Mining Customer Survey Data for SFO Airport Authority

Ankit Jain,
50097432

Deepak Veerupapuram
50098125

Milky Sahu
50096350

Abstract— Customer feedback is a marketing term that describes the process of obtaining a customer's opinion about a business, product or service. Customer feedback is so important because it provides marketers and business owners with insight that they can use to improve their business, products and/or overall customer experience. Measuring customer satisfaction is typically based on self-declared or interview-based questionnaires where users or consumers are asked to express opinions on statements, or satisfaction scales, mapping out various interactions with the service provider or product supplier. This paper presents an approach to extract dependency patterns from the customer survey dataset by using data mining techniques. With the proposed approach, features can be identified from the dataset in the form of both statistical results and visualizations that have strong influence on overall customer satisfaction. These visualizations and results help to generate a variety of management insights which in turn help to improve overall customer satisfaction.

Keywords— Ordinal Data, SFO Airport Customer survey dataset.

I. INTRODUCTION

Customer satisfaction is the utmost priority of every business in every domain. Be it banks, retailers, food and beverage manufacturers, transportation or service providers, all conduct self-declared or interview-based surveys on their customers to know their level of satisfaction of different grounds. Customers are requested to fill in questionnaires with typically 10 to 100 questions. The survey produces responses that can be considered as random variables. Some of these variables are responses to questions on Overall Satisfaction, Recommendation or Repurchasing Intention, which are considered target variables. Responses to the other questions can be analyzed under the hypotheses that they are positively dependent with the target variables. The combinations formed are either positive dependent or independent for each pair of variable. In general, dependency patterns can be extracted from data by using data mining techniques.

In this paper we analyze San Francisco International Airport's Customer Satisfaction survey data for SFO Airport Authority collected in the year 2013. The qualitative outcomes and derived conclusions are obtained using models like Bayesian Networks, Decision tree and K-Means Clustering.

Airport administrators and researchers often face problems in drawing conclusions from the information already present either in documented form or as undocumented experience and practice. The information may be fragmented, scattered, and unevaluated. As a consequence, full knowledge of what has been learned about a problem may not be brought to bear on its solution. Costly research findings may go unused, valuable experience may be overlooked, and due consideration may not be given to recommended practices for solving or alleviating the problem. SFO Airport being one of the busiest airports of the world caters to a wide range of customers and is liable to face such challenges. It is strongly needed that the authority keeps a log of the feedback of the customers and analyzes the data with powerful analysis techniques to overcome any prevalent or predictable problems. This paper examines the growing strategic importance of customer service and how SFO airport is measuring the quality of customer service. The analysis provides insights into functioning and performance of the airport, the weak areas which need special attention and the improvements needed to satisfy the customers belonging to different classes.

The survey initially consisted of 68 features and 3872 customer survey responses, out of which we chose 14 significant features to conduct the analysis, the major one being the assessment of Overall Satisfaction, with 13 other specific variables viz. Restaurants, Retail Shops, Signs and Directions, Escalators, Screen Information, Information Booth, Airport Parking, AirTran, Rental Car, Cleanliness, Safety Finding Way and Security Screening, evaluated on a seven-point anchored scale.

Novelty of our idea

Although there are different packages and algorithms available for analyzing customer survey data, but the thing that makes our idea unique is the selection and implementation of specific and highly efficient models to our problem. Similar type of analysis have been done for other survey datasets but there are no existing procedures to analyze airport data and improve customer satisfaction. We have tried to solve the problem in a way that provides both the pictorial (visualizations) and numerical results which makes it easy to be interpreted by business executives as well as subject matter expert.

II. ALGORITHMS AND RESULTS

Data mining using nominal, interval and ratio data is generally straightforward and transparent. Analysis of ordinal data (survey data), particularly as it relates to Likert or other scales in surveys, is not. An ordinal variable is one where the order matters but not the difference between values. For example, we might ask patients to express the amount of pain they are feeling on a scale of 1 to 10. A score of 7 means more pain that a score of 5 and that is more than a score of 3. But the difference between the 7 and the 5 may not be the same as that between 5 and 3. The values simply express an order.

Problems while formulating algorithms

1. The response variables are discrete, not continuous. Hence a regression model would not be appropriate.
2. An ordered choice model not only models the problem correctly, but estimating such a model will give us some idea of the significance of the independent variables.

We have used more than one data mining technique to a given data set as it increases knowledge about the results. In other words, combining models increases the derived utility from the application of such models to a certain data set.

A. Bayesian Networks (BN)

Bayesian networks implement a graphical model structure known as a directed acyclic graph (DAG). BNs are both mathematically rigorous and intuitively understandable. They enable an effective representation and computation of the joint probability distribution over a set of random variables.

Why Bayesian Networks?

1. Cause-and-effect diagram and is easy to interpret.

2. BN summarizes subject-matter knowledge and data-derived information.

3. Various possible improvement scenarios can be easily simulated and evaluated.

A BN can therefore be considered an innovative approach to support strategic decisions.

Initially, the BN is unknown and we needs to learn it from the data. This problem is known as the BN learning problem, which can be stated informally as follows: Given training data and prior information (e.g., expert knowledge, casual relationships), estimate the graph topology (network structure) and the parameters of the JPD in the BN. We have used *bnlearn* package from R to apply Bayesian Network model to our dataset.

The structure of a DAG is defined by two sets: the set of nodes (vertices) and the set of directed edges. The nodes represent random variables and are drawn as circles labeled by the variable names. The edges represent direct dependence among the variables and are drawn by arrows between nodes. In particular, an edge from node Xi to node Xj represents a statistical dependence between the corresponding variables. Thus, the arrow indicates that a value taken by variable Xj depends on the value taken by variable Xi, or roughly speaking that variable Xi "influences" Xj . Node Xi is then referred to as a parent of Xj and, similarly, Xj is referred to as the child of Xi. The BN analysis provides a visual causality map linking the various survey variables and target variable i.e. Overall Satisfaction from Airport facilities.

The objective is to understand what dimensions have a direct influence on Overall Satisfaction. The data is analyzed with a score-based algorithm (AIC criterion)

Figure 1 represent the BN of variables representing overall satisfaction from the various survey features.

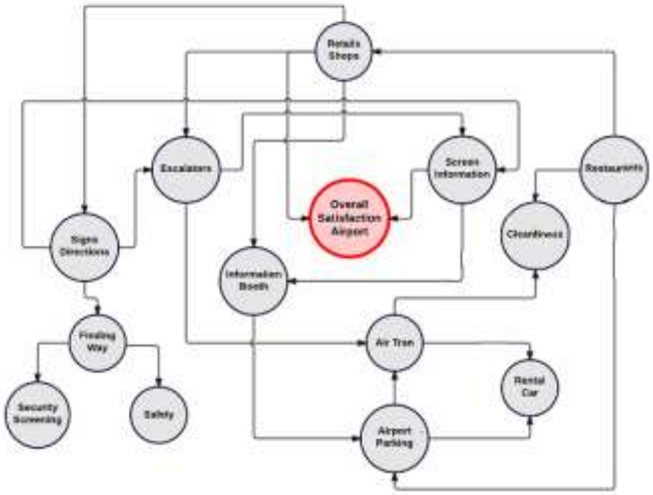

Figure 1

By studying the above network we can infer that an intervention to improve satisfaction levels from Screen Information at Airport or Retail Shops will increase Overall Satisfaction of customers at airport. The implication is that if the SFO Airport Authority increases the percentage of customers with top-level satisfaction from Screen Information at airport by introducing more screens or enriching the content displayed on screen, overall satisfaction of customers will also increase. The visual display of a BN makes it particularly appealing to decision makers who feel uneasy with mathematical models.

We further performed various simulations and assessed various scenarios. As an example, we changed the satisfaction profile of screen information and found out how the overall satisfaction level changed.

This model helped us to get an overview of dependency between the feature variables. The deeper analysis of this result was being done by using Tree Based methods.

B. Tree Based Methods (Optimal CART Tree)

Decision trees provide conceptually simple ways of understanding and summarizing the main features of the data; in particular, they exploit tree-graphs to provide visual representations of the rules underlying a given data set. The target variable is considered to be Overall Satisfaction with Airport facilities.

Classification Algorithms and Regression Trees (CART) model is used next on the same dataset. Dichotomization of target variable is obtained by aggregating the two highest levels on the one hand and the other lower ones on the other hand. The categories thus obtained are labeled ‘yes’ and ‘no’, respectively. We have used *rpart* and ‘*partykit*’ package from R to generate the below tree. The *rpart* package implements the CART methodology using cross validation techniques in the third step of the pruning procedure (Figure 2)

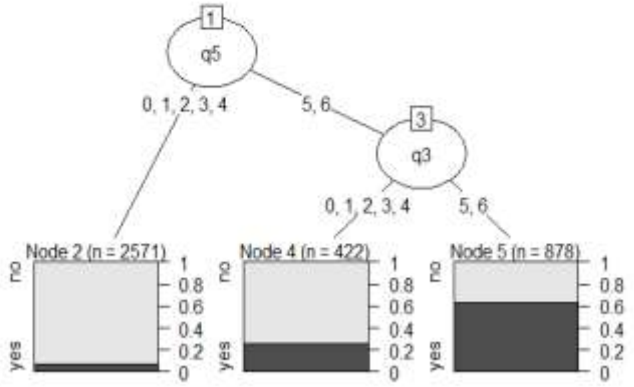

Figure 2

Here, the distribution of the dichotomized overall satisfaction with airport facilities within each terminal node is graphically represented.

This tree exploits the overall levels of satisfaction with Screen Information (q5) and Signs and Directions (q3) to partition the customers into three groups.

Node 5 contains customers with high levels of satisfaction with Signs and Directions and Screen Information; in this group the percentage of customers who are overall satisfied with SFO Airport Facilities is 61% (against 22.05% in the learning set as a whole).

Node 4 contains customers with high levels of satisfaction with Screen Information and low levels of satisfaction with Signs and Directions. In this group the percentage of customers satisfied with SFO Airport Facilities is around 21%.

Finally, customers with very low or intermediate levels of satisfaction with Screen Information are assigned to Node 2. In this group the percentage of customers satisfied with SFO Airport Facilities is close to 0%.

From the above figure 2, it is clear that the overall satisfaction with SFO Airport can be further increased if the satisfaction level can be increased with Signs and Directions.

C. K Means Clustering

The most common partitioning method is the K-means cluster analysis. We have used *flexclust* package from R to perform clustering for our dataset. For the first analysis, we have used this algorithm to divide the complete dataset into three clusters viz less satisfied, moderately satisfied and highly satisfied customers. (Figure 4)

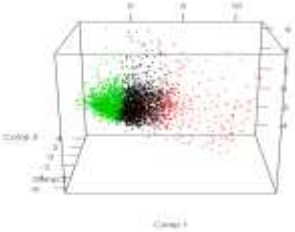

Figure 3

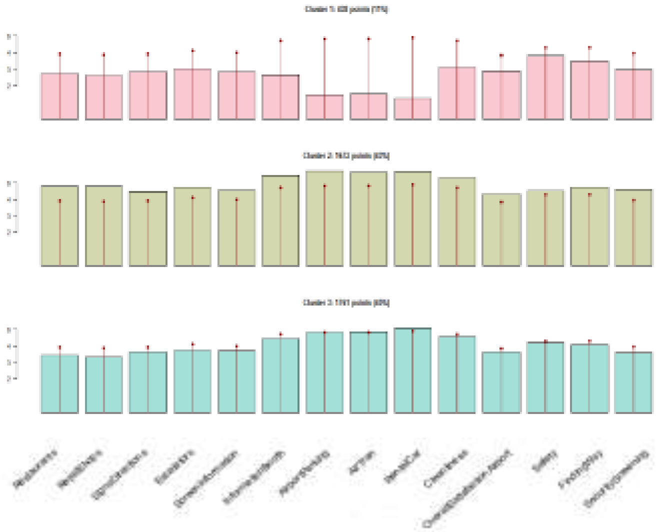

Figure 4

In the second analysis, we have done clustering on different terminals dataset to identify the terminal (Figure 5) with highest level of unsatisfaction. Once the terminal is identified, we have tried to identify the airport facilities in that terminal that contribute to the unsatisfaction level. Management decisions can then be taken to improve that particular feature and thus the overall customer satisfaction.

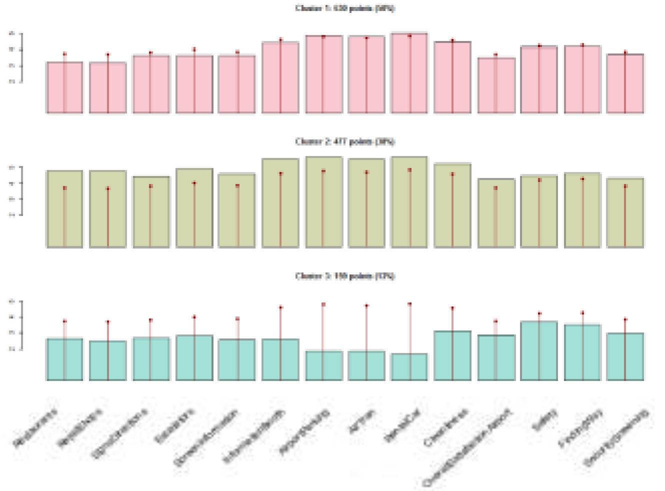

Figure 5

Figure 5, 6 and 7 shows the clustering plots for Terminal 1, 2 and 3 respectively. From these figures, it is clear that unsatisfaction with the airport facilities is highest in Terminal

1 and it is the same for Terminal 2 and 3 (cluster density of cluster 3 in all the three figures). For terminal 1, the cluster 3 i.e cluster with least satisfaction level responses is further partitioned at feature level to determine the feature that needs attention of SFO authority for improvement. In this case, average satisfaction level for airtran, airport parking and rental car is least, but considering the fact that airport parking and rental car facility is not directly dependent on terminal, so airtran (connectivity) needs attention. Improving the airtran facility will lead to improvement of overall customer satisfaction level.

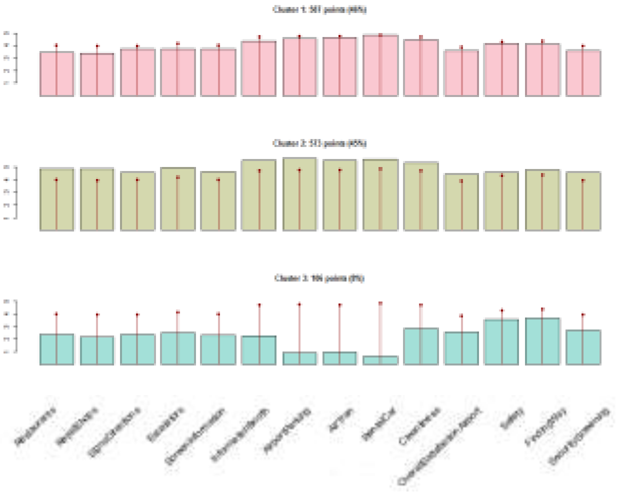

Figure 6

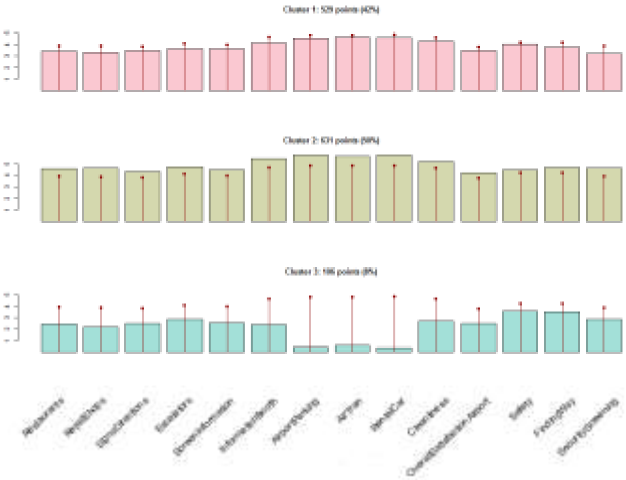

Figure 7

III. CONCLUSION

Managing airports today involves complex partnerships between public and private entities. From the customer's view, an airport is a monolithic organization. Airport staff or volunteers are often the first point of call when there is a problem. Increased use of self-service technology to check in, change seats, find the gate, order food, or return rental cars has sharply reduced passenger contact with airport and airline representatives and other business partners. The distributed nature of responsibility for customer service also makes it difficult for airports to know precisely how many resources (time, people, and money) are directed at customer service.

The survey on SFO airport shows direct dependency of overall satisfaction of its customers on the services and facilities it offers. Screen information and retail shops directly influence overall customer satisfaction. Signs and directions can be further improved to facilitate easy moving of the travellers. Parking facilities, rental cars at the terminal fail to satisfy customers.

IV. FURTHER RESEARCH

The measurement of customer service performance is based on an evolving understanding of what is really important to a customer's experience. Research suggests that cleanliness, courtesy of staff, processing times, gate experience, and concession choices are the most important factors that contribute to a customer's experience. However, it appears likely that factors not under the airport's control— delays getting to the airport, parking congestion, slow shuttle buses, bad weather, and flight cancellations— can drastically alter a customer's experience.

Further research could test these hypotheses and confirm which performance measures are most important to track. Airports are beginning to look in more detail into what customers want and what they are willing to spend. Most of the literature on customer satisfaction focuses on basic experiences at an airport, ranking cleanliness and fast processing through security checkpoints as important contributors to customer satisfaction. But it is also likely that what the business traveler perceives as excellent customer service is different from that of economy.

As customer service evolves at airports, so will measurement of performance. This synthesis provides a snapshot of current practice. The subject area invites further research in a number of areas:

1. *Passenger segmentation:* How different passenger groups rank satisfaction measures and how this understanding can improve the program design and delivery of parking products and other passenger services.

2. *Investment:* How airports evaluate the return on various customer service initiatives and make decisions about where best to allocate resources.

3. *Planning:* How information about customer satisfaction can improve efficiency in the terminal area and increase revenues from concessions and other airport services.

4. *Brand:* How the airport can integrate a culture of customer service and communicate what the airport will consistently deliver.

5. *Handbook:* A low-cost, customer service management strategy and service quality measurement system for small and non-hub airports.

V. REFERENCES

1. AIRPORT COOPERATIVE RESEARCH PROGRAM, ACRP SYNTHESIS 48 Sponsored by the Federal Aviation Administration
2. Modern analysis of customer satisfaction surveys: comparison of models and integrated analysis Ron S. Kenetta,
3. http://www.r-project.org/conferences/DSC-2003/Proceedings/BottcherDethlefsen.pdf
4. http://www.jstatsoft.org/v08/i20/paper
5. http://www.r-statistics.com/