Prepared by : DharmaTeja Chelamalasetty, id : DXC210022 Date : 07/06/2022

BUAN 6341 Applied Machine Learning

ASSIGNMENT NO 2

Seoul Rental Bike prediction Part II

Executive Summary

- **Converted the data set to 2 classes based on Rented Bike Count median value, if <= Median it is treated as class 0, otherwise class 1.**
- **After Experimenting with SGD classifier and observed at alpha=0.001 and tol=1e-6 model is doing best with train and test data with accuracies of 91.7 and 91.4 with AUC of 0.91 .**
- **After Experimenting with SVM classifier it is observed that at C=50.0 and kernel = RBF model is doing best with train and test accuracies of 97.8 and 92.5 with AUC of 0.93 .**
- **After Experimenting with DT using pre pruning and post pruning it is observed that max_depth=10 and cp=0.01 model is doing best with accuracies 88.8 and 88.8 with AUC of 0.89 .**
- **It is observed that DT is doing best when we consider Type 1 Error and SVM is doing best when we consider Type II Errors .**
- **Using K-Fold Cross validation technique it is concluded that there is variation in SGD and SVM classifier with unseen data but DT is consistent with unseen data or production data .**
- **So Far after all experimentation with careful consideration we conclude that DT is doing best w.r.t Type 1 Errors, AUC almost similar and with unseen data or production data .**

Introduction

In this project, the objectives were to convert the data set from assignment 1 to a binary classification problem by thresholding the output to a class label and implement Logistic, SVM, Decision Tree Models with different hyper parameters for Logistic and with different kernel ,penalty parameters for SVM and performing pre pruning and post pruning techniques to prune the decision tree and chose the best model with through experimentation for predicting the class 0 or 1 which is (0 is less than the median value or 1 is greater than the median value) .

About the Data

The dataset consists of 14 features and 8760 records. The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. In order to remove the serial correlation with time, hours variable is hot encoded and converted to dummies and ran the model .

Project Outline

The Project is outlined to have 4 parts:

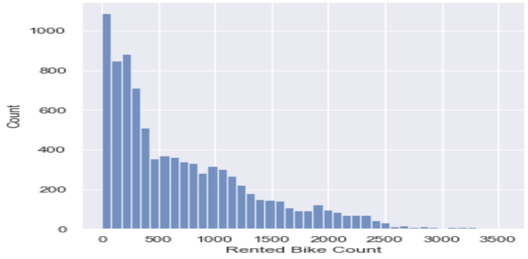
Part 1 : Convert the data set from assignment 1 to a binary classification problem by thresholding the output to a class label and implement logistic regression and experiment with different hyperparameters .

Part 2 : Implement SVM Classifier and experiment with different hyperparameters such as kernel and with different regularization parameter .

Part 3: Implement Decision tree and experimenting with different hyperparameters in pre pruning and post pruning techniques to prune the tree .

Part 4: Use Cross validation technique with optimized models of Logistic Regression, SVM Classifier, Decision using K-Fold technique to check the accuracies of the models in order to deploy in production .

Part 1:

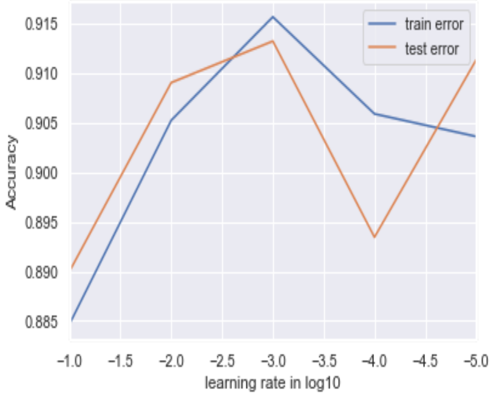


We can observe from the figure that the data of the rented bike count is right skewed. So, we apply a median technique to divide the data in 2 distributions and to handle outliers.
Since the Median value is 504.5., Less than or equal to the value is considered as class 0 and greater than the value is considered as class 1.

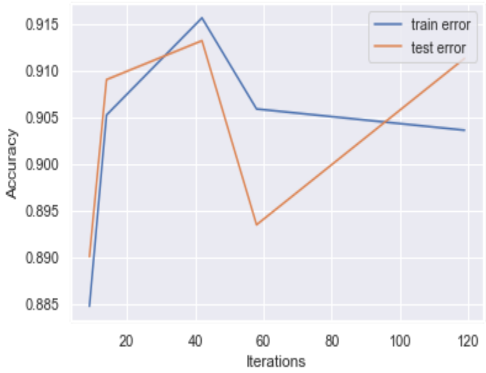
SGD Classifier :

We will use SGD classifier as logistic regressor .

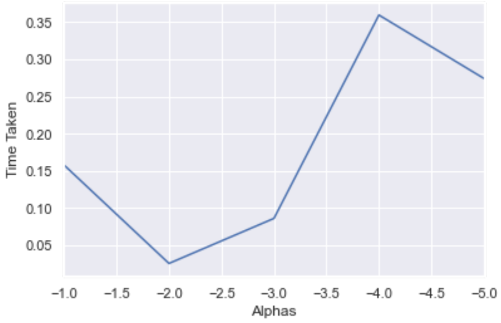
After thorough experimentation keeping threshold at 1e-3 and with learning rate alpha, below are the observations :



From the experimentation it is observed that both train error and test error are travelling in same direction, with decrease in alpha, accuracies for train and test are improving till 0.001 and post that train accuracy is still fine but test accuracy is decreasing and then again increasing and making model leading to **high variance** situation.
When **alpha = 0.001** both train and test accuracies seem to be better and converged.

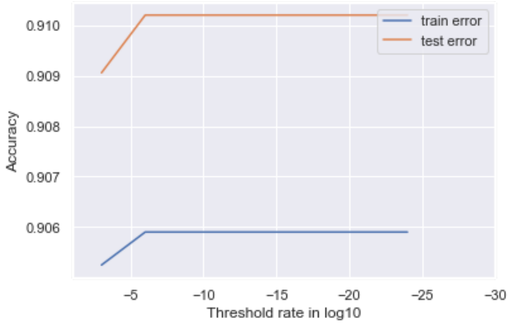


We can observe that with increase of iterations train accuracy is travelling from ideal to underfit or bias situation, whereas test accuracy is travelling from ideal to high variance situation and making model instable.
At 40 iterations we can observe that train and test accuracies seems ideal and converged.

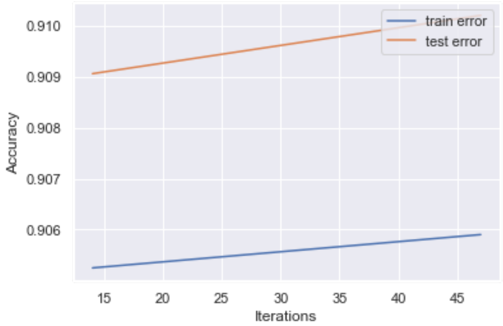


As Alpha decreases, we can observe that time taken for convergence is increased except at alpha = 0.01.
We can say that at **alpha = 0.001** model has ideal time to converge.

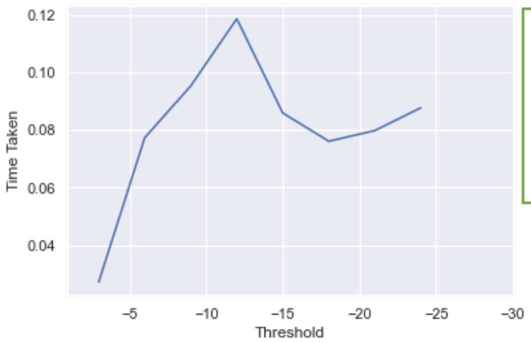
After thorough experimentation with best alpha = 0.001 and various values of threshold, below are the observations :



We can see that after threshold value 1e-6 there is not much difference in the train and test accuracies.
So we can conclude that model is converged at **tol=1e-6**.



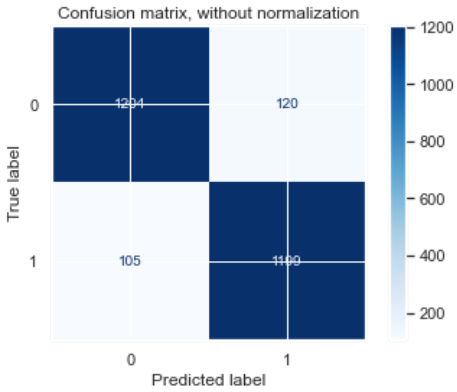
From the graph we can see that there is not much difference in accuracies with increase in the nr of iterations.

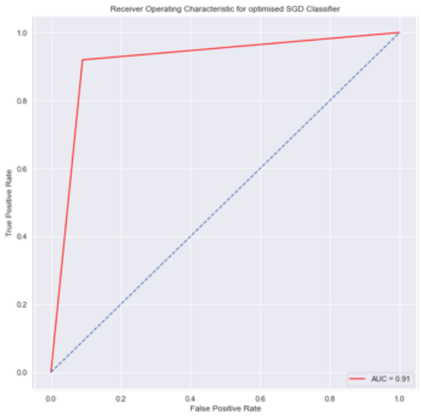


We can observe that when threshold value increases time taken for the convergence increase but there is not much difference in the accuracy of train and test data sets.
We can conclude that at tol = 1e-6 model has converged.

Conclusion :

1. From the above experiment with SGD classifier we can conclude that at alpha = 0.001 and tol = 1e-6 model has better accuracy for train and test data sets .
2. Confusion Matrix, Accuracies & AUC for optimized SGD Classifier :-





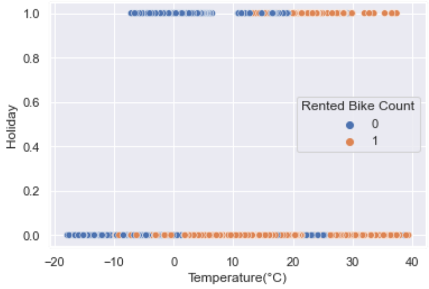
After optimizing the hyperparameters we can conclude that at alpha = 0.001 and tol = 1e-6 model has a better convergence and below are the accuracies for the same.

Model/Accuracy	Train	Test
Optimised SGD Classifier	91.7	91.4

We can conclude that there are 120 Type 1 Errors and 105 Type II Errors, since this is a prediction on rental bike count so, technically we can consider Type 1 Errors should be minimal.
When we see the ROC and AUC curve, we can observe that optimized model using SGD classifier has AUC of 0.91.

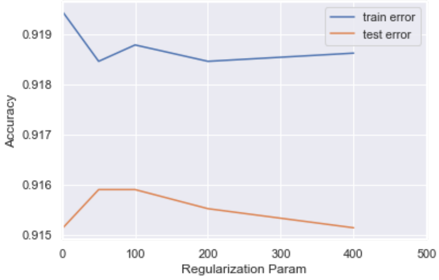
Part 2:

We will use SVM SVC classifier from scikit-learn package for the experimentation .

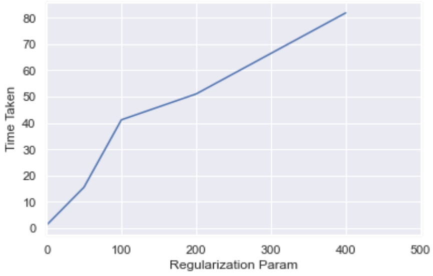


From the scatter plot we can conclude that there is no linear separable between the classes and need to go for kernel method to convert to a plan and make a separation of classes, but still, we will keep linear kernel and experiment with regularization parameter/penalty.

Keeping kernel as linear and experimenting with different values of regularization parameter/penalty :

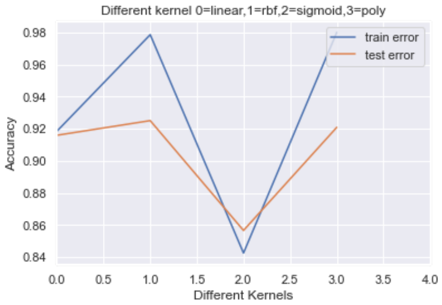


From the experimentation we can conclude that at regularization parameter/penalty at **C= 50** model is converging.

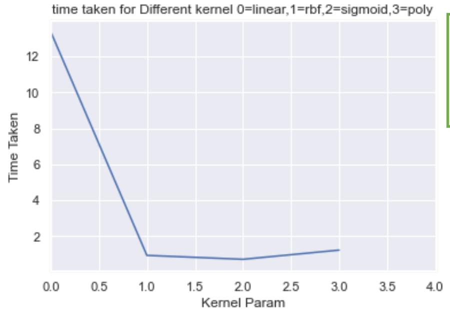


From the experimentation we can conclude that increase in regularization parameter time taken for convergence also increases.

Keeping C=50.0 and experimenting with different values of kernel methods :



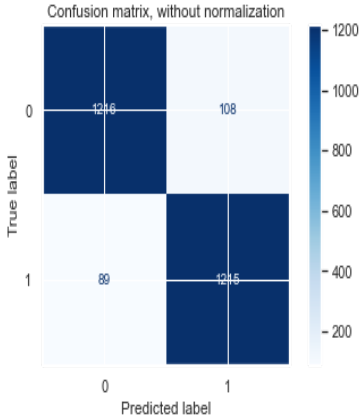
Keeping C=50.0 and kernel = linear model is good w.r.t train and test.
Keeping C=50.0 and kernel = rbf model is better compared to linear method.
Keeping C=50.0 and kernel = sigmoid model converges but accuracy goes down.
Keeping C=50.0 and kernel = poly model converges better than linear but not better than rbf.
So, we can conclude that **kernel = rbf** is best model.

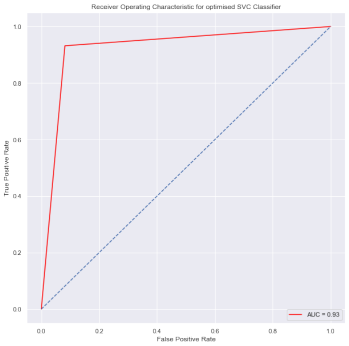


We can conclude that at kernel method = rbd model has converged faster.

Conclusion :

1. From the above experiment we can conclude that kernel method = rbf and at C=50.0 model has converged for train and test data sets .
2. Confusion Matrix, Accuracies & AUC for optimized SVM Classifier :-





After optimizing the hyperparameters we can conclude that at kernel method = rbf and C=50.0 model has a better convergence and below are the accuracies for the same.

Model/Accuracy	Train	Test
Optimised SVM Classifier	97.8	92.5

We can conclude that there are 108 Type 1 Errors and 89 Type II Errors, since this is a prediction on rental bike count so, technically we can consider Type 1 Errors should be minimal.

When we see the ROC and AUC curve, we can observe that optimized model using SGD classifier has AUC of 0.93.

Part 3:

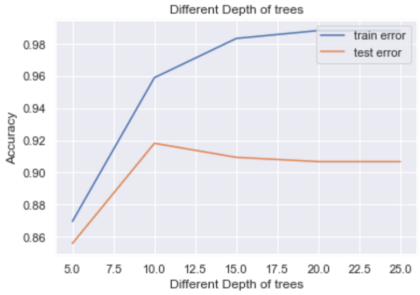
We will use Decision Tree classifier from scikit-learn package for the experimentation .

We will perform experimentation for pruning the tree in 2 phases pre pruning and post pruning and identify the right set of hyperparameters .

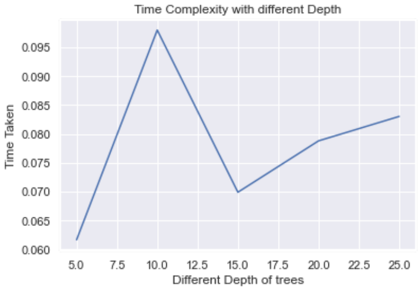
We will use Gini criterion for split information of nodes as we are only dealing with prediction of classes and not dealing with probabilities of classes .

Pre-Pruning :

1. Keeping **min_samples_split** at 5 and **min_samples_leaf** at 1 and varying **max_depth:**

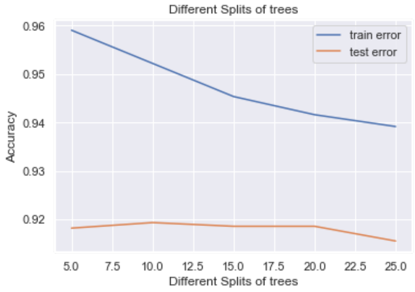


We can conclude that when depth of the tree increases model accuracy for train data tends towards overfit, and test data set accuracy decreases which is creating bias situation.
So max_depth at 10 is ideal for both train and test data sets.

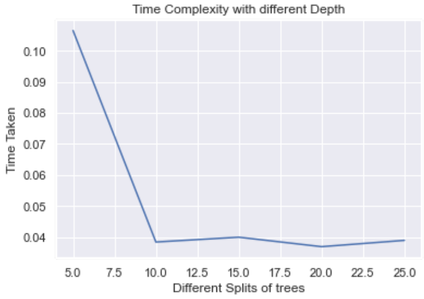


We can observe that at different depths time taken model for convergence is different and there Is no pattern observed, however at depth = 10.0 model took more time for convergence.

2. Keeping **max_depth** at 10 and **min_samples_leaf** at 1 and varying **min_samples_split :**

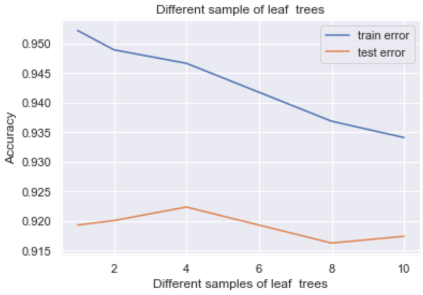


We can conclude that at max_depth =10 varying min_samples_split has very minimal effect on accuracies, however we can conclude **min_samples_split = 10** is best.

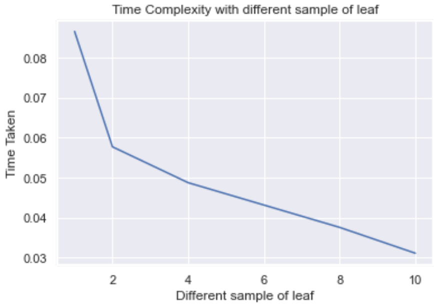


Varying min_samples_split has no effect on model time taken for convergence .

3. Keeping **max_depth** at 10 and **min_samples_split** at 10 and varying **min_samples_leaf :**



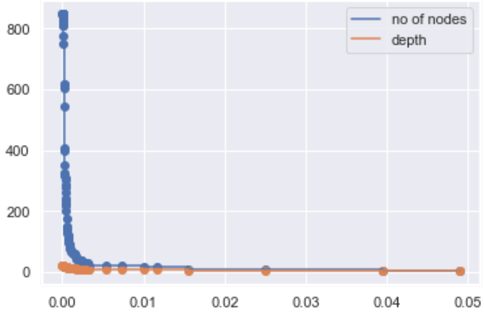
At min_samples_leaf = 4 both train and test accuracies are better however there is no much difference at max_depth = 10.
We can conclude that no much impact.



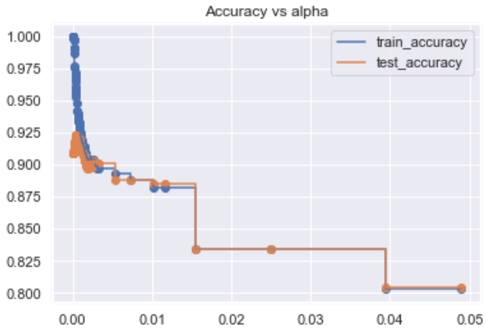
Varying min_samples_leaf has not much effect on model time taken for convergence. (negligible)

Post-Pruning :

In this experiment we will check cost complexity parameter varying w.r.t depth,nr of nodes, and accuracies .

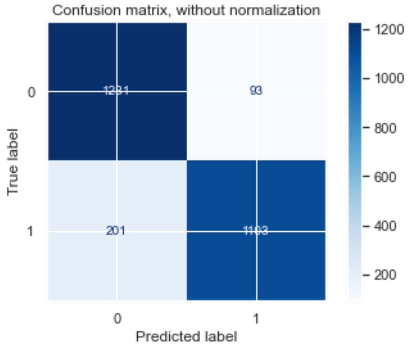


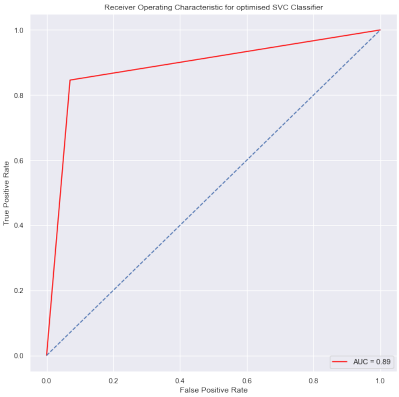
We can conclude that at cp=0.01 we have better accuracies w.r.t train and test data sets.



Conclusion :

1. After the experimentation we can conclude that at parameters max_depth=10,min_samples_split=10,min_samples_leaf=4,cp=0.01 model accuracies for train and test are better .
2. Confusion Matrix, Accuracies & AUC for optimized Decision Tree Classifier :-





After optimizing the hyperparameters we can conclude that at max_depth=10, min_samples_split=10, min_samples_leaf=4, cp=0.01 model has a better convergence and below are the accuracies for the same.

Model/Accuracy	Train	Test
Optimised DT Classifier	88.8	88.8

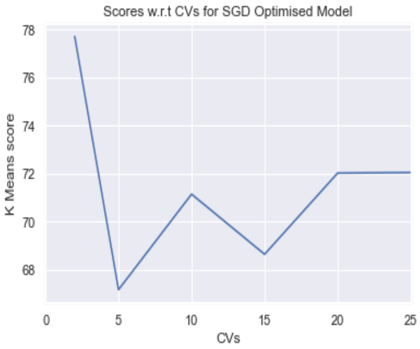
We can conclude that there are 93 Type 1 Errors and 201 Type II Errors, since this is a prediction on rental bike count so, technically we can consider Type 1 Errors should be minimal.
When we see the ROC and AUC curve, we can observe that optimized model using SGD classifier has AUC of 0.89.

Part 4:

Now we will use K-Fold Cross Validation Techniques to check the model performance for the optimized SGD,SVM and Decision Tree Classifiers .

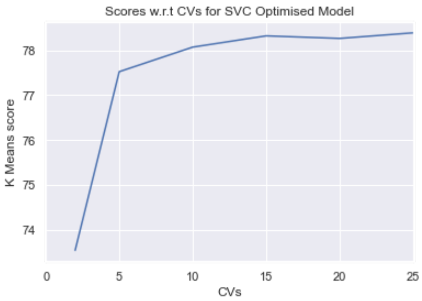
SGD :

We will conduct experiment with different K-Folds : 2,5,10,15,20,25



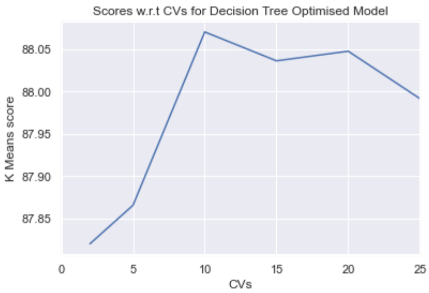
With K Fold 2 we found cross validation score of ~77
With K Fold 5 we found cross validation score of ~67
With K Fold 10 we found cross validation score of ~71
With K Fold 15 we found cross validation score of ~69
With K Fold 20 we found cross validation score of ~72
With K Fold 25 we found cross validation score of ~72
We can observe that after k Fold 20 no change in score
On Avg score is varying between 67~77 and **K-Fold of 2 or 10** can be baselined.

SVM :



With K Fold 2 we found cross validation score of ~73
With K Fold 5 we found cross validation score of ~78
With K Fold 10 we found cross validation score of ~78
With K Fold 15 we found cross validation score of ~79
With K Fold 20 we found cross validation score of ~79
With K Fold 25 we found cross validation score of ~79
We can observe that after k Fold 10 no change in score
On Avg score is varying between 73~79 and **K-Fold of 10** can be baselined.

Decision Tree :



With K Fold 2 we found cross validation score of ~87
With K Fold 5 we found cross validation score of ~87
With K Fold 10 we found cross validation score of ~89
With K Fold 15 we found cross validation score of ~88
With K Fold 20 we found cross validation score of ~88
With K Fold 25 we found cross validation score of ~88
We can observe that scores are consistent with the model.
K-Fold of **2 or 5** can be baselined .

Conclusion :

1. After experimenting we observed that SGD classifier cross validation score is ranging between 67~77 when unseen data is seen in the production .
2. After experimenting we observed that SVM classifier cross validation score is ranging between 73~79 when unseen data is seen in the production , though consistent at K-Fold > 5 .
3. After experimenting we observed that DT classifier cross validation score is ranging between 87~88 when unseen data is seen in the production , and is consistent at Nr of K-Folds

Results :

1. **We have divided the data set in to 2 classes based on the median as we have the data to be right skewed:**
   Median value is 504.5 if Rented Bike Count <=504.5 then considered as Class 0
   If Rented Bike Count > 504.5 then considered as Class 1 .
2. **After Thorough Experimentation with hyperparameter and with different K-Folds :**

Model/Accuracy	Train	Test	Type 1	Type II	AUC	CV Score
Optimised SGD Classifier	91.7	91.4	120	105	0.91	67~79
Optimised SVM Classifier	97.8	92.5	108	89	0.93	73~79
Optimised DT Classifier	88.8	88.8	93	201	0.89	87~88

i) Overall SVM model has better accuracies on train and test when compared with SGD and DT .

ii) Since this is a prediction of classes and business focus is to predict rented bike count we take Type 1 Errors to be less and when compared DT has the lowest Type 1 Errors , Whereas SVM classifier has the lowest Type II Errors .

iii) After Cross Validation we can observe that the variation in SGD and SVM models are more when compared to DT and DT is almost consistent with unseen data or with production data .

3. Overall DT is the best model to choose for deployment even though it slightly performed lower when compared with SVM and SGD but it is great with unseen data .