Community Issues in American Metropolitan Cities:
A Data Mining Case Study

Brooke Sullivan, California State University, Fullerton, CA, USA

Sinjini Mitra, California State University, Fullerton, CA, USA

EXECUTIVE SUMMARY

The city of San Francisco in California has 826,000 residents and is growing slowly compared to other large cities in the western United States, facing concerns such as an aging population and flight of families to nearby suburbs. This case study investigates the social and demographic factors that are causing this phenomenon based on data that were collected by San Francisco's city controller's office in its annual survey to residents. By using data analytics, we can predict which residents are likely to move away, and this help us infer which factors of city life and city services contribute to a resident's decision to leave the city. Results of this research indicate that factors like public transportation services, public schools, and personal finances are significant in this regard, which can potentially help the city of San Francisco to prioritize its resources in order to better retain its locals.

Keywords: *Classification, Data Mining, Data Visualization, Demographic Analysis, Public Administration, Public Services, Resident Satisfaction, Survey*

INTRODUCTION

Cities are sometimes thought of as the "closest, most intimate form of government" (Funkhouser, 2013). Cities are therefore expected to have a greater ability to understand and satisfy residents' needs for public goods and services (Montalvo, 2009). For metropolitan cities, some of the important services include urban transportation systems such as subways and trains, and parks which break up the urban atmosphere. However, without tools that can interpret and explain residents' perceptions of their city, it can be difficult for city officials to judge which issues should be addressed first. Towards this end, cities often conduct surveys among their residents to seek their opinions about certain specific city services with the goal of determining which group of people are likely to move away. Once the survey results are obtained, the main challenge is to analyze the data using suitable methods and interpret them in order to gain valuable insights. The latter will help city officials to make important

DOI: 10.4018/jcit.2014010103

decisions about allocating resources towards specific improvement projects that may help retain more residents in the area.

Many statistical and analytical tools are available today for analyzing survey data. The term, "analytics," is the most current term for pattern recognition and knowledge discovery in large datasets (Duval, 2012). This can be done through data mining, predictive analytics, and statistics (Gartner, n.d.). In businesses, analytics tools are often used to predict which customers are at risk to stop buying, identify segments of customers, forecast business outcomes, and ultimately lead to business decisions. One form of business analytics is data mining, which refers to finding patterns in large amounts of data (Sharda, Delen, & Turban, 2013; Azzalini & Scarpa, 2012; Pearce, 2011). Data mining incorporates concepts from different fields, including mathematics, statistics, artificial intelligence, and database management (Sharda, Delen, & Turban, 2013). There are several industry buzz words that are going around currently and hence are worth discussing, including *business intelligence, big data*, and *business analytics*. "Business intelligence" refers to using collected data to present results through queries, reports and data mining (Gittlen, 2012). Another one is "Big data", which relates to the phenomenon that due to cheaper data storage and faster processing, organizations gather data at huge rates that previously could not have been comprehended (Barton & Court, 2012; AWS, n.d.). "Business analytics" is a term similar to business intelligence, but is more powerful because it does not rely on historical data (Gittlen, 2012). Analytics and data mining have existed under different names since the 1980s (University of North Carolina, 2003). Analysts predict that this field will continue to develop, and employers currently face a shortage of talent with skills in this field (Violino, 2012). Some popular examples of analytics are the recommender systems of Netflix and Amazon. Amazon has been using data mining tools since the late 1990s to create lists of bestselling books for special populations of people, such as those at the same workplace (Streitfield, 1999). Today, Amazon recommends products to shoppers based on associations in the purchase patterns of products (Furnas, 2012). Techniques from analytics and data mining are also present in the fields of healthcare, scientific research, banking, and retailing (Suh, 2012).

Analytics research can be useful in public administration applications because it has the ability to deliver meaningful results that can be used to improve the lives of people and give representatives a better picture of reality. Using analytics in the public sector can provide decision makers with information that can inspire new city projects and suggest initiatives that can be voted on, or even predict whether a ballot measure will be passed based on the sentiments of constituents. This type of research in cities is not common, however it holds great promise as a tool to better serve urban dwellers and perhaps even those in smaller cities.

San Francisco, a city known for its progressive policies, is currently facing issues that involve the flight of certain groups of residents to other cities in the Bay Area (Hartman & Carnochan, 2002). It is concerning that these key population of citizens might move away because it could cause San Francisco to lose part of its identity. For example, many families with children are moving away from San Francisco and wealthier residents are moving in. As a result, fewer residents will have grown up in San Francisco as children. Due to the cycle of wealthy residents moving to the city later in life, there may be fewer people that feel they have "roots" in their hometown. Finding out why these issues are occurring, why families and the working class are not satisfied with San Francisco, can suggest how city planners might address these problems and prevent San Francisco from losing its city identity. This case study will use data from San Francisco's biannual city survey in order to present how metropolitan cities can use analytics techniques to improve their operations and decision making.

The rest of the paper is organized as follows. Organization Background presents a brief background of the organization being studied here, namely the city of San Francisco. Setting the Stage covers San Francisco's current