Predict Customer for Loan Default

Using Python-Scikit

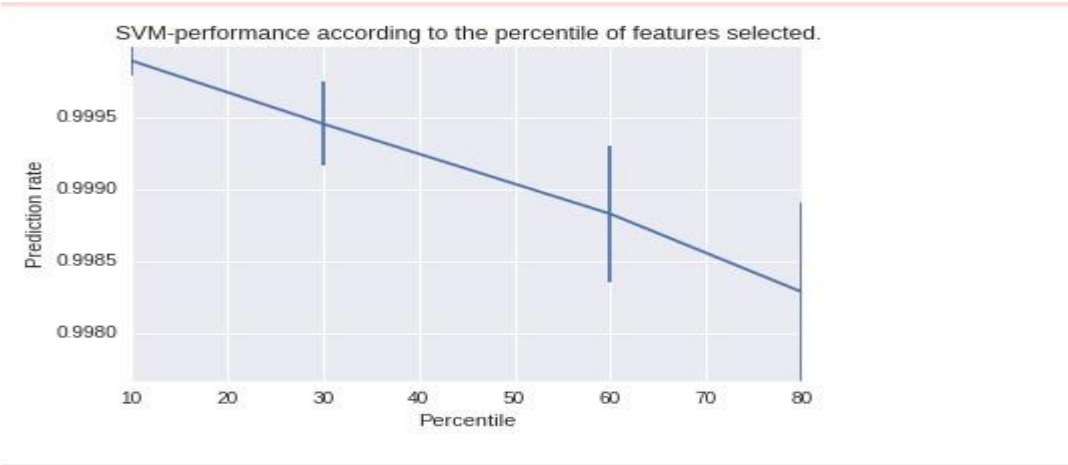
Steps that took in order to complete the challenge

Steps

In this code I have fit a SVM model on the Loans Dataset. Below are the steps that I have used in the process.

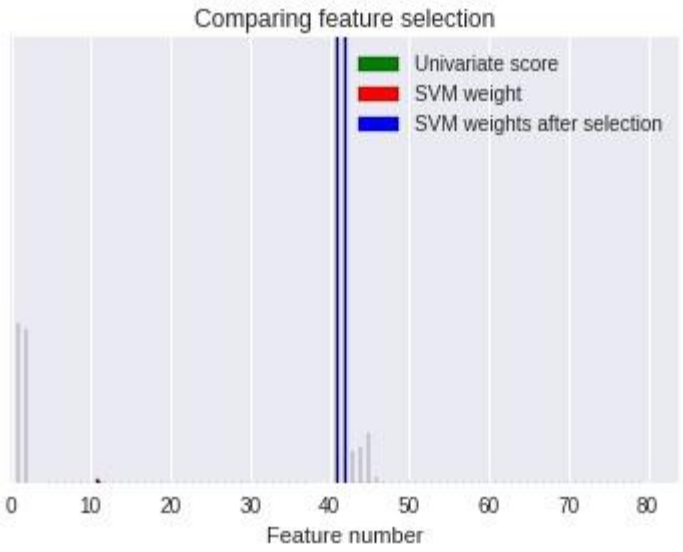
- Imported the dataset and filtered only the 36 months records for further analysis.
- On initial review, removed the unnecessary columns such as dates, descriptions, urls etc. (Please see the comments against the code)
- Used an encoding technique to encode all the factors to int's (grade , emp_length etc).
- Printed the initial ratio for the (Target = 1 / Total observations) to see if I need to use a stratified sampling approach
- Substituted all the Nan's in the data frame to -1. We Could use better techniques such as mean/median/Regression.
- Normalized the columns since all the predictor variables had different range of values.
- Now, we are ready with the data i.e our X matrix and the Y matrix which contains the actual values.
- Just to see what percentile of the features we need to included I plotted a graph for the Percentile Features Vs the Prediction Rate.
- From Point 8. I got the Percentile values as 10%. Thus I proceeded with this percent of features.
- Used the Feature selection technique with 10% features from the scrubbed X matrix.
- Fit a SVM model to see which are the features that we need to use for prediction.
- . Used a K-Cross fold validation with a stratified sampling approach since the target variables had very few proportions of '1' than '0' .
- One more advantage that stratified sampling had was that it reduced the bias when a set of data was selected for cross validation.
- Plotted a Confusion matrix after fitting the model for review.

Below are the plots that I have included in the code.



Percentile Features used Vs the Prediction Rate

2. Comparing the Features by plotting a graph between the Features used and the SVM weights.



3. Confusion Matrix for the SVM Model

Confusion Matrix