Team Project

The Role of Organizational Factors in Nursing Home Quality

BUDT 733 – Data Mining for Business

Erica Eisenhart
Alon Gotesman
Audra Johnson
Ben Meadema
Shalin Saini

EXECUTIVE SUMMARY

Thousands of Americans reside in nursing homes across the US, with facilities spanning a wide range of sizes, occupancy rates, staffing levels and expertise, available services, business models and managerial expertise. This great variability between each facility results in a highly divergent quality levels, creating a substantial challenge for families across the country: which home should they choose for their elderly relativesThe Centers for Medicare & Medicaid Services (CMS) aims to introduce transparency into this process. The agency collects, measures, analyzes, rates and publishes a wealth of data on US nursing homes operating under the federal Medicare or Medicaid programs (these account for the vast majority of US nursing homes). But while intuition suggests the information should contain predictors to quality, it was unclear to our group which ones were truly significant.Our group set out to identify facility attributes that best explain CMS quality rating – scores that measure the physical and clinical conditions of nursing home residents. Using this information we believed prospective patients and their families would be best equipped to make educated decisions on nursing home selection.Interestingly, this task turned out to be more difficult than expected, as initial attempts at both data visualization and modeling were futile; none of the facility attributes provided by the CMS could be used to explain the variability in CMS quality ratings. Looking for a new approach we enhanced our dataset by adding a substantial amount of demographics information based on nursing home location. While not sufficiently explaining all variability, some useful results emerged as it became clear that demographic attributes such as average income and education level were significant in explaining variability in quality ratings.Once the logistic model was run, there were some non-intuitive signs for some of the coefficients of the predictor variables, as described in the table above. We assessed these non-intuitive results and determined that several variables could be correlated to the patient mix that a particular nursing home has. Nursing home clinical quality depends on the health outcomes and care of the home’s patients. Patients who are sicker may have poorer health outcomes and be more difficult to care for, resulting in potentially lower quality scores. Therefore, variables that may be impacted by the patient mix in a nursing home may not be showing the true effect of that organizational characteristic. The Multiple R-squared value as well as the error rate on the validation data set suggests that while this model provides us with some valuable explanation the overall fit of the model is not complete and that there are several other factors that affect the quality rating of nursing homes. Factors related to the patient may be overshadowing the positive organizational characteristics of the nursing home. Given the direction that sicker patients may be driving the quality score more than the operational efficiencies and other quality characteristics of the nursing home, we

suggest risk adjusting the quality scores and then completing the analysis to better understand the role of organizational characteristics.

Technical Summary

The goal of our analysis was to identify the factors that most influence nursing home quality, using data published by the Centers for Medicare and Medicaid Services (CMS). The dataset includes many operational characteristics and clinical quality measures of nursing homes, and must be submitted to CMS by all organizations participating in Medicare and Medicaid. The data is then combined into an aggregate measure of clinical quality, ranked from 1 (lowest) to 5 (highest) stars. CMS publishes all results online, freely available to consumers in the process of selecting a nursing home. Additionally, our group augmented the dataset with information about the environment in which the nursing homes operate, including county-level data from the census. For each nursing home, we appended county demographic data including income, education, population, and age distribution, as we hypothesized that these factors may also influence the types of residents at each nursing home. By extension, we believed the residents' characteristics could influence the severity of their illnesses, which would likely impact the clinical quality ratings of facilities due to circumstances beyond those they can control (staffing levels, organizational strategies, etc).To improve the quality of care delivered in nursing homes, CMS and other interested payers would need to identify the key factors that shape the quality of care that is delivered. With the understanding of important factors, CMS and other payers could create incentives for nursing homes to change their characteristics or offer technical assistance to improve in those areas. At the beginning of the project, we clearly defined our task of understanding the predictors of quality as an explanatory exercise. Throughout our tasks of visualization, data preparation, iterative model development, model comparison, and final model selection, we hoped to create a parsimonious model with interpretable results. The following sections describe the process we used to select our final model, as well as the interpretation of the results.

Data Acquisition and Preparation

CMS data is freely available online and is impressively usable, so very little formatting or other preparation work was needed. After combining several tables together using simple join operations we were left with a data set that included 19 attributes for 15,710 nursing home records. CMS data includes organizational characteristics such as facility size and occupancy rates, staffing levels, ownership type, and business status/governance. Using CMS's five star quality ratings system, we categorized nursing homes into two groups: low-quality (1-3 stars) and high-quality (4-5 stars). This task slightly reduced our data set, as several homes did not have a quality rating.

Visualization

In completing our visualizations, we used many plots and charts to analyze each available variable against nursing home quality type. We were able to identify a few variables displaying a large difference between the low and high quality nursing homes. Some of these have been included in Exhibit A of the Appendix. Additionally, we suspected that at least some variables that displayed only small differences may still be significance, as we reasoned that the large size of data set may have made small differences difficult to see. Furthermore we hypothesized that additional factors may be influential in determining nursing home quality, although this required additional data.

More Data Preparation and Visualizations

Our group believed that demographics information based on the location of each facility would potentially provide further explanatory power to our visualizations and modeling. After some work we located this information in the form of census data aggregated on a county level. As our CMS dataset included county name we were able to append the demographics information relatively easily. Once completed, we began transforming some attributes to help with interpretation and ensure parsimony. These transformations included binning facilities into one of ten geographic regions and creating dummy variables for the ten nursing home ownership types. Later we explored combined the nursing home ownership categories into three broad areas: for-profit, non-profit, and government-owned. We also considered potential interaction terms that could be used, but ultimately did not include these due to difficulties in interpretation. Finally, we rescaled some variables for a more meaningful interpretation, including transforming income per capita to a scale measured in thousands. With a fully prepared dataset we created training and validation sets -- despite the fact we pursued an explanatory task this step was both necessary (due to XLMiner's 10,000 record limit) and useful (with the luxury of a large dataset we were able to gauge the model's predictive powers).

Selection of an Analysis Method

To select our analysis method, we reviewed the available tools for explanation. We determined that discriminant analysis and logistic regression were most likely to be good tools for our problem, given our data size and number of variables. We ran a preliminary model of each type to determine which could be most useful. Discriminant analysis did not produce a significant difference between the high quality and low quality nursing homes. The inputs for the classification score for each variable had very small differences across the high and low quality nursing home types. Results for our test discriminant analysis are included in Exhibit B.However, the logistic regression model did have potentially positive results. Several variables had coefficients that were statistically significant, with a p-value less than .05. Based on these results, we decided to pursue a logistic regression model.

Iterative Output

Once we decided on using a logistic regression model, we set out to find the model with the greatest explanatory power. We conducted several iterations of the model using different explanatory variables. Our initial analysis was limited to 30 variables through XLMiner. To begin, we used variables from the CMS data set, along with census data variables that we believed to be important given our domain knowledge. A list of the different models we ran is included in the Appendix as Exhibit C, where significant variables are designated with an asterisk. Variables related to nursing home size, whether the nursing home was part of a continuing care retirement community, or whether the nursing home had a resident or family council to help voice individual residents' concerns were not found to be significant. Several variables were found to be significant, such as the type of insurance accepted, the region in which the nursing home is located, the staffing levels to care for residents, the type of ownership, and income and education of people in the county in which the nursing home is located. Through our iterations, we found that explanatory power, measured by the multiple R-squared, increased when we added variables from the census data to the variables from the original nursing home data set. This is likely due to the impact of adding more variables to the model.

Final Model Selection and Results

For our final model, we included all variables that were significant in previous models. In this model, each variable or at least one in a family of dummies was found to be significant. This model had the highest R-squared of all of the models we tested. The model and its results are included in the Appendix in Exhibit D. The R-squared value for the model is 0.0427. This shows that even though several variables are significant, there are several other factors that contribute to nursing home quality. Significance may also be attributed to the large amount of data included in our data set. The model does have some parsimony, as we used the binned, broader nursing home ownership categories, rather than the specific type. However, as there are many dummy variables, the total number of variables is still high. The coefficients and results of the model are interpreted below.

Interpretation and Recommendations

The final model as attached in Exhibit D is a logistic regression model that was run with the following predictors –

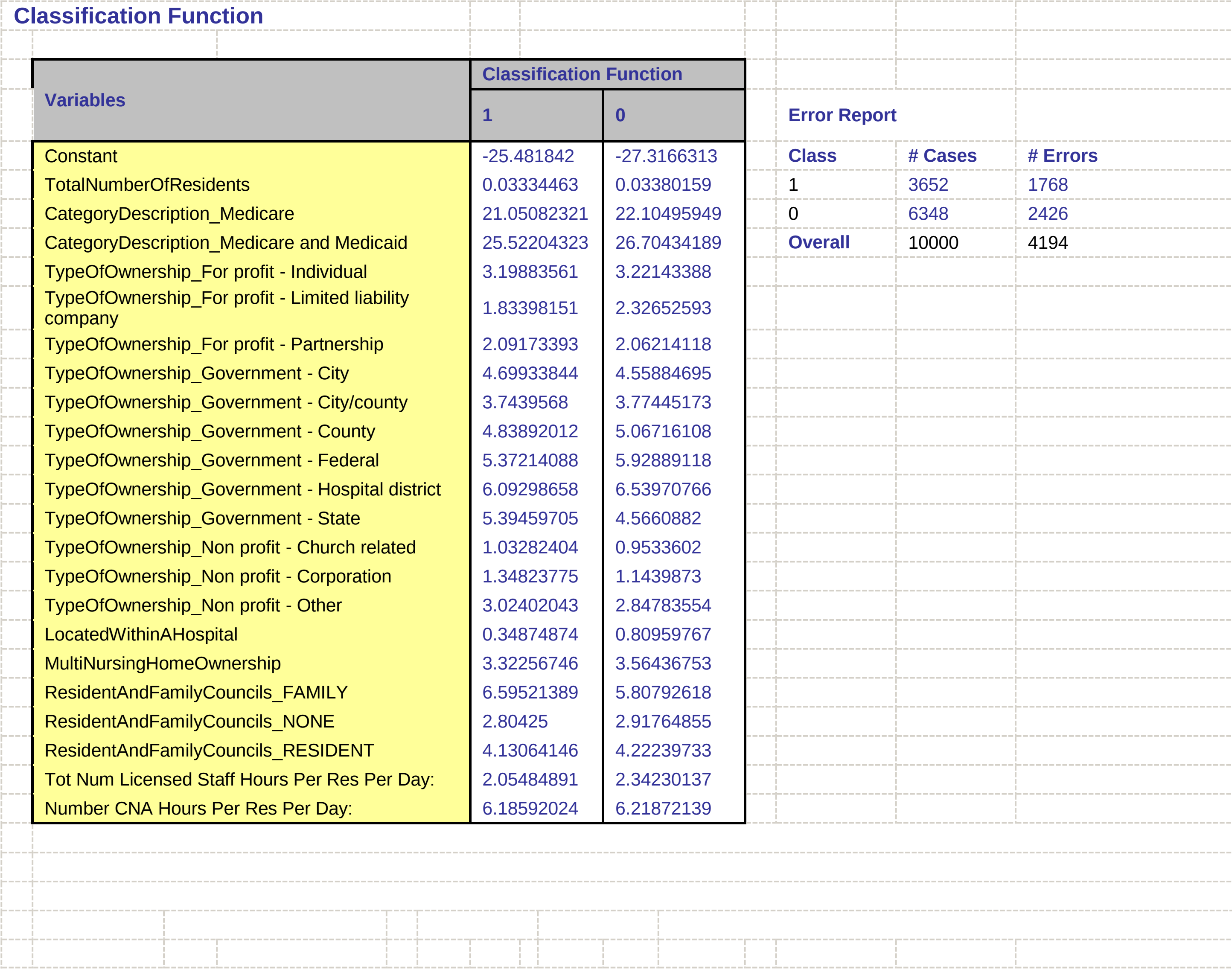
Predictors	Definition	Coefficient Interpretation
Region Category	Logical grouping of the US states as determined by the first digit of the postal zip code. For example Region_1 contains the states Delaware (DE), New York (NY), Pennsylvania (PA).	Region 9 and Region 5 are statistically significant region category variables. Region 9 has a negative coefficient suggesting a lower quality rating for nursing homes located on the west coast states (including Alaska and Hawaii) while the positive coefficient sign for Region 5 suggests a higher quality rating for nursing homes located in the mid-west states, compared to Region 10 (the northwest).
Medicare/Medicaid Category	Determines whether the nursing home either accepts Medicare only, Medicaid only, or both Medicare and Medicaid.	Nursing homes that accept both Medicare and Medicaid or just Medicare have a higher quality rating as compared to those that accept Medicaid only. This is an interesting finding as Medicare is considered to be a preferable payer with higher reimbursement than Medicaid, which would suggest that quality should be higher at homes that take Medicare. This finding may demonstrate that characteristics about Medicaid and Medicare patients may be different and driving the clinical quality score of homes.
Broader Ownership Category	Determines whether the nursing home is Government, non-profit, or for-profit owned .	Government owned and non-profit owned nursing homes are more likely to be of a higher quality than for-profit nursing homes.
Located Within A Hospital	Determines whether the Nursing home is located within a hospital.	The odds of a nursing home to be categorized as high quality decreases by a factor of 0.6289 if it is located within a hospital. This direction is also interesting, as one would expect higher quality clinical care in a nursing home at a hospital. However, the direction may also be impacted by patient factors, for example, patients in nursing homes attached to hospitals may be sicker and may bring down the clinical quality of the home.
MultiNursingHomeOwnership	Determines whether the nursing home is owned by an entity that owns multiple nursing homes	The odds of a nursing home to be categorized as high quality decreases by a factor of 0.9005 if it is one of many homes owned by a single entity.
Tot Num Licensed Staff Hours Per Res Per Day	Numeric variable containing the total number of licensed staff hours per resident per day.	The odds of a nursing home to be categorized as high quality decreases by a factor of 0.6421 per one unit of increase in the total number of licensed staff hours per day. This finding is counterintuitive. One would think that increasing staffing hours per resident would improve quality of care. However, this variable may also be impacted by the patient population. Sicker patients may require more staffing levels, and sicker patients may have poorer clinical outcomes, leading to lower clinical quality scores.
Percent with HS education	Educational attainment - persons 25 years and over - percent high school graduate or higher	The odds of a nursing home to be categorized as high quality decreases by a factor of 0.9628 with a unit increase in the percent of high school graduates or higher within the county where the nursing home is located.
Percent with college education	Educational attainment - persons 25 years and over - percent bachelor's degree or higher	The odds of a Nursing home to be categorized as high quality increases by a factor of 1.01928842 with a unit increase in the percent of bachelor's degree or higher within the county where the nursing home is located.
Income per capita	Per capita income in a county where the nursing home is located	The odds of a Nursing home to be categorized as high quality increases by a factor of 13.45 with a $1000 increase in the revised average household income of the county where the nursing home is located.

Once the logistic model was run, there were some non-intuitive signs for some of the coefficients of the predictor variables, as described in the table above. We assessed these non-intuitive results and determined that several variables could be correlated to the patient mix that a particular nursing home has. Nursing home clinical quality depends on the health outcomes and care of the home's patients. Patients who are sicker may have poorer health outcomes and be more difficult to care for, resulting in potentially lower quality scores. Therefore, variables that may be impacted by the patient mix in a nursing home may not be showing the true effect of that organizational characteristic. The Multiple R-squared value as well as the error rate on the validation data set suggests that while this model provides us with some valuable explanation the overall fit of the model is not complete and that there are several other factors that affect the quality rating of nursing homes. Factors related to the patient may be overshadowing the positive organizational characteristics of the nursing home. Given the direction that sicker patients may be driving the quality score more than the operational efficiencies and other quality characteristics of the nursing home, we suggest risk adjusting the quality scores and then completing the analysis to better understand the role of organizational characteristics.

Appendix

Exhibit A. Visualization Examples Demonstrating Few Differences Across High and Low Quality

Exhibit B. Initial Model Output: Discriminant Analysis Results: Success Class- High quality nursing home



Classification Function

Variables	Classification Function	
	1	0
Constant	-25.481842	-27.3166313
TotalNumberOfResidents	0.03334463	0.03380159
CategoryDescription_Medicare	21.05082321	22.10495949
CategoryDescription_Medicare and Medicaid	25.52204323	26.70434189
TypeOfOwnership_For profit - Individual	3.19883561	3.22143388
TypeOfOwnership_For profit - Limited liability company	1.83398151	2.32652593
TypeOfOwnership_For profit - Partnership	2.09173393	2.06214118
TypeOfOwnership_Government - City	4.69933844	4.55884695
TypeOfOwnership_Government - City/county	3.7439568	3.77445173
TypeOfOwnership_Government - County	4.83892012	5.06716108
TypeOfOwnership_Government - Federal	5.37214088	5.92889118
TypeOfOwnership_Government - Hospital district	6.09298658	6.53970766
TypeOfOwnership_Government - State	5.39459705	4.5660882
TypeOfOwnership_Non profit - Church related	1.03282404	0.9533602
TypeOfOwnership_Non profit - Corporation	1.34823775	1.1439873
TypeOfOwnership_Non profit - Other	3.02402043	2.84783554
LocatedWithinAHospital	0.34874874	0.80959767
MultiNursingHomeOwnership	3.32256746	3.56436753
ResidentAndFamilyCouncils_FAMILY	6.59521389	5.80792618
ResidentAndFamilyCouncils_NONE	2.80425	2.91764855
ResidentAndFamilyCouncils_RESIDENT	4.13064146	4.22239733
Tot Num Licensed Staff Hours Per Res Per Day:	2.05484891	2.34230137
Number CNA Hours Per Res Per Day:	6.18592024	6.21872139

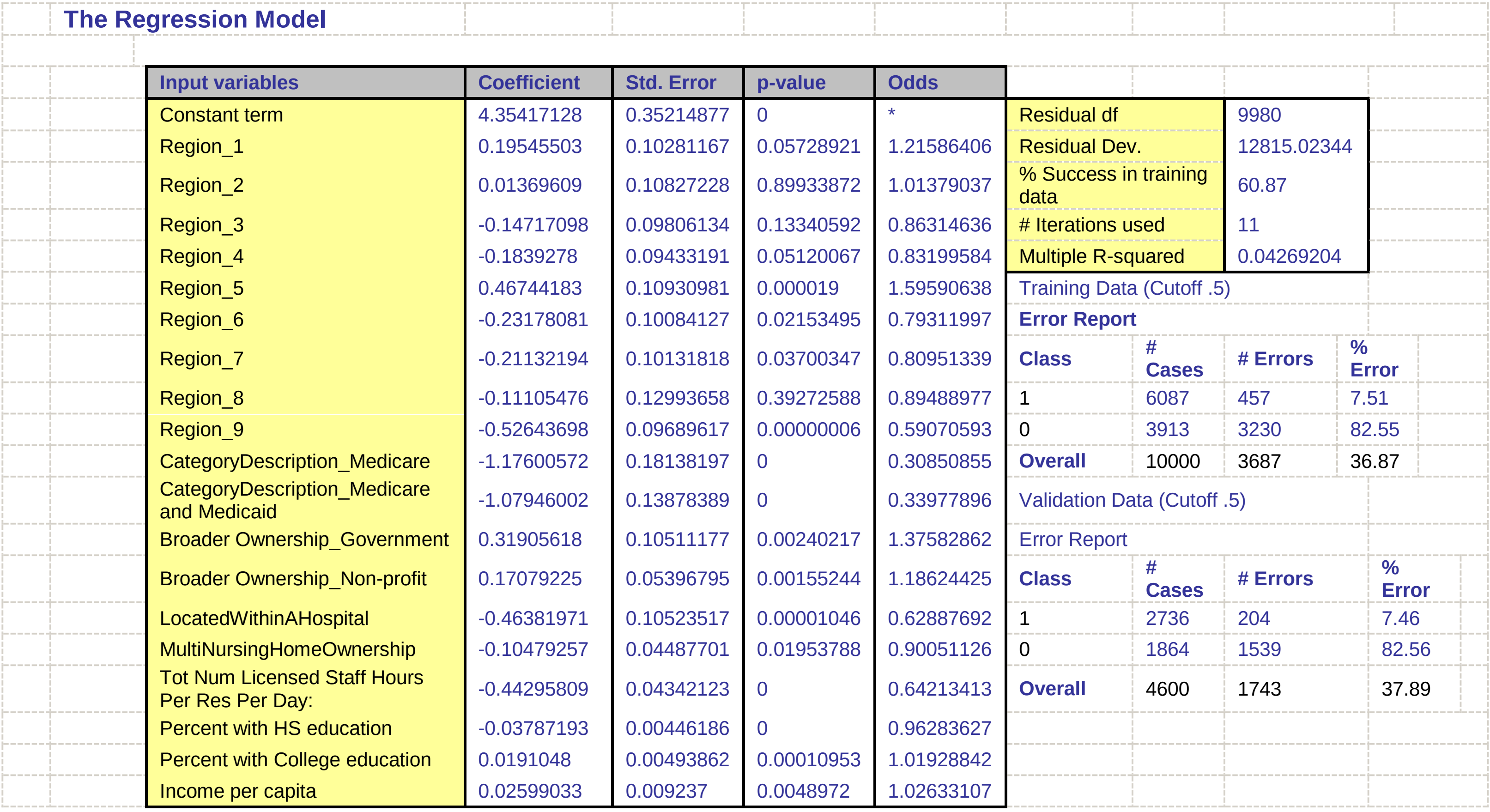
Error Report

Class	# Cases	# Errors
1	3652	1768
0	6348	2426
Overall	10000	4194

Exhibit C. Interim Models- Variables with asterisks are significant

Model 1 (R-squared: 0.0215)	**Model 2 (R-squared: 0.0279)**	**Model 3 (R-squared: 0.0275)**	**Model 4 (R-squared: 0.0344)**	**Model 5 (R-squared: 0.0397)**
• Region*• Acceptance of Medicare/Medicaid• Nursing home ownership type• Located within a hospital*• Multi-nursing home ownership*• Presence of resident and family council• Located in a continuing care retirement community• Licensed staff hours per resident per day*• Nursing assistant hours per resident per day• Per capita income	• Total number of residents• Acceptance of Medicare/Medicaid*• Nursing home ownership type*• Located within a hospital*• Multi-nursing home ownership*• Located in a continuing care retirement community• RN hours per resident per day*• LPN hours per resident per day*• Nursing assistant hours per resident per day• Per capita income*	• Acceptance of Medicare/Medicaid*• Nursing home ownership type*• Located within a hospital*• Multi-nursing home ownership*• RN hours per resident per day*• LPN hours per resident per day*• Nursing assistant hours per resident per day• Per capita income*	• Acceptance of Medicare/Medicaid*• Nursing home ownership type*• Located within a hospital*• Multi-nursing home ownership*• RN hours per resident per day*• LPN hours per resident per day*• Nursing assistant hours per resident per day• Per capita income*• Percent of population over age 65• Percent of population over age 25 with a high school diploma*	• Region*• Nursing home ownership type*• Located within a hospital*• Multi-nursing home ownership*• RN hours per resident per day*• LPN hours per resident per day*• Nursing assistant hours per resident per day• Per capita income*• Percent of population over age 65• Percent of population over age 25 with a high school diploma*• Percent of population over 25 with a college degree*

Exhibit D. Final Model Output



The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	4.35417128	0.35214877	0	*
Region_1	0.19545503	0.10281167	0.05728921	1.21586406
Region_2	0.01369609	0.10827228	0.89933872	1.01379037
Region_3	-0.14717098	0.09806134	0.13340592	0.86314636
Region_4	-0.1839278	0.09433191	0.05120067	0.83199584
Region_5	0.46744183	0.10930981	0.000019	1.59590638
Region_6	-0.23178081	0.10084127	0.02153495	0.79311997
Region_7	-0.21132194	0.10131818	0.03700347	0.80951339
Region_8	-0.11105476	0.12993658	0.39272588	0.89488977
Region_9	-0.52643698	0.09689617	0.00000006	0.59070593
CategoryDescription_Medicare	-1.17600572	0.18138197	0	0.30850855
CategoryDescription_Medicare and Medicaid	-1.07946002	0.13878389	0	0.33977896
Broader Ownership_Government	0.31905618	0.10511177	0.00240217	1.37582862
Broader Ownership_Non-profit	0.17079225	0.05396795	0.00155244	1.18624425
LocatedWithinAHospital	-0.46381971	0.10523517	0.00001046	0.62887692
MultiNursingHomeOwnership	-0.10479257	0.04487701	0.01953788	0.90051126
Tot Num Licensed Staff Hours Per Res Per Day:	-0.44295809	0.04342123	0	0.64213413
Percent with HS education	-0.03787193	0.00446186	0	0.96283627
Percent with College education	0.0191048	0.00493862	0.00010953	1.01928842
Income per capita	0.02599033	0.009237	0.0048972	1.02633107

Residual df	9980
Residual Dev.	12815.02344
% Success in training data	60.87
# Iterations used	11
Multiple R-squared	0.04269204

Training Data (Cutoff .5)

Error Report

Class	**# Cases**	**# Errors**	**% Error**
1	6087	457	7.51
0	3913	3230	82.55
Overall	10000	3687	36.87

Validation Data (Cutoff .5)

Error Report

Class	**# Cases**	**# Errors**	**% Error**
1	2736	204	7.46
0	1864	1539	82.56
Overall	4600	1743	37.89