Using Machine Learning Algorithms to Identify Physician-Hospital Integration Based on Physician Characteristics Data

MSc Scientific and Data Intensive Computing

A dissertation submitted for the degree of

Master of Science

Of

University College London

Student ID: 20068937

Student Name: Chen, Yuanzhen

Supervisor: Lina Dahye Song

Co-supervisor: Qinquan Cui

Word count: 9063

Department of Physics and Astronomy

August, 2021

I, Yuanzhen Chen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the dissertation.

Abstract

In recent years, an increasing number of physicians are vertically consolidated with hospitals, and this phenomenon has aroused the interest of scholars and health care workers. Besides, in order to make sure that the medical resources are rationally allocated, it is necessary for policy makers to track the financial and organizational changes in structures of the health care institutions such as integration, merger and closure. Given that few of the previous studies concentrated on identifying physician-hospital integration, this project aims to identify and predict integration, applying seven machine learning methods. The data sources for this project are the claims data from CMS and the NPPES provider data from 2012 to 2015. The project compares two classes of models, one with claims data and the other without claims, and it turns out that the former performed much better but existed severe multicollinear problems. So this project uses the claim-free models as the main models, and found that Support Vector Machine, Decision Tree, K-Nearest Neighbour and Random Forest algorithms had the highest accuracy and f1 score of over 90%. Predicting the physicians integration status using these models, the integration trend from 2015 to 2018 is visualized in a line chart which demonstrates that there is a fluctuated tendency with slight rises of physician-hospital integration. However, due to the limitations on data, variables, models and parameters, this result is out of the expectation that the integration rate would increase by over 10% annually according to the evidence from other studies. For future study implications, analyzing the influences of each physician characteristic variable on predicting physician integration status or the construction of more reliable models of identification may contribute a lot to this topic.

Key words: vertical integration; physician characteristics; machine learning methods; identifying integration.

Acknowledgement

First and foremost, I would like to express my deep gratitude to my project supervisor Professor Lina Dahye Song. She has provided plenty of resource and detailed interpretation that help me generate enough knowledge on the filed I learn about for the first time. At the weekly meetings, after discussing the work I have done for that week, she would give suggestions on the objectives I was expected to achieve by the next meeting. Her constructive instruction enabled me to have a clear thought on this project so that I have generated confidence to finish such a tough task.

Secondly, I would like to thank my coursemates Fangwei Dong and Yisu Zong who had similar research topic to mine. We usually shared things like learning outcomes and discussed directions of our projects. They are generous and kind and I am glad to see that we are all making progresses in this project.

Last but not the least, I would like to express my sincere thanks to all of the faculties of University College London who persevere their posts and provide supports all the time during this special but tough academic year. Hope that everyone is doing well!

Contents

1 Introduction

1.1 Research Background

The consolidation between health care participants has been gradually developed into a trend. There are at least two drivers. First, consolidation can result in less transaction cost or bring the benefits of economies of scale, so the organizations co-operate to reduce relative cost and increase market share. Second, the health care providers integrate with each other to help patients so that they can receive a coordinated process of services (Raus et al., 2020).

In the health care provider market, there are two types of consolidation of medical institutions: horizontal integration and vertical integration. The former refers to the merger of over two economic entities that do similar business, say hospital mergers, whereas the latter means integration between the entities that are adjacent to each other along a supply chain, that is , one needs to purchase the semi-manufactured products or services that produced by the other. For example, the common participants of a health care supply chain are primary care physicians, specialists, hospitals and rehabs. Moreover, forward integration occurs when a firm takes control of its downstream firms (demanders) and backward integration means that it consolidates with its upstream firms (suppliers) (Gaynor, 2006). This project will concentrate on the vertical integration of hospitals and their suppliers - physicians (backward integration for hospitals), and this can be done in the forms of hospital acquisition of physician practices or direct employment of specialists (Robinson & Casalino, 1996). The phenomenon of integration between physicians and hospitals has been on the rise and it is becoming a noticeable trend in the health care market, which has aroused a number of scholars' interest in research. Yet most of the empirical papers concentrated on the impact of integration, such as spending and quality of care (Cuellar et al., 2006; Baker et al., 2014; Ho et al., 2000). The related literature lacks of research on identifying and predicting physician-hospital consolidation, both individually and macroscopically. Thus the aim of this project is to predict physicians' vertical integration with hospitals based on physician characteristics.

1.2 Research Aim and Objectives

Based on the research aim mentioned above, this project has four specific steps. The first one is data cleaning and merging of three datasets, including the Medicare claims dataset, the provider dataset and the physician characteristics dataset. The second objective is to identify vertical integration using the chosen variables by applying different machine learning algorithms, taking the indicator of integration from the gold standard data for reference. The gold standard data here was created by applying the gold standard rule made by Hannah T. Neprash, and the integration

indicator is obtained using her identification method (Neprash et al., 2015). Apart from this, I also compared the performance of the claim-based models and claim-free models. Thirdly, I selected the claim-free models with the best performance and used these models to predict physician integration in the "future". Owing to the limitation of accessing more recent data, I applied the selected model to "predict" physician integration status for the years from 2015 to 2018. Lastly, I predicted the trend of hospitals that consolidated with their providers from 2015 to 2018 for each chosen algorithm.

1.3 Research Necessity and Contributions

In recent years, the vertical integration between health care institutions has become an significant issue within the health care industry. Researchers, health care suppliers and policy makers are getting concerned about the integration status of physicians and hospitals (Ho et al., 2020). For policy makers, they take charge of fairly allocating the medical resources including hospitals and specialties for each region, and they need to invest once the change in the distribution of resources breaks the balance between medical workers and patients of that region (Raus et al., 2020). Thus, it is necessary for them to know about the integration status or tendency of the institutions.

However, due to the limitation on the existing accessible data, the change in hospital financial and operational structure is difficult to observe. Therefore, some scholars has started working on identifying consolidation using hospital-level data such as inpatient care data and Medicare claims data and found out that these datasets can be a reliable source to track integration status (Neprash et al., 2015; Ho et al., 2020). Nevertheless, the models based on spending of patients or Medicare claims can not be used to identify integration status until the payments or bills occur, hence these models fail to predict collaborations between medical institutions in the future. So this project will provide an idea on the prediction of integration.

There are three main contributions of this project to the analysis of collaboration between health care institutions. First of all, given that there are few studies on the identification of integration, this project summarizes the existing algorithms proposed by other scholars and puts forward several identification models based on machine learning methods. Secondly, this project gives comparisons on the machine learning algorithms from several aspects, regarding their properties and performances. Thirdly, to cope with the difficulty in obtaining claims data, this project attempts to use physician-level data that are more accessible and cost-effective, providing a convenient way to track the integration between physicians and hospitals in terms of physician characteristics, thereby makes it possible to predict future integration tendency of physicians using machine learning methods.

The codes for the implementation of the machine learning algorithms are in the Github repository: https://github.com/Yuanzhen-Chen/PHAS0077_Project_2021.git

1.4 Project Structure

This project is comprised of five chapters. Following the chapter of introduction, Chapter 2 will give a review on related literature that studying vertical integration between hospitals and physicians, classifying into impact analysis and identification of vertical integration. Chapter 3 and Chapter 4 are the highlights of the project: Chapter 3 illustrates the whole process of machine learning application including data processing, implementation of algorithms and model validation results, whereas Chapter 4 concentrates on the prediction of physician integration trend and the comparisons of prediction results. Chapter 5, as a conclusion of the project results, summarizes the limitations and provides an idea for future study.

2 Literature Review

2.1 Impact of Vertical Integration

In the last few decades, the financial collaboration between hospitals and physicians has shown an upward tendency in the United States (Post, 2017; Cutler, 2020), particularly in recent years the percentage of physicians integrated with hospitals increased by over 10% from 2010 to 2016 (Scheffler et al., 2018). This phenomenon has influenced hospital management and operation. Previous studies have focused on certain aspects of impact of vertical integration such as spending, prices, quality of care and efficiency.

There are a number of scholars who are interested in the changes of health spending and prices after hospital taking the ownership of physician practices. Nevertheless, whether the price rises or is not significantly influenced sparks controversy. For instance, Scheffler et al. (2018) found an upward trend in specialty and primary care prices and Baker et al. (2014) argue that physician-hospital integration would lead to higher health spending , but Cuellar et al. (2006) convince that the effect is not significant because the result varies due to different data source and difference measuring approaches. Similarly, some scholars deem that integration would increase the quality of care (Baker et al., 2014), while the others find no impact in it (Ho et al., 2000).

As for efficiency, conceptually, consolidation between hospitals and physicians will reduce transaction cost between these two parties and help remove unnecessary processes like duplication of efforts, and therefore effectiveness is improved (Robinson et. al., 2014). Based on this link to hospitals, physicians may tend to refer their patients to the hospital they integrated with, reducing the agency problem to some extent. However, some papers point out that some unscrupulous hospitals may stealthily pay physicians for referrals, which exacerbates the agency problem (Baker, 2016) .

Judging from the current analysis, whether the impacts of vertical integration between hospitals and physicians are positive or not are worth further exploration, and these are partly dependent on how the entities integrate and to what extent, so it is of tremendous significance to learn vertical integration deeply. However, fewer studies focus on finding the best approaches to identify physician-hospital integration or predicting which providers are going to consolidate with

hospitals. One of the reasons is that the researchers were interested in analyzing the consolidations that have been occurred and the approaches of identification of integration proposed by them were accurate enough for the analysis.

2.2 Previous Identification Approaches

This section is a summary of the approaches of identifying integration by other researchers. Although few studies are purely concerned about the identification methods, some researchers provided the idea on how they measure the integration of physicians and hospitals in order to analyse the influence.

According to Neprash et al. (2015), if a physician's practice is owned by a hospital, Medicare will pay less on the professional fee but pay much more on facility for the same Medicare service, leading to a higher proportion on hospital outpatient claims. Using the Medicare data, they computed the proportion of outpatient claims and compared this percentage with any of the cutoffs including 25%, 75% and 95%, then the physician will be flagged as integrated if the percent of outpatient claims is bigger than the cutoff and flagged as independent otherwise. This serves as the gold standard rule and the indicator of integration for physician is obtained by applying the rule, which acts as the gold standard data for this project.

Craig et al. (2020) proposed a method of identification using data of higher dimension by applying machine learning algorithms. Using the Medicare data on Provider Practice and Specialty (MD-PPAS) data, they determined the integration status for each physician and assigned the ones who are labeled integrated to the hospitals they work for in terms of the taxpayer identification numbers (TIN). Choosing the variables including NPI, practice address, specialty and so on, they ran the tuned Random Forest model on the data and applied cross-validation to improve the model ability of classification.

As for Ho et al. (2020), based on the MD-PPAS claims data, they conducted internet search of TIN legal names for all TINs of a physician, equivalently all TINs with the same NPI in the dataset, and determined for each TIN whether the practice is owned by a hospital or physician. As long as there exists one TIN legal name that is like a hospital's name, the physician will be categorized as vertically integrated. For reference, they used the Blue Cross Blue Shield Texas (BCBSTX) claims data recording contracts of physicians and hospitals as gold standard data, and classified physicians as integrated if they had a claim record submitted for hospital services according to the contract.

3 Methodology：Identify Vertical Integration

3.1 Data and Variables

3.1.1 Data Sources

This project focuses on only one specialty, Gastroenterology (GI). This is not merely because GI takes the largest proportion in the outpatient setting, but the percentage of consolidation between GI and hospitals is increasing (Song et al., 2020). In this project, the source of original data used are Medicare Provider Utilization and Payment data files and NPPES provider data files , which are from the Centers for Medicare and Medicaid Services (CMS). The CMS Medicare data ranging from the years 2012 to 2018 contains provider characteristics such as providers' names, provider practices' address in details and Medicare claims data. Each NPI record that consists of the variables mentioned above is stored in this dataset as a single row, and there are several records for every NPI number. Originating from the Medicare data, the physician integration data for GI is in line with the gold standard rule put forward by Neprash (Neprash et al., 2015), and it contains indicator of each physician integration status with hospitals during the years between 2005 and 2015, with the integration indicator deducing from Medicare claims data in accordance with Neprash's method (Neprash et al., 2015). As for the NPPES provider data, it includes complete and thorough information on provider details for the physicians censused from May 2005 to February 2010. Apart from this, I also use rural-urban-codes file 2013 version that describe counties in the United States as metro or non-metro in terms of FIPS code and a file to convert provider ZIP code to FIPS code for reference.

3.1.2 Variables Description

The aim of this project is to predict integration status using physician characteristic data. But due to the reason that the integration indicator in the gold standard data is based on claims data, so it is necessary to figure out to what extent will the claims data make influence on these models. In order to compare the performance, I trained two classes of models that identifying physician-hospital integration, one class includes Medicare claims data, while the other does not. The other variables are the same for both classes of models, hence the claim-based models serve as the control group. The followings are descriptions of a list of variables I used:

Independent variables:

- **Claims:** Claims data are from the gold standard dataset, which have been employed when applying to the method determining integration status. There are 5 columns of claims data in total, and I included the ones other than total claims in the models to avoid problem of collinearity despite that only outpatient claims and total claims are used for the indicator of integration. The others are claims of inpatient, claims of office and claims of Ambulatory

Surgical Centers (ASC).

- **Gender:** The gender code of physicians is from Medicare data, using initial capital letters of male and female to indicate gender. I convert it into a dummy variable to make it applicable, with binary values 0 and 1 representing female and male respectively.
- **States:** The state of physicians is also from Medicare data, and similarly, dummy variables are created to quantify it. Yet what is different from gender is that the number of dummy variables is the same as that of categories of states, 53 in total.
- **Years of experience:** The years of experience of physicians are counted from the year of provider enumeration date that are recorded in the NPPES provider data, then the number of experience years of each physician is simply the difference between the physician enumeration year and the occurrence year of Medicare claims issued by this physician, with the latter obtaining from Medicare utilization and payment files.
- **Rural/urban indicator:** This is a location variable that indicate the development of the county where the physician perform medical operations. I used Rural Urban Code (RUCC) to classify the counties into 9 levels according to the population, which decreases gradually from level 1 to level 9. The first 3 levels represent metro areas, and the cutoffs of population are 1 million and 250 thousands. On the other hand the rest are non-metro areas classifying into 3 groups with 2 levels in each group, truncating by 20 thousand and 2,500. The counties of level 4, 6 and 8 are adjacent to metro areas, while the others are not. Nevertheless, there is no such variable in the Medicare data and the provider location ZIP code need to be converted into FIPS code that refers to each county, thereby the RUCC code comes correspondingly.

Outcome variable:

- **Indicator of integration:** For each physician, the indicator of integration status with hospitals is obtained using Neprash method (Neprash et al., 2015). and the value 1 means that the physician is vertically integrated with hospital, while 0 has the opposite meaning.

So as to demonstrate the physician characteristics intuitively, the summary statistics for physician characteristics are shown in the following tables.

Table 1 shows the numbers of physicians in each administrative area for each year between 2012 and 2015. Obviously, the top three states with the most physicians in the samples are Texas, New York and North Carolina, regardless of the year time. Table 2 illustrates other characteristics of physicians, including the percentage of males, the percentage of physician that situated in metro areas, the percentage of physician that are vertically integrated with hospitals and the average years of experience of these physicians in each year. Interestingly, male physicians account for a large proportion and the majority of physicians work in a non-metro area. What worth noting is that there is a positive relationship between Integration and Metro and the percentage of physicians that integrated with hospitals presents a downward trend, declining by 1.7% in three years and dropping to 23.3% in 2015. However, the sample size is terribly limited except for the year 2014, and further analysis is needed to figure out the relationships of these variables between each others.

Table 1 Physician characteristic: Administrative location

State	2012	2013	2014	2015	State	2012	2013	2014	2015

AZ	2	2	36	2	NJ	0	0	9	1
CA	1	2	12	2	NY	3	4	52	3
FL	3	3	51	0	OK	1	1	16	1
GA	1	1	0	1	OR	1	1	8	0
IA	1	1	15	1	PA	2	2	32	2
IL	1	0	12	0	PR	1	0	6	0
IN	0	0	1	1	SC	1	1	11	1
MA	1	1	9	1	TN	1	1	11	1
MD	2	2	10	1	TX	5	5	88	6
MI	1	0	0	1	VA	2	3	51	2
NC	2	2	54	3	Other states	0	0	0	0

Note that for each year from 2012 to 2015, no physician is located in the other 32 states that are not shown in the table.

Table 2 Physician characteristics

Characteristics	**2012**	**2013**	**2014**	**2015**
Male(%)	90.63	81.25	82.85	86.67
Metro(%)	46.88	46.88	41.53	40.00
Integration(%)	25.00	25.00	23.35	23.33
Average experience years	5.94	6.94	7.97	9.00
Sample size	32	32	484	30

3.1.3 Data Processing

In this section, I am going to illustrate the data cleaning and merging in details, and the programming language I used is Python 3 mainly.

For the Medicare Utilization and Payment files, I used a short version file for each year in which every NPI number only appears once, that is to say, every physician remained only one Medicare operation record in order to avoid the problem of multimatching. First and foremost, the columns that need to be used are stored in an object, and so did the physician integration of GI file. Then I filtered out the records in terms of the year corresponding to that of the physician integration file and merged them according to the NPI number so that the merged data acted as the sample data.

Secondly, the enumeration date from the NPPES provider data was stored in the kernel and its data type was changed to date data. Having been separated out, the enumeration year for each physician was subtracted from the year set in the previous step to compute the years of experience of physicians. Furthermore, I merged the data file that converts provider ZIP code into FIPS code and the other file that categorizes FIPS, equivalently counties, as RUCC code, hence a converter from ZIP code to RUCC code was obtained. Then I merged the variables years of experience and RUCC code into the sample data in terms of NPI number and ZIP code respectively.

Thirdly, I selected the gender variable and state variable in succession and create dummy

variables for them before concatenating these binary values with the sample data and removing the original gender and state columns in the meantime. At this point, there would be some missing values in the sample data, so I dropped the rows if there existed any missing value in that row.

Finally, I divided the sample data into outcome predictor (y) and independent variables (X), namely the indicator of integration and all of the rest. Then I split the data into 75% training set and 25% testing set randomly and ready for applying to the machine learning algorithms. Finally, I checked the capability of categorization of the models with the full sample data, because by checking the results of classification of the training set I could deduce the goodness of fit of the model, whereas by checking the results of testing set I would form a general cognition about how well will the model performs on the out-of-sample data.

3.2 Application of Machine Learning Methods

3.2.1 General Implementation of Algorithms

Deciding whether a physician or its practice is owned by a hospital or not is a binary classification problem, which can be solved using regression or classification algorithms. The application for the latter is undoubtedly applicable, so considering the regression methods, it suffices to set a cutoff to distinguish between the two classes.

In this project, seven machine learning methods (including Linear Regression, Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbour (KNN), Naive Bayes together with an ensemble method Random Forest) were applied to discriminate and classify the GI physicians as integrated or independent. The implementation of these methods were mainly done by the algorithms featuring in Scikit-learn library in Python, adopting the data from 2012 to 2015. Each of the data set was split into 75% training set and 25% testing set. In each year, two of the models, including one claim-based model and one without claims data, were fitted and tuned separately with the training data of that year. Here, the model tuning tool of grid search that looked for the optimal hyper-parameters was applied so that the possibility of model under-fitting or over-fitting can be reduced and that the efficiency of the model will be improved.

All of the data will then be reused in order to validate the model, in particular the testing data which is the out of the sample, and the predicted indicator of integration will be obtained by employing the corresponding independent variables to the models of the same year as that of the data. Compared with the integration status indicators obtained from the gold standard data, says the authentic values of the indicator, the accuracy score or f1 score of the comparison results acted as a reference to determine the performance for each algorithm.

3.2.2 Implementation Details

When implementing Linear Regression, I also called the Ordinary Least Squares(OLS) module from statsmodels library instead of Scikit-learn in Python so as to get the regression result of each model. After constructing the linear regression model, I concatenated the independent variables with

the outcome predictor and fitted the model with training data set. Indicating from the values of adjusted R squared for each of all years, claim-based models with R squared of over 0.9 performed much better than those without claims data, whose values of R squared were only slightly more than 0.5. The next step is to predict the outcome variable using the above models. Since the predicted values are continuous float numbers, a cutoff is set as a decision boundary to categorize the physician integration status. That is to say, if the predicted value is lager than the cutoff then the physician will be labeled as integrated and the predicted indicator of integration will be set to 1, otherwise set to 0 if the physician is labeled as independent with predicted value less and equal to 0.5.

Among all machine learning methods implementing in this project, SVM is one of the most effective classification methods. Having tried four kinds of kernel, the ones with best performance are Radial Basis Function (RBF) kernel and polynomial kernel, both of which created a curved boundary, comparing to the performance of linear kernel and sigmoid kernel. For each implementation, I used grid search to determine the penalty parameter C and besides, to determine gamma for RBF kernel and degree for polynomial kernel.

Considering Naive Bayesian methods, GaussianNB, MultinomialNB and BernoulliNB are three of the classification algorithms in Scikit-learn. Among them, I applied the Bernoulli Naive Bayes model whose prior is Bernoulli distribution, assuming that all of outcome features are binary and have only two values in total. Comparing to decision trees, KNN and other algorithms, Naive Bayes has fewer parameters to pay attention to so that it is much easier to handle. The only parameter to be adjusted is the smoothing factor alpha that reduce zero probability, and the probability tends to 0.5 as alpha increases. In particular, that alpha equals to 1 represents Laplace smoothing, while that alpha equals to 0 means no smoothness is added to the data. As before, I used grid search on alpha to search for the optimal value.

In regard to the ensemble method, Random Forest is a variant of Bagging Algorithm that trains a number of Decision Tree models using random sample and decides the one with best performance by voting. This method does well in avoiding the problem of over-fitting and therefore ensures the accuracy of prediction. Again, I used grid search to choose the optimal parameters of number of estimators, maximum depth, minimum sample split and minimum sample leaf, the influence of which to the model reduces successively.

Other methods I used are Logistic Regression, Decision Tree and K-Nearest Neighbour (KNN) algorithms. The implementation of these algorithms are summarized in the previous section, so I won't go into details here.

All of the above machine learning algorithms other than the ensemble method ran twice on the provider data yearly from 2012 to 2015, once with Medicare claims data and the other time without claims. Due to the limitation on the ZIP-FIPS conversion file, although the original data scale is huge, there were a large amount of the records do not contain a provider ZIP code of 5 digits and hence were not able to be converted to RUCC, which is considered as an indispensable variable to the models.

3.3 Results and Discussion

3.3.1 Model Validation Results

After data cleaning, all records containing missing values were removed, thus the sample left for each year have the size of 32, 32, 484 and 30 respectively, which is extremely limited except for the year 2014. This is because each physician has more than one records in the 2014 short version data file, yet I am not going to remove the duplicates since these samples can be used to train and improve the models.

Comparing the values predicted by the models with the authentic values, the results can be concluded in a confusion matrix which counts the numbers of True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). According to the results, I calculated the accuracy and f1 score for claim-based models and claim-free models for data of each year, as shown in the following tables (Table 3 to Table 6). These tables are of the same structure, and they record accuracy or f1 score as percentage, separating the validation set.

Table 3 Accuracy for claim-based models

ML methods	2012		2013		2014		2015	
	Training set	Testing set	Training set	Testing set	Training set	Testing set	Training set	Testing set
Linear Regression	100.00%	87.50%	100.00%	87.50%	100.00%	100.00%	100.00%	100.00%
Logistics Regression	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
SVM - RBF kernel	100.00%	87.50%	100.00%	87.50%	100.00%	100.00%	100.00%	87.50%
Decision Tree	100.00%	100.00%	100.00%	87.50%	100.00%	100.00%	100.00%	100.00%
K-Nearest Neighbour	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Naive Bayesian	95.83%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Sample size	24	8	24	8	363	121	22	8

Table 4 Accuracy for claim-free models

ML methods	2012		2013		2014		2015	
	Training set	Testing set	Training set	Testing set	Training set	Testing set	Training set	Testing set
Linear Regression	100.00%	62.50%	95.83%	87.50%	87.88%	90.91%	86.36%	100.00%
Logistics Regression	70.83%	87.50%	70.83%	87.50%	87.05%	89.26%	72.73%	87.50%
SVM - RBF kernel	70.83%	87.50%	70.83%	87.50%	96.42%	96.69%	72.73%	87.50%
Decision Tree	100.00%	75.00%	95.83%	100.00%	96.42%	96.69%	77.27%	87.50%
K-Nearest Neighbour	79.17%	87.50%	62.50%	87.50%	96.42%	95.87%	81.82%	87.50%
Naive Bayesian	95.83%	62.50%	75.00%	87.50%	86.23%	88.43%	72.73%	87.50%
Sample size	24	8	24	8	363	121	22	8

Table 5 F1 score for claim-based models

ML methods	2012		2013		2014		2015	
	Training set	Testing set	Training set	Testing set	Training set	Testing set	Training set	Testing set
Linear Regression	100.00%	0.00%	100.00%	0.00%	100.00%	100.00%	100.00%	100.00%
Logistics Regression	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
SVM - RBF kernel	100.00%	0.00%	100.00%	0.00%	100.00%	100.00%	100.00%	66.67%
Decision Tree	100.00%	100.00%	100.00%	0.00%	100.00%	100.00%	100.00%	100.00%
K-Nearest Neighbour	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Naive Bayesian	92.31%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Sample size	24	8	24	8	363	121	22	8

Table 6 F1 score for claim-free models

ML methods	2012		2013		2014		2015	
	Training set	Testing set	Training set	Testing set	Training set	Testing set	Training set	Testing set
Linear Regression	100.00%	0.00%	92.31%	0.00%	73.17%	75.56%	72.73%	100.00%
Logistics Regression	0.00%	0.00%	0.00%	0.00%	71.86%	72.34%	0.00%	0.00%
SVM - RBF kernel	0.00%	0.00%	0.00%	0.00%	92.72%	92.86%	0.00%	0.00%
Decision Tree	100.00%	0.00%	92.31%	100.00%	92.74%	92.86%	90.91%	0.00%
K-Nearest Neighbour	44.44%	0.00%	0.00%	0.00%	92.74%	90.91%	54.55%	0.00%
Naive Bayesian	92.31%	0.00%	25.00%	0.00%	62.12%	63.16%	0.00%	0.00%
Sample size	24	8	24	8	363	121	22	8

Note that f1 score equals to 0 if the model categorizes all sample into the same class, either class-0 or class-1.

In order to obtain a more reliable classifier for reference, I also applied the ensemble method on the 2014 data without Medicare claims based on the comparison about the model discussing in the next section. Table 7 demonstrates that nearly all of the scores for Random Forest are over 90%, and its performance was similar to that of SVM, Decision Tree and K-Nearest Neighbour methods.

Table 7 Scores for Random Forest on 2014 data without claims

Random Forest	Accuracy	Precision	Recall	F1 score
Training set	96.42%	90.22%	95.40%	92.74%
Testing set	95.87%	86.21%	96.15%	90.91%

Although I had several trails with different parameters on some of the algorithms, the above tables only recorded the ones with highest accuracy and f1 score among all of the trails for each method. Intuitively, there were a number of models correctly classified the vertical integration status of every physician, no matter the training set or the testing set. Without misclassification, both of the accuracy and f1 score were 100% for these models.

3.3.2 Result Analysis and Discussions

For further analysis, this section will compare the model validation results from several dimensions.

Comparing between accuracy and f1 score, the former is rather high and none of the values are less than 70 percent, while the latter seems to have an elastic range from 0 to 100 percent. And sometimes the accuracy can be extremely high even though the f1 score is 0. This is because the accuracy may be "fake" when the sample is imbalanced or the sample size is too small. Take the testing set of 2012 as an example, there is merely 8 physician to be categorized. In accordance with the true indicator, all of them are labeled as independent. Despite that some of the algorithms performed perfectly on them, the models still got 0 in f1 score since the model assessing criteria filtered out the models that was thought fail to distinguish the other class of vertically integrated. Therefore, in general, f1 score acts as a trade-off between precision and recall is much more informative than accuracy, and it will provide a more objective evaluation on the models. So in this project, f1 score rather than accuracy served as a significant reference when selecting algorithms.

Owing to the limitation of sample size, the models may not be well-trained, leading to biasedness in prediction. In particular, the change in a single predicting value resulted in a substantial fluctuation in accuracy or f1 score for small sample, so the percentages in the above tables become unreliable in evaluation of the models. In comparison, the models that based on the data of 2014 were more robust and hence the corresponding accuracy and f1 score were more significant and credible.

Focusing on the data of 2014, considering the variables included in the models. Initially, I added the claims occurred in all settings and calculated the Variance Inflation Factor (VIF) for each of them to check if there exists problem of multicollinearity in these variables. In accordance with the rule of thumb, there was no collinearity and it is acceptable if VIF is less than 10. So from table 8 below I can deduce that there exists severe problem of collinearity in all of the claims variables except for the claims of inpatient. One effective way to deal with this problem is to remove one of the variables with VIF larger than 10 and check the VIF again for the rest until no variable can be removed. Then as shown in table 9, after removing the claims of total alone, the VIF for the rest indicated that the effect that one of them making on the predictor variable was almost undisturbed by the others. Equivalently, the problem of collinearity had been eliminated. So I included the claims of all settings other than the claims of total in the models. Not surprisingly, those consisted of Medicare claims data and other physician characteristics perform much better than the models without claims. Due to the reason that the authentic indicator of integration is obtained in terms of the proportion of outpatient claims, the claims data were closely related to the indicator to some extent. In the light of the association, the Medicare claims data may support the models in making a more effective and accurate decision on classification. Nevertheless, in terms of table 4 and table 6, the models behaved well enough with high accuracy and f1 score even though they were without the claims data and so they were adequate for the task of identifying physician-hospital integration status.

Table 8 Initial VIF table

Variables	VIF
claim_office	inf
claim_inpatient	1.07
claim_outpatient	inf
claim_asc	inf
claim_total	inf

Note that "inf" refers to infinity.

Table 9 Final VIF table

Variables	VIF
claim_office	5.58
claim_inpatient	1.07
claim_outpatient	1.54
claim_asc	5.68

According to the columns under 2014 in table 6, the methods SVM with RBF kernel, decision tree as well as K-nearest neighbour had the highest f1 score, which were more than 90% for both of training set and testing set. This manifested that these algorithms are not sensitive to the associations among between variables, so compared with other three algorithms they were well-behaved in classifying the physician-hospital integration status using only physician characteristics even when running on out-of-sample data. Additionally, it is not surprising if the f1 score for testing set was slightly higher than that of training set, as a result of the small sample size. From Table 7, the high percentages in all of the scores for training set and testing set respectively certified the excellent classification ability and predictability of the Random Forest method. For further comparisons, practically the predicting results of these four algorithms were almost the same, so it seemed that f1 score fail to be a criteria for the selection of algorithm to predict physicians' tendency of consolidating with hospitals in the following years.

Regarding to the other algorithms, in this paragraph I am going to discuss and explain the reasons why they have worse performance even when they run on the training set data. Considering the linear regression model, although I have eliminated the problem of multicollinearity in the claims variables, there may still exist collinearity in the dummy variables, physicians' years of experience, location variable of RUCC code and the combination of all these variables together with claims. However, the physician characteristics would make a significant contribution to the prediction of integration status and thus it did not worth the loss to remove any of these indispensable variables. Letting alone the collinear relationship, the standard error of the predictor variable got larger and the statistical significance of each independent variable was undermined, leading to lower prediction accuracy and f1 score. As for the Logistic Regression model, it is also sensitive to the problem of multicollinearity. Due to the simpleness of this algorithm, its accuracy rate was not very high, despite that it had excellent performance when solving linear problems. When it comes to the Bernoulli Naive Bayesian method, it is based on the assumption that the sample properties are independent with each other. In the Medicare data, however, the relationships such as the multicollinearity among some of the variables could hardly be totally cleared, leading to

the failure of this model and the corresponding low prediction accuracy and f1 score. Evidently, it is the collinear relationship that brought about the bad performance of these algorithms.

4 Prediction of Future Trend

4.1 Prediction Process

The data cleaning and merging procedures were similar to that of model validation. One of the differences is that I loaded the full version of CMS Medicare data file and removed duplicates. Because in the full version file there were several Medicare records for each physician, each of the NPI number repeated for several times. It is unnecessary to repeatedly predict integration status for the same physician, so I remained the first record corresponding to each NPI number and deleted the others. For the other thing, so as to store the predicting results for all of the years from 2015 to 2018, I created a function for prediction. The function will load all data files that needed and merge them as before, then removed rows containing missing values and duplicate rows. The returning value for this function is a data frame consisting of all of the independent variables that were going to be set into the model for prediction. The numbers of sample left for prediction are 189, 186, 235 and 1387 for each year respectively.

According to the results of model validation in section 3.3.1, the methods of SVM, Decision Tree, KNN as well as Random Forest were about the same but excellent goodness of fit and predicting ability for claim-free data of 2014. Thus in this part I applied all of these four algorithms to predict the integration status of each physician from the year 2015 to 2018 so that I will be able to compare the predicting results and to choose the ones that mirror reality as much as possible. After getting a binary list of integration indicator for the remaining physicians, I computed the percentages of physicians that are predicted to be vertically integrated with hospitals in each year for all of the above methods. Since the list only contains 0 or 1, the value of integration percentage is equivalent to the mean value of the list and I did so to obtain the percentages and store the ones predicted by the same algorithm in a new list in the order of predicting year. For visualization, I drew the line charts of the integration percentage for all methods in the same figure (see Figure 1 below) in order to analysis the trend of integration between physicians and hospitals. According to the results, the methods KNN and Random Forest had almost the same percentage of integration for each predicting year, so I calculated the coincidence rates for these two methods using the accuracy score function embedded in Scikit-learn. And therefore I was able to estimate the prediction accuracy for KNN and Random Forest methods by the coincidence rates.

4.2 Prediction Results and Discussions

Demonstrating from Figure 1 below, the Decision Tree classifier categorized the most physicians as integrated, which accounted for around a half. The second most of the integration percentage was predicted by SVM for the first to years, but as the percentage dives rapidly to 11% in 2017, both of KNN and random Forest methods took the second place. With slight fluctuations,

the integration percentages kept stable for all algorithms except for SVM that keeps levels on over 30% for the first two years and showed a downward trend from 2016 to 2018. Moreover, it is unexpected that the percentage of integration for SVM dropped considerably to 4.7% in 2018, so this method is likely to claim failure in predicting future integration tendency for physician. In regard to KNN and Random Forest algorithms, the numbers of physicians that were labeled as integrated with hospitals by them were nearly the same, which were between 20% and 25%, and the range was conformed with the integration percentages of physician characteristic from 2012 to 2015. Besides, the coincidence rates for these two algorithms for all years were around 85%, suggesting that high prediction accuracy was estimated for both methods. Since the ensemble method was expected to be robust and well-predictive, both of KNN and Random Forest algorithms were considered realizing the prediction of physician-hospital integration with high effectiveness. As the most reliable method, the Random Forest method had forecast ascending trend for vertical integration, while the KNN method suggested vibrated situation.

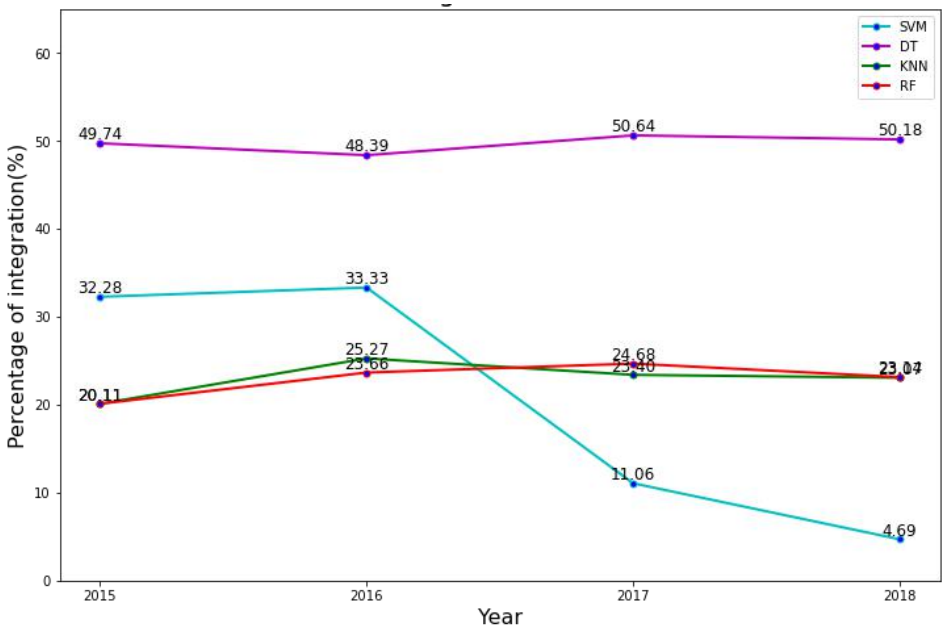

Figure 1 Trend of physician integration

Note that the integration percentage for K-Nearest Neighbour and Random Forest from 2015 to 2018 are [20.11, 25.27, 23.40, 23.07] and [20.11, 23.66, 24.68, 23.14] respectively.

Further analysis on the properties of the above algorithms would be needed to figure out the reasons behind the large gap in the predicted percentages of integration obtained from different algorithms. So in the next two paragraphs I am going to discuss the advantages and/or disadvantages for each of the four algorithms and that how will these properties influence the predictive ability of the models.

First, considering the algorithm that predicted the largest proportion of physicians that integrating with hospitals, Decision Tree was powerful in dealing with both of numerical data and categorical data and consistently the variables used in this project are either numerical or binary. Since the model can be well-trained using a smaller sample size, Decision Tree was easily encountered with the problem of over-fitting and therefore it is likely to construct a model with superfluous complexity, which had poor generalization. In addition, the model inferred by this

algorithm was unstable in the sense that even a tiny disturb added to the data can result in a model that would be totally different from the previous one. Owing to the limited sample size, the Decision Tree model probably over-fitted the data that used to train the model itself, leading to weak prediction ability. As for SVM, its classification idea is simple but effective when dealing with linear problem. Its validity to small sample size makes it suitable for our physician characteristic data but its sensitiveness to kernel type and parameter may also result in over-fitting because any change in the data is likely to affect the choice of optimal parameters. Hence, similar to Decision Tree model, SVM model might not behave well in categorizing the physician integration status in the predicting years on account of the changes in physician characteristics, especially the changes in the variables physicians' years of experience and the rural-urban code. As far as I am concerned, the large deviation of prediction results of Decision Tree and SVM was mainly caused by the over-fitness of the models.

Compared to Decision Tree and SVM, KNN algorithm was considered as the best-behaved one in predicting physician integration status, which has high tolerance to outliers and noise. The main problems of KNN are its computational inefficiency and the difficulty in choosing the value of K, and for the latter, I applied grid search to search for the optimal value of K to make sure the model is validated. As mentioned in the previous section, the ensemble machine learning methods have the collection advantages of all methods that it consists of, so it is expected to have better performance in model validation as well as prediction. In this project, the Random Forest algorithm, an extension of bagging algorithm that randomly chose subsets of features and created a number of Decision Tree classifiers, acting as a reference of the prediction results for other algorithms. The reason why it was more accurate than the other methods is that the Random Forest model took the mode of prediction values that made by all of the classifiers for each prediction, reducing the occasionality to a large extent. However, the predicted result can be ostensible if there were many resembling Decision Tree models that provide the same predicting results, increasing the possibility of this value to be output by Random Forest algorithm. From the analysis above, it can be said that both of KNN and Random Forest predictors succeed in predicting the tendency that physician consolidating with hospitals in the following years from 2015 to 2018.

Concerning the trend of physician-hospital integration, conceptually, the percentage of physician who chose to be owned by a hospital is anticipated to rise year by year and the empirical evidence from Medicare FFS claims data shows that for each year from 2008 to 2014 the percent of GI physicians who were newly integrated with hospitals increased by more than 10% (Song et al., 2020). Further more, in accordance with the American Hospital Association Hospital Statistics Report (2012), the number of hospital recruitment of physicians has increased by over 30% since 2000 (Cho et al., 2018) and presently the majority of the physicians were employed by hospitals (Kocher et al., 2011). Nevertheless, as observed in the physician characteristics from CMS data, the integration percentage kept descending during the years between 2012 and 2015, and then in respect to the prediction results above it vibrated with a slight rise in the following years until 2018.

5 Conclusions

5.1 Summary

To sum up, following the objectives, I cleaned and merged the CMS Medicare claims dataset and the NPPES provider dataset. To evaluate the classification ability for physician characteristics, I created another class of models which contains Medicare claims data and compared the performances between the two classes of models with each class contains 7 models of distinct machine learning algorithms. Not surprisingly, based on the 2014 provider data, the claim-based models behaved so well that they perfectly classified the integration status for almost all of the physicians, referring to the indicator of integration that obtained in accordance with the gold standard rule proposing by Neprash et al. (2015). Yet the models without claims data also had high accuracy and f1 score of over 90 percent, especially the methods such as SVM, Decision Tree, KNN and Random Forest.

Therefore, in the following section of predicting integration, I applied these four models trained by 2014 data to predict the integration status for physicians in the years between 2015 to 2018 (owing to the limitation on more recent data, I "predict" integration for the years after 2014 regardless of the authentic integration status). However, the trends indicated by Decision Tree and SVM are not practical. The former has classified about half of the physicians as integrated and the percentage remained unchanged until the last predicting year 2018, while the latter showed a sharp dive from 2016 to 2017 and the percentage dropped to only 4% in 2018. by contrast, the prediction ability for KNN and Random Forest seems more stable with integration rate increased slightly during these years. Despite all this, the results of this project are out of expectation comparing to the evidences from other studies, because there are some limitations for this project which will be demonstrate in the next section.

5.2 Limitations

Considering the reasons why the prediction results are out of expectation, there exists several limitations for this project but here I summarized them as two main limitations, one is on the data and variables, and the other one is on the models and relevant parameters.

The first limitation is related to the limitation on datasets and selection of variables. I have used physicians' gender, state, years of experience and rural-urban code as independent variables in the models. When calculating the experience years for physicians, the information on provider enumeration date needed for calculation can only be found in the NPPES provider data, which contains merely the records from 2005 to 2015. When assigning the RUCC code for physicians, the file used to matching contains only the ZIP code of 5 digits, whereas the ZIP code recorded in the CMS Medicare dataset for most of the physicians are of 9 digits. That is to say, the project did not consider the physicians who have provided a non-5-digit ZIP code or have registered in the years beyond the decade from 2005 to 2015. In spite of the richness of the original data source, rare of the physicians are eligible for the conditions above imputing to the limitation on the intermediate files.

Moreover, other variables might contribute more on predicting integration status for physicians, and they can be chosen to avoid making use of the intermediate files.

As for the second limitation, the machine learning algorithms I used are not targeted in order to find out a simple but efficient way to identify physician-hospital consolidation. The only adjustment on these general models is the values of the hyper-parameters. However, the optimal value is selected via applying the method of grid search, which only concerned about a limited list of values. Thus, the values of parameters chosen by grid search are probably not the best choices out of all possible options. What's worse, the models trained by the limited sample data are likely to be biased and over-fitted, which are not generalized enough to predict future integration status for physicians.

5.3 Study Implications

Corresponding to the limitations discussed above, the recommendations for this project can be concluded from the following two aspects. Firstly, more variables on physician characteristics can be added to make the models more reliable, but the choice of these variables needs to be deliberate so that the models are able to make full use of these variables and avoid unnecessary computations. Secondly, it is highly recommended to try more ensemble methods that reduce the possibility of contingency. When searching for the optimal values of parameters, a more intelligent alternative is the method of random search that attempt diversified combinations of parameters.

For further study implication, it will be contributive to find out some targeted models and the best combination of independent variables on physician characteristics for these models, including analysis on how each of these variables suggest the integration status for physicians. Besides, based on the results of this project, K-Nearest Neighbour and Random Forest models have the best performance on both of validation and prediction, therefore more ensemble machine learning algorithms related to these two methods worth trying to construct an effective model for predicting physician-hospital integration in the future. These make it convenient for researchers and policy makers to get access to the integration status of physicians so that they can do further research and make political decisions on allocation of medical resources respectively using the classified results.

Reference

· Baker, L. C., Bundforf, M. K.,Kessler, D. P. (2014) 'Vertical integration: hospital ownership of physician practices is associated with higher prices and spend'. *Health Affairs,* 33(5), pp.756-763.

· Baker, L. C., Bundorf, M. K., Kessler, D. P. (2016) 'The Effect of Hospital/Physician Integration on Hospital Choice'. *Journal of Health Economics,* 50, pp.1-8.

· Cho, N., Lee, S., Lee, J. D. (2018) 'Economic evaluation of the impact of physician-hospital integration and physician boards on hospital expenditure per patient'. *Medicine (Baltimore)*. 97(41):e12812

· Cuellar, A. E., & Gertler, P. J. (2006) 'Strategic integration of hospitals and physicians'. *Journal of Health Economics*, 25(1), pp.1-28.

· Cutler, D. M., Dafny, L., Grabowski, D. C., Lee, S, Ody, C. (2020) 'Vertical Integration of

Healthcare Providers Increases Self-Referral and Can Reduce Downstream Competition: The Case of Hospital-Owned Skilled Nursing Facilities'.
· Gaynor, M. (2006) 'Is vertical integration anticompetitive? Definitely maybe (but that's not final)'. *Journal of Health Economics*, 25, pp.175-180.
· Ho, V., Hamilton, B. (2000) 'Hospitals Mergers And Acquisitions: Does Market Consolidation Harm Patients?'. *Journal of Health Economics*, 19(5), pp.767-791.
· Kocher R., Sahni N. R. (2011) 'Hospitals race to employ physicians—the logic behind a money-losing proposition'. *Journal of Medicine*, 364, pp.1790–1793.
· Post, B., Buchmueller, T., Ryan, A. M. (2017) 'Vertical Integration of Hospitals and Physicians: Economic Theory and Empirical Evidence on Spending and Quality'. *Sage Journals*, 75(4), pp.399-433.
· Raus, K., Mortier, E., Eeckloo, K. (2020) 'Challenges in Turning a Great Idea into Great Health Policy: The Case of Integrated Care'. *BMC Health Services Research*, 20(1), pp.130.
· Robinson, J. C., Casalino, L. P. (1996) 'Vertical integration and organizational networks in health care'. *Health Affairs*, 15(1), pp.7-22.
· Robinson, J. C., Miller, K. (2014) 'Total expenditures per patient in hospital-owned and physician-owned physician organizations in California'. *Journal of the American Medical Association*, 312, pp.1663-1669.
· Scheffler, R. M. , Arnold, D. R., Whaley, C. M. (2018) 'Consolidation Trends In California's Health Care System: Impacts On ACA Premiums And Outpatient Visit Prices'. *Health Affairs*, 37(9), pp.1409-1416.
· Song, L. D., Saghafian, S. , Newhouse J. P., Landrum M. B., Hsu J. (2020) 'The Impact of Vertical Integration on Physician Behavior and Healthcare Delivery: Evidence from Gastroenterology Practices'.