OR-568 Final Project Report

Young People Survey Analysis: Predictive Modelling on Youngster's Age

Sujan Chava | Vardan Chennupati | Vamshidhar Gunnala | Sahithi Lakamana | Kranthi Rallabandi

Abstract

With the advancement of technology, it is no surprise that one can be able to predict the minds of youngsters. Every generation of teens are shaped by social, political and economic events of the day. Since these factors cannot be the same in every era, youth of every generation have different mindset. The dataset "Young People Survey" consists of data about the habits and survey about the interests of the students at a university in the U.K. This gives you detailed information about the demographics, spending habits, personality traits, views on life and opinions, health habits, phobias, hobbies and interests, movie preferences and music preferences. These features can be used to cluster data of similar behavior, hypothesis testing, visualization, predictive modelling etc.

Literature Review

This dataset was previously used to analyze the gender differences, i.e. trying to find out the differences in interests of the male and female based on their demographics, music interests and so on. It was also used to predict about the willingness to put money into healthy food based on the hobbies and interests, health habits and spending habits. In addition to this, dataset was used to find out the habituate of people based on gender in prevalence to phobias.

Data

Background and History

This dataset extracted by conducting a survey Statistical class's student studying in FSEV UK. Though, real data has an advantage of its veracity, on the other side, it requires high amount of pre-processing. However, this dataset was first used in the year 2013 for the analysis of the healthy habits based on the gender and also prediction of the shopping habits, when trying to find out human behavior given different circumstances. Followingly, another project came up in 2018 to find out the classification of number of siblings based on personality. (Henry, 2018)

Exploration

The dataset is taken from Kaggle which has 1010 rows and 150 columns. This data was collected by the students of Statistics class at FSEV UK as a part of the survey. This dataset has 139 numerical columns and 11 categorical columns, where the entire data set is divided into 8 groups where each group contain certain number of columns in it and each group represent different categories or interests of the students.
The 8 groups contained in the dataset are:

1) Music Preferences: This group has 19 items (columns) i.e first 19 columns says about music preferences of people of different age group and gender.
2) Movie Preferences: This group has 12 items (columns) in it i.e next 12 columns says about movie preferences of people of different age group and gender.
3) Hobbies & Interests: This group has 32 items (columns) in it and says about hobbies and interests of different people.
4) Phobias: This group has 10 items (columns) in it and says about different kinds of phobias people have.
5) Health Habits: This group has 3 items (columns) in it and it says about smoking and drinking habits in different age groups and gender.
6) Personality traits, views on life, & Opinions: This group has 57 items (columns) in it and tells about personality and opinions of different kinds of people.

7) Spending Habits: This group has 7 items (columns) and it tells about the spending habits of the different kinds of people in different aspects.
8) Demographics: This group has 10 items (columns) in it and it tells about age , height, weight, gender, number of siblings etc.

The sample dataset looks like the one in following image:

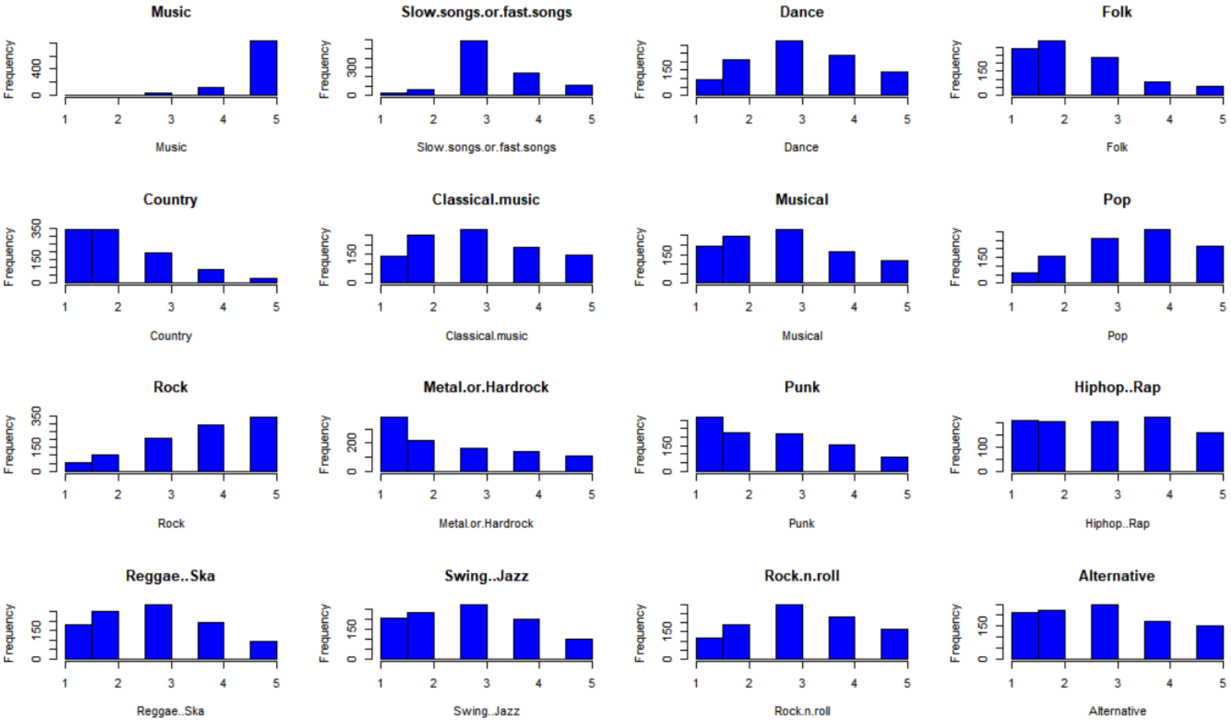


Fig 1.1: The young people survey dataset

Using the above dataset, we can analyze lot of details corresponding to people of different age groups and gender. We have used age as predictor variable in our analysis i.e. using all other different attributes present in the above 8 groups we have tried to predict the age of the person. For proceeding to modelling, firstly we have visualized how age is correlated with all the attributes present in each group and how smoking and drinking habits are related with age of a person and gender of a person. These plots helps us to identify what are the attributes that are highly correlated which can be used for building our models in predicting the age of a person.

Correlation plots for age VS different attributes in each group and visualization:

1) Age Vs Different Grouping correlation plot:

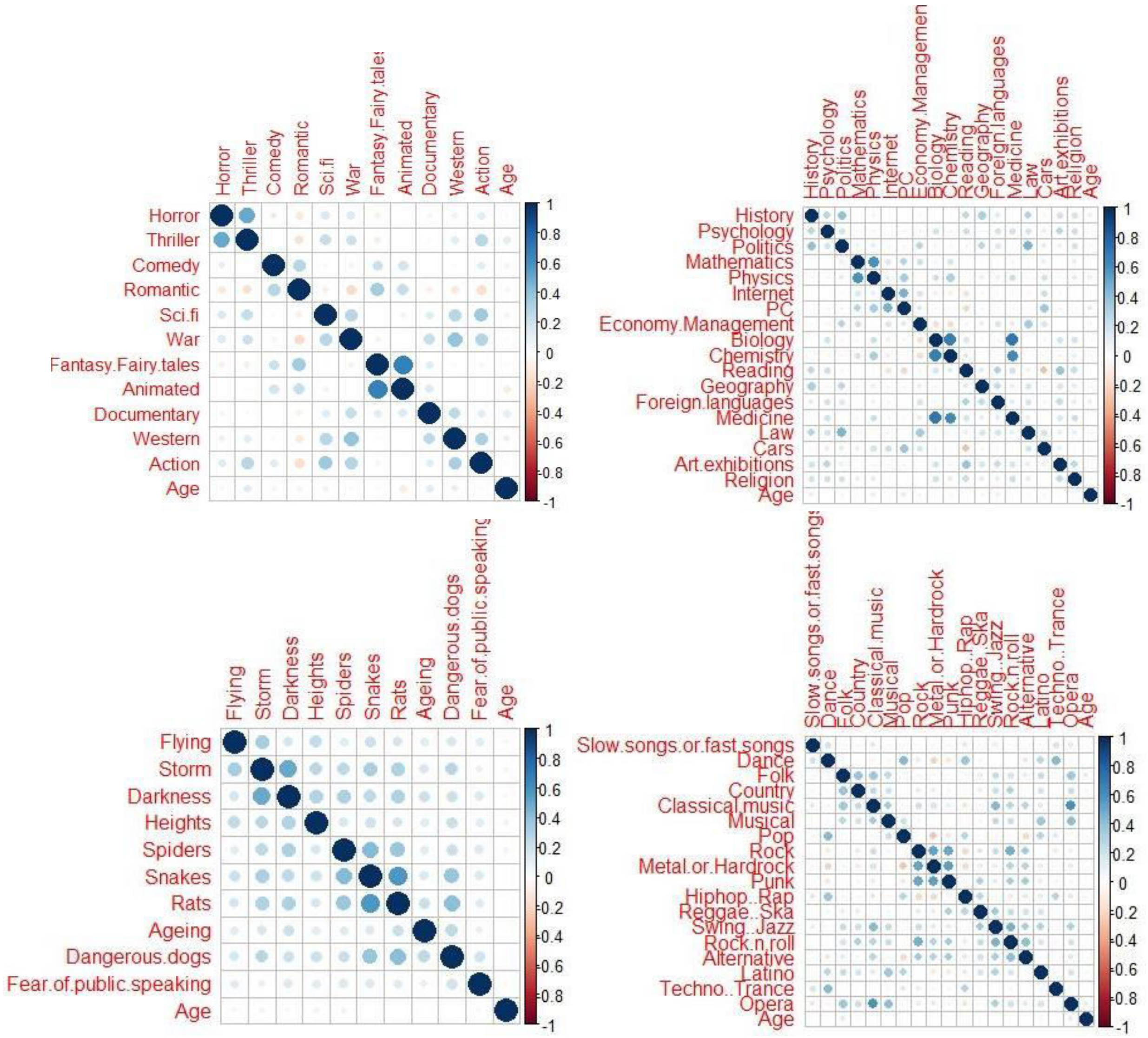


Fig. 2.1: Correlation Visualization by grouping movies, interests, hobbies and music.

The value of the attributes correlated to age can be seen in the scale on the right of the figure. But these are not highly correlated with age because none of them have bigger circle. Here in interest's plot, the correlated variables are History, Mathematics, Physics, PC, Chemistry, Foreign Languages, Cars. In the same plot, it is observed that chemistry and physics subjects were highly correlated; medicine and biology which are easily understood by human interpretation.

From movie's plot we understood that most of the animated movies were fantasy and fairy tales, while, horror movies were watched by almost every age group. Coming to music's plot, correlated attributes are Country, Opera, Hip hop…Rap etc. People don't change their musical choices often as they grow older. People were usually like metal music, likes to listen punk too. On the other hand, classical music, Folk and Opera form a team.

2) Age Vs Phobia correlation plot:

All the attributes in fig 2.5 are highly correlated with each other. Few attributes include Flying, Storm, Darkness, Heights, Spiders Few visualizations are created to find to out how many people have smoking and drinking habits and how these habits vary in different genders.

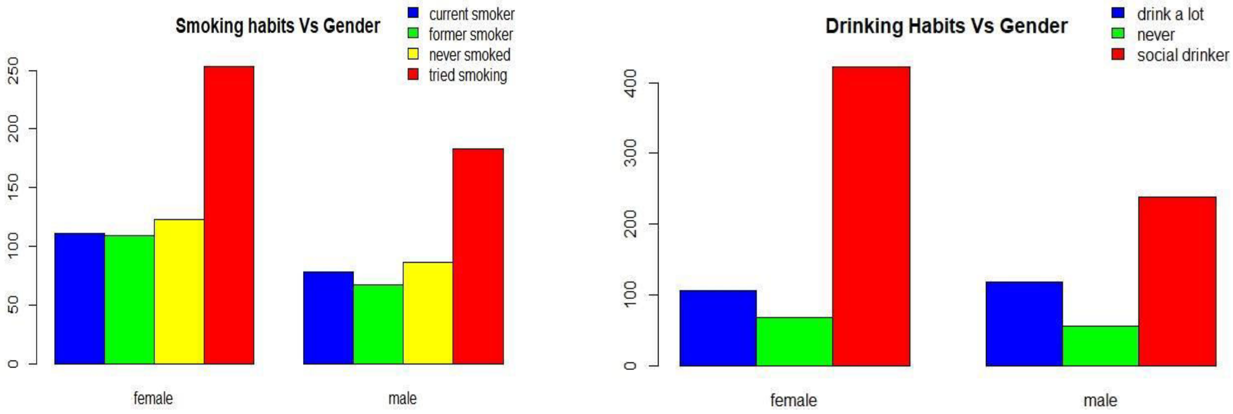


Fig: 2.2: Smoking habits & Drinking habits

From the figure above it is surprising to see that there are a greater number of females who have tried smoking than males and also overall it is females who does more smoking when compared to males. Surprising results from the above figure is that most of the social drinkers are female than males and males drink a lot when compared to females.

Modelling

3.1 Preprocessing:

There are null values in the dataset in different columns, which are to be handled for using the data for modelling. We can cannot delete the row or column with null values because valuable information in those columns or rows may be lost. Therefore, we have used K-nearest neighbor algorithm to replace the missing values which replaces the missing values in numerical data with mean of the column and mode of the column for the one with categorical data. Also, the columns which are categorical data in the dataset are to be converted to levels of factor variable using as.factor() in R.

3.2: Linear regression with age as predictor variable:

a) **Using all the attributes in the dataset:**
We try to build a linear regression model using all the attributes present in the dataset and try to predict age using this model and then we check the accuracy of the model using AIC, BIC, MSE and R-squared value. By observing the p-value of the F-statistic for the overall model it is less than 2.2e-16 which says that there are significant variables in the dataset which can be used to predict the age. But there are 150 attributes in total in the dataset, which makes it very difficult to select all the significant variables for the model, therefore we use backward selection method to select the optimal model with significant attributes.

b) **Using backward selection for the modelling:**
The backward selection makes modelling with every attribute starting from the end of the dataset and repeats the process till it reach starting attribute and prune the insignificant attributes from the model. The R-squared for optimal model i.e. backward selection is 0.6233 and p-value of F-statistic is less than 2.2e-16 and the significant variables in the model is shown in the figure above. Total of 50 attributes are significant among 150 attributes in the dataset.

Comparing results of model with all attributes and model with backward selection:
The MSE of 1st model is 5.39 and AIC, BIC of 1st model are 2671.94, 3368.23 also R-squared is 0.65. The MSE of second model is 4.81 and AIC, BIC are 2526.76, 2795.58 and R-squared is 0.62. Since the values of MSE, AIC, BIC are lesser for second model and also there is not much difference b/n R-squared. Therefore, we can say that model with backward selection is best for predicting the age.
We now try to improve the model performance of backward selection by applying PCA on the original dataset.

c) **Principal Component Analysis (PCA):**

PCA is a dimensionality reduction technique where principle components equal to number of columns are generated where each principal component are linear combination of all the variables and 1st PC explains more variance in the data and next PC explains slightly less variance in data and so on. More the variance explained in the data more is the amount of information or features known. Therefore 1st principal component always explains more variance than remaining components in the PCA.

For applying PCA on the data our data has to be normalized first because if we apply PCA on un-normalized data, principle components results in large loading of the data of the variables which have high variance, resulting in biased principle components towards high variance variables.

Now for dimensionality reduction we look at the cumulative sum of variance plot.

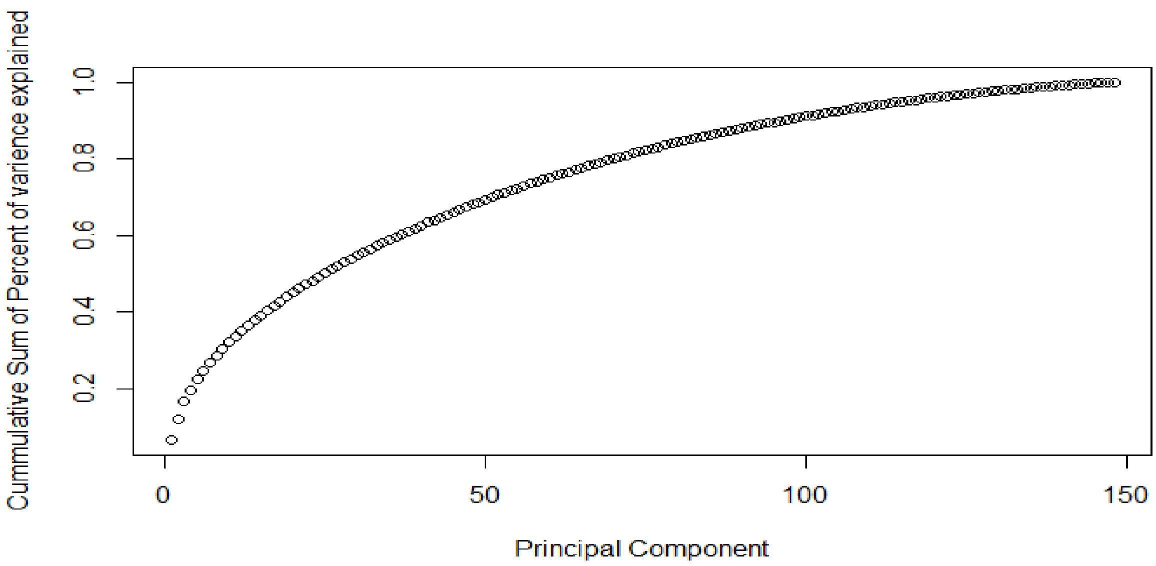


Fig.3.2.7: (c) Cumulative sum of variance VS No. of Principle Components

From the above figure we can say that after approximately 100 principle components the sum of variance explained is almost constant and 100 principal components explains about 90% variance in the dataset. Therefore, we select first 100 principle components for our model.
Now we apply linear regression with age as predictor variable and these 100 principle components as independent variables. Now when we apply test dataset on this model for predicting the age and if we calculate the MSE, AIC and BIC values of this model, as 1.76, 1647.09, 2096.59.

Comparing results of PCA model and backward selection model:
The R-squared value of PCA model is 0.922 and AIC, BIC and MSE on test are 1647.09, 2096.59 and 1.764.
The R-squared value of backward model is 0.62 and AIC, BIC and MSE on test are 2526.76, 2795.58 and 4.81. Therefore, by observing the above results PCA model has better performance due to low AIC, BIC

and MSE values and high value of R-square. The R-square says that about 92.2% of the data exactly fits the fitted line of the model.

3.3 Random Forest Regression:

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. This is performed on the significant variables in this project.

Model:

A random forest model was built using all of the prediction. A random forest was attempted with the random Forest function from the random Forest package. The random Forest method ran in under two minutes, and thus the random Forest method was used to build the random forest.

Results:

The resulting of Mean of squared residual 4.285593 and the %var is 46.09**.**

Distribution of minimal depth

The plot below shows the distribution of minimal depth among the trees of your forest. Note that:

- the mean of the distribution is marked by a vertical bar with a value label on it (the scale for it is different than for the rest of the plot),
- the scale of the X axis goes from zero to the maximum number of trees in which any variable was used for splitting.

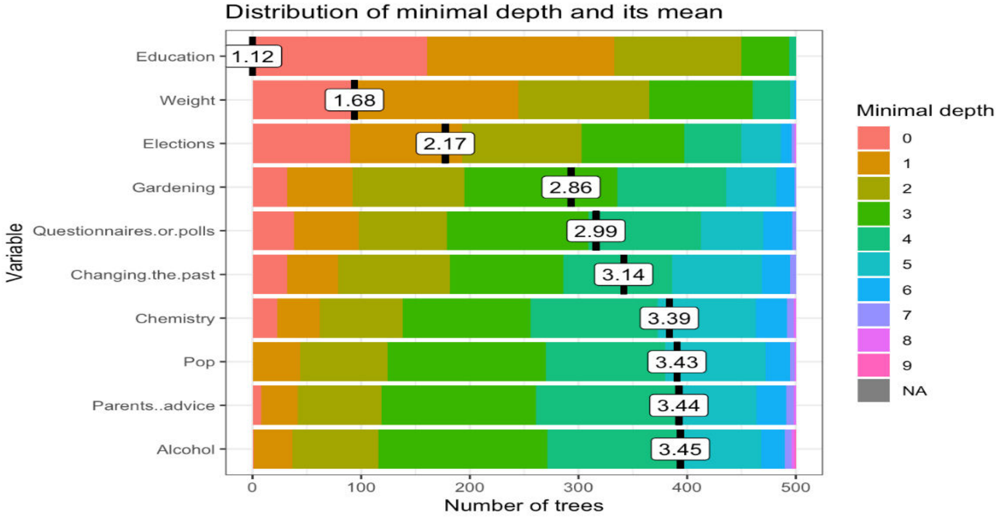


Fig.3.3.1: Distribution of minimal depth and its mean

Minimal depth for a variable in a tree equals to the depth of the node which splits on that variable and is the closest to the root of the tree. If it is low than a lot of observations are divided into groups on the basis of this variable.

Variable Importance:

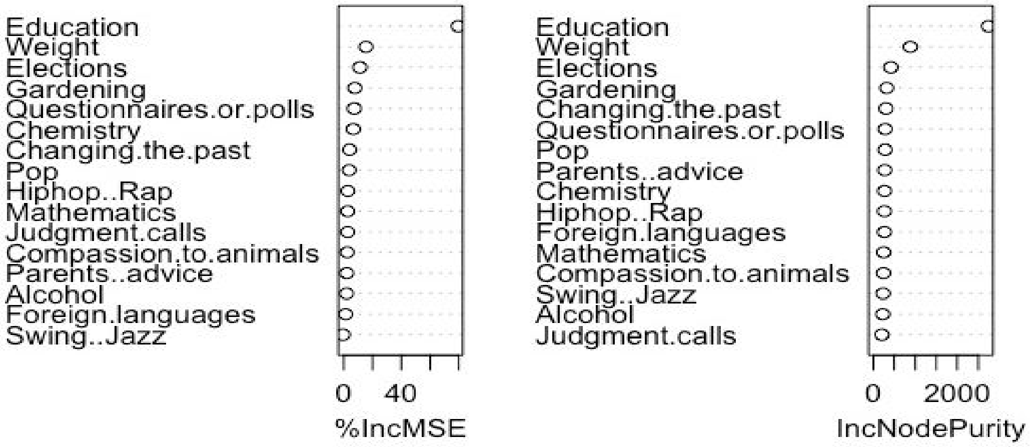


Fig.3.3.2: Variable Importance plot

From the above variable importance plot we can say that Education, Weight and Elections are the top three attributes in predicting the age of person.

Multi-way importance plot:

The multi-way importance plot shows the relation between three measures of importance and labels 10 variables which scored best when it comes to these three measures (i.e. for which the sum of the ranks for those measures is the lowest).

The first multiway importance plot focuses on three importance measures that derive from the structure of trees in the forest:

- mean depth of first split on the variable,
- number of trees in which the root is split on the variable,
- the total number of nodes in the forest that split on that variable.

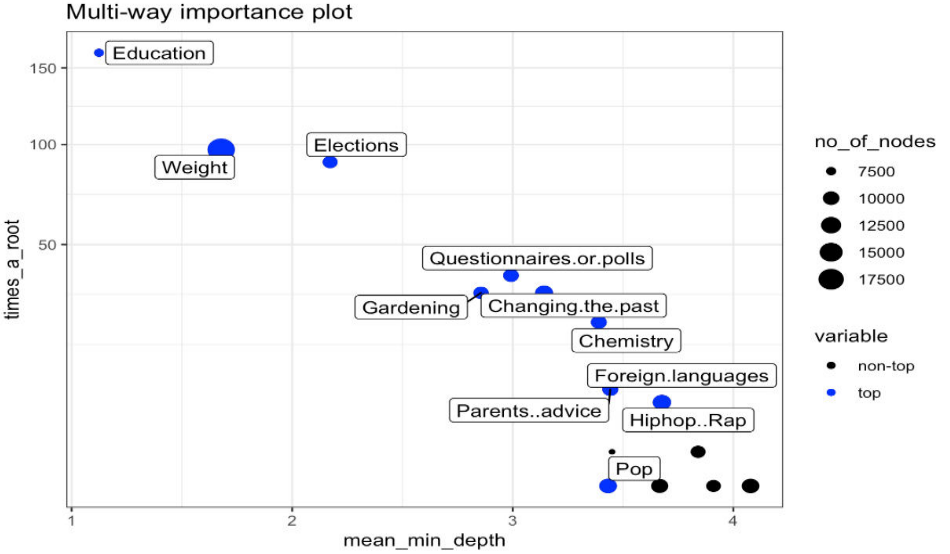


Fig.3.3.3: Multi-way importance plot

The second multi-way importance plot shows two importance measures that derive from the role a variable plays in prediction: with the additional information on the pp-value based on a binomial distribution of the number of nodes split on the variable assuming that variables are randomly drawn to form splits (i.e. if a variable is significant it means that the variable is used for splitting more often than would be the case if the selection was random).

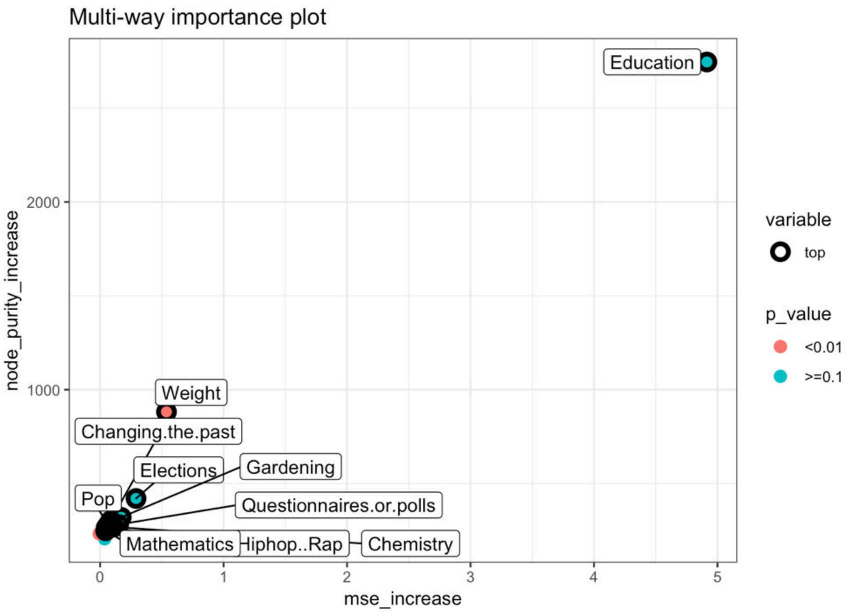


Fig.3.3.4: Multi-way importance plot

Compare importance measures

The plot below shows bilateral relations between the following importance measures: if some variables are strongly related to each other it may be worth to consider focusing only on one of them. “times_a_root” is correlated with “mean_min_depth” negatively while positively correlated with “mse_increase”.

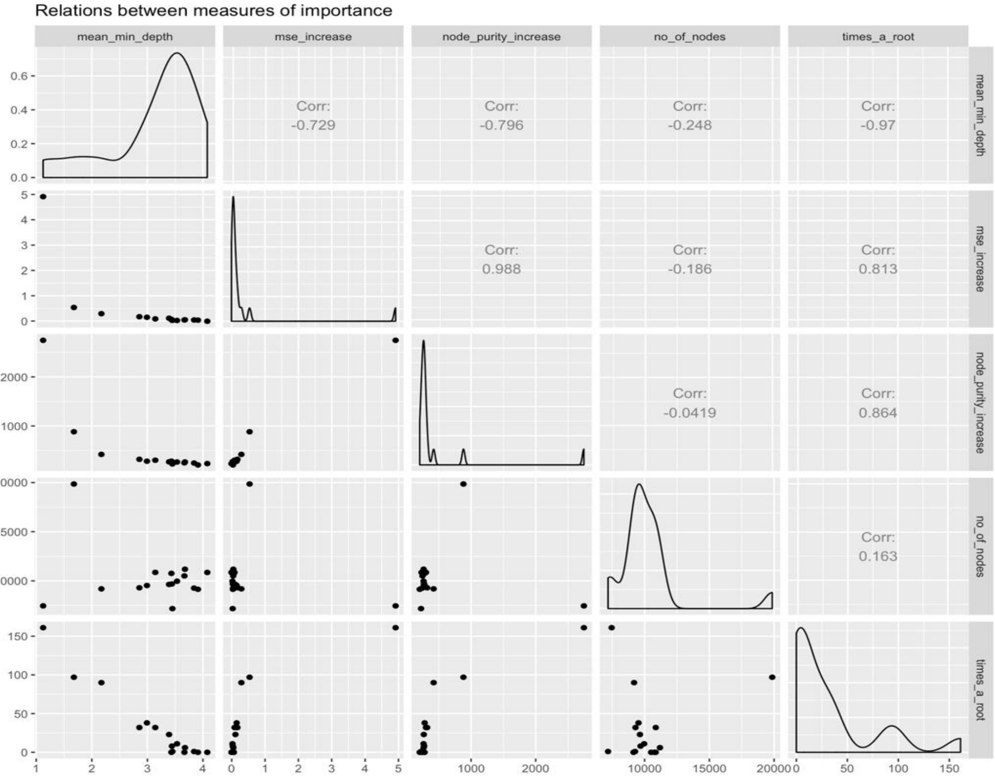


Fig.3.3.5: Relations b/m measures of importance

Compare rankings of variables

The plot below shows bilateral relations between the rankings of variables according to chosen importance measures. This approach might be useful as rankings are more evenly spread than corresponding importance measures. This may also more clearly show where the different measures of importance disagree or agree. In the following plot, “mean_min_depth” linearly increasing with increase in “node_purity_increase”. “node_purity_increase” is also positively correlated with “times_a_root”.

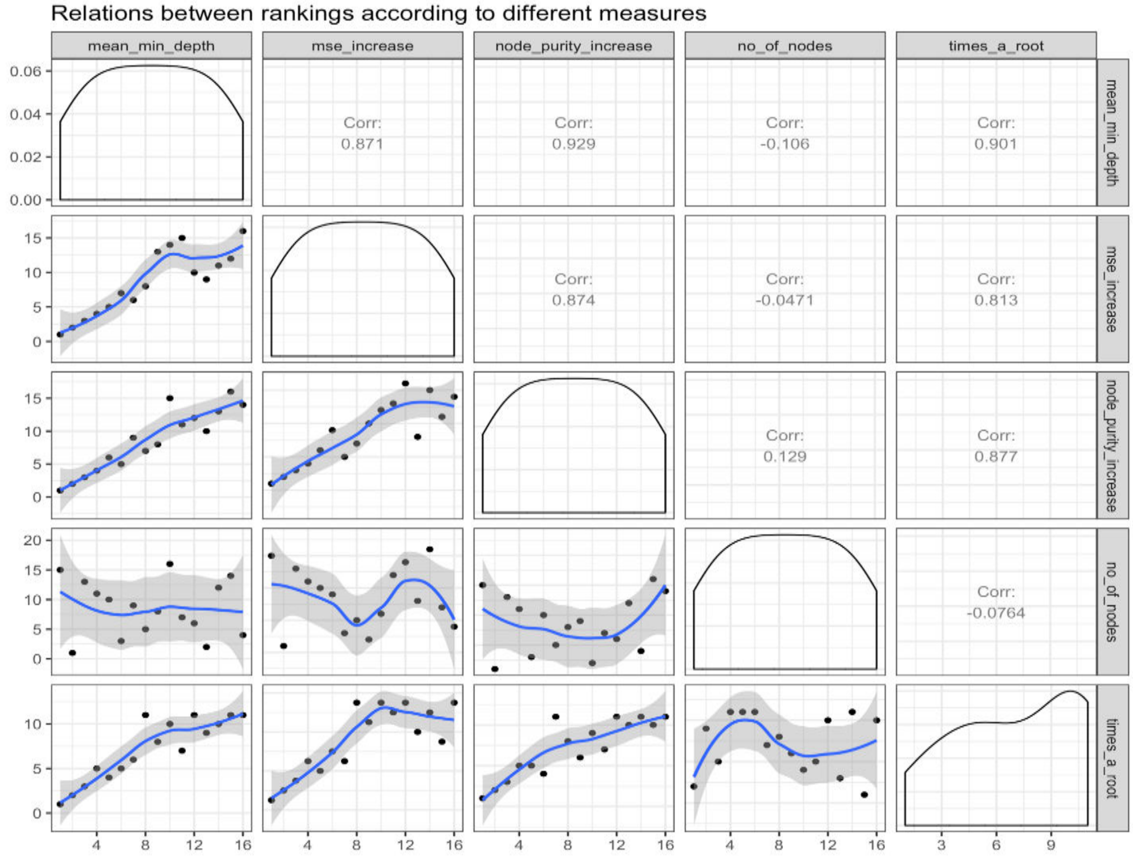


Fig.3.3.6: Relations b/n rankings according to different measures.

Variable interactions

Conditional minimal depth

The plot below reports 30 top interactions according to mean of conditional minimal depth – a generalization of minimal depth that measures the depth of the second variable in a tree of which the first variable is a root (a subtree of a tree from the forest). In order to be comparable to normal minimal depth 1 is subtracted so that 0 is the minimum.

For example, value of 0 for interaction x: y in a tree means that if we take the highest subtree with the root splitting on x then y is used for splitting immediately after x (minimal depth of x in this subtree is 1). The values presented are means over all trees in the forest.

Things to be noted that, the plot shows only 30 interactions that appeared most frequently. Secondly, the horizontal line shows the minimal value of the depicted statistic among interactions for which it was calculated. Thirdly, the interactions considered are ones with the following variables as first (root variables): and all possible values of the second variable.

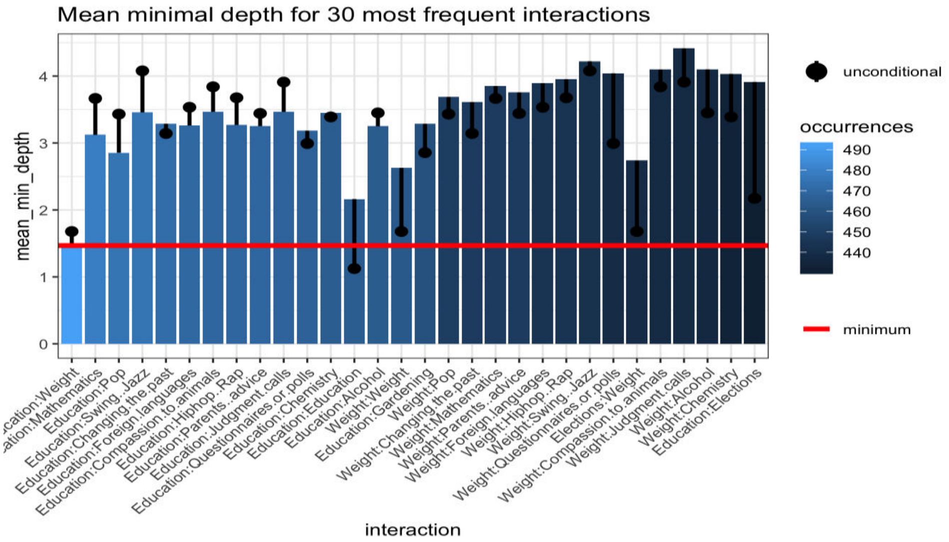


Fig.3.3.7: Mean minimal depth for 30 most frequent interactions

3.4 SVM and Ensemble models:

SVM:

Support Vector Machine is again another robust and flexible machine learning algorithm that helps in predicting Age in the young this dataset with the help of set of dependent variables. Generally, svm is used for classification purpose where hyperplane is used for dividing the classes of predictor variable. However, it can also be used for regression methodology, keeping complexity of algorithm in mind.

SVM model has been initially built on 148 variables initially, where the RMSE value is very high and low R-squared value. Hence, top 100 principle components were used in practice to build model. After looking at the prediction from train data only, it is found out that the RMSE value has been dropped to 2.41.

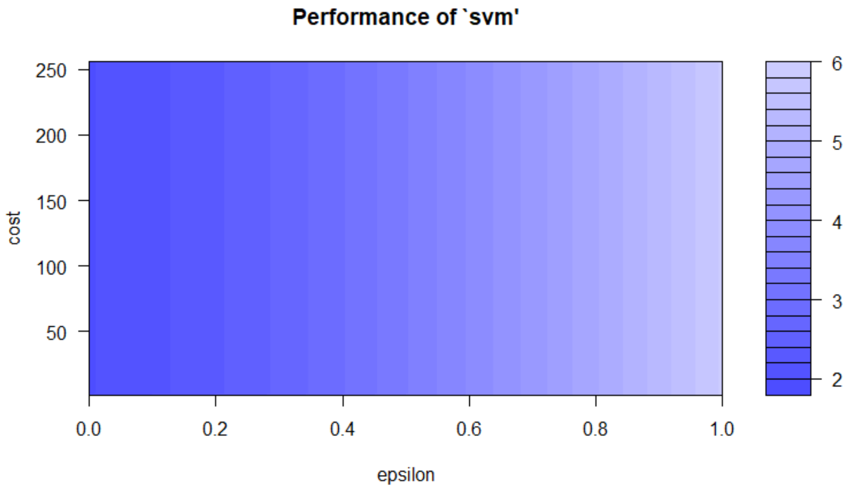


Fig.3.4.1: Performance on SVM

Model has been tuned for identifying parameters for best model shown in Fig.3.4.1 and performance depreciates with increase in epsilon value. It is observed that cost and epsilon were the important parameters that define performance of a model. With this, the best model is observed at 0 epsilon value and 4 cost. Now using tuned parameters another model has been created and has least RMSE scores compared to previous three models valued to 1.41.

3. 5 Comparing the models:

All best performing regression models like linear model, random forest, SVM and additionally knn were used for comparing and following has been obtained.

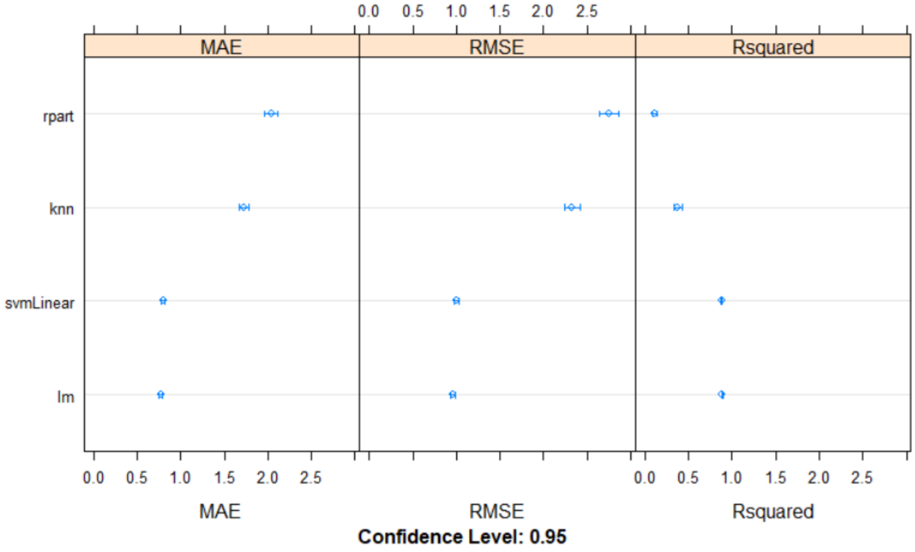


Fig 3.5.1: Comparison of models

A box plot has been created shown in fig 3.5.1 for performance measuring parameters and different models and found out that most suitable models were svm and linear model, while, least preferred model is random forest.

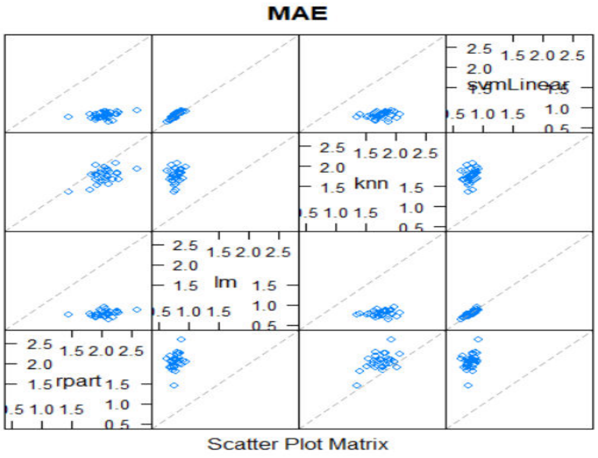


Fig 3.5.2: Scatter plot for MAE values for different models

On observing the scatter plot matrix for MAE values shown in fig 3.5.2, it is observed that linear model and svm linear are highly correlated compared to other three. While, knn and linear model has inverse relationship and were concentrated more in the fourth quadrant. Random forest continues to have higher MAE value and plot depicts that all the values lies between 2.0 to 2.5 irrespective of other three models.

3.6 Ensemble

Though, comparing models allow us to select best model for predicting the age, ensemble modeling allows us to combine the results of all previously used models and build another machine learning algorithm over it. It basically generates two levels of models by considering linear models, random forest and svm as predictors and build random forest on top of it. This model selects the best outcome from predictors at every datapoint and decreases variance and improves bias of total model.

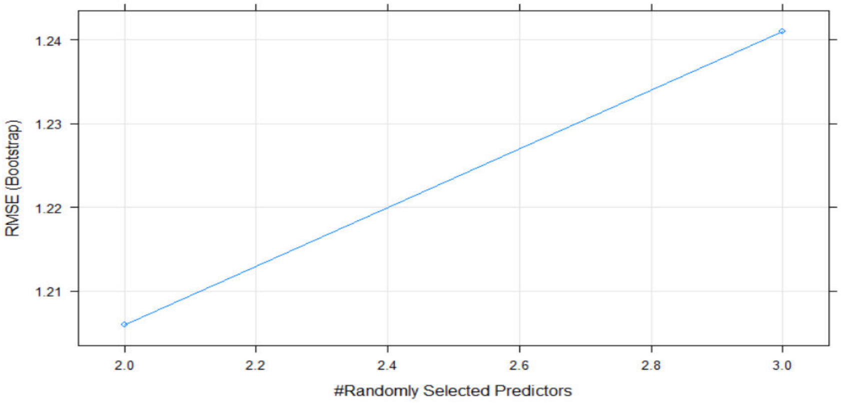


Fig 3.6.1: Graph plotted between Randomly selected predictors and RMSE values

From the fig 3.6.1, it is observed that RMSE values linearly increasing with number of predictors used for building a model having 1.24 RMSE value as its highest and around 0.4 as its least. However, prediction levels were best when two predictors were used to build a model and RMSE scores estimated to decrease to 0.42 for 2 predictors.

3.7 Classification Process

Regression analysis has been performed with significant results and capable of predicting the age. On the other side, in order to understand the efficiency of the attributes to predict young people age group, a new attribute has been created named “Age5” (Age Groups) wherein ages were divided into age and classification analysis has been performed on it. Generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables. We tried using logistics regression to find how different attributes are associated with different age groups of youth. We used clustering for grouping the target variable “Age”. Clustering helped us to find out the best possible number of splits that age can be divided into.

Clustering

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

Clustering Method

Partitioning Method: These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example *K*-means. Using Nbclust package from R we obtained the optimal K value.

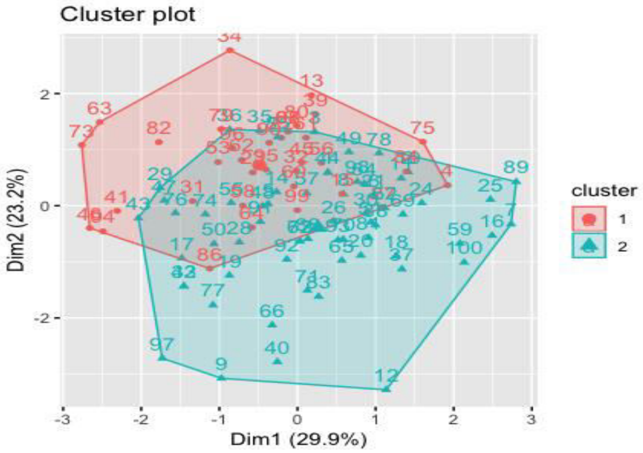
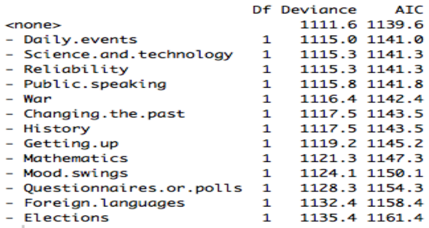


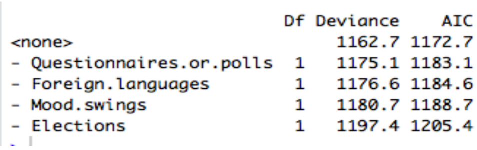
Fig.3.7.1: Clusters formed

The optimal K value for clustering of age group is 2. Using this as a base we built logistic regression model.

Backward Selection: We used backward selection to find prominent attributes that are more closely associated with the target variable "Age".



```
                          Df Deviance    AIC
<none>                         1111.6 1139.6
- Daily.events             1   1115.0 1141.0
- Science.and.technology   1   1115.3 1141.3
- Reliability              1   1115.3 1141.3
- Public.speaking          1   1115.8 1141.8
- War                      1   1116.4 1142.4
- Changing.the.past        1   1117.5 1143.5
- History                  1   1117.5 1143.5
- Getting.up               1   1119.2 1145.2
- Mathematics              1   1121.3 1147.3
- Mood.swings              1   1124.1 1150.1
- Questionnaires.or.polls  1   1128.3 1154.3
- Foreign.languages        1   1132.4 1158.4
- Elections                1   1135.4 1161.4
```



```
                          Df Deviance    AIC
<none>                         1162.7 1172.7
- Questionnaires.or.polls  1   1175.1 1183.1
- Foreign.languages        1   1176.6 1184.6
- Mood.swings              1   1180.7 1188.7
- Elections                1   1197.4 1205.4
```

Fig.3.7.2: Significant variables **Fig.3.7.3: AIC values**

From the above process the prominent attributes that are more closely associated with the age based on AIC values are presented in fig.3.7.3. We built Logistics regression model using these prominent predictor variables with age as target variable.

Logistic Regression Equation:

We grouped the age of youth into two groups. The age group from 14 – 19 as 0 and 20 – 30 as 1. We tried different age groups with different trials in clustering, but this was the best.

(p/ 1 – p) = 1.8455 - 0.2256 × Questionnaires.or. polls- 0.28482 × Mood.swings + 0.2617 × Elections – 0.2309 × Foreign.languages

Model Evaluation and Visualizations

For evaluating the model, we use confusion matrix to obtain accuracy. Using cut off value of 0.5, we built confusion matrix.

The accuracy if the model if found to be **68.3%**. We tried logistics regression but could not get better accuracy.

ROC Curve

To produce the ROC, curve the final logistic regression model was run on the full data set. The resulting ROC AUC on the validation set was 0.6885. The overall performance of a classifier, summarized over all possible thresholds, is given by the Area Under the ROC curve (AUC). ROC curves are useful for comparing different classifiers, since we consider all possible thresholds. An ideal ROC curve will hug the top left corner, so the larger area under the ROC curve the AUC the better the classifier. From fig.3.5.4, it is observed that model has lower sensitivity irrespective of the specificity values and it is expected for a classifier that performs well to have an AUC larger than 0.5.

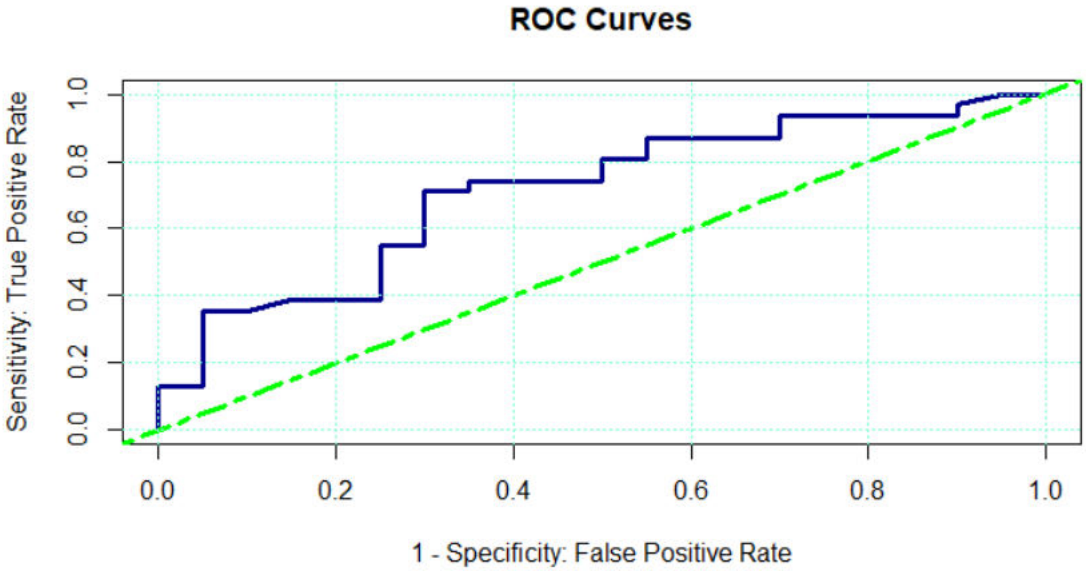


Fig.3.7.4: ROC Curves

Lift Chart

Lift charts is a graphical evaluative method for assessing and comparing the usefulness of classification models. It seeks to compare the response rates with and without using the classification model. The lift chart of the logistic model explains that lift value to be one for random data point while has value of 1.6 for first 0.2 percentile of the population. Lift value eventually drops down to the random lift value at 1% of population. From this we can say that the response rate has been improved by building up a model rather than choosing a data point at random.

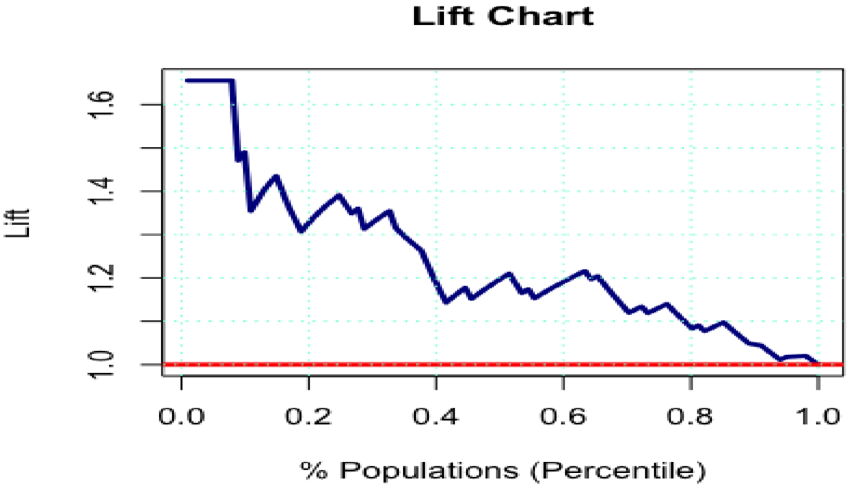


Fig.3.7.5: Lift chart

Conclusion

Hence, when the project was viewed in terms of a predictive model building; machine learning algorithms of both regression and classification models were implemented. Regression model, when implemented with pre-processed attributes, gives better results than the one with actual attributes. Regression models show a higher accuracy of predicting the age and ensemble 4 models by creating an upper-level regression model provided enhanced results with lower root mean square error values. An attempt has been made to perform classification analysis and this project has a future scope of improving the classification error rate.

While looking in the view of the dataset, we can conclude that awareness for education has been improved at a higher rate. Students improving their educational standards by not just stopping at secondary education. Secondly, the opinion for elections was changing time to time in a student life cycle. They are having different opinions at different phases of their teenage. Followingly, teens view on gardening and idea of changing the past play a prominent role in identifying the age of an individual.

Steps for executing the code

a) Download data from source mentioned in references (1)
b) Load it into system using code in line 2
c) Understand the data structure using preliminary descriptive analysis steps from line 3.
d) Perform pre-processing steps from line 19.
e) Run codes from line 136 to execute models and obtain results.

REFERENCES:

1. Data Source: https://www.kaggle.com/miroslavsabo/young-people-survey#responses.csv
2. “Machine Learning Techniques” by Adam Novotny published on 12th August, 2018 in https://medium.com/coinmonks/machine-learning-tutorial-2-training-f6f735830838
3. “Predicting Young People” by Henry Quan and Evan Feieresel published published on 1st February, 2018 in http://chaspari.engr.tamu.edu/wp-content/uploads/sites/147/2018/01/2_9.pdf
4. “7 Steps to Mastering Data Preparation” by Matthew Mayo published in 7th June, 2017 in https://www.kdnuggets.com/2017/06/7-steps-mastering-data-preparation-python.html
5. “How To Build an Ensemble of Machine Learning Algorithms in R” by Jason Brownlee published on 8th February, 2016 in https://machinelearningmastery.com/machine-learning-ensembles-with-r/
6. “Building Regression Models in R Using Support Vector Regression” published by Chaitanya Sagar in https://www.kdnuggets.com/2017/03/building-regression-models-support-vector-regression.html
7. “Support Vector Regression With R” by Alexandre Kowalczyk published on 23rd October, 2014 in https://www.svm-tutorial.com/2014/10/support-vector-regression-r/
8. “An Introduction To Clustering And Different Methods Of Clustering” by Saurav Kaushik published on 3rd November, 2016 in https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
9. “Practical Guide to Principle Component Analysis (PCA) in R and Python.” by Analytics Vidya. https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/
10. “K means Clustering in R -by Teja Kodali”- R bloggers. https://www.r-bloggers.com/k-means-clustering-in-r/