Using Clustering to Predict Next Year Admission with MEPS Data

Prepared for
The Final Project of STA 588

By
Tatiana Romanchishina
Smita Sukhadeve
5/16/2015

Introduction

The issue of hospital readmission is now front and center in the national conversation. Higher rate of hospital readmission denotes poor quality of the health care service. As per American Hospital Association, nearly 34 million patients are admitted to the hospital each year. There remains continuing interest in identifying patients at risk of future hospital admissions to target care coordination and management strategies that may potentially reduce future inpatient expenditures. The Hospital Readmission reduction Program is in effect from october 2012 which focuses on implementing strategies to reduce unnecessary hospital readmissions.

A number of research and studies are carried out in the past to identify possible causes of hospital readmissions and to predict the patients with high risk of readmissions using healthcare data (Soley-Bori, 2015). Early identification of such population will not only help in reduction of unnecessary admission expenditure but will also improve healthcare quality. This study is an attempt to identify patients with the high risk of admission in the next year using previous year health, demographic and financial data.

The dataset used during this study is taken from Medical Expenditure Panel Survey (MEPS, 2016). It is a set of surveys of individuals and their medical providers across the United States. We use the longitudinal file from 2012-2013 which contains more than 3500 variables. This dataset contains health care, health status, utilization and expenditures information for the persons who participated in the Medical Panel Survey for the two-year period. Subsequent section focuses on detailed description of variables used during study. We use K-Means Clustering to look into the association among the selected quantitative variables. We demonstrate the use of Logistic Regression and Random Forests classifiers to predict the next year hospital admissions and compare their results.

Dataset Description

The dataset we used in this study contains 17932 observations of 3496 variables many of which are quantitative variables. We include only quantitative variables in this study because we plan to use K-Means clustering which limits us to using only quantitative variables. We plan to predict whether a patient will be admitted in 2013 using the data from year 2012. Some of our numerical variables describe different types of income and expenses of the participants such as total income, family's total income, medical expenses, family practitioner visits charges over the year, amount of money spent on medicines, hospital admission charges, hospital inpatient and outpatient charges, etc (MEPS Documentation, 2016).

Figure 1 shows distribution of six quantitative variables (scaled) from our dataset. The data in our dataset describe health-related expenses and events. Only a small portion of our participants had any healthcare events, while the overwhelming majority have no associated events or expenses. Therefore, as we can see in Figure 1, some of the distributions are skewed.

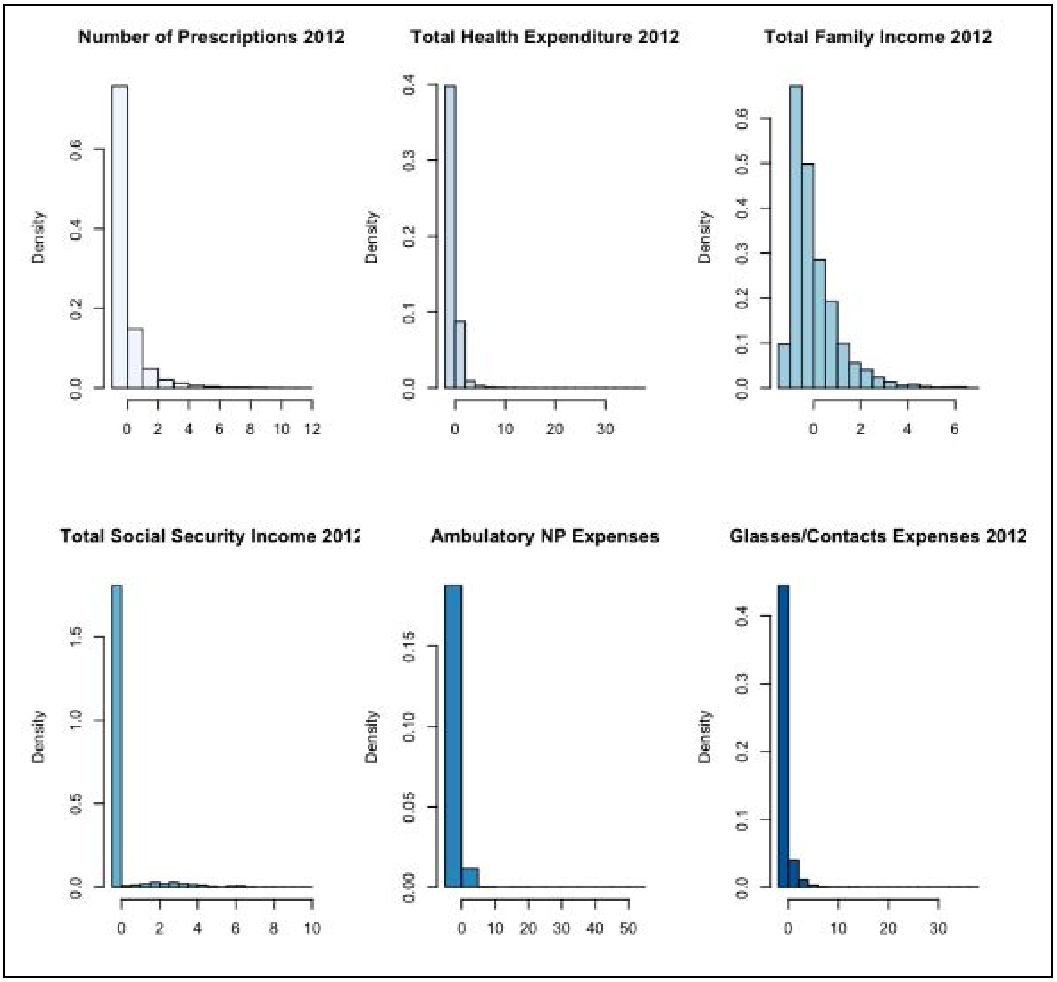

Figure 1. Distribution of some of the quantitative variables.

Missing values in the dataset stand for different possible scenarios: participant skipped or did not answer the question, for example. We plan to identify and remove variables that have more than 500 missing values. Then we intend to remove all the rows that contain any missing values. From the remaining data we plan to select all the numerical variables that pertain to 2012. We want to perform the best subset selection procedure to find an optimal subset, that will be used as the predictors in our classifiers.

We also plan to use K-Means clustering on the final subset of quantitative variables. As a result of the clustering, each observation will be assigned a ClusterID that will show which cluster it belongs to. Thus, we will add one categorical variable to our subset of numerical variables. This new variable ClusterID will have the number of classes, or levels, equal to the number of clusters found by the K-Means clustering algorithm.

The only other categorical variable in our dataset is the response. The response is also the only variable in our work dataset that pertains to year 2013. We derive the response from the variable in the MEPS longitudinal file, that represents the number of nights spent in a hospital in 2013. We transform this variable into a binary one by making all non-zero counts equal to 1. Thus, being admitted to the hospital in 2013 is coded as 1, and not being admitted in 2013 is coded as 0. As we can see in Figure 2 the distribution of the response variable is imbalanced, which is something we need to keep in mind when evaluating our models and their prediction results.

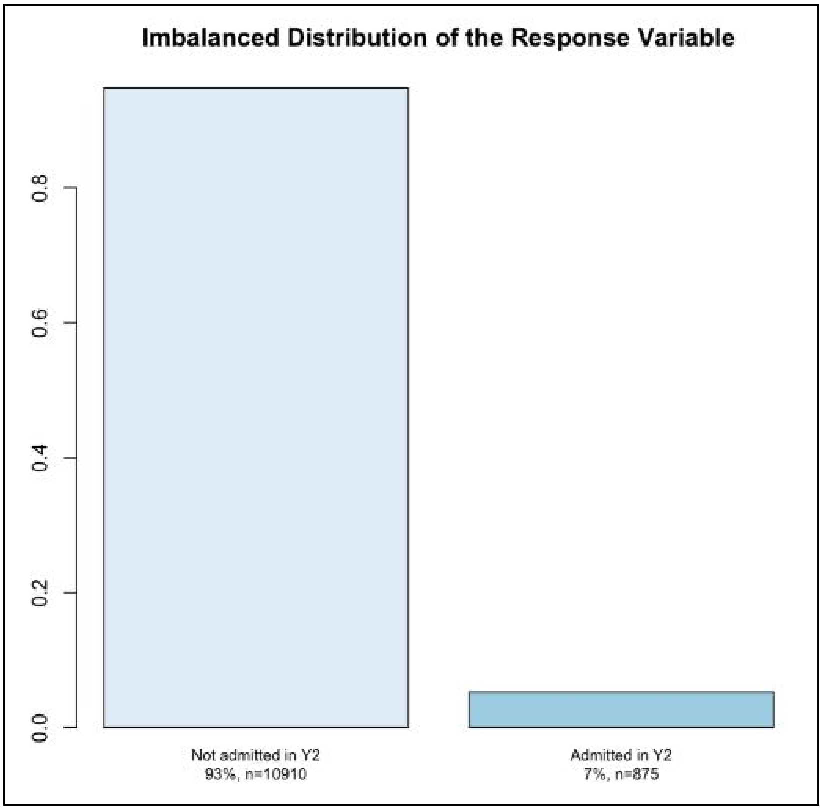

Figure 2. Imbalanced distribution of the response variable

Methodology

The longitudinal dataset used in this study consists of 3496. After removing variables with that contain missing values. Then we focus on only quantitative variables from 2012, which is still a great number of variables. For the model interpretability, it is necessary to identify irrelevant variables and remove them to get better estimates. We will use the stepwise selection approach to find the best subsets of the predictors to estimate our response. We use a hybrid approach, a combination of forward and backward selection methods, which uses all the predictors to choose an optimal subset that is related to the response. It uses Akaike Information Criteria (AIC) with minimum information loss. The best model has the lowest AIC, given by:

$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$ (James et al., 2013)

where n is the number of observations, $\hat{\sigma}^2$ is an estimate of the variance of error associated with each response measurement, *RSS* is the residual sum of squares and *d* is the number of explanatory variables.

In a subsequent section, we will demonstrate K-Means clustering approach for partitioning the data into distinct clusters. In order to perform K-Means clustering, we must first identify the optimal number of clusters K. Then the K-Means algorithm assigns each observation to one of the K cluster. We will use function kmeans() from the R *cluster* package. The 'Haritan-Wong' K-Means algorithm will be used to partition the observation into K groups such that sum of squares of the observations of their assigned cluster center is minimized. We will try using different values of K to select the best value based on the minimum within-cluster sum of squares. The within-cluster variation for cluster C_k is a measure of $W(C_k)$ of the amount by which each observation within a cluster differ from each other. Minimum value of within-cluster variation indicates good clustering.

Each observation will be assigned to cluster C_k with smallest value of within cluster sum of squares $W(C_k)$, expressed by the squared Euclidean distance (James, 2013):

$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{i'j})^2$, where $|C_k|$ is the number of observations in the kth cluster.

Next, we will introduce the cluster ID as one of predictors for further modelling. First, we use a Logistic Regression classifier to predict our binary response variable. The performance of the model is evaluated based on the sensitivity and specificity. Sensitivity denotes the proportion of correctly identified 1's, which in our case denotes getting admitted to the hospital next year. Whereas, specificity is the proportion of correctly identified 0's, which denotes here not being admitted to the hospital next year. We would like to see high sensitivity as we are interested in finding these generally rare cases. We look at different probability cut-off points to find an optimal point.

Then, we use the tree-based Random Forests classifier. We will also use different cut-off probabilities to maximize the sensitivity of prediction, as we are primarily interested in finding the participants who are going to be admitted next year. We will use ROC curves to measure the prediction power of both models. ROC curve is the plot between (1-specificity, sensitivity). Each point on the ROC curve denotes the cut-off probability. Area Under the ROC Curve will be the measure of choosing best model out of the two. As a control measure, we will compare both to a random guess models.

Results

Variable Selection

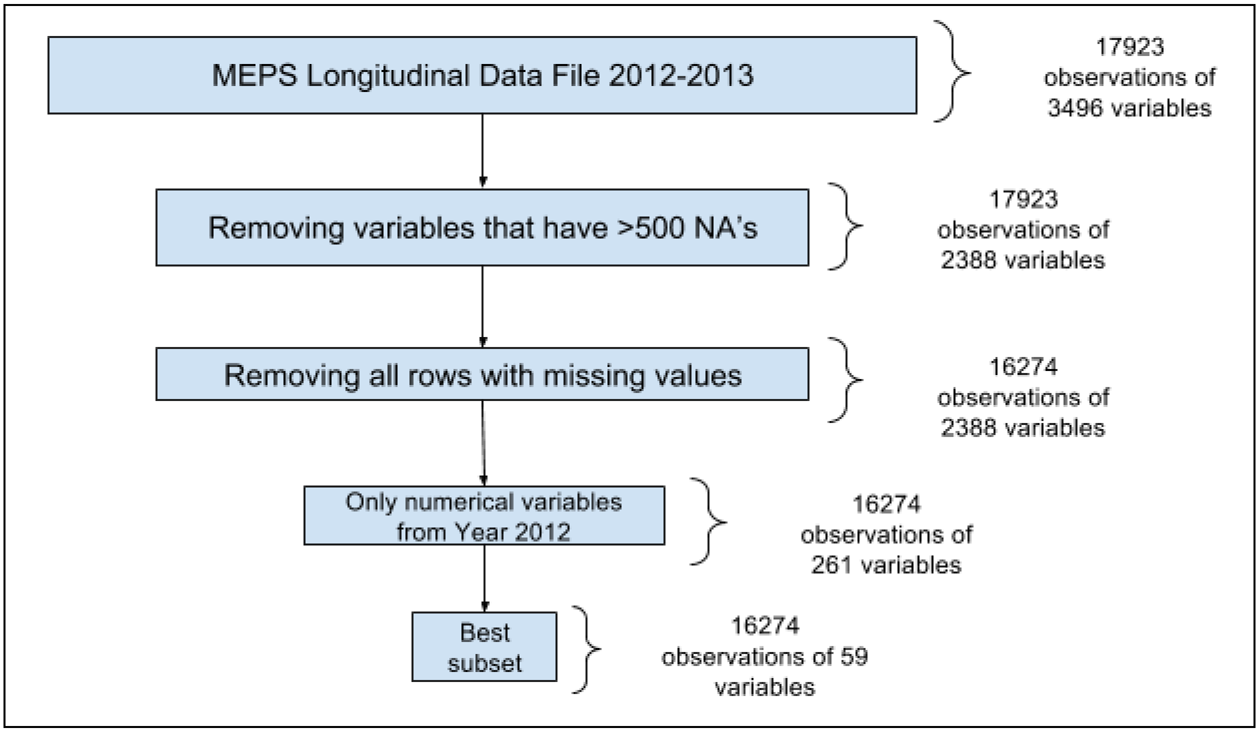
MEPS Longitudinal Data File 2012-2013
17923 observations of 3496 variables
Removing variables that have >500 NA's
17923 observations of 2388 variables
Removing all rows with missing values
16274 observations of 2388 variables
Only numerical variables from Year 2012
16274 observations of 261 variables
Best subset
16274 observations of 59 variables

Figure 3. Dataset Preparation.

Figure 3 summarizes the process of selecting the variables for the final models. We use the MEPS longitudinal data file from 2012-2013 in this study, thus we start out with 17923 observations of 3496 variables. Then we identify the variables that contain more than 500 missing values and remove them, which leaves us with 2388 variables. We proceed by removing all observations with missing values, which brings the number of observations down to 16274. Then we subset all numerical variables that pertain to year 2012, since we want to predict admission in year 2013. From the remaining 261 variables we select the best subset of 59 variables, using R functions for stepwise selection procedures. We describe the variables in more detail as we learn of their significance or importance in the following sections.

Clustering (Withinness)

We use the KMeans algorithm for clustering on the 59 selected numerical variables to create a new categorical variable to add to the final model. First, we decide on the best number of clusters for these data by examining the within groups sum of squares for different numbers of clusters. We want to see a steady decrease in the within groups sum of squares, that would be defined by one turning point, after which there is a definite consistency in the measure. As we can see in Figure 4 there is no clear elbow or bend that would indicate the optimal number of clusters.

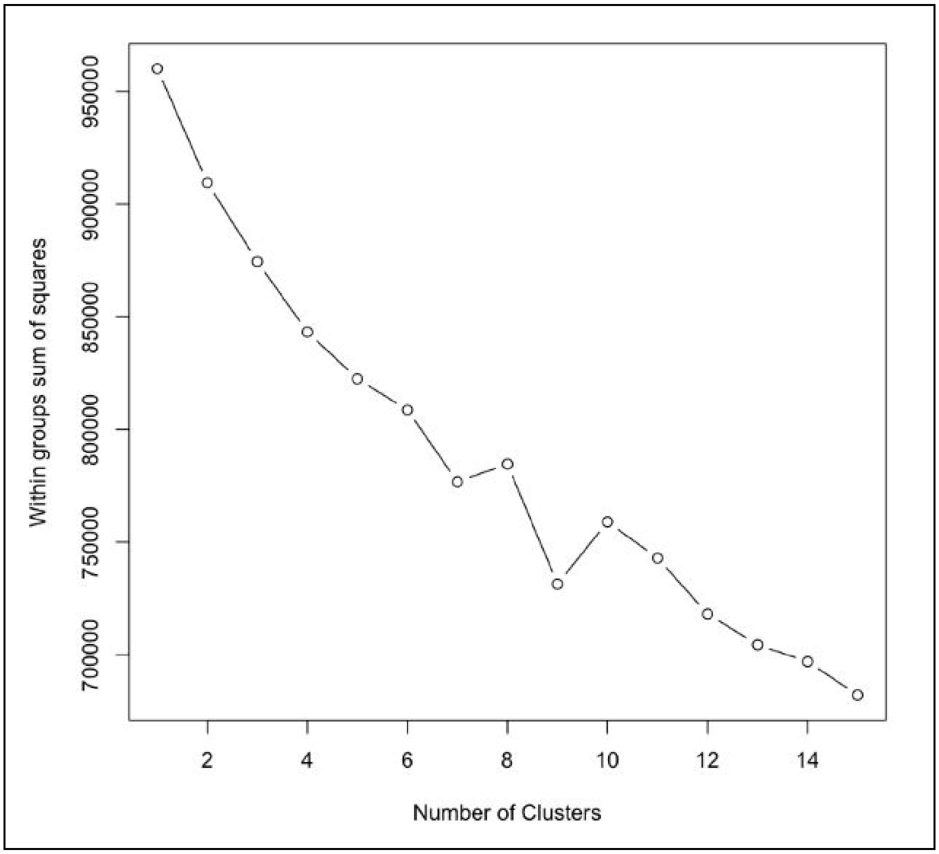

Figure 4. Within groups sum of squares for different number of clusters.

It seems that the within groups sum of squares decreases as the number of clusters increases. We considered every number from 2 to 15 as the number of clusters, and the lowest within groups sum of squares was recorded for 15 clusters. However, our plan is to turn the results of this clustering into a categorical variable, and we do not want it to have too many levels. Thus, we decide to perform the K-Means clustering algorithm with 7 clusters. Figure 5 demonstrates the results of K-Means clustering, visualized in terms of the total number of prescriptions in year 2012 and total health expenditure for year 2012.

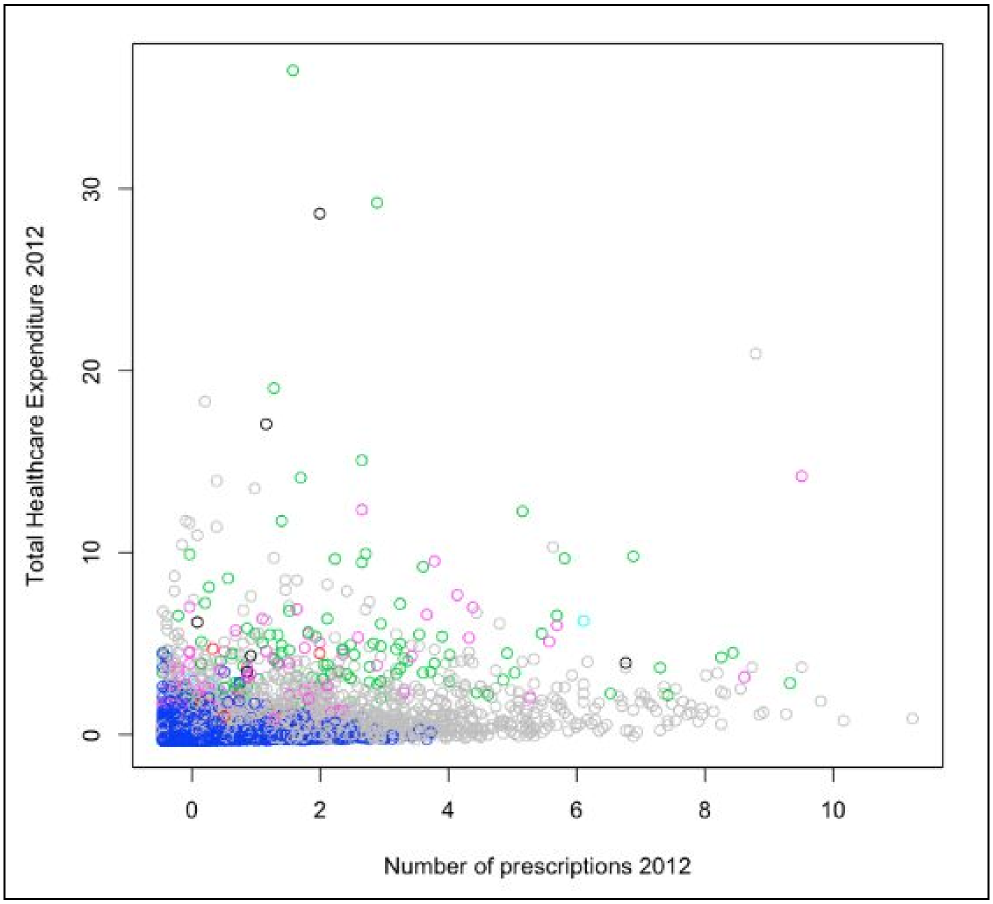

Figure 5: Results of K-Means Clustering with 7 clusters on two variables.

Classification

Using a subset of 59 quantitative variables and the categorical variable ClusterID derived from the results of clustering, we fit a Logistic Regression and a Random Forests model on 67% of the data. We look at what variables are found significant. We test the resulting models on the remaining 33% and look at their sensitivity and specificity levels to select the best one. In this study, we are more interested in identifying the participants that will be admitted, than those that will not. We use sensitivity and specificity to find a probability cut-off point that maximizes sensitivity, while still maintaining high specificity. We select the model that achieved better results in maintaining the balance between sensitivity and specificity.

Logistic Regression

The logistic regression model achieved sensitivity of 0.47 and specificity of 0.85 at 0.05 probability that the participant will be admitted in the next year. The sensitivity rapidly decreases when we increase the probability cut-off point. Thus, 0.05 probability seems to be the optimal cut-off point. Table 1 shows the sensitivity and specificity values for different probability values.

Probability Cut-off	Sensitivity	Specificity
0.05	0.47	0.85
0.1	0.35	0.92
0.15	0.26	0.96
...	...	...
> 0.95	0.00	1.00

Table 1. Sensitivity and specificity of the Logistic Regression model.

The following variables, used to fit the model, were found significant:

- Cluster ID, level 5, *p = 0.001*,
- Number of prescriptions in Year 1, *p = 7.96e-06,*
- Total healthcare expenditure in year 1, *p = 0.002,*
- Participant's Social Security income, *p = 0.0003,*
- Non-Physician medical charges, *p = 0.03,*
- Inpatient Hospital Medicaid amount, *p = 0.03,*
- ER Doctor visit charges, *p = 0.003,*
- ER Other Unclassified amount, *p = 0.016,*
- Participant's public assistance in year 1, *p = 0.018,*
- Inpatient Hospital Private and Tricare amount, *p = 0.032,*
- Participant's trust/rent income in year 1, *p = 0.005,*
- Family's total income in year 1, *p = 0.02,*
- Total outpatient doctor expenses in year 1, *p = 0.007,*
- Outpatient Doctor visit charges, *p = 0.013,*
- Participant's total outpatient provider expenditure, *p = 0.018,*
- Outpatient Doctor Visits Private and Tricare, *p = 0.016,*
- All Outpatient Doctor Visits Private Insurance, *p = 0.019.*

Random Forests

The random forests model performed better than the logistic regression model, achieving sensitivity of 0.6 while maintaining specificity of 0.76 at 0.05 probability of being admitted next year. Table 2 shows the sensitivity and specificity results for different probability cut-off points.

Probability Cut-off	Sensitivity	Specificity
0.05	0.60	0.76
0.1	0.44	0.87
0.15	0.33	0.93
...	...	...
> 0.60	0.00	1.00

Table 2. Sensitivity and specificity of the Random Forests model.

A random forest model is hard to interpret due to its nature, but it provides information about which variables contributed most to classification. The following variables were found most important by the Random Forests model (in descending order of importance):

- Total healthcare expenditure in year 1
- Family’s total income in year 1
- Total amount paid by participant
- Total Prescribed medicines in year 1
- Total Office based provider visits
- Total office based non-Dr expenditure
- Emergency room Doctor’s charges in year 1
- Office-based medicare charges
- Participant’s Social Security income
- Outpatient department Visit charges
- Total Glasses/contact lens expenditure in year 1
- ClustID
- Total amount paid by other sources
- Emergency room medical Aid amount
- Other equipment and supply charges

We plot ROC curves for both classification models and compare them, while also using a random guess estimate as control. We can see that both models are undoubtedly better than random guess. We can also tell, that even though the models are quite similar the random forests model is more accurate.

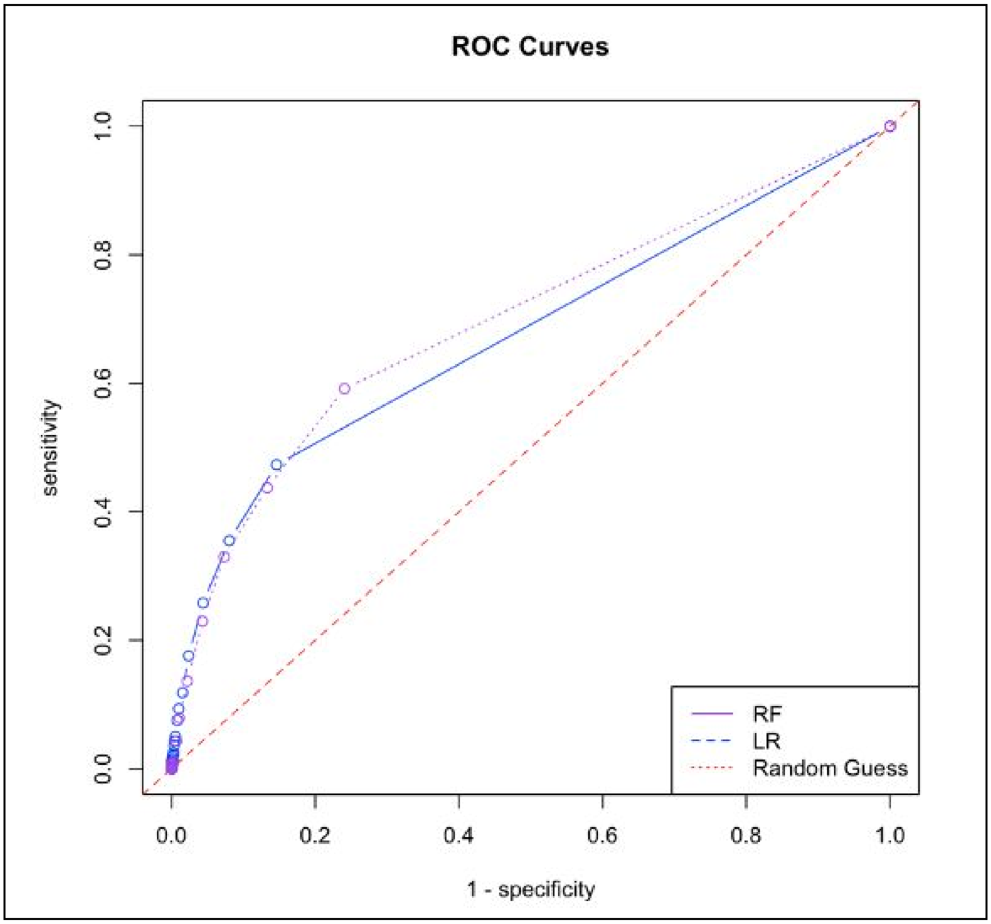

Figure 6. ROC curves of the Logistic Regression and Random Forests models.

Discussion

As discussed in the Methodology Section, the Random Forests model has proven its efficiency over the Logistic Regression model. In Random Forests, we grow multiple trees and take the mode of the most popular class aggregating all the decision trees. Random selection of input variables during tree growing process reduces the correlation between the trees and further improves the accuracy of result. Obviously, results of logistic regression are much more interpretable than Random forest. However, the variable importance information obtained using the Gini index function helps us understand what the important predictors are that contribute towards the increasing the likelihood of hospital admission. Moreover, the dataset is highly imbalanced and finding the rare case is very challenging. We can increase the accuracy of our models by using different techniques, such as over-sampling of the rare case, sample stratification, etc.

Conclusions

In this study we focused on the problem of predicting next year admission using information from previous year. We used the dataset derived from the 2012-2013 MEPS longitudinal file. We reduced the number of variables from 3496 to 59 quantitative variables. Then we performed K-Means clustering to assign observations to seven clusters, which we had estimated to be an optimal number of clusters. We used these results to create a new categorical variable ClusterID which we added to the predictors.

We fit a Logistic Regression and a Random Forests model to predict next year admission. We looked at sensitivity and specificity of their prediction, as we are interested in having high sensitivity while maintaining good specificity. We found the the Random Forests model we created achieved higher sensitivity overall, achieving the best result 0.05 probability.

References

- James, G. et al. (2013). An introduction to statistical learning (Vol. 112). New York: Springer.

- Medical Expenditure Panel Survey (MEPS) Documentation. (2016). http://meps.ahrq.gov/mepsweb/data_stats/download_data/pufs/h156/h156cb.pdf

- Dr. Cheng Peng. (2016) Logistic Regression Model - STA 588. University of Southern Maine.

- Dr. Cheng Peng. (2016) ROC Analysis on Logistic Classifier - STA 588. University of Southern Maine.

- Dr. Cheng Peng. (2016) K-mean Clustering - STA 588. University of Southern Maine.

- Soley-Bori, M. et al. (2015). Functional Status and Hospital Readmissions Using the Medical Expenditure Panel Survey. Journal of general internal medicine, 30(7), 965-972. https://www.researchgate.net/profile/Marina_Soley-Bori/publication/272513564_Function

al_Status_and_Hospital_Readmissions_Using_the_Medical_Expenditure_Panel_Survey
/links/54f09ea70cf2b36214aab2d2.pdf

Appendix A: Dataset summary (scaled)

```
'data.frame':	16274 obs. of  61 variables:
 $ ClustID : Factor w/ 7 levels "1","2","3","4",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ RXTOTY1 : num  -0.453 -0.453 -0.453 1.634 -0.453 ...
 $ TOTEXPY1: num  -0.323 -0.323 -0.315 -0.238 -0.309 ...
 $ SSECPY1X: num  -0.289 -0.289 -0.289 -0.289 -0.289 ...
 $ ERTMCRY1: num  -0.0703 -0.0703 -0.0703 -0.0703 -0.0703 ...
 $ OPOTCHY1: num  -0.0689 -0.0689 -0.0689 -0.0689 -0.0689 ...
 $ IPDMCDY1: num  -0.0782 -0.0782 -0.0782 -0.0782 -0.0782 ...
 $ ERDTCHY1: num  -0.166 -0.166 -0.166 -0.166 -0.166 ...
 $ OBDMCRY1: num  -0.133 -0.133 -0.133 -0.133 -0.133 ...
 $ OBDWCPY1: num  -0.0284 -0.0284 -0.0284 -0.0284 -0.0284 ...
 $ OBCPRVY1: num  -0.0646 -0.0646 -0.0646 -0.0646 -0.0646 ...
 $ OTHSLFY1: num  -0.038 -0.038 -0.038 -0.038 -0.038 ...
 $ ERTOSRY1: num  -0.0498 -0.0498 -0.0498 -0.0498 -0.0498 ...
 $ PUBPY1X : num  -0.0853 -0.0853 -0.0853 -0.0853 -0.0853 ...
 $ AMTTCHY1: num  -0.058 -0.058 -0.058 -0.058 -0.058 ...
 $ IPFPTRY1: num  -0.0738 -0.0738 -0.0738 -0.0738 -0.0738 ...
 $ TOTSLFY1: num  -0.309 -0.309 -0.253 -0.273 -0.305 ...
 $ RXSTLY1 : num  -0.0413 -0.0413 -0.0413 0.1211 -0.0413 ...
 $ TOTOSRY1: num  -0.0694 -0.0694 -0.0694 -0.0694 -0.0694 ...
 $ OPFSLFY1: num  -0.0873 -0.0873 -0.0873 -0.0873 -0.0873 ...
 $ ERDMCDY1: num  -0.0938 -0.0938 -0.0938 -0.0938 -0.0938 ...
 $ HHATCHY1: num  -0.0651 -0.0651 -0.0651 -0.0651 -0.0651 ...
 $ HHAEXPY1: num  -0.062 -0.062 -0.062 -0.062 -0.062 ...
 $ VISEXPY1: num  -0.221 -0.221 -0.221 -0.221 -0.221 ...
 $ OTHTCHY1: num  -0.044 -0.044 -0.044 -0.044 -0.044 ...
 $ TRSTPY1X: num  -0.07 -0.07 -0.07 -0.07 -0.07 ...
 $ FAMINCY1: num  0.447 0.447 0.447 -0.629 -0.629 ...
 $ OBTOTVY1: num  -0.429 -0.429 -0.429 0.152 -0.313 ...
 $ OBCSLFY1: num  -0.087 -0.087 -0.087 -0.087 -0.087 ...
 $ OBOEXPY1: num  -0.162 -0.162 -0.162 -0.162 -0.162 ...
 $ AMNMCRY1: num  -0.0325 -0.0325 -0.0325 -0.0325 -0.0325 ...
 $ OBTTCHY1: num  -0.0741 -0.0741 -0.0741 -0.0741 -0.0741 ...
 $ OBVVAY1 : num  -0.0495 -0.0495 -0.0495 -0.0495 -0.0495 ...
 $ TOTWCPY1: num  -0.0473 -0.0473 -0.0473 -0.0473 -0.0473 ...
 $ OPPEXPY1: num  -0.069 -0.069 -0.069 -0.069 -0.069 ...
 $ OPOPRVY1: num  -0.0822 -0.0822 -0.0822 -0.0822 -0.0822 ...
 $ SSIPY1X : num  -0.156 -0.156 -0.156 -0.156 -0.156 ...
 $ TOTOPUY1: num  -0.0246 -0.0246 -0.0246 -0.0246 -0.0246 ...
 $ IPDMCRY1: num  -0.0736 -0.0736 -0.0736 -0.0736 -0.0736 ...
 $ IPFMCRY1: num  -0.0766 -0.0766 -0.0766 -0.0766 -0.0766 ...
 $ DVOTCHY1: num  -0.0949 -0.0949 -0.0949 -0.0949 -0.0949 ...
 $ OBTPTRY1: num  -0.0458 -0.0458 -0.0458 -0.0458 -0.0458 ...
 $ AMTPRVY1: num  -0.0508 -0.0508 -0.0508 -0.0508 -0.0508 ...
 $ OPVMCDY1: num  -0.064 -0.064 -0.064 -0.064 -0.064 ...
 $ OPSPTRY1: num  -0.0835 -0.0835 -0.0835 -0.0835 -0.0835 ...
 $ TOTVAY1 : num  -0.0419 -0.0419 -0.0419 -0.0419 -0.0419 ...
 $ OPDEXPY1: num  -0.128 -0.128 -0.128 -0.128 -0.128 ...
 $ OPTPTRY1: num  -0.0974 -0.0974 -0.0974 -0.0974 -0.0974 ...
 $ OPTPRVY1: num  -0.0961 -0.0961 -0.0961 -0.0961 -0.0961 ...
 $ OPOPTRY1: num  -0.0836 -0.0836 -0.0836 -0.0836 -0.0836 ...
 $ OBVTRIY1: num  -0.0449 -0.0449 -0.0449 -0.0449 -0.0449 ...
 $ OTHEXPY1: num  -0.0464 -0.0464 -0.0464 -0.0464 -0.0464 ...
 $ OBDOPRY1: num  -0.0428 -0.0428 -0.0428 -0.0428 -0.0428 ...
 $ OBVOPRY1: num  -0.0405 -0.0405 -0.0405 -0.0405 -0.0405 ...
 $ DIVDPY1X: num  -0.0823 -0.0823 -0.0823 -0.0823 -0.0823 ...
 $ OPFTCHY1: num  -0.125 -0.125 -0.125 -0.125 -0.125 ...
 $ OBCTCHY1: num  -0.095 -0.095 -0.095 -0.095 -0.095 ...
 $ OBCEXPY1: num  -0.0904 -0.0904 -0.0904 -0.0904 -0.0904 ...
 $ OBNSLFY1: num  -0.0733 -0.0733 -0.0733 -0.0733 -0.0733 ...
 $ AMNEXPY1: num  -0.0638 -0.0638 -0.0638 -0.0638 -0.0638 ...
 $ Response: Factor w/ 2 levels "NotAdmittedY2",..: 1 1 1 1 1 1 1 1 1 1 …
```

(MEPS documentation)

Appendix B: Technical Section

- Logistic Regression estimates the probability P(Y =1| $(\vec{X})$) by using the following formula:

$$log\left(\frac{p(\vec{x})}{1-p(\vec{x})}\right) = \beta_0 + \sum_{i=1}^{p} \beta_j X_j$$

Where j = 1, 2 .. .. p and left hand side is called as logit.

Above formula is transformed in terms of $P(\vec{X})$ to estimate the probabilities.

- For the classification problem, we use Cut-off probability π_0 to decide the class boundaries given by the following formula:

$$\pi_0 = \frac{exp\ (\hat{\beta}_0 + \hat{\beta}_1\, x_0 + \ .... \ + \hat{\beta}_p\, x_p)}{1 + exp\ (\hat{\beta}_0 + \hat{\beta}_1\, x_0\ _{.....} + \hat{\beta}_p x_p)}$$

- For Random Forest algorithm finds the class for the observation ‘x’ using following formula:

$$\hat{C}\,^{B}_{rf}(x) \ = majority\ vote\ \left\{\hat{C}_b(x)\right\}\,^{B}_{1}$$

Where C(x) denotes the class assigned to the observation x

- The Random Forests algorithm uses Gini Index as a measure of total variance across the K classes for model evaluation:
  $G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$, where $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class (James, 2015).