CLASSIFICATION MODEL COMPARISON

THIS PROJECT COMPARES LOGISTIC REGRESSION, K NEAREST NEIGHBOR, NAVIE BAYES AND SUPPORT VECTOR MACHINE MODELS

Babatunde Racheal

05/12/23

DATA SET OVERVIEW

The data set contains a sample of 300 of a stock named WAB. WAB return contains 300 samples of return values taken from 2006 to now. For dates recorded for this period, our analysis is investigating the percentage returns for two trading days which is represented as Lag1 return and Lag2 return and also the WAB risk for these two trading days is included.

The response variable these classifications will be predicting will be the WAB return and WAB risk while the independent variables are the two trading days which is Lag1 return, Lag2 return and Lag1 risk, Lag2 risk. The date entry of these two data sets will be drop due it being irrelevant to this study.

LOGISTIC REGRESSION

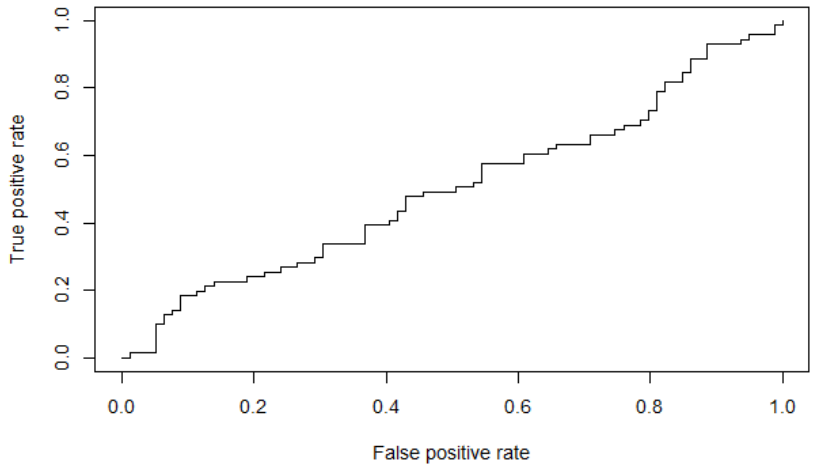
Logistic Regression is a statistical approach that is used for classification problems and is based on the concept of probability. It is used when the dependent variable (target) is categorical.

It is widely used when the classification problem at hand is binary; true or false, yes or no and in the case of this study high return, low return and high risk, low risk

WAB RETUN RATE

ACCESSING THE LOGISTIC MODEL VISUALLY

ROC PLOT



The ROC curve shows the tradeoff between sensitivity and 1- specificity. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. The AUC which is

equivalent to the probability that a randomly chosen positive instance that is high return is ranked higher than a randomly chosen negative instance.

The Area Under the ROC curve (AUC) is an aggregated metric that evaluates how well a logistic regression model classifies positive and negative outcomes at all possible cutoffs. It can range from 0.5 to 1, and the larger it is the better.

AUC SUMMARY FOR WAB RETURN

Area under the curve: 0.4994

Two different extractor functions have been used to display our result. The first gives what amounts to regression coefficients with the standard errors and z test. Accessing the deviance of the two-predictor variable, one of the coefficients is significantly different from zero. The total deviance of both predictor variable is 3.185 points on 2 degree of freedom, for a p value is 0.203.

Overall this logistic model seems to have performed poorly, showing no significant reduction in the deviance (no significant difference from the null model)

WAB RISK RATE

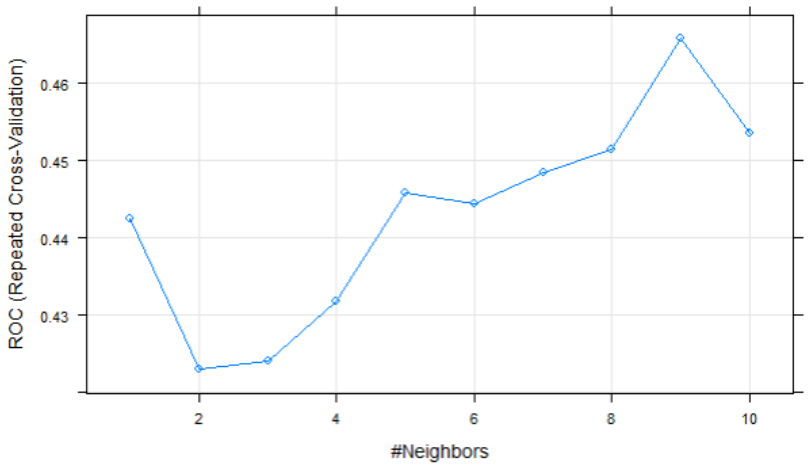
The deviations recorded from both predictor variables of WAB risk rate shows a significant reduction in the deviation and both predictor variables also showed very small p values (2.2e-16, 15.871e-07).

Area under the curve: 0.507

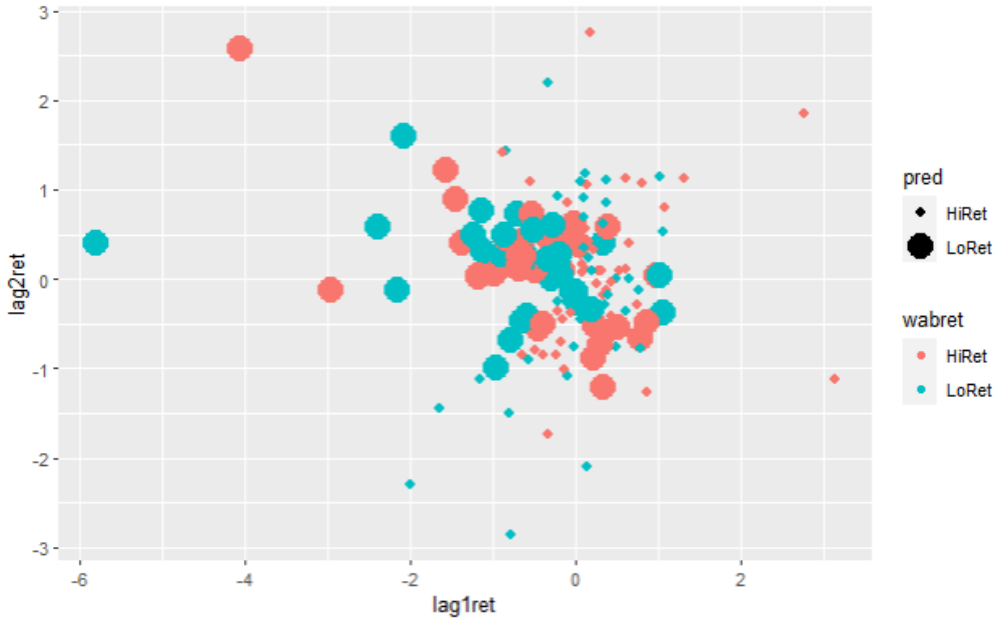
This implies that the logistic regression model performs slightly better for the WAB risk data set than the WAB return data set

K -NEAREST NEIGHBOR CLASSIFICATION MODEL

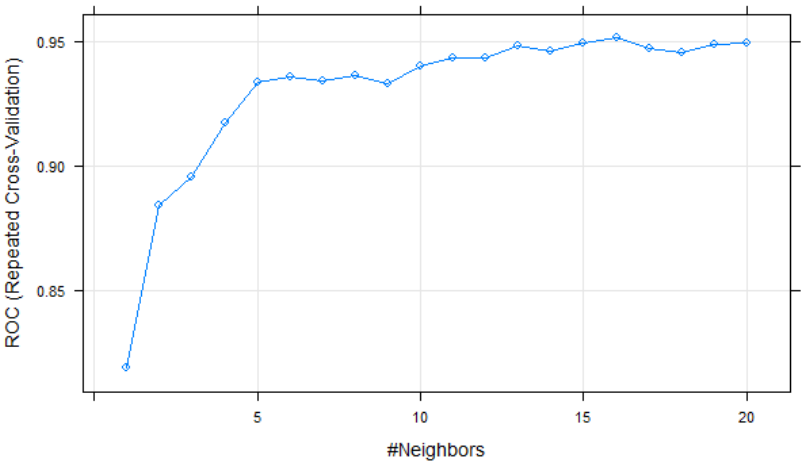
K-NN algorithm assumes the similarity between the new data and available cases and put the new case into the category that is most similar to the available categories.



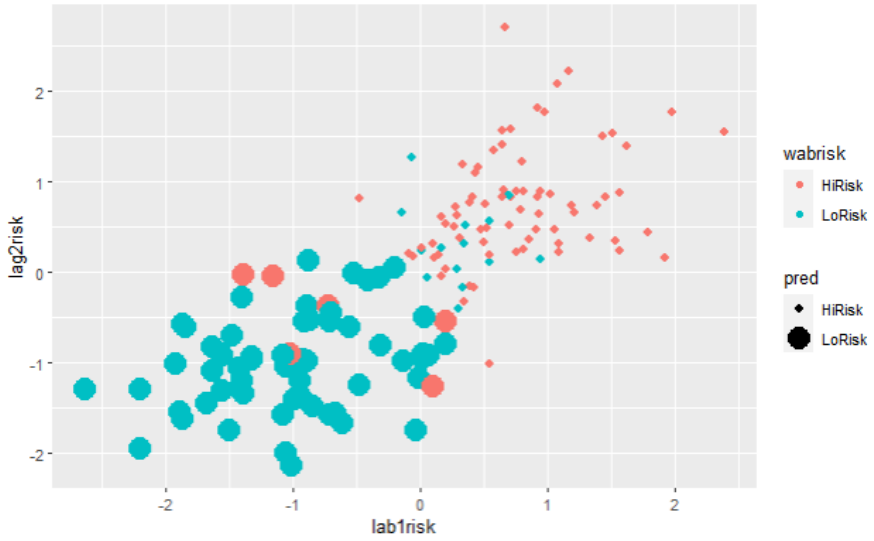
The K = 9 value chosen for KNN tells us that the model chose 9 closet points to say High return, and hence Low return will be predicted High using the majority vote of (5:4) The model accuracy is 47.33% with a specificity of 45.71%. ROC when k =9 is 0.4658



K-NN for WAB RISK

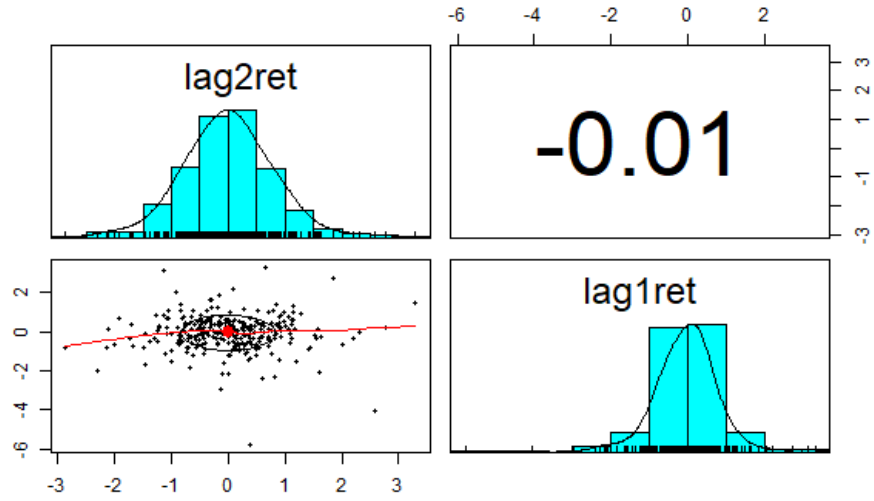


ROC was used to select the optimal model using the largest value. The final value used for the model was k = 16. The model accuracy is 95% with a specificity of 84%. ROC when k = 16 is 0.9522

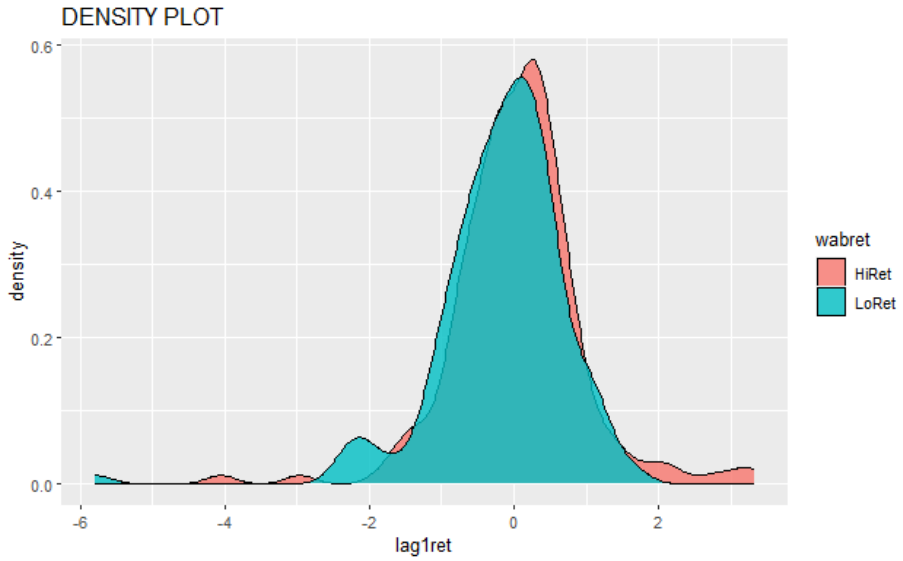


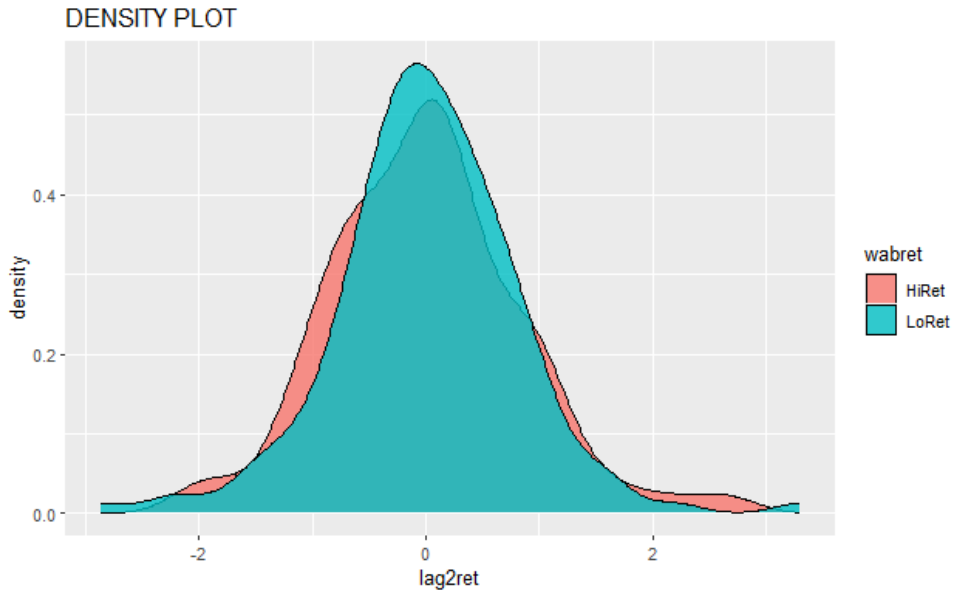
NAVIE BAYES CLASSIFIER FOR WAB RETURN

In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This is not a realistic view because most independent variables have some form of correlation hence the name “naïve”.



The pair panels display the correlation between the two independent variables lag1ret and lag2ret and it can be confirmed that correlation between these two variables is weak which is in contrary to our naïve bayes assumptions.

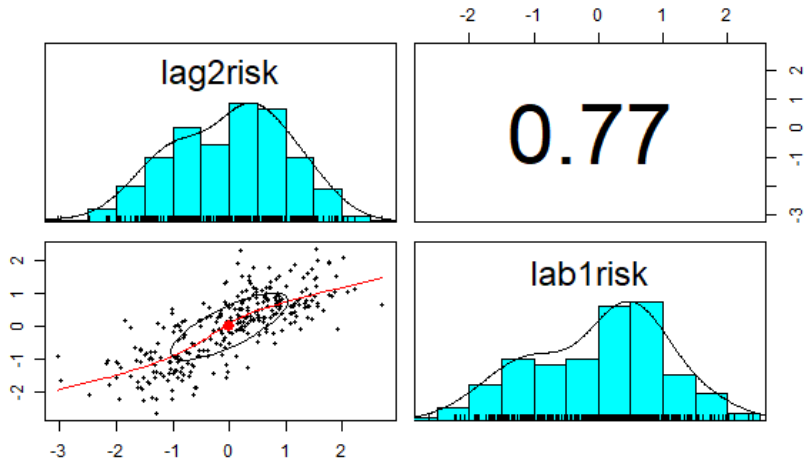




Looking at the density plots for lag1ret and lag2ret there is a significant amount of overlap and for lag2ret low return is higher than higher return and the inverse is for lag1ret.These plots shows that there is more potential to develop a classification model but the model is likely to be less accurate as a result of this overlap

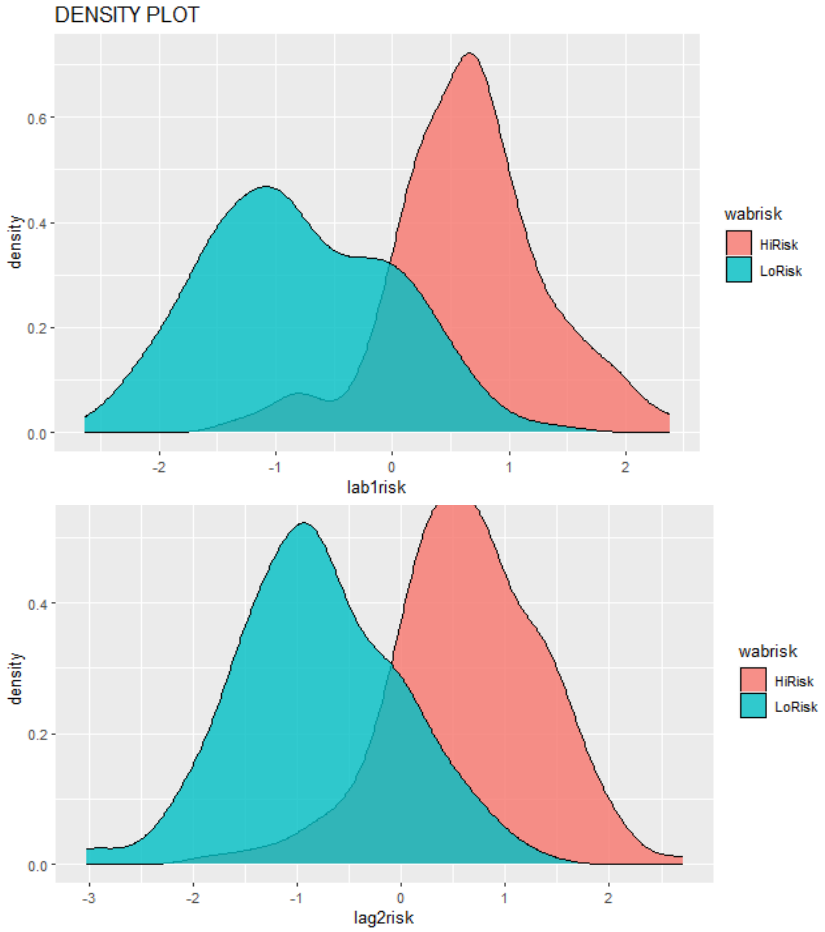
Computing the accuracy for naive bayes classification using cross validation. Accuracy was used to select the optimal model using the largest value. The final values used for the model were fL = 0, use kernel = TRUE and adjust = 1

NAÏVE BAYES FOR WAB RISK RATE



The correlation between predictors lab1 risk and lab 2 risk is strong which implies that the **naïve** bayes classifier will predict values of WAB risk rate more accurately.

The density plot displays clearly that there is a minimal overlap between the two predictor variables



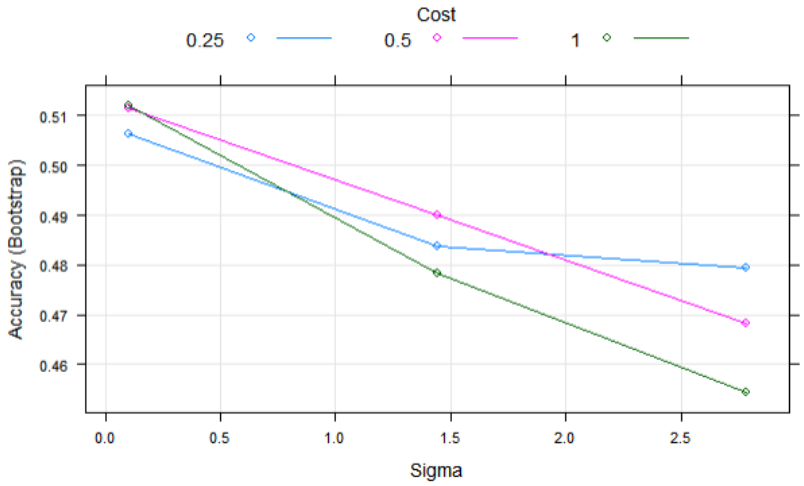
SUPPORT VECTOR MACHINE

SVM offers very high accuracy compared to other classifiers such as logistic regression, and decision trees. It is known for its kernel trick to handle nonlinear input spaces. It is used in a variety of applications such as face detection, intrusion detection, classification of emails, news articles and web pages, classification of genes, and handwriting recognition.

The data points that are closest to the hyperplane (a decision plane which separates between a set of objects having different class memberships).are called support vectors. By computing margins, these points will better define the dividing line. These points are more relevant to the construction of the classifier. The data points that are closest to the hyperplane are called support vectors. By computing margins, these points will better define the dividing line. These points are more relevant to the construction of the classifier. The data points that are closest to the hyperplane are called support vectors. By computing margins, these points will better define the dividing line. These points are more relevant to the construction of the classifier. The data points that are closest to the hyperplane are called support vectors. By computing margins, these points will better define the dividing line. These points are more relevant to the construction of the classifier.

The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset

SVM MODEL FIT FOR WAB RETURN

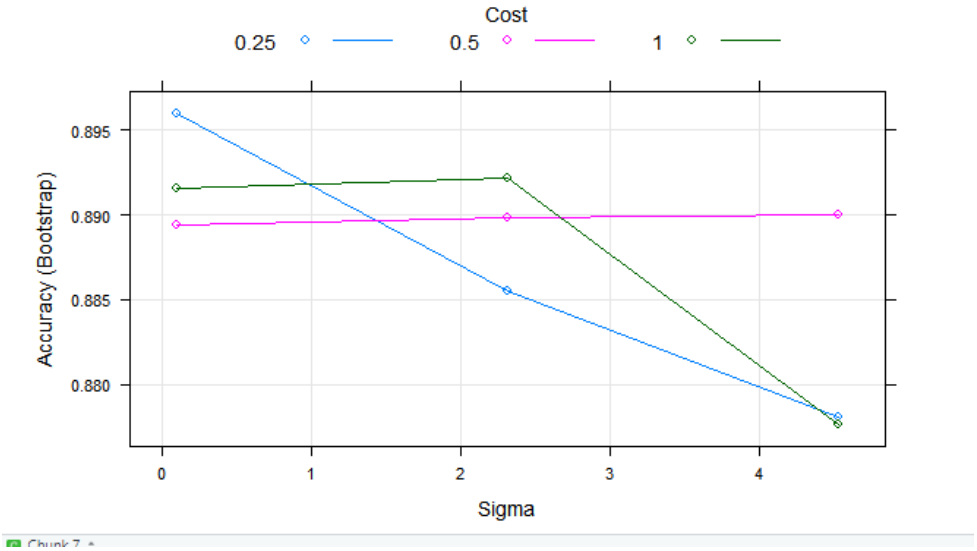


we are getting the accuracy of 54. 67% which has the cost hyper parameter is 1. After tuning our hyper parameters, we got an optimal value for sigma = 0.1040489 and C = 1. This value of grid radial will be given to the train's method's tune grid parameter.

For svm radial classifier, its accuracy is 52. 67%. So, it shows that the radial classifier is not giving us a better result as compared to the linear classifier even after tuning it. This can be as a

result of overfitting (this happens when the model is memorizing the data it as seen and it is unable to generalize on data unseen)

SVM MODEL FIT FOR WAB RISK



For WAB risk, we got the accuracy of 82. 67% which has the cost hyper parameter is 1. After tuning our hyper parameters, we got an optimal value for sigma = 1.679779 and C = 0.5. This value of grid radial will be given to the train's method's tune grid parameter.

For svm radial classifier, its accuracy is 88. 7%. So, it shows that the radial classifier is not giving us a better result as compared to the linear classifier even after tuning it. This can be as a

MODEL ACCURACY SUMMARY FOR WAB RETURN	
Forecast accuracy/Model	
Logistic regression	49.94%
Naïve Bayes	42.07%
K-NN classification	52.00%
Support vector machine	54.67%

Comparing these four models together, we can see that overall the model performed fairly and the most efficient model for predicting the WAB return rate is the support vector machine

MODEL ACCURACY SUMMARY FOR WAB RISK	
Forecast accuracy/Model	
Logistic regression	49.94%
Naïve Bayes	87.42%
K-NN classification	84.67%
Support vector machine	88.7%

Looking at the summary for WAB risk, the Naïve Bayes classifier is seen to have an high accuracy result in line with the svm model. The reason for this can be suspected to be the assumptions of the naïve bayes classifier, which also follows the pattern of the data set WAB risk, hence the high accuracy.