	Model Development	64 days	1/23	4/22
3.1	Get Access To SAP	14 days	1/23	2/11
3.2	Get Data	14 days	1/23	2/11
3.3	Determine Approach	10 days	2/12	2/25
3.4	Analyze Data	40 days	2/12	4/8
3.5	Develop Model	40 days	2/26	4/22
3.6	Sensitivity Analysis	10 days	3/21	4/3
4	Website	6 days	4/28	5/5
5	Final Report	14 days	4/16	5/5
6	Final Presentation	29 days	4/1	5/9

5.2 Project Risks

Risk 1: Data Delivery

Due to the data being shipped, there is risk that the data will arrive too late in the semester for the team to develop a fully functional model. If the team does not have access to the data by February 14th, the scope of the model may have to be adjusted.

Risk 2: Data Access

Depending on the installation of the data there is a change that the data will only be accessible at the SAP Reston office. The team would need to be escorted while in the office and due to schedule limitations of the project team there would be limited availability to work on data analysis and model.

Risk 3: Big Data Expertise

The team does not have experience with SAP HANA or PAL analytics capabilities, or much exposure to data mining techniques. In order to mitigate this risk the team is consulting with Professors at the George Mason University and Subject Matter Experts at SAP.

Appendix A: WBS

Outline Number	Task Name	Duration	Start	Finish
1	Project Management	44 days	1/23	3/25
1.1	Project Kickoff	1 day	1/23	1/23
1.2	Create Presentation	3 days	1/23	1/27
1.3	Preliminary Project Description Presentation	0 days	1/28	1/28
1.4	Create Presentation	3 days	1/29	1/31
1.5	Problem Definition and Scope Presentation	0 days	2/4	2/4
1.6	Draft Project Proposal	3 days	2/4	2/6
1.7	Project Proposal	1 day	2/10	2/10
1.8	Create Progress Presentation 1	5 days	2/25	3/3
1.9	Progress Presentation 1	1 day	3/4	3/4
1.10	Create Progress Presentation 2	5 days	3/19	3/25
1.11	Progress Presentation 2	1 day	3/25	3/25
1.12	Spring Break	5 days	3/10	3/14
2	Research	14 days	1/23	2/11
2.1	Mobile Phone Use Demographics	14 days	1/23	2/11
2.2	Big Data Tools	14 days	1/23	2/11
3	Model Development	64 days	1/23	4/22
3.1	Get Access To SAP	14 days	1/23	2/11
3.2	Get Data	14 days	1/23	2/11
3.3	Determine Approach	10 days	2/12	2/25
3.4	Analyze Data	40 days	2/12	4/8
3.5	Develop Model	40 days	2/26	4/22
3.6	Sensitivity Analysis	10 days	3/21	4/3
4	Website	6 days	4/28	5/5
4.1	Create Website	5 days	4/28	5/2
4.2	Website Due	1 day	5/5	5/5

5	**Final Report**	**14 days**	**4/16**	**5/5**
5.1	**Draft**	**10 days**	**4/16**	**4/29**
5.2	**Review**	**2 days**	**4/30**	**5/1**
5.3	**Tech Edit**	**1 day**	**5/2**	**5/2**
5.4	**Final Report Due**	**1 day**	**5/5**	**5/5**
6	**Final Presentation**	**29 days**	**4/1**	**5/9**
6.1	**Draft 1**	**5 days**	**4/1**	**4/7**
6.2	**Meet with Professor**	**1 day**	**4/8**	**4/8**
6.3	**Meet with Professor**	**1 day**	**4/15**	**4/15**
6.4	**Draft 2**	**2 days**	**4/16**	**4/17**
6.5	**In Class Dry Run**	**1 day**	**4/22**	**4/22**
6.6	**In Class Dry Run**	**1 day**	**4/29**	**4/29**
6.7	**Final Draft**	**2 days**	**4/30**	**5/1**
6.8	**Final Presentation**	**1 day**	**5/9**	**5/9**

www.sap.com