# **FINAL PROJECT REPORT**

# **Build Classification Model for Insurance Company (CoIL 2000) Dataset**

Navya Sai Paladugu (nxp153030)

Raghunadh Metta (rxm153530)

**Purpose:** The main aim of this project is to build an accurate classification model that will predict the target attribute of the given dataset.

**Dataset(s):**  We used the following data set for this assignment:

1. **Insurance Company Benchmark (CoIL 2000) Dataset:** This dataset is available from <https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+%28COIL+2000%29> .

**Approach:**

## **Data Cleaning and Preprocessing:**

Below is the procedure followed for data cleaning and pre-processing of data.

1. We downloaded the data and gave the colnames() so as it will be easy for the further processing. The dataset has 86 variables. So instead of writing all the 86 column names in colnames() we got to find that the same dataset is present in ISLR library, after some internet search, so we just took the column names of that dataset and applied it to our dataset.
2. Out of 86 variables not all variables are useful. So as to select the useful variables from both the sociodemographic data (customer data) and product ownership data, we analyzed the dataset using pie charts and bar plots.
3. We observed that out of 5822 customer records most of them does not have the insurance policy. This is found from the target variable (86th variable) ‘CARAVAN’ (See Figure-01)

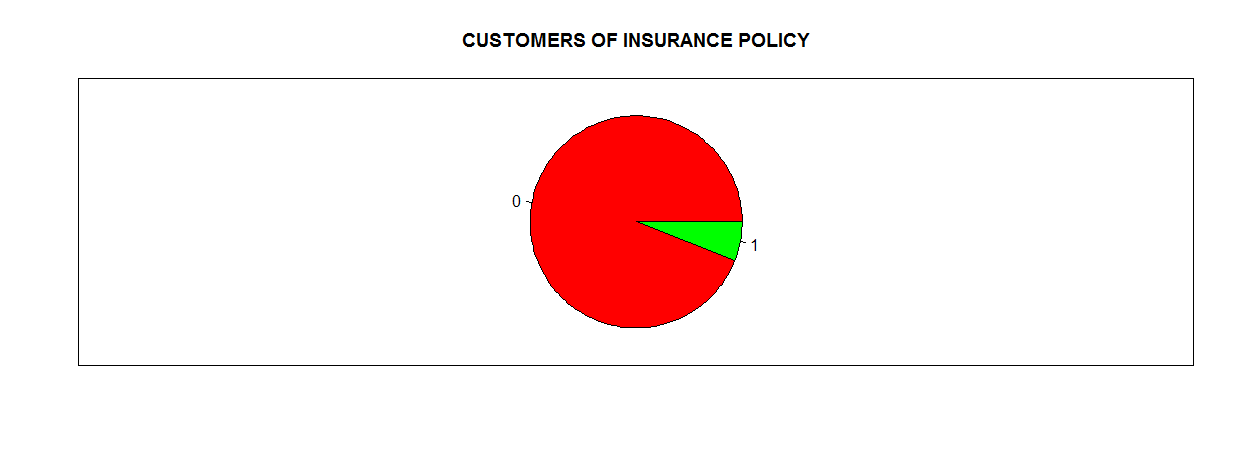


Figure-01: Summary of selected features of Dataset – New Dataset

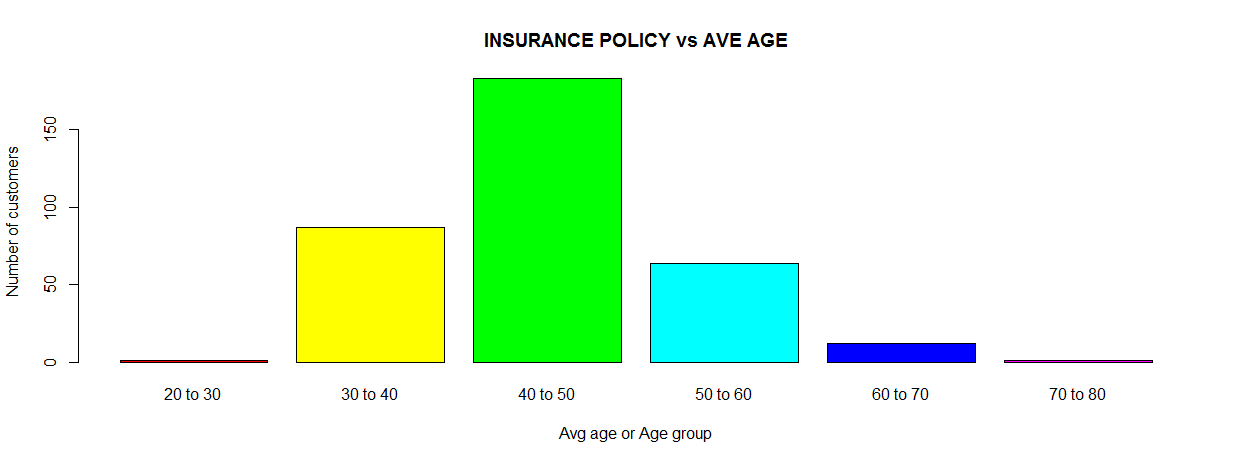
1. We analyzed other customer data variables (1-43) and product ownership data variables (44-85) and how these variables effect target variable ‘CARAVAN: Number of mobile home policies’ (86). Some of the barplots are shown below. Figure-02 to Figure-05. 

Figure-02: Barplot of Number Insurance Policy holders vs. Average Age group

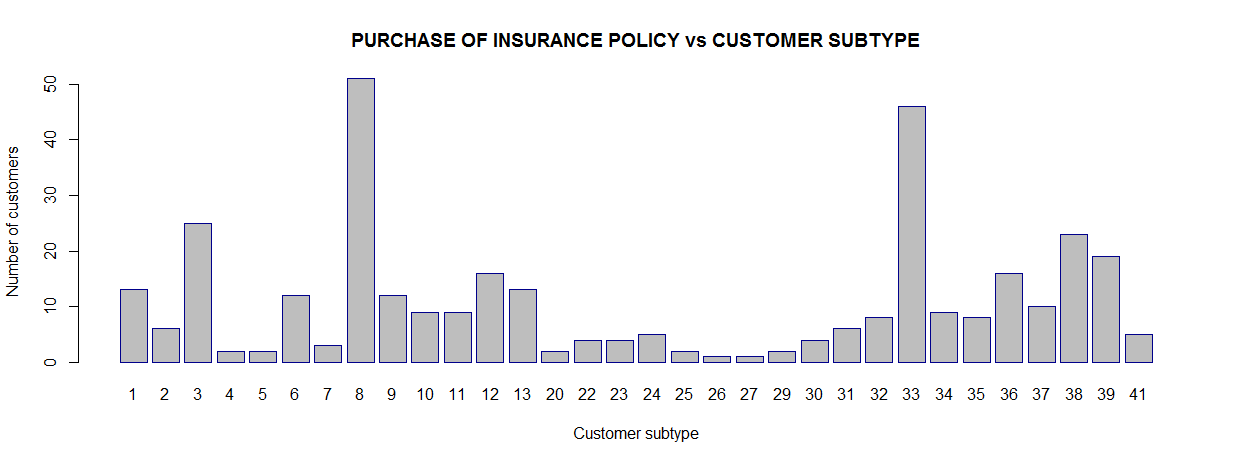


Figure-03: Barplot of Number Insurance Policy holders vs. Customer Subtype

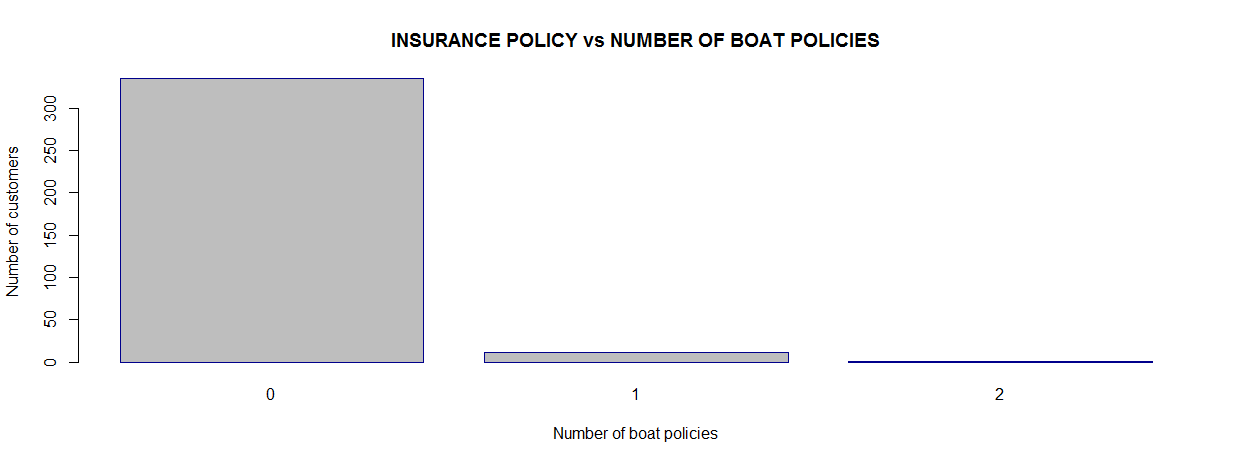


Figure-04: Barplot of Number Insurance Policy holders vs. No. of Boat Policies

1. The new dataset after selecting 9 variables (customer data - 5, product data - 4) is formed.
2. The variables selected are: ‘MOSTYPE – Customer Subtype’, ‘MGEMLEEF – Average Age Group’, ‘MKOOPKLA – Purchasing Power Class’, ‘MINKGEM – Average Income’, ‘MOSHOOFD – Customer Main Type’, ‘APLEZIER – No. of Boat Policies’, ‘ABYSTAND – No. of Social Security Insurance Policies’, ‘PPERSAUT – Contribution Car Policies’, ‘ABRAND – No. of Fire Policies’. Summary is shown in Figure-05.

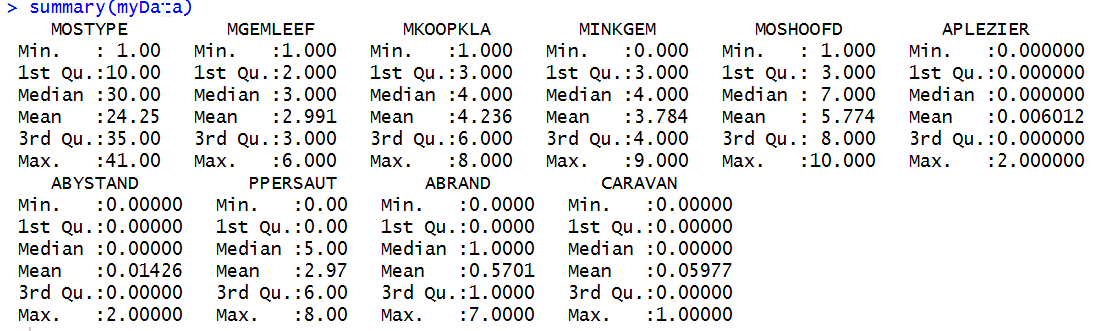


Figure-05: Summary of selected features of Dataset – New Dataset

1. We selected the same variables for the test data provided from ‘ticeval2000.txt’ file. See Figure-06.

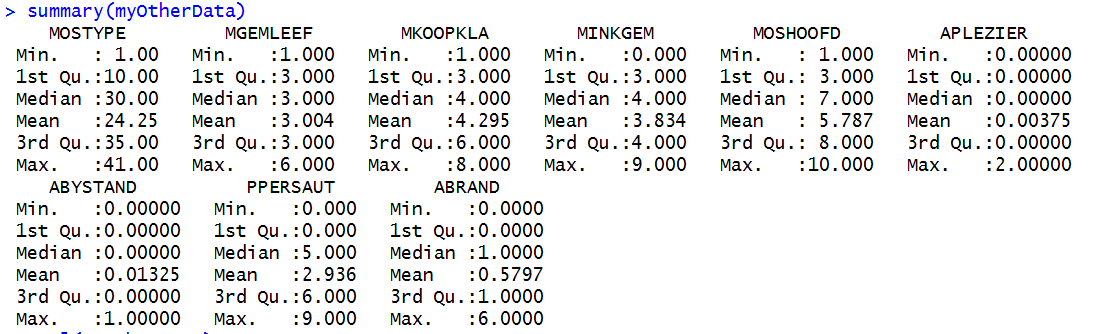


Figure-06: Summary of selected features of Train Dataset from ticeval2000.txt – New train Dataset

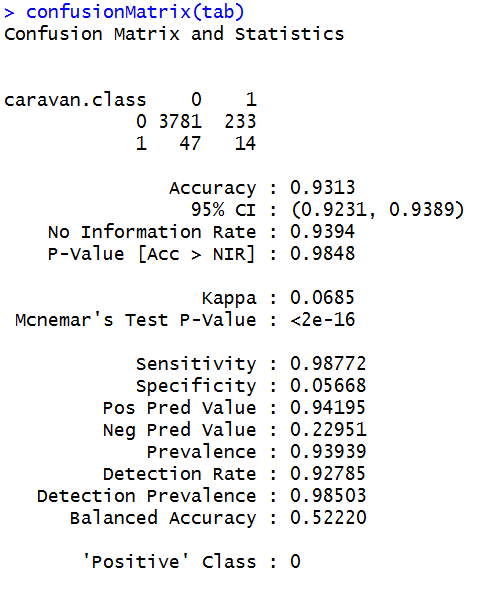
1. Out of the 5822 customer records, we considered 4075 as training data and remaining 1747 as test data wherever necessary in the classification models. As there are no missing values as described in the description of dataset we considered every record.
2. The model is cross-validated and also the model is applied on the given testing dataset which is compared with the given target data.
3. Here on we will use the following notion:
   * *myData* – data obtained from ticdata2000.txt file i.e training data.
   * *myData train* – 4075 records of the myData
   * *myData test* – 1747 records of the myData
   * *testData* – data obtained from ticeval2000.txt i.e test data
   * *target* – target data from tictgts2000.txt i.e target data for the myOtherData

## **Modelling:**

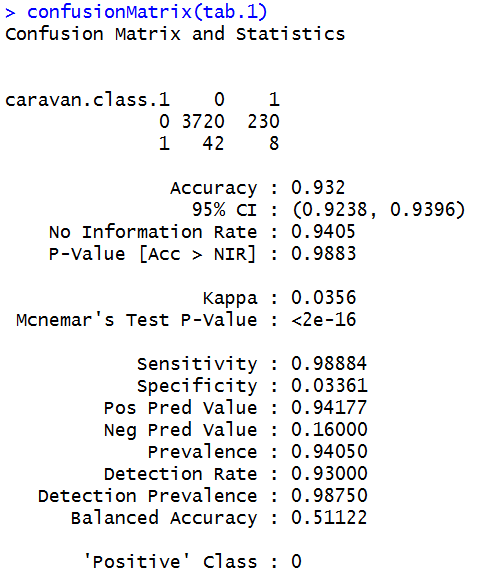
We applied following models on both myData and testData dataset.

### **LDA:**

* We cross-validated LDA on the ‘*myData*’ and obtained the following confusion matrix with accuracy 93.13%:

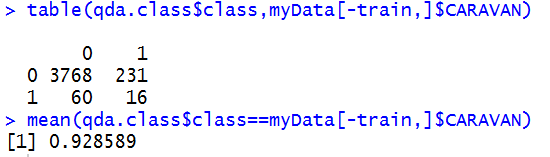


* We applied the same on the ‘testData’ and obtained following confusion matrix with accuracy 93.2%:

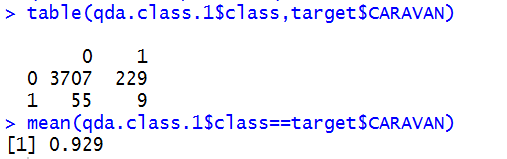


### **QDA:**

* We cross-validated QDA on the ‘*myData*’ and obtained the following with accuracy 92.86%:

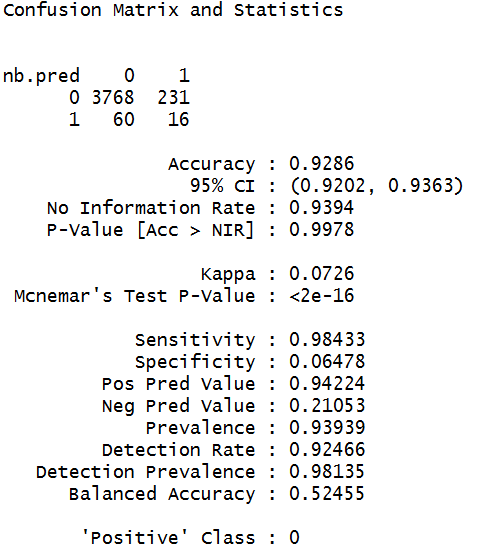


* We applied the same on the ‘testData’ and obtained following with accuracy 92.9%:

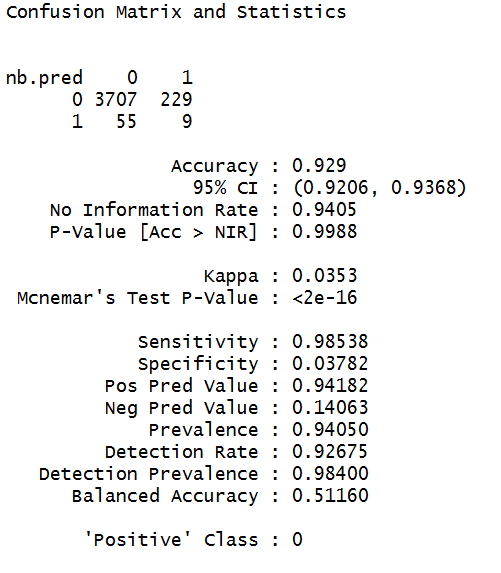


### **Naïve Bayes:**

* We cross-validated Naïve Bayes on the ‘*myData*’ and obtained the following with accuracy 92.86%:

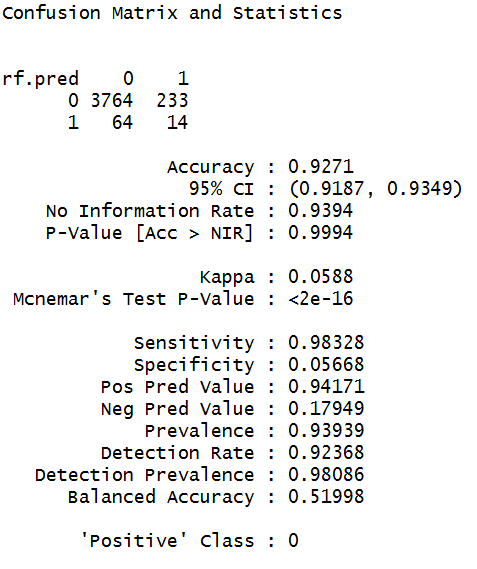


* We applied the same on the ‘testData’ and obtained following with accuracy 92.9%:

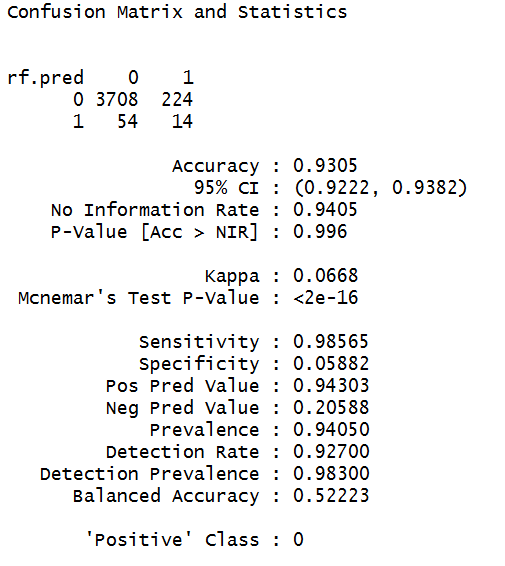


### **Random Forest:**

* We cross-validated Random Forest on the ‘*myData*’ and obtained the following with accuracy 92.71%:

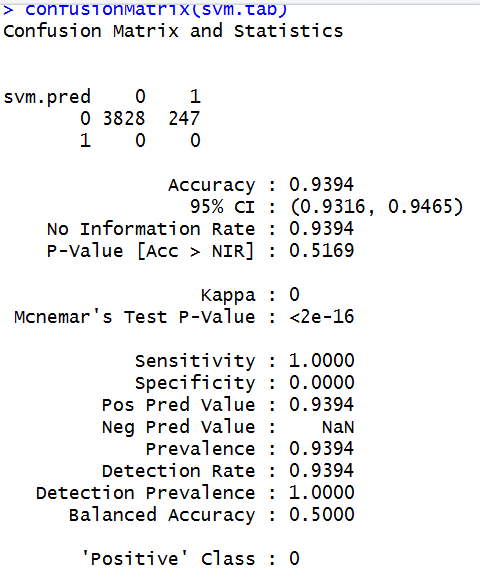


* We applied the same on the ‘testData’ and obtained following with accuracy 93.05%:

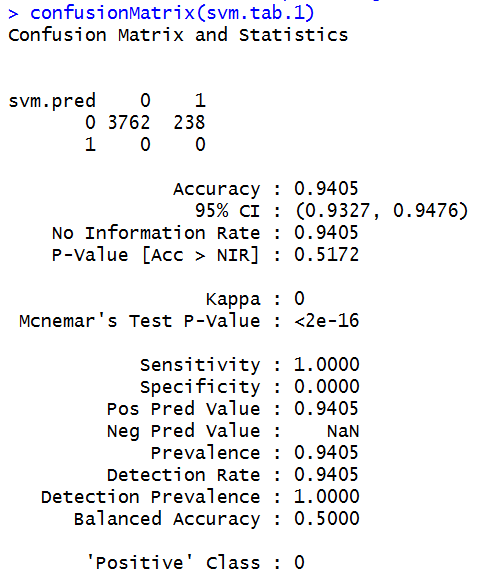


### **SVM – Linear Kernel:**

* We cross-validated Linear Kernel on the ‘*myData*’ and obtained the following with accuracy 93.94%:

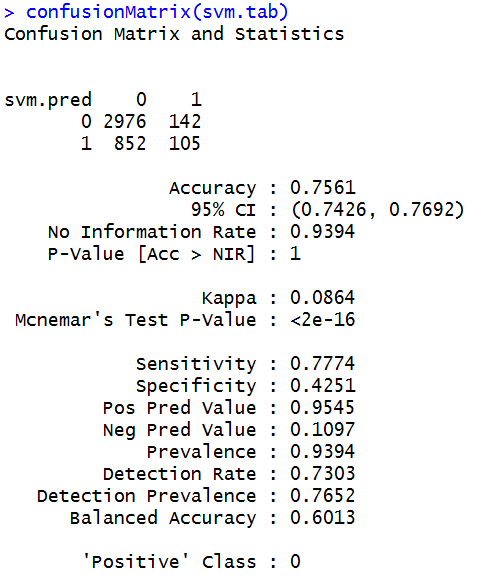


* We applied the same on the ‘testData’ and obtained following with accuracy 94.05%:

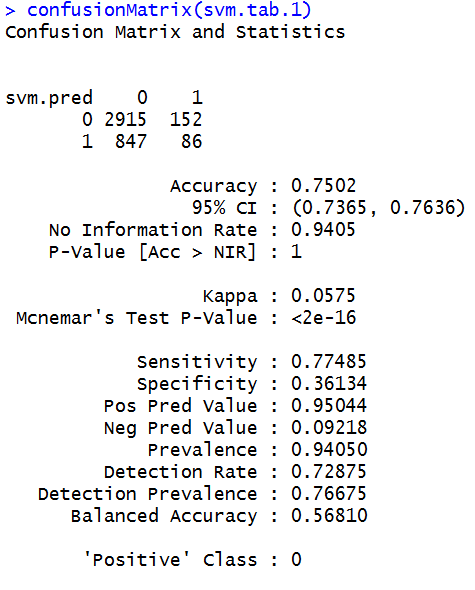


### **SVM – Polynomial Kernel:**

* We cross-validated Polynomial Kernel on the ‘*myData*’ and obtained the following with accuracy 75.61%:

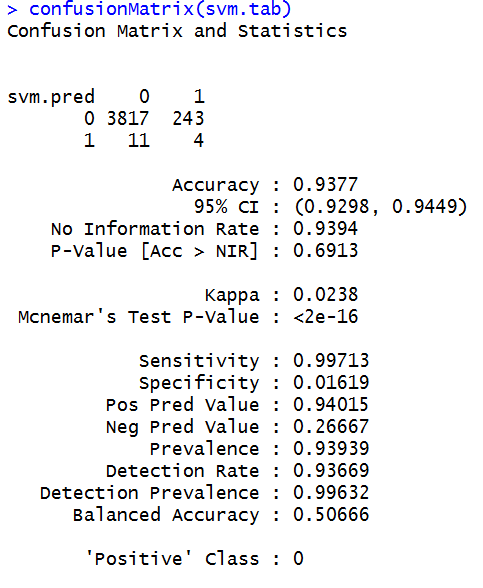


* We applied the same on the ‘testData’ and obtained following with accuracy 75.02%:

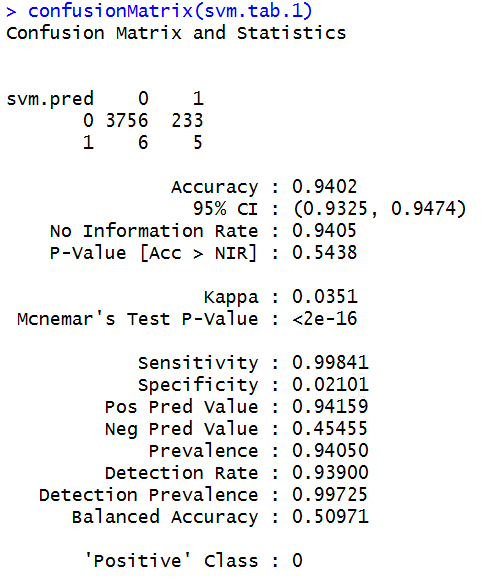


### **SVM – Radial Kernel:**

* We cross-validated Radial Kernel on the ‘*myData*’ and obtained the following with accuracy 93.77%:



* We applied the same on the ‘testData’ and obtained following with accuracy 94.02%:



### **Neural Net:**

* We cross-validated Neural net of one hidden layer and 3 nodes on the ‘*myData*’ and obtained the following with RMSE 0.305:
* We applied the same on the ‘testData’ and obtained following with RMSE 0.32

## **Conclusion:**

The following table shows the summary of the accuracy obtained from different models:

|  |  |  |
| --- | --- | --- |
| Models | Cross Validation Data Set (%) | Test Data Set (%) |
| LDA | 93.13 | 93.2 |
| QDA | 92.86 | 92.9 |
| Naïve Bayes | 92.86 | 92.9 |
| Random Forest | 92.71 | 93.05 |
| SVM – Linear | 93.94 | 94.05 |
| SVM – Polynomial | 75.61 | 75.02 |
| SVM – Radial | 93.77 | 94.02 |

By comparing the accuracy obtained from different models examined for our test dataset, we concluded that SVM – Linear is the best model suitable for modelling this dataset as the accuracy obtained from this model is 93.94% (cross-validation accuracy) or 94.05% (test data accuracy) which is highest among all the models.

Apart from SVM -Linear Kernel model, LDA and Random Forest can also be used as models for this dataset.

*The observations made from this project are:*

* *The Selection of the features which contribute most to the prediction of the data is very important.*
* *After analyzing all the different combination features, we thought the 9 features we chose were sufficient as even with more features there was no change in the final best model selection. Hence, we used only 9 features.*
* *We chose to run the cross-validation on the myData because we considered testData to be the data for which we should finalize the model for. So we finalized the model based on cross-validation accuracy (SVM -Linear – 93.94%) rather than test data accuracy (as in table above). Though SVM- linear model gave us both accuracies high.*