Week 5 (18.11 - 21.11.2019)

**Part 1**

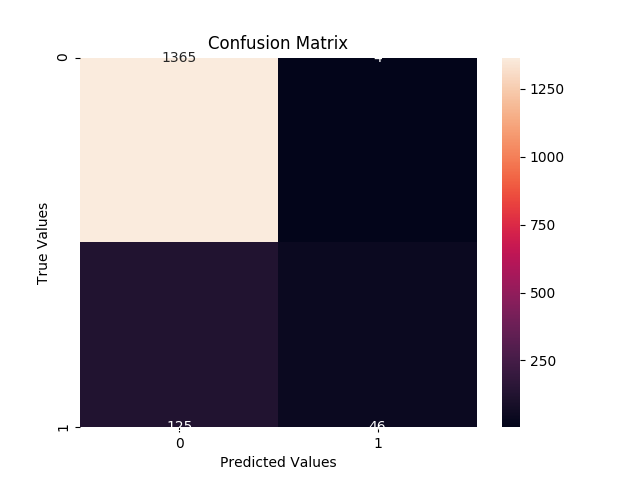
**Comparison before and after feature selection based on various metrics applied to different classifiers:**

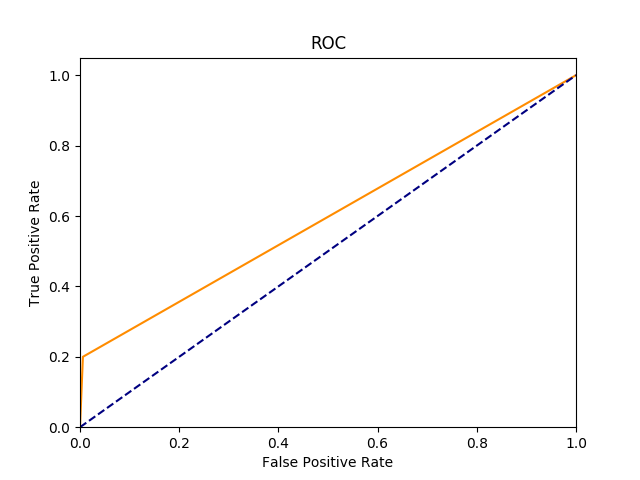
**Before Feature Selection:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Random Forest** | **Gradient Boosting** | **SVM Classifier** |
| **Accuracy** | 0.89 | 0.89 | 0.89 |
| **Precision** | 0.73 | 0.75 | 0.67 |
| **Recall** | 0.32 | 0.35 | 0.23 |
| **F1 Score** | 0.45 | 0.47 | 0.35 |
| **ROC-AUC Score** | 0.65 | 0.66 | 0.61 |

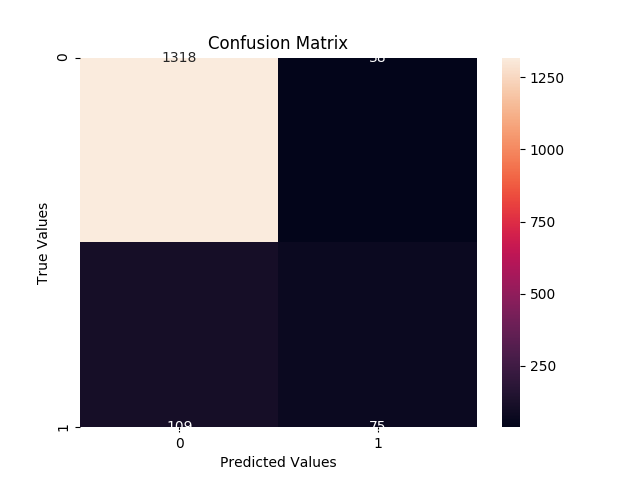
Confusion Matrix and ROC (The labels on the confusion matrix have been cropped; I couldn’t find a solution to it. But I have printed it out in the code)

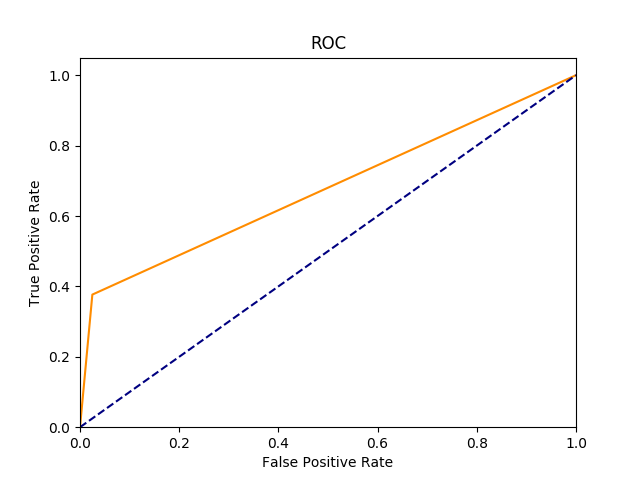
**Random Forest:**



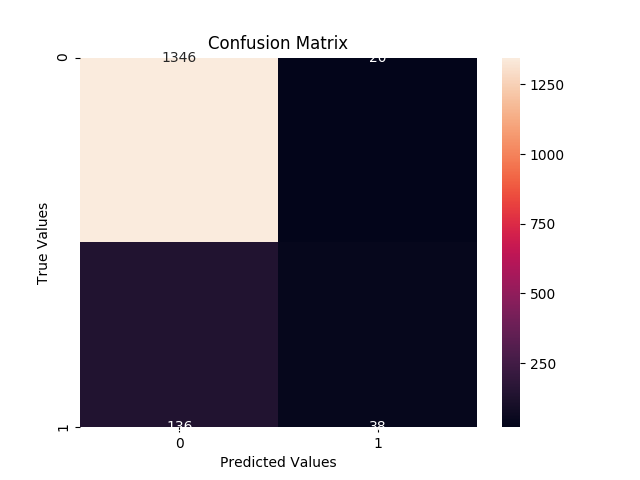


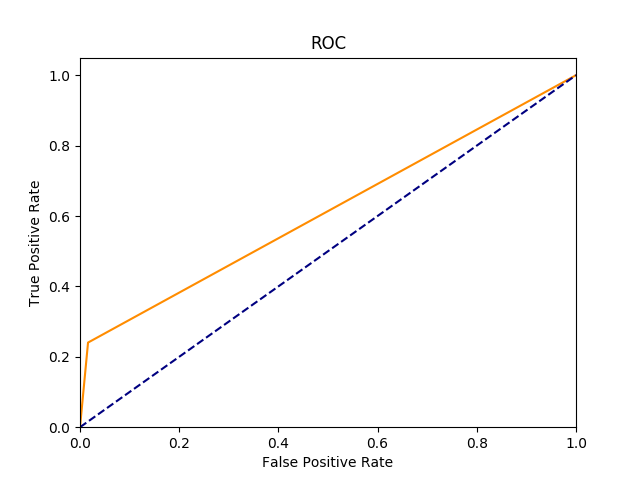
**Gradient Boost:**





**SVM Classifier:**



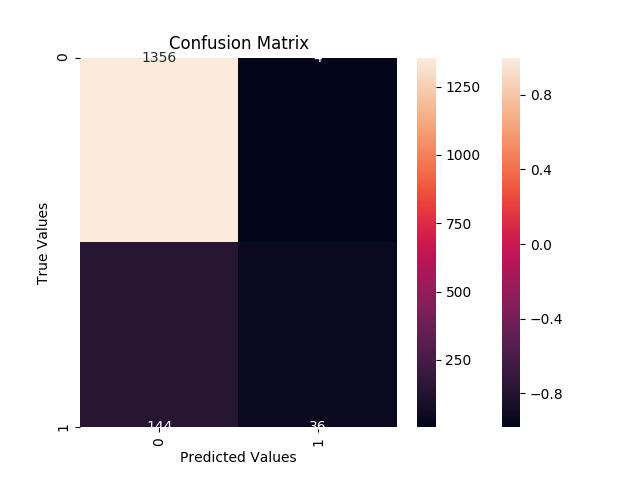


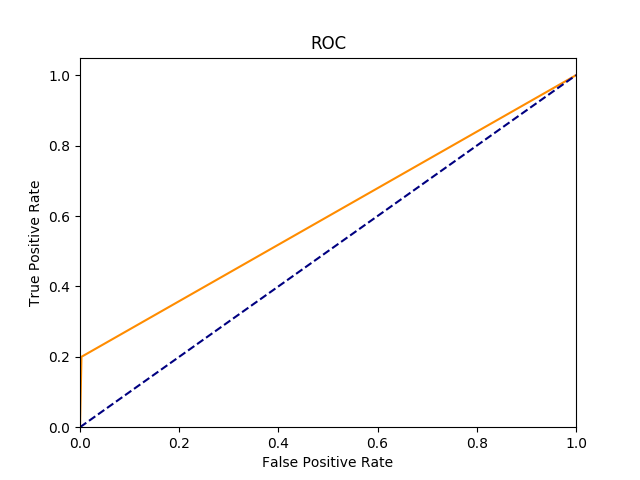
**After Feature Selection:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Random Forest** | **Gradient Boosting** | **SVM Classifier** |
| **Accuracy** | 0.90 | 0.89 | 0.90 |
| **Precision** | 0.90 | 0.71 | 0.77 |
| **Recall** | 0.24 | 0.38 | 0.27 |
| **F1 Score** | 0.38 | 0.50 | 0.40 |
| **ROC-AUC Score** | 0.61 | 0.68 | 0.63 |

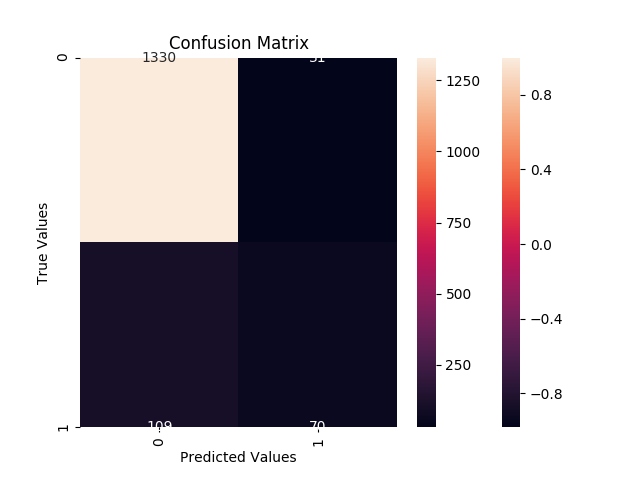
Confusion Matrix and ROC

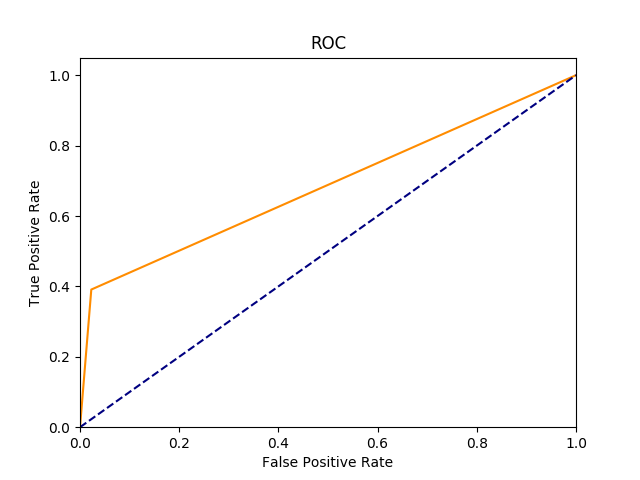
**Random Forest:**



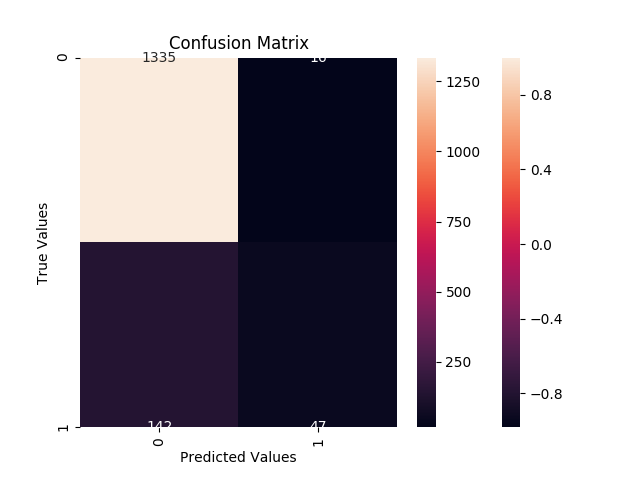


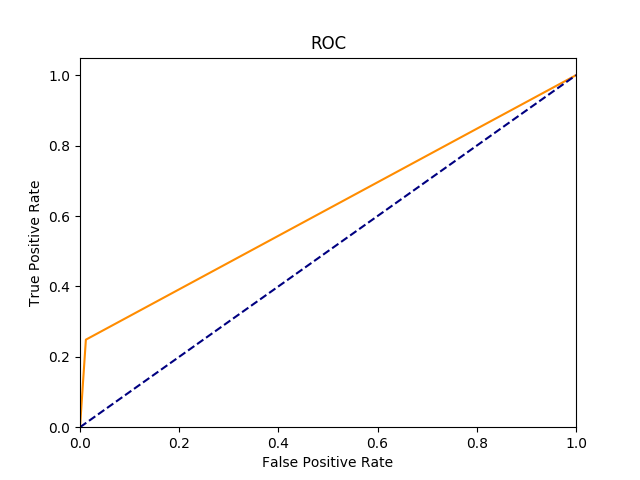
**Gradient Boost:**





**SVM Classifier:**





**Other changes made:**

1. Earlier, one feature out of a pair of highly correlated features was randomly discarded. Now, the feature with the higher average correlation with all the remaining features is discarded.
2. Imputation is done after removing features based on statistical tests like NaN value ratio, single unique values etc in the columns.