**Steps:**

1. Download data from freddie mac website automatically based on quarters entered.

2. Preprocess data : Read data from the text file as pandas dataframe. Handle missing values based on variable document provided by freddie Mac.

Example if the Net\_Sales\_prceeds is C and U then Actual\_loss\_calculation is set as 0.

**Assumptions and Handling for missing data:**

'ZERO\_BALANCE\_CODE'-If missing replaced with 0. As we cannot assign default ratio if the value is missing.

'CURRENT\_INTEREST\_RATE',-Replaced by ffill method

CURRENT\_DEFEREED\_UPB-Replaced by Zero

#'MI\_RECOVERIES',

#'NET\_SALES\_PROCEEDS',

NON\_MI\_RECOVERIES

EXPENSES-Replaced with Zero

'ACTUAL\_LOSS\_CALCULATION',-Replaced by Zero if the Net\_Sales\_Proceeds is C and U.

|  |  |  |
| --- | --- | --- |
| Column | Reason | Value |
| ZERO\_BALANCE\_CODE | 3,6, and 9 imply default in loan .Hence missing value replaced by 0. Assumption loan still being payed | 0 |
| CURRENT\_INTEREST\_RATE | Used ffill value to replace it | propagate last valid observation forward |
| MI\_RECOVERIES | Assuming as Zero for missing as per documentation | 0 |
| ACTUAL\_LOSS\_CALCULATION | If net\_sales\_proceeds is C and U then it is set as 0 as ffill to set the value | 0 and propagate last valid observation forward |
| EXPENSES | Assuming as Zero for missing as per documentation | 0 |

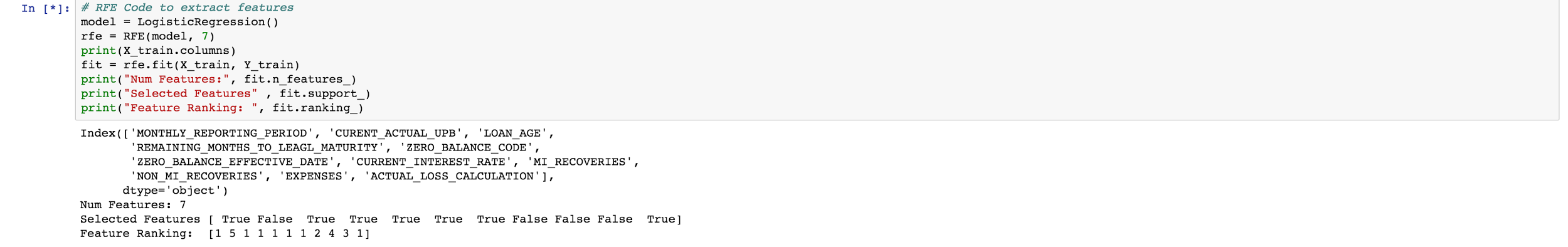
Dropped Net Sales proceeds as it was used in calculating actual\_loss\_calculation

Dropped Misc expenses, Modification cost,legal costs, Maintenance and preservation costs and taxes and insurance as they contribute to other columns

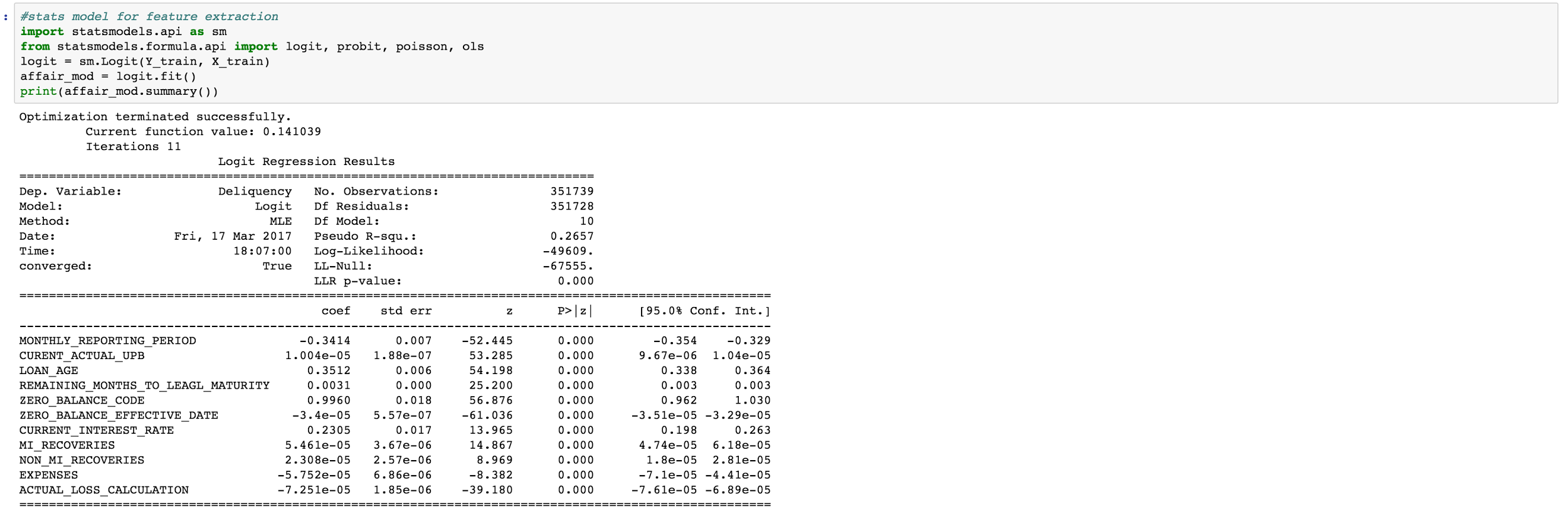
**Feature Selection**

# **Recursive feature elimination**

Recursive feature elimination is based on the idea to repeatedly construct a model (for example an SVM or a regression model) and choose either the best or worst performing feature (for example based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. Features are then ranked according to when they were eliminated. As such, it is a greedy optimization for finding the best performing subset of features.

The stability of RFE depends heavily on the type of model that is used for feature ranking at each iteration. Over analysis in various iterations I came up with 7 features contributing to delinquency.

Also used statsmodel logit to get p-value for features.The resultant feature and output was not optimal.



Features selected

'MONTHLY\_REPORTING\_PERIOD',

'LOAN\_AGE',

'REMAINING\_MONTHS\_TO\_LEAGL\_MATURITY',

'ZERO\_BALANCE\_CODE',

'ZERO\_BALANCE\_EFFECTIVE\_DATE',

'CURRENT\_INTEREST\_RATE',

EXPENSES

'ACTUAL\_LOSS\_CALCULATION',

**Algorithm used**

Logistical regression

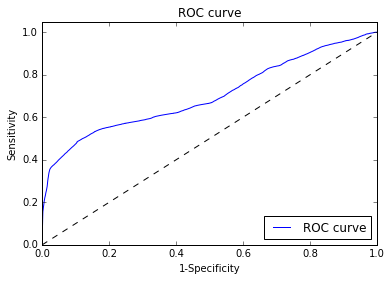
Random Forest

Neural Network

SVN

**Algorithm performance**

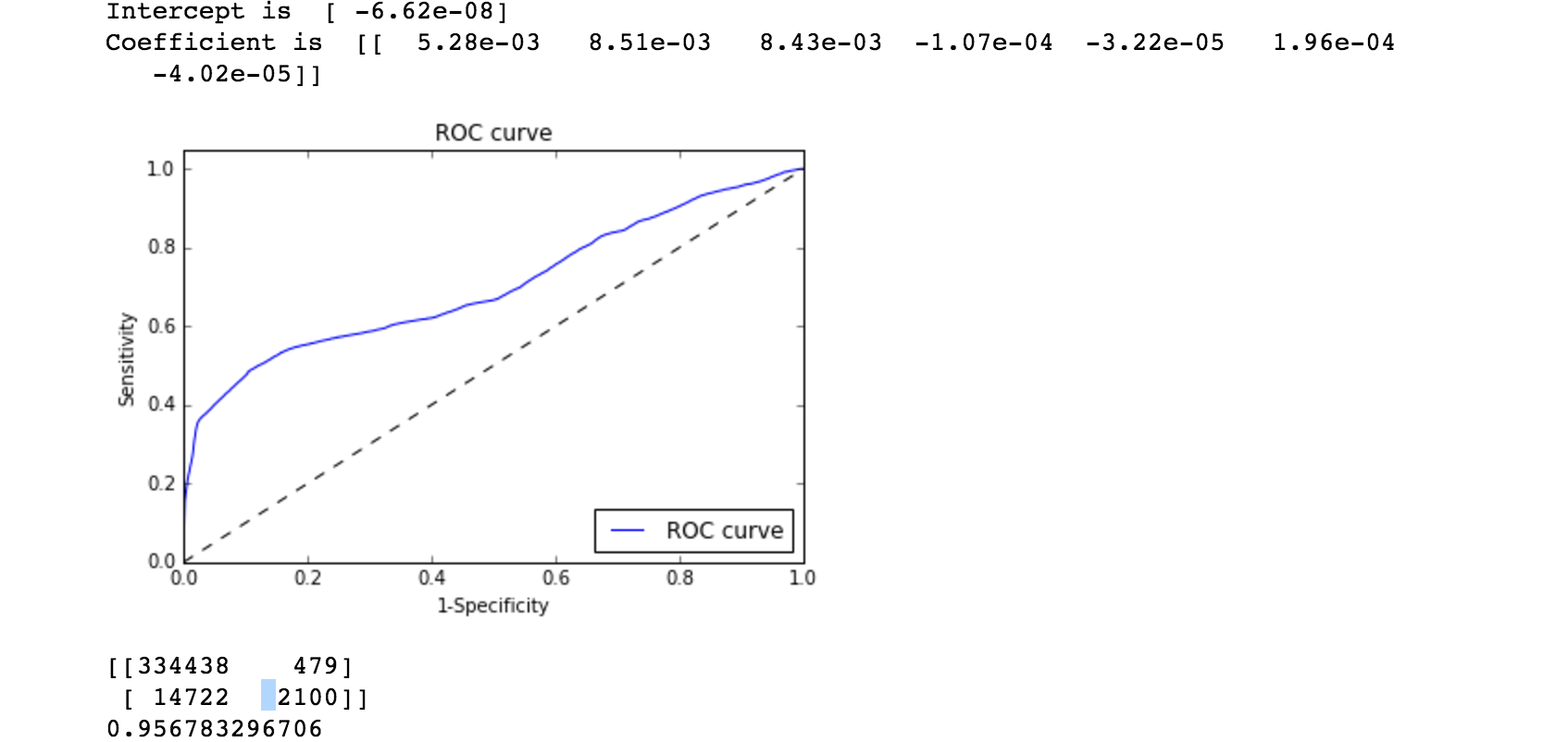
**Logistic Regression**



Model Accuracy -0.956783296706

Confusion Matrix

[[334438 479]  
 [ 14722 2100]]

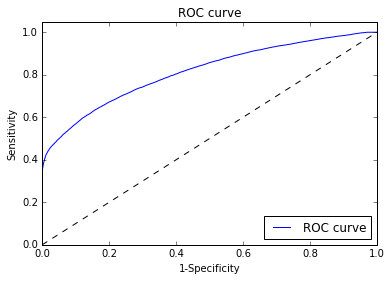


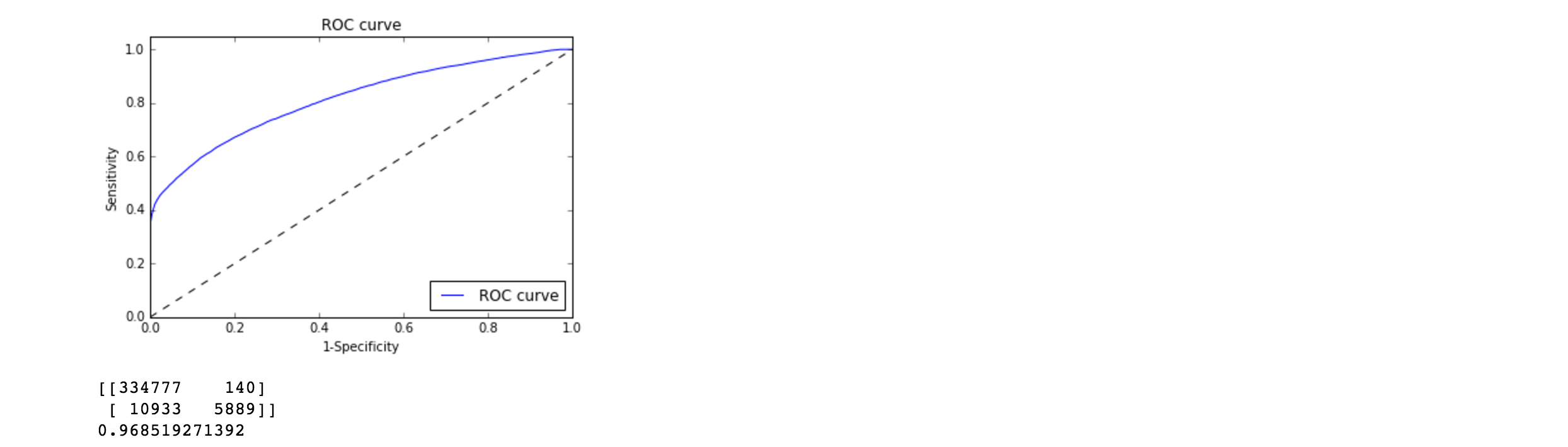
Neural Network

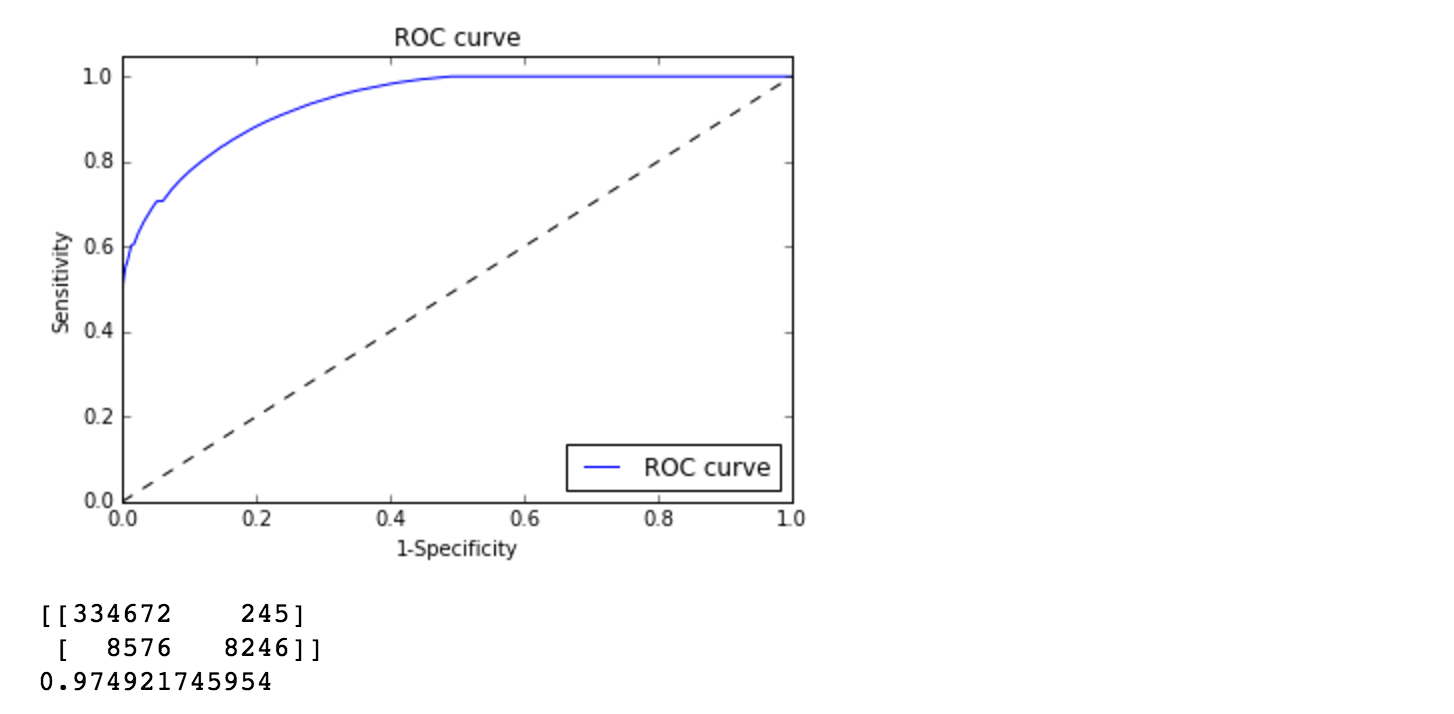
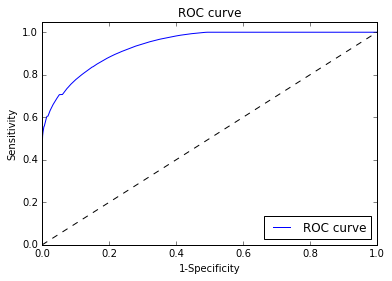
Accuracy-0.968519271392

Confusion Matrix

[[334777 140]  
 [ 10933 5889]]





Random Forest

Accuracy 0.974921745954

Confusion Matrix

[[334672 245]  
 [ 8576 8246]]

SVN

I tried it on the several times on same dataset but it was not completed.

Final model Selected based on analysis on dataset is Random Forest