Week 5 Assignment- Predictive Analytics Case Study

Guruprasad Velikadu Krishnamoorthy

College of Science and Technology, Bellevue University

DSC630-T302: Predictive Analysis

Professor. Fadi Alsaleem

January 12, 2024

Week 5 Assignment - Survey Analysis Case Study

1. Introduction

The case study is about the survey analysis project for the Worldwide Youth Organization (YMCA), from the Applied Predictive Analytics textbook (Abbott, 2014). YMCA, sometimes called Y, is one of the most prominent nonprofit organizations in the world that aims to strengthen the community by inspiring young people and improving the well-being of people of all ages with various fitness activities and programs to build strong and healthy minds and bodies. This case study focuses on the annual survey results across nationwide branches from 2001 and the steps and methods involved in extracting meaningful insights from the survey results.

1.1 What was the problem being solved?

The survey contained 48 multiple-choice questions targeting various areas such as the member's experiences with the club, staff competency, cleanliness, being valued or not, etc. The questions were rated from 1 to 5, with one being highly satisfied and five being highly dissatisfied. Out of the 48 questions, three questions were identified as the target variables, each focused on the Satisfaction factor at the club, the willingness to renew the membership, and if they were willing to recommend it to other friends. The three target variables had different degrees of association with each other. Hence, a new linear combination of the target variables called the Index of Excellence (IOE) was derived from these target variables.

Though regression models were built, the problem being solved was not to predict the combined form of three target questions- the IOE, but to identify the member attitudes associated with the target variables. The problem that the project was trying to solve was to determine the IOE of the current year and compare it with the previous year, measure the areas where the club had performed better, and identify the areas where it could be improved.

1.2 Why was the problem significant to solve:

The YMCA is a membership-based club where fees are collected from members every month. Hence, it is essential to ensure the members feel welcome in the club and are willing to extend their membership yearly. The project focuses on analyzing the performance of individual clubs by comparing their performance with previous years and with their peers. It is essential to measure the performance of the club and identify the areas of improvement to retain the club members and to improve their satisfaction factor.

1.3 How was the data acquired?

The data was acquired through the annual survey conducted each year in the 2500 branches across the country. For this project, 32811 survey responses collected for the year 2001 were used as inputs. The survey consisted of 48 multiple-choice questions, each coded from 1 to 5.

2. Methods and Results

2.1 What steps were taken to prepare the data?

To prepare the data for building the model, the questions were first classified based on the classification labels such as staff-related, Building-related, Equipment-related questions, etc. The number of missing responses, which was only 0.7 percent of the entire data population, was then validated. As the responses were in the range of 1 to 5, there were no outliers in the dataset. However, using 0 to represent the missing values may lead to incorrect outcomes, especially when the survey responses were Ordinal ranging from 1 to 5. So, the missing data was replaced by data imputation techniques, and a value that did the least harm was identified rather than just using zeros or the mean or median values. For this reason, the imputed values were determined using the linear fit method and were used to replace the 0 or missing values in the data. Further, for the Dimensionality reduction, the

Factor Analysis method was adopted to extract the factors with the variance cutoff, and ten factors were extracted for the model building.

2.2 What Modeling techniques were used?

Two modeling techniques were used to build the model. In the first approach, Stepwise Linear Regression was adopted to identify the factors that resulted in better accuracy in predicting the target value IOE. Instead of using the factors identified in the factor analysis method, a much better approach of using the questions that represented the question better was chosen to build the model. With this approach, rather than using all the data from the survey, only the key questions that contributed the most were used, which resulted in better R2 values.

In the second model-building approach, the Decision Tree algorithm was used to gain insights from the data due to its ease of interpretation. Different models were built for each target value, such as the satisfaction factor, intention to renew the membership, and willingness to recommend it to a friend.

2.3 Why did the team choose the models/methods they did?

In the first model using Stepwise Linear Regression, seven primary questions were identified, and visuals were built to represent how well the club performed compared to its peers. The intention was to highlight the areas of improvement needed for the club, such as staffing, better facilities, upgrading equipment, etc., thus increasing the Index of Excellence (IOE) for the club. The visuals also had data points from the previous year that helped understand how the club performed compared to the previous year. However, the decision-makers felt that the visualization was too complex to understand and preferred an easy-to-understand solution. Hence, the Decision tree algorithm was chosen to build the second model. It was not only easy to implement but also easy to gain insights from as the visuals are represented in a tree-like structure. The decision-makers felt that the rules described

in plain English from the visuals were easier to understand than the Mathematical equations. Hence, the Decision tree model was finalized for this project.

2.4 What metrics were used to evaluate the results? Why was this metric chosen?

In the first model where the traditional approach of stepwise linear regression was used, metrics such as R-squared values, standard errors, and the respective p-values were used to determine the significant predictors of the target variable. Also, AIC (Akaike Information Criterion) was used to exclude the features that resulted in the reduction of accuracy. In the Decision Tree model, accuracy was used as a metric to evaluate the model based on the various rules derived from the trees. Alternatively, key statistics, such as the percentage of the population in the terminal node and the percentage of all highly satisfied populations in the terminal nodes, were used to compute the accuracy of the model.

2.5 How was the problem solved?

In the traditional approach of using stepwise linear regression, the problem was solved by using visuals to represent the results. The seven key questions identified by the model were represented on the x-axis in order of their importance. The importance of these questions in determining the target variable IOE was represented on the y-axis. The questions on the x-axis were each described as balls, and the color of the ball decided if the performance was better or worse compared to the previous year. The size of the ball represented the relative importance of the question to IOE for that club. Also, the measures of the prior year were described as small circles on the graph to estimate the relative performance compared to the previous year. Due to the complexity of understanding these visuals, the decision-makers opted to build an easy-to-understand visualization using Decision Trees.

In the second approach, the machine learning approach of Decision trees was used to build the model, and the problem was solved by displaying the outcome of the model as rules in plain English. Various key statistics were created from the models to extract insights from them. For instance, the

Satisfaction model was used to identify the top three rules that constitute the most highly satisfied members from the key statistics that the previous model did not provide. Thus, the top three nodes and the respective questions that included most of the variability in the target population were identified, and the interaction factor between them was analyzed. The interaction effect between the features that were not available in the previous model, provided more insights and recommendations. For instance, one of the outcomes of the model was that the members were most satisfied when they felt welcome and when the facilities were clean.

3. Conclusion

3.1 How were the results/model implemented:

The outcome of the models built using the Decision tree was analyzed, and the key questions that constituted most of the variance in the target variable were extracted for further analysis. The key questions from each model were compared to identify how they provided insights about the attitudes of the members. For instance, the feeling of being welcome in the club was found to be more important than the parking, facilities of the club, competence of the staff, etc. So, more emphasis was placed on improving the feeling of being welcome.

3.2 What were the actionable consequences of the case study?

As the model identified key questions for each target variable, the clubs chose to utilize the key areas where improvements can be achieved. For instance, if the member satisfaction rates were low in a club, they would focus on making the members feel welcome by training the staff and improving the cleanliness of the club. Though the model was not implemented across all the clubs across the country, the clubs chose the extent of improvement required in their club and focused on those key parameters. After a few years of implementation, the model resulted in about 32 percent improvement in the satisfaction rate, and Recommend to Friends improved by about 6 percent.

3.3. What did the team learn from the case study?

The team learned that though two different models were developed, there was no right or wrong answer to either of them. The first model was built using the traditional approach of stepwise linear regression focused on providing insights about the individual club and its relative performance nationwide alongside the comparison from the previous year. The other approach using a Decision tree was adopted due to the complexity of understanding the results of the first one. Hence, it was better to build visuals that are easy to explain and interpret. Another learning from the study, which was already implemented in the project, was the importance of identifying the correct Target variables, as the execution of the entire project relies upon determining the appropriate target variables.

3.4 How should or would the team approach the problem differently in the future?

The model built using the first approach was discontinued due to its complexity, though it provided many valuable insights into the attributes of the club compared to its peers. So, a few things can be done differently in the future, such as building visuals that are easy to understand and interpret. Also, the data can be tested on a limited population, and based on the response, further effort can be put into further expanding or deploying the model. Another thing that can be done differently is to combine insights from all the models built and generate insights from all of them.

4. References

Abbott, D. (2014). Applied predictive analytics: principles and techniques for the professional data analyst. Wiley & Sons.