University of Waterloo

Faculty of Engineering

Department of Management Sciences

MSCI 723: Final Project Report

U.S Home Health Care Data Analysis

Prepared by:

Weiyi Huang ID: 20426153

Ke Chen ID: 20456026

April 3rd, 2017

Table of Contents

1. Introduction

Abstract: The goal of our project is to find which quality measures are important to determine the overall care rating of home health care agencies. In specific, we are going to analyze quality measures from two aspects. First we build Decision Tree models to find out the important features by comparing prediction accuracies. Second we apply the association rules to further analyze the relationship between some quality measures.

Health care is a popular topic nowadays since an increasing number of people start to pay attention to their health problems. This social phenomenon has given rise to many health care services provided by different institutions. For example, the Home health care that is a wide range of health care services are provided to patients at home by professional health team. The goal of home health care is to treat an illness or injury while offering patients the most convenience. [1] The Home health care is well known for being less expensive and as effective as care that patients receive in a hospital. There are many factors that can affect the quality of care provided by home health care agencies. Therefore, the purpose of this project is to analyze the U.S. home health data to find out the quality measures that matter most to patients and how they influence the overall performance of the home health care agency.

Our motivation for doing this project contain two aspects. One aspect is that we are interested in studying health care problems. As mentioned earlier, home health care is very popular in today's society and thus it is a meaningful topic to discover. And our analysis from this project could provide some insight as how to improve the health care agencies' provided care services. The other motivation is that we want to apply the big data analytics knowledge that we obtained from the course on this project and gain some practical machine learning experiences.

The project analysis is beneficial to health care agencies who want to improve their care services. By examining the importance of quality measures and the correlation between the measures and the final overall care rating, health agencies can better understand the needs of patients and improve the quality of health care accordingly. For example, if an agency is experiencing a poor business time, this agency could refer to our output to check there might exist some important services that they are not performing well or they should introduce some other services that they may have missed in order to improve their overall performance.

We propose that well communication between the home health team and patients would have a positive impact on health agencies' overall performance. This is because clear communication would help patients follow doctor's' orders precisely and help the health team understand patient's' emerging needs. In addition, we hypothesize that the provided patient drug education service is influential in determining the overall care performance. This is because teaching patients about their drugs can help patients to develop a good drug-use habit, which is beneficial for their recovery process.

According to our research, we have found some related scholarly work done on the home healthcare topic. The research paper "A data mining approach in home healthcare: outcomes and service" adopted a data mining approach to investigate the drivers of home healthcare service outcomes in the forms of discharge and length of stay under three specific conditions (chronic obstructive pulmonary disease, heart failure and hip replacement). The paper concluded that "Patients at age 85 or older was a driving force in discharge destination and length of stay for all three conditions." The other paper "Home Health Agency Work Environments and Hospitalizations" explored the effect of work environments on the 2 home health care quality measures. It concluded that good work environments could pose positive effect on the patience outcome and lead to fewer nurse burnout.

Both of these two papers discussed the different factors that influence on the home healthcare service outcomes while our project focuses on comparing the different home health agencies across U.S. and predicting the overall care rating of the health care agencies.

2. Exploratory Data Analysis

1) Data Description

Our project data (Home Health Compare data) was downloaded from the Data.Medicare.gov website which is the Official U.S. Government Site for Medicare. (The data source link is: https://data.medicare.gov/data/home-health-compare)

The dataset was the official dataset used on the Medicare.gov Home Health Compare Website and allow people to compare the quality of care provided by home health agencies across the United States. It was collected from the period of July 1, 2015 - June 30, 2016 and was compiled in a zip CSV file format. The original downloaded dataset included eight CSV files and only two of them were of interest to us. Among the two files, the first file contains information on the home health agency and the agency performance on quality measures. (The file name is: HHC_SOCRATA_PRVDR.csv) The second file contains information on the Patient Experience of Care Survey results for each home health agency. (The file name is: HHC_SOCRATA_HHCAHPS_PRVDR.csv) We then combined these two files as the input to our project since these two files record the performance evaluation for the same home health agencies in the U.S.

After removing the footnote columns in the file, our combined file included 12058 rows and 42 columns. The features include both categorical and numeric data types. Further, they can be classified into six major groups and details are presented in Table A-1 in the appendix.

2) Data Cleaning

After reading our data into a python data frame, we start our data cleaning process by checking data values inconsistencies. Upon a quick scan of the input data, we notice that there is some inconsistent data content. The patients' survey results columns contain 'Not Available' text values to indicate missing data while other columns contain 'NaN' for missing data. Thus, we need to standardize data by converting 'Not Available' text values to none values (NaN) in order to keep a consistent format for missing data.

We then check columns data types and found some numeric columns are recorded as string in the data frame. Thus, we need to convert the string to numeric type for those numeric columns.

After completing the data type checking and conversion, it is important to obtain an overview of the missing values information of dataset. Based on the generated summary table (Please see the Figure A-1 in appendix), we could note that there are columns that have a large amount of missing data. So we decided to first drop columns and rows that contain less than 70% of data. The reason for choosing 70% as the threshold is because 70% is a commonly used threshold number in the data cleaning process. Also, since we want to use patient care star rating as our class variable, we dropped the missing values in that column as well for more accurate and reliable star rating information. Then we fill in remaining missing data with the column average values. After data cleaning process, our data reduced to 7802 rows and 37 columns.

3) Exploratory data analysis

We explore our dataset along different aspects. The first aspect is to study the number of health agencies group by state as we want to know which states have large number of health agencies. The obtained bar chart is presented below.

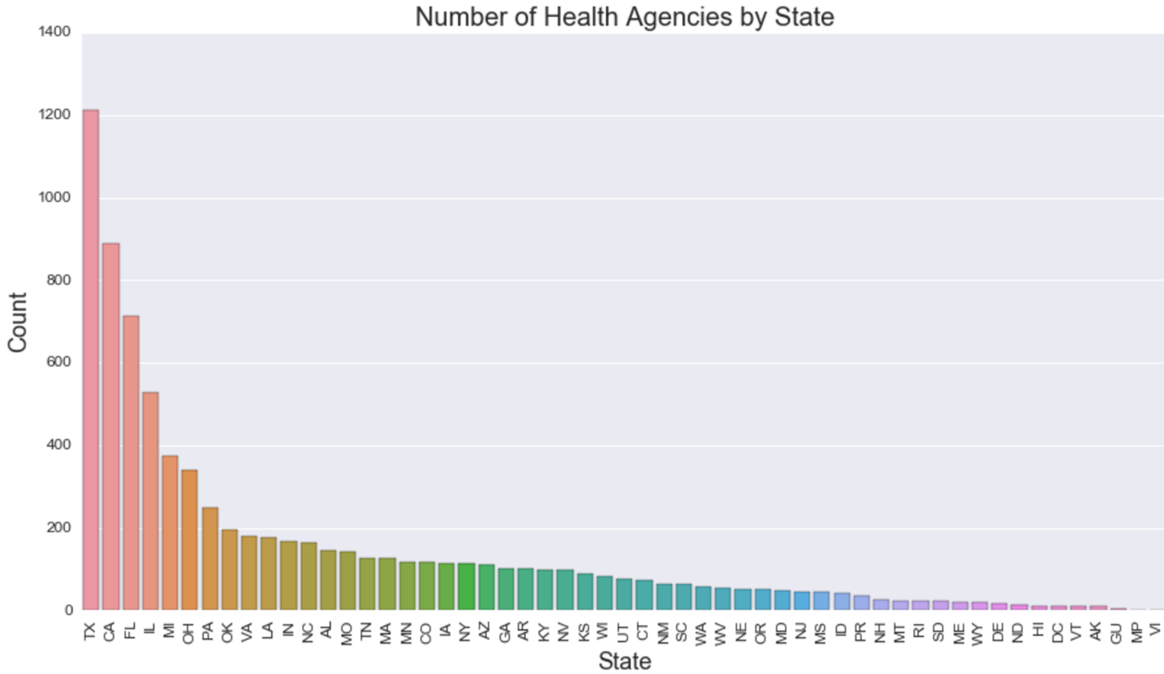
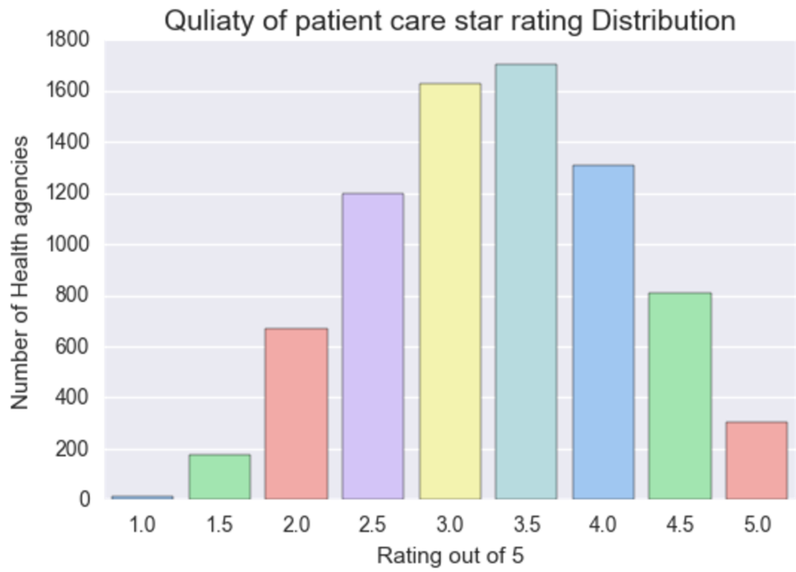


Figure 1: Number of health agencies by state

The above graph shows that state TX (Texas) has the highest number of home health service providers across U.S. with around 1200 home health agencies. Further, the top 10 states that have the most number of health agencies in the U.S. are TX, CA, FL, IL, MI, OH, PA, OK, VA, LA.

After checking on the number of health agencies per state, we are interested in studying on the distribution of the overall quality of patient care star rating of all health agencies. The reason to study on this feature is that overall care star rating is the class variable that we want to predict. The output graph is presented below.



count	7802.000000
mean	3.309344
std	0.833432
min	1.000000
25%	2.500000
50%	3.500000
75%	4.000000
max	5.000000

Figure 2: Quality of patient care star rating distribution

Based on the generated bar chart above on the left, it shows that rating between 3.0 to 4.0 take up the majority of health care agencies. In particular, rating of 3.5 is the most common star rating with around 1700 of health agencies. The second and third common star rating is 3.0 and 4.0 with around 1600 and 1300 health agencies respectively. The figure on the right presents some statistics measures of the overall care star rating. It indicates that the mean of rating is around 3.3 with standard deviation around 0.83.

We then would like to explore the average of overall care star rating on state level that is to find out the distribution of star ratings group by state. The output is presented as follows.

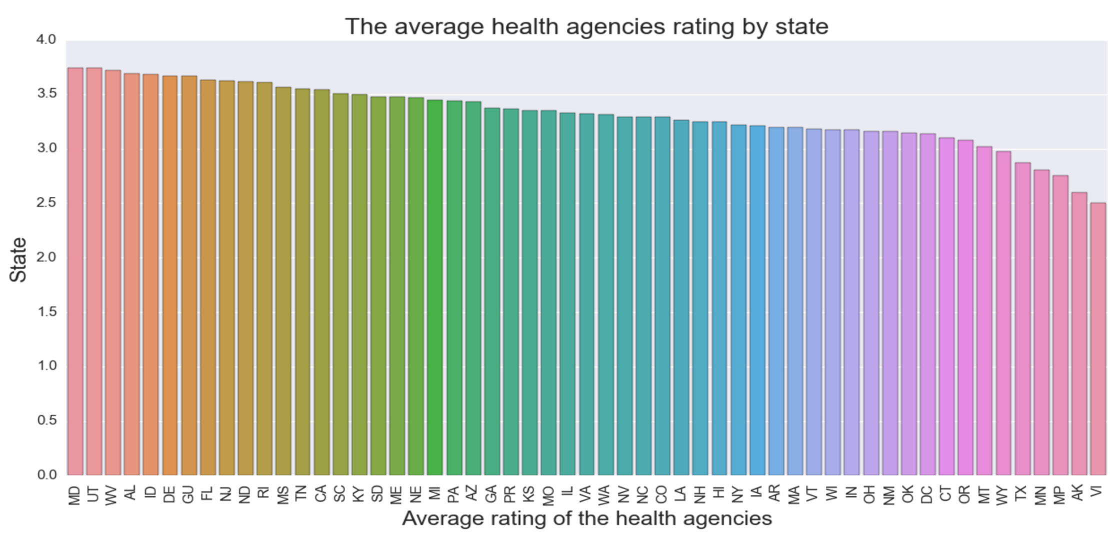


Figure 3: The average health agencies star ratings by state

Based on the graph above, we can see that state MD (Maryland) has the highest average health care star rating with rating around 3.7. Further, the states with the highest 10 average ratings are MD, UT, WV, AL, ID, DE, GU, FL, NJ, ND. The states with the lowest 10 average ratings are DC, CT, OR, MT, WY, TX, MN, MP, AK, VI. It is interesting to note that state TX (Texas) that has the highest number of home health agencies as mentioned earlier receives a low star rating (rating less than 3). We are interested to find out what causes this to happen. Therefore, we are motivated to study the variances of the star ratings by generating box plots of ratings for the top 10 states that have the most number of health agencies as shown below.

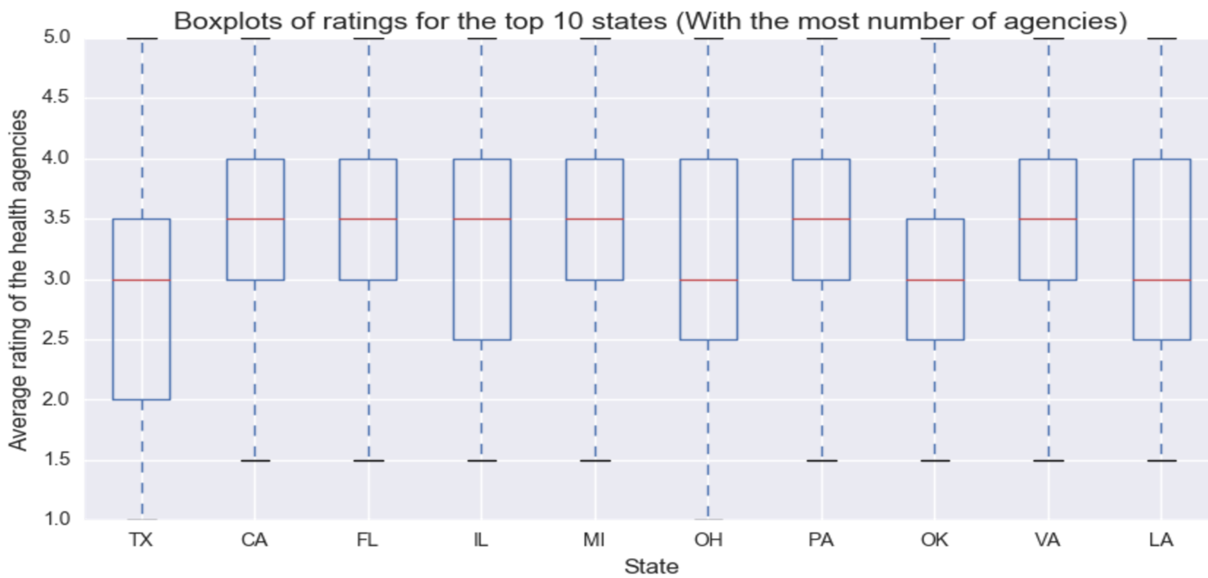


Figure 4: Boxplots of ratings for the top 10 states with most number of agencies

From the generated boxplot above, we can see state Texas that is on the leftmost of the graph, has the median at 3.0 and it has a large data spread with Interquartile range equals to 1.5. Also, the star ratings distribution for state Texas is skewed, with most of the ratings concentrate on the values from 3.0 to 3.5.

In addition, we also want to check the variances for the top 10 states with the highest average star ratings. The output graph is shown below.

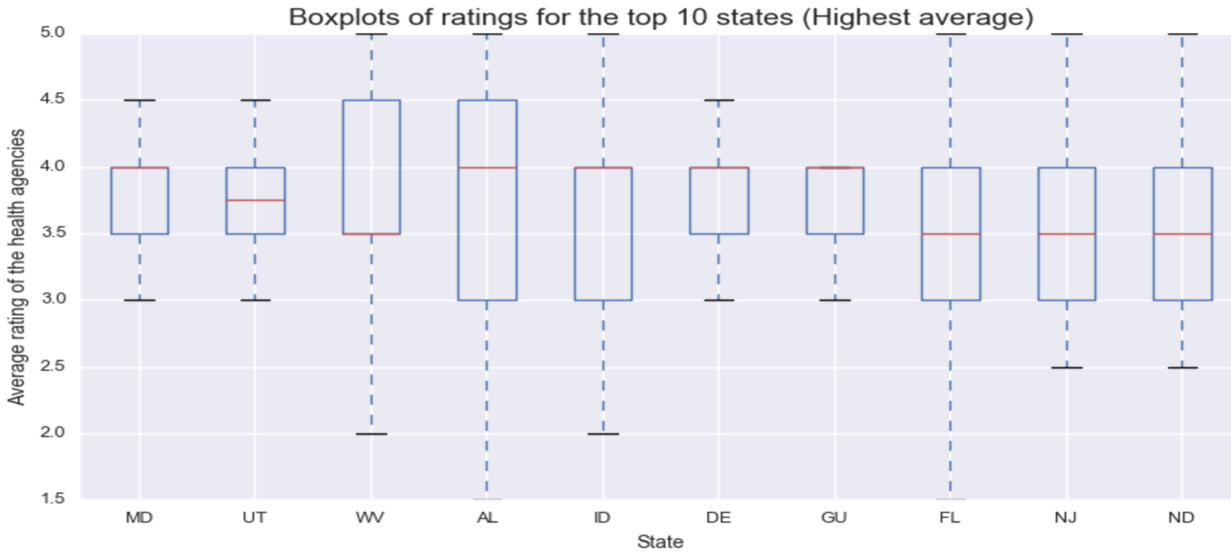


Figure 5: Boxplots of ratings for the top 10 states with the highest average rating

Some interesting observations can be made from the boxplot above. For states MD, ID and DE, their medians overlap with the upper quartile. This indicates that they have highly repeated star ratings. For example, these three states all have ratings concentrate on value of 4. Also, for state WV, its median overlaps with the lower quartile. This indicates that it has a lot of repeated ratings of 3.5. Furthermore, state GU has median overlap with the upper quartile and it does not even have an upper whisker. It means that the ratings are highly concentrated on value 4.

Last but not least, we want to do a deeper analysis of the average rating of the health agencies by going down to city level. We choose state MD (Maryland) as our focus of analysis since it has the highest average star rating of all states in the U.S.

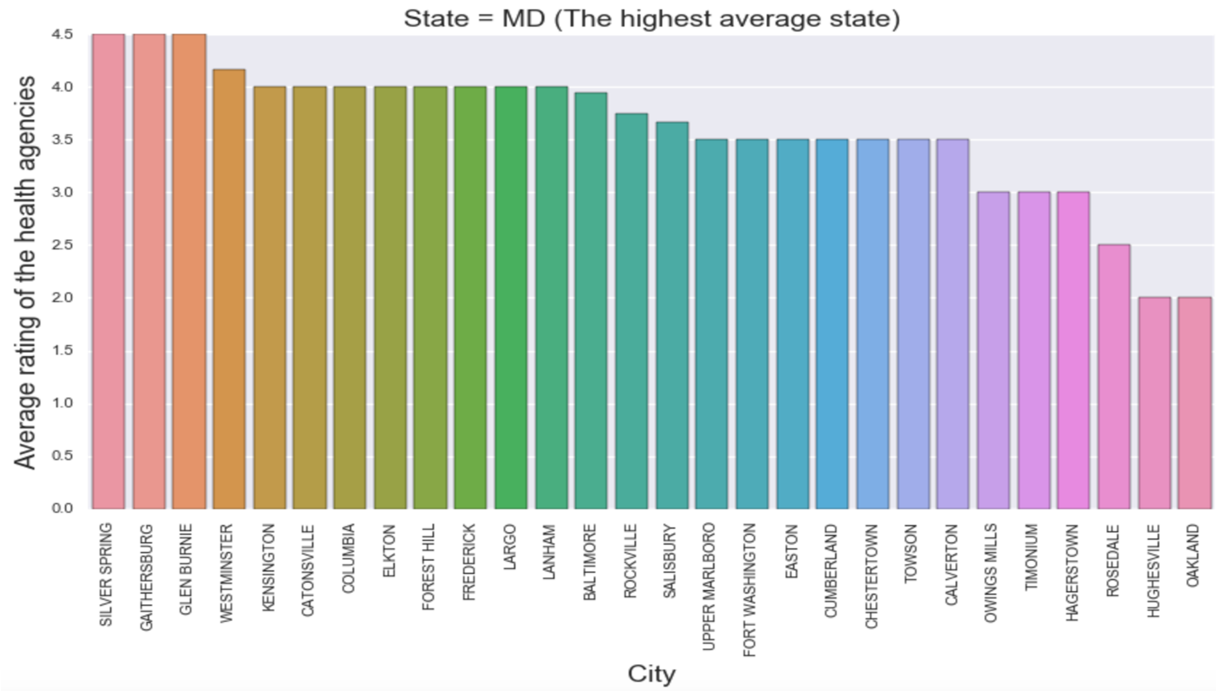


Figure 6: The average health agencies star ratings by city in State Maryland

From the output bar chart, it is noted that city of Silver Spring, Gaithersburg, Glen Burnie all have the highest average star rating of 4.5. Also, city of Hughesville, Oakland have the lowest average star rating of 2.0

3. Classification

1) Motivation

Our primary goal of doing this project is to determine important features which would have significant influence on overall quality of care star rating. We plan to find those features by building the prediction models, and we can compare the results of the prediction models to analyze feature selection. The reason why we focus on classification in our supervised learning part is that our class variable star rating is of discrete data type. Overall, in order to achieve the final goal of this project, our procedure is as follows: data pre-processing, modelling, evaluation of features by prediction results and analysis.

2) Sample of labelled training dataset

The project data includes numeric and categorical variables. In python, we mainly used package "pandas" for data processing. Since our goal is to select the important features which have an impact on class variable, we keep all features in our training dataset and use them for modelling and feature selections. (Please see the sample of labelled training dataset in Figure A-2 in appendix)

3) Data pre-processing: modifications made to the data to suit python

In the data pre-processing process, we did the following steps:

a) We replaced the current feature names with shorter representations. In the original dataset, feature names are given in questionnaire sentences which is long and difficult for future implementation; therefore, we need to convert them to short name form. For example, the original feature name is "Offers Nursing Care Services", and we replace it with "Nursing_Care". The figure below shows the sample of features' names before and after converting.

```
[('Offers Nursing Care Services', 'Nursing_Care'),
 ('Offers Physical Therapy Services', 'Physical_therapy'),
 ('Offers Occupational Therapy Services', 'Occupational_Therapy'),
 ('Offers Speech Pathology Services', 'Speech_pathology'),
 ('Offers Medical Social Services', 'Medical_Social'),
 ('Offers Home Health Aide Services', 'HomeHealth_Aide'),
 ('Quality of patient care star rating', 'Star_rating'),
 ('How often the home health team began their patients\x92 care in a time
ly manner',
  'Timely_manner'),
```

Figure 7: converting feature names results

b) Then we implemented "LabelBinarizer" and "One-hot Encoding" on this dataset in python. Training dataset:

   1. In feature dataset: there are some category features given in string format "yes" and "no", then we convert them to binary features for modelling. Then, we apply "reshape" function on each column converting them into proper format for modelling step.

   2. Class variable: The star rating is displayed as discrete data from 1 to 5. So we used "bin" and "cut" function to rename the class variable as "Low", "Okay", "Average", "Good" and "Great". We also use "reshape" function to convert class variable into proper format for modelling step.

4) Model and hyper-parameters:

We built 16 models in total by implementing Decision Tree and Naive Bayes algorithms. The following section discusses the models for both algorithms in detail.

Decision Trees

Overview:
We chose Decision Trees as our primary model since it is an interpreted and human-readable algorithm. We used information theoretic based Decision Tree, which relies on the basics of entropy and information gain. We can produce the rank of important features based on information entropy. For instance, the root node of the Decision Tree is the feature with the least uncertainty to predict the class variable, so we will consider this feature as the most significant feature. As a result, in python implementation, we set criterion equals “entropy”.

Furthermore, since Decision Tree is likely to have overfitting problem, we use the tuning hyper-parameter “min_samples_leaf” which is the minimum number of samples required to be at a leaf node.[4] Thus, when choosing small value of “min_samples_leaf”, the accuracy would become high but it is likely to cause overfitting problem.
The following section discusses the detail steps of how we choose subset of features and hyper-parameters.

a) First, We use all 37 features to build the full model by using Decision Tree, and we use “feature_importances_” function to produce the importance of features and summarized them in Table A-2 in appendix. Based on this table, we can see the root node is “bathing”, and all features that follow are sorted by their importance in decreasing order.

b) Then, after we have the importance of features, we want to know which combination of features is significant to determine the class variable. Thus, we built another 7 models by including different number of features:

- Model 1: Decision Tree with most 2 important features; such as “bathing” and “Lesspain_Walking”. (For feature full names, please see Appendix Table A-3; For feature importance, please see Appendix Table A-2.)

- Model 2: Decision Tree with most 5 important features; such as “bathing”, “Lesspain_Walking”, “Timely_manner”, “Flu_shot” and “Drug_taught”.

- Similarly, Model 3 contains 7 important features, model 4 contains 9 important features, model 5 contains 10 important features, model 6 contains 11 important features, model 7 contains 12 important features and full model contains all features by Decision Trees respectively.

c) Choosing Hyper-parameters: Since we build 8 models in total including different features, we want to choose the appropriate value of hyper-parameter “min_samples_leaf”. So We loop through the “min_samples_leaf” value from 1 to 100 for each of the 8 models and produce 8 figures(See Figure A-3 in Appendix). Based on these 8 figures, we decide to choose “min_samples_leaf” value equal to 90 for all 8 models. Each figures shows the training score and cross validation score versus min_samples_leaf, and we are expecting to find the point where training score and validation score intersect. As we can see from the 8 plots, two lines are asymptotic to each other gradually, almost all 8 models have training score and cross validation score intersected around "min_samples_leaf" equals to 90. Thus this is a reasonable value to use for the "min_samples_leaf" hyper-parameter. The figure below shows an example of hyper-parameter selection.

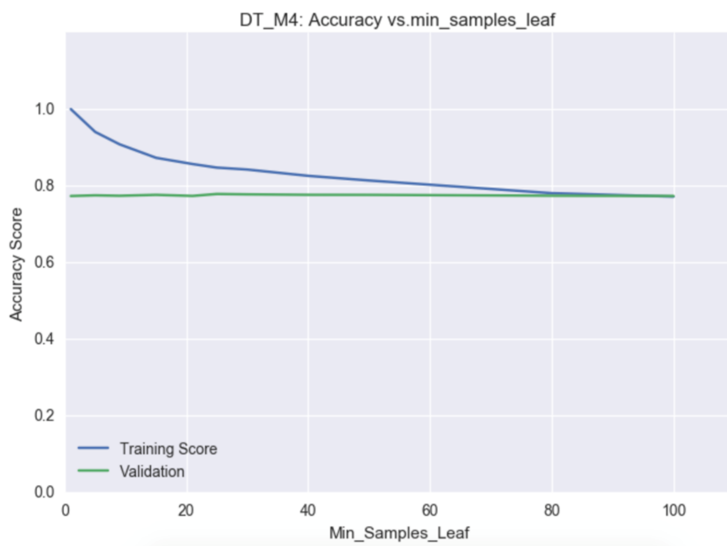


Figure 8: Min_samples_leaf hyper-parameter selection

Gaussian Naive Bayes
In order to do the deeper analysis, we consider applying Gaussian Naive Bayes algorithm to further verify the decision trees result. The process of applying Naive Bayes is much easier than Decision Trees since Naive Bayes does not contain hyper-parameters. Also, we build 8 models in Naive Bayes and each model contains the same training dataset as those 8 models implemented in Decision Trees. For example:

- Model 1: Naive Bayes with most 2 important features; such as "bathing" and "Lesspain_Walking".
- Model 2: Naive Bayes with most 5 important features; such as "bathing", "Lesspain_Walking", "Timely_manner", "Flu_shot" and "Drug_taught" (For feature full names, please see Appendix Table A-3; For feature importance, please see Appendix Table A-2.)
- Similarly, Model 3 contains 7 important features, model 4 contains 9 important features, model 5 contains 10 important features, model 6 contains 11 important features, model 7 contains 12 important features and full model, contains all features by Naive Bayes respectively.

In summary, we have built 16 models in total by using Decision Trees and Naive Bayes with the chosen subset of features and hyper-parameters.

5) Model Evaluation:

Since we build 8 models by using Decision Trees and 8 models by using Naive Bayes. Thus, we are going to compare the results by producing and analyzing the summary table and Accuracy score vs the number of features that are presented below.

Accuracy Score vs Cross-Validation Score				
	DT(min_samples_leaf = 90)		**NB**	
	Training score	**Cross-validation**	**Training score**	**Cross-validation**
Full model	0.7786	0.75	0.6574	0.64
(36 features)				
Model 1 (2 features)	0.7149	0.58	0.6973	0.7
Model 2 (5 features)	0.7735	0.72	0.7351	0.73
Model 3 (7 features)	0.7794	0.76	0.7584	0.75
Model 4 (9 features)	0.7786	0.78	0.7719	0.77
Model 5 (10 features)	0.7786	0.77	0.7616	0.76
Model 6 (11 features)	0.7786	0.77	0.754	0.75
Model 7 (12 features)	0.7786	0.77	0.7529	0.75

Table 1: Accuracy Score Vs Cross validation Score for DT and NB

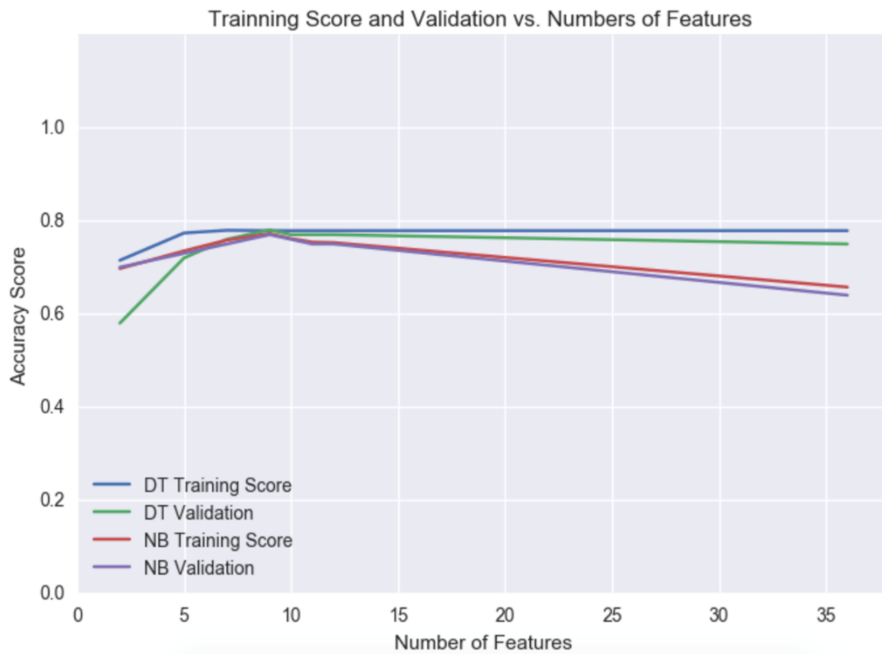


Figure 9: Accuracy Score and Cross Validation Score VS Number of Features

Decision Trees Evaluation

First, we focus on analyzing Decision Tree results. Since we selected the optimal hyper-parameters "min_samples_leaf" equals to 90, the overfitting is controlled generally. As can be seen from Table 1 above, Training Score increases from 0.7149 to 0.7786 with the number of important features increases from model 1 to model 4. The training score remains unchanged after 9 features which corresponds to model 4. The validation score follows the same pattern, it reaches the highest accuracy at model 4 (9 important features). In addition, the difference between training score and validation score decreases from model 1 to model 4 and increases after model 4, which means model 4 has the least overfitting. Overall, for Decision Tree, model 4 with 9 important features has the best performance.

Naive Bayes Evaluation

The training score of Naive bayes increases from model 1 with two features to model 4 with 9 features, and the maximum training score is 0.7719 in model 4. Then, by adding more features, the training accuracy decreases. The validation score follows the same pattern as training score, and the model 4 with 9 important features shows the peak accuracy of 0.77. On the other hand, the difference between training score and validation score in model 4 is 0.0019 which is very small. According to the Naive Bayes accuracy evaluation, the finding verifies the results produced by Decision Tree. The model 4 with 9 important features has the best performance.

Overall, the model 4 with 9 important features reaches the best training score, validation score, and least overfitting. Moreover, we will analyze the result based on this evaluation outcome.

6) Prediction

The figure below captures a part of the decision tree from our final selected model.

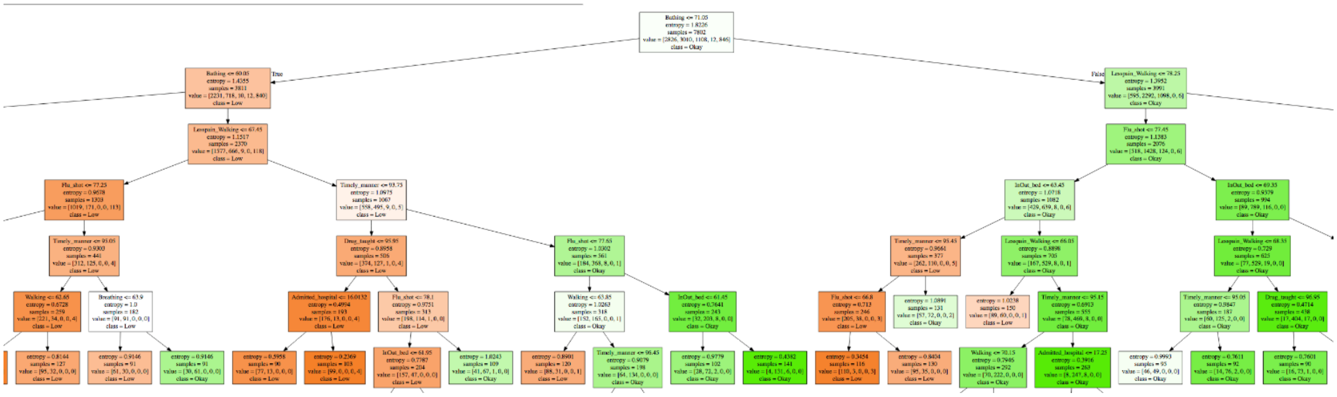


Figure 10: Graph of the partial Decision Tree

7) Analysis of result

Based on the evaluation results, the model 4 with 9 important features has the best performance because adding more other features or deleting existing features will decrease the accuracy and increase overfitting. Recall our business goal which is to find the influential features, thus we will focus on analyzing features in model 4. We summarized those features in Table A-4 in appendix. From this table, we have the most significant quality measures sorted from the most important to the least important including "Bathing" , "Lesspain_Walking", "Flu_shot", "Timely_manner", "In_Out_bed", "Walking", "Drug_taught", "Breathing" and "Admitted_hospital". And "Bathing" is our decision tree's root node. Please refer to Table A-3 in appendix for the full name of each feature.

According to our output, Home Health Care Agencies should improve their medical equipment and services on these 9 features. For example, from the summary table we can see "Bathing" has the highest importance value of 0.47. It means that the health care agency should put lot of effort to improve their bathing facilities and bathing service. Patients' feeling of bathing service highly impact on the agency overall quality of care star rating. For another example, "Lesspain_Walking" that is whether the patients had less pain when moving around, is the second important feature and has importance value of 0.18. The health care agency should pay more attention to patients' recovery from injury when they are moving around.

4. Association

1) Binning numeric variables

Our group used Association Rule Mining for the unsupervised learning part. The rationale behind choosing association method is that the output rules can give us more insight into the hidden patterns of our health care dataset.

Before applying association to the dataset, we need to convert numeric features to categorical. Upon studying the distribution of the numeric variables, we can observe there are two types of numeric columns. One type has continuous numeric range from 0 – 100 and it includes the quality measures columns; the other type contains discrete values from 1- 5 and it includes the star ratings columns. So we have decided to classify the star ratings columns into 5 categories with each category corresponds to a discrete number. (i.e., category 1 contains rating 1 and category 2 contains rating 2, etc.).

For the remaining numeric columns that have values ranging from 0 -100, we decided on the number of required bins and bin size by generating and inspecting the histograms of numeric columns. Based on those generated histograms, it is noted that the features can be further divided into 4 groups and they are summarized in the table below. Please see the sample of generated histograms in Fig A-4 in appendix.

Columns	Bins interval	Explanations
Check risk of falling, Checking depression, Teach about foot care, Teach about drug use, provide care in a timely manner	[0, 90, 92.5, 95, 97.5, 100]	They have most values fall within the range of 90-100. So make the first bin on values < 90 and the rest of bins on values between 90 – 100 with bin size = 2.5.
Admittance to hospital Admittance to ER	[0,10,15,20,25,50]	They have values range from 0 – 50 with data concentrated around 10 to 25. So make the first and last bin on values < 10 and > 25 respectively. The rest bins on values from 10 to 25 with bin size = 5.
Recommendation	[0, 60, 70, 80, 90, 100]	Values concentrated around 60 – 100, so make the first bin on values < 60, and the rest bins on 60 -100 with bin size = 10

Remaining columns	[0, 20, 40, 60, 80, 100]	Classify the values into 5 groups with bin size = 20

Table 2: Summary of the groups

Note that all the numeric features are classified into 5 groups and are assigned with corresponding category values "Low", "Okay", "Average", "Good", "Excellent".

Further, the Admittance to Hospital and Admittance to ER columns are assigned with category values from "Excellent", "Good", "Average", "Okay", "Low". This is because these two columns are the frequency of patients being admitted to the hospital and ER. So we assume that the lower the frequency value, the better the Health care service. In other words, if patients receive good quality health service, then they do not need to be admitted to the hospital or ER room for further treatment. As a result, we gave the first class of values from 0 – 10 the name "Excellent", indicating the highest quality of health care service provided by agency. While assign the last class containing values from 40 to 50 with the name "Low".

2) Feature selection and results comparison

Since our dataset contains a lot of features, it is not quite feasible to pass all the features at once into the association Apriori model. Therefore, we need to select few features and pass them to the model to test our hypotheses. Our hypotheses are derived from the different aspects that variables measure. For example, we want to explore how the provided care service affects the outcome of care measures. The detail explanation of hypotheses and the model results are presented below.

a) **Hypothesis 1**: There is a strong correlation between how often the home health team determined patients received a flu shot and how often the team made sure patients received a pneumococcal vaccine. Also, they are important factors in determining the quality of patient care star rating.

Since these two features are both about checking patients' taking shots, we assumed they are related with each other. In addition, we want to explore whether these two features have high association with patient care star rating. So we start with support = 0.15 and confidence = 0.90 since this is a commonly used combination. We would change the threshold each run based on the returned rules. The used hyper-parameters and the output rules are summarized in the table as follows. (Note: In the table, S stands for the support and C stands for Confidence)

Input features: Flu_shot, Pneumococcal_vaccine, Star_Rating	S = 0.15 C= 0.90	S = 0.10 C= 0.90	S = 0.10 C = 0.85

Rules output	No rules returned	No rules returned	('Excellent_flu',) ==> ('Excellent_vaccine',) , 0.851
Rationale for the chosen hyper-parameters	Starting threshold	No rules returned from the previous run, so we decreased the support.	Still no rules returned from the previous run, so we decreased the confidence threshold.

Table 3: Summary of the run for Hypothesis 1.

Results: Based on the table above, we can see if the health team check patients' flu shot very frequently, then they will check patients' pneumococcal vaccine very frequently as well.

b) **Hypothesis 2**: There is a strong correlation between how often the health team provided patients with drug use education and how often patients got better at taking their drugs correctly by mouth. Also, the drug education factors are important to determine the quality of patient care star rating.

Since these two features both involve the drug use aspect, one is about providing drug education service while the other measures the outcome of the provided service. Thus, we assumed they are correlated with each other. In addition, we want to explore whether these two features have high association with patient care star rating. The used hyper-parameters and the output rules are summarized in the table as follows.

Input features: Drug_taught, Drug_bymouth, Star_Rating	S = 0.15 C= 0.90	S = 0.10 C= 0.90	S = 0.05 C = 0.90
Rules output	No rules returned	('Excellent_starR',) ==> ('Excellent_drug',) , 0.921	('Excellent_starR',) ==> ('Excellent_drug',) , 0.921 ('Good_drugMouth', 'Excellent_starR') ==> ('Excellent_drug',) , 0.940
Rationale for the chosen hyper-parameters	Starting threshold	No rules returned from the previous run, so we decreased the support.	Further decreased support to see if there are any additional rules returned.

Table 4: Summary of the run for Hypothesis 2.

Results: Based on the table above, we can see if the quality of patient care star rating is excellent which the highest rating 5 is, then the drug education provided by the home health team is excellent as well. Further, it is observed that if patients are good at taking drugs correctly by mouth and the quality of care star rating is excellent, then the provided drug education is excellent.

c) **Hypothesis 3**: There is a strong correlation between how often the health team taught patients about foot care and how often patients got better at walking and had less pain when moving around. Also, the foot care and foot outcome measures are important to determine the quality of patient care star rating.

The reason for choosing this group of features is that they are all about foot care aspect. One measures the foot care service provided by heath team while the other two features are the service outcome measures. Thus, we assumed they are correlated with each other. In addition, we want to explore whether these features have high association with patient care star rating. The used hyper-parameters and the output rules are summarized in the table as follows.

Input features: 'Foot_care, Walking, Lesspain_Walking, Star_Rating	S = 0.10 C= 0.90	S = 0.10 C= 0.80	S = 0.10 C = 0.70
Rules output	No rules returned	('Excellent_starR',) ==> ('Excellent_foot',) 0.844 ('Excellent_starR',) ==> ('Excellent_lessPain')0.821 ('Good_starR',) ==> ('Good_walking',) , 0.859	('Excellent_lessPain',) ==> ('Excellent_foot',) , 0.730 ('Excellent_lessPain', 'Good_walking') ==> ('Excellent_foot',) , 0.708
Rationale for the chosen hyper-parameters	Starting threshold	No rules returned from the previous run, so we decreased the confidence.	Further decreased confidence to see if there are any additional rules returned.

Table 5: Summary of the run for Hypothesis 3.

Results: Based on the table above, we can see if the quality of patient care star rating is excellent which is the highest rating 5, then the foot care service and patients had less pain when moving around are excellent as well. Also, we can notice that if the patients feel better and less pain when moving around, then they receive the good foot care service with confidence around 0.7.

d) **Hypothesis 4**: There is a strong correlation between how often the health patients had to be admitted to the hospital and how often patients receiving home health care needed urgent, unplanned care in the ER (Emergency Room) without being admitted. Also, these two factors are strongly associated with the quality of patient care star rating.

The reason for choosing this group of features is that they are all about patients required acute care hospitalization. In addition, we assumed they are important factors to determine the patient care star rating. The used hyper-parameters and the output rules are summarized in the table as follows.

Input features: Admitted_hospital, Urgentcare_ER, Star_Rating	S = 0.10 C= 0.90	S = 0.05 C= 0.90	S = 0.05 C = 0.60
Rules output	No rules returned	No rules returned	('Average_Hospital',) ==> ('Good_ER',) , 0.602 ('Good_ER', 'Good_starR') ==> ('Average_Hospital',) , 0.616 ('Average_Hospital', 'Good_starR') ==> ('Good_ER',) , 0.627
Rationale for the chosen hyper-parameters	Starting threshold	No rules returned from the previous run, so we decreased the support.	Still no rules returned from the previous run, so we decreased the confidence threshold.

Table 6: Summary of the run for Hypothesis 4.

Results: Based on the output above, we can see if the hospital admittance rate is average then the ER rate is good with confidence 0.602. It indicates that there is not a very strong connection between being admitted to hospital and being admitted to ER.

e) **Hypothesis 5**: There is a strong association between outcome of care measures including patients got improved at walking, getting in and out of bed, bathing by themselves, less pain in moving around and breathing and the quality of care star rating.

The rationale behind choosing this group of features is that they all measure the outcome of care services provided by health agency. We assumed they play a critical role in determining the agency overall performance. The used hyper-parameters and the output rules are summarized in the table as follows.

Input features: InOut_bed, Lesspain_Walking, Drug_bymouth, Walking, Bathing, Breathing, Star_Rating	S = 0.10 C= 0.90	S = 0.10 C= 0.80	S = 0.10 C = 0.70
Rules output	('Good_bathing', 'Good_lessPain', 'Good_starR', 'Good_bed') ==> ('Good_walking',) , 0.962	('Average_drugMouth' , 'Good_breath', 'Good_starR') ==> ('Good_bathing',) , 0.895	('Good_drugMouth', 'Good_lessPain', 'Good_bed') ==> ('Good_starR',) , 0.744
Rationale for the chosen hyper-parameters	Starting threshold	Decreased the confidence to see if there are any additional rules returned.	Further decreased confidence to see if there are any additional rules returned.

Table 7: Summary of the run for hypothesis 5.

Results: Based on the output above, we can see if the patients got better at taking drug correctly by mouth and at getting in and out of bed, plus they had less pain when moving around, then the overall care star rating is good with confidence 0.744. It indicates that patients taking drug correctly by mouth, had less pain moving around and getting better at getting in and out of bed determine the good care star rating to some extent.

3) Results Analysis

This section discusses the insight into the results returned from the above models. According to the results returned, we can note that if the health team ensured patients received flu shot, they would be likely to make sure that patients received pneumococcal vaccine for the current season as well.

Further, it is observed that there is very strong correlation between quality care star rating and the drug education provided by the home health team. It implies that if the quality care star rating is very high, then the agency is likely doing a good job in educating patients about use of drug. It is quite surprising to see that taking drugs correctly by mouth cannot directly imply that the agency provides an excellent drug education service. Instead, only when the quality of care star rating is excellent can we see the excellent drug education service has an impact on the patients' good drug-use habit. The implication is that there may be additional factors that contribute to the patients' good drug-use habit.

In addition, it can be seen that that there is a strong relationship between quality care star rating and the provided foot care service and patients had less pain when moving around measure. It implies that if the quality care star rating is very high, then the agency is providing good foot care service and the patients had less pain when moving around. Also, there is some correlation between patients good walking outcomes and they receive the good food care service.

Then, it also indicates that there is not a very strong connection between the rate of being admitted to hospital and the rate of being admitted to ER. The result is somewhat surprising but also reasonable. Patients visit ER (Emergency Room) for acute and urgent care while go to hospitals for chronic and less urgent care. In other words, these two aspects differ from the degree of illness severity. Therefore, it is reasonable to notice there is weak connection between these two measures.

Last but not least, it shows that patients' good drug-use habit, walking outcomes, get better at getting in and out of bed are three significant measures to impact on the overall quality of care star rating. Therefore, it is suggested that health agencies could place more emphasis on these three outcome measures in order to obtain a higher overall care star rating.

5. Conclusion and Discussion

In conclusion, our project implemented supervised learning including Decision Trees and Naïve Bayes as well as unsupervised learning including Association rules to find out the important quality measures. Based on supervised learning results, the most significant quality measures ordered from the most important to the least important are Bathing, Lesspain_Walking, Flu_shot, Timely_manner, In_Out_bed, Walking, Drug_taught, Breathing and Admitted_hospital.(Please see Table A-3 in appendix for the full name of the measures). According to Association rules, Drug education has a strong correlation with the overall star rating. Also, patients' good drug-use habit, walking outcomes, get better at getting in and out of bed are three important quality measures to influence the overall quality of care star rating.

Our results provide the direction for the home health agencies to work on in the future. In particular, the health care agencies are recommended to focus on improving their bathing services, drug education and pay close attention to patients' walking experience especially check whether they have less pain when moving around among all the quality measures in order to achieve a higher overall care star rating.

References

[1] ‘What’s home health care?’ https://www.medicare.gov/what-medicare-covers/home-health-care/home-health-care-what-is-it-what-to-expect.html

[2] E. A. Madigan and O. L. Curet, "A data mining approach in home healthcare: Outcomes and service use," *BMC Health Services Research*, vol. 6, Feb. 2006.

[3] O. Jarrín, L. Flynn, E. T. Lake, and L. H. Aiken, "Home health agency work environments and hospitalizations" *Medical Care*, vol. 52, no. 10, pp. 877–883, Oct. 2014.

[4] "sklearn.tree.DecisionTreeClassifier — scikit-learn 0.18.1 documentation", *Scikit-learn.org*, 2017. [Online]. Available: http://scikit learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier. [Accessed: 03- Apr- 2017].

[5] "What's home health care? | Medicare.gov", *Medicare.gov*, 2017. [Online]. Available: https://www.medicare.gov/what-medicare-covers/home-health-care/home-health-care-what-is-it-what-to-expect.html. [Accessed: 03- Apr- 2017].

Appendix

Groups	Columns	Data Type
Basic info of health agency	Agency Name, CMS Certification Number (CCN), State, Address, City, Zipcode, Type of Ownership.	Text
Types of services offered to patients	Offers Nursing Care services	Binary (Yes/No)
	Offers Physical Therapy Services	
	Offers Occupational Therapy Services	
	Offers Speech Pathology Services	
	Offers Medical Social Services	
	Offers Home Health Aide Services	
Star Rating	Quality of patient care star rating	Discrete (from 1 - 5 with 0.5 increment)
Process of care measures	How often the agency gave patient care in a timely manner	Numeric (from 0 - 100)
	How often the agency provided patient/caregiver drug education on all medications	
	How often the agency ensured patients received flu and vaccine for the current season	
	With diabetes, how often the home health team got doctor's orders, gave foot care, and taught patients about foot care	
	How often the home health team checked patients for depression	
	How often the home health team checked patients' risk of falling	
Outcome of care measures	How often the patientg got better at walking or moving around	Numeric (from 0 - 100)
	How often the patients got better at getting in and out of bed	
	How often the patients got better at bathing themselves	
	How often the patients had less pain when moving around	
	How often the patients got better at taking their drugs correctly by mouth	
	How often the patients' breathing improved	
	How often home health patients had to be admitted to the hospital	
	How often home health patients needed urgent, unplanned care in the ER without being admitted	
Patient experience of care measures	HHCAHPS Survey Summary Star Rating	Discrete (from 1 - 5)
	How often the home health team gave care in a professional way	Percent of patients measures is of numeric type (from 0 - 100) Star Rating is of discrete type (from 1 - 5)
	How well did the home health team communicate with patients	
	Did the home health team discuss medicines, pain, and home safety with patients	
	Star Rating for how patients rated overall care from agency	
	Percent of patients who would recommend the home health agency to friends and family	

Table A-1: Data features summary

Note that in the last group of patient experience of care measures, the professional, communication, discussing medicines measures for the home health team consist of two columns. For example, for the communication measure, the data includes the star rating for health team communicated well with them and percent of patients who reported that their home health team communicated well with them.

Feature Name	Importance
Bathing	0.32717387
Lesspain_Walking	0.1397533
Flu_shot	0.08778397
Timely_manner	0.08077286
InOut_bed	0.0770425
Walking	0.06383152
Drug_taught	0.05065568
Breathing	0.04907809
Admitted_hospital	0.04733465
Urgentcare_ER	0.00875753
Drug_bymouth	0.00865971
Depression_check	0.00853156
Recom_percent	0.00776014
Foot_care	0.00712703
pneumococcal_vaccine	0.00711932
Rating9or10_percent	0.00443492
Medicine_percent	0.00437708
Professional_percent	0.00397165

Feature Name	Importance
Falling_check	0.00349977
Communicate_percent	0.00323177
Communicate_rating	0.00272705
Overall_rating	0.0017042
Medicine_rating	0.00132452
Medical_Social_no	0.00105458
Survey_Sumrating	0.0008086
Professional_rating	0.00076259
Medical_Social_yes	0.00072154
Nursing_Care_yes	0
Physical_therapy_no	0
Physical_therapy_yes	0
Occupational_Therapy_no	0
Occupational_Therapy_yes	0
Speech_pathology_no	0
Speech_pathology_yes	0
HomeHealth_Aide_no	0
HomeHealth_Aide_yes	0

Table A-2: Feature importance summary

Coding Name	Original Dataset Name
Bathing	How often patients got better at bathing
Lesspain_Walking	How often patients had less pain when moving around
Flu_shot	How often the home health team determined whether patients received a flu shot for the current flu season
Timely_manner	How often the home health team began their patients care in a timely manner
In_Out_bed	How often patients got better at getting in and out of bed
Walking	How often patients got better at walking or moving around
Drug_taught	How often health team taught patients about their drugs
Breathing	How often patients' breathing improved
Admitted_hospital	How often home health patients had to be admitted to the hospital

Table A-3: Feature full names

Column Coding Name	Importance
Bathing	0.47227525
Lesspain_Walking	0.18028254
Flu_shot	0.09911997
Timely_manner	0.08221419
In_Out_bed	0.08023729
Walking	0.03705954
Drug_taught	0.03125152
Breathing	0.01392587
Admitted_hospital	0.00363384

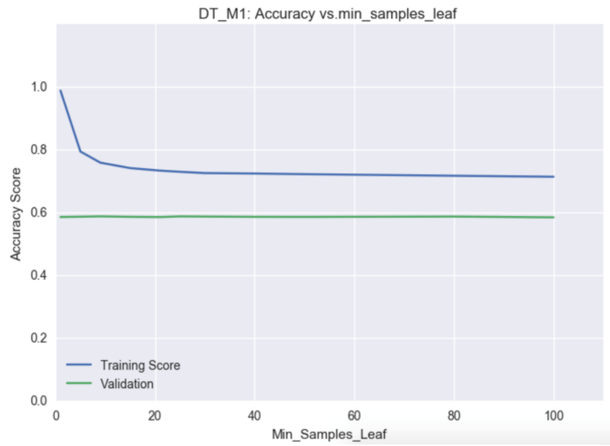
Table A- 4: Model 4 feature importance

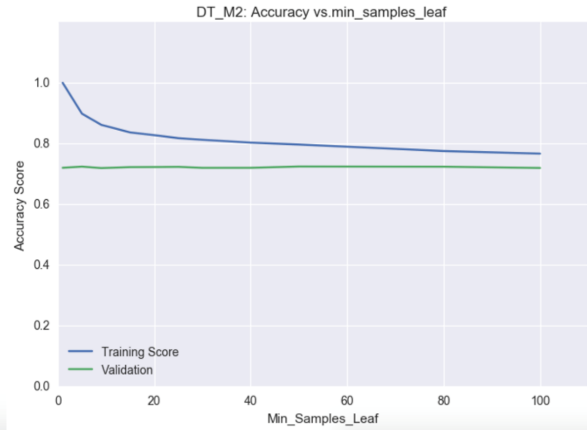
```
Out[3]: State
                                                    0
        CMS Certification Number (CCN)
                                                    0
        Provider Name
                                                    0
        Address
                                                    0
        City
                                                    0
        Zip
                                                    0
        Type of Ownership
                                                    0
        Offers Nursing Care Services
                                                    0
        Offers Physical Therapy Services
                                                    0
        Offers Occupational Therapy Services
                                                    0
        Offers Speech Pathology Services
                                                    0
        Offers Medical Social Services
                                                    0
        Offers Home Health Aide Services
                                                    0
        Quality of patient care star rating
                                                 2838
        How often the home health team began their patients' care in a timely manner
                                                 2069
        How often the home health team taught patients (or their family caregivers) about their drugs
                                                 2084
        How often the home health team checked patients' risk of falling
                                                 2289
        How often the home health team checked patients for depression
                                                 2087
        How often the home health team determined whether patients received a flu shot for the currnet flu season
                                                 2467
        How often the home health team made sure that their patients have received a pneumococcal vaccine (pneumonia shot)
                                                 2118
        With diabetes, how often the home health team got doctor's orders, gave foot care, and taught patients about foot car
```

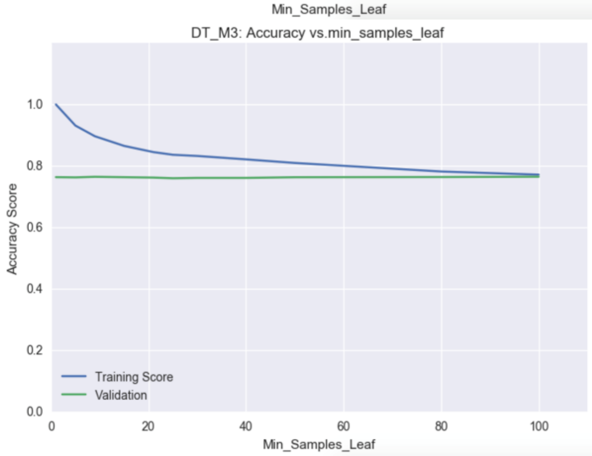
Figure A-1: Missing data summary before cleaning

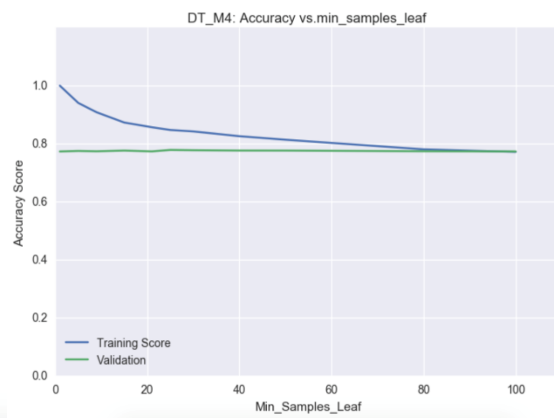
Offers Nursing Care Services	Offers Physical Therapy Services	Offers Occupational Therapy Services	...	HHCAHPS Survey Summary Star Rating	Star Rating for health team gave care in a professional way	Percent of patients who reported that their home health team gave care in a professional way	Star Rating for health team communicated well with them	Percent of patients who reported that their home health team communicated well with them	Star Rating team discussed medicines, pain, and home safety
Yes	Yes	Yes	...	4	4	90	4	88	3
Yes	Yes	Yes	...	5	5	91	5	89	4
Yes	Yes	Yes	...	5	5	93	5	92	5

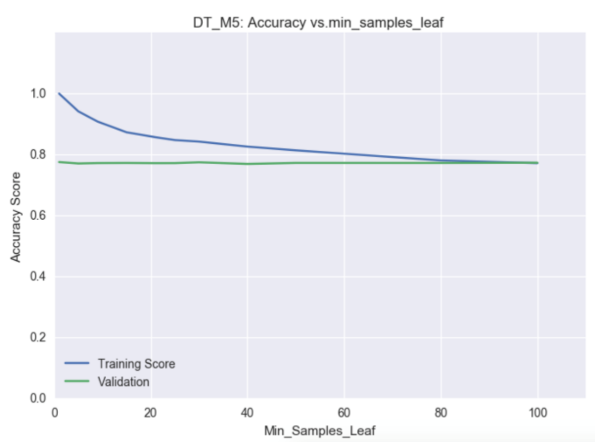
Figure A-2: Labelled training dataset

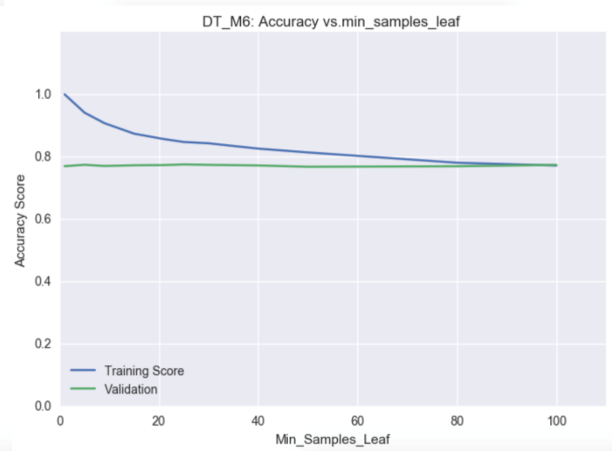

DT_M1: Accuracy vs.min_samples_leaf
Accuracy Score
Min_Samples_Leaf
Training Score
Validation


DT_M2: Accuracy vs.min_samples_leaf
Accuracy Score
Min_Samples_Leaf
Training Score
Validation


DT_M3: Accuracy vs.min_samples_leaf
Accuracy Score
Min_Samples_Leaf
Training Score
Validation


DT_M4: Accuracy vs.min_samples_leaf
Accuracy Score
Min_Samples_Leaf
Training Score
Validation


DT_M5: Accuracy vs.min_samples_leaf
Accuracy Score
Min_Samples_Leaf
Training Score
Validation


DT_M6: Accuracy vs.min_samples_leaf
Accuracy Score
Min_Samples_Leaf
Training Score
Validation

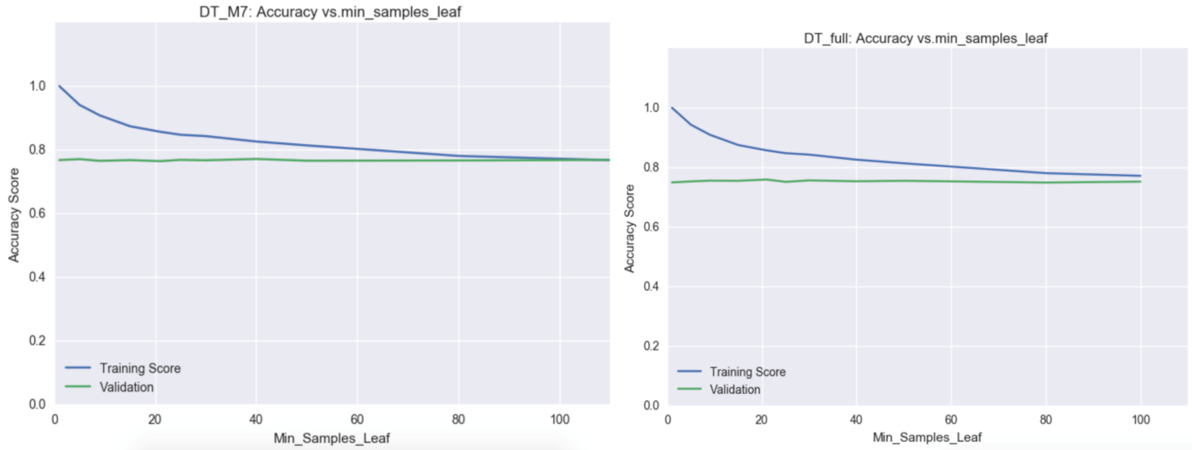
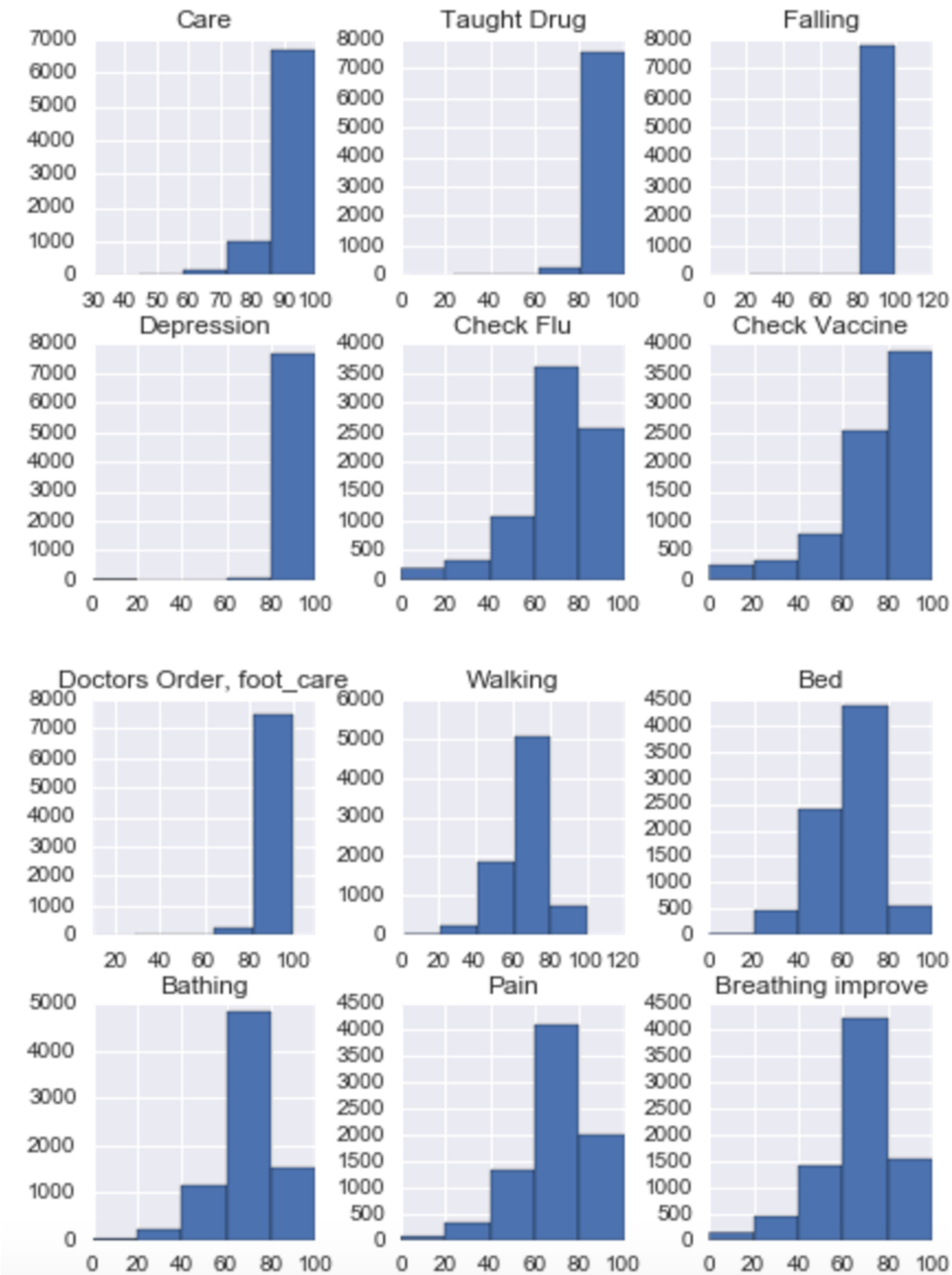


Figure A-3: Hyper-parameter selection



Care
Taught Drug
Falling
Depression
Check Flu
Check Vaccine
Doctors Order, foot_care
Walking
Bed
Bathing
Pain
Breathing improve

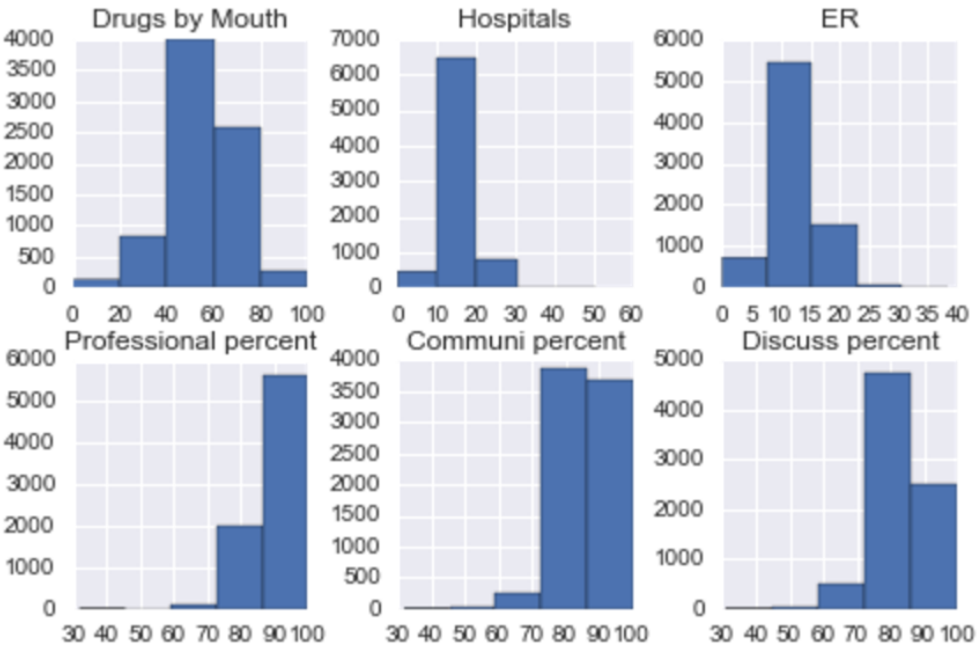


Figure A-4: Histograms of numeric variables