Multiple Sclerosis Statistics Analysis

Kim Brooks, Vibhuti Chandna, Dong Hyun Jeon, Alexey Luchinsky

Bowling Green State University

Abstract

This paper is devoted to the analysis of Multiple Sclerosis pattern in the United States. To perform the analysis, the data provided by Medical Expenditure Panel Survey was used. Methods such as Decision Tree, Logistic Regression, Neural Network, k-Nearest Neighbors, Random Forest, Adaptive Boosting, and Linear Regression were implemented to predict the probability of a person being diagnosed with MS and determine which demographic factors are important for answering this question. The data set was imbalanced, so undersampling and oversampling methods were used to get a more balanced data set for the models. Each model's performance was evaluated with various cut off values, and the best result was obtained from Random Forest, with respect to the *kappa* value.

Contents

1. Introduction

Multiple sclerosis (MS), also known as encephalomyelitis disseminata, is a demyelinating disease in which the insulating covers of nerve cells in the brain and spinal cord are damaged [1] This damage disrupts the ability of parts of the nervous system to transmit signals, resulting in a range of signs and symptoms, including physical, mental, and sometimes psychiatric problems[2, 3]. According the the Altas of MS report [4] currently there are more then 2.8 people living with MS worldwide, including about 90,000 in the United States. Obviously, this disease significantly changes the life of a large number of people, decreasing their quality of life, working productivity, etc. In addition, MS affects mainly young people (peak age of the onset is approximately 30 years [5]), so it impacts the economy of the country. All of these reasons make the study of factors correlated with MS interesting.

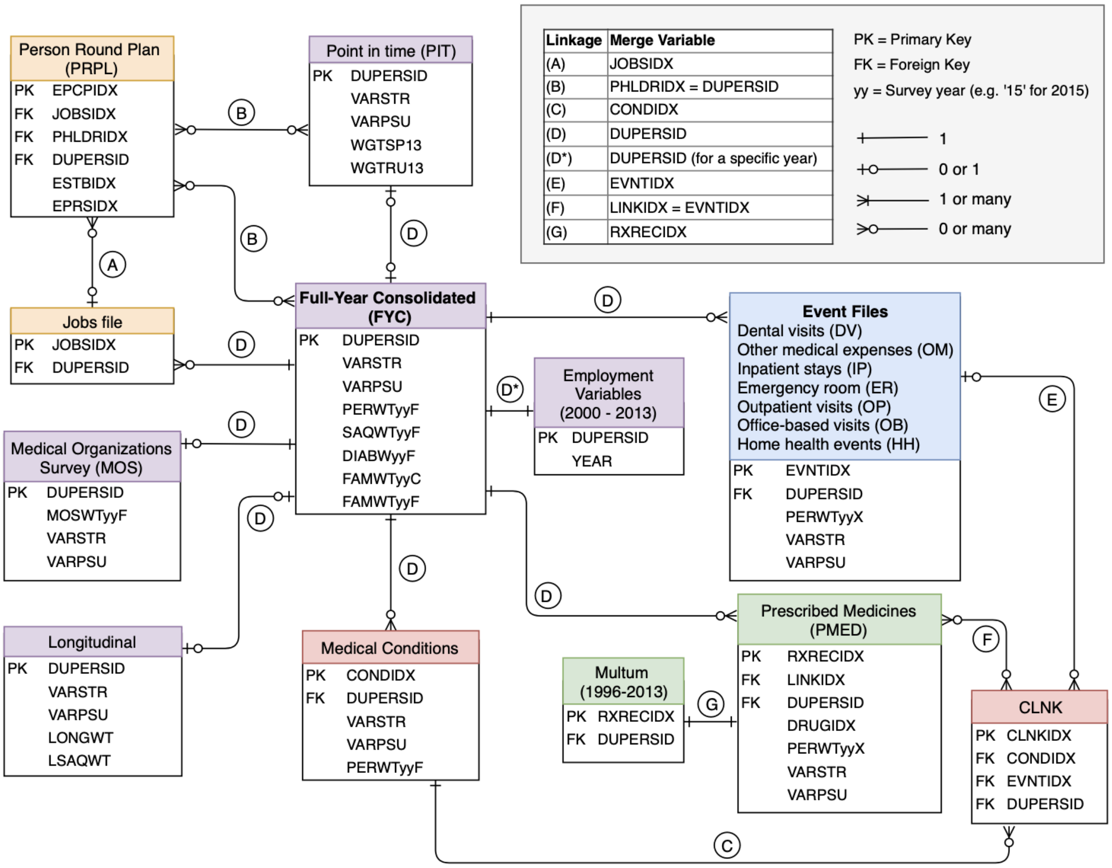
The most reliable source of information would be the results of some medical tests, hospital records, etc. Unfortunately, none of the members of our group has the required medical background, so other approached will be used. Demographic data (like gender, age, etc), educational level, geographical positions, etc. will be considered. Influence of some of these factors has already been studied in literature (see, for example, mentioned above article on the age distribution 'of the MS patients), but it could be interesting to put all these factors into analysis. This research will focus on inclusion of the aforementioned factors.

A good database, containing much of the demographic data that is suitable for solving our project, is the Medical Expenditure Public Survey data set [6]. This is the most complete source of data on the cost and use of health care and contains a representative sample of the whole US population. To complete this analysis, all of the data was downloaded but only the information that is required for this analysis was extracted. The necessary adjustments were made and used to obtained result for the subsequent analysis.

The rest of the paper is organized as follows. The next section gives a brief description of the MEPS data set paying special attention to the fields that were used in our work and adjustments that were made. In section 3, all methods used in the project analysis techniques are described. Results are presented and discussed in section 4 and the final section is reserved for a short conclusion.

2. Dataset Description

The presented analysis is based on data provided by Medical Expenditure Panel Surveys (MEPS). This is the most complete source of data on the cost and use of health care and consists of a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. A complete detailed description of this data set can be found elsewhere (see, for example, [6]). In this section, a brief description of the features that are important for this study is given.



Linkage	Merge Variable
(A)	JOBSIDX
(B)	PHLDRIDX = DUPERSID
(C)	CONDIDX
(D)	DUPERSID
(D*)	DUPERSID (for a specific year)
(E)	EVNTIDX
(F)	LINKIDX = EVNTIDX
(G)	RXRECIDX

Figure 1: Structure of the MEPS data set [7]

The MEPS has two major components: the insurance component, and the Household Component, which is of main interest. The household component data (HC) collects a data from a nationally representative sample of individual households and their members. It has been collected over the last 20 years (1996-2019) and is ongoing. Each year the data set is organized into more than 10 tables (the exact number depends on the specific year) and contains a detailed information for each person in fields such as demographic characteristics, health conditions, changes, and sources of medical payment, etc. A brief structure of the data set is shown in figure 1. Described in more detail below are the fields that will be used in the analysis.

As can be seen from the Figure 1, the main tables that could be interesting to the research question are Full Year Consolidated Data File (FYC) and Medical Condition File. From the first file the following features were selected:

- **DUPERSID**: The unique (throughout the year) identifier of the person. It is a primary key that links together all tables;
- **SEX***: the gender of the person;

Gender	RACES	REGION	INSCOV
Male	AMERICAN INDIAN	NORTHEAST	Private
Female	ALEUT	MIDWEST	Public
	ASIAN	SOUTH	Uninsured
	BLACK	WEST	
	WHITE		
	OTHER		
	MARRY	HIDEG	
	MARRIED	NO DEGREE	
	WIDOWED	GED	
	SEPARATED	HIGH SCHOOL	
	NEVER MARRIED	BACHELORS	
	UNDER 16	MASTERS	
		DOCTORATE	
		OTHER	
		UNDER 16	

Table 1: Description of the categorical variables

- **RACEX**: The race of the person;
- **REGION**: Number of the region;
- **DOBMM**, **DOBYY**: birth month and year respectively;
- **POVCAT**: Poverty category;
- **CONDBEGY**, **AGEDIAG**: The year when the medical condition has started or the age when the person was diagnosed respectively. FYC for years 1996-2006 contain only CONDBEGY field, while for 2007-2017 records only AGEDIAG field is presented. Modifications were made in the analysis to ensure the field was coded uniformly;
- **INSCOV**: Insurance coverage;
- **MARRY**: Martial status;
- **HIDEG**: Highest degree obtained;

From the MCF table, in addition to mentioned above primary key DUPERSID, only the information about the disease of the person is required, which is coded is ICD9CODX field (see reference [8] for information about ICD codes).

The majority of the selected variables were categorical and coded as integers in the original data set. To make the interpretation of the final results easier, the corresponding fields were transformed into factors and assigned the required levels in accordance with the table 1.

After processing the downloaded data from MEPS archive files, only listed above variables were extracted. All necessary corrections and adjustments were made and the resulting data set was saved for future analysis.

	1996 – 2001	2001 – 2019
1	AMERICAN_INDIAN	WHITE
2	ALEUT	BLACK
3	ASIAN	AMERICAN_INDIAN
4	BLACK	ASIAN
5	WHITE	NATIVE_HEWAIIAN
6		MULTIPLE
91	OTHER	OTHER

Table 2: RACE

Before performing the analysis, preprocessing of the data was necessary. One of the fields (RACE), was coded with differing codes for years before and after 2005 (see table 2). During preprocessing, the codes were adjusted and put in accordance with the one presented in table 1. Another issue that required some correction is the age when the condition started. The field (AGEDIAG) was presented only in surveys from year 2002. For earlier data sets instead of it the year when the condition started, the year when the condition was reported was used, so the reported birth year of the person was used to calculate the age of the first diagnosis.

It should be noted that the original data set is strongly imbalanced: less then 1% of persons that participated in surveys were diagnosed with MS. To deal with this issue, both undersampling and oversampling approaches were used. Undersampling was done by taking all MS records from the data set and randomly selecting the same number of non-MS records. The oversampling approach, on the other hand, did not yield reasonable results (see the last section for more details), therefore, only undersampling was used for the analysis.

3. Methodology

3.1. Decision Tree

A decision tree was built to predict which of the attributes used in our analysis contributed to a patient's multiple sclerosis applying the classification tree on the data set including all the study variables described in the previous section using the *rpart* (Recursive Partitioning and Regression Trees) package in R. In brief, a tree was built using the following process:

- identification of the single variable that, when used to split the data set into two groups ("nodes"), best minimized impurity of MS status in each child node, according to the Gini impurity criterion;
- repetition of the partitioning process within each child node and subsequent generations of nodes ("recursive partitioning" or "branching");
- cessation at "terminal" nodes when no additional variables achieve further reductions in node impurity.

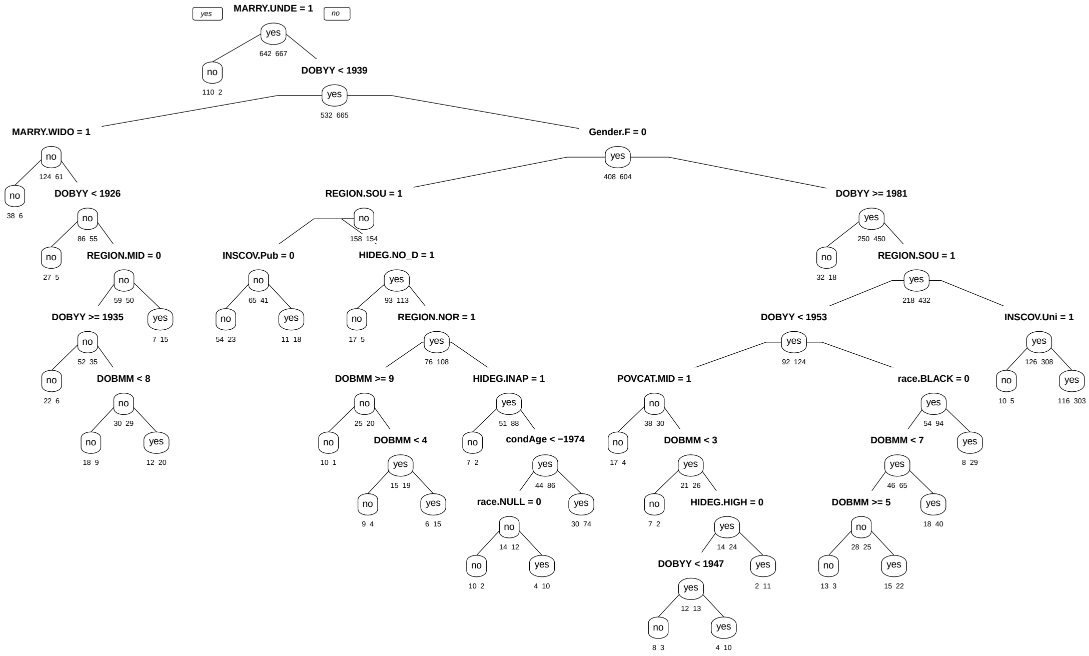


Figure 2: Decision Tree

However, since the tree structure can be quite unstable, shifting substantially depending on the sample chosen, and the fully grown tree leads to overfitting, therefore, cost complexity pruning was employed. The *caret* package was used for cost complexity pruning using 10-fold cross-validation, wherein the largest value of the *kappa* was achieved at $cp = 3.43 \times 10^{-3}$

Some of the major predictors obtained from the final classification tree using binary recursive partitioning included marital status, year of birth, gender, region of the US, race, and education level of the patient. As can be seen in the Figure 2, the first question in the tree, also called the root node, asked: is the patient under the age of 16? There are no predicted cases of MS for people under the age of 16. The second issue, under the right branch (not under 16), asked whether the patient's year of birth is less then 1939. There are very few predicted cases of MS for people born before 1939. The next question asked is if gender is male. There are a lot of MS positive cases among women. The next split is made at if the year of birth greater or equal to 1981. People born between 1939 and 1981 have a higher percentage of MS cases. Further splits inquired as to whether the patient's home area in the United States was the South. Patients who met the requirements were classified as MS positive, while those who did not meet the criteria were classified as MS negative. Patients with MS who were not from the South were often covered by insurance. Another division was made for patients from the South, with the tree asking if the patient's birth year was 1953. If yes,

	Estimate	Standard Error	Pr	
(Intercept)	-37.37	9.364	6.59×10^{-5}	***
Gender.M.	-0.4904	1.358	0.00030	***
race.ASIAN	-1.886	0.614	0.0043	**
REGION.SOUTH	-0.6935e	0.1756	7.82×10^{-5}	***
DOBMM	0.04952	0.01.834	0.006929	**
DOBYY	0.01906	0.004.757	6.15×10^{-5}	***
INSCOV.Public	0.9125e	0.2916	0.001754	**
MARRY.UNDER_16	-4.184	1.159	0.000307	***
HIDEG.INAPPLICABLE	-1.044	0.3490	0.002771	**

Table 3: Extract from the summary of logistic regression table

patients were graded as non-MS and if no, they were graded as MS positive. Further splits were made for if the patient's race is black and then the month of birth.

3.2. Logistic Regression

In most cases, the logistic regression is used to model the relationship between the independent variables and the dependent variable, which is a binary response. In our analysis, the relationship between each study covariate and MS status was evaluated using full logistic regression, as summarized by odds ratios and corresponding 95% confidence intervals. The 10-fold cross-validation method of the *caret* package in R was used to train the model, wherein the largest value of the metric kappa was selected.

As can be seen from the table 3, in the logistic regression analysis, 8 study variables were significantly associated with MS-positive at a significance level of 0.05. Gender, race, area of the United States, month and year of birth, insurance coverage, marital status, and education level were the factors that most strongly contributed to the patients getting MS. Women have been shown to have more MS cases than men in the gender sector. Similarly, a higher percentage of Americans with multiple races are diagnosed with MS as compared to White-Americans, and Native Hawaiians are least likely to develop the disease. As compared to the residents of the West part of the United States, those in the South have the lower rate of MS-positive incidents. When compared to those who are uninsured and protected by private medical insurance, people who are covered by public insurance have a higher percentage of MS cases registered. Furthermore, more people with MS are widow, and young people under the age of 16 have least number of cases. People with a higher education degree that is not ascertained have the most incidents of MS.

3.3. Neural Network

The popular multi-layer architecture of feed-forward neural network was employed to produce a system that was able to predict the MS status given all the study variables in our data set. The *caret* package was used for building the neural net model on the training data, and the 10-fold cross validation approach helped

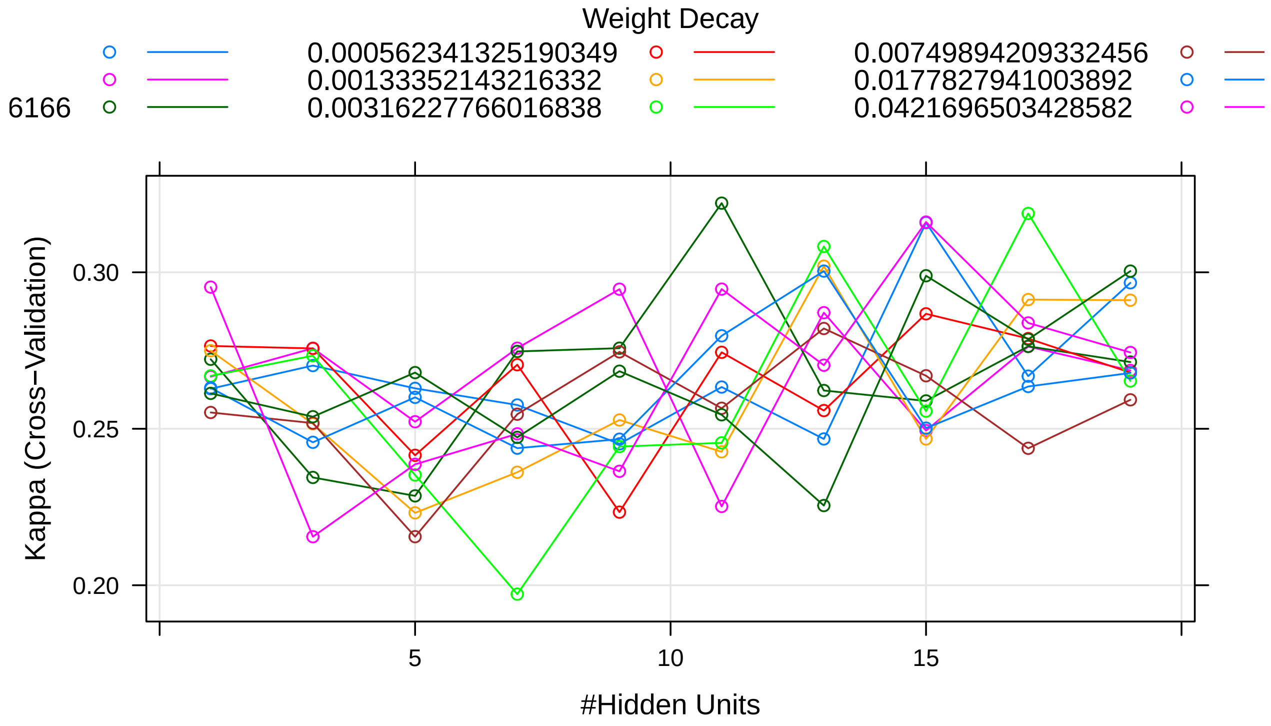


Figure 3: Determining the optimal tuning parameters using cross-validation

us to determine the optimal value of the number of hidden nodes and learning rate (decay) parameter. Figure 3 shows the plot for the tuning parameters.

The optimal tuning parameters chosen to fit our neural network model were one hidden layer with 11 nodes and learning rate (decay) of 0.0002371.

3.4. KNN

As one of the most used clustering methods, k-Nearest Neighbor (k-NN) algorithm was used widely with a simple logic. The algorithm uses a k number of data around multiple centroids and calculates distances between data using the Euclidean distance. The model was trained with the 10-cross validation model, and tune length of 20. Different number of k was tested with its accuracy and kappa value. For this project, the kappa value was used as the main tuning parameters, and the optimal model was analyzed to be k of 31, and kappa value of 0.25. k-NN algorithm output suggested that the data has a great Sensitivity of 0.8284, but the lowest Kappa value of 0.25, resulting as the worst model out of 6 different models tested. Not only it was the worst model, but one challenge was the difficulty of analyzing the output data. With the models such as Logistic regression or Decision tree, it was possible to analyze the result easily with specific variables and numerical outputs provided. It was possible to get the best number of k's according to the best kappa value. However, analyzing farther details was not possible without getting many details from the result.

k	Accuracy	Kappa
31	0.6295092	0.2538403

Table 4: Results from KNN

3.5. RandomForest

Discovered by Leo Breiman, a professor from UC Berkeley, Random Forest is one of the most loved clustering algorithms even in this day. Taking account of 50 different decision trees, prediction was done to analyze if a person is multiple sclerosis or not with all the variables we had in the data. As the name 'Random Forest' suggests, the variables were randomly chosen when creating each Decision Trees. Selecting only few of all the variables considered, each decision tree was created to predict the probability of having MS the patient may have. Like the setting used for all other models, 10-fold cross-validation was used for cost complexity pruning. The tuning parameters were specified using the tune grid option, sequencing from 1 through 10 by 1. The training set was predicted with the test data with the cutoff of 0.5. The output of the model was exceptionally good among 6 different models, with 0.72 accuracy, 0.45 kappa, 0.79 sensitivity, and 0.66 specificity value. The output suggested that the top three variables impacted the result were 'Date of birth: Year' with 100, 'Condition of Age' with 83.11, and 'Date of birth: Month' with 59.52. The decision was made as the Random forest to be the best model of all, when cutoff value was specifically set to 0.5.

	Overall
DOBYY	100
condAge	83.10
DOBMM	59.51

Table 5: Top 3 results from Random Forest

3.6. AdaBoost

As an approach to create a highly accurate prediction rule Adaptive Boosting collects many relatively weak and inaccurate rules to create a strong classifier. Each of the weak classifiers train any random N data with exact same 1/N weights. At the first round of clustering, there are some data that are not clustered sufficiently. By the error rate, the weight of the data increases, and trains the data again. This is done adaptive, which named this model Adaptive Boosting. Repeating this multiple times, the performance of the strong classifier increases. Like other models, 10-fold cross-validation was used as a train control, and tune grid method was used to train the model. The result of AdaBoost suggest that the three variables that impacted model the most were married under 16 with value 100, gender of female and male both having value of 90.2.

	Importance
MARRY.UNDER16	100
Gender.M	90.20059
Gender.F	90.20059

Table 6: Top 3 results from AdaBoost

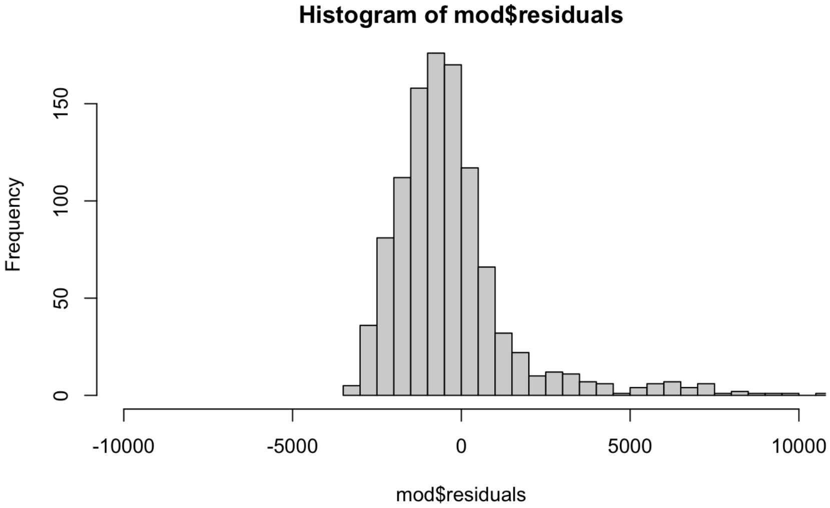


Figure 4: Histogram of mod

3.7. Linear Regression

The second main question to be answered through the analysis was if any association existed between the total out of the pocket costs and MS. TO achieve the goal, a linear relationship was established between the target variable TOTSLF, and the predictor variables of Gender, race, REGION, DOBMM, DOBYY, POVCAT, condAge, INSCOV, MARRY, HIDEG, and isMS. The data was split into a training data, and test data. The model was created with the linear regression, and the multiple R-squared value came out to be 0.09789. As can be seen from the histogram of Mod on Residuals (Figure 4), the residuals show a huge variability ranging from -5000 to 10000 so it can be said that the linear regression model did not correctly analyze, and produced the output. Since the multiple R-squared value did not come out to be close enough to the value 1, it can be inferred that the total out of pocket costs for patients associated with MS cannot be predicted.

	ME	RMSE	MAE
Test set	32.62251	3298.17	1679.946

Table 7: Results from linear regression

4. Results

For each of the models, using a cut score of 0.5, a confusion matrix was constructed and the values for Kappa, accuracy, sensitivity, and specificity were calculated. Results are shown in Table 8.

Model	Kappa	Accuracy	Sensitivity	Specificity
DT	0.327	0.660	0.810	0.522
GLM	0.290	0.642	0.750	0.543
NN	0.312	0.654	0.720	0.594
KNN	0.240	0.613	0.828	0.416
RF	0.450	0.724	0.795	0.659
ADA	0.312	0.654	0.731	0.584

Table 8: Performance metrics with cutoff = 0.5

Since the maximum Kappa value was only 0.45, two other cut scores were generated for each model. The first was the cut score that maximized Kappa and the second found the cut score that balanced specificity and sensitivity. Results are shown in Table 9.

	Model	cutKappa	Kappa	cutCross	Sensitivity
1	DT	0.370	0.330	0.690	0.650
2	GLM	0.450	0.320	0.570	0.610
3	NN	0.480	0.310	0.710	0.640
4	KNN	0.520	0.280	0.570	0.630
5	RF	0.560	0.470	0.560	0.720
6	ADA	0.450	0.360	0.560	0.660

Table 9: Different cut values

The undersampling method that was used to balance the data selected the same number of non-MS patients for the training data set as MS patients existed. The modeling was done, not for the purpose of making or confirming a diagnosis, but for finding demographic factors that are correlated with MS. For this reason, a balance between specificity and sensitivity was selected to optimize the models. Results are shown in Table 10. The random forest model produced the best results with a moderate Kappa score and the best values for accuracy, sensitivity and specificity.

Model	Cut	Kappa	Accuracy	Sensitivity	Specificity
DT	0.690	0.294	0.647	0.646	0.649
GLM	0.570	0.246	0.624	0.608	0.638
NN	0.710	0.272	0.636	0.638	0.635
KNN	0.570	0.220	0.610	0.627	0.594
RF	0.560	0.471	0.736	0.724	0.747
ADA	0.560	0.322	0.661	0.657	0.656

Table 10: Cut Decision

Two other tools were used to compare the performance of the models, lift charts and ROC (Receiver Operating Characteristic) curves. The lift charts shown in Figure 5 are all fairly similar, but the lift chart

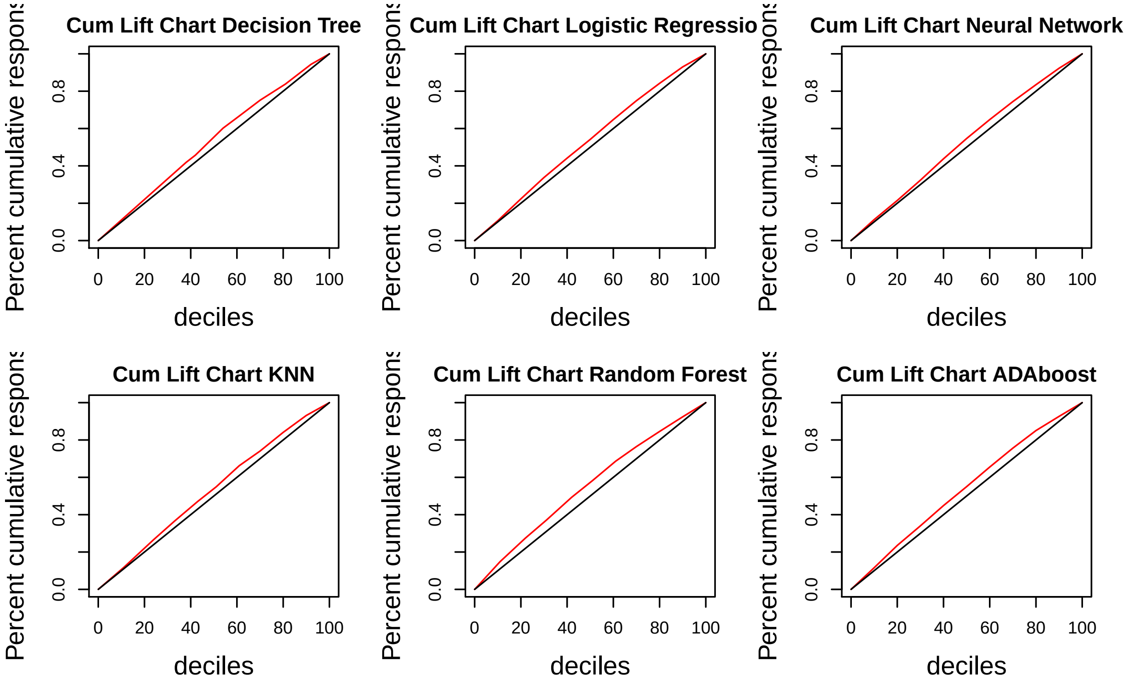


Figure 5: Lift charts

for the random forest model has the greatest area between the diagonal and the percent cumulative response plot against deciles. The ROC curve plots sensitivity against specificity and as shown in Figure 6, confirms that the random forest performs the best of the selected models with an AUC (area under the curve) of 0.8.

5. Limitations and Future Work

The MEPS data set used for this analysis is very large with almost 300,000 observations. Of those observations, there were only 935 instances of patients with MS. The analysis and subsequent results shown in this paper were computed using undersampling to balance the data in the training data set. The SMOTE (Synthetic Minority Oversampling TEchnique) algorithm [9] was also used to generate more MS observations using the k-nearest neighbors. All models were run and the values are shown below in Table 11. The highest Kappa value came from the random forest model, but the value was only 0.001486. A Kappa value this low suggests very slight agreement between the factors and the response variable.

To improve model performance in the future, the N-tree variable could be increased from 50 in the random forest model. Perhaps that would bring the top indicators more in agreement with the top indicators from the other models. There were also many more variable to select from in the data set. It is possible that there are many that are correlated with a patient having MS.

Clustering models could also be explored. It may be interesting to select a few other diseases and make

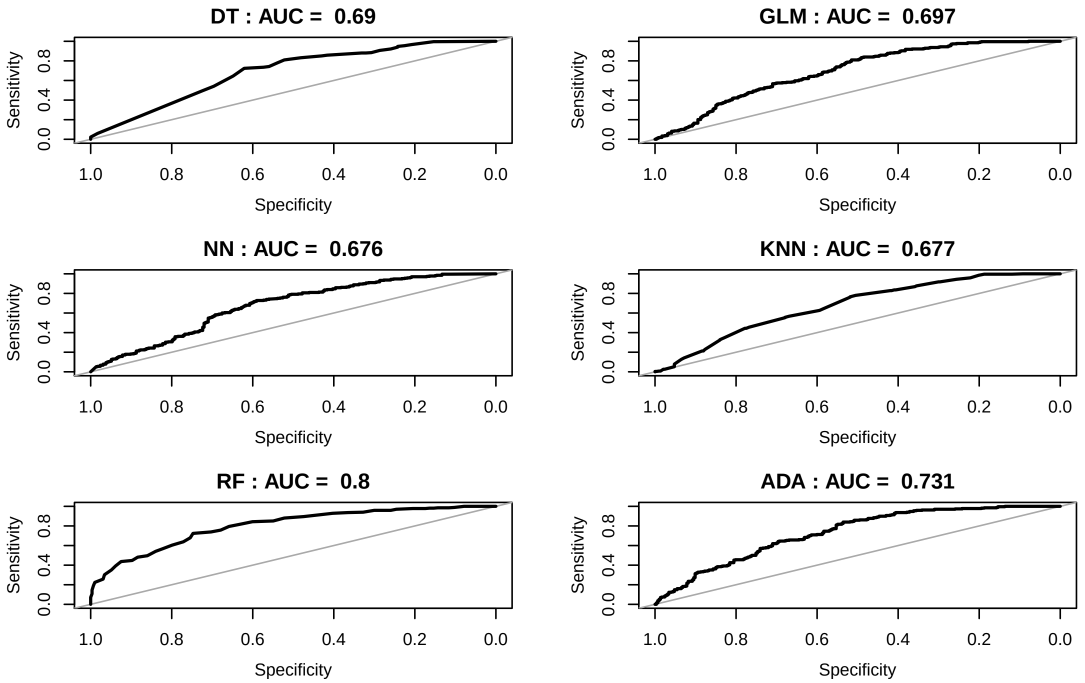


Figure 6: ROC plots

Model	Cut	Kappa	Accuracy	Sensitivity	Specificity
DT	0.250	0.000810	0.654982	0.695853	0.654965
GLM	0.470	0.000648	0.641848	0.649770	0.641845
NN	0.390	0.000770	0.664953	0.658986	0.664956
KNN	0.400	0.000504	0.647053	0.576037	0.647081
RF	0.380	0.001486	0.739589	0.746544	0.739586
ADA	0.410	0.000975	0.691305	0.686636	0.691306

Table 11: Results Using SMOTE

comparisons of mean profiles to look for differentiators.

References

[1] Ninds multiple sclerosis information page. https://web.archive.org/web/20160213025406/http://www.ninds.nih.gov/disorders/multiple_sclerosis/multiple_sclerosis.htm. Accessed: 2021-04-14.

[2] Alastair Compston and Alasdair Coles. Multiple sclerosis. *The Lancet*, 372(9648):1502–1517, Oct 2008.

[3] Alastair Compston and Alasdair Coles. Multiple sclerosis. *The Lancet*, 359(9313):1221–1231, Apr 2002.

[4] Atlas of ms. https://www.atlasofms.org/map/global/epidemiology/number-of-people-with-ms. Accessed: 2021-04-14.

[5] Sara Olofsson, Anne Wickström, Anna Höger Glenngård, Ulf Persson, and Anders Svenningsson. Effect of treatment with natalizumab on ability to work in people with multiple sclerosis. *BioDrugs*, 25(5):299–306, Oct 2011.

[6] Medical expenditure panel survey. `https://meps.ahrq.gov/mepsweb/`. Accessed: 2021-04-07.

[7] Meps, github repository. `https://github.com/HHS-AHRQ/MEPS`. Accessed: 2021-04-07.

[8] Icd9 codes. `http://www.icd9data.com/`. Accessed: 2021-04-07.

[9] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.