**CREDIT LOAN APPLICANTS – MODEL REPORT**

**PROBLEM IDENTIFICATION OVERVIEW**

* **DEFINE THE QUESTION SPECIFIC TO MODELING ACTIVITIES—DEVICE FAILURES PREDICTION**

The SVM or Support Vector Machine is well suited for data input where the data frame consists of at least several hundred observations if not more. It also works well with a set of minimal features anywhere between a few to several. Dependent on the relationship between the features and values within the features, a kernel will need to be selected that is tailored to address the complexities of that data set such as using the sigmoid kernel for binary classifications. In addition, null values also need to be cleaned before feeding the data set into the SVM or it will fail to work.

* **IDENTIFY THE DATA NEEDED AND OR AVAILABLE**

Several features are employed but two types of features are needed to address the binary classification problem. One feature which contains the binary labels and the remainder of the features consisting of features containing the values to base the SVM’s classification according to the values to predict the appropriate label. For the data sets named checking status, credit history, employment data, and merged credit data, the feature containing the label feature is class, and the attribute features are checking status, credit history, and employment data and their respective variations in respect to their original features which have been one-hot encoded for each value to be represented with its own feature accounting for a total of 12 attribute features.

* **DEFINE THE DATA TIME FRAME**

2014

* **DESCRIBE THE MODELING RESPONSE**

Binary. Binary Classification: Good; Bad.

* **UNSUPERVISED OR SUPERVISED CLASSIFICATION OR REGRESSION MODEL**

Supervised Classification.

* **WHAT DELIVERABLES WILL BE GENERATED**

A total of four data sets were created with the objective of testing how each of the original attribute features would influence the SVM in its binary classification and then a final combined data set of all 3 one-hot encoded data sets to verify how each of the attributes in question would wield influence on the prediction of the SVM. For each of the data set runs, a confusion matrix and a classification report are provided. The confusion matrix details the number of predicted classifications for each label while the classification report provides a reading of each of the metrics used and their accuracies in determination of its classification and the SVM’s efficacy in doing so given the data set it is working with.

* **DATA PREPROCESSING STEPS OF NOTE**

1. Created three new data sets each composed of a label feature and one original attribute feature.
2. Merged those three data sets into one composite data set containing all prior features.
3. One-hot encoded all the attribute features for each data set to convert each value of any column into its own feature.

* **MODEL DESCRIPTION**

1. One thousand rows. Fifteen columns. The features are as follows: Class; Checking\_Status\_0<=X<200; Checking\_Status<0; Checking\_Status>=200; Checking\_Status\_No Checking; Credit\_History\_All Paid; Credit\_History\_Critical/Other Existing Credit; Credit\_History\_Delayed Previously; Credit\_History\_Existing Paid; Credit\_History\_No Credits/All Paid; Employment\_1<=X<4; Employment\_4<=X<7; Employment\_<1; Employment\_>=7; Employment\_Unemployed.
2. Model Algorithm: SVC – Support Vector Classifier.

Parameters: Kernel – Kernel SVM type used for complex and non-linear problems; assigned ‘rbf’ which stands for Gaussian as the argument for the parameter.

1. Three additional and newly created data sets were also fed into the SVC to determine each attribute feature’s impact upon the classification of each observation.

* **MODEL PERFORMANCE**

|  |
| --- |
| precision recall f1-score support |
|  |
| bad 0.49 0.38 0.43 60 |
| good 0.76 0.83 0.79 140 |
|  |
| accuracy 0.69 200 |
| macro avg 0.62 0.61 0.61 200 |
| weighted avg 0.68 0.69 0.68 200 |

* **MODEL FINDINGS**

The label feature ‘class’ is critical in its binary classification of ‘good’ and ‘bad’ in predicting whether a credit applicant is too high a risk or not. The three main and original attribute features of checking status, credit history, and employment data were discovered to be amongst the most relevant of features and thus were sifted from the rest of the original data set to determine just how much of a bearing they hold in the binary classification of each observation. Including all the data sets that were ran, the accuracy fell anywhere from 64-71%.