**Survey Analysis Project for the YMCA**

**What was the problem being solved?**

Identify survey questions related to the three target questions so that individual YMCA branches can evaluate how they can improve member satisfaction, likelihood to renew, and recommendations to friends.

**Why was this problem important to solve?**

This problem is important because this helps in evaluating how well the YMCA is meeting the needs and expectations of its members there by improving the lives of its members and to build

communities.

**How was the data acquired?**

The YMCA developed an annual member survey for use by its branches. For the Analysis project, 32,811 surveys with responses were made available for the year 2001. Modeling data from the survey contained 48 multiple choice questions coded with values between 1 and 5, where 1 was the most positive response and 5 the most negative. Questions were primarily attitudinal, related to the member’s experience with the Y, though four questions were demographic variables.

**Methods and Results**

**What steps were taken to prepare the data?**

For the Regression model, any NULL values were recoded as 0. The decision tree algorithm used in the analysis could handle missing values, no data preparation was done to recode NULLs or 0s. However, for each question, dummy variables were created to indicate if the responder checked the “1” box or not;

**How was this problem solved?**

This problem was solved with the help of decision trees which are often used to gain insights into data because rules are easier to interpret than mathematical equations, especially if the rules are simple.

**What modeling techniques were used?**

A linear regression model with a new target variable as a combination of the three target values to characterize members with a highly positive option of Y branch was used as an initial model. Factor Analysis was applied to determine the top 10 factors. An alternative approach of using the questions that loaded the highest on each factor was also used.

Decision trees were built using a CART-styled algorithm and complexity was determined using cross-validation, a standard CART algorithm practice. Models were built for each of the three target variables: Satisfaction, intend to renew, and Recommend to a friend.

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

After a comparison of the two approaches mentioned in the stepwise linear regression, using individual questions as inputs rather than the factors, generated higher R-squared, and therefore was the model selected for use.

Also, decision-makers found the visualizations based on the regression models were too complex; it wasn’t clear to them exactly what the information meant for their individual branch.

Therefore, decision trees were created to uncover key questions in the survey related to the three target variables that provided a more transparent interpretation of the survey.

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

R2 was used to determine the best model in the two approaches in the linear regression.

R-Squared or the coefficient of determination is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model.

Also, Ease of understanding the visualization was used to decide to go with the Decision tree model. Decision trees are often used to gain insights into data because rules are easier to interpret than mathematical equations.

**Conclusion**

**How were the results or model implemented?**

In the linear regression models, a new derived target variable: a simple linear combination of the three questions, called the Index of Excellence (IOE) was created. Feature Creation and Selection was done through Factor Analysis. Factor Analysis was used to determine the top ten factors that were interesting and comprised enough of the variance to be a good cutoff. Each of the factors was named according to the questions that loaded highest on the factors. 50% of data was used for training and the remaining 50% was used for the out-of-sample testing.

A traditional approach of stepwise linear regression to predict IOE with a few key questions and the factor analysis projections as inputs were used. Decision trees were built using a CART-styled algorithm and determined complexity using cross-validation, a standard CART algorithm practice. The algorithm identified surrogate splits in the tree which were used to help understand the variables that were the most important to predict to target variables. Models

were built for each of the three target variables: Satisfaction, intend to renew, and Recommend to a friend.

**What were the actionable consequences of the case study?**

Individual branches decided to what degree the models from the case study were used to influence what changes were made within their branches. After years of using models, a 32 percent improvement in satisfaction was measured, which clearly indicates improvement at the branch level in meeting the expectations of members. Additionally, the Recommend to a friend improved by 6 percent, easily a statistically significant improvement based on the number of surveys, though operationally not much higher than the original value.

**What did the team learn from the case study?**

Seer Analytics, the company who did the Analysis for Y, learned that neither of the models were right or wrong and that they provided different ways of understanding the survey. The team has progressed well beyond the models mentioned in the case study and has developed a framework that addresses the target variable from a different perspective; they readdressed the objectives of modeling to better match the business needs. Instead of predicting influencers for factors that drive satisfaction, they now define six key components of the member experience and create a pyramid of factors that are hierarchical.

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

The two model approaches that are dealt in this case study are neither right nor wrong. They just provide different ways to interpret the survey. Other models could have been tried including the combination of the two models that are implemented in the case study if they have resources and time.