Predicting Patient Noncompliance Based On Geographic Location

Case History

Category: *Healthcare*

Methods: *Predictive Modeling, Logistic Regression, Random Forest™, Advanced Analytics*

Summary

Decision Analyst explored the use of self-reported behavior from surveys to develop predictive models of noncompliant patients (those who have been told to take a prescribed medication for a chronic illness, but are not currently taking the medication).

The National Health and Nutrition Examination Survey (NHANES), published by the National Center for Health Statistics (NCHS) at www.cdc.gov, provided the noncompliance indicator variable for hypertension (high blood pressure). If the patients had been told on two or more different visits to a health professional that they had hypertension, and had been told to take a prescribed medicine for the condition, but were not currently taking the medication, then we considered them to be noncompliant for this research.

Predictive models were optimized using best-practice, cross-validation methods to select the best predictor variables. Two model types were investigated: logistic regression and Random Forest™ (an ensemble model that combines many models to produce more robust predictions). The study found that the Random Forest™ model with 14 variables delivered the highest overall predictive accuracy.

Census-tract TIGER/Line® shapefiles available at www.census.gov/geo/maps-data/ include several demographic variables that were also available within the NHANES data (namely, age, gender, family size, and the interaction of age with gender). This set of common variables was used to develop a geographic-variables-only model of noncompliance, which was optimized with logistic regression and included 10 age categories plus selected interactions of age categories with gender.

The optimized geodemographic-variables-only model was then used to simulate relative noncompliance rates by census tract in order to deliver heat maps to display results for specific metropolitan areas, scoring census tracts with predicted rates of noncompliance.

Model performance for (a) the optimized model based on individual demographics and behaviors was compared to (b) the optimized geodemographic-variables-only model. Future modeling extensions for improved geography-based performance were identified.

Copyright © 2016 Decision Analyst. All rights reserved.

Strategic Issues

The ability to geographically identify populations based on health conditions and associated behaviors can aid in the targeting of healthcare delivery and outreach services. This research demonstrates that high-risk behavior (such as noncompliance with prescribed medication) can be identified at the census-tract level by applying analytical modeling to free, publicly available data.

Health systems, ACOs (Accountable Care Organizations), public health managers, payers, patient advocacy groups, and other entities focused on population-level health improvement could benefit from using this type of approach. Results could be used to select locations for new facilities or to target educational or promotional health campaigns or direct mail. All these activities can benefit from geographic identification of high-risk populations, whether behaviorally specific, disease-state specific, or environmentally at risk.

While this case demonstrates that modeling using public data is possible, perhaps even better models can be developed by augmenting the public data with primary research. Surveys, ethnographic studies, profiling of lifestyle characteristics, and other primary research methods can potentially enhance the acuity of the models and the "reachability" of the target populations. In this way, additional primary research can leverage health outcomes.

Research Objectives

The objective of this research was to explore the value and feasibility of using survey-based predictive models of noncompliant patient behaviors as a tool to better geographically target both education and delivery of compliance-enhancing and, hopefully in the future, adherence-enhancing treatment methods for chronic illnesses.

To satisfy this objective, we used best-practice, cross-validation techniques and ensemble modeling (drawn from the field of predictive modeling) to optimize the performance of the predictive models, given the available data—in this case, the NHANES survey administered by the NCHS. Then these optimized models were used to identify geographic targets within metropolitan areas with an expected high incidence of noncompliant patient behavior.

In particular, we set out to:

- Gather publicly available data for noncompliance and potential predictors of noncompliance, such as patient demographics, possession of health insurance, comorbidities, and geographic data.
- Optimize the performance of alternative predictive models of noncompliance, using models of varying complexity (logistic regression and Random Forest™).
- Measure the relative performance of an optimized model using all publicly available data vs. an optimized model using geographic variables only.

Copyright © 2016 Decision Analyst. All rights reserved.

- Apply the model to simulate alternative degrees of noncompliant behavior by geographic areas within a large metropolitan area.

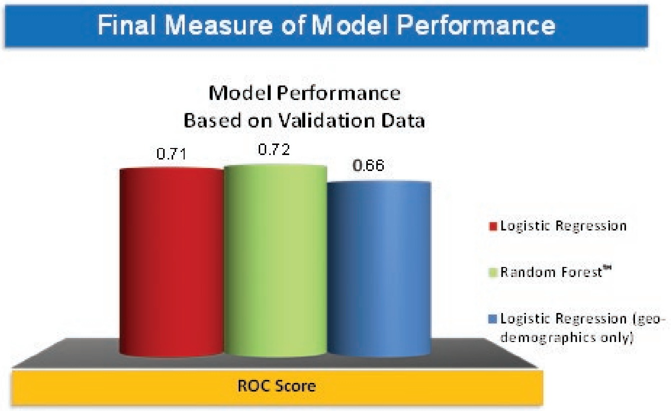
Research Design and Methods

Based on NHANES survey data, the sampling frame was selected to be a representative sample of those individuals who had been told on two or more different visits to a health professional that they had hypertension and/or had been told to take a prescribed medicine for the condition. The dependent variable was binary, based on whether the individual self-reported to be currently taking a medication for hypertension.

The final modeling data set included 1,937 respondents and 55 potential predictor variables. Variables that were found to be either linear combinations of, or highly correlated with, other variables, or which had variability too low to make them useful as predictors of store sales, were eliminated. The final reduced data contained 35 potential predictor variables. In a parallel exercise, an alternative modeling data set was produced for the same 1,937 respondents using only geodemographic variables available from census-tract shapefiles.

These two final modeling data sets were randomly split into training data and validation data. The Validation Data was not used for model development, but was set aside to validate results of the models that were created.

Using the Training Data, cross-validation techniques were applied to determine how many and which predictor variables optimized the logistic regression and Random Forest™ models. The optimized models were then tested for predictive accuracy using the validation data. These procedures were conducted in parallel using the full data set and the geodemographic-variables-only data set.



Finally, the geodemographic-variables-only model was used to simulate noncompliance rates by census tract and was dynamically linked to Google Maps to demonstrate how such a model could be used to produce heat maps, where the census tract was shaded based on the predicted percentage of noncompliance. In addition, a dynamic simulation engine was developed to display the relative average noncompliance percentage for surrounding census tracts around a specified address. The delivery platform was a browser page that included a Google Map and a dashboard of simulation results.

Copyright © 2016 Decision Analyst. All rights reserved.

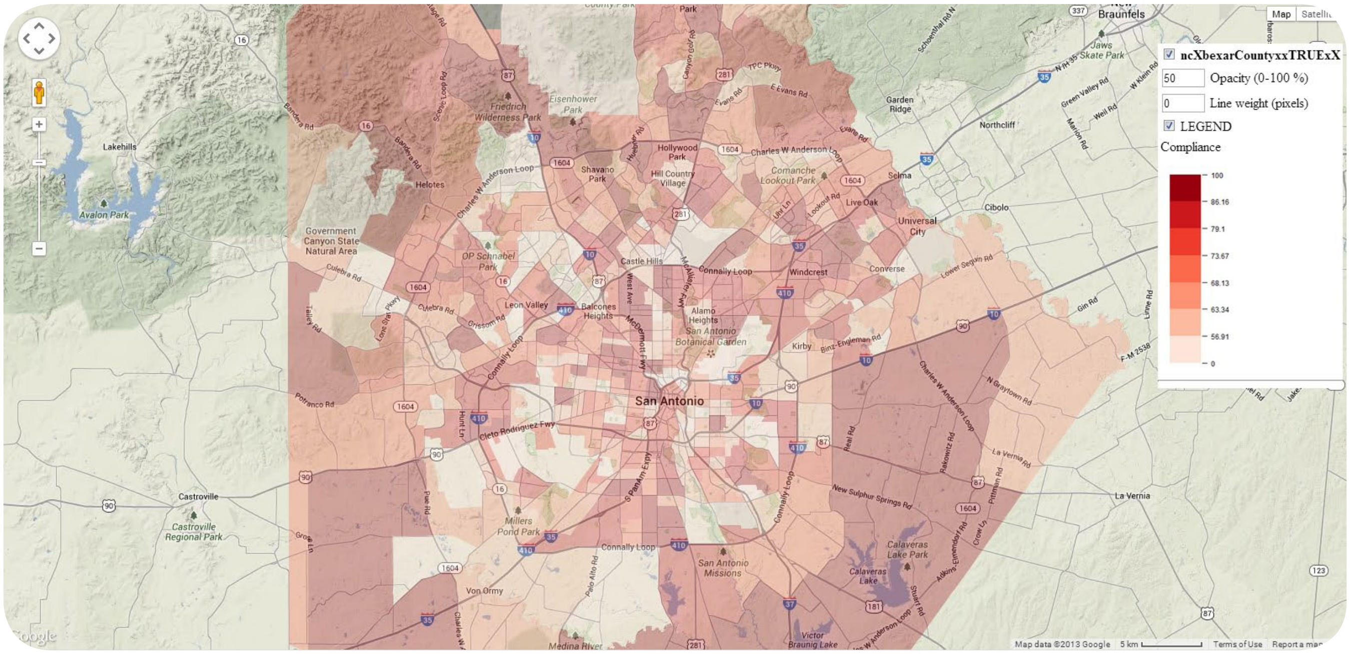
Results

The optimal Random Forest™ model with 14 variables had the greatest predictive accuracy within the validation data (ROC Score of 0.72). The optimal logistic regression with 5 predictor variables performed almost as well with a ROC Score of 0.71. The ROC Score measures model performance, where a 0.5 results if the model is no better than chance and a 1.0 results if the model predicts perfectly. The optimal geodemographic-variables-only model (a logistic regression with 10 age-category predictors plus selected interactions of age categories with gender) attained a ROC Score of 0.66, indicating an 8% decline in model performance, suggesting that demographic variables available at the census-tract level (or at the zip-code level in future studies) can be effective predictors of patient noncompliance.

The geodemographic-variables-only model was used to simulate predicted percentage of noncompliance by census tract and mapped via Google Maps (as in this map example for Bexar County, Texas).

Plans are being made to extend the methodology explored in this case study. Possible next steps include:

- Incorporating insurance claims data to obtain a more reliable measure of patient noncompliance.
- Incorporating additional geodemographic variables available from the U.S. Census Bureau to potentially improve the predictive accuracy of models based on geodemographic variables.
- Exploring model predictive accuracy for smaller geographic areas; e.g., approximately 220,000 block groups vs. the 74,000 census tracts used in this analysis.



☑ ncXbexarCountyxxTRUExX
50 Opacity (0-100 %)
0 Line weight (pixels)
☑ LEGEND

Decision Analyst
strategic research ■ analytics ■ modeling ■ optimization

604 Avenue H East • Arlington, TX 76011-3100, USA
1.817.640.6166 or 1.800.ANALYSIS • www.decisionanalyst.com

Copyright © 2016 Decision Analyst. All rights reserved.