

­­

Project

Applications of Machine Learning

Universität Bayreuth

Fakultät für Mathematik, Physik, Informatik

Institut für Informatik

Lehrstuhl für Angewandte Informatik IV

By

Aneeba Khalid

Daniyal Bin Sajid

Muhammad Talal Nasir

Owais Anwar

Shayan Mahmood

From Bayreuth

1. Gutachter: Prof. Dr.-Ing. Stefan Jablonski

2. Gutachter: Dr. Stefan Schönig

Supervisor: Martin Käppel und Nicolai Schützenmeier

Table of Contents

[2 Introduction 25](#_Toc16452965)

[3 Motivation 26](#_Toc16452966)

[4 Tools and Technologies 26](#_Toc16452967)

[5 Data Description 27](#_Toc16452968)

[5.1 Dimensions 27](#_Toc16452969)

[5.2 Data Types 27](#_Toc16452970)

[6 Data Preprocessing 30](#_Toc16452971)

[6.1 Feature Scaling: 30](#_Toc16452972)

[6.2 Data Cleaning: 30](#_Toc16452973)

[6.2.1 Missing Values (imputation) 30](#_Toc16452974)

[7 Machine Learning Algorithms 31](#_Toc16452975)

[7.1 Regression: 31](#_Toc16452976)

[7.1.1 Linear Regression 31](#_Toc16452977)

[7.1.2 Ridge Regression 31](#_Toc16452978)

[7.1.3 Lasso Regression 31](#_Toc16452979)

[7.1.4 Support Vector Machine (Regressor) 32](#_Toc16452980)

[7.1.5 Guassian Naive Bayes 32](#_Toc16452981)

[7.1.6 Decision Tree Regressor 32](#_Toc16452982)

[7.1.7 Random Forest Regressor 32](#_Toc16452983)

[7.1.8 K-Nearest Neighbor Regressor 32](#_Toc16452984)

[7.1.9 Neural Network Regressor 32](#_Toc16452985)

[7.2 Classification: 33](#_Toc16452986)

[7.2.1 Logistic Regression 33](#_Toc16452987)

[7.2.2 Support Vector Machine (Classifier) 33](#_Toc16452988)

[7.2.3 Guassian Naive Bayes 33](#_Toc16452989)

[7.2.4 Decision Tree Classifier 33](#_Toc16452990)

[7.2.5 Random Forest Classifier 33](#_Toc16452991)

[7.2.6 K-Nearest Neighbor (Classifier) 33](#_Toc16452992)

[7.2.7 Neural Network Classifier 33](#_Toc16452993)

[7.3 Dimensionality Reduction: 34](#_Toc16452994)

[7.3.1 Principle Component Analysis 34](#_Toc16452995)

[8 Feature Selection 34](#_Toc16452996)

[8.1 Overview of Feature Selection 34](#_Toc16452997)

[8.2 Feature Selection Methods 34](#_Toc16452998)

[8.3 Feature Selection Techniques Used 35](#_Toc16452999)

[8.3.1 Sequential Forward Selection 35](#_Toc16453000)

[8.3.2 Recursive Feature Elimination 35](#_Toc16453001)

[8.3.3 Correlation Method 35](#_Toc16453002)

[8.3.4 LASSO Feature Selection 36](#_Toc16453003)

[8.3.5 Feature Importance 36](#_Toc16453004)

[9 Data Splitting 36](#_Toc16453005)

[10 Validation Techniques 36](#_Toc16453006)

[10.1 K-Folds 36](#_Toc16453007)

[10.2 Stratified K-folds 37](#_Toc16453008)

[10.3 Tuning HyperParameters 37](#_Toc16453009)

[10.3.1 Grid Search CV 37](#_Toc16453010)

[11 Evaluation Metrics 37](#_Toc16453011)

[11.1 Clasiification Metrics 37](#_Toc16453012)

[11.1.1 Confusion Matrix 37](#_Toc16453013)

[11.1.2 Accuracy Score 38](#_Toc16453014)

[11.1.3 Precision Score 38](#_Toc16453015)

[11.1.4 Recall Score 38](#_Toc16453016)

[11.1.5 F1\_Score 38](#_Toc16453017)

[11.1.6 Roc\_auc Curve 38](#_Toc16453018)

[11.2 Regression Metrics 38](#_Toc16453019)

[11.2.1 Root Mean Square Error (RMSE) 38](#_Toc16453020)

[11.2.2 Mean Squared Error 38](#_Toc16453021)

[11.2.3 Mean Absolute Error 38](#_Toc16453022)

[11.2.4 R2\_Score 39](#_Toc16453023)

[12 Problems 39](#_Toc16453024)

[12.1 Problem 1 39](#_Toc16453025)

[12.1.1 Explanation: 39](#_Toc16453026)

[12.1.5 Predictors: 41](#_Toc16453032)

[12.1.6 Model Evaluation: 41](#_Toc16453033)

[12.1.7 Parameter Tuning: 42](#_Toc16453039)

[12.1.8 Model Evaluation after Tuning: 42](#_Toc16453040)

[12.1.9 Results: 43](#_Toc16453041)

[12.2 Problem 2. 43](#_Toc16453042)

[12.2.1 Explanation: 43](#_Toc16453043)

[12.2.2 Target Feature: 43](#_Toc16453045)

[12.2.3 Predictors: 43](#_Toc16453047)

[12.2.4 Train Test Split: 43](#_Toc16453049)

[12.2.5 Feature Selection: 43](#_Toc16453051)

# Introduction

With the evolution in technology over the period of 10 years, Data has become an essential part of businesses. Regardless of the type of industry, Data has the key role to play with success of every business. With the enormous increase in data it has become difficult for human to process and analyze it and to overcome this thing an idea of Machine Learning came into existence. Machine Learning is a sub-domain of computer science which is principally based on the study of pattern recognition and Artificial Intelligence.

Machine Learn in the same way as Humans do but the only difference is that, it is much faster and cater much data in one time to process which has made it extremely popular among businesses. The core principles are same as that of statistics and mathematics behind this revolutionary technology. Machine Learning is classified in three categories. Supervised Learning, Unsupervised Learning and Reinforcement Learning.

Supervised Learning is the oldest yet popular technique of machine learning. It has labelled data in it and the analysis is performed accordingly to forecast future trends using present and past scenarios. In layman’s term, Supervised Learning is known as the process of concept learning.

In this project we will work on supervised learning concept of Machine Learning to analyze different trends and to come up with certain results which can be used to solve different problem statements on given dataset.

In order to work under the umbrella of Machine Learning, there should be a problem statement which tells what exactly we want to extract out of this beautiful invention. So, our major goal is to analyze relationship between financial crisis and the voting behavior of the people of different countries or different regions. This is entirely the scope of our project and we have to perform data analysis on given dataset using Machine Learning. During the analysis process we will come up with different trends and thoughtful information which can be really helpful in making the decision making process of concern authorities more strong and they can come up with solutions which can help them preventing the financial crisis and to examine the voting trends which has a great impact on country or region stability.

# Motivation

Our main goal is to incorporate our working on Machine Learning to help authorities of different nations to make their decision-making process strong on certain issues. With our results they might be able to understand the current situation of financial crisis or voting trends and how they are affected or evolved from past. Some steps could also be taken to overcome the issues for future perspective. Our analysis can even bring certain imbalances among regional economies and how to overcome them to make this world peaceful for human race. Governments can have indicators to determine what steps should be taken to make their citizen happy and to strengthen democracy. Banks will be able to avoid financial crisis and political leaderships will be able to make decisions which make them strong to face unseen threats. Impact of recessions can be reduced by just predicting the future trends from past experiences.

In simple words, Data analysis using technology can help society to make predictions about future and to overcome the mistakes done before also can make our present more pleasant.

# Tools and Technologies

We consider PyCharm and Jupyter as our IDE and using Python Programing language including different libraries which are as follows;

Pandas: it is an open source library. It provides data analysis tools, high performance and easy to use data structures for python programing language. We are using pandas for reading our dataset.

NunPy: it is the fundamental package for scientific computing in python. We are using it for some scientific functions like mean, mode and median etc.

Matplotlib: it is a Python 2D plotting library, you can generate plots, histograms bar charts scatter plots, etc. we are using matplotlib for plotting different type of visualization figures, like ROC curve etc.

Seaborn: it is a python data visualization library based on matplotlib. It provides a high-level interface for visualization of informative statistical graphs. We are using seaborn for different type of plotting.

Mlxtend: it is a python library of useful tools for the day-to-day data science tasks. We are using mlxtend for our feature selection (sequential forward selection) and plotting of the selection.

Sklearn: it is very important library for machine learning, built on NumPy, SciPy and matplotlib. It contains wide range of models, preprocessing techniques and model selection. We are using Sklearn for our models, model selection and preprocessing.

# Data Description

We are given a data set with electoral and financial features of 20 different countries from all over the world. This data is from year 1870 to 2014.

Countries that are included in this data are USA, Australia, Japan, Italy, France, Germany, Belgium, Switzerland, Belgium, Denmark, Netherlands, Britain, Sweden, Norway, Finland, Spain, Ireland, Greece, Portugal and Spain.

## Dimensions

Shape of this data is (2900,28), 2900 rows and 28 columns.

## Data Types

Continuous

Discrete

Categorical

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Description | | Missing Values | |
| ccode | Numeric country code | | 0 | |
| iso | Three letter WB-ISO code | | 0 | |
| year | 1870-2014 | | 0 | |
| govvote | Government vote share in most recent general  election to the (lower) legislative chamber | | 822 | |
| oppvote | Opposition vote share in most recent general election to the (lower) legislative chamber | | 822 | |
| frac | Fractionalization of (lower) legislative chamber | | 390 | |
| partycount | Number of parties in (lower) legislative chamber | | 405 | |
| right | Sum of far-right votes in most recent general election to the (lower) legislative chamber | | 1106 | |
| left | Sum of far-left votes in most recent general election to the (lower) legislative chamber | | 1106 | |
| extr | Sum of far-right and far-left votes in most recent general election to the (lower) legislative chamber | | 1106 | |
| protests | Number of street protest incidents per year | | 1173 | |
| PROTESTDEV | Street protest incidents per year, percentage deviation from trend | | 1159 | |
| DEMOSDEV | Demonstrations per year, percentage deviation from trend | | 1159 | |
| RIOTSDEV | Violent riots per year, percentage deviation from trend | | 1159 | |
| sTRIKESDEV | General strikes per year, percentage deviation from trend | | 1159 | |
| RGDP | Real GDP per capita | | 53 | |
| GDPPEAK | “1” if real GDP per capita peak in this year | | 0 | |
| crisisJST | “1” if systemic financial crisis in this year | | 0 | |
| pk\_fin | “1” if financial recession starts in this year | | 0 | |
| pk\_norm | “1” if non-financial recession starts in this year | | 0 | |
| pk\_dis | “1” if non-financial macroeconomic disaster starts in this year | | 0 | |
| CPI | Consumer Price Index (inflation rate) | | 307 | |
| govcris | Number of major government crises per year | | 1173 | |
| vetopl | Veto player index | 1954 | |
| dict | “1” if dictatorship in this year | 0 | |
| election | “1” if election in this year | 148 | |
| election\_year | Year of most recent election | 148 | |
| turnover | “1” if change in the effective executive in this year | 328 | |

# Data Preprocessing

## Feature Scaling:

In some problems we need to scale the data to make it more useful and to reduce the anomalies where they are not needed. Like for dimensionality reduction scaling of the data is a necessity. For scaling we are using Standard Scaling.

## Data Cleaning:

Data cleaning is one of the major parts of the machine learning, as with unsuitable and missing values and string values we cannot apply any models on the data. For that we need to convert these values into numerical values. In our dataset all the values are already numerical or binary but where we are using some external data set, we converted continuous values into binary values using if conditions.

### Missing Values (imputation)

In our data set we handled the missing values according to the target value. If its continuous than we used mean or Interpolation and where problem statement works on classification, we used median and mode.

#### Mean

This function takes the average value of all the available values in that column and fill all the missing values with that average value.

#### Median

This function arranges all the values in ascending order than take the middle value of that column and fill all the missing values with that number.

#### Mode

This function takes the more often repeated value in that column and fill all the missing values with that number.

#### Interpolation

This function finds the estimated values between the two available points.

# Machine Learning Algorithms

There are different classes for Machine learning algorithms like Classification, Regression, Dimensionality Reduction and Clustering. In our scope of problems only classification, regression and dimensionality reduction techniques are used.

## Regression:

Regression algorithms are used to find the relationship between dependent variables and independent variable. It is used for Prediction and Forecasting of dependent variable. It deals with the continuous type of dependent values

### Linear Regression

It is a supervised Machine Learning model. This model is used to predict the continuous dependent variable using the independent variable’s (x). X variables and y predictors should have a linear relationship. It is a regression algorithm.

Y=mx+c. With x as independent variable, y as dependent variable, c as y-intercept and m as slope of the line.

### Ridge Regression

It is regression model like Linear Regression with x as independent continuous variable and y as dependent. This model is used to reduce the effect of overfitting by modifying the slope of the line. (y=mx+c). This model multiply (lambda) with the slope m of the line to keep it fitted to the data.

### Lasso Regression

It is regression model used for supervised learning. It works like linear and ridge regression with continuous variables. This model works better if some independent variables are not related to the y variable. As it multiply (lambda) with the slope of the line to reduce overfitting and take the absolute value for the x variable and eliminate the unrelated features. That is why it is also used as the feature selection model.

### Support Vector Machine (Regressor)

Support vector regression follows the same procedure as SVM use in classification. For continuous values this model set a limit (epsilon). This model solve regression problem as optimization problem by making a hyperplane on epsilon values with best continued value function.

### Guassian Naive Bayes

A Naive Bayes algorithm (NB) is a probabilistic machine learning model. As by its name, NB algorithm works on Naïve approach that the presence of one particular feature does not affect the other. It works on Bayes theorem of probability to predict the class of unknown data sets. Naive Bayes model is easy to build and particularly useful for very large data sets



* P(A|B) is the posterior probability of class (A, target) given predictor (B, attributes).
* P(A) is the prior probability of class.
* P(B|A) is the possibility which is the probability of predictor given class.
* P(A) is the preliminary probability of predictor.

we find the probability of occurrence of A, provided that B has occurred. B is the feature and A is the hypothesis.

### Decision Tree Regressor

Decision tree builds regression model in a form of tree structure that help to breakdown a dataset into smaller and smaller subset and an associated decision tree is incrementally developed. In the end giving a tree with decision nodes and leaf nodes. The target Variable can take the continuous values typically real numbers as it defines ranges to separate the branches of the tree.

### Random Forest Regressor

As the name says this model is a collection of trees. Each tree works as a decision tree and gives a result and all the results given by these committee of trees are evaluated and result with the highest weight is selected. Random forest regressor deals with the continuous features as it defines some limits to differentiate the value between branches of tree.

### K-Nearest Neighbor Regressor

K-nearest neighbors computes the Euclidean Distance to find similarity and average to predict an unseen value. the output value is the average of the values of k nearest neighbors.  
for example if k=3, then output value will be the average of that 3 neighbor’s value.

### Neural Network Regressor

Neural Networks models are designed inspired by human brain. With nodes as neurons and they are connected using edges. These edges will be given weights as it proceeds to learn. It has different layers of nodes. It also runs a backpropagation algorithm to check its answers and learn from it and then again weight the edges accordingly. Neural network Regressor deals with the continuous features and target values.

## Classification:

Classification is a task used for Descriptive and Predictive Models. These algorithms classify data into different categories. A target value that is a discrete data set is evaluated using the known features (x values). It deals with the categorical type of dependent values

### Logistic Regression

It is a classification model used in binary classification. As this model divide the feature values into two classes. This model sets a threshold if a value is smaller than that threshold it classifies it as class one and if value is greater than classify it as class two. It is used in supervised learning

### Support Vector Machine (Classifier)

This is supervised learning-based classification model. This model defines a hyperplane between the different classes to categorize them. This model not only use a line as a hyperplane but also a circular plane if features are mingled in each other. This model finds the best dimension where it is easier to categorise the features into different classes and make the best fit hyperplane between these classes for forecasting of new values.

### Guassian Naive Bayes

It works similar as for regression problem, Naive Bayes model is easy to build and particularly useful for very large data sets. Along with the simplicity, Naive Bayes is known to outperform even highly complex classification methods.

### Decision Tree Classifier

Decision tree algorithms works on the supervised learning. It can be used to solve the classification problem in which the data is split continuously according to certain parameter, it is also consists of Edges/Branch. The outcome is a variable like “fit” or “unfit” so the decisions variable can be Categorical or discrete

### Random Forest Classifier

As the name says this model is a collection of trees. Each tree works as a decision tree and gives a class as an output and all the outputs given by these committee of trees are evaluated and output with the highest weight is selected. Then the class given by most of the trees is considered as result. This evaluation by different trees reduce the mistakes made by trees individually. It is supervised learning-based model.

### K-Nearest Neighbor (Classifier)

K Nearest Neighbor (KNN) is a very simple, easy to understand model. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.  
for example if k = 1, then the object is simply assigned to the class of that single nearest neighbor.

### Neural Network Classifier

Neural Networks models are designed inspired by human brain. With nodes as neurons and they are connected using edges. These edges will be given weights as it proceeds to learn. It has different layers of nodes. It also runs a backpropagation algorithm to check its answers and learn from it and then again weight the edges accordingly. Neural Network classifier is designed to deal with categorical and binary features.

## Dimensionality Reduction:

Dimensionality reduction models are used to reduce the number of features and extract the more related ones. It is also used for visualization of results when it is more than 3 dimensional

### Principle Component Analysis

This method is used to reduce the dimensions of data to make it simpler to interpret, keeping the information it has. It uses orthogonal transformation to reduce the data features and combine the most correlated features to make a new feature. These new features are known as principle components.

# Feature Selection

## Overview of Feature Selection

Feature Selection is the process to select those features for a model that better represent a problem. Feature selection is used to reduce the number of explanatory features that describe a response feature. The main reasons why feature selection is used are as follows:

1. Make the model easier to explicate, removing features that are unneeded and do not add any knowledge to the data.
2. Reduce the complexity of the problem and enable models to work faster
3. Reduce overfitting.

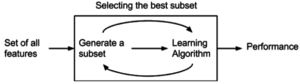
## Feature Selection Methods

In literature, there are several types of methods for feature selection (Filter, Wrapper and Embedded Methods).

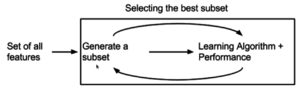
Filter Methods select the features by ranking them on how useful they are for the model, to compute the usefulness score, different statistical test and correlation results are taken into account (e.g. Chi-square, ANOVA, Pearson’s correlation).



Wrapper Methods generate different subsets of features, each subset is then used to train the learning algorithm and build a model. The best feature subset is determined by testing the model. To determine the feature subsets, different techniques are used (e.g. Forward feature selection, Backward feature elimination, Recursive feature elimination).



Embedded Methods (EM) are a combination between the filter and Wrapper methods, EM learn which features contribute best to the performance of the algorithm while the model is being created and determine the best feature. Regularization methods (e.g. L1 LASSO regularization) are used in which additional constraints are introduced into predictive algorithm (such as a regression algorithm) optimization that bias the model toward lower complexity (fewer coefficients).



## Feature Selection Techniques Used

### Sequential Forward Selection

SFS is an iterative method which start by having no features. It adds the features on each iteration that maximizes the performance of our model in accordance with the dependent variable. SFS performs best when the optimal subset is not huge.

### Recursive Feature Elimination

RFE gives the solution by developing the subsequent models with features until all the features are explored. It works by recursively removing attributes and building a model on those attributes that remain. It uses the accuracy of model to identify which features (and combination of features) most contribute in predicting the target feature.

### Correlation Method

Correlation explains how the features are in relation (Positive or negative). Positive Correlation explains the increase in one value of the feature increases the value of dependent feature. Negative Correlation explains increase in one value feature decreases the value of dependent variable.

### LASSO Feature Selection

Lasso Feature Selection is one of the simple approach to reduce model complexity and prevent over-fitting which may result regression. By Cross validation. It selects the penalty factor that ensures the model will generalize well for future data samples

### Feature Importance

Sklearn ensembes come with Tree based Classifiers (e.g. Random Forest or Extra Trees) with an inbuilt class of Feature importance that computes the relative importance of each attribute. Importance values are used to inform a feature selection process.

# Data Splitting

To train different algorithms on the dataset, data is split into training and testing set. We are using ‘train\_test\_split’ function of library Sklearn. Its split arrays or matrices into random train and test subsets. we are splitting our dataset into 30% of testing set and 70% of training set and using ‘random state’ parameter as ‘42’ which depend order of splitting of our data. Training set is used to train our model and testing set is used to test our trained model

# Validation Techniques

It is used to validate our model and to reduce over fitting from our model. There are some type of validation which we have used in our experiments.

## K-Folds

We use k-folds cross validation to reduce the overfitting, cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. it will split original data set into k subsets and uses one of the subset as testing set and other remaining as training set. This process end’s when every subset is used as testing set.

## Stratified K-folds

Stratification is the process of rearranging data, it ensures that each fold is good representation of data, generally a better approach when dealing with bias and variance.

In stratified k-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds. In the case of binary classification, this means that each fold contains roughly the same proportions of the two types of class labels

## Tuning HyperParameters

Parameter can be tuned to optimize the model which is done by parameter tuning methods.

### Grid Search CV

This technique is used to tune the parameters of models so that they will perform better on dataset. It explores the range of parameters and find the best combination for a given model, for examples; parameters which can be tuned in support vector machine are solver, kernel and gamma values.

The goal is then to train each of these models and evaluate them e.g. using cross-validation and select the best performed model.

# Evaluation Metrics

Evaluating a machine learning algorithm is an important part of any project. Your Model then can give you satisfying results when its evaluated, following are the Evaluation Techniques that we use in our project

## Clasiification Metrics

Following are the classification metrics in Scikit-Learn in Python

### Confusion Matrix

The table of confusion matrix is often used for explaining the performance of a classification model or for classifier on the set of test data for which the true values are to be known. It also helps to visualize the performance of an algorithm

### Accuracy Score

The accuracy classification score is a multilabel classification in which the functions computes the subset of the accuracy. The set of labels that are predicted for a sample must exactly match the corresponding sets of labels in y-true

(TruePostive + TrueNegative)/

(TruePostive + TrueNegative+FalsePostive+FalseNegative)

### Precision Score

Precision refers to the percentage of your result that are relevant, it’s the ratio of correctly predicted positive observation to the total predicted positive observation.

### Recall Score

Recall refers to the percentage of total relevant result correctly classified by our algorithms,

TruePostive/ (TruePostive+ FalseNegative)

### F1\_Score

Using of F1\_score given advantage that it incorporates both precision and recall into a single metric, with a high F1\_score means that it’s a sign pf well-performing model even you might have imbalanced data. It is a Harmonic mean of precision and recall

### Roc\_auc Curve

The Receiver Operating Characteristic Curve defines the plot of the true positive versus the false positive rate of the prediction of a model for the multiple thresholds between 0.0 and 1.0

## Regression Metrics

In a regression task the model learns to predict numeric scores e.g. predicting the price of stock on future days by the given past history or the information about the previous price of stock and the market prices.

### Root Mean Square Error (RMSE)

The most frequently used metric for regression tasks is RMSE, it’s defined as the square root of the average squared distance between the predicted score and the actual score

### Mean Squared Error

It tells us how close a regression is to a set of points. It works by taking the distance from the point to the regression line (distance are the errors) and squaring them.

### Mean Absolute Error

In statistics MAE defines the measure of difference between the two continuous variables, as the name defines the mean absolute error is an average of the absolute errors.

### R2\_Score

It’s a statistical measure of how close the data is fitted to the regression line also called as the coefficient of determination. It’s the percentage of the response variable variation that is explained by the linear model.

R-squared = explained variation / Total Variation.

# Problems

## Problem 1 (AK)

Can we predict financial recession in the region of Europe before hand?

### Explanation:

I deduced a data frame with 17 European countries:

Italy, France, Germany, Belgium, Switzerland, Belgium, Denmark, Netherlands, Britain, Sweden, Norway, Finland, Spain, Ireland, Greece, Portugal and Spain.

I handled the missing values using median function which fills all the not available values with the median value of that specific column. I use median because target value is discrete with 0 and 1 values so it will fill values with float.

### Target Feature:

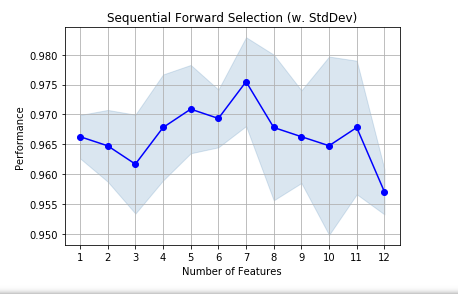
Financial Recession (pk\_fin). We will be applying Classification Models as our target is binary value.

### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 75% and Testing set 25%.

### Feature Selection:

As my question is about financial recession prediction in Europe, so I use the best features selected by Sequential Forward Selection

In the start I gave Sequential Feature Selection Algorithm 12 features to check and select the best ones

This feature selection technique is forward selection. It adds and check features one by one in a list and gives accuracy score at each step to give you elaborated idea of feature selection there is a table step wise calculation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature Index | Feature Names | Average Score |
| Step 1 | 0 | Govcris | 96.6% |
| Step 2 | 0 , 9 | Govcris, Turnover | 96.4% |
| Step 3 | 0 , 1 , 9 | Govcris, CrisisJST, Turnover | 96.1% |
| Step 4 | 0 , 1 , 6 , 9 | Govcris, CrisisJST, GDPpeak, Turnover | 96.7% |
| Step 5 | 0 , 1 , 3 , 6 , 9 | Govcris, CrisisJST, Right, GDPpeak, Turnover | 97% |
| Step 6 | 0, 1, 2, 3, 6, 9 | Govcris, CrisisJST,  Partycount, Right, GDPpeak, Turnover | 96.9% |
| Step 7 | 0, 1, 2, 3, 4, 6, 9 | Govcris, CrisisJST, Partycount, Right, Left, GDPpeak, Turnover | 97.5% |

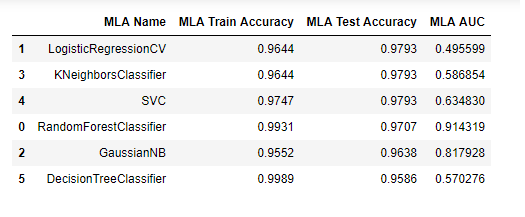
At step 7 it gives the best features with maximum average score, so the sequence of feature I used in algorithms are 7.

### Predictors:

Government crisis, Crisis JST, Partycount, Right, Left, GDP peak, Turnover

### Model Evaluation:

Algorithms I used are for classification because my Target values is discrete value, Financial recession has the value 0 (there was no recession) and 1 (there was a recession).



This table is arranged according to best Test Accuracy values because good Test accuracies gives the best prediction of the answers.

#### Roc Auc curve

Roc\_Auc curve is the graphical representation of confusion matrix with False Positive Rate at x-axis and True Positive Rate at y-axis and from (0,0) to (1,1). AUC stands for area under the curve which show how well this algorithm distinguish between target values there is any financial recession or not. As area under the curve is greater it shows that algorithm works better as is can give more accurate results.

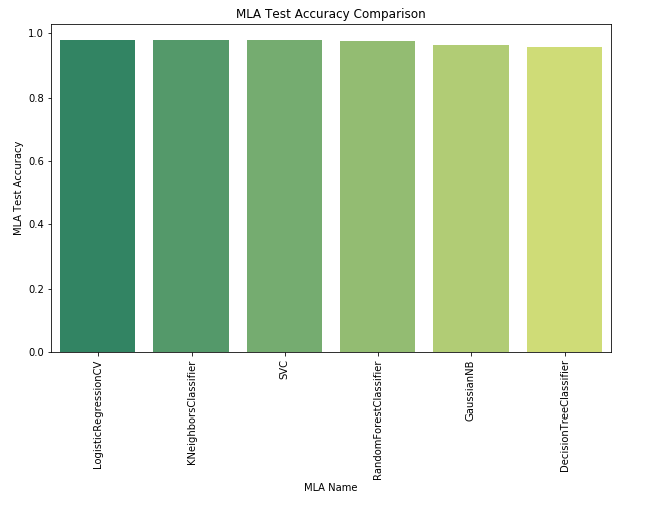


Figure 1.4

### Parameter Tuning:

In Parameter tuning I used Scoring type as roc auc score and Cross validation value of 12.

### Model Evaluation after Tuning:

Table 1.2 This table shows that score for Logistic regression reduces greatly after parameter tuning. It shows that the effect of overfitting was giving the higher percentage for results, but parameter tuning decreases the effect of overfitting. While Decision tree gives 95% after parameter tuning.

|  |  |  |
| --- | --- | --- |
| Algorithm | Training Score | Testing Score |
| Logistic Regression |  |  |
| Before Tuning | 96% | 98% |
| Best Parameters | Penalty : l2, Random\_state : 80, Solver : lbfgs |  |
| After Tuning | 56% | 50% |
| Decision Tree |  |  |
| Before Tuning | 100% | 96% |
| Best Parameters | Criterion : Gini, max\_depth : 3, splitter : best |  |
| After Tuning | 95% | 95% |

### Results:

This shows that if we apply Decision Tree to find the answer of this question that there will be Financial Recession in this year or not than it will give the right answer 95% of the times on the given data which is a good percentage for prediction of an event to occur. This prediction could help different countries to avoid being in the conditions that raise Financial recession.

## Problem 2.(AK)

On which features performance of Government depend before WWII, after WWII and today? so we can analyses the factors that affect Govt Performance.

### Explanation:

I deduced a data frame of USA with both electoral and financial features. I handled the missing values using median because the target feature is a discrete value with 0 and 1 so if we take the mean it will give result in float that will be a problem. For this problem I split the data into 3 different eras before WW2, after WW2 and resent years. I check how different features have impact on the target features at these eras and how this information can be used in future to stabilize the recession.

### Target Feature:

Normal Recession (pk\_norm). It is a discrete value with 0 and 1 values so I will be using Binary Classification Models.

### Predictors:

Party count, Protest, Govt.vote, Opp.vote, Turnover, realGDP, Frac

### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 70% and Testing set 30%.

### Feature Selection:

As in this question my main task is to select the most effected features from Normal recession. So, for applying Feature Selection Decision Tree Classifier is considered to be the best choice. Problem is about financial recession prediction in Europe, so I use the best features selected by Sequential Forward Selection

#### Before World War 2 (1990-1935)

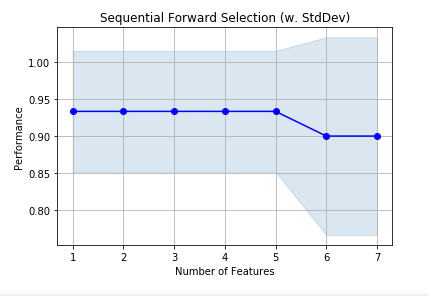


Figure 2.1

Table 2.1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature Index | Feature Names | Average Score |
| Step 1 | 0 | Party count | 84% |
| Step 2 | 0 , 1 | Party count, Protest | 84% |
| Step 3 | 0 , 1 , 4 | Party count, Protest, Govt.vote | 84% |
| Step 4 | 0 , 1 , 4, 5 | Party count, Protest, Govt.vote, Opp.vote | 84% |
| Step 5 | 0 , 1 , 4 , 5 , 6 | Party count, Protest, Govt.vote, Opp.vote, Turnover | 84% |
| Step 6 | 0, 1, 4, 5, 6, 2 | Party count, Protest, Govt.vote, Opp.vote, Turnover, realGDP | 84% |
| Step 7 | 0, 1, 4, 5, 6, 2, 3 | Party count, Protest, Govt.vote, Opp.vote, Turnover, realGDP, Frac | 79% |

Before World War 2 the most important features that effect the normal recession to occur are Party count, Protests, Government Vote, Opposition Vote, Turnover, Real GDP with percentage of 84%.

#### After World War 2

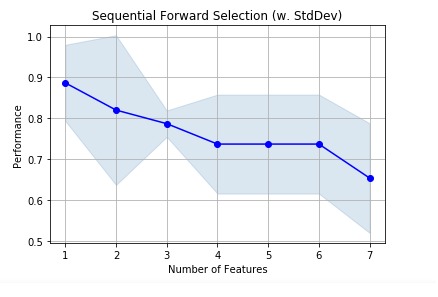


Figure 2.2

Table 2.2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature Index | Feature Names | Average Score |
| Step 1 | 0 | Party count | 84% |
| Step 2 | 0 , 1 | Party count, Turnover | 84% |
| Step 3 | 0 , 1 , 4 | Party count, Opp.vote, Turnover, | 79% |
| Step 4 | 0 , 1 , 4, 5 | Party count, Protest, Govt.vote, Opp.vote | 74% |
| Step 5 | 0 , 1 , 4 , 5 , 6 | Party count, Frac, Protest, Govt.vote, Opp.vote, Turnover | 74% |
| Step 6 | 0, 1, 4, 5, 6, 2 | Party count, Frac, RealGDP, Govt.vote, Opp.vote, Turnover | 63% |
| Step 7 | 0, 1, 4, 5, 6, 2, 3 | Party count, Protest, Govt.vote, Opp.vote, Turnover, realGDP, Frac | 67% |

After World War 2 most of the features that were having affect on normal recession before war are not affecting it any more. Features that have effect are Party count, Turnover with maximum percentage of 84%.

#### Now (1984-2014)

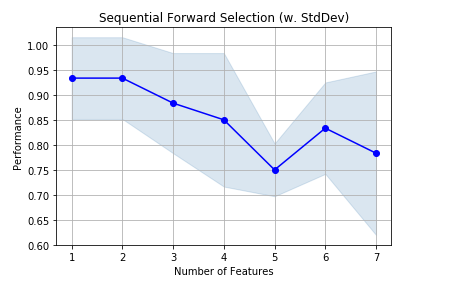


Figure 2.3

Table 2.3

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature Index | Feature Names | Average Score |
| Step 1 | 0 | Party count | 93% |
| Step 2 | 0 , 6 | Party count, Turnover | 93% |
| Step 3 | 0 , 2 , 6 | Party count, Frac, Turnover, | 88% |
| Step 4 | 0 , 1 , 2, 6 | Party count, Frac, Turnover,  Protest | 85% |
| Step 5 | 0 , 1 , 2 , 3 , 6 | Party count, Frac, Protest,  realGDP, Turnover | 75% |
| Step 6 | 0, 1, 2, 3, 4, 6 | Party count, Frac, RealGDP,  Govt.vote, Turnover | 83% |
| Step 7 | 0, 1, 2, 3, 4, 5, 6 | Party count, Protest,  Govt.vote, Opp.vote, Turnover, realGDP, Frac | 78% |

From year 1980-2014 dependency of features like Party count, Fractionalization, Turnover, Protests are affecting the occurrence of normal recession.

### Results:

In first 35 years there was more possibility to predict the occurrence of event like recession using the given features but after World War 2 there was chaos and it was not possible to predict using these features and adding there features to predict the result is decreasing the accuracy of result. In last 35 years of data it is more dependent on these features and results are more predictable. So, using these results we can select best features for an event to occur and prevent future recessions by keeping the situation more in control.

## Problem 3 (AK)

Is it possible to predict opposition vote by applying machine learning prediction algorithms?

### Explanation:

I deduced a data frame of 6 different countries with both electoral and financial features. Those countries are Germany, Great Britain, Austria, Belgium, Italy and Sweden. Fill the missing values using mean function, as my target feature is a continuous value.

### Target Feature:

Number of opposition vote (oppvote). It is a continuous value so I will be using Regression Models.

### Predictors:

Party count, Right wing, Left Wing, Extremists, Protests, Riotsdev, Dictatorship, Turnover, CrisisJST.

### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 70% and Testing set 30%.

### Model Evaluation:

Table 3.1 Mean absolute error values of respective algorithms are:

|  |  |
| --- | --- |
| Algorithm | Mean absolute error |
| Linear Regression | 8.20 |
| Decision Tree Regressor | 4.68 |
| Gaussian Process Regressor | 44.11 |
| Random Forest Regressor | 4.74 |

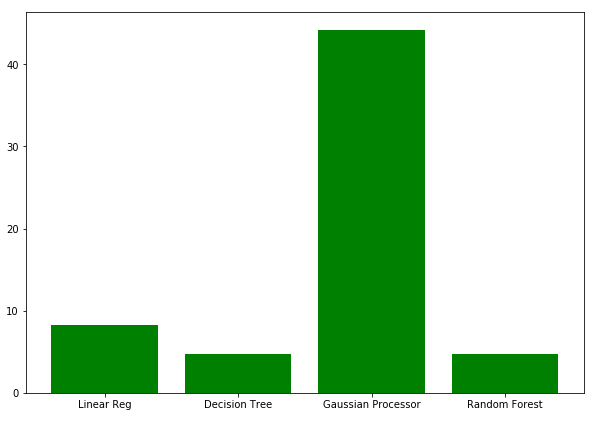


Figure 3.1

This table is a bar plot of Mean absolute error of 4 different regressor algorithms to show the difference in their performance to predict the real value.

### Feature Selection:

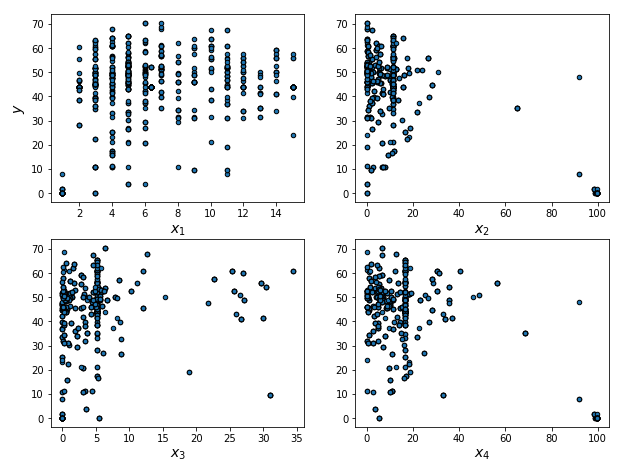
For feature Selection I used Mutual Information Regression and F-Test Regression values to see the relation of features with the target value.

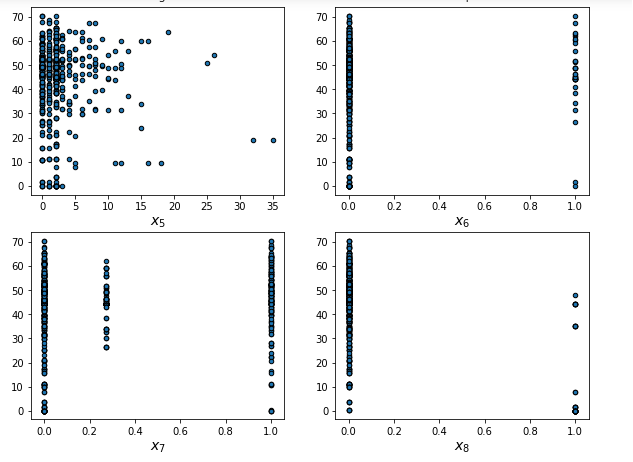
#### F-Test Regression:

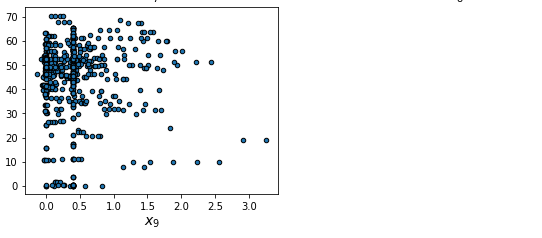
This value gives only linear dependency of the feature with the target value. 1 as maximum and 0 as minimum dependency.

#### Mutual Information Regression:

This value gives any kind of regression dependency between feature and target values. 1 as maximum and 0 as minimum dependency





Figure 3.2

These figures show the scatter plot of every feature we have selected on x axis and target value on y-axis to check its correlation with the target value. X1(partycount) X2(right) X3(left) X4(extremist) X5(protests) X6(crisisJST) X7(turnover) X8(dict) X9(riotsdev)

Table 3.2 for comparison of F-test and mutual information values of every feature

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Name | Feature Index | Mutual info | F-Test |
| Party Count | 0 | 0.45 | 0.100 |
| Right | 1 | 0.88 | 1.0 |
| Left | 2 | 0.83 | 0.05 |
| Extremists | 3 | 1.0 | 0.72 |
| Protests | 4 | 0.14 | 0.00 |
| Crisis JST | 5 | 0.0 | 0.00 |
| Turnover | 6 | 0.07 | 0.03 |
| Dictatorship | 7 | 0.05 | 0.80 |
| Riots Deviation | 8 | 0.22 | 0.00 |

#### Predictors:

Right

Extremists

Dictatorship

### Model Evaluation after Feature Selection:

Table 3.3 Mean absolute error values of respective algorithms are:

|  |  |
| --- | --- |
| Algorithm | Mean absolute error |
| Linear Regression | 8.11 |
| Decision Tree Regressor | 2.95 |
| Gaussian Process Regressor | 10.79 |
| Random Forest Regressor | 4.15 |

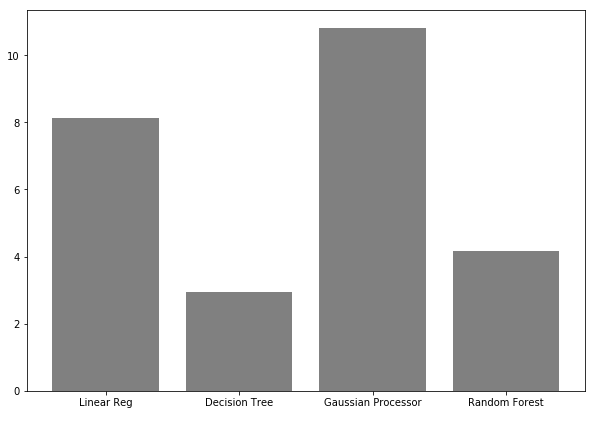


Figure 3.2

This table is a bar plot of Mean absolute error of 4 different regressor algorithms after feature selection to show the difference in their performance to predict the real value.

### Results:

To predict the number of oppositions vote we check different regression algorithm. Before feature Selection all these algorithms were giving comparatively large values of mean absolute error in the prediction. But after feature selection mean absolute error reduced. Decision Tree Algorithm gives the lowest value of error, so it is the best algorithm to find the result. Number of Opposition votes predicted from this conclusion will not be very accurate, but authorities can estimate the future situations of any country.

## Problem 4. (AK)

Does GDP and protests have any effect on the economic as well as non-economic disasters? If yes than can we predict is there going to be any kind of disaster in the near future of the world.

### Explanation

Data set with both electoral and financial features of all 20 countries. Fill the missing values using median because the target feature is a discrete value with 0 and 1 so if we take the mean it will give result in float that will be a problem.

### Target Feature

Financial Recession, Non-Financial Recession and Macro Economic Disaster (pk\_fin, pk\_norm, pk\_dis)

#### Dimensionality Reduction:

I take OR of these 3 target features and reduce these three columns into one target value to check if there is any of these recession/disasters in this year or not.It is a discrete value so I will be using Classification Models.

### Predictors

Real GDP, Inflation Rate (cpi), Government Crisis, Protests and GDP peak

### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 80% and Testing set 20%.

### Model Evaluation:

K Nearest Neighbour

Decision Tree Classifier

Gaussian Naïve Bayes

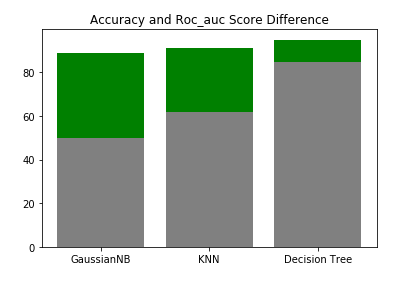


Figure 4.1

In this bar plot grey part of the bar shows precision/recall score of the algorithms and green part show its difference with the accuracy score. It shows that all algorithms have big difference in these values it shows these are not accurate enough to predict from this result. We always have to distinguish between these two values that which one is better for the type of result we are looking for. Accuracy score gives the overall percentage of accuracy of right result while roc auc score gives the cut points where the specific algorithm gives the best result.

#### Confusion Matrices:

Confusion Matrices have 4 values:

TP True Positive

FP False Positive

FN False Negative

TN True Negative

Gaussian NB KNN

|  |  |
| --- | --- |
| TP 518 | FP 2 |
| FN 60 | TN 0 |

|  |  |
| --- | --- |
| TP 517 | FP 3 |
| FN 45 | TN 15 |

Decision Tree Classifier

|  |  |
| --- | --- |
| TP 511 | FP 9 |
| FN 16 | TN 44 |

### Parameter Tuning:

Parameter tuning using Scoring as accuracy and Cross validation value with 10.

Table 4.1

|  |  |  |
| --- | --- | --- |
| Algorithm | Accuracy Score | Roc auc Score |
| K Nearest Neighbour |  |  |
| Before Tuning | 91% | 62% |
| Best Parameters | n\_neighbors = 3,Random\_state = 80, |  |
| After Tuning | 93% | 69% |
| Decision Tree Classifier |  |  |
| Before Tuning | 95% | 85% |
| Best Parameters | Criterion : Entropy, max\_depth : 3,  splitter : best |  |
| After Tuning | 97% | 94% |

### Model Evaluation after Tuning:

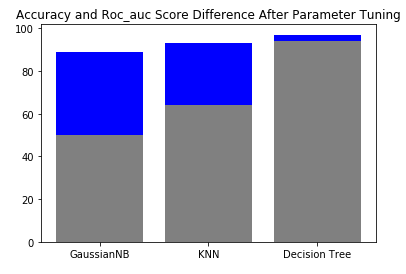


Figure 4.2

In this bar plot grey part of the bar shows roc\_auc score of the algorithms and blue parts show its difference with the accuracy score. It shows that KNN and Decision Tree Classifier algorithm have comparatively less difference in the values.

#### Confusion Matrices after Parameter Tuning:

Gaussian NB KNN

|  |  |
| --- | --- |
| TP 518 | FP 2 |
| FN 60 | TN 0 |

|  |  |
| --- | --- |
| TP 511 | FP 4 |
| FN 37 | TN 23 |

Decision Tree Classifier

|  |  |
| --- | --- |
| TP 511 | FP 4 |
| FN 37 | TN 23 |

### Results:

After all the analysis it shows that GDP and Protests have great effect on the number of financial and non-financial disasters. The best algorithm that can be used to predict that is there any disaster in coming years or not is Decision tree with the given parameters in Table 4.1

## Problem 5 (AK)

Prediction of Fractionalization without knowing opposition and government vote.

### Explanation:

I deduced a data frame of 6 different countries with both electoral and financial features. Fill the missing values using interpolation method that takes the nearest available value to the other sides nearest available value and if needed takes mean. I also Scaled the data to make the results more accurate using Standard Scaler.

### Target Feature:

Fractionalization of Legislature Chamber (frac). It is a continuous value so I will be using Regression Models.

### Predictors:

Right, Left, Extremists, Government Crisis, Protests, Party count, demonstration deviation, Riots dev, Strikes dev, Turnover, Veto Players, Dictatorship.

### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 70% and Testing set 30%.

### Feature Selection:

For the solution of this query I selected the most correlated features to the target value using Correlation method of ‘Pearson’ and Heatmap.

### Model Evaluation:

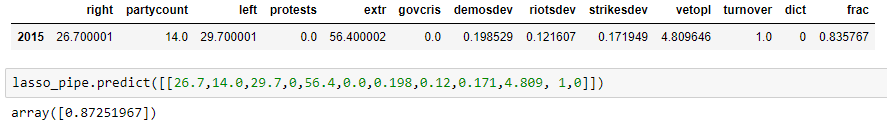
RMSE: Root Mean Square Error

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms | Training Score | Testing Score | RMSE |
| Linear | 91% | 77% | 0.11 |
| Lasso | 92% | 93% | 0.06 |
| Ridge | 88% | 72% | 0.12 |

Table 5.1

### Results:

As the Table 5.1 shows that the Lasso gives the best percentage of testing score. As LASSO not only perform the regression the given features but also reduce the number of features by eliminating the unrelated ones and reduce overfitting. So, It is best idea to use this algorithm for prediction

This result shows that the real value of fractionalization is 0.84 while the predicted value using Lasso algorithm is 0.87 which is a pretty good value in predicting some unknown important figures.

## Problem 6. (AK)

Can we predict Systematic Financial Crisis in the future of the world? Only using the Government related features and government performance.

### Explanation:

I used the given dataset with 20 different countries and only their electoral and government related features so we can analyze their importance in the Systematic crisis to take place. This analysis will help governments of these countries to stabilize the situation of their countries. I fill the missing values with median function because my target feature is a binary value.

### Target Feature:

Systematic Financial Crisis (crisisJST). We will be applying some Classification Models on this problem as our target feature is a discrete value

### Predictors:

Government Vote, Turnover, Opposition Vote, Party count, Right, Dictatorship, Left, No. of Veto Players, Fractionalization, Government Crisis and Extremists

### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 70% and Testing set 30%.

### Feature Selection:

Question is about Systematic financial crisis prediction only using the Government related features. So I selected 11 from 28 features of the data which represent the government related oinformation so I can have the result that is it effection the financial situation of the countries.

### Model Evaluation:

Algorithms I used are all for classification because my Target values is discrete value, Systematic Financial crisis has the value 0 ( if there was no crisis) and 1 (if there was a crisis)

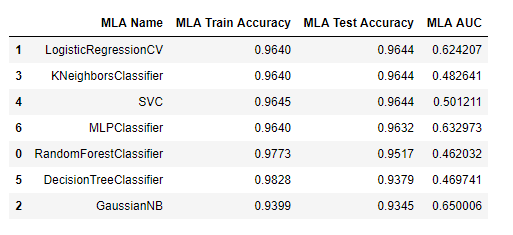


Figure6.2

This table is arranged according to best Test Accuracy values because good Test accuracies gives the best prediction of the answers.

#### Roc Auc curve

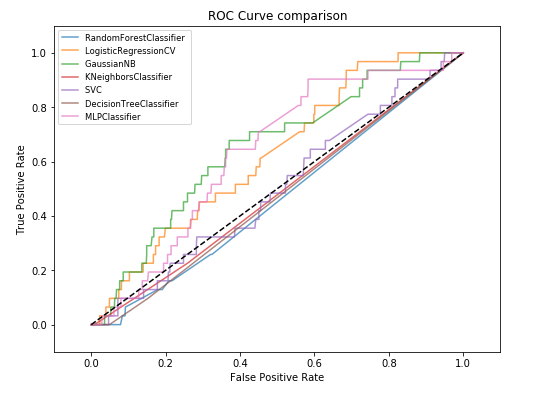


Figure 6.3

Roc\_Auc curve is the graphical representation of confusion matrix with False Positive Rate at x-axis and True Positive Rate at y-axis and from (0,0) to (1,1). AUC stands for area under the curve which show how well this algorithm distinguish between target values there is any systematic financial crisis or not.

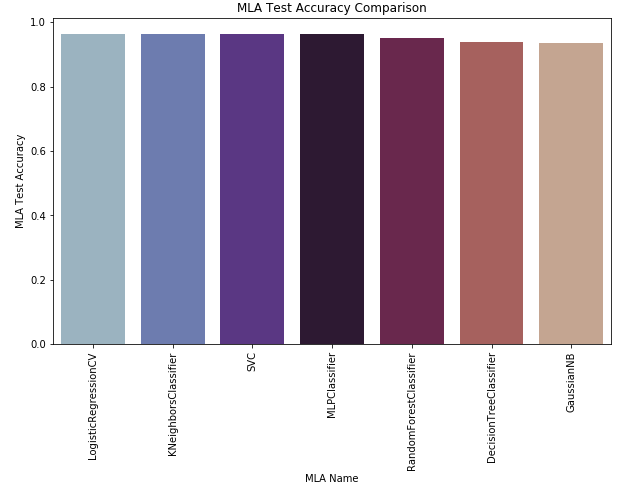


Figure 6.4

### Parameter Tuning:

Parameter tuning using with Scoring as accuracy and Cross validation value of 12 and 10.

### Model Evaluation after Parameter Tuning

Table 1.2

|  |  |  |
| --- | --- | --- |
| Algorithm | Training Score | Testing Score |
| Logistic Regression |  |  |
| Before Tuning | 96% | 96% |
| Best Parameters | Penalty : l2, Random\_state : 80,  Solver : lbfgs |  |
| After Tuning | 98% | 94% |
| Decision Tree |  |  |
| Before Tuning | 96% | 96% |
| Best Parameters | Criterion : Gini, max\_depth : 1,  splitter : random |  |
| After Tuning | 96% | 96% |
| Neural Network |  |  |
| Before Tuning | 96% | 96% |
| Best Parameters | Activation= identity, Hidden\_layer\_sizes= 5,Random\_state= 1,Solver= lbfgs |  |
| After Tuning | 96% | 96% |

### Results:

This shows that if we apply Neural Network MLP Classifier to find the answer of this question that there will be Financial Crisis in this year or not only using the limited features related to government related information.

Than it will give the right answer 96% of the times which is pretty good percentage but as we can see area under the curve is not giving very good values. AUC is giving value 50 which means these models are overfitting because we mostly have 0 value in the target feature. This shows we do not have enough data to predict this event in future but at least we can analyse the importance of government related features in this regard.

## Problem 7 (AK)

Does unemployment effect the stability of the countries and lead them to much worst conditions like normal recession?

### Explanation:

Data set with both electoral and financial features of USA. I merged the data set of unemployment of USA with the given data set on years from 1941 to 2010. Fill the missing values using median.

### Target Feature:

Non-Financial Recession (pk\_norm). It is a discrete value so I will be using Classification Models.

### Predictors

Unemployed Percentage, Unemployed Number, Employed Percentage, Not in Labor, Population, Labor Force, Employed Number

### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 70% and Testing set 30%.

### Feature Selection:

Table 7.1 Using Feature Importance Method. These are Percentage share of every predictor from total of 100%

|  |  |
| --- | --- |
| Features | Percentage from total of 100 |
| Unemployed Percentage | 20% |
| Unemployed No. | 18% |
| Employed Percentage | 13% |
| Not in Labor | 13% |
| Population | 12% |
| Labor Force | 12% |
| Employed No. | 10% |

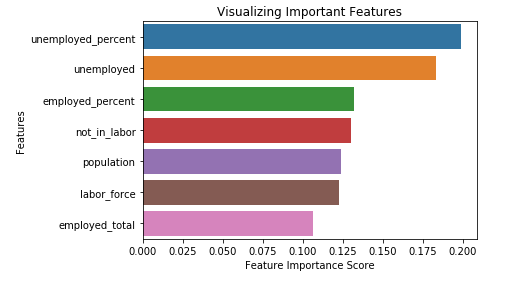


Figure 7.1

#### Predictors:

Unemployed Percentage

Unemployed Number

Employed Percentage

Not in Labor

### Model Evaluation:

Table 7.2

|  |  |  |
| --- | --- | --- |
| Algorithm | Training Score | Testing Score |
| Logistic Regression | 86% | 82% |
| Decision Tree Classifier | 100% | 77% |
| K Nearest Neighbor | 88% | 82% |
| Random Forest Classifier | 100% | 73% |
| Naïve Bayes | 67% | 36% |
| SVM | 100% | 82% |

This table shows that there is the case of overfitting in most of the algorithms as they are giving 100% Training accuracy and comparatively very low testing accuracy so to reduce the overfitting.

### Parameter Tuning:

Parameter tuning using Scoring as accuracy and Cross validation value of 10.

Table 7.3

|  |  |
| --- | --- |
| Algorithm | Best Parameters |
| Logistic Regression | Penalty= l2, Random\_state= 80, Solver= sag |
| Decision Tree Classifier | Criterion= gini, Max\_depth= 3, Splitter= random |
| Random Forest Classifier | Max\_depth= 8, n\_estimators= 2 |
| K Nearest Neighbour | n\_neighbors= 2, Metric= minkowski |

### Model Evaluation after Tuning:

Table 7.3

|  |  |  |
| --- | --- | --- |
| Algorithm | Training Score | Testing Score |
| Logistic Regression | 88% | 82% |
| Decision Tree Classifier | 90% | 82% |
| K Nearest Neighbor | 88% | 82% |
| Random Forest Classifier | 92% | 82% |
| Naïve Bayes | 67% | 36% |

You can clearly see the effect of overfitting is reduced by Parameter tuning. Algorithms that were giving 100% of accuracy due to overfitting is reduced to 94% and 92% of training accuracy.

### Results:

All these algorithms Logistic Regression, Decision Tree, Random Forest and K nearest Neighbour are giving 82% of accuracy for crisis of a country. This concludes that unemployment and large labour force with no jobs have great effect on the stability of the country. So to Unemployment plays a large roll in the rise of a recession or vice versa.

## Problem 8 (AK)

Is it possible to get the prediction about growth rate of GDP that will be positive or negative without knowing the real GDP? And is it possible to answer this question using this dataset? If not than how we will solve this problem of not enough data?

### Explanation:

I deduced a data frame of USA with both electoral and financial features. Fill the missing values using median.

#### Data Merging:

Merged the given data set with a new data set of USA from year 1929 to 2014. To check the validity of the datasets I draw a simple plot of Real GDP from given dataset and bar plot of GDP growth rate from the new dataset using Matplotlib Library.

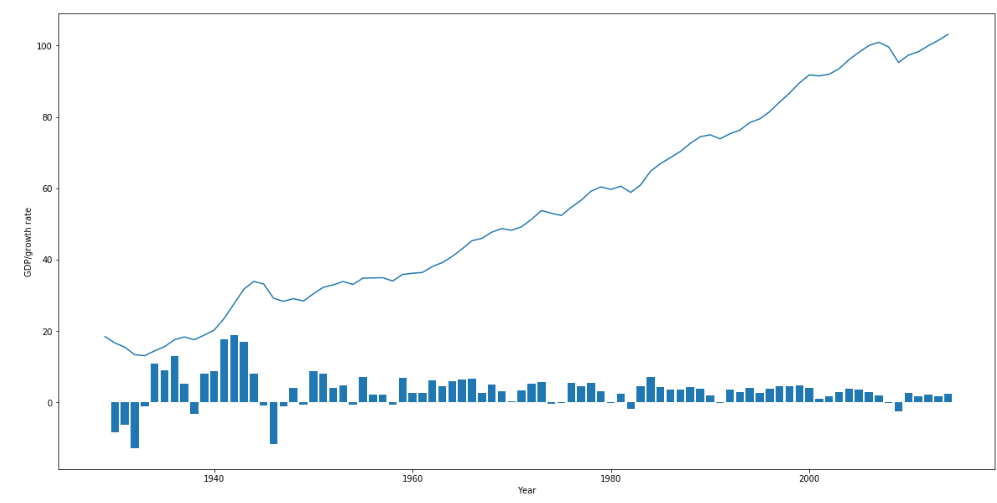


Figure 8.1

This figure clearly shows that both datasets are correctly representing each other. As the value of growth rate is in negative you can see the real GDP graph also drops.

#### Type Conversion:

In the new data set GDP growth rate are positive and negative both increasing and decreasing values. To make the data more useable according to my question, I converted the column GDP growth rate (gdpgrowthrate) into a new discrete value column with 0 (if value is below 0) and 1(if value is equal or above 0).

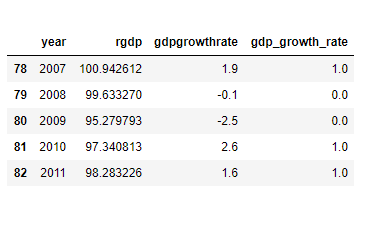


Table 8.1

### Target Feature:

GDP Growth Rate (gdp\_growth\_rate). It is a discrete feature with 0 and 1 values so to predict this as a target value I will be using Classification Models.

### Predictors:

Strikes dev, Riots dev, Protests dev, Demonstration dev, Protests, Turnover, pk\_norm, pk\_fin, pk\_dis, govcrisis, and crisisJST

### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 70% and Testing set 30%.

### Feature Selection:

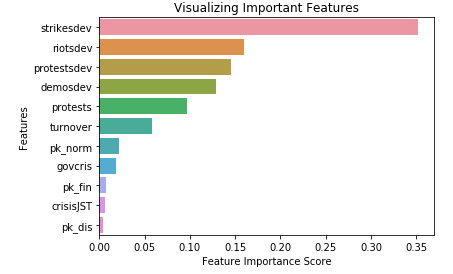


Figure 8.2

Predictors after Feature Selection are:

Table 7.1

|  |  |
| --- | --- |
| Features | Percentage from total of 100 |
| Strikes Deviation | 35% |
| Riots Deviation | 16% |
| Protests Deviation | 15% |
| Demonstration Deviation | 13% |

### Model Evaluation:

Table 7.2 RMSE: Root Mean Square Error

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Training Score | Testing Score | RMSE |
| Logistic Regression | 82% | 81% | 0.44 |
| Decision Tree Classifier | 98% | 69% | 0.55 |
| K Nearest Neighbor | 87% | 73% | 0.52 |
| Random Forest Classifier | 98% | 73% | 0.52 |
| Naïve Bayes | 82% | 65% | 0.59 |
| SVM | 78% | 81% | 0.44 |
| Neural Network | 87% | 77% | 0.48 |

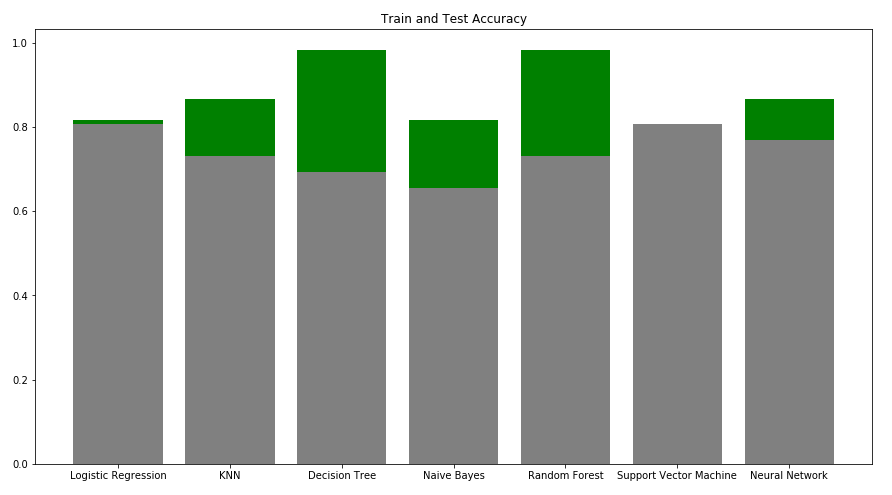


Figure 8.3

This figure shows the Testing accuracies of all the models in grey and their difference with training accuracies to show how good that models works on this data for prediction of answer.

### Results:

For the classification that there will be positive growth rate of GDP or not, 2 algorithms Logistic Regression and Support Vector Machine Algorithm are giving comparatively better results than others and they have less difference in their training and testing scores. So without knowing the real GDP it is an interesting result to predict the increasing or decreasing GDP growth rate.

## Problem 9 (AK)

On which features performance of Government depend on tree different eras of times in Europe? so we can analyse the factors that affect Govt Performance.

### Explanation:

I deduced a data frame of European 17 countries with both electoral and financial features. Those 17 countries are all from the region of Europe.

Italy, France, Germany, Belgium, Switzerland, Belgium, Denmark, Netherlands, Britain, Sweden, Norway, Finland, Spain, Ireland, Greece, Portugal and Spain

Fill the missing values using median.

### Target Feature:

Normal Recession (pk\_norm). It is a discrete value with 0 and 1 values so I will be using Classification Models.

### Predictors:

Right, Left, Extremists, Protests, RGDP, opp vote, Financial recession

### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 70% and Testing set 30%.

### Feature Selection:

As in this question main task is to select the most effected features for Normal recession. So, for applying Feature Selection Decision Tree Classifier is considered to be the best choice. Question is about financial recession prediction in Europe, so I use the best features selected by Sequential Forward Selection

#### Before World War 2 (1900-1934)

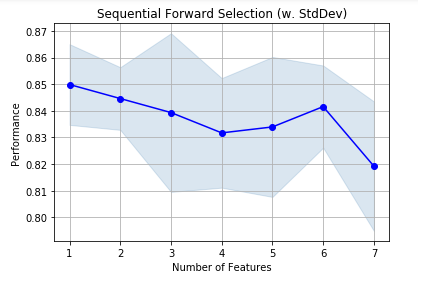


Figure 9.1

Table 9.1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature Index | Feature Names | Average Score |
| Step 1 | 6 | Financial recession | 85% |
| Step 2 | 0 , 6 | Right, Financial recession | 84% |
| Step 3 | 0 , 2, 6 | Right, Extremists, Financial recession | 84% |
| Step 4 | 0 , 2 , 3, 6 | Right, Extremists, Protests, Financial recession | 84% |
| Step 5 | 0 , 1 , 2 , 3 , 6 | Right, Left, Extremists, Protests, Financial recession | 83% |
| Step 6 | 0, 1, 2, 3, 4, 6 | Right, Left, Extremists, Protests, RGDP, Financial recession | 84% |
| Step 7 | 0, 1, 2, 3, 4, 5, 6 | Right, Left, Extremists, Protests, RGDP, opp vote, Financial recession | 82% |

Before World War 2 the most important features that effect the normal recession to occur are Financial Recession, Right wing, Extremists, Left wing, Protests, Real GDP with percentage of 84%

#### After World War 2 (1945-1984)

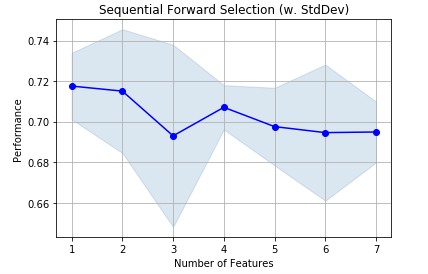


Figure 9.2

Table 9.2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature Index | Feature Names | Average Score |
| Step 1 | 6 | Financial recession | 72% |
| Step 2 | 3 , 6 | Extrimists, Financial recession | 72% |
| Step 3 | 2, 3, 6 | Extremists, Protests, Financial recession | 70% |
| Step 4 | 1 , 2 , 3, 6 | Left, Extremists, Protests, Financial recession | 71% |
| Step 5 | 1 , 2 , 3 , 5, 6 | Left, Extremists, Protests, opp vote, Financial recession | 70% |
| Step 6 | 1, 2, 3, 4,5, 6 | Left, Extremists, Protests, RGDP, opp vote, Financial recession | 69% |
| Step 7 | 0, 1, 2, 3, 4, 5, 6 | Right, Left, Extremists, Protests, RGDP, opp vote, Financial recession | 69% |

### After World War 2 most of the features that were having effect on normal recession before war are not affecting it any more. Features that have effect are Extremists and Financial Recession with maximum percentage of 72% which is comparatively very low as compared to before and after WW2. During war situation was not in control.

#### Now (1985 - 2014)

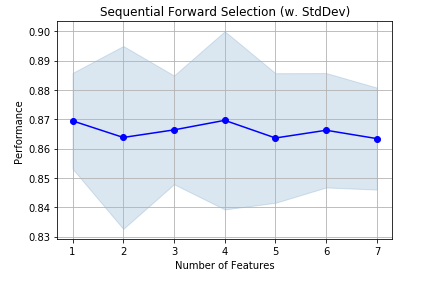


Figure 9.3

Table 9.3

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature Index | Feature Names | Average Score |
| Step 1 | 6 | Pk\_fin | 87% |
| Step 2 | 2, 6 | Extrimist, pk\_fin | 86% |
| Step 3 | 2 ,4, 6 | Extrimists, RGDP, pk\_fin | 87% |
| Step 4 | 2, 4, 5, 6 | Extrimists, RGDP, opp vote, pk\_fin | 87% |
| Step 5 | 1, 2, 4, 5, 6 | Left, Extrimists, RGDP, opp vote, pk\_fin | 86% |
| Step 6 | 1,2,3,4,5,6 | Left, Extrimists, Right, RGDP, opp vote,protests, pk\_fin | 87% |
| Step 7 | 0, 1, 2, 3, 4, 5, 6 | Right, Left, Extrimists, RGDP, opp vote, pk\_fin | 86% |

From year 1980-2014 dependency of features are:

Left wing, Right wing, Extremists, Real GDP, Opposition vote, Financial Recession, Protests

Are affecting the occurrence of normal recession with maximum of 87% which show that now conditions are in control of government so they can improve their performance by improving these features.

### Results:

In first 35 years there was more possibility to predict the occurrence of event like recession using the given features but after World War 2 there was chaos and it was not possible to predict using these features and adding more features to predict the result is decreasing the accuracy of result.

In last 35 years of data it is more dependent on these features and results are more predictable. So, using these results we can select best features for an event to occur and prevent future recessions by keeping the situation more in control. These results also conclude that how War effects the condition of not only the country that is in war but also derail the whole region both economically and systematically.

## Problem 10 (AK)

Can we predict the GDP growth rate of USA as it goes negative when decreases and positive when increases so, to predict it is going to increase or decrease in future?

### Explanation:

I deduced a data frame of USA with both electoral and financial features. Fill the missing values using mean as the target feature is a continuous value with data type of float so mean is the best method here as it will give a value in float.

#### Data Merging:

Merged the given data set with a new data set of USA from year 1929 to 2014.

### Target Feature:

GDP Growth Rate (gdpgrowthrate). It is a continuous value so I will be using Regression Models.

### Predictors:

Government Vote, Opposition Vote, Fractionalization, Party count, Protests, Protests dev, Demonstration dev, Strikes dev, real GDP', Inflation Rate, No. of government crisis.

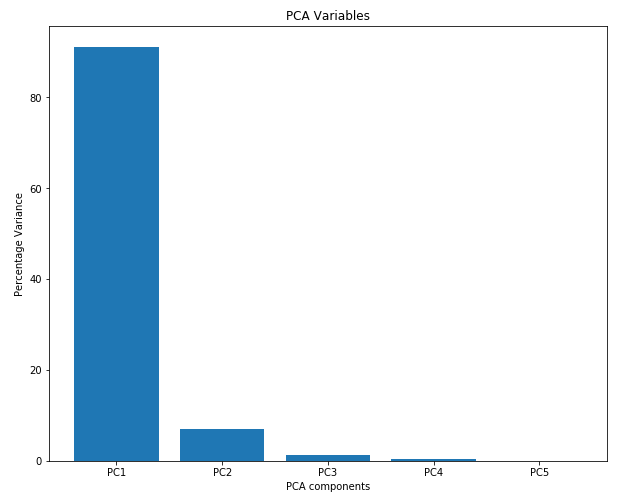
### Train Test Split:

I use Train Test Split method to divide the data into training and testing parts with Training set 75% and Testing set 25%.

### Feature Selection:

#### Dimensionality Reduction.

For the selection of the most related features I applied Principle Component Analysis PCA. Reduce the features to 5 and select the most related one from 5 components.



### Model Evaluation:

RMSE: Root Mean Square Error

|  |  |
| --- | --- |
| Algorithm | RMSE |
| Linear Regression | 3.74 |
| Lasso Regression | 3.74 |
| Decision Tree Regressor | 3.84 |
| Random Forest Regressor | 3.64 |
| MLP Regressor | 3.83 |

Table 10.1

### Results:

In this Question I consider the root mean square error to compare the performance of Algorithms. As the Table 10.1 shows that the Random Forest Regressor gives the lowest Root mean square error score. So, it is best idea to use this algorithm for prediction of GDP growth rate. Prediction results shows that it gives the right answers about the negative and positive values of the continuous value of GDP growth rate. So this method can be used for the prediction of GDP growth rate although it would not give the accurate result but we analyze the trend.

### Explanation

### Target Feature

### Predictors

### Train/Test/Split

### Feature Selection

### Model Evaluation

### Parameter Tuning

### Model Evaluation after Parameter Tuning

### Results

# Conclusion

# Literature

[1] T. Davenport and J. Short, “The New Industrial Engineering: Information Technology and Business Process Redesign,” *Sloan Management Review*, vol. 31, no. 4, pp. 11–27, 1990.

[2] <https://www.bea.gov/data/gdp/gross-domestic-product>

[3] <https://data.oecd.org/unemp/unemployment-rate.htm>