World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:16, No:8, 2022

Optimizing Data Evaluation Metrics for Fraud Detection Using Machine Learning

Jennifer Leach, Umashanger Thayasivam

Abstract—The use of technology has benefited society in more ways than one ever thought possible. Unfortunately, as society's knowledge of technology has advanced, so has its knowledge of ways to use technology to manipulate others. This has led to a simultaneous advancement in the world of fraud. Machine learning techniques can offer a possible solution to help decrease these advancements. This research explores how the use of various machine learning techniques can aid in detecting fraudulent activity across two different types of fraudulent datasets, and the accuracy, precision, recall, and F1 were recorded for each method. Each machine learning model was also tested across five different training and testing splits in order to discover which split and technique would lead to the most optimal results.

Keywords—Data science, fraud detection, machine learning, supervised learning,

I. Introduction

SINCE technology has become a fundamental part of how society runs and operates, it has become even easier for people to utilize these new advancements to manipulate others. This is where the area of fraud detection starts to have greater importance within the foundations of our society.

A. Fraud Detection within Society

As technology has advanced, the various ways people can be manipulated has also advanced with it. Today fraud can be seen in multiple areas within society. It can be found in the financial field by the manipulation of one's banking accounts to obtain monetary gains. Also, one can see fraud within the security field through the use of identity theft. Fraud is even found in everyday email inboxes through the use of spam or phishing emails. However, through the use of fraud detection, one is able to minimize the number of people who are affected by the perpetrators of fraud.

B. Usefulness of Machine Learning

Machine learning techniques are a crucial part of the field of data science. The various techniques within machine learning help aid with problems found in multiple areas of study. Within the field of fraud detection, machine learning can pose a potential method of identifying fraudulent activity or perpetrators of fraud. This can drastically minimize the number of individuals negatively impacted by fraud [1], [2].

Jennifer Leach and Umashanger Thayasivam are with Department of Mathematics, Rowan University, Glassboro, NJ 08028 USA (e-mail: Leachj15@students.rowan.edu).

C. Summary of Analysis to Be Done

For this analysis, the use of various machine learning techniques were analyzed across two types of fraud: credit card fraud and cyber-attack fraud. These diverse types of fraudulent activity were then analyzed across five different machine learning techniques: Logistic Regression, Random Forest, Bagging, Support Vector Machine, and k-Nearest Neighbors [3]–[5].

D. Objectives of Research

The objective of this research is, first, to find which machine learning techniques optimize the accuracy, precision, recall, and F1 of the various fraud detection data; secondly, to find the optimal training and testing split that will give the most efficient model of detecting fraudulent activity.

E. Evaluation Metrics

In order to evaluate the results of each of the five machine learning techniques, the methods of accuracy, precision, recall, and F1 were used [6], [7]. Each of the four methods uses the following four types of predictions in order to calculate the appropriate results:

- True Positive (TP): Where the model predicted the outcome was fraudulent, and it was, in fact, fraudulent.
- False Negative (FN): Where the model predicted the outcome was not fraudulent, but it in fact was fraudulent.
- False Positive (FP): Where the model predicted the outcome was fraudulent, and it was not actually fraudulent.
- True Negative (TN): Where the model predicted the outcome was not fraudulent, and it was not actually fraudulent.

1) Accuracy: Accuracy is the most common evaluation method. It calculates the percentage of observations that were correctly predicted by using the formula:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

2) Precision: Precision calculates the amount of correct positive predicted outcomes compared to the total positive predicted outcomes. This is done through:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

3) Recall: Recall calculates the amount of correct positive predicted outcomes compared to the total accurate predicted outcomes. This is done through:

$$Precision = \frac{TP}{TP + FN} \tag{3}$$

4) F1: F1 calculates the harmonic mean of the precision and recall through:

$$Precision = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

II. Summary of Datasets

For this research, the use of machine learning techniques were analyzed across two different fraud detection datasets.

A. Cyber-Attack Dataset

The first dataset [8] looks to identify cyber-attacks that occurred in the province of Elaziğ in Turkey between 2015 and 2019. It consisted of 901 cases of cyber-attacks which were analyzed across 11 unique features. Those features are all categorical and consist of: Crime, Gender, Age, Income, Job, Marital Status, Education, Harm, Attack, Attack Method, and Perpetrator.

The 'Crime' attribute describes the crime done by the attacker and consists of the following three categories: misuse of debit/credit card, through informatics theft, and hacking into the information system and capturing data. The next attribute is 'Gender' and is categorized as either male or female. 'Age' describes the age of the attacker and consists of the following four categories: 27 years old and under, between 28 and 37 years old, between 38 and 50 years old, and 51 years old and older. Next, the attribute 'Income' is the income level of the cyber-attacker and is categorized as Low, Medium, or High. The attribute 'Job' describes the job the attacker had during the time of their attack and consists of the following nine categories: other, student, retired, justice and security, health sector manager, housewife, education, technical, and finance sector. 'Marital Status' is the marital status of the attacker, labeled as either single or married. 'Education' is the highest educational status of the attacker and is categorized as: primary education, high school, undergraduate, or graduate. The attribute 'Harm' is categorized into seven categories. 'Attack' is categorized into five categories. The feature 'Attack Method' can be used as either an attribute or a response variable and is categorized into five categories. The description of the categories for the three aforementioned variables can be found in the journal [8]. For the sake of this analysis, 'Attack Method' was used as one of the attributes. Finally, 'Perpetrator' can also be used as either an attribute or a response variable and describes whether the cyber-attacker was either Known or Unknown.

For this analysis, 'Perpetrator' was used as the response variable. Therefore, for this analysis of cyber-attacks, there was a total of 10 attributes across 901 different cyber-attack cases, with the intention to try to predict whether one can identify the perpetrator of the attack or not.

B. Credit Dataset

The second dataset is a R dataset under the CASdatasets [9]–[11] package. It consisted of 1,000 credit records which were analyzed across 21 unique features. Those features are a combination of categorical and numerical variables and consist of: Credit Status, Duration, Credit History, Purpose, Credit Amount, Savings, Employment, Installment Rate, Personal Status, Other Parties, Residence Since, Property Magnitude, Age, Other Payment Plans, Housing, Existing Credits, Job, Number of Dependents, Telephone, Foreign Worker, and Class.

The "Checking Status" attribute is a categorical variable that describes the status of the existing checking account, with the following categories: Less than 0, from 0 to 200, more than 200, or no running account/unknown account. "Duration" is a numerical variable explaining the credit duration in months. "Credit History" is a categorical variable that consists of the following categories: delay in paying off in the past, critical account, no credits taken or all credit paid back duly, existing credits paid back duly till now, all credits at this paid back duly. The attribute "Purpose" is a categorical variable describing the purpose of the credit. It has the following categories: new car, used car, item of furniture/equipment, radio/television, domestic household appliances, repairs, education, vacation, retraining, business, and others. "Credit Amount" is a numerical variable stating the credit amount in Deutsch marks. "Savings" is a categorical variable with the following categories: less than 100, from 100 to 500, from 500 to 1,000, more than 1,000, and no savings/unknown account. "Employment" is a categorical variable describing how long the person has been employed. It consists of the following: unemployed, less than 1 year, from 1 to 4 years, from 4 to 7 years, and more than 7 years. "Installment Rate" describes the person's installment rate in percentage of disposable income with the following categories: greater than 35, between 25 and 35, between 20 and 25, and less than 20. "Person Status" is a categorical variable explaining the person's marital status and sex. It consists of the following: male: divorced/separated, female: divorced/separated/married, male: single, male: married/widowed, and female: single. "Other Parties" describes any other debtors or guarantors with the following options: none, co-applicant, and guarantor. "Resident Since" is a categorical variable broken up by: less than 1 year, from 1 to 4 years, from 4 to 7 years, and more than 7 years. "Property Magnitude" describes the person's most valued property with the following categories: real estate, savings contract with building society/life insurance, car or other, and unknown/no property. "Age" is the age of the person in years. "Other Payment Plans" consists of: at other bank, at department store or mail order house, and no further running credits. "Housing" is the type of housing the person has from the following: rented flat, owner-occupied flat, and free apartment. "Existing credits" states the number of existing credits the person has at this bank. It consists of the following categories: one, two or three, four or five, and six or more. "Job" consists of the following: unemployed/unskilled with no permanent residence, unskilled with permanent residence, skilled worker/skilled employee/minor civil servant, and executive/self-employed/higher civil servant. "Number of Dependents" is a categorical variable explaining the number of dependents for which the person is liable to provide maintenance. It has the following categories: zero to two,

and three and more. “Telephone” is a categorical variable consisting of either “none” or “yes, registered under the customer’s name.” “Foreign Worker” is whether or not the person is a foreign worker. Finally, class is a binary variable where 0 represents good, and 1 represents bad.

For this analysis, “Class” was used as the response variable. So, this credit data analysis had a total of 20 attributes across 1,000 credit records. It looked to predict whether a specific credit report is good or bad.

C. Splitting of Credit Dataset

Since the credit dataset was a combination of numerical and categorical variables, it was then split into its corresponding numerical and categorical parts. This was to see how the different variable types affect the outcomes of the machine learning models.

1) Numerical Variables: For this analysis, only the numerical variables within the credit dataset were analyzed. Therefore, the dataset consisted of the following variables: ”Duration,” ”Credit Amount,” ”Installment Rate,” ”Residence Since,” ”Age,” ”Existing Credits,” ”Number of Dependents,” and the categorical response variable, ”Class.” So, the analysis had a total of 8 attributes across 1,000 credit records.

2) Categorical Variables: For this analysis, only the categorical variables within the credit dataset were explored. Therefore, the dataset consisted of the following categorical variables: ”Checking Status”, ”Credit History”, ”Purpose”, ”Savings”, ”Employment”, ”Personal Status”, ”Other Parties”, ”Property Magnitude”, ”Other Payment Plans”, ”Housing”, ”Job”, ”Telephone”, ”Foreign Worker” and the response variable, ”Class”. So, the analysis had a total of 14 attributes across 1,000 credit records.

III. Programming Software Used

The analysis of the aforementioned datasets were mainly done using R and RStudio to analyze the different variables across five different training and testing splits. JMP Pro was also used to help assist with some of the analysis.

IV. Results of Cyber-Attack Dataset

The Cyber-Attack dataset used five machine learning techniques to help predict whether or not one could detect the perpetrator of the fraud. The model was created using the following training splits of the original dataset: 70%, 75%, 80%, 85%, and 90%. Then, the model created from those splits was applied to the corresponding testing splits for each of the five machine learning techniques analyzed. Table I and Figs. 1-4 show the accuracy, precision, recall, and F1 obtained when applying those models.

TABLE I
Results of Cyber-Attack Data by Machine Learning Type

Cyber-Attack Results	Machine Learning Techniques				
Training/ Testing Split	*Logistic Regression*	*Random Forest*	*Bagging*	*SVM*	*KNN*
Accuracy %					
70% / 30%	64.8%	62.6%	63.0%	68.9%	**86.7%**
75% / 25%	63.3%	65.0%	61.9%	67.7%	**85.4%**
80% / 20%	62.2%	58.9%	58.9%	62.8%	**81.1%**
85% / 15%	65.9%	63.0%	58.5%	72.6%	**88.9%**
90% / 10%	61.1%	63.3%	61.1%	67.8%	**87.8%**
Precision %					
70% / 30%	60.8%	57.3%	57.5%	70.4%	**85.8%**
75% / 25%	58.6%	61.2%	57.1%	69.2%	**86.5%**
80% / 20%	59.6%	52.7%	52.6%	59.3%	**78.2%**
85% / 15%	62.7%	59.6%	52.6%	72.6%	**90.7%**
90% / 10%	56.7%	57.9%	55.0%	69.2%	**85.0%**
Recall %					
70% / 30%	53.0%	53.8%	55.6%	48.7%	**82.9%**
75% / 25%	52.0%	53.1%	49.0%	45.9%	**78.6%**
80% / 20%	39.7%	50.0%	51.3%	44.9%	**78.2%**
85% / 15%	54.2%	47.5%	50.8%	54.2%	**83.1%**
90% / 10%	43.6%	56.4%	56.4%	46.2%	**87.2%**
F1 %					
70% / 30%	56.6%	55.5%	56.5%	57.6%	**84.3%**
75% / 25%	55.1%	56.8%	52.7%	55.2%	**82.4%**
80% / 20%	47.7%	51.3%	51.9%	51.1%	**78.2%**
85% / 15%	58.2%	52.8%	51.7%	63.4%	**86.7%**
90% / 10%	49.3%	57.1%	55.7%	55.4%	**86.1%**

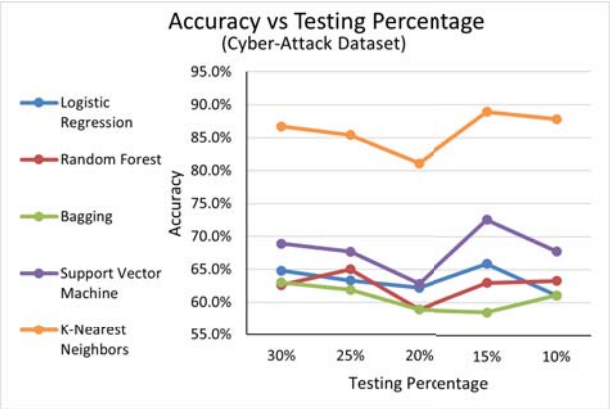

Fig. 1 Accuracy of Each Machine Learning Across Five Testing Splits for Cyber-Attack Dataset

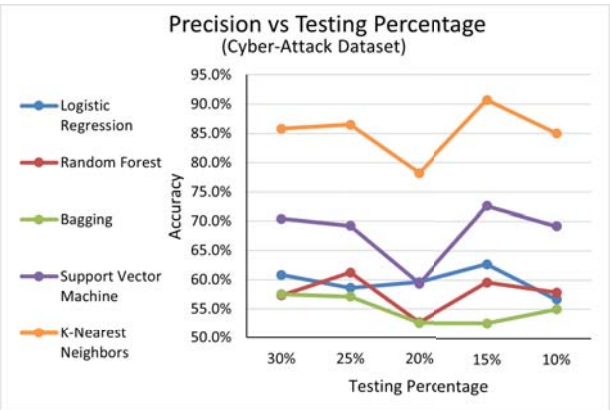

Fig. 2 Precision of Each Machine Learning Across Five Testing Splits for Cyber-Attack Dataset

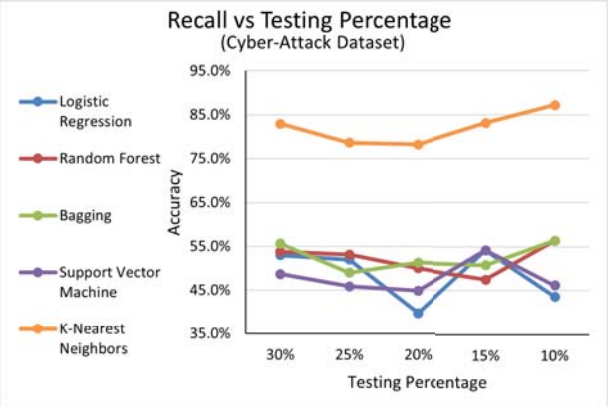

Fig. 3 Recall of Each Machine Learning Across Five Testing Splits for Cyber-Attack Dataset

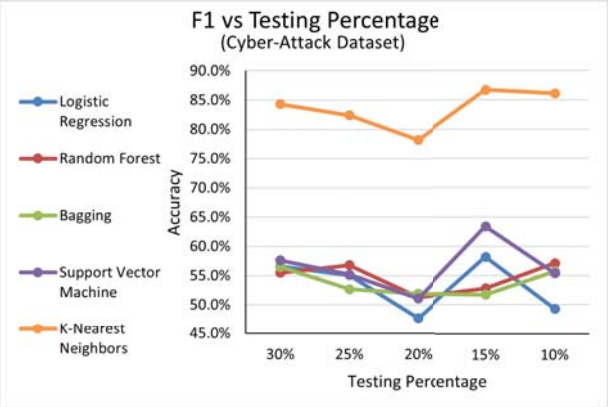

Fig. 4 F1 of Each Machine Learning Across Five Testing Splits for Cyber-Attack Dataset

For the Cyber-Attack dataset, k-Nearest Neighbors produced the greatest accuracy, precision, recall, and F1 across all training and testing splits. When analyzing the specific training and testing splits done with KNN, overall, the split of 85% training and 15% testing produced the optimal results.

V. Comparing Cyber-Attack Data

Since the Cyber-Attack Dataset was obtained from an academic journal [8], the results from this analysis were compared to the results within that journal. Both analyses have the following machine learning techniques in common: Logistic Regression, Random Forest, and KNN. Also, the academic journal where the Cyber-Attack dataset came from only performed the aforementioned machine learning techniques at an 80% Training 20% Testing split. Table II and Figs. 5-8 show the accuracy, precision, recall, and F1 obtained from both analyses at the 80% Training/20% Testing split, where "Journal" represents the original journal from which the Cyber-Attack dataset was obtained.

TABLE II
Comparing Cyber-Attack Data by Machine Learning Type

Cyber-Attack Results	Machine Learning Techniques					
Measurement Type	*Logistic Regression*	*Logistic Regression (Journal)*	*Random Forest*	*Random Forest (Journal)*	*KNN*	*KNN (Journal)*
Accuracy %	62.2%	65.40%	58.9%	63.54%	*81.1%*	64.57%
Precision %	59.6%	60.67%	52.7%	63.91%	*78.2%*	56.85%
Recall %	39.7%	60.22%	50.0%	63.54%	*78.2%*	56.91%
F1 %	47.7%	59.14%	51.3%	62.92%	*78.2%*	56.85%

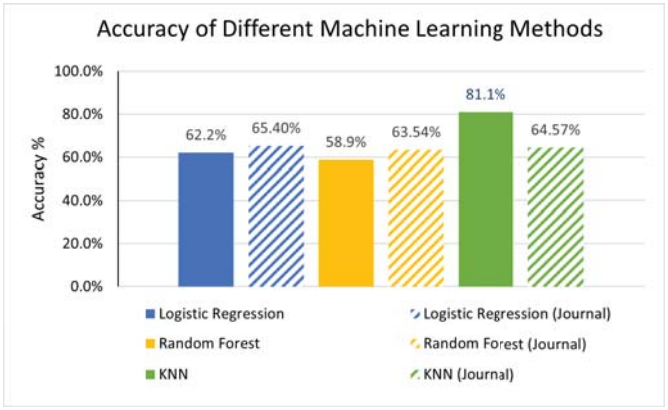

Fig. 5 Comparing Accuracy of Each Machine Learning Technique Across Two Analyses

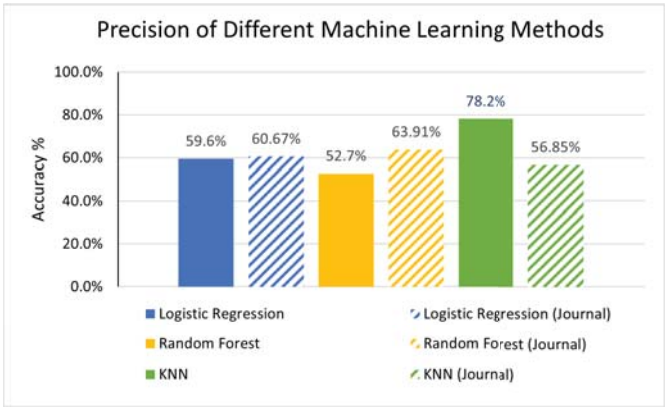

Fig. 6 Comparing Precision of Each Machine Learning Technique Across Two Analyses

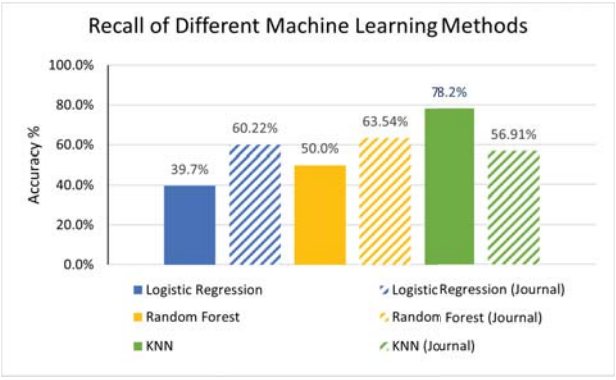

Fig. 7 Comparing Recall of Each Machine Learning Technique Across Two Analyses

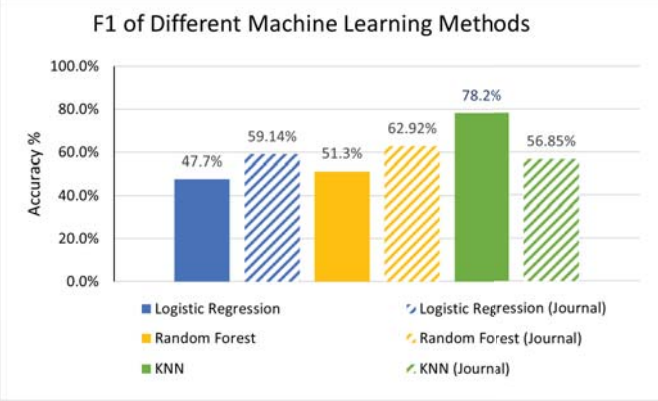

Fig. 8 Comparing F1 of Each Machine Learning Technique Across Two Analyses

Logistic Regression and Random Forest produced similar results as the original journal for all four types of measurements. The only discrepancy between the original journal and this analysis was in K-Nearest Neighbors. The K-Nearest Neighbor in this analysis was higher than the ones obtained in the original journal. The reason for this difference could be due to the different programming softwares used between the two analyses or due to the selected k-value used for the analysis. The original journal used Python to run their analyses, while this research used R. The way the different programs perform the K-Nearest Neighbor analysis may be the reason for the discrepancies. Also, this research selected the k-value for the analysis by sequentially picking values for k until they found a value that produced the highest accuracy. The process/value of k used in the original journal is unknown. This may also explain the discrepancy between the two results.

VI. Results from Credit Dataset (Full Dataset)

The credit dataset used five different machine learning techniques in order to help predict whether or not a specific credit report was good or bad. The model was created using the following training splits of the original dataset: 70%, 75%, 80%, 85%, and 90%. Then, the model created from those splits was applied to the corresponding testing splits for each of the five machine learning techniques analyzed. Table III and Figs. 9-12 show the accuracy, precision, recall, and F1 obtained when applying those models.

For the Full Credit dataset, in general, Logistic Regression, Support Vector Machines, and Random Forest produced the highest accuracy, precision, recall, and F1 across all training and testing splits. Support Vector Machines produced the highest or about equal to the other aforementioned machine learning techniques. In conclusion, Support Vector Machine performs best when using both categorical and numerical variables to predict whether a credit report is good or bad. Across all four tests of measurements, the optimal training and test was 85% Training and 15% Testing.

TABLE III
Results of Full Credit Data by Machine Learning Type

Credit Data Results (Full Dataset)	Machine Learning Techniques				
Training/ Testing Split	*Logistic Regression*	*Random Forest*	*Bagging*	*SVM*	*KNN*
		Accuracy %			
70% / 30%	73.3%	**75.0%**	72.3%	72.0%	68.3%
75% / 25%	78.0%	78.0%	75.2%	**79.2%**	66.4%
80% / 20%	75.0%	**76.5%**	71.5%	75.0%	69.5%
85% / 15%	**80.7%**	78.7%	78.0%	**80.7%**	71.3%
90% / 10%	76.0%	**78.0%**	73.0%	74.0%	71.0%
		Precision %			
70% / 30%	**79.3%**	78.2%	78.5%	77.6%	72.5%
75% / 25%	**81.6%**	79.4%	79.6%	80.6%	71.6%
80% / 20%	**78.5%**	78.2%	77.1%	76.8%	73.4%
85% / 15%	**82.2%**	78.3%	80.5%	81.1%	73.5%
90% / 10%	80.3%	79.3%	**81.2%**	76.8%	73.0%
		Recall %			
70% / 30%	83.8%	**89.0%**	83.3%	84.3%	88.1%
75% / 25%	88.6%	**92.6%**	86.9%	**92.6%**	86.3%
80% / 20%	88.6%	**92.1%**	84.3%	**92.1%**	88.6%
85% / 15%	92.4%	**96.2%**	90.5%	94.3%	92.4%
90% / 10%	87.1%	**92.9%**	80.0%	90.0%	**92.9%**
		F1 %			
70% / 30%	81.5%	**83.3%**	80.8%	80.8%	79.6%
75% / 25%	84.9%	85.5%	83.1%	**86.2%**	78.2%
80% / 20%	83.2%	**84.6%**	80.5%	83.8%	80.3%
85% / 15%	87.0%	86.3%	85.2%	**87.2%**	81.9%
90% / 10%	83.6%	**85.5%**	80.6%	82.9%	81.8%

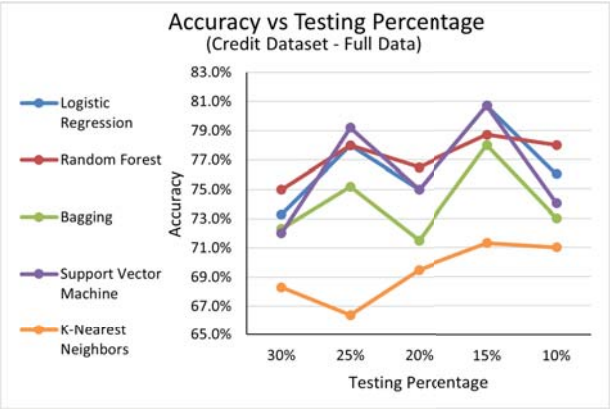

Fig. 9 Accuracy of Each Machine Learning Across Five Testing Splits for Full Credit Dataset

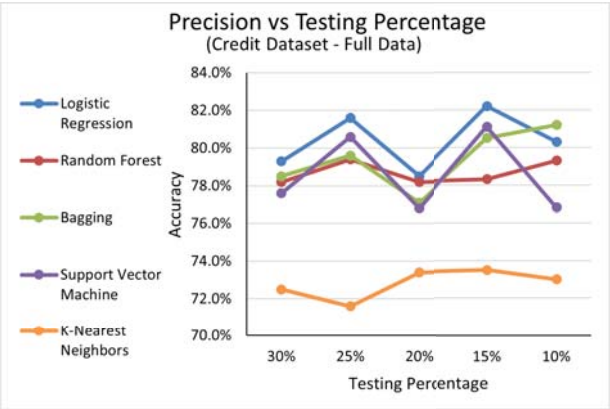

Fig. 10 Precision of Each Machine Learning Across Five Testing Splits for Full Credit Dataset

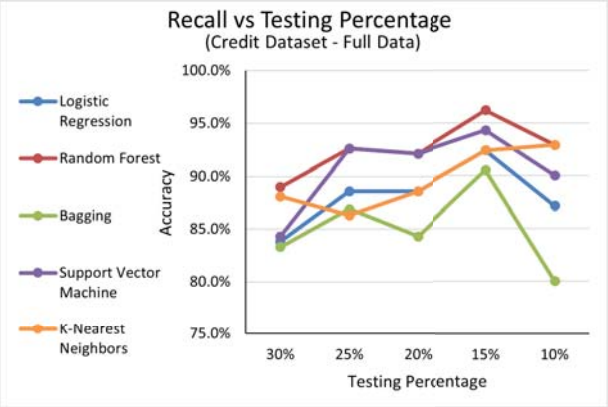

Fig. 11 Recall of Each Machine Learning Across Five Testing Splits for Full Credit Dataset

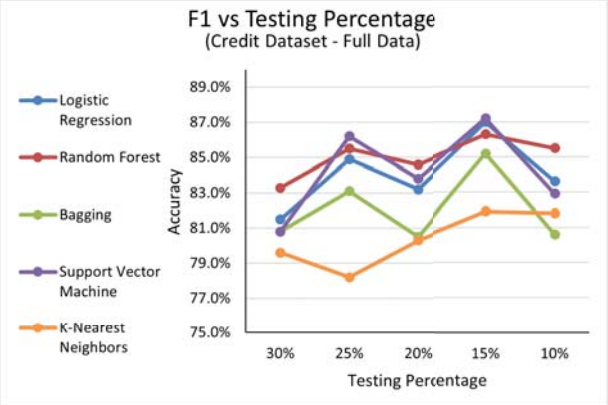

Fig. 12 F1 of Each Machine Learning Across Five Testing Splits for Full Credit Dataset

VII. Results From Credit Dataset (Only Numerical Variables)

After running the analyses on the Full Credit model, the credit model was then split into only numerical values and only categorical variables in order to predict whether a specific credit report was good or bad. Therefore, the model was created using the same training splits on the dataset for only numerical variables: 70%, 75%, 80%, 85%, and 90%. Then, the model created from those splits was applied to the corresponding testing splits for the five machine learning techniques analyzed. Table IV and Figs. 13-16 show the accuracy, precision, recall, and F1 obtained when applying those models to only the numerical variables.

TABLE IV
Results of Credit Data by Machine Learning Type
(Numerical Variable Only)

Credit Data Results (Numerical Only)	Machine Learning Techniques				
Training/ Testing Split	*Logistic Regression*	*Random Forest*	*Bagging*	*SVM*	*KNN*
		Accuracy %			
70% / 30%	69.3%	**70.7%**	65.0%	70.0%	68.3%
75% / 25%	**71.2%**	**71.2%**	65.6%	70.8%	66.4%
80% / 20%	69.5%	**72.0%**	67.0%	70.5%	69.0%
85% / 15%	69.3%	68.7%	68.0%	70.0%	**72.7%**
90% / 10%	72.0%	73.0%	**74.0%**	71.0%	72.0%
		Precision %			
70% / 30%	71.9%	**74.6%**	73.8%	70.5%	72.5%
75% / 25%	72.5%	**73.5%**	71.9%	70.7%	71.6%
80% / 20%	71.1%	**74.1%**	73.7%	70.8%	73.2%
85% / 15%	71.2%	72.3%	**74.8%**	72.1%	73.9%
90% / 10%	71.9%	74.2%	**76.8%**	70.7%	73.3%
		Recall %			
70% / 30%	92.4%	88.1%	77.6%	**98.1%**	88.1%
75% / 25%	94.9%	92.0%	83.4%	**99.4%**	86.3%
80% / 20%	95.0%	92.1%	82.1%	**98.6%**	87.9%
85% / 15%	**94.3%**	89.5%	81.9%	93.3%	**94.3%**
90% / 10%	98.6%	94.3%	90.0%	100.0%	94.3%
		F1 %			
70% / 30%	**80.8%**	**80.8%**	75.6%	80.7%	79.6%
75% / 25%	**82.2%**	81.7%	77.2%	78.9%	78.2%
80% / 20%	81.3%	82.2%	77.7%	**86.3%**	79.9%
85% / 15%	81.1%	80.0%	78.2%	79.0%	**82.8%**
90% / 10%	**83.1%**	83.0%	82.9%	79.2%	82.5%

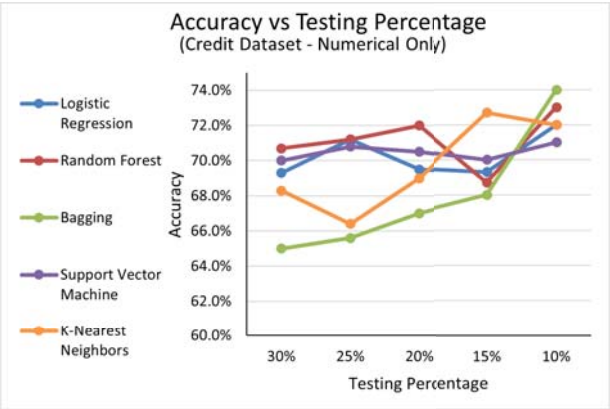

Fig. 13 Accuracy of Each Machine Learning Across Five Testing Splits for Numerical Values in Credit Dataset

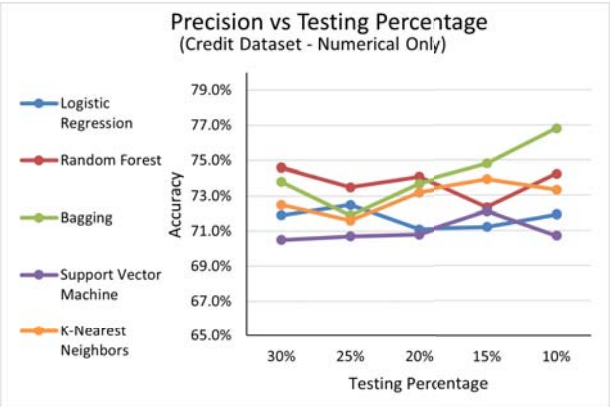

Fig. 14 Precision of Each Machine Learning Across Five Testing Splits for Numerical Values in Credit Dataset

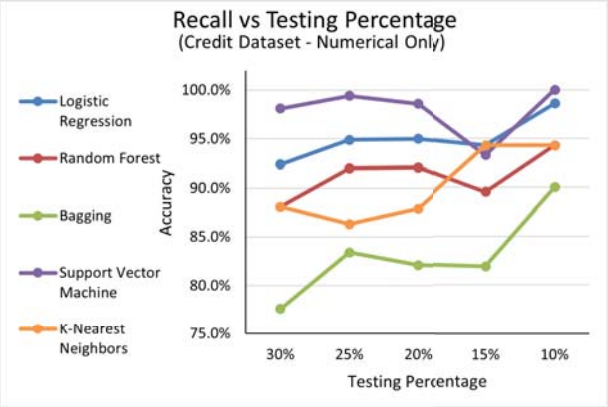

Fig. 15 Recall of Each Machine Learning Across Five Testing Splits for Numerical Values in Credit Dataset

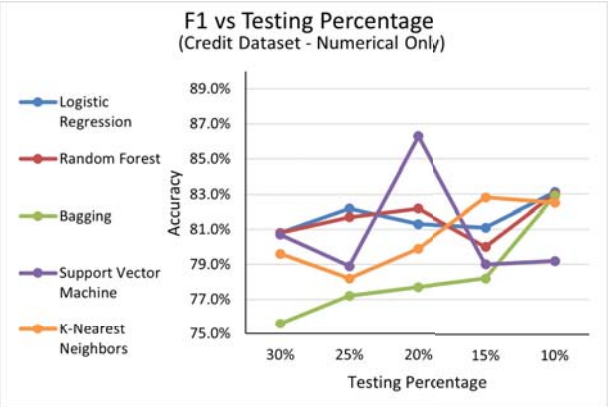

Fig. 16 F1 of Each Machine Learning Across Five Testing Splits for Numerical Values in Credit Dataset

For the Credit dataset for only the numerical variables, Bagging and Random Forest provided the largest accuracy and precision at the 90% Training 10% Testing split. Meanwhile, Logistic Regression and Support Vector Machine produced the highest Recall and F1 across four of the five training and testing splits but with moderate accuracy and precision. Therefore, when analyzing the numerical variables within the Credit dataset, Boosting and Random Forest performed the best when predicting whether a credit report is good or bad, with an optimal training test split of 90% Training 10% Testing.

VIII. Results From Credit Dataset (Only Categorical Variables)

After running the analyses on the Full Credit model, the credit model was then split into only numerical values and only categorical variables in order to predict whether a specific credit report was good or bad. Therefore, the model was created using the same training splits on the dataset for only categorical variables: 70%, 75%, 80%, 85%, and 90%. Then, the model created from those splits was applied to the corresponding testing splits for each of the five machine learning techniques analyzed. Table V and Figs. 17-20 show the accuracy, precision, recall, and F1 we obtained when applying those models to only the categorical variables.

TABLE V
Results of Credit Data by Machine Learning Type
(Categorical Variables Only)

Credit Data Results (Categorical Only)	Machine Learning Techniques				
Training/ Testing Split	*Logistic Regression*	*Random Forest*	*Bagging*	*SVM*	*KNN*
		Accuracy %			
70% / 30%	72.0%	72.7%	70.7%	72.3%	**80.7%**
75% / 25%	74.4%	75.6%	71.6%	74.4%	**83.6%**
80% / 20%	74.5%	75.0%	74.0%	74.0%	**84.5%**
85% / 15%	78.0%	77.3%	78.7%	77.3%	**86.0%**
90% / 10%	72.0%	71.0%	68.0%	70.0%	**84.0%**
		Precision %			
70% / 30%	77.9%	77.8%	78.2%	76.8%	**83.0%**
75% / 25%	78.5%	77.1%	77.1%	77.6%	**85.6%**
80% / 20%	76.6%	77.1%	78.9%	74.4%	**85.2%**
85% / 15%	78.1%	78.9%	81.2%	77.5%	**86.8%**
90% / 10%	76.3%	75.3%	75.0%	75.0%	**85.5%**
		Recall %			
70% / 30%	83.8%	85.2%	80.5%	86.7%	**91.0%**
75% / 25%	87.4%	**92.6%**	84.6%	89.1%	92.0%
80% / 20%	91.4%	91.4%	85.7%	**95.7%**	94.3%
85% / 15%	**95.2%**	92.4%	90.5%	**95.2%**	94.3%
90% / 10%	87.1%	87.1%	81.4%	85.7%	**92.9%**
		F1 %			
70% / 30%	80.7%	81.4%	79.3%	81.4%	**86.8%**
75% / 25%	82.7%	84.2%	80.7%	83.0%	**88.7%**
80% / 20%	83.4%	83.7%	82.2%	83.8%	**89.5%**
85% / 15%	85.8%	85.1%	85.6%	85.5%	**90.4%**
90% / 10%	81.3%	80.8%	78.1%	80.0%	**89.0%**

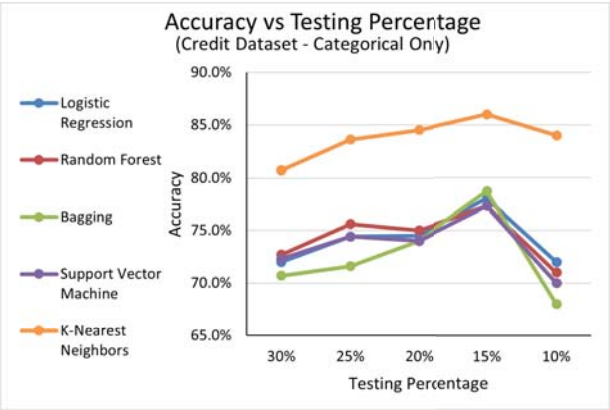

Fig. 17 Accuracy of Each Machine Learning Across Five Testing Splits for Categorical Variables in Credit Dataset

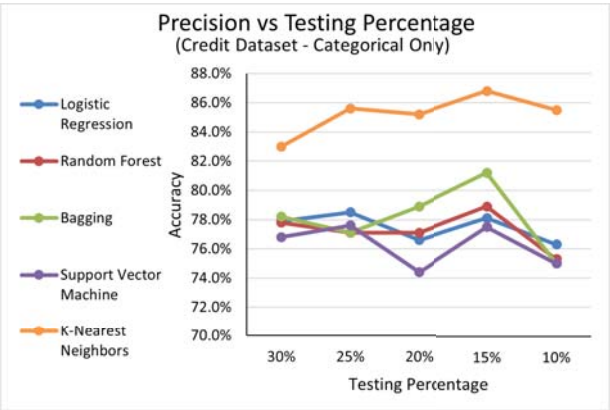

Fig. 18 Precision of Each Machine Learning Across Five Testing Splits for Categorical Variables in Credit Dataset

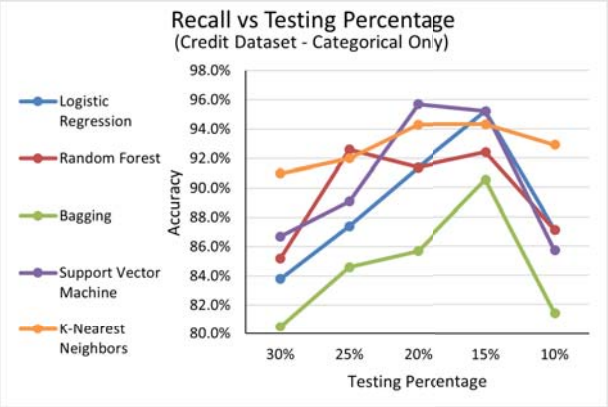

Fig. 19 Recall of Each Machine Learning Across Five Testing Splits for Categorical Variables in Credit Dataset

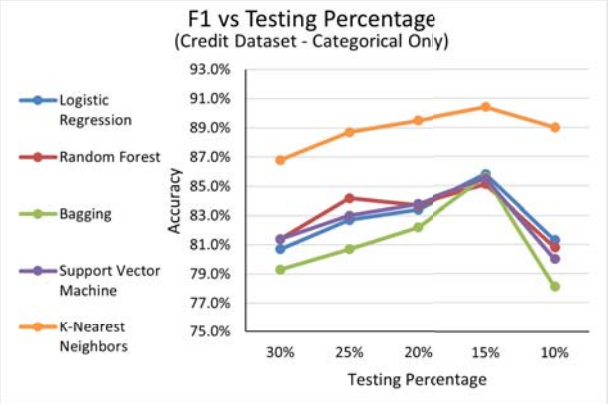

Fig. 20 F1 of Each Machine Learning Across Five Testing Splits for Categorical Variables in Credit Dataset

For the Credit dataset for only the categorical variables, K-Nearest Neighbors produced the highest accuracy, precision, and F1 across all training and testing splits compared to the other four machine learning models. As for recall, the machine learning algorithms that produced the highest results were K-Nearest Neighbors, Support Vector Machine, and Logistic Regression. Within those, K-Nearest Neighbors produced higher or approximately equal results to the other two aforementioned machine learning techniques across all training and testing splits. Therefore, for analyzing only categorical variables, K-Nearest Neighbors performed the best for predicting whether a credit report is good or bad, with an optimal training and testing split being 85% Training and 15% Testing.

IX. Conclusions and Future Research

Across all data sets, K-Nearest Neighbors, Support Vector Machines, Bagging, and Random Forest were the optimal models for predicting fraudulent activity. K-Nearest Neighbor performed the best when used for analyzing fully categorical data. Bagging or Random Forest performed the best when used for analyzing entirely numerical data. Finally, Support Vector Machine performed the best when used for analyzing data with a combination of numerical and categorical variables. Also, in general, the optimal training and testing split was 85% Training and 15% Testing across the various data sets. The only exception to this was when analyzing only numerical values, then the optimal training and testing split was 90% Training and 10% Testing.

Future research direction of this study is to further analyze the various datasets, including techniques such as deep learning and other ensemble learning. Also, to perform the analysis and optimize performance evaluation analysis between different software programs, such as Python.

References

[1] James, Gareth, et al. An Introduction to Statistical Learning with Applications in R. Springer, 2021.
[2] William Ezekiel and Umashanger Thayasivam. "A Comparison of Supervised Learning Techniques for Clustering" Neural Information Processing Vol. 9489 (2015) p. 476 - 483
[3] Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, Gianluca Bontempi, Combining unsupervised and supervised learning in credit card fraud detection, Information Sciences, Volume 557, 2021, Pages 317-331, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.05.042. (https://www.sciencedirect.com/science/article/pii/S0020025519304451)
[4] Bilen, A., & Ahmet Bedri Özer. (2021). Cyber-attack method and perpetrator prediction using machine learning algorithms. PeerJ Computer Science, doi:http://dx.doi.org/10.7717/peerj-cs.475
[5] Siddhant Bagga, Anish Goyal, Namita Gupta, Arvind Goyal, Credit Card Fraud Detection using Pipeling and Ensemble Learning, Procedia Computer Science, Volume 173, 2020, Pages 104-112, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.06.014. (https://www.sciencedirect.com/science/article/pii/S1877050920315167)
[6] Serhiy Hnatyshyn, Umashanger Thayasivam, Vasil Hnatyshin and Curtis White. "Machine learning algorithms for metabolomics applications" LondonIdentification and Data Processing Methods in Metabolomics (2015) p. 96 - 110 Available at: http://works.bepress.com/umashanger-thayasivam/12/
[7] Hajjami, S. , Malki, J. , Bouju, A. , Berrada, M.. "Machine Learning Facing Behavioral Noise Problem in an Imbalanced Data Using One Side Behavioral Noise Reduction: Application to a Fraud Detection". World Academy of Science, Engineering and Technology, Open Science Index 171, International Journal of Computer and Information Engineering (2021), 15(3), 194 - 205.
[8] Bilen, Abdulkadir and Ahmet Bedri Özer. 2021. "Cyber-Attack Method and Perpetrator Prediction using Machine Learning Algorithms." PeerJ Computer Science (Apr 09). doi:http://dx.doi.org.ezproxy.rowan.edu/10.7717/peerj-cs.475. http://ezproxy.rowan.edu/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fscholarly-journals%2Fcyber-attack-method-perpetrator-prediction-using%2Fdocview%2F2510490837%2Fse-2%3Faccountid%3D13605.
[9] Fahrmeir, L. and Tutz, G. (1994), Multivariate Statistical Modelling Based on Generalized Linear Models, Springer.
[10] Nisbet, R., Elder, J. and Miner, G. (2011), Handbook of Statistical Analysis and Data Mining Applications, Academic Press.
[11] Tuff'ery, S. (2011), Data Mining and Statistics for Decision Making, Wiley.