Introduction

The analysis of real-world data set contributes to solving the problem in real-world such as medical section. Diabetes can be a large problem in the world about healthcare. Diabetes can be a dangerous and chronic disease because complications can be occurred. The diabetic patients suffer from the side effect, so they re-admit the hospital. The 130 US hospital collected the patient’s information related to diabetes. In order to analyse above, Machine Learning and Deep Learning can be used. Machine Learning contributes to the prediction of readmission such as logistic regression, SVM. Moreover, when the readmission can be predicted, the doctor can be easy to diagnose and treat diabetic patients. For the patients, it can be identified how to deal with their health because diabetes is a lifestyle disease. Thus, this paper will approach identifying risk factor re-admission and predicting. Some public paper can make a request to be able to solve the diabetic problem by using Machine Learning or Deep Learning algorithms.

Methods

The logistic regression is able to predict a qualitative response as an instance of the classification technique. The logistic regression can measure the relationship between the dependent and one or more independent variables about categorical data. The logistic regression can be a binary logistic regression as Binomial Logistic Regression, thus logistic regression models can be the probability that readmission belongs to a particular category. However, the readmission category in this data set separates from three parts. One is patients do not re-admit and another is patients re-admit less than 30 days. The other is more than 30 days. Thus, the logistic regression model can be necessary to be prepared two different models.

The support vector machine can be related with learning algorithms by classifying and analysing regression for analysing data set. Like logistic regression, the data set separates from the three parts. The models are divided by two. By using the SVM, the models can be analysed thus the objective of this algorithm is able to find a hyperplane in an n-dimensional space that can classify the data points. As a result, when maximizing the margin distance can be found, the maximizing the margin distance supports some reinforcement. The p-value can be shown for identifying the validation of the prediction by using chi-square test.

The random forest can be ensemble learning method by classifying, analysing regression and constructing a multitude of decision tree for analysing data setRandom forest can be a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. By using the random forest, the models can be analysed without hyper-parameter tuning. The random forest can perform in order to understand and use the various options. Most of the options depend on two data objects generated. Like logistic regression, the data set separates from the three parts and models are divided by two. The p-value can be shown for identifying the validation of the prediction by using chi-square test.

For the ROC curve of Model 1, these plot can be analysed and identified the model performance can be significant by using methods such as Logistic regression, Support vector machine and Random Forest by train and test, thus this curve can depict the accuracy of a single test. In addition, the areas under ROC curve (AUC) can describe to compare the usefulness of tests. This curve use by plotting the true positive rate and false positive rate. The curve can be composed of sensitivity and 1-specifjcity. When the threshold can be close to coordinate (0, 1), the curve approach to be perfect classifier. In the upper left corner, 100% sensitivity and specificity are represented. As a result, when the sensitivity and specificity are higher, this model can be significant. The cut point can be decided through some methods such as maximum value of sum of sensitivity and specificity, Youden index and Euclidean distance. For the cutoff values can be shown below table.

For this Model 1, the ROC curve can be moderate and the model can be necessary to advance and compare to other methods in order to find the optimal methods. The AUC is able to identify accuracy of model. The criteria of AUC is the following sentences. When AUC is equal to 0.5, the model has no discrimination. When AUC can be located in between 0.5 and 0.7, it is less significant model. Moreover, more than 0.7 and less than 0.9, this model is significant. Finally, more than 0.9, it is strongly significant.

The one of the model as Model 1 can be whether patients re-admit less than 30 days (class 1) or not readmission and readmission more than 30 days (class 0). This dataset can be imbalanced so accuracy of this model can be calculated by confusion matrix. So the train data set is 35,000 and test set is 15,000 because all of data set can be approximately 100,000 and validation part is about 50,000 as 50% of all of data set.

Next, the other model as Model 2 can be whether patients re-admit less than or more than 30 days by using methods such as Logistic regression, Support vector machine and Random Forest by train and test. The not readmission designates NA and then NA is omitted. So the train data set is 15,995 and test set is 6855 because all of data set is 45700 and validation part is 22850 as 50% of all of data set.

Prevalence of class in Model 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Train | Percentage | Test | Percentage |
| 0 (Admission>30 days or no re-admission) | 30,951 | 88.43% | 13,225 | 88.17% |
| 1 (Admission in <30 day) | 4,049 | 11.57% | 1,775 | 11.83% |
| Sum | 35,000 | 100% | 15,000 | 100% |

Prevalence of class in Model 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Train | Percentage | Test | Percentage |
| 0 (Admission>30 days) | 4,162 | 26.02 | 1506 | 21.97% |
| 1 (Admission in < 30 days) | 11,833 | 73.98% | 5349 | 78.03% |
| Sum | 15,995 | 100% | 6855 | 100% |

Logistic Regression in Model 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std.Error | z value | Pr(>|z|) |
| (Intercept) | 7.1692 | 113.6334 | 0.063 | 0.949694 |
| race | 0.019895 | 0.023808 | 0.836 | 0.40336 |
| gender | 0.004546 | 0.034348 | 0.132 | 0.894703 |
| age | 0.002078 | 0.00117 | 1.777 | 0.075613 |
| admission\_type\_id | 0.001129 | 0.011916 | 0.095 | 0.924503 |
| discharge\_disposition\_id | 0.010825 | 0.002631 | 4.114 | 3.88E-05 |
| admission\_source\_id | -0.00689 | 0.004319 | -1.594 | 0.110946 |
| time\_in\_hospital | 0.020469 | 0.006251 | 3.275 | 0.001058 |
| num\_lab\_procedures | 0.001944 | 0.001084 | 1.794 | 0.072788 |
| num\_procedures | -0.057 | 0.011975 | -4.76 | 1.94E-06 |
| num\_medications | 0.01058 | 0.002671 | 3.962 | 7.44E-05 |
| number\_outpatient | 0.015431 | 0.019076 | 0.809 | 0.418573 |
| number\_emergency | 0.04805 | 0.024451 | 1.965 | 0.049399 |
| number\_inpatient | 0.257515 | 0.011414 | 22.562 | 2.00E-16 |
| number\_diagnoses | 0.052068 | 0.010049 | 5.181 | 2.20E-07 |
| max\_glu\_serum | -0.00922 | 0.037585 | -0.245 | 0.806244 |
| A1Cresult | -0.04481 | 0.019618 | -2.284 | 0.022375 |
| metformin | -0.09208 | 0.025842 | -3.563 | 0.000366 |
| repaglinide | 0.19583 | 0.072329 | 2.707 | 0.006779 |
| nateglinide | -0.12276 | 0.159327 | -0.77 | 0.441004 |
| chlorpropamide | -0.41613 | 0.289125 | -1.439 | 0.15007 |
| glimepiride | -0.05379 | 0.045443 | -1.184 | 0.236532 |
| acetohexamide | -5.02365 | 98.48385 | -0.051 | 0.959318 |
| glipizide | 0.009273 | 0.027582 | 0.336 | 0.736718 |
| glyburide | -0.00595 | 0.028167 | -0.211 | 0.832769 |
| tolbutamide | -0.21118 | 0.520149 | -0.406 | 0.68475 |
| pioglitazone | 0.017521 | 0.037016 | 0.473 | 0.635967 |
| rosiglitazone | -0.06159 | 0.036291 | -1.697 | 0.089651 |
| acarbose | -0.15093 | 0.173226 | -0.871 | 0.3836 |
| miglitol | 0.456173 | 0.566278 | 0.806 | 0.420495 |
| troglitazone | -4.84271 | 56.67931 | -0.085 | 0.931911 |
| tolazamide | -0.21394 | 0.36325 | -0.589 | 0.555888 |
| examide | NA | NA | NA | NA |
| citoglipton | NA | NA | NA | NA |
| insulin | 0.021657 | 0.023285 | 0.93 | 0.352309 |
| glyburide.metformin | -0.00857 | 0.234139 | -0.037 | 0.9708 |
| glipizide.metformin | NA | NA | NA | NA |
| glimepiride.pioglitazone | NA | NA | NA | NA |
| metformin.rosiglitazone | NA | NA | NA | NA |
| metformin.pioglitazone | NA | NA | NA | NA |
| change | 0.054017 | 0.047558 | 1.136 | 0.25603 |
| diabetesMed | 0.168729 | 0.05761 | 2.929 | 0.003402 |

Logistic Regression in Model 2

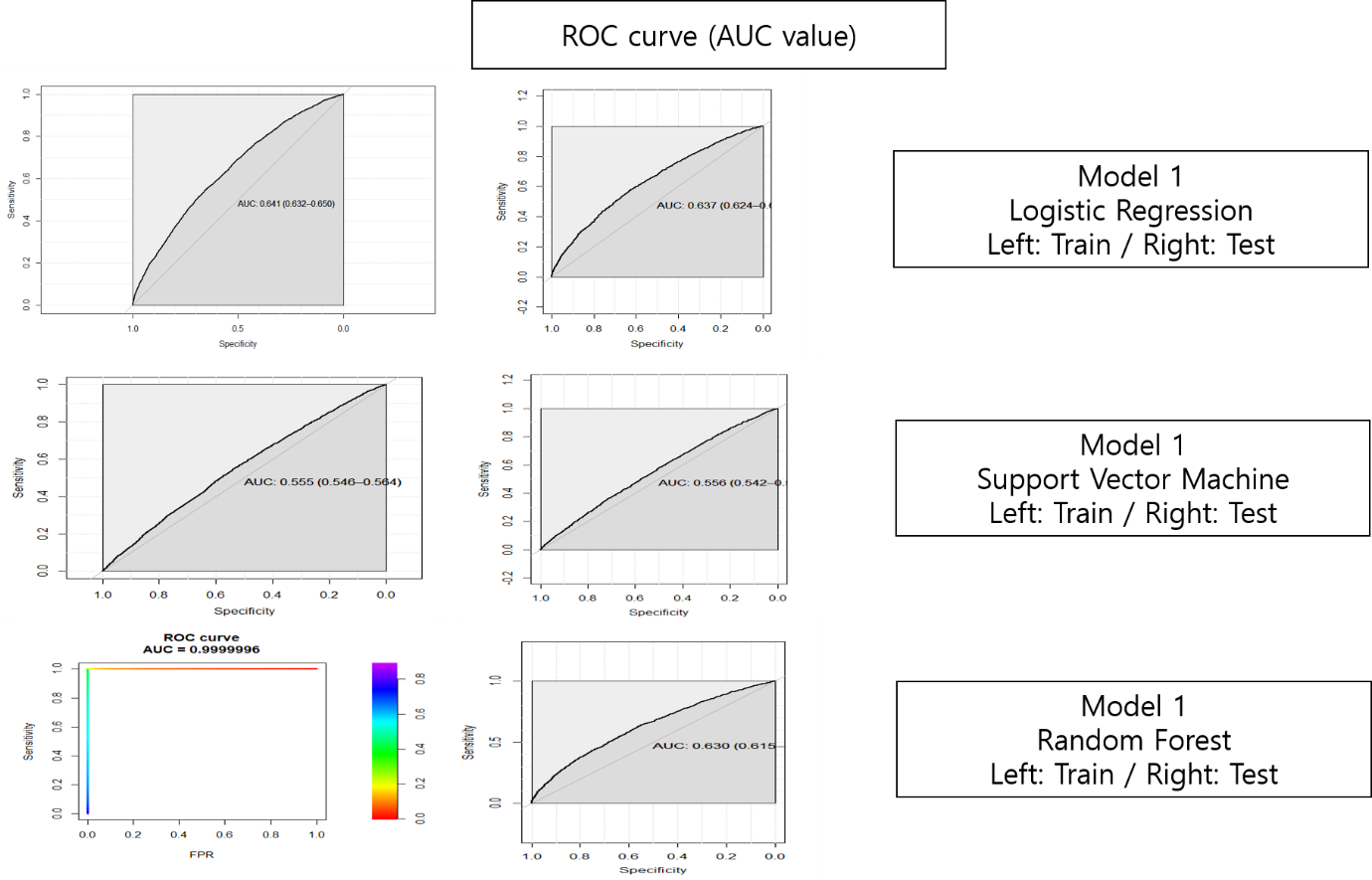
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std.Error | z value | Pr(>|z|) |
| (Intercept) | 9.70046 | 120.584 | 0.08 | 0.93588 |
| race | -0.02698 | 0.02599 | -1.038 | 0.29932 |
| gender | 0.02317 | 0.036919 | 0.628 | 0.53028 |
| age | 0.003319 | 0.001262 | 2.629 | 0.00856 |
| admission\_type\_id | -0.05376 | 0.0124 | -4.335 | 1.46E-05 |
| discharge\_disposition\_id | 0.020214 | 0.002854 | 7.081 | 1.43E-12 |
| admission\_source\_id | -0.01399 | 0.004453 | -3.142 | 0.00168 |
| time\_in\_hospital | 0.010922 | 0.006786 | 1.61 | 0.10748 |
| num\_lab\_procedures | -0.00205 | 0.001166 | -1.76 | 0.0784 |
| num\_procedures | -0.01292 | 0.012643 | -1.022 | 0.30682 |
| num\_medications | 0.014337 | 0.002959 | 4.845 | 1.26E-06 |
| number\_outpatient | -0.05663 | 0.020933 | -2.705 | 0.00682 |
| number\_emergency | 0.020255 | 0.023983 | 0.845 | 0.39836 |
| number\_inpatient | 0.132712 | 0.011837 | 11.212 | 2.00E-16 |
| number\_diagnoses | -0.00672 | 0.010809 | -0.622 | 0.53391 |
| max\_glu\_serum | 0.073749 | 0.038282 | 1.926 | 0.05404 |
| A1Cresult | -0.04028 | 0.020816 | -1.935 | 0.05298 |
| metformin | -0.05636 | 0.027508 | -2.049 | 0.04048 |
| repaglinide | 0.19188 | 0.078731 | 2.437 | 0.0148 |
| nateglinide | -0.20712 | 0.150716 | -1.374 | 0.16937 |
| chlorpropamide | -0.50116 | 0.295535 | -1.696 | 0.08993 |
| glimepiride | -0.03541 | 0.048763 | -0.726 | 0.46771 |
| acetohexamide | -5.59577 | 98.48385 | -0.057 | 0.95469 |
| glipizide | 0.003856 | 0.029242 | 0.132 | 0.89509 |
| glyburide | 0.007625 | 0.030208 | 0.252 | 0.80072 |
| tolbutamide | 0.168421 | 0.613109 | 0.275 | 0.78355 |
| pioglitazone | -0.00737 | 0.039132 | -0.188 | 0.85062 |
| rosiglitazone | -0.0841 | 0.038374 | -2.192 | 0.02841 |
| acarbose | -0.18329 | 0.181818 | -1.008 | 0.3134 |
| miglitol | 0.55313 | 0.685106 | 0.807 | 0.41946 |
| troglitazone | -5.25848 | 69.57159 | -0.076 | 0.93975 |
| tolazamide | -0.23212 | 0.378779 | -0.613 | 0.54 |
| examide | NA | NA | NA | NA |
| citoglipton | NA | NA | NA | NA |
| insulin | 0.025637 | 0.024527 | 1.045 | 0.29591 |
| glyburide.metformin | 0.059198 | 0.262081 | 0.226 | 0.8213 |
| glipizide.metformin | NA | NA | NA | NA |
| glimepiride.pioglitazone | NA | NA | NA | NA |
| metformin.rosiglitazone | NA | NA | NA | NA |
| metformin.pioglitazone | NA | NA | NA | NA |
| change | 0.009877 | 0.050169 | 0.197 | 0.84392 |
| diabetesMed | 0.030923 | 0.061576 | 0.502 | 0.61554 |

Identifying each Methods by model

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model 1 | Cutoff Point | AUC | Accuracy | Sensitivity | Specificity | PPV (precision) | Beta value | P-value |
| Logistic Regression  (Model 1 / Train) | 0.6867 | 0.641 | 0.8846 | 0.9997 | 0.0047 | 0.8848 | 7.1692 | 0.9497 |
| Logistic Regression  (Model 1 / Test) | 0.6867 | 0.637 | 0.8817 | 1 | 0 | 0.8817 | 7.1692 | 0.9497 |
| Support Vector Machine  (Model 1 / Train) | 0.1006 | 0.555 | 0.8843 | 1 | 0 | 0.8843 |  | 0.4945 |
| Support Vector Machine  (Model 1 / Test) | 0.1006 | 0.556 | 0.8817 | 1 | 0 | 0.8817 |  | 0.4962 |
| Random Forest  (Model 1 / Train) | 0.3369 | 1.000 | 0.0001 | 0.0001 | 0.0002 | 0.0000 |  | 0.0000 |
| Random Forest  (Model 1 / Test) | 0.3369 | 0.630 | 0.8817 | 0.9999 | 0.0001 | 0.8817 |  | 0.2884 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model 2 | Cutoff Point | AUC | Accuracy | Sensitivity | Specificity | PPV (precision) | Beta value | P-value |
| Logistic Regression  (Model 2/ Train) | 0.5735 | 0.598 | 0.2595 | 0.9942 | 0.0011 | 0.2593 | -0.1073 | 0.8831 |
| Logistic Regression  (Model 2 / Test) | 0.5735 | 0.575 | 0.2120 | 0.9994 | 0.0005 | 0.2196 | -0.1073 | 0.8831 |
| Support Vector Machine  (Model 2 / Train) | 0.9008 | 0.556 | 0.2602 | 1 | 0 | 0.2602 |  | 0.4941 |
| Support Vector Machine | 0.9008 | 0.581 | 0.2197 | 1 | 0 | 0.2197 |  | 0.4943 |
| Random Forest  (Model 2 / Train) | 0.5857 | 1.000 | 0.2602 | 1 | 0 | 0.9989 |  | 0.0000 |
| Random Forest  (Model 2 / test) | 0.5857 | 0.595 | 0.7707 | 0.0202 | 0.9961 | 0.6087 |  | 0.374 |

ROC curve of Model 1



Confusion Matrix of train set in Model 1 (LR)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission>30 days or no re-admission) | 1 (Admission in <30 day) |
| Prediction | 0 (Admission>30 days or no re-admission) | 30942 | 4030 |
| 1 (Admission in <30 day) | 9 | 19 |

Confusion Matrix of test set in Model 1 (LR)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission>30 days or no re-admission) | 1 (Admission in <30 day) |
| Prediction | 0 (Admission>30 days or no re-admission) | 13225 | 1775 |
| 1 (Admission in <30 day) | 0 | 0 |

Confusion Matrix of train set in Model 1 (SVM)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission>30 days or no re-admission) | 1 (Admission in <30 day) |
| Prediction | 0 (Admission>30 days or no re-admission) | 30951 | 4049 |
| 1 (Admission in <30 day) | 0 | 0 |

Confusion Matrix of test set in Model 1 (SVM)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission>30 days or no re-admission) | 1 (Admission in <30 day) |
| Prediction | 0 (Admission>30 days or no re-admission) | 13225 | 1775 |
| 1 (Admission in <30 day) | 0 | 0 |

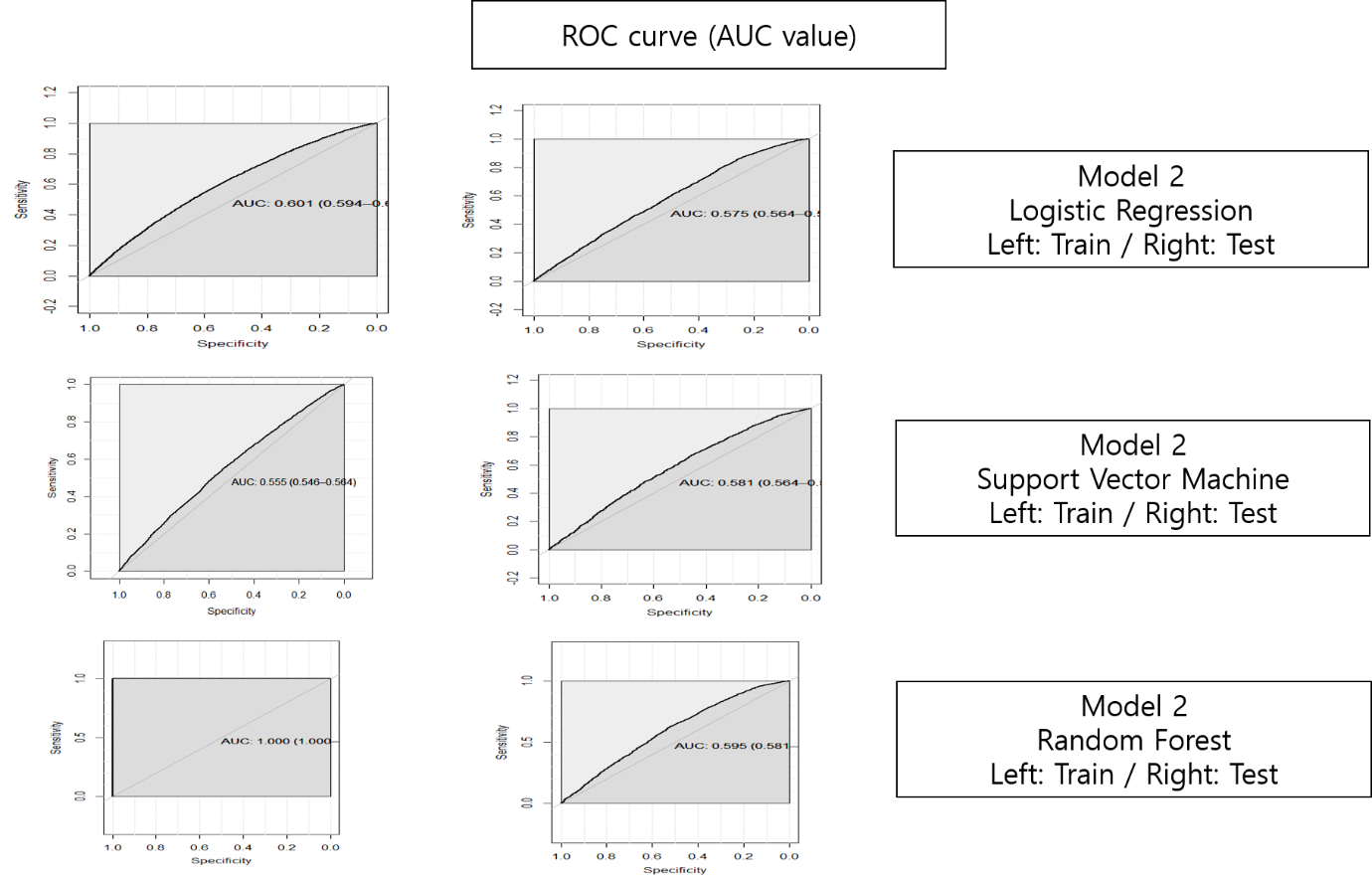
Confusion Matrix of train set in Model 1 (RF)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission>30 days or no re-admission) | 1 (Admission in <30 day) |
| Prediction | 0 (Admission>30 days or no re-admission) | 4 | 4048 |
| 1 (Admission in <30 day) | 30947 | 1 |

Confusion Matrix of test set in Model 1 (RF)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission>30 days or no re-admission) | 1 (Admission in <30 day) |
| Prediction | 0 (Admission>30 days or no re-admission) | 13224 | 1774 |
| 1 (Admission in <30 day) | 1 | 1 |

ROC curve of Model 2



Confusion Matrix of train set in Model 2 (LR)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission in > 30 day) | 1 (Admission in < 30 days) |
| Prediction | 0 (Admission in>30 day) | 4138 | 11820 |
| 1 (Admission in < 30 days) | 24 | 13 |

Confusion Matrix of test set in Model 2 (LR)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission in > 30 day) | 1 (Admission in < 30 days) |
| Prediction | 0 (Admission in>30 day) | 1505 | 5346 |
| 1 (Admission in < 30 days) | 1 | 3 |

Confusion Matrix of train set in Model 2 (SVM)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission in > 30 day) | 1 (Admission in < 30 days) |
| Prediction | 0 (Admission in>30 day) | 4162 | 11833 |
| 1 (Admission in < 30 days) | 0 | 0 |

Confusion Matrix of test set in Model 2 (SVM)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission in > 30 day) | 1 (Admission in < 30 days) |
| Prediction | 0 (Admission in>30 day) | 1506 | 5349 |
| 1 (Admission in < 30 days) | 0 | 0 |

Confusion Matrix of train set in Model 2 (RF)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission in > 30 day) | 1 (Admission in < 30 days) |
| Prediction | 0 (Admission in>30 day) | 3519 | 4 |
| 1 (Admission in < 30 days) | 0 | 9997 |

Confusion Matrix of test set in Model 2 (RF)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | 0 (Admission in > 30 day) | 1 (Admission in < 30 days) |
| Prediction | 0 (Admission in>30 day) | 42 | 27 |
| 1 (Admission in < 30 days) | 2037 | 6894 |

As a result, as can be shown the results’ table, the logistic regression and SVM by applying Model 1 can be higher but when by applying Model 2, only random forest is able to have high accuracy. Moreover, as shown the AUC values, the logistic regression can be high so these data set by applying both models, the logistic regression performance can be significant.