https://doi.org/10.48009/2_iis_2020_229-237

Issues in Information Systems
Volume 21, Issue 2, pp. 229-237, 2020

BUILDING AN ANALYTICAL MODEL TO PREDICT WORKFORCE OUTCOMES IN MEDICAL EDUCATION

Logan Butler, University of Minnesota, lbutler@umn.edu
Manjeet Rege, University of St. Thomas, rege@stthomas.edu

ABSTRACT

Across the United States, there is a shortage of physicians providing care in rural areas. This shortage means patients living in rural communities must travel further with fewer care options. The purpose of this study is to ultimately fill the gap in rural workforce outcomes by identifying students that are likely to practice in rural areas once they complete medical school and residency/fellowship programs. These students may be identified through use of predictive analytics techniques. By identifying these students, we can provide informational material and optional programs to further foster interest in rural care.

Through techniques such as feature extraction, resampling, and data imputation, we prepare data for various machine learning classifiers. These models allow us to identify features common to urban providers and rural providers. Seventy percent of rural providers were correctly identified as practicing in rural areas, while 25% of their urban counterparts were classified as rural. One characteristic difference between the groups shows rural providers have high average scores through medical school courses, while urban providers have higher standardized test scores.

Keywords: Medical Education, Rural Workforce Outcomes, Predictive Analytics, Machine Learning

INTRODUCTION

The United States has a shortage of medical care providers in rural areas. This is not a new problem, it has been written about in reports both from a century ago (Pusey, 1925), and more recently (Rabinowitz, Diamond, Markham, & Santana, 2011). Twenty percent of the US population lives in rural areas of the country, while reports predict between 9-12% of providers are practicing in these same rural areas (Hancock, Steinbach, Nesbitt, Adler, & Auerswald, 2009; MacQueen et al, 2018). A study of travel times between patient and Medicare healthcare provider shows rural patients travel 2-3 times further to see medical providers than those living in urban areas (Chan, Hart, & Goodman, 2006). Urban areas of the country are medically underserved, and generally harder to access what care is available. We hope to understand this gap in coverage by identifying features of a population of providers practicing in rural areas and using those features to predict future rural providers.

Related Works

Past efforts to increase rural provider output include medical school admissions policies which favor applicants who come from a rural background (Hancock, Steinbach, Nesbitt, Adler, & Auerswald, 2009; Rabinowitz, 1983), which has previously been identified as a good influencer for rural practice both within the United States (Parlier, Galvin, Thach, Kruidenier, & Fagan, 2018) and outside (Magnus & Tollan, 1993). Additional work has determined the influence some programs have on positive rural workforce outcomes (Halaas, Zink, Finstad, Bolin, & Center, 2008). Through polling matriculating students, another study shows interest in practicing family medicine is strongly related to practice in rural areas (Rabinowitz, Diamond, Markham, & Santana, 2012).

These described studies and programs show correlation between rural upbringing and rural practice, but no previous work has been done to identify those individuals through predictive analytics on the full educational process of each medical student. We are building a model to aggregate all data through the spectrum of a learners' education, from undergraduate studies, application, and medical school entrance data, through to the beginning of residency. We describe new trends we see and discuss building upon these studies in our *Descriptive analysis* section below.

Objectives

The objective of these analyses was to build a predictive model using machine learning methods and descriptive analyses to identify future rural providers from current medical school students. This model will help us identify trends in past graduates and apply those trends to current medical school students. We will describe rural and urban providers to help identify features of each group. As our data grows, we hope to apply these analyses to identify future rural providers while they are in medical school. By identifying these students, we can provide them opportunities to participate in optional rural clerkship programs, and provide other material that may further foster interest in rural practice.

METHODS

Data was collected from the undergraduate education, medical school application, medical school course work and clerkships, and residency stages in a learners' career. These collected data points come from internal and external systems which have been integrated thanks to the Medical Education Outcomes Center (MEOC) to make medical education data available to University of Minnesota Medical School (UMMS) faculty and staff requesters for research and operational purposes.

Through the education and career of a provider, the first data we collect is during the medical school application process. UMMS provides an online application portal for applicants to provide demographic data, such as their city and state of residency, race & ethnicity, and sex, among many other data points. We also collect undergraduate information from each applicant, such as the institution they attended, year of graduation, field of study, degree(s) earned, and their GPA.

These application data points can be enhanced with demographic data available through IPUMS Geomarker (Oakes, Riper, Kugler, & Goldberg, 2019). We combine the application data with IPUMS data to learn more about the demographics of where each student grew up. These demographic data points include population density, percent of residents under the poverty line, the percent of homes owned by the residents living in them, and median income, among others.

We begin collecting medical education data once an applicant is accepted into UMMS. During the first two years of medical school, a student is enrolled in courses in a classroom, from which we collect course performance data. Additionally, we collect data from students when they take the standardized United States Medical Licensing Examination (USMLE) exams Step 1, and Step 2. These are exams each medical school student takes across the country, usually after the second and fourth years of medical school, respectively.

After two years of course work, students enroll in two years of various clerkships, some optional, some required. Some students participate in optional longitudinal integrated clerkships (LIC), which are generally programs that allow cohorts of students to focus their learning on specific topics. Through these clerkships, we collect data on elective clerkship choices, and performance.

The described application and medical school data are supplemented by data from the American Medical Association (AMA) Physician Masterfile (2018). The AMA Masterfile data provides practice specialty and practice location data for graduated medical school students. We use these data to ultimately determine if a graduate is practicing in a rural or urban setting. Because our goal is to predict this data point, we only use the AMA Masterfile data as the target variable in our analyses.

We limit our dataset to medical school matriculating classes from 1999 to present, because our primary internal systems first began collecting data in 1999. Additional data systems collect data prior, but these data do not contain many of the data elements of interest. The resulting dataset contains 2000 providers (rural = 309, urban = 1691).

Defining Rural

The United States Census Bureau defines rural by what is not considered to be in an urban area (Ratcliffe, Burd, Holder, & Fields, 2016). To define rurality in our analyses, we utilized the practice location provided by the AMA Masterfile dataset to identify in which zip code each provider is practicing. These zip codes were used to identify rural versus urban practice location through the US Census Urban Area (UA) geographic code to zip code approximation. These codes provide a more granular level of detail to the rurality of each region over other available US Census codes. Other geographic codes, such as the Rural-Urban Continuum Codes (RUCC) or Rural-Urban Commuting Codes (RUCA) classify rurality at the county level and census tract level. These regions are much larger than zip codes, so we choose to use the more specific UA definition.

Imputing Missing Values

The data systems we combine in order to form a profile for each graduate have varying levels of data available and have varying ranges of years available. Because of this, some data points are sparse. In order to perform analyses on these data, we must impute missing values or remove them. One way of imputing is by calculating the mean or median value for the available data in the column and using that value for any missing values. In our project this is beneficial for data such as age at the end of residency. This field is centered around a specific mean with minimal distribution (mean = 32.8 years, standard deviation = 3.27 years), but more importantly, these numbers remain evenly distributed among other factors such as graduating year, so we know the average age of our graduating class does not change significantly over time.

Another way we can generate data for missing values is by finding another variable without null data that is highly correlated. We calculate linear regression and impute using slope and intercept from this calculation. To do this, we look at a subset of features to impute. In *Table 1* below, we gather a group of features which all include a milestone year. The features include the year of med school graduation, birth year, matriculation year, and undergrad finish year. Three of these do not contain a null value while the undergrad graduation year does. Because these features also broadly represent when an event occurred and at what age, we will look at the correlation between *undergrad finish year* and the other three features to see how we can best replace missing values.

Table 1. Null counts for selected features

Feature	Null values
med school grad year	0
birth year	0
med school matriculation year	0
undergrad finish year	875

We compute the correlation for this subset of data to see if any relationship exists which can help us impute the missing values for *undergrad finish year*. In *Table 2* we see the correlation between *undergrad finish year*, the year medical school students complete their undergraduate education, and other selected features. We see the highest correlation between *undergrad finish year* and *birth year*. This indicates there is a relationship between the year a student graduates university and the year they were born.

Table 2. Undergrad finish year correlation

Feature	Correlation
med school grad year	0.688
med school matriculation year	0.716
birth year	0.888

Our median university completion year is 2008, but using this median value as we do above with other features will cause issues, such as when a student begins medical school in 2006, 2 years before they finish their undergraduate education. In this case, we compute the university completion year using birth year. These features have a correlation of 0.89 and an r-squared of 0.82. A correlation of 0.89 means these two features are highly correlated, and thus using regression to impute these missing values is more beneficial compared to using the mean or median value.

Encoding Categorical Features

Many of our data points are categorical features. Some of our machine learning models will not work with unaltered categorical data, so we must encode these categorical variables into numeric variables. One method of encoding these variables is known as One-Hot Encoding. This process splits a single column of categorical data into a single column for each category.

An example from our dataset is the residency specialty chosen by each provider. This data is originally stored in one column with values such as “Pediatrics”, “Surgery”, and others. When we One-Hot Encode this column, we end up with a single column designated for “Pediatrics”, another for “Surgery”, and one for each other specialty available in our dataset. From this encoding, we now have a binary representation of the specialty for each provider.

Balancing Mismatched Classes

Our dataset contains 309 providers in rural areas, and 1691 providers in urban areas. Because these numbers imbalanced, both oversampling and under sampling techniques were used in addition to analyzing the unaltered data. Synthetic Minority Oversampling Technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) was used to oversample the rural data. This method of over-sampling creates “synthetic” data points which follow the same pattern as the existing data points. Because we are only trying to learn the association between features, these new data points boost the weight of the rural provider class so our predictor does not choose urban for all providers.

An under-sampling technique was also used to under sample the urban class. This reduced the number of urban data points down to match the number of rural data points. Both resampling techniques and the original unaltered dataset were used while predicting. Generally, the SMOTE oversampled data performed better with the under-sampled dataset performing almost as well with a few models. Our data is split for training and testing, so below numbers reflect the 70% reserved for training. To ensure we are testing on unchanged data, the remaining 30% reserved for testing is not resampled.

Table 3. Training Data after Re-sampling

	Urban	Rural	Training Samples
Original training data	1195	205	1400
SMOTE over-sampling	1195	1195	2390
Under-sampling	205	205	410

Feature Elimination

After encoding categorical variables and imputing values for missing data, we eliminate features which are not significant to producing a better prediction model. Through the process of encoding data, we lose some human-recognizable meaning to each feature, but we build better indicators to predict on. As we grow the number of features through encoding processes, we observe some features which do not provide significance in predicting our ultimate outcome, and only provide additional model complexity. Principal Component Analysis (PCA) is used to find the number of features which explain 99% of the variability. After encoding variables, our dataset has 857 columns of categorical and numeric data. Once we run PCA on these data, we find 99% of the variability is contained within a subset of 645 features. These 645 features were selected for these analyses.

Descriptive Analysis

In the process of building a predictive model, we find an interesting, divide in these two groups. On average, rural providers perform better in the classroom, while urban providers tend to perform better on standard tests, such as the United States Medical Licensing Examination (USMLE) Step 1. Rural providers have higher average class scores (mean = 0.9335, standard deviation = 0.0217) than their Urban provider counterparts (mean = 0.9155, standard deviation = 0.0327).

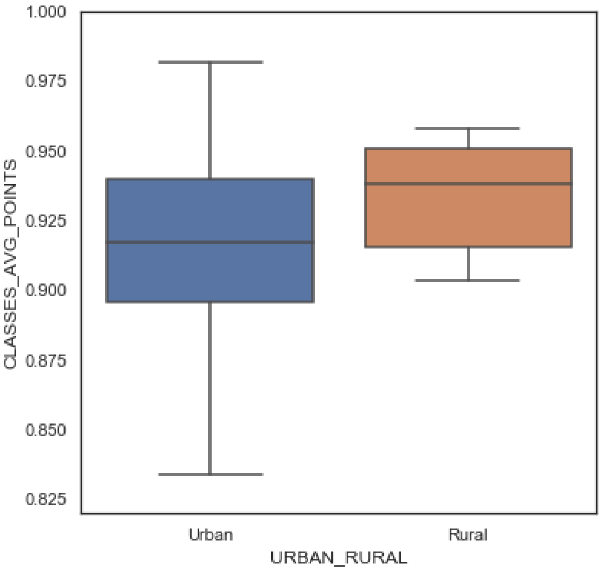

Figure 1. Average Course Scores in Years 1 and 2

Inversely, Urban providers perform better on standard tests, such as the USMLE Step 1 exam. They perform slightly better (mean = 224.99, standard deviation = 18.7) than their rural provider counterparts (mean = 221.96, standard deviation = 17.94. These described differences are small enough and limited to only recent data available in our systems, thus they are not statistically significant. These differences show a trend that allow us to improve prediction. We hope to study these indicators more in future years as more data becomes available.

Additionally, we notice a similar trend in the rural upbringing of our providers. Providers who end up in rural practice more frequently have a rural upbringing (47.8%) than those who practice in urban areas (26.5%). We also observe the group of rural providers tends to be whiter (87.4%) compared to urban providers (78.5%). However, these groups are similar in both sex and average age when they finish residency.

Model Selection

From the scikit-learn (Pedregosa et al, 2011) python project, 7 classifiers (Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, Gradient Boost, K-Nearest Neighbors, and Support Vector Machine) are tested. XGBoost (Chen & Guestrin, 2016) is also tested from the XGBoost python project. Each classifier was tested with the original data, over-sampled data, and the under-sampled data.

Defining Performance of Predictive Models

In order to choose which model performs the best, we must decide how to evaluate the success of each model. Because our data is imbalanced, we cannot use an evaluation metric like accuracy. If we use accuracy, we benefit from predicting all students will practice in an urban setting. The dataset is imbalanced, so we must consider other options. We will favor the ability to correctly classify rural providers over the ability to identify urban providers. Additionally, because there is a low barrier to providing more information and resources to those we believe may be influenced to practice in rural areas, we will not penalize urban providers classified as rural providers as much. Performance of machine learning models are generally evaluated using a confusion matrix, as shown in *Table 4*.

Table 4. Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Because our objective is to identify as many of the rural providers correctly, we chose to use recall as our method of evaluation. We can calculate the recall of each provider type by dividing the number of providers correctly identified (True Positives) by the total number of providers (True Positive + False Negative). For all models, we choose to prioritize the recall for classifying rural providers, with a secondary goal of the maximizing recall of the urban class.

$$Recall\ = \frac{TP}{TP + FN}$$

RESULTS

Our best performing model is a Support Vector Machine (SVM) model using a radial basis function (RBF) kernel using the SMOTE oversampled data. This model correctly identified 73 of 104 (70.19%) total rural providers in the testing set (a 30% sample of the total dataset). While identifying 73 rural providers, we also misclassify 133 of 496 (26.81%) of urban providers.

Other models which had similar success include Gradient Boost, and Logistic Regression, both of which use the SMOTE oversampled data. Gradient Boost identifies 66 of 104 (63.46%) rural providers correctly, and misclassified 189 of 496 (38.10%) urban providers. Logistic Regression classifies 52 of 104 (50%) rural providers, and misclassified 96 of 496 (19.35%) urban providers.

For our SVM model, our rural class has a recall of 0.70 (73 / 106), and the urban class has a recall of 0.73 (362 / 496). We show these two metrics in *Table 5* with an average recall for both classes calculated in each model.

Table 5. Model Performance

Model	Recall (Rural)	Recall (Urban)	Average Recall
SVM w/ SMOTE	0.70	0.73	0.715
Gradient Boost w/ SMOTE	0.63	0.62	0.625
Logistic Regression w/ SMOTE	0.50	0.81	0.655

DISCUSSIONS

Challenges and Limitations

One major challenge for these analyses is limited data. The internal student databases only have data going back to the mid-2000's, which leaves only a decade of data once the years spent in residency and fellowship programs are accounted for. Additionally, many of the data points collected were frequently null as data systems used to collect these data were implemented at different times over the past two decades.

The decision to choose practicing medicine in one location over another has many factors at play that are not available in a database or easily modelled. One data point which we are unable to access for these analyses is the offer of student loan repayment programs made available to providers who practice in rural areas. Many rural hospitals and clinics offer loan repayment plans for providers to move to the area and practice for a number of years. We hope to study the duration providers practice in rural locations in future work.

One part of the decision to practice medicine rurally that other studies discussed is the intention to practice family medicine, a common specialty among rural providers. We do not currently collect these data points, but implementing surveys with these questions may help us better identify the students we hope to.

Next Steps

Next steps include re-training these data each year as additional information on graduates becomes available. This study was conducted on approximately ten years of data from graduated students. As cohorts of students graduate and begin practicing each year, we will add these new providers to the analyses to improve our predictions.

We notice demographic differences and similarities between the two groups and will study these further to observe the changes in these trends over time. Additionally, we find a difference in the course performance versus standardized test performance that indicates some underlying difference between these groups could be studied further.

CONCLUSION

These results show promising steps towards using predictive analytics for filling gaps in workforce outcomes both by identifying common demographic differences between the groups, but also through our predictive model. We have been able to correctly classify a majority of the rural providers, while misclassifying significantly less than half of the urban providers with a small, and sparse, dataset. There is room for improvement in the model, and we expect changes each year with an increased availability of data to build our model.

By identifying potential students who may eventually practice in a rural location, we can provide information and optional clerkship programs that foster interest in rural practice. We will also use these insights to study those that practice in rural areas that do not fit the predictions to hopefully gain insight into other aspects of the decision to practice in rural areas that we have not yet identified.

REFERENCES

AMA Physician Masterfile (2018). [Data set]. American Medical Association.

Chan, L., Hart, L. G., Goodman, D. C. (2006). Geographic Access to Health Care for Rural Medicare Beneficiaries. *Rural Health Association* 22(2), 140-146.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. Vol 1 6. 321-357.

Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.

Halaas, G. W., Zink, T., Finstad, D., Bolin, K., Center, B. (2008). Recruitment and Retention of Rural Physicians: Outcomes from the Rural Physician Associate Program of Minnesota. *The Journal of Rural Health*, 24(4), 345-352.

Hancock, C., Steinbach, A., Nesbitt, T. S., Adler, S. R., Auerswald, C. L. (2009). Why doctors choose small towns: A developmental model of rural physician recruitment and retention. *Social Science & Medicine* 69 1368-1376.

MacQueen, I. T., Maggard-Gibbons, M., Capra, G., Raaen, L., Ulloa, J. G., Shekelle, P. G., Miake-Lye, I., Beroes, J. M., Hempel, S. (2018). Recruiting Rural Healthcare Providers Today: A Systematic Review of Training Program Success and Determinants of Geographic Choices. *Journal of General Internal Medicine* 33(2), 191-199.

Magnus, J. H., Tollan, A. (1993). Rural Doctor Recruitment: Does Medical Education in Rural Districts Recruit Doctors to Rural Areas? *Medical Education*, 27(3), 250-253.

Oakes, J. M., Riper, D. V., Kugler, T. A., Goldberg, D. (2019). *IPUMS GeoMarker* (Version 1.0) [Data set].

Parlier, A. B., Galvin, S. L., Thach, S., Kruidenier, D., Fagan, E. B. (2018). The Road to Rural Primary Care: A Narrative Review of Factors that Help Develop, Recruit, and Retain Rural Primary Care Physicians. *Academic Medicine*, 93(1), 130-140.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825-2830.

Pusey, W. A. (1925). Medical Education and Medical Service. I: The Situation. *JAMA*, 84(4), 281–285.

Rabinowitz, H. K. (1983). A Program to Recruit and Educate Medical Students to Practice Family Medicine in Underserved Areas. *JAMA*, 249(8), 1038-1041.

Rabinowitz, H. K., Diamond, J. J., Markham, F. W., Santana, A. J. (2011). Increasing the Supply of Rural Family Physicians: Recent Outcomes from Jefferson Medical College's Physicians Shortage Area Problem (PSAP). *Academic Medicine* 86(2), 264-269.

Rabinowitz, H. K., Diamond, J. J., Markham, F. W., Santana, A. J. (2012). The relationship between matriculating medical students' planned specialties and eventual rural practice outcomes. *Academic Medicine*, 87(8), 1086-1090.

Ratcliffe, M., Burd, C., Holder, K., Fields, A. (2016). Defining Rural at the U.S. Census Bureau. Retrieved May 27, 2020, from https://www.census.gov/library/publications/2016/acs/acsgeo-1.html.