**Before Feature Engineering**

KNN:

Test set accuracy: 0.94

Decision Tree:

Accuracy on training set: 0.918

Accuracy on test set: 0.910

**Random Forest:**

**Accuracy on training set: 1.000**

**Accuracy on test set: 0.974**

Naive Bayes:

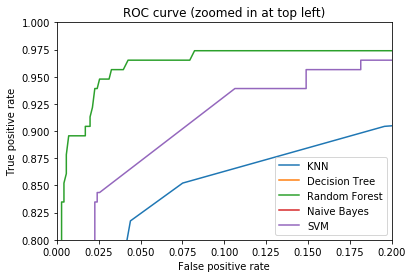
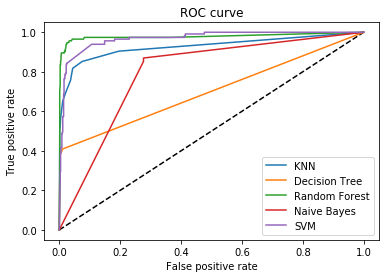
Accuracy on training set: 0.805

Accuracy on test set: 0.741

SVM:

Accuracy on training set: 1.000

Accuracy on test set: 0.890



**After Feature Engineering**

**Univariate statistics:**

**Percentile = 1 (i.e 15 features selected out of 1559)**

KNN:

Test set accuracy: 0.94

**Decision Tree:**

**Accuracy on training set: 0.974**

**Accuracy on test set: 0.961**

**Random Forest:**

**Accuracy on training set: 0.976**

**Accuracy on test set: 0.959**

Naive Bayes:

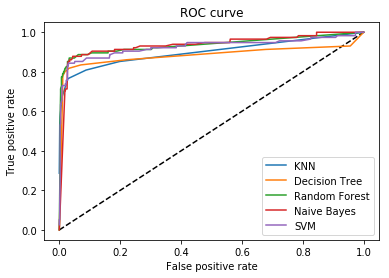
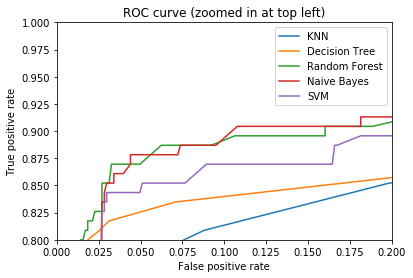
Accuracy on training set: 0.942

Accuracy on test set: 0.941

SVM:

Accuracy on training set: 0.975

Accuracy on test set: 0.948



**Model Based Selection:**

Random Forest Classifier used with n\_estimators=10, random\_state=42,max\_features=15

KNN:

Test set accuracy: 0.94

Decision Tree:

Accuracy on training set: 0.983

Accuracy on test set: 0.966

**Random Forest:**

**Accuracy on training set: 1.000**

**Accuracy on test set: 0.980**

Naive Bayes:

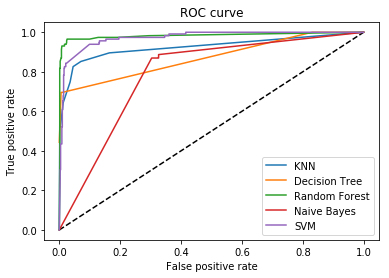
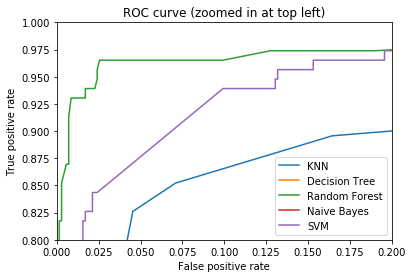
Accuracy on training set: 0.774

Accuracy on test set: 0.721

SVM:

Accuracy on training set: 0.999

Accuracy on test set: 0.894



**Recursive Feature Elimination (Iterative Feature Selection):**

Random Forest Classifier used with **n\_estimators=10, random\_state=42, max\_features=15**

RFE used with **n\_features\_to\_select=15, step=50**

KNN:

Test set accuracy: 0.94

Decision Tree:

Accuracy on training set: 0.980

Accuracy on test set: 0.970

**Random Forest:**

**Accuracy on training set: 0.993**

**Accuracy on test set: 0.972**

Naive Bayes:

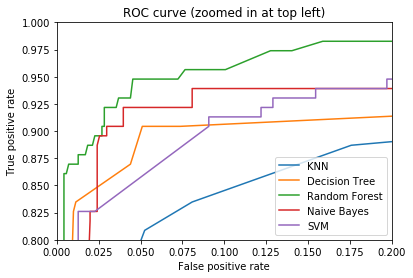
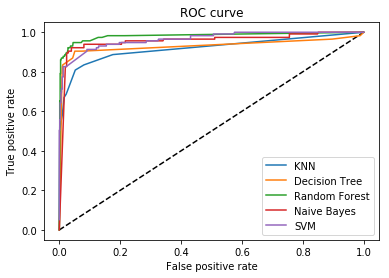
Accuracy on training set: 0.958

Accuracy on test set: 0.957

SVM:

Accuracy on training set: 0.992

Accuracy on test set: 0.952



**Conclusion:**

By Observing the training and testing accuracies and roc curves during all the above-mentioned procedures the best one is **Random Forest after Feature Selection via Recursive Feature Elimination with 15 features.**

Training Accuracy of 99.3%

Testing accuracy of 97.2%

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Without Feature Selection | | Select Percentile | | Select From Model | | RFE | |
| Dimensions | (3279, 1559)  1559 - Features | | (3279, 15)  15 - Features | | (3279, 779)  779 - Features | | (3279, 15)  **15 - Features** | |
|  | **Percentage Accuracy** | | **Percentage Accuracy** | | **Percentage Accuracy** | | **Percentage Accuracy** | |
|  | Train | Test | Train | Test | Train | Test | Train | Test |
| KNN | - | 94 | - | 93 | - | 94 | - | 94 |
| Decision Tree | 91.8 | 91 | 98.1 | 96.7 | 98.3 | 97 | 98 | 97.1 |
| Random Forest | 100 | 97.4 | 98.6 | 96.6 | 100 | 97.4 | **99.3** | **97.3** |
| Naïve Bayes | 80.5 | 74.1 | 96.3 | 95.6 | 96.7 | 94.6 | 95.7 | 95.6 |
| SVM | 100 | 89 | 98.5 | 94 | 99.9 | 89.1 | 99.2 | 95.2 |

**BEST MODEL: RANDOM FOREST**

|  |  |
| --- | --- |
| **BEFORE FEATURE SELECTION**  **1559 features** | **AFTER FEATURE SELECTION** |
| Training Accuracy: **1.0**  Testing Accuracy: **0.974** | **Univariate: (15 features)**  Training Accuracy: **0.976**  Testing Accuracy: **0.959**  **Select From Model: (779 features)**  Training Accuracy: **1.0**  Testing Accuracy: **0.98**  **RFE: (15 features)**  Training Accuracy: **0.993**  Testing Accuracy: **0.972** |