GSK - Predictive Modelling for Improved Sales Growth and Efficiency

ANALYTICS PRACTICUUM
PROJECT PROPOSAL

ANLY482 AY2016-17 Term 2

Team Kes MY JX
Kesmeen Tan Jia Min
Matthew Yee Guan Feng
Sim Jing Xiang

Table of Contents

Project Overview

Sponsor Background

GlaxoSmithKline (GSK) is a British pharmaceutical company, focusing on manufacturing products for illnesses such as asthma, cancer, infections, diabetes and mental health[1] that come in the form of medicines, vaccines and other various consumer health products. GSK has a strong interest in the Singaporean market, as can be seen from its investment position, with a staggering 1.5 billion invested in the Singaporean biomedical industry. Singapore is the Asia Pacific headquarters for GSK and the company has two global manufacturing supply sites and a vaccines manufacturing facility locally.

Project Background

GlaxoSmithKline supplies its products to various clinics in Singapore. Periodically, salespersons are sent to clinics of interest with the aim of promoting new/existing drugs. Other avenues of promotion include advertising activities via the media. GlaxoSmithKline wishes to discover more about the effectiveness of their marketing and sales strategies in order to improve the company's top-line performance.

Project Objectives

There are two main project objectives:

1. Visualise the consumer purchasing patterns and behaviour within Singapore
2. Evaluate the effectiveness of different sales and marketing strategies within different areas of Singapore

Through these, some of the business applications that could potentially be improved are:

1. Strategic prioritization of sales and marketing strategies (proportions for optimal investment of resources
2. Frequency of visits of salespersons to a singular customer of interest for effective usage of manpower
3. Placement of products in sales catalogues

Methodology

Data Collection

The data given by GSK are mainly in the form of flat files (Excel). Each contains 1 or more sheets with multiple columns. Hence the data is very high in dimensionality. Metadata is not yet available, but from column headers and the conversation with the sponsor, we have an idea on which ones will be more

[1] https://en.wikipedia.org/wiki/GlaxoSmithKline

relevant to us. Such data include sales information, competency and results of sale staff, and data on the methods of the salespeople. These data have been promised to us.

To discover potential insights through spatial clustering analysis of sale territories, we also intend to collect spatial data from its vertical industries: hospitals, clinics and retail pharmacies. This can be easily collected from Singapore's public data website, Data.gov.sg, in SHP or KML formats.

Data Preparation

The stage of data preparation (or data wrangling, newly termed as data preparation taken to the next level[2]) would involve employing techniques of ETL (Extract, Transform, Load) to form an Analytics Sandbox used for further exploratory analysis purposes. To better facilitate future analysis, we will be conducting ETL process and exploratory data analysis cyclically such that if the latter is not satisfactory, we will go back to revise the former. The entire process of data preparation will be done using JMP Pro 13, which supersedes its predecessor SAS Enterprise Guide and Miner and has capabilities in the fields of descriptive and predictive modelling required by our team.

Data Consolidation (Extract)

Extraction is the primary step of the ETL process, whereby data are pulled from various different sources and aggregated into a single environment for further manipulation. That being said, it is important to note that most data could come in highly varied structure, data storage types and from different periods of time. As mentioned previously, the data given by GSK are mainly in flat file format (Excel) and spatial data collected from Data.gov.sg is in SHP or KML formats. These are also known as structured data, which are easier to process as compared to unstructured data like multimedia image or video. At this stage, it is important to check whether the data type for each variable is correctly classified (be it categorical or continuous) in JMP.

Data Cleaning & Transformation (Transform)

The next step would involve cleaning the data. We would need to explore the data iteratively to identify anomalous patterns which we can then eliminate. For example, there could be many different versions of records that all refer to the same thing. "GSK", "GlaxoSmithKline", "GlaxoSmithKline plc" all refer to the same entity. Techniques such as fuzzy cleaning and if-else rules can be implemented for standardization of variables.

Missing values will also be handled in this stage. The exact way we handle them will be determined once we take a look at the data. Our decision will be based on factors such as what data is missing, at what proportion, etc. We can choose to omit the rows with missing data from our analysis, or perhaps interpolate and impute the missing data with estimated ones. New interpreted variables (data columns) can also be created to enhance understanding and improve efficiency for further analysis.

[2] https://tdwi.org/Articles/2015/01/13/Introduction-to-Data-Wrangling.aspx?Page=1

Data Reduction (Load)

We will also need to determine which columns to focus our analysis on. This will be done in conversations with our sponsor as we seek to understand the data. Once we have understood the metadata, we will then be able to pull out the sales and other relevant data to begin exploratory data analysis. The reason for selecting only a portion of the data is that the large dimensionality would strain computer hardware and slow analysis. Additionally, there is a large amount of data that would not be in the scope of our project. We will be focusing on sales methods and results. To streamline analysis and boost runtime, we will create individual data marts for each type of analysis that we are going to carry out.

Exploratory Data Analysis

A descriptive analytics dashboard will be created via JMP Pro. We will seek to uncover patterns and anomalies. We will perform scatter plots and histograms to identify trends. For example, if we find that certain teams have very little face-to-face interactions with customers, they may require more confidence training or the client they have been assigned is less receptive to face-to-face meetings. Any assumptions that we have, either by preconceived notions or passed to us by GSK will also be tested in this phase.

Methods of Analysis

Correlations

Some questions we hope to answer include what should the business invest in in order to achieve higher efficiency and growth and which sales method is the most efficient. For this, we could look at correlations between sales revenue and inputs. While correlation is not indicative of causation, it can be highly suggestive.

Cluster Analysis + Machine Learning (Artificial Neural Networks)

Depending on quality of data and conversations in future, we also hope to create a machine learning model that will be able to do some predictive analytics. For example, by predicting how would performance vary if we change an input resource.

We could do clustering on the client data, and then for each client cluster, we can train an artificial neural network (ANN) on the sales inputs, client characteristics and resulting revenue and thereby predict results based on sales input. This is to create a predictive model for each type of client.

After the clustering, we could also compare the revenue to the sales input to identify the more efficient teams or methods and recommend GSK to analyze them in future to uncover the reasons behind the efficiency and to spread them as best practices through the organization.

Survival Analysis

Survival Analysis is a statistical technique used to analyze the expected duration of time until an event occurs and also one of the cornerstones of customer analytics[3]. An event in our project context can be customer attrition (where existing customers turnover to other companies) or inventory depletion (where certain pharmaceutical products run dry). An understanding of when customer is most likely turnover or when inventory needs to be replenished enables GSK to plan in advance churn prevention efforts and engage in proactive customer communication to effectively improve sales.

An important aspect of survival analysis is the censoring of observations in which information about their survival time may not be complete. Censored data represents a type of missing data and is required to avoid bias in survival analysis — reason why linear regression is not used to model survival time. In our context, we will be handling right-censored data such as newly subscribed customers in which there are incomplete information on when they will churn.

We will be looking into Kaplan-Meier estimator, which is a non-parametric statistic used to estimate survival function from lifetime data, usually illustrated through a survival curve of survival percentage over periods of time. For multiple sales territories, we can conduct a Log-rank test to determine differences in survivability over time and hence, effectively enhancing sales strategies across different sales teams.

Review of Existing Work

Relationship marketing and sales has been the standard approach to pharmaceutical companies achieving sales to doctors and physicians (Wright and Lundstrom 29-38). However, we can find no instance of application of advanced analytics to this field, except for basic descriptive analytics. Advanced predictive analytics has been successfully applied in other industries. For example, a neural network has been successfully applied to sales forecasting in a fashion retail setting before (Sun et al.), but to the best of our knowledge has not been applied to pharmaceutical sales.

Tools and Technologies

We intend to use the following tools and technologies for our project

- JMP Pro 13 - for its interactive data visualization
- Python - needed for machine learning purposes by Matthew
- R - to provide statistical analysis

[3] http://www.optimove.com/blog/how-to-perform-customer-survival-analysis

Project Management

Project Timeline

S/N	Task Name	In Charge	Man Day	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16
	Project Preparation																		
1	Source for Sponsors	All	3																
2	Confirm Project Sponsors	All	1																
3	Gather Data	Matthew	1																
4	NDA Confirmation	Matthew	1																
5	Preliminary Data Analysis	All	2																
	Project Proposal																		
1	Define Project Scope	All	1																
2	Proposal Report	All	3																
3	Wiki Page Design Set-up	Jing Xiang	1																
4	Wiki Page Content Update	Kesmeen	2																
	Project Proposal Due - 15 Jan 2017																		
	Discovery & Data Preparation																		
1	Data Collection & Consolidation	All	3																
2	Data Cleaning	All	8																
3	Data Transformation	All	2																
4	Data Reduction	All	2																
5	Analytical Sandbox Preparation	All	2																
	Model Planning																		
1	Data Exploration	All	1																
2	Model Selection	All	1																
	Data Analysis & Model Building																		
1	Cluster Analysis	Kesmeen	12																
2	Survival Analysis	Jing Xiang	12																
3	Machine Learning	Matthew	12																
	Mid-Term Review																		
1	Interim Report	All	6																
2	Interim Presentation Preparation	All	3																
3	Wiki Page Content Update	Kesmeen	3																
	Interim Report Due - 19 Feb 2017, Interim Project Presentation - 20 to 24 Feb 2017																		
	Communicate Results																		
1	Further Analysis from Feedbacks	All	3																
2	Create Visualizations	Jing Xiang	5																
3	Formulate Recommendations	Matthew	5																
4	Presentation to Sponsor	All	1																
	Final Project Review																		
1	Abstract & Full Paper Preparation	All	14																
2	Wiki Page Content Update	Kesmeen	3																
3	Project Poster	Matthew, Jing Xiang	3																
4	Final Presentation Preparation	All	5																
5	Final Paper Preparation	All	5																
	Undergraduate Conference Abstract & Full Paper Submission - 2 April 2017, Final Paper Submission - 16 April 2017, Conference Day - 22 to 23 April 2017																		

Scope of Work

This project will focus on the following scope:

- Data Collection & Consolidation
- Data Cleaning
- Data Transformation
- Data Exploration
- Cluster Analysis
- Machine Learning (Artificial Neural Networks)
- Survival Analysis

Milestones and Deliverables

- Project Proposal (15 Jan 2017)
    - Proposal Report
    - Wiki Page
- Interim Report (19 Feb 2017) & Project Presentation (20 to 24 Feb 2017)
    - Interim Report
    - Interim Presentation Slides
    - Wiki Page
- Abstract & Full Paper Submission (2 April 2017)
    - Abstract & Full Paper
- Final Paper Submission (16 April 2017)
    - Final Paper
    - Project Poster
    - Wiki Page
- Conference Day (22 to 23 April 2017)
    - Presentation Slides
    - Findings for Sponsors

Stakeholders

The primary stakeholders of this project are:

- Sponsor: Ms Elaine Tan, Asia Pacific Program Manager - Business Intelligence, GlaxoSmithKline
- Project Supervisor: Prof Kam Tin Seong, Associate Professor of Information Systems (Practice)

References

D. (n.d.). Analyzing Customer Churn – Basic Survival Analysis. Retrieved January 13, 2017, from http://daynebatten.com/2015/02/customer-churn-survival-analysis/

Despa, S. (n.d.). Censored Data. StatNews, (67). Retrieved January 10, 2017, from https://www.cscu.cornell.edu/news/statnews/stnews78.pdf.

Despa, S. (n.d.). What is Survival Analysis? StatNews, (78). Retrieved January 10, 2017, from https://www.cscu.cornell.edu/news/statnews/stnews78.pdf.

Sun, Z., Choi, T., Au, K., & Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. Decision Support Systems, 46(1), 411-419. doi:10.1016/j.dss.2008.07.009