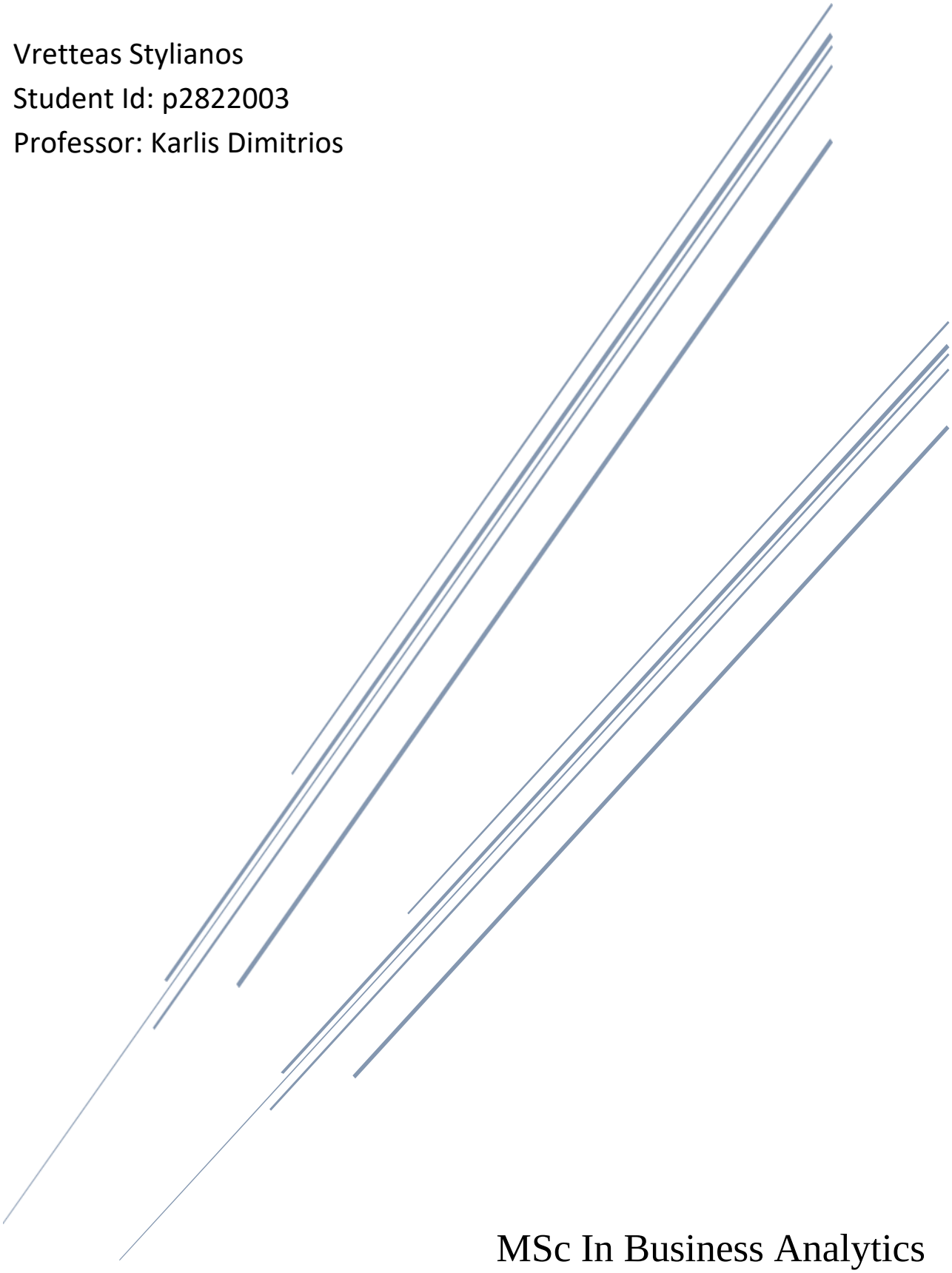
STATISTICS FOR BUSINESS ANALYTICS II
Project II 2020-2021

Vretteas Stylianos
Student Id: p2822003
Professor: Karlis Dimitrios



MSc In Business Analytics

Athens University of Economic and Business

Table of Contents

Introduction

This project consists of two parts and deals with Classification and Clustering Machine Learning problems, moreover from the first Project where we implement a logistic regression model in order to unsterstand the voting behavior between Bernie Sanders and Hillary Clinton. Specifically, the two parts are problems of Supervised learning and Unsupervised learning respectively.

In the first part we deal with the supervised learning problem, we have already information about the two classes we want to predict, eg. the winner of the elections between Hillary Clinton and Bernie Sanders. This means that we will construct several models with various methods which will try to predict the voting behavior having as input the demographic and socioeconomics variables of the counties and try to predict the winner of the elections. As said in this situation we know the classes which we want to predict and try to learn from our data in order construct a good classifier with a good predicting ability.

In the second part we deal with an unsupervised learning problem and try to implement a Cluster analysis. Cluster analysis is a branch of machine learning that groups the data that has not been labelled, classified or categorized, eg. now we do not take into consideration the elections and the votes. We have to construct a clustering algorithm in order to group the data based on their demographic characteristics. Although in this situation we have

as input the demographic information of the counties, we do not know the clusters, we will try to find similarities among them based on their “distance” and we will try to create clusters with the counties.

At the end of each problem we will interpret and evaluate our results and methods and propose solutions in order to further improve them in order to have better results both in the classification and the clustering problem.

Part 1 - Classification

In this part, as previously said we will create several models in order to predict the behavior of the voters between Hillary Clinton and Bernie Sanders.

We will use the cleaned dataset from the previous project, which consists of 2802 observations and 30 variables and 52 variables. From these 51 variables are exploratory and one variable the Response is the column which indicates the winner. Value 1 is stands for “Hillary Clinton” while value 2 stands for “Bernie Sanders”, from the first look we see that from the original dataset 1153 vote for Bernie Sanders and 1649 voted for Hillary Clinton. See below table.

Table 1. Total Votes

Bernie Sanders	Hillary Clinton
1153	1649

For starters, before we begin to construct models, we have to split our dataset into training dataset and test dataset. We will use the first dataset in order to train and tune our models and then we will test them in the second dataset which will serve as unseen data. We do this in an attempt to protect our models from the overfitting problem.

Overfitting happens when the models' algorithms "see" all the data during their training session, so they "learn" from all the data and provide very good accuracy estimates. This is wrong because these models have not seen any other data during their training and will perform poorly into unseen datasets.

In addition, for the training session we will use the repeated k-fold cross validation procedure for better results. We decide that the number of the folds will be ten (k=10) and the procedure will be repeated twenty (repeats = 20) times. This trainControl method will be implement into all models.

Method 1 – Logistic Regression

The first method we are going to use is to build a logistic regression model. Logistic regression estimates the probabilities between the two outcomes, so we will classify this probability between Bernie Sanders and Hillary Clinton.

First of all, we construct a general linear model with all the predictors which we call "full_model" and will use the Akaike information criterion (AIC) method to estimate the likelihood of a model to predict/estimate the future values. A good model is the one that has minimum AIC among all the other models so the algorithm will propose us the model with the minimum AIC. This criterion can be applied to perform automatic variable selection.

Table 2. AIC results

Call:

glm(formula = Response ~ +PST045214 + PST040210 + PST120214 + AGE135214 + AGE295214 + AGE775214 + SEX255214 + RHI325214 + RHI525214 + RHI625214 + RHI625214 + RHI725214 + RHI825214 + POP645213 + EDU635213 + EDU685213 + EDU685213 + VET605213 + HSG445213 + HSD410213 + INC910213 + BZA010213 + BZA110213 + NES010213 + SBO515207 + POP060210, family = "binomial", data = data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.2594	-0.7406	0.0141	0.5860	3.5706

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3796.1 on 2801 degrees of freedom
Residual deviance: 2193.1 on 2777 degrees of freedom
AIC: 2243.1

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.506e+01	2.327e+00	10.770	< 2e-16	***
PST045214	-8.090e-05	1.765e-05	-4.583	4.59e-06	***
PST040210	8.971e-05	1.850e-05	4.850	1.23e-06	***
PST120214	1.272e-01	2.325e-02	5.473	4.43e-08	***
AGE135214	-3.066e-01	1.058e-01	-2.899	0.003742	**
AGE295214	2.402e-01	4.202e-02	5.716	1.09e-08	***
AGE775214	2.491e-01	2.439e-02	10.211	< 2e-16	***
SEX255214	4.944e-02	3.330e-02	1.485	0.137607	
RHI325214	-2.424e-01	1.826e-02	-13.272	< 2e-16	***
RHI525214	5.192e-01	3.905e-01	1.330	0.183674	
RHI625214	-1.016e+00	8.982e-02	-11.316	< 2e-16	***
RHI725214	-1.970e-01	1.744e-02	-11.296	< 2e-16	***
RHI825214	-2.696e-01	1.774e-02	-15.191	< 2e-16	***
POP645213	-1.094e-01	2.544e-02	-4.300	1.71e-05	***
EDU635213	-9.582e-02	1.457e-02	-6.575	4.88e-11	***
EDU685213	-9.376e-02	1.524e-02	-6.152	7.63e-10	***
VET605213	1.553e-04	3.701e-05	4.195	2.72e-05	***
HSG445213	-4.336e-02	1.163e-02	-3.729	0.000192	***
HSD410213	-6.983e-05	2.325e-05	-3.004	0.002666	**
INC910213	1.805e-04	2.354e-05	7.667	1.77e-14	***
BZA010213	-1.266e-03	2.168e-04	-5.838	5.29e-09	***
BZA110213	3.202e-05	8.496e-06	3.769	0.000164	***
NES010213	4.184e-04	7.127e-05	5.870	4.36e-09	***
SBO515207	-4.302e+00	3.251e+00	-1.323	0.185739	
POP060210	4.340e-04	2.454e-04	1.768	0.077004	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2 shows the proposed model with the lowest AIC rate (2243.1).

We could do further improvisations into this model but we will not proceed into further improvements because we do not care about inference. Theoretically we can “accept” a model with not statistically important variables because we actual care for prediction.

Now that we have our model we can proceed into training and tuning methods in order to train our model into the data and then evaluate its ability to predict on the unseen data.

Table 3. method1 - Accuracy

Generalized Linear Model

1962 samples
24 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 20 times)
Summary of sample sizes: 1766, 1766, 1766, 1766, 1766, 1766, ...
Resampling results:

RMSE	Rsquared	MAE
0.3659009	0.4484567	0.263764

After the training session we evaluate our model and see that its RMSE is 36% a rate which is quite low.

Now we are going to test the model into the unseen data.

We use the confusion matrix in order to evaluate the predicting ability of the model. The model accuracy on the test dataset is 80.24 % which is a very good rate. In addition, the accuracy of the model is estimated between 77.38% and 82.88% in a confidence interval of 95%. It is a good rate too because there is not much distance between these two values. Finally, we plot the ROC curve of the model in order to estimate the true positive rate vs the false positive rate of the classification. We want the ROC curve to be stretched as far as possibly to the upper left corner, the higher to the left is the curve then the percentage of the false classified observations is lower. Below are represented the detailed confusion matrix results and the ROC curve plot.

Table 4 - Confusion Matrix Logistic Regression

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	273	92
1	74	401

Accuracy : 0.8024
95% CI : (0.7738, 0.8288)
No Information Rate : 0.5869
P-Value [Acc > NIR] : <2e-16

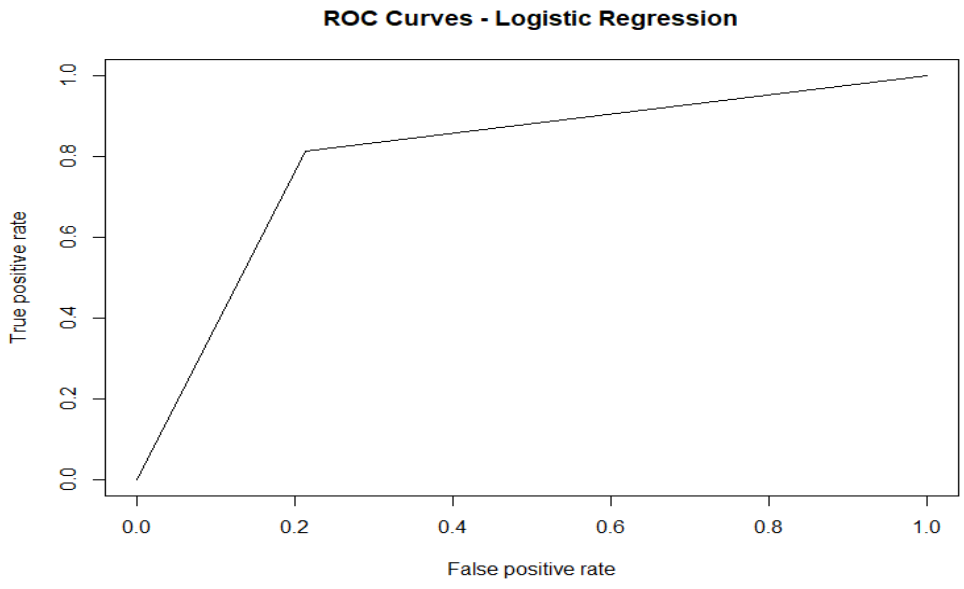
Kappa : 0.5956

Mcnemar's Test P-Value : 0.187

Sensitivity : 0.7867
Specificity : 0.8134
Pos Pred Value : 0.7479
Neg Pred Value : 0.8442
Prevalence : 0.4131
Detection Rate : 0.3250
Detection Prevalence : 0.4345
Balanced Accuracy : 0.8001

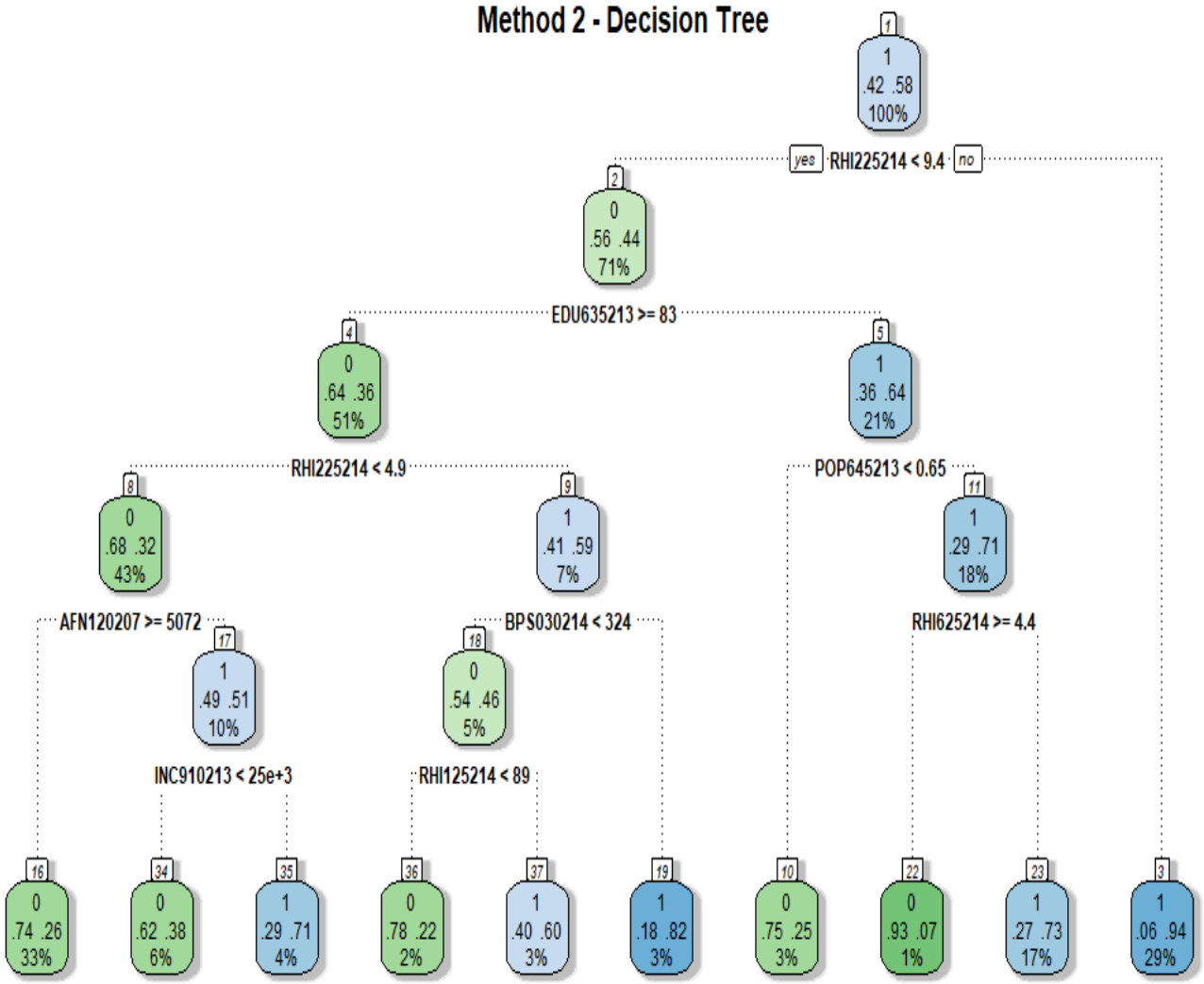
'Positive' Class : 0

Figure 1



Method 2 – Decision Tree

The second method we are going to use in order to predict the outcome of the elections is name Decision Tree. Decision tree is a type of supervised learning algorithm that can be used in classification problems, the main idea behind the algorithm is to ask questions and regarding the answer to proceed into the next step. When you have a good number of questions, you can classify each observations of the dataset in the wright partition. Below is the decision tree we constructed based on our train dataset.



From the above decision tree, we could understand the following.

Response is 0.07 when	
RHI225214 < 9.4	Black or African American alone, percent, 2014
EDU635213 < 83	High school graduate or higher, percent of persons age 25+, 2009-2013
POP645213 >= 0.65	Foreign born persons, percent, 2009-2013
RHI625214 >= 4.4	Two or More Races, percent, 2014
Response is 0.22 when	
RHI225214 is 4.9 to 9.4	Black or African American alone, percent, 2014
EDU635213 >= 83	High school graduate or higher, percent of persons age 25+, 2009-2013
BPS030214 < 324	Building permits, 2014
RHI125214 < 89	White alone, percent, 2014
Response is 0.25 when	
RHI225214 < 9.4	Black or African American alone, percent, 2014
EDU635213 < 83	High school graduate or higher, percent of persons age 25+, 2009-2013
POP645213 < 0.65	Foreign born persons, percent, 2009-2013
Response is 0.26 when	
RHI225214 < 4.9	Black or African American alone, percent, 2014
EDU635213 >= 83	High school graduate or higher, percent of persons age 25+, 2009-2013
AFN120207 >= 5072	Accommodation and food services sales, 2007 ($1,000)
Response is 0.38 when	
RHI225214 < 4.9	Black or African American alone, percent, 2014
EDU635213 >= 83	High school graduate or higher, percent of persons age 25+, 2009-2013
AFN120207 < 5072	Accommodation and food services sales, 2007 ($1,000)
INC910213 < 25102	Per capita money income in past 12 months (2013 dollars), 2009-2013
Response is 0.60 when	
RHI225214 is 4.9 to 9.4	Black or African American alone, percent, 2014
EDU635213 >= 83	High school graduate or higher, percent of persons age 25+, 2009-2013
BPS030214 < 324	Building permits, 2014
RHI125214 >= 89	White alone, percent, 2014
Response is 0.71 when	
RHI225214 < 4.9	Black or African American alone, percent, 2014
EDU635213 >= 83	High school graduate or higher, percent of persons age 25+, 2009-2013
AFN120207 < 5072	Accommodation and food services sales, 2007 ($1,000)
INC910213 >= 25102	Per capita money income in past 12 months (2013 dollars), 2009-2013
Response is 0.73 when	
RHI225214 < 9.4	Black or African American alone, percent, 2014
EDU635213 < 83	High school graduate or higher, percent of persons age 25+, 2009-2013
POP645213 >= 0.65	Foreign born persons, percent, 2009-2013
RHI625214 < 4.4	Two or More Races, percent, 2014
Response is 0.82 when	
RHI225214 is 4.9 to 9.4	Black or African American alone, percent, 2014
EDU635213 >= 83	High school graduate or higher, percent of persons age 25+, 2009-2013
BPS030214 >= 324	Building permits, 2014
Response is 0.94 when	
RHI225214 >= 9.4	Black or African American alone, percent, 2014

The possibility of the predicted values is 7% when the Black or African American alone, percent, 2014 is less than 9.4 and the High school graduate or higher, percent of persons age 25+, 2009-2013 is less than 83, Foreign born persons, percent, 2009-2013 is greater than 0.65 and
Two or More Races, percent, 2014 is bigger than 4.4.

The possibility of the predicted value is 22% when Black or African American alone, percent, 2014 is between 4.9% and 9.4%, High school graduate or higher, percent of persons age 25+, 2009-2013 is greater than 83%, building permits of 2014 is less than 324 and white alone people percent is less than 89%.

The possibility of the predicted value to be correct is 60% when black or African alone, percent is less than 9.4%, high school graduate or higher, percent of persons age 25+ for 2009-2013 is less than 83% and foreign-born people, percent for 2009 – 2013 is less than 65%.

The possibility of the predicted value to be correct is 60% when black or African alone, percent is less than 9.4%, high school graduate or higher, percent of persons age 25+ for 2009-2013 is less than 83% and foreign-born people, percent for 2009 – 2013 is less than 65%.

The possibility of the predicted value to be correct is 73% when black or African alone, percent is less than 9.4%, high school graduate or higher, percent of persons age 25+ for 2009-2013 is less than 83%, foreign-born people, percent for 2009 – 2013 is less than 65% and two or more races' percent is less than 4.4%

The possibility of the predicted value to be correct is 82% when black or African alone, percent is less than 9.4% and greater than 4.9%, high school graduate or higher, percent of persons age 25+ for 2009-2013 is greater than 83% and building permits is less than 4.4%.

The possibility of the predicted value to be correct is 94% when black or African alone, percent is less than 9.4%, this is the variable with the most information gain.

We see that as we as ask more questions then our possibility to be correct is augmented.

The most influential variables regarding the estimations made by the decision tree are the following. We can understand this from three nodes.

- Black or African American alone, percent, 2014
- High school graduate or higher, percent of persons age 25+, 2009-2013
- Foreign born persons, percent, 2009-2013
- Accommodation and food services sales, 2007 ($1,000)
- Building permits, 2014

Regarding its predicting ability, we are going evaluate the decision tree using by calculating its confusing matrix. Below the detailed table of Decision Tree method.

Table 5- Desicion Tree method

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	244	122
1	93	381

Accuracy : 0.744
95% CI : (0.7131, 0.7733)
No Information Rate : 0.5988
P-Value [Acc > NIR] : < 2e-16

Kappa : 0.4747

Mcnemar's Test P-Value : 0.05619

Sensitivity : 0.7240
Specificity : 0.7575
Pos Pred Value : 0.6667
Neg Pred Value : 0.8038
Prevalence : 0.4012
Detection Rate : 0.2905
Detection Prevalence : 0.4357
Balanced Accuracy : 0.7407

'Positive' Class : 0

As we see the accuracy of the tree model on unseen data is 74.40%.
An estimation is made also at 95% CI and the values are between 71% and 77%.
It is worse than the previous method but it is not so bad.

We could boost its predicting ability by apply Random Forest method which applies many decision trees and takes the information for all of them but there is always the problem of overfitting.

Method 3 – Support Vector Machines

The third method we are going to use is called SVM which stands for Support Vector Machines. The basic idea behind this technique is to find the optimal hyperplane, which in a two dimensions environment it is a simple straight line. For the SVM the hyperplane it's the decision boundary that maximizes the margins from both tags (in our case – Response variable 0 & 1) i.e. the distance between the separation boundary and the points that are closest to it.

Regarding the predictive ability of the SVM model its accuracy is 78% and the estimation at a CI of 95% is 76% between 81% which is fair good.

Table 6 - SVM method

Confusion Matrix and Statistics

Prediction	Reference 0	Reference 1
0	267	107
1	70	396

Accuracy : 0.7893
95% CI : (0.7601, 0.8164)
No Information Rate : 0.5988
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5692

Mcnemar's Test P-Value : 0.006811

Sensitivity : 0.7923
Specificity : 0.7873
Pos Pred Value : 0.7139
Neg Pred Value : 0.8498
Prevalence : 0.4012
Detection Rate : 0.3179
Detection Prevalence : 0.4452
Balanced Accuracy : 0.7898

'Positive' Class : 0

Comparison of the Methods.

Roc Curve for all models.

We plot ROC curves in a single plot for all three methods in order to understand which is better. From the below plot we can say that SVM and Logistic Regression methods perform quite similar while the decision tree method has poorer results than the previous two. Specifically, the SVM method has a slightly bigger true positive rate.

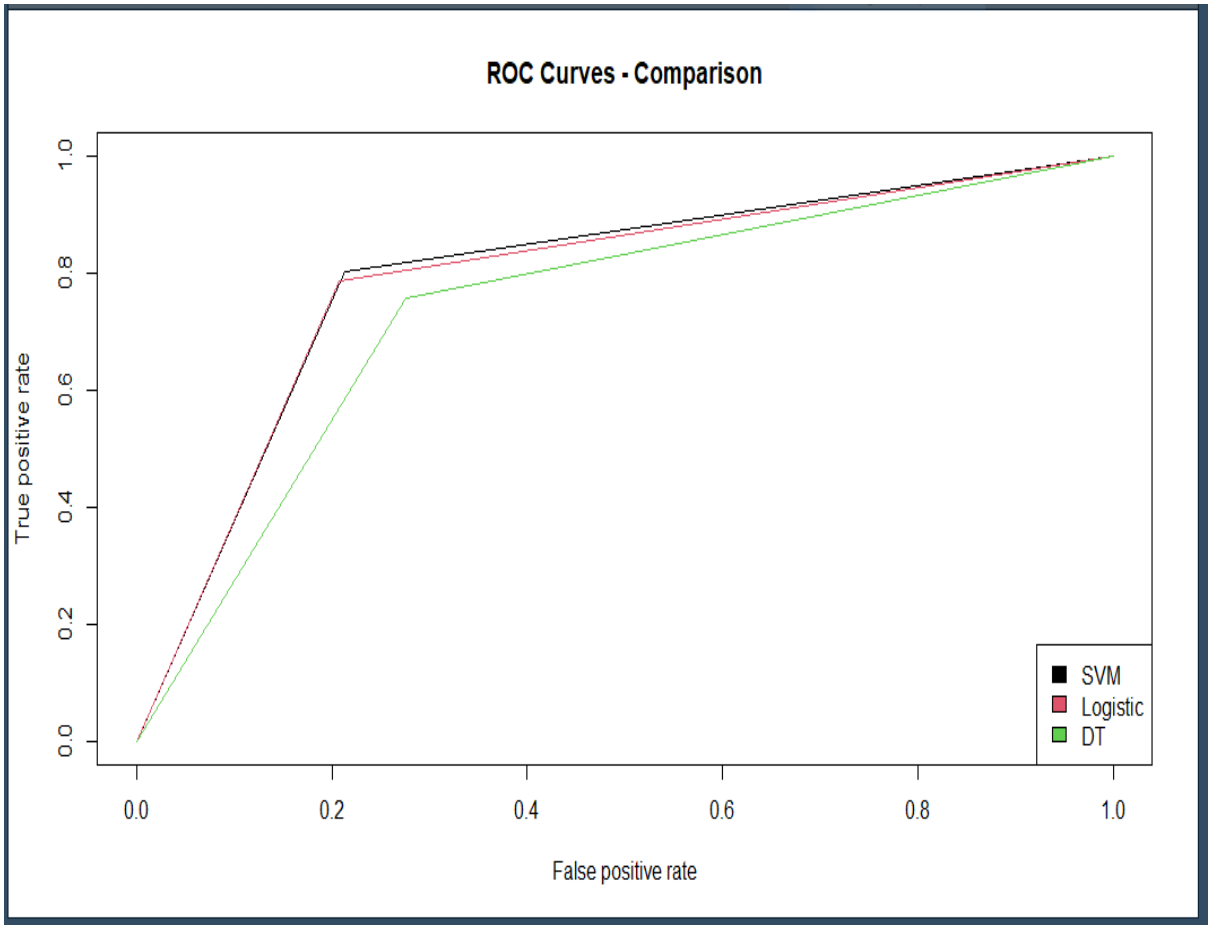


Figure 2- ROC Curves Comparison

Part 2 – Clustering

In this part, we are into the unsupervised learning side and we are going to implement some clustering methods in order to extract information from our dataset. This time we do not have any knowledge about the labels of the dataset and we will try to split the data into groups ("clusters ") where each cluster will consist of observations who share similarities among them.

We will use K-Mean's algorithm, K-Means is an unsupervised learning algorithm whose aim is to partition and allocate the n observations of the dataset into k clusters in which each observation belongs to the cluster with the nearest mean. The basic idea behind K-Means is to is to define cluster in such a way that the total intra-cluster variation is minimized. The standard algorithm defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid.

First of all, in order to use the Euclidean distance, we will scale our dataset before the use of the algorithm. If the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms. Since K-Means computes means it is very important to scale our data before we fed them into the algorithm. With this transformation we ensure that we bring the data into the same format in order to have a better calculation of their datapoints distance.

After the scaling we have to decide the optimal number of clusters.
We are going to use the elbow method for this task in order to find the optimal number of clusters. The elbow method looks at the percentage of variance explained as a function of the number of clusters. The optimal number of clusters is the margin number when adding another one cluster then the modelling of the data is not improved.
Specifically, if we plot the percentage of variance explained by the clusters against the number of clusters, we will see that the first clusters add information (explain a lot of variance), there is one point when if you add another one cluster the marginal information gain will drop (in the graph is like an elbow).

The optimal number of clusters is chosen at this point, hence the “Elbow method”.

Below the graph of the elbow method for our data. We see that the optimal number of clusters is 4.

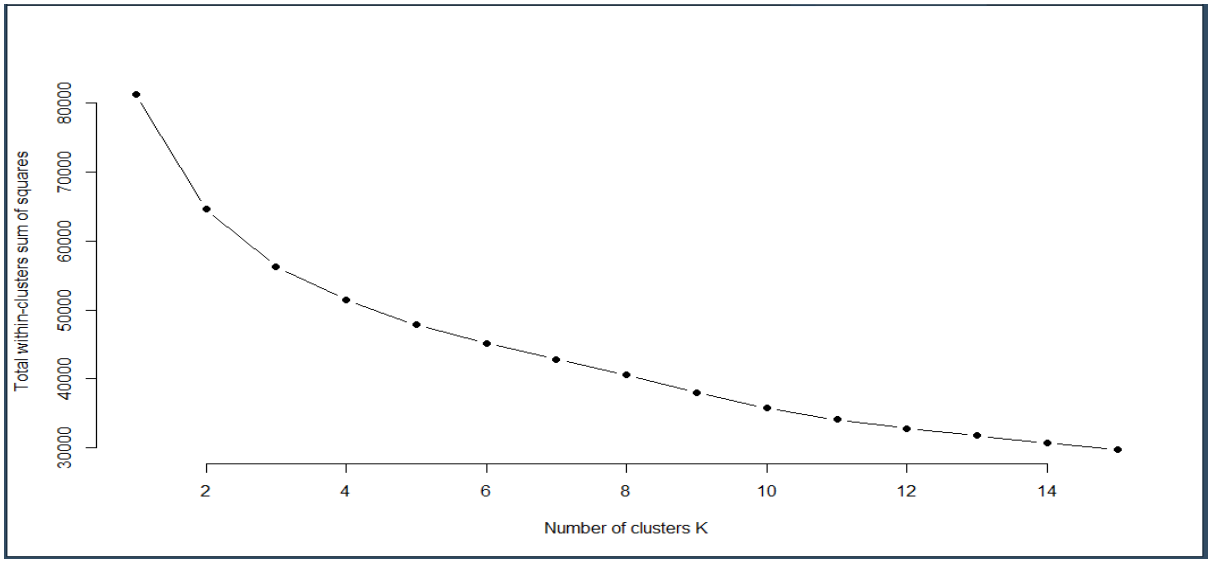


Figure 3 - Elbow method

Specifically, we see that we earn more information in the four first clusters. When we add a fifth cluster there is not so much gain, thus we decide to implement K-Means with 4 clusters.

In the next table, we present the detailed cluster analysis results.

We see that the Within Sum of Squares by cluster is 36.7 % which is quite low. The within-cluster sum of squares is a measure of the variability of the observations within each cluster.

The next table is the detailed R output of the cluster means results.

Table 7- K-Means Results

K-means clustering with 4 clusters of sizes 5, 680, 2043, 74

Cluster means:

	LFE305213	HSG010214	HSG445213	HSG096213	HSG495213	HSD410213	HSD310213
1	0.81485327	13.9040486	-2.8682278	3.8473768	3.7546449	13.9070026	0.87228851
2	0.01782785	0.2592129	-0.3957496	0.8164635	0.8526737	0.2527376	0.21844358
3	-0.03536874	-0.2471138	0.1934216	-0.3710533	-0.3646437	-0.2441177	-0.09850361
4	0.75758276	3.5009253	-1.5095885	2.4814836	1.9780522	3.4775114	0.65324044

	INC910213	INC110213	PVY020213	BZA010213	BZA110213	BZA115213	NES010213
1	2.3217048	1.3788410	-0.08811707	15.1763872	15.9191265	0.40943027	14.4916966
2	0.8756093	0.9203788	-0.55417126	0.2138392	0.1932329	0.15665643	0.1621349
3	-0.3522370	-0.3589201	0.19946150	-0.2275370	-0.2229197	-0.06235435	-0.2031529
4	1.5215865	1.3584050	-0.40841634	3.2914252	3.3031220	0.25427600	3.1396104

	SBO001207	SBO315207	SBO115207	SBO215207	SBO515207	SBO415207	SBO015207
1	14.7404897	0.4915715	0.23261404	4.3856444	0.33122565	2.0678192	0.8582206
2	0.1885847	0.2162810	-0.03521789	0.3729472	-0.03001924	0.2386357	0.6447436
3	-0.2136920	-0.1116186	0.00766741	-0.2580949	-0.04794208	-0.1292959	-0.2495689
4	3.1707124	1.0609171	0.09622397	3.4021017	1.57706248	1.2370278	0.9074654

	MAN450207	WTN220207	RTN130207	RTN131207	AFN120207	BPS030214	LND110210
1	14.3184201	17.84002590	14.8397060	1.0587408	13.7298244	13.2020595	1.62771106
2	0.1997737	0.05555762	0.2489355	0.7511937	0.1405229	0.2643712	0.05634278
3	-0.1820660	-0.15127078	-0.2412354	-0.2792919	-0.1849804	-0.2230752	-0.03459714
4	2.2232775	2.46036330	3.3698557	0.7363232	2.8879777	2.8372950	0.32743674

	POP060210
1	8.6000216
2	0.1005910
3	-0.1171943
4	1.7300811

Within cluster sum of squares by cluster:

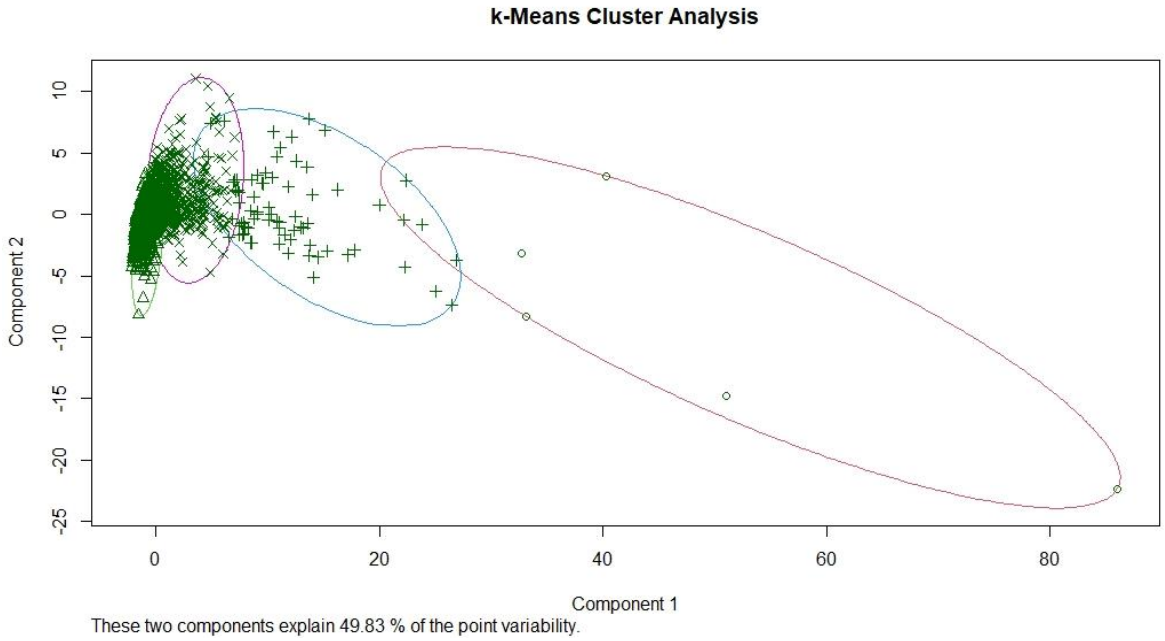
[1] 5115.520 12620.466 23774.122 9940.781

(between_SS / total_SS = 36.7 %)

From the above table we look into each cluster means.

In this way we can get insights from our dataset and find similarities regarding the counties. From the above table we understand how the algorithm assigned the counties by comparing their means into clusters.

Finally, we plot the outcome of the K-Means.



In this plot we see how K-Means assigned its data point to separate clusters. We see that there is one cluster for the observations that are extreme outliers. One cluster for observations that are not so much extreme outliers and finally in the right part of the plot there are two clusters for the rest observations.

Conclusion

Part 1- Classification
Regarding classification we train and test three methods and we compare the results. In order to be more sure we could implement more methods such as Naive Bayes and KNN to have a more spherical approach. We could also boost the results of decision tree with the Random Forest method.

Part 2- Clustering
We implemented only K-Means algorithm, and the results were not satisfying. In order to have a better clustering approach we could implement PAM

algorithm or CLARA which use distances based on medoids and not centroids to see how the clustering will perform.