2018

# Classification Problem Project

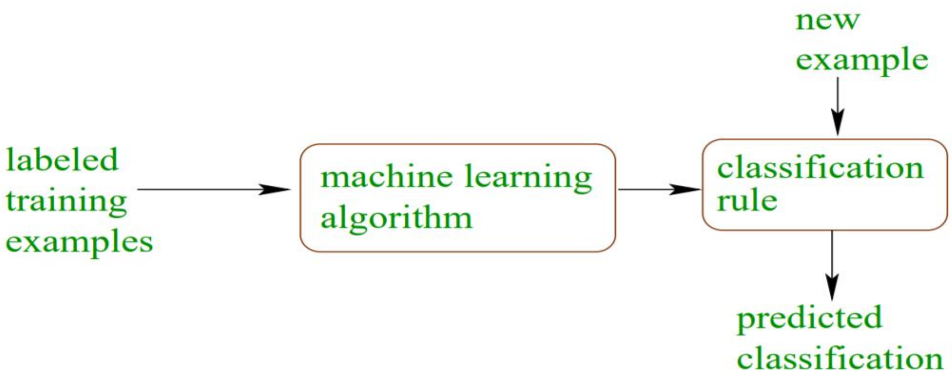
BY
HARICHARAN ADHYAM VITTAL
HARISHKARTHIK GUNALAN
KEDAR WAGHOLIKAR
POOJA SINHA
NAMRATHA KONAGAVALLI
SHUBHANGI JAIN
SREERAM DANTU

SUPPLY CHAIN ANALYTICS-BADM 590 | UIUC

## INTRODUCTION

**Question 1: Classification:**

Classification is about predicting the exact category among a set of categories for new observation, based on a trained data model that knows the data corresponding to each category. In statistical terms, the core goal of classification is to predict a response variable y from some set of predictor variables x. For instance, the classification approach in a supervised machine learning is judicious in the supply chain to categorize the defaulting suppliers from the non-defaulting one's.




**Question 2: Objective of the Project:**

Supply Chains are prone to get affected due to bankruptcy of Suppliers. It is important to assess the suppliers risk of bankruptcy and plan for alternatives. The objective of this project is to build a risk assessment model for potential bankruptcy of suppliers. The risk assessment in this project is done by building classification models based on four variable groups: Competitiveness of Suppliers, Supplier Technological Investments, Supplier Financial Health and Supplier Country Location. Based on the data available related to past instances of bankruptcy consisting of both bankrupt and solvent suppliers, we proceed with building our model to evaluate the probability of bankruptcy using Logistic Regression, Random Forest and SVM methods. We have also built the model finding how each group of variables affect the accuracy, sensitivity and Specificity of the model.

**Question 3: Why is classification the right approach?**

A Data Classification is an extremely important step in building a secure organization. Classifying data is the process of categorizing data based on nominal values according to its sensitivity. Classification can help an organization to meet legal and regulatory requirements for retrieving specific information in a set timeframe.

**Question 4: Describe briefly the steps followed**

We split the sample data (Supplier_Bankruptcy_Risk.csv) into train set and test set. We build a logistic regression model with the train set in competitiveness of suppliers group variables

(**Bankruptcy ~ NumberOfCustomers" +AverageShareofBusiness + AverageSizeofCustomers + AverageSizeofSuppliers**).

We predicted the probability of bankruptcy with binomial response for supplier's test set. Furthermore, we have added variable for the competitiveness of suppliers and supplier technological investment groups together.

For each model, based on the predicted probability, we were able to find AUC and plot AUCs for Random Forest and Support Vector Machine models.

Finally, we calculated decision threshold to determine a predict a supplier as solvent. The optimal decision threshold is obtained by minimizing the joint squared errors

$Error2=(1-sensitivity)2+(1-specificity)2$ to check how well the model can detect bankruptcy.

**Observation And Results**

As can be observed from the table under Exhibit 1 attached herein, the AUC of the models, and hence the performance of the models becomes better if we add more predictors to the model. This trend can be seen for the generalized linear model, Random Forest model as well as the Support Vector Machine (SVM) model. Exhibit 2, Exhibit 3, Exhibit 4 and Exhibit 5 attached herein show the Receiver Operating Curve

(ROC) plots for step 1, step 2, step 3 and step 4, respectively, where the different parameters associated with suppliers are added in each step.

Further, if we compare the performance of the different algorithms used by us, we observe that the Generalized linear model - Logistic Regression performed better than Random Forest and SVM and gave an AUC of 0.9729 for step 4. The optimal decision threshold point for the generalized linear model is 0.1354, corresponding to a sensitivity of 0.94 and specificity of 0.88. Likewise, for the Random Forest model, the optimal threshold point is 0.2339, corresponding to a sensitivity of 0.85 and specificity of 0.89. For the SVM model, the optimal threshold point is 0.09719, corresponding to a sensitivity of 0.91 and specificity of 0.86.

**Question 4: Discussion**

**How can this project help management of supply chain risk?**

This project can help in identifying potential suppliers who are at the verge of facing bankruptcy risk. Identification of suppliers at the initial stage can help them in taking required precaution at the right time. Identifying and finding alternative suppliers for the important components is also one of the good way to mitigate risk. Classification of suppliers based on different attributes can provide various advantages. By creating supplier segments, group of suppliers can be removed who are not meeting the required specification. It also very helpful if there is large set of supplier's data for different varieties of product. By eliminating the set of weak suppliers, it also contributes in selecting desired set of suppliers as per the requirement. Providing alternative suppliers contributes in unforeseen interruption of raw materials and prevent from making the process slow. All these mentioned advantages will help in management of supply chain by preventing risk for the company. This process can also help in developing plans and strategies to make successful supplier management programs for their respective organization. In these recent times including varieties of supplier's evaluation and selection criteria, clustering in an efficient tool to break

down complex and huge data set and extracting useful insights from it for the management of supply chain risk.

We can also use four important performance aspects including **quality, cost, delivery lead-time, and flexibility**. Firstly, in measuring quality performance, Defect Rates are used to monitor supplier delivery quality in manufacturing industry where as the equivalent measure would be Non-conformance Rate in service industry. Secondly for cost performance measure, suppliers undertaking higher effort levels for cost reduction are preferred more though their initial offered price are high. Thirdly, measuring lead times of deliveries from suppliers is an important performance measure where lengthy lead times or fluctuations in lead times are generally not preferred. Lastly, Flexibility of suppliers is helpful in selecting and retaining suppliers specially in industry having high variety of product and frequent change in product including volume flexibility and variety flexibility.

## Key learning from this project:

We learnt how in classification problems we are trying to predict a discrete number of values. The labels(y) generally comes in categorical form and represents a finite number of classes.

**Types of classification**

1. **Binary classification**—in which there is only two classes to predict, usually 1 or 0 values.
2. **Multi-Class Classification**—When there are more than two class labels to predict we call multi-classification task. E.g. While predicting 3 types of image species from image classification problems where there are more than thousands classes(cat, dog, fish, car,…).

We also learn how Classification differs from Regression:

- Classification is the task of predicting a discrete class label.
- Regression is the task of predicting a continuous quantity.

In regression problems we trying to predict continuous valued output, take this example. Given a size of the house predict the price (real value).Some of the Regression Algorithms are Linear Regression, Regression trees, SVM.

But there is some overlap between the algorithms for classification and regression; For example: A classification algorithm may predict a continuous value, but the continuous value is in the form of a probability for a class label. Whereas a regression algorithm may predict a discrete value, but the discrete value in the form of an integer quantity.

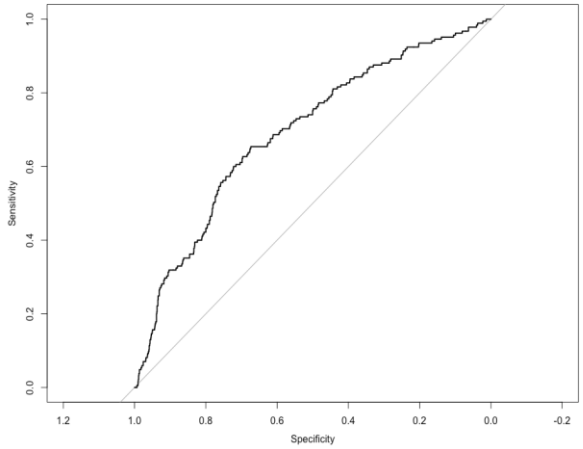
## APPENDIX

## EXHIBIT 1

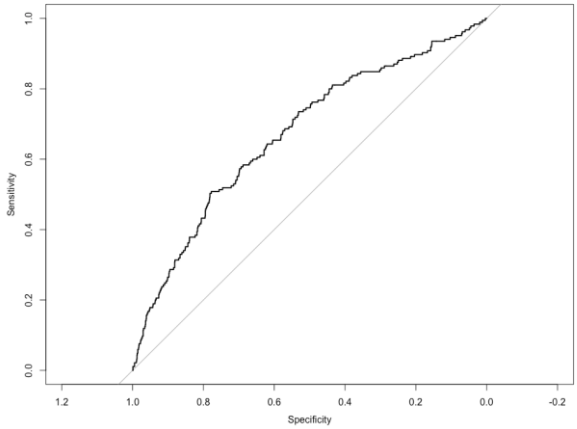
| **Steps** | **Logistic Regression** | **Random Forest** | **SVM** |
|---|---|---|---|
| **Step 1.** Competitiveness of Suppliers | 0.6918 | 0.6718 | 0.6209 |
| **Step 2.** Competitiveness of Suppliers + Supplier Technological Investments | 0.7241 | 0.6895 | 0.6234 |
| **Step 3.** Competitiveness of Suppliers + Supplier Technological Investments + Supplier Financial Health | 0.8449 | 0.8136 | 0.7239 |
| **Step 4.** Competitiveness of Suppliers + Supplier Technological Investments + Supplier Financial Health + Supplier Country Location | 0.9729 | 0.9484 | 0.9533 |

## EXHIBIT 2

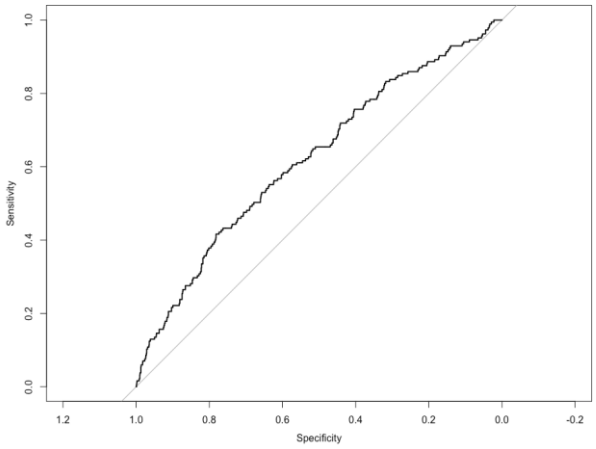
ROC plot - GLM for step 1



ROC plot - Random Forest for step 1

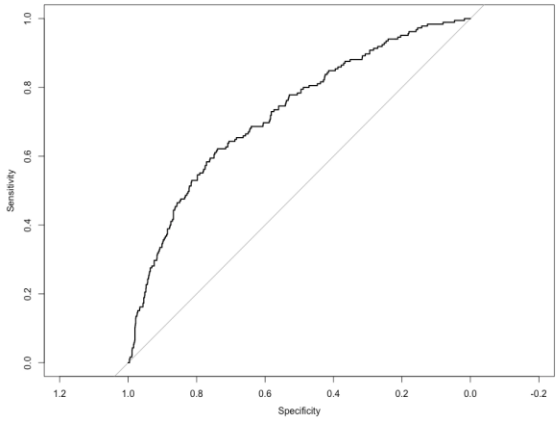



ROC plot - SVM for step 1

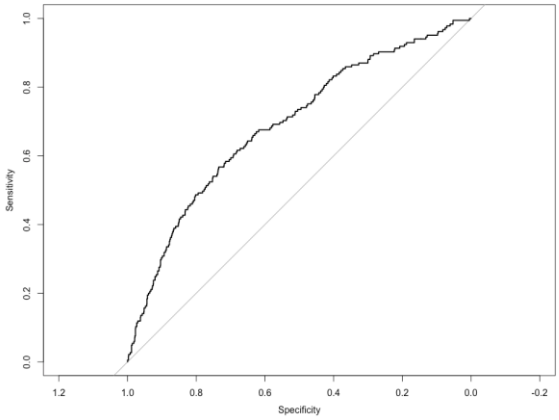


## EXHIBIT 3

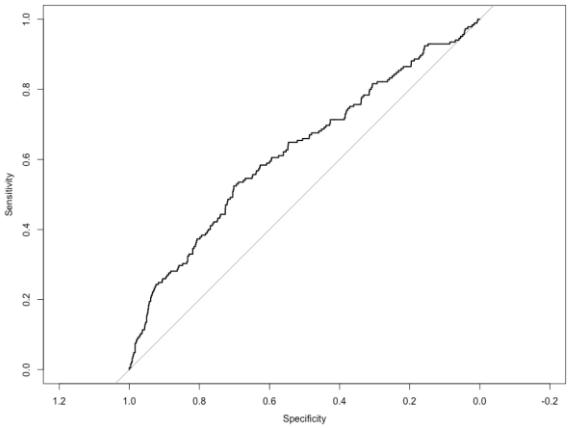
ROC plot - GLM for step 2




ROC plot - Random Forest for step 2

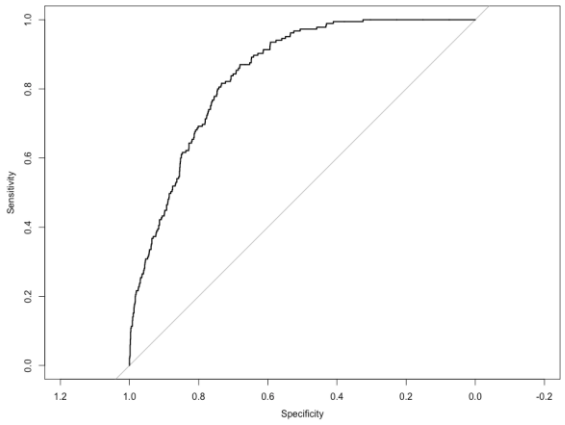



ROC plot - SVM for step 2

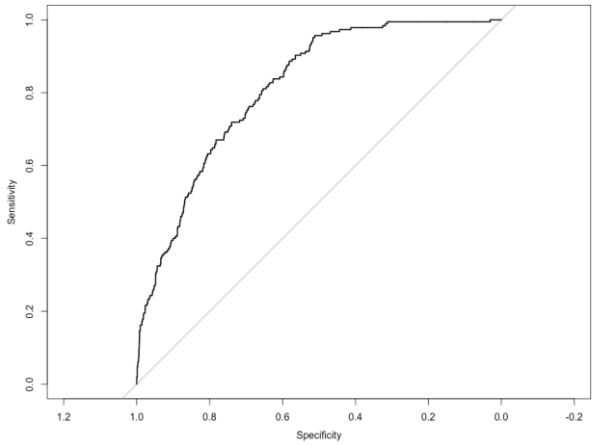



## EXHIBIT 4

ROC plot - GLM for step 3




ROC plot - Random Forest for step 3




ROC plot - SVM for step 3

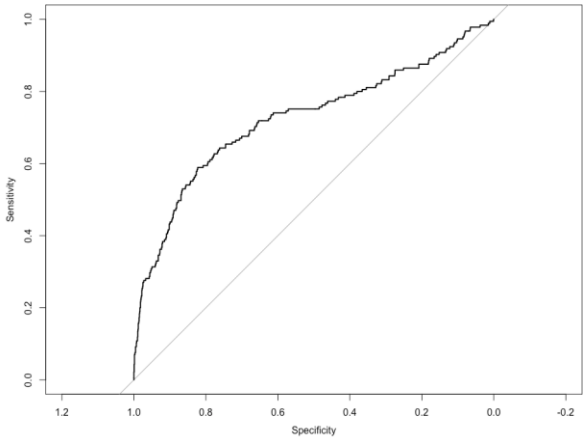
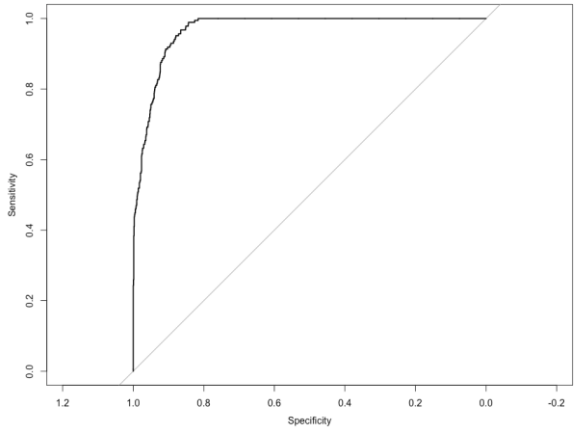
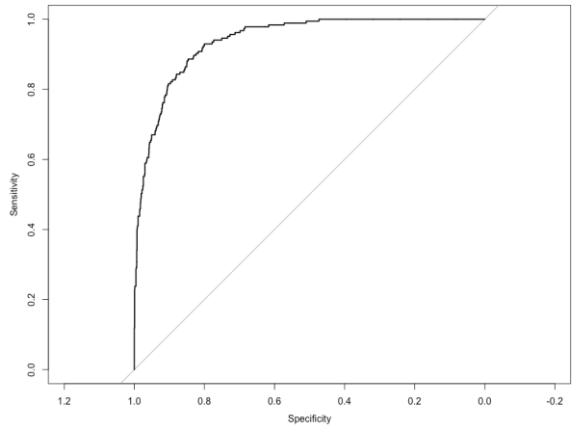



**EXHIBIT 5**

ROC plot - GLM for step 4




ROC plot - Random Forest for step 4




ROC plot - SVM for step 4